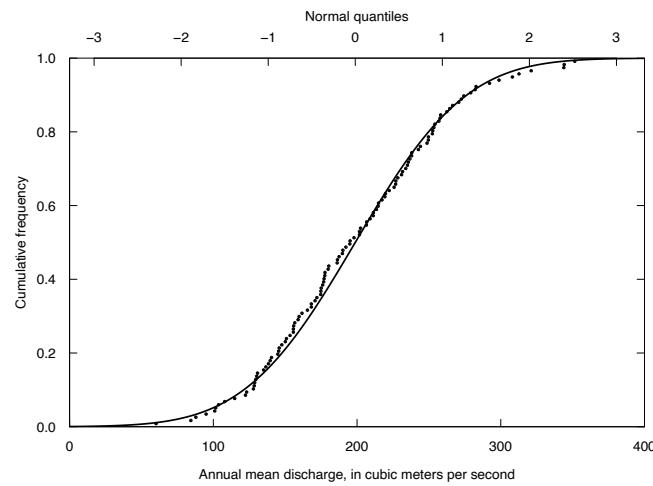
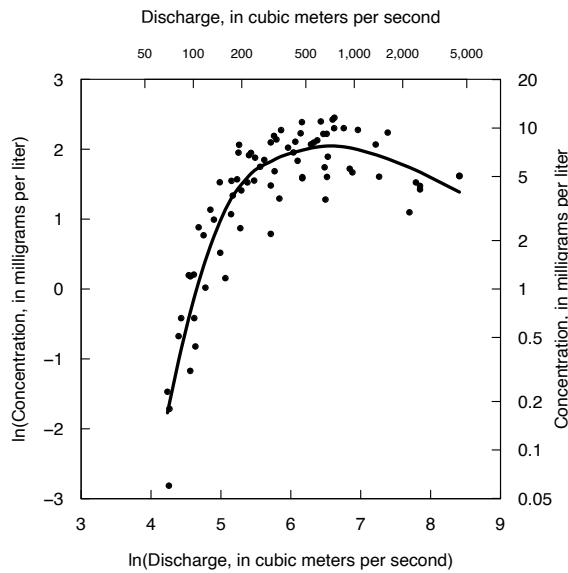


# Statistical Methods in Water Resources

Chapter 3 of  
 Section A, Statistical Analysis  
**Book 4, Hydrologic Analysis and Interpretation**



**Techniques and Methods 4–A3**  
 Supersedes USGS Techniques of Water-Resources Investigations, book 4,  
 chapter A3

**Cover:** Top Left: A loess smooth curve of dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September.

Top Right: U.S. Geological Survey scientists Frank Engel (left) and Aaron Walsh (right) sample suspended sediment at the North Fork Toutle River near Kid Valley, Washington, upstream of a U.S. Army Corps of Engineers sediment retention structure. This gage is operated in cooperation with the U.S. Army Corps of Engineers. Photograph by Molly Wood, U.S. Geological Survey, March 7, 2018.

Bottom Left: Jackson Lake Dam and USGS streamgage 13011000, Snake River near Moran, Wyoming, within Grand Teton National Park. Photograph by Kenneth M. Nolan, U.S. Geological Survey.

Bottom Right: Overlay of James River annual mean discharge (1900–2015) and standard normal distribution quantile plot.

# **Statistical Methods in Water Resources**

By Dennis R. Helsel, Robert M. Hirsch, Karen R. Ryberg, Stacey A. Archfield, and Edward J. Gilroy

Chapter 3 of  
Section A, Statistical Analysis  
**Book 4, Hydrologic Analysis and Interpretation**

Techniques and Methods 4–A3  
Supersedes USGS Techniques of Water-Resources Investigations, book 4,  
chapter A3

**U.S. Department of the Interior**  
DAVID BERNHARDT, Secretary

**U.S. Geological Survey**  
James F. Reilly II, Director

**U.S. Geological Survey, Reston, Virginia: 2020**

First release: 1992 by Elsevier, in print

Revised: September 2002 by the USGS, online as Techniques of Water-Resources Investigations (TWRI), book 4, chapter A3, version 1.1

Revised: May 2020, by the USGS, online and in print, as Techniques and Methods, book 4, chapter A3

Supersedes USGS Techniques of Water-Resources Investigations (TWRI), book 4, chapter A3, version 1.1

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit <https://www.usgs.gov> or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit <https://store.usgs.gov>.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Suggested citation:

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chap. A3, 458 p., <https://doi.org/10.3133/tm4a3>. [Supersedes USGS Techniques of Water-Resources Investigations, book 4, chap. A3, version 1.1.]

Associated data for this publication:

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources—Supporting materials: U.S. Geological Survey data release, <https://doi.org/10.5066/P9JWL6XR>.

ISSN 2328-7047 (print)

ISSN 2328-7055 (online)

## Contents

<b>Chapter 1 Summarizing Univariate Data.....</b>	<b>1</b>
1.1 Characteristics of Water Resources Data .....	2
1.2 Measures of Central Tendency.....	5
1.2.1 A Classical Measure of Central Tendency—The Arithmetic Mean.....	5
1.2.2 A Resistant Measure of Central Tendency—The Median.....	6
1.2.3 Other Measures of Central Tendency .....	7
1.3 Measures of Variability.....	8
1.3.1 Classical Measures of Variability .....	8
1.3.2 Resistant Measures of Variability.....	9
1.3.3 The Coefficient of Variation—A Nondimensional Measure of Variability .....	10
1.4 Measures of Distribution Symmetry.....	11
1.4.1 A Classical Measure of Symmetry—The Coefficient of Skewness.....	11
1.4.2 A Resistant Measure of Symmetry—The Quartile Skew .....	12
1.5 Other Resistant Measures of Symmetry.....	12
1.6 Outliers .....	12
1.7 Transformations .....	13
1.7.1 The Ladder of Powers.....	14
<b>Chapter 2 Graphical Data Analysis.....</b>	<b>17</b>
2.1 Graphical Analysis of Single Datasets.....	18
2.1.1 Histograms.....	18
2.1.2 Quantile Plots .....	20
2.1.3 Boxplots.....	22
2.1.4 Probability Plots .....	26
2.1.5 Q-Q plots as Exceedance Probability Plots.....	28
2.1.6 Deviations from a Linear Pattern on a Probability Plot .....	28
2.1.7 Probability Plots for Comparing Among Distributions.....	30
2.2 Graphical Comparisons of Two or More Datasets.....	30
2.2.1 Histograms.....	32
2.2.2 Dot-and-line Plots of Means and Standard Deviations .....	33
2.2.3 Side-by-side Boxplots.....	34
2.2.4 Q-Q Plots of Multiple Groups of Data .....	36
2.3 Scatterplots and Enhancements.....	38
2.3.1 Evaluating Linearity.....	39
2.3.2 Evaluating Differences in Central Tendency on a Scatterplot.....	42
2.3.3 Evaluating Differences in Spread .....	42
2.4 Graphs for Multivariate Data .....	45
2.4.1 Parallel Plots.....	46
2.4.2 Star Plots.....	46
2.4.3 Trilinear and Piper Diagrams .....	48
2.4.4 Scatterplot Matrix.....	50
2.4.5 Biplots of Principal Components.....	51

2.4.6 Nonmetric Multidimensional Scaling.....	52
2.4.7 Three-dimensional Rotation Plots.....	53
2.4.8 Methods to Avoid.....	55
<b>Chapter 3 Describing Uncertainty.....</b>	<b>57</b>
3.1 Definition of Interval Estimates .....	57
3.2 Interpretation of Interval Estimates.....	58
3.3 Confidence Intervals for the Median .....	61
3.3.1 Nonparametric Interval Estimate for the Median .....	61
3.3.2 Parametric Interval Estimate for the Median .....	65
3.4 Confidence Intervals for the Mean.....	67
3.4.1 Symmetric Confidence Interval for the Mean .....	67
3.4.2 Asymmetric Confidence Interval for the Mean .....	68
3.4.3 Bootstrap Confidence Interval for the Mean for Cases with Small Sample Sizes or Highly Skewed Data.....	68
3.5 Nonparametric Prediction Intervals.....	70
3.5.1 Two-sided Nonparametric Prediction Interval.....	70
3.5.2 One-sided Nonparametric Prediction Interval .....	72
3.6 Parametric Prediction Intervals .....	73
3.6.1 Symmetric Prediction Interval.....	73
3.6.2 Asymmetric Prediction Intervals .....	74
3.7 Confidence Intervals for Quantiles and Tolerance Limits.....	75
3.7.1 Confidence Intervals for Percentiles Versus Tolerance Intervals.....	76
3.7.2 Two-sided Confidence Intervals for Percentiles.....	77
3.7.3 Lower One-sided Tolerance Limits .....	81
3.7.4 Upper One-sided Tolerance Limits .....	84
3.8 Other Uses for Confidence Intervals.....	88
3.8.1 Implications of Non-normality for Detection of Outliers.....	88
3.8.2 Implications of Non-normality for Quality Control .....	90
3.8.3 Implications of Non-normality for Sampling Design.....	91
<b>Chapter 4 Hypothesis Tests .....</b>	<b>93</b>
4.1 Classification of Hypothesis Tests .....	93
4.1.1 Classification Based on Measurement Scales .....	93
4.1.2 Divisions Based on the Method of Computing a <i>p</i> -value.....	95
4.2 Structure of Hypothesis Tests .....	97
4.2.1 Choose the Appropriate Test.....	97
4.2.2 Establish the Null and Alternate Hypotheses .....	100
4.2.3 Decide on an Acceptable Type I Error Rate, $\alpha$ .....	101
4.2.4 Compute the Test Statistic and the <i>p</i> -value .....	102
4.2.5 Make the Decision to Reject $H_0$ or Not .....	103
4.3 The Rank-sum Test as an Example of Hypothesis Testing.....	103
4.4 Tests for Normality .....	107
4.5 Other Hypothesis Tests .....	111

<b>4.6 Considerations and Criticisms About Hypothesis Tests .....</b>	<b>111</b>
4.6.1 Considerations .....	111
4.6.2 Criticisms.....	112
4.6.3 Discussion.....	112
<b>Chapter 5 Testing Differences Between Two Independent Groups.....</b>	<b>117</b>
5.1 The Rank-sum Test .....	118
5.1.1 Null and Alternate Hypotheses for the Rank-sum Test.....	118
5.1.2 Assumptions of the Rank-sum Test .....	118
5.1.3 Computation of the Rank-sum Test.....	119
5.1.4 The Large-sample Approximation to the Rank-sum Test.....	122
5.2 The Permutation Test of Difference in Means.....	123
5.2.1 Assumptions of the Permutation Test of Difference in Means .....	124
5.2.3 Computation of the Permutation Test of Difference in Means.....	124
5.3 The <i>t</i> -test.....	126
5.3.1 Assumptions of the <i>t</i> -test .....	126
5.3.2 Computation of the Two-sample <i>t</i> -test Assuming Equal Variances.....	127
5.3.3 Adjustment of the <i>t</i> -test for Unequal Variances.....	127
5.3.4 The <i>t</i> -test After Transformation Using Logarithms.....	129
5.3.5 Conclusions as Illustrated by the Precipitation Nitrogen Example.....	130
5.4 Estimating the Magnitude of Differences Between Two Groups.....	130
5.4.1 The Hodges-Lehmann Estimator of Difference in Medians .....	131
5.4.2 Confidence Interval for the Hodges-Lehmann Estimator, $\hat{\Delta}$ .....	132
5.4.3 Estimate of Difference Between Group Means .....	133
5.4.4 Parametric Confidence Interval for Difference in Group Means .....	134
5.4.5 Bootstrap Confidence Interval for Difference in Group Means .....	134
5.4.6 Graphical Presentation of Results .....	135
5.4.7 Side-by-side Boxplots.....	135
5.4.8 Q-Q Plots .....	135
5.5 Two-group Tests for Data with Nondetects .....	137
5.6 Tests for Differences in Variance Between Groups .....	138
5.6.1 Fligner-Killeen Test for Equal Variance .....	140
5.6.2 Levene's Test for Equal Variance .....	141
<b>Chapter 6 Paired Difference Tests of the Center .....</b>	<b>145</b>
6.1 The Sign Test.....	146
6.1.1 Null and Alternative Hypotheses .....	146
6.1.2 Computation of the Exact Sign Test.....	147
6.1.3 The Large-sample Approximation to the Sign Test.....	149
6.2 The Signed-rank Test.....	150
6.2.1 Null and Alternative Hypotheses for the Signed-rank Test.....	150
6.2.2 Computation of the Exact Signed-rank Test.....	151
6.2.3 The Large-sample Approximation for the Signed-rank Test .....	152
6.2.4 Permutation Version of the Signed-rank Test .....	153
6.2.5 Assumption of Symmetry for the Signed-rank Test .....	154

6.3 The Paired <i>t</i> -test .....	155
6.3.1 Null and Alternate Hypotheses .....	155
6.3.2 Computation of the Paired <i>t</i> -test.....	156
6.3.3 Permutation Test for Paired Differences .....	157
6.3.4 The Assumption of Normality for the Paired <i>t</i> -test.....	158
6.4 Graphical Presentation of Results .....	159
6.4.1 Boxplots.....	159
6.4.2 Scatterplots with a One-to-one Line .....	159
6.5 Estimating the Magnitude of Differences Between Two Groups .....	161
6.5.1 The Median Difference.....	161
6.5.2 The Hodges-Lehmann Estimator.....	161
6.5.3 Mean Difference.....	163
<b>Chapter 7 Comparing Centers of Several Independent Groups .....</b>	<b>165</b>
7.1 The Kruskal-Wallis Test.....	167
7.1.1 Null and Alternate Hypotheses for the Kruskal-Wallis Test.....	167
7.1.2 Assumptions of the Kruskal-Wallis Test .....	167
7.1.3 Computation of the Exact Kruskal-Wallis Test.....	168
7.1.4 The Large-sample Approximation for the Kruskal-Wallis Test.....	169
7.2 Analysis of Variance .....	171
7.2.1 Null and Alternate Hypotheses for Analysis of Variance .....	172
7.2.2 Assumptions of the Analysis of Variance Test .....	172
7.2.3 Computation of Classic ANOVA .....	173
7.2.4 Welch's Adjusted ANOVA .....	174
7.3 Permutation Test for Difference in Means .....	175
7.3.1 Computation of the Permutation Test of Means.....	176
7.4 Two-factor Analysis of Variance.....	177
7.4.1 Null and Alternate Hypotheses for Two-factor ANOVA.....	177
7.4.2 Assumptions of Two-factor ANOVA .....	177
7.4.3 Computation of Two-factor ANOVA.....	178
7.4.4 Interaction Effects in Two-factor ANOVA .....	181
7.4.5 Two-factor ANOVA on Logarithms or Other Transformations.....	182
7.4.6 Fixed and Random Factors.....	183
7.5 Two-factor Permutation Test .....	184
7.6 Two-factor Nonparametric Brunner-Dette-Munk Test.....	185
7.7 Multiple Comparison Tests.....	185
7.7.1 Parametric Multiple Comparisons for One-factor ANOVA.....	186
7.7.2 Nonparametric Multiple Comparisons Following the Kruskal-Wallis Test .....	188
7.7.3 Parametric Multiple Comparisons for Two-factor ANOVA.....	192
7.7.4 Nonparametric Multiple Comparisons for Two-factor BDM Test .....	193
7.8 Repeated Measures—The Extension of Matched-pair Tests .....	193
7.8.1 Median Polish.....	194
7.8.2 The Friedman Test .....	198
7.8.3 Computation of the Friedman Test .....	199
7.8.4 Multiple Comparisons for the Friedman Test .....	199

7.8.5 Aligned-ranks Test.....	.200
7.8.6 Multiple Comparisons for the Aligned-ranks Test.....	.202
7.8.7 Two-factor ANOVA Without Replication .....	.203
7.8.8 Computation of Two-factor ANOVA Without Replication .....	.204
7.8.9 Parametric Multiple Comparisons for ANOVA Without Replication.....	.205
7.9 Group Tests for Data with Nondetects.....	.206
<b>Chapter 8 Correlation.....</b>	<b>.209</b>
8.1 Characteristics of Correlation Coefficients.....	.209
8.1.1 Monotonic Versus Linear Correlation .....	.210
8.2 Pearson's $r$ .....	.212
8.2.1 Computation.....	.212
8.2.2 Hypothesis Tests.....	.212
8.3 Spearman's Rho ( $\rho$ ).....	.214
8.3.1 Computation of Spearman's $\rho$ .....	.215
8.3.2 Hypothesis Tests for Spearman's $\rho$ .....	.215
8.4 Kendall's Tau ( $\tau$ ).....	.217
8.4.1 Computation of Kendall's $\tau$ .....	.217
8.4.2 Hypothesis Tests for Kendall's $\tau$ .....	.218
8.4.3 Correction for Tied Data when Performing Hypothesis Testing Using Kendall's $\tau$ .....	.220
<b>Chapter 9 Simple Linear Regression.....</b>	<b>.223</b>
9.1 The Linear Regression Model.....	.224
9.1.1 Computations.....	.225
9.1.2 Assumptions of Linear Regression .....	.228
9.1.3 Properties of Least Squares Solutions .....	.228
9.2 Getting Started with Linear Regression .....	.229
9.3 Describing the Regression Model .....	.231
9.3.1 Test for Whether the Slope Differs from Zero.....	.233
9.3.2 Test for Whether the Intercept Differs from Zero .....	.233
9.3.3 Confidence Intervals on Parameters .....	.234
9.4 Regression Diagnostics.....	.236
9.4.1 Assessing Unusual Observations .....	.238
9.4.1.1 Leverage.....	.238
9.4.1.2 Influence .....	.238
9.4.1.3 Prediction, Standardized and Studentized Residuals .....	.240
9.4.1.4 Cook's $D$ .....	.241
9.4.1.5 DFFITS.....	.241
9.4.2 Assessing the Behavior of the Residuals .....	.242
9.4.2.1 Assessing Bias and Homoscedasticity of the Residuals .....	.242
9.4.2.2 Assessing Normality of the Residuals .....	.244
9.4.2.3 Assessing Serial Correlation of the Residuals .....	.246
9.4.2.4 Detection of Serial Correlation.....	.246
9.4.2.5 Strategies to Address Serial Correlation of the Residuals .....	.248
9.4.3 A Measure of the Quality of a Regression Model Using Residuals: PRESS .....	.248

9.5 Confidence and Prediction Intervals on Regression Results .....	248
9.5.1 Confidence Intervals for the Mean Response .....	248
9.5.2 Prediction Intervals for Individual Estimates of $y$ .....	250
9.5.2.1 Nonparametric Prediction Interval.....	252
9.6 Transformations of the Response Variable, $y$ .....	254
9.6.1 To Transform or Not to Transform.....	254
9.6.2 Using the Logarithmic Transform.....	255
9.6.3 Retransformation of Estimated Values .....	255
9.6.3.1 Parametric or the Maximum Likelihood Estimate of the Bias Correction Adjustment .....	256
9.6.3.2 Nonparametric or Smearing Estimate of the Bias Correction Adjustment.....	256
9.6.4 Special Issues Related to a Logarithm Transformation of $y$ .....	261
9.7 Summary Guide to a Good SLR Model.....	263
<b>Chapter 10 Alternative Methods for Regression .....</b>	<b>267</b>
10.1 Theil-Sen Line.....	267
10.1.1 Computation of the Line.....	268
10.1.2 Properties of the Estimator .....	270
10.1.3 Test of Significance for the Theil-Sen Slope.....	276
10.1.4 Confidence Interval for the Theil-Sen Slope.....	277
10.2 Alternative Linear Equations for Mean $y$ .....	278
10.2.1 OLS of $x$ on $y$ .....	279
10.2.2 Line of Organic Correlation.....	280
10.2.3 Least Normal Squares .....	283
10.2.4 Summary of the Applicability of OLS, LOC, and LNS .....	284
10.3 Smoothing Methods.....	285
10.3.1 Loess Smooths .....	287
10.3.2 Lowess Smooths.....	288
10.3.3 Upper and Lower Smooths .....	289
10.3.4 Use of Smooths for Comparing Large Datasets .....	290
10.3.5 Variations on Smoothing Algorithms.....	291
<b>Chapter 11 Multiple Linear Regression .....</b>	<b>295</b>
11.1 Why Use Multiple Linear Regression?.....	295
11.2 Multiple Linear Regression Model .....	296
11.2.1 Assumptions and Computation .....	296
11.3 Hypothesis Tests for Multiple Regression.....	296
11.3.1 Nested $F$ -Tests.....	296
11.3.2 Overall $F$ -Test.....	297
11.3.3 Partial $F$ -Tests .....	297
11.4 Confidence and Prediction Intervals.....	298
11.4.1 Variance-covariance Matrix.....	298
11.4.2 Confidence Intervals for Slope Coefficients .....	299
11.4.3 Confidence Intervals for the Mean Response .....	299
11.4.4 Prediction Intervals for an Individual $y$ .....	299

11.5 Regression Diagnostics.....	.299
11.5.1 Diagnostic Plots.....	.300
11.5.2 Leverage and Influence.....	.300
11.5.3 Multicollinearity.....	.308
11.6 Choosing the Best Multiple Linear Regression Model.....	.313
11.6.1 Stepwise Procedures .....	.314
11.6.2 Overall Measures of Quality .....	.315
11.6.3 All-Subsets Regression .....	.317
11.7 Summary of Model Selection Criteria .....	.320
11.8 Analysis of Covariance .....	.320
11.8.1 Use of One Binary Variable.....	.320
11.8.2 Multiple Binary Variables.....	.323
<b>Chapter 12 Trend Analysis .....</b>	<b>.327</b>
12.1 General Structure of Trend Tests.....	.328
12.1.1 Purpose of Trend Testing.....	.328
12.1.2 Approaches to Trend Testing.....	.331
12.2 Trend Tests with No Exogenous Variable .....	.332
12.2.1 Nonparametric Mann-Kendall Test.....	.332
12.2.2 Ordinary Least Squares Regression of $Y$ on Time, $T$ .....	.335
12.2.3 Comparison of Simple Tests for Trend .....	.335
12.3 Accounting for Exogenous Variables.....	.336
12.3.1 Mann-Kendall Trend Test on Residuals, $R$ , from Loess of $Y$ on $X$ .....	.338
12.3.2 Mann-Kendall Trend Test on Residuals, $R$ , from Regression of $Y$ on $X$ .....	.339
12.3.3 Regression of $Y$ on $X$ and $T$ .....	.340
12.4 Dealing with Seasonality or Multiple Sites .....	.342
12.4.1 The Seasonal Kendall Test.....	.343
12.4.2 Mixed Method—OLS Regression on Deseasonalized Data .....	.345
12.4.3 Fully Parametric Model—Multiple Regression with Periodic Functions.....	.345
12.4.4 Comparison of Methods for Dealing with Seasonality.....	.346
12.4.5 Presenting Seasonal Effects .....	.347
12.4.6 Seasonal Differences in Trend Magnitude.....	.347
12.4.7 The Regional Kendall Test.....	.349
12.5 Use of Transformations in Trend Studies.....	.349
12.6 Monotonic Trend Versus Step Trend.....	.352
12.6.1 When to Use a Step-trend Approach.....	.352
12.6.2 Step-trend Computation Methods .....	.353
12.6.3 Identification of the Timing of a Step Change.....	.355
12.7 Applicability of Trend Tests with Censored Data .....	.355
12.8 More Flexible Approaches to Water Quality Trend Analysis .....	.357
12.8.1 Methods Using Antecedent Discharge Information.....	.357
12.8.2 Smoothing Approaches.....	.358
12.9 Discriminating Between Long-term Trends and Long-term Persistence.....	.359
12.10 Final Thoughts About Trend Assessments .....	.362

<b>Chapter 13 How Many Observations Do I Need?</b> .....	<b>365</b>
13.1 Power Calculation for Parametric Tests .....	365
13.2 Why Estimate Power or Sample Size? .....	369
13.3 Power Calculation for Nonparametric Tests.....	370
13.3.1 Estimating the Minimum Difference PPlus.....	370
13.3.2 Using the power.WMW Script.....	374
13.4 Comparison of Power for Parametric and Nonparametric Tests .....	376
 <b>Chapter 14 Discrete Relations</b> .....	<b>385</b>
14.1 Recording Categorical Data .....	385
14.2 Contingency Tables .....	386
14.2.1 Performing the Test for Association.....	386
14.2.2 Conditions Necessary for the Test.....	389
14.2.3 Location of the Differences .....	389
14.3 Kruskal-Wallis Test for Ordered Categorical Responses .....	390
14.3.1 Computing the Test.....	390
14.3.2 Multiple Comparisons.....	392
14.4 Kendall's Tau for Categorical Data .....	393
14.4.1 Kendall's $\tau_b$ for Tied Data .....	393
14.4.2 Test of Significance for $\tau_b$ .....	395
14.5 Other Methods for Analysis of Categorical Data .....	399
 <b>Chapter 15 Regression for Discrete Responses</b> .....	<b>401</b>
15.1 Regression for Binary Response Variables.....	401
15.1.1 The Logistic Regression Model.....	402
15.1.2 Important Formulae for Logistic Regression.....	403
15.1.3 Estimation of the Logistic Regression Model .....	403
15.1.4 Hypothesis Tests for Nested Models .....	404
15.1.5 Amount of Uncertainty Explained, $R^2$ .....	404
15.1.6 Comparing Non-nested Models .....	405
15.2 Alternatives to Logistic Regression.....	409
15.2.1 Discriminant Function Analysis.....	409
15.2.2 Rank-sum Test.....	409
15.3 Logistic Regression for More Than Two Response Categories.....	410
 <b>Chapter 16 Presentation Graphics</b> .....	<b>413</b>
16.1 The Value of Presentation Graphics.....	413
16.2 General Guidelines for Graphics.....	414
16.3 Precision of Graphs.....	415
16.3.1 Color.....	415
16.3.2 Shading.....	416
16.3.3 Volume and Area .....	418
16.3.4 Angle and Slope.....	418
16.3.5 Length.....	420
16.3.6 Position Along Nonaligned Scales .....	420

16.3.7 Position Along an Aligned Scale.....	422
16.4 Misleading Graphics to Be Avoided .....	423
16.4.1 Perspective.....	423
16.4.2 Graphs with Numbers .....	425
16.4.3 Hidden Scale Breaks.....	425
16.4.4 Self-scaled Graphs.....	427
16.5 Conclusion .....	429
References Cited.....	431
Index.....	451

## Figures

1.1. Graphs showing a lognormal distribution.....	3
1.2. Graphs showing a normal distribution .....	4
1.3. Graph showing the arithmetic mean as the balance point of a dataset .....	7
1.4. Graph showing the shift of the arithmetic mean downward after removal of an outlier.....	7
2.1. Four scatterplots of datasets that all have the same traditional statistical properties of mean, variance, correlation, and x-y regression intercept and coefficient .....	18
2.2. Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015 .....	19
2.3. Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015 .....	19
2.4. Quantile plot of annual mean discharge data from the James River, Virginia, 1900–2015 .....	21
2.5. A boxplot of annual mean discharge values from the James River, Virginia, 1900–2015 .....	23
2.6. Boxplot of the unit well yields for valleys with unfractured rocks from Wright.....	25
2.7. Boxplot of the natural log of unit well yield for valleys with unfractured rocks from Wright .....	25
2.8. Overlay of James River annual mean discharge and standard normal distribution quantile plots.....	27
2.9. Probability plot of James River annual mean discharge data .....	27
2.10. Exceedance probability plot for the James River annual mean discharge data .....	28
2.11. A normal probability plot of the Potomac River annual peak discharge data .....	29
2.12. Normal probability plot of the natural log of the annual peak discharge data from the Potomac River at Point of Rocks, Maryland, streamgage .....	31
2.13. Boxplot of the natural log of the annual peak discharge data from the Potomac River at Point of Rocks, Maryland, streamgage .....	31
2.14. Histograms of unit well yield data for valleys with fractures, and valleys without fractures .....	32
2.15. Dot-and-line plot of the unit well yield datasets for areas underlain by either fractured or unfractured rock.....	33
2.16. Side-by-side boxplots of the unit well yield datasets for areas underlain by either fractured or unfractured rock .....	34
2.17. Boxplots of ammonia nitrogen concentrations as a function of location on two transects of the Detroit River.....	35

2.18.	Side-by-side boxplots by month for dissolved nitrate plus nitrite for the Illinois River at Valley City, Illinois, water years 2000–15.....	37
2.19.	Probability plots of the unit well yield data .....	37
2.20.	Q-Q plot of the well yield data in fractured and unfractured areas .....	38
2.21.	Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, for the months of June, July, August, and September of 1990–2008 .....	39
2.22.	Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September .....	40
2.23.	Loess smooths representing dependence of log(As) on pH for four areas in the western United States .....	41
2.24.	Scatterplot of water-quality measures draining three types of upstream land use .....	43
2.25.	Polar smooths with 75 percent coverage for the three groups of data seen in figure 2.24, from Helsel .....	43
2.26.	Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September or the months of January, February, March, and April .....	44
2.27.	Absolute residuals from the loess smooth of ln(NO <sub>2</sub> 3) concentrations versus ln(discharge), Iowa River at Wapello, Iowa, for the warm season 1990–2008.....	45
2.28.	Parallel plot of six basin characteristics at a low salinity site and a high salinity site .....	47
2.29.	Parallel plot of six basin characteristics at the 19 sites of Warwick.....	47
2.30.	Stiff diagrams used to display differences in water quality in the Fox Hills Sandstone, Wyoming .....	48
2.31.	Star plots of site characteristics for 19 locations along the Exe estuary.....	49
2.32.	Trilinear diagram for groundwater cation composition in four geologic zones of the Groundwater Ambient and Monitoring Assessment Program Sierra Nevada study unit.....	49
2.33.	Piper diagram of groundwater from the Sierra Nevada study unit of the Groundwater Ambient Monitoring and Assessment Program.....	50
2.34.	Scatterplot matrix showing the relations between six site characteristics.....	51
2.35.	Principal component analysis biplot of site characteristics along the Exe estuary.....	53
2.36.	Nonmetric multidimensional scaling showing the relations among sites, and between sites and variables, using the six site characteristics of Warwick.....	54
2.37.	Two three-dimensional plots of the site characteristics data of Warwick.....	54
2.38.	Stacked bar charts of mean percent milliequivalents of anion and cations within the four Groundwater Ambient Monitoring and Assessment Program lithologic zones of Shelton and others .....	55
3.1.	Ten 90-percent confidence intervals for normally distributed data with true mean = 5 and standard deviation = 1, in milligrams per liter.....	59
3.2.	Boxplot of a random sample of 1,000 observations from a lognormal distribution.....	60
3.3.	Ten 90-percent confidence intervals around a true mean of 1, each one based on a sample size of 12.....	61
3.4.	Boxplots of the original and log-transformed arsenic data from Boudette and others used in example 3.1.....	62
3.5.	Plot of the 95-percent confidence interval for the true median in example 3.1 .....	64

3.6.	Histogram of bootstrapped estimates of the mean of arsenic concentrations used in example 3.1 .....	69
3.7.	Example of a probability distribution showing the 90-percent prediction interval, with $\alpha=0.10$ .....	71
3.8.	Histogram of the James River annual mean discharge dataset .....	71
3.9.	Example of a probability distribution showing the one-sided 90-percent prediction interval .....	72
3.10.	A two-sided tolerance interval with 90-percent coverage, and two-sided confidence interval on the 90th percentile .....	76
3.11.	Confidence interval on the $p$ th percentile $X_p$ as a test for $H_0: X_p = X_0$ .....	79
3.12.	Lower tolerance limit as a test for whether the percentile $X_p > X_0$ .....	82
3.13.	Upper tolerance limit as a test for whether percentile $X_p < X_0$ .....	85
3.14.	Boxplot of a sample of size 25 from a lognormal distribution .....	90
4.1.	Five classes of hypothesis tests .....	94
4.2.	Four possible results of hypothesis testing .....	101
4.3.	Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5 .....	105
4.4.	Probabilities of occurrence for a rank-sum test with sample sizes of 2 and 5 .....	106
4.5.	Probability plots for yields of wells in unfractured and fractured rock, with probability plot correlation coefficient (PPCC) correlation coefficient ( $r$ ) .....	110
4.6.	Cartoon showing what can easily happen when running multiple hypothesis tests, $p$ -hacking or $p$ -fishing .....	113
5.1.	Box plots of concentration data for two groups and the logarithms of the same data .....	119
5.2.	Boxplots of ammonia plus organic nitrogen from the precipn data .....	121
5.3.	Histogram showing the distribution of the exact test statistic $W_{rs}$ and its fitted normal approximation for $n=10$ and $m=10$ .....	122
5.4.	Histogram showing 10,000 permuted differences in group means for the precipn dataset, computed by rearrangement of the group assignments .....	126
5.5.	Q-Q plot of the precipitation nitrogen data from example 5.1 .....	136
5.6.	Q-Q plot of the logs of the precipitation nitrogen data from example 5.1 .....	136
5.7.	Box plots of two groups of 50 samples each of randomly generated data from a single lognormal distribution .....	139
6.1.	Boxplots of mayfly nymph counts at two different sites, and the differences .....	148
6.2.	Boxplot of the differences of the natural logarithms of the mayfly data from example 6.1 .....	154
6.3.	Histogram of permuted differences and the observed mean difference from the logs of mayfly data from example 6.1 .....	158
6.4.	Scatterplot of the mayfly data from example 6.1 .....	160
6.5.	Mayfly data from example 6.1 .....	160
7.1.	Boxplots of specific capacity of wells in four rock types .....	169
7.2.	Quantile plots of the natural log of specific capacity for the four rock types from Knopman .....	170
7.3.	Hypothetical data for three groups .....	171
7.4.	Hypothetical data for three groups .....	172
7.5.	Histogram of $F$ -statistics for 10,000 permutations of the specific capacity group assignments from example 7.1 .....	176
7.6.	Boxplots of iron concentrations at low flow from Helsel .....	179

7.7.	Q-Q plot showing the non-normality of the ANOVA residuals of the iron data from example 7.5 .....	180
7.8.	Interaction plot presenting the means of data in the six treatment groups from example 7.5 showing no interaction between the two factor effects .....	181
7.9.	Interaction plot showing interaction by the large nonparallel increase in the mean for the combination of abandoned mining history and sandstone rock type .....	182
7.10.	Boxplots of the natural logarithms of the iron data from example 7.5.....	183
7.11.	Natural logs of specific capacity of wells in four rock types in Pennsylvania.....	188
7.12.	The 95-percent Tukey family confidence intervals on differences in group means of the data from Knopman.....	189
7.13.	Boxplots showing mercury concentrations in periphyton along the South River, Virginia, from upstream to downstream .....	195
7.14.	Residuals from the median polish of periphyton mercury data from Walpole and Myers.....	198
8.1.	Plot showing monotonic, but nonlinear, correlation between $x$ and $y$ .....	210
8.2.	Plot showing monotonic linear correlation between $x$ and $y$ .....	211
8.3.	Plot showing nonmonotonic relation between $x$ and $y$ .....	211
8.4.	Plot of example 8.1 data showing one outlier present .....	213
9.1.	Plot of the true linear relation between the response variable and the explanatory variable, and 10 observations of the response variable for explanatory variable values at integer values from 1 through 10 .....	224
9.2.	Plot showing true and estimated linear relation between the explanatory and response variables using the observations from figure 9.1.....	225
9.3.	Plot of true and estimated linear relations between $x$ and $y$ from different sets of 10 observations all generated using the true relation and sampling error.....	226
9.4.	The bulging rule for transforming curvature to linearity .....	230
9.5.	Scatterplot of discharge versus total dissolved solids concentrations for the Cuyahoga River, Ohio, over the period 1969–73.....	231
9.6.	Scatterplot of discharge versus total dissolved solids concentrations after the transformation of discharge using the natural log.....	232
9.7.	Plots of four different datasets fit with a simple linear regression .....	237
9.8.	Influence of location of a single point on the regression slope.....	239
9.9.	Plot of 95-percent confidence intervals for the mean total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1 .....	249
9.10.	Plot of 95-percent prediction intervals for an individual estimate of total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1 .....	251
9.11.	Plot of 95-percent parametric and nonparametric prediction intervals for an individual estimate of total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1 .....	253
9.12.	Comparison of the relation between discharge and total phosphorus concentration for the Maumee River, in original units .....	257
10.1.	Computation of the Theil-Sen slope .....	269
10.2.	Plot of the Theil-Sen and ordinary least-squares regression fits to the example data .....	271
10.3.	Probability density functions of two normal distributions used by Hirsch and	

others, the first with mean = 10 and standard deviation = 1; the second with mean = 11 and standard deviation = 3 .....	272
10.4. Probability density function of a mixture of data (95 percent from distribution 1 and 5 percent from distribution 2) .....	272
10.5. Probability density function of a mixture of data (80 percent from distribution 1 and 20 percent from distribution 2) .....	273
10.6. Relative efficiency of the Theil-Sen slope estimator as compared with the ordinary least squares slope represented as the ratio of the root mean square error of the Theil-Sen estimator to the OLS estimator.....	274
10.7. Scatterplot of total phosphorus concentrations for the St. Louis River at Scanlon, Minnesota, 1975–89 with ordinary least squares regression and Theil-Sen fitted lines .....	276
10.8. Plot of three straight lines fit to the same data.....	280
10.9. Characteristics of four parametric methods to fit straight lines to data .....	281
10.10. Boxplot of the original residue on evaporation data in comparison to boxplots of predicted values from the line of organic correlation and regression lines .....	282
10.11. Plot of four straight lines fit to the same data.....	284
10.12. Nitrate concentrations as a function of daily mean discharge during the months of June through September of the years 1990–2008 for the Iowa River at Wapello, Iowa, showing linear and quadratic fit, estimated as the natural log of concentration as a function of natural log of discharge.....	286
10.13. Graph of the tri-cube weight function, where $d_{max} = 10$ and $x^* = 20$ .....	288
10.14. Graphs of smooths of nitrate concentration as a function of daily mean discharge during the months of June through September of the years 1990–2008 for the Iowa River at Wapello, Iowa .....	289
10.15. Graph of annual mean daily discharge of the Colorado River at Lees Ferry, Arizona, for water years 1922–2016 .....	290
10.16. Plot of lowess smooths of sulfate concentrations at 19 stations, 1979–83.....	291
11.1. Scatterplot matrix for the variables listed in table 11.1 .....	301
11.2. Rotated scatterplot showing the position of the high leverage point.....	303
11.3. Partial-regression plots for concentration as a function of distance east, distance north, and well depth .....	304
11.4. Partial-regression plots for concentration as a function of distance east, distance north, and well depth, with outlier corrected .....	306
11.5. Component + residual plots for concentration as a function of distance east, distance north, and well depth, with outlier corrected .....	307
11.6. Plots of the magnitude of adjusted R-squared, Mallow's $C_p$ , BIC, and residual sum of squares for the two best explanatory variable models as a function of the number of explanatory variables .....	318
11.7. Plot of regression lines for data differing in intercept between two seasons .....	321
11.8. Plot of regression lines differing in slope and intercept for data from two seasons.....	323
12.1. Map showing trend analysis results for specific conductance for the time period 1992–2002 .....	329
12.2. Plot of annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013, shown with the Theil-Sen robust line.....	333
12.3. Plot of the natural log of annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013, shown with the Theil-Sen robust line .....	334
12.4. Plot of the annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013,	

shown with the transformed Theil-Sen robust line based on slope of the natural log discharge values .....	334
12.5. Graph of chloride concentration in the month of March for the Milwaukee River, at Milwaukee, Wisconsin.....	336
12.6. Graph of the relation between chloride concentration and the natural log of discharge, Milwaukee River at Milwaukee, Wisconsin, for samples collected in March 1978–2005 .....	338
12.7. Graph of concentration residuals versus time for chloride concentrations in the Milwaukee River at Milwaukee, Wisconsin, for samples collected in March, from 1978 through 2005 .....	339
12.8. Graph of log concentration residuals versus time for chloride concentrations in the Milwaukee River at Milwaukee, Wisconsin, for samples collected in March, from 1978 through 2005 .....	340
12.9. Graph of curves that represent median estimates of chloride concentration as a function of time, from the Milwaukee River at Milwaukee, Wisconsin.....	341
12.10. Graph of monthly trends in discharge, Sugar River near Brodhead, Wisconsin, for water years 1952–2016.....	348
12.11. Graph showing trend in annual mean discharge, Big Sioux River at Akron, Iowa.....	350
12.12. Graph of trend in the natural log of annual mean discharge, Big Sioux River at Akron, Iowa.....	351
12.13. Graph of annual peak discharge, North Branch Potomac River at Luke, Maryland....	354
12.14. Graph of the natural log of annual peak discharge, North Branch Potomac River at Luke, Maryland .....	355
12.15. Graph of contoured surface describing the relation between the expected value of filtered nitrate plus nitrite concentration as a function of time and discharge for the Choptank River near Greensboro, Maryland.....	358
12.16. Graph of annual peak discharge, Red River of the North, at Grand Forks, North Dakota, 1940–2014, and a loess smooth of the data .....	360
12.17. Graph of annual peak discharge, Red River of the North, at Grand Forks, North Dakota, 1882–2014, and a loess smooth of the data .....	360
13.1. Boxplots of the two groups of molybdenum data from exercise 2 of chapter 5 .....	372
13.2. Graph showing the effect on <i>PPlus</i> of adding 1.0 to the downgradient observations .....	372
13.3. Graph showing the effect on <i>PPlus</i> of adding 2.0 to the downgradient observations .....	373
13.4. Graph showing the effect on <i>PPlus</i> of adding 3.0 to the downgradient observations .....	373
13.5. Graph showing <i>GMratio</i> versus <i>PPlus</i> using the observed standard deviation of logarithms of 0.70 and 0.87 from the molybdenum dataset of chapter 5.....	375
13.6. Graph showing power to differentiate downgradient from upgradient molybdenum concentrations for various sample sizes for dissimilar data versus quite similar data .....	376
14.1. Structure of a two variable, 2x3 matrix of counts .....	385
14.2. The 2x3 matrix of observed counts for the data from example 14.1.....	387
14.3. The 2x3 matrix of expected counts for the data from example 14.1.....	388
14.4. A 2x3 matrix for a Kruskal-Wallis analysis of an ordered response variable.....	391
14.5. The 2x3 matrix of observed counts for example 14.2.....	392
14.6. Diagram showing suggested ordering of rows and columns for computing $\tau_b$ .....	394

14.7.	Diagrams of 3x3 matrix cells contributing to P .....	394
14.8.	Diagrams of 3x3 matrix cells contributing to M.....	395
14.9.	The 3x3 matrix of observed counts for example 14.3.....	396
15.1.	Graph of logistic regression equations; solid curve has a positive relation between the explanatory variable and $p$ , the dashed curve has a negative relation.....	402
15.2.	Graph of estimated trichloroethylene detection probability as a function of population density, showing 95-percent confidence intervals.....	407
16.1.	Map showing total offstream water withdrawals by state, 1980, from Solley and others.....	416
16.2.	Map of withdrawals for offstream water use by source and state, from Solley and others .....	417
16.3.	Graph of seasonal flow distribution for the Red River of the North at Grand Forks, North Dakota, for water years 1994–2015.....	418
16.4.	Bar chart of seasonal flow distribution for the Red River of the North at Grand Forks, North Dakota, for water years 1994–2015, using the same data as in figure 16.3.....	419
16.5.	Graph of measured and simulated streamflow .....	419
16.6.	Graph demonstrating that judgment of length is more difficult without a common scale.....	420
16.7.	Graph showing how framed rectangles improve figure 16.6 by adding a common scale.....	420
16.8.	Stacked bar charts demonstrating the difficulty of comparing data not aligned with the y-axis .....	421
16.9.	Grouped bar charts display the same data as in figure 16.8.....	421
16.10.	Boxplots of nitrate concentrations by land use and sewerage .....	422
16.11.	Boxplots of nitrate concentrations in milligrams per liter as N, Iowa River at Wapello, Iowa, water years 2000–16.....	423
16.12.	Pie chart in perspective view, showing the drainage area of the five largest monitored tributaries in the Chesapeake Bay watershed .....	424
16.13.	Bar chart of water use data, in perspective view.....	424
16.14.	Map of water use data in perspective view.....	425
16.15.	Graph of simulated concentration and depth data plotted with no scale breaks, a scale break indicated by a zigzag line at the break, and a full-scale break.....	426
16.16.	Two graphical presentations of the same hypothetical dataset .....	427
16.17.	Time series of nitrate concentration data for the Appomattox River at Matoaca, Virginia, and Patuxent River near Bowie, Maryland .....	428
16.18.	Graphs showing the same datasets as in figure 16.17, but with the scaling of both axes identical across the two graphs .....	429

## Tables

1.1.	Ladder of powers as modified from Velleman and Hoaglin.....	14
2.1.	Quantile plot values for streamflow data from the James River, Virginia, 1900–2015.....	21
2.2.	Definitions and comments on eight possible plotting position formulas, based on Hyndman and Fan and Stedinger and others.....	23
3.1.	Ten replicate datasets of 12 samples each of chloride concentrations, each with mean = 5 and standard deviation = 1 .....	59
3.2.	Arsenic concentrations for groundwaters of southeastern New Hampshire, ranked in ascending order.....	62
3.3.	Log-transformed arsenic concentrations for groundwaters of southeastern New Hampshire, ranked in ascending order .....	66
3.4.	Comparison of 95-percent confidence interval estimators for various measures of central tendency for the arsenic data, in parts per billion .....	69
4.1.	Guide to the classification of some hypothesis tests with continuous response variables.....	97
4.2.	Probabilities and one-sided <i>p</i> -values for the rank-sum test with <i>n</i> =2 and <i>m</i> =5 .....	105
4.3.	Unit well yields from Virginia, in gallons per minute per foot.....	109
5.1.	Hypothesis test methods in this chapter and their characteristics .....	117
6.1.	Paired difference tests of this chapter and their characteristics .....	145
7.1.	Hypothesis tests with one factor and their characteristics .....	166
7.2.	Hypothesis tests with two factors and their characteristics .....	166
7.3.	Hypothesis tests for repeated measures and their characteristics .....	167
7.4.	Kruskal-Wallis test statistic computation for fecal coliform counts.....	168
7.5.	Schematic of a one-factor ANOVA table.....	174
7.6.	Sums of squares definitions for two-factor ANOVA.....	178
7.7.	Schematic for a two-factor ANOVA table .....	178
7.8.	Mercury concentrations, in micrograms per gram, in periphyton .....	194
7.9.	Data from table 7.8 aligned by subtraction of row medians.....	196
7.10.	Data from table 7.9 after subtraction of the median of row medians .....	196
7.11.	Data from table 7.10 after subtractions of column medians from their respective column's data .....	196
7.12.	First polish of the periphyton data of Walpole and Myers.....	197
7.13.	Aligned ranks of the aligned periphyton mercury data from table 7.8.....	201
7.14.	Sums of squares definitions for two-factor ANOVA.....	204
7.15.	Analysis of variance table for two factors without replication.....	205
7.16.	Mercury concentrations, in micrograms per liter, in periphyton.....	206
9.1.	Formulas utilized in ordinary least squares linear regression .....	227
9.2.	Assumptions necessary for the purposes to which ordinary least squares regression is applied. ....	228
9.3.	Comparison of results of regression of $\ln(C)$ on $\ln(Q)$ versus $\ln(L)$ on $\ln(Q)$ .....	261
10.1.	Intercepts and slopes for the four lines of figure 10.11 .....	285
11.1.	Data and diagnostics for chemical concentrations used in example 11.1 .....	302
11.2.	Data for example 11.2.....	311
11.3.	Results of forward selection procedure for example 11.3.....	315
11.4.	Results of stepwise selection procedure for example 11.3.....	315
11.5.	Statistics for several multiple regression models of Haan's data .....	317

12.1.	Probabilities associated with possible outcomes of a trend test.....	330
12.2.	Classification of five types of tests for trend.....	331
12.3.	Model coefficients and their <i>t</i> -statistics and <i>p</i> -values .....	341
12.4.	General categories of options for dealing with seasonality in conducting trend tests.....	343
12.5.	Methods for characterizing seasonal patterns .....	347
12.6.	Step-trend tests that do not consider seasonality.....	353
12.7.	Step-trend tests that consider seasonality .....	353
12.8.	Classification of monotonic trend tests for censored data .....	357
13.1.	Power for three samples sizes for the specific capacity data using the <i>t</i> -test, <i>t</i> -test on the logarithms, and the rank-sum test .....	382
15.1.	Trichloroethylene data in the Upper Glacial Aquifer, Long Island, New York.....	405
16.1.	Precision of perceptual tasks summarized from figure 1 of Cleveland and McGill.....	415

## Preface

This book began as a collection of class notes for a course on applied statistical methods for hydrologists taught by Dennis Helsel, Robert Hirsch, Ed Gilroy, and others at the U.S. Geological Survey (USGS) National Training Center. The first course was offered in 1986 and still continues at the USGS in a modified form more than 30 years later. Course material was formalized and organized into a textbook, first published in 1992 by Elsevier as part of their Studies in Environmental Science series. The first hardback contained an actual “floppy disk” in the back! The paperback that followed contained a 3.5-inch diskette, as technology swiftly changed. In 2002, the text was republished by the USGS in its Techniques of Water-Resources Investigations series (book 4, chapter A3, version 1.1). The 2002 republished version was made freely available online in digital form, though initially as individual chapter files in case the full 12 MB file might overwhelm the reader’s download capability. Both the hardback version published in 1992 and the republished version in 2002 are considered the first edition of this book since the republished version in 2002 is identical to the 1992 version except for five small errors fixed in the more recent version.

Our book was originally intended to be a text in an applied statistics course in hydrology, environmental science, environmental engineering, geology, or biology. For too long, scientists had been asked to take examples from business, agronomy, public health, or other areas and apply them to water resources science. Difficulties in doing so included the fact that water resources data tend to be more skewed than data from other disciplines, and the common expectation that “the assumption of a normal distribution is appropriate” was clearly not sufficient. Our book was never intended to be a stand-alone text on statistics or a text on statistical hydrology. There were (and are) excellent texts already available on probability theory and the statistics of extreme events.

For this update, much has changed and much has stayed the same. We again chose to emphasize topics not always found in introductory statistics textbooks and often not adequately covered in statistical textbooks for scientists and engineers. We also point scientists toward robust and nonparametric statistics, and to exploratory data analysis. Major changes are the result of advances in computer technology now available to scientists. Less familiar but very important resampling methods such as bootstrap and permutation tests have been added, much as smoothing and Kendall-based trend methods were new for many readers back in 1992. As before, exercises are included at the end of chapters.

The textbook now utilizes R, a programming language and free software environment for statistical computing and graphics (<https://www.r-project.org/>). Text in the book shown in the font **Consolas** denotes commands, functions, inputs, or outputs from R. More specifically, text shown in the font **Consolas** and preceded by the cursor symbol (>) are R commands, followed by R output generated by these commands. When an R command was too long to fit on one line of text, the next line begins with a “+” symbol to denote the continuation of the R command. This symbol must be deleted when copying and pasting the full command into R or the command will fail to execute. Supplemental material (SM) for each chapter are available at <https://doi.org/10.5066/P9JWL6XR> to re-create all examples and figures, and to solve the exercises at the end of each chapter, with relevant datasets provided in an electronic format readable by R. The SM, defined and referred to in each chapter as SM.X (where X is the chapter

number) provide (1) datasets as .Rdata files for immediate input into R, (2) datasets as .csv files for input into R or for use with other software programs, (3) R functions that are used in the textbook but not part of a published R package, (4) R scripts to produce virtually all of the figures in the book, and (5) solutions to the exercises as .html and .Rmd files. The suffix .Rmd refers to the file format for code written in the R Markdown language; the .Rmd file that is provided in the SM was used to generate the .html file containing the solutions to the exercises. Unless otherwise noted, all data used in the in-text examples, figures, and exercises are downloaded from the National Water Information System (U.S. Geological Survey, 2016) by means of the `dataRetrieval` R package (De Cicco, 2016).

With a few exceptions of reprints, graphs in the text were plotted in R and are reproduced exactly as output by the R code published in the SM for this book. The graphics are not always uniform in their formatting to show variation in graphical options in R. Seeing these different formats and how they are created may help readers to select the outputs most useful for their work. Because the graphs are provided as examples of output and for instructional purposes only, they have not been redrafted to follow the USGS standards for page-size illustrations in terms of fonts and line weights.

Many contributed to the first edition, including other instructors for the USGS course. Ed Gilroy critiqued and improved much of the original material, and now has added his own. Tim Cohn contributed in several areas, particularly to the sections on bias correction in regression and methods for data below the reporting limit. Richard Alexander added to the trend analysis chapter, and Charles Crawford contributed ideas for regression and analysis of variance; their work has undoubtedly made its way into this new edition. Ken Potter (University of Wisconsin) and Gary Tasker (USGS) reviewed the original manuscript, spending long hours with no reward except the knowledge that they have improved the work of others. Madeline Sabin carefully typed original drafts of the class notes on which the first edition was based.

For the second edition, three new authors were added, including Karen Ryberg and Stacey Archfield, who are presently among the instructors of the current version of the USGS course in Environmental Data Analysis. Ed Gilroy expanded the book's reach considerably by teaching it to more than 1,000 students in Federal and state agencies and private firms after his retirement from USGS.

Ken Potter (University of Wisconsin) and William Farmer (USGS) provided a review of the book in its entirety and, as with the first edition, gave countless hours to read and improve the text. We are also grateful to the numerous reviewers who carefully evaluated individual chapters and provided invaluable comments: Brian Cade (USGS), Michael Chimney (South Florida Water Management District), James Durant (Agency for Toxic Substances and Disease Registry), William Farmer (USGS), Gregory Granato (USGS), Brian Gray (USGS), Tara Gross (USGS), Margaret Guyette (St. Johns River Water Management District of Florida), Julie Kiang (USGS), Kelsey Kolars (USGS), Jennifer Kostrzewski (Metropolitan Council of Twin Cities), Sally Letsinger (Indiana University-Bloomington), Dendy Lofton (LimnoTech), Graham McBride (National Institute of Water and Atmospheric Research, New Zealand), Doug McLaughlin (National Council for Air and Stream Improvement), Wesley Newton (USGS), Tom Nolan (USGS), Thomas Over (USGS), Valerie Partridge (Washington State Department of Ecology), Matt Pocernich (Oracle Data Cloud), Nick Procopio (New Jersey Department of Environmental Protection), Emily Read (USGS), Steven Saiz (Central Coast Regional Water Quality Control Board of California), Mark Sandstrom (USGS), Keith Sawicz (AIR Worldwide), Lori Sprague (USGS), Paul Stackelberg

(USGS), Michael Tomlinson (University of Hawai‘i at Mānoa), Gregg Wiche (USGS), Aldo (Skip) Vecchia (USGS), and Helen Yu (San Diego Regional Water Resources Control Board of California). We also extend our gratitude to Ian Willibek-Lemair (Virginia Polytechnic Institute and State University), who provided essential support in managing the many in-text citations and bibliographic records and for converting the text and many equations from the first edition into modern word-processing formats, and to Benjamin York (USGS), for his help in standardizing the R code and for his review of the supporting information.

We are indebted to the many hydrologists and hydrologic technicians of the USGS and from other institutions who have created the data that these methods were designed to analyze, and to the talented software engineers who have curated many of the datasets we use. We are indebted to the many students we have taught in our various courses and those who have contacted us over the years asking interesting questions about how to evaluate new and different types of datasets. Their questions help to keep us on our toes and continuing to seek new and better approaches to data analysis. As always, the responsibility for all errors is ours alone.

# Chapter 1

## Summarizing Univariate Data

---

A dataset is a collection of measurements that are used to learn about a population. For example, the population might be the sodium concentration of all of the water in an aquifer, or the instantaneous discharge values for river water passing a streamgage over a 50-year period, or the number of a particular species of invertebrates on a streambed over a reach of river that is 10 kilometers in length. In each case there is a population, which is very large, and only a part of that population is available to us. Our data are the measurements that we take—the sample—and we use those data to try to characterize the overall population. In this chapter we will only consider the univariate problem. That is, we just have one random variable we are trying to characterize. In later chapters we will extend this topic to multiple variables with the goal of characterizing how they vary in relation to each other. The choice of statistical methods to be used to characterize a population based on the data we have in hand should be built on what we know about the characteristics of the population. This statement involves the scientific process of model building using both **inductive** (developing broad generalizations from specific examples) and **deductive** (deriving conclusions from a general statement or hypothesis) reasoning. To select the right statistical method we need to know something about the characteristics of the data. A goal of this chapter is to use common experiences with hydrologic datasets to point to tools that are likely to work relatively well for certain types of hydrologic data. We may read in the statistical literature about the optimality of a specific approach to identifying characteristics of a population, but that optimality depends on an assumption that the population has certain characteristics (things like normality and independence). Little is gained by employing analysis procedures that assume the data conform to certain assumptions about their characteristics, when, 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. Knowing these basic characteristics of the data is crucial to selecting appropriate data analysis procedures.

One of the most frequent tasks when analyzing data is to describe and summarize those data in forms that convey their important characteristics. “What is the sulfate concentration one might expect at this location?” “How variable is hydraulic conductivity?” “What is the size of the flood that has an annual probability of 1/100 (often called the 100-year flood)?” Estimation of these and similar summary statistics are fundamental to understanding the population we are interested in. Characteristics often described include a measure of the center of the population, a measure of the variability of the population, a measure of the symmetry of the probability distribution around its center, and perhaps estimates of extreme quantiles of the population such as the 10-year low flow or 100-year flood. This chapter discusses methods for summarizing a univariate dataset for the purposes of shedding light on the characteristics of the population from which it was sampled.

This first chapter also demonstrates one of the major themes of the book—the use of robust statistical techniques. A robust technique is one that works reasonably well over a wide range of situations (for example, populations that have different probability distributions), in contrast to a technique that might be optimal for some particular situation (for example, a normally distributed population) but works poorly in other situations that one might encounter in practice. The reasons why one might prefer to use a robust measure, such as the median, as opposed to a more classical measure, such as the mean, are explained.

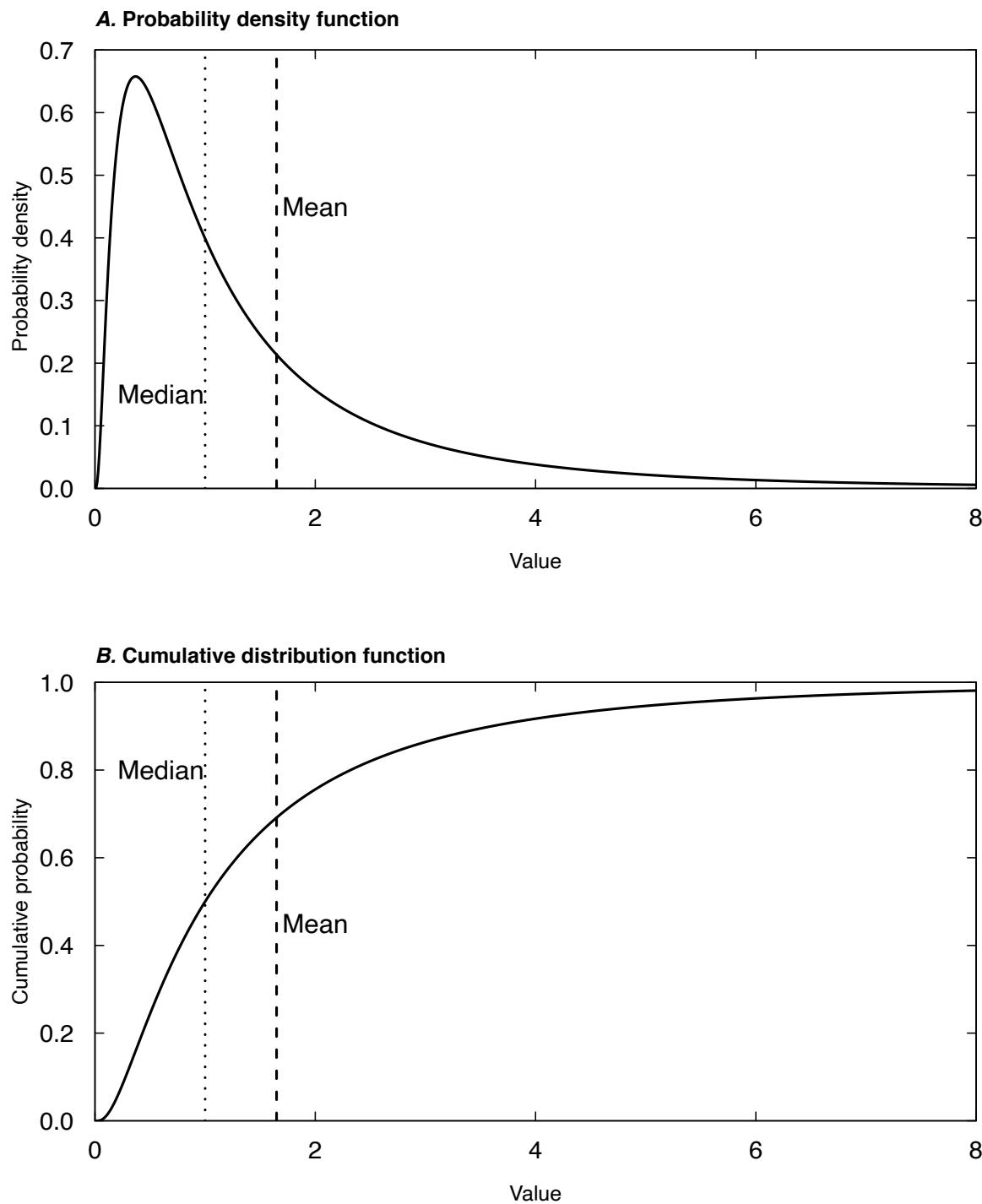
In most cases, the hydrologist is given a finite sample (the dataset) to characterize a target population, which is typically infinite in size. The population might be concentrations in all waters of an aquifer or stream reach, or all streamflows over some time at a particular site. Rarely are all such data available to the scientist. An example of a finite population would be the set of all lakes that exist within a geographical area, but such finite populations are not commonly encountered. In the typical case of an infinite population it would be physically impossible to collect all data of interest (all the water in a stream over the study period), or if the population is finite, but very large, it may be too costly to collect a sample that includes

every member of the population. Typically, a subset of the population, called the sample, is selected and measured in such a way that conclusions about the sample may be extended to make inferences about the characteristics of the population, such as its central tendency, variability, or the shape of the distribution. Measures of central tendency (sometimes called location) are usually the sample mean or sample median. Common measures of spread include the sample standard deviation and sample interquartile range. Use of the word “sample” before each statistic conveys the idea that these are only estimates of the population value. For example, the sample mean is an estimate of the population mean. Generally, we compute statistics based on a sample and not the whole population. For this reason, 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.

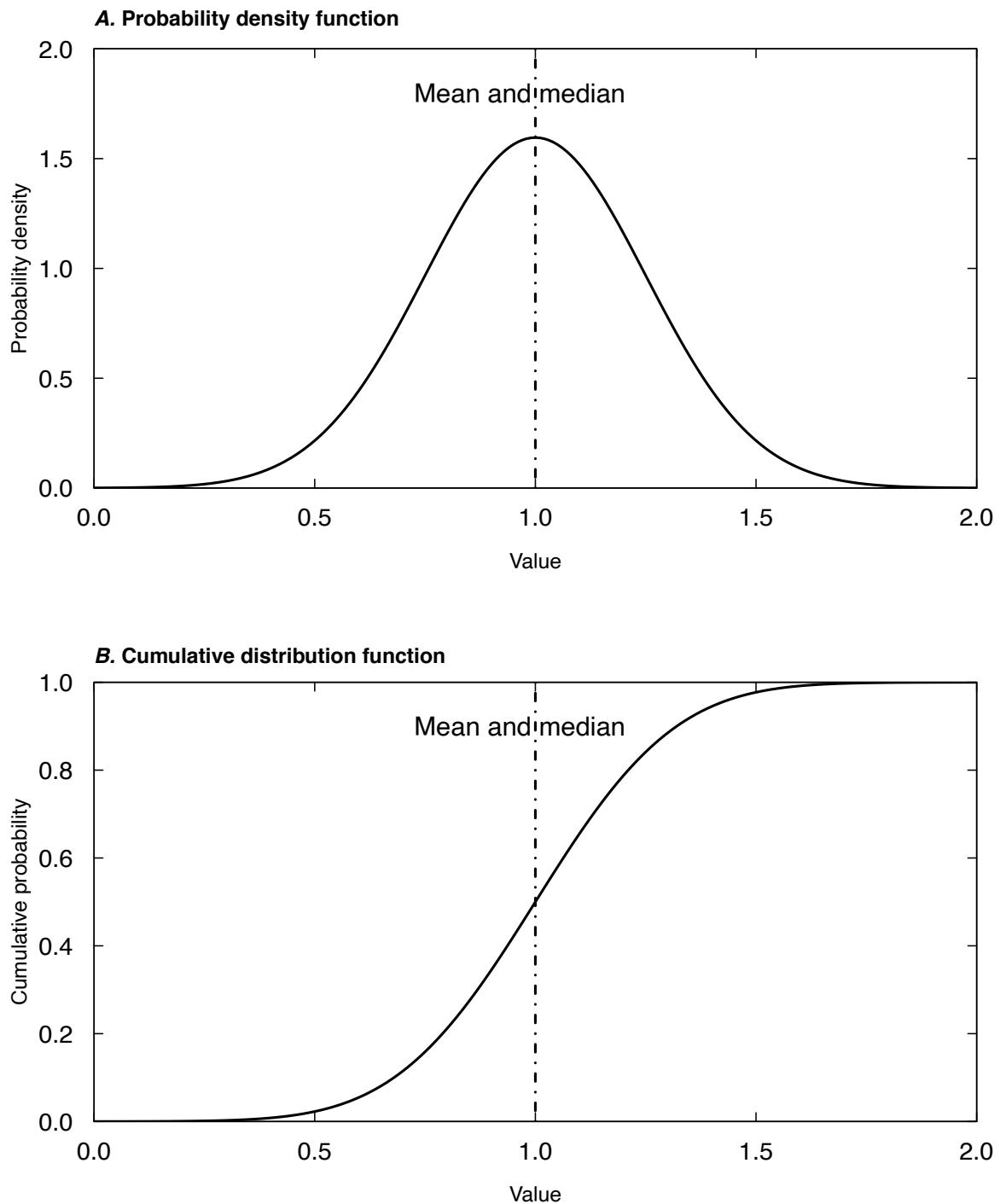
## 1.1 Characteristics of Water Resources Data

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

1. **A lower bound of zero.** Negative values are rarely possible. There are certainly examples of variables that can take on negative values, such as hydraulic heads measured against some datum, temperatures, or flows in situations where flow reversals are possible (for example, backwater from a larger river or from tidal waters), but in most cases hydrologic variables have a lower bound of zero.
2. **The presence of outliers,** observations that are considerably higher or lower than the vast majority of the data. High outliers (such as flood discharges that are vastly larger than typical annual floods) are more common in water resources than low outliers.
3. **Positive skewness,** which is typically a result of the properties listed in points 1 and 2 above. Skewness can be expected when the values that are farthest from the center of the distribution occur primarily on one side of the center rather than on both sides. An example of a positive-skewed distribution, which is often a good representation of the population of some hydrologic variables, is the lognormal distribution. The probability density function (pdf) of a lognormal distribution is shown in figure 1.1A. In a pdf, the horizontal axis covers the values that the random variable might take on. The vertical axis is a measure of the probability that a given observation of that random variable will take on that specific value. Most readers will be familiar with a histogram, which shows the frequency with which the values of a random sample fall into each of several class intervals. A pdf conveys the same kind of information, but does so with class intervals that are infinitely narrow and for which the entire theoretical population is perfectly known. The area under the curve always equals 1, for any pdf. A cumulative distribution function (cdf) of this same distribution is shown in figure 1.1B. In a cdf, the vertical axis is a measure of the probability that a given observation of that random variable will be less than or equal to that specific value. Thus, the vertical axis is bounded by zero and one. The cdf is the integral of the pdf (or conversely the pdf is the first derivative of the cdf).
4. **Non-normal distribution of data.** The three points mentioned above (lower bound of zero, outliers, and positive skewness) constitute one possible set of causes for data to be non-normal. The pdf of a normal distribution and the cdf of the same distribution are shown in figure 1.2. Many classical statistical methods assume that the population follows a normal distribution and, although in many fields of science the assumption of normality is often a very defensible assumption, normality in hydrologic data may be more the exception than the rule. Even in cases where the pdf is symmetric, the normal assumption may be poor, because extreme values can be more common than one would expect from a normal distribution. Such distributions are often called heavy tailed distributions.
5. **Data reported only as below or above some threshold.** In statistics these are known as censored data. Examples include concentrations of a chemical or particles that are reported as being below a laboratory reporting limit (for example, arsenic concentration in this sample is <0.001 milligrams per liter [mg/L]), annual flood discharges that are known only to be lower than a level that would have caused the creation of a long-lasting public record of the flood (for example, the annual flood of 1888 is <20,000 cubic meters per second [ $m^3/s$ ]), and hydraulic heads that are known to have been above the land surface at some point in time because they are shown as flowing artesian wells on old maps (for example, head >800 meters above mean sea level in 1910).



**Figure 1.1.** Graphs showing a lognormal distribution. *A*, the probability density function (pdf) of a lognormal distribution, showing the location of the population mean and median. *B*, the cumulative distribution function (cdf) of the same lognormal distribution.



**Figure 1.2.** Graphs showing a normal distribution. *A*, the probability density function (pdf) of a normal distribution showing the mean and median of the distribution, which are identical. *B*, the cumulative distribution function (cdf) of the same distribution.

6. **Seasonal patterns.** Values tend to be higher or lower in certain seasons of the year. If these regular seasonal patterns are not considered in the analysis, this part of the variation is likely to be viewed as random noise, even though it is highly predictable.
7. **Autocorrelation.** Observations tend to be strongly correlated with other observations that are nearby. In the case of time series data the term “nearby” refers to close in time, which means that high values tend to follow high values and low values tend to follow low values. In the case of time series, this autocorrelation is also known as serial correlation. In the case of spatial data (for example, water levels or chemical concentrations in a set of wells) the term “nearby” indicates geographic proximity. This is the tendency for high values to be near other high values and low values to be near other low values. One of the consequences of autocorrelation is that the accuracy of any statistical estimates will be overstated if this property is ignored. For example, 100 observations of a random variable should provide a fairly accurate estimate of the population mean, but if the samples collected were spaced very close in time and the serial correlation was strong, the accuracy of the estimate of the population mean may be no better than what could be derived from a set of only 10 uncorrelated observations from that population.
8. **Dependence on other variables.** For example, the probability distribution of chemical concentrations in a stream can change dramatically with water discharge, or the distribution of hydraulic conductivity can be very different for lithologies with different particle size distribution. Failure to recognize and deal with the dependencies can greatly diminish the ability to describe and understand the variation in the variable of interest.

Methods for analysis of water resources data, whether the simple summarization techniques mentioned in this chapter or the more complex procedures of later chapters, should be selected based on consideration of these common properties. Failure to consider them can lead to mistaken conclusions or can result in analyses that are relatively ineffective at extracting accurate or meaningful inferences from the data.

## 1.2 Measures of Central Tendency

The mean and median are the two most commonly used measures of central tendency (sometimes known as 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 A Classical Measure of Central Tendency—The Arithmetic Mean

The arithmetic mean ( $\bar{X}$ ), here referred to simply as the mean, is computed as the sum of all data values  $X_i$ , divided by the sample size  $n$ :

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n} . \quad (1.1)$$

For data that 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:

$$\bar{X} = \sum_{i=1}^k \bar{X}_i \frac{n_i}{n} , \quad (1.2)$$

where  $\bar{X}_i$  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

$$\bar{X} = \bar{X}_{(j)} \frac{(n-1)}{n} + X_j \cdot \frac{1}{n} = \bar{X}_{(j)} + \left( X_j - \bar{X}_{(j)} \right) \cdot \frac{1}{n} , \quad (1.3)$$

where  $\bar{X}_{(j)}$  is the mean of all observations excluding  $X_j$ . Each observation's influence on the overall mean  $\bar{X}$  is  $(X_j - \bar{X}_{(j)}) \cdot \frac{1}{n}$ , the distance between the observation and the mean, excluding that observation divided by the sample size  $n$ . 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  $\bar{X}$  than does a more typical observation, one closer to its  $\bar{X}_{(j)}$ .

Another way of illustrating this influence is to imagine that the mean is the balance point of the data, when each point is stacked on a number line (fig. 1.3) and is assumed to have equal weight. 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 dataset in balance, but if one outlying value were removed, the balance point would shift dramatically (fig. 1.4). The mean is a summary statistic that is not resistant to changes in the presence of, or to changes in the magnitudes of, a few outlying observations. Thus, we may want to use other measures of central tendency that are more resistant to the influence of outlying observations. It may be the case that we truly want to use the mean as our measure of central tendency, because we are interested in a variable that is going to be expressed as a sum. An example is the case where we want to know the mean of the flux of some material (for example, a nutrient or suspended sediment) into a receiving water body. In this case, we truly want to know the mean value of the flux. In contrast, where we are looking to characterize typical values of some variable, we may want to consider other more resistant statistics for the central tendency. The median (discussed in section 1.2.2.) is a great example of a resistant estimator of central tendency; another is the mode (section 1.2.3.).

## 1.2.2 A Resistant Measure of Central Tendency—The Median

The median, or 50th percentile ( $P_{0.50}$ ), is the central value of the distribution when the data are sorted by magnitude. For an odd number of observations, the median is the data point that has an equal number of observations both above and below it. For an even number of observations, it is the arithmetic mean of the two central-most observations. To compute the median, first sort the observations from smallest to largest, so that  $X(1)$  is the smallest observation and  $X(n)$  is the largest observation. Then

$$\text{median} = P_{0.50} = \begin{cases} X\left(\frac{n+1}{2}\right) & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X\left(\frac{n}{2}\right) + X\left(\frac{n}{2}+1\right)\right) & \text{when } n \text{ is even} \end{cases}. \quad (1.4)$$

The median is only minimally affected by the magnitude of any single observation. 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 dataset (a) of 7 observations was multiplied by 10 to obtain dataset (b):

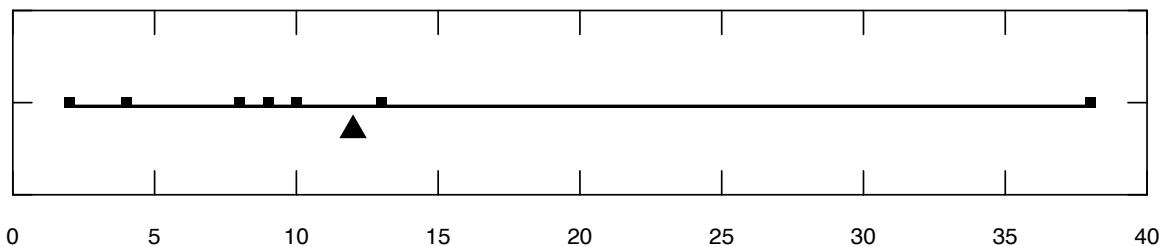
### Example 1.1. Resistance of the mean and median

Dataset (a) 2 4 8 9 11 11 12     $\bar{X} = 8.1$      $P_{50} = 9$

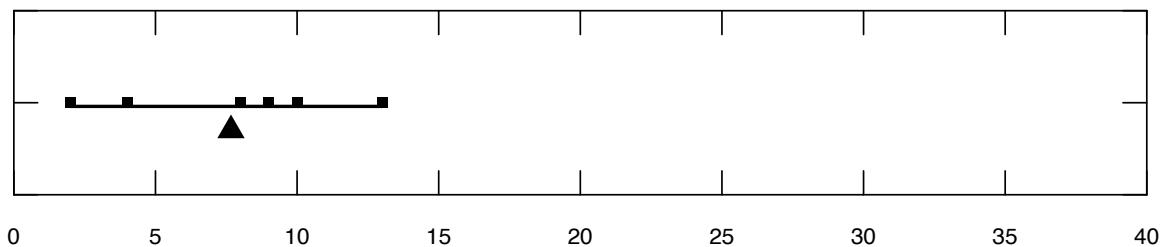
Dataset (b) 2 4 8 9 11 11 120     $\bar{X} = 23.6$      $P_{50} = 9$

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

When a summary value of the central tendency is desired that is not strongly influenced by a few extreme observations, the median is preferable to the arithmetic 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; this would not be the case for the median.



**Figure 1.3.** Graph showing the arithmetic mean (triangle) as the balance point of a dataset. The mean is 12.



**Figure 1.4.** Graph showing the shift of the arithmetic mean (triangle) downward after removal of an outlier. The mean is 7.67.

### 1.2.3 Other Measures of Central Tendency

Three other measures of central tendency are less frequently used: the mode, the geometric mean, and the trimmed mean. The mode is defined as the most frequently observed value. It is more applicable with discrete data (where the only possible data values are integers). It can be computed for continuous data, but the user must define a bin size to sort the data into. For example, the bins might be values from 0.5 to 1.499, 1.5 to 2.499, and so forth. Another example might be values from 0 to 9.99, 10 to 19.99, and so on. It is very easy to obtain, but a poor measure of location for continuous data because its value depends on the definition of the bins.

The geometric mean ( $GM$ ) is often reported for positively skewed datasets. It is only defined in cases where all data values are positive. By definition, it is the  $n$ th root of the product of the  $n$  values in the sample.

$$GM = \sqrt[n]{X_1 \cdot X_2 \cdot \dots \cdot X_n} \quad (1.5)$$

A simple way to calculate it is to take the mean of the logarithms of the data and then transform that value back to the original units.

$$GM = \exp(\bar{Y}), \quad (1.6)$$

where

$$Y_i = \ln(X_i); \text{ and}$$

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}.$$

Note, in this book the natural (base e) logarithm will be abbreviated `ln`, and its inverse  $e^x$  will be abbreviated `exp(x)`. For positively skewed data, the geometric mean is usually quite close to the median (not so for negatively skewed data). 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 arithmetic mean logarithms are equal, as in figure 1.2. When transformed back to original units, the geometric mean continues to have half the observations below it and half above, and so it is located at the median and is lower than the arithmetic mean (fig. 1.1). The geometric mean will always be lower than or equal to the arithmetic mean. This point becomes important in later chapters.

A compromise between the median and mean is the trimmed mean, which is the arithmetic mean computed after trimming off equal numbers of the lowest and highest observations. 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. A 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 referred to the trimmed mean, but is more precisely the 25-percent trimmed mean. A 0-percent trimmed mean is the arithmetic mean itself, and a 50-percent trimmed mean is the same as the median. Percentages of trimming should be explicitly stated when used. The trimmed mean is a resistant estimator of central tendency, as it is not strongly influenced by outliers. 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.

In R, the function for determining the mean is `mean`; for the median, the function is `median`. The trimmed mean can be computed with the function `mean`, using the `trim` argument. For example, a trimmed mean with 25 percent trimming on each side (or 25-percent trimmed mean) would be computed as `mean(x, trim = 0.25)`. The computation of the geometric mean is given in the solutions to exercise 1 at the end of this chapter. The mode can be computed in a two-step process.

```
> y <- table(as.vector(x))
> modeX <- as.numeric(names(y) [y == max(y)])
```

Take note here that there may be more than one value returned as the mode and also note that calculation of the mode depends on the extent to which the values are rounded.

## 1.3 Measures of Variability

It is just as important to know how variable the data are as it is to know their central tendency or location. Variability in the data is often called spread, and there are several measures of variability that can be used.

### 1.3.1 Classical Measures of Variability

The sample variance and its square root, the sample standard deviation, are the classical measures of variability. Like the mean, they are strongly influenced by extreme values. The sample variance is denoted as  $s^2$ ; its square root, denoted  $s$ , is the sample standard deviation.

$$s^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{(n-1)} \quad (1.7)$$

The values are computed using the squares of deviations of data from the sample mean, so that extreme values influence their magnitudes even more so than for the mean. When extreme values are present, these measures are unstable and inflated. They may give the impression of much greater variability than is indicated by the majority of the dataset. In R, the standard deviation may be computed with the function `sd` and the variance can be computed as the square of the value resulting from `sd` or simply by using the function `var`.

### 1.3.2 Resistant Measures of Variability

The interquartile range (IQR) is the most commonly used resistant measure of variability. 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.

The IQR is defined as the 75th percentile minus the 25th percentile. The 75th, 50th (median), and 25th percentiles split the data into equal-sized quarters. The 75th percentile ( $P_{0.75}$ ), also called the upper quartile, is a value that exceeds no more than 75 percent of the data and is therefore exceeded by no more than 25 percent of the data. The 25th percentile ( $P_{0.25}$ ), or lower quartile, is a value that exceeds no more than 25 percent of the data and is therefore exceeded by no more than 75 percent. Consider a dataset ordered from smallest to largest:  $X_i$ ,  $i=1, 2, \dots, n$ . Percentiles ( $P_j$ ) are computed using equation 1.8

$$P_j = X_{(n+1) \cdot j} , \quad (1.8)$$

where  $n$  is the sample size of  $X$ , and  $j$  is the fraction of data less than or equal to the percentile value (for the 25th, 50th, and 75th percentiles,  $j=0.25, 0.50$ , and  $0.75$ , respectively).

For the datasets used in example 1.1,  $n=7$ , and therefore the 25th percentile is  $X_{(7+1)\cdot 0.25}$  or  $X_2=4$ , the second lowest observation. The 75th percentile is  $X_6$ , the sixth lowest observation, or 11. The IQR is therefore  $11-4=7$ . When values of  $(n+1) \cdot j$  are not integers, then some type of interpolation method is needed to compute the percentiles. There are several different ways to do this computation. The preference for this book is to use the R function `quantile`. This function allows the user to specify the type of interpolation. There are nine possible types of interpolation available in R for this function. With large sample sizes (greater than about 100 observations), the choice of type is of very little consequence. The choice preferred here, `type = 6`, is commonly known as the Weibull plotting position, which has a long history in hydrology (Weibull, 1939). In hydrology the term “plotting position” comes from the rules for constructing an empirical cumulative distribution function, which plots the individual observations in the sample (in ranked order) against an estimated probability of nonexceedance. That estimated probability is called the plotting position. Hydrologists have historically also used the Hazen plotting position (which corresponds to `type = 5`) (Hazen, 1914) and the Blom plotting position (Blom, 1958), (`type = 9`), which are both used in flood frequency analysis and in distribution fitting methods (as discussed in chap. 4). For an extensive discussion of these algorithms, see Hyndman and Fan (1996), which explains each `type` using the same system for the numbering for the choices of `type` that is used in R. Using our preferred choice, `type = 6`, the R commands for computing the IQR of a dataset (with the data stored as a vector called `x`) are the following:

```
> quant <- as.numeric(quantile(x, type = 6))
> IQR <- quant[4] - quant[2]
```

Note that the default for the `quantile` command is that it returns a set of five values representing, in this order, the minimum, lower quartile, median, upper quartile, and maximum. One of the arguments to the `quantile` function is `probs`, which can be either a scalar or vector for the probabilities we wish to estimate. When the argument `probs` is set equal to some other sequence of values (which are in the range of 0 to 1), then the function returns a set of values for each specified probability value. An alternative way to compute the IQR would be in a single line:

```
> IQR <- quantile(x, probs = 0.75, type = 6) - quantile(x, probs =
0.25, type = 6)
```

There is a standard R function for the IQR (which is called `IQR`). When the default values of the function are used, the function will return a different value than what is defined above because the default is `type = 7`, known as the Gumbel plotting position. See section 2.1.2. for a discussion of these choices. However, it can return exactly the value of the IQR as defined here by calling it in this manner:

```
> IQR(x, type = 6)
```

In most cases, the difference between the results will be quite small if the sample size is larger than about 100 observations.

Another resistant estimator of variability is the median absolute deviation, or MAD. The MAD is computed by first computing the median of the dataset and then computing the absolute value of all differences,  $|d_i|$ , between each observation ( $X_i$ ) and the median. The MAD is the median of these absolute differences.

$$MAD(X) = \text{median} |d_i| , \quad (1.9)$$

where

$$d_i = X_i - \text{median}(X) .$$

In R, the function `MAD` computes the median absolute deviation of a dataset.

We now compare the estimates of spread for the datasets introduced in example 1.1. First we will compute the three measures of spread (IQR, MAD, and standard deviation) with dataset (a). The code and output are shown here:

```
> x <- c(2, 4, 8, 9, 11, 11, 12)
> quant <- as.numeric(quantile(x, type = 6))
> iqR <- quant[4] - quant[2]
> xbar <- mean(x)
> SD <- sqrt(sum((x - xbar)^2) / (length(x) - 1))
> med <- median(x)
> MAD <- median(abs(x - med))
> cat("IQR =", iqR, ", MAD =", MAD, ", Standard Deviation =", SD)
IQR = 7 , MAD = 2 , Standard Deviation = 3.804759
```

Now we will change the last value in the dataset from 12 to 120 to form dataset (b) and run it again.

```
> x <- c(2, 4, 8, 9, 11, 11, 120)
> quant <- as.numeric(quantile(x, type = 6))
> iqR <- quant[4] - quant[2]
> xbar <- mean(x)
> SD <- sqrt(sum((x - xbar)^2) / (length(x) - 1))
> med <- median(x)
> MAD <- median(abs(x - med))
> cat("IQR =", iqR, ", MAD =", MAD, ", Standard Deviation =", SD)
IQR = 7 , MAD = 2 , Standard Deviation = 42.65699
```

The results show that when we change the last value in the dataset from 12 to 120, the IQR and MAD do not change at all, but the standard deviation increases from 3.8 to 42.7. This demonstrates that both the IQR and MAD are resistant to outliers but the standard deviation is highly sensitive to them, suggesting that the IQR or MAD might be more reliable measures of spread. Note that the value returned here for IQR is different from the value one would get from the `IQR` function and the MAD value is also different from the value one would get from the `MAD` function in R.

### 1.3.3 The Coefficient of Variation—A Nondimensional Measure of Variability

One more measure of variability is the coefficient of variation (CV). The sample CV is defined as the standard deviation of the sample divided by the mean of the sample. It is dimensionless and it can be very useful in characterizing the degree of variability in datasets. For example, when comparing the distributions of some random variable related to streamflow or stream transport rates we might expect to

find that mean flow or transport increases with the size of the drainage area. However, we can obtain a more meaningful perspective by comparing the CV of flow or of transport across multiple sites in order to gain an understanding of the way that variability changes across multiple stream sites in a region.

## 1.4 Measures of Distribution Symmetry

Hydrologic data are typically skewed, which means that their distribution is not symmetric around the mean or median, with extreme values extending out farther in one direction than the other. The density function for a lognormal distribution, shown in figure 1.1, illustrates an asymmetric distribution; in this case, one with positive 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 arithmetic mean is not expected to equal the median. Rather, the arithmetic mean is on the side of the median with the longer tail of the distribution. Thus, when data have positive skewness, the mean typically exceeds more than 50 percent of the data (the mean is larger than the median), as in figure 1.1. The standard deviation is also inflated by the long tail. Therefore, tables of summary statistics including only the mean and standard deviation (or variance) are of questionable value for water resources data. These types of data often have considerable skewness and, thus, the mean and standard deviation reported may not describe the data very well, as both will be inflated by outlying observations. Summary tables that include the median and other percentiles have far greater applicability to skewed data. Skewed data also call into question the applicability of many hypothesis tests (discussed in chap. 4) that are based on assumptions that the data follow 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 symmetric. Later chapters will discuss this in detail and suggest several solutions.

### 1.4.1 A Classical Measure of Symmetry—The Coefficient of Skewness

The coefficient of skewness ( $g$ ) is the skewness measure used most often. In statistical terms, it is the centralized third moment (moment is generally defined as a sum of the data values raised to some specified power) divided by the cube of the standard deviation (where the mean and variance are the first and second moments, respectively):

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \frac{(X_i - \bar{X})^3}{s^3} . \quad (1.10)$$

A right-skewed distribution has a positive  $g$ ; a left-skewed distribution has a 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 datasets introduced in example 1.1, the skewness coefficient increases from  $-0.84$  to  $2.6$  when the last data point is changed from  $12$  to  $120$ . Extensive Monte Carlo testing has been used to explore the accuracy of sample skewness coefficients (Wallis and others, 1974) and it has shown that with sample sizes typical in hydrology, often less than  $100$  observations, the skewness coefficient can be highly biased; this means that the expected value of the sample statistic is much smaller, in absolute value, than the true value of the statistic, which we can call the population skewness, and has a great deal of sampling variability. Kirby (1974a) showed that the skewness coefficient has an algebraic bound, meaning that for any given sample size, the absolute value of the sample skewness coefficient has a finite upper bound, and this bound may be less than the absolute value of the true population coefficient of skewness. Only when the population value of skewness is zero is the sample coefficient of skewness an unbiased estimate. An alternative less biased approach to describing skewness that is more resistant to outliers is the L-moment approach developed by Hosking (1990). Discussions of this method are beyond the scope of this book. The important point is that unless sample sizes are large (well above  $100$  samples), skewness coefficients computed using equation (1.10) are not very informative except to the extent that they may distinguish between right-skewed and left-skewed populations.

### 1.4.2 A Resistant Measure of Symmetry—The Quartile Skew

A resistant measure of symmetry is the quartile skew  $qs$  (Kenney and Keeping, 1954):

$$qs = \frac{(P_{0.75} - P_{0.50}) - (P_{0.50} - P_{0.25})}{P_{0.75} - P_{0.25}} \quad (1.11)$$

defined as the difference in distances of the upper and lower quartiles from the median, divided by the IQR. A right-skewed distribution again has a positive  $qs$ ; a left-skewed distribution has a negative  $qs$ . Similar to the 25-percent trimmed mean and IQR,  $qs$  uses the central 50 percent of the data. For the example dataset,  $qs = (11 - 9) - (9 - 4) / (11 - 4) = -0.43$  for both datasets (a) and (b). Note that this resistance may be a liability if sensitivity to a few observations is important.

In R, the  $qs$  value can be computed from the data (here stored in a vector called  $x$ ), with the following two lines of code.

```
> pvals <- as.numeric(quantile(x, probs=c(0.25, 0.5, 0.75), type=6))
> qs <- ((pvals[3]-pvals[2]) - (pvals[2]-pvals[1])) /
  (pvals[3]-pvals[1])
```

## 1.5 Other Resistant Measures of Symmetry

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:

$$qs_{0.10} = \frac{(P_{0.90} - P_{0.50}) - (P_{0.50} - P_{0.10})}{P_{0.90} - P_{0.10}} \quad (1.12)$$

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). Measures based on quartiles have generally become 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 chap. 2) and are familiar to most data analysts.

## 1.6 Outliers

Outliers, observations whose values are quite different than others in the dataset, often cause great interest or alarm. Their presence raises the questions (1) Did they arise as a result of some error (for example, instrument malfunction or data entry error) or (2) do they represent a reasonably accurate observation of an unusual situation? Outliers are often dealt with by removing them prior to describing data or prior to applying some of the hypothesis test procedures discussed in later chapters. One of the goals of this book is to present methods that are relatively resistant to the influence of outliers so that there is no need to delete them from the dataset in order to conduct a meaningful analysis. Outliers may be the most important points in the dataset, and should be investigated further. If outliers are deleted, it creates the risk that those who use the dataset will only see what they expected to see and may miss gaining important new information. Outliers typically have one of these three causes:

1. A measurement or recording error;
2. An observation from a different population than most of the data, such as a flood caused by a dam break rather than by precipitation or a concentration resulting from a brief chemical spill into a river; or
3. A rare event from a single population; for example, if floods are always caused by rainfall events, the outlier may arise simply because the rainfall was extreme.

The graphical methods presented in chapter 2 are very helpful for identifying outliers. Whenever outliers occur, the analyst should first verify that the value recorded is not simply an error in locating the decimal point or some other kind of transcription error. If this type of error is found then the value should be corrected, if possible. The effort that should be invested in verification, such as rerunning the sample in the laboratory, will depend on the benefit gained versus the cost of verification. It may not be possible to duplicate past events. 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 fit nicely to a preconceived theoretical distribution. There is no reason to suppose that they should! The entire dataset 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, outlier-resistant methods should be employed. If computing a mean appears of little value because of an outlier, the median is 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 dataset, 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. Sensitivity studies based on multiple alternative values and (or) explanations of the most extreme values in the dataset may be options for datasets with extreme values.

## 1.7 Transformations

There are three common reasons to consider transformations of the data (and often more than one of them are involved):

1. To make data distributions more symmetric,
2. To make relations between variables more linear, and
3. To make variability more constant.

Many effective statistical methods (for example, linear regression or analysis of variance) are only appropriate when the data (and in some cases, model errors) follow a symmetric distribution, relations among variables are linear, and errors are homoscedastic (have a constant variance over the range of predicted values). Transformations of the original data can sometimes produce these characteristics even when the original data do not possess these qualities. Thus the use of transformed variables enables the analyst to use a set of useful statistical tools that might not be appropriate if the original data were not transformed. However, using a transformation requires some special considerations in the interpretation of results (for example, retransformation bias correction, which is discussed in chap. 9). Selection of an appropriate transformation is not an arbitrary choice but needs to be guided by the data and by some theoretical considerations.

Transformations can help to create a variable that has better statistical properties than the original measured variable. For example, the negative logarithm of hydrogen ion concentration, pH, is as valid a measurement system of hydrogen ion concentration itself and tends to produce a nearly symmetrical, rather than skewed, distribution. 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 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 transformation of frequency. Each time a piano is played, a logarithmic transform is employed! Similarly, the Richter scale for earthquakes, graphs of long-term price variations of stock market indices, and f-stops for camera exposures 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. This is discussed more in the chapters on regression and trend analysis (chaps. 9, 11, and 12). We must also remember that the use of many resistant techniques such as percentiles and nonparametric test procedures (to be discussed later) are invariant to measurement scale.

### 1.7.1 The Ladder of Powers

Transforming or re-expressing data in new units is a common approach to making asymmetric distributions more symmetric. These new units alter the distances between observations on a line plot. The effect is to either expand or contract these distances between extreme observations on one side of the median to make it look more like the behavior on the other side. The most commonly used transformation in water resources is the logarithm. Statistical analyses are performed on the logarithms of water discharge, hydraulic conductivity, or concentration rather than on the raw data values.

Most other transformations usually involve power functions of the form  $y=x^\theta$ , where  $x$  is the untransformed data;  $y$ , the transformed data; and  $\theta$ , the power exponent. In table 1.1, the values of  $\theta$  are listed in the ladder of powers introduced by Velleman and Hoaglin (1981), a useful structure for determining a proper value of  $\theta$ .

As can be seen from the ladder of powers, any transformations with  $\theta$  less than 1 may be used to make right-skewed data more symmetric. Constructing a boxplot or Q-Q plot (see chap. 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  $\theta$  closer to 1 should be employed instead, such as a square-root ( $\theta=1/2$ ) or cube-root ( $\theta=1/3$ ) transformation. Transformations with  $\theta>1$  will aid in making left-skewed data more symmetric.

The tendency to search for the best transformation should be avoided. For example, when dealing with several similar datasets, it is probably better to find one transformation that works reasonably well for all, rather than using slightly different ones for each. It must be remembered that each dataset 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 with great precision is an approach that is rarely worth the effort.

**Table 1.1.** Ladder of powers as modified from Velleman and Hoaglin (1981).

[ $\theta$ , the power exponent; -, not applicable]

$\theta$	Transformation	Name	Comment
Used for negatively skewed distributions			
$i$	$x^i$	$i$ th power	-
3	$x^3$	Cube	-
2	$x^2$	Square	-
Original units			
1	$x$	Original units	No transformation.
Used for positively skewed distributions			
1/2	$\sqrt{x}$	Square root	Commonly used.
1/3	$\sqrt[3]{x}$	Cube root	Commonly used. Approximates a gamma distribution.
0	$\log(x)$	Logarithm	Very commonly used. Holds the place of $x^0$ .
-1/2	$-1/\sqrt{x}$	Negative square root	The minus sign preserves the order of observations.
-1	$-1/x$	Negative reciprocal	-
-2	$-1/x^2$	Negative squared reciprocal	-
$-i$	$-1/x^i$	Negative $i$ th reciprocal	-

## Exercises

1. Yields in wells penetrating rock units without fractures were measured by Wright (1985), and are given below. Calculate the values for the terms listed below. Then compare these estimates of location. Why do they differ?

- A. Mean
- B. Trimmed mean (at 10-percent trimmed and 20-percent trimmed)
- C. Geometric mean
- D. Median

<b>Unit well yields (in gallons per minute per foot)</b>					
0.001	0.030	0.10	0.003	0.040	0.454
0.007	0.51	0.49	0.020	0.077	1.02

2. For the well yield data of exercise 1, calculate the values for the terms listed below. Then discuss the differences in the values for a through c.

- A. Standard deviation
- B. Interquartile range
- C. Median absolute deviation
- D. Skew and quartile skew

3. Ammonia plus organic nitrogen (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?

<b>Ammonia plus organic nitrogen</b>									
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 near the front of a text on statistical methods. We believe this is very appropriate, as graphs provide crucial information that is difficult to obtain in any other way. For example, figure 2.1 shows four scatterplots, all of which have exactly the same correlation coefficient (a correlation coefficient is a measure of the degree of association between two variables, discussed in detail in chap. 8). 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. Given the capabilities of modern statistical software there is no basis for any hydrologist to say, “I didn’t have time to plot my data.” Plotting the data is an essential step in data analysis.

A good set of graphs is essential for two purposes:

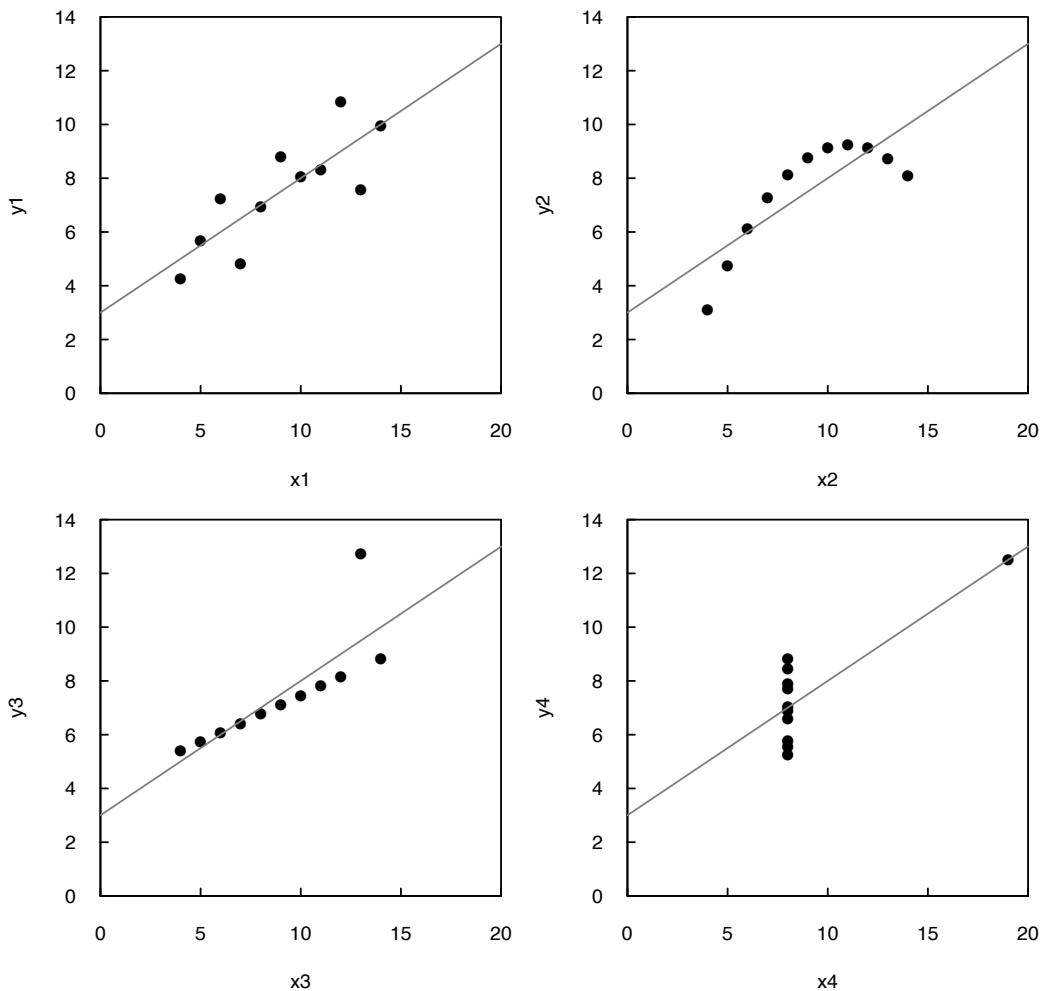
1. To provide the analyst insight into the data under scrutiny, and
2. To illustrate important concepts when presenting the results to others.

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 that have concisely summarized the information for the analyst will also provide insight into the data for the reader or audience. The use of graphics for presentation purposes is the subject of chapter 16. For readers of this text who are interested in following the R scripts provided, we have written the R scripts for our graphics using commands that provide a polished graphic, which should be suitable for presentations. As a result, our commands are somewhat lengthy and complex. In many cases, graphics that are entirely suitable for exploratory data analysis purposes can be done with much simpler versions of the same commands that rely on default values for many arguments. We have generally not shown these simpler versions. We warn the readers that a simple graphic, suitable for quick looks in the EDA process, should not be used in making presentations and for that reason we provide more elaborate scripts here than what might be used for EDA.

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

Two datasets will be used to compare and contrast the effectiveness of each graphical method throughout sections 2.1 and 2.2. These are annual streamflow in cubic meters per second ( $m^3/s$ ) for the James River at Cartersville, Va., for water years 1900–2015, and unit well yields (in gallons per minute per foot of water-bearing material) for valleys without fracturing in Virginia (Wright, 1985).



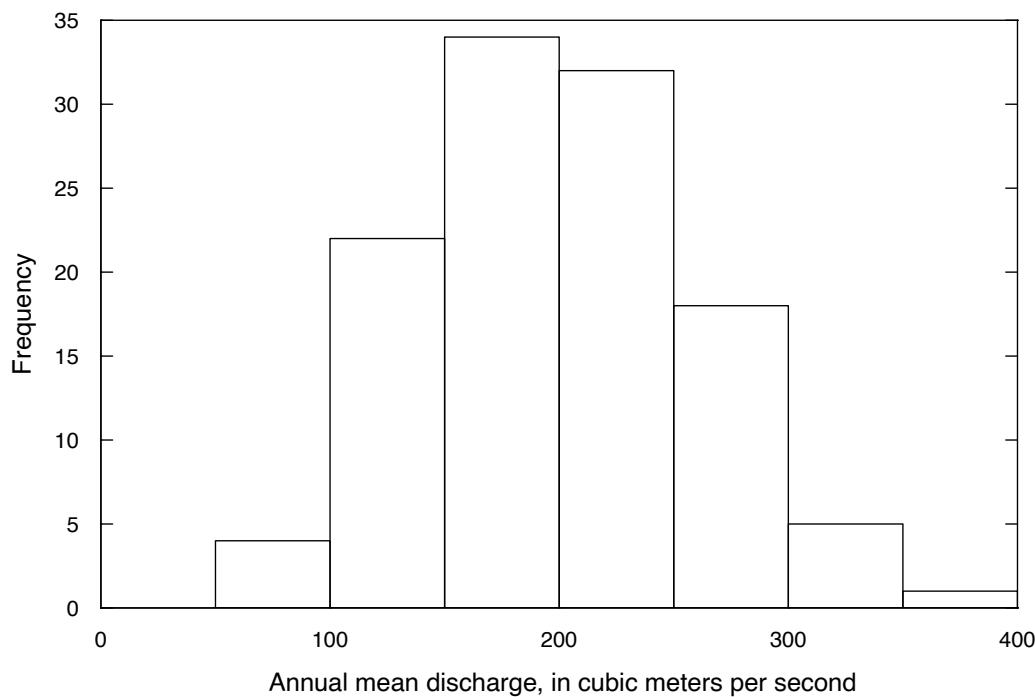
**Figure 2.1.** Four scatterplots of datasets that all have the same traditional statistical properties of mean, variance, correlation, and x-y regression intercept and coefficient. These datasets are known as Anscombe's quartet (Anscombe, 1973) and are available in R.

## 2.1 Graphical Analysis of Single Datasets

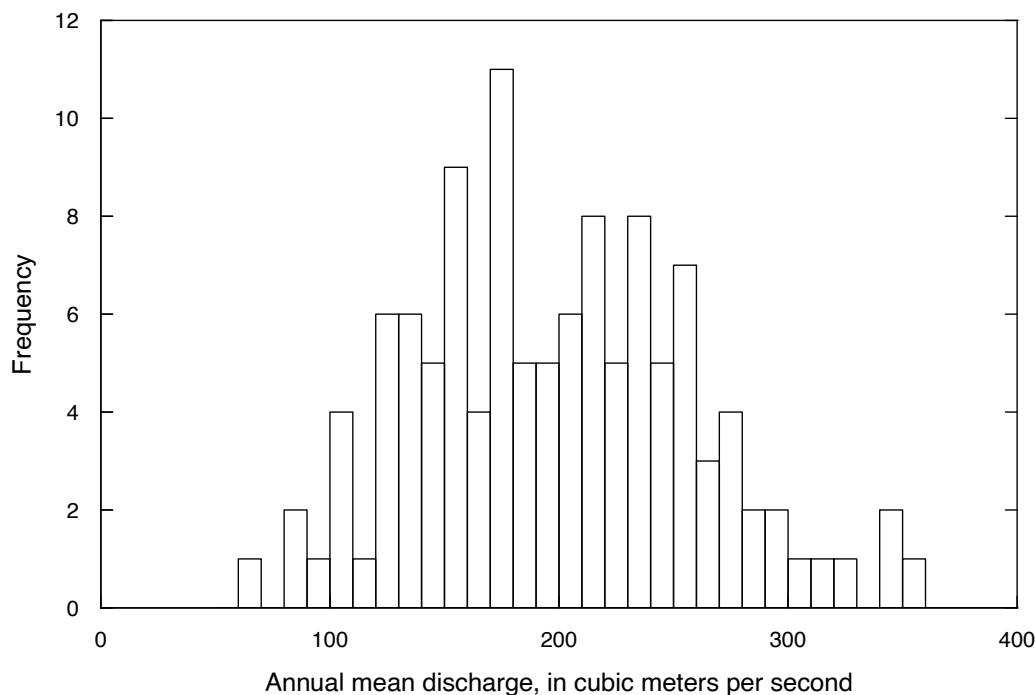
### 2.1.1 Histograms

Histograms are familiar graphics, and their construction is described in numerous introductory texts on statistics. They portray the central tendency, variability, and symmetry of the dataset. If the sample was infinitely large then they would converge to being the probability density function of the population (such as those depicted in figures 1.1 and 1.2). The process of creating a histogram is simple. For a sample of  $n$  values, the data are sorted into a set of categories of equal width and the number of observations falling in each category ( $n_i$ ) is the number in the  $i$ th category. A series of bars are drawn, where the bar height is either,  $n_i$  the number of observations in each category, or  $n_i/n$ , the fraction of data falling into each of the several categories (fig. 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 figures 2.2 and 2.3. Both are histograms of the same data: annual streamflow for the James 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. Figure 2.3 shows much higher frequencies in the 270–280 m<sup>3</sup>/s range as compared to the 260–270 range or 280–300 range, but these differences are simply manifestations of random variations in



**Figure 2.2.** Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015.



**Figure 2.3.** Histogram of annual mean discharge for the James River at Cartersville, Virginia, 1900–2015. Annual streamflow data are the same as shown in figure 2.2, but with different interval divisions.

the number of events in each of these narrow bins. The primary information in the dataset is best seen in figure 2.2, which shows that the central tendency is in the bins from 150–250 m<sup>3</sup>/s, that the distribution is fairly symmetrical, and it ranges no more than about 200 m<sup>3</sup>/s either side of the central values. Figures 2.2 and 2.3 were constructed using the R function `hist`. In the case of figure 2.2, the function set the bin locations and widths automatically. Note that figure 2.2 uses 7 bins and this is quite consistent with the guidance mentioned above ( $2^7 = 128$  which is fairly close to the sample size of 116 values). Figure 2.3 uses 30 bins, which would only be appropriate if the sample size was in the neighborhood of about a billion observations! For preliminary data exploration, simply using the `hist` function with all arguments set to their default values (that would simply be `hist(Q)` where `Q` is the name of the vector of data values) will generally produce an informative histogram. Only when the goal is to produce a histogram suitable for presentation or publication is it necessary to add more specific argument values to the command (which was done to create these figures).

Histograms are quite useful for providing a general impression about the central tendency, variability, and degree of symmetry of the data. They cannot be used for more precise judgments such as depicting individual values. For example, from figure 2.2 we do not know the minimum value in the record, but we do know that it is between 50 and 100 m<sup>3</sup>/s.

For data measured on a continuous scale (for example, streamflow or concentration) histograms are not the best method for graphical analysis, as the process of forcing continuous data into discrete categories may obscure important characteristics of the distribution. However, histograms are excellent when displaying discrete data (for example, 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) and they can be valuable for presentation to audiences that are not accustomed to more complex graphical presentations.

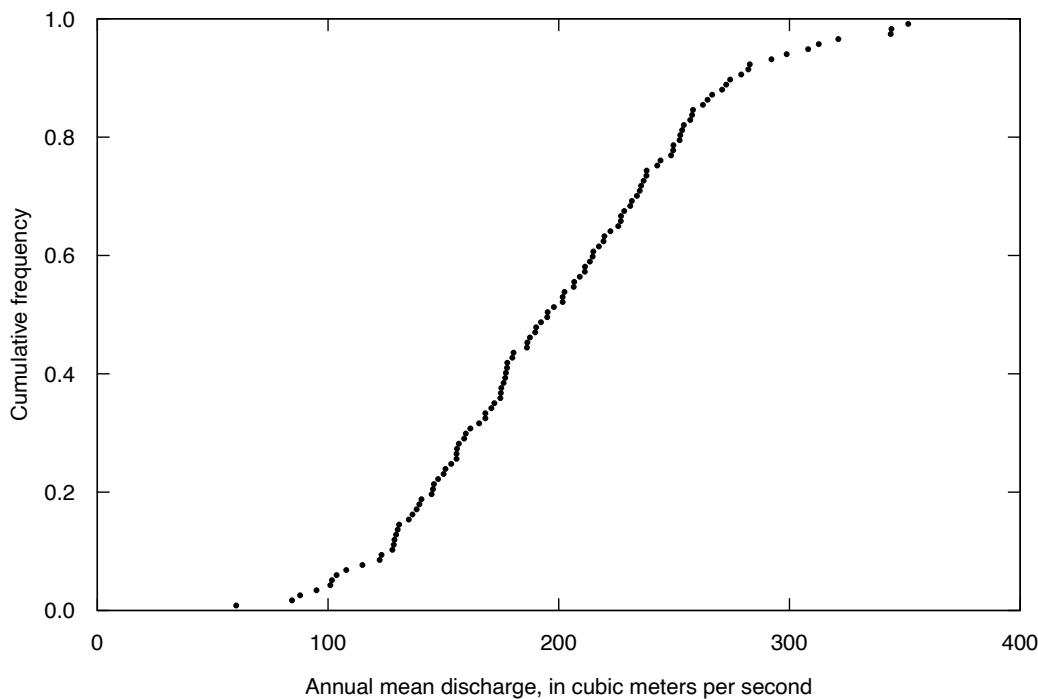
## 2.1.2 Quantile Plots

As discussed in the previous section, histograms are a sample-based approximation of a probability density function (pdf). Another way to display information about a distribution is to use the integral of the probability density function, which is called the cumulative distribution function (cdf). The cdf is a plot of the probability that the random variable will be less than some specific quantity. The vertical scale of a cdf ranges from 0 (for the smallest possible value of the random variable) to 1 (for the largest possible value). Quantile plots are approximations of the cdf based on the sample data; they are often called empirical cumulative distribution functions (ecdf) and visually portray the quantiles, or percentiles (which equal the quantiles multiplied by 100), of the distribution of sample data. Quantiles of importance, such as the median, are easily determined (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 over the alternative methods of portraying a sample such as the histogram and the boxplot (boxplots are described in section 2.1.3.):

1. Arbitrary categories are not required, as they are with histograms.
2. All of the data are displayed, unlike a boxplot (see section 2.1.3.).
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 include the fact that there are three tightly clustered high values (around 350 m<sup>3</sup>/s) but no single high year that can be considered extreme. At the low end of the distribution we see that there is one very low value (around 60 m<sup>3</sup>/s), which is much lower than the second lowest value. The percent of data in the sample less than a given discharge value can be read from the graph with much greater accuracy than from a histogram.

To construct a quantile plot, the data are ranked from smallest to largest. The smallest data value is assigned a rank  $i=1$ , and the largest is assigned a rank  $i=n$ , where  $n$  is the sample size of the dataset. 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 section 1.3.2. of chapter 1, the Weibull plotting position,  $p_i = (i) / (n+1)$ , is generally used in this book. The first and last 5 of the 116 data pairs used in construction of figure 2.4 are listed in table 2.1. When tied data values are present, each is assigned a separate plotting position (the plotting positions are not averaged); therefore, tied values are portrayed as a vertical cliff on the plot.



**Figure 2.4.** Quantile plot of annual mean discharge data from the James River, Virginia, 1900–2015.

**Table 2.1.** Quantile plot values for streamflow data from the James River, Virginia, 1900–2015.

[ $x$ , annual mean discharge in cubic meters per second ( $\text{m}^3/\text{s}$ );  $p$ , cumulative frequency values;  $i$ , rank of the observation; dots indicate data not shown in table but included in figure 2.4]

$i$	$x_i$	$p_i$
1	60.2	0.0085
2	84.4	0.0171
3	87.9	0.0256
4	95.0	0.0342
5	101.0	0.0427
.	.	.
112	312.7	0.9573
113	321.2	0.9658
114	343.9	0.9744
115	344.2	0.9829
116	351.5	0.9915

Variations of quantile plots are used for three purposes:

1. To compare two or more data distributions (a Q-Q plot),
2. To compare data to a normal distribution (a normal probability plot, a specialized form of the Q-Q plot), and
3. To calculate frequencies of exceedance (for example, a flow-duration curve used to evaluate streamflow data).

Historically, in the field of hydrology and in statistics in general, different plotting positions have been used to construct quantile plots. The choice of plotting positions used depends on these purposes, but also on tradition or on automatic selection of methods in various statistical software packages. Most plotting positions have the general formula  $p = (i - \alpha)/(n - \alpha - \beta + 1)$  where  $\alpha$  and  $\beta$  are constants. Commonly used formulas are listed in table 2.2.

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 17C, the standard reference for determining flood frequencies in the United States (England and others, 2018). The Weibull formula's primary advantage over the Parzen and Gumbel formulas (**type 4** and **type 7**) is that it recognizes the existence of a nonzero probability of exceeding the maximum observed value. As such, the plotting position of the maximum value is less than 1.0. Positions such as **type 4** or R's default position of **type 7** set the plotting position of the maximum observation to 1.0, implying that the probability of exceeding the maximum observed value is zero. This is unrealistic unless the entire population is sampled (a census) and is especially unrealistic with small sample sizes. The Weibull formula (R's **type 6**) is therefore our preferred plotting position and will be used for quantile plots and most other graphs.

One other plotting position used in this text is the Blom (1958) formula (**type 9**). It is used in the probability plot correlation coefficient test (introduced in chap. 4) and in graphical comparisons of a dataset to the normal distribution. It is the standard plotting position for normal probability plots, which use the inverse of the normal cdf. Excellent discussions of plotting position formulas include Stedinger and others (1993) and Hyndman and Fan (1996).

### 2.1.3 Boxplots

A useful and concise graphical display for summarizing the distribution of a dataset is the boxplot (fig. 2.5). Boxplots provide visual summaries of

1. The center of the data (the median—the centerline of the box);
2. The variation or spread (interquartile range—the box height);
3. The skewness (quartile skew—the relative size of box halves); and
4. The presence or absence of unusual values and their magnitudes (outliers).

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

The origin of boxplots traces back to the box-and-whisker plot defined by Tukey (1977) and then refined by Cleveland (1985). Many variations on the general theme have been defined and used since that time. This text will not attempt to compare and contrast these variations and will use the operational definition of the boxplot used in the R function `boxplot`. The elements of a boxplot include

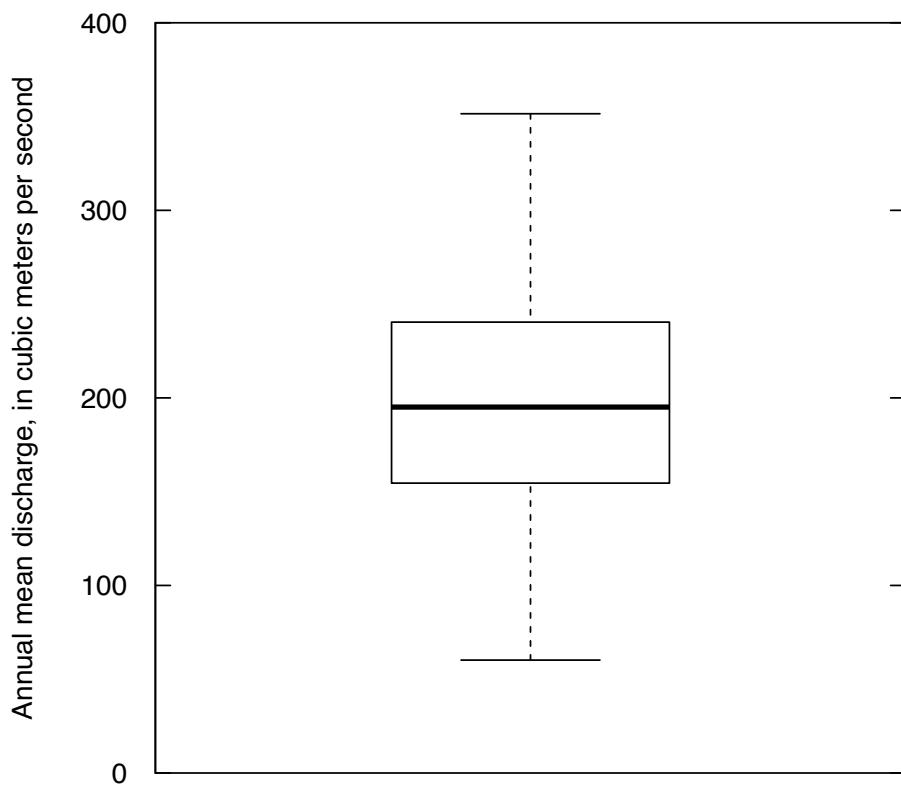
1. A box that delineates the middle 50 percent of the data;
2. Hinges, the top and bottom of the box that are approximately equal to the 75th and 25th percentiles of the sample, respectively (there is a slight difference in the definition of the hinges and the definition of the quartiles see discussion below);
3. A line within the box representing the median of the sample;
4. Whiskers extending outward from the box to indicate the extent of the data beyond the middle 50 percent (see below for a precise definition of how the whisker lengths are determined); and
5. The outside values, observations that lie beyond the limits of the whiskers, shown as individual symbols.

Using the default values in the `boxplot` function, the whiskers extend to the most extreme data point that is no more than 1.5 times the length of the box away from the box. The outside values are all of the data that lie beyond the whiskers. The choice of the value 1.5 is related to the quantiles of a normal distribution. Specifically, in a very large sample from a normal distribution we expect that about 5 percent of all observations will be outside values (2.5 percent on the upper end and 2.5 percent on the lower end), and 95 percent of the observations will fall between the two whiskers. Thus, when the boxplot appears to be roughly symmetrical and there are roughly an equal number of outside values above and below the box, if substantially more than 5 percent of the observations are plotted individually, we can infer that the sample comes from a distribution that has a heavier tail than a normal distribution. If there are substantially fewer

**Table 2.2.** Definitions and comments on eight possible plotting position formulas, based on Hyndman and Fan (1996) and Stedinger and others (1993).

[NA, not applicable;  $i$ , rank of the observation;  $n$ , sample size;  $p_i$ , calculated probability for the  $i$ th ranked observation;  $p_n$ , rank of the largest observation]

Reference	$\alpha$	$\beta$	Formula $p_i =$	Type in R quantile function	Comments
Parzen (1979)	0	1	$i/n$	4	$p_n = 1.0$ , poor choice: suggests largest observation can never be exceeded
Hazen (1914)	1/2	1/2	$(i - (1/2))/n$	5	Traditional in hydrology
Weibull (1939), also Gumbel (1958)	0	0	$(i)/(n + 1)$	6	Unbiased exceedance probabilities
Gumbel (1958)	1	1	$(i - 1)/(n - 1)$	7	$p_n = 1.0$ , poor choice: suggests largest observation can never be exceeded
Reiss (1989)	1/3	1/3	$(i - (1/3))/(n + (1/3))$	8	Median unbiased quantiles
Blom (1958)	3/8	3/8	$(i - (3/8))/(n + (1/4))$	9	Unbiased quantiles for normal
Cunnane (1978)	2/5	2/5	$(i - (2/5))/(n + (1/5))$	NA	Approximate quantile unbiased
Gringorten (1963)	0.44	0.44	$(i - 0.44)/(n + 0.12)$	NA	Optimized for Gumbel distribution



**Figure 2.5.** A boxplot of annual mean discharge values from the James River, Virginia, 1900–2015.

than 5 percent outside values in a symmetrical boxplot, we can infer that the distribution has tails that are lighter than a normal. The quantitative measure of the heaviness or lightness of the tails of a distribution is called kurtosis.

The rules for computing the location of the hinges are somewhat complex and for large samples they are functionally equivalent to the upper and lower quartiles of the sample. To be more precise about the definitions of the hinges (for a sample size of  $n$ ), the following can be stated.

- If  $n$  is odd, then the hinges are exactly equal to the 25th and 75th percentiles of the sample, as it would be computed using the quantile function with the plotting position set to `type = 7`.
- If  $n$  is divisible by four ( $n=4, 8, 12, 16, \dots$ ) then the hinges are the average of two adjacent observations in a ranked list of the observations. For example, if  $n=12$ , the lower hinge is the average of the third and fourth smallest observations. If  $n=16$ , it is the average of the fourth and fifth smallest observations. The upper hinge would be the mirror image of these. For example, if  $n=12$  the upper hinge would be the average of the third and fourth largest observations.
- If  $n$  is divisible by two but is not divisible by four ( $n=6, 10, 14, 18, \dots$ ) then the hinges are exactly equal to one of the observations. For example, if  $n=10$  the lower hinge would be the third smallest observation and if  $n=14$  the lower hinge would be the fourth smallest observation.

In cases where the user wants to see the actual values that are used in plotting the boxplot, those can be obtained by setting the argument `plot = FALSE` in the boxplot function. Instead of a plot, R will return the values on which the boxplots are based.

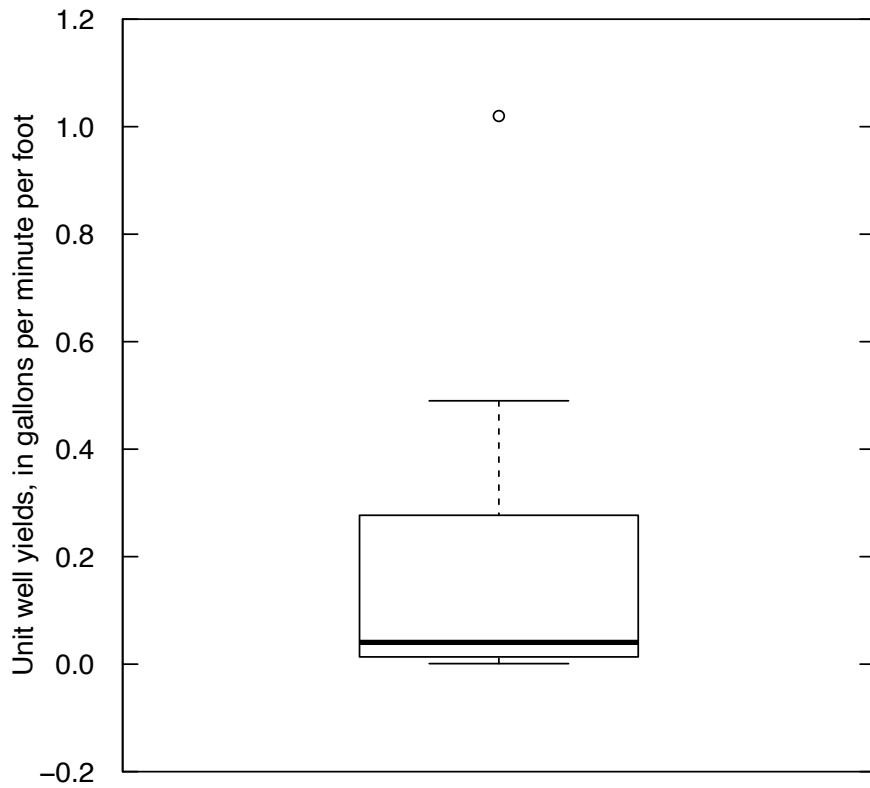
Boxplots are a valuable means of obtaining a simple overview of data, but they do not provide much insight on characteristics of the middle portion of the data. For example, if the distribution were bimodal (in other words, the probability density function has more than one peak), it is unlikely that a boxplot would give us the slightest hint at this feature of the dataset; in particular, the data between the upper and lower quartiles would be obscured within the box. Nevertheless, boxplots provide what is generally considered to be the most effective overall summary of a dataset.

Annual mean discharge data for the James River are shown in a boxplot in figure 2.5; this is the same data that are plotted as a histogram in figure 2.2 and as a quantile plot in figure 2.4. Several things are made clear by this plot; the distribution is highly symmetrical, both in terms of the middle 50 percent of the data (those depicted by the box) and by the tails of the distribution (noting the roughly equal length of the two whiskers). Also notable is that there are no outside values. Given that the sample size is 116 observations, if this were a sample from a normal distribution we would expect about 6 outside values. Not all symmetric data follow a normal distribution, and this dataset appears to have somewhat light tails. Whether this is a serious departure from normality depends on the purpose of the analysis.

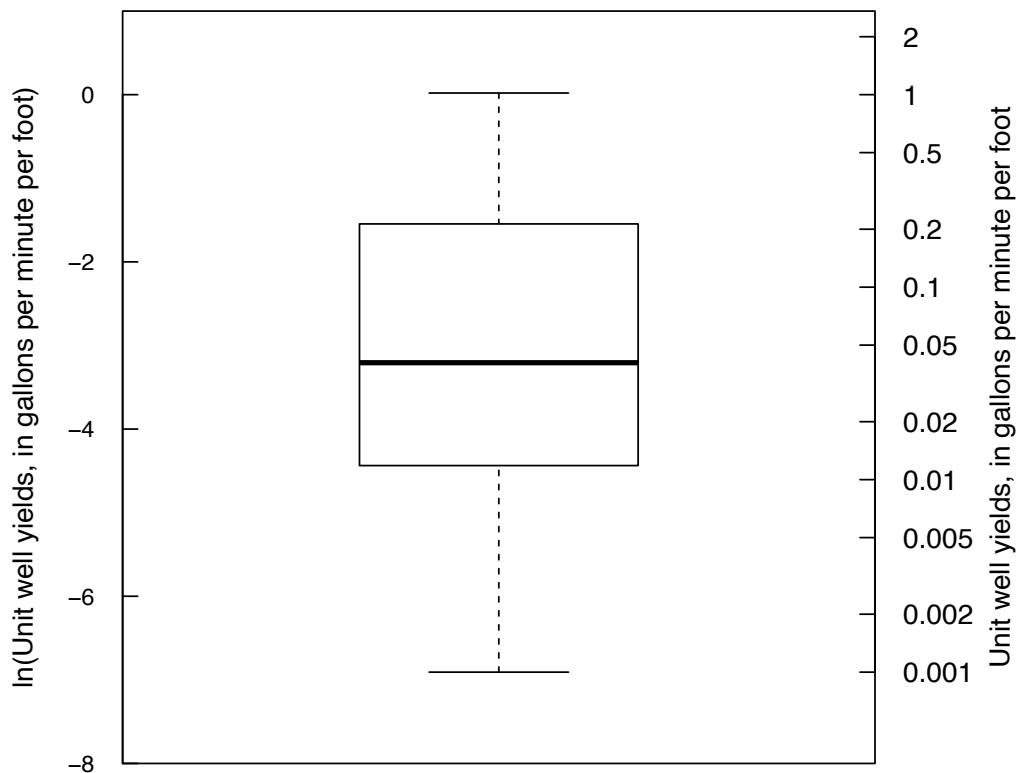
In contrast to this example of a highly symmetrical dataset, we consider well-yield data for unfractured conditions from Wright (1985). This is a small dataset ( $n=12$ ), but it is highly skewed, with many values close to zero. The boxplot for this dataset is shown in figure 2.6.

In figure 2.6, we depart from the usual practice of using the value of zero as the lower limit on the y-axis, because the lower whisker and the lower hinge would plot virtually on top of the x-axis. The dataset shows strong signs of asymmetry. The median is far from being midway between the hinges and is very close to the lower hinge; this suggests a good deal of skewness in the body of the dataset. The fact that the upper whisker is much longer than the lower whisker is also an indication of asymmetry (in this case in the tails) and the single outlier is far from the upper whisker. All of this suggests that a normal distribution would be highly inappropriate for representing the data and that a log transformation may be more suitable. The boxplot of the transformed data is shown in figure 2.7.

Having made the log transformation, the dataset now appears to be highly symmetrical both in terms of the middle 50 percent of the data as well as the extremes. Note that this representation of the data is not the same as taking the original boxplot and simply plotting it using a log transformed scale. The difference lies in how the upper and lower quartiles, including outliers, are represented. The rules for drawing the whiskers and the outside values are expressed in relation to the dimensions of the box, and simply replotting the boxplot with a log scale will cause these features of the boxplot to violate the rules by which boxplots are drawn. In short—if transformations are to be considered—the data should be transformed and then the boxplot created from the transformed data. When this is done, it is important that the plot axes should show both the original units and the transformed units, as is done in figure 2.7. Generally, axes of a graph should make it possible for the person reading the graph to interpret the data in their original units.



**Figure 2.6.** Boxplot of the unit well yields for valleys with unfractured rocks from Wright (1985).



**Figure 2.7.** Boxplot of the natural log of unit well yield for valleys with unfractured rocks from Wright (1985).

There is a good deal of literature about the uses and interpretation of boxplots, and there are many variations on the basic approach to boxplots described here. Some of the relevant references on this topic include McGill and others (1978), Chambers and others (1983), Frigge and others (1989), and Krzywinski and Altman (2014).

## 2.1.4 Probability (Q-Q) Plots

Two sets of quantiles are plotted against one another in a Q-Q plot, one on the vertical axis and the other on the horizontal axis. The second set of quantiles in a Q-Q plot can be a cdf for a theoretical distribution (see fig. 2.9). This type of Q-Q plot is also called a probability plot (discussed in this section). Another type of Q-Q plot compares two empirical distributions, in order to illustrate their similarity (see fig. 2.20 presented in section 2.2.4.).

Probability plots are used to determine how well data fit a theoretical distribution such as the normal, lognormal, or gamma distributions. Determining fit could be attempted by visually comparing histograms of sample data to the probability density function of the theoretical distributions as seen in figures 1.1 and 1.2. It could also be determined as in figure 2.8, where the quantile plot of the James River annual mean streamflow data (lower scale) is overlaid with the S-shaped quantiles of the standard normal distribution (upper scale) where a standard normal quantile of 0 is placed at the sample mean and a standard normal quantile of 1 is placed at the mean plus one standard deviation. However, research into human perception has shown that departures from straight lines are discerned more easily than departures from curvilinear patterns (Cleveland and McGill, 1984a; also see several references in chap. 16). By expressing the theoretical distribution as a straight line, departures from the distribution are more readily apparent, as is the case with probability plots.

To construct this version of a probability plot, the cumulative frequency values (shown on fig. 2.8) are re-expressed as standard normal quantiles such that a normal distribution will plot as a straight line. Figure 2.9 shows a normal probability plot of the James River streamflows. This figure shows that James River dataset is highly consistent with a normal distribution because the points are so close to the straight line. One could argue that it departs very slightly from a normal distribution, particularly in the upper tail. The slight divergence of the highest few data points indicates that the high end of the dataset is slightly more extreme than we should expect given the mean and standard deviation computed from the dataset, but in practical terms these are very small departures. In chapter 4, we will introduce a formal hypothesis test that can be used to examine the assumption that this sample could have been a sample from a normal distribution.

The construction of this figure can be described as follows (the R code that produced it is shown in the supplemental material (SM) for chapter 2 [SM.2]). Assume we have  $n$  observations  $Q_i$  where  $i=1, 2, \dots, n$  sorted from the smallest value,  $Q_1$ , to the largest,  $Q_n$ . The straight line is defined by

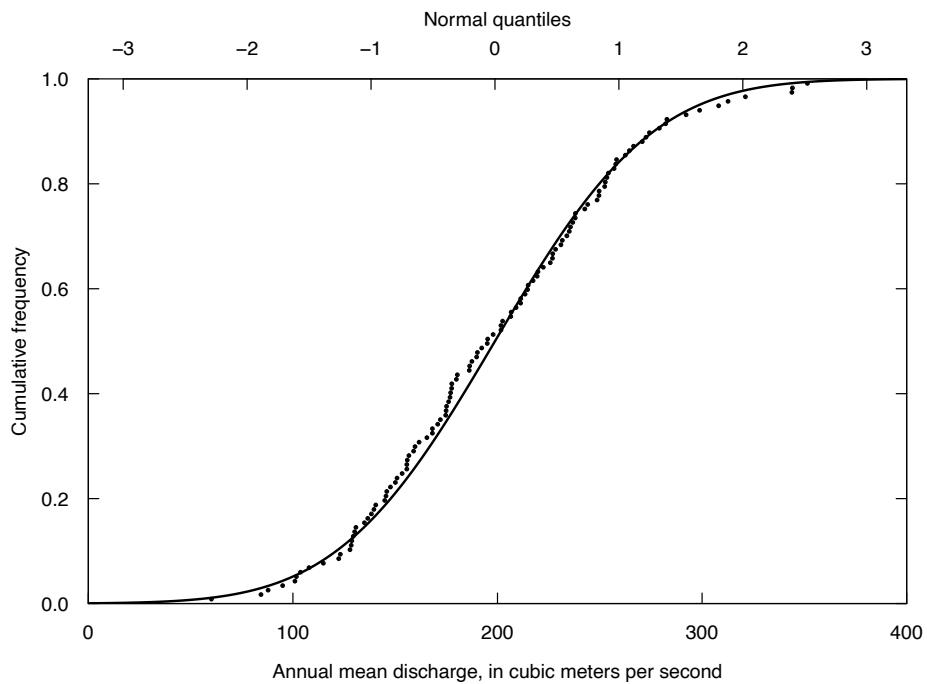
$$Q = \bar{Q} + (z \cdot s_Q) , \quad (2.1)$$

where  $\bar{Q}$  is the sample mean of the  $Q_i$  values, and  $s_Q$  is the sample standard deviation of the  $Q_i$  values. The  $n$  individual points are plotted at  $(Z_i, Q_i)$ , where

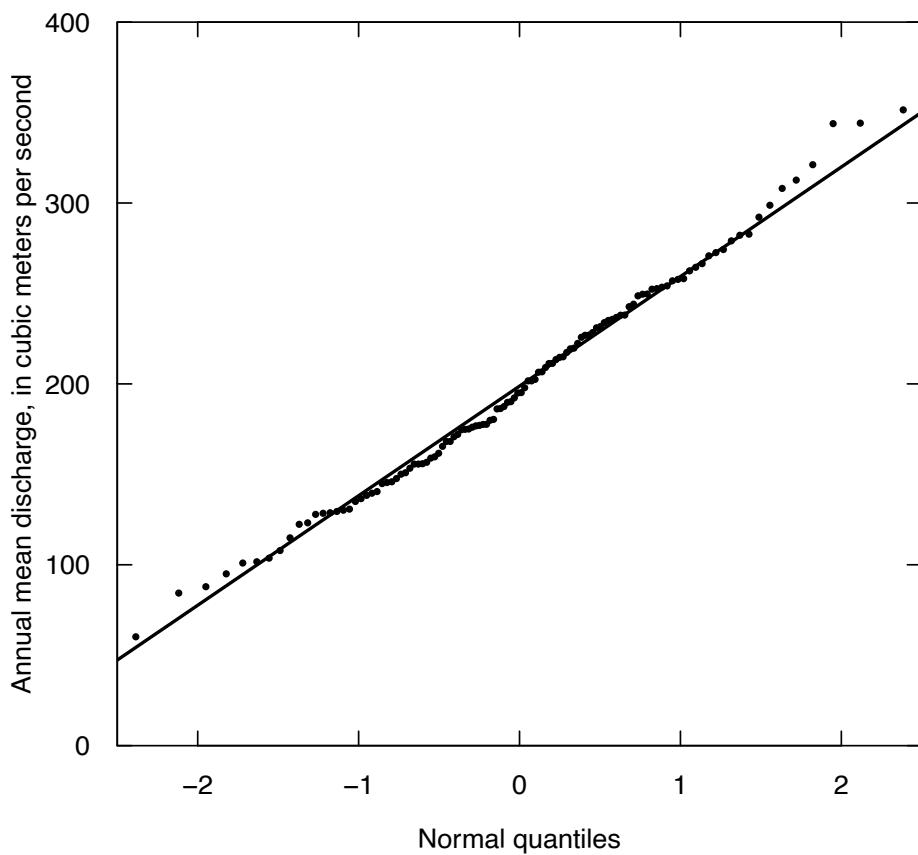
$$Z_i = F_N^{-1}(p_i) , \text{ and} \quad (2.2)$$

$$p_i = \frac{i}{n+1} , \quad (2.3)$$

where  $F_N^{-1}(p_i)$  is the inverse of the cumulative distribution function for the standard normal distribution (mean = 0, standard deviation = 1). Note that in R this function is called qnorm. The formula used here for  $p_i$  follows the Weibull plotting position formula (type = 6 in the quantile function). However, because the Blom plotting position (type = 9 in the quantile function) is unbiased for the normal distribution, it can be argued that the Blom plotting position is preferable in this kind of plot. The difference in the appearance of the graph between these two choices of plotting position is very small.



**Figure 2.8.** Overlay of James River annual mean discharge (1900–2015) and standard normal distribution quantile plots.



**Figure 2.9.** Probability plot of James River annual mean discharge data (1900–2015).

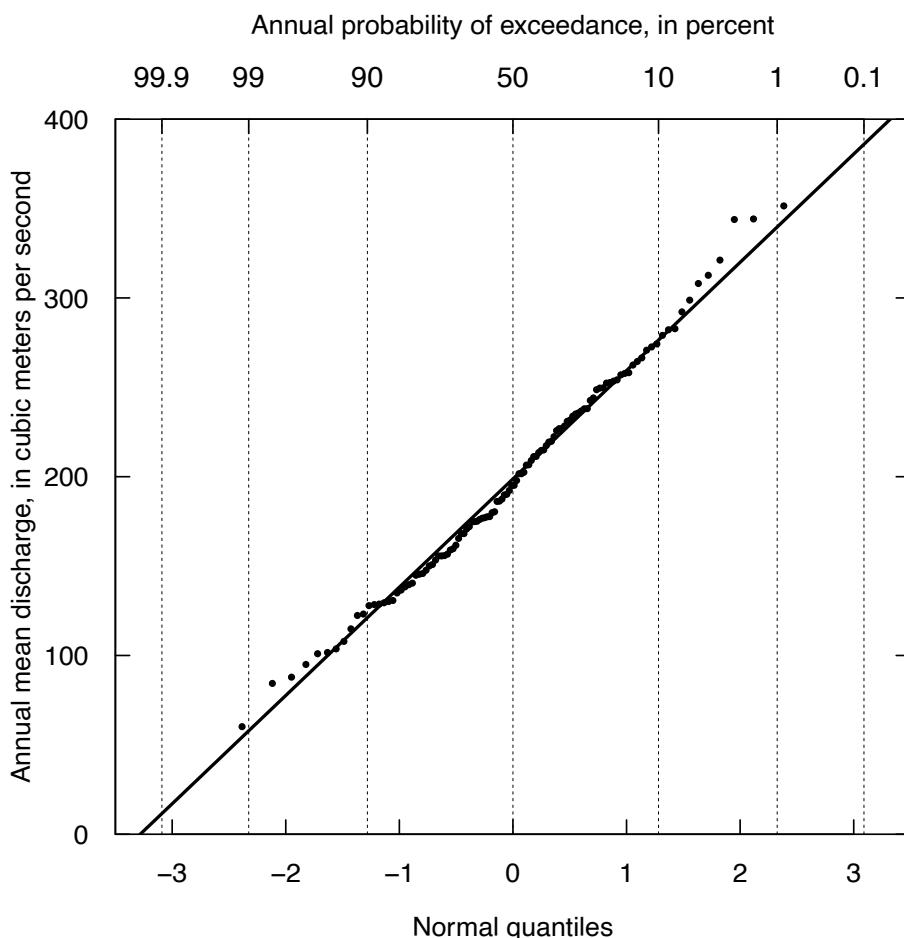
## 2.1.5 Q-Q plots as Exceedance Probability Plots

In water resources, a second horizontal scale is sometimes added to a probability plot to show the probability of exceedance on the horizontal axis. This is simply a restatement of the information provided by the quantiles, so the normal quantile scale can be deleted. An example of such a plot is shown in figure 2.10. The plot is identical to the probability plot shown in figure 2.9, but a horizontal scale of exceedances probability has been added. The addition of the exceedances probability scale makes the graphic more easily understood. This approach can be applied with any distribution. Before statistical software became common, hydrologists would draw exceedances probability plots using a specialized probability paper which was designed so that a normal distribution would plot as a straight line.

## 2.1.6 Deviations from a Linear Pattern on a Probability Plot

The deviations from a linear pattern on a probability plot can be useful in identifying the nature and severity of departures from the selected theoretical distribution. In particular, the deviations can identify the overall asymmetry or skewness of the distribution, the presence of outliers, and the heaviness of the tails (kurtosis) of the distribution.

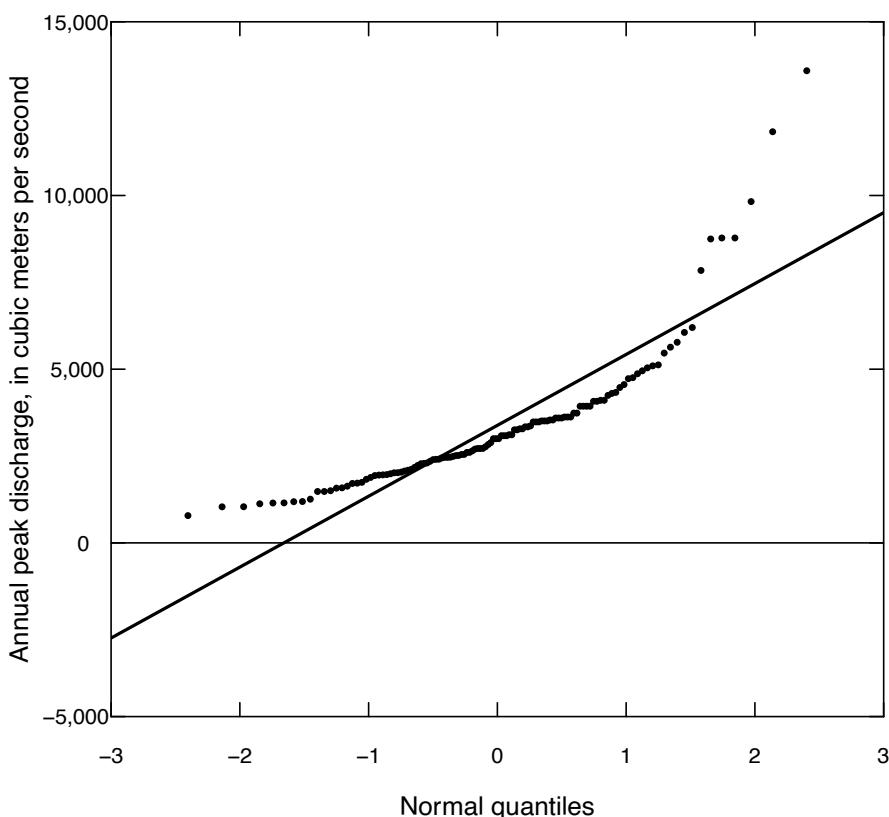
The probability plot of the James River annual discharge data in figure 2.10 shows a modest amount of skewness. It has a slight positive skew, meaning that values in the right tail of the distribution (say  $z > +2$ ) are somewhat farther from the mean than the values in the left tail (say  $z < -2$ ). Positive skewness is indicated by the concave upward shape of the data as compared to the theoretical normal distribution that we see in figure 2.10. Negative skewness would be indicated by a convex-upward shape. The probability plot is another way of seeing the modest asymmetry we saw in the histogram (fig. 2.2), but in



**Figure 2.10.** Exceedance probability plot for the James River annual mean discharge data (1900–2015).

the probability plot it is directly compared to a normal distribution. Note that in the boxplot of this same dataset (fig. 2.5) it is nearly impossible to see the asymmetry because the plot does not convey any detailed information about the distribution shape inside the range of the upper and lower quartiles. However, when skewness is more pronounced, the boxplot provides good visual evidence for it. Outliers do not appear to be present in the dataset based on visual appearance of the probability plot, histogram, or boxplot. The probability plot also shows this by the fact that the maximum and minimum values both plot close to the theoretical line for the normal distribution. Additionally, we can see that the tails of the distribution are not significantly heavier or lighter than those of the normal distribution. Data having a slope steeper than the normal distribution in both the high and low portions would indicate high kurtosis (heavy tails). Data having a more gentle slope in the extremes at both the high and low ends would indicate low kurtosis (light tails).

Annual peak discharge values for the Potomac River at Point of Rocks, Maryland, are shown in a probability plot in figure 2.11. The dataset covers the 122-year period from 1895 to 2016. As before, the solid line represents a normal distribution with the same mean and standard deviation as the dataset. The first thing to note about the line is that for reasonable values of  $z$  it extends well below zero, yet we know that negative discharge values are impossible. This alone is a sufficient argument for rejecting the idea that these data follow a normal distribution. Even if the negative value problem did not exist, we can see that the normal distribution is a poor fit given the strongly concave-upward shape of the plotted data, which shows a strong positive skewness. This plot suggests that a transformation is needed to make the data approximate a normal distribution. A good candidate for such a transformation is the log transformation (indicated by the concave-upward shape and negative values).



**Figure 2.11.** A normal probability plot of the Potomac River annual peak discharge data (1895–2016).

A normal probability plot of the natural log of the annual peak discharge data is shown in figure 2.12. Note that the right axis is modified here to show the discharge values that are equivalent to the log values used to make the graph. In general, when transformations of any kind are used to make a graphic, the analyst should present a scale that allows the reader to translate the plotted results in to their original units. Taking the log eliminates both of the problems: negative values and the asymmetry. The quantile plot shows a very good fit to a normal distribution except that the seven largest values are all higher than we would expect from a normal distribution. This suggests that the log discharge data have a small amount of positive skewness. We can gain similar insights from a boxplot of these log-transformed discharges, as shown in figure 2.13. What is particularly striking about this boxplot is how close to symmetric it is, both in terms of the hinges and in terms of the outliers.

Given that both the box and the whiskers are nearly symmetric, the plot suggests that no power transformation, such as those in the ladder of powers, would produce a more nearly normal distribution. Although in this case the log transformation results in a good approximation to normality, there are many cases where transformations that render the sample nearly symmetric have an excess of extreme values compared to what a normal distribution can be expected to produce. Although transformations can be helpful, they are not a panacea because of the tendency for datasets to have heavy tails. Consequently, this book emphasizes the use of nonparametric and permutation tests, both of which are designed to be robust against departures from normality.

### 2.1.7 Probability Plots for Comparing Among Distributions

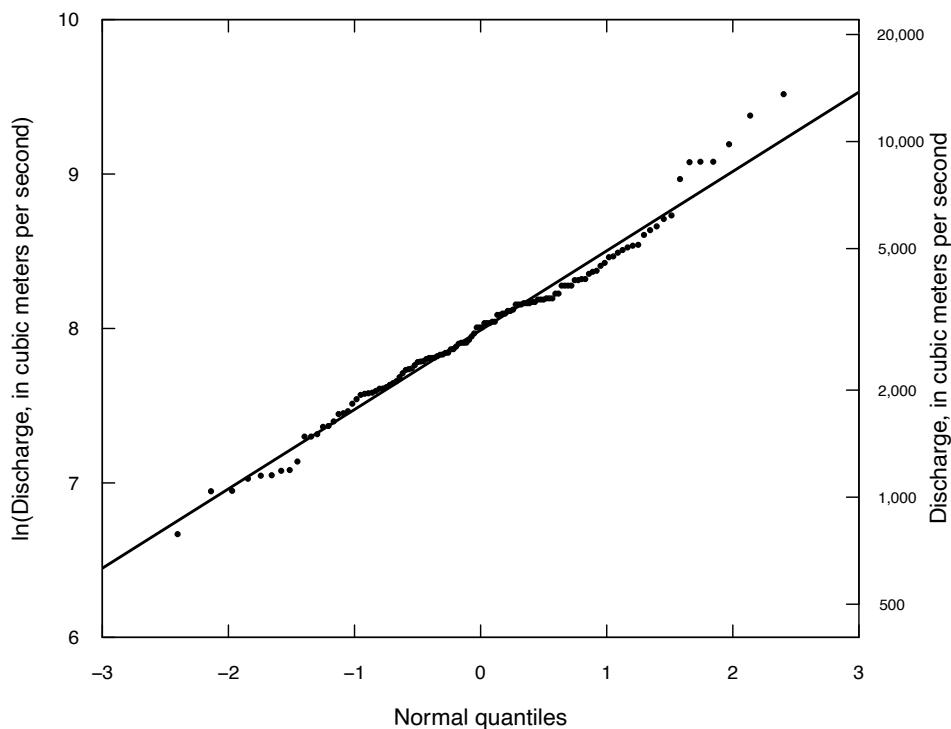
In addition to a normal probability plot, quantiles may be computed and probability plots constructed for any theoretical probability distribution. A good visual check of the appropriateness of a distribution is to make probability plots for several distributions, where each plot is designed to show the selected distribution as a straight line. In general, a distribution can be selected based on the similarity of the data quantiles to this straight line. This approach is formalized as a hypothesis test, known as the probability plot correlation coefficient test, which is discussed in chapter 4.

There is a rich literature in hydrology discussing selection of distribution types and parameter estimation for both flood flows and low flows. This text does not pursue those ideas in any depth. Examples of the use of probability plots for frequency analysis can be found in references such as Vogel (1986), Vogel and Kroll (1989), and Stedinger and others (1993).

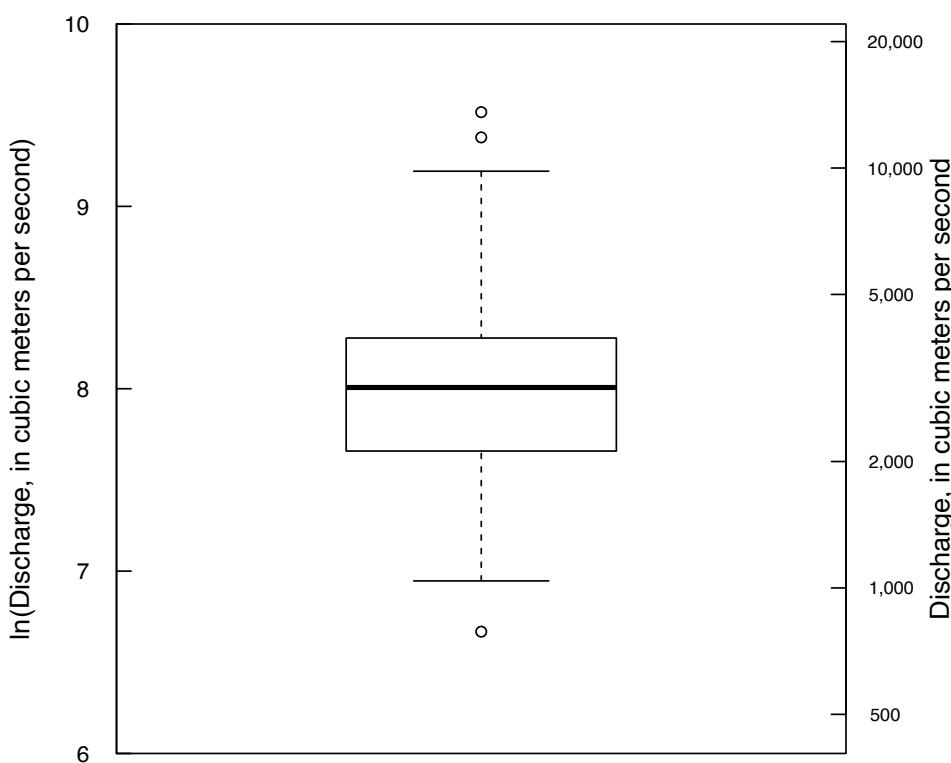
## 2.2 Graphical Comparisons of Two or More Datasets

Each of the graphical methods discussed thus far can be, and have been, used for comparing multiple groups of data. However, each is not equally effective. As the following sections show, histograms are not capable of providing visual comparisons between datasets 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 among many groups of data. A quantile-quantile (Q-Q) plot (similar to a quantile plot but for multiple datasets), provides additional information about the relation between two datasets.

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



**Figure 2.12.** Normal probability plot of the natural log of the annual peak discharge data from the Potomac River at Point of Rocks, Maryland, streamgage (1895–2016).

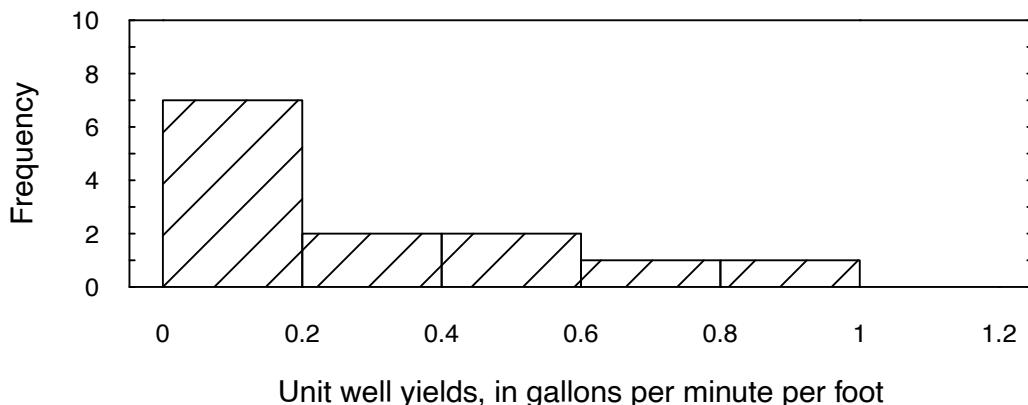


**Figure 2.13.** Boxplot of the natural log of the annual peak discharge data from the Potomac River at Point of Rocks, Maryland, streamgage (1895–2016).

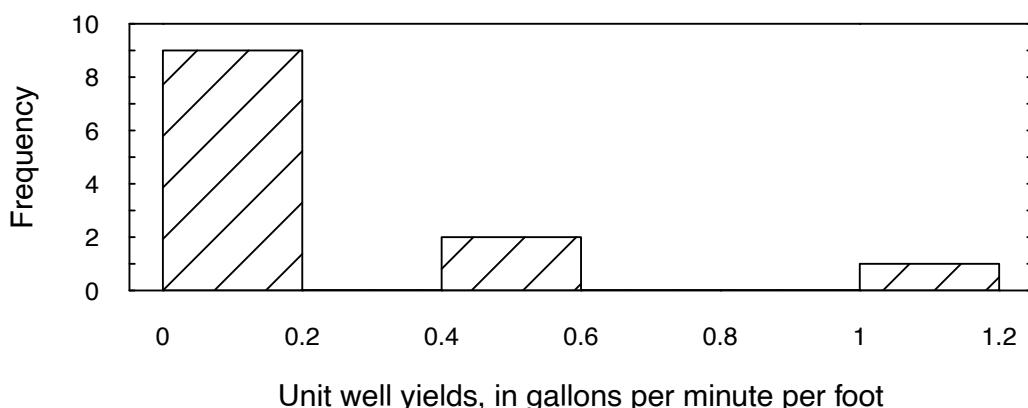
## 2.2.1 Histograms

Histograms for the two sets of well yield data from Wright (1985) are shown in figure 2.14. The right-skewness of each dataset 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. Note that for comparability the bins on the x-axis were designed to be the same and they are plotted one above the other. Even with these two features designed to facilitate comparisons, this is a poor way to make a comparison of the two groups. There are also examples where graphics have been designed to superimpose two histograms on one set of axes, but these tend to be highly confusing and uninformative. Thus, we will not present any of these multiple histogram approaches.

**A. Valleys with fractures**



**B. Valleys without fractures**



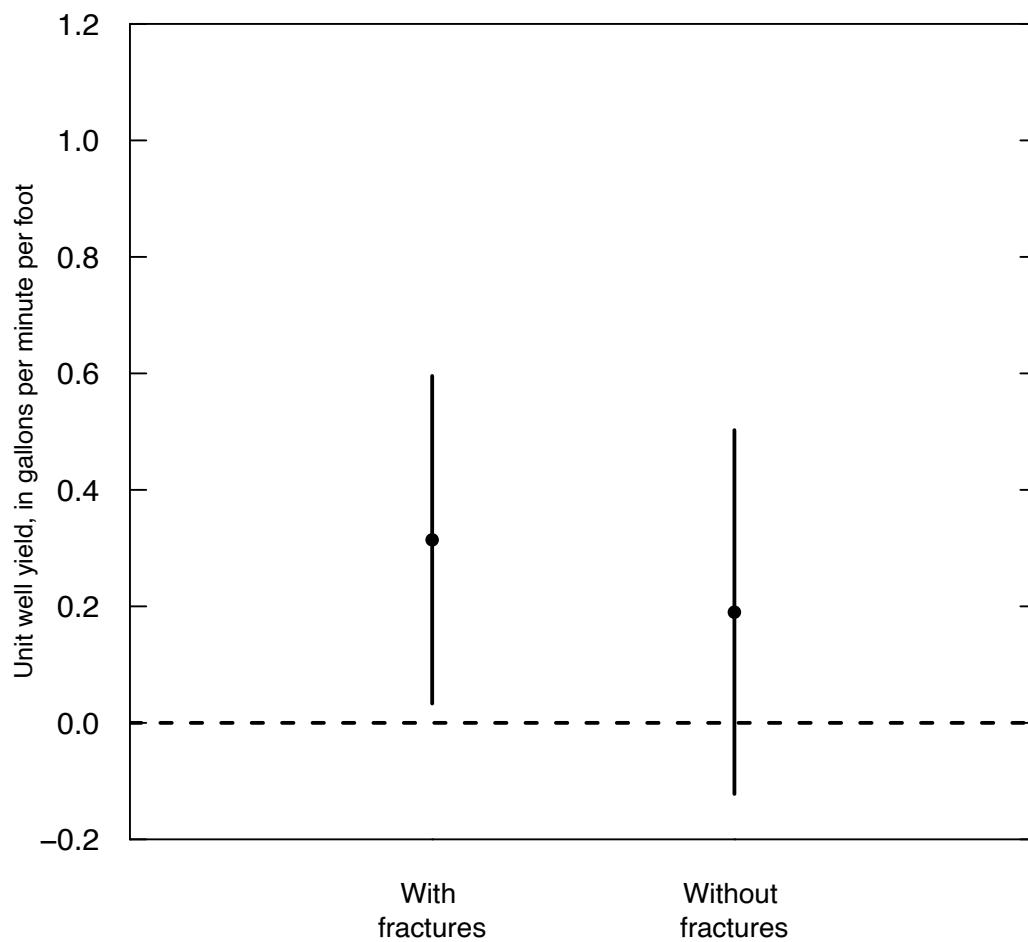
**Figure 2.14.** Histograms of unit well yield data for (A) valleys with fractures, and (B) valleys without fractures.

## 2.2.2 Dot-and-line Plots of Means and Standard Deviations

Dot-and-line plots, used by some to compare datasets, are shown in figure 2.15. We include them here as a contrast with side-by-side boxplots, which are discussed in the next section. Dot-and-line plots are much less meaningful than boxplots and fail to show many important features of the datasets. Each dot represents the mean of a dataset. The bars extend to plus and minus one standard deviation beyond the mean, though two standard deviations or standard errors of the mean have also been used. This plot displays differences in mean yields, but little else.

There are also several deficiencies to this plot. First, is the assumption of symmetry; this causes the lower end of the “Without fractures” line to fall below zero, which is physically impossible. Second, no information on the symmetry of the data or presence of outliers is available. Third, there is little information given on the spread of the data, as the same portrayal of standard deviation may represent the spread of most of the data or may be strongly influenced by skewness and a few outliers.

As will be shown in the next section, these two groups of data are much better described and contrasted with each other through the use of side-by-side boxplots.

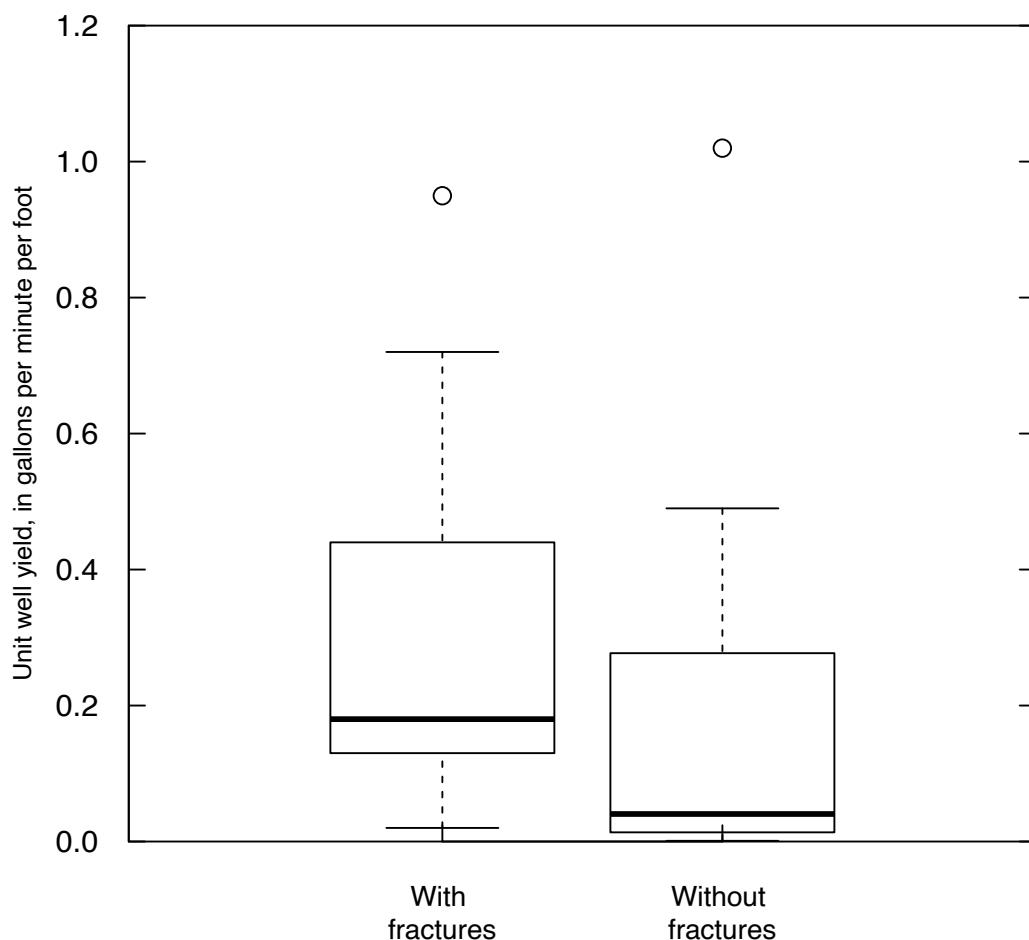


**Figure 2.15.** Dot-and-line plot of the unit well yield datasets for areas underlain by either fractured or unfractured rock.

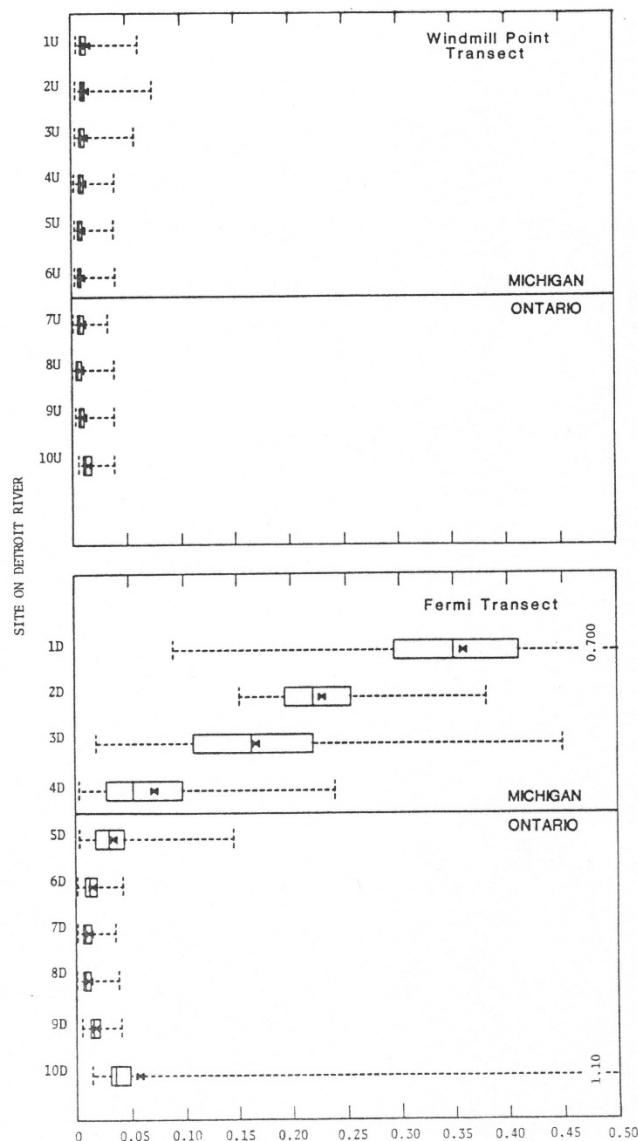
### 2.2.3 Side-by-side Boxplots

Figure 2.16 presents side-by-side boxplots of the same well yield data used in figure 2.15. 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 datasets are seen to be right-skewed. Clearly, a large amount of information is contained in this concise illustration. The mean yield, particularly for wells without fractures, is undoubtedly inflated owing to skewness, and differences between the two groups of data will, in general, be larger than indicated by the differences in their mean values. The same characteristics that make boxplots useful for inspecting a single dataset make them even more useful for comparing multiple datasets. They are valuable guides in determining whether central tendency, variability, and symmetry differ among groups of data. At a glance we can determine the approximate difference between their medians as well as the differences between their upper quartiles or between their lower quartiles.

Side-by-side boxplots display the essential characteristics of numerous groups of data in a small space. For example, the 20 boxplots of figure 2.17 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 stream water quality is succinctly summarized in this relatively small figure.



**Figure 2.16.** Side-by-side boxplots of the unit well yield datasets for areas underlain by either fractured or unfractured rock.



**Figure 2.17.** Boxplots of ammonia nitrogen concentrations as a function of location on two transects of the Detroit River (from Holtschlag, 1987). The Windmill transect lies upstream of Detroit and the Fermi transect lies downstream of Detroit.

Another appropriate use of side-by-side boxplots is the comparison of observations from different months or seasons of the year. The example shown in figure 2.18 is a set of dissolved nitrate-plus-nitrite sample values from the Illinois River at Valley City, Illinois. This is just a subset of the available data that covers water years 2000 through 2015. The boxplots reveal a number of features. There are large differences across seasons. The months of December through June have relatively high concentrations and among those months there are no particularly strong differences. Each of the boxplots is either approximately symmetrical or has some positive skewness and their outliers are rather symmetric around the middle 50 percent of the distribution. The concentrations are much lower in the summer and fall, with July being a transition from the high values of the winter and spring. August and September are the months with the lowest concentrations and the October and November boxplots show a transition back to the higher winter values. August and September show some right skewness but the others are relatively symmetrical. Not surprisingly, the months with the lowest median concentrations also show lower variability. One could also plot these data again using the logs of the concentration values. The result (not shown) indicates that the logs of the concentration values have similar interquartile ranges across the months, suggesting that variability is proportional to the magnitude of the means. The explanation for this pattern is related to the timing of nitrogen fertilizer application in this watershed (late fall and spring) and the fact that biological uptake and denitrification, which are both most active in the warm summer months, have a strong effect on reducing nitrate concentrations in the stream as compared to the winter and spring when these processes are less effective. This figure is an important reminder that although temperatures (air or water) often are well approximated by a sinusoidal curve with a period of one year, some hydrologic variables such as nitrate concentrations can have a seasonal pattern that is more complex because of the dependence on physical processes, such as rainfall or snowfall, as well as biological processes and timing of human activities in the watershed.

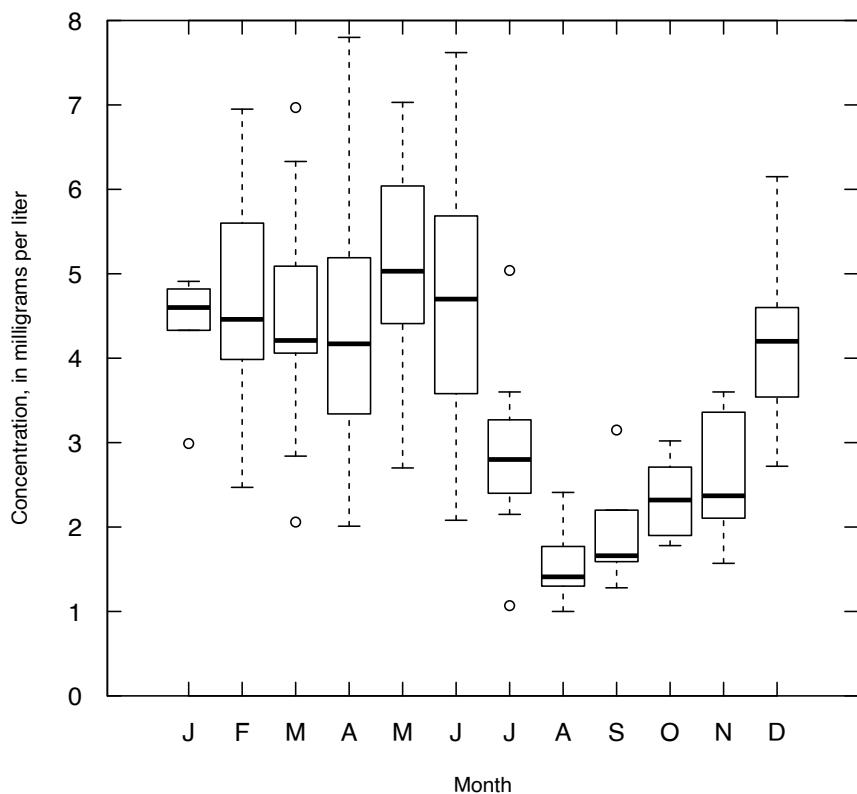
## 2.2.4 Q-Q Plots of Multiple Groups of Data

Q-Q plots (first discussed in section 2.1.4. and used to compare an empirical distribution to a theoretical distribution) are also useful graphics for comparing two empirical distributions. Characteristics evident in boxplots are also seen using Q-Q plots, though in a different format. Comparisons of each quantile, not just the boxplot quartiles, can be made.

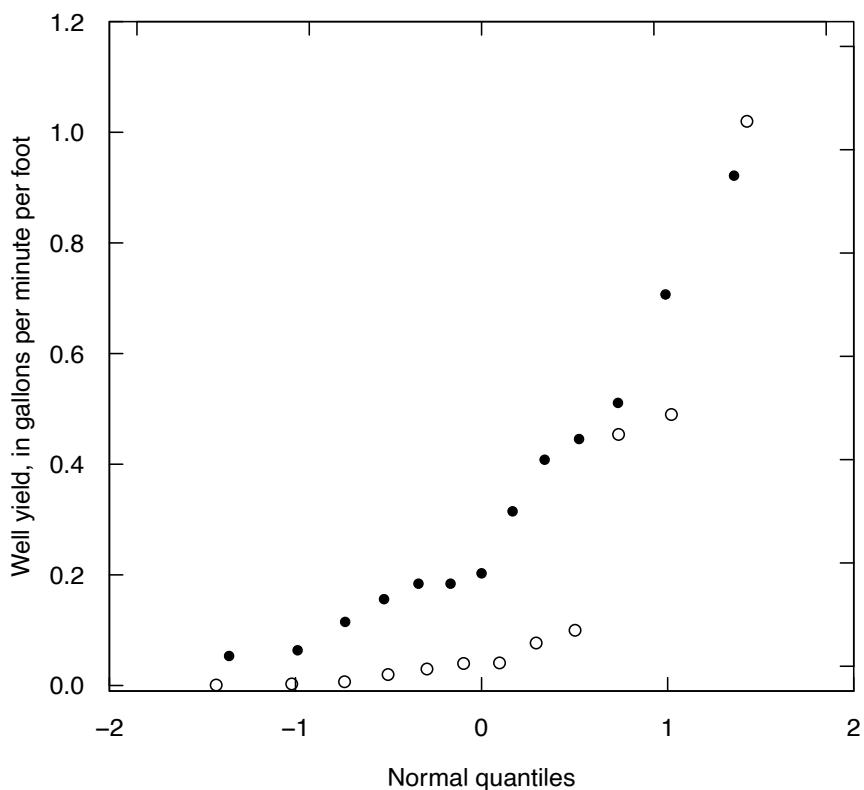
A probability plot of the two well yield datasets is shown in figure 2.19. The right-skewness of each dataset is shown by their concave-upward shapes. Wells without fractures have greater skewness, as shown by their greater concavity on the plot. Quantiles of the wells with fractures are higher than those without, indicating generally higher yields. Figure 2.19 shows that the lowest yields and the highest yields in each group are similar even though the middle part of the distributions are rather different from each other. Comparisons between median values are simple to do—just travel up the normal quantile = 0 line. We see that the median for the fractured rock group is much higher than the median for the unfractured group.

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, Q-Q plots are useful graphics. Both boxplots and Q-Q plots have many advantages over histograms or dot-and-line plots.

Direct comparisons can be made between two datasets by graphing the quantiles of one versus the quantiles of the second (Chambers and others, 1983). If the two datasets came from the same distribution, the quantile pairs would plot along a straight line with  $Y_p = X_p$ , where  $p$  is the plotting position and  $Y_p$  is the  $p$ th quantile of  $Y$ . In this case, it would be said that the median, the quartiles, the 10th and 90th percentiles, and so forth, of the two datasets were equal. If one dataset had the same shape as the second, differing only by an additive amount (each quantile was 5 units higher than for the other dataset, for example), the quantile pairs would fall along a line parallel to, but offset from, the  $Y_p = X_p$  line, also with slope = 1. If the datasets differed by a multiplicative constant ( $Y_p = 5 \cdot X_p$ , for example), the quantile pairs would lie along a straight line with slope equal to the multiplicative constant. Relations that are more complex will result in pairs of quantiles that do not lie along a straight line. The question of whether or not datasets differ by additive or multiplicative relations will become important when hypothesis testing is conducted.



**Figure 2.18.** Side-by-side boxplots by month for dissolved nitrate plus nitrite for the Illinois River at Valley City, Illinois, water years 2000–15.



**Figure 2.19.** Probability plots of the unit well yield data. Solid circles, wells located in areas of fractured rock; open circles, wells located in areas without fracturing.

A Q-Q plot of the two groups of well yield data is shown in figure 2.20. Several aspects of the relation between the two datasets are immediately clear. First, the lowest nine quantile pairs appear to fall along a straight line with a slope greater than 1, not parallel to the  $Y_p = X_p$  line shown as a reference. This indicates a multiplicative relation between the data, with  $Y \cong 4.4 \cdot X$ , where 4.4 is the approximate slope of those data on the plot. Therefore, the well yields in fractured areas are generally 4.4 times those in unfractured areas for the lowest 75 percent of the data. The three highest quantile pairs return near to the  $Y=X$  line, indicating that the higher yields in the two datasets 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. Therefore, the Q-Q plot becomes a valuable tool in understanding the relations between datasets before performing any hypothesis tests.

## 2.3 Scatterplots and Enhancements

The two-dimensional scatterplot is one of the most familiar graphical methods for data analysis and illustrates the relation between two variables. Of usual interest are three types of questions:

1. What is the shape of the relation? Does it appear to be linear, curved, or piecewise linear?
2. When the data come from two different groups (where a group might be defined by the area from which the samples were collected, or the time period during which it was collected) does the relation between the two variables appear to be the same for the two groups or are they different?
3. Is the variability, or spread, in the relation between the two variables constant over the range of data?

In each case, an enhancement called a smooth (short for smooth curve) may enable 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.

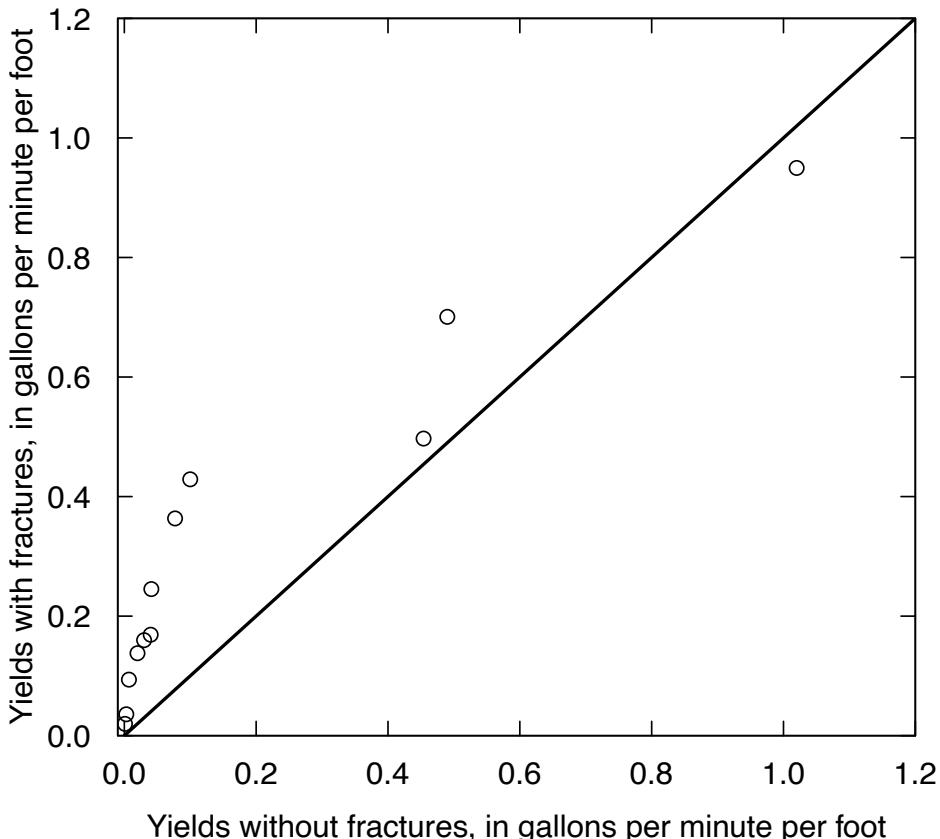
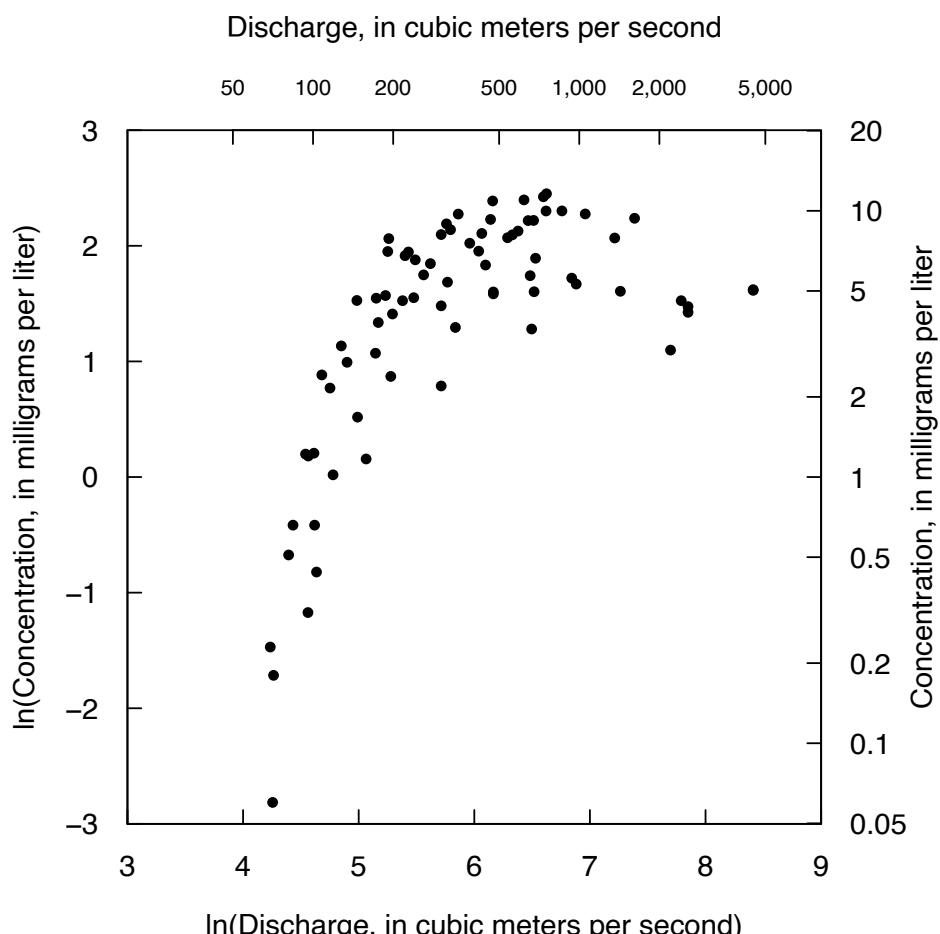


Figure 2.20. Q-Q plot of the well yield data in fractured and unfractured areas.

### 2.3.1 Evaluating Linearity

A scatterplot of the natural log of the concentration of dissolved nitrate plus nitrite (for simplicity we will call that “NO<sub>23</sub>”) for the Iowa River, at Wapello, Iowa, for the months of June, July, August, and September of 1990–2008 are plotted against the natural log of the mean daily discharge for the day on which the sample was collected in figure 2.21. Looking at the scatterplot we can easily see that there is a strong relation between the two variables. We can also see that it would be inaccurate to describe a linear relation between the log of concentration and the log of discharge; however, we might consider a quadratic relation (log of concentration as a linear function of the log of discharge and the square of the log of discharge). In chapters 9 and 11, other ways to answer this question will be presented, but many assessments of linearity are appropriately made solely based on a scatterplot. Superimposing a smooth on the data may also improve the ability to make such assessments.

A smooth is a resistant centerline that is fit to the data whose level and slope varies locally in response to the data themselves. The word “locally” is intended to imply that the location of the curve for any given  $x$  value is determined by the  $y$  values observed at  $x$  values that are close to that  $x$  and not by those that are far away. The purpose of the smooth is to highlight the central tendency of the relation and without being overly influenced by the spread or the extremes of the dataset (either in the  $x$  or  $y$  direction). This approach stands in contrast to using linear regression or multiple linear regression on various transformations of  $x$  (for example using  $x$  and  $x^2$  as predictors of  $y$ ). The regression approach assumes that the relation of  $x$  and  $y$  follow a relation that is specified by a particular mathematical function (for example, linear or quadratic). A consequence of this reliance on the particular functional form selected for the regression is that the fitted value for any given  $x$  value can vary substantially as a result of the  $y$  values that are associated with  $x$  values far from the given  $x$  value. For example, we might observe that suspended sediment concentrations

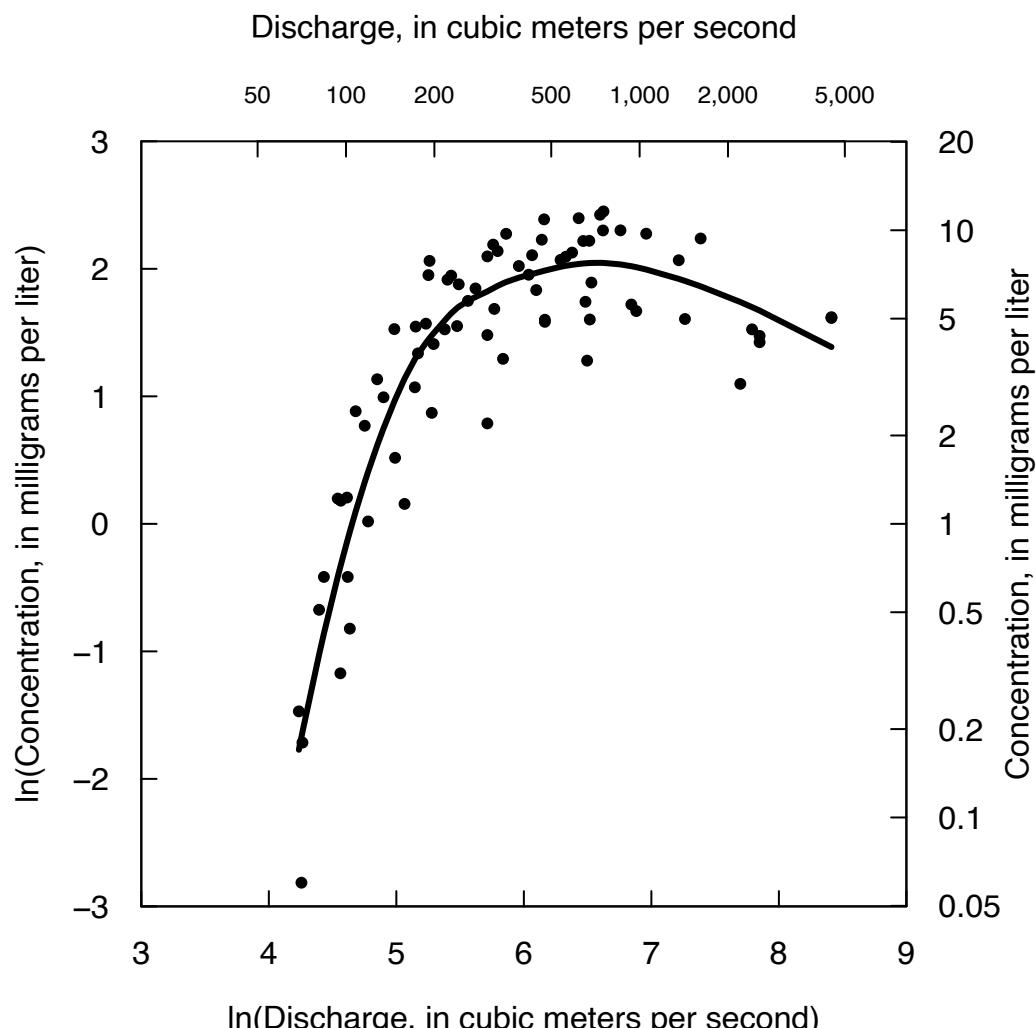


**Figure 2.21.** Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, for the months of June, July, August, and September of 1990–2008.

are strongly related to discharge, but in a regression approach our estimate of central tendency of sediment concentration for a discharge of 1,000 m<sup>3</sup>/s, could be very much influenced by observations of concentration that were at discharges such as 1 m<sup>3</sup>/s or 10 m<sup>3</sup>/s. Smoothing methods avoid this problem and base the estimated central tendency of concentration at 1,000 m<sup>3</sup>/s only on data that are nearby, for example, between 500 m<sup>3</sup>/s and 2,000 m<sup>3</sup>/s.

Many methods are available for constructing this type of centerline, of which the moving average is probably the most familiar. However, the moving average is very sensitive to the influence of outliers and thus not very robust. We discuss computation of smooths in chapter 10, but for now we will merely illustrate their use as aids to graphical data analysis. The smoothing procedure we prefer is called Local Polynomial Regression Fitting, which is implemented by the function `loess` in R (Cleveland and others, 1992). We refer to it as “loess” because it is not strictly an acronym, but it is often called LOESS. A closely related method, the predecessor of loess is LOWESS or LOcally WEighted Scatterplot Smoothing (Cleveland and McGill, 1984a; Cleveland, 1985), which is an iterative process designed to be particularly robust to extreme outliers. An example of loess is shown in figure 2.22.

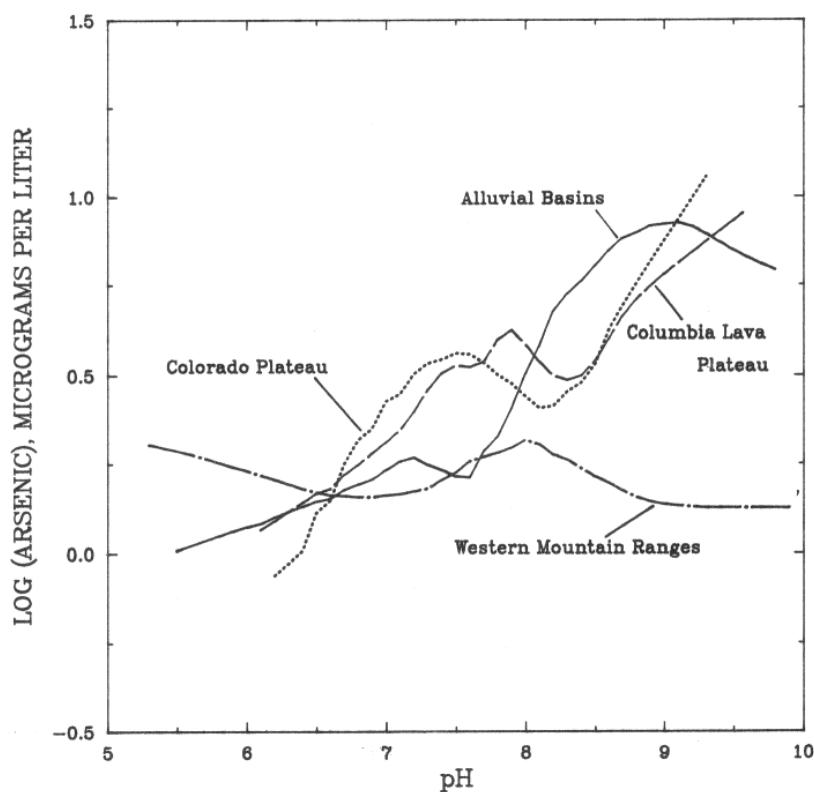
Figure 2.22 shows the same dataset as in figure 2.21, but with the addition of the loess smooth. Note that the relation between  $\ln(\text{concentration})$  and  $\ln(\text{discharge})$  appears relatively linear for discharge values of as much as about 200 m<sup>3</sup>/s, then the relation becomes much less steep (but still increasing) to about



**Figure 2.22.** Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September. The curve represents a loess smooth of these data.

1,000 m<sup>3</sup>/s, and then at higher discharge values concentrations appear to decline with increasing discharge. This curvature is easier to see with the superimposed smooth. It is important to remember that no single model, such as a linear or quadratic function, is assumed before computing a smooth (although smoothing methods may utilize linear or quadratic functions within the smoothing process). The smoothed pattern is derived from the pattern of the data and may take on any shape. As such, smooths are an exploratory tool for discerning the form of the relation between  $y$  and  $x$ . Seeing the pattern of figure 2.22, the smooth suggests that the real pattern does not have the symmetry that arises from a quadratic (a perfect parabola) but rather the upper part (in terms of  $x$ ) of the relation has a much lower negative slope than the positive slope that would fit the lower part of the relation.

This kind of plot, which shows the data plus the loess smooth, can be used when analyzing data on scatterplots and when presenting those data to others. Because no model form is assumed, the plots allow the data to 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 the log of arsenic concentration on pH for thousands of groundwater samples throughout the western United States (fig. 2.23). By using smooths (and eliminating the plotting of the individual data points), the relation between pH and arsenic was seen to differ between the four western provinces.



**Figure 2.23.** Loess smooths representing dependence of  $\log(\text{As})$  on pH for four areas in the western United States (from Welch and others, 1988).

### 2.3.2 Evaluating Differences in Central Tendency on a Scatterplot

A scatterplot of conductance versus pH for samples collected at low-flow in small streams within the coal mining region of Ohio is seen in figure 2.24 (Helsel, 1983). Each stream was classified by the type of land it drained—unmined land, lands mined and later reclaimed, and lands mined and then abandoned without reclamation.

To see the three groups more clearly, a smooth can be constructed for each group that encloses either 50 or 75 percent of the data. This type of smooth is called a polar smooth (Cleveland and McGill, 1984b). To construct it, the data are transformed into polar coordinates, a loess smooth is computed and then is re-transformed back into the original units. A polar smooth enclosing 75 percent of the data for each of the three types of upstream land uses is plotted in figure 2.25 (from Helsel, 1983). These smooths are not limited to a prior form, such as 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 2.25 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 sandstone. Therefore, the type of geologic unit should be included in any analysis or model of the behavior of chemical constituents for these data.

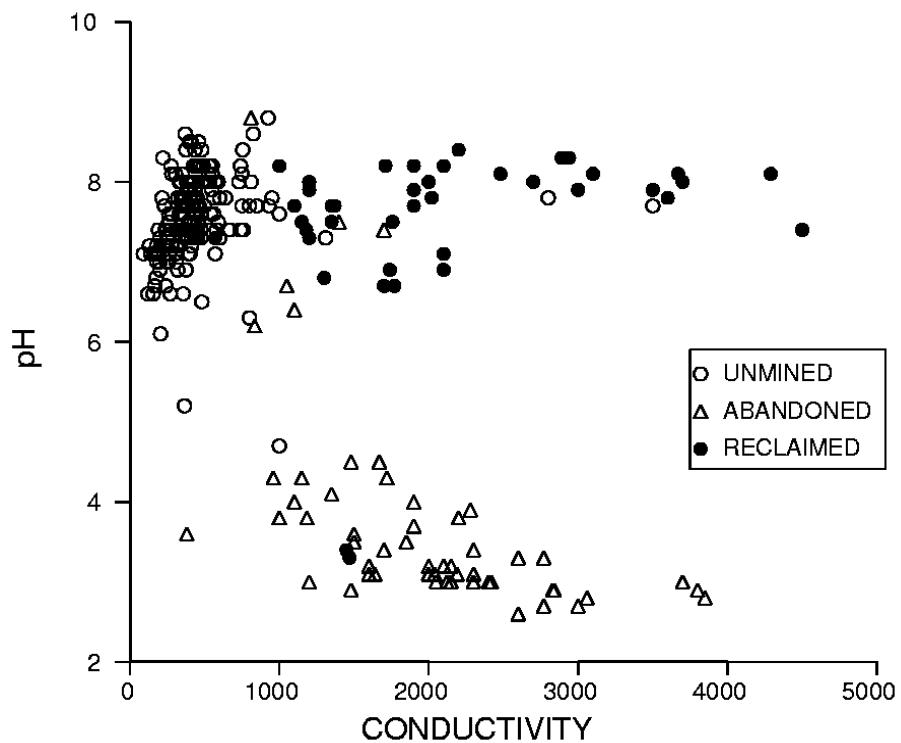
Polar smooths are especially helpful when there is a large amount of data to be plotted on a scatterplot. In such situations, the use of different symbols for distinguishing between groups will be ineffective, as the plot will be too crowded to see patterns in the locations of symbols. Indeed, in some locations it will not be possible to distinguish which symbol is plotted. Plots presenting just the polar smooths, as in figure 2.25, provide far greater visual differentiation between groups.

Returning to the stream NO23 concentration for the Illinois River, shown in figures 2.21 and 2.22, we can pose the question: does the relation between NO23 concentration and discharge vary across different seasons of the year? Here we add in a dataset for the colder months of January, February, March, and April. Figure 2.26 shows the warm season (solid circles) and cold season (open circles) and the loess smooths of both groups superimposed on them. The figure demonstrates how different the concentrations are for discharges below about 150 m<sup>3</sup>/s, but also the relative similarity of concentrations at discharges between about 200 and 1,000 m<sup>3</sup>/s. When dealing with multiple datasets on the same plot, it becomes difficult to discern the patterns of each because the data points may be very much intermingled. Using smooths for each dataset can provide a much clearer picture of the differences than what we can perceive from the two superimposed scatterplots. Recognition of these kinds of differences in patterns can be important for selecting analysis methods that are more flexible (such as Weighted Regressions on Time, Discharge, and Season [WRTDS] introduced in chap. 12) and do not assume that similar relations between  $x$  and  $y$  persist over time or across different seasons. These plots are a useful tool to identify when such approaches are needed.

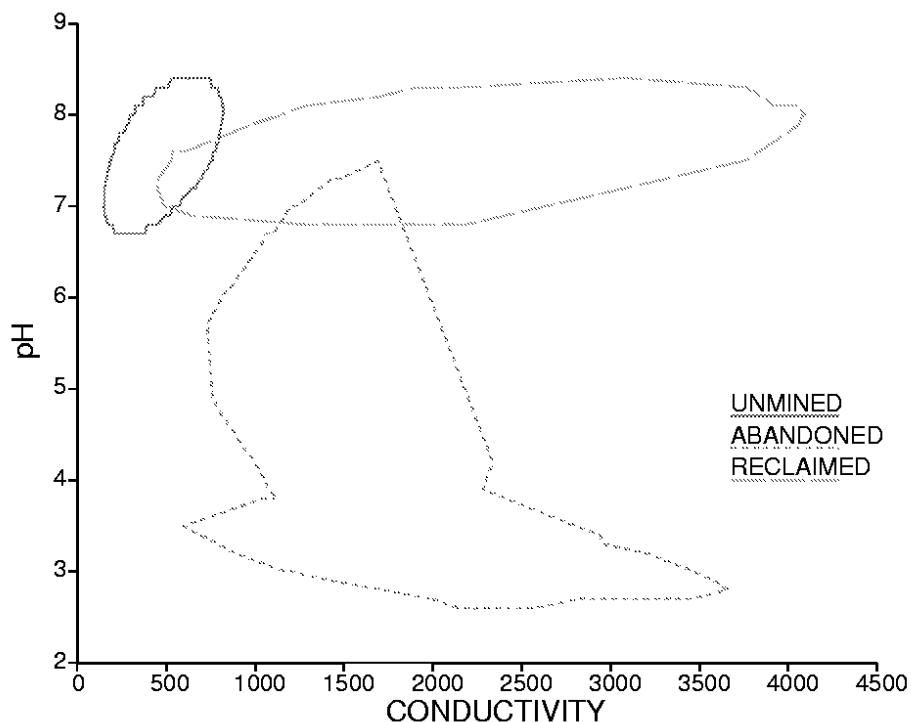
### 2.3.3 Evaluating Differences in Spread

In addition to understanding where the middle of the 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 can aid in judging differences in data variability and are often used for this purpose.

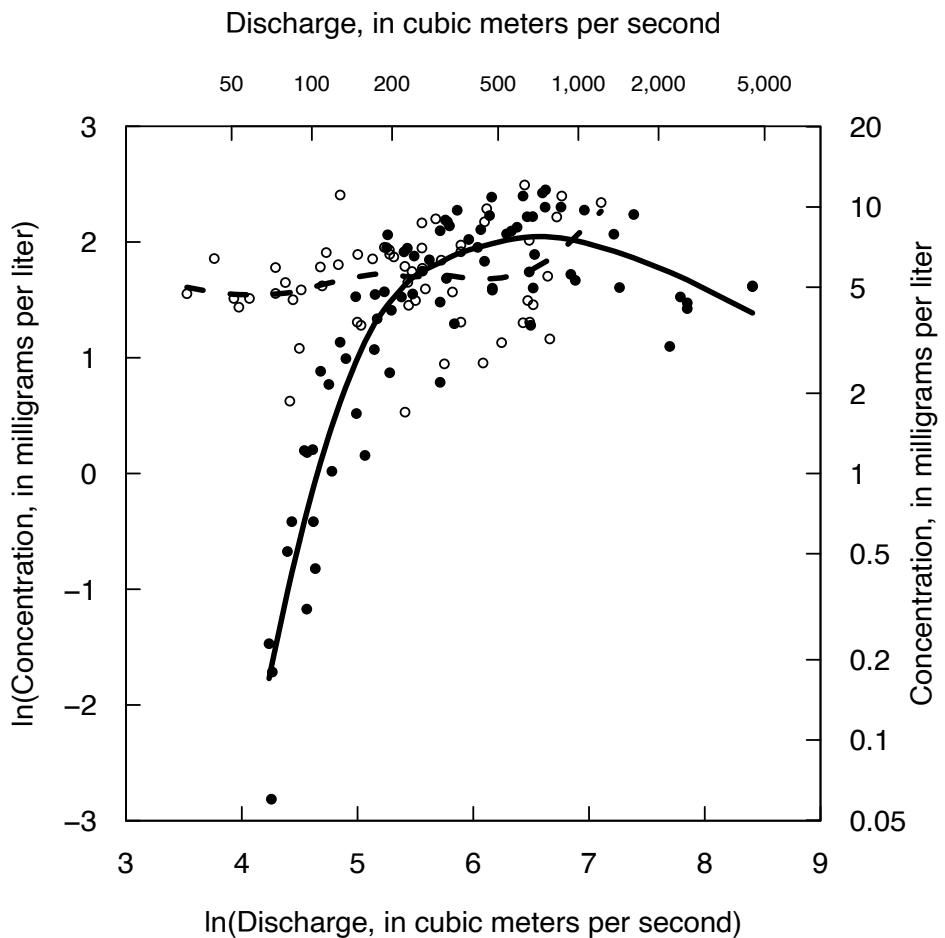
A major problem with judging the changing spread on a scatterplot is that the eye is sensitive to seeing the range of data. The presence of a few unusual values may therefore incorrectly trigger a perception of changing spread. This is especially a problem when the density of data changes across a scatterplot, a common occurrence. Assuming the distribution of data to be identical across a scatterplot, and that no changes in variability 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. Another problem is that the correct way to assess variability on a scatterplot is to measure the vertical distances between the data points and the central value (such as the loess smooth line),



**Figure 2.24.** Scatterplot of water-quality measures draining three types of upstream land use (from Helsel, 1983).



**Figure 2.25.** Polar smooths with 75 percent coverage for the three groups of data seen in figure 2.24, from Helsel (1983).



**Figure 2.26.** Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September (filled circles) or the months of January, February, March, and April (open circles). The solid curve is a loess smooth of the warm season data; the dashed curve is a loess smooth of the cold season data.

but the eye has a tendency to see the distances between the points and the line as the normal distance, which is the shortest distance from the point to the line rather than the vertical distance. Thus, it can be helpful to have a flexible quantitative method for describing the variability of a dataset shown in a scatterplot.

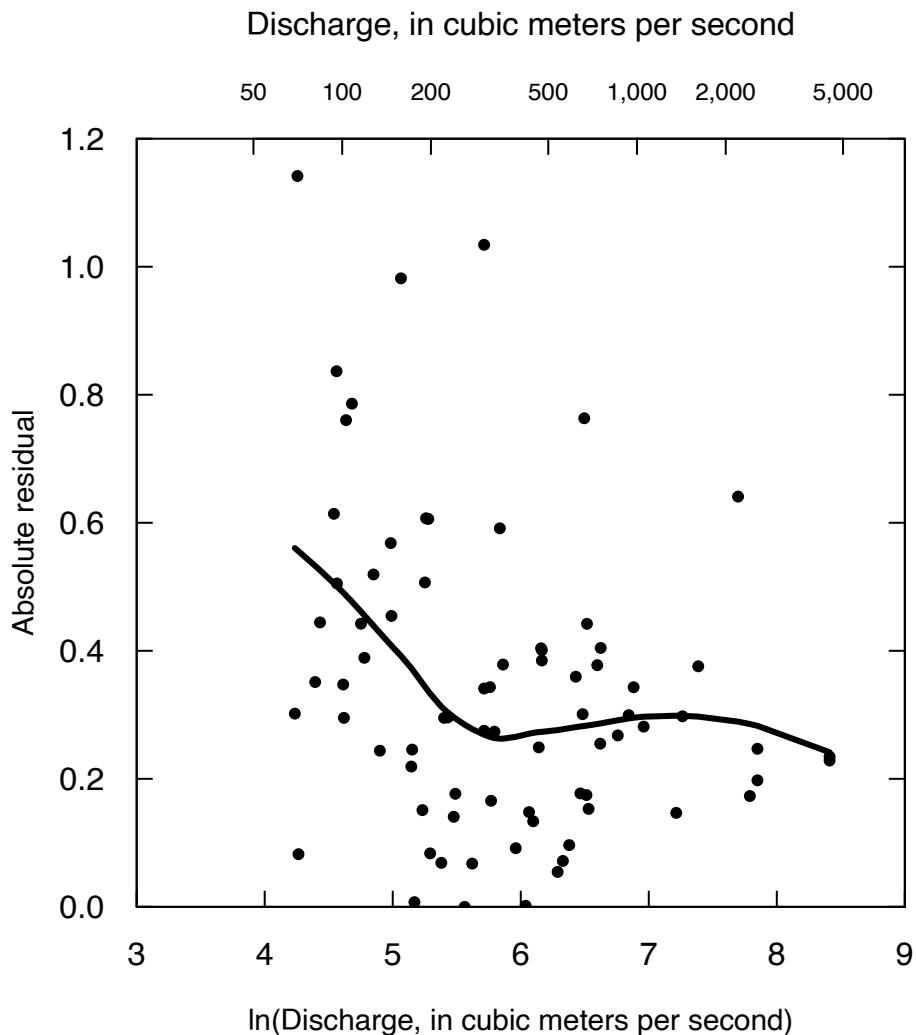
One graphical means of determining changes in spread has been given by Chambers and others (1983). First, as in figure 2.22, a smooth is computed using loess or some other smoothing method. For our purposes here we will call this the middle smooth. The absolute values of differences  $d_i$  between each data point and the smooth at its value of  $x$  is a measure of spread

$$d_i = |y_i - l_i| , \quad (2.4)$$

where

$l_i$  is the value for the loess smooth at  $x_i$ , and  
 $y_i$  is the true value at  $x_i$ .

By graphing the absolute differences  $d_i$  versus  $x_i$ , changes in spread will show as changes in the central tendency of the absolute differences. A smooth of the absolute differences can be used to make the pattern clearer, which is done in figure 2.27, a plot of the absolute differences between the log of concentration and its loess smooth, for the warm season NO23 data from the Iowa River at Wapello, Iowa. Note that at low discharge values, below about  $200 \text{ m}^3/\text{s}$ , the magnitude of the  $d_i$  decreases with increasing



**Figure 2.27.** Absolute residuals from the loess smooth of  $\ln(\text{NO}_2)$  concentrations versus  $\ln(\text{discharge})$ , Iowa River at Wapello, Iowa, for the warm season (June, July, August, and September) 1990–2008.

discharge, but above about  $200 \text{ m}^3/\text{s}$  it is roughly constant. This suggests that any statistical model that depends on the assumption of homoscedastic errors should not be applied here because the errors are quite clearly heteroscedastic. Again, the more free-form WRTDS water quality statistical model is one way to accommodate this departure from the common assumption of homoscedasticity that is fundamental to ordinary least squares regression.

## 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 relations between two variables and accentuate points that appear unusual in their  $x$ - $y$  relation. Yet, there are numerous situations where relations between more than two variables should be considered simultaneously. Similarities and differences between groups of observations based on three 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 relations. They supplement and enhance the understanding provided by formal hypothesis test procedures. Two multivariate graphical methods 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 and others (1983) or the textbook by Everitt (2007).

### 2.4.1 Parallel Plots

Parallel plots, also known as profile plots, 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. Comparisons between observations are made by comparing profiles.

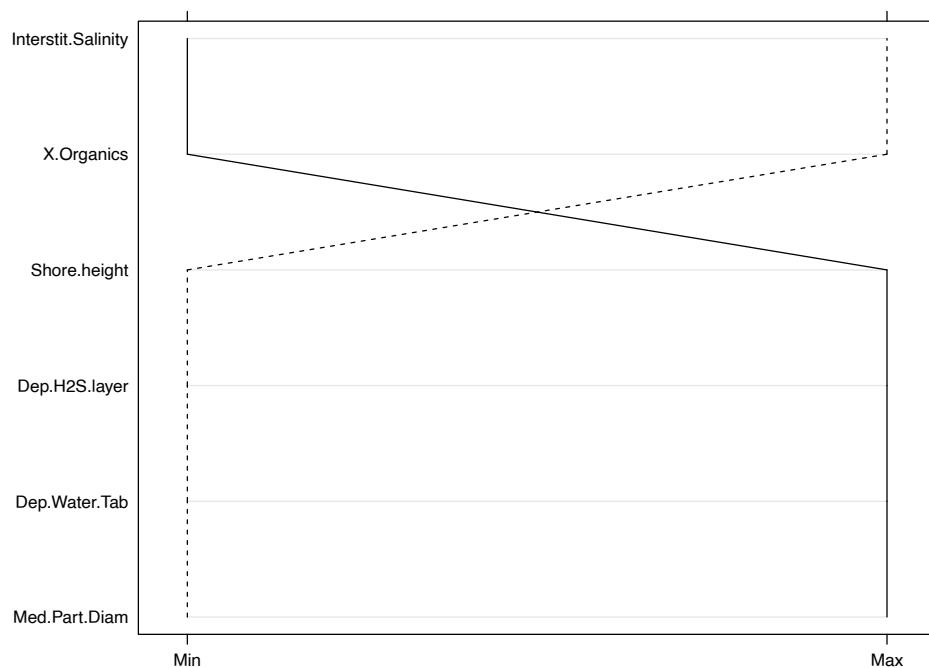
As an example, Warwick (1971) measured 6 site characteristics for 19 stream locations along the Exe river estuary. The types of sites can be classified based on the physical and chemical characteristics of those sites. Of interest were any relations between the 6 site characteristics measured and the types of nematodes found at each site. A parallel plot for 2 of the 19 sites is shown in figure 2.28. Each of the six characteristics is internally scaled from highest to lowest, so that the minimum value on the horizontal scale is the minimum value in the dataset and the maximum values on the horizontal scale is the maximum value for that characteristic; individual values are interpolated between these two values. These two sites show very different profiles. The site drawn as a solid line is low in salinity (Interstit.Salinity) and percent organics (X.Organics), and high in the rest. The site displayed with a dashed line has the opposite profile.

A parallel plot of all 19 sites in figure 2.29 shows why these plots are sometimes called spaghetti plots. They show two groups of multiple sites with differing profiles (the dashed and dotted line sites), as well as two outlier sites plotted with solid lines. The effectiveness of this type of plot depends on the order in which the characteristics are organized. Comparisons of nearby characteristics are easier than comparisons between those that are far apart.

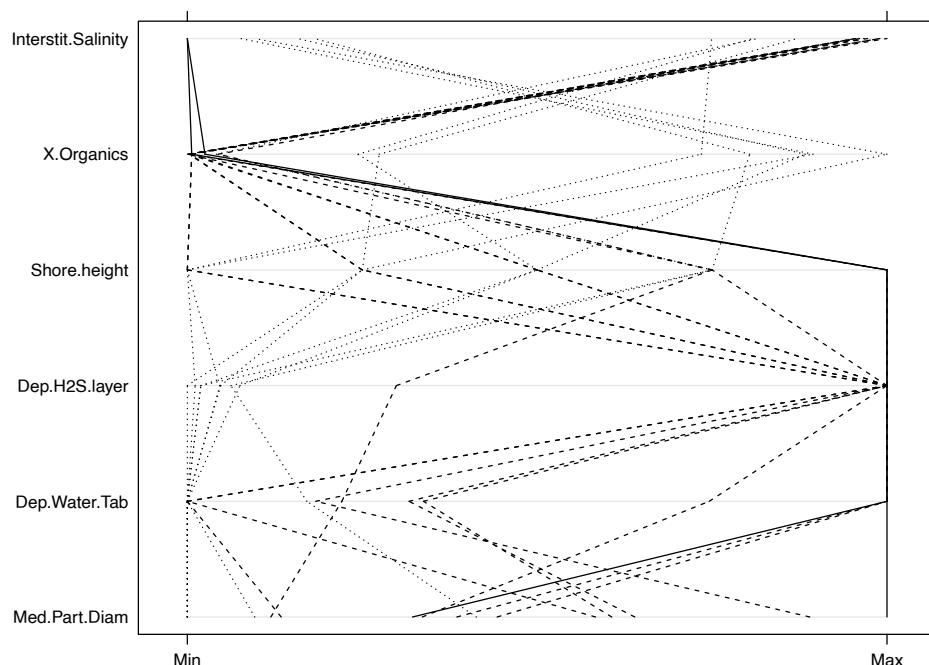
Stiff diagrams (Hem, 1985) are a specific type of parallel plot sometimes used in water quality applications. In a Stiff diagram, the milliequivalents of major water-quality constituents are plotted for a single sample, with the cation profile plotted to the left of the centerline, and the anion profile to the right. Comparisons among several samples based on multiple water-quality constituents are then done by comparing the shapes of the Stiff diagrams. One such comparison for 14 groundwater samples from the Fox Hills Sandstone in Wyoming (Henderson, 1985) is shown in figure 2.30. This facilitates the subjective grouping of similar water types and identifying gradients between sites.

### 2.4.2 Star Plots

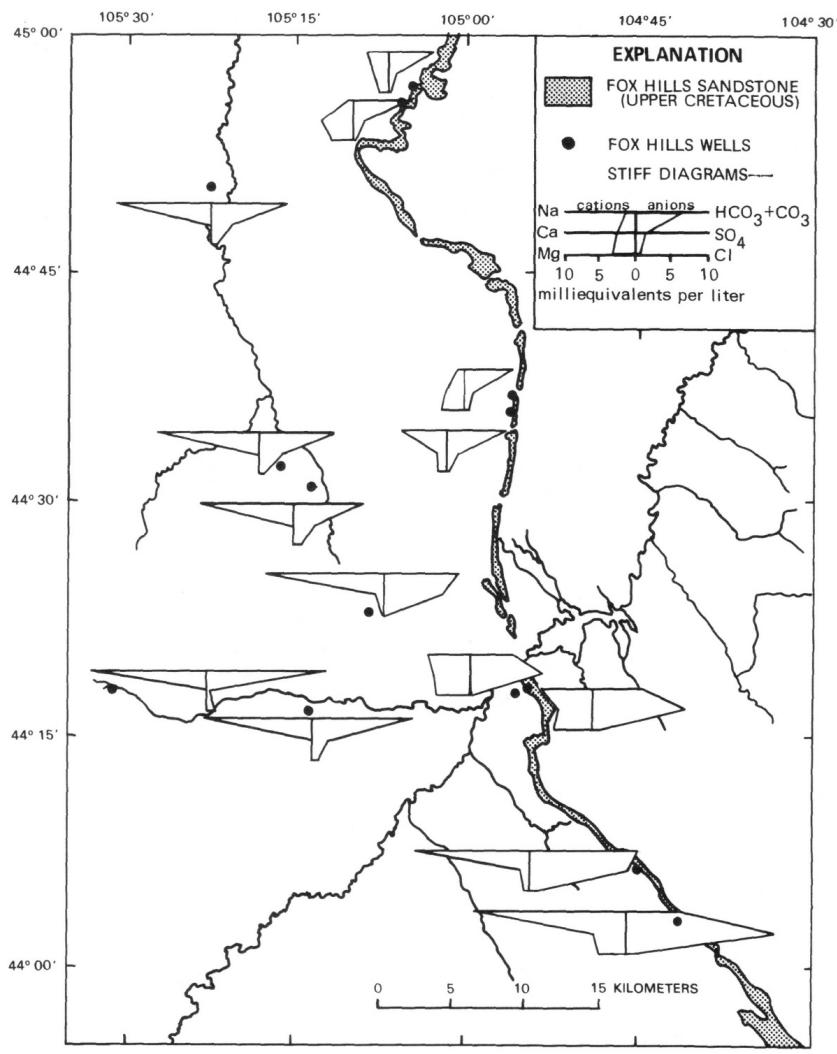
A second method of displaying multiple axes is to have them radiate from a central point, rather than aligned parallel as in a profile plot. Again, one observation is represented by a point on each axis, and these points are connected by line segments. The resulting figures resemble a star pattern and are called star plots. Angles between rays of the star are  $360^\circ/k$ , 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 that looks quite different from the others, perhaps having an unusually long or short ray. Site characteristics for stream locations along the Exe River (Warwick, 1971) are displayed as star plots in figure 2.31. The dramatically different shapes indicate the differences in site conditions.



**Figure 2.28.** Parallel plot of six basin characteristics at a low salinity site (solid line) and a high salinity site (dashed line) (from Warwick, 1971).



**Figure 2.29.** Parallel plot of six basin characteristics at the 19 sites of Warwick (1971).



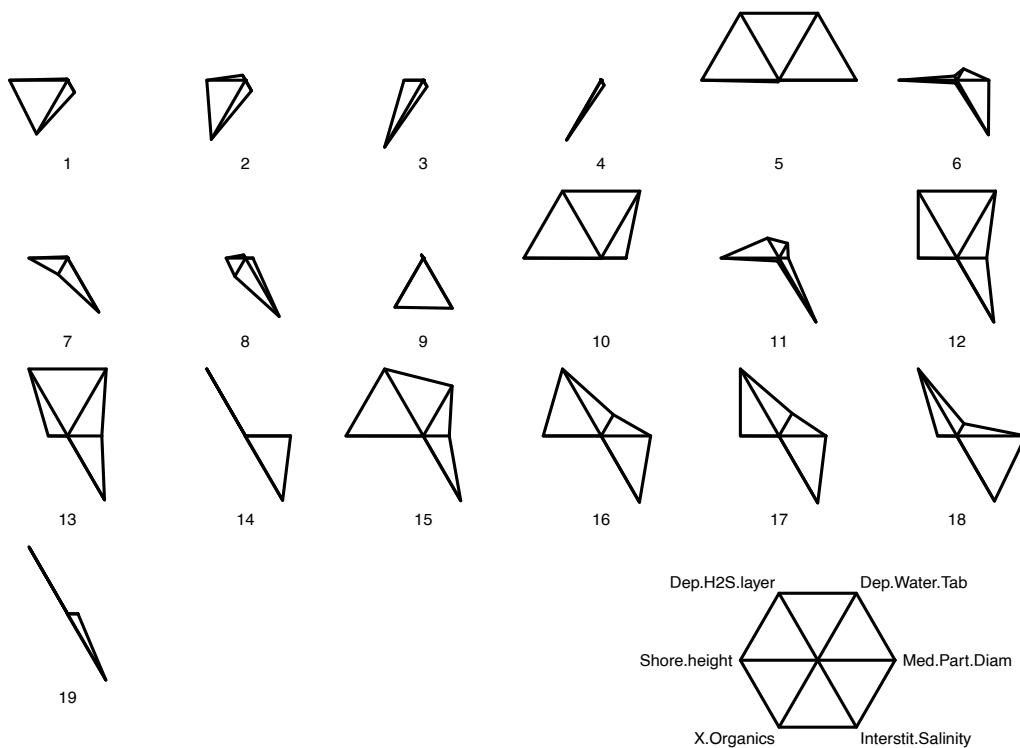
**Figure 2.30.** Stiff diagrams used to display differences in water quality in the Fox Hills Sandstone, Wyoming (from Henderson, 1985).

### 2.4.3 Trilinear and Piper Diagrams

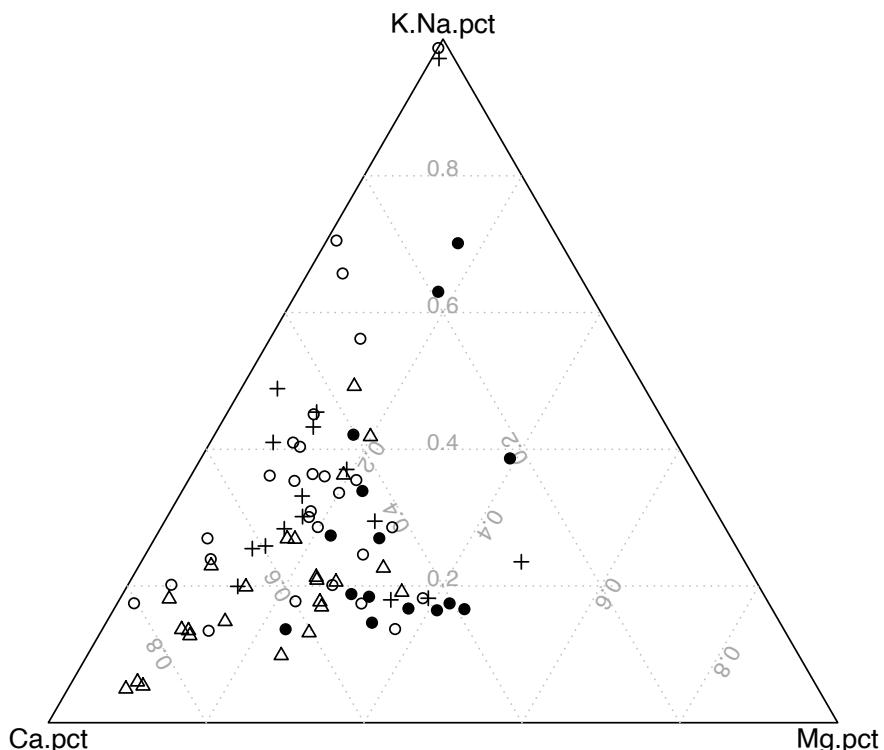
Trilinear diagrams have been used within the geosciences since the early 1900s. When three variables for a single observation sum to 100 percent, they can be represented as one point on a triangular (trilinear) diagram. An example of a trilinear diagram is shown in figure 2.32, where the cation compositions of the four geologic zones of the Groundwater Ambient Monitoring and Assessment (GAMA) Program conducted by the USGS in California. For the Sierra Nevada study unit of GAMA, the groundwater quality data are plotted against three major cation axes (Shelton and others, 2010). The concentration of cation  $i$ , expressed in milliequivalents, is denoted  $m_i$  for the three most prevalent cations in the sample study. The percentage composition of cation  $i$ , in the sample, is denoted as  $c_i$ , which is computed as

$$c_i = 100 \cdot m_i / (m_1 + m_2 + m_3) . \quad (2.5)$$

For example, if  $\text{Ca} = 0.80 \text{ meq}$ ,  $\text{Mg} = 0.26 \text{ meq}$ , and  $\text{Na} + \text{K} = 0.89 \text{ meq}$ , the percentages are  $\text{Ca} = 41$  percent,  $\text{Mg} = 13$  percent, and  $[\text{Na} + \text{K}] = 46$  percent of total milliequivalents. 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 2.31.** Star plots of site characteristics for 19 locations along the Exe estuary (from Warwick, 1971). Outliers such as sites 5 and 10 are seen to differ from the remaining sites owing to their low values for both interstitial salinity (Interstit.Salinity) and percent organics (X.Organics) composition.



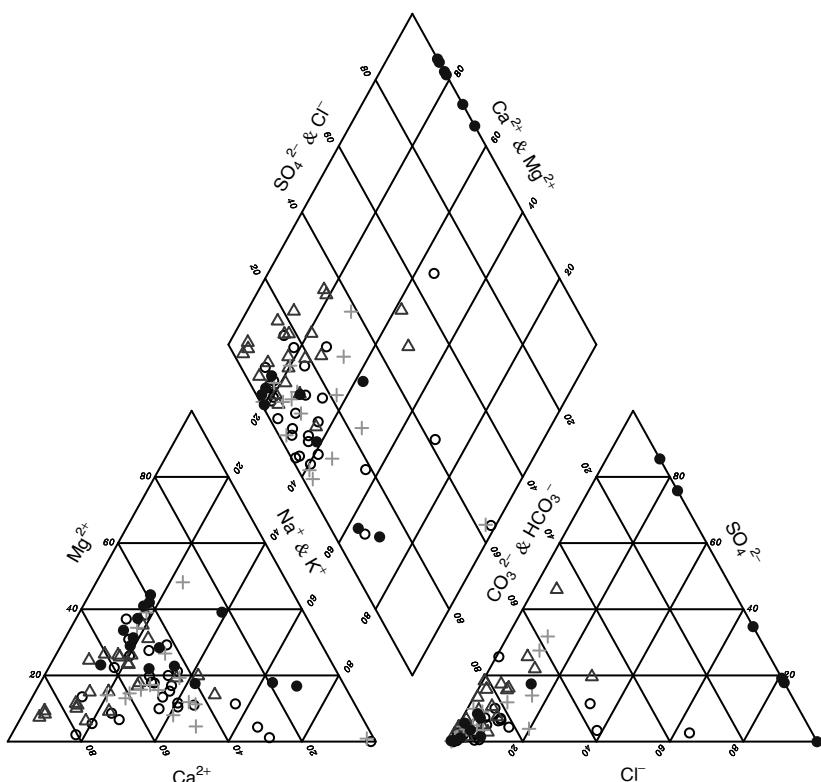
**Figure 2.32.** Trilinear diagram for groundwater cation composition in four geologic zones (each shown with a different symbol) of the Groundwater Ambient and Monitoring Assessment (GAMA) Program Sierra Nevada study unit (from Shelton and others, 2010). Units are percent milliequivalents (pct).

Piper (1944) applied these trilinear diagrams to both cation and anion compositions of water-quality data. He also combined both trilinear diagrams into a single summary diagram with the shape of a diamond (fig. 2.33). This diamond has four sides, two for cations and two for anions. However, it has only two independent axes, one for a cation (say  $\text{Ca}^{2+}$ + $\text{Mg}^{2+}$ ), and one for an anion (say  $\text{Cl}^-$ + $\text{SO}_4^{2-}$ ). If the ( $\text{Ca}^{2+}$ + $\text{Mg}^{2+}$ ) percentage is known, so is the ( $\text{Na}^+$ + $\text{K}^+$ ) percentage, as one cation is equal to 100 percent minus the other. The same is true for the anions. The collection of these three diagrams in the format shown in figure 2.33 is called a Piper diagram.

Piper diagrams have an advantage over Stiff and star diagrams in that observations are shown as points on a measured scale rather than as polygons. Therefore, similarities and differences in composition between numerous observations are 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 diagram to use will depend on their intended purpose.

## 2.4.4 Scatterplot Matrix

When there are multiple ( $p$ ) variables, all of their pairwise relations can be visualized by plotting a scatterplot for each of the  $p \cdot (p - 1)/2$  possible pairs of variables. These are then displayed together in a matrix. Obviously little detail can be discerned on any single plot within the matrix, but variables that are related can be grouped, linear versus nonlinear relations discerned, and so forth. Chambers and others (1983) describe the production and utility of scatterplot matrices in detail. They can be produced using the `pairs` function in base R.



**Figure 2.33.** Piper diagram of groundwater from the Sierra Nevada study unit of the Groundwater Ambient Monitoring and Assessment (GAMA) Program (Shelton and others, 2010).

A scatterplot matrix for the site characteristic data of Warwick (1971) is shown in figure 2.34. In the plot of median particle diameter (Med.Part.Diam) versus depth to sulfide layer (Dep.H2S.layer)—the two variables shown to have the highest correlation—there is a group of sites with low values for both variables. There is also a larger group of sites with varying, but generally higher, values for both median particle diameter and depth to sulfide layer. The Dep.H2S.layer could perhaps be modeled as two groups rather than as a continuous variable. The plot of Med.Part.Diam and percentage organics (X.Organics) appears to show an inverse relation between the variables rather than a linear one.

## 2.4.5 Biplots 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 (Borcard and others, 2011; Everitt and Hothorn, 2011). The following discussion provides an introduction on the use of principal components but does not describe how they are calculated. There are several R functions that serve this purpose; a simple one is `prcomp` in the base R package.

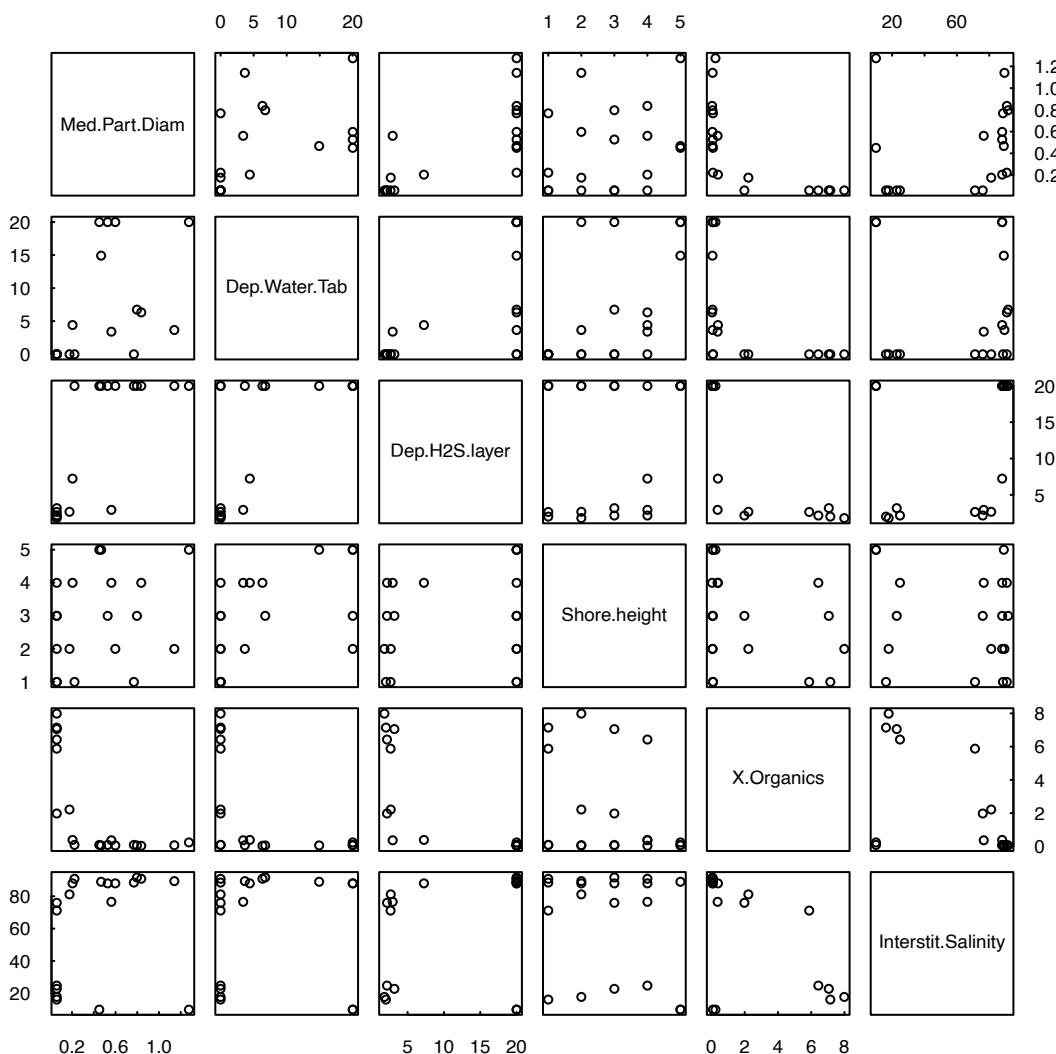


Figure 2.34. Scatterplot matrix showing the relations between six site characteristics from Warwick (1971).

Principal components are linear combinations of the  $p$  original variables to 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 through multidimensional space. The second principal component explains more of the remaining variance than any other axis that 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:

$pc1 = 0.75 \text{ Ca} + 0.80 \text{ Mg} + 0.10 \text{ Na} + 0.06 \text{ K}$	a calcareous axis?
$pc2 = 0.17 \text{ Ca} + 0.06 \text{ Mg} + 0.60 \text{ Na} + 0.80 \text{ K}$	a Na+K axis?
$pc3 = 0.40 \text{ Ca} - 0.25 \text{ Mg} - 0.10 \text{ Na} + 0.10 \text{ K}$	a Ca versus Mg axis?
$pc4 = 0.05 \text{ Ca} - 0.10 \text{ Mg} + 0.10 \text{ Na} + 0.20 \text{ K}$	residual noise

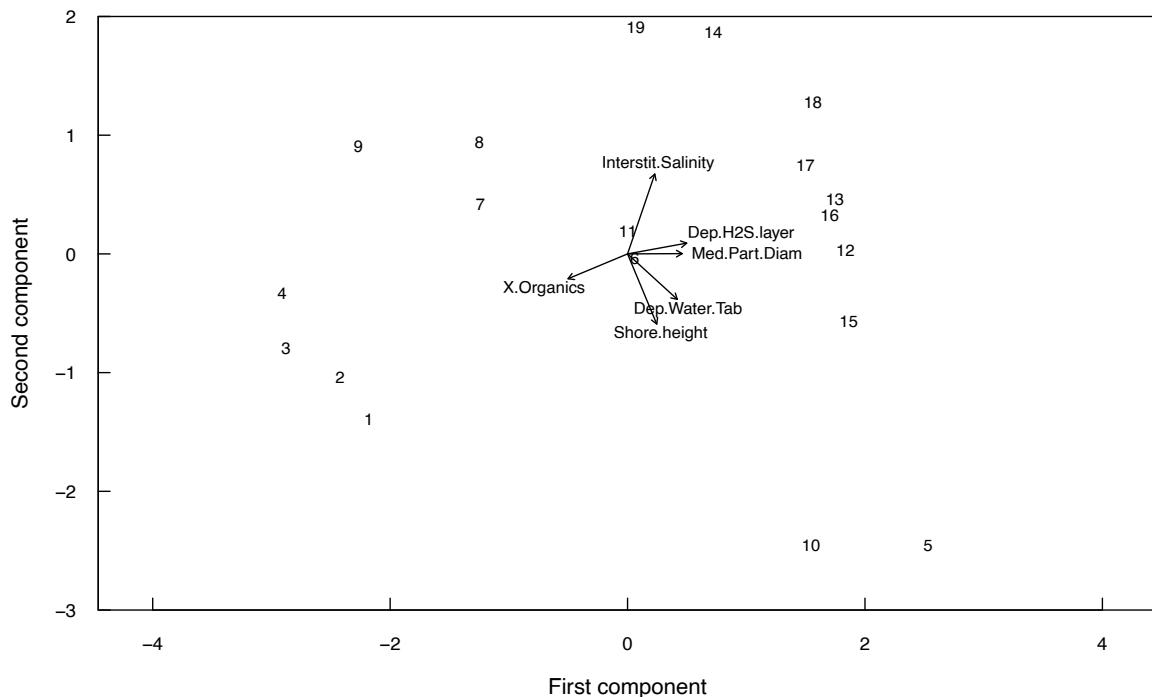
An observation that had milliequivalents of Ca=1.6, Mg=1.0, Na=1.3, and K=0.1 would have a value on  $pc1$  equal to  $(0.75 \cdot 1.6 + 0.8 \cdot 1.0 + 0.1 \cdot 1.3 + 0.06 \cdot 0.1) = 2.136$ , and similarly for the other new variables. At this point no reduction in dimensions has taken place, as each observation still has values along the  $p=4$  new  $pc$  axes, as they did for the four original axes.

Now, however, plots can be made of the locations of the observations (also called scores) oriented along the first two principal components axes. This is considered the most important view of data in multivariate space. A principal components biplot of the site characteristics data from Warwick (1971) is shown in figure 2.35. Two types of data are shown, hence the name biplot. The first type is the locations of observations on the plot, as shown by the numbers plotted for each observation. Observations near each other on the biplot have similar characteristics in multivariate space. For example, sites 5 and 10 are located near each other on the plot, and far away from the other locations. They were the outliers identified in the star chart of figure 2.31.

The second type of information on the biplot is the vector representing the direction of variables used in the analysis. Six attributes of each location were recorded in the field or were the result of lab measurements. These six variables are the basis for the similarities and differences between sites represented on the plot. The PCA biplot represents a slice through six-dimensional space, as defined by the six axes. Variables whose vectors are close to each other are strongly correlated, such as depth to the sulfide layer (Dep.H2S.layer) and median particle diameter (Med.ParDiam). Axes point in the direction of increasing value, so that as depth to the sulfide layer increases at the sites, so does the median particle diameter of the substrate. These two variables are also reasonably correlated with the percent organics (X.Organics) at the site, though the vector for percent organics is heading at an angle almost 180 degrees from the previous two, indicating a negative correlation. All three variables are correlated with each other, and so form one principal component or major direction for the data. Plotting both types of information on the same plot illustrates their relations—points (site locations) listed in the same direction the vector for percent organics is pointing have high percent organics compared to the other sites, in this case sites 1 through 4. These sites are in the opposite direction from the median particle diameter and depth to the sulfide layer variables because sites 1 to 4 have relatively low values for those variables. Sites 1 to 4 are organic rich silt/clay substrate locations that go quickly anoxic (high H<sub>2</sub>S) with depth.

## 2.4.6 Nonmetric Multidimensional Scaling

Nonmetric multidimensional scaling (NMDS) was developed by Kruskal (1964) for application to psychology, but has found frequent use in recent years in the discipline of ecology (Borcard and others, 2011). It is an alternative to, and often similar to, a PCA biplot. Its advantage over the PCA biplot is that all of the information in all of the variables is used to construct the plot, rather than using only the two best axes as defined by two principal components (Everitt and Hothorn, 2011). Its disadvantage in comparison to a PCA biplot is that distances are not measured in original scales, but in their ranks. This results in scales for the  $x$  and  $y$  axes that are arbitrary and may not even be shown on the plot. Because of this, NMDS is often considered a sketch of data relations rather than a scatterplot with defined scales.

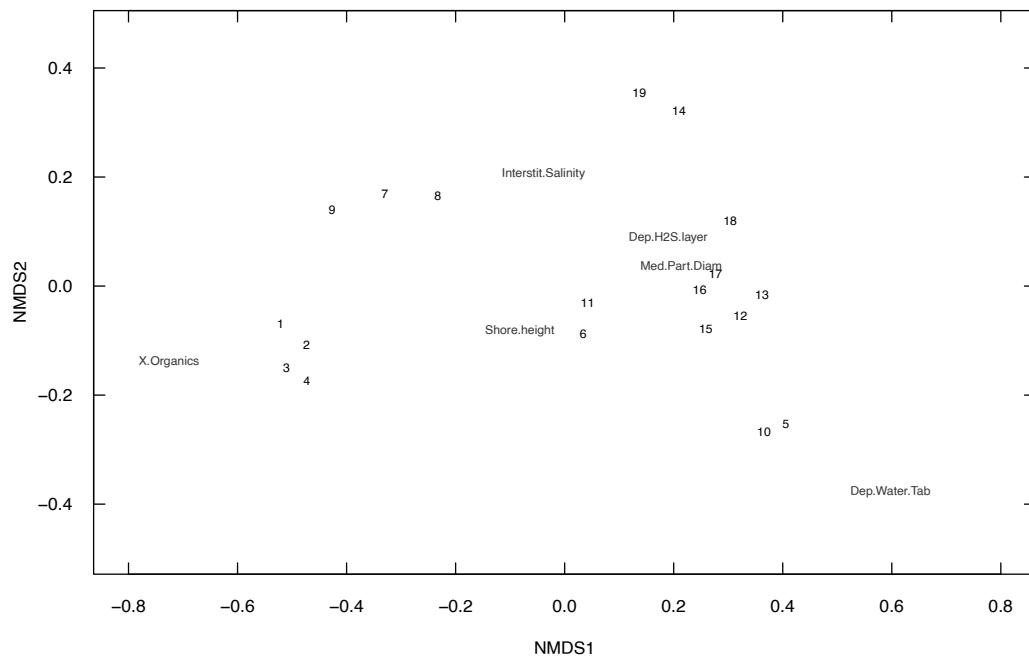


**Figure 2.35.** Principal component analysis (PCA) biplot of site characteristics along the Exe estuary (Warwick, 1971).

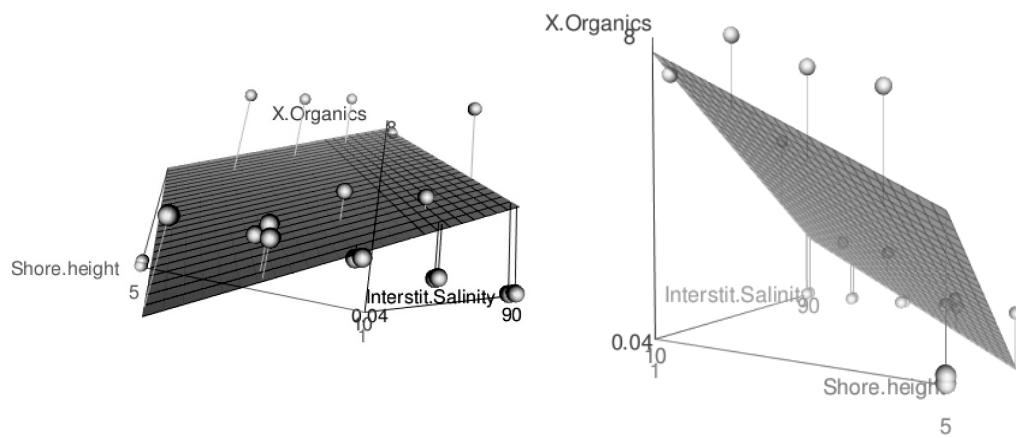
An NMDS for the site characteristic data of Warwick (1971) is shown in figure 2.36. Variable names indicate the direction of highest values for those variables, similar to the point of each vector shown on a biplot. The NMDS clearly shows sites 5 and 10 as outliers in terms of depth to the water table (Dep.Water.Tab). It shows the cluster of sites 1 to 4 as locations with high percentage organics. Data towards the center, near the (0,0) point on the map, are average in most variables. Clusters of sites indicate similarities in characteristics among those sites, and differences from characteristics at the other sites. NMDS and a PCA biplot are two of the best ways to visualize data in multivariate space, and together provide a valuable first look at complex datasets. They can both be computed using the `vegan` package of R (Oksanen and others, 2016).

## 2.4.7 Three-dimensional Rotation Plots

If only three variables are under consideration, software packages often will plot data in pseudo-three dimensions, and allow the axes to be rotated in space along all three dimensions. In this way the inter-relations between the three variables can be visually observed, data can be visually clustered into groups of similar observations, and outliers discerned. In figure 2.37, two of the many possible orientations for viewing three of the Exe site characteristic variables from Warwick (1971) are presented. The figures were generated using the R command `scatter3d` from the `car` package (Fox and Weisberg, 2011) which requires the `rgl` package (Adler and others, 2018). By rotating data around their three axes, patterns may be seen which would not be evident without a three-dimensional perspective and greater insight into the data is obtained.



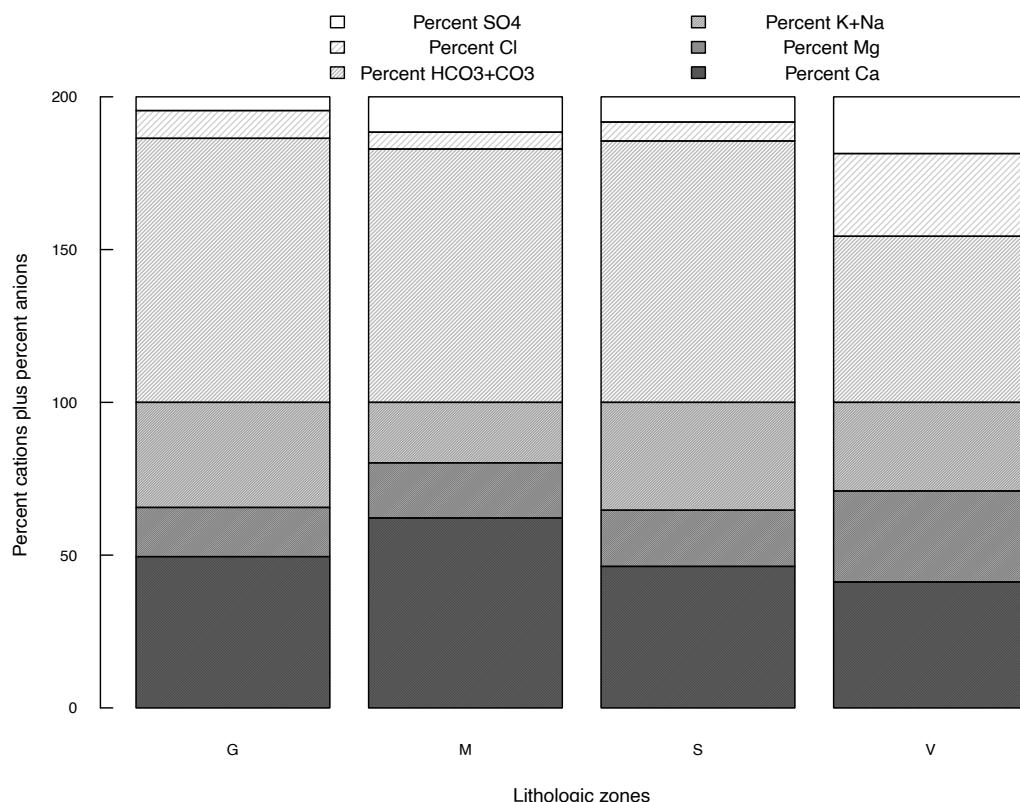
**Figure 2.36.** Nonmetric multidimensional scaling showing the relations among sites, and between sites and variables, using the six site characteristics of Warwick (1971).



**Figure 2.37.** Two three-dimensional plots of the site characteristics data of Warwick (1971).

## 2.4.8 Methods to Avoid

Two commonly used methods should generally be avoided, as they provide little ability to compare differences between groups of data; these are stacked bar charts and multiple pie charts. Both methods allow only coarse discrimination to be made between segments of the plot. For example, figure 2.38 is a stacked bar chart of the GAMA water-quality data previously shown as a trilinear plot (fig. 2.32). Note that only large differences between categories within a bar are capable of being discerned. Only the lowest anion (percent  $\text{HCO}_3 + \text{CO}_3$ ) and cation (percent Ca) categories have a common datum, so judgment of relative heights for the others is difficult for the human eye. For example, any difference in magnitude of percent K+Na in the G and S lithologic zones cannot be differentiated using 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. In figure 2.38 the mean values for each lithologic zone were computed, not for individual locations. Showing a separate bar for each of the many dozen locations in a given lithologic zone would be confusing. Multiple pie charts require similarly imprecise and difficult judgments of differences. Both stacked bar charts and pie charts can be replaced by any of the other methods shown in this chapter, improving insight and usefulness for data analysis.



**Figure 2.38.** Stacked bar charts of mean percent milliequivalents of anion and cations within the four Groundwater Ambient Monitoring and Assessment (GAMA) Program lithologic zones of Shelton and others (2010).

## Exercises

1. Annual peak discharges for the Otter Creek, at Middlebury, Vermont, are provided in the dataset `OtterCreek.RData`. For these data, draw the following plots.

- A. A histogram
- B. A boxplot
- C. A quantile plot using the Weibull plotting position  $k/(n+1)$ .

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

2. Arsenic concentrations (in parts per billion) were reported for groundwaters of southeastern New Hampshire. For the data shown below, draw the following plots.

- A. A boxplot
- B. A probability plot

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

<b>Arsenic concentration, in parts per billion</b>							
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							

3. Feth and others (1964) measured chemical compositions of waters in springs draining different rock types. Compare chloride concentrations from two of these rock types using a Q-Q plot. Also, plot another type of graph. Describe the similarities and differences in chloride between these two rock types. What characteristics are evident in each graph?

<b>Chloride concentration, in milligrams per liter</b>					
<b>Granodiorite</b>					
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
<b>Quartz monzonite</b>					
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	

# Chapter 3

## Describing Uncertainty

---

*The mean nitrate concentration in a shallow aquifer under agricultural land was calculated as 5.1 milligrams per liter (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 five-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 a result of the acid treatment, we wonder how unusual is it to have a well with a specific capacity of 15 gallons per minute per foot given our observations about the distribution of specific capacity values we see in the wells we have sampled?*

*An estimate of the 100-year flood, the 99th percentile of annual flood peaks, was determined to be 1,000 cubic meters per second ( $m^3/s$ ). 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 that described key attributes of a dataset, including sample estimates such as  $\bar{x}$  and  $s^2$ , of true and unknown population parameters, such as  $\mu$ , the population mean, and  $\sigma^2$ , the population variance. In this chapter, descriptions of the uncertainty or reliability of sample estimates are 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 prespecified value.

### 3.1 Definition of Interval Estimates

The sample median and sample mean are two types of estimates of the central tendency 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 datasets—X and Y—exist, both have a sample mean of 5 and contain the same number of observations. The Y data all cluster tightly around 5; the X data are much more variable. The point estimate of 5 for dataset X is less certain than the point estimate for dataset Y because of the greater variability in the X data. 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 that have a stated probability of containing the true population value. In general, we will be presenting two-sided intervals (where the probability of the true value being higher than the upper limit is assumed to be equal to the probability of its being lower than the lower limit). There are also one-sided intervals and these are discussed in later sections of this chapter. The intervals are wider for datasets having greater variability and the same number of data points. Thus, in the above example, an interval between 4.7 and 5.3 may have a 95-percent probability of containing the (unknown) true population mean of dataset 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 dataset 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 (often called tolerance intervals) will be discussed in section 3.7.

## 3.2 Interpretation of Interval Estimates

Suppose that the true population mean  $\mu$  of concentration of chloride in an aquifer was 5 milligrams per liter (mg/L). Also suppose that the true population variance  $\sigma^2$  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 1 per month) during a year, and the days on which sampling occurs are randomly chosen. From these 12 samples  $\bar{x}$  and  $s$  (the square root of  $s^2$ ) are computed. Although in reality only one set of 12 samples would be taken each year, using a computer simulation that has the true population characteristics and an assumption of normality, 12 days can be selected multiple times to illustrate the concept of an interval estimate. For each of 10 independent sets of 12 randomly selected samples, a two-sided confidence interval on the mean is computed using equations given in section 3.4.1. The results are shown in table 3.1 and figure 3.1. Note that just as the sample mean varies from sample to sample, so will the end points of the interval.

These 10 intervals are 90-percent 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. Thus, for the 10 intervals in the table, 9 are expected to include the true value and 1 is not. As shown in figure 3.1, this is in fact what happened. The sixth dataset drawn has a confidence interval from 5.23 to 5.97 and thus does not include the true mean of 5. When a one-time sampling occurs, the computed interval may or may not include the true, unknown population mean. The probability that the interval does include the true value is called the confidence level. The probability that this interval will not cover the true value is called the significance level,  $\alpha$ , which is computed as

$$\alpha = 1 - \text{confidence level} . \quad (3.1)$$

The width of a confidence interval is a function of the shape of the data distribution (its variability and skewness), the sample size, and 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 smaller interval. Thus, a 95-percent confidence interval will be wider than a 90-percent interval for the same data.

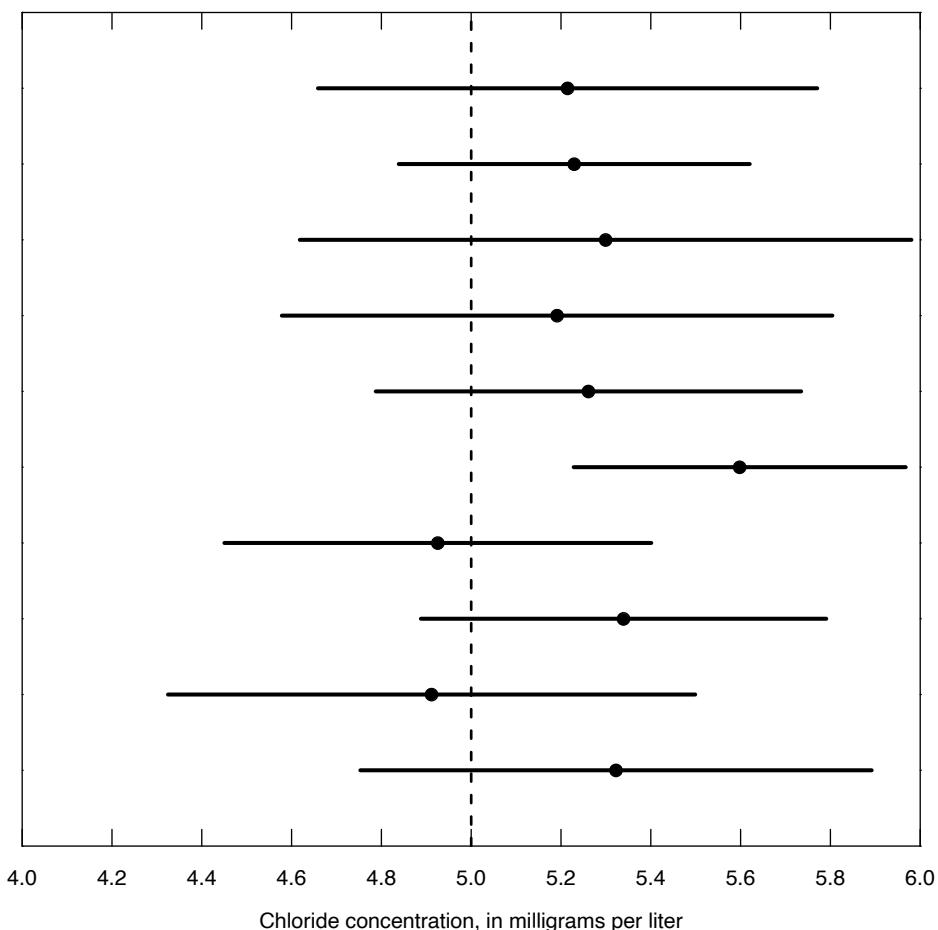
Symmetric confidence intervals on the mean are commonly computed when it is assumed that the data follow a normal distribution (see section 3.4.1). Even if the data are not normally distributed, the distribution of the sample mean will be approximately normal as long as sample sizes are large (say  $\geq 70$  observations for typical highly skewed environmental data; see U.S. Environmental Protection Agency [2002] or Boos and Hughes-Oliver [2000]). Confidence intervals assuming normality will then include the true mean  $100 \cdot (1 - \alpha)$ -percent 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 highly skewed and sample sizes are  $\leq 70$ , symmetric confidence intervals may not contain the mean  $100 \cdot (1 - \alpha)$ -percent of the time. Symmetric confidence intervals are computed for a skewed distribution as illustrated by the boxplot in figure 3.2.

Plots constructed from 10 datasets of 12 samples of chloride concentration, each sampled from a log normal distribution, show that the confidence intervals miss the true value of 1 more frequently than they should (4 times in 10 trials) in figure 3.3. 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.

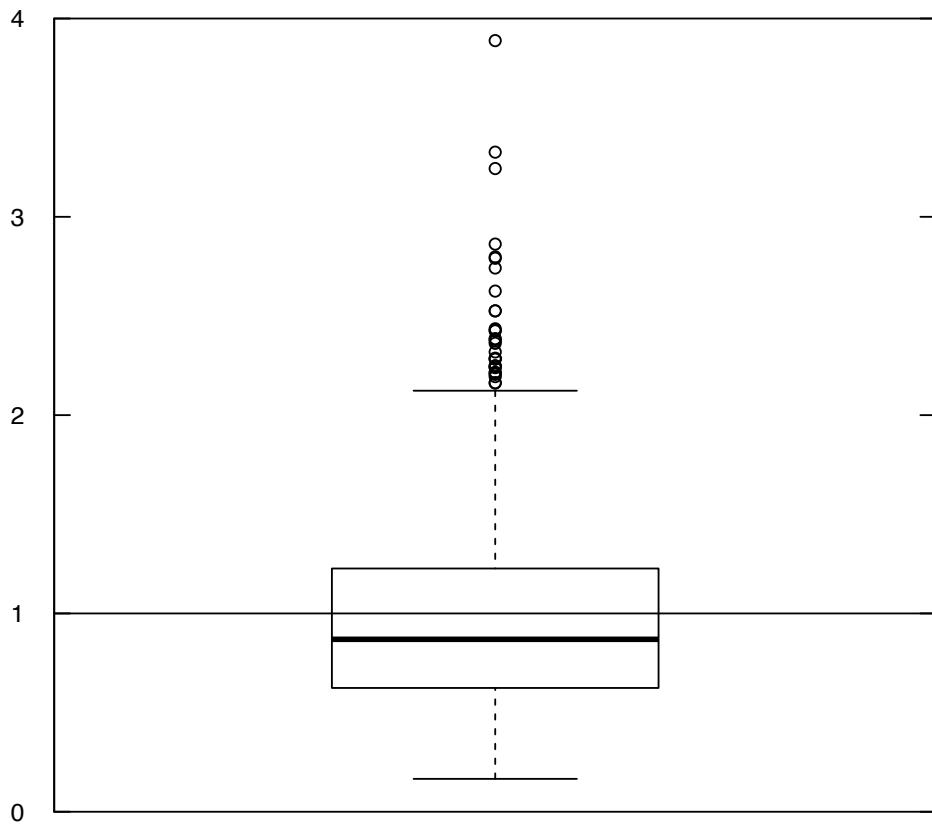
**Table 3.1.** Ten replicate datasets of 12 samples each of chloride concentrations, each with mean = 5 and standard deviation = 1. All units are in milligrams per liter.

[Xbar, sample mean; sd, sample standard deviation; lcl, lower confidence limit for the mean; ucl, upper confidence limit for the mean, where the confidence interval is a 90-percent two-sided interval]

Replicate	xbar	sd	lcl	ucl
1	5.21	1.072	4.66	5.77
2	5.23	0.754	4.84	5.62
3	5.30	1.314	4.62	5.98
4	5.19	1.182	4.58	5.80
5	5.26	0.914	4.79	5.73
6	5.60	0.713	5.23	5.97
7	4.93	0.917	4.45	5.40
8	5.34	0.871	4.89	5.79
9	4.91	1.132	4.32	5.50
10	5.32	1.098	4.75	5.89



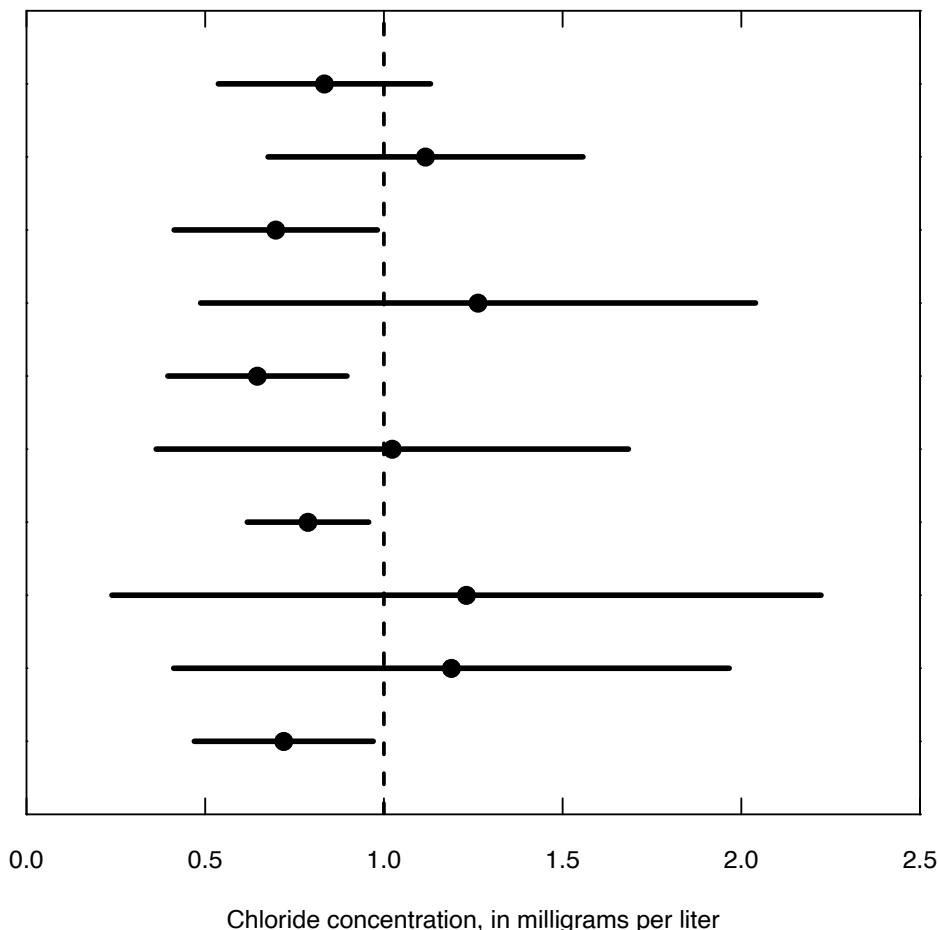
**Figure 3.1.** Ten 90-percent confidence intervals for normally distributed data with true mean = 5 and standard deviation = 1, in milligrams per liter. Dots indicate the sample mean from each sample.



**Figure 3.2.** Boxplot of a random sample of 1,000 observations from a lognormal distribution. Population mean = 1, population coefficient of variation = 1. The horizontal line that crosses the entire plot is the true population mean value. For definition of the boxplot features see chapter 2, section 2.1.3.

Throughout our discussion of confidence intervals we will only describe two-sided confidence intervals. For two-sided confidence intervals it is always assumed that the probability of the true value lying below  $c_{low}$  will be equal to the probability of the true value lying above  $c_{up}$ . There can also be one-sided confidence intervals, where our concern is strictly with trying to determine whether the true value is greater or less than some particular value. When we determine a one-sided confidence interval, the critical values we use are the cumulative distribution function evaluated at either  $\alpha$  or  $1 - \alpha$  rather than at  $\frac{\alpha}{2}$  or  $1 - \frac{\alpha}{2}$ .

A distribution that is commonly used in the definition of confidence intervals is the Student's  $t$ -distribution, which is based on the sampling properties of sample mean values. The Student's  $t$ -distribution has a parameter called the degrees of freedom which is generally  $n - 1$  where  $n$  is the sample size. Several of the formulas used in this chapter reference the critical levels of the  $t$ -distribution. For example,  $t_{\left(\frac{\alpha}{2}, n-1\right)}$  denotes the critical value of the Student's  $t$ -distribution for  $n - 1$  degrees of freedom, for which the probability of that variable being less than this value is  $\frac{\alpha}{2}$ . We can think of it as being the point on the distribution at which the tail area to the left of it is  $\frac{\alpha}{2}$ .



**Figure 3.3.** Ten 90-percent confidence intervals around a true mean of 1, each one based on a sample size of 12. Data are from a log normal distribution of mean = 1.0 and coefficient of variation = 1.0. Dots indicate the sample mean values. Four out of the 10 intervals do not include the true value.

### 3.3 Confidence Intervals for the Median

A confidence interval for the true population median may be computed in two ways: (1) without assuming the data follow any specific distribution (nonparametric; section 3.3.1.), or (2) assuming they follow a distribution such as the lognormal (parametric; section 3.3.2.).

#### 3.3.1 Nonparametric Interval Estimate for the Median

We will consider two nonparametric approaches to interval estimates for the median. The first is based on the binomial distribution and the second is a bootstrap method, which is a general approach to many estimation problems. Neither approach requires assumptions about the distribution of the random variable.

For the binomial approach we start by selecting the desired significance level  $\alpha$ , which is the acceptable risk of not including the true median. One-half of this risk ( $\alpha/2$ ) is assigned to each end of the interval. To compute the confidence interval for the median we use the cumulative distribution function (cdf) of the binomial distribution (in R that is the function `qbinom`). To determine the  $100 \cdot (1 - \alpha)$ -percent confidence interval we use the `qbinom` function to determine the critical values of the ranks of the dataset that correspond to  $\alpha/2$  and  $1 - (\alpha/2)$  points on the binomial cumulative distribution function. These critical

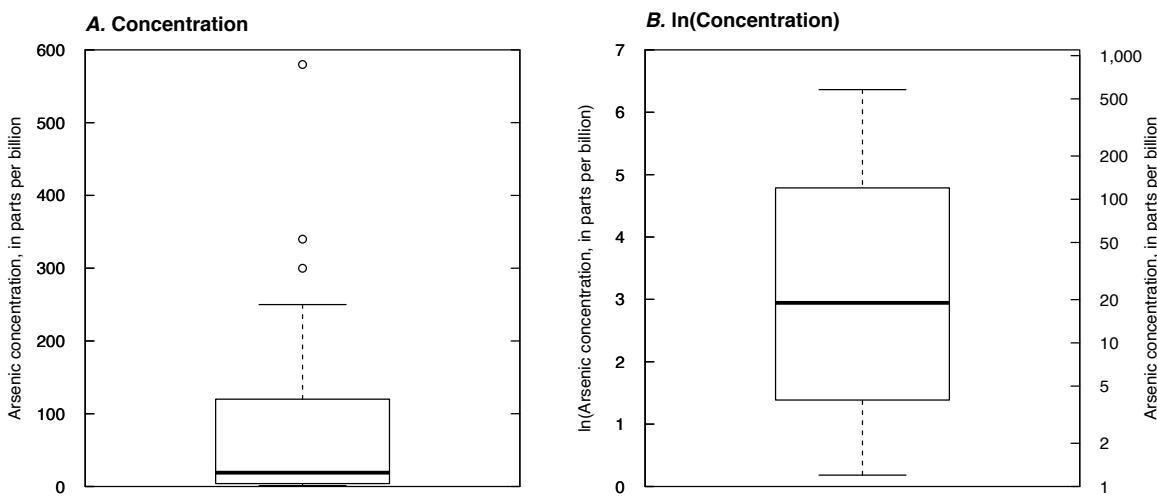
values of the ranks are then associated with their data values to form the upper and lower confidence limits for the median. We use the binomial distribution to answer the following question: How likely is it that the true population median,  $c_{0.5}$ , would be such that  $k$  of the  $n$  observed data would be above  $c_{0.5}$  and  $n-k$  below  $c_{0.5}$ , where for example,  $k$  could be 0, 1, 2, 3,..., 25 out of  $n=25$ ? The binomial distribution with  $\text{prob} = 0.5$  is used because the probability of being above the population median is 0.5. The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data. Nonparametric intervals cannot exactly produce the desired confidence level when the sample sizes are small. This is because the possible values are discrete, jumping from one data value to the next at the ends of the intervals. However, confidence intervals close to those desired are available for all but the smallest sample sizes. The process of computing the confidence interval is best illustrated with an example.

### Example 3.1. Nonparametric interval estimate of the median.

The 25 arsenic concentrations in table 3.2 were reported for groundwaters of southeastern New Hampshire (Boudette and others, 1985). A boxplot of the data is shown in figure 3.4. Compute the  $\alpha=0.05$  interval estimate of the median concentration.

**Table 3.2.** Arsenic concentrations (in parts per billion) for groundwaters of southeastern New Hampshire (from Boudette and others, 1985), ranked in ascending order.

Rank	Value	Rank	Value	Rank	Value
1	1.3	10	9.5	19	120
2	1.5	11	12	20	190
3	1.8	12	14	21	240
4	2.6	13	19	22	250
5	2.8	14	23	23	300
6	3.5	15	41	24	340
7	4.0	16	80	25	580
8	4.8	17	100		
9	8.0	18	110		



**Figure 3.4.** Boxplots of the (A) original and (B) log-transformed arsenic data from Boudette and others (1985) used in example 3.1.

The sample median  $\hat{c}_{0.5} = 19$ , which is the 13th observation ranked from smallest to largest in this sample size of 25. The binomial distribution is used to determine the 95-percent confidence interval for the true median concentration,  $c_{0.5}$ . We obtain the critical values from the `qbinom` function as `qbinom(p=c(0.025, 0.975), size=25, prob=0.5)`. Because we are focused on the median here, the `prob` value is always 0.5 for this calculation. For the population median, half the population values are above the median and half below. The  $\alpha$  value here is 0.05 and thus the end points of the confidence intervals are at  $\alpha/2$  and at  $1-(\alpha/2)$ , which are 0.025 and 0.975, respectively. The values returned by this function are 8 and 17, which are the ranks of the two end points. We can then compute the concentration values that are associated with these two ranks as follows.

```
> x <- c(1.3, 1.5, 1.8, 2.6, 2.8, 3.5, 4.0, 4.8, 8.0, 9.5, 12, 14,
+       19, 23, 41, 80, 100, 110, 120, 190, 240, 250, 300, 340, 580)
> indexRanks <- qbinom(c(0.025, 0.975), length(x), prob = 0.5)
> indexRanks
[1] 8 17
> x <- sort(x) # not actually needed, values are sorted already
> x[indexRanks]
[1] 4.8 100.0
```

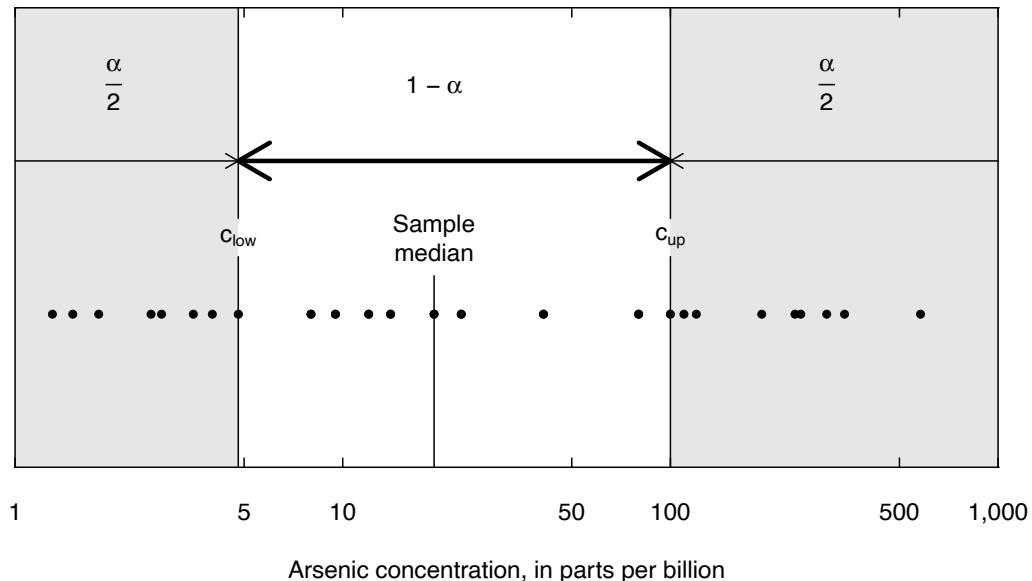
This code indicates that the lower and upper confidence intervals are at ranks 8 and 17, and that these translate to concentration values of 4.8 and 100 (the 8th and 17th values on the sorted list of concentration values in table 3.2). Because the sample size is relatively small ( $n=25$ ) we know that the interval will not be an exact 95-percent confidence interval. We can compute the probability that the interval will contain the true value using the `dbinom` function. This function returns the probability density of the binomial distribution, and we can sum the density values from 8 through 17 to determine the probability of the true median being in the range of the 8th through 17th values in the sorted vector of values.

```
> sum(dbinom(8:17, 25, 0.5))
[1] 0.9567147
```

The result tells us that the true probability for this range is 0.9567, which is very close to the desired probability of 0.95. Thus, one could say the closed interval [4.8, 100] is the best approximation to a 95-percent confidence interval for the median. Note that the bracket notation ([]) means greater than or equal to, whereas an open parenthesis is used to indicate greater than values. This means that the probability that  $c_{0.5}$  will be less than the 8th ranked sample value is  $\leq 0.025$  and similarly the probability that  $c_{0.5}$  will be greater than the 17th ranked sample value is also  $\leq 0.025$ . Thus, we can state that a 95-percent confidence interval for the median is [4.8, 100] because 4.8 and 100 are the 8th and 17th ranked values in the sample. The results of these computations are shown in figure 3.5; note the substantial amount of asymmetry in the confidence interval, which is what we would expect given the asymmetry of the full sample.

An alternative to the binomial distribution-based approach is to use bootstrapping. Bootstrapping (Efron and Tibshirani, 1994) is one of many computer-intensive methods that uses the observed data to represent the probability distribution from which the data were drawn, rather than assuming a normal or other theoretical distribution. Because of its robustness and coverage accuracy, the bootstrap method is commonly preferred for computing a confidence interval, especially when data are skewed or the distribution is unknown. The method also works well for data that do follow a specific distribution—in that case it returns interval endpoints very similar to those based on a distribution, for example *t*-intervals for the normal distribution (discussed in section 3.4).

The bootstrap method consists of repeatedly resampling from the dataset, with replacement. The term “with replacement” means that each observation that has been selected is replaced in the dataset, which means that it can be selected multiple times in the sampling process. Two thousand to 10,000 resamples of the data are commonly used, and for many statistics it takes a small amount of computational time.



**Figure 3.5.** Plot of the 95-percent confidence interval for the true median in example 3.1. The dots represent the 25 observations in the arsenic dataset on a logarithmic scale. The sample median (19 parts per billion [ppb]) and upper ( $c_{up}$ ) and lower ( $c_{low}$ ) bounds on the confidence interval for the true median (4.8 and 100 ppb) are shown. The confidence interval is computed using  $\alpha=0.05$ . Thus, the probability that the population median will be greater than  $c_{up}$  is 0.025 and the probability that the population median is less than  $c_{low}$  is also 0.025.

For more complex problems, the bootstrap process can take substantial amounts of time and selecting the number of resamples entails a trade-off between computer time and accuracy. Each resample is a random selection (with replacement) of the identical number of observations found in the original data. Each observation in the dataset has an equal probability of being chosen for each resample. For example, with 25 observations a bootstrap replicate is generated consisting of 25 observations, all of which came from the original 25. Some of the original 25 may have been chosen more than once and some may not have been chosen at all. The difference between the original data and the many resampled datasets becomes a representation of the noise in the data. For each resample, the median, mean, or other statistic to be estimated is calculated and stored, resulting in 2,000–10,000 estimates of the desired statistic. These estimates represent the sampling distribution of that statistic and from the entire set a confidence interval can be computed.

The bootstrap method discussed here is the percentile bootstrap; there are also other variations of bootstrap intervals. The percentile bootstrap gets its name because it uses a percentile of the thousands of estimates of the desired statistic for its final estimate. For example, suppose we want to estimate the two-sided 95-percent confidence interval for the median and that we chose to use 2,000 bootstrap replicate estimates of the median. We would compute the 2,000 estimates and sort them from smallest to largest. In this situation the lower confidence limit ( $c_{low}$ ) would be the 50th ranked value (because  $50=0.025 \cdot 2,000$ ) and the upper confidence limit ( $c_{up}$ ) would be the 1,950th ranked value (because  $1,950=0.975 \cdot 2,000$ ). These values can be determined using the `quantile` function in R. The selection of a plotting position (set by the `type` argument in the `quantile` function) has a very small effect on the result. Note that for a variable that is always positive the  $c_{low}$  will always be positive.

Bootstrapping requires no assumption of a distributional shape, but it does require sufficient data to represent the population well. The data itself becomes the estimate of the distribution, and like all statistics, it performs better with more data.

When there is a high positive skewness the percentile bootstrap value for  $c_{up}$  is often too low. This can be corrected by using an adjustment for skewness and bias called the bca bootstrap (Efron and Tibshirani, 1994); see the book for a description of the method for adjustment. Here we will demonstrate bootstrapping using the percentile bootstrap.

A bootstrap method for obtaining confidence intervals for the median uses the `BootMedian` code found in the supplemental material for chapter 3 (SM.3). The `BootMedian` code is designed to simplify obtaining the bootstrap confidence interval for the median by calling the `boot.ci` function in the R `boot` package in a manner that is tailored to the problem of obtaining confidence intervals for the median. The `boot.ci` function is very general and can be used to estimate confidence intervals on many different statistics. Below is an example for determining the 95-percent confidence interval using the arsenic data from example 3.1.

**Example 3.2. Bootstrap confidence interval for the median.**

```
> library(boot)
> source("BootMedian.R")
> arsenic <- c(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)
> BootMedian(arsenic)
```

Bootstrap Confidence Intervals of the Median of arsenic

Using boot package in R

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 10000 bootstrap replicates

CALL :

```
boot.ci(boot.out = B, conf = Conf, type = TYPE)
```

Intervals :

Level	Percentile
-------	------------

95%	( 4.8, 110.0 )
-----	----------------

Calculations and Intervals on Original Scale

In this particular case the bootstrap produces a result that is very close to the nonparametric interval estimate calculated in example 3.1.

### 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 (eqs. 3.2 and 3.3). These are parametric estimates of the median and its confidence interval because they are made using an assumption about the underlying distribution (in this case, the assumed distribution is lognormal). When the data are truly lognormal, the geometric mean and interval would be more efficient (a shorter interval) measures of the median and its confidence interval than the nonparametric sample estimates of section 3.3.1. The sample median and its interval are more appropriate and more efficient if the logarithms of the data still exhibit skewness and (or) outliers.

$$\text{Let } GM_x = \exp(\bar{y}), \quad (3.2)$$

where

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}, \text{ and}$$

$$y_i = \ln(x_i);$$

then the lower and upper confidence intervals for median are

$$\exp(\bar{y} - t_{(a/2,n-1)} \sqrt{s_y^2 / n}) \leq GM_x \leq \exp(\bar{y} + t_{(a/2,n-1)} \sqrt{s_y^2 / n}), \quad (3.3)$$

where  $s_y^2$  is the sample variance of  $y$ , and  $t_{(a/2,n-1)}$  is the critical value of the  $t$ -distribution with  $n-1$  degrees of freedom and a cumulative probability value of  $\alpha/2$  (see section 3.4.1. for more discussion of the use of the  $t$ -distribution). Using the same arsenic concentration dataset as in table 3.2, we now take the natural logarithms of each value (table 3.3).

The mean of the log-transformed data is 3.17, with a standard deviation of 1.96. Box plots of the original and log-transformed data are shown in figure 3.4. Clearly, the log-transformed data are much closer to being symmetric and are well approximated by a normal distribution.

From equations 3.2 and 3.3, the geometric mean and its 95-percent confidence interval are

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

$$\exp(3.17 - 2.064 \cdot \sqrt{1.96^2 / 25}) \leq GM_C \leq \exp(3.17 + 2.064 \cdot \sqrt{1.96^2 / 25}),$$

$$\exp(2.36) \leq GM_C \leq \exp(3.98), \text{ and}$$

$$10.6 \leq GM_C \leq 53.5.$$

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.

**Table 3.3.** Log-transformed arsenic concentrations (in parts per billion) for groundwaters of southeastern New Hampshire (from Boudette and others, 1985), ranked in ascending order.

Rank	Value	Rank	Value	Rank	Value
1	0.262	10	2.251	19	4.787
2	0.405	11	2.485	20	5.247
3	0.588	12	2.639	21	5.481
4	0.956	13	2.944	22	5.521
5	1.030	14	3.135	23	5.704
6	1.253	15	3.714	24	5.829
7	1.387	16	4.382	25	6.363
8	1.569	17	4.605		
9	2.079	18	4.700		

## 3.4 Confidence Intervals for the Mean

Interval estimates may also be computed for the true population mean  $\mu$ . These are appropriate if the center of mass of the data is the statistic of interest (see chap. 1). Intervals symmetric around the sample mean  $\bar{X}$  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 increase, even though the data may not be normally distributed. This property is called the Central Limit Theorem (Conover, 1999) and it holds for data that follow a distribution having finite variance. As such, the theorem includes most distributions of interest in water resources. For smaller sample sizes, however, the mean will not be normally distributed unless the data themselves are normally distributed. As skewness of the data increases, 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 as many as 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 equation 3.4.

$$\bar{x} + t_{\left(\frac{\alpha}{2}, n-1\right)} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + t_{\left(1 - \frac{\alpha}{2}, n-1\right)} \cdot \sqrt{\frac{s^2}{n}} \quad (3.4)$$

If  $1-\alpha$  were the desired confidence level and the sample size was  $n$ , then the critical  $t$ -values would be  $t_{\left(\frac{\alpha}{2}, n-1\right)}$  and  $t_{\left(1 - \frac{\alpha}{2}, n-1\right)}$ . For example, if one wanted a 95-percent confidence interval, then  $\alpha=0.05$ . The critical values can be found on  $t$ -distribution tables or could be computed with the `qt` function in R. If the sample size,  $n$ , was 25, they would be `qt(0.025, 24)` and `qt(0.975, 24)` which are -2.064 and +2.064 respectively. The width of the confidence interval is a function of these critical values, the sample standard deviation of the data, and the sample size. When the sample size is small ( $n < 70$ ) and the data are highly 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. In some cases, the lower limit of the confidence interval may be less than zero. A negative endpoint for a confidence interval for data that cannot be negative is a clear signal that the assumption of a symmetric confidence interval is not appropriate. For such data, assuming a lognormal distribution as described in section 3.4.2. will probably result in more realistic confidence intervals.

#### Example 3.3. Symmetric confidence interval for the mean.

Using the data from example 3.1, the sample mean arsenic concentration,  $\bar{x} = 98.4$  parts per billion (ppb), is the point estimate for the true unknown population mean,  $\mu$ . The standard deviation of the arsenic concentrations,  $s$ , is 144.7 ppb. Using equation 3.4, a 95-percent confidence interval ( $\alpha=0.05$ ) for the true mean,  $\mu$ , is

$$98.4 - 2.064 \cdot \sqrt{\frac{144.7^2}{25}} \leq \mu \leq 98.4 + 2.064 \cdot \sqrt{\frac{144.7^2}{25}},$$

$$38.7 \leq \mu \leq 158.1.$$

Thus, there is a 95-percent probability that the interval between 38.7 and 158.1 ppb contains the true population mean assuming that a symmetric confidence interval is appropriate. Note that this confidence interval is, like the sample mean  $\bar{x}$ , sensitive to the highest data values. If the largest value of 580 ppb were changed to 380 ppb, the median and its 95-percent confidence interval would be unaffected. However,  $\bar{x}$  would change to 90.4 ppb, with a 95-percent interval estimate for  $\mu$  from 40.7 ppb to 140.1 ppb.

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

Means and confidence intervals may also be computed for the case where the logarithms  $y = \ln(x)$  approximate a normal distribution. If the logarithmically transformed data are approximately normal, this approach will give a more reliable (lower variance) estimate of the mean than will computation of the usual sample mean without transformation. There may be other transformations available that can be considered if the logarithm doesn't cause the data to be approximately normal.

An estimate of the mean of the original  $x$  variable is  $\hat{\mu}_x = \exp(\bar{y} + 0.5 \cdot s_y^2)$  where  $\bar{y}$  is the sample mean of the logarithms of  $x$ , and  $s_y^2$  is the sample variance of the logarithms of  $x$  (Aitchison and Brown, 1981). This only holds true when the logarithms are normally distributed. For small sample sizes and (or) large estimated variances this estimate of the mean is biased (Bradu and Mundlak, 1970). However, for small  $s_y^2$  and large sample sizes the bias is negligible. See chapter 9 for more information on the bias of this estimator.

The confidence interval around  $\hat{\mu}_x$  is not the interval estimate computed for the geometric mean in equation 3.3. It cannot be computed simply by exponentiating the end points of 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. A better estimator of the confidence interval about the mean for skewed data is given by bootstrapping, which is discussed in section 3.4.3.

### 3.4.3 Bootstrap Confidence Interval for the Mean for Cases with Small Sample Sizes or Highly Skewed Data

Just as we used the bootstrap to develop confidence intervals for the median, in section 3.3.1., we can also use the bootstrap to develop confidence intervals for the mean. The following example illustrates how that is done.

#### Example 3.4. Confidence intervals for the mean, using the bootstrap method.

Using the `boot` library in R, 25 values from the 25 observations in the arsenic dataset used in example 3.1 are selected and their mean is computed (sampling with replacement). This is repeated 2,000 times and a two-sided 95-percent confidence interval for the mean is the 0.025 · 2,000th and 0.975 · 2,000th ordered resample estimates for the mean.

The R code that could be used to develop the percentile estimate of the confidence interval is as follows.

```
> x <- c(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)
> library(boot)
> mean1 <- function(x, i) {mean(x[i])}
> set.seed(1)
> boot1 <- boot(x, statistic = mean1, R = 2000)
> boot.ci(boot1, conf = 0.95, type = "perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates
```

CALL :

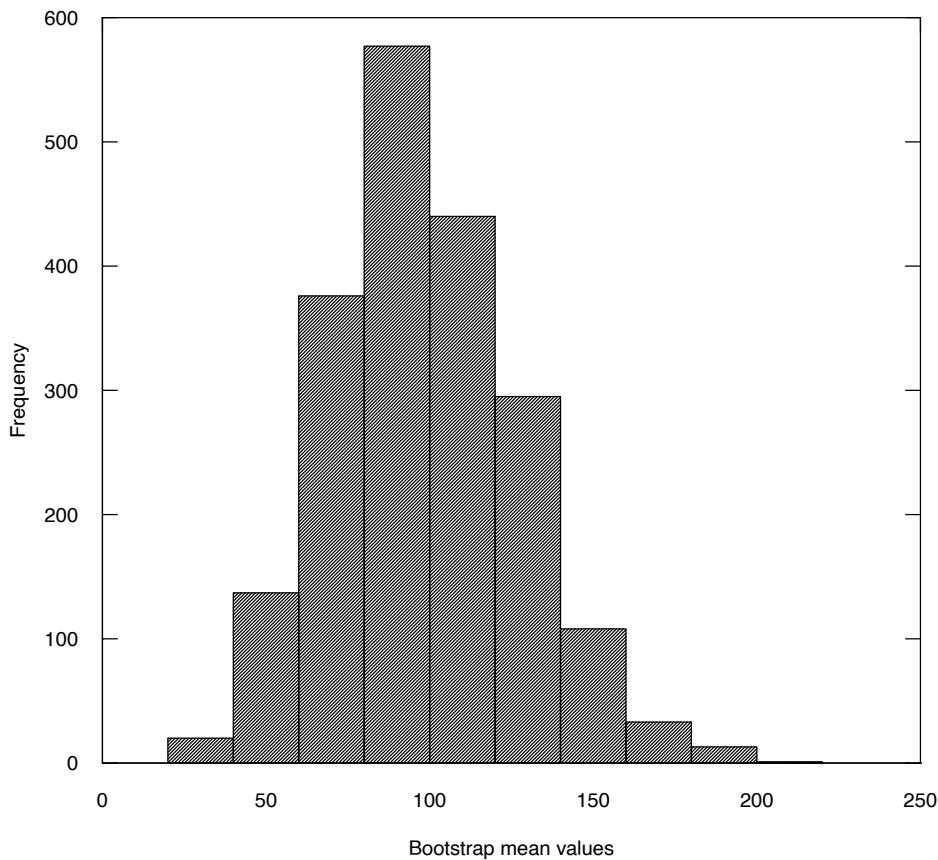
```
boot.ci(boot.out = boot1, conf = 0.95, type = "perc")
```

Intervals :

Level	Percentile
-------	------------

95%	( 47.78, 159.70 )
-----	-------------------

Calculations and Intervals on Original Scale



**Figure 3.6.** Histogram of bootstrapped estimates of the mean of arsenic concentrations used in example 3.1.

**Table 3.4.** Comparison of 95-percent confidence interval estimators for various measures of central tendency for the arsenic data, in parts per billion.

[ $c_{low}$ , lower confidence interval;  $c_{up}$ , upper confidence interval]

Parameter and estimation method	Estimate	$c_{low}$	$c_{up}$
Mean using $t$ -interval	98.4	38.7	158.1
Median using binomial confidence interval	19	4.8	100
Geometric mean based on retransformation of $t$ -interval estimates on the logs	23.8	10.6	53.5
Mean using percentile bootstrap	98.4	47.78	159.70

In this case the output says that the interval is from 47.78 ppb to 159.70 ppb, although another run of this procedure with a different random number seed (for example, using `set.seed(2)`) or with a different number of repetitions (say `R = 10000`) will generate slightly different results. This demonstrates one disadvantage of bootstrap methods, which is that they do not produce the exact same results each time they are used. The bootstrap results can be visualized with a histogram of the 2,000 bootstrap replicate values for the mean (fig. 3.6). In practice, for simple types of statistics such as the mean, doing 10,000 replicates is sensible because it takes up very little computer time. Bootstrap estimation for very complex statistical calculations may need to use fewer replicates to be practical, but using a smaller number of replicates will cause a slight loss of precision. A list of the estimates and 95-percent confidence limits for various measures of central tendency for this dataset are listed in table 3.4.

Bootstrap estimation is an important, but complex, tool in statistics. There are many options and variations on the technique and this text will not attempt to cover these, but this small example may help demonstrate the value of the bootstrap concept for highly skewed data, which are common in environmental and water resources data. The bootstrap method can be applied to any function of the data.

## 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. The scientist must select an  $\alpha$  level for the interval (the probability that a value falls outside the interval). Prediction intervals contain  $100 \cdot (1 - \alpha)$  percent of the data distribution where  $100 \cdot \alpha$  percent are outside of the interval. If a new observation comes from the same distribution as previously measured data, there is a  $100 \cdot \alpha$  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 under certain assumptions about the distribution. How likely this is depends on the choice of  $\alpha$  made by the scientist. Prediction intervals should never be used to exclude data points from an analysis, they merely help to indicate whether the data points are unusual in light of some existing set of data.

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 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 methods for estimating confidence intervals are mistakenly used to estimate 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. Nonparametric intervals do not require 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 distribution. Both two-sided and one-sided prediction intervals are described in section 3.6. 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. Such comparisons require hypothesis tests, which are introduced in chapter 5.

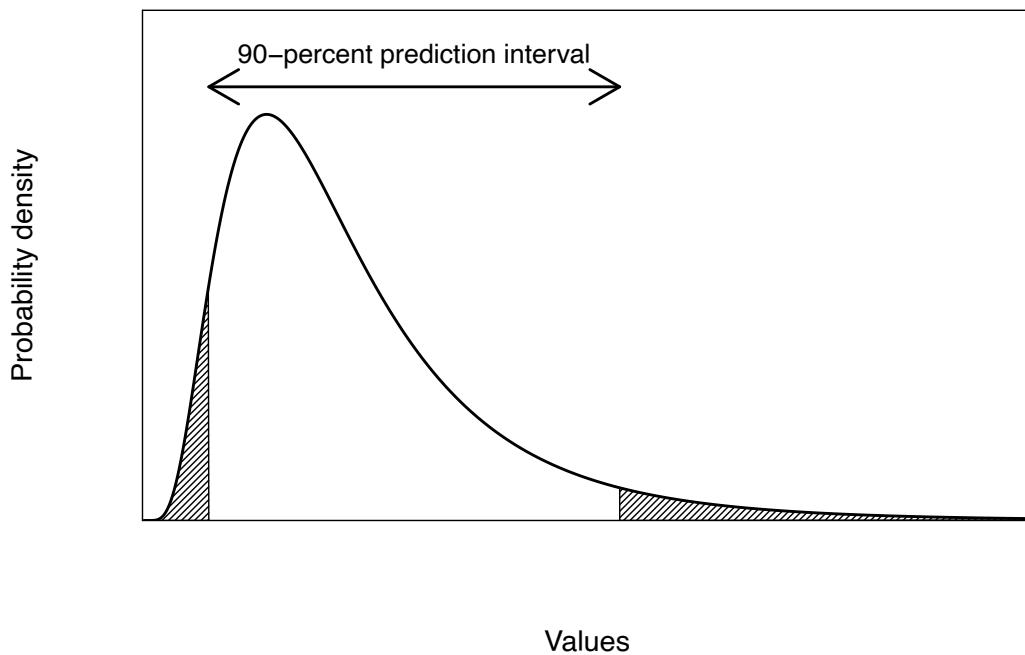
### 3.5.1 Two-sided Nonparametric Prediction Interval

The nonparametric prediction interval is simply the interval between the  $\alpha/2$  and  $1 - (\alpha/2)$  percentiles of the distribution (fig. 3.7). This interval contains  $100 \cdot (1 - \alpha)$  percent of the data, where  $100 \cdot \alpha$  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  $(100 \cdot \alpha)$ -percent chance that it will lie outside of the prediction interval. The interval will reflect the shape of the data it is developed from and no assumptions about the distribution need be made. In R we can use the `quantile` function to compute the prediction interval directly from the data. For instance, using the arsenic data from example 3.1 with  $\alpha=0.1$  and the Weibull plotting position (`type = 6`, as discussed in chap. 1, section 1.3.2., and chap. 2, section 2.1.1.), the interval can be computed as

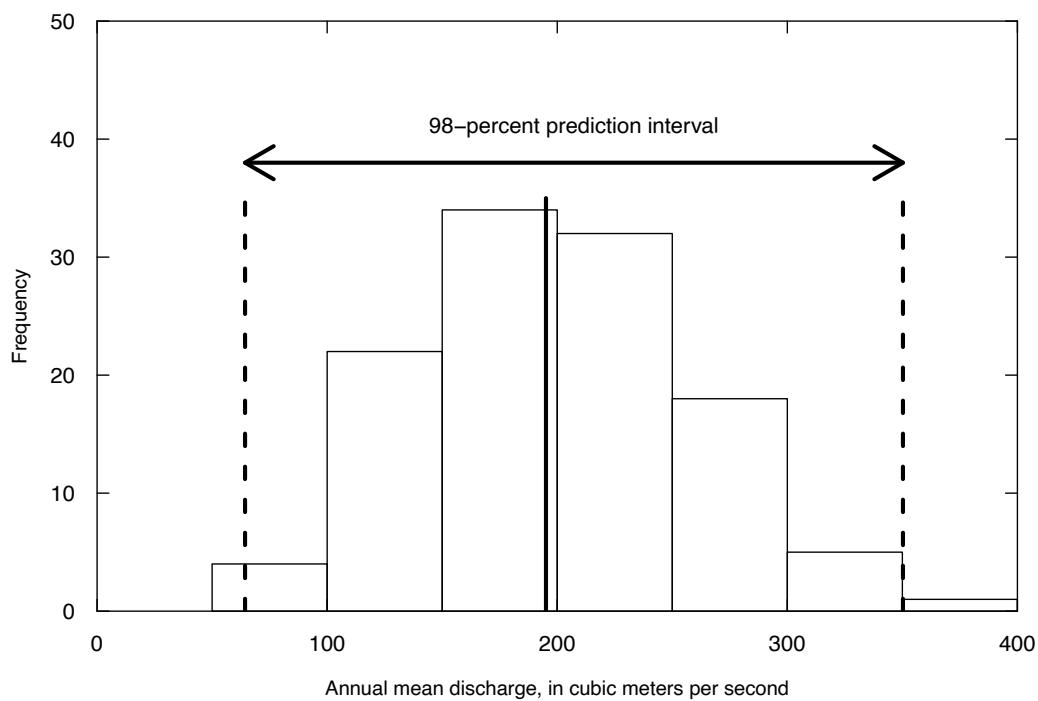
```
> quantile(x, prob = c(0.05, 0.95), type = 6)
```

With the arsenic dataset, the lower limit of the prediction interval is 1.36 ppb and the upper limit of the prediction interval is 508 ppb. If we conclude that any value that is less than 1.36 ppb or greater than 508 ppb comes from a different distribution than our original data, then there is a 10 percent chance of drawing that conclusion if, in fact, all of the values did come from the same population.

Figure 3.8 uses the James River discharge data introduced in chapter 2 to show the histogram of the dataset as well as the 98-percent prediction interval for annual mean discharge. The sample median is 195 cubic meters per second ( $\text{m}^3/\text{s}$ ) (shown with the solid line in fig. 3.8). The prediction interval includes



**Figure 3.7.** Example of a probability distribution showing the 90-percent prediction interval, with  $\alpha=0.10$ . Each of the two shaded areas in the tails has a probability of  $\alpha/2=0.05$ .



**Figure 3.8.** Histogram of the James River annual mean discharge dataset (see chap. 2). The solid vertical line is the sample median and the two vertical dashed lines are the lower and upper bound of the 98-percent prediction interval.

the interval from 64 to 350 m<sup>3</sup>/s. This implies that values less than 64 m<sup>3</sup>/s have a probability of about 1 percent of occurring in any given year and values greater than 350 m<sup>3</sup>/s also have a probability of about 1 percent of occurring in any given year. The prediction interval is the range of values between the two vertical dashed lines computed with `quantile(Q, prob = c(0.01, 0.99), type = 6)`. Note that the prediction interval is not symmetrical around the median.

### Example 3.5. Nonparametric prediction interval for the median.

Compute a 90-percent ( $\alpha=0.10$ ) prediction interval for the arsenic data used in example 3.1 without assuming the data follow any particular distribution. This can be computed in R as

```
> predInt <- quantile(x, prob = c(0.05, 0.95), type = 6)
```

The results (stored in `predInt`) are 1.36 ppb and 508 ppb. Note the high degree of asymmetry in this case. The median is 19 ppb, so the difference between the median and the lower prediction limit is about 18 ppb, but the difference between the median and the upper prediction limit is 489 ppb. Thus, a new observation less than 1.36 ppb or greater than 508 ppb can be considered as coming from a different distribution at a 10-percent significance level ( $\alpha=0.10$ ).

## 3.5.2 One-sided Nonparametric Prediction Interval

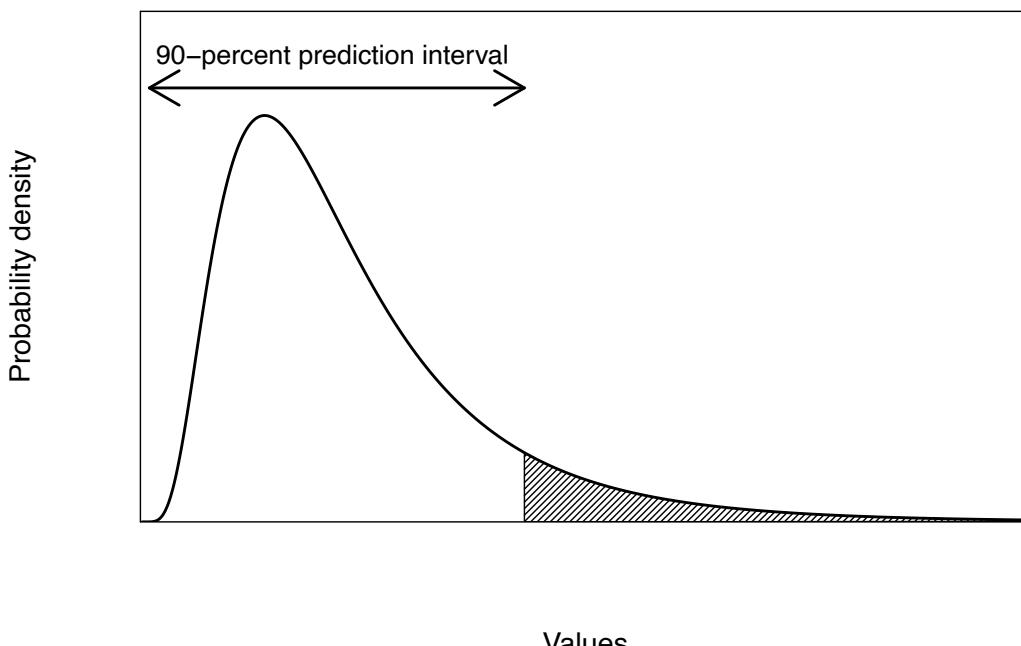
One-sided prediction intervals are appropriate if the scientist is interested in whether a new observation is larger than existing data or smaller than existing data, but not both. The decision to use a one-sided interval must be based entirely on the question of interest. It should not be determined 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  $\alpha$  rather than  $\alpha/2$  as the error risk, placing all the risk on one side of the interval (fig. 3.9).

If the 90-percent prediction interval is on the right tail, it would be the interval from PI to  $\infty$ , where PI is determined as:

```
> PI <- quantile(Q, prob = 0.9, type = 6)
```

If the 90-percent prediction interval is on the left tail, it would be the interval from 0 to PI, where PI is determined as:

```
> PI <- quantile(Q, prob = 0.1, type = 6)
```



**Figure 3.9.** Example of a probability distribution showing the one-sided 90-percent prediction interval ( $\alpha=0.10$ ). The shaded area in the right tail has a probability of 0.10.

### Example 3.6. One-sided nonparametric prediction interval.

An arsenic concentration of 350 ppb is found in a southeastern New Hampshire groundwater sample. Does this indicate a shift in the distribution to larger values as compared to the distribution of concentrations for this dataset (from example 3.1)? Use  $\alpha=0.10$ .

As only large concentrations are of interest, the new data point will be considered larger than the original dataset if it exceeds the  $\alpha=0.10$  one-sided prediction interval, or the upper 90th percentile of the existing data. Using the `quantile` function in R we compute this upper 90th percentile as `quantile(x, prob = 0.9, type = 6)`, which has a value of 316 ppb.

A concentration of 350 ppb is considered to come from a different distribution than the existing data at an  $\alpha$  level of 0.10. However, 316 ppb or greater will occur approximately 10 percent of the time if the distribution of data has not changed; therefore, a concentration of 350 ppb is considered larger than the existing data at an  $\alpha$  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 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

If the assumption is that the data follow a normal distribution, prediction intervals are then constructed to be symmetric around the sample mean and wider than the confidence intervals on the mean. The equation for this interval (3.5) differs from that for a confidence interval around the mean (eq. 3.4) by adding a term  $\sqrt{s^2} = s$ , the standard deviation of individual observations around their mean:

$$\text{PI} = \bar{X} + t_{(\alpha/2, n-1)} \cdot \sqrt{s^2 + (s^2/n)} \quad \text{to} \quad \bar{X} + t_{(1-\frac{\alpha}{2}, n-1)} \cdot \sqrt{s^2 + (s^2/n)} . \quad (3.5)$$

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

### Example 3.7. Two-sided parametric prediction interval.

Using the arsenic data from example 3.1, we will proceed as if the data were symmetric (which we know is a poor assumption). Using that assumption and  $\alpha=0.10$ , how would we answer the question: Is a concentration of 370 ppb different (not just larger) than what would be expected from the previous distribution of arsenic concentrations?

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

$$98.4 + t_{(0.05, 24)} \cdot \sqrt{144.7^2 + \frac{144.7^2}{25}} \quad \text{to} \quad 98.4 + t_{(0.95, 24)} \cdot \sqrt{144.7^2 + \frac{144.7^2}{25}} ,$$

$$98.4 - 1.711 \cdot 147.6 \quad \text{to} \quad 98.4 + 1.711 \cdot 147.6 , \text{ and}$$

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

The simple answer would be that it falls outside of the 90-percent prediction interval, and as such we would consider 370 ppb to be an unlikely value if it came from the same distribution at  $\alpha=0.10$ . However, the interval we compute is seriously flawed because it includes values less than zero. This is a clear indication that this prediction interval based on symmetry is not appropriate because concentrations can never be negative. To avoid an endpoint as unrealistic as this, an asymmetric prediction interval (discussed above) should be used instead. Another alternative would be to make the computations for a symmetrical interval based on the logarithms of the data.

On the other hand, if we were evaluating a prediction interval for the James River discharge data from chapter 2—which we can see in figure 3.8 is approximately symmetric—the computation of the 98-percent prediction interval would be as follows

$$198.8 - 2.36 \sqrt{60.6^2 + \frac{60.6^2}{116}} \text{ to } 198.8 + 2.36 \sqrt{60.6^2 + \frac{60.6^2}{116}},$$

$$198.8 - 2.36 \cdot 60.86 \text{ to } 198.8 + 2.36 \cdot 60.86, \text{ and}$$

$$55.2 \text{ to } 342.4.$$

These results are rather similar to the nonparametric prediction interval presented above, which was 64 to 350.

### 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 appropriate than a symmetric interval when the underlying data are not symmetric, as is the case for the arsenic data in example 3.1. As stated in chapter 1, most water resources data and indeed most environmental data show positive skewness. Thus, datasets should be modeled using asymmetric intervals. Symmetric prediction intervals should be used only when the data are highly consistent with a normal distribution because prediction intervals deal with the behavior of individual observations. Therefore, the Central Limit Theorem does not apply. Data must be assumed to be non-normal unless shown otherwise. It is difficult to disprove normality using hypothesis tests (chap. 4) owing to the small sample sizes common to environmental datasets. It is also difficult to see non-normality with graphs unless the departures are strong. It is unfortunate that though most water resources datasets 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,  $x$ , follow a lognormal distribution. Thus from equation 3.5

$$PI = \exp \left[ \bar{y} + t_{(\alpha/2, n-1)} \sqrt{s_y^2 + (s_y^2/n)} \right] \text{ to } \exp \left[ \bar{y} + t_{(1-\frac{\alpha}{2}, n-1)} \sqrt{s_y^2 + (s_y^2/n)} \right], \quad (3.6)$$

where

$$y = \ln(x),$$

$\bar{y}$  is the mean, and

$s_y^2$  is the variance of  $y$ .

### Example 3.8. Asymmetric prediction intervals.

An asymmetric parametric prediction interval is computed using the logs of the arsenic data from example 3.1. A 90-percent prediction interval becomes

$$PI = \exp\left[3.172 - 1.71\sqrt{1.96^2 + \frac{1.96^2}{25}}\right] \text{ to } \exp\left[3.172 + 1.71\sqrt{1.96^2 + \frac{1.96^2}{25}}\right],$$

$$PI = \exp[3.172 - 1.71 \cdot 2.00] \text{ to } \exp[3.172 + 1.71 \cdot 2.00],$$

$$PI = \exp(-0.248) \text{ to } \exp(6.592), \text{ and}$$

$$PI = 0.78 \text{ to } 729.2.$$

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 (shown above) would be preferred if the assumption of a lognormal distribution is believed. The nonparametric 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 Quantiles and Tolerance Limits

Quantiles have traditionally been used in water resources to describe the frequency of flood events and flow-duration curves. They differ from percentiles only by being on a scale from 0 to 1 rather than 0 to 100. Thus the 100-year flood is the 99th percentile (0.99 quantile) of the distribution of annual flood peaks. It is the flood magnitude having an annual probability of exceedance of 1 percent. The 20-year flood is of a magnitude having an annual probability of exceedance of 5 percent and so is the 95th percentile of annual peaks. Similarly, the 2-year flood is the median or 50th percentile (0.50 quantile) of annual peaks. Flood quantiles are determined assuming that peak flows follow a specified distribution. The log-Pearson Type III is often used in the United States (see England and others [2018]). Historically, European countries have used the Gumbel (extreme value) distribution, though the generalized extreme value (GEV) distribution is now more common (Ponce, 1989; Stedinger and others, 1993).

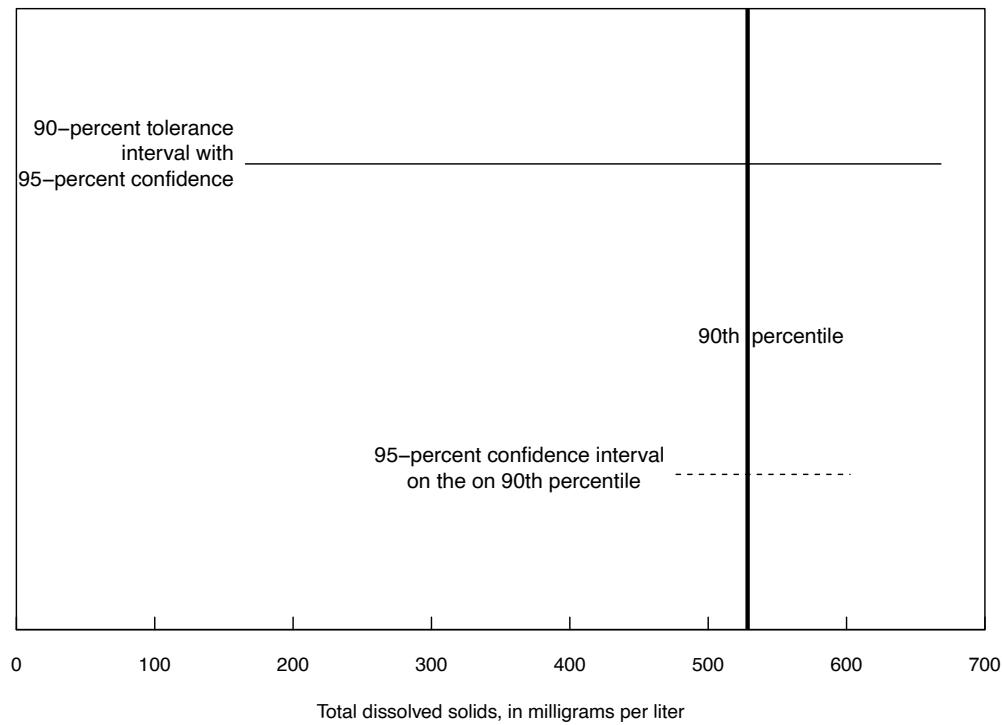
The most commonly reported statistic for analyses of low flows is also based on quantiles, the 7-day 10-year low flow, or 7Q10. The 7Q10 is the 10th percentile of the distribution of annual values of Y, where Y is the lowest average of mean daily flows over any consecutive 7-day period for that year.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

Percentiles of water-quality records are becoming more important in a regulatory framework. Crabtree and others (1987) have reported an increasing reliance on percentiles for developing and monitoring compliance with water quality standards. In light of the ever-increasing use of percentiles in water resources applications, understanding their variability is quite important, especially when comparing percentiles to a health criterion or legal standard. In section 3.7.1., confidence intervals on percentiles will be differentiated from tolerance intervals. In section 3.7.2., two-sided confidence intervals on percentiles are discussed. In section 3.7.3. the uses of one-sided lower confidence limits on a percentile, also known as lower tolerance limits, are discussed. In section 3.7.4. the uses of one-sided upper confidence limits on a percentile, also known as upper tolerance limits, are discussed. In each section both nonparametric and distributional intervals or limits are demonstrated. The usefulness of one-sided tolerance limits when comparing percentiles to criterion or standards is included in sections 3.7.3. and 3.7.4.

### 3.7.1 Confidence Intervals for Percentiles Versus Tolerance Intervals

A two-sided tolerance interval is an interval computed to contain, with a specified confidence ( $1 - \alpha$ ), a certain proportion, P, of the population between its two limits. P is called the coverage of the tolerance interval. Note that tolerance intervals involve two proportions, the coverage and the confidence level. This can be confusing. A two-sided tolerance interval covering the central 90 percent of total dissolved solids data (90-percent coverage) with 95-percent confidence from the Cuyahoga River, at Independence, Ohio, is shown in figure 3.10. The central 90 percent of observed data is between the 5th to the 95th percentiles—this 90 percent of the data is within the 90-percent tolerance interval. As with other types of statistical intervals, tolerance intervals also consider the sampling error, the error resulting from measuring a few observations from a much larger population, by using a confidence coefficient. An interval with 95-percent confidence of containing 90 percent of the population's values is a bit wider than the range of the observed 5th to 95th percentiles and depends on the number of observations sampled. Applications of two-sided tolerance intervals are few in water resources, although they are used in quality control and industrial applications.

A two-sided confidence interval around a percentile (dashed line in fig. 3.10) expresses the precision with which the percentile of the population has been estimated, similar to a confidence interval around a mean or median. An example application would be when the observed 10-year flood (90th percentile of annual peak flows) is compared to a design value  $X_0$  derived from a regional frequency analysis. If  $X_0$  is inside the (say 95 percent) confidence interval around the 90th percentile, then the observed percentile appears not to differ significantly from the design value. As shown in figure 3.10, two-sided tolerance intervals differ from two-sided confidence intervals on a percentile.



**Figure 3.10.** A two-sided tolerance interval with 90-percent coverage (solid line), and two-sided confidence interval on the 90th percentile (dashed line). Both were computed using a 95-percent confidence coefficient. The solid vertical line is the sample 90th percentile. The data are total dissolved solids concentrations for the Cuyahoga River, at Independence, Ohio (70 observations from 1969–73).

One-sided tolerance limits are far more commonly used than two-sided tolerance intervals in water resources and water quality applications; they are discussed in detail in sections 3.7.3 and 3.7.4. One-sided limits are used when the interest is only in how high or low the population percentile might be, but not both. They would appear as either the upper or lower half of the dashed confidence interval line in figure 3.10. One-sided tolerance limits are identical to one-sided confidence limits on a percentile. Examples of the use of one-sided limits include an upper confidence limit on a flood percentile to be used for culvert or bridge design, or to compare the tolerance limit for concentration or load against a water quality standard to determine whether the standard has been exceeded in more than  $(1-P) \cdot 100$  percent of measurements.

### 3.7.2 Two-sided Confidence Intervals for Percentiles

A two-sided confidence interval around a percentile (fig. 3.10) expresses the precision with which the percentile of the population has been estimated, similar to a confidence interval on a mean or median. It may be computed without assuming a distribution by counting out to observations on either side of the percentile, just as was computed in section 3.3.1 for the median. However, to do this requires a large sample size, with more data required for more extreme percentiles. Alternatively, a distribution may be used as a model for the shape of the data distribution, allowing an upper limit of the interval to perhaps extend higher than the largest current observation. The validity of the distributional model depends on whether values in the population actually follow that distribution. For smaller sample sizes, using a distributional model may be the only choice available if an interval around a high percentile must be made.

Nonparametric confidence intervals can be developed for any percentile analogous to those developed in section 3.3 for the median. Lower ( $R_L$ ) and upper ( $R_U$ ) ranks corresponding to data points at the ends of the confidence interval are found using the `qbinom` function in R by entering the probability associated with the percentile ( $p=0.75$  for the 75th percentile and so forth). A value of 1 should be added to the lower rank, as the desired confidence coefficient (say 95 percent) is the probability of exceeding the lower ranked observation, not including it.

#### Example 3.9. Nonparametric two-sided confidence interval around the 20th percentile.

For the arsenic concentrations of Boudette and others (1985) used in example 3.1, we determine a 95-percent confidence interval on  $C_{0.20}$ , the 20th percentile of concentration ( $p=0.2$ ). This percentile was chosen for illustration because it is not too extreme. A sample size of  $n=25$  would be insufficient to compute a 95-percent nonparametric interval around a high percentile such as the 90th.

```
> x <- c(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)
> quantile(x, 0.2, type = 6)
20%
2.94
```

The sample 20th percentile  $\hat{C}_{0.20}=2.94$  ppb, the  $0.20 \cdot (25+1)=5.2$ th smallest observation, or two-tenths of the distance between the 5th and 6th smallest observations. The order statistics corresponding to  $\alpha/2=0.025$  are at ranks 1 and 9 in the dataset.

```
> qbinom(c(0.025, 0.975), length(x), 0.2)
[1] 1 9
```

Adding 1 to the lower rank produced by the `qbinom` function and summing the probabilities of inclusion for the 2nd through 9th ranked observations yields

```
> sum(dbinom(2:9, length(x), 0.2))
[1] 0.9552784
```

The interval between and including the 2nd and 9th ranked observations (1.5 ppb to 8 ppb) contains the true population 20th percentile of arsenic with confidence 95.5 percent. Note that the asymmetry around  $\hat{C}_{0.20}=2.94$  reflects the asymmetry of the data.

An alternate method to compute a nonparametric interval is bootstrapping. The function `BootTOL`, provided in SM.3, will bootstrap this 20th percentile. It requires the package `boot`.

```
> library(boot)

> BootTOL(x, p = 20, conf = 95, R = 10000, TYPE = "perc")
```

```
Bootstrap Confidence Intervals of the 20-th percentile of x
Using boot in R
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

```
Based on 10000 bootstrap replicates
```

CALL :

```
boot.ci(boot.out = B, conf = Conf, type = TYPE)
```

Intervals :

Level	Percentile
-------	------------

95%	( 1.56, 9.20 )
-----	----------------

Calculations and Intervals on Original Scale

The bootstrap endpoints 1.56 to 9.20 are similar to the previous nonparametric estimate.

For sample sizes larger than 20, a large-sample (normal) approximation to the binomial distribution is a third method to obtain nonparametric interval estimates for percentiles. Ranks corresponding to the upper and lower confidence limits are determined by equations 3.7 and 3.8 using quantiles of the standard normal distribution,  $z_{\alpha/2}$  and  $z_{[1-\alpha/2]}$ . Those ranks are

$$R_L = np + z_{\alpha/2} \sqrt{np(1-p)} + 0.5 , \quad (3.7)$$

$$R_U = np + z_{[1-\alpha/2]} \sqrt{np(1-p)} + 0.5 , \quad (3.8)$$

whereas in example 3.9,  $n$  is the sample size and  $p$  is the quantile value for the percentile around which the interval is computed.

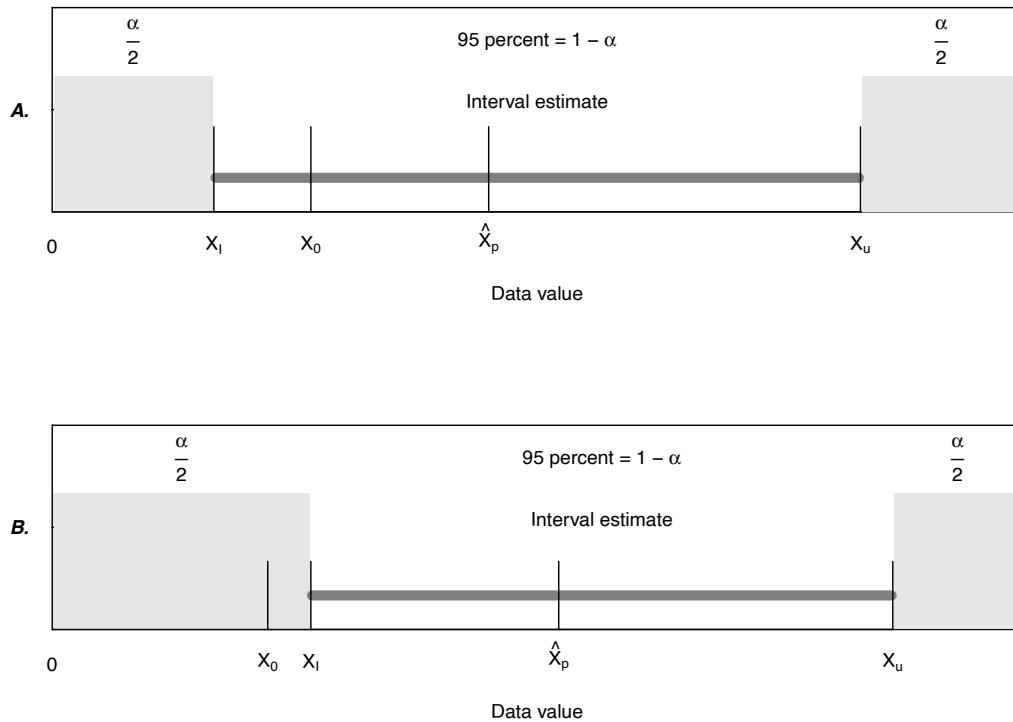
The 0.5 terms added to equations 3.7 and 3.8 reflect a continuity correction (see chap. 5) of 0.5 for the lower bound and -0.5 for the upper bound (which otherwise would be a value of +1). The computed ranks  $R_U$  and  $R_L$  are rounded to the nearest integer.

$$R_L = 25 \cdot 0.2 + (-1.96) \cdot \sqrt{25 \cdot 0.2(1-0.2)} + 0.5 = 5 - 1.96 \cdot 2 + 0.5 = 1.6$$

$$R_U = 25 \cdot 0.2 + 1.96 \cdot \sqrt{25 \cdot 0.2(1-0.2)} + 0.5 = 5 + 1.96 \cdot 2 + 0.5 = 9.4$$

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

For a test of whether a percentile significantly differs (either larger or smaller) from a prespecified value  $X_0$ , simply compute a  $(1-\alpha)$ -percent two-sided confidence interval for the percentile. If  $X_0$  falls within this interval, the percentile does not significantly differ from  $X_0$  at a significance level  $\alpha$  (fig. 3.11A). If  $X_0$  is not within the interval, the percentile significantly differs from  $X_0$  at the significance level of  $\alpha$  (fig. 3.11B).



**Figure 3.11.** Confidence interval on the  $p$ th percentile  $X_p$  as a test for  $H_0: X_p = X_0$ . *A.*  $X_0$  inside the interval estimate,  $X_p$  not significantly different from  $X_0$ . *B.*  $X_0$  outside the interval estimate,  $X_p$  significantly different from  $X_0$ .

**Example 3.10. Nonparametric test for whether the 80th percentile of annual peak flows differs from a design value.**

In the R dataset `Saddle` located in SM.3 are annual peak discharges for the Saddle River at Lodi, New Jersey from 1925 to 1989 in cubic meters per second. Of interest is the 5-year flood, the flood that is likely to be equaled or exceeded once every 5 years (20 times in 100 years), and so is the 80th percentile of annual peaks. Let's assume that using some method of calculation (such as a regional statistical model or a deterministic rainfall-runoff model) it has been determined that the 5-year design flood is  $37 \text{ m}^3/\text{s}$ . We would like to consider if the actual data are consistent with this estimate, at  $\alpha=0.05$ . The 5-year flood is equivalent to the 80th percentile of the population of floods because  $0.8=1-(1/5)$ .

The 80th percentile is estimated from the 65 values between 1925 and 1989 as follows

```
> load("Saddle.RData")
> Q <- Saddle$Q
> n <- length(Q)
> quantile(Q, probs = 0.8, type = 6)
80%
69.7727
```

Therefore,  $\hat{Q}_{0.8} = 69.77 \text{ m}^3/\text{s}$ . Following equations 3.7 and 3.8, a two-sided confidence interval on this percentile is

$$R_L = np + z_{\alpha/2} \sqrt{np(1-p)} + 0.5 ,$$

$$R_U = np + z_{1-\alpha/2} \sqrt{np(1-p)} + 0.5 .$$

The R commands and results are as follows

```
> RL <- n * 0.8 + qnorm(0.025) * sqrt(n * 0.8 * 0.2) + 0.5
> RL
[1] 46.17931
> RU <- n * 0.8 + qnorm(0.975) * sqrt(n * 0.8 * 0.2) + 0.5
> RU
[1] 58.82069
> quantile(Q, probs = c(RL/n, RU/n), type = 6)
71.04509% 90.49338%
60.28367 90.27553
```

Thus, the 95-percent confidence interval for the 5-year flood lies between the 46.2th and 58.8th ranked peak flows, or  $60.3 < Q_{0.8} < 90.3$ . The interval does not include the design value  $X_0 = 37 \text{ m}^3/\text{s}$ . Therefore the 20-year flood does differ from the design value at a significance level of  $\alpha = 0.05$ .

Smaller datasets may not have sufficient numbers of observations to estimate intervals around high or low percentiles without assuming a distributional shape. When used with too few observations, the nonparametric methods described above can incorrectly default to the minimum and (or) maximum values as interval endpoints (be careful to watch for this if your software does not warn you of this situation). In that case a distribution must be assumed. 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 assumed distribution does not fit the data well, the resulting estimates are less accurate than if no distribution were assumed. Unfortunately, the situation in which an assumption is most needed, that of small sample sizes, is where it is most difficult to determine whether the data follow the assumed distribution.

Similar to the nonparametric intervals described above, determining if the distribution-based confidence interval around a percentile includes a design value  $X_0$  is a parametric test for whether the observed percentile significantly differs from  $X_0$ . As an example, the computation of point and interval estimates for percentiles assuming a lognormal distribution is straightforward. Let  $y = \ln(x)$  where the  $x$  values are the original units. The sample mean of the  $y$  values is denoted  $\bar{y}$  and sample standard deviation of the  $y$  values is  $s_y$ . The point estimate of any percentile is then

$$\hat{X}_p = \exp(\bar{y} + z_p \cdot s_y) , \quad (3.9)$$

where  $z_p$  is the  $p$ th quantile of the standard normal distribution.

For percentiles other than the median, confidence intervals are computed using the noncentral  $t$ -distribution (Stedinger, 1983). The confidence interval on  $X_p$  is

$$CI(X_p) = \left[ \exp\left(\bar{y} - n^{1/2} \cdot \zeta_{[1-\frac{\alpha}{2}]} \cdot s_y\right), \exp\left(\bar{y} - n^{1/2} \cdot \zeta_{[\alpha/2]} \cdot s_y\right) \right] , \quad (3.10)$$

where  $\zeta_{\alpha/2}$  is the  $\alpha/2$  quantile of the noncentral  $t$ -distribution with  $n-1$  degrees of freedom and noncentrality parameter  $-n^{1/2} \cdot z_p$  for the desired percentile with sample size of  $n$ . Using R, the  $z_p$  values are

computed with the function `qnorm` and the  $\zeta_{\alpha/2}$  values are computed with the function `qt`. The EnvStats package in R (Millard, 2013) will compute percentiles and their confidence intervals for several other commonly used distributions.

**Example 3.11. Two-sided confidence interval for percentiles assuming a lognormal distribution.**

Compute a 90-percent confidence interval for the 90th percentile of the arsenic concentrations found in example 3.1, assuming the data are lognormal.

The 90th percentile, assuming concentrations are lognormal, is as given in equation 3.9:

$$\hat{C}_{0.90} = \exp(\bar{y} + z_{0.90} \cdot s_y) = \exp(3.173 + 1.282 \cdot 1.960)$$

$$= 294.1 \text{ ppb} .$$

Note this is lower than the sample estimate of 316 ppb obtained without assuming the data are lognormal. The noncentrality parameter is  $-5 \cdot 1.282 = -6.4$ .

The corresponding 90-percent confidence interval estimate from equation 3.10 is

$$\exp\left(3.173 - \frac{1}{\sqrt{25}} \cdot -4.489 \cdot 1.96\right) < C_{0.90} < \exp\left(3.173 - \frac{1}{\sqrt{25}} \cdot -9.19 \cdot 1.96\right), \text{ and}$$

$$138.7 < C_{0.90} < 875.3 .$$

This estimate would be preferred over the nonparametric estimate if it is believed that the data were truly lognormal, otherwise a nonparametric interval would be preferred when there are sufficient numbers of observations (doubtful for  $n=25$ ). Using the bootstrap approach, the 90-percent confidence interval is 210 ppb to 580 ppb. The upper bound in this case is the highest data value, indicating that there are too few observations to use a nonparametric estimation method. With parametric approaches, interval endpoints can extend well beyond the data. In the SM.3 we have included an R function called `CIqt1LN` that computes confidence intervals for a lognormal random variable. It computes a two-sided interval as well as one-sided intervals. Interval estimates for percentiles of the log-Pearson Type III distribution are computed in a similar fashion, see Stedinger (1983) for details on that procedure.

### 3.7.3 Lower One-sided Tolerance Limits

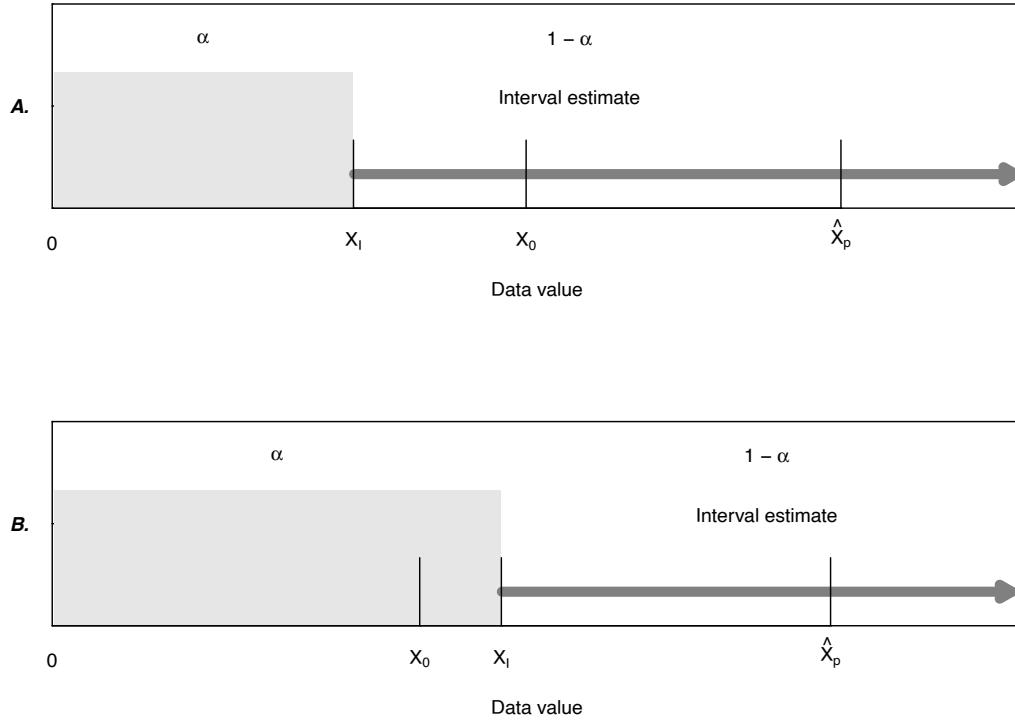
A  $(1-\alpha)$  lower tolerance limit (LTL) with coverage =  $p$  is identical to a one-sided lower  $(1-\alpha)$  confidence limit on the  $p$ th percentile. Lower tolerance limits for any percentile are computed in the same way as for the two-sided confidence limits of section 3.7.2., except that the entire error probability  $\alpha$  is assigned to the lower end. Computing an LTL declares that the upper side is of no interest and no upper bound is computed (software often represents it as positive infinity). Lower tolerance limits are often used in water resources to determine if a percentile has significantly exceeded a stated standard or guidance criterion. If the LTL exceeds the criterion there is evidence with  $(1-\alpha)$  confidence to state that more than  $(1-p) \cdot 100$  percent of concentrations in the population (stream, aquifer) exceed the criterion.

A nonparametric  $(1-\alpha)$ -percent lower tolerance limit on the  $p$ th percentile can be computed using the large-sample approximation. The rank of the observation corresponding to the lower confidence limit on the percentile (lower tolerance limit) is

$$R_L = np + z_\alpha \sqrt{np(1-p)} + 0.5 . \quad (3.11)$$

Equation 3.11 is just equation 3.7 using  $\alpha$  instead of  $\alpha/2$ .

To test whether a percentile  $X_p$  significantly exceeds a specified criterion or standard  $X_0$ , compute the lower tolerance limit on the percentile.  $X_p$  will be considered significantly higher than  $X_0$  if its lower tolerance limit lies entirely above  $X_0$  (fig. 3.12).



**Figure 3.12.** Lower tolerance limit as a test for whether the percentile  $X_p > X_0$ . *A.*  $X_0$  above lower limit  $X_l$ ;  $X_p$  not significantly greater than  $X_0$ . *B.*  $X_0$  below lower limit  $X_l$ ;  $X_p$  significantly greater than  $X_0$ .

**Example 3.12. Nonparametric lower tolerance limit as a test for whether the 90th percentile exceeds a standard.**

A water-quality standard states that the 90th percentile of arsenic concentrations in drinking water shall not exceed 10 ppb. Has this standard been violated at the  $\alpha=0.05$  confidence level by the New Hampshire arsenic data in example 3.1?

The 90th percentile of the arsenic concentrations is

```
> x <- c(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)
> quantile(x, 0.9, type = 6)
90%
316
```

or by hand

$$\hat{C}_{0.90} = (25+1) \cdot 0.9 = 23.4\text{th data point}$$

$$= 300 + 0.4(340 - 300)$$

$$= 316 \text{ ppb} .$$

Following equation 3.11, the rank of the observation corresponding to a one-sided 95-percent lower confidence bound on  $C_{0.90}$  is

$$\begin{aligned}
 R_L &= np + z_\alpha \sqrt{np(1-p)} + 0.5 \\
 &= 25 \cdot 0.9 + z_{0.05} \cdot \sqrt{25 \cdot 0.9(0.1)} + 0.5 \\
 &= 22.5 + (-1.64)\sqrt{2.25} + 0.5 \\
 &= 20.5 ,
 \end{aligned}$$

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 greater than  $X_0 = 10$  and therefore the standard has been proven to be exceeded at the 95-percent confidence level. The lower tolerance limit may also be computed using the `eqnpar` function in the `EnvStats` package (Millard, 2013). The function uses a nonlinear interpolation between two observations, providing a slightly different answer (211.88 ppb) than the 215 ppb computed above.

```
> eqnpar(conc, p = 0.9, type = 6, ci = TRUE, ci.type = "lower")
```

```
Results of Distribution Parameter Estimation
-----
Assumed Distribution: None
Estimated Quantile(s): 90'th %ile = 316
Quantile Estimation Method: Nonparametric
Data: conc
Sample Size: 25
Confidence Interval for: 90'th %ile
Confidence Interval Method: interpolate (Nyblom, 1992)
Confidence Interval Type: lower
Confidence Level: 95%
Confidence Limit Rank(s): 20 21 NA NA
Confidence Interval: LCL = 211.8822
UCL = Inf
```

Parametric lower tolerance limits are computed by assuming that data follow a specific distribution. They are particularly helpful for smaller datasets where nonparametric estimates may simply stop at the lowest observation in the dataset, resulting in an incorrect  $\alpha$  if that observation is used as the lower limit. Parametric limits will be accurate when data follow the assumed distribution, but may be quite inaccurate when data do not follow the assumed distribution. A probability plot and goodness-of-fit test such as the probability plot correlation coefficient (PPCC) test (see chap. 4) of the assumed distribution should be used to check the fit of the distribution to the data.

Parametric lower tolerance limits are used to conduct a parametric test for whether a percentile exceeds a specified value  $X_0$ . The error level  $\alpha$  is placed entirely on the lower side before conducting the (one-sided) test, and if the lower tolerance limit exceeds  $X_0$  then the percentile significantly exceeds  $X_0$  at the  $(1 - \alpha)$ -percent confidence level.

**Example 3.13. Parametric lower tolerance limit as a test for whether the 90th percentile exceeds a standard.**

Test whether the 90th percentile of arsenic concentrations in example 3.1 exceeds 10 ppb at the  $\alpha=0.05$  significance level, assuming the data are lognormal.

The 90th percentile of arsenic concentration was previously computed as 294.1 ppb assuming a lognormal distribution. The one-sided 95-percent lower confidence limit for the 90th percentile is the same as a two-sided 90-percent lower limit, as the same 5-percent error probability is below this estimate. The two-sided 90-percent value was previously computed as 138.7 ppb. This limit exceeds 10 ppb, therefore the standard has been proven to be exceeded at the 95-percent confidence level. This same result can be computed using the `eqlnorm` function in the `EnvStats` package of R (Millard, 2013). The package includes functions for other standard distributions as well.

```
> eqlnorm (conc, p = 0.9, ci = TRUE, ci.type = "lower")
```

Results of Distribution Parameter Estimation

Assumed Distribution:	Lognormal
Estimated Parameter(s):	meanlog = 3.172667
	sdlog = 1.959582
Estimation Method:	mvue
Estimated Quantile(s):	90'th %ile = 294.1155
Quantile Estimation Method:	qmle
Data:	conc
Sample Size:	25
Confidence Interval for:	90'th %ile
Confidence Interval Method:	Exact
Confidence Interval Type:	lower
Confidence Level:	95%
Confidence Interval:	LCL = 138.6747
	UCL = Inf

### 3.7.4 Upper One-sided Tolerance Limits

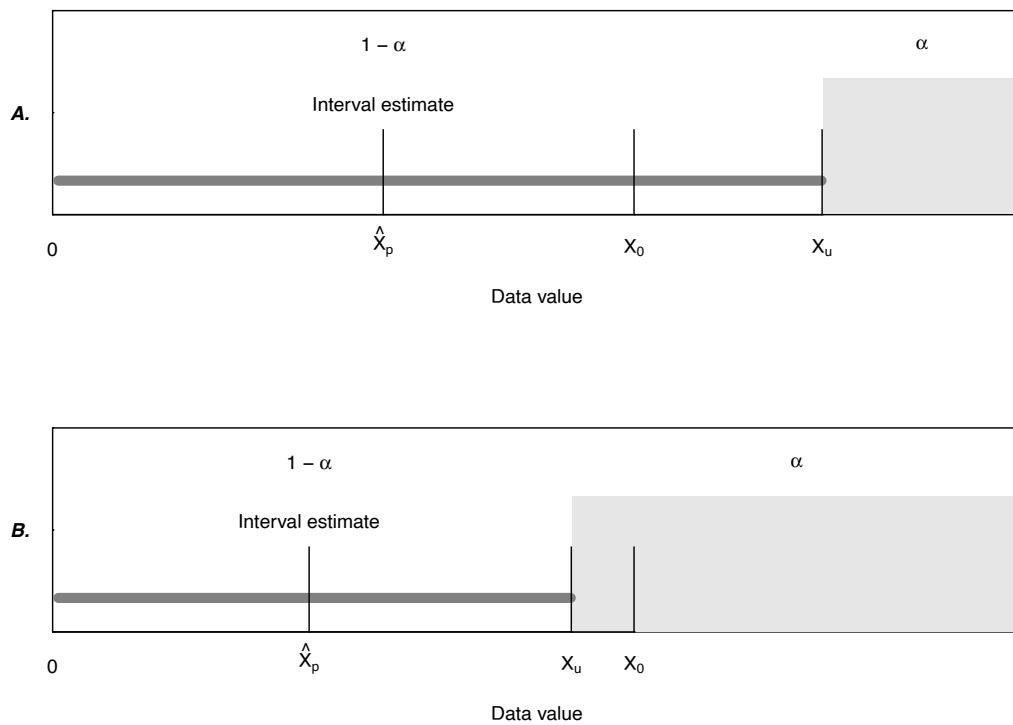
A  $(1-\alpha)$  upper tolerance limit (UTL) with coverage  $=p$  is identical to a one-sided upper  $(1-\alpha)$  confidence limit on the  $p$ th percentile. Computing a UTL declares that the lower side is of no interest and no lower bound is computed (software often represents it as negative infinity). UTLs are often used in natural resource studies to define a limit above which only  $(1-p) \cdot 100$  percent of new observations are expected to occur with  $(1-\alpha)$ -percent confidence. Multiple new observations are compared to the single UTL as opposed to the singular or few observations that can be compared to a prediction limit. An exceedance of the UTL by more than  $(1-p) \cdot 100$  percent of new observations indicates that conditions have changed from those used to compute the tolerance limit. One example of its use has been to compute a baseline 90th percentile criterion for sediment loads or chemical concentrations in total maximum daily load studies.

A nonparametric  $(1 - \alpha)$ -percent upper tolerance limit on the  $p$ th percentile can be computed using the large-sample approximation. The rank of the observation corresponding to the upper confidence limit on the percentile (upper tolerance limit) is

$$R_u = np + z_{[1-\alpha]} \sqrt{np(1-p)} + 0.5 . \quad (3.12)$$

Equation 3.12 is just equation 3.8 using  $\alpha$  instead of  $\alpha/2$ .

To test whether a percentile  $X_p$  is significantly less than  $X_0$ , compute the upper confidence limit, placing all error  $\alpha$  on the side above  $\hat{X}_p$  (fig. 3.13).  $X_p$  will be considered as significantly less than  $X_0$  if its upper confidence limit (upper tolerance limit) is entirely below  $X_0$ .



**Figure 3.13.** Upper tolerance limit as a test for whether percentile  $X_p < X_0$ . *A.*  $X_0$  below the upper tolerance limit  $X_u$ ;  $X_p$  not significantly less than  $X_0$ . *B.*  $X_0$  above the upper tolerance limit  $X_u$ ;  $X_p$  significantly less than  $X_0$ .

**Example 3.14. Nonparametric upper tolerance limit as a test for whether the 90th percentile is below the regional  ${}_{7}Q_{10}$ .**

We have 68 values of the annual 7-day minimum flows for climate years 1942–2009 (a climate year runs from April 1 to March 31) on the Little Mahoning Creek at McCormick, Pennsylvania, expressed in cubic meters per second (in the SM.3 dataset `LittleMA.RData`). We can load and view the data and then estimate the 7-day, 10-year, low flow ( ${}_{7}Q_{10}$ ) as follows

```
> load("LittleMA.RData")
> Q <- sort(Little$Q)
> n <- length(Q)
> Q
[1] 0.0194 0.0227 0.0368 0.0392 0.0413 0.0429 0.0498 0.0522
[9] 0.0603 0.0704 0.0708 0.0801 0.0825 0.0841 0.0858 0.0874
[17] 0.0926 0.0930 0.1048 0.1080 0.1141 0.1157 0.1201 0.1205
[25] 0.1230 0.1258 0.1270 0.1359 0.1388 0.1418 0.1598 0.1622
[33] 0.1634 0.1683 0.1707 0.1711 0.1731 0.1744 0.1748 0.1853
[41] 0.2013 0.2225 0.2253 0.2265 0.2421 0.2468 0.2557 0.2629
[49] 0.2743 0.2767 0.2840 0.2880 0.3014 0.3107 0.3236 0.3277
[57] 0.3600 0.3803 0.4284 0.4450 0.4482 0.5582 0.5663 0.6392
[65] 0.6432 0.7484 0.8050 1.1735
> xHatP <- quantile(Q, probs = 0.1, type = 6)
> xHatP
10%
0.04911
```

The estimate of the  ${}_{7}Q_{10}$  based on the data (with no distributional assumptions) is  $0.04911 \text{ m}^3/\text{s}$ . However, based on a regional regression model of  ${}_{7}Q_{10}$  values versus watershed characteristics the  ${}_{7}Q_{10}$  was expected to equal  $0.06 \text{ m}^3/\text{s}$  (call that  $X_0=0.06 \text{ m}^3/\text{s}$ ). The question is: Should we reject at  $\alpha=0.05$  the null hypothesis that the  ${}_{7}Q_{10}$  is  $0.06 \text{ m}^3/\text{s}$  versus the alternate hypothesis that it is below  $0.06 \text{ m}^3/\text{s}$ ? To answer this question we need to compute the 95-percent upper confidence limit for the true 0.1 percentile of the distribution of annual minimum 7-day low flows. We can compute that as follows

$$R_U = np + z_{1-\alpha} \sqrt{np(1-p)} + 0.5$$

$$R_U = 68 \cdot 0.1 + 1.644 \sqrt{68 \cdot 0.1 \cdot (0.9)} + 0.5$$

$$R_U = 11.369 .$$

The rank of the upper 95-percent confidence bound on the 10th percentile of the distribution of annual 7-day minimum flows is the 11.369th rank out of the 68 observed values. The R code to determine that is

```
> p = 0.1
> z95 <- qnorm(0.95)
> RU <- n * p + z95 * sqrt(n * p * (1 - p)) + 0.5
> RU
[1] 11.36914
```

We calculate the discharge associated with this fractional rank as xUpper:

```
> xUpper <- quantile(Q, probs = RU/n, type = 6)
> xUpper
16.71933%
0.07578793
```

Our best estimate based only on the quantiles of the data at the site is 0.04911 m<sup>3</sup>/s. We cannot reject the hypothesis that the true value of the  ${}^7Q_{10}$  is 0.06 m<sup>3</sup>/s because the upper 95-percent confidence bound for the  ${}^7Q_{10}$  is above 0.06, at 0.076 m<sup>3</sup>/s. These calculations could also have been performed using the eqnpar function in the EnvStats package (Millard, 2013):

```
> eqnpar(Q, p = 0.1, type = 6, ci = TRUE, ci.type = "upper")
```

#### Results of Distribution Parameter Estimation

---

Assumed Distribution:	None
Estimated Quantile(s):	10'th %ile = 0.04911
Quantile Estimation Method:	Nonparametric
Data:	Q
Sample Size:	68
Confidence Interval for:	10'th %ile
Confidence Interval Method:	interpolate (Nyblom, 1992)
Confidence Interval Type:	upper
Confidence Level:	95%
Confidence Limit Rank(s):	NA NA 12 11
Confidence Interval:	LCL = -Inf UCL = 0.07546245

Parametric UTL computations are beneficial with smaller sample sizes, but are accurate only when the data closely follow the assumed distribution. The EnvStats package in R contains functions to compute tolerance limits on percentiles for several standard distributions.

**Example 3.15. Parametric upper tolerance limit as a test for whether the 90th percentile is below the regional  ${}_7Q_{10}$ .**

Using the annual 7-day minimum flows for climate years 1942–2009 on the Little Mahoning Creek at McCormick, Pennsylvania, should we reject, at  $\alpha=0.005$ , the hypothesis that  ${}_7Q_{10}$  is  $0.06 \text{ m}^3/\text{s}$  versus being below  $0.06 \text{ m}^3/\text{s}$ ? Assume that flows follow a lognormal distribution.

The 95-percent upper tolerance bound on the 10th percentile is computed for a lognormal distribution of flows with the `eqlnorm` function of EnvStats (Millard, 2013):

```
> eqlnorm (Q, p=0.1, ci=TRUE, ci.type="upper")
```

Results of Distribution Parameter Estimation

---

Assumed Distribution:	Lognormal
Estimated Parameter(s):	meanlog = -1.8065945
	sdlog = 0.8580029
Estimation Method:	mvue
Estimated Quantile(s):	10'th %ile = 0.0546848
Quantile Estimation Method:	qmle
Data:	Q
Sample Size:	68
Confidence Interval for:	10'th %ile
Confidence Interval Method:	Exact
Confidence Interval Type:	upper
Confidence Level:	95%
Confidence Interval:	LCL = 0.00000000
	UCL = 0.06746206

Assuming a lognormal distribution of lows, the best estimate of the site's 10th percentile is  $0.0547 \text{ m}^3/\text{s}$ . We cannot reject the hypothesis that the true value of the  ${}_7Q_{10}$  is  $0.06 \text{ m}^3/\text{s}$  because the upper 95-percent confidence bound for the  ${}_7Q_{10}$  is just above  $0.06$ , at  $0.067 \text{ m}^3/\text{s}$ .

## 3.8 Other Uses for Confidence Intervals

Confidence intervals are used for purposes other than interval estimates. Three common uses are (1) to detect outliers, (2) for quality control charts, and (3) for determining sample sizes necessary to achieve a stated level of precision. However, the implications of data non-normality for the three applications are often overlooked; these issues are discussed in the following sections.

### 3.8.1 Implications of Non-normality for Detection of Outliers

An outlier is an observation that appears to differ in its characteristics from the bulk of the dataset to which it is assigned. It is a subjective concept; different people may define specific points as either outliers or not. When an outlier is observed in a dataset, the analyst must make an effort to try to evaluate what the observed value represents. There are three possibilities. The first is that it represents what actually happened and that it is simply a very extreme value from the same population as all the other values in the dataset. An example of the first possibility is the annual peak discharge record for Rapid Creek at Rapid City, South Dakota. This is a record of 73 years in which the largest flood was estimated to be about  $1,400 \text{ m}^3/\text{s}$  (in 1972), the second largest flood was less than  $100 \text{ m}^3/\text{s}$ , and the vast majority of annual floods were less

than 25 m<sup>3</sup>/s. It is quite clear that the extreme 1972 flood is real (not some data processing or measurement error) and the result of a stalled storm system over this watershed. The observation needs to be considered in any flood risk analysis. The second possibility is that the reported value represents what actually happened, but the circumstances under which it occurred would cause us to consider it to be outside of the population represented by the bulk of the data. An example of this second possibility is an extreme flood caused not by a precipitation event or snow melt event, but rather by a dam failure upstream. Another example would be a measured concentration of a solute that is the direct result of a catastrophic accident at an upstream facility. We may decide to leave these data out of the analysis, but we need to be clear that the analysis only considers events that were not a result of a catastrophic failure of some engineered system (a dam, a treatment plant, or a factory). The third case is one where the recorded value is simply one that is greatly in error. Examples of this could include instrument failure or transcription error (for example a series of pH measurements that are mostly around 6.0 followed by one that is recorded as 60). The process of determining the nature of the outlier will not always result in a clear determination of the cause and a clear determination of the proper course of action for the analyst, but it is important to take thoughtful steps to make such determinations.

Outliers are sometimes deleted from a dataset 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 valid. There is no law stating that observed data must follow some specific distribution. Outlying observations are often the most important data collected, providing insight into extreme conditions or important causative relations. Deleting outliers (unless they can clearly be shown to be in error) is unwarranted because procedures not requiring an assumption of normality are both available and powerful. Many of these tools are discussed in the following chapters.

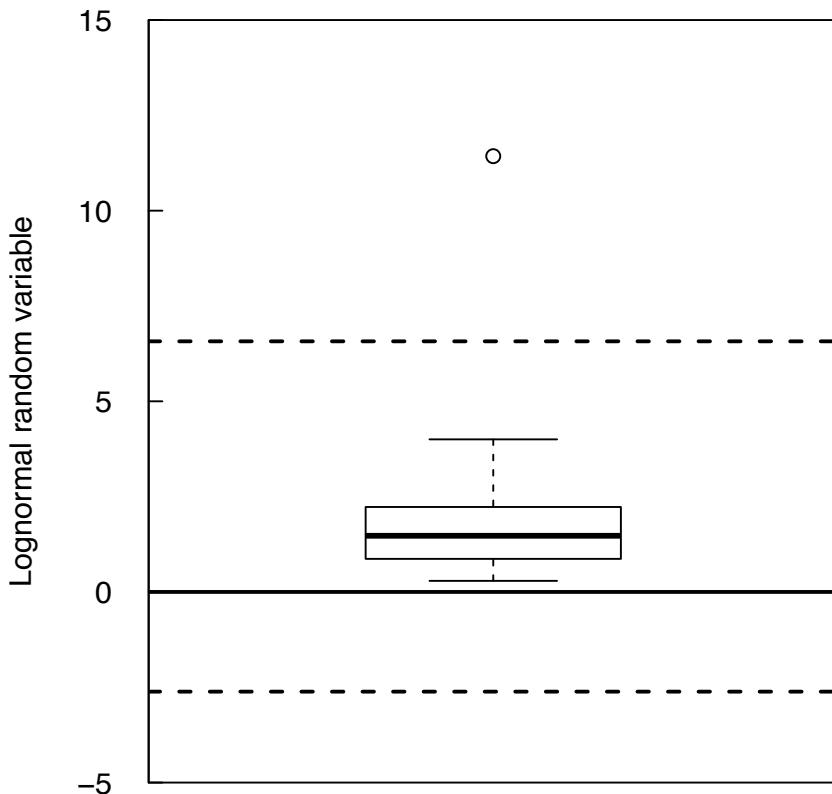
In order to delete an outlier, an observation must first be declared to be one. Rules or tests for outliers have been used for years, as surveyed by Beckman and Cook (1983). The most common tests are based on a *t*-interval and assume that data follow a normal distribution. Points beyond the prediction interval calculated using equation 3.5 are declared as outliers only because they were unlikely to originate from a normal distribution.

Real world data may not follow a normal distribution. Even though the mean of a large dataset is generally approximately normal, there is no reason to assume that the data themselves are normal. 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). We repeat—designation as an outlier by a statistical outlier test does not mean that the data are necessarily bad. Deletion of points only on this basis, without any corroboration from scientific knowledge, is what we have called dangerous.

To illustrate this concept, a boxplot of 25 observations generated from a lognormal distribution, as well as the upper and lower  $\alpha=0.05$  prediction interval limits assuming a normal distribution are shown in figure 3.14. Many water resources datasets appear similar in shape to a lognormal distribution. All data outside of these limits will be designated as outliers by this prediction limit outlier test.

First note that the lower limit is less than zero, yet we know that the data have a lower bound of zero by virtue of being from a lognormal distribution. The outlier test (prediction limit) boundary is unrealistic at the lower end. Next note that the largest observation greatly exceeds the upper prediction limit and thus the test would declare it an outlier, yet we know it to be a valid observation generated from the same lognormal distribution as generated the remaining observations. Outlier tests only check for whether data are likely to have been produced by the distribution (almost always the normal distribution) assumed by the test. Outlier tests cannot in themselves test for bad data.

Multiple outliers cause additional problems for outlier tests based on normality (Beckman and Cook, 1983). They may inflate the estimated standard deviation such 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 retesting after deletion causes the second largest to be rejected, and so on. Replication of the test may eventually discard a substantial part of the dataset. The choice of how many times to apply the test is entirely arbitrary. A variation on this approach has been developed for hydrologic applications (Cohn and others, 2013) that tests multiple outliers together rather than deleting them one at a time. This method, called the multiple Grubbs-Beck test, has proven useful in flood frequency analysis, where for some streams (particularly in arid regions) the annual peak discharge is zero or very close to zero and thus not truly a member of the population of flood events to begin with. This procedure separates the population of nonfloods from true floods in the annual peak streamflow series so that the frequency estimation for the population of true floods can be carried out without being contaminated by the presence of these nonflood events.



**Figure 3.14.** Boxplot of a sample of size 25 from a lognormal distribution. Dashed lines represent the upper and lower 95-percent prediction limits assuming normality.

### 3.8.2 Implications of Non-normality for Quality Control

One visual presentation of confidence intervals used extensively in industrial processes is a control chart (Montgomery, 1991). A small number of products are sampled at a given point in time and their mean is 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 a specific type of control chart called the  $\bar{x}$  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 per day over  $m$  time intervals (for example, weeks where one day per week is sampled), the total sample size  $N = n \cdot m$ . 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 centerline of the chart. A  $t$ -confidence interval on  $\bar{X}$  (eq. 3.4) uses the sample size  $n$  of each daily mean value. Those intervals are added as parallel lines to the quality control chart. Daily mean values will, on average, plot outside of these boundaries only  $\alpha \cdot 100$  percent of the time if the means are normally distributed. Means falling outside the boundaries more frequently than this are taken to indicate that something in the process has changed.

If  $n$  is large (say 70 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 very limited ability to detect departures from the expected mean value more frequently than if the data were not skewed.

Control charts are also produced to illustrate process variance and use either 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 charts will also 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 centerline estimating the geometric mean (an estimate of the median for lognormal data) in original units, with multiplicative variation represented by the confidence bands of section 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-pseudosigma of Hoaglin (1983); this was done by Schroder and others (1987). The F-pseudosigma 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. Before proceeding with this type of process, the analyst must first 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 3.4 is solved for  $n$  to produce

$$n = \left( \frac{t_{\alpha/2,n-1} s}{\Delta} \right)^2, \quad (3.13)$$

where  $s$  is the sample standard deviation and  $\Delta$  is one-half the desired interval width. Sanders and others (1983) and other authors have promoted this equation. As discussed above, this calculation may have large errors for sample sizes ( $n$ ) less than about 70 with strongly skewed data. Estimates of  $s$  will be inaccurate and strongly inflated by any skewness and (or) outliers; the resulting estimates of  $n$  will therefore be large. For example, Håkanson (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 a symmetric distribution, if not a normal distribution. For example, based on equation 3.13 logarithms will drastically lower estimated sample sizes for skewed data. Resulting samples sizes would allow the median (geometric mean) to be estimated within a multiplicative tolerance factor equal to  $\pm 2\Delta$  in log units.

Kupper and Hafner (1989) point out a second problem with using equations like 3.13 for estimating sample size even when data follow a normal distribution. They show that equation 3.13 underestimates the true sample size needed for a given level of precision, even for estimates of  $n \geq 40$ . This is because equation 3.13 does not recognize that the standard deviation  $s$  is only an estimate of the true value  $\sigma$ . They suggest adding a power criterion to equation 3.13 so that the estimated interval width will be at least as small as the desired interval width with some stated probability (say 90 or 95 percent). For example, when  $n$  would equal 40 based on equation 3.13, the resulting interval width will be less than the desired width  $2\Delta$  with only about 0.42 probability! The sample size should instead be 53 in order to ensure the interval width is within the expected range with 90 percent probability. Kupper and Hafner conclude that equation 3.13 and similar equations that do not consider a power criteria "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 3.13. 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. More discussion of power and sample size is given in chapter 13.

## Exercises

1. Compute both nonparametric and parametric 95-percent interval estimates for the median of the granodiorite data of chapter 2, exercise 3 in the dataset `grano.RData` located in the supplemental material for chapter 2 (SM.2). Which interval estimate is more appropriate for these data? Why?
2. 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 the SM.3 dataset `VAwells.Rdata` or does it represent something larger? Answer by computing appropriate 95-percent parametric and nonparametric intervals. Which intervals are more appropriate for these data?
3. Construct the most appropriate 95-percent interval estimates for the mean and median annual streamflows for the Conecuh River at Brantley, Alabama (dataset `Conecuh.Rdata` in SM.3). Include a bootstrap approach for the median in addition to the standard method using the binomial distribution. Use the parametric method for the mean. Also consider a bootstrap approach for the mean.
4. A water intake is located on the Potomac River at Little Falls, just above Washington, D.C. We want to select a design discharge for this water intake so that in only 1 year out of 10 does the 1-day minimum discharge of the river go so low that the intake becomes inoperable because it is not submerged in the river. But, we want to be conservative in our selection of this design discharge because we know that an estimate of the 0.1 frequency annual minimum flow is somewhat uncertain. We want to set the design discharge such that there is only a 5-percent probability that the true 0.1 frequency annual minimum flow is below the design discharge. Our dataset (`PotomacOneDayLow.RData` in SM.3) consists of the annual 1-day minimum discharge at this site for the 84-year period 1932–2015. Use the concept of a nonparametric one-sided confidence interval for the 0.1 quantile on the distribution of annual minimum discharges.

# 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 a quantitative framework to determine objectively whether or not a signal (nonzero difference between groups, correlation between variables) is present in the data.*

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. Historical 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. Hypothesis tests have at least two advantages over educated opinion:

1. They ensure that every analysis of a dataset using the same methods will arrive at the same result because computations can be checked and agreed upon by others.
2. They present a quantitative measure of the strength of the evidence (the  $p$ -value), allowing the decision to reject a hypothesis to be augmented by the risk of an incorrect decision.

In this chapter, we introduce the basic structure of hypothesis testing and classifications for appropriate usage and application. The rank-sum test is used to illustrate this structure, as well as to illustrate the origin of  $p$ -values for exact test results. Tests for normality are also discussed. Concepts and terminology found here will be used throughout the rest of the book. Finally, we end the chapter by discussing some of the criticisms, misuse, and misinterpretation of hypothesis tests.

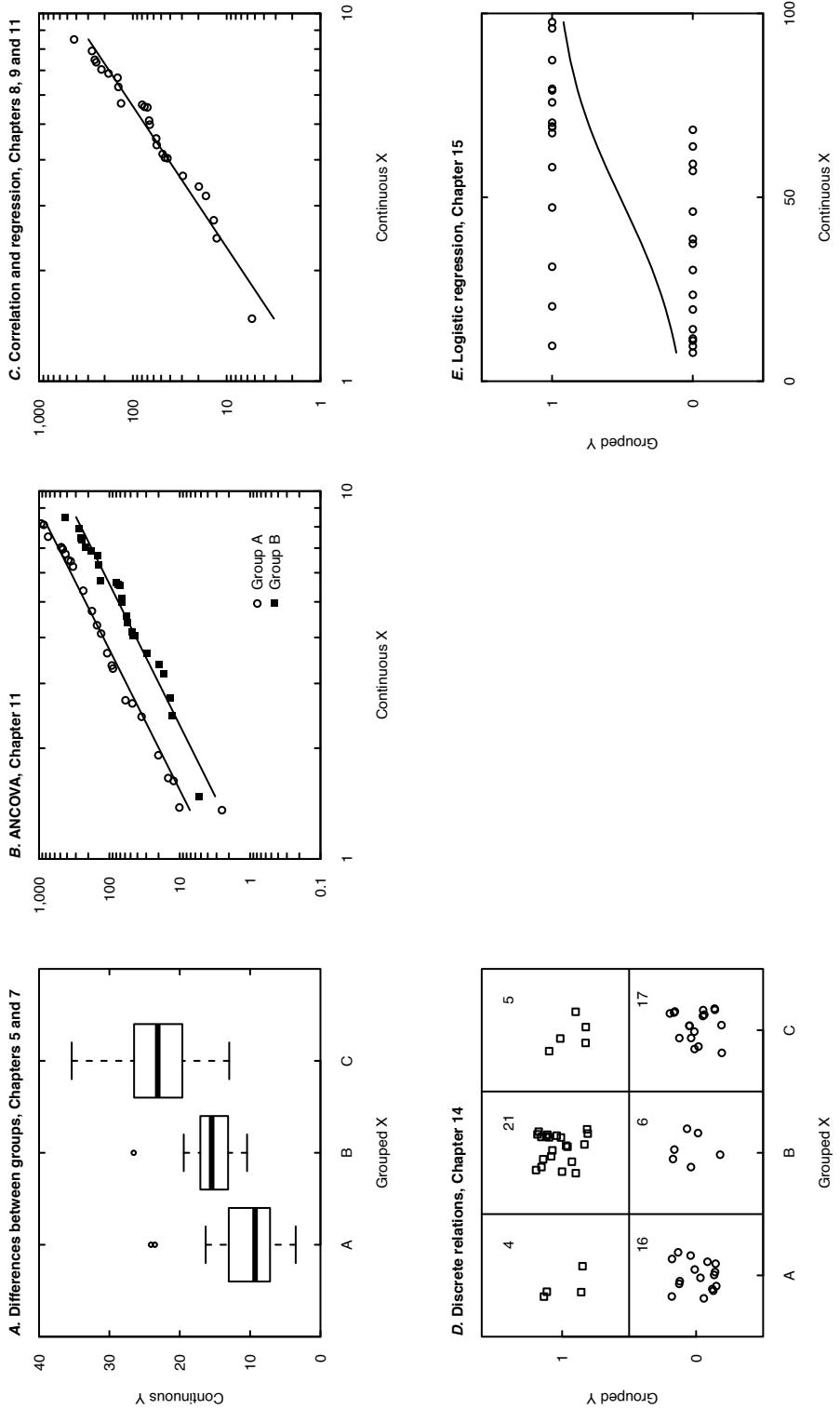
### 4.1 Classification of Hypothesis Tests

The choice of which test to use among the numerous hypothesis tests available to scientists often causes unnecessary confusion. Tests can be classified into the five types shown in figure 4.1, based on the measurement scales of the data being tested. Within these classes there are three major divisions of hypothesis tests: parametric, nonparametric, and permutation tests. These classes differ in the way that their  $p$ -values are computed. The nature of the data and the objectives of the study largely determine which class and division of hypothesis test 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, sometimes called an independent variable, is one used to explain why and how the magnitude of the response variable changes. When testing for a difference in central tendency between two populations, for example, the explanatory variable designates the population from which a sample was drawn.

#### 4.1.1 Classification Based on Measurement Scales

The five classes of test procedures are represented by the five graphs in figure 4.1. 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, and continuous data can be grouped into categories (for example, the analysis of covariance graph in fig. 4.1B). Parametric, nonparametric, and permutation tests may be found within a given class of hypothesis tests.



**Figure 4.1.** Five classes of hypothesis tests. (A) Differences between groups (see chaps. 5 and 7), (B) analysis of covariance (ANCOVA, see chap. 11), (C) correlation and regression (see chaps. 8, 9, and 11), (D) discrete relations (see chap. 14), and (E) logistic regression (see chap. 15).

Hypothesis tests represented by the three graphs in figure 4.1A–C 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 properties and concentrations measured by water resources scientists. In contrast, tests represented by the two graphs in figure 4.1D–E 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. 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. Categorical variables used primarily as explanatory variables include aquifer type, month, land-use group, and station number, and are often character variables.

The boxplots in the figure 4.1A represent the two- and multi-sample hypothesis tests such as the rank-sum test, the *t*-test, and the Kruskal-Wallis test. 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 graph in figure 4.1C represents two often-used methods—linear regression and correlation, including their variations and alternatives. 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, and is discussed in chapter 12.

The graph in figure 4.1B is a blend of these two approaches, called analysis of covariance (ANCOVA). A continuous response variable is related to two or more explanatory variables, some of which are continuous and some categorical.

The graph in figure 4.1D represents a situation similar to that for use of *t*-tests or analysis of variance, except that the response variable is categorical. Contingency tables appropriately measure the association between two such categorical variables. One example is to determine whether the probability of finding a volatile organic compound above the reporting limit (*y*) varies by land-use group (*x*).

The graph in figure 4.1E shows that a regression-type relation can be developed for the case of a categorical response variable. Probabilities of occurrence are modeled using logistic regression. For example, perhaps the proportion of concentrations of a pesticide or other constituent 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, or irrigation intensities. 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 Divisions Based on the Method of Computing a *p*-value

The three major divisions of hypothesis tests—parametric tests, nonparametric tests, and permutation tests—differ in how they obtain *p*-values. All hypothesis tests have underlying assumptions that should be reviewed before performing the tests. Parametric tests typically assume that the data have a particular distribution (such as a normal distribution, as in fig. 1.2), the data are independent and identically distributed (the data are not serially correlated, see chap. 8), and that when comparing multiple groups the groups have the same variance (there are statistical adjustments one can make if they do not). These assumptions were very helpful in creating statistical tests that were practical to implement before the advent of modern computers. Test statistics and *p*-values for the tests could be calculated based on a set of easily computed parameters such as the sample mean or variance, or the co-variance between two different variables. Parametric tests are appropriate when their distributional assumptions are met by the data. However, when the data substantially depart from the distributional assumptions of the test, parametric tests can fail to represent what is actually occurring in the data and thus they may lack sensitivity (power) to detect signals (effects) in the data.

The greatest strength of parametric procedures is in modeling and estimation, especially for complex designs. Relations among multiple variables can be described and tested that are difficult, if not nearly impossible, to describe and test with nonparametric methods. Parametric analysis of variance (ANOVA) studies more complex than what we cover in this text can be designed and tested in ways not yet possible for nonparametric or permutation methods. Parametric multiple-regression equations (chap. 11) can model more complex situations than what methods from the other two divisions currently accomplish. Time series models, kriging, and other parametric methods outside of the scope of this book model correlation structures in space and time that are difficult to otherwise compute (permutation methods are gradually encroaching into these areas). As always, care must be taken to use data that meet the requirements of the methods, and this often involves transforming variables. Aho (2016), among other texts, provides a good description of parametric tests, including some complex ANOVA designs.

Transformations are used to make data more normally distributed or linear in pattern before performing a parametric test. A possible pitfall in using transformations is that they change the parameter being tested. Testing in log units, for example, produces tests on the geometric means in original units, rather than means. Regression equations using the log of the response ( $y$ ) variable predict the geometric mean (median) of  $y$ , not the mean. As transformations change the parameter being tested, the investigator must be careful to match the test's objectives to what they want.

Nonparametric tests represent the observed data by their ranks. Ranks are a transformation of percentiles; these tests answer frequency questions like, "Does the frequency of exceeding a standard differ between groups?" or "Do high values occur more frequently in one group than the other?" No assumption of the shape of the data distribution is required—the tests are distribution-free methods. However, that does not mean that there are no underlying assumptions to the test of choice. Nonparametric tests assume independent, random samples, as do parametric tests, and may be subject to additional assumptions.

The  $p$ -value in nonparametric tests is computed by determining all possible outcomes of the test and determining the probability of obtaining the single observed outcome or a more extreme outcome. 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." Nonparametric tests efficiently extract information on the relative magnitudes (ranks, percentiles) of data, providing the same answer both before and after data transformations for transformations that preserve the order of the data, such as logarithmic transformations. They are most useful for simpler hypothesis tests between groups or regression type relations between two (but not easily with more than two) variables. Hollander and Wolfe (1999), among other texts, provide a detailed overview of nonparametric methods.

Permutation tests compute  $p$ -values by randomly selecting several thousand outcomes from the many larger number of outcomes possible that represent the null hypothesis. The  $p$ -value is then the proportion of outcomes that are equal to, or more extreme than, the one obtained from your data. Their greatest use is perhaps to test for differences in means without assuming a normal distribution as permutation tests are also distribution-free. However, like nonparametric tests, this does not mean permutation tests are free of underlying assumptions, including that the data are random. Permutation tests may be used to test any null hypothesis, so they can also test for differences in standard deviations or for the contingency table setup of chapter 14. These computer-intensive methods, envisioned by the pioneers in statistics in the early 1900s, require computing power that was not available until the late 1980s. Good (2005) and Manly (2007) provide an overview of permutation test methods. All three divisions of tests are presented in the upcoming chapters.

## 4.2 Structure of Hypothesis Tests

All hypothesis tests follow the same six steps, which are discussed in the following sections:

1. Choose the appropriate test and review its assumptions.
2. Establish the null and alternative hypotheses,  $H_0$  and  $H_A$ .
3. Decide on an acceptable error rate,  $\alpha$ .
4. Compute the test statistic from the data.
5. Compute the  $p$ -value.
6. Reject the null hypothesis if  $p \leq \alpha$ ; do not reject if  $p > \alpha$ .

### 4.2.1 Choose the Appropriate Test

Test procedures should be selected based on the data characteristics and study objectives. Figure 4.1 presents the first selection criterion—the measurement scales of the data. The second criterion 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, the two-sample  $t$ -test, the rank-sum test, or a two-sample permutation test might be selected (see table 4.1). Of importance is whether the central value is better defined by the mean (center of mass) or median (center of frequency). Subsequent chapters are organized by test objectives, with several alternate tests discussed in each.

**Table 4.1.** Guide to the classification of some hypothesis tests with continuous response variables.

[-, not applicable]

Parametric	Nonparametric	Permutation
Two independent data groups (chap. 5)		
Two-sample $t$ -test	Rank-sum test (two-sample Wilcoxon; Mann-Whitney test)	Two-sample permutation test
Matched pairs of data (chap. 6)		
Paired $t$ -test	Signed-rank test, sign test	Paired permutation test
Three or more independent data groups (chap. 7)		
Analysis of variance	Kruskal-Wallis test	One-way permutation test
Three or more dependent data groups (chap. 7)		
Analysis of variance without replication	Friedman test, aligned-rank test	-
Two-factor group comparisons (chap. 7)		
Two-factor analysis of variance	Brunner-Dette-Munk (BDM) test	Two-factor permutation test
Correlation between two continuous variables (chap. 8)		
Pearson's $r$ (linear correlation)	Spearman's $\rho$ or Kendall's $\tau$ (monotonic correlation)	Permutation test for Pearson's $r$
Model of relation between two continuous variables (chaps. 9 and 10)		
Linear regression	Theil-Sen line	Bootstrap of linear regression

The third selection criteria is the choice between parametric, nonparametric, or permutation tests. This should again be based on the objectives of your study—parametric and nonparametric tests look at different characteristics. For example, parametric tests that assume data follow a normal distribution are built around the mean as the measure of center. The mean is a standardized measure of the total and is most appropriate when you are interested in summing the data. For example, if the interest is in the total load transported down a river or the cumulative exposure to a contaminant that an organism or human has experienced over time, the mean is the appropriate per-sample or per-time unit of measurement to represent that total (Helsel and Griffith, 2003). On the other hand, if typical or commonly occurring differences were of interest, a test based on frequency statistics is more relevant. Tests between groups would determine if concentrations are generally the same in all groups or if they are generally higher in some. Nonparametric tests address this question of difference in frequencies. Finally, if a mean is of interest but the data distribution is asymmetrical or prone to outliers, the power (probability of making the correct decision if the alternative hypothesis is true) of parametric tests to reject  $H_0$  when  $H_0$  is false can be quite low and type II errors (failing to reject  $H_0$  when  $H_0$  is false) commonly result (Bradley, 1968). This loss of power is the primary concern when using parametric tests. The loss of power in parametric tests is avoided by using a permutation test on differences in means.

Arguments for and against use of parametric versus nonparametric tests have been around for decades. Texts such as Hollander and Wolfe (1999), Conover (1999), and Bradley (1968) base the advantages of nonparametric tests on considerations of power and invariance to measurement scale. We provide a further discussion of power in chapter 13. The importance of potential for loss of power when using parametric tests is not always fully appreciated by some in water resources, ecology, or other applied disciplines. For example, Johnson (1995) argues for a wide use of parametric methods over nonparametric methods in ecology.

The first argument is that “parametric methods do not require data to follow a normal distribution, because their sample means will have a distribution that follows a normal distribution for most data” (Johnson, 1995). This argument relies upon the Central Limit Theorem. The number of observations required for the Central Limit Theorem to hold true is a function of the data’s asymmetry, also called skewness. The more skewness, the more observations are necessary to expect sample means to follow a normal distribution (Boos and Hughes-Oliver, 2000). Johnson (1995) used the example of a symmetric uniform distribution to support the use of the Central Limit Theorem. For water resources, symmetry is uncommon and asymmetric data distributions prevail. The assessment by Singh and others (1999) for data skewed to the extent commonly found in water resources showed that sample sizes of about 100 observations were required for the Central Limit Theorem to hold, and for correct coverage of  $t$ -based confidence intervals to be obtained. A similar result for asymmetric data in a completely different field was found by Pocock (1982). In other words, most water resources projects do not have sufficient numbers of observations to rely on the Central Limit Theorem in order to say “it’s not normally distributed, but it doesn’t matter.” Using asymmetric data in a test requiring a normal distribution will often result in a loss of power and a failure to see differences that are present.

Johnson’s second argument is that “nonparametric tests require an assumption of equal variance of groups, just as parametric tests do, to test for differences in means.” Rank-based nonparametric tests do not test for differences in means, but percentiles, a shift in the distribution of the two datasets. A nonparametric test could be used to test for difference in means only if the distributions of both groups are symmetric and the mean and median are the same in each group. The nonparametric test itself does not require an assumption of equal variance, but if the intent is to test for differences in means, differing variance will affect any and all tests because of the strong effects of outliers on the mean. The scientist must decide whether the mean of asymmetric data with outliers is the appropriate statistic for which to test.

The third argument is that when variances differ, “the Welch-Satterthwaite version of the  $t$ -test performs well.” We agree that the Welch-Satterthwaite correction to the  $t$ -test can perform well and state in chapter 5 that the Welch-Satterthwaite adaptation of the  $t$ -test for differing variances should always be used whenever a  $t$ -test is performed. This adaptation of the  $t$ -test is the default in statistics software and far better than the original  $t$ -test, which can lead to incorrect outcomes when variances differ. The adaptation does, however, have a power cost— $p$ -values can be quite high in comparison to a permutation test for differences in means on the same data. We give an illustration of this in chapter 5.

The fourth argument for parametric tests is that the *t*-test is robust and this is true. However, nonstatisticians may understand robust to be a general term meaning that it is applicable in a variety of situations. Its technical meaning in statistics is that type I errors (false positives, rejecting  $H_0$  when  $H_0$  is true) occur no more often than expected. Robustness as a statistical term does not address the loss of power (false negatives) such as when applying parametric tests to skewed data. This misunderstanding of the definition of robust was discussed in the 1980s by Blair and Higgins (1980) and Helsel and Hirsch (1988).

One question that often arises is how non-normal a data distribution must be in order for nonparametric tests to be preferred over tests that rely on an assumption of normality. Blair and Higgins (1980) gave insight into this question by mixing data from two normal distributions, 95 percent from a normal distribution and 5 percent from a second normal distribution with a 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. A blending of two normal distributions with 5 percent from one and 95 percent from another may not be detectable by a graph or test for normality. Yet, when comparing two groups of this type, Blair and Higgins 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 were found to be not significantly different by the *t*-test. Blair and Higgins (1980) is recommended for additional details and study.

The final argument given against use of nonparametric tests is that, “By their very nature, nonparametric methods do not specify an easily interpreted parameter...Parameters are generally of most interest, so we should provide estimates of those parameters that are meaningful and applicable to making real decisions.” Johnson (1995) then recommends transformations followed by parametric tests for skewed data so that the results can still be interpreted using parameters. However, a median is a parameter, as are median differences between groups (the nonparametric Hodges-Lehmann estimator in chap. 5). Nonparametric tests compute parameters, just not moment statistics (means and standard deviations). If you transform data using a transformation that preserves the order of the values (logarithmic transformation, for example) and perform a *t*-test on the transformed units, this is not a test for difference in means in the original units (chap. 1) but the geometric means, and therefore medians of the untransformed values. Know what parameters you are actually testing and make sure what you test for fits the goals of your study.

Our approach in this book agrees with E.J.G. Pitman’s 1948 paper in the Annals of Mathematical Statistics (quoted in Salsburg, 2001), who found that, “...with only slight deviations from the parametric model, the nonparametric tests were vastly better than the parametric ones.” We provide an example of this in chapter 10. Our approach agrees with Higgins (2003), who stated in regard to the two-sample tests of chapter 5, “...the Wilcoxon [rank-sum] test can never be much less efficient than the *t*-test, but it has the potential of being infinitely more efficient.” Given that we discuss relatively simple situations in this book as compared to complex statistical models, the quote from Hahn and Meeker’s (1991) classic text on computing intervals (see chap. 3) applies: “One might ask ‘When should I use distribution-free statistical methods?’ The answer, we assert, is ‘Whenever possible.’ If one can do a study with minimal assumptions, then the resulting conclusions are based on a more solid foundation.”

Older guidance documents often used a flowchart recommending the choice of test be based on a prior test of whether data follow a specific distributional shape (usually a normal distribution). However, different divisions of tests have different objectives and one needs to carefully note the null and alternative hypotheses. If the interest is in testing whether one group tends to have higher values than the others, a nonparametric test addresses that objective. If the interest is in testing to determine whether the means (total amounts) are the same in all groups, a parametric test or a permutation test on the mean addresses that objective. A parametric test after taking logarithms of the data does not test for differences in means of data in their original units. In short, your choice should be informed by your objective, not just a reflection of the distributional shape of your data.

Select test procedures that have greater power for the types of data expected to be encountered. To obtain a test for means when data are asymmetric or contain outliers, a permutation test should be strongly considered, as the parametric test may suffer from a loss of power. Comparisons of the power of two test procedures, one parametric and one nonparametric, can be based on the tests’ asymptotic relative efficiencies (ARE; the computation of which is beyond the scope of this text), a property of their behavior with large sample sizes (Bradley, 1968). A test with larger ARE will have generally greater power. For

many cases with data that do not meet distributional assumptions, especially the common situation where data are highly asymmetrical or prone to outliers, the ARE of nonparametric tests can be many times those of parametric tests (Hollander and Wolfe, 1999). Thus, the power of nonparametric tests to reject  $H_0$  when it is truly false is generally much higher (as much as 3 times, or 300 percent) in the presence of outliers or skew. For example, the rank-sum test has a larger ARE (more power) than the *t*-test for distributions containing outliers (Conover, 1999). Permutation tests also will have greater power than parametric tests when data do not meet specific distributional assumptions. When data follow a normal distribution, nonparametric tests have slightly lower (5–15 percent) ARE than parametric tests (Hollander and Wolfe, 1999). Therefore, in the presence of skewness and outliers—precisely the characteristics often shown by water resources data—nonparametric and permutation tests commonly exhibit greater power than do parametric tests.

## 4.2.2 Establish the Null and Alternate Hypotheses

The null and alternate hypotheses should be established before collecting data when one has designed the study or, in the case where one is using retrospective data, before analysis of the data. The hypotheses are a concise summary of the study objectives and will keep those objectives in focus during data collection and analysis without being affected by unconscious bias arising from the data or desired results.

The null hypothesis ( $H_0$ ) is what is assumed to be true about the system, before collection of new data. 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 the alternative hypothesis, depending on one's vantage point. However, 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 the concentrations are the same upgradient and downgradient (the company did the original site design and hazardous waste has not contaminated downgradient wells). In either case, the null hypothesis assumed to be true is the same: concentrations are similar in both groups of wells.

The alternate hypothesis ( $H_A$ , sometimes represented as  $H_1$ ) is the situation anticipated to be true if the evidence (the data) show that the null hypothesis is unlikely. In some cases,  $H_A$  is the negation of  $H_0$ , such as “the 100-year flood is not equal to the design value.”  $H_A$  may also be more specific than just the negation of  $H_0$ , such as “the 100-year flood is greater than the design value.” Alternate hypotheses come in two general types: one-sided and two-sided. The associated hypothesis tests are called one-sided and two-sided tests.

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 suggests that the 100-year flood is different from the design value in either direction (larger or smaller), this would provide evidence against the null hypothesis of the 100-year flood equaling the design flood, thus the test is two-sided. Most tests in water resources are of this kind.

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  $H_0$  if 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 a specific culvert be replaced. The null hypothesis would be stated as “the 100-year flood is less-than or equal to the design flood,” and 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, before looking at any data, it cannot be stated 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 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 the direction of the alternative hypothesis is specified as one-sided tests include testing for (1) decreased annual floods or downstream sediment loads after completion of a flood-control dam; (2) decreased nutrient loads or concentrations because of a new sewage treatment plant or a newly incorporated best management practice; and (3) an increase in concentration when comparing a suspected contaminated site to an upstream or upgradient control site.

### 4.2.3 Decide on an Acceptable Type I Error Rate, $\alpha$

The  $\alpha$ -value, or significance level, is the probability of incorrectly rejecting the null hypothesis (rejecting  $H_0$  when it is in fact true). This is one of four possible outcomes of a hypothesis test, as shown in figure 4.2. The significance level is the risk of a type I error deemed acceptable by the decision maker. Statistical tradition uses a default of 0.05 (5 percent) or 0.01 (1 percent) for  $\alpha$ , but there is no reason why other values should not be used. For example, suppose that an expensive cleanup process will be mandated if the null hypothesis that there is no contamination is rejected. The  $\alpha$ -level for this test might be set very small (such as 0.01) 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 before further analysis of the high sites. In this case, the  $\alpha$ -level might be set to 0.10 or 0.20, so that all sites with high values would likely be retained for further study.

Given that  $\alpha$  represents one type of error, why not keep it as small as possible? One way to do this would be to never reject  $H_0$ . The chance of making a type I error,  $\alpha$ , would then equal zero. Unfortunately, this would lead to a large probability of error of a second type—failing to reject  $H_0$  when it was in fact false. This type of error is called a type II error (a false negative), and the probability of a type II error is designated by  $\beta$  (fig. 4.2). The power of a test is the probability of making a correct decision when  $H_0$  is false, or  $1 - \beta$ . Both type I and type II errors are of concern to practitioners and both will have some finite probability of occurrence. Once a decision is made as to an acceptable type I error probability,  $\alpha$ , two steps can be taken to concurrently reduce the probability of a type II error,  $\beta$ :

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

		$H_0$ is true	$H_0$ is false
Decision	Fail to Reject $H_0$	Correct decision $\text{Prob}(\text{correct decision}) = 1 - \alpha$ <b>Confidence</b>	Type II error $\text{Prob}(\text{type II error}) = \beta$
	Reject $H_0$	Type I error $\text{Prob}(\text{type I error}) = \alpha$ <b>Significance level</b>	Correct decision $\text{Prob}(\text{correct decision}) = 1 - \beta$ <b>Power</b>

**Figure 4.2.** Four possible results of hypothesis testing.

For water quality applications, null hypotheses are often stated as “no contamination” or “no difference” between groups. 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, consider type II error when determining the sample size (see chap. 13). Power is also sacrificed when data having the characteristics outlined in chapter 1, such as outliers and skewness, are analyzed with tests requiring a normal distribution. With parametric tests, test sensitivity decreases as skewness and the number of outliers increase.

#### 4.2.4 Compute the Test Statistic and the *p*-value

Test statistics summarize the information contained in the data. If the test statistic is not substantially 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. Note that in R software, test statistics are often adjusted by subtracting the lowest possible value for the test statistic with the current sample size(s). This is of no consequence whatsoever, besides that R’s reported statistic may differ from that output by commercial software. The *p*-values are not affected.

The *p*-value is the probability of obtaining the computed test statistic, or one even more extreme, when the null hypothesis is true. It is derived from the data and concisely expresses the evidence against the null hypothesis contained in the data. The *p*-value measures the believability of the null hypothesis; the smaller the *p*-value, the less likely the observed test statistic is when  $H_0$  is true, and, therefore, the stronger the evidence for rejection of the null hypothesis. The *p*-value is also referred to as the “attained significance level.”

How do *p*-values differ from  $\alpha$ -levels? The  $\alpha$ -level does not depend on the data but states the risk of making a type I error that is acceptable to the scientist or manager. The  $\alpha$ -level is the critical value that allows a yes/no decision to be made—the treatment plant has improved water quality, nitrate concentrations in the well exceed standards, and so forth. The *p*-value provides more information—the strength of the scientific evidence. Reporting the *p*-value allows analysts with different risk tolerances (different  $\alpha$ ) to make their own decision.

For example, consider a one-sided test of whether downgradient wells have higher contaminant concentrations than upgradient wells. 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  $\alpha = 0.01$  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; there is 1-percent probability that remediation would be performed when in fact it is not required. Reporting only reject or do 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 (less concern about making a type I error;  $\alpha = 0.05$ , perhaps) to decide for or against remediation.

It should be noted that there are three methods of computing *p*-values for nonparametric tests:

1. Exact test. Exact versions of nonparametric tests provide *p*-values that are exactly correct. They are computed by comparing the test statistic to a table of all possible results for the sample sizes present. In the past, an extensive set of tables was required, one for every possible combination of sample sizes. Today, software can compute these until sample sizes become very large. When sample sizes are small, the exact version provides the most accurate results.
2. Large-sample approximation (LSA). To avoid computing test statistics for large datasets (lots of computing time), approximate *p*-values are obtained by assuming that the distribution of the test statistic can be approximated by some common distribution, such as the chi-square or normal distribution. This does not mean the data themselves follow that distribution, but only that the test statistic does. A second reason for using an LSA is that when ties occur in the data, exact tests cannot be computed. 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. Commercial computer software predominantly uses large sample approximations when reporting *p*-values, whether or not the sample sizes are sufficient to warrant using them.

3. Permutation test. Large datasets may require enumerating many thousands to millions of possible outcomes to define the exact distribution of the test statistic. Permutation tests compute a smaller random sample of all the possible outcomes, representing the null hypothesis by the random sample. The  $p$ -value is the probability of the single observed test statistic, or one more extreme, being produced when the null hypothesis is true. This method approximates the  $p$ -value and is considered a more accurate way to approximate than assuming that the test statistic follows a specific distribution.

#### 4.2.5 Make the Decision to Reject $H_0$ or Not

When the  $p$ -value is less than or equal to the decision criteria (the  $\alpha$ -level),  $H_0$  is rejected. When the  $p$ -value is greater than  $\alpha$ ,  $H_0$  is not rejected. The null hypothesis is never accepted or proven to be true, it is assumed to be true until proven otherwise and is not rejected when there is insufficient evidence to do so. In short, reject  $H_0$  when the  $p$ -value  $\leq \alpha$ . However, it is good practice to report the  $p$ -values for those that may want to make decisions with a different significance level and type I error rate.

### 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 five samples,  $y_i$ , from aquifer  $Y$  ( $m=5$ ) for a total of seven samples ( $N=n+m=7$ ). Also suppose that there is a prior reason (that likely motivated the sampling) to believe that  $X$  values tend to be lower than  $Y$  values: aquifer  $X$  is deeper and likely to be uncontaminated. The null hypothesis ( $H_0$ ) and alternative hypothesis ( $H_A$ ) of this one-sided test are as follows:

$H_0$ :  $x_i$  and  $y_i$  are samples from the same distribution, or

$H_0$ :  $Prob(x_i \geq y_i) = 0.5$ ,  $i=1, 2, \dots, n; j=1, 2, \dots, m$ .

$H_A$ :  $x_i$  is from a distribution that is generally lower than that of  $y_j$ , or

$H_A$ :  $Prob(x_i \geq y_j) < 0.5$ .

Remember that with one-sided tests such as this one, data indicating differences opposite in direction to  $H_A$  ( $x_i$  frequently larger than  $y_j$ ) 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 type I error probability,  $\alpha$ , must be set. As in a court of law, innocence is assumed (concentrations are similar) unless evidence is collected to show beyond a reasonable doubt that aquifer  $Y$  has higher concentrations (that is, differences observed are not likely to have occurred by chance alone). The reasonable doubt is set by  $\alpha$ , the significance level.

If mean concentrations are of interest and the  $t$ -test is chosen as the test procedure, each data group should be tested for normality. However, sample sizes of two and five are too small for a reliable test of normality. Here the objective is instead to test whether concentrations in one group are higher than the other, so the nonparametric rank-sum test is more appropriate. This test procedure entails ranking all seven values (lowest concentration has rank = 1, highest has rank = 7) and summing the ranks of the two 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 of two and five as an example.

What are the possible values  $W$  may take, given that the null hypothesis is true? The collection of all 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, and so on. Each 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 in this case, also the 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 listed in table 4.2 where probability is expressed as a fraction. For example, of 21  $W$  values, one is equal to three, therefore, the probability that  $W=3$  is 1/21.

What if the data collected produced two  $x$  values having ranks 1 and 4? Then  $W$  would be 5 (or 2 using R, as in table 4.2), lower than the expected value  $E[W]=8$ . If  $H_A$  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 fig. 4.3). 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 uncommon (about 1 chance in 5). Thus, the evidence against  $H_0$  is not too convincing. If the ranks of the two  $x$ s had been 1 and 2, then  $W=3$  (0 using R) and the  $p$ -value would be  $1/21=0.048$ . This result is much less likely than the previous case but is still about 5 percent. In fact, owing to such a small sample size the test can never result in a highly compelling case for rejecting  $H_0$ .

The one-sided rank-sum test performed in R using the Wilcoxon rank-sum test, `wilcox.test` function, on randomly generated data looks like the following:

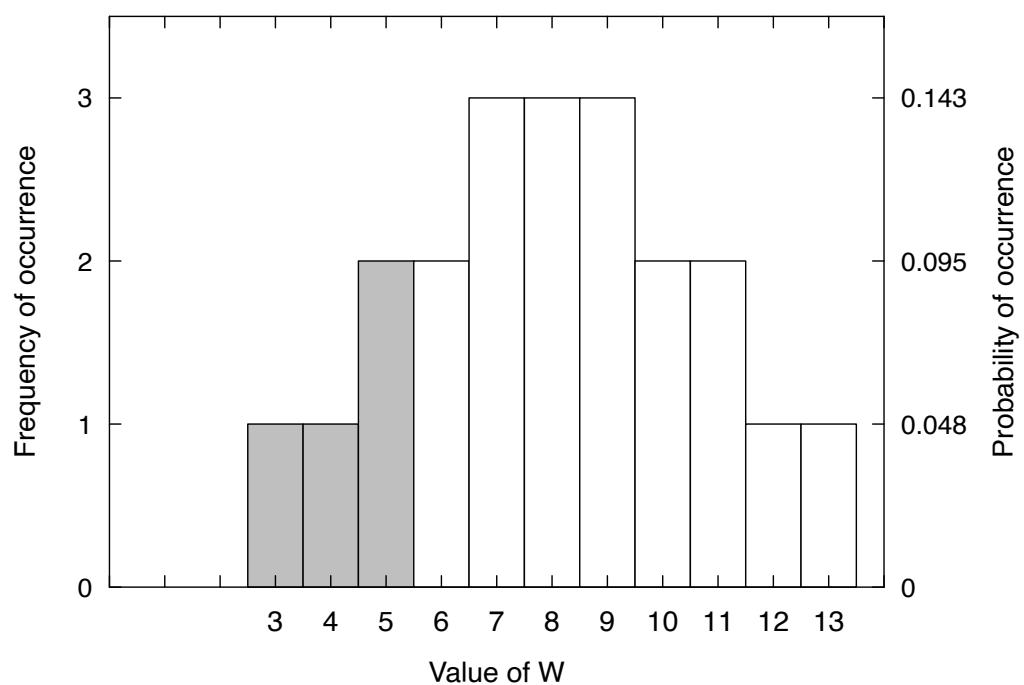
```
> # x is a random normal variable with mean 40 and standard
> # deviation of 5
> # y is a random normal variable with mean 50 and standard
> # deviation of 5
> # set.seed ensures that the authors and users will
> # have the same set of randomly generated numbers
> set.seed(100)
> x <- rnorm(2, mean=40, sd=5)
> y <- rnorm(5, mean=50, sd=5)
>
> wilcox.test(x, y, alternative="less")
```

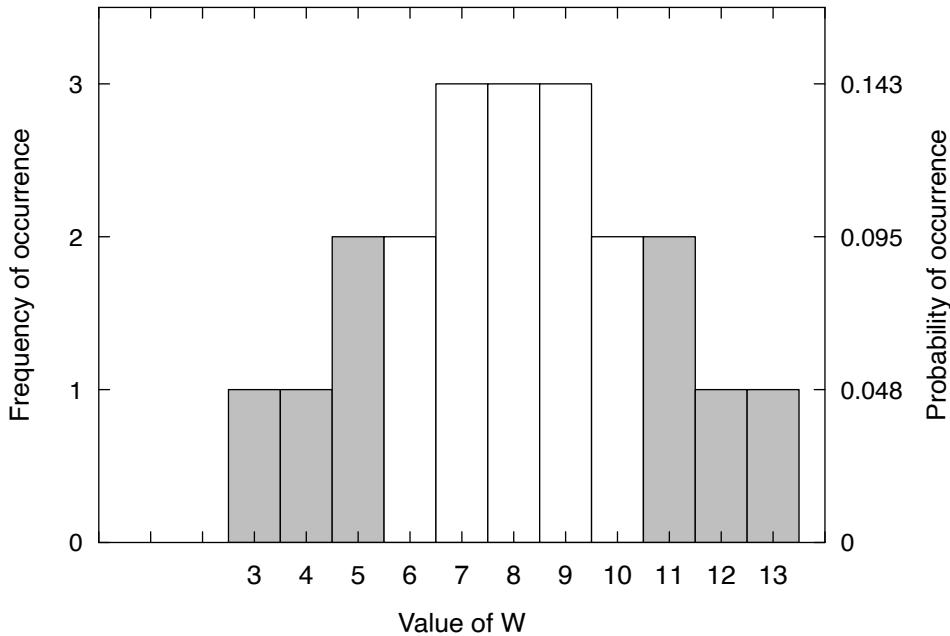
#### Wilcoxon rank sum test

```
data: x and y
W = 0, p-value = 0.04762
alternative hypothesis: true location shift is less than 0
```

**Table 4.2.** Probabilities and one-sided  $p$ -values for the rank-sum test with  $n=2$  and  $m=5$ .

$W$	Rescaled $W$ from R	$\text{Prob}(W)$	$\text{Prob}(\leq W)$
3	0	1/21	0.048
4	1	1/21	0.095
5	2	2/21	0.190
6	3	2/21	0.286
7	4	3/21	0.429
8	5	3/21	0.571
9	6	3/21	0.714
10	7	2/21	0.810
11	8	2/21	0.905
12	9	1/21	0.952
13	10	1/21	1.00

**Figure 4.3.** 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.



**Figure 4.4.** 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.

This example has considered only the one-sided  $p$ -value, which is appropriate when there is some prior notion that  $X$  tends to 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  $H_A$  is now that  $x_i$  and  $y_i$  are from different distributions, or

$$H_A: \text{Prob}(x_i \geq y_i) \neq 0.5 .$$

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 (see fig. 4.4). It is

$$\text{Prob}(W \leq 5) + \text{Prob}(W \geq 11) .$$

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

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

The two-sided rank-sum test performed in R using the `wilcox.test` function, on the same randomly generated data as the previous example but with a different alternative ( $H_A$ ) yields

```
> wilcox.test(x, y, alternative="two.sided")
```

```
Wilcoxon rank sum test

data: x and y
W = 0, p-value = 0.09524
alternative hypothesis: true location shift is not equal to 0
```

This example used a symmetric distribution, but the test can also be used with asymmetric distributions, in which case the probabilities in the two tails would differ. Fortunately, modern statistical software can handle symmetric or asymmetric distributions for us and reports two-sided *p*-values as the default, with a user option to select a one-sided alternative. For the alternative group A > group B, the option in R is `alternative = "greater"` for `wilcox.test`. For the reverse, `alternative = "less"` is the option.

To summarize, *p*-values describe the probability of calculating a test statistic as extreme or more extreme as the one observed, if  $H_0$  were true. The lower the *p*-value the stronger the case against the null hypothesis.

Now, let us look at an  $\alpha$ -level approach. Return to the original problem, the case of a one-sided test. Assume  $\alpha$  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 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. Searching table 4.2 for possible  $W$  values and their probabilities,  $W^* = 4$  because  $Prob(W \leq 4) = 0.095 \leq 0.1$ . If  $\alpha = 0.09$  had been selected then  $W^*$  would be 3.

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

$$Prob(W \leq W_L^*) + Prob(W \geq W_U^*) \leq \alpha \text{ and}$$

$$W_U^* - E[W] = E[W] - W_L^* .$$

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  $\alpha$ , without exceeding it, under the assumption that  $H_0$  is true. In the case at hand, if  $\alpha = 0.1$ , then  $W_L^* = 3$  and  $W_U^* = 13$  because

$$Prob(W \leq 3) + Prob(W \geq 13) = 0.048 + 0.048 = 0.095 \leq 0.1 .$$

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  $\alpha$ -level.

It is important to recognize 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.

## 4.4 Tests for Normality

If the objectives of a study are such that the analyst desires to use a parametric test that relies on an assumption of a normal distribution, a normality test should be used to determine whether data meet the test's requirements or whether a permutation test might be preferred. 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 follow a normal distribution, especially for small sample sizes. It simply says normality cannot be rejected with the evidence at hand.

Two tests for normality are used in this book, the probability plot correlation coefficient (PPCC) test introduced by Filliben (1975) and discussed by Looney and Gulledge (1985), and the Shapiro-Wilk test (Shapiro and others, 1968), as modified by Shapiro and Francia (1972). Both tests are related to and illustrated by a probability plot, a plot of data quantiles versus quantiles of the assumed theoretical distribution. These are two of the more powerful tests for normality available and both are reasonable choices for testing for normality. The PPCC requires the installation of an additional R package and can test hypotheses related to a number of other statistical distributions (Pohlert, 2017). The Shapiro-Wilk test (`shapiro.test`) is included in the base R installation (R Core Team, 2016).

Remember from chapter 2 that the more normal a dataset is, the closer it plots to a straight line on a normal probability plot. To determine normality, this linearity is tested by computing a PPCC test or a slightly modified  $R^2$  (Shapiro-Wilk test) between data and their standard normal quantiles (or normal scores, the linear scale on a probability plot). Samples from a normal distribution will have a correlation coefficient and  $R^2$  very close to 1. As data depart from normality, these statistics will decrease below 1. To perform a test of  $H_0$  (the data are normally distributed) versus  $H_A$  (the data are not normally distributed), the statistics are analyzed to see if they are significantly less than 1. These tests for normality use the Blom plotting position (see chap. 2 for more on plotting positions).

To illustrate these tests, table 4.3 and figure 4.5 show the unit well-yield data from chapter 2 and their probability plots. For the wells in unfractured rock, the probability plot correlation coefficient  $r^*=0.805$  (fig. 4.5A; see chap. 8 on correlation). This is the correlation coefficient between the yields ( $y_i$ ) and their associated plotting positions. For wells in fractured rock, the probability plot correlation coefficient  $r^*=0.943$  (fig. 4.5B).

The R code used to calculate the two tests for normality is given below. The `ppccTest` reports the probability correlation coefficient, the sample size, and a  $p$ -value. Using a 5-percent significance level,  $\alpha=0.05$ , and with a sample size of 12, the hypothesis that the yield of unfractured wells is normally distributed is rejected,  $p$ -value  $<0.05$ . We can see this visually in figure 4.5A. Using the same significance level and with a sample size of 13, the hypothesis that the yield of fractured wells follows a normal distribution is not rejected,  $p$ -value  $>0.05$ .

The function `shapiro.test` reports a test statistic,  $W$ , and a  $p$ -value. At a 5-percent significance level, the results are the same as for the PPCC test—the hypothesis that the yield of unfractured wells is normally distributed is rejected and the hypothesis that the yield of fractured wells is normally distributed is not rejected.

```
> library(ppcc)
> unit.well.yields <- data.frame(y = c(0.001, 0.003, 0.007, 0.020,
+ 0.030, 0.040, 0.041, 0.077, 0.100, 0.454,
+ 0.490, 1.02, 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),
+ frac = c(rep(0, 12), rep(1, 13)))
>
> unfrac <- subset(unit.well.yields, frac == 0)
> frac <- subset(unit.well.yields, frac == 1 )
> ppccTest(unfrac$y, qfn = "qnorm")
```

#### Probability Plot Correlation Coefficient Test

```
data: unfrac$y
ppcc = 0.80508, n = 12, p-value = 1e-04
alternative hypothesis: unfrac$y differs from a Normal distribution

> ppccTest(frac$y, qfn = "qnorm")
```

#### Probability Plot Correlation Coefficient Test

```

data: frac$y
ppcc = 0.94258, n = 13, p-value = 0.0886
alternative hypothesis: frac$y differs from a Normal distribution

> shapiro.test(unfrac$y)

Shapiro-Wilk normality test

data: unfrac$y
W = 0.66039, p-value = 0.0003593

> shapiro.test(frac$y)

Shapiro-Wilk normality test

data: frac$y
W = 0.88531, p-value = 0.08418

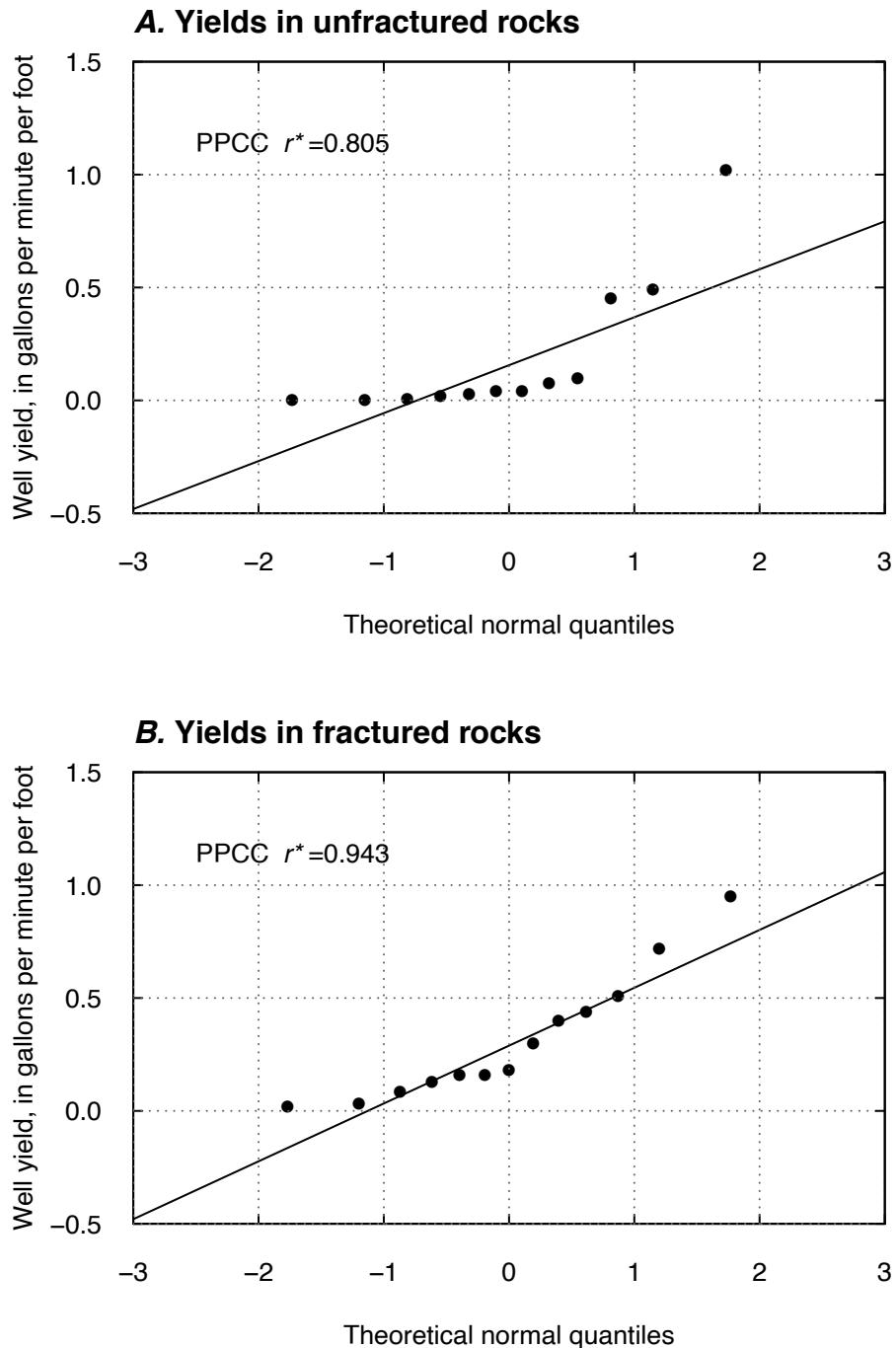
```

The PPCC and Shapiro-Wilk tests have an important graphical analog, the probability plot, which illustrates the test results. The probability plot provides information on how the data depart from normality, such as whether a transformation for skewness will improve the fit (or not), something not provided by any test statistic. A plot can be worth a thousand test statistics! Note the nonlinearity of the data on the normal probability plots of figure 4.5. The probability plot for the well yields in fractured rock (fig. 4.5B) shows a closer adherence to a straight line (normality not rejected) than for the well yields in unfractured rock (fig. 4.5A) where normality is rejected at  $\alpha=0.05$ .

**Table 4.3.** Unit well yields ( $y_i$ ) from Virginia, in gallons per minute per foot (Wright, 1985).

[-, no data]

Wells in unfractured rock ( $y_i$ )	Wells in fractured rock ( $y_i$ )
0.001	0.020
0.003	0.031
0.007	0.086
0.020	0.13
0.030	0.16
0.040	0.16
0.041	0.18
0.077	0.30
0.10	0.40
0.454	0.44
0.49	0.51
1.02	0.72
-	0.95



**Figure 4.5.** Probability plots for yields of wells in (A) unfractured and (B) fractured rock, with probability plot correlation coefficient (PPCC) correlation coefficient ( $r$ ). Data from Wright (1985).

The most common test for normality is the Shapiro-Wilk test, as its power to detect non-normality is as good or better than other tests (Thode, 2002). A table of quantiles for this test statistic is available for  $n < 50$  (Conover, 1999). Shapiro and Francia (1972) modified the Shapiro-Wilk test for all sample sizes and statistical software usually performs this form of the test, including the `shapiro.test` command in R. In this book, we have adopted the common practice of calling the Shapiro-Francia modification produced by the `shapiro.test` command the Shapiro-Wilk test. Power characteristics and  $p$ -values for the Shapiro-Wilk and PPCC tests should be similar.

Tests for normality not related to probability plots include the Kolmogorov-Smirnov (one-sample `ks.test` in R) and chi-square tests (`chi-sq.test` with  $x$  as a numeric vector and no  $y$  in R), described in more detail by Thode (2002). Both tests have lower power than the Shapiro-Wilk test to detect non-normality when data are continuous (Thode, 2002). Kolmogorov-Smirnov and chi-square tests are more useful for data that are ordinal (data recorded only as low/medium/high, for example), but this makes them less powerful than the probability plot tests for testing continuous data for normality (Shapiro and others, 1968). The Anderson-Darling test (Thode, 2002) is useful for testing a wide variety of distributions in addition to the normal. The Anderson-Darling test is not based on probability plots, but its power characteristics are similar to the PPCC and Shapiro-Wilk tests.

## 4.5 Other Hypothesis Tests

Many other hypothesis tests exist for questions beyond central tendency and distribution; table 4.1 of this chapter lists some other tests and the chapters in which they are introduced. Additional tests include tests for proportions, tests of independence, tests related to variance or spread, and tests related to skew and kurtosis, among many others. There is an extensive discussion of tests for equality of variances in chapter 5. Readers may consult specialized texts such as Sheskin (2011) for details about many more hypothesis tests.

## 4.6 Considerations and Criticisms About Hypothesis Tests

Though widely used, hypothesis tests and  $p$ -values are subject to misinterpretation and much criticism. Over the decades many have tried to discourage the use of  $p$ -values in published research, yet they remain widely reported because they are simply an expression of strength of evidence shown by the data. In some instances, if authors do not report a  $p$ -value with their results, reviewers will ask for a  $p$ -value as a measure of the statistical significance of the results.

### 4.6.1 Considerations

The  $p$ -values reported when performing hypothesis tests are an indicator of statistical, or probabilistic, significance but are not a measure of hydrologic importance. Authors should consider both statistical and hydrological (or practical) significance. As McCuen (2016) stated, “Widely accepted criteria are usually not available to assess hydrological significance, which unfortunately means that greater weight is often placed on statistical decision criteria; however, efforts should always be made to assess hydrological significance.” For example, one may find a statistically significant difference in the concentration of calcium in stream samples collected at two different sites. However, that difference might not have human or aquatic health effects.

The  $p$ -value is often misused and misinterpreted (Wasserstein and Lazar, 2016), sometimes as the probability of the null hypothesis being true, sometimes as the probability of some event, or sometimes as the likelihood of an outcome. The  $p$ -value is the probability of calculating a test statistic as extreme or more extreme as the one observed, if  $H_0$  were true. It is the probability of declaring that there is a signal in the data when one does not exist (a false positive). “The smaller the  $p$ -value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the  $p$ -value hold true” (Wasserstein and Lazar, 2016); however, there is no definitive point at which statistical incompatibility can be declared the truth.

When reporting hypothesis test results, analysts should also report sample size (as results can be sensitive to sample size and in some cases small samples require a special hypothesis test or a modification to a hypothesis test); the null and alternative hypotheses (so that readers know explicitly what is being tested); the significance level used to determine results, such as  $\alpha=0.05$ ; and the actual  $p$ -value, not just  $p<0.05$ , so that readers may make their own judgments about significance.

## 4.6.2 Criticisms

$P$ -values have been “vigorously excoriated” (Rozeboom, 1960; Nuzzo, 2014) for decades, yet remain widely used. Some of these criticisms are discussed below.

1.  $P$ -values do not convey the magnitude or importance of an effect.
2.  $P$ -values do not provide the range of plausible values that confidence or prediction intervals do.
3. There can be a bias against publishing results that fail to reject the null hypotheses. A nonstatistically significant result may be considered by some to be nonsignificant scientifically; however, a result of no change can tell us something important about the environment.
4. Researchers sometimes try multiple hypothesis tests, removal of outliers, or collection of more data to achieve statistically significant results. This process was termed “ $p$ -hacking” by Uri Simonsohn (Nuzzo, 2014).

In response to criticisms to 1 and 2,  $p$ -values were never designed to estimate the effect size. Other methods, such as the Hodges-Lehmann estimator (Hodges and Lehmann, 1963; Hollander and Wolfe, 1999; discussed in chap. 5) have been designed to estimate effect size. Estimators and  $p$ -values are complementary and both insufficient by themselves. The problem with estimating a mean effect and leaving it at that, is that no one knows whether the magnitude of the reported effect is anything other than noise.

We agree with criticism 3 entirely. No change is sometimes the most welcome result or can inform those in water resources that some action did not have the hoped-for results. There is a growing interest in science in general in publishing nonsignificant results because this can still inform future work (Charlton, 2004; Kotze and others, 2004; Levine, 2013; Goodchild van Hilten, 2015; World Health Organization, 2015; Lederman and Lederman, 2016).

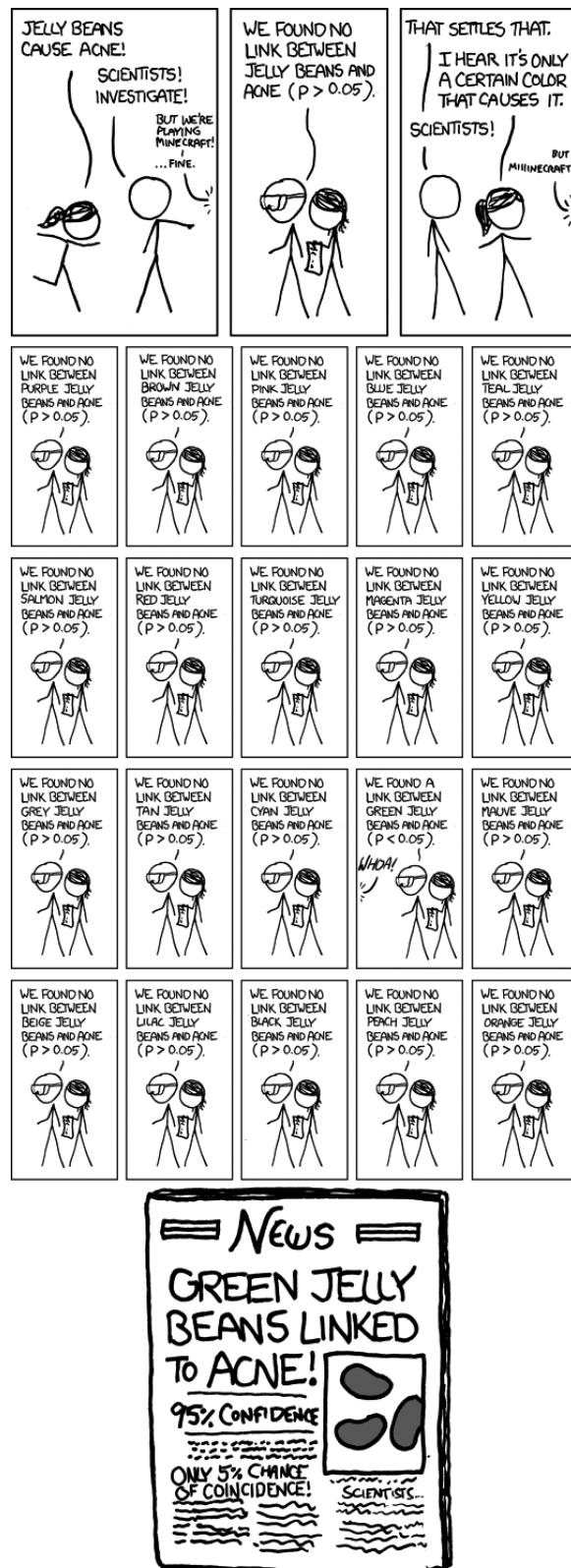
We agree that criticism 4 is a problem but banning  $p$ -values is not the solution. A humorous illustration of the all-too-common problem of multiple hypothesis tests is shown in figure 4.6. Always remember that  $\alpha=0.05$  states that rejection of the null hypothesis can be expected 1 out of 20 times (5 percent) simply by chance.

There are many ways that researchers can misrepresent their data, sometimes on purpose, but often through ignorance. The authors of this book have never advocated removing outliers to obtain a desired result; the discipline of statistics does not advocate this either. The remedy for the problems with  $p$ -values and hypothesis testing is better education on statistical theory and methods for environmental scientists. Hypothesis tests are just the inverse of confidence intervals, so the call to ban tests on one hand while advocating more reporting of confidence intervals on the other hand is ironic. A  $p$ -value is simply one minus the confidence level for the widest possible confidence interval showing a nonzero effect size.

## 4.6.3 Discussion

Because of criticisms and concerns summarized above, the American Statistical Association (ASA) for the first time ever took a position on specific matters of statistical practice and published a statement on statistical significance and  $p$ -values (Wasserstein and Lazar, 2016). Wasserstein and Lazar did not call for a ban on the use of  $p$ -values, but instead chose to clarify widely agreed upon principles underlying the use and interpretation of  $p$ -values, quoted directly here:

1.  $P$ -values can indicate how incompatible the data are with a specified statistical model;
2.  $P$ -values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone;



**Figure 4.6.** Cartoon showing what can easily happen when running multiple hypothesis tests,  $p$ -hacking or  $p$ -fishing. Figure from xkcd.com (Munroe, 2016), used under creative commons attribution-noncommercial license.

3. *Scientific conclusions and business or policy decision should not be based only on whether a p-value passes a specific threshold;*
4. *Proper inference requires full reporting and transparency;*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result;*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

We agree with these principles and have provided a few examples to consider the role of *p*-values in hydrological studies. Statistical tests often are conducted to help make a decision about the need for action. For example, we may be interested in determining if exposure to some chemical at a particular concentration has an impact on reproductive success in a fish species. Let us say that the *p*-value from a one-sided test is 0.20. Generally, one might set  $\alpha$  at 0.05, but the 0.20 value is telling us that there is reasonably strong evidence that the chemical has a negative effect although we still have some degree of doubt about it. The precautionary principle might suggest action to control the level of that chemical because it is more likely than not that it has a negative effect.

Another example in which the use of *p*-values can be harmful to the interpretation of results is the case when many similar hypothesis tests are being conducted and we want to use them to evaluate many sites. We may be conducting similar tests at many locations, such as trends in chloride in rivers, trends in nitrate in wells, or trends in floods at many streamgages. It may be the case that few or even none of the sites in the study may have significant trends ( $p \leq \alpha$ ), but it may be that many sites had trends in the same direction and many of those were of a moderate level of significance (say  $p \leq 0.2$ ). It is common in such situations that when the results are not statistically significant the author does not provide information about the magnitude, sign, or *p*-value. The consequence is a considerable loss of information to the reader of the report. The full results from the test (magnitude, sign, and *p*-value) are all very useful information. For example, knowing that most or all test statistics were of one sign is a useful piece of information even though no individual test statistic is significant. See Hirsch and Ryberg (2012) for an example of such an analysis. There are formal tests for tests conducted across many sites (see chap. 12, Regional Kendall test for example), but even informally, showing all results regardless of significance is a good practice. However, one should consider the spatial correlation in the data. If one has many sites and some are on the same river or near each other, they may be spatially correlated and the results for 10 such sites do not contain as much unique information as 10 independent sites scattered across the country would.

As the ASA's statement on *p*-values says, "The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves" (Wasserstein and Lazar, 2016). Graphical analyses can complement quantitative assessments that include statistical hypothesis tests. Consider how the graphics shown throughout this text may support a finding of statistical significance or nonsignificance or may support the hydrologic importance of a finding.

## Exercises

1. The following are annual streamflows, in cubic feet per second, for the Green River at Munfordville, Kentucky. Beginning in 1969 the stream was regulated by a reservoir.

Test both before and after datasets for normality using the PPCC test. If either is non-normal, transform the data and retest in order to find a scale that appears to be close to a normal distribution.

Year	Value	Year	Value
Before		After	
1950	4,910	1969	1,350
1951	3,660	1970	2,350
1952	3,910	1971	3,140
1953	1,750	1972	3,060
1954	1,050	1973	3,630
1955	2,670	1974	3,890
1956	2,880	1975	3,780
1957	2,600	1976	3,180
1958	3,520	1977	2,260
1959	1,730	1978	3,430
1960	2,340	1979	5,290
1961	2,600	1980	2,870
1962	3,410		
1963	1,870		
1964	1,730		
1965	2,730		
1966	1,550		
1967	4,060		
1968	2,870		

2. Test the arsenic data and transformed data of chapter 2, exercise 2 for normality.



# Chapter 5

## Testing 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 the concentrations greater at the  $\alpha = 0.01$  significance level? Does the magnitude of the difference warrant the expense of cleanup?

Measurements of a biological diversity index are made on 16 streams. Eight of the streams represent natural conditions and the other eight have received urban runoff. Is the biological quality of the urban streams estimated to be degraded in comparison to the natural streams?

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

The examples given above compare 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, and so forth. Where such a pairing does exist, methods for matched pairs discussed in chapter 6 should be used. Data should also be independent in the sense that the two groups represent different conditions—neither observations nor the population they represent should be in both groups.

This chapter will discuss nonparametric, permutation, and parametric tests for whether two independent groups differ in central location (see chap. 4 for a definition of these three classes of hypothesis tests). Graphical presentations of the test results will be quickly surveyed, methods for estimating the magnitude of the difference between the two groups provided, and methods for testing differences in the variability of two groups described. An overview of the types of tests considered in this chapter is given in table 5.1.

**Table 5.1.** Hypothesis test methods in this chapter and their characteristics.  $H_A$  is the alternative hypothesis, the signal to be found if it is present.

Objective ( $H_A$ )	Test	Class of test	Distributional assumption	Estimator of difference
Data values in one group are frequently higher than those in the other group	Wilcoxon rank-sum test	Nonparametric	None	Hodges-Lehmann estimate
One group has a higher mean	Two-sample <i>t</i> -test	Parametric	Normal distribution. Differences additive	Mean difference
	Two-sample permutation test	Permutation	Same distribution as in the other group	Mean difference
One group has higher variability	Fligner-Killeen	Nonparametric	None	Difference in median absolute distance from the median
	Levene's	Parametric	Normal distribution	Difference in group variance

## 5.1 The Rank-sum Test

The rank-sum test goes by many names. The test 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 (1947) and the test statistics can be derived one from the other; thus the test is also known as the Mann-Whitney test. The combined name of Wilcoxon-Mann-Whitney rank-sum test has also been used, as has the Two-sample Wilcoxon test. Regardless of the many names, two end results are important. Does the test conclude there is a significant difference between the two groups for a given level of significance? If so, what is the magnitude of that difference? Answers to these questions for the rank-sum test are given in the following discussion.

### 5.1.1 Null and Alternate Hypotheses for the Rank-sum Test

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

$$H_0: \text{Prob}(x_i > y_j) = 0.5, i=1,2,\dots,n; j=1,2,\dots,m,$$

where the  $x_i$  are from one group and the  $y_j$  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

$$H_{A1}: \text{Prob}(x_i > y_j) \neq 0.5 \quad (\text{Two-sided test, } x \text{ might be larger or smaller than } y).$$

$$H_{A2}: \text{Prob}(x_i > y_j) > 0.5 \quad (\text{One-sided test, } x \text{ is expected to be larger than } y).$$

$$H_{A3}: \text{Prob}(x_i > y_j) < 0.5 \quad (\text{One-sided test, } x \text{ is expected to be smaller than } y).$$

The rank-sum test is often presented as a test for difference in group medians. This follows from the form of the hypotheses above when the two groups have the same distributional shape. However, the test is more general than a test for differences in medians. For example, suppose the lower half of two sites' concentration distributions were similar, but a contaminant elevated the upper 40 percent of one site's concentrations. Group medians might not significantly differ, but the rank-sum test may find a significant difference because the upper 40 percent of concentrations at the contaminated site were higher than the upper 40 percent at the uncontaminated site.

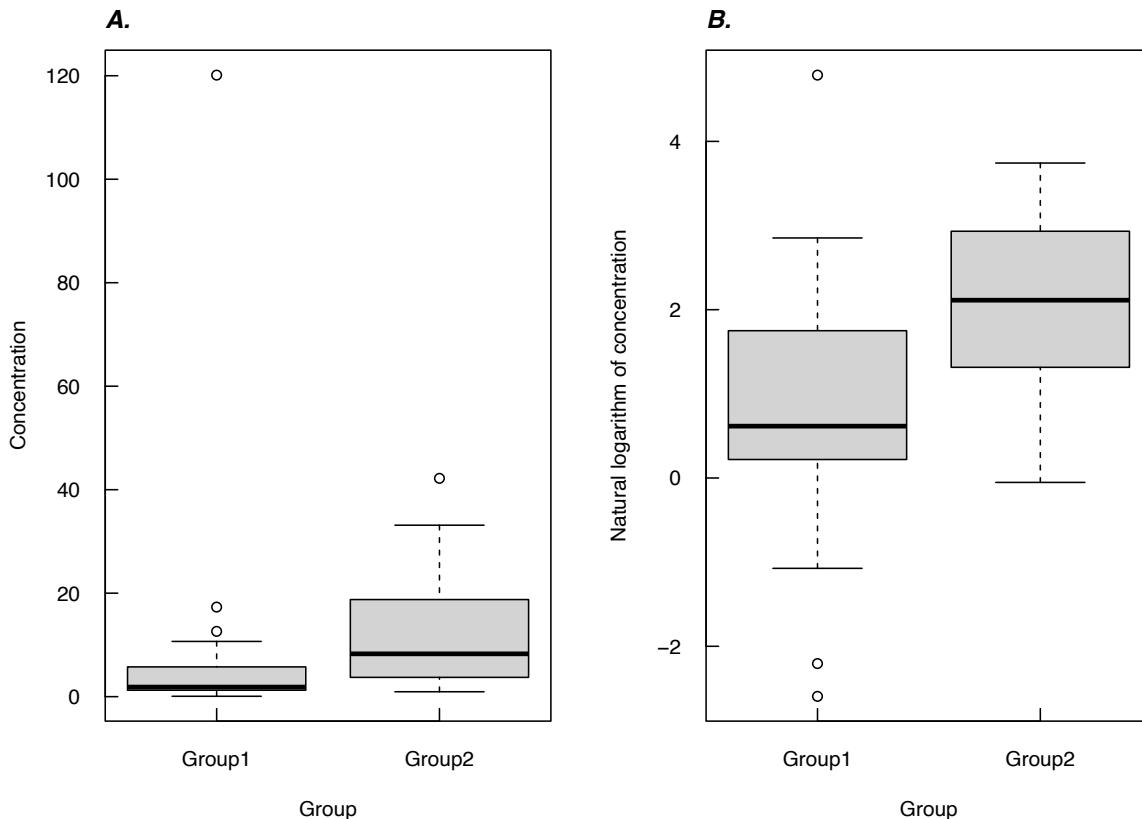
### 5.1.2 Assumptions of the Rank-sum Test

There are three assumptions for the rank-sum test (Conover, 1999):

1. Data in both groups are random samples from their respective populations.
2. In addition to independence of data within each group, there is mutual independence between the two groups. For example, data from the same sampling unit (and certainly the exact same observations) should never be present in both groups.
3. The measurement scale is at least ordinal.

There is no requirement of equal variances or normality of the distribution of data. 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 objective is to determine whether one group tends to produce generally higher observations than the other, 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 5.14. On the original scale, the data have different sample medians and interquartile ranges (IQRs), as shown by the



**Figure 5.1.** Box plots of (A) concentration data for two groups and (B) the logarithms of the same data. The pattern in A is typical of lognormal distributions (on the original scale) showing a multiplicative difference between groups. This produces the additive relation seen in B.

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? The logarithms of the data appear to have different medians (fig. 5.1B) but similar IQRs, and thus the logs of the data appear to differ only in central tendency. The test statistic and  $p$ -value for a rank-sum test computed on these transformed data is identical to that for the original scale! Nonparametric tests possess the useful property of being invariant to positive power transformations such as square and cube roots, as well as logarithms. Only the data, or any power transformation of the data, need be similarly shaped (except for their central location) to use the rank-sum test, so it is applicable in many situations. Unlike the  $t$ -test, the rank-sum test can discern multiplicative differences between groups, such as  $y=3 \cdot x$  (see section 5.3.2.).

### 5.1.3 Computation of the Rank-sum Test

For sample sizes  $n$  and  $m$  where  $n < m$ , and  $x_i, i=1, 2, \dots, n$  and  $y_j, j=1, 2, \dots, m$  are the two data groups, compute the joint ranks  $R_k$ :

$R_k = 1$  to  $(N=n+m)$ , using average ranks in case of ties. Then the test statistic

$W_{rs} = \text{sum of ranks for the group having the smaller sample size, or}$   
 $= \sum R_i$  from  $i=1, 2, \dots, n$  (using either group with equal sample sizes  $n=m$ ).

A one-sided or one-tailed alternative should be chosen when one group is expected to be higher or lower (but not both!) than the second group prior to observing the data. For example,  $y$  is a background site with lower concentrations expected than for a possibly higher-concentration site  $x$ . Determine the  $p$ -value associated with  $W_{rs}$ . Reject  $H_0$  when  $p < \alpha$ .

R provides the exact  $p$ -value for small to moderate sample sizes unless there are ties, in which case the large-sample approximation is provided. Commercial statistics packages typically report  $p$ -values using the large-sample approximation for all sample sizes.

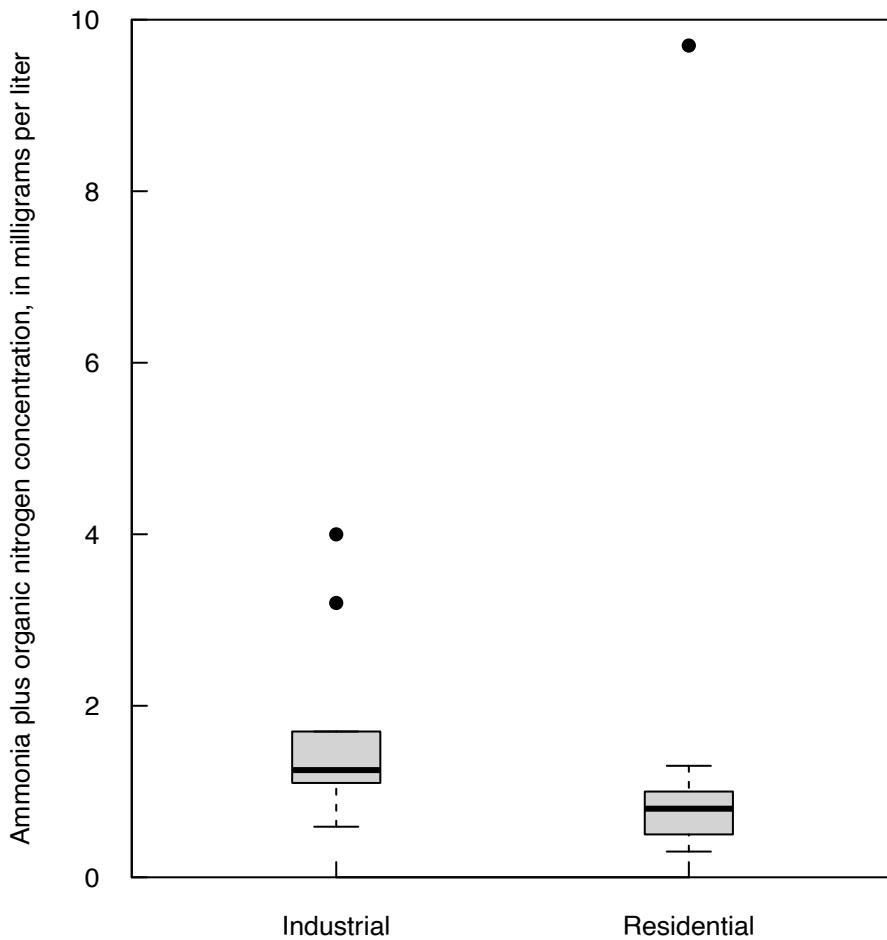
**Example 5.1. Precipitation nitrogen—Rank-sum test.**

Precipitation quality was compared at sites with different land uses by Oltmann and Shulters (1989). Ten concentrations of organic plus ammonia nitrogen (NH4orgN) at each site are listed below, along with their group location as the variable "where". The rank command below computes the joint ranks of concentrations from 1 to 20 only to illustrate what is performed internally by the `wilcox.test` rank-sum command. Note that three pairs of concentrations (at 0.7, 1.1, and 1.3 milligrams per liter [mg/L]) are tied, and so are assigned tied ranks equal to the average of their two individual ranks.

```
> load (precipn.RData)
> attach (precipn)
> precipn$rankN <- rank(precipn$NH4orgN)
> print (precipn)
```

	NH4orgN	where	rankN
1	0.59	indust	4.0
2	0.87	indust	7.0
3	1.10	indust	11.5
4	1.10	indust	11.5
5	1.20	indust	13.0
6	1.30	indust	14.5
7	1.60	indust	16.0
8	1.70	indust	17.0
9	3.20	indust	18.0
10	4.00	indust	19.0
11	0.30	residen	1.0
12	0.36	residen	2.0
13	0.50	residen	3.0
14	0.70	residen	5.5
15	0.70	residen	5.5
16	0.90	residen	8.0
17	0.92	residen	9.0
18	1.00	residen	10.0
19	1.30	residen	14.5
20	9.70	residen	20.0

Boxplots for the groups are shown in figure 5.2.



**Figure 5.2.** Boxplots of ammonia plus organic nitrogen from the precipn data. Data from Oltmann and Shulters (1989) by land use type: industrial or residential.

Median concentrations for the industrial and residential sites are 1.25 and 0.80 mg/L, respectively. The rank-sum test determines if ammonia plus organic nitrogen concentrations ( $\text{NH}_4\text{orgN}$ ) differ significantly ( $\alpha=0.05$ ) between the industrial (`indust`) and residential (`residen`) sites. The null ( $H_0$ ) and alternate ( $H_A$ ) hypotheses are

$$H_0: \text{Prob}(\text{concentration [industrial]} \geq \text{concentration [residential]}) = 0.5.$$

$$H_A: \text{Prob}(\text{concentration [industrial]} \geq \text{concentration [residential]}) \neq 0.5.$$

The test statistic is the sum of ranks in the group with fewer observations. Here either group could be used because sample sizes are equal. Choosing the residential group, the sum of ranks is 78.5. An exact test cannot be computed with a fractional test statistic, so the large-sample approximation form of the test will automatically be computed. Note that R subtracts the smallest possible test statistic prior to reporting the result. Here the smallest possible value for  $W_{rs}$  equals 2, so the test statistic reported by R equals 76.5.

```
>wilcox.test(NH4orgN~where, conf.int=TRUE)
```

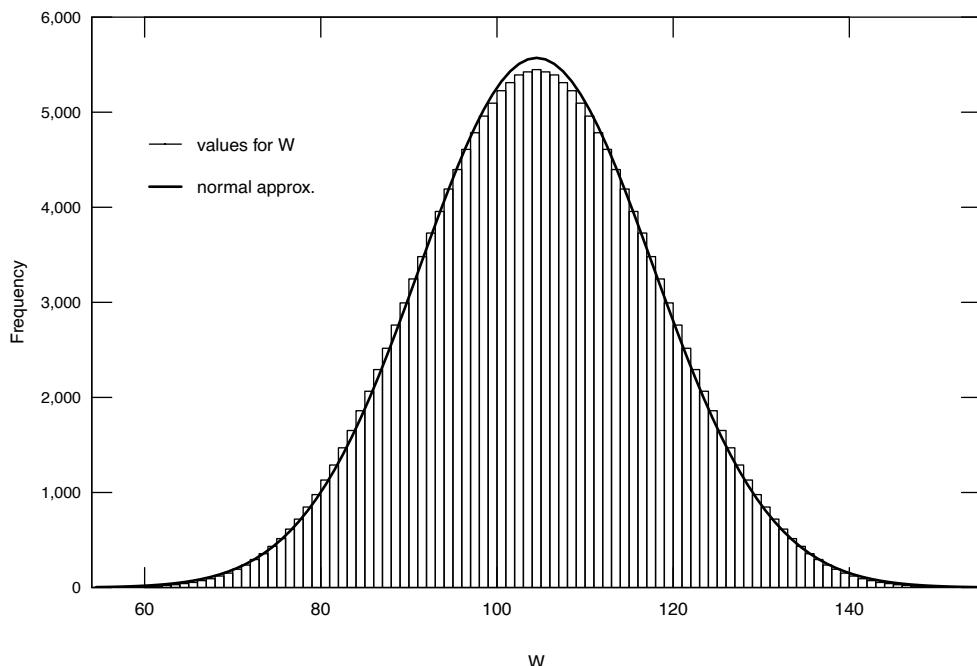
```
Wilcoxon rank sum test with continuity correction
```

```
data: NH4orgN by where
W = 76.5, p-value = 0.04911
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
3.948029e-05 1.099984e+00
sample estimates:
difference in location
0.5040993
```

The conclusion is that ammonia plus organic nitrogen concentrations from industrial precipitation differ significantly from those in residential precipitation at these locations by a median difference of 0.504. This estimate is the Hodges-Lehmann estimate discussed later in section 5.5 and is presented along with its confidence interval when specifying the option `conf.int=TRUE`.

### 5.1.4 The Large-sample Approximation to the Rank-sum Test

For the rank-sum test, the distribution of the exact test statistic  $W_{rs}$  is closely approximated by a normal distribution when the sample size for each group is 10 or more (fig. 5.3). With  $n=m=10$ , there are 184,756 possible arrangements of the data ranks (this can be computed with the `choose(20, 10)` command in R). The sum of ranks for one of the two groups for all arrangements comprises the exact distribution of



**Figure 5.3.** Histogram showing the distribution of the exact test statistic  $W_{rs}$  and its fitted normal approximation for  $n=10$  and  $m=10$ .

$W_{rs}$ , shown as bars in figure 5.3, with a mean of 105. Superimposed on the exact distribution is the normal distribution that closely approximates the exact values. This demonstrates how well the  $p$ -values can be approximated even for relatively small sample sizes. The 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 and the assumptions of  $H_0$  are valid,  $W_{rs}$  has a mean,  $\mu_W$ , and standard deviation,  $\sigma_W$ , of

$$\mu_W = n \cdot (N+1) / 2 \quad (5.1)$$

$$\sigma_W = \sqrt{n \cdot m \cdot (N+1)/12} , \quad (5.2)$$

where

$$N = n + m.$$

The  $p$ -value from the large-sample approximation is computed by standardizing  $W_{rs}$  and making a continuity correction. The continuity correction shifts the normal distribution to fit halfway through the top of the bars of the exact test statistic distribution. 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

$$Z_{rs} = \begin{cases} \frac{W_{rs} - \frac{d}{2} - \mu_W}{\sigma_W} & \text{if } W_{rs} > \mu_W \\ 0 & \text{if } W_{rs} = \mu_W , \\ \frac{W_{rs} + \frac{d}{2} - \mu_W}{\sigma_W} & \text{if } W_{rs} < \mu_W \end{cases} \quad (5.3)$$

$Z_{rs}$  is the quantile of the standard normal distribution from which the  $p$ -value is computed. For the precipitation nitrogen in example 5.1 the approximate  $p$ -value is 0.0491 (see the R output in the previous section). Reporting the  $p$ -value shows how close the risk of type I error is to 0.05.

Note that a tie correction for the standard deviation of the large-sample test statistic  $\sigma_W$  is necessary when ties occur and tied ranks are assigned (Conover, 1999). The formula below for  $\sigma_W$  should be used for computing the large-sample approximation rather than the uncorrected  $\sigma_W$  whenever ties occur. This is done for you in statistical software, including R.

$$\sigma_{Wt} = \sqrt{\frac{nm}{N(N-1)} \sum_{k=1}^N R_k^2 - \frac{nm(N+1)^2}{4(N-1)}} , \quad (5.4)$$

where

$$N = n + m.$$

## 5.2 The Permutation Test of Difference in Means

Permutation tests solve the long-standing riddle of how to test for differences between means for skewed, moderately sized datasets. They compute the  $p$ -value using computer-intensive methods (see section 5.2.2.) rather than assuming data follow normal distributions (see section 4.1.2.). Permutation tests are also called resampling methods (Good, 2001), randomization tests (Manly, 2007), and observation

randomization tests (Brown and Rothery, 1993). Although they were conceived of in the early 1900s, software for quickly computing them became available around the late 1980s.

A user-friendly way to compute the two-sample permutation test in R is provided by the script `perm2` provided in supplemental material (SM.5). The test can also be computed using the `permTS` command in the `perm` package of R (Fay and Shaw, 2010). If you are selecting a commercial software program, the authors of this book strongly recommend that you look for one that provides permutation tests as alternatives to traditional parametric methods.

### 5.2.1 Assumptions of the Permutation Test of Difference in Means

A two-sample permutation test for differences in means avoids the assumptions of the parametric *t*-test (section 5.3). The *t*-test requires that the data from each group follow a normal distribution and that the groups have the same variance. Violation of these assumptions leads to a loss of power, raising *p*-values and failing to find differences between group means when they occur. These assumptions are avoided by using a permutation test. The permutation test assumes only that the data from the two are exchangeable (Good, 2001). The exchangeable assumption is that any value observed in one group may belong in the population of either group.

### 5.2.2 Computation of the Permutation Test of Difference in Means

Permutation tests calculate either all of the possible test results that could be computed for the observed data or a large random selection of those results, and then determine what proportion of the computed results are equal to or more extreme than the one result obtained using the dataset tested. That proportion is the *p*-value of the test.

For a two-sample permutation test of means, the test statistic is the observed difference in the two group means,  $\bar{x} - \bar{y}$ . If the null hypothesis is true, the group assignment is arbitrary, as there is no difference in the means and the data in essence come from the same population. Therefore, the data are rearranged regardless of group assignment in either all possible rearrangements or in several thousand randomly selected rearrangements. This produces a different set of numbers assigned to the two groups in each rearrangement. The difference in group means is computed and stored after each rearrangement, representing the distribution of differences to be expected when the null hypothesis is true. The proportion of differences from the rearrangements that equal or exceed the one observed difference from the original data is the permutation *p*-value of the test.

#### Example 5.2. Precipitation nitrogen—Permutation test of difference in means.

With  $n=m=10$ , there are 184,756 possible rearrangements of assigning data to groups. As an example, one rearrangement for the precipitation nitrogen data is found in the third column of R output below. Instead of computing all of these assignments, the permutation procedure will randomly rearrange the group assignment many thousands of times and compute the difference in the resulting means.

NH4orgN	where	rearrangement of where
1 0.59	indust	indust
2 0.87	indust	indust
3 1.10	indust	residen
4 1.10	indust	residen
5 1.20	indust	residen
6 1.30	indust	indust

7	1.60	indust	residen
8	1.70	indust	indust
9	3.20	indust	residen
10	4.00	indust	residen
11	0.30	residen	indust
12	0.36	residen	indust
13	0.50	residen	indust
14	0.70	residen	residen
15	0.70	residen	residen
16	0.90	residen	indust
17	0.92	residen	residen
18	1.00	residen	residen
19	1.30	residen	indust
20	9.70	residen	indust

The default number of rearrangements used in the perm2 script is R = 10000. Increasing the number of permutations increases the precision of the reported *p*-value.

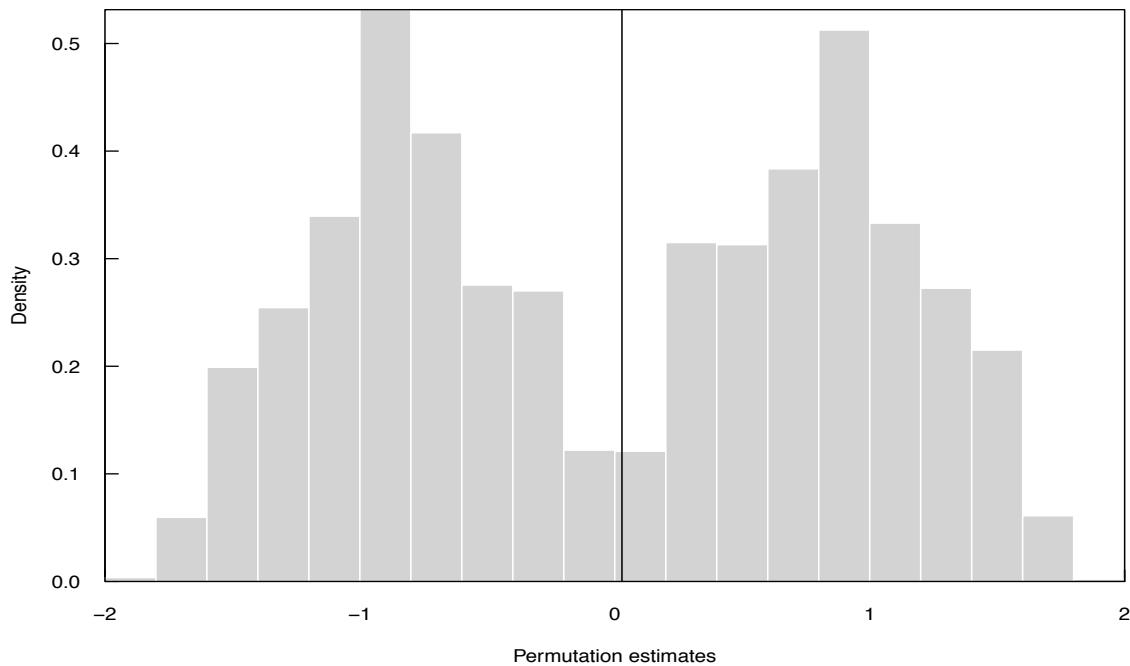
```
> perm2(NH4orgN,where)
```

```
Permutation Test of Difference Between 2 Group Means
Data: NH4orgN by where
Number of Possible Permutations is greater than 1000
```

```
R = 10000 pvalue = 0.9955
Alt Hyp: true difference in means is not equal to 0
```

```
sample estimates:
mean of indust = 1.666   mean of residen = 1.638
Diff of means = 0.028
95 percent confidence interval
-2.052  1.445
```

Out of 10,000 possible rearrangements of the where column, 99.5 percent of the absolute value of the estimated differences equaled or exceeded the observed difference of 0.028 (fig. 5.4). Therefore, the observed difference in means is not unusual at all and the permutation *p*-value is far greater than any reasonable significance level. The conclusion is to fail to reject  $H_0$ . There is little evidence that the group means differ. The advantage of this test over a *t*-test is that there is no concern that the nonsignificant result might be a result of the unfulfilled requirement that the input data follow a normal distribution.



**Figure 5.4.** Histogram showing 10,000 permuted differences in group means for the precipn dataset, computed by rearrangement of the group assignments. The observed difference in means from the original data is the solid vertical line.

### 5.3 The *t*-test

The *t*-test has been the most widely used method for comparing two independent groups of data and 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 rank-sum or permutation tests. These are

1. Lack of power when applied to skewed data,
4. Dependence on an additive model,
5. Lack of applicability for censored data,
6. Assumption that the mean is a good measure of central tendency for skewed data, and
7. 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).

#### 5.3.1 Assumptions of the *t*-test

In order to compute an accurate *p*-value the *t*-test assumes that both groups of data are normally distributed around their respective means. The test originally also assumed that the two groups have the same variance—a correction for unequal variance was added later. 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 assumptions of normality and equal variance are rarely satisfied with water resources data. The null hypothesis is stated as

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

If rejected, the alternative hypothesis is either two-sided or one-sided:

$$H_0: \mu_x \neq \mu_y \text{ (two-sided)}$$

$$H_0: \mu_x > \mu_y \text{ (one-sided)}$$

### 5.3.2 Computation of the Two-sample *t*-test Assuming Equal Variances

Two independent groups of data are to be compared. Each group is assumed to be normally distributed around its respective mean value, with each group having the same variance. The sole difference between the groups is that their means may not be the same—one is an additive shift from the other.

The test statistic (*t* in eq. 5.5) is the difference in group means, divided by a measure of noise:

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n} + \frac{1}{m}}} , \quad (5.5)$$

where

$\bar{x}$  is the sample mean of data in the first group  $x_i$  from  $i=1, 2, \dots, n$ , and

$\bar{y}$  is the sample mean of data in the second group  $y_j$  from  $j=1, 2, \dots, m$ .

$s$  is the pooled sample standard deviation and is estimated by assuming that each group's standard deviation is identical (eq. 5.6).

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} . \quad (5.6)$$

If the null hypothesis of equal group means,  $H_0: \mu_x = \mu_y$ , is rejected because the two-sided *p*-value  $< \alpha$ , the two-sided alternative that the group means do not differ is  $H_1: \mu_x \neq \mu_y$ . Similarly, the one-sided alternative  $H_1: \mu_x \neq \mu_y$  is employed when the mean of group X is expected to be greater than the mean of group Y prior to seeing any data. The null hypothesis is rejected in favor of the one-sided alternative when the one-tailed *p*-value is less than  $\alpha$ . These *p*-values are accurate if the data from each group follow the test's assumptions. If data are skewed or of unequal variance the *p*-values are expected to be too large and a false tendency to not find differences occurs.

### 5.3.3 Adjustment of the *t*-test for Unequal Variances

When two groups have unequal variances, the *t*-test's degrees of freedom should be adjusted using Satterthwaite's approximation (here called the Welch's *t*-test), which was developed in the 1940s. The degrees of freedom will be lowered, changing the *p*-value and penalizing the test because it is being applied to data that do not meet the *t*-test's assumptions. Statistics software correctly performs the Welch/Satterthwaite version of the test by default. Unless you have a clear reason for doing so (and we doubt that there is one), do not remove this adjustment by performing the *t*-test using the pooled standard deviation or with the option to assume equal variance. Always assume unequal variances. There is no benefit to performing the pre-1940s unadjusted test, as the adjustment goes to zero when sample variances are identical. Using the unadjusted test on data with unequal variance will likely provide an incorrect *p*-value that may be either too small or too large.

**Example 5.3. Precipitation nitrogen—The Welch's *t*-test (with Welch correction).**

The Shapiro-Wilk test of normality for each of the two groups of ammonia plus organic nitrogen from example 5.1 show that neither group follows a normal distribution at the  $\alpha=0.05$  level.

```
> shapiro.test(NH4orgN[where == "indust"])
```

```
Shapiro-Wilk normality test
data: NH4orgN[where == "indust"]
W = 0.80346, p-value = 0.01597
```

```
> shapiro.test(NH4orgN[where == "residen"])
```

```
Shapiro-Wilk normality test
data: NH4orgN[where == "residen"]
W = 0.46754, p-value = 1.517e-06
```

As this dataset is small and nowhere near the requirement for the Central Limit Theorem to hold, we should expect some loss of power, inflating the *p*-value of the *t*-test. Testing for unequal variance, both the parametric Levene's and nonparametric Fligner-Killeen tests (Aho, 2016)—discussed in section 5.6.1—find no difference in the variances of the two groups, though 10 observations is a small amount of data to work with.

```
> require(car)
> leveneTest(NH4orgN, where, center = median)
```

```
Levene's Test for Homogeneity of Variance (center = median)
```

Df	F	value	Pr(>F)
group	1	0.2242	0.6415

18

```
fligner.test(NH4orgN, where)
```

```
Fligner-Killeen test of homogeneity of variances
data: NH4orgN and where
Fligner-Killeen:med chi-squared = 0.067548,
df = 1, p-value = 0.7949
```

The *t*-test is computed with the default two-sided alternative using the *t.test* command in R. The Welch's *t*-test is used by default:

```
> t.test(NH4orgN~where, alternative = "two.sided")
```

Welch Two Sample *t*-test

```
data: NH4orgN by where
```

```
t = 0.029044, df = 11.555, p-value = 0.9773
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.081479 2.137479
sample estimates:
mean in group indust mean in group residen
1.666      1.638
```

From the  $p$ -value of 0.977, the null hypothesis of no difference cannot be rejected. There is essentially no evidence that the means differ using the  $t$ -test. However, because the test assumptions were violated the  $t$ -test may not be able to find differences that are there. To obtain a definitive answer to whether the group means differ, perform a permutation test instead.

### 5.3.4 The $t$ -test After Transformation Using Logarithms

A  $t$ -test on logarithms of data has been a popular approach to use when data are skewed. Water resources data more often appear closer to the shape of a skewed lognormal distribution than to a normal distribution. In log units, skewed data often appear close to a normal distribution with equal group variance. Not all who use it realize that by transforming with logs, the test determines whether the geometric means, and not arithmetic means, of the two groups differ. When the logarithms of data follow a normal distribution, the geometric mean estimates the sample median of the data. If a different transformation produces a distribution similar in shape to the normal, the  $t$ -test on transformed units can be considered a test for difference in group medians. Results of the  $t$ -test on data transformed to symmetry are often similar to those of the rank-sum test, as both are tests for differences in medians. However, the rank-sum test does not require the analyst to spend time determining what an appropriate transformation to symmetry might be.

#### Example 5.4. Precipitation nitrogen— $t$ -test on logarithms.

A  $t$ -test on the natural logarithms of the nitrogen data concludes that the difference in group means is not significant at  $\alpha=0.05$ .

```
> t.test(log(NH4orgN) ~ where)
```

Welch Two Sample t-test

```
data: log(NH4orgN) by where
t = 1.3578, df = 14.684, p-value = 0.195
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.275160 1.236132

sample estimates:
mean in group indust mean in group residen
0.3518488 -0.1286373
```

### 5.3.5 Conclusions as Illustrated by the Precipitation Nitrogen Example

Several tests were run on the precipitation nitrogen data from example 5.1, with varying outcomes. This is because not all of the tests have the same objectives, and not all of the tests have the same requirements.

1. The most important decision to make, before running a statistical test, is to determine what the correct statistic is for the question being asked. Does one group have higher values than the other? This is a frequency question and is best answered by a test on frequency measures (percentiles) such as medians. Concentrations in the industrial group are more often higher than those in the residential group. This is what was tested by the rank-sum test and seen in the boxplots of figure 5.2. Running *t*-tests on logarithms may approximately test the same hypothesis, but there is no advantage to using them versus the actual rank-sum test.
2. When the interest is in means, because the objective is to test the cumulative amounts in each group (mass, volume, cumulative exposure), use a permutation test instead of the *t*-test. The lack of power encountered when a *t*-test is applied to non-normal and unequal variance data is overcome by permutation methods. Skewness and outliers inflate the sample standard deviation used in the *t*-test and it often fails to detect the differences present that could be seen with a permutation test.
3. Both permutation tests and *t*-tests determine whether the total amounts (standardized by sample size  $n$ ) are the same or different. For the precipitation nitrogen data, the total in each group is about the same owing to the one large value in the residential group. Most of the nitrogen present came in that one precipitation event, such data often deserve closer scrutiny and may provide information about the processes occurring. Don't throw away outliers in order to meet the requirements of a substandard test. Use a better test and learn from the entire dataset.
4. Decide which type of test to use based on the study objectives rather than on the shape of the data distribution. For questions of whether one group has higher values than the other, compute the rank-sum test. For concerns about totals or mass, use a permutation test to judge differences in group means while protecting against the *t*-test's potential loss of power due to non-normal and unequal variance data.
5. 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. Helsel (2012) provides several better methods for examining censored data. If the question is whether one group shows higher values than another, all data below the highest reporting limit can be assigned a tied rank and the rank-sum test computed, without making any distributional assumptions or assigning arbitrary values to the data (see section 5.6).

## 5.4 Estimating the Magnitude of Differences Between Two Groups

After completion of a hypothesis test comparing the central tendency of two groups of data, the logical next step is to determine by how much the two groups differ. This can then be compared to the effect size, the amount the investigator believes is scientifically important. It should always be remembered that statistical significance is not the same as practical significance. Having a significant test result with an observed difference smaller than what is important may indicate the observed difference is actually unimportant. Of course it also may indicate a valuable early warning that the difference is trending towards a level of importance. And the level of importance for another purpose may be smaller than for the current study, making the small difference important for that other purpose. Therefore reporting the observed difference is always a good idea.

### 5.4.1 The Hodges-Lehmann Estimator of Difference in Medians

One nonparametric estimate of the difference between two independent groups is a Hodges-Lehmann estimator,  $\hat{\Delta}$  (Hodges and Lehmann, 1963; Hollander and Wolfe, 1999). This estimator is the median of all possible pairwise differences between the  $x$  values and  $y$  values:

$$\hat{\Delta} = \text{median}[x_i - y_j] \text{ for } x_i, i = 1, 2, \dots, n \text{ and } y_j, j = 1, 2, \dots, m . \quad (5.7)$$

There will be  $n \cdot m$  pairwise differences. 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.

The estimator 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 the hand computation example below,  $(x_{med} - y_{med}) = 10.5$ . 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 has a larger confidence interval than that of the Hodges-Lehmann estimate.

#### Example 5.5. Precipitation nitrogen—Hodges-Lehmann estimator.

The Hodges-Lehmann estimate of the median difference in ammonia+organic nitrogen between the industrial and residential groups from example 5.1 is computed by specifying `conf.int = TRUE` when performing the `wilcox.test` command. The Hodges-Lehmann estimate of median difference between the two groups is 0.504.

```
> wilcox.test(NH4orgN ~ where, conf.int = TRUE)
```

```
Wilcoxon rank sum test with continuity correction
data: NH4orgN by where
W = 76.5, p-value = 0.04911
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 3.948029e-05 1.099984e+00
sample estimates:
difference in location
0.5040993
```

**Example 5.6. Hand computation of the Hedges-Lehmann estimator.**

Suppose we had a sample of 3 values for the  $x$  group and they were values of 15, 17, and 25, and we had a sample of 4 values for the  $y$  group and they were the values 8, 27, 3, and 5. We can compute the Hedges-Lehmann estimate by hand by enumerating all possible values of the differences as shown here.

$x_i$	$y_j$	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  $3 \cdot 4 = 12$  pairwise differences are

$$-12, -10, -2, 7, 9, 10, 12, 14, 17, 20, 22.$$

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 to outliers.

**5.4.2 Confidence Interval for the Hedges-Lehmann Estimator,  $\hat{\Delta}$** 

A nonparametric interval estimate for  $\hat{\Delta}$  illustrates how variable the median difference between groups might be. No distribution is assumed for this interval; it is computed using the process for the binomial confidence interval on the median described in chapter 3 by finding appropriate rank positions from among the ordered  $n \cdot m$  pairwise differences that represent the ends of the confidence interval.

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

$$R_l = \frac{N - z_{\alpha/2} \cdot \sqrt{\frac{N(n+m+1)}{3}}}{2} \quad (5.8)$$

$$R_u = N - R_l + 1 . \quad (5.9)$$

When the exact test is used for smaller sample sizes, the quantiles for the rank-sum test statistics having a  $p$ -value nearest to  $\alpha/2$  and  $1-(\alpha/2)$  are used to find the lower and upper ends of the confidence limit for  $\hat{\Delta}$ . The lower limit uses the lower  $\alpha/2$  quantile. The upper limit uses the upper  $\alpha/2$  quantile plus 1.

The confidence interval around  $\hat{\Delta}$ , regardless of sample size, is computed by the same option `conf.int = TRUE` to the `wilcox.test` command that computed the estimate itself. R will compute the appropriate version regardless of whether the exact test (smaller sample sizes) or large-sample approximation (larger sample sizes) was used.

**Example 5.7. Precipitation nitrogen—Confidence interval for the Hedges-Lehmann estimator.**

The confidence interval on the median difference in nitrogen between the residential and industrial groups was determined in example 5.5 by the `wilcox.test` command with the option `conf.int = TRUE`.

95-percent confidence interval:

3.948029e-05 1.099984e+00

or 0.00004 to 1.1 mg/L.

**Example 5.8. Hand computation of the confidence interval for the Hodges-Lehmann estimator.**

The  $N=12$  possible pairwise differences between  $x$  and  $y$  are

$-12, -10, -2, 7, 9, 10, 12, 12, 14, 17, 20, 22$ .

To determine an  $\alpha \approx 0.10$  confidence interval for  $\hat{\Delta}$ , the quantiles for the rank-sum statistic at  $\alpha/2=0.05$  and  $1-(\alpha/2)=0.95$  can be provided by the `qwilcox` command with  $n=3$  and  $m=4$ .

```
> qwilcox(c(0.05,0.95),3,4)
[1] 1 11
```

The ranks 1 and  $(11+1)=12$  are the ranks of the ordered differences representing the endpoints of the confidence limit. With such a small dataset, the  $\alpha=0.10$  confidence limit for  $\hat{\Delta}$  is the entire range of the differences, or  $-12 \leq \hat{\Delta} \leq 22$ . It is easier to use the `wilcox.test` command:

```
> x.ex2=c(15, 17, 25)
> y.ex2=c(8, 27, 3, 5)
> wilcox.test(x.ex2, y.ex2, exact = TRUE, conf.int = TRUE)
```

Wilcoxon rank sum test

```
data: x.ex2 and y.ex2
W = 9, p-value = 0.4
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
-12 22
sample estimates:
difference in location
11
```

### 5.4.3 Estimate of Difference Between Group Means

Where group means are of interest, the difference between the means of the two groups  $\bar{x} - \bar{y}$  is the most efficient estimator of the mean difference between groups. This value is output by the `t.test` command in R. For the precipitation nitrogen data from example 5.1, the output in the *t-test* example in section 5.3.3. shows that an estimated difference in group means equals  $1.666 - 1.638 = 0.028$ , which was not significantly different from zero.

Perhaps it is obvious that when  $x$  and  $y$  are transformed prior to performing the *t*-test the difference in means in the transformed units does not estimate the difference between group means on their original scale. Less obvious is that the retransformation of the difference back to the original scale also does not estimate the difference between group means, but is closer to a function of group medians. For the log transformation, the difference in group means in log units when retransformed would equal the ratio of the geometric means of the two groups. How close any retransformation 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 Parametric Confidence Interval for Difference in Group Means

A  $t$ -confidence interval around the mean difference between groups  $\bar{x} - \bar{y}$  is output by the `t.test` command. It is appropriate in situations where the  $t$ -test may be used—when both data groups closely follow a normal distribution. For the most common situation, where the standard deviations of the two groups are dissimilar and should not be pooled, the confidence interval is

$$CI = \bar{x} - \bar{y} \pm t_{\alpha/2,(df)} \cdot \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} \quad (5.10)$$

where  $df$  is the degrees of freedom used in the Welch's  $t$ -test. For the uncommon situation, where the variances of the two groups can be assumed to be similar and the pooled standard deviation  $s$  is used in the test, the confidence interval simplifies to

$$CI = \bar{x} - \bar{y} \pm t_{\alpha/2,(n+m-2)} \cdot s \sqrt{\frac{1}{n} + \frac{1}{m}} \quad (5.11)$$

For the precipitation nitrogen data in example 5.1, the output in section 5.3.3. shows that the 95 percent  $t$ -confidence interval on the estimated difference in group means of 0.028 spans from -2.081 to 2.137.

#### 5.4.5 Bootstrap Confidence Interval for Difference in Group Means

Regardless of the distribution of data in either group, the computer-intensive method of bootstrapping can compute a confidence interval around the difference in group means. This should be the preferred method of obtaining an interval when a permutation test is used to test for a difference in group means or when data do not appear to come from a normal distribution. Bootstrap intervals also work well for data that follow a specific distribution. For example, bootstrap intervals will be quite similar to  $t$ -intervals when data follow a normal distribution. Bootstrap confidence intervals were described in chapter 3, where the percentile bootstrap method was introduced.

To compute the bootstrap interval, observations are repeatedly and randomly resampled from the original data with replacement for each group. Each resample contains only values found in the group's original data, but not necessarily in the same proportions—observations may be randomly selected in different frequencies than they originally occurred. The process is repeated thousands of times, each time resulting in an estimate of the difference in group means. An  $\alpha=95$ -percent bootstrap confidence interval is found by going to the 2.5 and 97.5 percentiles of the thousands of resampled differences. This is called the percentile bootstrap method (Efron and Tibshirani, 1994). A bootstrap confidence interval on the difference between group means is computed using the `perm2` script. For the precipitation nitrogen data from example 5.1, the `perm2` script output in section 5.2.2., repeated below, gave a 95-percent confidence interval on the mean difference from -2.041 to 1.443.

```
Permutation Test of Difference Between 2 Group Means
```

```
Data: NH4orgN by where
```

```
Number of Possible Permutations is greater than 1000
```

```
R = 10000 pvalue = 0.9954
```

```
Alt Hyp: true difference in means is not equal to 0
```

```
sample estimates:
```

```
mean of indust = 1.666 mean of residen = 1.638
```

```
Diff of means = 0.028
```

95 percent confidence interval  
 -2.041 1.443

### 5.4.6 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 the results of hypothesis tests and visually allowing a judgment of whether data fit the assumptions of the test being employed. This is illustrated using the precipitation nitrogen data.

### 5.4.7 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. Boxplots of the precipitation nitrogen data from example 5.1 were presented in figure 5.2. Note the difference in medians is clearly displayed, as well as the similarity in spread (IQR). The rejection of normality by Shapiro-Wilk tests is seen in the presence of skewness (industrial group) and the one large outlier (residential group). Side-by-side boxplots are an effective and concise method for illustrating the basic characteristics of data groups and of differences between those groups.

### 5.4.8 Q-Q Plots

Another method for illustration of rank-sum results is the quantile-quantile (Q-Q) plot described in chapter 2, where 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 values from each sample can be sorted separately from 1 to  $n$ , and the Q-Q plot is simply a scatterplot of the ordered data pairs  $(x_1, y_1), \dots, (x_n, y_n)$  where  $x$  and  $y$  designate the two samples. When sample sizes are not equal ( $n < m$ ), the quantiles from the smaller dataset are used as is, and the  $m$  corresponding quantiles for the larger dataset are interpolated.

It is always helpful in a Q-Q plot to graph the  $y=x$  line, the line with identical values for  $x$  and  $y$  and therefore having a slope of 1. A Q-Q plot of the precipitation nitrogen data is shown in figure 5.5, where 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 relation. 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 relation between two sets of quantiles on a Q-Q plot can be estimated using the median of all possible ratios  $(y_i/x_j)$ ,  $i=1, 2, \dots, n$  and  $j=1, 2, \dots, m$ . This is a type of Hodges-Lehmann estimator, as discussed in the previous section. The median ratio equals 0.58, and the line  $y=0.58 \cdot x$  (or residential =  $0.58 \cdot$  industrial) is shown in figure 5.5. Note the resistance of the median ratio to the one large outlier.

Second, the data are crowded together at low concentrations but spread further apart at higher concentrations—a pattern indicating right-skewness. To remedy both skewness and nonadditivity a power transformation was chosen, the natural logarithmic transform. A Q-Q plot of data logarithms is shown in figure 5.6. Note that the data are now more constant in variance 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 relation on the original scale has become an additive relation in logarithmic units, with the Hodges-Lehmann estimate of the difference between the natural logarithm of  $x$  and the natural logarithm of  $y$ ,  $\hat{D}$ , equal to  $-0.5447$ . Note that  $\hat{D}$  is the natural logarithm of the Hodges-Lehmann estimate of the ratios on the original scale,  $\ln(0.58)=-0.5447$ . The dashed line plotted in figure 5.6 is parallel to  $y=x$ , with an offset in intercept of  $-0.545$ . A  $t$ -test would now be appropriate for the logarithms, assuming each group's

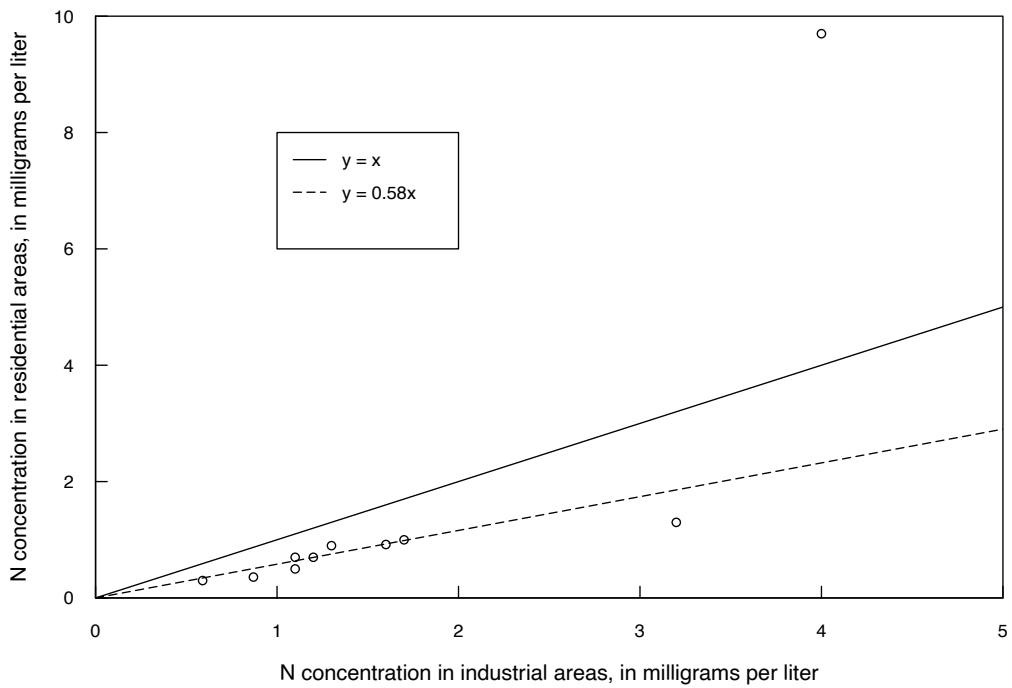


Figure 5.5. Q-Q plot of the precipitation nitrogen data from example 5.1.

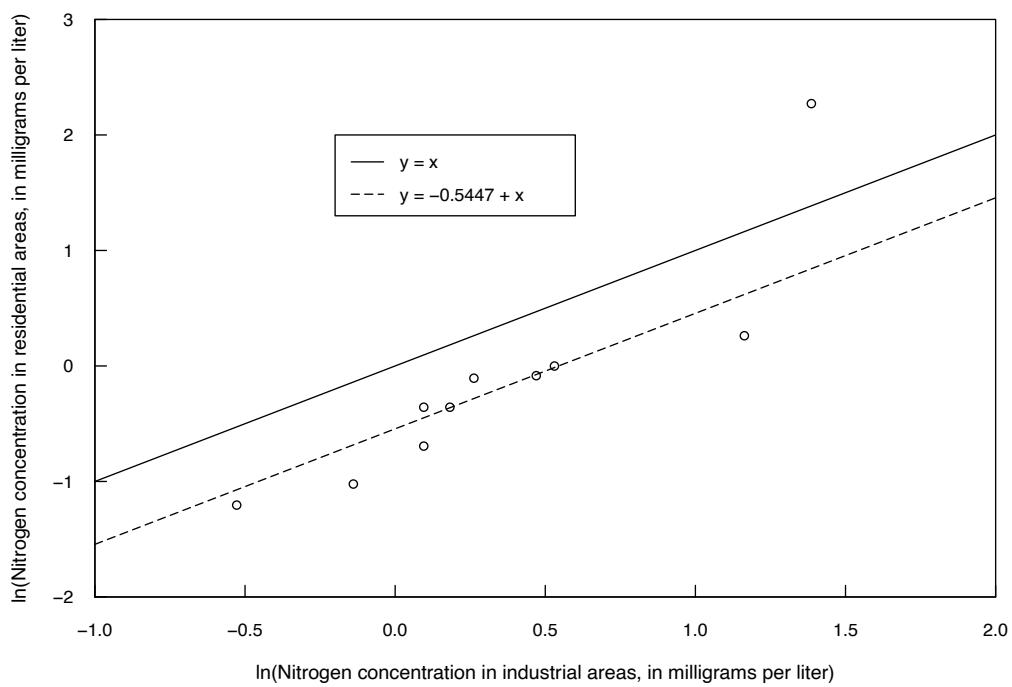


Figure 5.6. Q-Q plot of the logs of the precipitation nitrogen data from example 5.1.

transformed data were approximately normal, if estimating the difference in geometric means (medians) was desired.

In summary, Q-Q plots of the quantiles of two data groups illustrate the level of adherence to the assumptions of hypothesis tests (*t*-test or rank-sum), providing additional insight on which test procedures might be reasonable to employ. 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 relation between the two groups can easily be discerned.

## 5.5 Two-group Tests for Data with Nondetects

The two-sample *t*-test, after substituting one-half the reporting limit for nondetects, has unfortunately been a common procedure in water resources for testing differences between groups. This procedure may not find differences between groups that are there or may find differences that are not there (Helsel, 2012). There are better methods. The rank-sum test on data with one reporting limit provides much more power to detect differences than would substitution followed by a *t*-test. When there are multiple reporting limits, the data must first be recensored to show all data below the highest reporting limit as tied, both detected and nondetected observations, turning the data into a one-reporting limit format. The rank-sum test is then performed on the recensored values. Even with recensoring, this is a better procedure than using the *t*-test after substitution, as it avoids the consequences of falsely stating that the exact value for each nondetect is known.

As an example, a rank-sum test is computed on trichloroethylene (TCE) concentrations in groundwater measured by Eckhardt and others (1989) given in the dataset TCE2a. The original data had five reporting limits, at 1, 2, 3, 4, and 5 micrograms per liter ( $\mu\text{g/L}$ ). Concentrations have been recensored so that all values below 5 are designated as <5. In TCE2a the column TCECONC contains both concentrations at and above 5  $\mu\text{g/L}$ , and the reporting limit value of 5 for any data below 5. The indicator column LT5 is a 0/1 variable, where a 1 indicates a censored <5 and a 0 indicates a detected concentration. In the column HALF.DL one-half the reporting limit has been substituted for all nondetects.

The *t*-test on HALF.DL does not find a significant difference between the groups ( $p=0.91$ ). Primarily this is a result of the non-normality of any dataset with many values (here about 80 percent) so close to zero. However, the test also presumes that the analyst believes that 80 percent of the values are all at the same concentration of 2.5, and so the estimate of standard deviation used by the test is definitely not realistic.

```
> attach(TCE2a)
> t.test(HALF.DL ~ Density)

Welch Two Sample t-test

data: HALF.DL by Density
t = -0.11205, df = 200.79, p-value = 0.9109
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-8.185681 7.305413
sample estimates:
mean in group High mean in group Medium
```

8.752174      9.192308

The rank-sum test on the recensored data assigns all values below 5 as the same low rank, tied with one another and below the lowest detected value at or above 5 µg/L. This is an accurate representation of what is known about the data. The test indicates a significant difference in the distributions of the two groups; the medians of both groups are <5, but the upper ends differ. There are about 10 percent of values detected at or above 5 µg/L in the medium density group, whereas 20 percent of the high density group is at that level. This difference in the upper ends of the distribution is seen as significant by the rank-sum test.

```
> wilcox.test(HALF.DL ~ Density)
```

```
Wilcoxon rank sum test with continuity correction
```

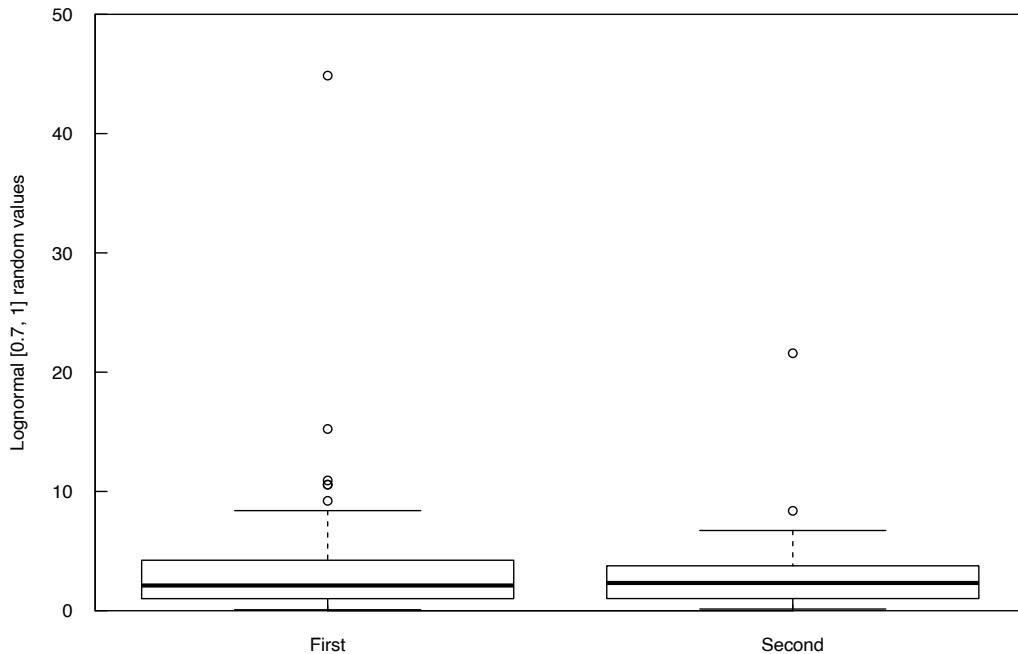
```
data: HALF.DL by Density
W = 6599.5, p-value = 0.02713
alternative hypothesis: true location shift is not equal to 0
```

The difference in these two test results is convincing evidence that a nonparametric test using recensored data is more powerful when compared to a *t*-test with half the reporting limit substituted for nondetects. Although better and more powerful tests are available from the field of survival analysis (Helsel, 2012) for censored data, when a simple test for group differences is needed the rank-sum test for data with one reporting limit or for data recensored at the highest reporting limit is a much better choice than the *t*-test.

## 5.6 Tests for Differences in Variance Between Groups

Differences in central location (mean, median) are not the only type of difference between groups that is of interest. Often important is whether the variability of data is the same in each group. Differences in variance are a violation of the assumption of the uncorrected *t*-test, so some analysts incorrectly test the variance of groups using Bartlett's test for unequal variance (sometimes called the *F*-test) prior to conducting a *t*-test, and if the variance is found to be significantly different, the Welch version of the *t*-test is used. If the variance is not significantly different, they use the uncorrected version. As stated in section 5.3, this is flawed reasoning as the Welch's correction can always be used to perform the *t*-test. As the difference in variance goes to zero, Welch's correction is minimized and the corrected test negligibly differs from the uncorrected version. Unfortunately, Bartlett's test is one of the most sensitive tests to the assumption of a normal distribution. Conover and Iman (1981) note that in several studies comparing tests for heteroscedasticity (unequal variance), Bartlett's test "is well known to be nonrobust and that none of the comparative studies recommends [it] except when the populations are known to be normal." Bartlett's test will too often lead to the conclusion that variances differ when in fact they do not. There are better tests for heteroscedasticity (changing variance) than Bartlett's test; these tests are discussed in the next two sections.

To demonstrate the deficiency of Bartlett's test, the short script `mcbart` below will compute thousands of repetitions of using Bartlett's test on data generated from a single lognormal distribution. It is a simple example of scripting in R that may be a guide for that topic, but we focus on the results here. The distribution from which data are generated have a mean logarithm of 0.7 and a standard deviation of



**Figure 5.7.** Box plots of two groups of 50 samples each of randomly generated data from a single lognormal distribution (and so have the same variance). Bartlett's test declares the variances different. The non-normality of the data is a violation of the test's (strict) assumption of a normal distribution.

logarithms of 1. The data generated (sample size of 50 observations in each group) have characteristics very similar to water quality and other water resources data, as shown in figure 5.7.

The Bartlett's test results for the data generated for figure 5.7 are

```
> set.seed(1832)
> val.expl <- rlnorm(100, 0.7, 1.0)
> group <- c(rep("First", 50), rep("Second", 50))
> b.result <- bartlett.test(val.expl ~ group)
> b.result
```

Bartlett test of homogeneity of variances

```
data: val.expl by group
Bartlett's K-squared = 23.16, df = 1, p-value = 1.491e-06
```

The rejection of the null hypothesis of equal variance is incorrect because the data for both groups were randomly generated from the same lognormal distribution. We would expect a test to incorrectly produce this result with a 5 percent probability if conditions for the test are followed by the data. To show the sensitivity of Bartlett's test to non-normality, the Monte Carlo script `mcbart` generates 2,000 sets of data from the lognormal (0.7, 1) distribution and runs the test on each set. The result of TRUE means that the test's  $p$ -value was below 0.05 and equality of variance is rejected. This is expected in 5 percent (or 100) of the 2,000 sets. We see that Bartlett's test rejects the null hypothesis in 1,114 out of 2,000 sets, or 56 percent of the cases generated! Bartlett's test rejects equal variance incorrectly far more than it should when data do not follow a normal distribution.

```
> mcbart # a Monte Carlo evaluation of Bartlett's test
function(nrep = 2000, lnmean = 2, lnsd = 1.5) {
  group <- c(rep("First", 50), rep("Second", 50))
  num.reject <- rep(1, times=nrep)      # initializes num.reject
  for (i in 1:nrep) {
    val.expl <- rlnorm(100, lnmean, lnsd)
    b.result <- bartlett.test(val.expl ~ group)
    num.reject[i] <- b.result$p.value < 0.05
  }
  ftable(num.reject) }
> # run the mcbart function, with mean of logs = 0.7
> # standard deviation of logs = 1.0
> set.seed(1832)
> mcbart(lnmean = 0.7, lnsd = 1.0)
num.reject     0      1
```

886 1114

Tests for equal variance are often used to determine important characteristics of the data, such as whether the precision (usually defined as the inverse of standard deviation) of groups is changing. The tests of the next two sections are better able to do this than Bartlett's test and can test for changing variance among two or more groups; they should therefore be used instead of Bartlett's test, and are appropriate for more than two groups as well as for a two-group test.

### 5.6.1 Fligner-Killeen Test for Equal Variance (Nonparametric)

Out of the 56 tests evaluated by Conover and Iman (1981), the Fligner-Killeen test was found to be the most robust for unequal variance when data are non-normally distributed. It begins by computing the absolute value of the residuals (AVR) from each group median. For  $j=1$  to  $k$  groups and  $i=1$  to  $n_j$  observations

$$AVR_{ij} = |x_{ij} - \text{median}_{j\cdot}| . \quad (5.12)$$

The test then ranks the AVR and weights each rank to produce a set of scores. A linear-rank test (a nonparametric test of location) is computed on the scores. The null hypothesis is that the average score is the same in all groups, indicating that the variances are the same in all groups. The alternative hypothesis is that at least one group's variance differs.

The Fligner-Killeen test correctly does not find a difference in variance between the two lognormal groups generated from the same lognormal (0.7,1) distribution.

```
> fligner.test(val.expl, as.factor(group))

Fligner-Killeen test of homogeneity of variances

data: val.expl and as.factor(group)
Fligner-Killeen:med chi-squared = 0.27908, df = 1, p-value = 0.5973
```

## 5.6.2 Levene's Test for Equal Variance (Parametric)

Levene's test (Conover and others, 1981) determines whether the average distance from the median is the same in all groups. Though it assumes that data follow a normal distribution, it is much less sensitive to that assumption than is Bartlett's test. It behaves as most parametric tests do, losing power (an increase in the *p*-value) when applied to data that are not shaped like a normal distribution. This is more acceptable to statisticians than the too-frequent rejections made by Bartlett's test.

Levene's test also computes the AVR for each observation. It then performs an analysis of variance (ANOVA, see chap. 7) on the AVRs. ANOVA computed for only two groups is very similar to a *t*-test. The null hypothesis is that the average absolute residual is the same in all groups because the variance is the same in all groups. The alternative hypothesis is that at least one group's variance differs. Levene's test is commonly found in statistics software and is recommended for use in many guidance documents, including U.S. Environmental Protection Agency (2009). Conover and others (1981) found that Levene's test performed better than other parametric tests of heteroscedasticity evaluated in their study.

Levene's test is found in the *car* package of R (Fox and Weisberg, 2011). Unlike Bartlett's test, it (appropriately) does not find a difference in variance between the two lognormal (0.7, 1) groups generated in the example above (*p*=0.238).

```
> leveneTest(val.expl, as.factor(group))

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)
group 1 1.4094 0.238
```

## Exercises

1. We wish to test for a change in concentration between sites of differing land uses. The rank-sum test will be used. Should the test have a one-sided or two-sided alternative?
2. A shallow aquifer is contaminated by molybdenum leachate from mine tailings. A remediation effort was begun to reduce the molybdenum concentrations in waters leaving the site. Post-remediation concentrations (in  $\mu\text{g/L}$ ) from 13 wells downgradient of the remediation process are listed below. Also shown are concentrations at 3 wells upgradient from, and unaffected by, the remediation process. Test whether the wells downgradient of remediation are significantly lower in molybdenum than are the upgradient wells. Also test whether the mean concentration has changed with remediation.

downgradient	upgradient
0.850	6.900
0.390	3.200
0.320	1.700
0.300	
0.300	
0.205	
0.200	
0.200	
0.140	
0.140	
0.090	
0.046	
0.035	

3. Annual streamflows for the Green River at Munfordville, Kentucky, were listed in exercise 4.1. Beginning in 1969 the stream was regulated by a reservoir.
  - A. Construct a Q-Q plot that compares the distributions of the two groups. Indicate whether the flows exhibit an additive or multiplicative relation, or neither.
  - B. Does there appear to be an additive or multiplicative change in the magnitude of annual flow itself? If so, explain why this might occur.
  - C. Test whether flows after the reservoir began operations differ from flows beforehand. (Any differences may be attributable to the effect of the reservoir, to climatic differences between the periods, and likely to both. This simple exercise will not allow you to separate the two possible causes).
  - D. Unit well yields, in gallons per minute per foot of water-bearing material, were contrasted for wells within valleys containing fractured rock versus valleys with no fracturing (Wright, 1985). Perform the appropriate  $\alpha=0.05$  test to discern whether fracturing is associated with higher mean unit well yield.

Well yields in fractured rock	Well yields in unfractured rock
0.95	1.02
0.72	0.49
0.51	0.454
0.44	0.10
0.40	0.077
0.30	0.041
0.18	0.040
0.16	0.030
0.16	0.020
0.13	0.007
0.086	0.003
0.031	0.001
0.020	-

Well yields in fractured rock have a probability plot correlation coefficient (PPCC) of 0.943 ( $p > 0.05$ ). Well yields in unfractured rock have a PPCC of 0.806 ( $p < 0.05$ ).

- 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  $H_A: \mu_x > \mu_y$  using the rank-sum test. By how much does the test statistic change? Are the results altered by the presence of a detection limit? Could a  $t$ -test be used in this situation?



# Chapter 6

## Paired Difference Tests of the Center

---

*To determine the effectiveness of an acid solution in developing wells in carbonate rock, yields of 20 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 24 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 portion of year-to-year variation in load due to differences in precipitation be compensated for in determining 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, then we should be able to do our analysis using data from both laboratories. The differences between labs must be discerned beyond the sample-to-sample differences.*

Each of the example situations mentioned above is addressed by using the matched-pair tests of this chapter. As opposed to the tests of chapter 5, we now consider data having a logical pairing of observations within each group. There may be a great deal of variability from one pair to another, as with the year-to-year pairs of sediment data in the second example above. Both basins may exhibit low yields in dry years and higher yields in wet years. This variability among pairs of observations is noise that would obscure the differences between the two groups being compared if the methods of chapter 5 were used. Instead, blocking is used to eliminate the influence of this noise by basing the analysis on the pairwise differences between the groups. Tests are then conducted on the set of differences to determine whether the two groups differ significantly (table 6.1). Two nonparametric tests, the sign test and the signed-rank test, determine whether one group's paired observation is generally higher than the other group's paired observation. Also presented is the paired *t*-test, the parametric test of whether the mean difference between the groups equals zero. The *t*-test is used when the mean is of interest and requires that the differences between paired observations be normally distributed. A permutation test for determining whether the mean difference equals zero is also presented as a more powerful and flexible alternative to the paired *t*-test when differences do not follow a normal distribution. After surveying graphical methods to illustrate the test results, estimators for the difference between the two groups are discussed.

**Table 6.1.** Paired difference tests of this chapter and their characteristics.

[For the sign test and the signed-rank test, the data from one group are frequently higher than the data from the other group. For the paired *t*-test and the permutation test on mean difference, one group has a higher mean.  $H_A$  is the alternative hypothesis, the signal to be found if it is present]

Characteristic	Sign test	Signed-rank test	Paired <i>t</i> -test	Permutation test on mean difference
Class of test	Nonparametric	Nonparametric	Parametric	Permutation
Distributional assumption for differences	None	Symmetry	Normal distribution	Symmetry
Estimator of difference	Median difference	Hodges-Lehmann estimate	Mean difference	Mean difference

For paired observations  $(x_i, y_i)$ , their differences

$$D_i = x_i - y_i , \quad (6.1)$$

where

$$i = 1, 2, \dots, n$$

are computed. The tests in this chapter determine whether  $x_i$  and  $y_i$  are from the same population—the null hypothesis—by analyzing  $D_i$ . If the median or mean paired difference,  $D_p$ , significantly differs from zero, the null hypothesis is rejected.

As with the tests of chapter 5, the most important determinant of which test to use is the study objective. If the study is trying to determine if the conditions represented by the two groups are the same, this is a frequency question and best answered by a nonparametric test. If groups are similar (the null hypothesis,  $H_0$ ), then the observation from one group in the pair will be higher than the paired observation in the other group approximately half the time. If groups are not the same, the observation from one group will be higher than the paired observation from the other group at a frequency greater than 50 percent, and the nonparametric test will pick up on this difference. We discuss two nonparametric tests: the sign and signed-rank tests. The sign test examines whether within an  $(x, y)$  pair, does  $x$  tend to be higher (or lower, or different) than  $y$ ? The sign test is very useful 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  $3$ . It is quite useful with censored data (exercises 4 and 5 at the end of this chapter). The sign test is often less powerful than the signed-rank test because the sign test uses only the algebraic sign (+ or -) of the difference, ignoring the magnitude. It treats a large difference as no different than a small difference. The signed-rank test is generally more powerful than the sign test because it uses the magnitudes of differences—a larger difference has more weight than a smaller difference. The signed-rank test's null hypothesis is that the frequency of  $x > y$  for pairs is 50 percent, and so the median of  $x$  equals the median of  $y$ . The alternative (two-sided) hypothesis is that the frequency of  $x > y$  is not 50 percent, and therefore the medians of  $x$  and  $y$  differ.

When the  $D_i$  values follow a normal distribution, a paired  $t$ -test can evaluate a different null hypothesis: The mean of the differences  $(x_i - y_i) = 0$ , and therefore the mean of  $x_i$  differs from the mean of  $y_i$ . This is also equivalent to testing that the sums of the  $x_i$  and the  $y_i$  are the same, as both groups of paired data have the same sample size. Permutation tests (see section 4.1.2.) on the mean difference are often more powerful alternatives to the paired  $t$ -test, as permutation tests are not impaired when the shape of the  $D_i$  distribution fails to follow a normal distribution.

Nonparametric tests determine whether differences in frequencies occur, such as the frequency of  $x_i > y_i$ . The paired  $t$ -test determines whether measures of mass (means) of two groups are the same or not. Tests on means and tests on frequencies address two different objectives. Consider the question of whether one laboratory method tends to report higher concentrations than the second method for a given bottle of water. For a series of submitted water samples, one method reported moderate concentrations, while the second reported mostly lower concentrations plus a few high values. The mean concentration for each method could be about the same, and the  $t$ -test and permutation test on means would find no difference between the two methods. A nonparametric test on the frequency of which method has higher concentrations would likely find a difference—for most pairs of measurements the method reporting moderate concentrations was higher than the method with lower concentrations plus a few outliers.

## 6.1 The Sign Test

For data pairs  $(x_i, y_i)$ ,  $i=1, 2, \dots, n$ , the sign test determines whether  $x$  is frequently larger (or smaller, or different) than  $y$ , without regard to whether that difference is additive or to the distributional shape of the differences.

### 6.1.1 Null and Alternative Hypotheses

The null and alternative hypotheses may be stated as

$$H_0: \text{Prob}[x > y] = 0.5,$$

versus one of the three possible alternative hypotheses:

$H_{A_1}$ :  $\text{Prob}[x > y] \neq 0.5$  (Two-sided test— $x$  might be larger or smaller than  $y$ ). Reject  $H_0$  when the two-sided  $p$ -value  $< \alpha$ .

$H_{A_2}$ :  $\text{Prob}[x > y] > 0.5$  (One-sided test— $x$  is expected to be larger than  $y$ ). Reject  $H_0$  when the one-sided  $p$ -value  $< \alpha$ .

$H_{A_3}$ :  $\text{Prob}[x > y] < 0.5$  (One-sided test— $x$  is expected to be smaller than  $y$ ). Reject  $H_0$  when the one-sided  $p$ -value  $< \alpha$ .

### 6.1.2 Computation of the Exact Sign 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 alternative hypotheses is true instead, more than half of the differences will tend to be either positive or negative. The significance level,  $\alpha$ , reported by software is the probability of obtaining the observed test statistic, or a result more extreme, when the null hypothesis is true (chap. 4).

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. R defaults to using exact tests for small sample sizes. Unfortunately, commercial software generally performs large sample approximations regardless of sample size.

**Computation:** Compute  $D_i = x_i - y_i$ . Ignore all tied data pairs (all  $D_i = 0$ ). Reduce the sample size of the test to the number of 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 pluses, the number of times  $x_i > y_i$ ,  $i = 1, 2, \dots, n$ .

**Decision rule:** To reject  $H_0$ :  $\text{Prob}[x > y] = 0.5$ , either

1.  $H_{A_1}$ :  $\text{Prob}[x > y] \neq 0.5$  (the  $x$  measurement tends to be either larger or smaller than the  $y$  measurement). Reject  $H_0$  when the two-sided  $p$ -value associated with  $S^+ < \alpha$ .
2.  $H_{A_2}$ :  $\text{Prob}[x > y] > 0.5$  (the  $x$  measurement tends to be larger than the  $y$  measurement). Reject  $H_0$  when the one-sided  $p$ -value associated with  $S^+ < \alpha$ .
3.  $H_{A_3}$ :  $\text{Prob}[x > y] < 0.5$  (the  $x$  measurement tends to be smaller than the  $y$  measurement). Reject  $H_0$  when the one-sided  $p$ -value associated with  $S^+ < \alpha$ .

#### Example 6.1. Mayfly nymphs—Exact sign test, small samples.

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  $\alpha$  of 1 percent is set as acceptable.

```
> Above <- c(12, 15, 11, 41, 106, 63, 296, 53, 20, 110, 429, 185)
> Below <- c(9, 8, 38, 24, 48, 17, 11, 41, 14, 60, 53, 124)
> signdiff <- sign(Above-Below)
> nymph.list <- data.frame(Above, Below, signdiff)
> nymph.list
```

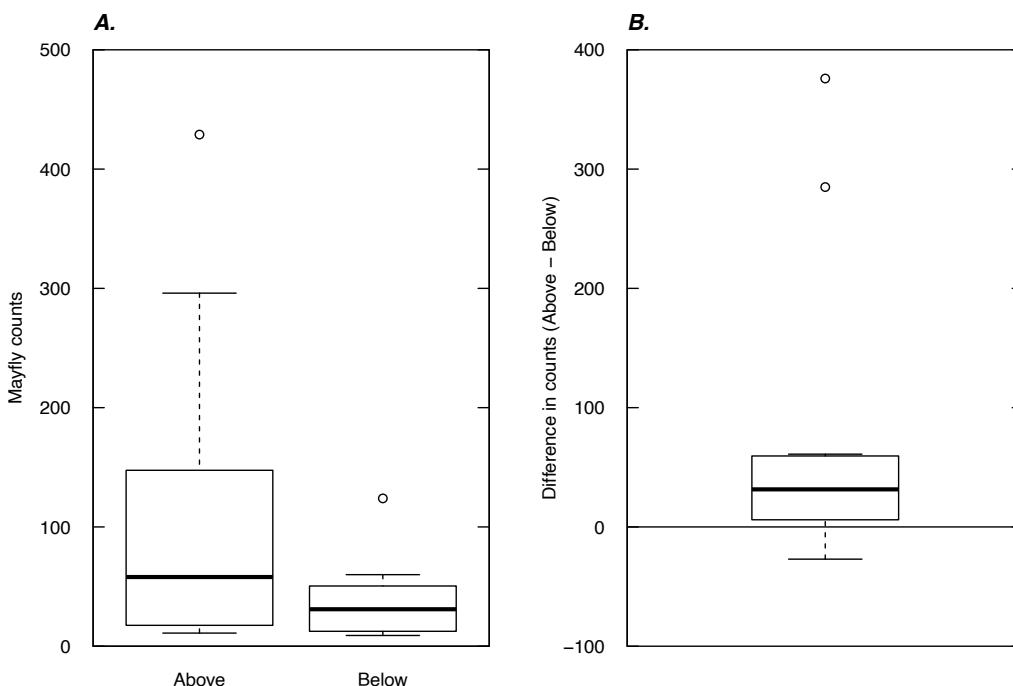
	Above	Below	signdiff
1	12	9	1
2	15	9	1
3	11	38	-1
4	41	24	1

5	106	48	1
6	63	17	1
7	296	11	1
8	53	41	1
9	20	14	1
10	110	60	1
11	429	53	1
12	185	124	1

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 two groups would be insufficient, 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 for which one is testing. The natural pairing of observations at the same stream can be used to block out the stream-to-stream variability by computing the difference in counts for each stream (fig. 6.1B). Two outliers are evident.

The null hypothesis  $H_0$  is that the mayfly counts above the outfalls are equally likely to be higher or lower than counts below the outfalls. The one-sided alternative hypothesis  $H_{A2}$  is that the counts above the outfalls are expected to be higher, so the Above-Below  $S^+$  statistic would be significantly greater than  $\frac{n}{2}$ .

Of the 12 pairs (trials), 11 are pluses, 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. R computes the exact sign test with the `binom.test` command, along with the observed proportion of pluses (probability of success) of 0.917 (or 11/12). The one-sided lower 95-percent confidence bound on that proportion is 0.66—the lower limit on the expected proportion of occurrences where **Above** is greater than **Below** in the population (the real world). There is no upper bound—1.000 is used here because this is the highest proportion possible



**Figure 6.1.** Boxplots of (A) mayfly nymph counts at two different sites, Above and Below, and (B) the differences ( $D_i = \text{Above}_i - \text{Below}_i$ ).

representing that no endpoint was determined. Using the alternative (`alt = "greater"`) option means that you are not asking for an upper bound. In routines where the scale is not bounded, a one-sided test or interval will report a value of infinity rather than 1.000 to represent that no endpoint was computed.

```
> binom.test(11, 12, alt = "greater")

Exact binomial test

data: 11 and 12
number of successes = 11, number of trials = 12, p-value
= 0.003174
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
0.6613193 1.0000000
sample estimates:
probability of success
0.9166667
```

The exact one-sided  $p$ -value for  $S^+ = 11$  is 0.003. Therefore reject that counts above and below the outfall are the same for the stated  $\alpha$  of 0.01.

### 6.1.3 The Large-sample Approximation to the Sign Test

For sample sizes of  $n > 20$ , the exact sign test statistic can be modified so that its distribution closely follows a normal or chi-square distribution, depending on the form of the approximation used. Again, this does not mean that the data or their differences require normality. It is only the modified test statistic that approximately follows a standard distribution.

The large-sample approximation for the sign test using a standard normal distribution takes the form

$$Z^+ \left\{ \begin{array}{ll} \frac{S^+ - \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ > \mu_{S^+} \\ 0 & \text{if } S^+ = \mu_{S^+} \\ \frac{S^+ + \frac{1}{2} - \mu_{S^+}}{\sigma_{S^+}} & \text{if } S^+ < \mu_{S^+} \end{array} \right. , \quad (6.2)$$

where

$$\mu_{S^+} = \frac{n}{2} \quad \text{and}$$

$$\sigma_{S^+} = \frac{1}{2}\sqrt{n} .$$

The  $1/2$  in the numerator of  $Z^+$  is a continuity correction (see section 5.1.4).  $Z^+$  is compared to quantiles of the standard normal distribution to obtain the approximate  $p$ -value. The square of  $Z$  is used when compared to a chi-square distribution with 1 degree of freedom. The chi-square approximation is available in R using the `prop.test` command.

**Example 6.2. Mayfly nymphs—Large-sample approximation for the sign test.**

```
> prop.test(11, 12, alternative = "greater")
```

```
1-sample proportions test with continuity correction
data: 11 out of 12, null probability 0.5
X-squared = 6.75, df = 1, p-value = 0.004687
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
0.6482637 1.0000000
sample estimates:
p = 0.9166667
```

The approximation  $p$ -value of 0.0046 is reasonably close to the exact  $p=0.003$ , though the exact test should be preferred with these small sample sizes. The availability of exact  $p$ -values in R software makes approximate methods far less necessary than in the past—there is no reason to use them if an exact test can be computed. If commercial software produces approximate  $p$ -values close to the agreed-upon risk  $\alpha$ , and the sample size is generally 20 or smaller, perform the exact test to get accurate  $p$ -values. Otherwise, an approximate procedure should be fine.

## 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 dataset is significantly different from zero.

### 6.2.1 Null and Alternative Hypotheses for the Signed-rank Test

For  $D_i = x_i - y_i$ , the null hypothesis for the signed-rank test is stated as

$$H_0: \text{median}[D] = 0.$$

The alternative hypothesis is one of three statements:

$$H_{A1}: \text{median}[D] \neq 0 \text{ (Two-sided test—}x\text{ might be larger or smaller than }y\text{).}$$

$$H_{A2}: \text{median}[D] > 0 \text{ (One-sided test—}x\text{ is expected to be larger than }y\text{).}$$

$$H_{A3}: \text{median}[D] < 0 \text{ (One-sided test—}x\text{ is expected to be smaller than }y\text{).}$$

The signed-rank test is usually stated as a determination of whether data from the two groups come from the same population (same median) or alternatively that they differ in location (median). If both groups are from the same population, regardless of the shape for both distributions, about half of the time their difference will be above 0 and half of the time their difference will be 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. Symmetry of the differences is a requirement of the signed-rank test, but they do not need to be anything like a normal distribution. 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 have a similar shape, with the only difference being that one box is offset from the other by the median difference between paired observations. The signed-rank test determines whether this offset is significantly different from zero. For an additive difference between groups, the assumption that the distribution of differences will be symmetric is valid, and 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 original scale, but symmetry in the differences can be achieved by a transformation of both datasets. If the transformation is with logarithms, a multiplicative relation on the original scale results in an additive relation in the logs. The  $y$  group has a higher median and variance, whereas the  $x$  (background) group has a lower median and variance. This is quite common in water-resources data. In the original scale, the differences between pairs are asymmetric. By taking logs prior to calculating differences, a symmetric distribution of data often results. The log transformation changes a multiplicative relation  $y_i = c \cdot x_i$ , to an additive one:  $\log(y_i) = \log(c) + \log(x_i)$ . The variances of the logs are often made similar by the transformation, so that the logs of the two groups differ only in central value (median). The transformed differences in log units are therefore much more symmetric than the differences in the original scale. The median difference in the logs can then be retransformed to estimate the median ratio on the original scale,

$$\hat{c} = \text{median}\left[\frac{y}{x}\right] = \exp\left(\text{median}[\log(y_i) - \log(x_i)]\right). \quad (6.3)$$

### 6.2.2 Computation of the Exact Signed-rank Test

The exact form of the signed-rank test is the best form for comparing 15 or fewer pairs of samples. With larger sample sizes, the large-sample approximation (section 6.2.3.) may be used.

**Computation:** Compute the absolute value of the differences  $|D_i|$ ,  $i=1, 2, \dots, N$ . Rank the  $|D_i|$  from smallest to largest. Delete any  $D_i = 0$  and adjust the sample size to  $n = N - [\text{number of } D_i = 0]$ . Compute the signed rank  $R_i = i = 1, 2, \dots, n$

$R_i$  = rank of  $|D_i|$  for  $D_i > 0$ , and

=  $-(\text{rank of } |D_i|)$  for  $D_i < 0$ .

When two nonzero differences are tied, assign the average of the ranks involved to all tied values.

**Test statistic:** The exact test statistic  $W^+$  (or  $V$  in the output from R's `wilcox.test` command) is the sum of all signed ranks  $R_i$  having a positive sign:

$$W^+ = \sum_{i=1}^n (R_i \mid R_i > 0), \quad (6.4)$$

where

| signifies "given that."

**Decision rule:** To reject  $H_0: \text{median}[D]=0$  when the  $p$ -value for  $W^+$ , either one- or two-sided as appropriate, is less than  $\alpha$ .

#### Example 6.3. Mayfly nymphs—Exact signed-rank test.

```
> D.i <- Above - Below
> SR.i <- rank(abs(D.i))*sign(D.i)
> nymph.sr <- data.frame(Above, Below, D.i, SR.i)
> nymph.sr
  Above Below D.i SR.i
1     12    9   3   1
2     15    8   7   3
3     11   38 -27  -6
4     41   24  17   5
5    106   48  58   9
6     63   17  46   7
7    296   11 285  11
```

```

8      53     41   12     4
9      20     14    6     2
10     110    60   50     8
11     429    53  376    12
12     185    124   61    10
> wilcox.test(Above, Below, alternative = "greater", paired = TRUE,
  exact = TRUE)

```

Wilcoxon signed rank test

```

data: Above and Below
V = 72, p-value = 0.003418
alternative hypothesis: true location shift is greater than 0

```

The test statistic  $V$ , the sum of the positive  $SR_i$  values, is 72. The R command for the signed-rank test is `wilcox.test`, specifying `paired = TRUE`. The exact test will be computed for small sample sizes unless there are ties in the differences. Here the exact test was specified by `exact = TRUE` to be sure the exact  $p$ -value, shown as 0.003, is produced.

### 6.2.3 The Large-sample Approximation for the Signed-rank Test

The large-sample approximation is computed by standardizing the exact test statistic; this is accomplished by subtracting its mean and dividing by its standard deviation. The distribution of the test statistic (not the data) was designed to be approximated by a standard normal distribution. This approximation is valid for sample sizes of  $n > 15$ . The large-sample approximation for the signed-rank test takes the form

$$Z_{sr}^+ = \begin{cases} \frac{W^+ - \frac{1}{2} - \mu_{W^+}}{\sigma_{W^+}} & \text{if } W^+ > \mu_{W^+} \\ 0 & \text{if } W^+ = \mu_{W^+} \\ \frac{W^+ + \frac{1}{2} - \mu_{W^+}}{\sigma_{W^+}} & \text{if } W^+ < \mu_{W^+} \end{cases}, \quad (6.5)$$

where

$$\begin{aligned} \mu_{W^+} &= \frac{n(n+1)}{4} \quad \text{and} \\ \sigma_{W^+} &= \sqrt{\frac{n(n+1)(2n+1)}{24}} . \end{aligned}$$

The  $1/2$  in the numerator of  $Z_{sr}^+$  is the continuity correction (chap. 4).  $Z_{sr}^+$  is compared to quantiles of the standard normal distribution, a normal distribution with mean of  $0$  and standard deviation of  $1$ , to obtain the approximate  $p$ -value for the signed-rank test. The large-sample approximation can be specified in R using the `exact = FALSE` option. There is no reason to do this other than for demonstration purposes. This is the version of the test that most commercial statistics software will run.

#### Example 6.4. Mayfly nymphs—Large-sample approximation to the signed-rank test.

```
> wilcox.test(Above, Below, alternative = "greater", paired = TRUE,
  exact = FALSE)
```

```
Wilcoxon signed rank test with continuity
correction
```

```
data: Above and Below
V = 72, p-value = 0.005394
alternative hypothesis: true location shift is greater than 0
```

The large-sample approximation  $p$ -value of  $0.005$  is similar to the exact test results ( $p=0.003$ ). If sample sizes are small and  $p$ -values are close to where a decision may change if the  $p$ -value changes by small amounts, use the exact test. Generally, let R perform the appropriate test automatically by not specifying the `exact=` option.

#### 6.2.4 Permutation Version of the Signed-rank Test

A permutation version of the signed-rank test (`wilcoxonsign_test` command using the `distribution = "approximate"` option) can be computed using the `coin` package of R (Hothorn and others, 2008). There is no advantage to this over using the exact test, but it will likely be a better approximation than the large-sample approximation of the previous section.

#### Example 6.5. Mayfly nymphs—Permutation version of the signed-rank test.

```
> require(coin)
> wilcoxonsign_test(Above ~ Below, alternative = "greater",
+   distribution = "approximate")
```

```
Approximative Wilcoxon-Pratt Signed-Rank Test
```

```
data: y by
  x (pos, neg)
stratified by block
Z = 2.5897, p-value = 0.0036
alternative hypothesis: true mu is greater than 0
```

The permutation  $p$ -value of  $0.0036$  is much closer to the exact  $p$ -value ( $p=0.003$ ) than was the large-sample approximation. Note the somewhat nonstandard input structure: for `wilcoxonsign_test`, the `~` sign does not indicate that a grouping variable follows, but that the paired columns are placed on either side of the `~` symbol.

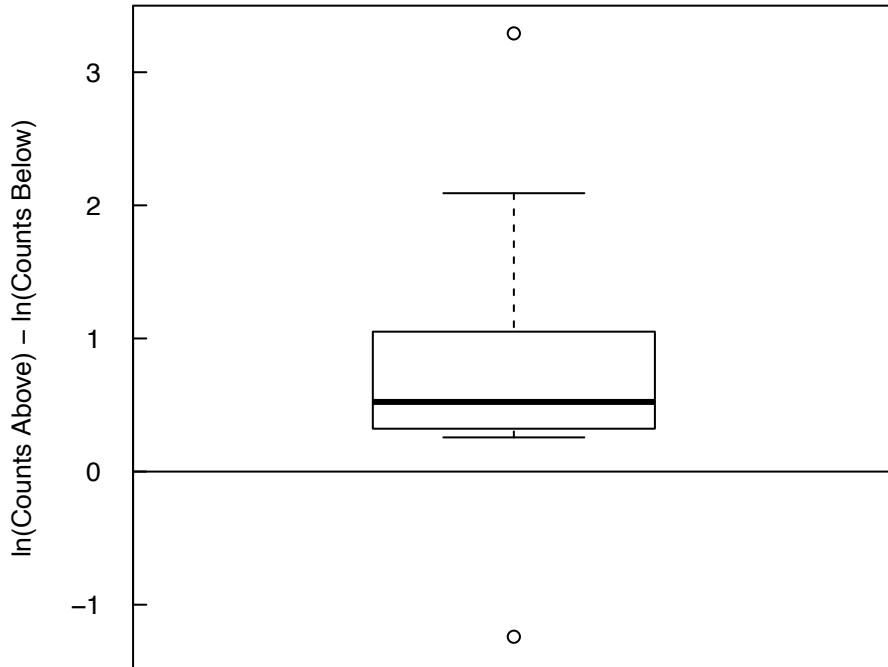
### 6.2.5 Assumption of Symmetry for the Signed-rank Test

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. Although even 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 the positive differences before a signed-rank test rejects  $H_0$  due solely to asymmetry. Outliers generally will have little effect on the signed-rank test, as it uses their rank and not their actual values for the computation. Violation of the symmetry assumption of the signed-rank test produces  $p$ -values only slightly lower than they should be, whereas 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. The permutation form of the test should be run if the symmetry assumption is strongly violated.

#### Example 6.6. Mayfly nymphs—Signed-rank test on logarithms.

The Above-Below differences are asymmetric in figure 6.1B, violating one of the signed-rank test's assumptions and indicating that the differences between the two groups may not be an additive one. Asymmetry can be expected when large values tend to produce large differences and smaller values smaller differences. This indicates that a multiplicative relation between the data pairs is more realistic. Here the natural logs of the data are calculated, and a new set of differences  $Dl_i = \log(x_i) - \log(y_i)$  are computed and shown in figure 6.2. Comparing figure 6.2 and 6.1B, note that differences in natural log units are much more symmetric than those in the original scale.

```
> ldiff <- log(Above)-log(Below)      # log = natural logs in R
> boxplot(ldiff, ylab = "ln(Counts Above) -ln(Counts Below)")
```



**Figure 6.2.** Boxplot of the differences of the natural logarithms of the mayfly data from example 6.1.

Computing the signed-rank test on the  $Dl_i$ , the exact  $p$ -value is 0.008.

```
> wilcox.test(log(Above), log(Below), paired = TRUE, alt = "greater",
+               exact = TRUE)
```

Wilcoxon signed rank test

```
data: log(Above) and log(Below)
V = 69, p-value = 0.008057
alternative hypothesis: true location shift is greater than 0
```

Inaccurate  $p$ -values for the signed-rank test are not the primary problem caused by asymmetry. The  $p$ -values for the mayfly data, for example, are not that different ( $p=0.003$  on the original scale and  $p=0.008$  for the natural 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 relation. Therefore, symmetry is especially important to check if the magnitude of the difference between data pairs is to be estimated. If there is a belief that a multiplicative change is more realistic than an additive change, using logarithms would model that relation. Checking the form of the relation between the two sets of data can be done using the scatterplots of section 6.4. If an additive difference is more realistic, use the permutation version of the test to avoid the consequences of asymmetry.

## 6.3 The Paired $t$ -test

The paired  $t$ -test evaluates whether the mean difference,  $\bar{D}$ , of matched pairs is zero. The test requires the paired differences,  $D_i$ , to follow a normal distribution. Logarithms may be taken to reduce asymmetry in the differences prior to running a paired  $t$ -test, but the results will not indicate whether means in the original scale are similar or not. Transformations change the meaning of a mean. After a log transformation, the  $t$ -test instead evaluates whether the ratio of the two geometric means on the original scale equals 1. If the means on the original scale are of interest and the paired differences do not follow a normal distribution, use a permutation test instead of a transformation.

### 6.3.1 Null and Alternate Hypotheses

The null hypothesis can be stated as

$H_0: \mu_x = \mu_y$  the means for the  $x_i$  and  $y_i$  are identical, or

$H_0: \mu_{[D]} = 0$  the mean difference between the  $x_i$  and  $y_i$  equals 0.

The three possible alternative hypotheses are

$H_{A1}: \mu_x \neq \mu_y$  the two group means differ, and both possible directions are of interest. Reject  $H_0$  if the two-sided  $p$ -value is less than  $\alpha$ ;

$H_{A2}: \mu_x > \mu_y$  prior to seeing any data,  $\mu_x$  is expected to be greater than  $\mu_y$ . Reject  $H_0$  if the one-sided  $p$ -value is less than  $\alpha$ ; and

$H_{A3}: \mu_x < \mu_y$  prior to seeing any data,  $\mu_y$  is expected to be greater than  $\mu_x$ . Reject  $H_0$  if the one-sided  $p$ -value is less than  $\alpha$ .

### 6.3.2 Computation of the Paired *t*-test

For two sets of data,  $x_i$  and  $y_i$ , paired by the attribute  $i=1, 2, \dots, n$ , compute the paired differences  $D_i = x_i - y_i$  and then  $\bar{D}$ , the sample mean of the differences.

**Test statistic:** The paired *t*-statistic is

$$t_p = \frac{\bar{D}\sqrt{n}}{s}, \quad (6.6)$$

where

$s$  is the standard deviation of the differences.

#### Example 6.7. Mayfly nymphs—Paired *t*-test.

As the *t*-test requires that the paired differences follow a normal distribution, test for normality of the differences ( $D_i$ ) prior to running the paired *t*-test. The null hypothesis for the Shapiro-Wilk test for normality is that data follow a normal distribution and a small *p*-value rejects this hypothesis (see section 4.4).

```
> shapiro.test(D.i)
```

```
Shapiro-Wilk normality test
```

```
data: D.i
W = 0.68339, p-value = 0.0005857
```

The very small *p*-value indicates that the paired differences of the mayfly data do not come from a normal distribution at an  $\alpha$  of 0.05.

The *t*-test is run below for demonstration purposes only, as the assumptions of the test are violated.

```
> t.test(Above, Below, paired = TRUE, alternative = "greater")
```

```
Paired t-test
```

```
data: Above and Below
t = 2.0824, df = 11, p-value = 0.03072
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
10.24942      Inf
sample estimates:
mean of the differences
```

```
74.5
```

The *p*-value produced is likely to be pushed upwards by the non-normality. Note that it is an order of magnitude higher than that for the signed-rank test.

In an attempt to obtain a distribution closer to normal, the logarithms of the data are computed. As with the signed-rank test, this implies that a multiplicative rather than an additive relation exists between the two sets of data. The Shapiro-Wilk test for normality of the differences between the logarithms has a *p*-value of 0.083, higher than the  $\alpha$  of 0.05, so normality in these units is not rejected. A paired *t*-test on the difference in logarithms should work well enough, but it will test whether the geometric means of the  $x_i$  and  $y_i$ , and not their arithmetic means, are similar.

```

> shapiro.test(ldiff)

Shapiro-Wilk normality test

data: ldiff
W = 0.87824, p-value = 0.08321

> t.test(log(Above), log(Below), paired = TRUE, alternative =
  "greater")

Paired t-test

data: log(Above) and log(Below)
t = 2.4421, df = 11, p-value = 0.01635
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.2054557      Inf
sample estimates:
mean of the differences
0.7764593

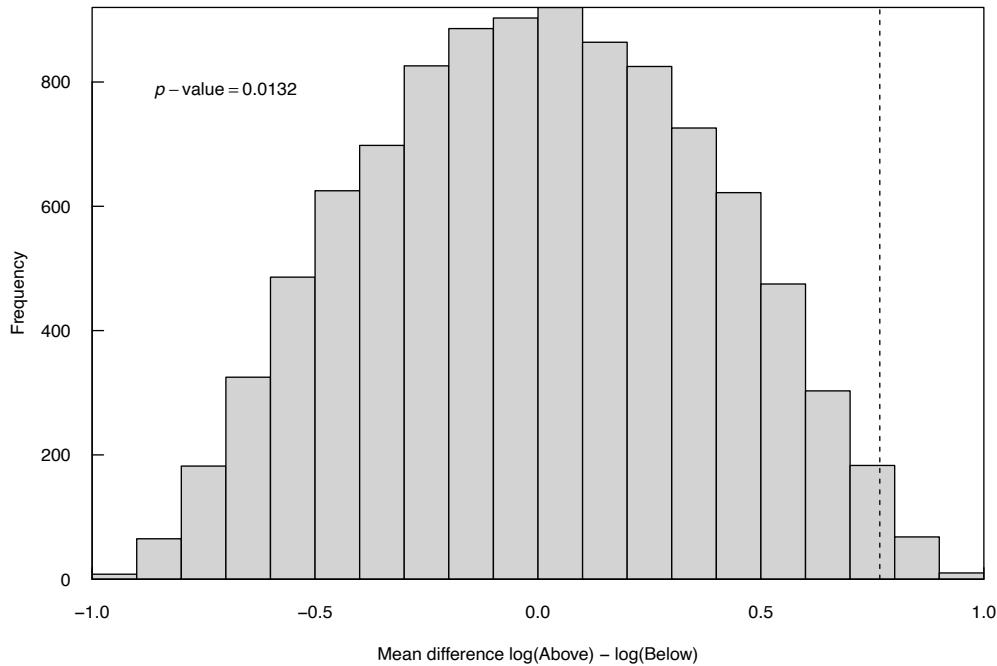
```

The one-sided  $p$ -value for  $t_p$  is 0.016. Therefore, reject that  $\mu_{lnx} = \mu_{lny}$  in favor of  $H_{A2}$ , the mean of the natural log of the  $x_i$  is greater than the mean of natural log of the  $y_i$ . Equivalently, reject the null hypothesis that the ratio of geometric means of the groups is 1.

### 6.3.3 Permutation Test for Paired Differences

Although the paired  $t$ -test assumes normality of the paired differences, a permutation test on the differences can be used to test whether the mean difference equals zero regardless of the distributional shape of the differences. Sutton (1993) found that these tests perform better than traditional  $t$ -tests in the presence of asymmetry. The use of paired differences in the original scale still assumes that an additive difference is appropriate, and that the mean difference is the best measure of difference between the two groups. If the variances of the two groups differ, and if the group with higher variance is also the group with the higher mean, a multiplicative difference is probably a better model. That appears to be the case for the mayfly data in figure 6.1, so we have taken natural logarithms prior to computing the paired differences to more accurately model the variation seen in the data.

The permutation test equivalent to the paired  $t$ -test assumes that the differences are symmetric because a mean is being tested, though a normal distribution is not required. The difference in logs of the mayfly data seems relatively symmetric (see fig. 6.2). An R script for the permutation test of mean differences of matched pairs (`permMatched.R`) is available in the supplemental material (SM.6). The `permMatched` script randomly assigns an algebraic sign to each of the observed differences and computes the test statistic, which is simply the mean difference. A distribution of test statistics representing the null hypothesis ( $H_0$ : the mean difference equals zero) is constructed by repeating this process several thousand times. The observed mean difference is compared to the distribution of test statistics, and the proportion of the distribution that is outside (above or below for a one-sided test) the observed mean difference is the permutation  $p$ -value of the test.



**Figure 6.3.** Histogram of permuted differences (representing the null hypothesis) and the observed mean difference from the logs of mayfly data from example 6.1. The proportion of the entire area at and to the right of the dashed line is the *p*-value of the test, which equals 0.013.

#### Example 6.8. Mayfly nymphs—Permutation test of paired differences.

```
> permMatched(log(Above), log(Below), alt = "g")
```

```
Permutation Matched-Pair Test R= 10000
log(Above) - log(Below) alternative = g
p-value = 0.0131
mean difference = 0.7764593
```

The permutation test *p*-value of 0.013 is slightly less than the normal-theory result of *p*=0.017, both of which reject the null hypothesis at a 5-percent significance level. A histogram of the permuted difference in means (fig. 6.3) shows that the distribution of differences between log counts looks fairly symmetric, though with a broader peak than a normal distribution would have. Even when differences appear to follow a normal distribution the permutation test works well, producing *p*-values similar to the paired *t*-test.

#### 6.3.4 The Assumption of Normality for the Paired *t*-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 the groups differ, it is only in their mean (central value).

When the  $D_i$  are not normally distributed, and especially when they are not symmetric, the *p*-values obtained from the *t*-test will not be accurate. Sutton (1993) states that it has been known since the 1920s that tests based on the *t*-statistic suffer from a loss of power when the data distribution has positive skewness. This effect is more severe for one-sided tests than two-sided tests but is present for both. When the  $D_i$  are asymmetric the mean will also not provide a good estimate of the center, as discussed in chapter 1.

To illustrate the importance of defining your objective for a test, a *t*-test on the original scale was computed above to see if the first group had more counts than the second, while ignoring the non-normality

of the differences. The test statistic of  $t=2.08$  had 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  $\alpha$  of 0.01 been chosen, the  $t$ -test would be unable to reject  $H_0$  whereas the sign test would easily reject it. It is important to clearly state your objective in order to choose the appropriate test. Even if the difference in means were the objective, 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 more difficult to discern. The measure of the confusion for any dataset is the difference in  $p$ -values between the normal-theory  $t$ -test and its permutation alternative. As permutation tests work well when data do follow a normal distribution, there is little reason to use the  $t$ -test when the mean difference is your objective.

The mean difference,  $\bar{D}$ , of 74.5 counts for the mayfly data is larger than 10 of the 12 paired differences listed in table 6.3. The mean is often not a typical value with skewed data—it has little usefulness as a measure of how many more mayfly nymphs are typically found above outfalls than below. If the objective is to determine whether one group is generally higher than the other, use a nonparametric test, with the associated Hodges-Lehmann estimate of median difference (section 6.5.2.) providing a more typical measure of difference between groups. Another drawback to the mean is that when transformations are used prior to computing a  $t$ -test, retransforming the estimate of the mean difference back into the original scale does not provide an estimate of the mean difference in the original scale.

## 6.4 Graphical Presentation of Results

Methods for illustrating matched-pair test results are described in chapter 2 for illustrating a single variable, as the differences between matched pairs are a single variable. A probability (Q-Q) plot of the paired differences shows whether or not those differences follow a normal distribution. Here we discuss two plots that illustrate both the test results and the degree of conformity to the test's assumptions.

### 6.4.1 Boxplots

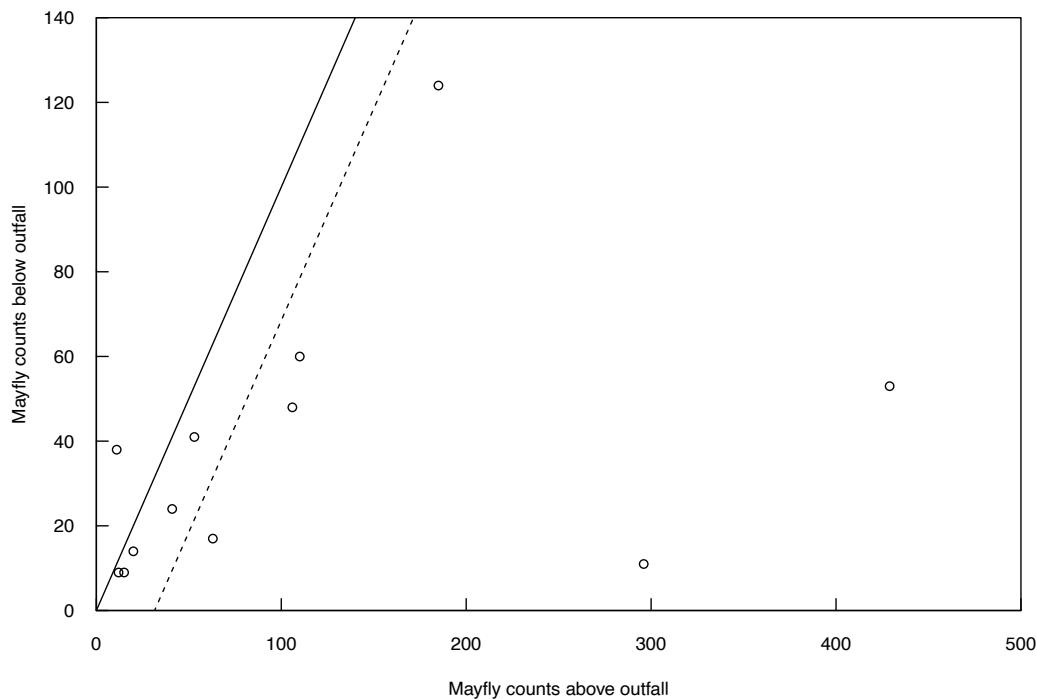
The best method for directly illustrating the results of tests in this chapter is a boxplot of the differences as in figure 6.1B, or as in figure 6.2 for the natural logarithms. The number of data above and below zero and the nearness of the median difference to zero are clearly displayed, as is the degree of symmetry of the  $D_i$ . Although a boxplot is an effective and concise way to illustrate the characteristics of the test for differences, boxplots of the original data for both groups are more intuitive (fig. 6.1A) and might be a good addition for presentations.

### 6.4.2 Scatterplots with a One-to-one Line

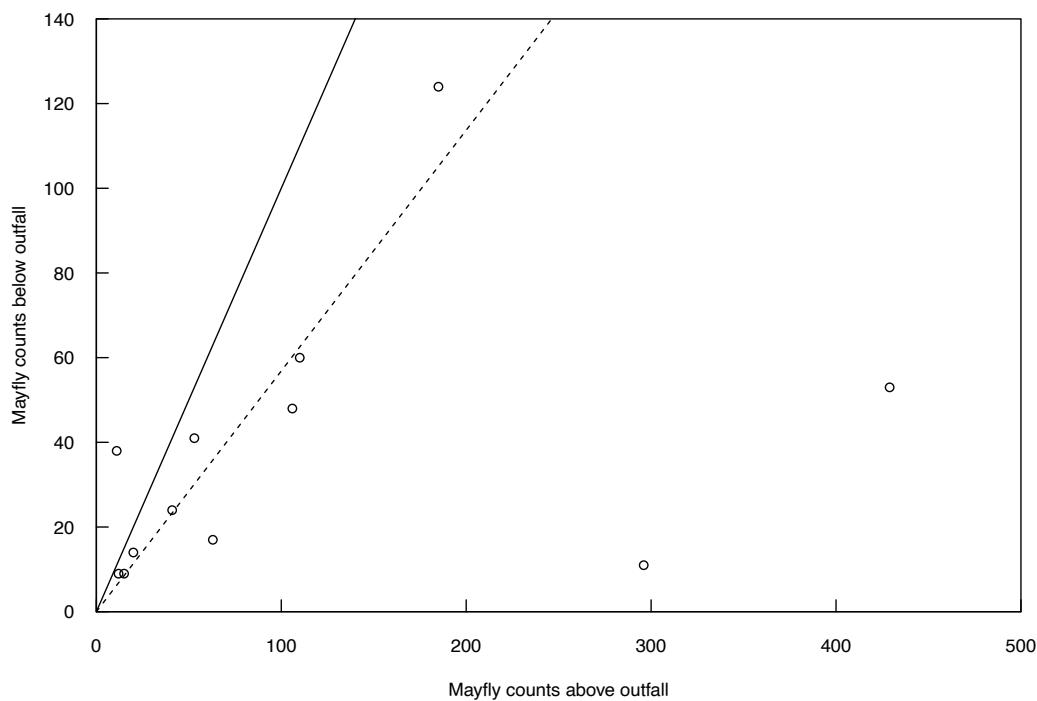
Scatterplots illustrate the relations between paired data (fig. 6.4). Each  $(x_i, y_i)$  pair is plotted on the scatterplot as a point. Similarity between the paired data is shown by the  $x=y$  line. If the variable plotted on the  $x$  axis is generally greater than the variable on the  $y$  axis, most of the data will fall to the right of, or below, the line. When  $y$  generally exceeds  $x$ , the data will lie largely to the left of, or above, the line.

The scatterplot illustrates that the Above data ( $x$ ) are generally greater than the Below data ( $y$ ), as all but one point falls to the right of the solid line. If there were an additive difference between data pairs, points would fall along a line pattern parallel to the  $x=y$  line. A line  $x=y+d$  is also plotted on the figure to illustrate the magnitude of the difference between  $x$  and  $y$ , where  $d$  is the mean or median estimate of the difference between data pairs, depending on objective. In figure 6.4 the dashed line  $y=x-31.5$  shows the median difference of 31.5 counts. For an additive relation the data points would scatter around this line.

Note that the dashed line in figure 6.4 also has a slope of 1, but not an intercept of 0. Determining whether a fitted line (say, a regression line) has slope equal to 1 is not the same as checking for similarity of paired data. A fitted line could have a slope equal to 1, but be offset by a significant difference from the  $x=y$  line—an additive difference that estimates the offset between group means or medians. The  $t$ -test and permutation tests look for an additive difference between paired observations. The sign and signed-rank tests evaluate differences more generally—whether one group tends to have higher values than the other. Thinking through the objective of your study is important when plotting data, just as much as when performing a numerical test.



**Figure 6.4.** Scatterplot of the mayfly data from example 6.1. Solid line is the  $x=y$  line. Dashed line is the  $y=x - \text{median difference}$  line,  $y=x - 31.5$ . Here  $x$  values generally fall to the right of the  $x=y$  line, showing that mayfly counts above the outfall are frequently greater than those below the outfall.



**Figure 6.5.** Mayfly data from example 6.1. The multiplicative relation  $y=0.555 \cdot x$  is shown as the dashed line. For reference, the solid line is the  $x=y$  line.

A multiplicative difference between the two datasets appears to be a better fit, and therefore using logarithms may be more appropriate. The line  $x = y \cdot f^{-1}(d)$  where  $d$  is an additive difference in transformed units can be plotted as an aid in visualizing the relation. For natural logs,  $f^{-1}(d) = \exp(d)$ . This model is illustrated in figure 6.5, where  $d = \hat{\Delta}$ , the Hodges-Lehmann estimator of group difference (see section 6.5.2.) in natural log of counts, which for the mayfly data equals  $-0.589$ . Converting to the original scale and plotting, the equation for the fitted dashed line is  $y = 0.555 \cdot x$ .

## 6.5 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 mean difference is considered a good central value, an efficient estimator is the mean difference,  $\bar{D}$ . This estimator is appropriate whenever the paired  $t$ -test, or its permutation test equivalent, is used. When outliers or non-normality are suspected, a more robust estimator is the Hodges-Lehmann estimator,  $\hat{\Delta}$ , the median of all possible pairwise averages of the differences. It is the appropriate measure of difference when the signed-rank test is used. 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,  $D_{med}$ .

### 6.5.1 The Median Difference (Following the Sign Test)

For the mayfly data, the median difference in counts,  $D_{med}$ , equals 31.5. As these data are asymmetric, there is no assumption that the two groups are related in an additive fashion. But subtracting the 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 on the original scale. Half of the differences are larger and half are smaller than the median.

A confidence interval on the median indicates the precision with which the difference between groups, as measured by the sign test, is known. It is simply the confidence interval on the median, which was previously presented in chapter 3.

### 6.5.2 The Hodges-Lehmann Estimator (Following the Signed-rank Test)

The estimate of difference between groups associated with the signed-rank test is the Hodges-Lehmann difference,  $\hat{\Delta}$ . When outliers or non-normality are suspected, it is a more robust estimator of the difference between groups than is the difference in means. 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  $n \cdot (n+1)/2$  possible pairwise averages:

$$\hat{\Delta} = \text{median}[A_{ij}], \quad (6.7)$$

where

$$A_{ij} = [(D_i + D_j)/2] \text{ for all } i \leq j.$$

Note that this version differs from the equation in chapter 5 because this estimator is based on paired differences. Chapter 5 presented a Hodges-Lehman estimator for two independent datasets.

The estimator is related to the signed-rank test in that subtracting  $\hat{\Delta}$  from all paired differences (or equivalently, from the  $x_i$  or  $y_j$ , whichever is larger) would cause the signed-rank test to have a test statistic  $W^+$  close to 0 and find no evidence of difference between data pairs. For the cases of symmetric differences where the signed-rank test is appropriate, the Hodges-Lehmann estimator  $\hat{\Delta}$  more efficiently measures the additive difference between two data groups than does the sample median of the differences,  $D_{med}$ . R computes  $\hat{\Delta}$ , calling it the (psuedo)median, when the `conf.int` option is specified as `TRUE` for the `wilcox.test` command. For the mayfly data,  $\hat{\Delta}$  of the natural logarithms = 0.589.

```
> wilcox.test(log(Above), log(Below), alternative = "greater",
+      paired = TRUE, conf.int = TRUE)
```

Wilcoxon signed rank test

```
data: log(Above) and log(Below)
V = 69, p-value = 0.008057
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
0.3566749      Inf
sample estimates:
(pseudo)median
0.5891098
```

```
> exp(-0.5891098)
[1] 0.554821
```

The log of upstream counts minus  $\hat{\Delta}$  models the log of the counts below the outfalls. Thus, the counts above the outfalls multiplied by  $e^{-0.589} = 0.555$  models the counts below the outfalls (the dashed line  $y=0.555 \cdot x$  in fig. 6.5).

The nonparametric confidence interval around  $\hat{\Delta}$  is computed by finding the ranks of the data points representing the ends of the interval. These are a function only of the sample size and  $\alpha$ . The pairwise average 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, quantiles for the signed-rank test at the  $\alpha$  nearest to  $\alpha/2$  and  $1-\alpha/2$  (two-sided interval) or at  $\alpha$  (one-sided interval) give the ranks  $R_u$  and  $R_l$  corresponding to the  $A_{ij}$  at the upper and (or) lower confidence limits for  $\hat{\Delta}$ . These limits are the  $R_l$ th ranked  $A_{ij}$  going from one or both ends of the sorted list of  $n \cdot (n+1)/2$  differences. In R, the `wilcox.test` command with `conf.int = TRUE` does this for you.

For larger sample sizes where the large sample approximation is used, quantiles of standard normal distribution provide the upper and lower ranks of the pairwise average differences,  $A_{ij}$ , corresponding to the ends of the confidence interval. Those ranks are

$$R_l = \frac{N - z_{\alpha/2} \cdot \sqrt{\frac{n(n+1)(2n+1)}{6}}}{2}, \text{ and} \quad (6.8)$$

$$R_u = \frac{N + z_{\alpha/2} \cdot \sqrt{\frac{n(n+1)(2n+1)}{6}}}{2} + 1 = N - R_l + 1, \quad (6.9)$$

where

$$N = n \cdot (n+1)/2.$$

For one-sided intervals, choose the appropriate lower or upper limit using  $\alpha$  instead of  $\alpha/2$ .

#### **Example 6.9. Mayfly nymphs—Estimate of median difference and two-sided confidence interval.**

For the  $n=12$  logarithms of the mayfly data in example 6.1, there are  $N=78$  pairwise averages. For an  $\alpha \approx 0.05$  two-sided confidence interval on the difference between groups, the 14th and 65th ranked averages (the 14th average in from either end) form the ends of the two-sided 95-percent confidence interval. For

the difference Above-Below in log units the interval is from 0.307 to 1.442, and the Hodges-Lehmann estimate at its center is 0.589.

```
> wilcox.test(log(Above), log(Below), paired = TRUE, conf.int=TRUE)
```

```
Wilcoxon signed rank test
```

```
data: log(Above) and log(Below)
V = 69, p-value = 0.01611
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
0.3066974 1.4417015
sample estimates:
(pseudo)median
0.5891098
```

### 6.5.3 Mean Difference (Following a *t*-test)

For the situation where the differences are not only symmetric but also normally distributed and the *t*-test is used, the most efficient (precise) estimator of the difference between the two groups is the mean difference,  $\bar{D}$ . However,  $\bar{D}$  is in this case 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  $\bar{D}$ . 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 field data, which is never exactly normal,  $\bar{D}$  has little advantage over  $\hat{\Delta}$ , whereas  $\hat{\Delta}$  is more appropriate in a wider number of situations—for data that are symmetric but not normal.

A confidence interval on  $\bar{D}$  is computed exactly like any confidence interval for a mean. For a two-sided interval

$$CI = \bar{D} \pm t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} , \quad (6.10)$$

where  $s$  is the standard deviation of the differences,  $D_i$ . A one-sided interval uses  $\alpha$  instead of  $\alpha/2$ . These are output by the *t.test* command in R.

**Example 6.10. Mayfly nymphs—Estimate of mean difference and two-sided confidence interval.**

```
> t.test(log(Above), log(Below), paired = TRUE)
```

```
Paired t-test
```

```
data: log(Above) and log(Below)
t = 2.4421, df = 11, p-value = 0.0327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.07665364 1.47626496
sample estimates:
mean of the differences
0.7764593
```

## Exercises

1. Which of the following are not matched pairs?
  - A. Analyses of a series of standard solutions, each sent to two different laboratories.
  - B. Same-day evaluations of sediment toxicity over several years at two different sites.
  - C. Nitrate analyses from randomly selected wells in each of two aquifers.
  - D. Contaminant concentrations measured above and below stormwater retention basins on several urban streams.
2. Specific conductance was measured on the two forks of the Shenandoah River in Virginia and the resulting data are included in the file `Shenandoah.rda`. Does the North Fork, which drains terrain with calcareous (more soluble) minerals have higher conductance than the South Fork, which drains terrain with quartz-rich (less soluble) minerals?
  - A. State the appropriate null and alternative 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  $\alpha=0.05$  test and report the results.
  - D. Estimate the amount by which the forks differ in conductance, regardless of the test outcome.
  - E. Illustrate and check the results with a plot.
3. Atrazine concentrations in shallow groundwaters were measured by Junk and others (1980) before (June) and after (September) the application season and the data are paired by well. Determine if concentrations of atrazine are higher in groundwater following surface application than before. All values of  $<0.01$  have been set to  $-0.01$  so that they can easily be ranked as the lowest values and tied with one another.
 

```
> June <- c(0.38, 0.04, -0.01, 0.03, 0.03, 0.05, 0.02, -0.01,
        -0.01, -0.01, 0.11, 0.09, -0.01, -0.01, -0.01,
        -0.01, 0.02, 0.03, 0.02, 0.02, 0.05, 0.03, 0.05, -0.01)
      > Sept <- c(2.66, 0.63, 0.59, 0.05, 0.84, 0.58, 0.02, 0.01,
        -0.01, -0.01, 0.09, 0.31, 0.02, -0.01, 0.5, 0.03,
        0.09, 0.06, 0.03, 0.01, 0.1, 0.25, 0.03, 88.36)
      > well.num <- c(1:24)
```
4. Compare mean atrazine concentrations of the data in exercise 3 using a *t*-test, setting all values below the detection limit to zero (not recommended here, just as an exercise!). Compare the results with those of exercise 3. Discuss why the results are similar or different.

# Chapter 7

## Comparing Centers of Several Independent Groups

---

*Concentrations of volatile organic compounds are measured in shallow ground waters across a multi-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 effect has reclamation had? Are there differences in chemical or biological quality owing to rock type separate and distinct from the effects owing to mining impacts?*

*Three methods for field sampling and extraction of an organic chemical are to be compared at numerous wells. Are there differences among concentrations produced by the three processes? These must be discerned above the well-to-well differences in concentration that contribute considerable noise to the data.*

The methods of this chapter, comparing centers of several independent groups, can be used to answer questions such as those above. These methods are extensions of the ones introduced in chapters 5 and 6; in this chapter more than two groups of data will be compared. The parametric technique in this situation is analysis of variance (ANOVA). More robust nonparametric and permutation techniques are also presented for the frequent situations where data do not meet the assumptions of ANOVA.

First consider the effect of only one grouping variable, also called a factor. A factor is a categorical variable suspected of influencing the measured data, analogous to an explanatory variable in regression. The factor is made up of more than one level and each level is defined by a group of observations. Levels may be an ordered low-medium-high change in intensity or unordered categories such as different locations or times that represent a change in underlying influences. The factor 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 aquifers. The various aquifers are the groups or levels. Within each group (aquifer) there are  $n_j$  observations (the sample size of each of the  $j$  groups is not necessarily the same). Observation  $y_{ij}$  is the  $i$ th of  $n_j$  observations in group  $j$ , so that  $i=1, 2, \dots, n_j$  for the  $j$ th of  $k$  groups  $j=1, 2, \dots, k$ . The total number of observations  $N$  is thus

$$N = \sum_{j=1}^k n_j , \quad (7.1)$$

which simplifies to  $N=k \cdot n$  when the sample size  $n_j=n$  for all  $k$  groups (equal sample sizes per group, also called a balanced design).

When data within each of the groups are normally distributed and possess identical variances, classical ANOVA can be used. Analysis of variance is a parametric test, determining whether all group means are equal. ANOVA is analogous to a  $t$ -test between three or more groups of data and is restricted by the same assumptions as the  $t$ -test. When data in each group do not have identical variance, an adjustment similar to the one for the  $t$ -test will improve on classical ANOVA (Welch, 1951). When data in each group do not follow a normal distribution a permutation test can check differences between group means. When the

objective is to determine whether some groups have higher or lower values than others and is not focused on the mean as a parameter, nonparametric tests such as the Kruskal-Wallis (KW) and Brunner-Dette-Munk (BDM) tests will have more power than parametric ANOVA methods (table 7.1).

When the null hypothesis is rejected, these tests do not state which group or groups differ from the others! We therefore discuss multiple comparison tests—tests for determining which groups differ from others. These methods are then expanded to evaluating the effects of two factors simultaneously (table 7.2). These factorial methods determine whether neither, or one or both of two factors significantly affect the values of observed data. Although higher numbers of factors can be evaluated, the design of those studies and the tests that follow are beyond the scope of this book.

We finish the chapter by discussing repeated measures designs, the extension of the matched-pairs tests of chapter 6 to situations where three or more related observations are taken on each subject or block (table 7.3).

**Table 7.1.** Hypothesis tests with one factor and their characteristics.

[ANOVA, analysis of variance; BDM, Brunner-Dette-Munk test; MCT, multiple comparison test.  $H_A$  is the alternative hypothesis, the signal to be found if it is present]

Objective of test ( $H_A$ )					
	Data from at least one group is frequently higher than the other groups		Mean of at least one group is higher than the mean of the other groups		
Test	Kruskal-Wallis test	BDM test	ANOVA	Welch's adjusted ANOVA	Permutation test on group means
<b>Class of test</b>	Nonparametric	Nonparametric	Parametric	Parametric	Permutation
<b>Distributional assumption for group data</b>	None	None	Normal distribution; equal variances	Normal distribution	Exchangeable
<b>Multiple comparison test</b>	Pairwise rank-sum tests or Dunn's test	Pairwise rank-sum tests or Dunn's test	Tukey's MCT	Tukey's MCT	Tukey's MCT

**Table 7.2.** Hypothesis tests with two factors and their characteristics.

[ANOVA, analysis of variance; BDM, Brunner-Dette-Munk test; MCT, multiple comparison test.  $H_A$  is the alternative hypothesis, the signal to be found if it is present]

Objective of test ( $H_A$ )				
	Data from at least one group is frequently higher than the other groups		Mean of at least one group is higher than the mean of the other groups	
Test	BDM two-factor test		Two-factor ANOVA	Two-factor permutation test
<b>Class of test</b>	Nonparametric		Parametric	Permutation
<b>Distributional assumption for group data</b>	None		Normal distribution; equal variances	Exchangeable
<b>Multiple comparison test</b>	Pairwise rank-sum tests		Two-factor Tukey's MCT	Two-factor Tukey's MCT

**Table 7.3.** Hypothesis tests for repeated measures and their characteristics.

[ANOVA, analysis of variance; MCT, multiple comparison test.  $H_A$  is the alternative hypothesis, the signal to be found if it is present]

Objective of test ( $H_A$ )			
	Data from at least one group is frequently higher than the other groups	Mean of at least one group is higher than the mean of the other groups	
Test	Friedman test	Aligned-rank test	ANOVA without replication
Class of test	Nonparametric	Nonparametric	Parametric
Distributional assumption for group data	None	Symmetry	Normal distribution; equal variances
Multiple comparison test	Paired Friedman comparison tests	Tukey's MCT on aligned ranks	Pairwise paired <i>t</i> -tests

## 7.1 The Kruskal-Wallis Test (One Factor)

### 7.1.1 Null and Alternate Hypotheses for the Kruskal-Wallis Test

The Kruskal-Wallis (KW) test objectives are stated by the null and alternate hypotheses:

$H_0$ : All groups of data have identical distributions.

$H_A$ : At least one group differs in its distribution.

For the tests of this chapter, the alternate hypothesis  $H_A$  is always two-sided; no prior direction of difference is hypothesized, but only whether group differences exist or not.

### 7.1.2 Assumptions of the Kruskal-Wallis Test

For the general objective of determining whether all groups are similar in value, or alternatively that one or more groups more frequently have higher or lower values than the other groups, no assumptions are required about the shape of the distributions. They may be normal, lognormal, or anything else. If the alternate hypothesis is true, they may have different distributional shapes. This difference is not attributed solely to a difference in median, though that is one possibility. The test can determine differences where, for example, one group is a control group with only background concentrations, whereas the others combine background concentrations with higher concentrations owing to contamination. For example, 35 percent of the data in one group may have concentrations indicative of contamination and yet group medians remain similar. The KW test can see this type of change in the upper 35 percent as dissimilar to the control group.

The test is sometimes stated with a more specific objective—as a test for difference in medians. This objective requires that all other characteristics of the data distributions, such as spread or skewness, be identical—though not necessarily on the original scale. This parallels the rank-sum test (section 5.1). As a specific test for difference in medians, the Kruskal-Wallis null and alternate hypotheses are

$H_0$ : The medians of the groups are identical.

$H_A$ : At least one group median differs from the others.

As with the rank-sum test, the KW test statistic and *p*-value computed for data that are transformed using any monotonic transformation give identical test statistics and *p*-values to those using data on the original scale. Thus, there is little incentive to search for transformations (to normality or otherwise) as the test is applicable in many situations.

### 7.1.3 Computation of the Exact Kruskal-Wallis Test

The exact method determines a  $p$ -value by computing all possible test statistics when the observed data are rearranged, calculating the probability of obtaining the original test statistic or those more extreme. It is needed only for quite small sample sizes—three groups with  $n_j \leq 5$ , or with four or more groups of size  $n_j \leq 4$  (Lehmann, 1975). Otherwise, the large sample approximations are very close to their exact values.

To compute the test, all  $N$  observations from all groups are jointly ranked from 1 to  $N$ , smallest to largest. These ranks  $R_{ij}$  are used to compute the average rank  $\bar{R}_j$  for each of the  $j$  groups, where  $n_j$  is the number of observations in the  $j$ th group:

$$\bar{R}_j = \frac{\sum_{i=1}^{n_j} R_{ij}}{n_j}. \quad (7.2)$$

Compare  $\bar{R}_j$  to the overall average rank  $\bar{R}_{ij} = (N+1)/2$ , squaring and weighting by sample size, to form the test statistic  $K$ :

$$K = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left[ \bar{R}_j - \frac{N+1}{2} \right]^2. \quad (7.3)$$

When the null hypothesis is true, the average rank for each group should be similar to one another and to the overall average rank of  $(N+1)/2$ . When the alternative hypothesis is true, the average rank for some of the groups will differ from others, some higher than  $(N+1)/2$  and some lower. The test statistic  $K$  will equal 0 if all groups have identical average ranks and will be positive if average group ranks differ. The null hypothesis is rejected when  $K$  is sufficiently large. Conover (1999) provided tables of exact  $p$ -values for  $K$  for small sample sizes. An example computation of  $K$  is shown in table 7.4. In past years,  $K$  would have been compared to the 0.95 quantile of the chi-squared distribution with  $k-1$  degrees of freedom, which for these data would be 7.815 (3 degrees of freedom). Because  $K$  in table 7.4 does not exceed the 7.815, the null hypothesis is not rejected at an  $\alpha$  of 0.05. Today, software will compute the proportion of the chi-square distribution that equals or exceeds the test statistic value of 2.66. That proportion is the  $p$ -value, here 0.44. Because 0.44 is higher than  $\alpha$ , the null hypothesis is not rejected and the values in each group are not considered to be different.

**Table 7.4.** Kruskal-Wallis test statistic computation for fecal coliform counts (Lin and Evans, 1980).

	Ranks $R_{ij}$						$\bar{R}_j$
<b>Summer</b>	6	12	15	18	21	24	16
<b>Fall</b>	5	8.5	11	14	19.5	22	13.3
<b>Winter</b>	2	4	8.5	13	16	19.5	10.5
<b>Spring</b>	1	3	7	10	17	23	10.2

$$\bar{R}_{ij} = \frac{16 + 13.3 + 10.5 + 10.2}{4} = 12.5$$

$$K = \frac{12}{24(25)} \sum 6(16 - 12.5)^2 + 6(13.3 - 12.5)^2 + 6(10.5 - 12.5)^2 + 6(10.2 - 12.5)^2$$

$$K = 2.66 \quad x_{0.95,(3)}^2 = 7.815 \quad p = 0.44.$$

### 7.1.4 The Large-sample Approximation for the Kruskal-Wallis Test

The distribution of  $K$  when the null hypothesis is true can be approximated quite well at small sample sizes 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 (section 3.2). If all data are divided by their overall mean to standardize the dataset, then when any  $k-1$  average group ranks are known, the final ( $k$ th) average rank can be computed from the others as

$$\bar{R}_k = \frac{N}{n_k} \cdot \left( 1 - \sum_{j=1}^{k-1} \frac{n_j}{N} \bar{R}_j \right). \quad (7.4)$$

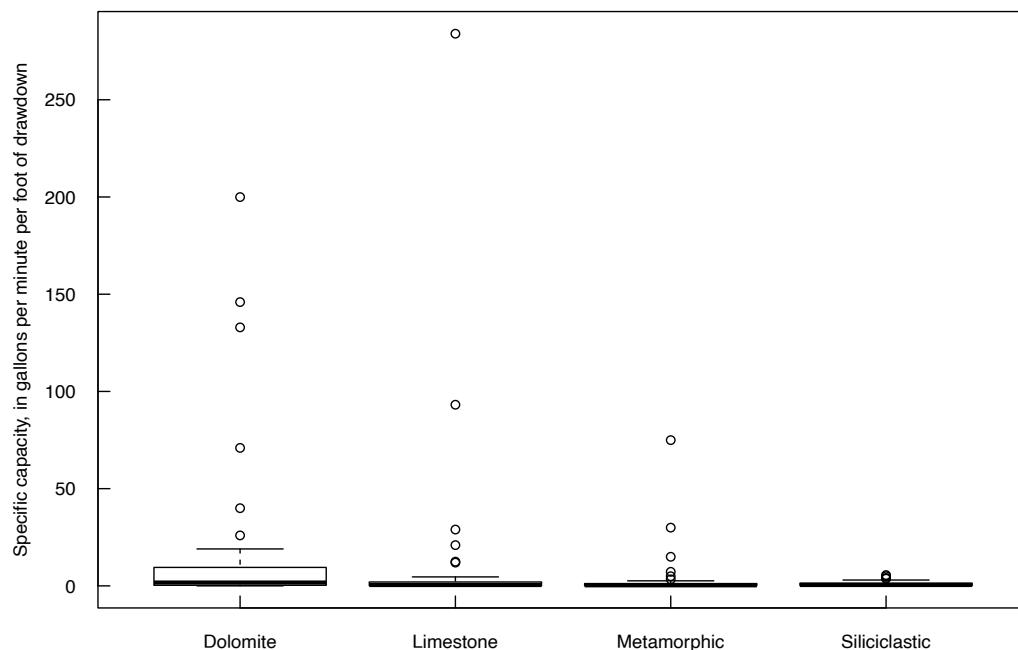
Therefore, there are actually only  $k-1$  independent pieces of information as represented by  $k-1$  average group ranks. From these and the overall average rank, the  $k$ th average rank is fixed. This is the constraint represented by the degrees of freedom.

The null hypothesis is rejected when the approximate  $p$ -value is less than  $\alpha$ . The R command `kruskal.test` computes the large-sample approximation results. Because of the precision of the approximation, most software does not compute an exact test.

#### Example 7.1. Specific capacity—The Kruskal-Wallis test.

Knopman (1990) reported the specific capacity (discharge per unit time per unit drawdown) of wells within the Piedmont and Valley and Ridge Provinces of Pennsylvania. Two hundred measurements from four rock types were selected from the report—see the `specapic.rda` dataset.

Boxplots for the four rock types are shown in figure 7.1. The fact that the boxes are flattened as a result of high outliers is important—the outliers may strongly affect the means and parametric tests, just as they affect your ability to see differences on the figure. Based on the outliers alone it appears that the variance differs among the groups, and all but the siliciclastic group are clearly non-normal. The null hypothesis  $H_0$  for the KW test on these data is that each of the four rock types has the same distribution (set of percentiles) for specific capacity. The alternate hypothesis  $H_A$  is that the distributions are not all the same, with at least one shifted higher than another (a two-sided test).



**Figure 7.1.** Boxplots of specific capacity of wells in four rock types (Knopman, 1990).

To run the Kruskal-Wallis test, use the `kruskal.test` command in R.

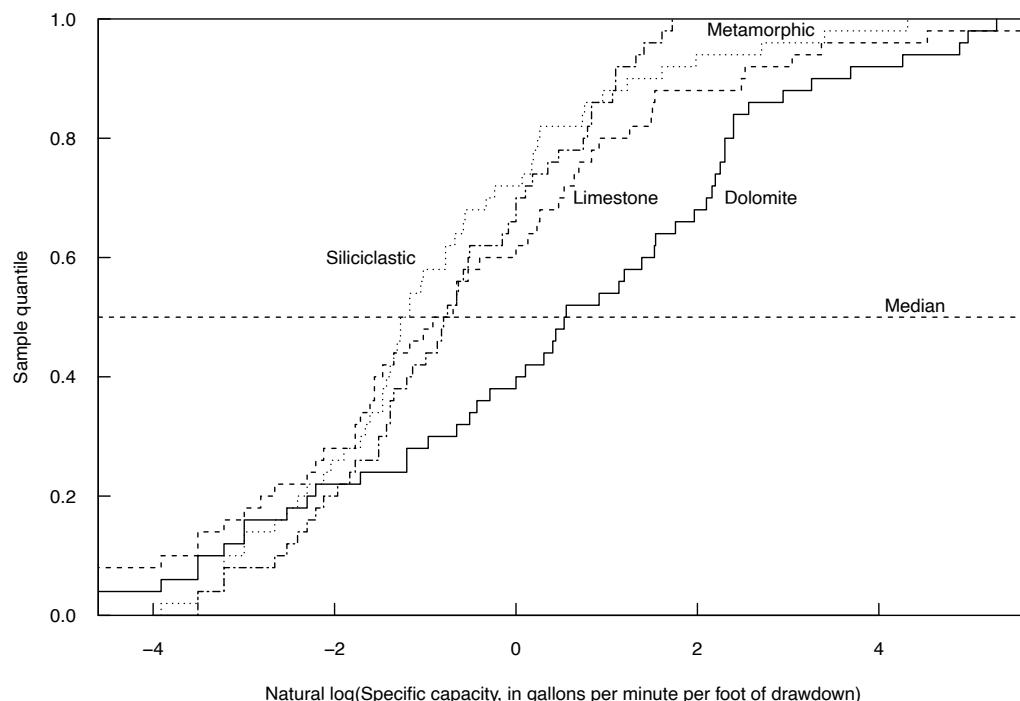
```
> kruskal.test(spcap~rock)
```

```
Kruskal-Wallis rank sum test

data: spcap by rock
Kruskal-Wallis chi-squared = 11.544, df = 3,
p-value = 0.00912
```

The small  $p$ -value leads us to reject the null hypothesis of similarity of group percentiles. At least one group appears to differ in its frequency of high versus low values. A graph that visualizes what the Kruskal-Wallis test is testing for is the quantile plot (fig. 2.4 in chap. 2). A quantile plot for natural logarithms of the four groups of specific capacity data is shown in figure 7.2. The dolomite group stands apart and to the right of the other three groups throughout most of its distribution, illustrating the Kruskal-Wallis conclusion of difference. Moving to the right at  $y=0.5$ , three groups have similar medians but the dolomite group median is higher. An experienced analyst can look for differences in variability and skewness by looking at the slope and shapes of each group's line. Boxplots are more accessible to nontechnical audiences, but quantile plots provide a great deal of detail while still illustrating the main points, especially for technical audiences.

Alternative nonparametric tests, none of which have significant advantages over Kruskal-Wallis, include computing Welch's ANOVA on the ranks of data (Cribbie and others, 2007); the normal-scores test, of which there are two varieties (Conover, 1999); and a one-factor version of the BDM test (Brunner and others, 1997). The Cribbie and others (2007) procedure is similar to a  $t$ -test on ranks and is only an approximate nonparametric test. The normal-scores tests perform similarly to Kruskal-Wallis with a slight



**Figure 7.2.** Quantile plots of the natural log of specific capacity for the four rock types from Knopman (1990). Three rock types have similar medians, but the dolomite group median is higher.

advantage over KW when data are normally distributed—water resources data rarely are. We discuss the two-factor version of the BDM test in section 7.6 as a nonparametric alternative to ANOVA.

## 7.2 Analysis of Variance (One-factor)

Analysis of variance (ANOVA) determines whether the mean of at least one group differs from the means for other groups. If the group means are dissimilar, some of them will differ from the overall mean, as in figure 7.3. If the group means are similar, they will also be similar to the overall mean, as in figure 7.4.

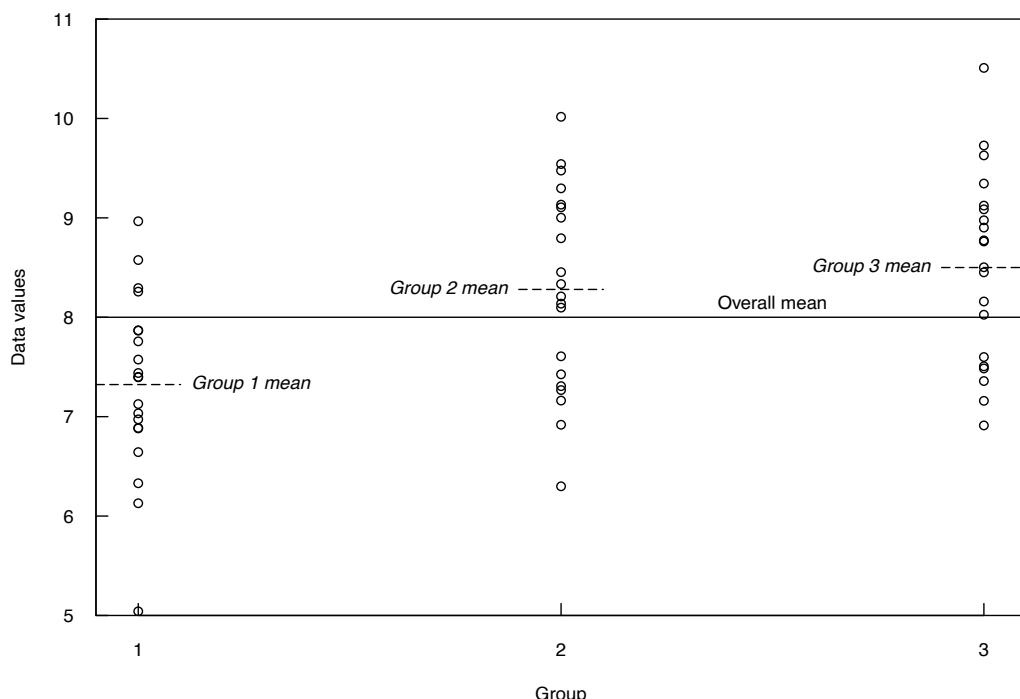
Why should a test of differences between means be named analysis of variance? In order to determine if the differences between group means (the signal) can be seen above the variation within groups (the noise), the total noise in the data as measured by the total sum of squares is split into two parts:

$$\begin{array}{lcl} \text{Total sum of squares} & = & \text{Factor sum of squares} \\ (\text{Overall variation}) & & (\text{Group means--overall mean}) + \text{Residual sum of squares} \\ & & (\text{Variation within groups}) \\ \\ \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 & = & \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \end{array}$$

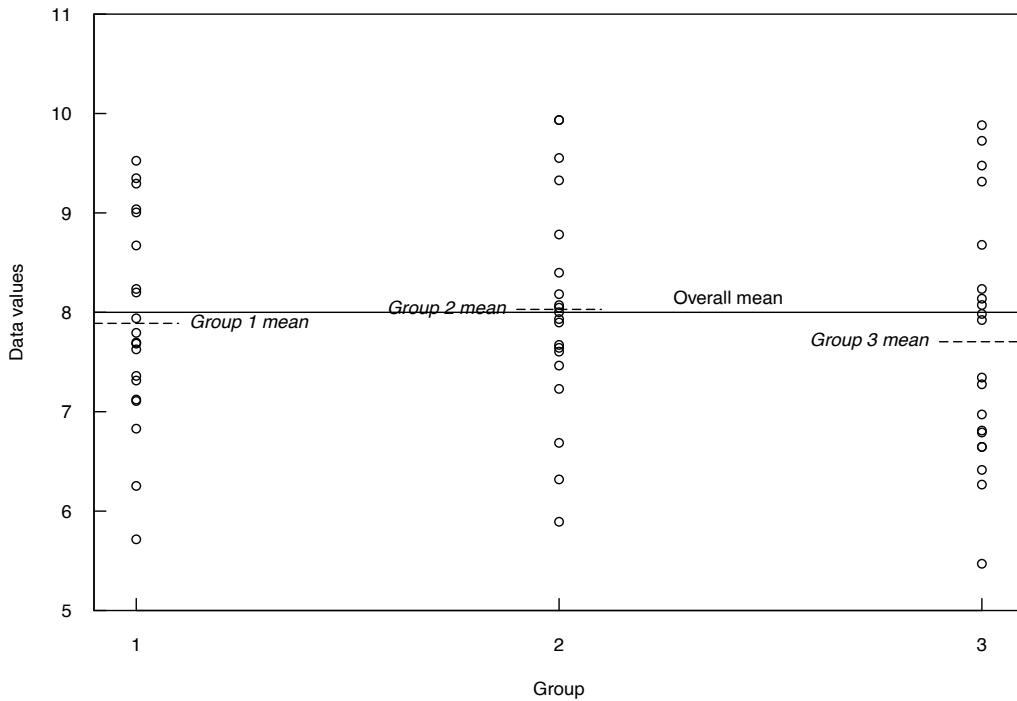
therefore,

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2, \quad (7.5)$$

where  $y_{ij}$  is the  $i$ th observation in the  $j$ th group, there are  $k$  groups, and  $n_j$  designates that sample sizes within the  $j$ th group may or may not be equal to those in other groups.



**Figure 7.3.** Hypothetical data for three groups. Factor mean square > residual mean square, and group means are found to differ.



**Figure 7.4.** Hypothetical data for three groups. Factor mean square  $\equiv$  residual mean square, and group means do not significantly differ.

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 (factor mean square, representing differences between groups) and the other measuring the noise (residual mean square, representing differences within groups). If the signal (factor mean square) is large compared to the noise (residual mean square), the means are found to be significantly different.

### 7.2.1 Null and Alternate Hypotheses for Analysis of Variance

The null and alternate hypotheses for the analysis of variance are

$H_0$ : The group means are identical  $\mu_1 = \mu_2 = \dots = \mu_k$ .

$H_A$ : At least one mean is different.

This is always a two-sided test.

### 7.2.2 Assumptions of the Analysis of Variance Test

ANOVA extends the  $t$ -test to more than two groups. 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,  $\sigma^2$ , estimated for each group by  $s_j^2$ .

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1} \quad (7.6)$$

Violation of either the normality or constant variance assumptions results in a loss of ability to see differences between means (a loss of power). Welch (1951) proposed an adaptation for unequal variance (or heteroscedasticity) in ANOVA analogous to that for the *t*-test, which should become the standard parametric procedure for one-factor ANOVA. We discuss Welch's procedure in section 7.2.4., and strongly prefer it over the classic ANOVA.

The power of ANOVA (classic or Welch's variety) to detect differences will decrease with non-normal data. Scientists often transform data by taking logarithms or a power function and then applying ANOVA. ANOVA on transformed data (using an order-preserving transformation such as logarithms, resulting in data that are approximately symmetric) determines differences in medians on the original scale, not means, which changes the objective. It is also often difficult to find a single transformation which, when applied to all groups, will result in each becoming normal with constant variance. The best alternative for testing means of non-normal data is the permutation test, and it should be the default method (rather than ANOVA) to test differences between group means. If data were normally distributed with equal variances, permutation tests would give very similar results to those of ANOVA, so there is no penalty for routine use of permutation tests. If the objective is to determine if one or more groups have higher or lower values than others, this is a frequency objective, which is tested directly by the Kruskal-Wallis test, not ANOVA. Define your objective and then run the most appropriate test to meet that objective.

### 7.2.3 Computation of Classic ANOVA

Each observation,  $y_{ij}$ , can be written as

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad (7.7)$$

where

- $y_{ij}$  is the  $i$ th individual observation in group  $j$ ,  $j=1, 2, \dots, k$ ;
- $\mu$  is the overall mean (over all groups);
- $\alpha_j$  is the group effect, or  $(\mu_j - \mu)$ , and
- $\epsilon_{ij}$  are the residuals or error within groups.

If  $H_0$  is true, all  $j$  groups have the same mean equal to the overall mean,  $\mu$ , and thus  $\alpha_j=0$  for all  $j$ . If group means differ,  $\alpha_j \neq 0$  for some  $j$ . To detect a difference between means, the variation within a group around its mean must be sufficiently small in comparison to the difference between group means so that the group means may be seen as different (see fig. 7.3). The noise within groups is estimated by the residual or error mean square (MSE), and the signal between group means is estimated by the factor or treatment mean square (MSF). Their computation is shown below.

The residual or error sum of squares

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2, \quad (7.8)$$

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 noise within a group. The factor or treatment sum of squares

$$SSF = \sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2, \quad (7.9)$$

estimates the factor effect using differences between group means,  $\bar{y}_j$ , and the overall mean,  $\bar{y}$ , weighted by sample size.

Each sum of squares has an associated number of degrees of freedom, or the number of independent pieces of information used to calculate the statistic. For the factor sum of squares this equals  $k-1$ , as when  $k-1$  of the group means are known, the  $k$ th group mean can be calculated. The total sum of squares has  $N-1$  degrees of freedom. The residual sum of squares has degrees of freedom equal to the difference between the above two, or  $N-k$ .

Dividing the sums of squares by their degrees of freedom produces variance estimates: the total variance, the MSF, and the MSE. The MSF estimates the variance as a result of any signal between groups plus the residual error variance. If the MSF is similar to the MSE, there isn't much signal and  $H_0$  is not rejected (fig. 7.4). If the MSF is sufficiently larger than the MSE, the null hypothesis will be rejected and at least one group has a mean different from the others (fig. 7.3). The test to compare the two estimates of variance, MSF and MSE, is whether their ratio equals 1:

$$F = \text{MSF} / \text{MSE}.$$

The test statistic  $F$  is compared to quantiles of an  $F$ -distribution and  $H_0$  is rejected for large  $F$ . Equivalently, reject  $H_0$  if the  $p$ -value for the test  $< \alpha$ .

The computations and results of an ANOVA are organized into an ANOVA table. Items usually provided in a one-way ANOVA table are shown in table 7.5. Note that the R `summary` command for analysis of variance does not display the Total row.

### Example 7.2. Specific capacity—Classic ANOVA.

Classic ANOVA is run on the specific capacity data of example 7.1 (Knopman, 1990) to illustrate the effects of its non-normality and unequal variance. There are 50 observations in each of the four groups.

```
> summary(aov(spcap~rock))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rock	3	6476	2158.5	2.512	0.0599 .
Residuals	196	168450	859.4		

The  $F$ -statistic of 2.51 is not significant ( $p=0.0599$ ) at an  $\alpha$  of 0.05. The temptation is to declare that no difference has been found. However, neither of the requirements of normality or equal variance was met. The data analyst should be worried about the effects of failing to meet these assumptions when using ANOVA, even with a dataset of 50 observations per group. Using the preferred Welch adjustment instead would address the issue of unequal variance, but not non-normality.

### 7.2.4 Welch's Adjusted ANOVA

Recognition of the loss of power for ANOVA as a result of heteroscedasticity has slowly made its way into statistics software—for example, Minitab® version 17 chose the Welch (1951) adjustment as its default for one-way ANOVA designs (Frost, 2016). This mirrors standard practice for the  $t$ -test, where Welch's adjustment is the default (see chap. 6). In R, Welch's adjustment is performed with the command `oneway.test`. The adjustment should be the default method for one-way ANOVA with water resources data, because heteroscedasticity is common.

Welch's  $F$ -statistic is computed by weighting each group's contribution to the MSF by  $n/s^2$ , so that groups with greater variability have lower weight. The MSE (or residual mean square) is computed using an adjusted degrees of freedom whose value decreases from the ANOVA residual degrees of freedom as group variances become dissimilar. The resulting  $F$ -test is more accurate for heteroscedastic data. There is little disadvantage to using the Welch adjustment as its correction to the classic ANOVA  $F$ -statistic is negligible when heteroscedasticity is not present.

**Table 7.5.** Schematic of a one-factor ANOVA table.

[df, degrees of freedom;  $k$ , number of groups;  $N$ , number of observations in all groups together; SS, sum of squares; SSF, SS for factor; SSE, SS for error; MS, mean square; MSF, MS for factor; MSE, MS for error; F, F test statistic; -, not applicable]

Source	df	SS	MS	F	p-value
Factor/Treatment	$(k-1)$	SSF	MSF	MSF/MSE	$p$
Residual error	$(N-k)$	SSE	MSE	-	-
Total	$N-1$	Total SS	-	-	-

### Example 7.3. Specific capacity—Welch's adjusted ANOVA.

The nonparametric Fligner-Killeen test of equal variance (section 5.6.1) rejects the null hypothesis that variances are equal, finding a difference between the group variances. Then the Welch adjusted ANOVA is performed using the `oneway.test` command.

```
> fligner.test(spcap, rock)

Fligner-Killeen test of homogeneity of variances
data: spcap and rock
Fligner-Killeen:med chi-squared = 39.458, df = 3, p-value
= 1.388e-08

> oneway.test(spcap~rock)

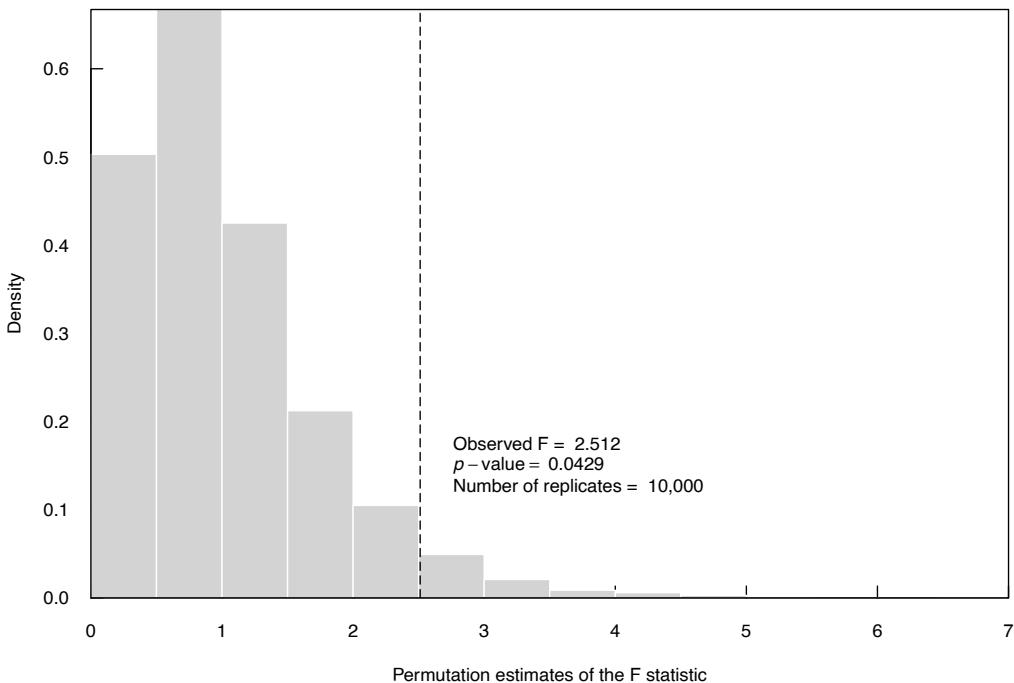
One-way analysis of means (not assuming equal variances)
data: spcap and rock
F = 3.4397, num df = 3.000, denom df = 82.541, p-value =
0.02052
```

Welch's adjustment increases the  $F$ -statistic signal to 3.4 from the classic ANOVA's 2.5 by reducing the residual MS originally inflated by unequal variance. The cost of adjustment is that the denominator (residual) degrees of freedom decreases from 198 to 82. The cost is small compared to the benefit, as the adjusted  $p$ -value of 0.02 is sufficiently lower than the unadjusted  $p$ -value to reject the null hypothesis, finding a difference between group means. This illustrates the power loss for this dataset by using classic ANOVA on data with unequal group variances. Even with the relatively large sample size of 50 observations per group, violation of the two primary assumptions of classical ANOVA can lead to a loss of power.

Water resources and other environmental data are known for their strong skewness, leading to violations of both normality and constant variance. If a test on means is the appropriate objective, use the Welch's adjusted ANOVA to correct for violation of equal variance, or use a permutation test (section 7.3) to avoid the interferences of both unequal variance and non-normality. If the objective is to determine whether at least one group has higher values than another, the Kruskal-Wallis test (section 7.1) addresses that frequency objective directly.

## 7.3 Permutation Test for Difference in Means (One-factor)

Permutation tests allow the means of skewed datasets to be tested without the loss of power inherent in parametric ANOVA procedures. One-factor permutation tests in R are provided by the `permKS` command in the `perm` package (Fay and Shaw, 2010). An R script for a more user-friendly version of the permutation test is `perm1way.R`, found in the supplemental material (SM.7) for this chapter. Permutation tests determine whether group means differ without requiring the assumption of normality and without suffering the same consequences of unequal variance that parametric tests have. Permutation tests require exchangeability (chap. 4)—any observation found in one group could have come from another group because they originate from the same population. This is simply a restatement of the null hypothesis—there is only one population from which all groups are sampled. Permutation tests on means do not require a mathematical adjustment for unequal variance as do parametric tests because they do not use a standard deviation or variance parameter to compute the test statistic. However, if the observed variances are unequal, the larger variance spills over into all groups during the permutation process. In short, permutation tests are less susceptible to loss of power caused by unequal variance than are parametric tests, but the



**Figure 7.5.** Histogram of  $F$ -statistics for 10,000 permutations of the specific capacity group assignments from example 7.1. The vertical dashed line at 2.51 is the  $F$ -statistic of the analysis of variance for the original data.

observed variability of a single group will be spread to all groups. There is really no way around this if the goal is to perform a test of means.

### 7.3.1 Computation of the Permutation Test of Means

Permutation tests compute either all test results possible for rearrangements of the observed data (exact test), or thousands of test results for a large random selection of possible rearrangements. The proportion of computed results equal to, or more extreme than, the one result obtained from the original data is the  $p$ -value of the test. For a one-factor permutation test, the simplest visualization is that the column of group assignments is randomly reordered thousands of times, and the ANOVA  $F$ -test statistic computed for each reordering. By reordering group assignments the number of observations per group stays the same, but the observations assigned to groups differ for each randomization. If the null hypothesis is true, each group has the same data distribution; each has the same mean and variance. Therefore, for the null hypothesis the group assignment is basically random—any observation from one group could have just as easily come from another group. The  $F$ -statistics from the thousands of random group assignments represent the distribution of  $F$ -statistics expected when the null hypothesis is true. This distribution may or may not resemble any specific shape, the data determine the shape of the test statistic distribution. A histogram of 10,000  $F$ -statistics from permutations of the specific capacity data from example 7.1 representing the null hypothesis of no group differences are shown in figure 7.5.

#### Example 7.4. Specific capacity—Permutation test on means.

The permutation test for differences in mean specific capacity is computed using the `perm1way` script. This script also produced figure 7.5.

```
> perm1way(spcap, rock)
```

```
Permutation One-Factor Test
spcap ~ rock
F= 2.51153    Permutation pval= 0.0429    Nrep= 10000
```

The  $p$ -value of 0.0429 states that 4.29 percent of the permutation  $F$ -test statistics equaled or exceeded the original observed  $F$  of 2.51. Therefore, the mean specific capacity is declared significantly different for the four rock types in example 7.1. Running the procedure again will produce a slightly different  $p$ -value, but using 10,000 rearrangements ensures that the variation in  $p$ -values will be small. A permutation test can always be used instead of classic (or Welch's) ANOVA when violations of normality or equal variance assumptions occur. Here the similarity between permutation and Welch's adjustment results indicates that unequal variance was the more important violation of assumptions for this dataset, pushing the classic ANOVA's  $p$ -value above 0.05.

## 7.4 Two-factor Analysis of Variance

Often more than one factor may simultaneously be influencing the magnitudes of observations. Multi-factor tests can evaluate the influence of all factors simultaneously, in a similar way to multiple regression. The influence of one factor can be determined while compensating for the others. This is the objective of a factorial analysis of variance and its nonparametric alternatives.

A factorial ANOVA occurs when none of the factors is a subset of the others. If subset factors do occur, the design includes nested factors and the equations for computing the  $F$ -test will differ from those here; see Aho (2016) for more information on nested ANOVA. We discuss two-factor ANOVA here—more than two factors can be incorporated but that is beyond the scope of this report.

### 7.4.1 Null and Alternate Hypotheses for Two-factor ANOVA

The first page of this chapter presented a two-factor ANOVA, the determination of chemical concentrations among stream 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.

Call the two factors A and B. There are  $i=1$  to  $a \geq 2$  categories of factor A, and  $j=1$  to  $b \geq 2$  categories of factor B. Treatment groups are defined as all the possible combinations of factors A and B, so there are  $a \cdot b$  treatment groups. Within each treatment group there are  $n_{ij}$  observations. The test determines whether mean concentrations are identical among all the  $a \cdot b$  treatment groups, or whether at least one differs.

$H_0$ : All treatment group means,  $\mu_{ij}$ , are equal.  $\mu_{11} = \mu_{12} = \dots = \mu_{ab}$

$H_A$ : At least one  $\mu_{ij}$  differs from the rest.

For the  $k=1, 2, \dots, n_{ij}$  observations in treatment group  $ij$ , the magnitude of any observation,  $y_{ijk}$ , differs from the overall mean,  $\mu$ , by being affected by several possible influences:

$$y_{ijk} = \mu + \gamma_i + \delta_j + \gamma\delta_{ij} + \epsilon_{ijk},$$

where

- $\gamma_i$  is the influence of the  $i$ th category of factor A;
- $\delta_j$  is the influence of the  $j$ th category of factor B;
- $\gamma\delta_{ij}$  is the interaction effect between factors A and B beyond those of  $\gamma_i$  and  $\delta_j$  individually for the  $ij$ th treatment group; and
- $\epsilon_{ijk}$  is the residual error, the difference between the  $k$ th observation ( $k=1, 2, \dots, n_{ij}$ ) and the treatment group mean  $\mu_{ij} = \mu + \gamma_i + \delta_j + \gamma\delta_{ij}$ .

The null hypothesis states that treatment group means  $\mu_{ij}$  all equal the overall mean,  $\mu$ . Therefore  $\gamma_i$ ,  $\delta_j$ , and  $\gamma\delta_{ij}$  all equal 0—there are no effects resulting from either of the factors or from their interaction. If any one of the  $\gamma$ ,  $\delta$ , or  $\gamma\delta$  effects are sufficiently nonzero, the null hypothesis is rejected and at least one treatment group mean significantly differs from the others.

### 7.4.2 Assumptions of Two-factor ANOVA

In two-factor ANOVA, the residuals  $\epsilon_{ij}$  from each treatment group mean  $\mu_{ij}$  (each combination of factors A and B) are assumed to be normally distributed with identical variance  $\sigma^2$ . The normality and

constant variance assumptions could be checked by inspecting separate boxplots of data for each treatment group, but more powerfully by testing the ANOVA residuals from all groups together,  $\epsilon_{ij}$ , using the Shapiro-Wilk test and plotting them on one normal probability plot or boxplot. The effect of violating these assumptions is the same as for one-way ANOVA, a loss of power leading to higher  $p$ -values and failure to find significant differences that are there. No convenient version of a Welch's adjustment exists for two or more factor designs. Brunner and others (1997) state that Welch adjustments to factorial ANOVA are "cumbersome." Instead, permutation tests are the primary method for computing factorial ANOVA on non-normal or heteroscedastic data.

### 7.4.3 Computation of Two-factor ANOVA

The influences of factors A, B, and their interaction are evaluated separately by partitioning the total sums of squares into component parts for each effect. After dividing by their respective degrees of freedom, the mean squares for factors A (MSA), B (MSB), and their interaction (mean square for interaction, MSI) are produced. As with a one-way ANOVA, these are compared to the MSE using  $F$ -tests to determine their significance.

The sums of squares for factor A (SSA), factor B (SSB), interaction (SSI), and error (SSE), assuming constant sample size  $n_{ij} = n$  per treatment group, are presented in table 7.6.

Dividing the sums of squares by their degrees of freedom produces the mean squares MSA, MSB, MSI, and MSE as in table 7.7. If  $H_0$  is true and  $\gamma_i$ ,  $\delta_j$ , and  $\gamma\delta_{ij}$  all equal 0, all variation is simply around the overall mean,  $\mu$ . The MSA, MSB, and MSI will then all approximate the MSE, and all three  $F$ -tests will

**Table 7.6.** Sums of squares definitions for two-factor ANOVA.

[SS, sum of squares; SSA, SS for factor A; SSB, SS for factor B; SSE, SS for error; SSI, SS for interaction]

Sums of squares formula	Effect
$SSA = \sum_a \frac{\left(\sum^b \sum^n y\right)^2}{bn} - \frac{\left(\sum^a \sum^b \sum^n y\right)^2}{abn}$	$\mu_i - \mu$
$SSB = \sum_b \frac{\left(\sum^a \sum^n y\right)^2}{an} - \frac{\left(\sum^a \sum^b \sum^n y\right)^2}{abn}$	$\mu_j - \mu$
$SSI = Total\ SS - SSA - SSB - SSE$	$\mu_{ij} - (\mu_i + \mu_j) + \mu$
$SSE = \sum^a \sum^b \sum^n (y)^2 - \sum^a \sum^b \frac{\left(\sum^n y\right)^2}{n}$	$y_{ijk} - \mu_{ij}$
$Total\ SS = \sum^a \sum^b \sum^n (y)^2 - \frac{\left(\sum^a \sum^b \sum^n y\right)^2}{abn}$	$y_{ijk} - \mu$

**Table 7.7.** Schematic for a two-factor ANOVA table.

[SS, sum of squares; SSA, SS for factor A; SSB, SS for factor B; SSE, SS for error; SSI, SS for interaction; MS, mean square; MSA, MS for factor A; MSB, MS for factor B; MSE, MS for error; MSI, MS for interaction; df, degrees of freedom; F, F-test statistic; -, not applicable]

Source	df	SS	MS	F	p-value
Factor A	( $a-1$ )	SSA	$MSA=SSA/(a-1)$	$F_A=MSA/MSE$	$p[F_A]$
Factor B	( $b-1$ )	SSB	$MSB=SSB/(b-1)$	$F_B=MSB/MSE$	$p[F_B]$
Interaction	( $a-1)(b-1$ )	SSI	$MSI=SSI/(a-1)(b-1)$	$F_I=MSI/MSE$	$p[F_I]$
Error	$ab(n-1)$	SSE	$MSE=SSE/[ab(n-1)]$	-	-
Total	$abn-1$	Total SS	-	-	-

have ratios similar to 1. However, when the alternate hypothesis  $H_A$  is true, at least one of the mean squares in the numerators will be significantly larger than the MSE, and the ratio  $MS_{\text{factor}}/\text{MSE}$  ( $F$ -statistic) will be larger than the appropriate quantile of the  $F$  distribution.  $H_0$  is then rejected and that factor is considered significant at a risk level  $\alpha$ .

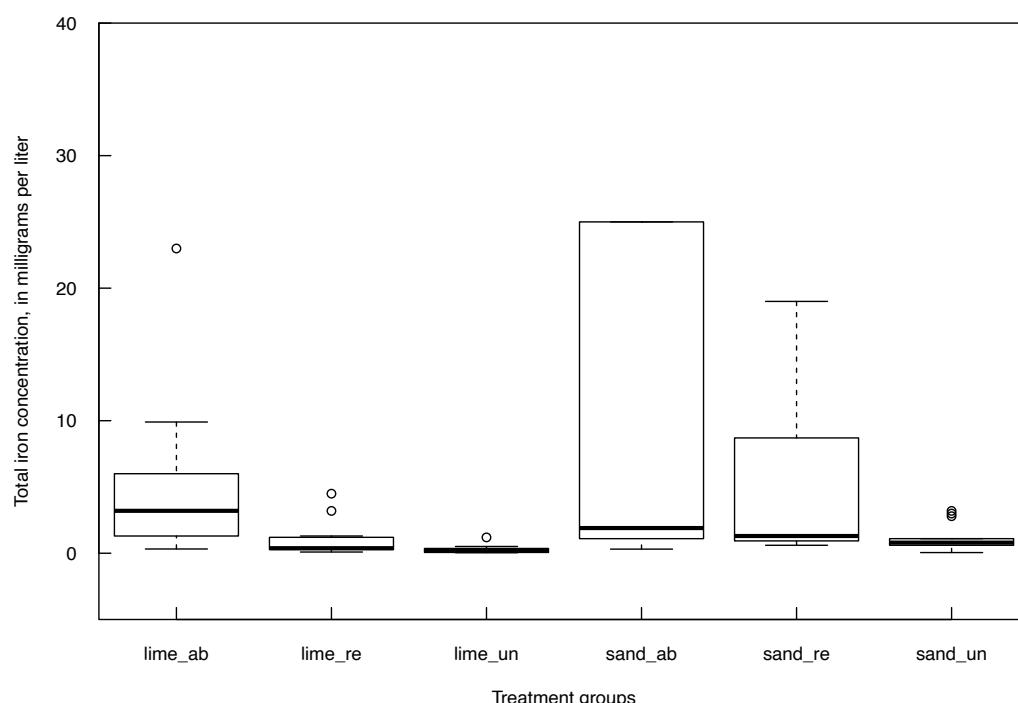
The formulae in the two-factor ANOVA table (table 7.7) are for an equal number of observations in each treatment group (all  $n_{ij} = n$ ). More complex formulae are involved when there are unequal numbers of observations (an unbalanced design). Note that the output for `summary(aov)` in R does not include the Total row at the bottom of table 7.7.

### Example 7.5. Iron at low flows—Two-factor ANOVA.

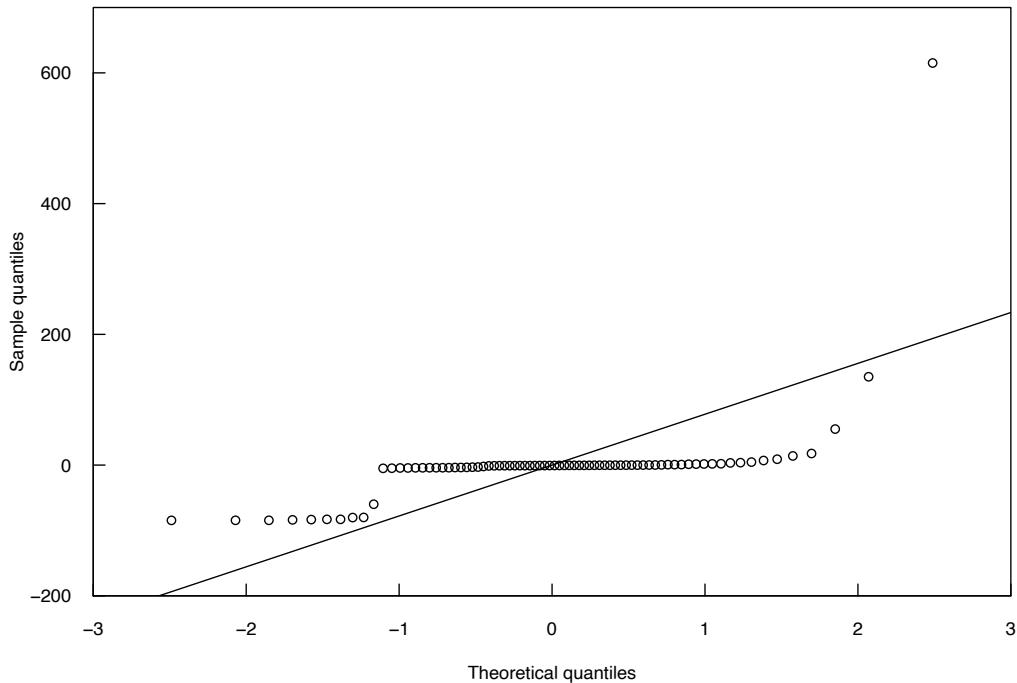
Iron concentrations were measured at low flow in numerous small streams in the coal-producing areas of eastern Ohio (Helsel, 1983). Each stream drains either an unmined area, a reclaimed coal mine, or an abandoned coal mine. Each site is also underlain by either a sandstone or limestone formation. The data are found in `iron.rda`. Are iron concentrations at low flow influenced by upstream mining history, by the underlying rock type, or by both?

Boxplots for total iron concentrations are shown in figure 7.6, where three outliers greater than 100 milligrams per liter in the sandstone, abandoned (sand\_ab) group are not shown. Note the skewness evidenced by the larger upper portions of several boxes. Also note the differences in variance as depicted by differing box heights. This exercise illustrates the problems incurred when parametric ANOVA is applied to data with (commonly occurring) non-normal, heteroscedastic characteristics.

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). Subtracting the group mean from each group's data, the Q-Q plot of residuals clearly shows the three high outliers and several large negative residuals resulting from subtracting the large mean of the abandoned sandstone group from its lower concentration data (fig. 7.7). The pattern is not consistent with a normal distribution. The ANOVA table is shown below. Tests are computed for the factors of mining history alone, rock type alone, and their interaction (`Mining:Rocktype`).



**Figure 7.6.** Boxplots of iron concentrations at low flow from Helsel (1983). Three outliers greater than 100 milligrams per liter are not shown. Lime\_ab, limestone, abandoned mine; lime\_re, limestone, reclaimed; lime\_un, limestone, unmined; sand\_ab, sandstone, abandoned mine; sand\_re, sandstone, reclaimed; sand\_un, sandstone, unmined.



**Figure 7.7.** Q-Q plot showing the non-normality of the ANOVA residuals of the iron data from example 7.5.

```
> fe.aov<-aov(Fe ~ Mining * Rocktype)
> shapiro.test(residuals(fe.aov))
```

```
Shapiro-Wilk normality test
data: residuals(fe.aov)
W = 0.33507, p-value < 2.2e-16

> summary(fe.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mining	2	32282	16141	2.493	0.0898 .
Rocktype	1	15411	15411	2.380	0.1273
Mining:Rocktype	2	25869	12934	1.997	0.1431
Residuals	72	466239	6476		

```
> qqnorm(residuals(fe.aov)) #Figure 7.7
```

```
> fligner.test(residuals(fe.aov)~group)
```

```
Fligner-Killeen test of homogeneity of variances
data: residuals(fe.aov) by group
```

```
Fligner-Killeen:med chi-squared = 23.781, df = 5,
p-value = 0.0002392
```

Neither factor nor the interaction appears 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, shown by the Shapiro-Wilk test and Q-Q plot (fig. 7.7), and of equal variance, shown in the boxplots (fig. 7.6) and by the Fligner-Killeen test, must not be ignored. Perhaps the failure to reject  $H_0$  is due not to a lack of an influence, but to the parametric test's lack of power to detect these influences because of the violation of test assumptions. We will examine that possibility in section 7.4.5. using a permutation test.

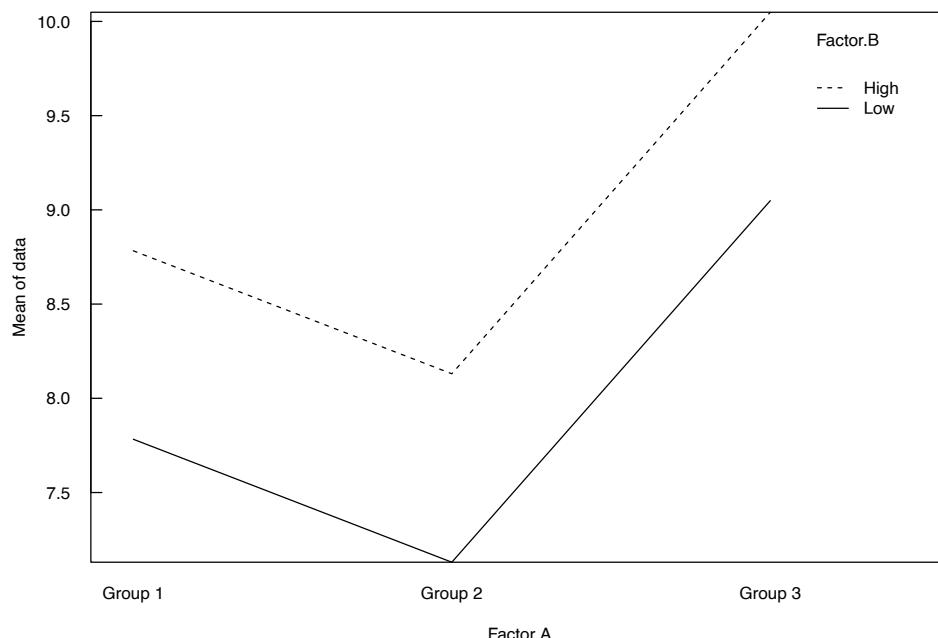
#### 7.4.4 Interaction Effects in Two-factor ANOVA

Interaction is a synergistic or antagonistic change in the mean for a combination of the two factors, beyond what is seen from individual factor effects. 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. Plotting the means of all  $a \cdot b$  groups, with factor A on the  $x$  axis and factor B represented by different connecting lines (an interaction plot—fig. 7.8), the lines are parallel, showing that there is no change in the effect of one factor based on the levels of the second factor and thus there is no interaction.

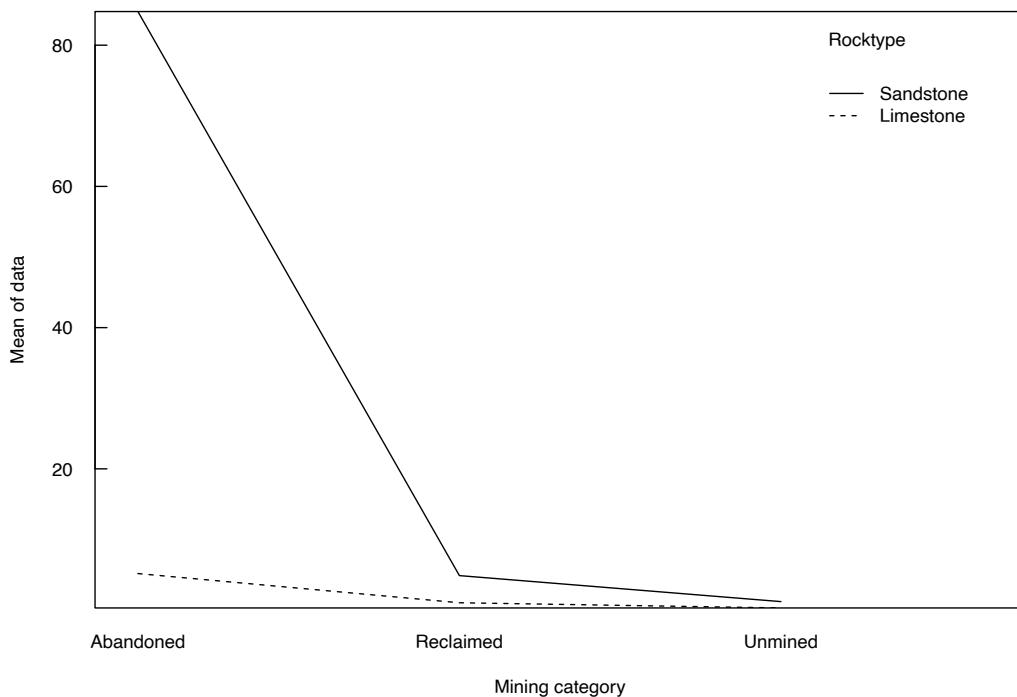
When interaction is present ( $\gamma\delta_{ij} \neq 0$ ), the treatment group means are not determined solely by the additive effects of factors A and B alone. Some of the groups will have mean values larger or smaller than those expected from the individual factors. The effect of factor A can no longer 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—the lines are not parallel. This is the pattern exhibited by the mining history and rock type effects of the example 7.5 data (fig. 7.9). To create the interaction plot in figure 7.9, use the R command

```
> interaction.plot(mining, rocktype, fe)
```

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.



**Figure 7.8.** Interaction plot presenting the means of data in the six treatment groups from example 7.5 showing no interaction between the two factor effects.



**Figure 7.9.** Interaction plot showing interaction by the large nonparallel increase in the mean for the combination of abandoned mining history and sandstone rock type.

#### 7.4.5 Two-factor ANOVA on Logarithms or Other Transformations

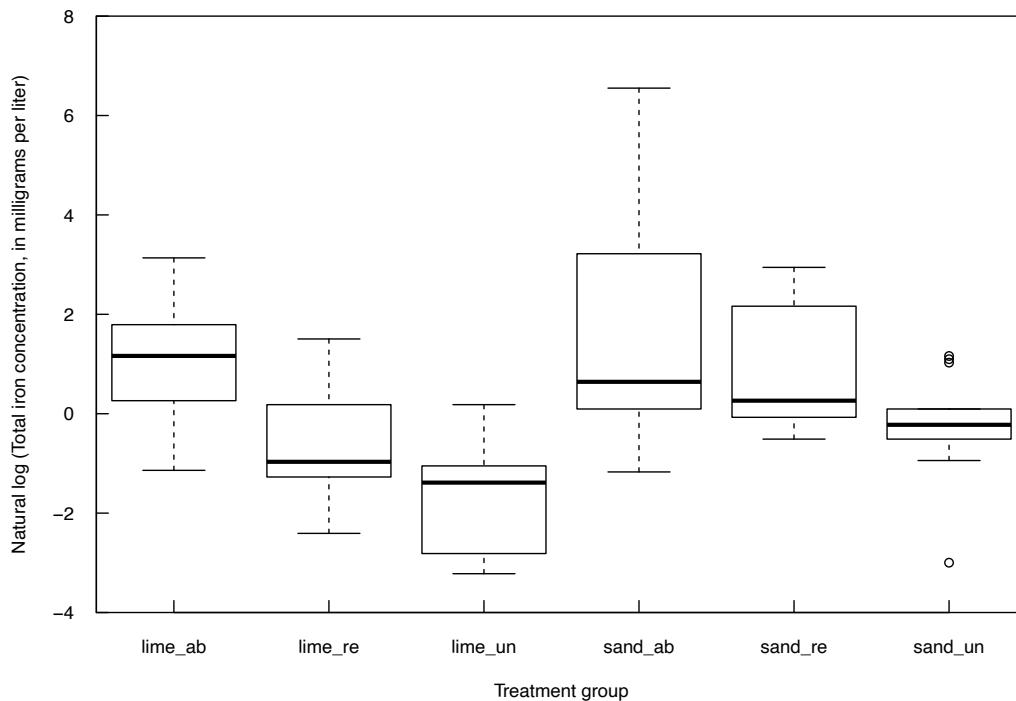
A common approach for analysis of the two-way design with non-normal and heteroscedastic data is to perform ANOVA on data transformed by a power transformation such as the logarithm. The purpose of the power transformation is to produce a more nearly normal and constant-variance dataset. Using logarithms models a multiplicative influence of each factor on the original scale, as the influences in log units are additive. However, an ANOVA on transformed data is no longer a test for differences in means on the original scale. Using logarithms, it is a test for differences in geometric means, an estimator of the median of data on the original scale. Transforming data requires the analyst to understand that the test no longer looks for differences in means. If the means are of interest, then a permutation test is a much better approach than altering the data scale with a transformation. ANOVA on logarithms is often a very good test for typical (median) differences between groups, though the assumptions about residuals for the log-transformed data should always be checked.

Natural logarithms of the low-flow iron concentrations from example 7.5 are shown in figure 7.10. Most of the treatment groups remain distinctly right-skewed even after the transformation, whereas the unmined limestone (lime\_un) group appears less symmetric following transformation! There is nothing magic in the log transformation. Any other transformation going down the ladder of powers (see chap. 1) might, or might not, remedy positive skewness or unequal variance. It may instead alter a symmetric group into one that is left-skewed, as with the lime\_un group here. The result could be that the assumptions of ANOVA are no better met after transformation. If a test on means is desired, use a permutation test instead.

#### Example 7.6. Iron at low flows—Two-factor ANOVA using logarithms.

A two-factor ANOVA on iron concentrations was previously not found to be significant, with obvious non-normality and unequal variance. To deal with these violations of assumptions, the ANOVA is performed on the logarithms (fig. 7.10), testing differences in effects on geometric means instead of means (on the original scale) for each factor.

```
> a2way=aov(log(fe)~mining*rocktype)
> summary(a2way)
```



**Figure 7.10.** Boxplots of the natural logarithms of the iron data from example 7.5. Lime\_ab, limestone, abandoned mine; lime\_re, limestone, reclaimed; lime\_un, limestone, unmined; sand\_ab, sandstone, abandoned mine; sand\_re, sandstone, reclaimed; sand\_un, sandstone, unmined.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mining	2	69.75	34.87	15.891	1.92e-06
rocktype	1	26.31	26.31	11.990	0.000904
mining:rocktype	2	2.44	1.22	0.556	0.575759
Residuals	72	158.01	2.19		

Both mining and rock type factors are significant, showing that some combinations of the two factors demonstrate higher geometric means, even though differences in means were not significant.

#### 7.4.6 Fixed and Random Factors

An additional requirement of the previously given  $F$ -test equations 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 the iron data, differences in chemistry are between the three specific mining histories. In contrast, a random factor would randomly select several groups out of the larger possible set to represent the overall factor. Inferences from the test results would extend beyond the specific groups being tested to the generic factor itself. With random factors there is no interest in attributing test results to a specific individual group, but only in ascertaining a generic effect caused by that factor.

As an example, suppose soil concentrations of a trace metal are to be compared between three particle size fractions statewide to determine which of the three fractions is most appropriate as a reconnaissance medium. Particle size is a fixed effect in this case—there is interest in those three specific sizes. However, there could also be a second, random factor. Suppose that there is only enough funding to sample sparsely if done statewide, so instead a random factor is incorporated to determine whether spatial differences occur. Seven 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 seven 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.

*F*-tests where all factors are random use the mean square for interaction as the denominator rather than the mean square for error. Designs with a mix of random and fixed factors (called a mixed effects design), as in the example above, have a mixture of types of mean squares as denominators. In general, the fixed factors in the design use the interaction mean squares as denominators, and the random factors use 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. A good discussion of a variety of ANOVA designs including mixed-effects models can be found in Aho (2016).

## 7.5 Two-factor Permutation Test

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, permutation (also called randomization) tests can evaluate both factor effects and interactions. Manly (2007) evaluated several proposed methods for randomizing observations or randomizing residuals by subtracting the  $a \cdot b$  group means, producing *p*-values more robust to violations of assumptions than classic ANOVA *F*-tests. The two most versatile of the tests evaluated by Manly (2007) are implemented in the **asbio** package of R (Aho, 2019). These methods randomize the original observations while preserving the number of observations allocated to each factor, generalizing the method used in the **perm1way** script. As with the **perm1way** script, they compute thousands of *F*-statistics after rearranging observations, and the proportion of *F*-statistics equal to or greater than the observed *F*-statistic from ANOVA is the permutation *p*-value for each factor. Tests for interactions as well as the primary factors are produced. Unlike nonparametric tests or ANOVA tests on logarithms, permutation tests can determine whether group means differ as a result of factor effects rather than group percentiles or geometric means. If the factor effects on group means are of interest, the permutation test should be preferred over classical two-factor ANOVA when non-normality or unequal variance appear. For water resources data, that is often the case.

### Example 7.7. Iron at low flows—Two-factor permutation test.

A two-factor permutation test on iron concentrations is performed using the **perm.fact.test** function in the **asbio** package. The current (v. 3.4 default of 100 permutation rearrangements is much too small. Using 5,000 rearrangements only takes a few seconds, yet it provides much greater precision and therefore repeatability of the resulting *p*-values.

```
> require(asbio)
> perm.fact.test(fe, mining, rocktype, perm=5000)
```

\$Table

	Initial.F	Df	pval
X1	2.492636	2	0.0008
X2	2.379906	1	0.0118
X1:X2	1.997428	2	0.0332
Residual	NA	72	NA

X1 is the generic name assigned to the first factor listed in the input command (**mining**). X2 is the generic name assigned to the second factor listed (**rocktype**). The interaction term is listed as X1:X2. Although ANOVA was unable to find significance for either factor or for the interaction, all three tests are significant using the permutation test. This agrees with what was obvious to the eye in figures 7.6 and 7.9, where group boxplots clearly differ for the three mining history categories, and the interaction plot of figure 7.9 exhibited a large difference in mean concentration for sandstone as compared to limestone for the abandoned mine sites. The loss in power resulting from the assumptions required for classic ANOVA is real and common for water resources data, as our field data is usually much more strongly skewed and non-normal than those of more controlled experiments in other disciplines. Permutation tests are vital, therefore, to not miss effects that are actually present in data.

## 7.6 Two-factor Nonparametric Brunner-Dette-Munk (BDM) Test

The BDM test is a nonparametric test for multi-factor ANOVA designs that is particularly adept at dealing with heteroscedastic data. The test determines whether the frequencies of high versus low values have dissimilar patterns attributable to two or more factors. As with other nonparametric tests, no assumptions of distributional shape are required. The test can evaluate one-factor and three-or-more-factor designs, but is applied here to the two-factor layout. It is more powerful than the two-way ANOVA on ranks method of Conover and Iman (1981) that was for many years the nonparametric test most familiar to scientists for evaluating two factors simultaneously.

In their simulation study, Brunner and others (1997) found that when all assumptions of ANOVA held, the BDM test had nearly the same power to detect factor effects as did classic ANOVA. When there is heteroscedasticity or non-normality, BDM has greater power. It works well even when there are small sample sizes in some combinations of the factors. Brunner and others summarize by saying that the test “represents a highly accurate and powerful tool for nonparametric inference in higher-way layouts.”

The null hypothesis of the BDM test is that there are no changes in the cumulative distribution function of data attributable to factor A, factor B, or to an interaction. After some interesting yet computationally simple matrix algebra on ranks (in essence, distribution percentiles), a shift in the distribution function attributable to factor A or B will cause an increase in the  $F$ -test statistic and rejection of the null hypothesis. A shift beyond what is a result of either factor in one or more of the  $a \cdot b$  group cells will cause the test for interaction to be significant. The procedure is quite analogous to ANOVA, with the difference that shifts in group distribution functions are being evaluated rather than shifts in group means. The BDM test is implemented in the `asbio` package of R (Aho, 2019) with command `BDM.2way`.

### Example 7.8. Iron at low flows—Two-factor BDM test.

The BDM test is conducted on the iron dataset from example 7.5.

```
> BDM.2way(fe, mining, rocktype)
```

Two way Brunner-Dette-Munk test				
	df1	df2	F*	P(F > F*)
X1	1.981511	64.36822	17.740921	7.885850e-07
X2	1.000000	64.36822	13.375242	5.152032e-04
X1:X2	1.981511	64.36822	3.709541	3.023646e-02

Again, X1 is the first factor listed in the command (`mining`), and X2 is the second (`rocktype`). The interaction term is listed as X1:X2. Although ANOVA was unable to find significance for either factor or for the interaction, all three tests are significant using the nonparametric procedure.

BDM is a better choice than ANOVA if data violate the normality and constant variance assumptions of ANOVA, and if the objective is to determine whether data values of groups differ from one another. If the objective is specifically to determine whether mean values differ between groups, a permutation test is a better choice than ANOVA when ANOVA assumptions are violated.

## 7.7 Multiple Comparison Tests

In most cases the analyst is interested not only in whether group medians or means differ, but which groups 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 (often, all possible) pairs of treatment group medians or means; there are both parametric and nonparametric MCTs. With all possible comparisons, interest is in the pattern of group medians or means, such as

$$\text{mean(group A)} = \text{mean(group B)} < \text{mean(group C)}.$$

However, if each pair of groups is tested using an  $\alpha_{pairwise}$  of 0.05 (the pairwise error rate), the overall probability of making at least one error, called the overall or family error rate,  $\alpha_{family}$ , will be much higher than 0.05. With comparisons among all  $k$  groups, the number of pairwise comparisons made is  $c=k(k-1)/2$  and the family error rate is

$$\alpha_{family} = 1 - (1 - \alpha_{pairwise})^c . \quad (7.10)$$

For example, when comparing six group means, there are  $(6 \cdot 5) / 2 = 15$  pairwise comparisons. If  $\alpha_{pairwise} = 0.05$  were used for each test, the probability of making at least one error in the pattern is  $\alpha_{family} = 1 - (1 - \alpha_{pairwise})^{15} = 0.54$ . There's about a 50 percent chance that at least one of the comparisons shown in the pattern of the six groups is incorrect. MCTs set the  $\alpha_{family}$  at the desired level such as 0.05, making the  $\alpha_{pairwise}$  much smaller than 0.05. Each individual test must produce a  $p$ -value smaller than  $\alpha_{pairwise}$  in order for the difference to be significant. The much-older Duncan's multiple range, Student-Newman-Keuls (SNK), and Least Significant Difference (LSD) MCTs incorrectly used the pairwise  $\alpha$  for each comparison, increasing the probability of finding false differences in the group pattern. This led to the recommendation that MCTs should be used only after a significant ANOVA or Kruskal-Wallis test was first found. That recommendation is not necessary for MCTs using the family error rate, though it is the typical order of testing even now.

Different MCTs have differing formulae to correct from the family to the pairwise error rate.

The simplest is the Bonferroni correction, where  $\frac{\alpha_{family}}{c} = \alpha_{pairwise}$ . For 6 group means, to achieve a Bonferroni family error rate of 0.05, each of the 15 pairwise group comparisons would need a  $p$ -value below  $(0.05) / 15 = 0.003$  to find a significant difference between each pair of group means. Most MCT software reports the result of pairwise tests after converting the  $p$ -value to the family rate equivalent, for this example by multiplying by 15. For example, if one of the 15 tests comparing 2 of the 6 group means achieved an  $\alpha_{pairwise}$  of 0.01, it would be reported as  $p=0.15$  on the output so that the user could compare it directly to their  $\alpha_{family}$  of 0.05, and know to not reject the null hypothesis.

If prior to testing it is known that only a few pairwise tests are of interest, for example, sites B, C, and D will be compared only to control site A but not to each other, the number of comparisons made ( $c=3$ ) is fewer than all possible, and  $\alpha_{family}$  will be more similar to  $\alpha_{pairwise}$ . These more specific tests are called contrasts.

Aho (2016) and Hochberg and Tamhane (1987) review many types of parametric MCTs. Hollander and Wolfe (1999) discuss nonparametric multiple comparisons. Benjamini and Hochberg (1995) developed a now widely used correction minimizing the false discovery rate rather than the family error rate. First developed for genomics, their approach is applicable to water resources, and environmental studies in general. It is discussed further in section 7.7.2.

### 7.7.1 Parametric Multiple Comparisons for One-factor ANOVA

Parametric MCTs compare treatment group means by computing a least significant range or LSR, the distance between any two means that must be exceeded in order for the two groups to be considered significantly different at a family significance level,  $\alpha_{family}$ . If  $|\bar{y}_1 - \bar{y}_2| > LSR = Q\sqrt{s^2/n}$ , then  $\bar{y}_1$  and  $\bar{y}_2$  are significantly different.

The statistic  $Q$  is analogous to the  $t$ -statistic in a  $t$ -test.  $Q$  depends on the test used (and is some function of either a  $t$ - or studentized range statistic,  $q$ , the error degrees of freedom from the ANOVA, and  $\alpha_{family}$ ). The variance,  $s^2$ , is the mean square for error (residual) from the ANOVA table.

A few MCTs are valid only for the restrictive case of equal sample sizes within each group. In contrast, the Tukey's Honest Significant Difference (HSD), Scheffe, and Bonferroni tests can be used with both equal and unequal group sample sizes. These MCTs compute one least significant range for all pairwise comparisons. The harmonic mean sample size

$$\text{harmonic mean of } n_1 \text{ and } n_2 = \frac{2n_1n_2}{n_1 + n_2} , \quad (7.11)$$

is substituted for  $n$  in the case of unequal group sample sizes. Of these tests, Tukey's has the most power as its correction from family to pairwise error rates is the smallest. Thus, Tukey's has become the standard procedure for parametric MCTs.

All parametric MCTs require the same assumptions as ANOVA—data within each group are normally distributed and have equal variance. Violations of these assumptions will result in a loss of power to detect differences that are actually present.

#### Example 7.9. Specific capacity—Tukey's multiple comparison test.

The specific capacity data from Knopman (1990) were found to be non-normal and unequal in variance. The natural logs  $y=\ln(\text{specific capacity})$  better met these two critical assumptions. Boxplots of the natural logs of the data are shown in figure 7.11. The ANOVA on logarithms found a significant difference between the four rock types ( $p=0.0067$ ). To determine which groups differ from others, Tukey's MCT is now computed.

```
> AnovaSC <-aov(log(spcap) ~ rock)
> summary(AnovaSC)

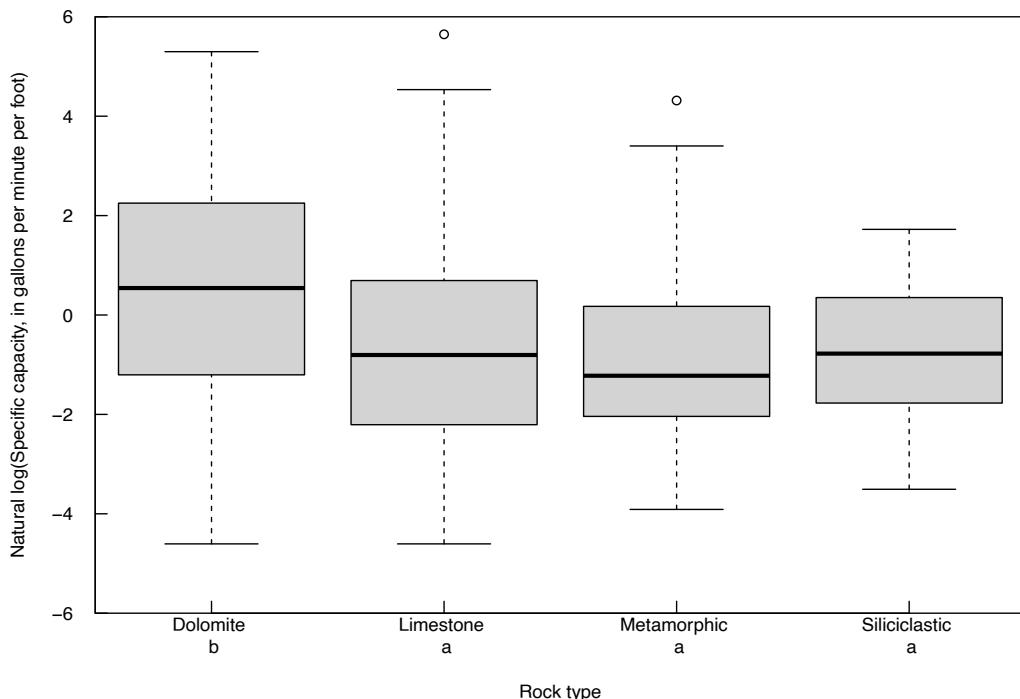
Df Sum Sq Mean Sq F value Pr(>F)
rock          3   54.0  18.010   4.192 0.00667 ***
Residuals    196  842.2   4.297

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = log(specapic$spcap) ~ rock.type)

$rock.type
      diff        lwr        upr      p adj
Lime-Dolo -1.09640600 -2.1706416 -0.02217037 0.0434850
Meta-Dolo -1.30175201 -2.3759876 -0.22751638 0.0104395
Sili-Dolo -1.16632001 -2.2405556 -0.09208439 0.0274778
Meta-Lime -0.20534601 -1.2795816  0.86888962 0.9600526
Sili-Lime -0.06991401 -1.1441496  1.00432161 0.9982893
Sili-Meta  0.13543200 -0.9388036  1.20966762 0.9879283
```

Pairwise  $p$ -values are presented in the  $p \text{ adj}$  column after adjusting them to compare to the  $\alpha_{\text{family}}$  of 0.05. For the six pairwise tests, Tukey's  $p$ -values are  $<0.05$  for differences between dolomite and the other three groups. The other three pairwise group comparisons were not significant. The  $\text{diff}$  column lists the differences in mean(log), where values for the nonsignificant tests are not different from zero. The results of Tukey's test can be summarized using the letters "b" for dolomite and "a" for limestone, metamorphic, and siliciclastic rock types (fig. 7.11).



**Figure 7.11.** Natural logs of specific capacity of wells in four rock types in Pennsylvania. Letters below rock type names designate group. Data from Knopman (1990).

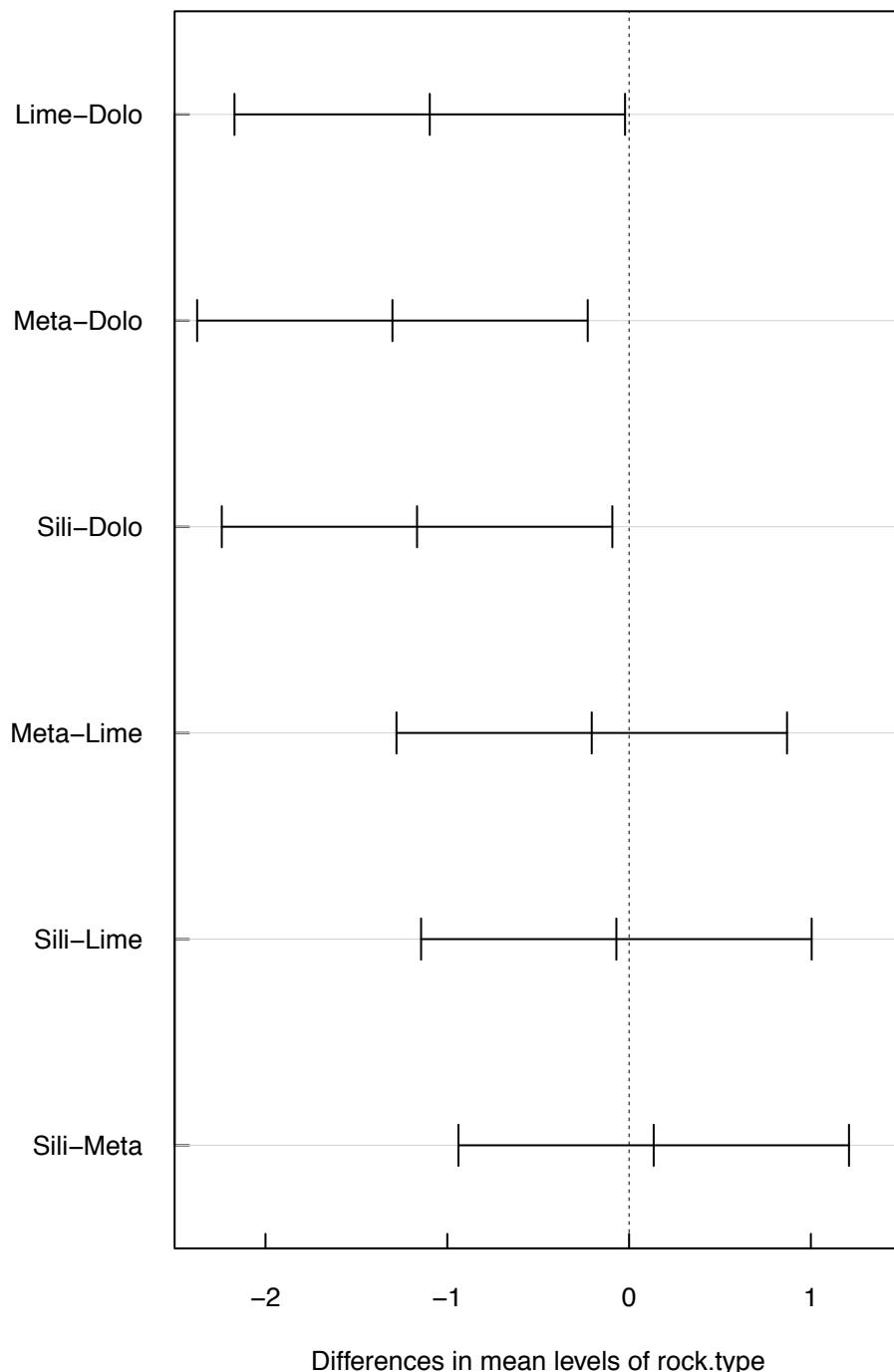
Dolomite has a letter “b” which differs from the other groups to show that dolomite has a mean log significantly different from the other groups. The other group mean logarithms do not significantly differ from one another, and so receive the same letter “a”. Another way to present these results is by plotting the confidence intervals around differences in group means (fig. 7.12). If the confidence intervals include zero there is no significant difference between group means.

### 7.7.2 Nonparametric Multiple Comparisons Following the Kruskal-Wallis Test

A conceptually simple nonparametric MCT to evaluate group patterns following a Kruskal-Wallis test is to compute all possible pairwise Wilcoxon rank-sum tests, setting the tests’ error rates to achieve the family error rate,  $\alpha_{\text{family}}$ , of 0.05. This option is available in the `pairwise.wilcox.test` command of R. This MCT creates separate rankings for each test of group pairs, which occasionally could lead to inconsistent results such as  $A > B$ ,  $B > C$ , but  $A \not> C$ . A strength of using separate ranks is that it allows specific contrasts to be easily computed.

A second common nonparametric MCT is Dunn’s test (Dunn, 1964). This test differs from pairwise rank-sum tests by using one set of joint (all-group) ranks to test each pairwise difference between groups. These are the same ranks used in the Kruskal-Wallis test. Joint ranking avoids the inconsistent test result pattern of the pairwise ranking scheme, and has slightly better power to detect extreme group differences, although the rank-sum approach has slightly better power to detect differences in adjacent groups (Hochberg and Tamhane, 1987). Critchlow and Fligner (1991) list several desirable attributes of MCTs, determining that the separate rankings of pairwise rank-sum tests exhibit more of these attributes than does Dunn’s test. Hochberg and Tamhane (1987) note that “separate rankings are still generally preferred in practice.”

Of great importance is how a family error rate is translated into individual pairwise error rates. Dunn’s test may be the most commonly available nonparametric MCT in software, but it is not the most powerful of the nonparametric MCTs because it uses the Bonferroni correction. The R package `PMCMR` (Pohlert, 2014) computes the original Dunn’s test and a few modifications, including using the BH correction of Benjamini and Hochberg (1995) instead of the Bonferroni correction.



**Figure 7.12.** The 95-percent Tukey family confidence intervals on differences in group means (log scale) of the data from Knopman (1990). Lime, limestone; dolo, dolostone; meta, metamorphic; sili, siliciclastic.

Benjamini and Hochberg (1995) developed a correction that controls the false discovery rate, the expected proportion of false positives among the rejected hypotheses. The false discovery rate is equivalent to the family error rate when all null hypotheses are true but is smaller otherwise. This criterion is less stringent than the family error rate but is perhaps a more logical objective than  $\alpha_{\text{family}}$  because it focuses only on comparisons that significantly differ. The resulting benefit is that the BH correction is more powerful than Bonferroni or other methods of adjusting  $p$ -values.

Here is how the BH correction works. Suppose we set the false discovery rate,  $q^*$ , to 0.05. If there are 6 groups there will be  $c = \frac{6(5)}{2} = 15$  pairwise comparisons. Compute all 15 comparisons and sort them by their  $p$ -values from low to high,  $i=1, 2, \dots, 15$ . Compare each  $p$ -value to the limit  $\frac{i}{c}(q^*)$ , starting at  $i=15$ , the largest  $p$ -value. The largest  $p$ -value is compared to a limit of  $\frac{15}{15}(0.05) = 0.05$ . The second largest  $p$ -value is compared to a limit of  $\frac{14}{15}(0.05) = 0.0467$ , and on down to the smallest  $p$ -value, which is compared to  $\frac{1}{15}(0.05) = 0.0033$ . The first computed  $p$ -value to fall below its limit is considered significant, as are all smaller  $p$ -values. Suppose this is at the  $i=4$ th lowest  $p$ -value of 0.010, which had been compared to  $\frac{4}{15}(0.05) = 0.0133$ . The pairwise comparisons with the four lowest  $p$ -values are therefore found to be significant at the false discovery rate of 0.05. In contrast, the Bonferroni correction would have compared all 15  $p$ -values to a limit of  $\frac{1}{15}(0.05) = 0.0033$ , resulting in fewer significant differences.

Controlling the false discovery rate is a reasonable goal for water resources. We recommend that the false discovery rate (the BH correction in R) should be the prevalent adjustment method for pairwise rank-sum or other nonparametric MCTs.

#### Example 7.10. Specific capacity—Nonparametric multiple comparisons.

The pairwise rank-sum test is computed for the specific capacity data using the BH false-discovery rate  $p$ -value adjustment:

```
> pairwise.wilcox.test(spcap, specapic$rock, p.adjust.method = "BH")
```

Pairwise comparisons using Wilcoxon rank sum test

data: spcap and specapic\$rock

	Dolomite	Limestone	Metamorphic
Limestone	0.043	-	-
Metamorphic	0.013	0.806	-
Siliciclastic	0.013	0.896	0.589

P value adjustment method: BH

The  $p$ -values for each paired comparison in the triangular format output have been adjusted to be comparable to the desired  $\alpha_{\text{family}}$ . For this  $\alpha_{\text{family}}$  of 0.05, medians for dolomite differ from all three of the other rock types because they are lower than  $\alpha_{\text{family}}$ . No other significant differences are observed. Though the output does not provide a letter diagram as did Tukey's MCT, these test results would match those of the Tukey's letter diagram previously given for the mean logarithms.

To compute Dunn's test using R, first load the PMCMR package. Execute the `posthoc.kruskal.dunn.test` command and select the `p.adjust.method="bonferroni"` option to run the 1964 test. Do not omit this option, or it will set  $\alpha$  (default = 0.05) as the uncorrected  $\alpha_{\text{pairwise}}$  instead of the  $\alpha_{\text{family}}$ , an incorrect setting for an MCT.

```
> library(PMCMR)
> posthoc.kruskal.dunn.test(spcap, specapic$rock,
+     p.adjust.method="bonferroni")

Pairwise comparisons using Dunn's-test for multiple
comparisons of independent samples

data: spcap and specapic$rock
```

	Dolomite	Limestone	Metamorphic
Limestone	0.066	-	-
Metamorphic	0.011	1.000	-
Siliciclastic	0.072	1.000	1.000

P value adjustment method: bonferroni

Two-sided  $p$ -values for each pairwise comparison are seen in the triangle of results. The  $p$ -values are adjusted to compare to the  $\alpha_{\text{family}}$ . Medians for dolomite differ from metamorphic at the 5 percent level. No other significant differences are observed. Fewer comparisons are significant than with pairwise rank-sum tests using the BH adjustment. To directly compare the Dunn's test results to pairwise rank-sum tests, use the BH adjustment for Dunn's test as well:

```
> posthoc.kruskal.dunn.test(spcap, specapic$rock, p.adjust.method =
  "BH")
```

```
Pairwise comparisons using Dunn's-test for multiple
comparisons of independent samples
```

data: spcap and specapic\$rock

	Dolomite	Limestone	Metamorphic
Limestone	0.024	-	-
Metamorphic	0.011	0.689	-
Siliciclastic	0.024	0.973	0.689

P value adjustment method: BH

Using the BH false discovery rate provides more power to see pairwise differences than the original Bonferroni correction, as seen by the lower BH  $p$ -values. The pattern of results with the BH adjustment is the same as with the rank-sum MCT and the Tukey's MCT on logarithms. It is the same pattern shown by the quantile plots in figure 7.2. Although there are other nonparametric MCTs within R, the pairwise rank-sum tests with BH correction for family error rate and Dunn's test using the BH correction seem to have the most power over a range of conditions.

### 7.7.3 Parametric Multiple Comparisons for Two-factor ANOVA

Tukey's MCT can be computed for one of two factors in a two-factor ANOVA by first adjusting the data for the other factor. If A is the factor to be tested, the data are first adjusted for factor B by subtracting the mean,  $\delta_j$ , of the  $j$ th category of factor B from each of the  $y$  values.

$$y_{adj_{ijk}} = y_{ijk} - \delta_j , \quad (7.12)$$

where

$\delta_j$  is the influence (mean) of the  $j$ th category of factor B.

Tukey's MCT is then computed on the adjusted  $y$  ( $y_{adj_{ijk}}$ ). The error degrees of freedom of the pairwise one-factor tests are also reduced to that of the two-way ANOVA to acknowledge the presence of factor B.

Tukey's MCT for a two-factor layout requires the same assumptions as the two-factor ANOVA itself—data within each of the  $a \cdot b$  treatment groups are required to be normally distributed and have equal variance. Violations of these assumptions will result in a loss of power to detect differences that are actually present.

#### Example 7.11. Iron at low flows—Tukey's multiple comparisons for two-factor ANOVA.

A two-factor ANOVA on the logarithms was previously computed in example 7.6, finding that both factors (mining and rock type) were significant. Tukey's MCT is now performed for the primary factor of interest, `mining`. To focus on this effect only, the name of the primary factor is listed as input to the `TukeyHSD` function below. The function then (internally) subtracts the mean `ln(fe)` for a rock type from all data of that rock type, to adjust for the effect of rock type. It then performs the MCT on the adjusted values by mining category, producing the results.

```
> TukeyHSD(a2way, "mining")

Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = log(fe) ~ mining * rocktype)
$mining

      diff      lwr      upr
Reclaimed-Abandoned -1.247771 -2.231040 -0.26450281
Unmined-Abandoned   -2.313903 -3.297172 -1.33063505
Unmined-Reclaimed    -1.066132 -2.049401 -0.08286381

      p adj
Reclaimed-Abandoned 0.0092200
Unmined-Abandoned    0.0000010
Unmined-Reclaimed    0.0304423
```

From the  $p$ -values and the direction of differences (diff column), the Tukey tests determine that each of the three mining categories have significantly different geometric means, in the order Abandoned > Reclaimed > Unmined.

The `TukeyHSD` function is part of the `base` R package. There are other Tukey tests for a variety of complicated ANOVA designs available in the `TukeyC` (Faria and others, 2019) package of R.

### 7.7.4 Nonparametric Multiple Comparisons for Two-factor BDM Test

A nonparametric MCT analogous to the Tukey's MCT of the previous section for parametric two-factor ANOVA is to compute all possible pairwise Wilcoxon rank-sum tests for the primary factor after subtracting medians of  $y$  for the secondary factor. Subtraction of medians defined by the second factor adjusts for differences attributable to that factor. This is an adjustment for `rocktype` in the next example.

#### Example 7.12. Iron at low flows—Factorial pairwise rank-sum tests.

Using R, first compute a numeric group indicator for `rocktype`, then compute median iron concentrations for each `rocktype`, and finally subtract the medians from the iron concentrations to produce the adjusted iron concentrations (`feadj`). The relative effects (`Rel.effects`) indicate the relative levels of adjusted iron concentrations for the three mining categories, in the order Abandoned > Reclaimed > Unmined.

```
> Gpind = 1 + as.numeric(rocktype == "sandstone")
> rockmed <- c(median(fe[Gpind == 1]), median(fe[Gpind == 2]))
> feadj = fe - rockmed[Gpind]
> BDM.test(feadj, mining) $Q
```

Levels	Rel.effects
1 Abandoned	0.6760355
2 Reclaimed	0.5140533
3 Unmined	0.3099112

Pairwise Wilcoxon rank-sum tests on the adjusted concentrations using the BH correction for the family error rate show that the three mining categories all significantly differ in their adjusted iron concentrations at  $\alpha_{\text{family}}=0.05$ .

```
> pairwise.wilcox.test(feadj, mining, p.adjust.method = "BH")
```

```
Pairwise comparisons using Wilcoxon rank sum test
data: feadj and mining

Abandoned Reclaimed
Reclaimed 0.01608  -
Unmined    0.00012  0.00511
P value adjustment method: BH
```

## 7.8 Repeated Measures—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 design was that it blocks out the differences from one matched pair (row) to another and so removes unwanted noise. Such matching (or blocking) schemes can be extended to test differences among more

than two groups. One observation is available for each combination of factor (columns) and block (rows) to test for factor effects. This commonly used design is called repeated measures, as well as a randomized complete block design, and in the case of parametric assumptions it is called the two-way ANOVA without replication.

One example at the beginning of this chapter—detecting differences between three sampling or extraction methods used at numerous wells—illustrates this design. The factor tested is the sampling or extraction method, of which there are three types. The blocking effect is the well location; the well-to-well differences are to be blocked out. One sample is analyzed for each sampling or extraction method at each well.

With this design, observations,  $y_{ij}$ , are broken down into four contributions:

$$y_{ij} = \mu + \gamma_j + \delta_i + \epsilon_{ij}, \quad (7.13)$$

where

- $y_{ij}$  is the individual observation in block  $i$  and group  $j$ ;
- $\mu$  is the overall mean or median (over all groups),
- $\gamma_j$  is the  $j$ th group effect,  $j=1, 2, \dots, k$ ;
- $\delta_i$  is the  $i$ th block effect,  $i=1, 2, \dots, n$ ; and
- $\epsilon_{ij}$  is the residual difference between the individual observation and the combined group and block effects.

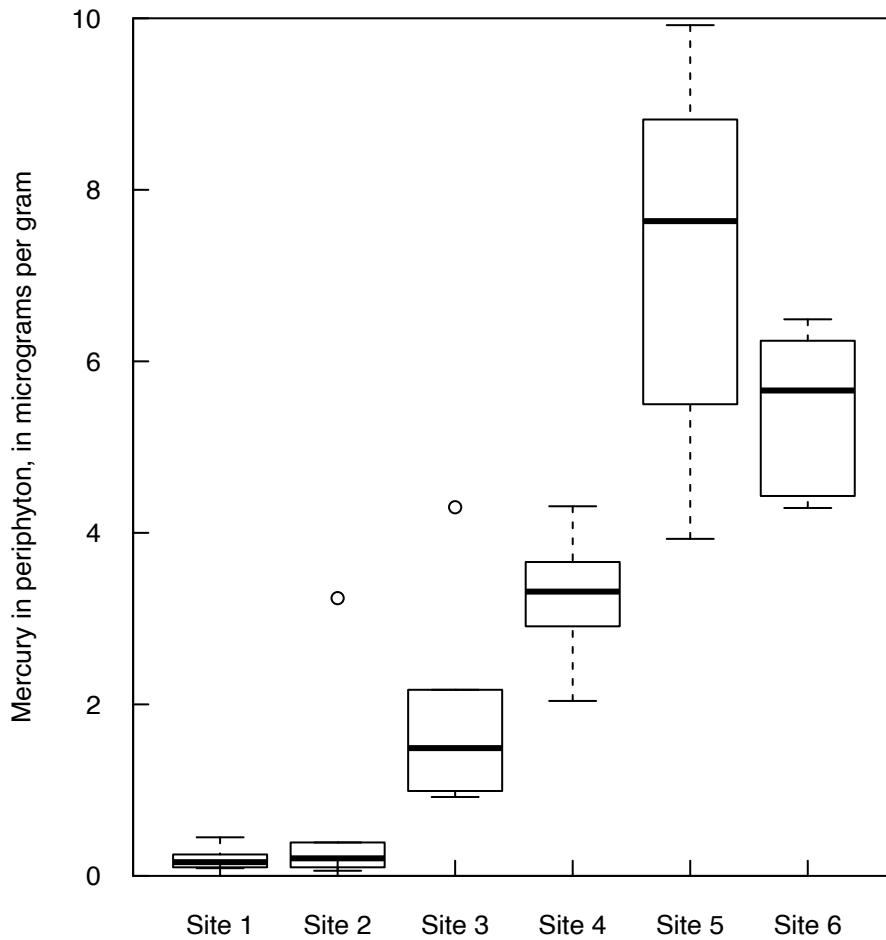
Median polish provides resistant estimates of group and block effects. It is an exploratory technique, not a hypothesis test. Related graphical tools determine whether the two effects are additive or not, and whether the  $\epsilon_{ij}$  are normally distributed, as assumed by an ANOVA. If not, a transformation should be employed to achieve additivity and normality before an ANOVA is performed. The Friedman and median aligned ranks tests (sections 7.8.2. and 7.8.5.) are nonparametric alternatives for testing whether the median factor effect is significant in the presence of blocking.

### 7.8.1 Median Polish

Median polish (Hoaglin and others, 1983) is an iterative process which provides a resistant estimate of the overall median,  $n$ , as well as estimates  $\alpha_j$  of the group effects,  $\alpha_j$ , and estimates  $b_i$  of the block effects,  $\beta_i$ . The usefulness of median polish lies in its resistance to the effects of outliers. The polishing begins by subtracting the medians of each block (shown as the rows in table 7.8) from the data, 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 and set aside. The median of the column medians is subtracted from the column medians and added to the previous estimate of 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  $\alpha_j$  and  $b_i$ ), and a table of residuals,  $e_{ij}$ , estimating the  $\epsilon_{ij}$ .

**Table 7.8.** Mercury concentrations, in micrograms per gram, in periphyton (Walpole and Myers, 1985).

Date	Site					
	1	2	3	4	5	6
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 7.13.** Boxplots showing mercury concentrations in periphyton along the South River, Virginia, from upstream (site 1) to downstream (site 6). Data from Walpole and Myers (1985).

#### Example 7.13. Mercury in periphyton—Median polish.

Mercury concentrations were measured in periphyton at six sites along the South River, Virginia, above and below a large mercury spill (Walpole and Myers, 1985). Measurements were made on six different dates. Of interest is whether the six sites 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. The dates are not randomly selected within each site but are the same for each site. Differences between the six sampling dates are unwanted noise that should be blocked out, hence date is a blocking effect. The data are presented in table 7.8 and boxplots by site in figure 7.13. Median polish will provide an estimate of the magnitude of row (block) and column effects, providing a magnitude of effects for the test results to follow. There appears to be a strong increase in mercury concentration going downstream from site 1 to site 6, reflecting an input of mercury along the way.

The first step in computing 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 7.9). 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 in table 7.10. The median of each column (Site) is then computed and subtracted from that column's data (table 7.11). The residuals from the subtractions remain in the table. Then the median of the column medians ( $-0.16$ ) is subtracted from each of the column medians and added to the overall median. The result is shown in table 7.12.

**Table 7.9.** Data from table 7.8 aligned by subtraction of row medians.

Date	Site						Row median ( $b_i$ )
	1	2	3	4	5	6	
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

**Table 7.10.** Data from table 7.9 after subtraction of the median of row medians.

[-, no data]

Date	Site						Row median ( $b_i$ )
	1	2	3	4	5	6	
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
						Overall median	$m=2.64$

**Table 7.11.** Data from table 7.10 after subtractions of column medians from their respective column's data.

Date	Site						Row effect ( $b_i$ )
	1	2	3	4	5	6	
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
Column effect	-2.38	-2.39	-1.02	0.71	5.30	2.92	

**Table 7.12.** First polish of the periphyton data of Walpole and Myers (1985).

Date	Site						Row effect ( $b_i$ )
	1	2	3	4	5	6	
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
Column effect	-2.22	-2.23	-0.86	0.87	5.46	3.08	$m=2.48$

The first polish of the data from Walpole and Myers (1985) is shown in table 7.12. Two or more polishes are performed in order to produce more stable estimates of the overall median  $m$ , as well as row and column effects. For a second polish, the above process is repeated on the table of residuals from the first polish (table 7.12). Median polish is accomplished in R by

```
> medpolish(Merc)
```

```
1: 31.28
```

```
2: 26.71
```

```
Final: 26.71
```

```
Median Polish Results (Dataset: "Merc")
```

```
Overall: 2.428438
```

Row Effects:

Date1	Date2	Date3	Date4	Date5	Date6
-0.0031250	0.3612500	-0.0946875	0.0031250	0.0528125	-0.2446875

Column Effects:

Site1	Site2	Site3	Site4	Site5	Site6
-2.2075000	-2.2025000	-0.8895313	0.9075000	5.2523438	3.2067187

Residuals:

	Site1	Site2	Site3	Site4	Site5	Site6
Date1	0.23219	3.01719	-0.20578	-1.29281	-3.74766	0.29797
Date2	-0.48219	-0.48719	-0.91016	0.61281	1.87797	0.49359
Date3	0.12375	0.11875	0.20578	-0.11125	-0.19609	-1.11047
Date4	-0.13406	-0.16906	-0.62203	0.32094	0.19609	0.60172
Date5	-0.12375	-0.11875	0.57828	0.11125	1.08641	-0.29797
Date6	0.19375	0.40875	3.00578	-0.18125	-1.93609	-1.10047

The overall, row, and column effects are those after several polishes were computed. The Merc format used is a data matrix, see the commands file in SM.7 to set up data in this format. The median polish shows that

1. The site (column) effects are large in comparison to the date (row) effects.
2. The site effects show a generally increasing pattern going downstream (Site 1 to Site 6), with the maximum at Site 5.
3. A large negative residual ( $-3.75$ ) occurs at Site 5 on Date 1. This is a smaller concentration than expected for this site if the site effect was consistent across all dates.

A boxplot of residuals,  $e_{ij}$ , (fig. 7.14) provides a look at the distribution of errors after the factor and block effects have been removed; the figure shows that the residuals from median polish are relatively symmetric. This is true after subtracting medians, but may not be true when subtracting group means using ANOVA. Median polish is helpful in deciding whether to use an aligned-ranks test (see section 7.8.5.), where symmetry is assumed.

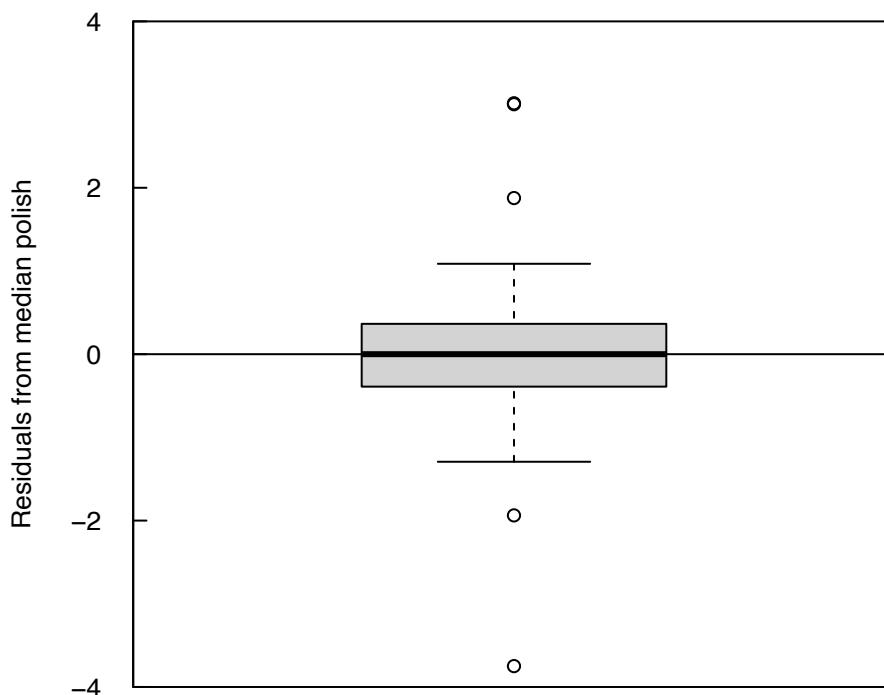
### 7.8.2 The Friedman Test

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 analysis of variance are the same as that of the sign test to the  $t$ -test. When the residuals,  $e_{ij}$ , can be considered normal with equal variance in each group, ANOVA will have more power. For the many situations where the residuals are not normal, the Friedman test will generally have greater power to detect differences between treatment groups and should be performed. The Friedman test is especially useful for ordinal data—data that can be ranked but differences between observations cannot be computed, as when comparing a 1 to a 5.

The Friedman test is used to determine whether

$H_0$ : The median values for all treatment groups are identical.

$H_A$ : At least one treatment group median is significantly different.



**Figure 7.14.** Residuals from the median polish of periphyton mercury data from Walpole and Myers (1985).

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 the associated multiple comparison test presented in section 7.8.4.

### 7.8.3 Computation of the Friedman Test

Understanding the test statistic provides insight into how the Friedman test works. Data are ranked only within each block, not by making any cross-rankings between blocks. With  $k$  treatment groups (columns), rank the data within each of the  $n$  blocks (rows) from 1 to  $k$ , from smallest to largest. If the null hypothesis is true, the ranks within each row will vary randomly with no consistent pattern. Second, sum the ranks for each group (column). When the null hypothesis is true, the average rank for each group will be close to the overall average rank of  $(k+1)/2$ . When the alternative hypothesis is true, the average group rank will differ from one another and from the overall average rank. Third, compute the test statistic  $X_f^*$ , which squares the differences between the average group rank,  $\bar{R}_j$ , and the overall rank to determine if the  $k$  groups differ in magnitude:

$$X_f^* = \frac{12n}{k(k+1)} \sum_{j=1}^k \left[ \bar{R}_j - \frac{k+1}{2} \right]^2. \quad (7.14)$$

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  $\leq 9$ . For larger sample sizes a large-sample approximation is sufficient. When observations are tied within a block, assign the average of their ranks to each.  $X_f^*$  must be corrected using equation 7.15 when ties within a block occur:

$$X_f^* = \frac{12n}{k(k+1) - \frac{1}{n(k-1)} \sum_{i=1}^n \sum_{j=1}^k (t_{ij} (j^3 - j))} \sum_{j=1}^k \left[ \bar{R}_j - \frac{k+1}{2} \right]^2, \quad (7.15)$$

where  $t_{ij}$  equals the number of ties of extent  $j$  in row  $i$ . When ties occur, the large-sample approximation using a chi-squared distribution with  $k-1$  degrees of freedom must be used. R computes the test as `friedman.test`.

#### Example 7.14. Mercury in periphyton—Friedman test.

Does the median mercury concentration in periphyton differ for the six sites along the South River of Virginia (fig. 7.13 and table 7.8)? There are sufficient columns and rows to employ the large-sample approximation, and because ties are present the approximation is required. The `friedman.test` function has the following order of arguments: (Data values, groups, blocks).

```
> friedman.test(Hg, Site.hg, Date.hg)

Friedman rank sum test
data: Hg, Site.hg and Date.hg
Friedman chi-squared = 25.577, df = 5, p-value = 0.0001078
```

The median mercury concentration differs significantly between the six sites.

### 7.8.4 Multiple Comparisons for the Friedman Test

The decision of which groups' data differ from others can be determined using a multiple comparison test. The MCT associated with Friedman's test (Hollander and Wolfe, 1999) controls the family error rate using Bonferroni's adjustment, so it will have less power than previous MCTs using the BH adjustment. The test uses the difference in the mean group rank, rejecting the null hypothesis ( $H_0$ : No difference in mean rank) when differences are larger than expected. As with Dunn's MCT, Friedman ranks are joint ranks,

so values for data in the  $(k-2)$  groups not being compared do affect the computation of each test. For this situation where there is only one observation per cell, an MCT using separate rankings for each pairwise comparison, such as a series of sign or signed-rank tests, would have little power unless there were many rows (blocks). For the mercury data in table 7.8, a sign test between the first and fifth columns would have only six pairs of observations to use. Even though the fifth column has concentrations higher than the first column for all six pairs, the resulting  $p$ -value will not be below 0.05 as a result only of the small sample size. For a two-factor analysis without replication, the joint Friedman ranks provide more information to determine group differences than would separate pairwise rankings.

#### Example 7.15. Mercury in periphyton—Pairwise Friedman comparison test.

The `pairw.fried` function in the `asbio` package (Aho, 2019) scales the reported  $p$ -values to compare to the family error rate, whose default  $\alpha_{\text{family}}$  is 0.05.

```
> pairw.fried(Hg, Site.hg, Date.hg, 6)
```

95% confidence intervals for Friedman's comparisons

	Diff	Lower	Upper	Decision	Adj.	P-value
Site 1-Site 2	-0.66667	-3.83704	2.50371	FTRH0	1	
Site 1-Site 3	-1.83333	-5.00371	1.33704	FTRH0	1	
Site 2-Site 3	-1.16667	-4.33704	2.00371	FTRH0	1	
Site 1-Site 4	-2.33333	-5.50371	0.83704	FTRH0	0.461303	
Site 2-Site 4	-1.66667	-4.83704	1.50371	FTRH0	1	
Site 3-Site 4	-0.5	-3.67038	2.67038	FTRH0	1	
Site 1-Site 5	-4.5	-7.67038	-1.32962	RejectH0	0.000465	
Site 2-Site 5	-3.83333	-7.00371	-0.66296	RejectH0	0.005801	
Site 3-Site 5	-2.66667	-5.83704	0.50371	FTRH0	0.20332	
Site 4-Site 5	-2.16667	-5.33704	1.00371	FTRH0	0.672934	
Site 1-Site 6	-3.66667	-6.83704	-0.49629	RejectH0	0.010307	
Site 2-Site 6	-3	-6.17038	0.17038	FTRH0	0.082178	
Site 3-Site 6	-1.83333	-5.00371	1.33704	FTRH0	1	
Site 4-Site 6	-1.33333	-4.50371	1.83704	FTRH0	1	
Site 5-Site 6	0.83333	-2.33704	4.00371	FTRH0	1	

For the periphyton mercury data from table 7.8, Site 5 differs from Sites 1 and 2, and Site 6 differs from Site 1.

#### 7.8.5 Aligned-ranks Test

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 treatment groups were symmetric. An extension to the signed-rank test for three or more treatment groups is the Aligned-ranks test (ART), one of several possible extensions—Quade's test (Conover, 1999) and Doksum's test (Hollander and Wolfe, 1999) are others. Groggel (1987) and Fawcett and Salter (1984) have shown that an aligned-rank method has substantial advantages in power over other signed rank extensions. For more information on ART methods, see the textbook by Higgins (2003), as well as papers by Mansouri and others (2004), Richter and Payton (2005), and Wobbrock and others (2011). Discussion of the MCT for the two-way design without replication is found in Barefield and Mansouri (2001).

Friedman's test computes within-block ranks, avoiding the confusion produced by block-to-block differences. ART instead allows comparisons across blocks by first subtracting the within-block mean from all of the data within that block. This adds additional information and degrees of freedom to tests, increasing the power over the Friedman within-block only approach. Subtracting the block mean aligns the data across blocks to a common center.

To compute aligned ranks, first subtract the  $i$ th block mean,  $\beta_i$ ,  $i=1, 2, \dots, n$ , from each observation,  $y_{ij}$ :

$$O_{ij} = (y_{ij} - \beta_i) , \quad (7.16)$$

where  $j=1, 2, \dots, k$  is the number of groups. Then the  $O_{ij}$  are jointly ranked from 1 to  $N$ , where  $N=n \cdot k$  is the number of observations, forming aligned ranks,  $AR_{ij}$ . A one-way ANOVA or BDM test is then performed on the  $AR_{ij}$ .

Aligned ranks are equivalent to the ranking of magnitudes of row-to-row differences in the signed ranks test. To derive the benefits of these cross-block comparisons, a cost is incurred. The cost is an assumption that the residuals,  $\epsilon_{ij}$ , from the ANOVA are symmetric. Symmetry can be evaluated by estimating the residuals using median polish, or by computing them in the ANOVA process and plotting them on a boxplot, as in figure 7.14.

The null and alternate hypotheses are identical to those of the Friedman test:

$H_0$ : The median values for all groups are identical.

$H_A$ : At least one group median is significantly different.

Though a one-way ANOVA on the aligned ranks is computed, the correct  $F$ -test will differ from the one determined for the group effect by ANOVA software. Instead, the error degrees of freedom must be  $(k-1) \cdot (n-1)$  because of blocking, not  $k \cdot (n-1)$  as for a one-way ANOVA. Variations of the test are available in the R package *ARTool* (Kay and Wobbrock, 2016).

Richter and Payton (2005) have shown that performing a BDM test on aligned ranks has greater power than using ANOVA on either the original data or on aligned ranks, though to date (2018) it has been less commonly used than ANOVA on aligned ranks.

#### Example 7.16. Mercury in periphyton—Aligned-ranks test.

The periphyton mercury data in table 7.8 are first aligned by subtracting block (row) means. These aligned concentrations (*ac*) are then ranked from 1 to  $N=36$  to form aligned ranks (table 7.13). Tied aligned concentrations are given tied ranks.

```
> ac = Merc
> for (i in 1:length(levels(Date.hg))) {
+   ac[i,] = Merc[i,] - mean(Merc[i,]) }
> alrk = rank(ac) # alrk are shown in Table 7.12
```

**Table 7.13.** Aligned ranks (*alrk*) of the aligned periphyton mercury data from table 7.8.

<b>Date</b>	<b>Site</b>					
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
1	12.0	21.0	14.0	17.0	24.0	32.0
2	1.5	1.5	8.0	23.0	36.0	30.0
3	9.5	9.5	15.0	20.0	33.0	27.0
4	6.0	5.0	13.0	22.0	34.0	31.0
5	3.0	4.0	16.0	19.0	35.0	28.0
6	7.0	11.0	26.0	18.0	29.0	25.0

A one-way analysis of variance is conducted on the aligned ranks. However, the standard ANOVA  $F$ -test from the summary command is incorrect, as its error degrees of freedom are  $n \cdot (k-1)=30$  and it does not correctly reflect the alignment process. The error degrees of freedom for ART should be  $(n-1) \cdot (k-1)=25$  rather than 30. To calculate the correct  $p$ -value, get the probability of equaling or exceeding the test statistic from the  $F$ -distribution (pf function), setting 25 as the degrees of freedom for the MSE:

```
> summary(aov(alrk ~ Site.hg))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Site.hg	5	3222	644.5	29.22	1.12e-10 ***
Residuals	30	662	22.1		

```
> Ftest = (644.5)/(662/25)
> pf(c(Ftest), df1 = 5, df2 = 25, lower.tail = FALSE)
[1] 7.342276e-09
```

The small  $p$ -value of 7.3e-9 shows that  $H_0$  can be strongly rejected, concluding that median mercury concentrations in periphyton differ among the six sites. To avoid the manual computation above, the ARTTool package will use the correct degrees of freedom for each test without requiring the above manual adjustment. ARTTool also computes aligned-rank tests as a nonparametric alternative to two-factor ANOVA when there are replicates in each cell, and for other ANOVA designs.

The BDM test on aligned ranks is an alternative to using ANOVA on aligned ranks. The tests provide similar results for the periphyton mercury data. Both are nonparametric alternatives to the parametric ANOVA of section 7.8.7. Neither aligned-ranks test provides information on which group medians significantly differ from others, that must come from a multiple comparison test.

```
> BDM.test(alrk, Site.hg)
```

```
One way Brunner-Dette-Munk test
df1      df2      F*      P(F > F*)
3.580639 20.0602 29.22015 6.914315e-08
```

## 7.8.6 Multiple Comparisons for the Aligned-ranks Test

Group multiple comparisons following ART take advantage of the data alignment by not requiring paired tests to be performed. A sequence of paired  $t$ -tests or paired Wilcoxon tests would be less powerful because there are few blocks (pairs) with which to conduct those tests. By using aligned-ranks, the block effect is factored out and the relations across blocks can be utilized, increasing degrees of freedom and the power of the test. Tukey's tests on aligned ranks are the natural follow-up to ANOVA on aligned ranks to determine which group levels differ from others.

```
> TukeyHSD(aov(alrk ~ Site.hg))
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = alrk ~ Site.hg)
$Site.hg
          diff        lwr        upr      p adj
Site 2-Site 1  2.166667 -6.0804028 10.413736 0.9654676
```

Site 3-Site 1	8.833333	0.5862639	17.080403	0.0302535
Site 4-Site 1	13.333333	5.0862639	21.580403	0.0003898
Site 5-Site 1	25.333333	17.0862639	33.580403	0.0000000
Site 6-Site 1	22.333333	14.0862639	30.580403	0.0000000
Site 3-Site 2	6.666667	-1.5804028	14.913736	0.1689032
Site 4-Site 2	11.166667	2.9195972	19.413736	0.0034203
Site 5-Site 2	23.166667	14.9195972	31.413736	0.0000000
Site 6-Site 2	20.166667	11.9195972	28.413736	0.0000004
Site 4-Site 3	4.500000	-3.7470695	12.747069	0.5675393
Site 5-Site 3	16.500000	8.2529305	24.747069	0.0000153
Site 6-Site 3	13.500000	5.2529305	21.747069	0.0003289
Site 5-Site 4	12.000000	3.7529305	20.247069	0.0014989
Site 6-Site 4	9.000000	0.7529305	17.247069	0.0261206
Site 6-Site 5	-3.000000	-11.2470695	5.247069	0.8748662

Tukey's tests find that four of the five adjacent site pairings do not differ—Site 4 does differ from Site 5. All nonadjacent sites differ from one another. If instead of ANOVA the BDM test was performed on aligned ranks, the natural follow-up is a series of Wilcoxon rank-sum tests to determine which groups differ from others.

```
> pairwise.wilcox.test(alrk, Site.hg, p.adjust.method = "BH")
```

```
Pairwise comparisons using Wilcoxon rank sum test
data: alrk and Site.hg
```

	Site 1	Site 2	Site 3	Site 4	Site 5
Site 2	0.8095	-	-	-	-
Site 3	0.0130	0.1074	-	-	-
Site 4	0.0046	0.0354	0.0812	-	-
Site 5	0.0046	0.0046	0.0072	0.0046	-
Site 6	0.0046	0.0046	0.0072	0.0046	0.1925

P value adjustment method: BH

The Wilcoxon tests find the same pattern of differences and similarities as did Tukey's test. Both are determining whether medians or cumulative distribution functions of the groups differ, as ranks were taken before computing the tests. The Wilcoxon MCT on aligned ranks has a bit more power than Tukey's MCT on aligned ranks (Barefield and Mansouri, 2001).

### 7.8.7 Two-factor ANOVA Without Replication

The parametric alternative to a Friedman's test for a complete block design is a two-factor ANOVA with only one observation per factor-block combination. The first factor is the effect of interest and the second factor is the block effect. The block effect is of no interest except to remove its masking of the factor effect, so no test for its presence is required. Because there is only one observation per cell, it is impossible to test for an interaction.

The hypotheses are similar to those of the Friedman and ART tests, except that treatment group means, rather than medians, are being tested.

$H_0$ : The treatment group means are identical,  $\mu_1 = \mu_2 = \dots = \mu_k$ .

$H_A$ : At least one mean is significantly different.

The ANOVA model without replication is

$$y_{ij} = \mu + \gamma_j + \delta_i + \epsilon_{ij}, \quad (7.17)$$

where

- $y_{ij}$  is the individual observation in block  $i$  and group  $j$ ;
- $\mu$  is the overall mean;
- $\gamma_j$  is the  $j$ th group effect,  $j=1, 2, \dots, k$ ;
- $\delta_i$  is the  $i$ th block effect,  $i=1, 2, \dots, n$ ; and
- $\epsilon_{ij}$  is the residual between the individual observation and the combined group and block effects.

It is assumed that the residuals,  $\epsilon_{ij}$ , follow a normal distribution. ANOVA does not provide information on which means differ from others; that must come from a multiple comparison test.

### 7.8.8 Computation of Two-factor ANOVA Without Replication

Sums of squares for factor, block, and error are computed using the following formulae (table 7.14). These are divided by their appropriate degrees of freedom to form mean squares.

The factor  $F$ -test is the MSF divided by the MSE, to be compared to quantiles of the  $F$ -distribution for evaluation of its significance. We aren't interested in the block effect, so its test can be ignored. The general structure for a two-factor ANOVA table without replication is found in table 7.15.

Reject the null hypothesis for the factor effect when the  $F$ -test with statistic MSF/MSE has a  $p$ -value less than the desired  $\alpha$ .

**Table 7.14.** Sums of squares definitions for two-factor ANOVA.

[SS, Sum of squares; SSF, SS for factor; SSB, SS for block; SSE, SS for error]

Sums of squares formula	Effect
$SSF = \frac{\sum^k \left[ \sum^n y \right]^2}{n} - \frac{\left[ \sum^k \sum^n y \right]^2}{kn}$	$\mu_j - \mu$
$SSB = \frac{\sum^n \left[ \sum^k y \right]^2}{k} - \frac{\left[ \sum^k \sum^n y \right]^2}{kn}$	$\mu_i - \mu$
$SSE = Total\ SS - SST - SSB$	$y_{ij} - \mu_i - \mu_j + \mu$
$Total\ SS = \sum^k \sum^n y^2 - \frac{\left[ \sum^k \sum^n y \right]^2}{kn}$	$y_{ij} - \mu$

**Table 7.15.** Analysis of variance (ANOVA) table for two factors without replication.

[df, degrees of freedom; SS, sums of squares; SSF, SS for factor; SSB, sum of squares for block; SSE, sum of squares for error; MS, mean square; MSF, mean square for factor; MSE, mean square for error; F, F-test statistic; -, not applicable]

Source	df	SS	MS	F	p-value
Factor/treatment	$k-1$	SSF	$SSF/(k-1)$	MSF/MSE	-
Block	$n-1$	SSB	$SSB/(n-1)$	-	-
Error	$(k-1) \cdot (n-1)$	SSE	$SSE/[(k-1) \cdot (n-1)]$	-	-

### Example 7.17. Mercury in periphyton—ANOVA without replication.

A two-factor ANOVA without replication is calculated directly on the periphyton mercury concentrations from table 7.8. The null hypothesis for the group (`Site.hg`) effect is soundly rejected. Note that no test for interaction is performed as there is only one observation per cell. The residuals are sufficiently normal—their Shapiro-Wilk test null hypothesis was not rejected.

```
> Hg.aov=(aov(Hg~Site.hg+Date.hg))
> summary(Hg.aov)

Df Sum Sq Mean Sq F value    Pr(>F)
Site.hg      5 230.13   46.03   26.14 3.54e-09 ***
Date.hg       5    3.26    0.65    0.37    0.864
Residuals    25  44.02    1.76

> shapiro.test(residuals(Hg.aov))

Shapiro-Wilk normality test
data: residuals(Hg.aov)
W = 0.9469, p-value = 0.08354
```

### 7.8.9 Parametric Multiple Comparisons for ANOVA Without Replication

Pairwise paired *t*-tests will take the blocking structure into account (the blocks form matched pairs) while comparing all pairs of group means. Because no alignment was performed, Tukey's test is not appropriate, as it doesn't take the blocking structure into account. The BH adjustment is used to minimize the false positive error rate.

**Example 7.18. Mercury in periphyton—Pairwise paired *t*-tests.**

```
> pairwise.t.test(Hg, Site.hg, p.adjust.method = "BH", paired=TRUE)
```

```
Pairwise comparisons using paired t tests
data: Hg and Site.hg
```

	Site 1	Site 2	Site 3	Site 4	Site 5
Site 2	0.32796	-	-	-	-
Site 3	0.03132	0.19636	-	-	-
Site 4	0.00227	0.03132	0.11517	-	-
Site 5	0.00253	0.00962	0.01181	0.00358	-
Site 6	0.00061	0.00227	0.01383	0.00608	0.11189

P value adjustment method: BH

Because the residuals follow a normal distribution and variances are not too dissimilar, the pattern of group differences and similarities is essentially the same for pairwise paired *t*-tests as they were for the nonparametric aligned-rank MCTs.

## 7.9 Group Tests for Data with Nondetects

The methods quickly described in this section extend those given for two groups in chapter 5 and are described in far more detail in the textbook by Helsel (2012). The most convenient and powerful procedure is to recensor data so that all observations below the highest detection limit (HDL) are noted as <HDL. No alterations are needed if only one detection limit is present, then any of the nonparametric methods of this chapter can be computed with little loss of information. This method has far more power to detect differences than would substitution followed by a parametric test, as observations small enough to be below detection are certainly nearing the lower bound of zero, resulting in an overall skewed distributional shape. The ranks of all <HDLs will be tied, so software must include tie corrections (as does R) to obtain accurate *p*-values.

For example, table 7.16 presents the mercury concentrations of table 7.8 where concentrations below 0.20 have been censored as <0.20. These could have come from data measured with detection limits of 0.10, 0.15, and 0.20.

**Table 7.16.** Mercury concentrations, in micrograms per liter, in periphyton (Walpole and Myers, 1985), altered to have a detection limit of 0.20.

Date	Site					
	1	2	3	4	5	6
1	0.45	3.24	1.33	2.04	3.93	5.93
2	<0.20	<0.20	0.99	4.31	9.92	6.49
3	0.25	0.25	1.65	3.13	7.39	4.43
4	<0.20	<0.20	0.92	3.66	7.88	6.24
5	<0.20	<0.20	2.17	3.50	8.82	5.39
6	<0.20	0.39	4.30	2.91	5.50	4.29

The Friedman test can be run on these data using any value less than 0.20 to represent each nondetect. This preserves any detected 0.20 values as higher than the nondetects. A suggested practice is to use something like a value of negative one so that this unusual value is not mistaken later for a numerical laboratory measurement.

```
> Hg.nd = Hg  
> Hg.nd[Hg < 0.20] = -1  
> friedman.test(Hg.nd, Site.hg, Date.hg)
```

Friedman rank sum test

```
data: Hg.nd, Site.hg and Date.hg  
Friedman chi-squared = 25.825, df = 5, p-value = 9.648e-05
```

The *p*-value is within 0.00001 of that for the Friedman test on the mercury data without censoring (*p*=0.0001). This demonstrates that the tied ranks obtained from the nondetects made little difference in the overall test result. Substitution followed by ANOVA, on the other hand, results in non-normality resulting from many tied and low concentrations and inaccurate measures of variance that change depending on which number below the detection limit is substituted (Helsel, 2012).

Although better and more powerful tests are available from the field of survival analysis for data with nondetects (Helsel, 2012), simple nonparametric tests for data with one reporting limit, or for data recensored to the highest reporting limit, perform much better than ANOVA and are simple to explain to others.

## Exercises

- Discharge from pulp liquor waste may have contaminated shallow groundwater with caustic, high pH effluent (Robertson and others, 1984). Determine whether the pH of samples taken from three sets of piezometers are identical. One piezometer group is known to be uncontaminated. If not identical, which groups are different from others? Which are contaminated? Be sure to check the normality and equal variance assumptions of ANOVA if using this parametric method.

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

- Feth and others (1964) measured chloride concentrations of spring waters draining three different rock types in the Sierra Nevada—granodiorites, quartz monzonites, and undifferentiated granitic rocks. Determine whether chloride concentrations differ among springs emanating from the three rock types. Check the assumptions of ANOVA before using it. Try the permutation test as well. The data are in `feth.rda`. If differences occur, which rock types differ from the others?
- The number of *Corbicula* (bottom fauna) per square meter for a site on the Tennessee River was presented by Jensen (1973). The data from Jensen's Strata 1 are found in `Corb.rda`. Test the *Corbicula* data to determine whether either season or year are significant factors for the number of organisms observed. If significant effects are found, test for which levels of the factor differ from others.
- Stelzer and others (2012) conducted fecal-indicator quantitative polymerase chain reaction (qPCR) assays to determine if three laboratories were providing similar results. Splits of 15 river samples and 6 fecal-source samples were sent to each of the 3 laboratories. Data for AllBac, a general fecal-indicator assay are found in `allbac.rda`. With the samples as blocks, perform Friedman's test, the ART, and two-way ANOVA without replication to decide whether at least one laboratory provided higher/lower results than the others. Use a multiple comparison test to determine which labs differ from others.

# Chapter 8

## Correlation

---

*Concentrations of atrazine and nitrate in shallow groundwaters are measured in wells over an area of several counties. For each sample, the concentration of atrazine is plotted versus the concentration of nitrate. As atrazine concentrations increase, so do nitrate. How might the strength of this association be measured and summarized? Can nitrate concentrations be used to predict atrazine concentrations?*

*Streams draining the Sierra Nevada in California usually receive less precipitation in November than in other months. Has the amount of November precipitation gradually changed over the past 70 years?*

*Streamflow observations at two streamgages appear to respond similarly to precipitation events over the same period of time. If one streamgage has more observations than the other, can the observations from that streamgage be used to fill in portions of the streamflow record missing from the other streamgage?*

These examples require a measure of the strength of association between two continuous variables. Correlation coefficients are one class of measures that can be used to determine this association. Three correlation coefficients are discussed in this chapter. Also discussed is how the significance of an association can be tested to determine whether the observed pattern differs from what is expected entirely owing to chance. For measurements of correlation between noncontinuous or grouped variables, see chapter 14.

Whenever a correlation coefficient is calculated, the data should first 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 relations can produce differing coefficients depending on the curvature of the relation. Recall that in figure 2.1, eight plots showed the relation between two variables, all with a linear correlation coefficient of 0.70; yet the data were radically different! It is important to never compute correlation coefficients without plotting the data first.

### 8.1 Characteristics of Correlation Coefficients

Correlation coefficients measure 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 covariation between two variables; that is, how one varies by the other. It does not provide evidence for causal relation between the two variables. A change in one variable may cause change in the other, for example, precipitation changes cause runoff changes. Two variables may also be correlated because they share the same cause, for example, changes to concentrations of two constituents measured at a variety of locations are caused by variations in the quantity or source of the water. In trend analysis (chap. 12), correlation is used to measure the change in one variable respective to time. Evidence for causation must come from outside the statistical analysis, through knowledge of the processes involved.

Measures of correlation have the characteristic of being dimensionless and scaled to lie between values of  $-1$  and  $1$ . When there is no correlation between two variables, correlation is equal to zero. When one variable increases as the second increases, correlation is positive. When the variables vary together but in opposite directions, correlation is negative. When using a two-sided test, the following statements about the null,  $H_0$ , and alternative hypotheses,  $H_A$ , are equivalent:

$H_0$ : No correlation exists between  $x$  and  $y$  (correlation = 0), or  $x$  and  $y$  are independent.

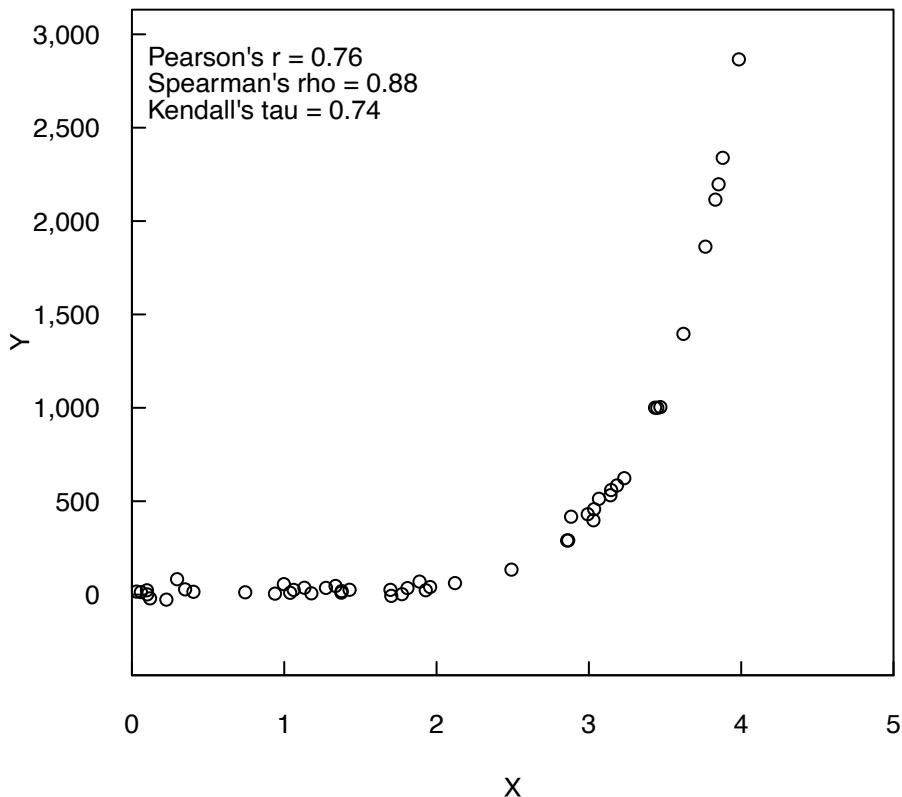
$H_A$ :  $x$  and  $y$  are correlated (correlation  $\neq 0$ ), or  $x$  and  $y$  are dependent.

### 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 defined as possessing a monotonic correlation. This correlation may be nonlinear; for example, when plotted they have exponential patterns, linear patterns, or patterns similar to power functions when both variables are nonnegative. A special case of monotonic correlation is linear correlation, where a plot of  $y$  versus  $x$  has a linear pattern.

Three measures of correlation are in common use—product moment or Pearson's  $r$ , Spearman's rho ( $\rho$ ), and Kendall's tau ( $\tau$ ). The more commonly used Pearson's  $r$  is a measure of linear correlation, whereas Spearman's  $\rho$  and Kendall's  $\tau$  measure monotonic correlation. The last two correlation coefficients are based on ranks and measure monotonic relations such as that in figure 8.1. These two metrics are also resistant to the effects of outliers because they are rank-based. Pearson's  $r$  is only appropriate when plots of  $x$  and  $y$  indicate a linear relation between the two variables, such as shown in figure 8.2. None of the measures are appropriate to assess nonmonotonic relations where the pattern doubles back on itself, like that in figure 8.3.

A monotonic, but not linear, association between two variables is illustrated in figure 8.1. If the Pearson's  $r$  correlation coefficient were calculated to measure the strength of the association between the two variables, the nonlinearity would result in a low value that would not reflect the strong association between the two variables that is apparent in the figure. This illustrates the importance of plotting the data before deciding which correlation measure would appropriately represent the relation between data.



**Figure 8.1.** Plot showing monotonic, but nonlinear, correlation between  $x$  and  $y$ .

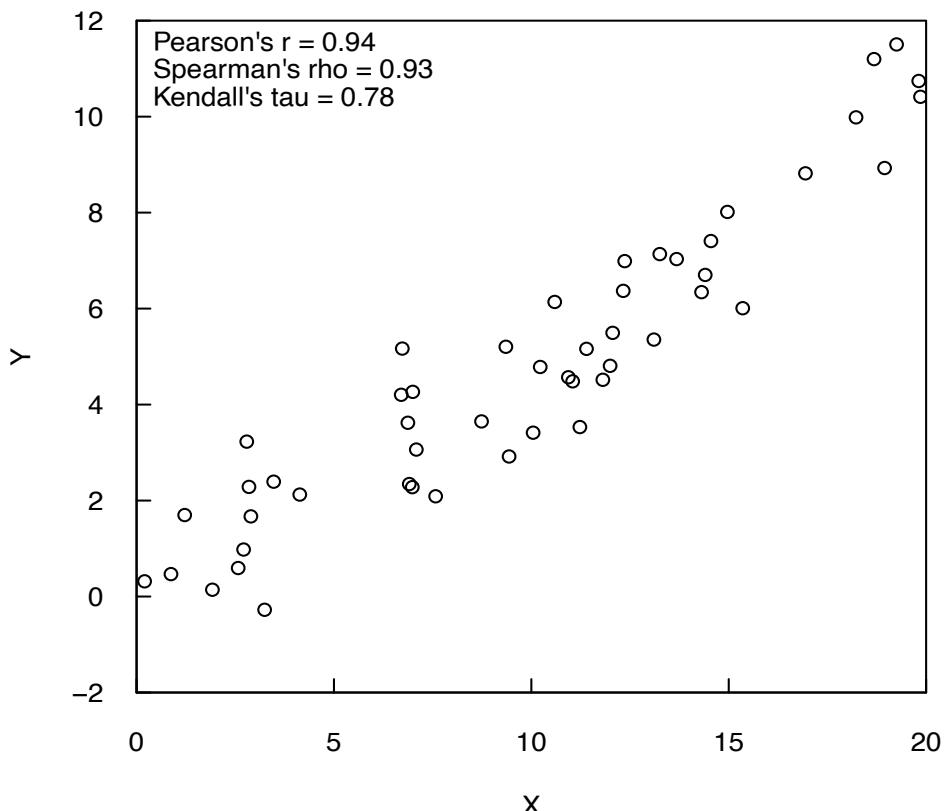


Figure 8.2. Plot showing monotonic linear correlation between  $x$  and  $y$ .

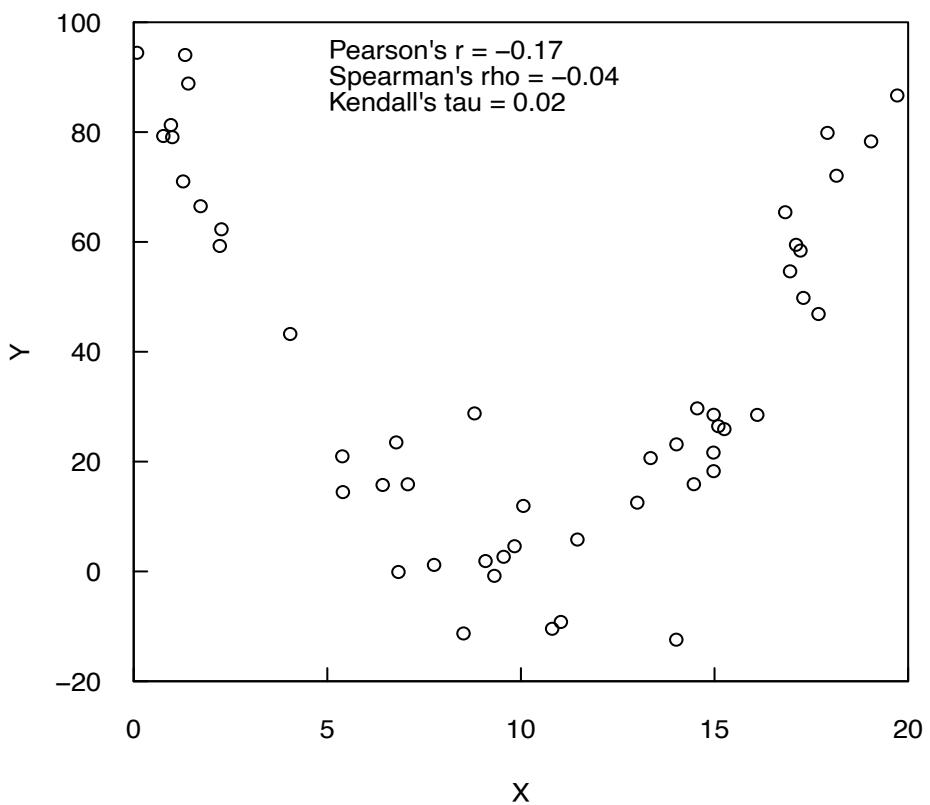


Figure 8.3. Plot showing nonmonotonic relation between  $x$  and  $y$ .

## 8.2 Pearson's $r$

Pearson's  $r$  is the most commonly used measure of correlation and sometimes called the linear correlation coefficient because  $r$  measures the linear association between two variables. If the data lie exactly along a straight line with a positive slope then  $r=1$ ; if the straight line has a negative slope then  $r=-1$ . When considering the use of Pearson's  $r$ , this assumption of linearity makes inspection of a plot even more important for  $r$  than for other correlation metrics, because a small value of Pearson's  $r$  may be the result of curvature or outliers. As in figure 8.1,  $x$  and  $y$  may be strongly related in a nonlinear fashion and the value of the Pearson's  $r$  measure may not be statistically significant.

### 8.2.1 Computation

Pearson's  $r$  is computed from equation 8.1:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right), \quad (8.1)$$

where

- $n$  is the number of observations;
- $\bar{x}$  and  $\bar{y}$  are the means of  $x$  and  $y$ , respectively;
- $s_x$  and  $s_y$  are the standard deviations of  $x$  and  $y$ , respectively; and
- $r$  is a dimensionless value that is not affected by scale changes in the  $x$  and  $y$  observations, for example, converting streamflows in cubic feet per second into cubic meters per second.

This dimensionless property results from dividing by  $s_x$  and  $s_y$ , the sample standard deviations of the  $x$  and  $y$  variables, respectively (eq. 8.1).

### 8.2.2 Hypothesis Tests

Pearson's  $r$  is not resistant to outliers because it is computed by using nonresistant measures—means and standard deviations. Pearson's  $r$  also assumes that the variability in  $y$  cannot increase (or decrease) with increasing  $x$ . In linear regression (chaps. 9 and 11), variables with the property of having constant variability in  $y$  with increasing  $x$  are said to be homoscedastic. Skewed variables often demonstrate outliers and increasing variance; thus  $r$  is often not useful for describing the correlation between skewed hydrologic variables. Transforming the data to reduce skewness and linearize the relation between  $x$  and  $y$  in order to compute Pearson's  $r$  is a common practice that is often used in hydrologic data analysis and is explored further in chapter 9. If these assumptions are met, the statistical significance of  $r$  can be tested under the null hypothesis that  $r$  is not significantly different from zero (that is, there is no correlation) or, in terms of the null and alternate hypothesis,  $H_0: r=0$  or  $H_A: r \neq 0$ . The test statistic  $t_r$  is computed by equation 8.2 and compared to a table of the  $t$ -distribution with  $n-2$  degrees of freedom.

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (8.2)$$

#### Example 8.1. Pearson's $r$

Ten pairs of  $x$  and  $y$  variables are used in this example and shown in order of increasing  $x$ . The data are then plotted in figure 8.4.

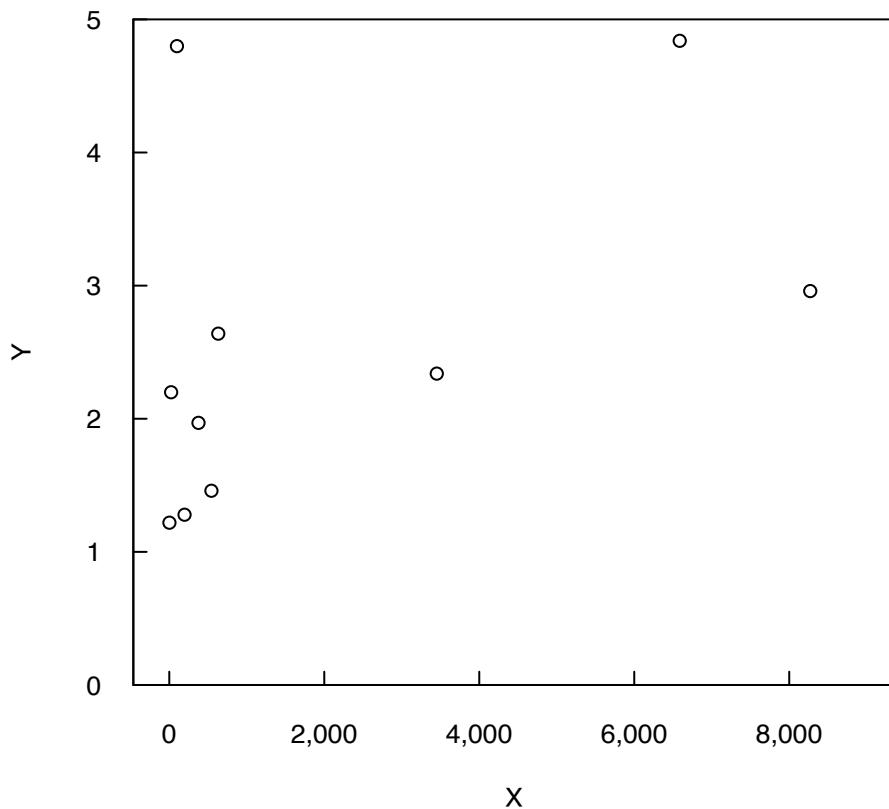
```
> load("example.RData")
> print(example.data)
```

x	y
1	2 1.22

```
2      24 2.20
3      99 4.80
4     197 1.28
5     377 1.97
6     544 1.46
7     632 2.64
8    3452 2.34
9   6587 4.84
10  8271 2.96
```

Compute the means and standard deviations of  $x$  and  $y$ :

```
> mean(example.data$x)
[1] 2018.5
> sd(example.data$x)
[1] 3052.459
> mean(example.data$y)
[1] 2.571
> sd(example.data$y)
[1] 1.314504
```



**Figure 8.4.** Plot of example 8.1 data showing one outlier present (the third value in the dataset (99, 4.80)).

These values can be used with equation 8.1 to compute Pearson's  $r$ ,

$$r = \frac{1}{9} \sum_{i=1}^{50} \left( \frac{x_i - 2018.5}{3052.459} \right) \left( \frac{y_i - 2.571}{1.314504} \right) = 0.457 ,$$

or using the command

```
> cor(example.data$x, example.data$y, method = "pearson")
[1] 0.4578309
```

To test for whether  $r$  is significantly different from zero, and therefore  $y$  is linearly dependent on  $x$ ,

$$t_r = \frac{0.4578309 \sqrt{(10-2)}}{\sqrt{1-(0.4578309)^2}} = 1.456563 ,$$

with a  $p$ -value of 0.18 from a table of the  $t$ -distribution or by using the command

```
> 2 * pt(1.456563, 10-2, lower.tail = FALSE)
[1] 0.1833357
```

where `pt` is the function that returns the area of the  $t$ -distribution that is greater than  $t_r$ , the test statistic value. The value returned by `pt` is doubled because the function `pt` returns the area under the tail for one side of the  $t$ -distribution.

In this example, a two-sided hypothesis test is used because the null hypothesis is simply that the correlation is not equal to zero. The hypothesis is not that the correlation is positive or negative, as that would be a one-sided test. Because this is a two-sided test, the area under both tails of the  $t$ -distribution must be included. In practice, the computation of the Pearson's  $r$  value and its significance can be evaluated using the following command:

```
> cor.test(example.data$x, example.data$y, method = "pearson")
```

#### Pearson's product-moment correlation

```
data: example.data$x and example.data$y
t = 1.4566, df = 8, p-value = 0.1833
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2413747  0.8441271
sample estimates:
cor
0.4578309
```

If our alpha was selected as 0.1,  $H_0: r=0$  is not rejected because the  $p$ -value is  $> 0.1$ , and, therefore we should conclude that  $y$  is not linearly dependent (or related) to  $x$ .

### 8.3 Spearman's Rho ( $\rho$ )

Spearman's  $\rho$  is a nonparametric, rank-based correlation coefficient that depends only on the ranks of the data and not the observations themselves. Therefore,  $\rho$  is resistant to outliers and can be implemented even in cases where some of the data are censored, such as concentrations known only as less than an

analytical detection limit. These properties are important features for applications to water resources. With  $\rho$ , differences between data ranked further apart are given more weight, similar to the signed-rank test discussed in chapter 6;  $\rho$  is perhaps easiest to understand then as the linear correlation coefficient computed on the ranks of the data rather than the data themselves.

### 8.3.1 Computation of Spearman's $\rho$

To compute  $\rho$ , the data for the two variables are separately ranked from smallest to largest. Ties in  $x$  or  $y$  are initially assigned a unique rank. The average rank is then computed from these unique ranks and assigned to each of the tied observations, replacing the unique ranks. Using the ranks of  $x$  and ranks of  $y$ ,  $\rho$  can be computed from the equation

$$\rho = \frac{\sum_{i=1}^n (Rx_i Ry_i) - n\left(\frac{n+1}{2}\right)^2}{n(n^2-1)/12}, \quad (8.3)$$

where  $Rx_i$  is the rank of  $x_i$ ,  $Ry_i$  is the rank of  $y_i$ , and  $(n+1)/2$  is the mean rank of both  $x$  and  $y$ . This equation can be derived from substituting  $Rx_i$  and  $Ry_i$  for  $x_i$  and  $y_i$  in the equation for Pearson's  $r$  (eq. 8.1) 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 paired with 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 8.3. Thus,  $\rho$  will be close to zero.

### 8.3.2 Hypothesis Tests for Spearman's $\rho$

To compute the test statistic,  $S$ , for the significance of the  $\rho$  value, the rank transform method is used. The values for each variable are ranked separately and the Pearson's  $r$  correlation is computed from the ranks. The test statistic,  $S$ , is then given by equation 8.4

$$S = \sum_{i=1}^n (Rx_i - Ry_i)^2, \quad (8.4)$$

where  $Rx_i$  is the rank of  $x_i$ , and  $Ry_i$  is the rank of  $y_i$ . The statistical significance of  $\rho$  can be tested under the null hypothesis that  $\rho$  is not significantly different from zero (that is, there is no correlation) or, in terms of the null and alternate hypothesis,  $H_0: \rho=0$  or  $H_A: \rho \neq 0$ . For large sample sizes ( $n > 20$ ),  $S$  follows a  $t$ -distribution with  $n-2$  degrees of freedom (the same distribution as the Pearson  $r$  test statistic). However, for small sample sizes ( $n < 20$ ), the rank-transformed test statistic does not fit the distribution of the Pearson's  $r$  test statistic well. There has been some work to define the exact probabilities associated with  $\rho$  values for small sample sizes (see Franklin [1988] and Maciak [2009] as examples), and in the example below one such implementation is used in the base R function `cor.test`.

#### Example 8.2. Spearman's $\rho$

Using the same data in example 8.1, rank the  $x$  and  $y$  data separately in ascending order and assign ranks to each observation or use the `rank` function:

```
> Rx<-rank(example.data$x, na.last = NA, ties.method = "average")
> Ry<-rank(example.data$y, na.last = NA, ties.method = "average")
> print(Rx); print(Ry)
[1]  1  2  3  4  5  6  7  8  9 10
[1]  1  5  9  2  4  3  7  6 10  8
```

To solve for  $\rho$ , multiplying the ranks above gives,

```
> Rx * Ry
[1] 1 10 27 8 20 18 49 48 90 80
> sum(Rx * Ry)
[1] 351
```

The sums of  $Rx_i$  multiplied by  $Ry_i$  can then be substituted into equation 8.3 to obtain the value of  $\rho$ :

$$\rho = \frac{351 - 10(5.5)^2}{990/12} = \frac{48.5}{82.5} = 0.588$$

Note that there are fewer than 20 samples in this example, so an exact test should be used in determining the significance of the result. Fortunately, the R function `cor.test` can ensure that the exact  $p$ -value is returned when dealing with small samples by using the arguments `exact = TRUE` and `continuity = TRUE`.

```
> cor.test(example.data$x, example.data$y, method = "spearman",
+           exact = TRUE, continuity = TRUE)
```

Spearman's rank correlation rho

```
data: example.data$x and example.data$y
S = 68, p-value = 0.08022
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.5878788
```

Note that using the large sample test statistic (named `t.distr.teststat` below) and the  $t$ -distribution, the approximate  $p$ -value for the Pearson's  $r$  based on the data ranks has a  $p$ -value = 0.074 and is different from the  $p$ -value obtained using the exact test:

```
> t.distr.teststat <- 0.588 / sqrt((1 - (0.588^2))/(10-2))
> 2 * pt(t.distr.teststat, 10-2, lower.tail = FALSE)
[1] 0.07380343
```

where `pt` is the R function that returns the density function for a  $t$ -distributed value with  $n-2$  degrees of freedom. We can also see this result using the R function `cor.test` with the argument `exact = FALSE`.

```
> cor.test(example.data$x, example.data$y, method = "spearman",
+           exact = FALSE, continuity = TRUE)
```

Spearman's rank correlation rho

```
data: example.data$x and example.data$y
S = 68, p-value = 0.07237
alternative hypothesis: true rho is not equal to 0
```

sample estimates:

```
rho
0.5878788
```

In this example, the argument `continuity = TRUE` is made. The `continuity = TRUE` argument is used with `exact = TRUE` to correct for the fact that the exact test uses a test-statistic distribution that is not continuous (see section 5.1.4. for more information). The authors recommend that this argument be added any time `cor.test` is used to compute the Spearman's  $\rho$  or Kendall's  $\tau$  (section 8.4) correlation. To see the documentation for this argument, type `?cor.test` in the R command window. If the alpha was selected as 0.1,  $H_0: \rho=0$  is rejected because the  $p$ -value is less than 0.1, and therefore we can conclude that  $y$  is monotonically dependent (or related) to  $x$ .

## 8.4 Kendall's Tau ( $\tau$ )

Kendall's  $\tau$  (Kendall, 1938, 1975), much like Spearman's  $\rho$ , measures the strength of the monotonic relation between  $x$  and  $y$  and is a rank-based procedure. Just as with  $\rho$ ,  $\tau$  is resistant to the effect of outliers and, because  $\tau$  also depends only on the ranks of the data and not the observations themselves, it can be implemented even in cases where some of the data are categorical, such as censored observations (for example, observations stated as less than a reporting limit for concentrations or less than a perception threshold for floods). See chapter 14 for more detail on analysis of categorical data. Despite these similar properties,  $\rho$  and  $\tau$  use different scales to measure the same correlation, much like the Celsius 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.

In general,  $\tau$  will be lower than values of  $r$  for linear associations for any given linearly related data (see fig. 8.2). Strong linear correlations of  $r=0.9$  (or above) typically 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. As it is a rank correlation method,  $\tau$  is unaffected by monotonic power transformations of one or both variables. For example,  $\tau$  for the correlation of  $\log(y)$  versus  $\log(x)$  will be identical to that of  $y$  versus  $\log(x)$ , and of  $y$  versus  $x$ .

### 8.4.1 Computation of Kendall's $\tau$

Kendall's  $\tau$  examines every possible pair of data points,  $(x_i, y_i)$  and  $(x_j, y_j)$ , to determine if the pairs have the same relation to one another—that is, if  $x_i$  is greater than  $y_i$  and  $x_j$  is greater than  $y_j$ , or if  $x_i$  is less than  $y_i$  and  $x_j$  is less than  $y_j$ . Each pair is assessed in this way, keeping track of the number of pairs that have the same relation to one another versus the number of pairs that do not.

Kendall's  $\tau$  is most easily computed by ordering all data pairs by increasing  $x$ . If a positive correlation exists, the  $y$  observations will increase more often than decrease as  $x$  increases. For a negative correlation, the  $y$  observations will decrease more often than increase as  $x$  increases. If no correlation exists, the  $y$  observations will increase and decrease about the same number of times. Kendall's  $\tau$  is related to the sign test in that positive differences between data pairs are assigned +1 without regard to the magnitude of those differences and negative differences are assigned -1.

The calculation of  $\tau$  begins with the calculation of Kendall's  $S$ , the test statistic (eq. 8.5). Kendall's  $S$  measures the monotonic dependence of  $y$  on  $x$  and the formula is

$$S = P - M . \quad (8.5)$$

The  $S$  statistic is simply the number of concordant pairs (denoted as  $P$ ) minus the number of discordant pairs (denoted as  $M$ ). We can think of this conveniently as “ $P$  for plus” when the slope between the two points is a positive value. We can think of “ $M$  for minus” when the slope between the two points is a minus value. A concordant pair is a pair of observations where the difference between the  $y$  observations is of the same sign as the difference between the  $x$  observations. A discordant pair is a pair of observations where the difference in the  $y$  observations and the difference in the  $x$  observations is of the opposite sign.

The computation can be simplified by rearranging the data pairs, placing the  $n$  observations in order based on the  $x$  observations with  $x_1$  being the smallest  $x$  to  $x_n$  being the largest  $x$ . After this rearrangement we consider all pairwise comparisons of the  $y$  observations, where the pairs are sorted by their  $x$  rank. If we compare  $(x_i, y_i)$  to  $(x_j, y_j)$  where  $i < j$ , then a concordant pair is the case where  $y_i < y_j$  and a discordant pair is the case where  $y_i > y_j$ .

Note that there are  $n \cdot (n-1)/2$  possible comparisons to be made among the  $n$  data pairs. If all  $y$  observations increased along with the  $x$  observations,  $S = n \cdot (n-1)/2$ . In this situation, the Kendall's  $\tau$  correlation coefficient should equal +1. When all  $y$  observations decrease with increasing  $x$ ,  $S = -n \cdot (n-1)/2$  and Kendall's  $\tau$  should equal -1. Therefore, dividing  $S$  by  $n \cdot (n-1)/2$  will give a value always falling between -1 and +1. This is the definition of Kendall's  $\tau$ , which measures the strength of the monotonic association between two variables

$$\tau = \frac{S}{n(n-1)/2} . \quad (8.6)$$

### 8.4.2 Hypothesis Tests for Kendall's $\tau$

To test for the significance of  $\tau$ ,  $S$  is compared to what would be expected when the null hypothesis is true for a given  $n$ . For a two-sided test,  $H_0: \tau=0$ , or  $H_A: \tau \neq 0$ . If  $\tau$  is further from 0 than expected,  $H_0$  is rejected. When  $n \leq 10$ , the table of exact  $p$ -values using the  $S$  and  $n$  values should be used; this is because the distribution of  $S$  for a given  $n$  at small sample sizes is not easily approximated. Such a table can be found in Hollander and Wolfe (1999).

When  $n > 10$ , a large-sample approximation can be used because the test statistic  $Z_s$  (a rescaled version of  $S$ ) closely approximates a normal distribution with mean,  $\mu_s$ , equal to zero and variance,  $\sigma_s$  (eq. 8.7). For the large-sample approximation,

$$Z_s = \begin{cases} \frac{S-1}{\sigma_s} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma_s} & \text{if } S < 0 \end{cases} \quad (8.7)$$

where  $\sigma_s = \sqrt{(n/18)(n-1)(2n+5)}$ ,  $n$  is the number of samples and  $S$  is defined in equation 8.5. Here the null hypothesis is rejected at significance level  $\alpha$  if  $|Z_s| > Z_{crit}$ , where  $Z_{crit}$  is the value of the standard normal distribution with a probability of exceedance of  $\alpha/2$  (for a two-sided test). Recall that  $\alpha$  is selected by the user and is the probability at which they think the null hypothesis can be rejected.

Just as when computing the test statistic for the rank-sum test (chap. 5), a continuity correction must be applied. This is reflected in equation 8.7 by the -1 or +1. In R, if the `cor.test` function is used, the argument `continuity = TRUE` must be written in the command line to include the continuity correction. The `cor.test` function will give the exact  $p$ -values for  $n < 50$ ; for  $n > 50$ , `cor.test` will give the  $p$ -value resulting from the large-sample approximation unless the argument `exact = TRUE` is specified.

It is worth noting that there is another package in R named `Kendall` (McLeod, 2011), which always includes the continuity correction in the calculation of the  $p$ -value for the test statistic. Additionally, the `Kendall` package will always give the exact  $p$ -value, even for large samples. The `Kendall` package is most useful in the case where some of the  $x$  or  $y$  observations are tied, which requires additional modification of the test statistic and because `cor.test` provides no such adjustment. This is discussed in the next section.

**Example 8.3. Kendall's  $\tau$** 

Recall the 10 pairs of  $x$  and  $y$  in example 8.1, ordered by increasing  $x$ :

```
> load("example.RData")
> print(example.data)

      x     y
1    2 1.22
2   24 2.20
3   99 4.80
4  197 1.28
5  377 1.97
6  544 1.46
7  632 2.64
8 3452 2.34
9 6587 4.84
10 8271 2.96
```

To compute  $S$ , first compare  $y_1 = 1.22$  with all subsequent  $y_i$  values where  $i > 1$ .

$2.20 > 1.22$ , score as +

$4.80 > 1.22$ , score as +

$1.28 > 1.22$ , score as +

$1.97 > 1.22$ , score as + ...continue through until  $i=n$

All subsequent  $y_i$  observations are larger, so there are nine pluses for  $i=1$ .

Move on to the next  $y(i=2)$  and compare  $y_2 = 2.20$  with all subsequent  $y_i$  observations where  $i > 2$ .

$4.80 > 2.20$ , score as +

$1.28 < 2.20$ , score as -

$1.97 < 2.20$ , score as -

$1.46 < 2.20$ , score as - and then continue to next  $i$  until  $i=n$

There are five pluses and three minuses 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 pluses and minuses 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 pluses ( $P=33$ ) and 12 minuses ( $M=12$ ). Therefore, according to equation 8.5,  $S=33 - 12 = 21$ . There are  $10 \cdot 9/2 = 45$  (that is,  $n(n-1)/2$ ) possible comparisons, so  $\tau = 21/45 = 0.47$ . This example could be considered a small sample size and it would therefore be advisable to use the table containing the exact  $p$ -values for a given  $S$  and  $n$ . For  $n=10$  and  $S=21$ , the exact  $p$ -value is  $2 \cdot 0.036 = 0.072$ .

For the purposes of this example, we wish to compare the exact  $p$ -value with the large-sample approximation. We compute the test statistic for the large-sample approximation

$$Z_s = \frac{(21-1)}{\sqrt{\left(\frac{10}{18}\right)(10-1)(20+5)}} = \frac{20}{(11.18)} = 1.79.$$

From a table of the standard normal distributions and a value of 1.79, the lower tail probability is 0.963, so that the  $p \geq 2 \cdot (1 - 0.963) = 0.074$ , slightly larger than the exact  $p$ -value of 0.072. The  $p$ -value is compared to the significance level to determine if the null hypothesis should be rejected. For a significance level of 0.1, the null hypothesis can be rejected.

In R, these calculations are performed using the commands

```
> cor.test(example.data$x, example.data$y, alternative =
+           "two.sided", method = "kendall", continuity = TRUE)
```

Kendall's rank correlation tau

```
data: example.data$x and example.data$y
T = 33, p-value = 0.07255
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.4666667
```

Note that although the argument `exact = TRUE` is not specified, the exact  $p$ -value is returned. This is because `cor.test` will compute an exact  $p$ -value if there are less than 50 paired samples regardless of whether or not this argument is specified. This is not the case for Spearman's  $\rho$ .

### 8.4.3 Correction for Tied Data when Performing Hypothesis Testing Using Kendall's $\tau$

When tied observations of either  $x$  or  $y$  are present in the data, they will produce a 0 rather than + or - when counting the number of  $P$ 's and  $M$ 's. If the ties are not accounted for when determining the significance of  $\tau$ , the variability of  $S$  (represented by  $\sigma_s$ ) will be an overestimate of the actual  $\sigma_s$  and an underestimate of the test statistic. Therefore, an adjustment is needed for  $\sigma_s$  in the equation for the test statistic  $Z_s$  (eq. 8.7) to account for the presence of ties in the data. Details of the adjustment to  $\sigma_s$  can be found in Kendall (1975). This adjustment is only applicable to the large-sample approximation.

There is no exact test for  $\tau$  when ties are present. When the `cor.test` function in R is given small ( $n < 50$ ) datasets with ties and attempts to compute the exact test, it reports an error stating "Cannot compute exact p-value with ties". It then computes the large-sample approximation test and  $p$ -value. This is similar to other nonparametric tests; the large-sample approximation test is the only available method and so the error message is actually just an informational message.

## Exercise

1. The Ogallala aquifer was investigated to estimate relations between uranium and other concentrations in its waters. Below are the concentrations of uranium and total dissolved solids. This set of data is also available in the supplementary material for chapter 8 (SM.8) and is called `urantds2.RData`.

Total dissolved solids, in milligrams per liter	Uranium, in parts per billion	Total dissolved solids, in milligrams per liter	Uranium, in parts per billion
682.65	0.93	1,116.59	7.24
819.12	1.94	301.20	5.72
303.76	0.29	265.45	4.74
1,151.40	11.90	295.88	2.81
582.42	1.57	442.36	5.63
1,043.39	2.06	342.71	3.09
634.84	3.89	361.30	3.58
1,087.25	0.98	262.07	1.77
1,123.51	1.94	546.22	11.27
688.09	0.44	273.89	4.98
1,174.54	10.11	281.38	4.08
599.50	0.76	588.86	14.63
1,240.81	6.86	574.11	12.38
538.35	0.48	307.09	1.53
607.75	1.15	409.37	4.46
705.89	6.09	327.07	2.46
1,290.57	10.88	425.69	6.30
526.09	0.15	310.05	4.54
784.68	2.67	289.75	0.97
953.14	3.09	408.18	2.16
1,149.31	0.76	383.04	8.38
1,074.22	3.71		

- A. What are the best correlation metric(s) to use in describing this relation?
- B. Are uranium concentrations correlated with total dissolved solids in the groundwater samples? If so, describe the strength of the relation.
- C. Is the relation significant for  $\alpha=0.1$ ?



# Chapter 9

## Simple Linear Regression

---

*The relation 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 measures of the significance and quality of the relation are desired so that the analyst can be assured that the predictions are meaningful.*

*Sediment concentrations in an urban river are investigated to determine if installation of detention ponds throughout the city have decreased sediment concentrations. Linear regression is first performed between sediment concentration and river discharge to remove the variation in concentrations owing to flow variations. After subtracting this relation from the data, the residual variation before and 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 relation. This coefficient is tested to see if it is significantly different than the slope coefficient obtained 5 years ago.*

The examples involve performing a linear regression between the same two variables, sediment concentration and water discharge, but for three different and commonly used 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.

This chapter focuses on the analysis of the linear relation between one continuous variable of interest, called the response variable, and one other variable, called the explanatory variable, by simple linear regression. The name simple linear regression (SLR) 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 Linear Regression). If the data are not found to be linearly related to one another, then one must decide whether to apply a transformation of the data or seek an alternate approach, discussed in chapter 10. In general, regression is performed to

1. Learn something about the relation between the two variables;
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.

SLR is an important tool for the statistical analysis of water resources data. It is used to describe the covariation between some continuous variable of interest and some other continuous variable.

## 9.1 The Linear Regression Model

The model for simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (9.1)$$

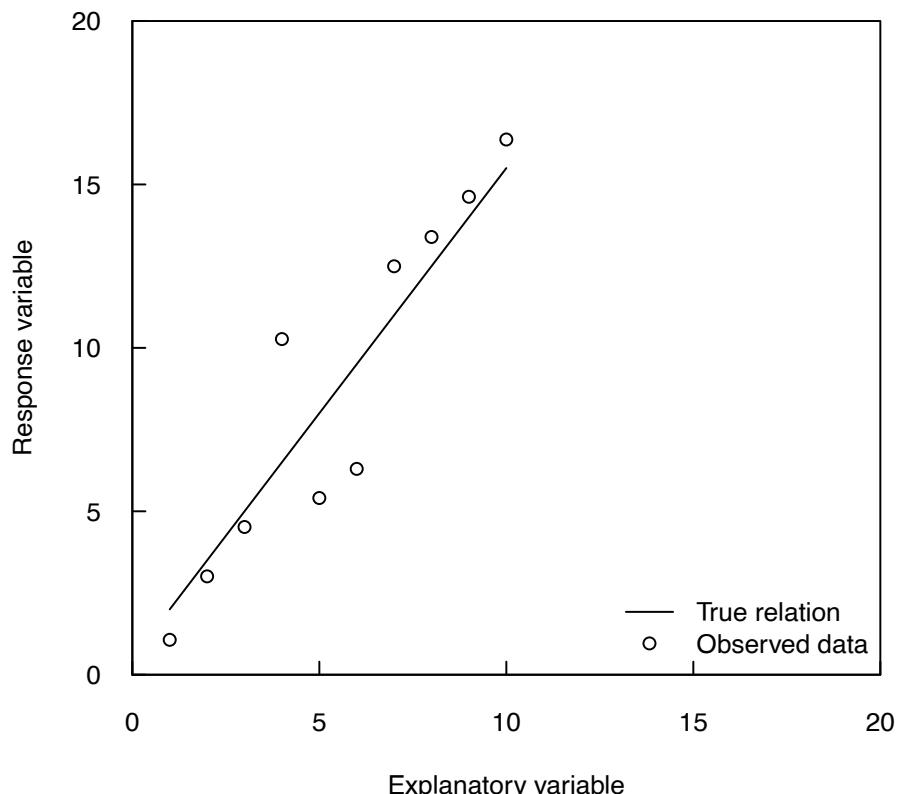
for  $i=1, 2, \dots, n$ ,

where

- $y_i$  is the  $i$ th observation of the response variable,
- $x_i$  is the  $i$ th observation of the explanatory variable,
- $\beta_0$  is the intercept,
- $\beta_1$  is the slope (the change in  $y$  with respect to  $x$ ),
- $\epsilon_i$  is the random error or residual for the  $i$ th observation, and
- $n$  is the sample size.

The error around the linear model,  $\epsilon_i$ , is a random variable. That is, its magnitude is the unexplained variability in the data. The values of  $\epsilon_i$  are assumed to have a mean of zero, and a constant variance,  $\sigma^2$ , that does not depend on  $x$ . The  $\epsilon_i$  values are assumed to be independent of  $x_i$ .

Regression is performed by estimating the unknown true intercept,  $\beta_0$ , and slope,  $\beta_1$ , with estimates  $b_0$  and  $b_1$ , respectively, that are computed from observed data. As an example, in figure 9.1 a solid line represents the true linear relation between an explanatory variable,  $x$ , and the response variable,  $y$ . Around the line are 10 observed data points that result from observing this relation, plus the random error,  $\epsilon_i$ , inherent in the natural system and the process of measurement. In practice, the true line is never known,



**Figure 9.1.** Plot of the true linear relation between the response variable and the explanatory variable, and 10 observations of the response variable for explanatory variable values at integer values from 1 through 10. Observations are the true relation plus random error with mean of zero and standard deviation of 2.

instead the analyst uses the observed data points to estimate a linear relation between  $x$  and  $y$ . The linear regression estimate developed from the 10 measurements is shown as the dashed line in figure 9.2. Note that the random error results in an estimate of the relation that differs from the true relation.

If 10 new data points were collected to estimate the true (solid line) relation and their linear regression line was computed, slightly different estimates of  $b_0$  and  $b_1$  would 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 relation than others. This example illustrates that a regression line should always be considered as a sample estimate of the true, but unknown, linear relation.

Another way of describing the SLR model is that it provides an estimate of the mean, also called the expected value,  $E[]$ , and variance,  $Var[]$ , of  $y$ , given some particular value of  $x$ . Conceptually, for a particular value of  $x_0$ , there is a distribution of  $y$  values having a mean and variance that described this distribution. In statistical notation, this is expressed as  $E[y|x_0]$  and  $Var[y|x_0]$ , which are called the conditional mean and variance, respectively, of the  $y$  values given a value of  $x_0$ . SLR estimates the conditional mean given for a given value of  $x_0$  and the condition variance of the distribution for all values of  $x$  is  $\sigma^2$ .

### 9.1.1 Computations

Linear regression estimation is nothing more than a minimization problem; that is, linear regression is the process of estimating the line that minimizes some measure of the distance between the line and the observed data points. In ordinary least squares (OLS) regression, the estimated line minimizes the sum of the squared vertical distances between the observed data and the line. OLS is by far the most common way to obtain the linear regression line. For this reason, in both this chapter and chapter 11, the linear relation between  $y$  and  $x$  is determined by OLS. Other computational methods exist to define the relation between  $x$  and  $y$  and are useful in certain circumstances (see chap. 10).

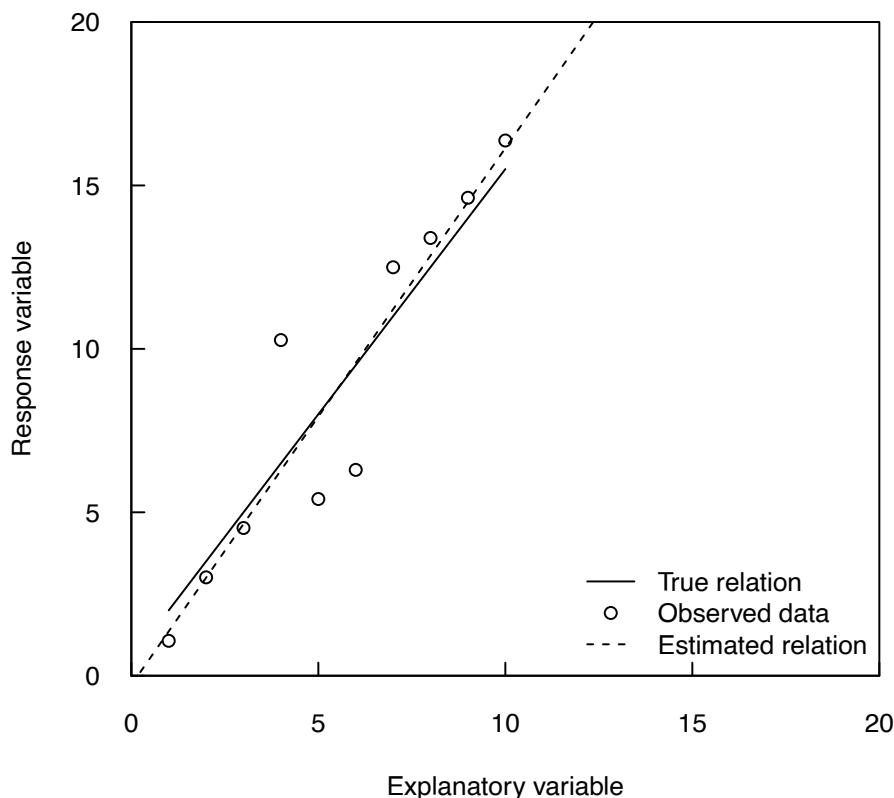
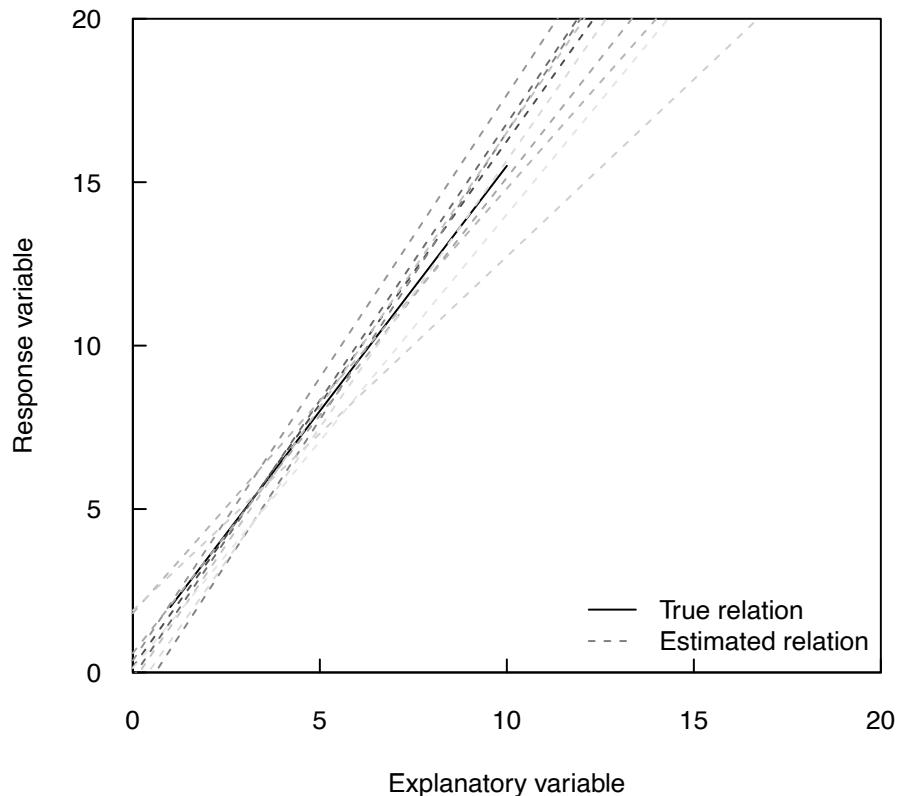


Figure 9.2. Plot showing true and estimated linear relation between the explanatory and response variables using the observations from figure 9.1.



**Figure 9.3.** Plot of true and estimated linear relations between  $x$  and  $y$  from different sets of 10 observations all generated using the true relation and sampling error.

The OLS solution can be stated as follows: find two estimates,  $b_0$  and  $b_1$ , such that the sum of the squared differences between the estimates and the observations is minimized. In mathematical terms,  $\sum_{i=1}^n (\hat{y}_i - y_i)^2$  is minimized, where  $\hat{y}_i$  is the OLS estimate of  $y$ :

$$\hat{y}_i = b_0 + b_1 x_i . \quad (9.2)$$

The minimization problem can be solved using calculus, and the solution is referred to as the normal equation. From the solution comes an extensive list of expressions used in regression analysis that are shown in table 9.1. Nearly every statistical software program, including R, calculates these statistics. They form the basis for many of the statistical tests associated with linear regression.

**Table 9.1.** Formulas utilized in ordinary least squares (OLS) linear regression.

Formula	Explanation
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	Mean of $x$ .
$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$	Mean of $y$ .
$SS_y = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n(\bar{y})^2$	Sum of squares (SS) of $y$ = Total sum of squares.
$SS_x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$	Sum of squares (SS) of $x$ .
$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$	Sum of $x, y$ cross products.
$b_1 = \frac{S_{xy}}{SS_x}$	The estimate of $\beta_1$ (slope).
$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}$	The estimate of $\beta_0$ (intercept).
$\hat{y}_i = b_0 + b_1 x_i$	The estimate of $y$ given $x_i$ .
$e_i = y_i - \hat{y}_i$	The estimated residual for observation $i$ .
$SSE = \sum_{i=1}^n e_i^2$	Error sum of squares.
$s^2 = \frac{(SS_y - b_1 S_{xy})}{(n-2)} = \frac{\sum_{i=1}^n e_i^2}{(n-2)} = \frac{SSE}{(n-2)}$	The estimate of $\sigma^2$ , also called mean square error (MSE).
$s = \sqrt{s^2}$	Standard error of the regression or standard deviation of residuals.
$SE(\beta_1) = \frac{s}{\sqrt{SS_x}}$	Standard error of $\beta_1$ .
$SE(\beta_0) = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$	Standard error of $\beta_0$ .
$r = \frac{S_{xy}}{\sqrt{SS_x SS_y}} = b_1 \sqrt{\frac{SS_x}{SS_y}}$	The (Pearson) correlation coefficient between $x$ and $y$ .
$R^2 = \frac{[SS_y - s^2(n-1)]}{SS_y} = 1 - \left( \frac{SSE}{SS_y} \right) = r^2$	The coefficient of determination, or the fraction of the variance explained by regression.

**Table 9.2.** Assumptions necessary for the purposes to which ordinary least squares (OLS) regression is applied.

[X, the assumption is required for that purpose; -, assumption is not required]

Assumption	Purpose				Test hypotheses, estimate confidence or prediction intervals
	Predict $y$ given $x$	Predict $y$ and a variance for the prediction	Obtain best linear unbiased estimator of $y$		
Model form is correct: $y$ is linearly related to $x$ .	X	X	X	X	X
Data used to fit the model are representative of data of interest.	X	X	X	X	X
Variance of the residuals is constant (homoscedastic). It does not depend on $x$ or on anything else such as time.	-	X	X	X	X
The residuals are independent of $x$ .	-	-	-	X	X
The residuals are normally distributed.	-	-	-	-	X

### 9.1.2 Assumptions of Linear Regression

There are five assumptions associated with OLS linear regression, which are listed in table 9.2. The necessity of satisfying them is determined by the intended purpose of the regression equation. The table indicates which conditions must be met for each purpose. Each assumption is described in detail in later sections of this chapter.

Note that the assumption of a normal distribution is involved only when testing hypotheses, where the residuals from the regression equation are required to be normally distributed. In this sense, linear regression is a parametric procedure; however, no assumptions are made concerning the distributions of either the explanatory ( $x$ ) or response ( $y$ ) variables. Normality of residuals is also required for the most important hypothesis test in regression—whether the slope coefficient is significantly different from zero, meaning the linear relation between  $y$  and  $x$  is significant. Normality of the residuals should be checked by a boxplot or probability plot. It should be noted that although the residuals must follow the condition of normality, there is no requirement that the distribution of  $x$  or  $y$  are normal. However, the regression line, because it is a conditional mean, is sensitive to the presence of outliers in much the same way as a sample mean is sensitive to outliers.

### 9.1.3 Properties of Least Squares Solutions

If the first four assumptions of table 9.2 are all met, then the following is true:

1. The estimators  $b_0$  and  $b_1$  are the minimum variance unbiased estimators of  $\beta_0$  and  $\beta_1$ , respectively. This means that  $b_0$  and  $b_1$  are not only unbiased estimators but also have the smallest variance of any other unbiased estimator of  $b_0$  and  $b_1$ .
2. The mean of the residuals ( $e_i$  values) is exactly zero.
3. The mean of the predictions ( $\hat{y}_i$  values) equals the mean of the observed responses ( $y_i$  values).
4. The regression line passes through the centroid of the data ( $\bar{x}$ ,  $\bar{y}$ ).
5. The variance of the predictions ( $\hat{y}_i$  values) is less than the variance of the observed responses ( $y_i$  values) unless  $R^2=1.0$ .

## 9.2 Getting Started with Linear Regression

A common first step in performing regression is to plug the data into a statistics software package and evaluate the results using the value of  $R^2$  (table 9.1), which is a measure of the variance explained by the model.  $R^2$  is ubiquitous in the environmental literature as a measure of the quality of a regression model because of its ease of understanding and simplicity; however, this could lead one towards a dangerous, blind reliance on the computer software values. For example, values of  $R^2$  close to 1 are often incorrectly deemed an indicator of a good model; it is possible for a poor regression model to result in an  $R^2$  near 1, and there are cases for which models resulting in low  $R^2$  values may often be preferable to models with higher  $R^2$  values. For example, a model may result in a low  $R^2$  value; however, the model may still be useful in determining the significance of the change in  $y$  for a given change in  $x$ .

To avoid reliance solely on  $R^2$ , this chapter outlines a series of steps that will generally lead to a good regression model. These steps will also help determine if the assumptions in table 9.2 are met by both the data and resulting model and, if they are not met, offer solutions for mitigation. One example is carried through the chapter to demonstrate these steps in practice.

The first step in regression model development is always to plot the data! Note from table 9.2 that the correct model form is critical to every application of linear regression. When viewing the plot of  $y$  versus  $x$ , there are two properties that must be evident before proceeding with the development of a regression model:

1. The relation must appear to be linear. If this is not the case and the problem is with the curvature of the data only, try to identify a new  $x$  which is a better linear predictor either through a transform of the original  $x$  or use another explanatory variable altogether. When possible, use the best physically based argument in choosing the correct  $x$ . It may be appropriate to resort to empirically selecting the  $x$  that works best (highest  $R^2$ ) from among a set of equally reasonable explanatory variables.
2. The relation must exhibit homoscedasticity (constant variance) in  $y$  across all values of  $x$ . If heteroscedasticity is present—in other words, the variance is not constant—or if both curvature and heteroscedasticity are present, then transforming  $y$ , or  $x$  and  $y$ , may mitigate this issue. Mosteller and Tukey (1977) provided a guide to selecting power transformations using plots of  $y$  versus  $x$  called the bulging rule (fig. 9.4). By comparing the curvature of a dataset to figure 9.4, one can determine what type of transformation may help linearize the relation between  $x$  and  $y$  and mitigate heteroscedasticity. Going “up” means exponentiating the  $x$  or  $y$  value by a power greater than 1 (for example, using  $x^2$  rather than  $x$ ). Going “down” means exponentiating the  $x$  or  $y$  value by a power less than 1 (for example, using the natural logarithm of  $x$ ,  $1/x$ , or  $\sqrt{x}$  rather than  $x$ ).

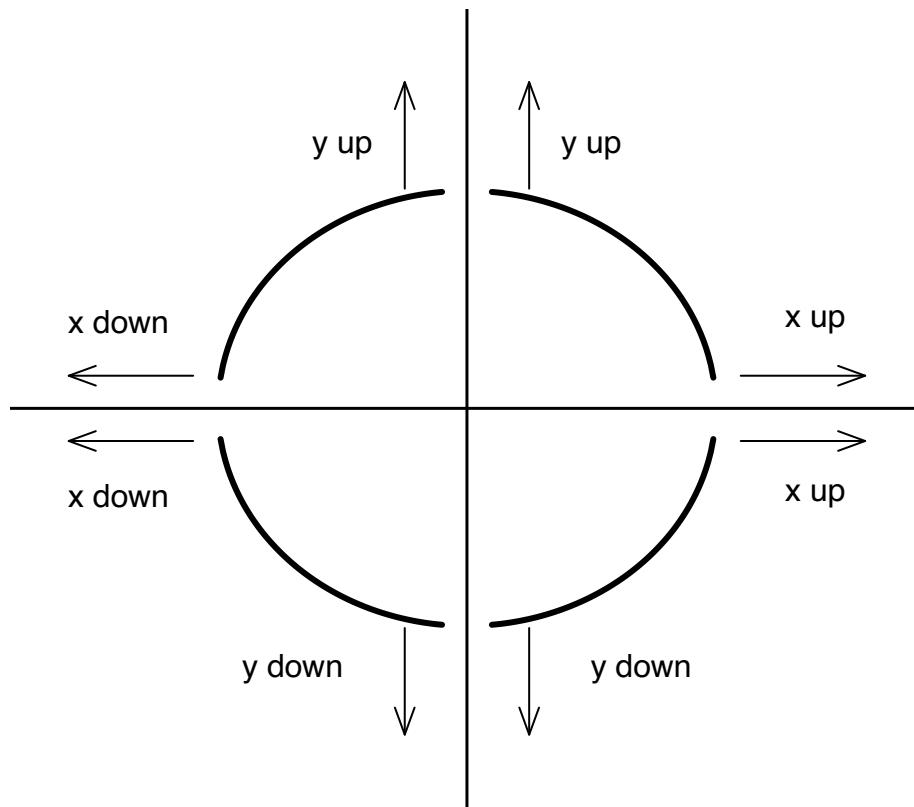
Until these conditions are met, it is not advisable to proceed with the development of a linear regression model. Additional information on the implications of variable transformation is provided in section 9.6.

### Example 9.1. Plotting the relation.

The example used in the following sections will relate the total dissolved solids (TDS) concentrations, in milligrams per liter (mg/L) to stream discharge at the U.S. Geological Survey monitoring site Cuyahoga River at Old Portage, Ohio, for the period 1969–73 as an example dataset. The TDS data are provided with the report as an .RData file but can also be downloaded from the U.S. Geological Survey National Water Information System using the site number 04206000 and parameter code 70300.

We load the data for the Cuyahoga River and then plot discharge versus TDS concentration:

```
> load("CuyaTDS.RData")
> par(tck = 0.02, las = 1, xaxs = "i", yaxs = "i")
> plot(cuya.tds$discharge_cms, cuya.tds$tds_mgL, xlim = c(0, 60),
+       xlab="Discharge, in cubic meters per second",
+       ylab="Total dissolved solids concentration, in milligrams per
liter",
+       ylim=c(0,800))
```

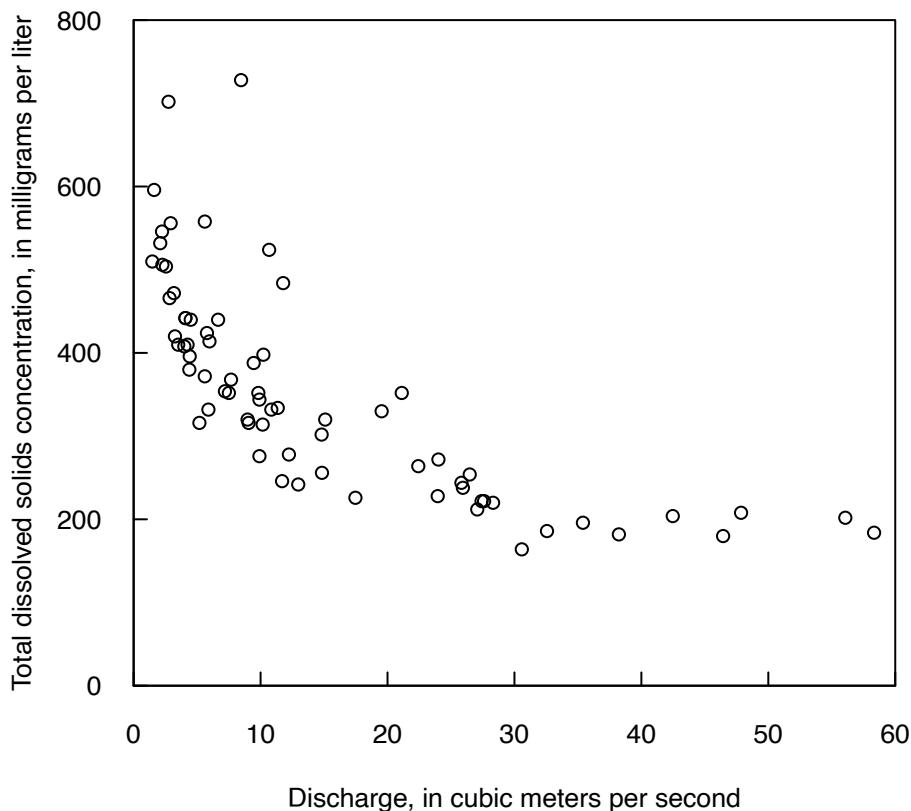


**Figure 9.4.** The bulging rule for transforming curvature to linearity (based on Mosteller and Tukey, 1977).

The nonlinearity of the TDS data as a function of discharge is obvious from figure 9.5, and some type of transformation—in this case, on discharge,  $Q$ —could be applied to attempt to linearize the relation and minimize heteroscedasticity before continuing. In this example, the plot in figure 9.5 has the shape of the lower left quadrant of the bulging rule (fig. 9.4). This means that a good transformation of  $Q$  would be one in which we exponentiate  $Q$  with an exponent less than 1. The most common transformation for environmental data with this nonlinear pattern is the logarithm transformation. In this example, we will choose to take the natural log of  $Q$  to linearize the relation and so we want to plot the natural logarithm of discharge versus TDS concentration to determine if this has the desired effect (recall that in R, the function `log` computes the natural (base e) logarithm as the default base):

```
> par(tck = 0.02, las = 1, xaxs = "i", yaxs = "i")

> plot(cuya.tds$discharge_cms, cuya.tds$tds_mgL,
+       xlim=c(1,100),
+       xlab="Discharge, in cubic meters per second",
+       ylab="Total dissolved solids concentration, in milligrams per
+             liter",
+       ylim=c(0,800), log = "x")
```

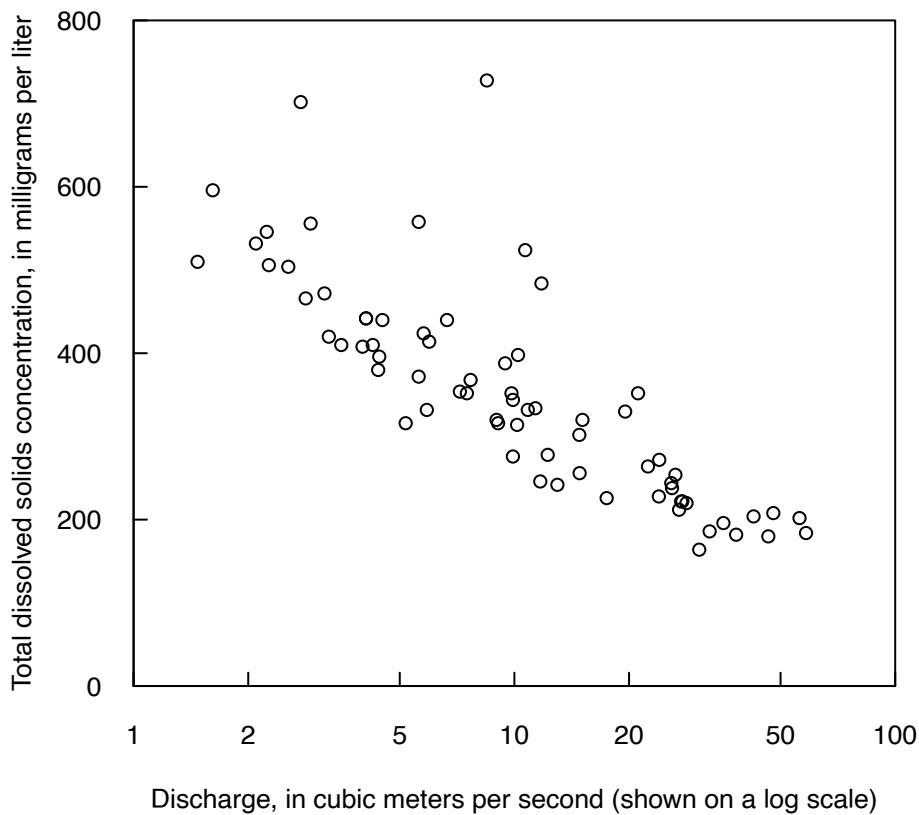


**Figure 9.5.** Scatterplot of discharge versus total dissolved solids concentrations for the Cuyahoga River, Ohio, over the period 1969–73.

The TDS data versus the natural log of  $Q$  are presented in figure 9.6, and we observe that approximate linearity is achieved. There is some hint of decreased variance in the higher  $Q$  values; however, based on the first set of plots, this transformation appears acceptable. Although examining the assumption of homoscedasticity is important, the most important assumption of simple linear regression is that the relation is linear. The analyst must first determine the case for linearity before proceeding with any regression model. As an exercise beyond this example, the reader is encouraged to try other transformations.

### 9.3 Describing the Regression Model

Recall that the regression model had two parameters that must be estimated, the slope,  $\beta_1$ , and the intercept,  $\beta_0$ , (eq. 9.2); however, the estimated values of  $\beta_1$  and  $\beta_0$  alone will not provide any information as to their usefulness. Fortunately, there are several hypothesis tests that can be used to evaluate  $b_1$  and  $b_0$  (the estimates of  $\beta_1$  and  $\beta_0$ ) and the resulting  $\hat{y}_i$  values. Note that all assumptions in table 9.2 must be met in order to apply hypothesis tests to the regression model. Section 9.4 describes in detail how to check these assumptions.



**Figure 9.6.** Scatterplot of discharge versus total dissolved solids concentrations after the transformation of discharge using the natural log.

#### Example 9.2. Fitting the regression model.

A regression line is fitted to the data using the natural log transformation of the discharge values and the summary output is shown below. The values for  $b_1$  and  $b_0$  are shown in the summary output under the Coefficients section. The value of  $b_1 = -111.631$  and  $b_0 = 609.549$ .

```
> cuya.lm <- lm(tds_mgL ~ log(discharge_cms), data = cuya.tds)
> summary(cuya.lm)
```

Call:

```
lm(formula = tds_mgL ~ log(discharge_cms), data = cuya.tds)
```

Residuals:

Min	1Q	Median	3Q	Max
-109.89	-42.95	-10.65	13.32	356.91

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	609.549	22.800	26.73	<2e-16 ***
log(discharge_cms)	-111.631	9.227	-12.10	<2e-16 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.57 on 68 degrees of freedom

Multiple R-squared: 0.6828, Adjusted R-squared: 0.6781

F-statistic: 146.4 on 1 and 68 DF, p-value: < 2.2e-16

### 9.3.1 Test for Whether the Slope Differs from Zero

The hypothesis test of greatest interest in regression is the test as to whether  $\beta_1$  is significantly different from zero. The null hypothesis for this test is

$$H_0 : \beta_1 = 0,$$

and the alternative hypothesis is

$$H_A : \beta_1 \neq 0.$$

If the null hypothesis is true, equation 9.2 reduces to  $\hat{y}_i = b_0$ . In other words, the value of  $y$  does not vary as a linear function of  $x$  if the null hypothesis is true. In the case of SLR only (one explanatory variable), there are two additional interpretations of the model results that follow from this hypothesis test: (1) whether the regression model has statistical significance, and (2) whether the linear correlation coefficient significantly differs from zero. Both interpretations will have the identical answers because the significance for (1) and (2) are identical in SLR. These latter two interpretations are not applicable for linear regression equations with multiple explanatory variables, which are discussed in chapter 11.

The test statistic computed is the  $t$ -ratio (the fitted coefficient divided by its standard error):

$$t = \frac{b_1}{\sqrt{\frac{s}{\sqrt{SS_x}}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (9.3)$$

$H_0$  is rejected if  $|t| > t_{crit}$ , where  $t_{crit}$  is the point on the  $t$ -distribution with  $n-2$  degrees of freedom and with a probability of exceedance of  $\alpha/2$ . If we select  $\alpha=0.05$ , then  $t_{crit}=1.96$ . A handy rule of thumb is to consider  $|t| > 2$  to be significant and  $|t| < 2$  to be nonsignificant. To compute the value of  $t_{crit}$  in R, enter `qt(1-(\alpha/2), n-2)`, where  $\alpha$  is the  $\alpha$  level and  $n$  is the number of observations.

This test for nonzero slope can also be generalized to testing the null hypothesis that  $b_1 = \beta_1^*$  where  $\beta_1^*$  is some prespecified value, although this test is used far less frequently in statistics. For this test, the statistic is defined as

$$t = \frac{b_1 - \beta_1^*}{\sqrt{\frac{s}{\sqrt{SS_x}}}}. \quad (9.4)$$

### 9.3.2 Test for Whether the Intercept Differs from Zero

Hypothesis tests on the intercept,  $\beta_0$ , can also be computed. The null and alternative hypotheses are

$$H_0 : \beta_0 = 0$$

$$H_A : \beta_0 \neq 0 .$$

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.

It can be dangerous to delete the intercept term from a regression model. The fact that the intercept term in a fitted model is not significantly different from zero is not a justification for removing the intercept term from the model (that is, setting it equal to zero). Regression statistics such as  $R^2$  and the  $t$ -ratio for  $\beta_1$  lose their usual meaning when the intercept term is dropped (set equal to zero); this is because  $R^2$  is comparing the linear regression model to an intercept-only model. If there is no intercept, then the comparison does not make sense. Recognition of a physical reason why  $y$  must be zero when  $x$  is zero is not a sufficient argument for setting  $\beta_0=0$ . 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 relation clearly looks linear from zero to the most extreme  $x$  values.
3. The null hypothesis that  $\beta_0=0$  is not rejected.
4. There is some economic or scientific benefit to dropping the intercept.

It is worth noting again that even if all conditions are met, one cannot report the  $R^2$  value or do further statistical testing on  $b_1$  when the intercept is not statistically different from zero.

### Example 9.3. Evaluating the slope and intercept.

The summary output for the Cuyahoga River TDS example provides the  $t$ -statistics and corresponding  $p$ -values for both  $b_1$  and  $b_0$ :

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	609.549	22.800	26.73	<2e-16 ***
log(discharge_cms)	-111.631	9.227	-12.10	<2e-16 ***

Observe that  $t$ -values for  $b_1$  and  $b_0$  were much greater than  $\pm 2$  (26.73 and -12.10, respectively). Therefore, the  $p$ -values were small and when compared to the predetermined  $\alpha$  level, both  $b_1$  and  $b_0$  were significant. Assuming the conditions in table 9.2 are met, these results indicate that there is a statistically significant linear, negative correlation between TDS and the natural log of discharge. Section 9.4 will discuss how to check if the assumptions of table 9.2 are met.

### 9.3.3 Confidence Intervals on Parameters

Confidence intervals for the individual parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$  indicate how well they can be estimated. The meaning of the  $(1-\alpha) \cdot 100$  percent 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  $\alpha$ . 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 percent of the time.

For the slope,  $\beta_1$ , the confidence interval (CI), as a function of the standard error of  $\beta_1$ , is

$$(b_1 - t SE(\beta_1), b_1 + t SE(\beta_1)) \quad (9.5)$$

or by substituting the equation for  $SE(\beta_1)$  from table 9.1

$$\left( b_1 - t \frac{s}{\sqrt{SS_x}}, b_1 + t \frac{s}{\sqrt{SS_x}} \right), \quad (9.6)$$

where  $t$  is the point on the  $t$ -distribution having  $n-2$  degrees of freedom with a probability of exceedance of  $\alpha/2$  and  $s$  is the standard error of the regression (also referred to as the standard deviation of the residuals; see table 9.1).

For the intercept,  $\beta_0$ , the CI, as a function of  $SE(\beta_0)$ —the standard error of  $\beta_0$ —is

$$\left( b_0 - t_{n-2, 1-\frac{\alpha}{2}} s SE(\beta_0), b_0 + t_{n-2, \frac{\alpha}{2}} s SE(\beta_0) \right), \quad (9.7)$$

where  $t$  is the point on the  $t$ -distribution having  $n-2$  degrees of freedom with a probability of exceedance of  $\alpha/2$  and  $s$  is the standard error of the regression, or by substituting the equation for  $SE(\beta_0)$  from table 9.1:

$$\left( b_0 - t_{n-2, 1-\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}, b_0 + t_{n-2, \frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}} \right). \quad (9.8)$$

For the variance,  $\sigma^2$ , of the residuals,  $e_i$ , the CI is

$$\left( \frac{(n-2)s^2}{\chi^2_{n-2, 1-\frac{\alpha}{2}}}, \frac{(n-2)s^2}{\chi^2_{n-2, \frac{\alpha}{2}}} \right), \quad (9.9)$$

where

$\chi^2_{n-2, 1-\frac{\alpha}{2}}$  is the quantile of the chi-square distribution having  $n-2$  degrees of freedom with exceedance probability of  $\alpha/2$ .

#### Example 9.4. Computing confidence intervals on the slope, intercept, and error variance.

The 95-percent confidence intervals for  $\beta_1$  and  $\beta_0$  using the Cuyahoga River dataset from example 9.1 are

$$\text{For } \beta_1 : \left( -111.631 - \frac{1.995 \cdot 72.57}{\sqrt{61.85}}, -111.631 + \frac{1.995 \cdot 72.57}{\sqrt{61.85}} \right) = (-130.0443, -93.21767)$$

$$\begin{aligned} \text{For } \beta_0 : & \left( 609.549 - 1.995 \cdot 72.57 \sqrt{\frac{1}{70} + \frac{2.285^2}{61.85}}, 609.549 + 1.995 \cdot 72.57 \sqrt{\frac{1}{70} + \frac{2.285^2}{61.85}} \right) \\ & = (564.0643, 655.0337). \end{aligned}$$

In R, both calculations can be made using one command, with differences between the results above and those from the R command only the result of rounding:

```
> confint(cuya.lm, level = 0.95)
              2.5 %    97.5 %
(Intercept)      564.0529 655.04457
log(discharge_cms) -130.0436 -93.21855
```

The 95% confidence intervals for  $\sigma^2$ :

$$\text{For } \sigma^2 : \left( \frac{(n-2) \cdot s^2}{\chi^2_{1-\frac{\alpha}{2}, n-2}}, \frac{(n-2) \cdot s^2}{\chi^2_{\frac{\alpha}{2}, n-2}} \right) = (3863.6, 7472.5)$$

Or in R,

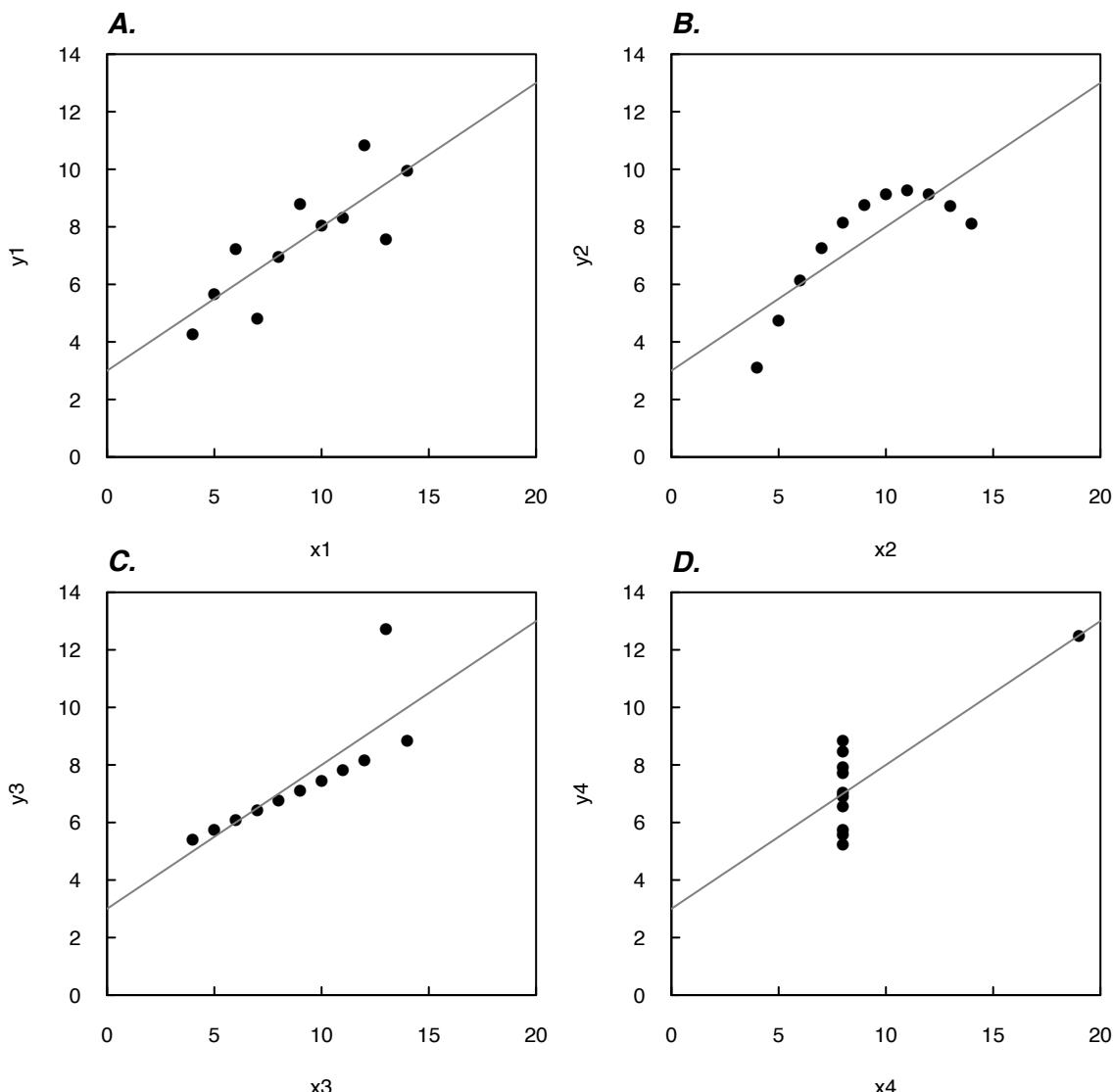
```
> alpha <- 0.05
> n <- 70
> s <- 72.57
> mse <- s^2
> CIlower <- ((n - 2) * mse) / qchisq(1 - (alpha/2), n - 2)
> CIupper <- ((n - 2) * mse) / qchisq(alpha/2, n - 1)
```

## 9.4 Regression Diagnostics

One common mistake in regression analysis is to base decisions about model adequacy solely on the regression summary statistics.  $R^2$  is a measure of the percent of the linear variation in the response variable ( $y$ ) that is accounted for by the variation in the explanatory variable ( $x$ ). The  $s$  term—the standard error of the regression or standard deviation of the residuals (table 9.1)—is a measure of the dispersion of the data around the regression line. Most regression programs also perform an overall hypothesis test to determine if the regression relation is statistically significant; that is, that the apparent linear relation between  $y$  and  $x$  is sufficiently strong that it is not likely to arise as a result of chance alone. As described in section 9.3, in SLR, this test is the same as a test to determine if the slope coefficient is significantly different from zero, and also the same as a test if the correlation coefficient between  $x$  and  $y$  is significantly different from zero.

These statistics provide substantial information about regression results. A regression 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 regression coefficients alone do not guarantee that the data have been fitted well by the model, as illustrated by figure 9.7 (Anscombe, 1973).

The data in the four graphs shown in figure 9.7 have exactly the same summary statistics and regression line (same  $b_0$ ,  $b_1$ ,  $s$ , and  $R^2$ ). The figure 9.7A is a perfectly reasonable regression model, an evidently linear relation having an even distribution of data around the least-squares line; however, the other models are not. The strong curvature in figure 9.7B suggests that a linear model is highly inadequate and that some transformation of  $x$  is necessary or that an additional explanatory variable is required. With these improvements perhaps more of the variance could be explained, but in its current state, it is not a valid model. The effect of a single outlier on regression is illustrated in figure 9.7C. The line does not properly fit the data, as the line is drawn towards the outlier. Such an outlier must be recognized and



**Figure 9.7.** Plots of four different datasets fit with a simple linear regression. All plots have the same regression summary statistics and regression line (based on Anscombe, 1973).

carefully examined to verify its accuracy if possible. If it is impossible to demonstrate that the point is erroneous, a more robust procedure than OLS should be utilized (see chap. 10). The regression slope in figure 9.7D is strongly affected by a single point (the large  $x$  value), with the regression simply connecting a single point plus a small cluster of points all at the same  $x$  value. Such situations often produce  $R^2$  values close to 1, yet may have little if any predictive power because the slope and  $R^2$  are totally controlled by the position of one point—an unstable situation. Had the outlying point been in a different location, the resulting slope would be totally 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.

Using statistical terminology, the plots in figure 9.7 demonstrate three situations that create problematic regression models for which summary statistics alone cannot characterize: curvature (fig. 9.7B), outlier or large residual (fig. 9.7C), and high influence and leverage (fig. 9.7D). These cases are generally easy to identify from plots of  $y$  versus  $x$  or residuals versus predicted  $y$  values in a linear regression with one explanatory variable. However, in multiple linear regression (chap. 11) 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 to overcome this challenge.

Regression diagnostics, which particularly relate to the behavior of the residuals, are required to meet the assumptions of table 9.2 and to ensure that regression coefficients represent a good estimate of the relation of  $x$  and  $y$ . This section describes the metrics and manner in which additional diagnostics are computed and evaluated. In assessing a regression model, it is important to remember that regression is an iterative and interpretive process that requires the evaluation of many different aspects of the regression model before one can determine if they have a good model or not. The analyst must weigh how the regression model will be used versus how strictly to enforce required assumptions (table 9.2). Graphical methods are essential tools for evaluating regression relations. Each of the following diagnostics and metrics are given here in terms of one explanatory variable but can be generalized using matrix notation to a larger number of dimensions for multiple linear regression, which is described in chapter 11.

### 9.4.1 Assessing Unusual Observations

There are several metrics, all based on residuals, that are used to assess the effects of individual observations on the regression model. Recall from table 9.1 that the model residual,  $e_i$ , is the difference between the observed and predicted value:

$$e_i = y_i - \hat{y}_i .$$

The values of the residuals in a regression model are critical to understanding the underlying model and ensuring the assumptions of the model are met (table 9.2). There are different formulations of the model residuals, all with their own advantages and purposes, which are discussed further in this section.

#### 9.4.1.1 Leverage

Leverage is a function of the distance from an individual  $x$  value to the middle (mean) of the  $x$  values. Leverage for a given  $x_i$  is usually denoted as  $h_i$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x} . \quad (9.10)$$

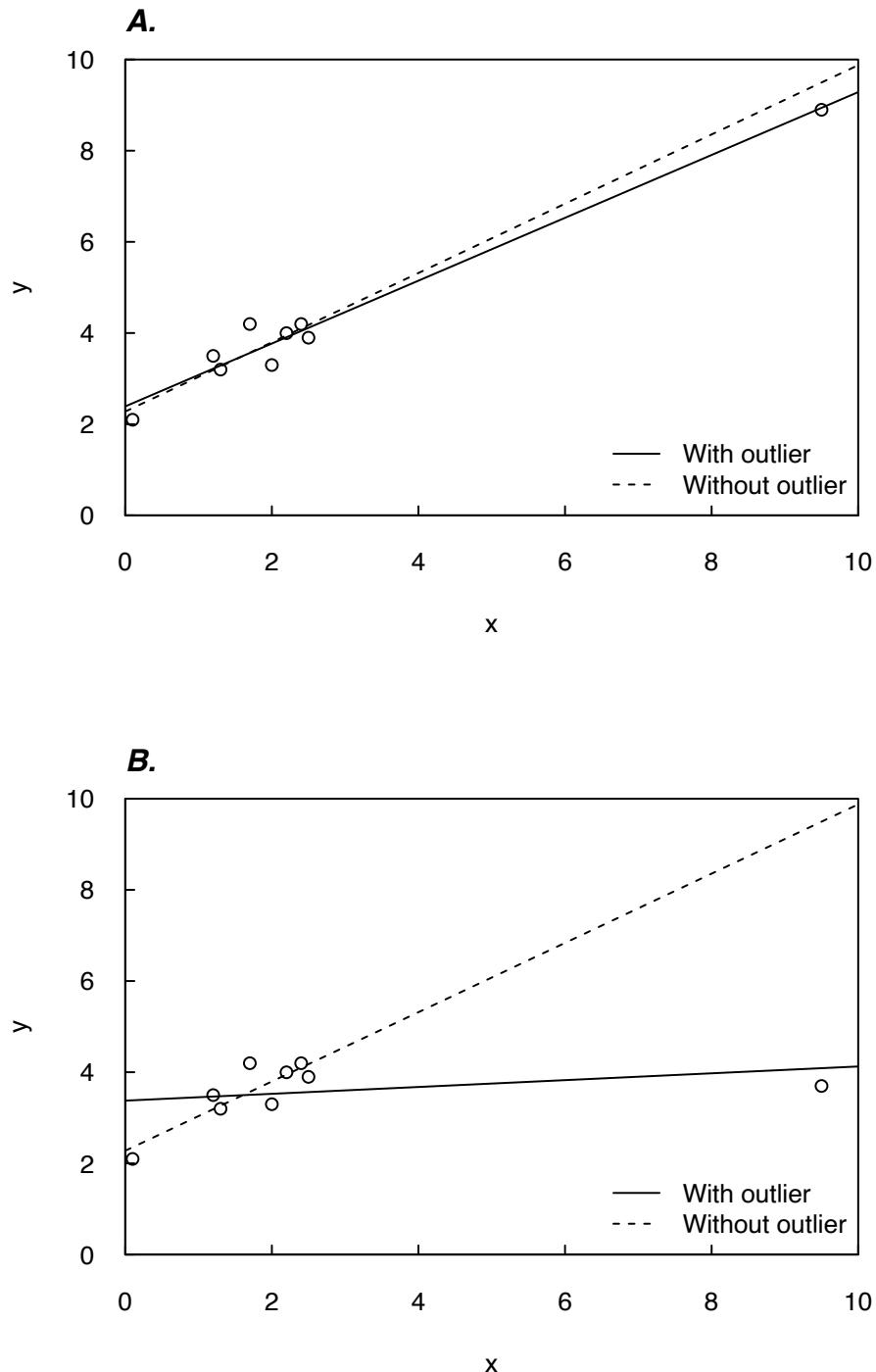
An observation with high leverage is one where  $h_i > \frac{3p}{n}$ , where  $p$  is the number of coefficients in the model. In SLR,  $p=2$  because there are two coefficients to be estimated:  $b_0$  and  $b_1$ . Although leverage is concerned only with the  $x$  direction, an observation with high leverage has the potential for exerting a strong influence on the regression slope. Observations with high leverage should be examined for errors; however, an observation with high leverage is not reason enough to remove this observation from the analysis.

#### 9.4.1.2 Influence

Observations identified as having high influence should lead to a very careful examination of the data value for possible errors or special conditions that might have prevailed at the time it occurred. If it can be shown that an error occurred, the observation 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 that better fits the observation is one option—either through transformation or the addition of multiple explanatory variables. The second option is to use a more robust procedure such as that based on Kendall's  $\tau$  or weighted least squares. Methods for robust regression are discussed in chapter 10. Weighted least squares regression is discussed in section 9.8. There are two commonly used statistics for identifying observations with high influence: Cook's  $D$  and DFFITS. They serve similar functions, although DFFITS is possibly easier to work with because the critical value is more easily computed.

Leverage and influence statistics provide a means to identify  $x$  and  $y$  values that lie outside of the bulk of the data and, therefore, have the potential to produce an unstable regression model, such as those shown in figure 9.7. Leverage is a measure of an outlier in the  $x$  direction, as shown in figure 9.8. Observations that exhibit high leverage may not necessarily affect the regression estimates much, such as shown in

figure 9.8A. Here, even though one observation is clearly further away from the others, the regression line does not change much when that observation is removed. Observations with high influence are those that have both high leverage and, when the observation is removed, substantially affect the estimated regression line (fig. 9.8B). In regression terminology, figure 9.8A shows an outlier with high leverage but low influence; figure 9.8B shows an outlier with both high leverage and high influence.



**Figure 9.8.** Influence of location of a single point on the regression slope. (A) The effect on the fitted line resulting from an outlier with high leverage but low influence; (B) the effect on the fitted line resulting from an outlier with high leverage and high influence.

### 9.4.1.3 Prediction, Standardized and Studentized (Deleted-t) Residuals

To compute the metrics that determine the influence of a particular observation, an alternative formulation of the residuals has been developed, which can be used in conjunction with other metrics to assess influence or as an additional metric to assess unusual observations.

The prediction residual,  $e_{(i)}$ , is 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 by leaving out the  $i$ th observation. Prediction residuals provide a means to assess how well the model performs in a prediction capacity because each observation is removed and not considered in fitting the model coefficients. Prediction residuals can also be calculated using leverage statistics, eliminating the need to perform a separate regression for each of the  $i$  observations:

$$e_{(i)} = \frac{e_i}{1-h_i}, \quad (9.11)$$

where

- $h_i$  is the leverage of observation  $i$  and
- $e_i$  is the residual.

Another measure of outliers in the  $y$  direction is the standardized residual,  $e_i^*$ , which is the actual residual  $e_i = y_i - \hat{y}_i$  standardized by its standard error:

$$e_i^* = \frac{e_i}{s\sqrt{1-h_i}}, \quad (9.12)$$

where

- $h_i$  is the leverage of observation  $i$ ,
- $e_i$  is the residual, and
- $s$  is the standard deviation of the residuals.

The studentized residual (also called the deleted  $t$ -residual),  $t_{(i)}$ , is also used as an alternate measure of outliers by some texts and computer software:

$$t_{(i)} = \frac{e_i}{s_{(i)}\sqrt{1-h_i}} = \frac{e_{(i)}\sqrt{1-h_i}}{s_{(i)}}, \quad (9.13)$$

where

$$s_{(i)} = \sqrt{\frac{(n-p)s^2 - \left[ \frac{e_{(i)}}{1-h_i} \right]^2}{n-p-1}}$$

- $n$  is the number of observations,
- $h_i$  is the leverage of observation  $i$ ,
- $e_i$  is the residual,
- $p$  is the number of estimated parameters in the model (for SLR,  $p=2$ ),
- $s^2$  is the variance of the residuals,
- $e_{(i)}$  is the prediction residual, and
- $s_{(i)}$  is the standard deviation of the prediction residuals.

Studentized residuals are often similar to the standardized residuals,  $e_i^*$ , but are computed using a variance,  $s_{(i)}^2$ , that does not include their own observation. Therefore, an unusually large observation does not inflate the estimate of variance that is used to determine whether the deleted observation is unusual, thus allowing outliers to be more easily detected.

### 9.4.1.4 Cook's D

The most widely used measure of influence is Cook's  $D$  (Belsley and others, 1980):

$$D_i = \frac{e_i^2 h_i}{ps^2 (1-h_i)^2} = \frac{e_{(i)}^2 h_i}{ps^2}, \quad (9.14)$$

where

- $h_i$  is the leverage of observation  $i$ ,
- $e_i$  is the residual,
- $p$  is the number of estimated parameters in the model (for SLR,  $p=2$ ), and
- $s^2$  is the variance of the residuals (table 9.1).

An observation,  $x_i$ , is considered to have high influence if  $D_i > F_{(p+1,n-p)}$  at  $\alpha=0.1$  where  $p$  is again the number of coefficients estimated in the regression (for SLR,  $p=2$ ),  $n$  is the number of observations, and  $F_{(p+1,n-p)}$  is the value of the  $F$ -distribution for  $p+1$  and  $n-p$  degrees of freedom. Note that, for SLR with more than about 30 observations, the critical value for  $D_i$  would be about 2.4, and with several explanatory variables, the critical value would be in the range of 1.6 to 2.0.

### 9.4.1.5 DFFITS

Another influence diagnostic is the DFFITS measure (Belsley and others, 1980):

$$DFFITS_i = \frac{e_i \sqrt{h_i}}{s_{(i)} (1-h_i)} = \frac{e_{(i)} \sqrt{h_i}}{s_{(i)}}, \quad (9.15)$$

where

- $h_i$  is the leverage of observation  $i$ ,
- $e_{(i)}$  the prediction residual, and
- $s_{(i)}$  is the standard deviation of the prediction residuals.

An observation is considered to have high influence if  $|DFFITS_i| \geq 2\sqrt{\frac{p}{n}}$  (Belsley and others, 1980).

#### Example 9.5. Leverage and influence statistics.

There are a number of functions and plots in R that can be used to evaluate leverage and influence. In working on regression problems, it is useful to know where certain statistics are stored or how they can be created. This example uses the Cuyahoga River dataset from example 9.1 to illustrate the ways variables can be stored and accessed. The residuals of the model,  $e_i$ , are stored as part of the regression model output in the vector `cuya.lm$residuals`. The predicted values,  $\hat{y}_i$ , are stored in the vector `cuya.lm$fitted.values`, and the coefficients are stored in the vector `cuya.lm$coefficients`. The leverage, Cook's  $D$ , and DFFITS for each residual value can be computed using the output from the stored `cuya.lm` model. The commands to compute leverage, Cook's  $D$ , and DFFITS values are `hatvalues(cuya.lm)`, `cooks.distance(cuya.lm)`, and `dffits(cuya.lm)`, respectively. To obtain the values for  $n$  and  $p$ , use the following R code with the stored `cuya.lm` model:

```
> n <- length(cuya.lm$residuals)
> p <- length(cuya.lm$coefficients)
```

To find any observations that exceed the criteria for having high leverage, use the command

```
> subset(hatvalues(cuya.lm), hatvalues(cuya.lm) > 3 * (p / n))
> named numeric(0)
```

In this case, no values were returned exceeding the criteria. To assess whether any observations exceed the criteria for Cook's  $D$  with  $\alpha=0.1$ , use the command

```
> subset(cooks.distance(cuya.lm), cooks.distance(cuya.lm) > qf(0.1, p,
+           n - p, lower.tail = FALSE))
> named numeric(0)
```

Again, no values were returned exceeding the criteria. However, for DFFITS, we find two observations—observations 2 and 29—are identified as being influential:

```
> subset(dffits(cuya.lm),dffits(cuya.lm) > 2 * sqrt(p / n))
2          29
0.7500530 0.6290945
```

## 9.4.2 Assessing the Behavior of the Residuals

Examination of the behavior of the regression model residuals is necessary to ensure that the regression equation meets the assumptions required for most applications of regression (table 9.2; assumptions 3–5). In general, the residuals must possess the following properties: (1) unbiased, (2) homoscedastic, (3) normally distributed, and (4) independent. This section discusses the ways in which residuals can be evaluated to explore their behavior and test whether the assumptions have been met.

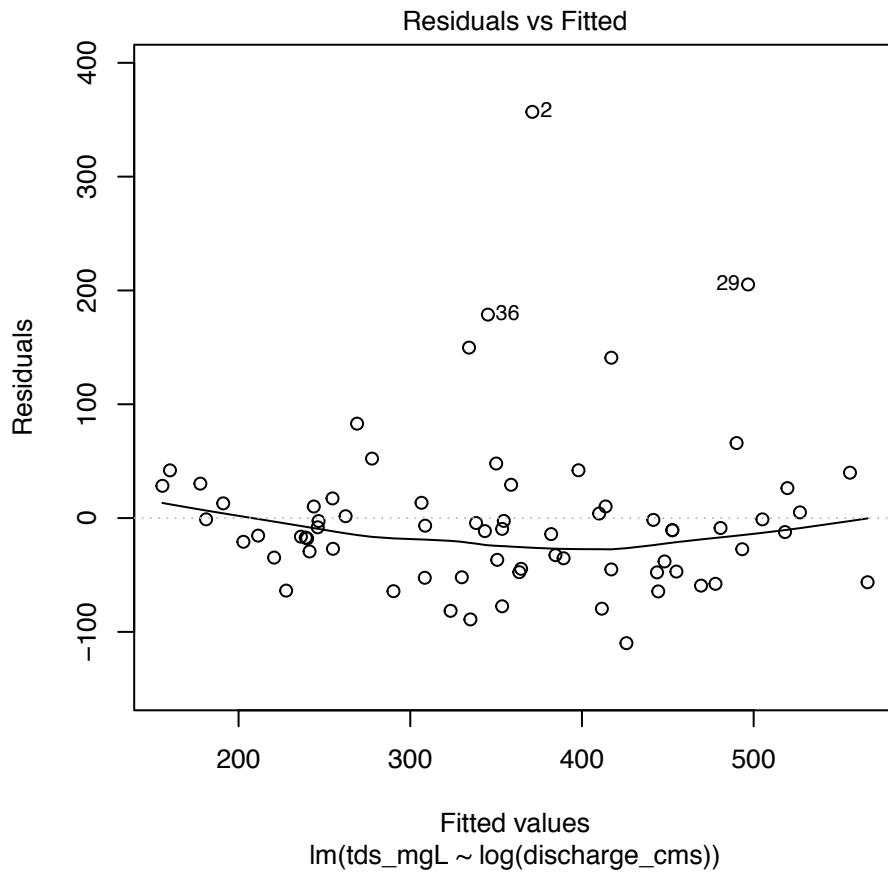
### 9.4.2.1 Assessing Bias and Homoscedasticity of the Residuals

Bias and homoscedasticity of the residuals can be evaluated visually by examining a plot of the residuals (or standardized residuals) versus predicted values ( $e$  versus  $\hat{y}$ ). Plotting the residuals in this way enhances the opportunity to more clearly evaluate the behaviors of the residuals as compared to plotting the original data,  $y$ , versus the predictions,  $\hat{y}$ . A plot of the residuals versus predicted values will allow for visual inspection of homoscedasticity (that is, that the variability in the residuals does not vary over the range of predicted values) and bias ( $e$  values generally plot equally above and below zero). For ideal regression model behavior, this plot should show a horizontal cloud of data rather than a pattern that has curvature, and the variability of that cloud of data should not substantially change as one scans from left to right across the graph. Examples of such graphs are given in example 9.6 that follows. We can also apply a formal test, the Breusch-Pagan test (Breusch and Pagan, 1979), to evaluate homoscedasticity. In this test, the null hypothesis is that the square of the residuals is independent of the fitted value. It can be run in R, using the `ncvTest` function in the `car` package (Fox and Weisberg, 2011). This test is applied in example 9.6.

There are two commonly used solutions that can mitigate poorly behaved residuals: (1) a transformation of the data (discussed in section 9.6), or (2) the use of a different independent variable to explain the variation in  $y$ . It is possible to read too much into these plots, however. Beware of apparent 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.

#### Example 9.6. Assessing bias and homogeneity of residuals.

Continuing with the example, the residuals can be plotted to determine if homogeneity exists and there is no relation between the residuals and the explanatory variable. Note that R has several plotting options available to assess a regression model. These plots are accessed by specifying a value for the `which` argument. When `which = 1`, a plot of the residuals versus the fitted data is created and this plot is useful for assessing the bias and homogeneity of the residuals. In the line of code below, the argument `ask = FALSE` prevents the user from being asked each time they would like to see the plot. For more information on this command, type `?plot.lm` at the R command prompt.

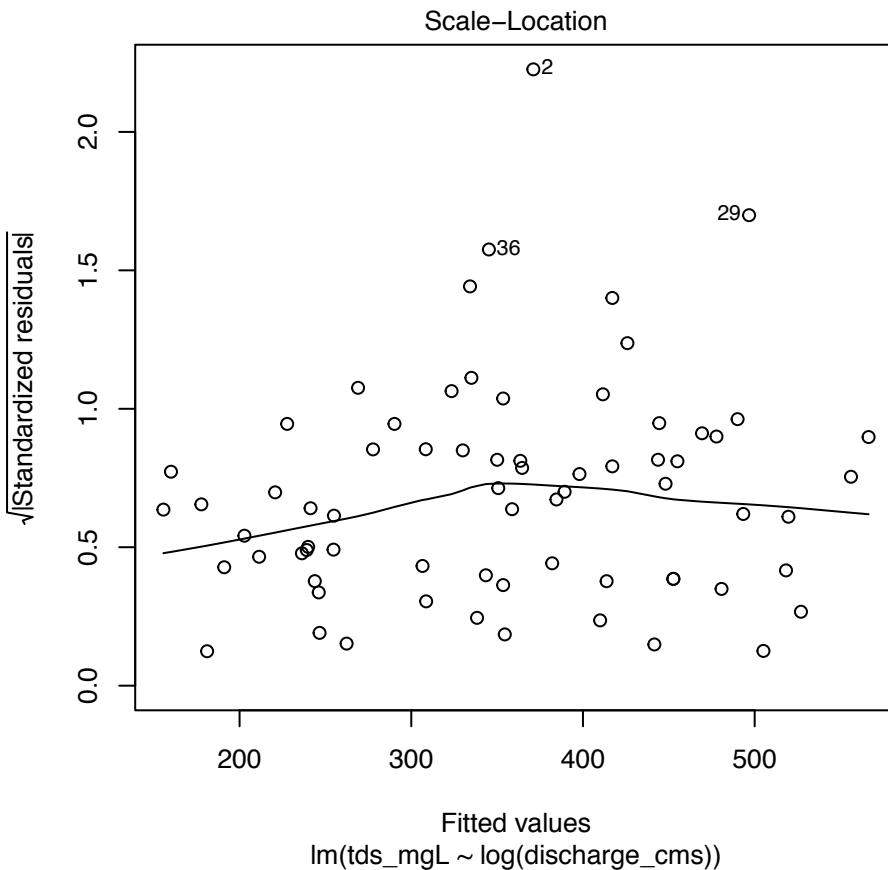


```
> plot(cuya.lm, which = 1, ask = FALSE)
```

In this example, the open circles are the residuals, the solid line is the loess-smoothed (see sections 2.3 and 10.3) line through the residuals, and the dashed line is horizontal at zero. Note that in the R output, the solid line will be colored red. The fitted values resulting in the three largest absolute standardized residuals are shown with an observation number. Notice that there is a curvature to the residuals (a pattern of high-low-high residual values going from low fitted values to high fitted values). This lack of fit is slightly worrisome but not egregious. In chapter 11 we will consider the possibility that a more complex model may improve the behavior of the residuals, such as using two explanatory variables. Other aspects of the regression diagnostics should be evaluated to determine if an alternate model should be pursued. To evaluate heteroscedasticity, we can look at the variability of the residuals across the fitted values. Homoscedastic residuals will have no trend in the variability of the points across the fitted values. In this example, aside from the outliers, the residuals have similar variability across the fitted values and are generally homoscedastic. An additional graphic that can help identify problems of heteroscedasticity is a plot of the square root of the absolute value of the standardized residuals versus the fitted values. This can be obtained with the command

```
> plot(cuya.lm, which = 3, ask = FALSE)
```

and results in the plot below, where the open circles are the residuals and the solid line is the smoothed line through the residuals. Note that in the R output, the solid line will be colored red.



Although we see some tendency for these values to rise and then fall, the smooth curve does not show a strong relation between fitted values and residual variance, we can apply the Breusch-Pagan test to evaluate homoscedasticity. The commands are

```
> install.packages(car) # To install the car package
> library(car) # To load the car package
> ncvTest(cuya.lm)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.652948 Df = 1 p = 0.103358
```

Thus, we see that although there is some indication of nonconstant variance, if we had selected  $\alpha=0.05$ , we would not reject the null hypothesis of constant variance. Although at  $\alpha=0.10$ , we would consider rejecting the null hypothesis and consider a transformation of  $y$ , a different transformation of  $x$ , or test a different explanatory variable. But, the violation of assumption of homoscedastic residuals is not so severe and we should not feel that it is a problem that must be solved in this case.

#### 9.4.2.2 Assessing Normality of the Residuals

Recall from table 9.2 that, in addition to ensuring that the residuals are unbiased and homoscedastic, the residuals must also be normally distributed in order to apply hypothesis tests and prediction and confidence intervals to the regression results. If the residuals depart substantially from a normal distribution, then the various confidence intervals, prediction intervals, and hypothesis tests will be inappropriate. Specifically, (1) hypothesis tests will have low power (slopes or explanatory variables may

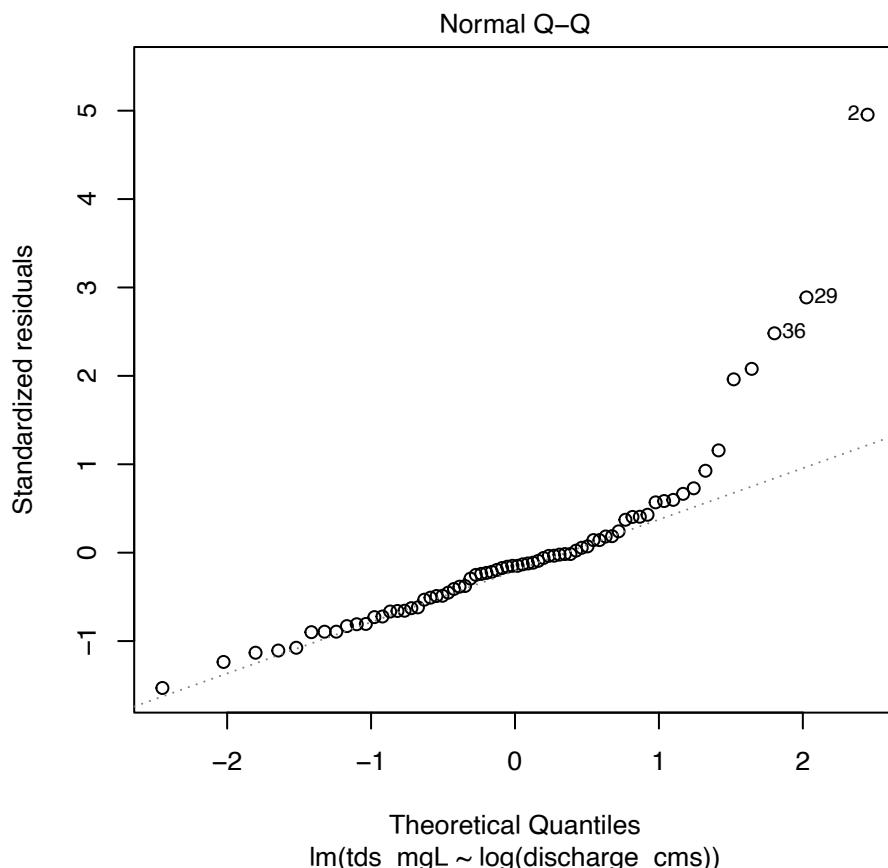
falsely be declared insignificant), and (2) confidence or prediction intervals will be too wide, as well as giving a false impression of symmetry. The determination of the normality of the residuals can be assessed graphically or through formal hypothesis tests. To determine if the residuals are normally distributed, most software programs, including R, use the standardized values of the residuals. The values of the standardized residuals are ranked and an empirical probability is computed based on each ranked value. These values are compared to the expected probabilities of the residuals if they had resulted from a normal distribution. The example, continued below, provides additional information as to how to interpret a probability plot of the residuals.

In addition to probability plots, other guidelines can be applied to determine if the residuals follow a normal distribution. If the residuals are normally distributed, values of  $|e_{si}| > 3$  should happen about 3 times in 1,000 observations and values of  $|e_{si}| > 2$  should happen about 5 times in 100 observations. Under a correct model with normal residuals, the deleted-*t* residuals,  $t_{(j)}$ , have the theoretical advantage in that they should follow a *t*-distribution with  $(n-1)-p$  degrees of freedom. Simple boxplots of the residuals can also indicate if the distribution of the residuals is skewed and, therefore, do not follow a normal distribution. There are formal statistical tests for normality that are discussed elsewhere in the text; however, small deviations from these assumptions are not likely to affect results and only severe violations of the normality of the residuals should be cause to abandon or change the regression model (Montgomery and others, 2012). For this reason, a graphical assessment of the values plotted on a normal probability plot is generally adequate.

#### Example 9.7. Assessing normality of the residuals.

The most common way to assess the normality of the residuals in a regression model is to plot the residuals on a normal probability plot:

```
> plot(cuya.lm, which = 2, ask = FALSE)
```



The resulting plot shows the standardized residuals plotted as circles, and the dashed line indicates where they would fall if the residuals were normally distributed. We see that this is not the case for all data points. There are several positive residuals that are far higher than we would expect given the sample size. In a case like this, we need to think carefully if we plan to use the regression results for hypothesis testing, prediction, or confidence intervals (table 9.2).

### 9.4.2.3 Assessing Serial Correlation of the Residuals

Another assumption of regression is that the residuals,  $e_i$ , are independent (table 9.2). Many hydrologic datasets on which regression is performed are actually pairs of data observed at the same time, for example precipitation and flow, flow and concentration, concentration of one constituent versus concentration of another. These individual time series (flow, concentration, precipitation) often exhibit some level of serial correlation, which is the correlation of one observation point in time with another observation point of the same series at some time apart. The important question for a regression application is whether the residuals are serially correlated. When two time series are regressed against one another and 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 estimates that provide the smallest possible variance (in statistical terms, this means the estimates of the coefficients are no longer the most efficient estimates possible), though they remain unbiased.
2. The value of  $s^2$  may seriously underestimate the true  $\sigma^2$ . This means that the test statistics underlying the hypothesis tests are not correctly estimated and the confidence and prediction intervals will be incorrect as well.

### 9.4.2.4 Detection of Serial Correlation

Correlation between residuals over time will not be evident from the  $e_i$  versus  $\hat{y}_i$  residuals plot but will stand out on a plot of  $e_i$  versus time. If there is a tendency for the residual values to clump in such a plot—positive residuals tending to follow positive residuals, negative residuals tending to follow negative residuals—this may mean there is serial dependence present in the residuals. The clumping could arise for four different reasons: (1) the presence of a long-term trend or cyclic patterns in the relation between the two variables, (2) dependence on some other serially correlated variable which was not used in the model, (3) serial dependence of residuals, or (4) some combination of these reasons. These correlation issues can also exist when the observations are distributed spatially. In these cases, a plot of  $e_i$  versus the spatial coordinate system can be used to identify spatial correlation patterns. This text will not deal with these spatial issues; they are similar to the temporal issues considered here, but the tools to detect and address this issue are more complex.

One can use graphical methods to explore the possibility that the residuals vary as a function of time (including time of day or season of the year). It is often the case that what may appear to be serial correlation is actually an artifact of these time or trend components. A good residuals pattern, one with no relation between residuals and these measures of time, will look like random noise. If time is measured as a categorical variable (for example, month or season), boxplots of residuals by category can be evaluated for patterns of regularity. If we see that the residuals are related to these time-related variables or categories, then more sophisticated modeling approaches may be called for (see chaps. 11 and 12) to attempt to remove this source of variability before dealing with any remaining serial correlation.

If plots of residuals versus time show a clumpy behavior (in other words, long runs of consecutive positive residuals alternating with long negative runs) there may be an additional explanatory variable that can remove this pattern (see chap. 12). The residuals from these new regressions can be plotted again to see what effect the additional variables had. The use of multiple explanatory variables is explained in chapter 11.

Serial dependence is generally quantified by the correlation coefficient between a data point and its adjacent point. The procedure described here is only appropriate when the data collected are truly a time series, that is, observations are equally spaced in time (for example, hourly, daily, weekly, monthly, or annual)

although it can be applied if there is a modest departure from regularity of sample collection (for example, if samples are approximately weekly but sometimes the spacing between samples is 5 days, or 6 days, or 8 days and so forth, or if the sampling is regular but there are a modest number of missing values). A lag-one serial correlation coefficient can be used to assess serial dependence. This is typically done using the Pearson's correlation coefficient, but a more robust approach to evaluating serial dependence could be done with Kendall's or Spearman's correlation coefficients (see chap. 8). To compute whether this serial dependence is in fact significant for the case of a regularly spaced time series

1. Compute the regression between  $y$  and  $x$ ;
2. Order the resulting residuals by the relevant time variable  $t_1$  to  $t_n$ ;
3. Offset or lag the vector of residuals to form a second vector, the lagged residuals;
4. The residuals pairs then consist of  $(e_i, e_{i-1})$  for all  $t_i$  from  $t_2$  to  $t_n$ ; and
5. Compute a measure of correlation between the pairs  $(e_i, e_{i-1})$ .

In R, this can be done as follows. Assume that the vector,  $e$ , is the time-ordered set of residuals. The lag-one serial correlation can be computed (and the associated hypothesis test done) with the commands

```
> n <- length(e)
> cor.test(e[2 : n], e[1 : n - 1], method = "kendall")
```

The words `spearman` or `pearson` could also be specified in the `method` argument to use the Spearman's  $\rho$  or Pearson's  $r$  correlation, respectively.

If the correlation is significant, the residuals are likely not independent, violating one of the assumptions listed in the last column of table 9.2. This does not mean that the regression analysis should not be carried through, but it does limit the kinds of questions the regression analysis can address.

A related statistic to assess serial correlation of the residuals is the Durbin-Watson statistic (Durbin and Watson, 1950). 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  $d < d_L$ . The value of  $d_L$  depends on the size of the dataset, the number of explanatory variables, and  $\alpha$ . However, a low value of  $d$  will not give any clue as to its cause. Thus, the graphical approach is vital, and the test should only be used as a check. The Durbin-Watson statistic requires data to be evenly spaced in time and with few missing values. Thus, the statistic is not always ideally suited for use with environmental data. In R, the function for this test is `durbinWatsonTest` in the `car` package (Fox and Weisberg, 2011).

The important point here is that when residuals show strong serial correlation the hypothesis tests and confidence intervals used in regression are no longer valid (see table 9.2). The information content of the dataset is lower than what we would derive from the sample size, and therefore uncertainties are larger than what is suggested by the dataset alone. This can become particularly problematic for datasets collected at daily time steps or at hourly or shorter time steps. If we considered a dataset of TDS concentrations and discharges collected at a daily interval for 20 years, we would have about 7,300 samples, but doing any of the computations for significance tests or confidence intervals using an  $n$  value of 7,300 would grossly overstate the true degree of certainty in our results.

The effect of serial correlation on estimation and hypothesis testing has a long history in the field of hydrology and is commonly referred to as “equivalent independent sample size” (Matalas and Langbein, 1962; Lettenmaier, 1976). For datasets that have high frequency sampling these issues become very important and require the use of time series methods that are beyond the scope of this book.

### 9.4.2.5 Strategies to Address Serial Correlation of the Residuals

There are several ways to address the presence of serial correlation in the residuals in a linear-regression framework:

1. Sample from the dataset. For example, if the dataset 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 of information is lost in doing this.
2. Group the data into time periods (for example, weeks or months) and compute a summary statistic for the period, such as a time- or volume-weighted mean or median, and then use these summary statistics in the regression. This tactic should only be applied when the sampling frequency has remained unchanged over the entire period of analysis. This ensures that each summary statistic was computed using the same number of observations for each period and, therefore, that the summary statistic has approximately equal variance across each period.
3. If a pattern over time is evident, add additional terms to the regression equation to account for seasonality or a long-term trend (see chap. 12).

Much more sophisticated approaches such as transfer function models (Box and others, 2015), regression with autoregressive errors (Johnston, 1984), or generalized additive models (GAMs) (Wood, 2017), are beyond the scope of this text and not discussed here.

### 9.4.3 A Measure of the Quality of a Regression Model Using Residuals: PRESS

One of the best measures of the quality of a regression equation is the PRediction Error Sum of Squares (PRESS statistic):

$$PRESS = \sum_{i=1}^n e_{(i)}^2 . \quad (9.16)$$

PRESS is a validation-type estimator of error that uses the deleted residuals to provide an estimate of the prediction error. When comparing alternate regression models, selecting the model with the lowest value of the PRESS statistic is a good approach because it means that the equation produces the least error when making new predictions. It is particularly valuable in assessing multiple forms of multiple linear regressions (see chap. 11), but it is also useful for simply comparing different options for a single explanatory variable in SLR.

## 9.5 Confidence and Prediction Intervals on Regression Results

If all assumptions in table 9.2 are met, one can then compute a confidence interval and prediction interval for a specified value of  $x$ .

### 9.5.1 Confidence Intervals for the Mean Response

If  $x_0$  is a specified value of  $x$ , then the estimate of the conditional mean of  $y_0$  at  $x_0$  is obtained by simply substituting  $x_0$  into the regression equation:  $\hat{y}_0 = b_0 + b_1 x_0$ , the value predicted from the regression equation. SLR estimates the mean response of  $y$  given  $x$  and, because the regression model is an estimate of the true relation between  $x$  and  $y$ , there is some uncertainty in the resulting values of  $\hat{y}$  which arise owing to the uncertainty in the regression parameters  $\beta_0$  and  $\beta_1$ .

The  $(1-\alpha) \cdot 100$  percent confidence interval for the condition mean,  $\hat{y}_0$ , for a specified  $x_0$  is then

$$\left( \hat{y}_0 - t_{\left(\frac{\alpha}{2}, n-1\right)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}, \hat{y}_0 + t_{\left(\frac{\alpha}{2}, n-1\right)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}} \right), \quad (9.17)$$

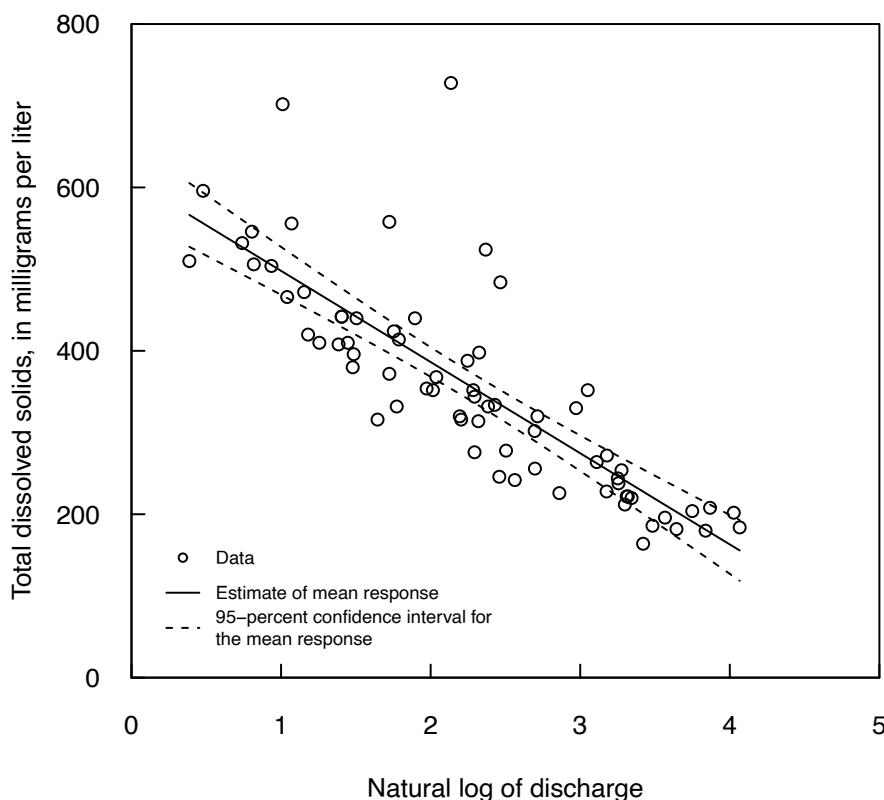
where  $t$  is the quantile of the Student's  $t$ -distribution having  $n-2$  degrees of freedom with probability of exceedance of  $\alpha/2$ . The other variables are the same as those defined in table 9.1. The 95-percent confidence interval is shown in figure 9.9. It is two-sided and symmetric around  $\hat{y}_0$ . Also note from the formula that the farther  $x_0$  is from  $\bar{x}$ , the wider the interval becomes. This is because the model will have less uncertainty near the central location of the  $x$  values than at the extremes. Note that this bow shape of the confidence interval shown in figure 9.9 agrees with the pattern seen in figure 9.3 for randomly generated regression lines, where the positions of the line estimates are more tightly clustered near the center than near the ends.

#### Example 9.8. Computing the confidence interval.

To continue with the Cuyahoga River data from example 9.1, the 95-percent confidence interval for the mean response,  $y$ , is calculated for two values of  $\ln(x_0)$ , 3.05 (a value near the mean computed from the natural log of values,  $\frac{\sum_{i=1}^n \ln(x_i)}{n}$ , which is 2.29) and 4.03 (a value much farther from the mean). Recall that values of  $x$  are the natural logarithm of the discharge and not the discharge values themselves because we transformed the  $x$  data. The sample size,  $n$ , is 70.

Written out, the confidence intervals for  $\ln(x_0)$  3.05 (observation #37 in the Cuyahoga River dataset) are

$$\begin{aligned} & \left( 269.03 - 1.99 \cdot 72.04 \sqrt{\frac{1}{70} + \frac{(3.05 - 2.29)^2}{61.85}}, \right. \\ & \quad \left. 269.03 + 1.99 \cdot 72.04 \sqrt{\frac{1}{70} + \frac{(3.05 - 2.29)^2}{61.85}} \right) = (246.94, 291.12) \text{ mg/L}, \end{aligned}$$



**Figure 9.9.** Plot of 95-percent confidence intervals for the mean total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1.

and for  $\ln(x_0) = 4.03$  (observation #62 in the Cuyahoga River dataset) are

$$\left( \begin{array}{l} 160.06 - 1.99 \cdot 72.04 \sqrt{\frac{1}{70} + \frac{(4.03 - 2.29)^2}{61.85}}, \\ 160.06 + 1.99 \cdot 72.04 \sqrt{\frac{1}{70} + \frac{(4.03 - 2.29)^2}{61.85}} \end{array} \right) = (123.6, 196.5) \text{ mg/L.}$$

Notice that the confidence interval has a width of approximately 44 mg/L at  $\ln(x_0) = 3.05$ , and a width of approximately 72 mg/L at  $\ln(x_0) = 4.03$ .

In R, the predict function can be used to obtain the confidence intervals. Using the stored cuya.lm regression model and the Cuyahoga River dataset, one could enter the following for  $\ln(x_0) = 3.0$  (observation #37 in the Cuyahoga River dataset):

```
> data.point.far <- cuya.tds$discharge cms[37]
> predict(cuya.lm,
+         newdata = data.frame(discharge cms = xdata.point.far),
+         interval = "confidence", level = 0.95)
      fit      lwr      upr
1 269.0263 246.7076 291.3449
```

and for  $\ln(x_0) = 4.03$  (observation #62 in the Cuyahoga River dataset):

```
> xdata.point.close <- cuya.tds$discharge cms[62]
> predict(cuya.lm,
+         newdata = data.frame(discharge cms = xdata.point.close),
+         interval = "confidence", level = 0.95)
      fit      lwr      upr
1 160.0602 123.6231 196.4974
```

### 9.5.2 Prediction Intervals for Individual Estimates of $y$

The prediction interval for  $y_0$ , the  $y$ -value associated with a specified value of  $x_0$ , the confidence interval for prediction of an estimate of an individual point, is often confused with the confidence interval for the conditional mean. This is not surprising, as the same regression model equation is used to obtain the best estimate for both the conditional mean of  $y$  given  $x_0$  and for an individual  $y_0$  value given  $x_0$ . However, in addition to uncertainties in the parameter estimates  $\beta_1$  and  $\beta_2$ , the prediction interval includes an extra term that incorporates the unexplained variability in  $y$ . The conditional mean results in an estimate of  $y$  given the  $x$  values used to develop the regression; prediction intervals provide an estimate of a future  $y_0$  value given the  $x_0$  value that has been observed. This means that the variability of both  $x$  and  $y$  need to be accounted for in the prediction interval computation. The  $(1-\alpha)$ -100 percent prediction interval for a specified  $x_0$  value is

$$\left( \hat{y}_0 - t_{\left(\frac{\alpha}{2}, n-2\right)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}, \hat{y}_0 + t_{\left(\frac{\alpha}{2}, n-2\right)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}} \right), \quad (9.18)$$

where  $t$  is the quantile of the Student's  $t$ -distribution having  $n-2$  degrees of freedom with probability of exceedance of  $\alpha/2$ . The other variables are the same as those defined in table 9.1. The term inside the square root is the same as in the confidence interval formula, except for the addition of the value 1 in the

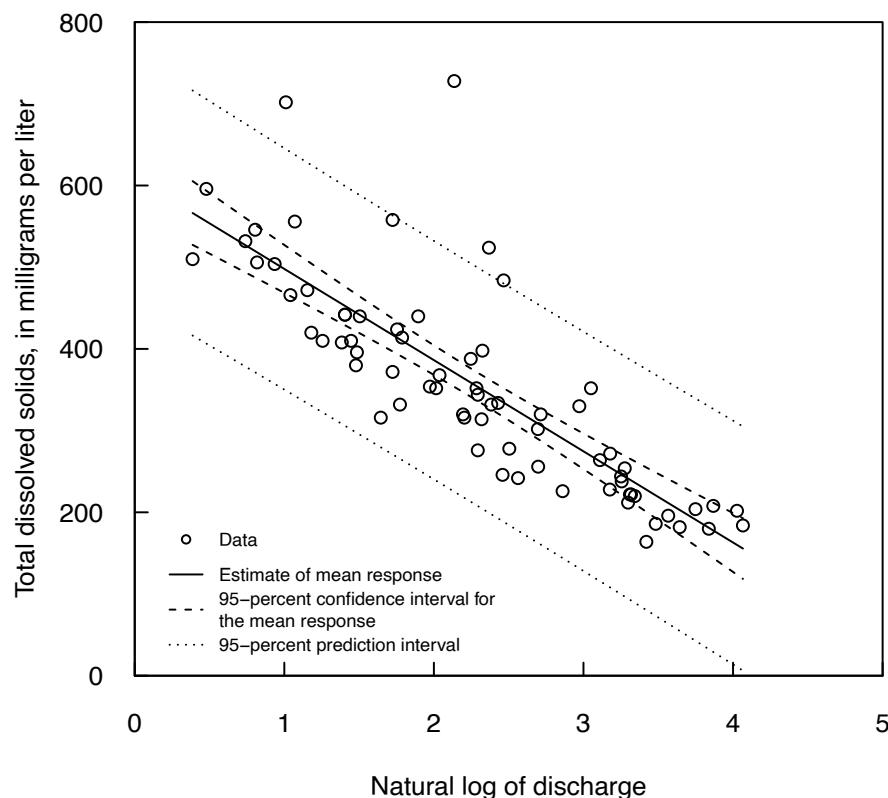
prediction interval formula, which adds a width of  $s$  to the overall width of the interval. This term increases the interval by the amount of the unexplained variability in  $y$ .

Note that the intervals widen as  $x_0$  departs from the center of the  $x$  values (fig. 9.10), but not nearly as markedly as the confidence intervals do. This is because the second and third terms inside the square root are small in comparison to the first term, provided the sample size is large. In fact, for large sample sizes a simple rough approximation to the prediction interval is just  $(\hat{y}_0 - t_{(\frac{\alpha}{2}, n-2)} \cdot s, \hat{y}_0 + t_{(\frac{\alpha}{2}, n-2)} \cdot s)$ , two parallel straight lines. The prediction intervals should contain approximately  $1-\alpha \cdot 100$  percent of the data within them, with  $\frac{\alpha}{2} \cdot 100$  percent of the data beyond each side of the intervals. They will do so if the residuals are approximately normal, independent of  $x$ , and homoscedastic.

### Example 9.9. Computing prediction intervals.

The 95-percent prediction intervals for the Cuyahoga River dataset from example 9.1 are computed below for  $\ln(x_0) = 3.05$  and  $4.03$ . For  $\ln(x_0) = 3.05$  (observation #37 in the Cuyahoga River dataset):

$$\begin{cases} 269.03 - 1.99 \cdot 72.04 \sqrt{1 + \frac{1}{70} + \frac{(3.05 - 2.29)^2}{61.85}}, \\ 269.03 + 1.99 \cdot 72.04 \sqrt{1 + \frac{1}{70} + \frac{(3.05 - 2.29)^2}{61.85}} \end{cases} = (122.5, 415.5) \text{ mg/L},$$



**Figure 9.10.** Plot of 95-percent prediction intervals for an individual estimate of total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1. Confidence intervals are shown for reference.

and for  $\ln(x_0)=4.03$  (observation #62 in the Cuyahoga River dataset)

$$\begin{cases} 160.06 - 1.99 \cdot 72.04 \sqrt{1 + \frac{1}{70} + \frac{(4.03 - 2.29)^2}{61.85}}, \\ 160.06 + 1.99 \cdot 72.04 \sqrt{1 + \frac{1}{70} + \frac{(4.03 - 2.29)^2}{61.85}} \end{cases} = (10.7, 309.4) \text{ mg/L.}$$

The 95-percent prediction interval has a width of approximately 293 mg/L at  $\ln(x_0)=3.05$  and a width of approximately 299 mg/L at  $\ln(x_0)=4.03$ . Notice 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. In R, the predict function can again be used to obtain the prediction intervals. Using the stored cuya.lm regression model and the Cuyahoga River dataset, one could enter the following for  $\ln(x_0)=3.05$  (observation #37 in the Cuyahoga River dataset)

```
> xdata.point.far <- cuya.tds$discharge_cms[37]
> predict(cuya.lm,
+   newdata = data.frame(discharge_cms = xdata.point.far),
+   interval = "prediction", level = 0.95)
      fit      lwr      upr
1 269.0263 122.5077 415.5449
```

and for  $\ln(x_0)=4.03$  (observation #62 in the Cuyahoga River dataset)

```
> xdata.point.close <- cuya.tds$discharge_cms[62]
> predict(cuya.lm,
+   newdata = data.frame(discharge_cms = xdata.point.close),
+   interval = "prediction", level = 0.95)
      fit      lwr      upr
1 160.0602 10.7376 309.3829
```

The small differences between the R results and the results obtained by hand are because of rounding.

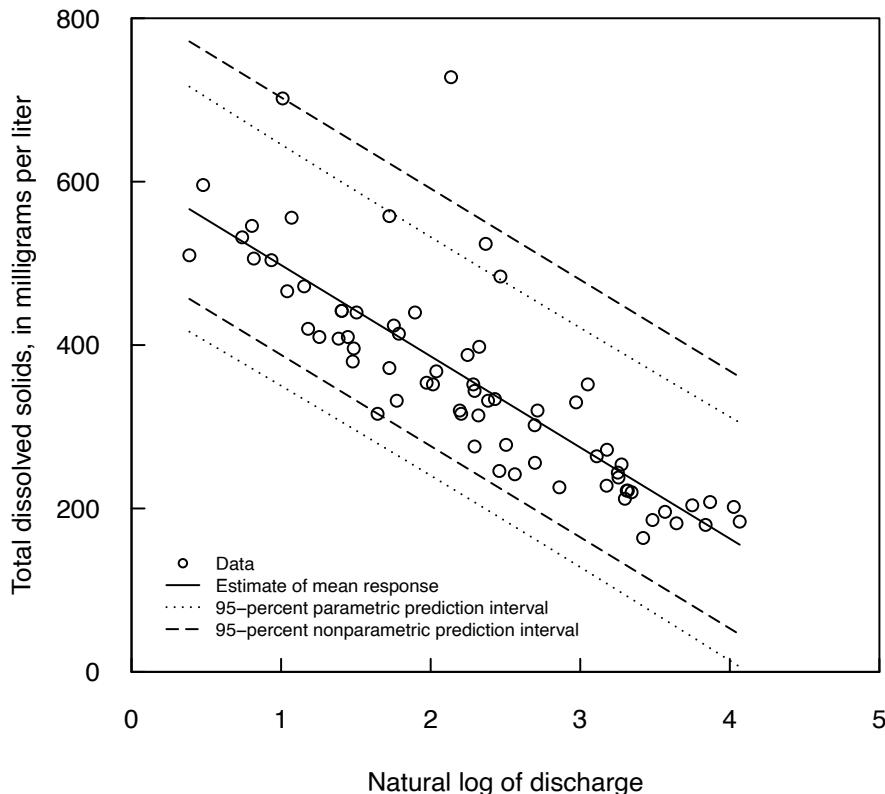
Note that all of the observations that are outside the prediction interval lie above the interval and none fall below it. This happens because the residuals for this model are clearly not normal (as we saw in section 9.4.2.2.). The prediction intervals computed here are not good representations of the behavior of the data because the assumptions of normality are not met (table 9.2). The next section provides one approach for developing a more robust prediction interval for cases like this.

### 9.5.2.1 Nonparametric Prediction Interval

There is a nonparametric version of the prediction interval, which can be used when the  $x, y$  data display a linear relation and residuals have constant variance (homoscedastic), but the distribution of the residuals appears non-normal. Typically, such departures from normality result in an excessive number of outside or far outside values or in the asymmetry of the distribution of residuals. Parametric prediction intervals are not able to capture these behaviors.

The nonparametric prediction interval is

$$(\hat{y}_0 + e_L, \hat{y}_0 + e_U) \quad (9.19)$$



**Figure 9.11.** Plot of 95-percent parametric and nonparametric prediction intervals for an individual estimate of total dissolved solids concentration resulting from the regression model fit between total dissolved solids and the log of discharge for the Cuyahoga River data from example 9.1.

where

- $e_L$  is the  $1-\alpha/2$  (lower limit) of the ranked residuals, and
- $e_U$  is the  $\alpha/2$  (upper limit) of the ranked residuals.

To compute the nonparametric prediction intervals, order the residuals from smallest to largest and assign a rank to each value (with the smallest value having rank 1 and the largest value having rank  $n$ ). Then compute the rank associated with  $e_L$  and  $e_U$  using the formulas:  $L=(n+1)\cdot\alpha/2$  and  $U=(n+1)\cdot(1-\alpha/2)$ . Choose the  $e_L$  and  $e_U$  values associated with the ranks  $L$  and  $U$ . When  $L$  and  $U$  are not integers, either the integer values closest to  $L$  and  $U$  can be chosen or the values of  $e_L$  and  $e_U$  can be interpolated between adjacent residuals. In figure 9.11, the nonparametric prediction interval is compared to the one previously developed assuming normality of residuals. Note that  $e_L$  and  $e_U$  will not change for any value of  $x_0$  once  $\alpha$  is chosen and the calculation of  $e_L$  and  $e_U$  is only dependent on  $n$  and  $\alpha$ . For this reason, the nonparametric interval follows parallel to the  $\hat{y}$  values and the intervals are asymmetric around the central regression line, reflecting the asymmetry of the residuals themselves.

#### Example 9.10. Computing nonparametric prediction intervals.

The 95-percent nonparametric prediction intervals for the Cuyahoga River dataset in example 9.1 are computed below for  $n=70$  and  $\ln(x_0)=3.05$  and  $4.03$ . Values for  $L$  and  $U$  are calculated as

$$L = (70+1)\cdot\frac{0.05}{2} = 1.78$$

$$U = (70+1)\cdot\left(1 - \frac{0.05}{2}\right) = 69.2$$

Either the 1st- and 69th-ranked residual can be selected, or values interpolated between the 1st- and 2nd-ranked residuals and the 69th- and 70th-ranked residuals. In R, the function `as.integer` will select the integer values of  $L$  and  $U$ :

```
> # This line saves the residuals from the regression model
> cuya.residuals <- as.vector(resid(cuya.lm))
> # This line sorts the residuals and stores them
> sort.residuals <- sort(cuya.residuals)
> # Compute L and U
> L <- as.integer((70 + 1) * (0.05 / 2))
> U <- as.integer((70 + 1) * (1 - (0.05 / 2)))
```

The values of  $L$  and  $U$  are then added to the estimate of  $\hat{y}_0$  at  $x_0$  for  $x_0 = 3.05$  with prediction intervals of (50.16, 365.30).

```
> 269.03 + sort.residuals[L]
[1] 159.1353
> 269.03 + sort.residuals[U]
[1] 474.2847
```

with prediction intervals of (159.14, 474.28) and for  $x_0 = 4.03$

```
> 160.06 + sort.residuals[L]
[1] 50.16534
> 160.06 + sort.residuals[U]
[1] 365.3047
```

## 9.6 Transformations of the Response Variable, $y$

The primary reason to transform the response variable,  $y$ , is because the residuals are heteroscedastic—their variance 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. In this case, a transformation could improve the behavior of the residuals. The two topics that require careful attention when transforming  $y$  are

1. Deciding if the transformation is appropriate, and
2. Interpreting the resulting estimates.

### 9.6.1 To Transform or Not to Transform

The decision to transform  $y$  should generally be based on graphs. The first step is to develop the best possible nontransformed model. The next step is to apply the methods outlined in section 9.4 to examine the behavior of the residuals: plot  $e_i$  versus  $\hat{y}_i$  to check for heteroscedasticity, make a probability plot for  $e_i$  to check for normality, and examine the function for unreasonable results such as predictions of negative values for variables for which negative values are physically impossible. If serious problems arise in any of these tests, then transformation of  $y$  according to the bulging rule and ladder of powers should proceed and the behavior of the residuals reevaluated. If both the transformed and untransformed scales have problems,

then either look for a different transformation or accept the limits of the regression model according to the assumptions met in table 9.2.

There are additional methods available to select the best transformation of  $y$ . This involves testing of a series of transformations and choosing the transformation that maximizes the probability plot correlation coefficient (PPCC) of the regression residuals. The implementation of this approach is beyond the scope of this text; however, regardless of the method used to transform the data, the final choice should be made only after looking at residuals plots, as described in section 9.4. Attempting to find the ideal transformation is not appropriate and one should be satisfied with simple ones such as the logarithm, square root, or reciprocal.

It is important to note that in the case of transformed  $y$ -variables, comparisons of  $R^2$ ,  $s$ , or  $F$ -statistics between transformed and untransformed models (or between two different transformations) should not be used to choose among them. Each model is attempting to predict a different variable (for example,  $y$ ,  $\log(y)$ , or  $1/y$ ). The above statistics measure how well these different dependent variables are predicted and so cannot be directly compared. Instead, the appropriate response variable is one that best fits the assumptions of regression (table 9.2). Once a hydrologist has developed some experience with certain kinds of datasets, it is quite reasonable to go directly to the appropriate transformation without a lot of investigation. One helpful generalization in deciding to transform the  $y$  variable is that any inherently positive  $y$  variable that covers more than an order of magnitude of values in the dataset probably needs to be transformed.

## 9.6.2 Using the Logarithmic Transform

A logarithmic transformation is the most common transformation used in hydrology. A common form of this transformation is to use the natural logarithm on both  $x$  and  $y$ :

$$\ln(y) = b_0 + b_1 \ln(x) + \varepsilon ,$$

where  $\ln$  is the natural log. Often in hydrology, a regression is performed on the log-transformed values; however, results need to be communicated in the original units. When this equation is expressed in terms of  $y$ , notice that the terms become multiplicative and not additive:

$$y = e^{b_0} x^{b_1} e^\varepsilon .$$

In transforming from logarithmic units back to the original units, there are important properties to note, the implications of which are described below and in the following sections.

- When the  $y$  values are appropriately log-transformed, the conditional distribution will be approximately normal and, for this reason, the mean and median of the log-transformed variable will be approximately equal to one another. However, this is not the case when the values are transformed back to their original units. This is because the transformed values follow a conditional normal distribution even though the original values do not. When back-transformed to the original units, the conditional mean will be greater than the conditional median because data that are log-normally distributed are positively skewed (the right tail is longer than the left).
- In regressions, the regression line estimates the mean response. Therefore, when the values are log-transformed, the regression equation estimates the mean response of  $\ln(y)$  conditioned on the value of  $\ln(x)$ . The resulting equation does not estimate the mean response of  $y$  conditioned on  $x$  but rather, the median response of  $y$  conditioned on  $x$ . When the dependent variable is transformed back from log space to real space, the resulting conditional distribution of  $y$  will be skewed to the right (because it follows an approximately log-normal distribution) and as such the mean will be greater than the median. Thus, simply back-transforming the estimates from log space to real space will result in underestimation of the true conditional mean.

## 9.6.3 Retransformation of Estimated Values

Simply transforming estimates from a log-regression equation back into the original units provides estimates of the medians of the  $y$  values, which are underestimates of the conditional means. This is

particularly important when mass conservation is relevant, such as when the estimates of  $y$  are to be summed to obtain an estimate of an aggregate value. For example, when summing daily estimates of a constituent flux to obtain a monthly flux value, the conditional mean is preferred. Ferguson (1986) points out for some very realistic cases where using this estimate for sediment loads will result in underestimates of the aggregated flux by as much as 50 percent.

Note that in some cases, the median value may be the appropriate measure of the typical response, but for cases where an estimate of the mean is deemed to be the preferred quantity, one must determine how to compensate or adjust for the back-transformation bias. The two methods described below attempt to correct for this bias of the estimate. The first method assumes that the log-transformed values follow a normal distribution, and the second method provides a nonparametric method that does not assume a distribution of the log-transformed values.

### 9.6.3.1 Parametric or the Maximum Likelihood Estimate (MLE) of the Bias Correction Adjustment

If the residuals in the log-transformed units are known to be normal and the parameters of the fitted regression model ( $\beta_0, \beta_1, \sigma^2$ ) are known without error, the theory of the lognormal distribution (Aitchison and Brown, 1981) provides the following results:

$$\begin{aligned} \text{Median of } L \text{ given } Q_0 &= \exp^{\beta_0 + \beta_1 \ln Q_0} = L_m \\ &= \exp^{\beta_0} Q_0^{\beta_1} \end{aligned} \quad (9.20)$$

$$\begin{aligned} \text{Mean of } L \text{ given } Q_0 &= E[L|Q_0] = \exp^{\beta_0 + \beta_1 \ln Q_0 + 0.5\sigma^2} \\ &= L_m \exp^{0.5\sigma^2} \end{aligned} \quad (9.21)$$

$$\begin{aligned} \text{Variance of } L \text{ given } Q_0 &= \text{Var}[L|Q_0] \\ &= (L_m \exp^{0.5\sigma^2})^2 \exp^{\sigma^2 - 1}. \end{aligned} \quad (9.22)$$

These equations would differ if base 10 logarithms were used, as was done in the first paper to address this topic in hydrology (Ferguson, 1986), but our presentation assumes natural logarithms.

Unfortunately, the true population values  $\beta_0, \beta_1$ , and  $\sigma^2$  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, which we call the maximum likelihood estimate (MLE), is then

$$\hat{L}_{MLE} = \exp^{b_0 + b_1 \ln Q_0 + 0.5s^2}. \quad (9.23)$$

When  $n$  is large ( $>30$ ) and  $s^2$  is small ( $<0.25$ ),  $\hat{L}_{MLE}$  is a very good approximation. However, when  $n$  is small or  $s^2$  is large, it can overestimate the true mean, in other words 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 owing to the complexity of the formula, but it has been implemented in the R package EnvStats (Millard, 2013) in the function `elnormAlt` and is denoted as the minimum variance unbiased estimate (MVUE) method. Its properties are discussed in Cohn (1988). This method has also been extended by Cohn (2005) to address the issue of censored data, a specific form of categorical data common in water quality sampling and is implemented in the software package LOADEST (Runkel and others, 2004).

### 9.6.3.2 Nonparametric or Smearing Estimate of the Bias Correction Adjustment

Even with the improvements to the parametric MLE method provided by the MVUE method, the validity of any parametric approach depends on normality of the residuals. There is an alternative approach that only requires the assumption that the residuals are independent and homoscedastic; they may follow

any distribution as well. This approach is the smearing estimate of Duan (1983). In the case of the log transform it is

$$\hat{L}_D = \exp^{b_0 + b_1 \ln Q_0} \frac{\sum_{i=1}^n \exp[e_i]}{n}. \quad (9.24)$$

The smearing estimate 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 each  $x_0$ . Even when the residuals in log units are normal, the smearing estimate performs nearly as well as Bradu and Mundlak's (1970) unbiased MVUE estimator. It also 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 estimate 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 (for example, square root, inverse, or log), then

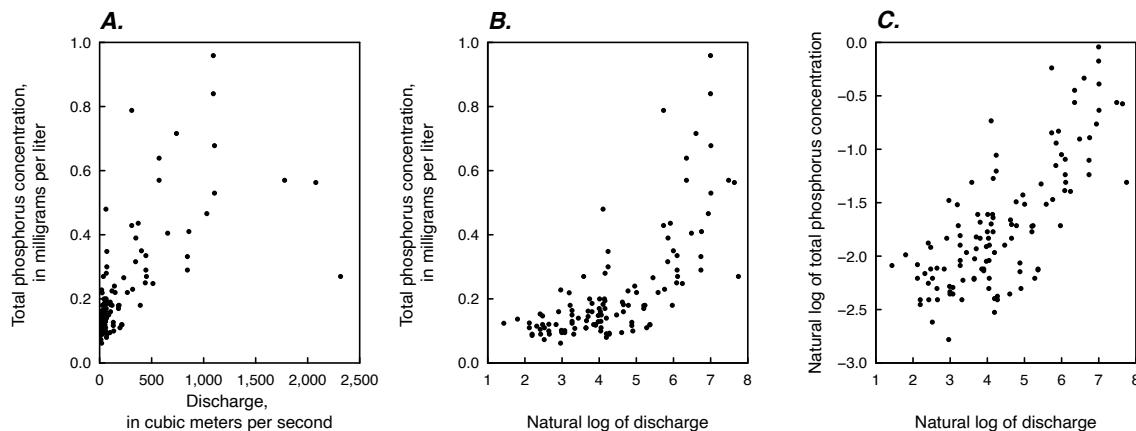
$$\hat{y}_D = \frac{\sum_{i=1}^n f^{-1}(b_0 + b_1 x_0 + e_i)}{n}, \quad (9.25)$$

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 (for example, square, inverse, or exponential, respectively), and  $x_0$  is the specific value of  $x$  for which we want to estimate  $y$ . Unlike the MVUE approach, no way has been developed to generalize the smearing estimate to work in cases where there are censored data.

#### Example 9.11. Log transformation of $y$ .

In previous examples in this chapter using the Cuyahoga River total dissolved solids dataset, only the  $x$  values were log-transformed. To show the potential impact of the transformation bias issue we will switch to a different dataset: total phosphorus (TP) concentrations for the Maumee River (Waterville, Ohio). The dataset consists of 117 samples collected by the U.S. Geological Survey from October 2010 through December 2015. The data are shown in figure 9.12A with neither discharge nor concentration transformed, then with a logarithm transformation on the discharge only (fig. 9.12B), and finally with a logarithm transformation on both concentration and discharge (fig. 9.12C).

With no transformation (fig. 9.12A) of either variable, the data strongly violate the requirements for linearity and homoscedasticity. When only the  $x$  values are transformed (fig. 9.12B) the data continue to



**Figure 9.12.** Comparison of the relation between discharge and total phosphorus concentration for the Maumee River (Waterville, Ohio), in original units (A), with a logarithmic transformation of the discharge values (B), and with a logarithmic transformation of the discharge and total phosphorus concentration values (C).

show considerable curvature and heteroscedasticity. However, when both the  $x$  and  $y$  values are transformed (fig. 9.12C), it appears that the data exhibit fewer and less extreme outliers, the errors appear roughly homoscedastic but the relation still exhibits some curvature. We will proceed with estimating a regression relation using transformed  $x$  and  $y$  values. In R, the regression resulting from the transformation of both the  $x$  and  $y$  values is shown in the following lines of code. The resulting model is called `tpMod`. Note that `tp` is the concentration of TP.

```
> load("maumeeTP.RData")
> tpMod <- lm(log(tp) ~ log(Q), data = maum)
> y(tpMod)

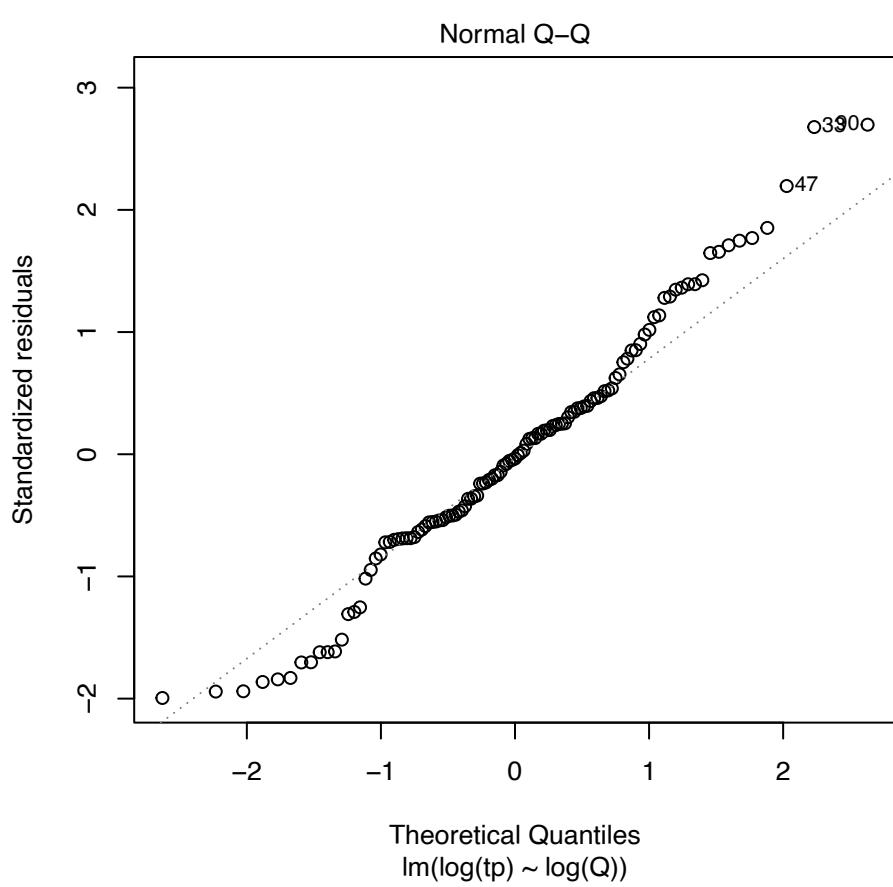
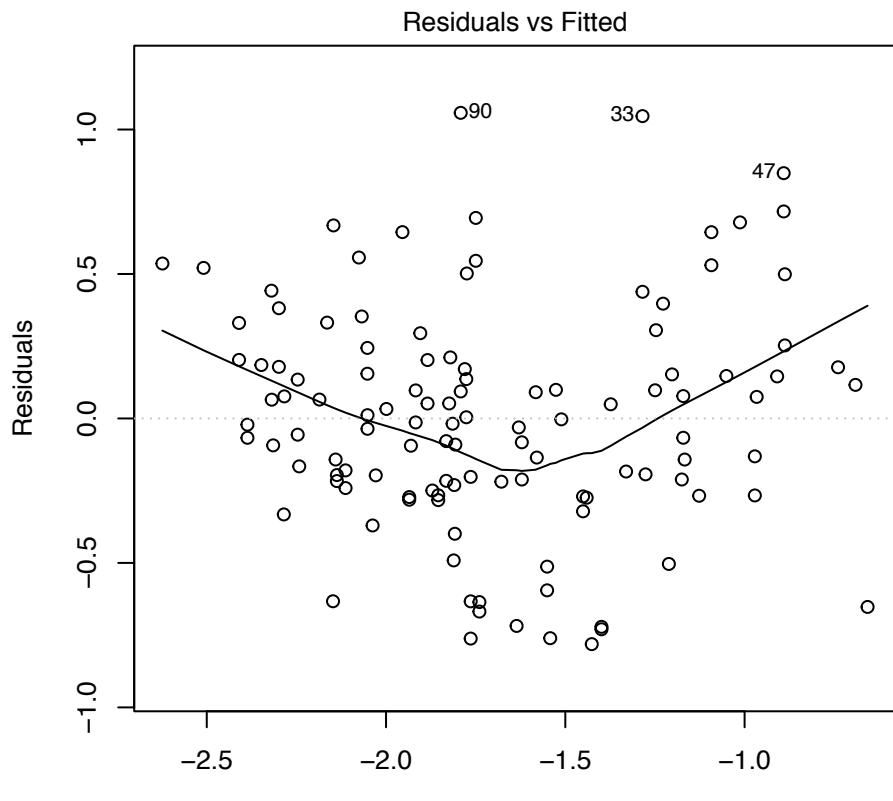
Call:
lm(formula = log(tp) ~ log(Q), data = maum)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.78108 -0.23052 -0.01374  0.20224  1.05776 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.07003   0.11484 -26.73   <2e-16 ***
log(Q)       0.31146   0.02487  12.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3939 on 115 degrees of freedom
Multiple R-squared:  0.577,    Adjusted R-squared:  0.5733 
F-statistic: 156.9 on 1 and 115 DF,  p-value: < 2.2e-16

> plot(tpMod, which = 1, ask = FALSE)
> plot(tpMod, which = 2, ask = FALSE)
```



The relation of the logarithm of concentration to the logarithm of streamflow is highly significant (note the  $t$ -value of 12.53). The residuals still show some problems: some curvature of the relation and a right tail of the distribution of the residuals that is somewhat thicker than a normal distribution, but the regression is reasonable and we will use it here. Methods described in the following two chapters could improve on it to a modest degree.

To illustrate the bias in total phosphorus concentrations, the mean resulting from each estimate is compared to the mean of the observed concentrations. We can complete these calculations in R using the results from the `tpMod` regression model.

The mean of the observed total phosphorus concentrations is 0.223 mg/L:

```
> mean(maum$tp)
[1] 0.2234103
```

We would like to have the estimates from the fitted model result in a mean value that is quite close to this true mean (recall that without a transformation of  $y$ , the mean of the estimates will exactly equal the mean of the observed  $y$  values). If one simply exponentiated the  $\hat{y}$  values and took the mean with no bias correction adjustment, the estimated mean concentration would be 0.202 mg/L:

```
> yhat <- exp(tpMod$fit)
> mean(yhat)
[1] 0.2023018
```

As expected, this is an underestimate of the mean of the observed data (about 10 percent below the observed mean). Applying the MLE estimate, the bias correction adjustment is a multiplier of 1.08 applied to each of the previous estimates, results in an estimated mean concentration of 0.219 mg/L:

```
> sse <- sum(tpMod$residuals^2)
> sSquared <- sse / (length(tpMod$residuals) - 2)
> biasAdj <- exp(0.5 * sSquared)
> biasAdj
[1] 1.080672
> yMLE <- yhat * biasAdj
> mean(yMLE)
[1] 0.2186219
```

This estimate is about 2 percent smaller than the observed mean. Applying the smearing estimate of the bias correction adjustment gives almost exactly the same result as the previous estimate, a mean concentration of 0.219 mg/L:

```
> smearAdj <- sum(exp(tpMod$residuals)) / length(tpMod$residuals)
> smearAdj
[1] 1.081462
> ysmear <- yhat * smearAdj
> mean(ysmear)
[1] 0.2187817
```

The fact that the smearing estimate and the MLE are so similar is because the residuals conformed well to a normal distribution. In this example either the MLE or the smearing estimate is a reasonable choice. The simple back transformation from the estimates in log space is not a good choice. If the residuals were strongly non-normal, then the smearing estimate would be preferred.

### 9.6.4 Special Issues Related to a Logarithm Transformation of $y$

When communicating about the quality of a regression, a common metric is the standard error of the residuals, the value denoted as  $s$  in table 9.1. In the usual case of regression where the  $y$  variable is not transformed, the units of  $s$  are the same as the units of  $y$ , and one can obtain an immediate understanding of the magnitude of the errors. But if  $y$  is the natural logarithm of the variable of interest, then  $s$  is the difference between the log of two numbers and, therefore, will be the log of the ratio of these two numbers. In this case, it is not easy to interpret the magnitude of the actual regression error. There is, however, another way to express  $s$  when the  $y$  variable is a log-transformation of the variable of interest. Assuming that the errors in the regression model of the natural logarithm of  $y$  are homoscedastic, we can express the standard error as a percent:

$$S_{pct} = 100 \cdot \sqrt{e^{s^2} - 1} \quad (9.26)$$

For example, if  $s$  is 0.1, then  $S_{pct}$  is 10.025 percent, and if  $s$  is 0.5 then  $S_{pct}$  is 53.29 percent. Given that for a normal distribution, about two-thirds of values are within plus or minus 1 standard deviation of the mean, then we can say that if  $s$  were 0.1, about two-thirds of the errors will be within a band of about 10 percent above or below their predicted values, and if  $s$  is 0.5 that about two-thirds of the errors will be within a band of about 53 percent above or below their predicted values.

Another question that comes up frequently is the appropriateness of using log-transformed regressions to compute river loads when discharge is the explanatory variable, for example, in equation 9.27:

$$\ln(L) = \beta_0 + \beta_1 \cdot \ln(Q) + \varepsilon \quad (9.27)$$

where the load  $L$  is defined as  $C \cdot Q \cdot k$  where  $C$  is concentration,  $Q$  is discharge, and  $k$  is a unit conversion (for example, if  $C$  were in mg/L and  $Q$  in cubic meters per second [ $m^3/s$ ] and  $L$  were in kilograms per day [kg/day], then  $k=86.4$ , which is the unit conversation needed to obtain kg/day). The argument is often made that this regression model is somehow inappropriate because the explanatory variable,  $Q$ , appears on both sides of the equation. This is often known as the problem of spurious correlation (Benson, 1965). The response to this concern is somewhat nuanced, depending upon the intended application of the regression equation.

**Table 9.3.** Comparison of results of regression of  $\ln(C)$  on  $\ln(Q)$  versus  $\ln(L)$  on  $\ln(Q)$ .

[ $C$ , concentration;  $Q$ , discharge;  $L$ , load]

Statistic	Model: $\ln(C)$ versus $\ln(Q)$	Model: $\ln(L)$ versus $\ln(Q)$
Intercept	-3.07	-1.39
Standard error of intercept	0.11	0.11
$t$ -value of intercept	-26.7	12.1
Slope	0.31	1.31
Standard error of slope	0.025	0.025
$t$ -value of slope	12.53	52.7
Residual standard error	0.39	0.39
$R^2$	0.58	0.96

Let's consider the total phosphorus data from the Maumee River that is discussed in example 9.11. We can compute  $L$ , in kg/day, and apply linear regression to estimate the natural logarithm of  $L$  as a function of the natural logarithm of  $Q$ .

```
> maum$L <- maum$Q * maum$tp * 86.4
> tpLoadMod <- lm(log(L) ~ log(Q), data = maum)
> summary(tpLoadMod)
```

Call:

```
lm(formula = log(L) ~ log(Q), data = maum)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.78108	-0.23052	-0.01374	0.20224	1.05776

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.38896	0.11484	12.10	<2e-16 ***
log(Q)	1.31146	0.02487	52.74	<2e-16 ***
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.3939 on 115 degrees of freedom

Multiple R-squared: 0.9603, Adjusted R-squared: 0.9599

F-statistic: 2781 on 1 and 115 DF, p-value: < 2.2e-16

Several statistics from the two regression models, (1) the model using  $Q$  to estimate  $C$ , and (2) the model using  $Q$  to estimate  $L$ , are shown in table 9.3.

Several aspects of table 9.3 are worth commenting on. The estimated slope coefficient of the second model is greater than the slope of the first model and they differ by exactly 1. In fact, all of the statistics of the second model can be derived from the first model. The  $t$ -values on both of the coefficients are different for the models because  $t$ -values are a ratio of the coefficient to its standard error, and the coefficients differ between the two models. Therefore, a test for the significance of the slope coefficient may have a different outcome for the two models (in this case, the slope coefficients are highly significant in both models).

One outcome that can arise from fitting one regression model to concentration versus streamflow and another model to load versus streamflow (but not in this case) is that we find a slope in the  $\ln(C)$  model that is not significantly different from zero, but the slope in the  $\ln(L)$  model is significantly different from zero. This makes sense if we imagine a case where concentration was truly unrelated to discharge, and it is only when we multiply the concentration by discharge to determine load, that the load is significantly related to discharge (the truly spurious case). Another outcome that can arise is when the slope coefficient in the  $\ln(C)$  model is near a value of -1 (this would happen if the relation between concentration and discharge was simply a dilution relation). This could be considered significant, but the slope of the  $\ln(Q)$  model would then be close to zero and it might turn out not to be significant. This underscores the importance of being specific about the question being asked in a hypothesis test before deciding which form of the model to use. The question of whether concentration is significantly related to discharge is different than whether load is significantly related to discharge. One simply needs to be specific about which question is being posed.

The residual standard errors are exactly the same ( $s=0.39$  in both models), the residual for each individual observation is exactly the same in both models, and the standard errors of the slopes are exactly the same. These measures of the quality of the regression will always be exactly the same for the  $\ln(L)$  model and the  $\ln(C)$  model. In fact, the percent standard error approach described above would result in the same answer. Using the `tpLoadMod` results we can compute the standard error in percent.

The mean square error is 0.155 and the standard error in percent is 40.97 percent for both models. One can verify the results using the residuals from the  $\ln(C)$  model (called `tpMod` in the previous example). In words, we can say the standard error of concentration estimates expressed in percent is 40.97 percent and the standard error of load estimates expressed in percent is also 40.97 percent.

The  $R^2$  values are very different in the two models (0.58 versus 0.96). Recall that  $R^2$  is a ratio of variance explained by the model to the total variance in the data. We can express  $R^2$  as

$$R^2 = \frac{Var(y) - s^2}{Var(y)}$$

In the case of the  $\ln(C)$  model, the  $Var(y)$  is 0.36, but the variance of  $\ln(L)$  values is much larger, 3.87. We might say that for estimating concentrations the model does only moderately well ( $R^2=0.58$ ), but for estimating loads the model has very good performance ( $R^2=0.96$ ). This result is not surprising, given that load is computed as a function of discharge. Therefore, the reporting of the  $R^2$  value of a  $\ln(L)$  model is not a very meaningful representation of the quality of the model, and only reporting the  $R^2$  value for such may undermine the credibility of the analysis being described. The best number to report for either model is the standard error in percent, in this case about 41 percent, which provides a clear indication that the errors in estimates of individual values are quite large, even though we can clearly show that the model improves estimates of either load or concentration, as compared to using no regression model at all.

To summarize this issue, fitting the  $\ln(L)$  model will look, to some readers of a report, as a form of cheating because  $L$  is, by definition, a function of  $Q$ , which means  $Q$  is both an independent and dependent variable in the regression. In actuality, there are good reasons to estimate  $L$  from  $Q$  using a regression equation, but the appearance of cheating is a good reason to avoid using it. Therefore, reporting is best done using the concentration model. Also, presenting scatter plots of  $\ln(C)$  versus  $\ln(Q)$  will make it easier to evaluate the quality of a model (for example, judging curvature or heteroscedasticity) because in plots of  $\ln(L)$  versus  $\ln(Q)$  the steepness of the curve can easily overwhelm more subtle effects that may be present.

## 9.7 Summary Guide to a Good SLR Model

The following is a brief guide to getting started on developing 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 ensure a good model. Many transformations can be rapidly checked with such statistics, but always look at a residual versus predicted plot before making a final decision. Transform  $x$  if the residuals plot appears nonlinear 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 linear, homoscedastic, and normal in its residuals.

The statistics  $R^2$ ,  $s$ , and  $t$ -statistics on  $\beta_0$  and  $\beta_1$  will not provide correct information for deciding if a transformation of  $y$  is required.

Should an estimate of the conditional mean of  $y$  be desired using SLR with transformed  $y$  units, the transformation bias must be compensated for by use of the smearing estimate or MLE estimate. However, no bias correction is required when an estimate of percentiles is the objective—the median of the natural  $\log(y)$  when exponentiated estimates the median of  $y$ , for example. Objectives where percentiles are of interest include estimating the conditional cumulative distribution function of  $y$ , or when a typical or median  $y$  is desired. This applies to any power function transformation of the data, as long as the residuals in transformed  $y$  units reasonably follow a normal distribution. Upon retransformation of the

regression fitted equation, fitted values estimate the median of  $y$ , though they will be biased estimates of the conditional mean of  $y$ . Upon retransformation of the percentiles of the assumed normal distribution of transformed values for a given  $x$ , the result is the cdf of  $y$  for that value of  $x$ .

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 guidelines are discussed in chapter 11.

## Exercises

1. Drainage area is often used to estimate a streamflow property of interest at an unmeasured river. This is accomplished by computing the streamflow property from a set of streamgages in the study area and then regressing the streamflow values against the drainage area contributing to the streamgage. This exercise asks you to explore the relation between drainage area and the Q90 streamflow value, the daily streamflow value expected to be exceeded about 90 percent of the time, and to determine if this relation has the potential to be a useful model for the northeastern United States. The Q90 is often used as a measure of low streamflow frequency in practice.
  - A. Determine if a linear relation between drainage area and Q90 is present. Try transformations of the two variables to see if this improves the relation.
  - B. Using the results from (A), create a regression equation that can be used to estimate Q90. Is drainage area a significant predictor of the Q90 streamflow? Provide evidence. Does the sign of the coefficient on drainage area match your intuition? Explain.
  - C. Is it appropriate to use the regression equation to make predictions?

Observation number	Streamgage	Drainage area, in square miles	Q90 streamflow, in cubic feet per second
1	01073000	12.2	1.2
2	01082000	67.0	16
3	01086000	146.0	19
4	01091000	104.0	13
5	01096000	64.1	15
6	01097300	12.8	1.2
7	01108000	261.0	71
8	01109000	43.6	6.8
9	01111300	15.6	1.7
10	01111500	91.2	26
11	01117500	99.3	50
12	01117800	35.2	16
13	01118000	74.2	35
14	01118300	4.0	0.56
15	01118500	293.9	126
16	01121000	27.1	4
17	01123000	29.6	8.8
18	01154000	72.2	11
19	01162500	19.2	2.4
20	01169000	89.7	22
21	01171500	54.0	14
22	01175670	8.8	1

Observation number	Streamgage	Drainage area, in square miles	Q90 streamflow, in cubic feet per second
23	01176000	149.5	38
24	01181000	94.0	18
25	01187300	20.7	2.4
26	01188000	4.1	1.3
27	01193500	104.7	18
28	01194500	22.4	3.4
29	01198000	51.3	7.3
30	01199050	29.6	8.9
31	01200000	200.0	38
32	01333000	42.4	11



# Chapter 10

## Alternative Methods for Regression

---

*Concentrations of a contaminant appear linearly related to distance down-dip in an aquifer. Regression residuals appear to be of generally constant variance. However, several outliers in the dataset inflate the standard error, and what appears graphically as a strong linear relation tests as being insignificant due to the influence from outliers. How can a more robust linear fit be obtained that is not overly sensitive to a few outliers, yet describe the linear relation between contaminant 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 a regression relating this and a nearby site having a 50-year record is used to generate a simulated 50-year stage record for the intake station. The 5th percentile of the simulated record is used as the intake elevation. Given that regression estimates are reduced in variance compared to actual data, this elevation estimate will not be as extreme as it should be. What alternatives to regression would provide better estimates?*

*Plots of concentration versus time are drawn for several regional watersheds. A summary of each relation is desired, but the best fit is a straight line on some plots and a curve on others. Transforming the concentration axis on some plots but not others is not an option. Is there some way to visually represent these relations that does not assume that each must be linear?*

The three examples above demonstrate characteristics of regression that are undesirable in specific situations. First, outliers may unduly influence both the estimated slope and the ordinary least squares (OLS) regression test for significance. Second, non-normality of residuals may cause a nonsignificant test result to be questioned. Third, OLS regression inherently reduces the variance of estimates as compared to the variance of the original data. When the variability of multiple estimates or an estimate of percentiles is required, OLS regression predictions will underestimate the variability and extremes that would have been found with original observations. Fourth, relations may be nonlinear, requiring a more flexible model than a straight-line relation. In these situations, alternative methods to OLS regression are better suited for fitting lines to data.

### 10.1 Theil-Sen Line

The Theil-Sen line (referred to as the Kendall-Theil line in the first edition) is a robust nonparametric model of the median of  $y$  given  $x$ . This line does not depend on the normality of residuals for validity of significance tests, and is not strongly affected by outliers, in contrast to OLS regression. However, the data should have a linear relation in order to use the Theil-Sen model.

The robust estimate of slope,  $\hat{b}_1$ , for this nonparametric median line was first described by Theil (1950). Estimates of the intercept,  $\hat{b}_0$ , are also available (Theil, 1950; Dietz, 1989; Conover, 1999).

Together, the slope and intercept estimate the median of  $y$  using a complete linear equation of the form:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 \cdot x . \quad (10.1)$$

The Theil-Sen line is closely related to Kendall's  $\tau$  (see chap. 8), in that the significance of the test for  $H_0: \text{slope } \beta_1 = 0$  is identical to the test for  $H_0: \tau = 0$ . In addition, the estimated slope,  $\hat{b}_1$ , will always have the same sign as the Kendall  $S$  statistic. Its primary application in water resources has been for trend analyses, where the hypothesis test associated with the line is called the Mann-Kendall test for trend (see chap. 12). The associated confidence interval for the slope and an adjustment for ties were defined by Sen (1968). The slope is sometimes referred to as the "Sen slope" though it seems more properly attributed to Theil, and so we call it the Theil-Sen slope in this text.

The Theil-Sen line has been regularly rediscovered as a useful method in many disciplines, including decision theory (Vannest and others, 2012) and chemistry (Lavagnini and others, 2011). The benefits it offers for data with non-normal error distributions and outliers, as well as those for small datasets—all characteristics common to water-resources data—have been known for some time. For example, Dietz (1987) found that Theil-Sen had far better error characteristics than did OLS regression for data with non-normal error residuals and outliers. Nevitt and Tam (1997) reaffirmed the “poor performance of OLS estimation” and “the merits of alternatives to OLS regression under non-ideal conditions.” Wilcox (1998) established the benefits of Theil-Sen—it is more resistant to the effect of multiple outliers and has a better efficiency (smaller confidence interval on the slope) for small datasets than several other robust methods of regression.

Various types of multivariate extensions to Theil-Sen have been proposed to produce a nonparametric analog to multiple regression. These include Dang and others (2008) and Libiseller and Grimvall (2002). The most common and important extension to Theil-Sen has been a blocking procedure associated with the Seasonal Kendall test (Hirsch and others, 1982); it is discussed in detail in chapter 12. Khalil and others (2012) proposed a version of Theil-Sen to address the issue of preservation of estimation variance, searching for a method to use for record extension that is also robust. Their simulations were limited to mixtures of normal distributions and found that the new method was more robust than the line of organic correlation (LOC, see section 10.2.2) in the presence of outliers, but required a hefty sample size in order to improve on LOC’s results. Not surprisingly, both LOC and the new method preserved estimation variance better than OLS regression or Theil-Sen.

Theil-Sen is computed in some commercial software packages, in several R packages (`rkt` [Marchetto, 2017]; `Kendall` [McLeod, 2011]); and `EcoGenetics` [Roser and others, 2017]), and was made freely available in a Visual Basic program by Granato (2006).

### 10.1.1 Computation of the Line

The Theil-Sen slope estimate,  $\hat{b}_l$ , is computed by comparing each data pair to all others in a pairwise fashion. A dataset of  $n$   $(x, y)$  pairs will result in  $n \cdot (n-1)/2$  pairwise comparisons. For each of these comparisons a slope,  $\Delta y / \Delta x$ , is computed (fig. 10.1A). The median of all pairwise slopes is taken as the nonparametric slope estimate,  $\hat{b}_l$  (fig. 10.1B).

$$\hat{b}_l = \text{median} \left( \frac{y_j - y_i}{x_j - x_i} \right) \quad (10.2)$$

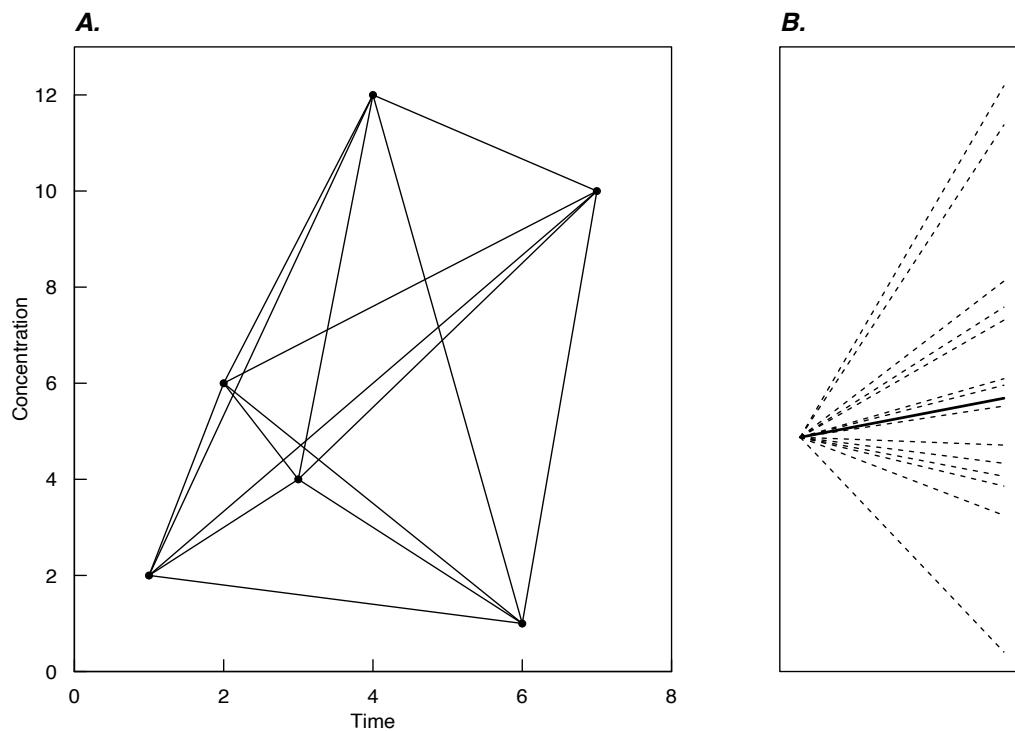
for all  $i < j$ ,  $i = 1, 2, \dots, (n-1)$ ,  $j = 2, 3, \dots, n$ .

The most common form of the intercept is defined as

$$\hat{b}_0 = y_{med} - \hat{b}_l \cdot x_{med}, \quad (10.3)$$

where  $x_{med}$  and  $y_{med}$  are the medians of  $x$  and  $y$ , respectively (Conover, 1999). 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  $(\bar{x}, \bar{y})$ .

Other estimates of intercept have been evaluated in simulation studies. In studies, Dietz (1989) found both the Conover estimator (eq. 10.3) and the median of the residuals,  $y - \hat{b}_l \cdot x_i$ , to have small mean square errors in the presence of outliers and non-normal residuals. The Conover estimator is recommended here, because of its robustness and efficiency, simplicity of computation, analogy to OLS, and historically wide use. If a confidence interval on the intercept is required, compute the intercept as the median of the residuals, and its confidence interval as that for the median (residual) as shown in chapter 3.



**Figure 10.1.** Computation of the Theil-Sen slope. *A*, Plot of the 15 possible pairwise slopes between 6 data points. In *B*, the ordered slopes are arranged to meet at a common origin. The thick line is the Theil-Sen slope, the median of the 15 slopes.

#### Example 10.1. Computation of the Theil-Sen line.

For example, review 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  $(7)(6)/2=21$  pairwise slopes. Comparing points 2 and 1, the slope = +1. Going down the column under point 1, comparing points 6 and 1, the slope = +3. After computing all possible slopes, they are ranked in 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 computed as in equation 10.2 from  $x_{med} = 4$  and  $y_{med} = 4$ , so that  $\hat{b}_0 = 4 - 1 \cdot 4 = 0$ . The line is easily computed using the senth R script in the supplemental material for chapter 10 (SM.10):

```
> y <- c(1, 2, 3, 4, 5, 16, 7)
> x <- c(1:7)
> senth(x,y)
```

```
Theil-Sen line
y = 0 + 1 * x

95 % Confidence interval on the slope
LCL = 1 Theil slope = 1 UCL = 4.333
```

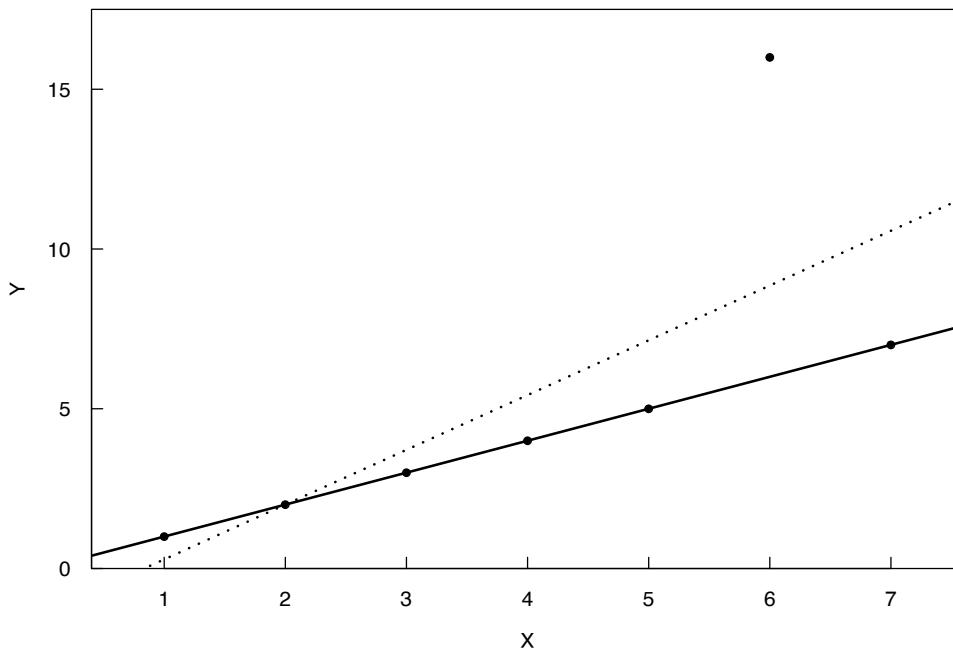
```
Kendall's rank correlation tau
data: x and y
T = 20, p-value = 0.002778
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.9047619
```

### 10.1.2 Properties of the Estimator

OLS regression for the data in example 10.1 produces a slope,  $b_1$ , of 1.71 (fig. 10.2). This differs substantially from the Theil-Sen slope,  $\hat{b}_1$ , of 1, because of 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, the OLS slope  $b_1$  would be greatly inflated, but the Theil-Sen slope  $\hat{b}_1$  would again remain at 1. The Theil-Sen slope  $\hat{b}_1$  is clearly resistant to outliers and responds to the bulk of the data.

The Theil-Sen slope  $\hat{b}_1$  is an unbiased estimator of the slope of a linear relation, as is the OLS slope  $b_1$ . However, the variance of the estimators differs. When the departures from the true linear relation (true residuals) are normally distributed, the OLS slope is slightly more efficient (slightly lower variance and smaller confidence interval) than the Theil-Sen slope. 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 Theil-Sen line has the desirable properties of a nonparametric estimator: it is almost as good (efficient) as the parametric estimator when all assumptions of normality are met, and much better when those assumptions are not met. It is less affected by the common problems of water-resources data (skewness, outliers) than is regression, and so provides a robust estimate of the typical slope, when the typical and not the mean slope is of interest. The efficiency of the Theil-Sen slope to the OLS slope is the same as that for the Hodges-Lehmann estimator (see chap. 5) in comparison to the mean, as the Theil-Sen slope estimate is in the class of Hodges-Lehmann estimators.

How much of a departure from a normal distribution is required before a nonparametric test has an advantage over its parametric counterpart? In the case of the Theil-Sen and OLS slope estimates, how non-normal must residuals be before the Theil-Sen 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 and others (1991) tested the



**Figure 10.2.** Plot of the Theil-Sen (solid) and ordinary least-squares (OLS) regression (dashed) fits to the example data. Note the effect of one outlier on the OLS regression line.

two slope estimators under one type of departure from normality—a mixture of two normal distributions. The two individual distributions are shown in figure 10.3; one distribution had a mean of 10 and a standard deviation of 1, and the second distribution had a mean of 11 and a standard deviation of 3. A mixture of the distributions—95 percent from the first distribution and 5 percent from the second—are shown in figure 10.4. Visual examination of figure 10.4 reveals only the slightest departure from symmetry.

Given sampling variability that would exist in an actual dataset, it would be exceedingly unlikely that samples from this distribution would be identified as non-normal. A more substantial departure from normality, a mixture of 80 percent of the first distribution and 20 percent of the second, is shown in figure 10.5. There is a difference in the shape of the two tails of the distribution, but again the non-normality is not highly noticeable.

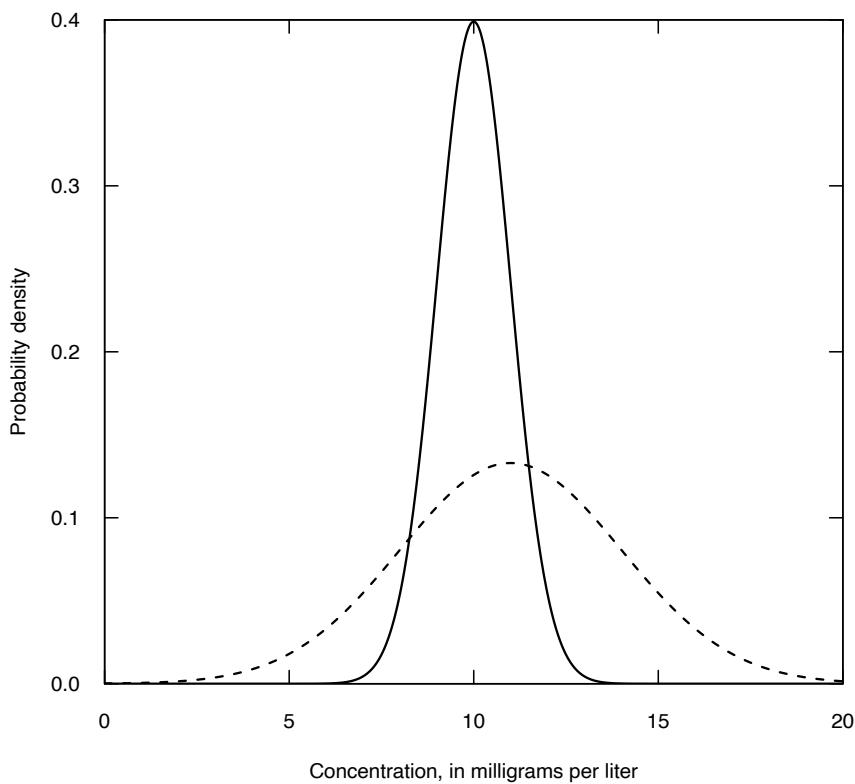
Random samples were generated from each of several different distribution mixtures containing between 0 and 20 percent of the second distribution. Data from each mixture were treated as a separate response variable in a regression versus an explanatory variable of random order. The true population slope is therefore zero. Both OLS and the Theil-Sen slope estimators were computed, and their errors around zero

recorded as root mean square error (RMSE),  $RMSE = \sqrt{\frac{\sum(b - 0)^2}{n}}$ , where  $b$  is the estimated slope using Theil-Sen or OLS, and  $n$  is the number of random samples.

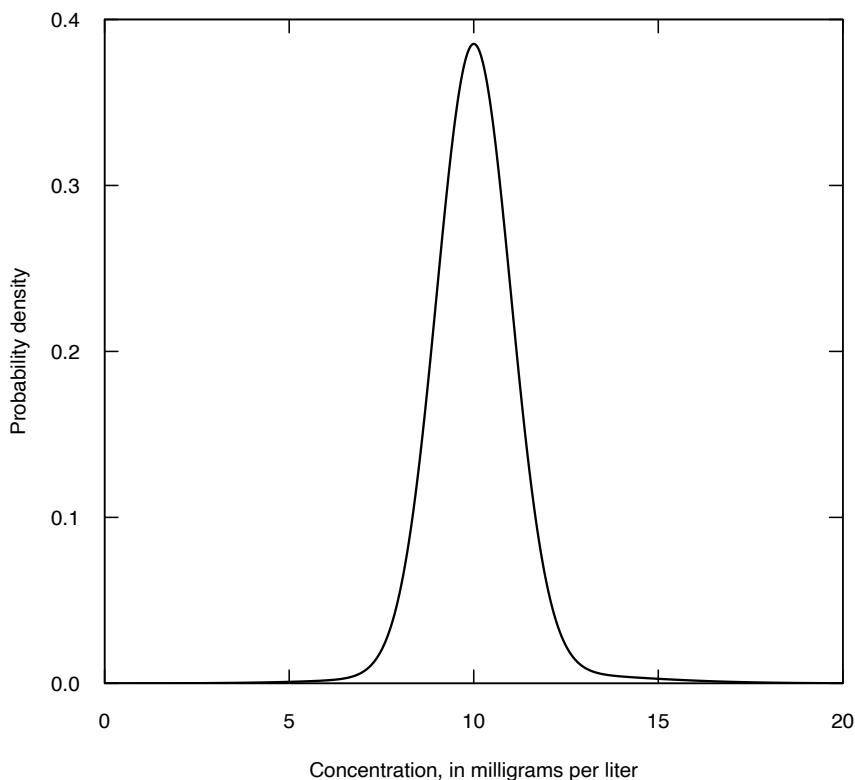
The results are presented in figure 10.6 as the ratio of the RMSE for the Theil-Sen estimator to the RMSE of the regression estimator (Hirsch and others, 1991). A value larger than 1 shows an advantage to OLS; smaller than 1 indicates the Theil-Sen estimate to be superior.

For the larger sample size ( $n=36$ , solid line) the OLS estimator was more efficient (by less than 5 percent) when the data are not mixed and therefore normally distributed. With even small amounts of mixtures the Theil-Sen estimator quickly becomes more efficient. At a 20 percent mixture the Theil-Sen estimator was almost 20 percent more efficient. When the sample size was very small ( $n=6$ , dashed line), efficiencies of the two methods were within 5 percent of each other, though the Theil-Sen estimator had the advantage for all but very small amounts of non-normality.

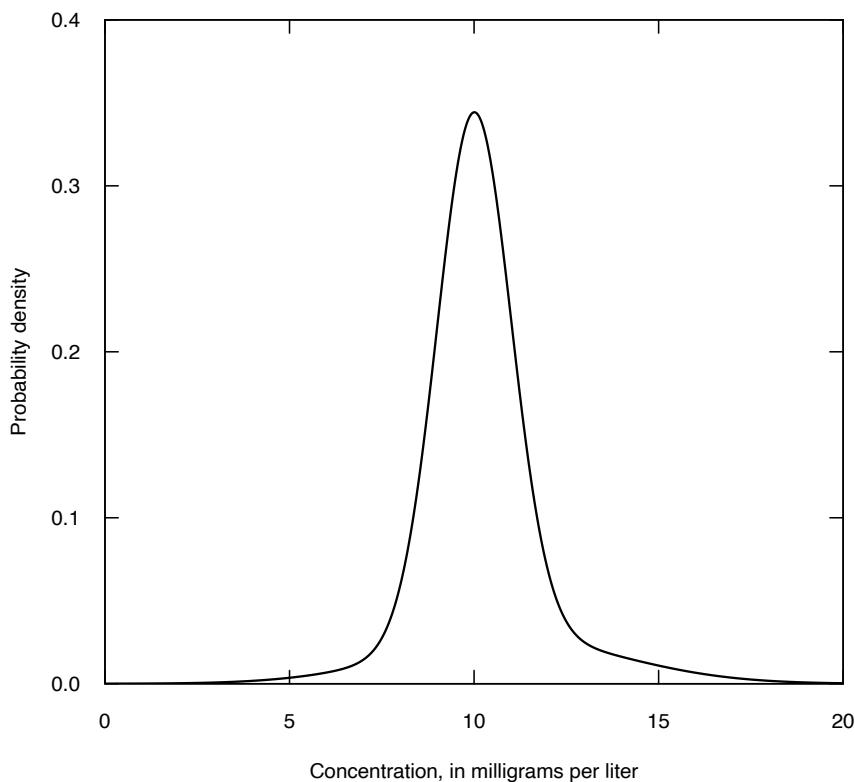
These results reinforce that the two methods will give nearly identical results when the data or their transformations exhibit a linear pattern, constant variance, and near-normality of residuals. The advantages of familiarity and availability of diagnostics may favor using OLS regression in that case. However, when



**Figure 10.3.** Probability density functions of two normal distributions used by Hirsch and others (1991), the first with mean=10 and standard deviation=1; the second with mean=11 and standard deviation=3.



**Figure 10.4.** Probability density function of a mixture of data (95 percent from distribution 1 and 5 percent from distribution 2) from Hirsch and others (1991).



**Figure 10.5.** Probability density function of a mixture of data (80 percent from distribution 1 and 20 percent from distribution 2) from Hirsch and others (1991).

residuals are not normally distributed, and especially when outliers are present, the Theil-Sen line has greater efficiency (lower variability and bias) than does OLS. Even small departures from normality (not always sufficient to detect with a test or boxplot of residuals) favor using Theil-Sen.

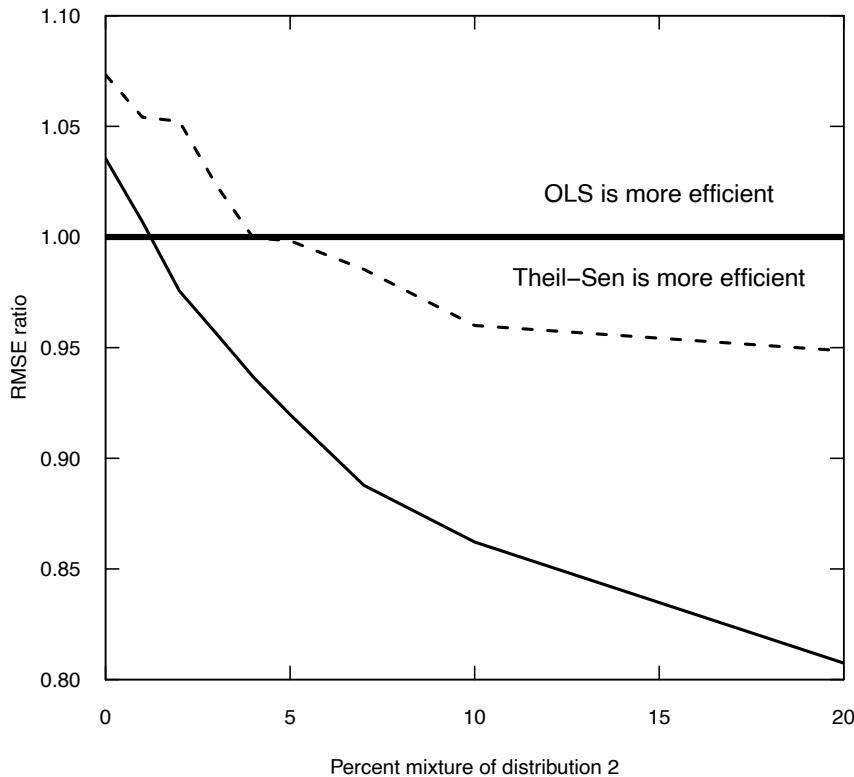
As a matter of course, one should check all outliers for error, as discussed in chapter 1. Do the outliers represent a condition different from the rest of the data? If so, they may be the most important points in the dataset. Outliers cannot and should not be automatically deleted -- often no error in them can be found. Robust methods like Theil-Sen provide protection against disproportionate influence by these distinctive, but perhaps perfectly valid and meaningful data points.

Possibly the two greatest uses for the Theil-Sen line are (1) in a large study where multiple equations representing multiple locations or variables are fit, without the capability for exhaustive checking of distributional assumptions or evaluations of the sensitivity of results to outliers; and (2) by users not trained in residuals plots and use of transformations to stabilize skewness and heteroscedasticity. A third use is for fitting lines to data where one does not wish to transform the  $y$  variable, perhaps due to the resulting transformation bias (chap. 11).

Theil-Sen estimates a median rather than a mean of  $y$ , which is a disadvantage when the latter is desired. A second possible disadvantage of Theil-Sen as compared to OLS is that there is currently no Theil-Sen line for multiple explanatory variables analogous to multiple regression.

#### Example 10.2. Theil-Sen line for trends in total phosphorus.

OLS and Theil-Sen lines for total phosphorus concentrations from 1975 to 1989 in the St. Louis River at Scanlon, Minnesota (see chap. 12 for more on trend tests), are shown in figure 10.7. The outliers are accurate values from floods and should not be ignored or deleted. The question is whether there is a significant linear trend in concentration over this 14-year period. We test this hypothesis using both the Theil-Sen line and OLS regression.



**Figure 10.6.** Relative efficiency of the Theil-Sen slope estimator as compared with the ordinary least squares (OLS) slope represented as the ratio of the root mean square error (RMSE) of the Theil-Sen estimator to the OLS estimator. When the RMSE ratio is greater than 1, OLS is more efficient. When RMSE ratio is less than 1, Theil-Sen is more efficient. Solid line:  $n=36$ . Dashed line:  $n=6$ . From Hirsch and others (1991).

```
> head(stLouisTP)
      Date DecYear   TP
1 1974-10-30 1974.829 0.04
2 1974-12-09 1974.938 0.04
3 1975-01-21 1975.056 0.05
4 1975-03-03 1975.168 0.04
5 1975-04-07 1975.264 0.07
6 1975-05-27 1975.401 0.03
```

```
> # to compute the Theil-Sen line
> senth(DecYear,TP)
```

Theil-Sen line

```
TP = 4.494457 -0.002248922 * DecYear
```

```
95 % Confidence interval on the slope
LCL = -0.003 Theil slope = -0.002 UCL = -0.001
```

```
Kendall's rank correlation tau
data: x and y
z = -4.4638, p-value = 8.052e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
```

-0.3170606

```
> # to compute the OLS regression
> summary(lm(TP ~ DecYear))
```

Call:

```
lm(formula = TP ~ DecYear)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.04374	-0.02422	-0.01298	0.00191	0.45527

Coefficients:

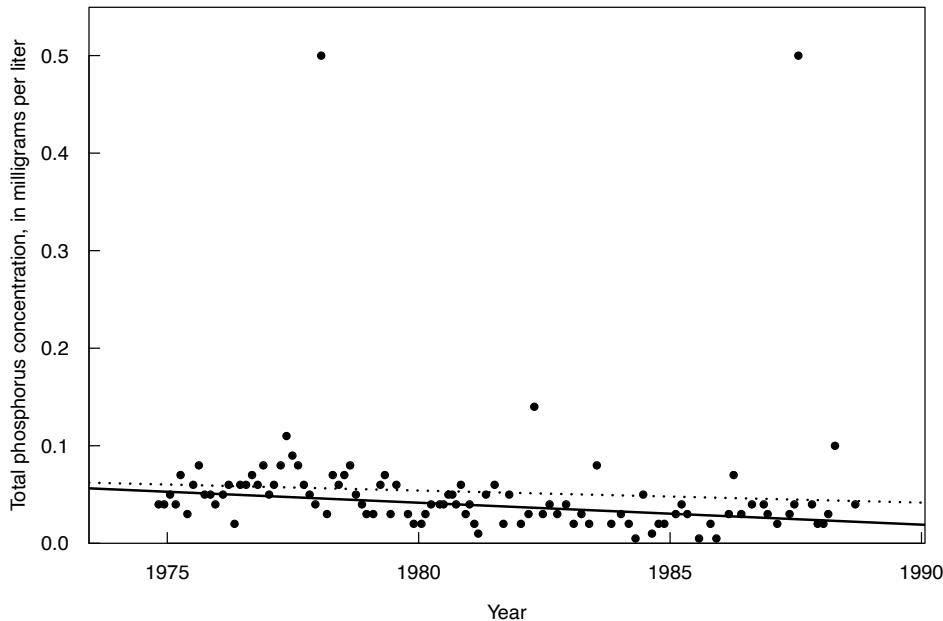
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.500856	3.354874	0.745	0.458
DecYear	-0.001236	0.001693	-0.730	0.467

Residual standard error: 0.06784 on 100 degrees of freedom

Multiple R-squared: 0.005297, Adjusted R-squared: -0.00465

F-statistic: 0.5325 on 1 and 100 DF, p-value: 0.4673

The estimated slopes are both negative, with the Theil-Sen estimate being twice the OLS regression slope estimate. The Theil-Sen slope is -0.0022 milligrams per liter per year (mg/L/yr) and OLS regression slope is -0.0012 mg/L/yr. The OLS regression slope is not significantly different from zero ( $p=0.467$ ). This is a result of the influence of the two extreme residuals on the standard error of the trend slope estimate. However, the Theil-Sen slope is highly significant ( $p=8e-06$ )—the significance test for the Theil-Sen slope is identical to the significance test for Kendall's  $\tau$  (see next section). The Theil-Sen line is not dependent on assumptions of normality that are strongly violated in this dataset and the line is highly resistant to the magnitude of the two extreme values in the dataset.



**Figure 10.7.** Scatterplot of total phosphorus concentrations for the St. Louis River at Scanlon, Minnesota, 1975–89 with ordinary least squares regression (dotted) and Theil-Sen (solid) fitted lines.

### 10.1.3 Test of Significance for the Theil-Sen Slope

The test for significance of the Theil-Sen slope is identical to the test for Kendall's  $\tau$ ,  $H_0: \tau=0$  (see chap. 8). The Theil-Sen slope,  $\hat{b}_1$ , is closely related to Kendall's  $S$  and  $\tau$  in the following ways.

1.  $S$  is the sum of the algebraic signs of the pairwise slopes.
2. If the product  $(\hat{b}_1 \cdot x)$  is subtracted from every  $y$  value, the new  $y$  values will have an  $S$  and  $\tau$  equal to zero, indicating no correlation.

If  $x$  is a measure of time, as it is for a trend test, subtracting  $(\hat{b}_1 \cdot x_i)$  yields a trend-free version of the  $y$  dataset.

#### Example 10.3. Exact and approximate tests for the Theil-Sen slope and $\tau$ .

For the dataset used in example 10.1, the test of significance is computed as follows.  $S$  equals the sum of the signs of pairwise slopes. There are  $n \cdot (n - 1) / 2 = 21$  slopes, 20 of which are positive and 1 is negative, so that  $S = 20 - 1 = 19$ . Kendall's  $\tau$  correlation coefficient equals  $19/21$  or 0.90. Using R, the exact two-sided  $p$ -value (here  $S=19$  and  $n=7$ ) is 0.0028:

```
> cor.test(x, y, method = "kendall")
```

```
Kendall's rank correlation tau

data: x and y
T = 20, p-value = 0.002778
alternative hypothesis: true tau is not equal to 0
```

sample estimates:

```
tau
0.9047619
```

Thus,  $y$  is significantly related to  $x$  in a linear fashion. In chapter 4 the benefits of exact tests over large-sample approximations for small sample sizes were discussed. Other software packages besides R often inappropriately use the large-sample approximation for all sample sizes, reporting a  $p$ -value of 0.0068 for this small sample size. This can be demonstrated (but should not be used) in R by forcing computation of the approximate test with the `exact=FALSE` option:

```
> cor.test(x, y, method = "kendall", exact=FALSE, continuity = TRUE)
```

Kendall's rank correlation tau

```
data: x and y
z = 2.7034, p-value = 0.006864
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.9047619
```

#### 10.1.4 Confidence Interval for the Theil-Sen Slope

Confidence intervals for the Theil-Sen slope,  $\hat{b}_t$ , can be computed for small sample sizes from a tabled distribution of the test statistic, such as table A30 in Hollander and Wolfe (1999). The critical value,  $X_u$ , for Kendall's  $\tau$  having a  $p$ -value nearest to  $\alpha/2$  is used to compute the ranks  $R_u$  and  $R_l$  of the  $n(n-1)/2=N$  pairwise slopes representing the upper and lower confidence limits for  $\hat{b}_t$ , respectively. These limits are the  $R_{lh}$  ranked data points going in from either end of the sorted list of pairwise slopes (eqs. 10.4 and 10.5). The resulting confidence interval will reflect the shape (skewed or symmetric) of the original data.

$$R_u = \frac{(N + X_u)}{2} \quad (10.4)$$

$$R_l = \frac{(N - X_u)}{2} + 1. \quad (10.5)$$

For sample sizes where  $n \geq 10$  the large-sample approximation can be used. If computing by hand, upper and lower limits are found corresponding to critical values at one-half the desired  $\alpha$  level. The critical value,  $z_{\alpha/2}$ , from quantiles of the standard normal distribution determines the upper and lower ranks of the pairwise slopes corresponding to the ends of the confidence interval. Those ranks are

$$R_u = \frac{N + z_{\alpha/2} \sqrt{\frac{n(n-1)(2n+5)}{18}}}{2} + 1, \quad (10.6)$$

$$R_l = \frac{N - z_{\alpha/2} \sqrt{\frac{n(n-1)(2n+5)}{18}}}{2}. \quad (10.7)$$

As an example, for  $n=20$  pairs of data there would be  $N=20 \cdot 19/2=190$  possible pairwise slopes. Therefore  $\hat{b}_l$  is the average of the 95th and 96th ranked slopes. For a 95-percent confidence interval on  $\hat{b}_l$ ,  $z_{\alpha/2}=1.96$  and the upper and lower bounds are

$$R_u = \frac{190+1.96\sqrt{950}}{2} + 1 = 126.2 ,$$

$$R_l = \frac{190-1.96\sqrt{950}}{2} = 64.8 ,$$

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  $\alpha=0.05$  confidence limit on  $\hat{b}_l$ . The `senth` script much more easily computes large-sample confidence intervals for the Theil-Sen slope. Further discussion of these equations is in Hollander and Wolfe (1999; p. 424–26).

#### Example 10.4. Computing the confidence interval for the Theil-Sen slope by hand.

The  $N=21$  possible pairwise slopes between the  $n=7$  data pairs for example 10.1 were:

−9	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
+1	+1	+1	+1	+1	+3	+3.5	+4.3	+6	+11	

The Theil-Sen slope,  $\hat{b}_l$ , was the median or 11th largest slope. To determine an exact (small-sample) confidence interval for  $\hat{b}_l$  with  $\alpha \approx 0.05$ , the critical value,  $X_{\nu}$ , nearest to  $\alpha/2=0.025$  is found to be 15 (using  $p=0.015$ , where 0.015 is as close to  $\alpha/2$  as we can get). The rank,  $R_u$ , of the pairwise slope corresponding to the upper confidence limit is therefore

$$R_u = \frac{(21+15)}{2} = 18 .$$

The rank  $R_l$  of the pairwise slope corresponding to the lower confidence limit is

$$R_l = \frac{(21-15)}{2} + 1 = 4 .$$

Therefore, an  $\alpha/2=2 \cdot 0.015=0.03$  confidence limit for  $\hat{b}_l$  is the interval between the 4th and 18th ranked pairwise slope (the 4th slope in from either end), or

$$+1 \leq \hat{b}_l \leq +3.5 .$$

The large-sample approximate confidence interval was previously computed by the `senth` script as  $+1 \leq \hat{b}_l \leq +4.33$ . The asymmetry around the estimate  $\hat{b}_l = 1$  reflects the low probability that the slope is less than 1, based on the data.

## 10.2 Alternative Linear Equations for Mean $y$

Hirsch and Gilroy (1984) described additional methods for estimating the mean of  $y$  by 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 applications such as calibration. When many predictions are to be made and the distribution of those predictions is important (predicted percentiles or spreads are more of interest than the predicted mean), the line of organic correlation (LOC) should be used instead of OLS. Either LOC or least normal squares (LNS) more appropriately incorporates the errors in both  $x$  and  $y$  when describing the

intrinsic relation between two variables without trying to predict one from the other. LNS is recommended when a geographic trajectory is to be computed.

### 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 the direction of  $y$  only, without regard to errors in the  $x$  direction. The equation may be written as

$$y_i = \bar{y} + r \frac{S_y}{S_x} (x_i - \bar{x}), \quad (10.8)$$

where  $r$  is Pearson's linear correlation coefficient,  $s_y$  and  $s_x$  are the standard deviations of the  $y$  and  $x$  variables,  $SS_y$  and  $SS_x$  are the sums of squared deviations from the means of  $y$  and  $x$ , respectively, and

$(r \cdot s_y / s_x) = (r \sqrt{SS_y} / \sqrt{SS_x}) = b_1$ , the OLS estimate of slope (see chap. 9). Assuming the linear form of the model is correct,  $x$  is measured without error and the distribution of observations around the line follows a normal distribution. OLS will lead to estimates of  $y_i$  for any given  $x_i$  that are unbiased and have minimum variance. In this situation OLS is the preferred method of estimating an expected mean  $y$  given  $x$ .

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. An example is in geomorphology, where 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 two chemicals in a sample, say for copper and lead concentrations. Either could be chosen to be predicted as a function of the other, and usually a description of their bivariate relation 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 10.8, reversing the usual order and setting  $x$  as the response variable, the resulting OLS equation will be

$$x_i = \bar{x} + r \cdot \frac{S_x}{S_y} (y_i - \bar{y}), \quad (10.9)$$

which when solved for  $y$  becomes

$$y_i = \bar{y} + \frac{1}{r} \cdot \frac{S_y}{S_x} (x_i - \bar{x}). \quad (10.10)$$

Let  $b_1' = \left( \frac{1}{r} \cdot \frac{S_y}{S_x} \right)$ , the slope of  $x$  on  $y$  re-expressed to compare with slope  $b_1$ . Contrasting equations 10.8 and 10.10, the slope coefficients  $b_1 \neq b_1'$ . The two regression lines will differ unless the correlation coefficient  $r$  equals 1.0. In figure 10.8, the two regression lines are plotted relating two measures of dissolved inorganic content, both measured from the same water sample (Hirsch and Gilroy, 1984). These two measures are total dissolved solids (TDS) and residue on evaporation (ROE). There is no reason to predict only one from the other, so it is unclear which of the two OLS lines is the true relation. Unless the correlation coefficient  $r=1$ , neither is the true relation.

The choice of which, if either, of the OLS lines to use follows a basic guideline. If one variable 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 intrinsic relation between the two variables is of interest, neither OLS line is the appropriate approach. Neither OLS line uniquely describes that relation, as the OLS slope is dampened (decreased by a factor equal to the correlation coefficient) to reduce error variance in predictions of  $y$ . A different linear model having a unique solution incorporating errors in both  $x$  and  $y$  should be used instead—either LOC or LNS.

### 10.2.2 Line of Organic Correlation (or Reduced Major Axis)

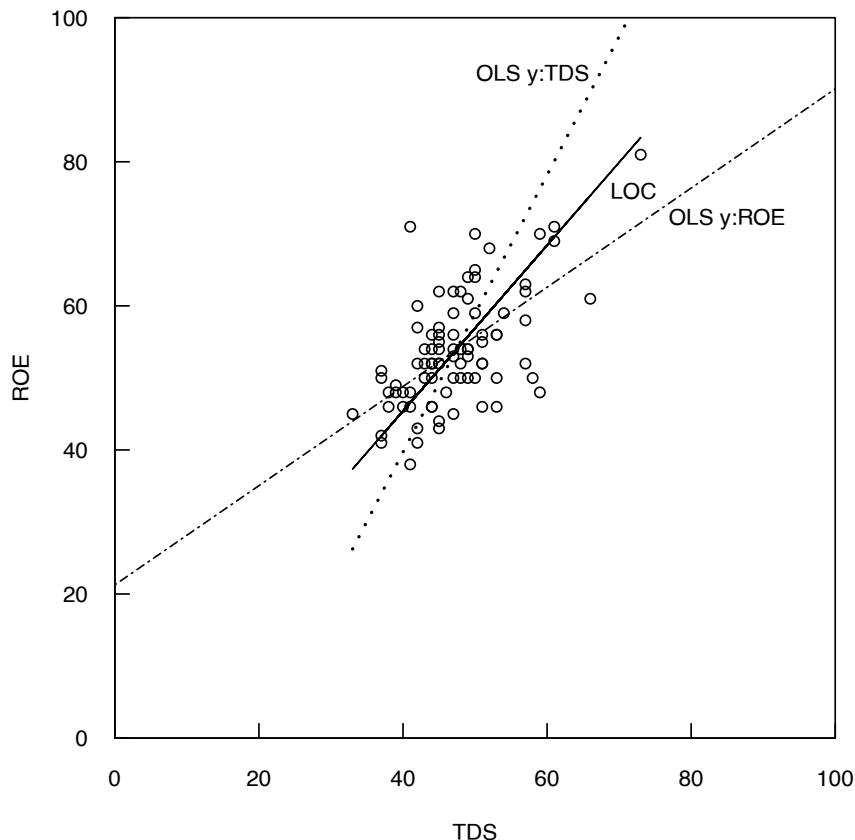
The line of organic correlation (LOC) has been discussed and used in many disciplines over many years. Samuelson (1942) presented its use in economics when both  $x$  and  $y$  variables were measured with error. Its theoretical properties were discussed by Kruskal (1953). It was proposed as a linear fitting procedure in hydrology by Kritskiy and Menkel (1968) and applied to geomorphology by Doornkamp and King (1971). The line of organic correlation also goes by many other names. In biological applications it is usually called the reduced major axis or RMA (Kermack and Haldane, 1950), or the standardized major axis or SMA (Warton and Weber, 2002). It has also been called the geometric mean functional regression (Halfon, 1985), the allometric relation (Teissier, 1948), impartial regression (Strömberg, 1940; Tofallis, 2002), and maintenance of variance-extension or MOVE (Hirsch, 1982).

LOC minimizes the sum of the areas of right triangles formed by horizontal and vertical lines extending from observations to the fitted line (fig. 10.9). By minimizing errors in both directions, it lies between the two OLS lines on a plot of  $y$  versus  $x$  (fig. 10.8). The LOC equation is

$$y_i = b_0^* + \text{sign}[r] \cdot \frac{s_y}{s_x} \cdot x_i , \quad (10.11)$$

where the slope of the LOC line  $b_1^*$  equals the geometric mean of the  $y$  on  $x$  and  $x$  on  $y$  OLS slopes:

$$b_1^* = \sqrt{b_1 \cdot b_1'} = \text{sign}[r] \cdot \frac{s_y}{s_x} . \quad (10.12)$$



**Figure 10.8.** Plot of three straight lines fit to the same data. Ordinary least squares (OLS)  $y$ : residue on evaporation (ROE) is the usual regression with ROE as the  $y$  variable (eq. 10.8). LOC is the line of organic correlation (eq. 10.11). OLS  $y$ : total dissolved solids (TDS) is a regression using TDS as the  $y$  variable (eq. 10.10).

LOC replaces the correlation coefficient in the equation for OLS slope with just the algebraic sign (+ or -) of the correlation coefficient. The magnitude of the LOC slope,  $b_1''$ , is determined solely by the ratio of standard deviations  $s_y/s_x$ , and so performing LOC of  $x$  on  $y$  produces the identical line (once converted back to the original scale) as the LOC of  $y$  on  $x$ .

The intercept of the LOC line  $b_0''$  is solved for by placing  $\bar{y}$  and  $\bar{x}$  into the LOC equation (eq. 10.11).

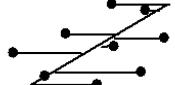
LOC possesses three characteristics preferable to OLS in specific situations:

1. LOC minimizes errors in both  $x$  and  $y$  directions,
3. LOC provides a single line regardless of which variable,  $x$  or  $y$ , is used as the response variable, and
4. The cumulative distribution function of the predictions, including the variance and probabilities of extreme events such as floods or percent exceedances of a numerical standard, estimates those of the actual records they are generated to represent.

LOC is therefore used for two purposes, corresponding to the three previously mentioned characteristics:

1. To model the intrinsic functional relation between two variables, both of which are measured with error, and
2. 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 (filling in missing record). Such estimates are important when multiple predictions are made, and it is the probability distribution (variance or percentiles) of the estimates, not just mean  $y$  or an individual estimate, which are to be interpreted and used.

Examples of the first use for LOC include the geomorphic relations 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. OLS slopes are not the physically intrinsic values but dampened toward zero to minimize the error in prediction of the mean.

Method	Minimizes	Slope	Scale change	Rotation
OLS $y$ on $x$		$b_1 = r \cdot \frac{s_y}{s_x}$	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 $A = \frac{1}{2} \left( \frac{s_x}{s_y} - \frac{s_y}{s_x} \right)$	Changes	Invariant

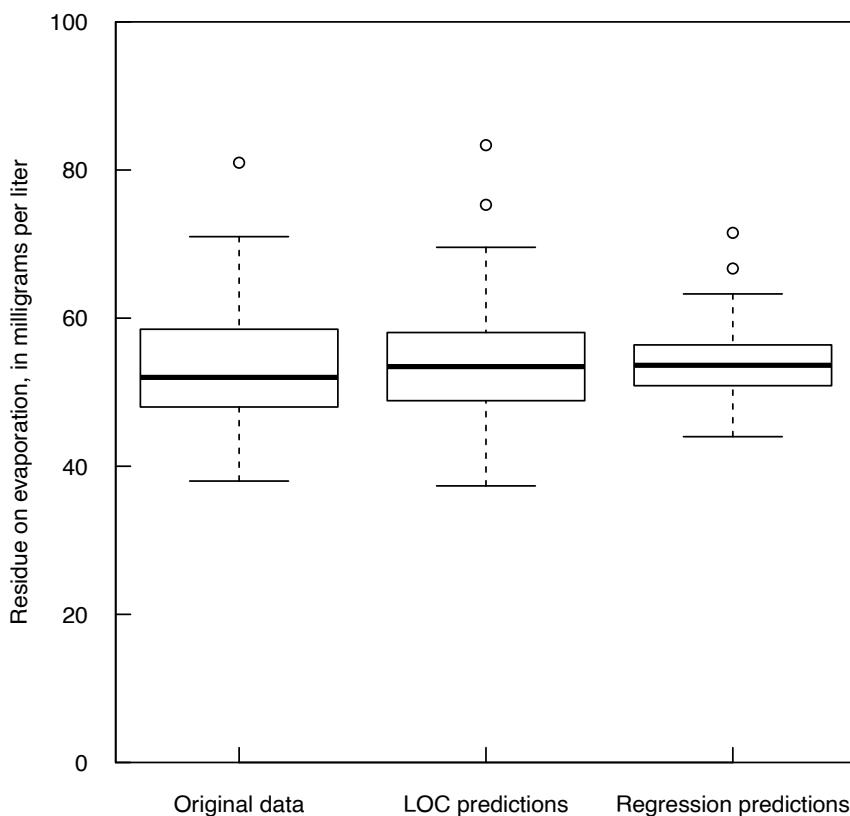
**Figure 10.9.** Characteristics of four parametric methods to fit straight lines to data. OLS, ordinary least squares; LOC, line of organic correlation; LNS, least normal squares.

The second use for LOC is for record extension, which 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 relation, a series of streamflow data at the shorter record site is estimated during an ungaged period from 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. Only when  $|r|=1$  do OLS estimates possess the same variance as would be expected for the original data.

To see this more clearly, take the extreme case where  $r=0$  and there is no relation between  $y$  and  $x$ . The slope then equals 0 and all OLS estimates would be identical and equal to  $\bar{y}$ . The variance of these 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 the 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. The reduction in variance in predictions of ROE by regression in comparison to those for the original data and for LOC predictions is shown in figure 10.10.

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 a water-quality guideline, inferences are made that depend on the probability distribution of the estimated data. In these cases LOC, rather than OLS, should be used to generate predictions. OLS estimates would substantially underestimate the variance because they do not include the variability of individual values around the regression line (Hirsch, 1982). Consequently, the frequency of extreme events such as floods, droughts, or exceedance of standards would be underestimated by OLS.

There are several variations on this technique that have been proposed and applied to hydrologic records, generally known as MOVE (maintenance of variance-extension) methods. These have been published by Vogel and Stedinger (1985) and Grygier and others (1989). There are differences in the details



**Figure 10.10.** Boxplot of the original residue on evaporation data in comparison to boxplots of predicted values from the line of organic correlation (LOC) and regression lines. The smaller box for regression predictions illustrates its reduction in variance.

of the various MOVE techniques, but all have the same goal, which is creating an estimated record that does not cause a reduction in the variance for the part of the record that has been estimated. In hydrology, where the behavior of the extremes of the distribution can be important, it is critical that the methods used be designed to correct for the tendency of any regression method to produce estimates that regress to the mean. The decision between using OLS or Theil-Sen versus MOVE should be based on whether the researcher is primarily interested in getting the best estimate of a single specific missing value, or primarily interested in getting the best collection (for several missing values) of estimates that preserves the overall properties of the dataset. If the goal is the former then OLS or Theil-Sen are best, but if it is the latter then the MOVE technique should be the preferred approach.

Marquet (2000) recommended using LOC rather than OLS to establish the intrinsic relation between body size and population density in ecological systems when neither is considered an independent variable appropriate to predict the other. Miller and Tans (2003) used LOC to obtain realistic estimates of uncertainty in the relation between CO<sub>2</sub> concentrations to isotopic fraction of carbon-13. In an important extension, Draper and Yang (1997) generalized LOC to more than two variables as an alternative to OLS multiple regression. Applying this, Lasi and others (2011) used multivariate LOC to model relations between turbidity, chlorophyll-a, and color (predictor variables) to the light extinction coefficient, K<sub>d</sub>, in lakes. All four variables were measured with error, and predictions for any one variable could be made from specific critical values of the other three. A robust version of LOC was proposed by Khalil and Adamowski (2012) which performed better than LOC itself for data with outliers, though not unless sample sizes were large. As simulations for the robust version of LOC used only mixtures of normal distributions, further evaluation on distributions more similar to hydrologic records (lognormal or gamma distributions) is needed before replacing LOC with its robust cousin.

All three of the parametric lines discussed thus far (LOC, OLS  $y$  on  $x$ , and OLS  $x$  on  $y$ ) have two identical characteristics. First, 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. Second, if the  $x$  and  $y$  axes are rotated in space and lines recomputed, the new lines when re-expressed into the original orientation will differ from the first. This change following rotation 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 (Major Axis)

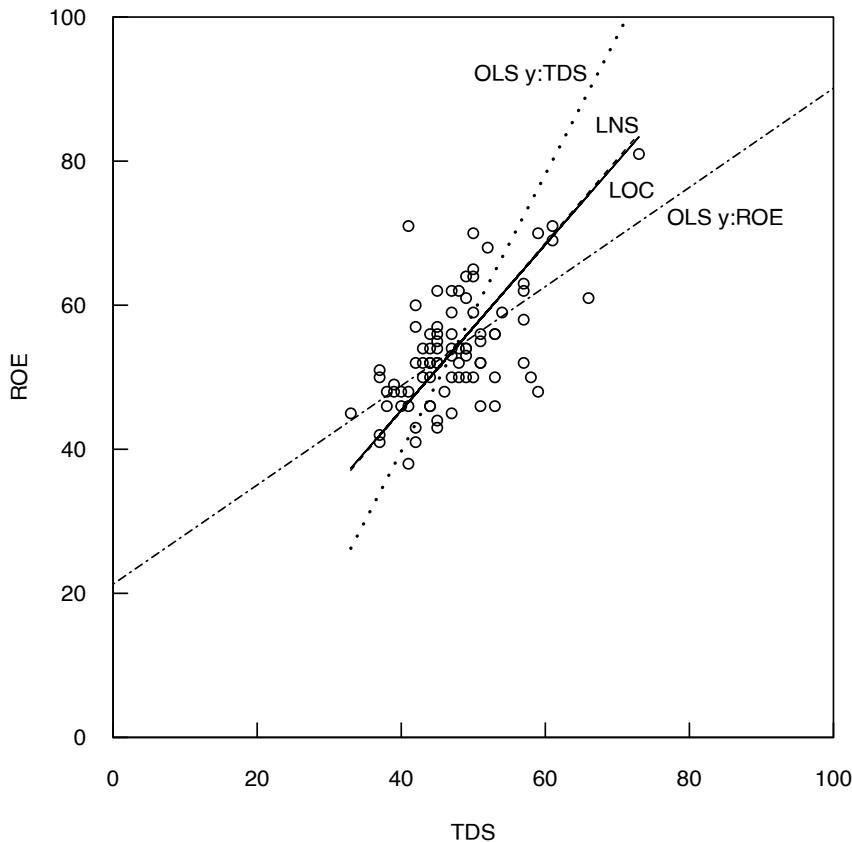
Least normal squares (LNS) is the line which minimizes the squared distances between observed points and the line, where distances are measured perpendicular (normal) to the line. It has also been called the major axis or orthogonal regression, and is identical to the first principal component for the two-dimensional dataset. The slope can be expressed as in equation 10.13:

$$b = -A + \frac{\sqrt{r^2 + A^2}}{r} , \quad (10.13)$$

where

$$A = \frac{1}{2} \left( \frac{S_x}{S_y} - \frac{S_y}{S_x} \right)$$

Though differing from LOC, the LNS line also lies between the two possible OLS lines (fig. 10.11), and in graphics experiments has been shown to be similar to a line drawn by eye when humans estimate the center of a linear field of data (Mosteller and others, 1981; Bajgier and others, 1989). 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 measures of 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. Where LNS is appropriate is in computing trajectories minimizing distances between observed points in space. Kirby (1974b) used LNS to compute the straight-line traverse of a ship from a set of coordinate locations taken along its trip. LNS is identical to the first principal component, so is easily extended to



**Figure 10.11.** Plot of four straight lines fit to the same data. Least normal squares (LNS) line added to the three lines from figure 10.8. LOC, line of organic correlation; OLS, ordinary least squares; ROE, residue on evaporation; TDS, total dissolved solids.

multivariate applications, and is available in any software performing principal component analysis. The LNS intercept is also solved for by placing the mean of  $y$  and  $x$  into their respective places in the linear equation. Therefore, all four parametric lines in figure 10.11 go through the point  $(\bar{x}, \bar{y})$ .

#### 10.2.4 Summary of the Applicability of OLS, LOC, and LNS

The application of each of the four parametric procedures is summarized as follows.

1. To estimate the expected (mean) value 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 or percentiles of the predicted data, use LOC. This preserves the characteristics of the entire distribution, avoiding the downward bias in variance of OLS estimates.
3. To describe the intrinsic relation between two variables with the primary interest in the slope coefficient, use LOC. The OLS slope will generally be biased towards zero compared to what is expected from physical or theoretical models between the variables.
4. To determine the geographic trajectory that minimizes differences from observed data, use LNS.

The four sets of slopes and intercepts relating ROE to TDS for each of the four lines discussed in this section are shown in table 10.1.

**Table 10.1.** Intercepts and slopes for the four lines of figure 10.11.

[OLS, ordinary least squares; ROE, residue on evaporation; TDS, total dissolved solids; LOC, line of organic correlation; LNS, least normal squares]

Method	Intercept	Slope
OLS $y$ :ROE (estimate ROE from TDS)	21.27	0.688
LOC	-0.621	1.150
OLS $y$ :TDS (estimate TDS from ROE and reconvert coefficients)	-37.21	1.923
LNS	-1.446	1.168

## 10.3 Smoothing Methods

Smoothing methods have been in use in statistics for decades. A smooth is a resistant centerline that is fit to the data whose level and slope varies locally in response to the data themselves (chap. 2). Some techniques make it possible to create a smooth in multiple dimensions (using several explanatory variables), although the discussion here will be limited to those using a single explanatory variable.

There are two primary reasons for creating a smooth of the dataset. The first is purely to create a visual representation of the dominant pattern to the data. This representation helps to remove incorrect impressions of the relation between  $x$  and  $y$  when extremes of  $y$  dominate the pattern. A smooth highlights the central tendency without giving excessive weight to the more extreme values of  $y$ ; it offers a great advantage over parametric methods such as linear regression because these methods require an assumed functional form (for example,  $y$  is linearly related to  $x$ , or  $y$  is related to  $x$  by a quadratic or cubic relation). Smoothing algorithms allow the shape of the relation to be dictated by the data and not some arbitrarily selected functional form. A second reason for constructing a smooth is to compute residuals that can then be studied as variables in their own right. For example, we may know that peak discharge from a storm varies strongly as a function of the maximum 6-hour intensity of the rainfall. But our research interest may be in determining if the peak-discharge for a given rainfall intensity changed over the period of record as a result of land-cover changes. We can examine that by building a smooth of discharge ( $y$ ) as a function of rainfall intensity ( $x$ ). Then we can take the residuals from this relation and see if they are changing over time or changing as a function of some land-cover variable (such as percent impervious surface). These techniques are discussed in chapter 12 in the context of trend analysis.

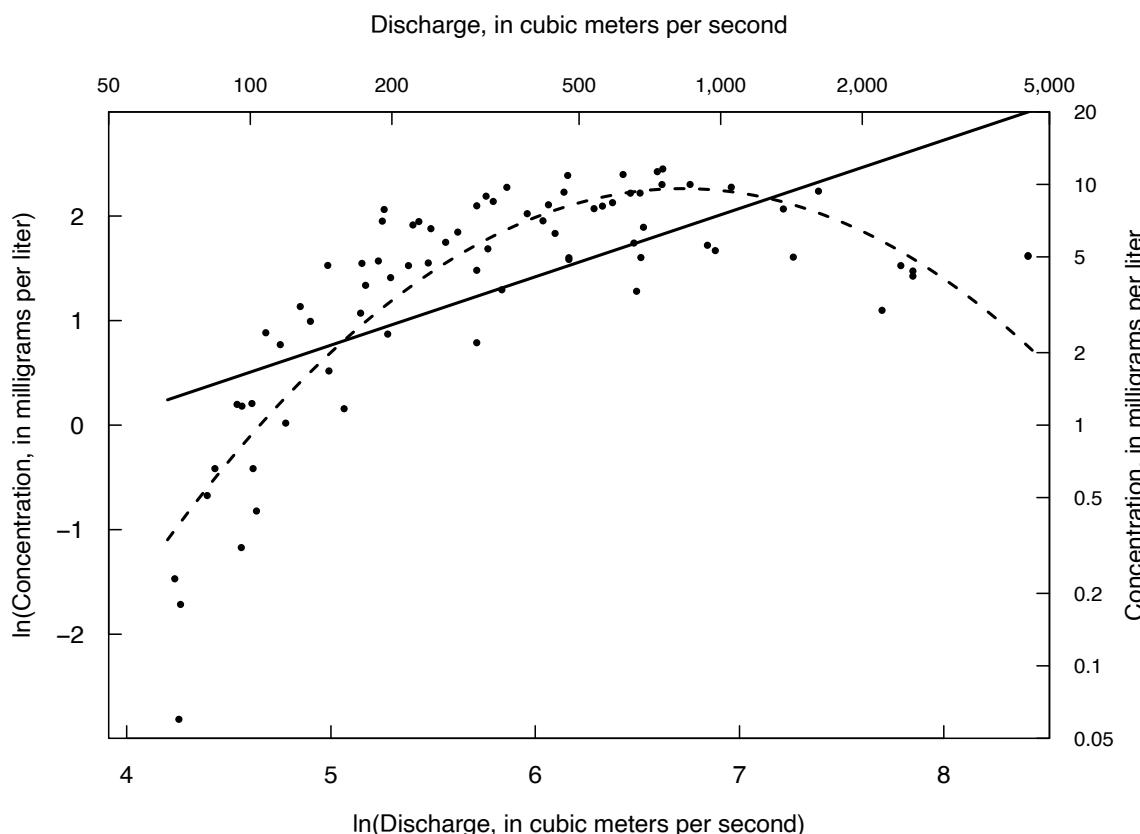
Before introducing a smoothing algorithm here, it is worth considering why we might want to use one as an alternative to regression. An example of nitrate concentrations and the associated daily mean discharge on the day the sample was collected is shown in figure 10.12. The data are from the months June through September of the years 1990–2008 for the Iowa River at Wapello, Iowa (shown in chap. 2 as well). The figure shows the data along with a linear fit of  $\log(C)$  to  $\log(Q)$  and a quadratic fit of  $\log(C)$  to  $\log(Q)$  and  $(\log(Q))^2$ .

The linear fit is clearly inappropriate; it is unable to accommodate the obvious curvature of the data and thus it seriously overestimates concentrations at low and high discharges, and underestimates concentrations in the middle discharge range. Estimates of the error variance (mean squared departure of observed minus predicted) will be too high, and any subsequent analysis of the residuals will likely be very misleading. The quadratic fit is certainly an improvement on the linear fit, but how appropriate it is will need to be determined. At the lower discharge values (say less than 90 cubic meters per second [ $m^3/s$ ]) virtually all of the residuals are negative. This is because the quadratic model doesn't have the flexibility to depict the rather steep positive slope that is apparent in the scatterplot at discharge values below about 150  $m^3/s$ . The reason for the poor fit in this range is that the quadratic model must be symmetrical, and it must also accommodate the small negative slope that is apparent at discharges above about 1,000  $m^3/s$ . In fact, we can argue that it may be seriously overstating the steepness of the relation at high discharges. This gets at one of the fundamental problems of any regression relation: the pattern of the data in a particular range of  $x$  values will influence the location of the fitted curve in a range of  $x$  values far removed from it. Simply put, in terms of this example, the extreme low concentration values at discharges around 75  $m^3/s$  are

actually influencing the location of the fitted curve at discharges of such as 2,000 or 5,000 m<sup>3</sup>/s. Smoothing techniques are designed to defeat this inadequacy of regression and assure that estimates at one extreme of the range of  $x$  values are not influenced by data from the other extreme, but rather are primarily influenced by those that are nearby. For example, Helsel and Ryker (2002) compared a multivariate loess smooth to kriging, finding that the loess surface was far less susceptible to the influence of distant outliers than was kriging.

How might we make such a smooth representation of  $y$  as a function of  $x$ ? Some of the first methods developed were the moving average (in R this is implemented in the function `rollmean` in the `zoo` package of Zeileis and Grothendieck [2005]) and moving median (in R this is implemented in the function `runmed` in R). These techniques are conceptually simple to explain but some of their properties are not highly desirable. They may result in a curve that can be quite jagged and they also have deficiencies in how they handle the end points of the curve. The methods have been replaced by a set of techniques that are generally known as locally weighted scatterplot smoothing.

The most widely used smoothing algorithm is loess (an acronym of LOcal regrESSion). It was first introduced by Cleveland (1979) and expanded by Cleveland and Devlin (1988) as a tool to compute both bivariate (one explanatory variable) and multivariate (multiple explanatory variables) surfaces. The `loess` function in R computes a loess smooth. A related method is called lowess (an acronym of LOcally WEighted Scatterplot Smoothing) with an associated function in R called `lowess`. The function called `lowess` is more complex than `loess`, employing an iterative method that weights the data based on the size of their residuals. In this text we describe both methods and make some observations about their proper use and variations, as well as extensions of the concept. There continues to be some confusion between the two methods, as both have been developed largely by the same group of people, and particularly because the pronunciation of the two names is identical.



**Figure 10.12.** Nitrate concentrations as a function of daily mean discharge during the months of June through September of the years 1990–2008 for the Iowa River at Wapello, Iowa, showing linear (solid line) and quadratic (dashed line) fit, estimated as the natural log of concentration as a function of natural log of discharge.

### 10.3.1 Loess Smooths

Assume we have a set of  $n$  measurements of the dependent variable,  $y$ , over a range of values of the explanatory variable,  $x$ . The  $n$  data pairs are sorted from the lowest  $x$  to the highest  $x$ . We will index the values as  $x_1, x_2, \dots, x_n$ , from lowest to highest values. The  $y$  values associated with these  $x$  values are  $y_1, y_2, \dots, y_n$ . If there are any ties in the set of  $x$  values they can be placed into the vector of  $x$  values in any order and it will have no consequences for the result of the computation. Now, let's assume we want to estimate the value of  $y$ , call it  $y^*$ , that would be associated with a value of  $x$  denoted as  $x^*$  (where  $x^*$  may be a vector of values for multiple explanatory variables). The values for  $x^*$  may be any arbitrary value within the range  $(x_1, x_n)$  and need not be equal to any specific observed  $x_i$  value. Conceptually,  $x^*$  could lie outside the range of the observed  $x$  values, inappropriately taking loess from the category of a smoothing algorithm into the realm of extrapolation, estimating  $y$  at values of  $x$  outside the range of our dataset. Given that loess is entirely empirical and has no fundamental connection to physical or chemical principles, the use of it for extrapolation is not recommended.

The estimate,  $y^*$ , is determined by the use of weighted regression, where each observation has a weight that determines its influence on the computed regression equation. For loess the weights on each of the  $n$  observations are determined by their distance from  $x^*$ . An observation at  $x^*$  (if such an observation exists) has a weight of 1, and weights decrease for observations further away in units of the explanatory variable(s).

The default in the `loess` function in R is that this weighted regression is a quadratic of  $y$  as a function of  $x$ , so we would write the model for one explanatory variable as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad (10.14)$$

In estimating a  $y^*$  for  $x^*$  there is a set of weights applied to all of the  $x$  values based on their proximity to  $x^*$ . The weights are determined using the tri-cubed weight function, although other weight functions are sometimes used. This function has a shape similar to that of a normal distribution, but it differs from the normal in that it goes to zero at large distances, rather than asymptotically approaching zero, and it is rather flat in the vicinity of its maximum value. The weight function requires the specification of a maximum distance,  $d_{max}$ , which is the distance at which the weight function goes to zero. This maximum distance,  $d_{max}$ , is sometimes called the half-window width. The form of the weight function is defined as the distance from observation  $i$  to  $x^*$  as  $d_i$

$$d_i = |x_i - x^*|. \quad (10.15)$$

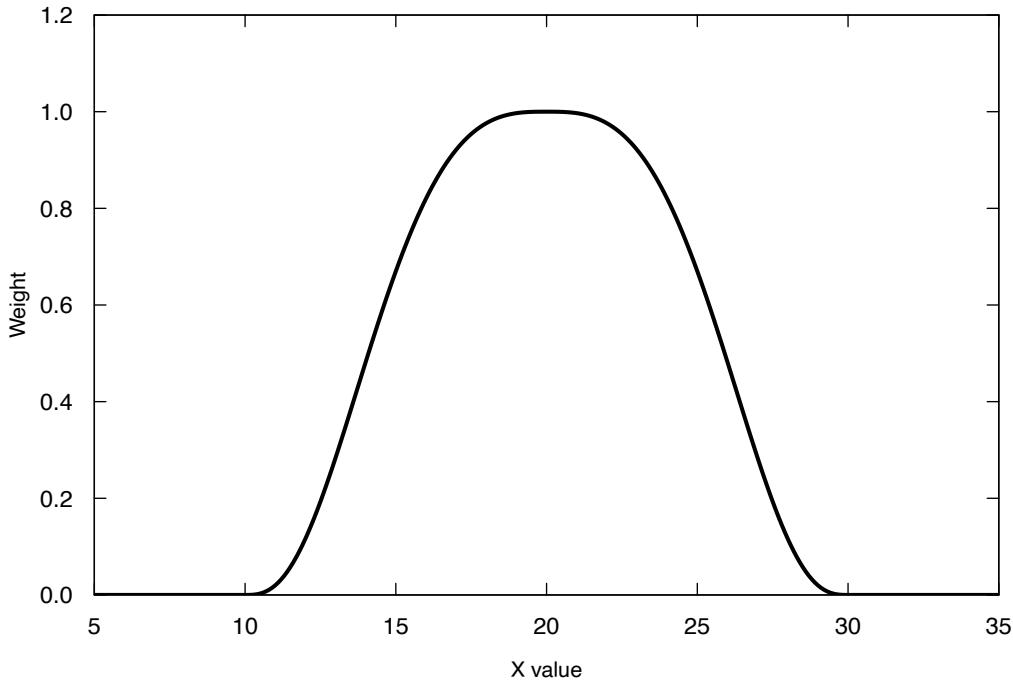
The weight for observation  $i$  is  $w_i$ ,

$$w_i = \begin{cases} 1 - \left( \frac{d_i}{d_{max}} \right)^3 & \text{if } d_i < d_{max} \\ 0 & \text{if } d_i \geq d_{max} \end{cases}. \quad (10.16)$$

The shape of the tri-cube weight function, in this case where  $d_{max}$  is set to a value of 10, is illustrated in figure 10.13.

Using those weights to estimate the parameters of the regression shown in equation 10.14, we can then take the estimated values of the three parameters and set  $x=x^*$  and determine the value of  $y^*$ . The process is repeated for a large set of  $x^*$  values, covering the full range from  $x_1$  to  $x_n$ , estimating new weights and new regression coefficients for each value of  $x^*$  in order to estimate a new value of  $y^*$  for each  $x^*$ . The weights and coefficients will all change with any change in  $x^*$ . Because the weight function is one that tapers gradually to zero, it guarantees that the set of values of  $y^*$  for a set of closely spaced  $x^*$  values will be quite smooth.

To implement this approach, we need some method for selecting the value of  $d_{max}$ . This value is set using a smoothing parameter called the span, which is denoted as `span` in the `loess` function of R. If  $\text{span} < 1.0$ , then  $d_{max}$  will be set so that  $n \cdot \text{span}$  observations have weights that are  $> 0$ , where  $n$  is the sample size of the dataset. Note that  $d_{max}$  will vary as a function of  $x^*$ . The default value for `span` is 0.75. Using



**Figure 10.13.** Graph of the tri-cube weight function, where  $d_{max}=10$  and  $x^*=20$ .

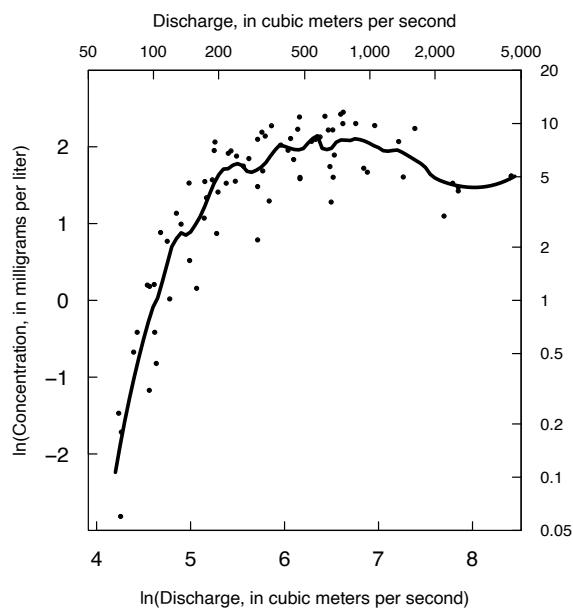
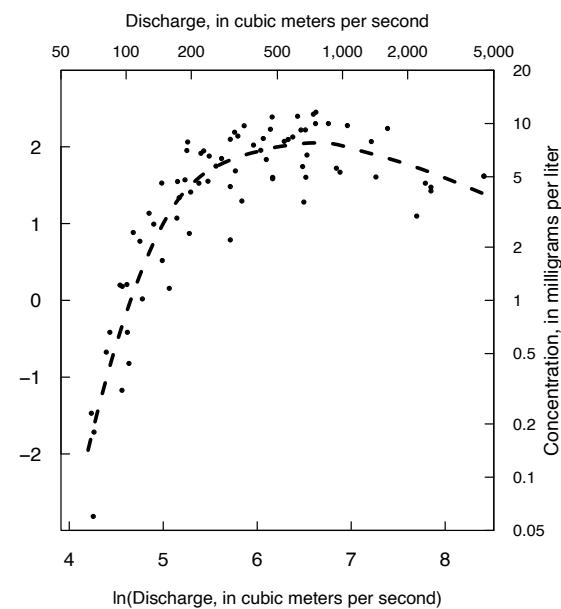
the example dataset shown in figure 10.12, with sample size 76, the  $d_{max}$  values will be set so that 57 of the 76 values will have distances of less than  $d_{max}$  from the given value of  $x^*$  and the remaining 19 observations will have zero weight. In the case where  $\text{span} \geq 1$ , the  $d_{max}$  value is  $\text{span}$  times the maximum  $d_i$  value in the dataset, which means that all observations are given some weight. The entire loess smooth curve is constructed by implementing this weighted regression process at a large number of  $x$  values that cover the range of the observed  $x$  values, and the curve is drawn by connecting the consecutive  $(x^*, y^*)$  pairs with straight lines.

An obvious question is how one should select the argument `span`. The selection of the `span` value is subjective. The larger the `span` value is, the smoother the curve will be. The goal should be to achieve a curve that is faithful to the overall shape of the dataset but devoid of oscillations which (in the opinion of the data analyst) are simply artifacts of the specific random set of observations (noise). Two panels are shown in figure 10.14, the left panel (*A*) uses a `span` value of 0.25 and the right panel (*B*) uses a `span` value of 0.75. The `span` value of 0.25 is clearly too small and results in a set of minor oscillations in the curve that are unlikely to be meaningful. In particular, it depicts a substantial upturn in the curve near the maximum  $x$  value. This is a result of the fit being too dependent on a few values that fall near that edge. The loess with `span`=0.75 looks reasonable. The default value often works very well but users can be more specific about finding the best `span` by trying a range of values and selecting the smallest value not subject to a series of wiggles caused by random artifacts of the sample.

### 10.3.2 Lowess Smooths

Originally named “robust locally weighted regression” (Cleveland, 1979), lowess begins by computing the loess smooth for every point  $x_i$ . Residuals from the loess estimates  $r_i$  in the  $y$  direction are then computed (eq. 10.17) and the bi-square weight function is used to compute robustness weights,  $R_i$  (eq. 10.18).

$$r_i = |y_i - y^*| \quad (10.17)$$

**A.****B.**

**Figure 10.14.** Graphs of smooths of nitrate concentration as a function of daily mean discharge during the months of June through September of the years 1990–2008 for the Iowa River at Wapello, Iowa. *A*, a loess smooth with  $\text{span} = 0.25$ ; *B*, a loess smooth with  $\text{span} = 0.75$  (the default value).

$$R_i = \begin{cases} \left\{ 1 - \left( \frac{r}{r_{\max}} \right)^2 \right\}^2 & \text{if } r_i < r_{\max}, \\ 0 & \text{if } r_i \geq r_{\max} \end{cases}, \quad (10.18)$$

where  $r_{\max} = 6 \cdot \text{median}|r_i|$  (or  $6 \cdot \text{MAD}$ , where MAD is the median absolute deviation). Multiplying the loess weights  $w_i$  (eq. 10.16) by the robustness weights  $R_i$  (from eq. 10.18) produces final weights used in a second tier of weighted regressions, one new weighted regression for every  $x_i$  in the dataset. The result of this second tier of weighted regressions is a lowess smooth of estimated  $y$  values. The purpose of lowess is to diminish the effect of outlying observations in the  $y$  direction as compared to their effect on a loess smooth (Chambers and others, 1983).

Commercial statistics software commonly includes lowess smooths. The biggest drawback of lowess is that it has never been extended beyond bivariate applications, whereas loess has. A second drawback of lowess is that estimates are not provided for observations other than for the original  $x_i$ . People have linearly interpolated between the predicted lowess values for each  $x_i$ , but the loess formula allows more direct computation of somewhat smoother estimates for new  $x$  values. For the case of one explanatory variable, and without noticeable outliers, loess and lowess will be nearly identical.

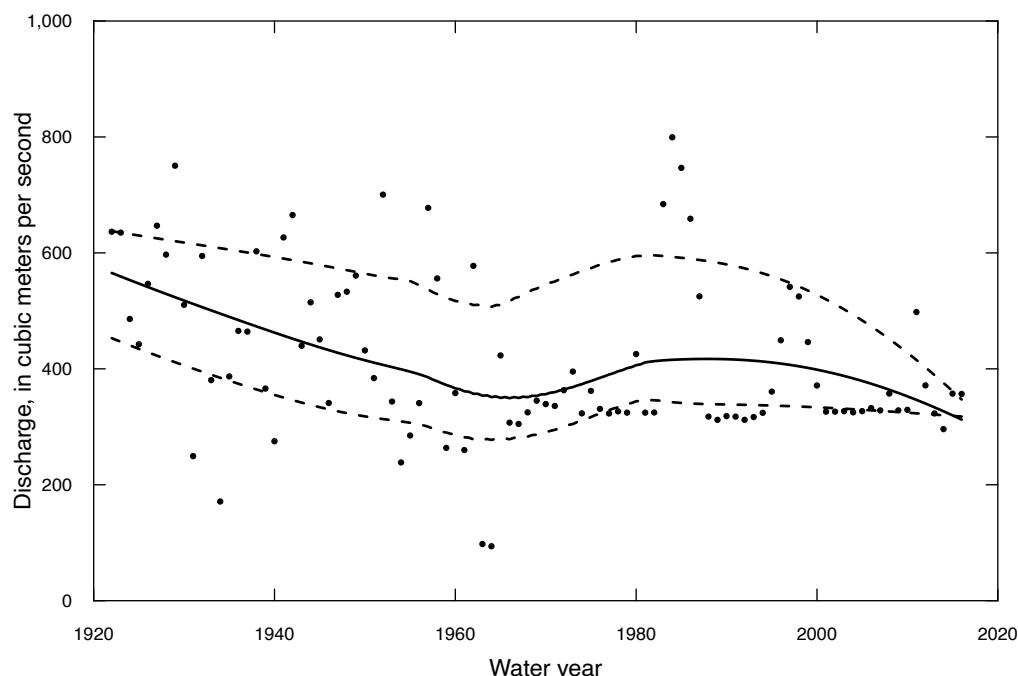
### 10.3.3 Upper and Lower Smooths

In addition to plotting the central smooth of the dataset, one may want a representation of the changing spread or symmetry of the data around the smooth. This is done with an upper and lower smooth (Cleveland and McGill, 1984b). These elements represent a smoothed version of upper and lower quartiles of the conditional distribution of  $y$  as a function of  $x$ . They are computed by separately smoothing the positive and negative residuals from the loess or lowess smooth, and adding these smooths to the central smooth. An example of this, a loess smooth of the annual mean daily discharge for the years 1922 through 2016 for the Colorado River at Lees Ferry, Arizona, is shown in figure 10.15.

The decreased mean flow over time is a function of increased consumptive use of water and climate change (decreasing snowpack in particular). The very low values in the early 1960s represent the years just after the completion of Glen Canyon Dam, located just upstream of the Lees Ferry streamgage, when a large part of the flow of the river was allowed to remain in storage behind the dam in order to build capacity for hydropower production. The upper and lower smooths are asymmetrical around the middle smooth, particularly since the dam was completed. This reflects the degree to which the dam is augmenting the flow of the river in low-flow years. Also notable is that during the most recent years (say 2000–16) not only is the middle smooth declining (a decrease in mean flow) but the variability has markedly decreased. This indicates that the reservoir is being used to control the delivery of water downstream during dry years, whereas wetter years continue to be more variable. Most recently, the interannual variability of discharge (as represented by the upper and lower smooths) has become much lower than in previous years and is converging on the smooth of the annual mean discharge. In short, streamflow is becoming less plentiful but also much less variable through time. This presentation of the data helps capture some of those details.

### 10.3.4 Use of Smooths for Comparing Large Datasets

A valuable use of smooths is to compare and contrast multiple, large datasets. Plotting all the data points in a scatter plot and using different symbols to represent different groups of data generally does not provide the clarity necessary to distinguish similarities and differences among groups. Plotting the individual data group's loess smooth curves (with some way of distinguishing between groups) without showing the actual data points, may provide insights into the group characteristics. For example, Welch and others (1988) used lowess to describe the relation between arsenic and pH in four physiographic regions of the western United States (fig. 2.23). Thousands of data points were involved; a scatterplot would have shown nothing but a blob of data. The smooths clearly illustrated that arsenic concentrations increased with increasing pH in three regions, and no increase was observed in the fourth. Schertz and Hirsch (1985) also used smooths to illustrate regional patterns in atmospheric precipitation chemistry. They used one smooth per station to display simultaneous changes in sulfate and other chemical concentrations occurring over broad regions of the country (fig. 10.16). These relations would have gone unnoticed using scatterplots—the underlying patterns would have been obscured by the proliferation and scatter of the data.



**Figure 10.15.** Graph of annual mean daily discharge of the Colorado River at Lees Ferry, Arizona, for water years 1922–2016. The graph shows the loess smooth for the data (solid curve) and the upper and lower smooths (dashed lines).

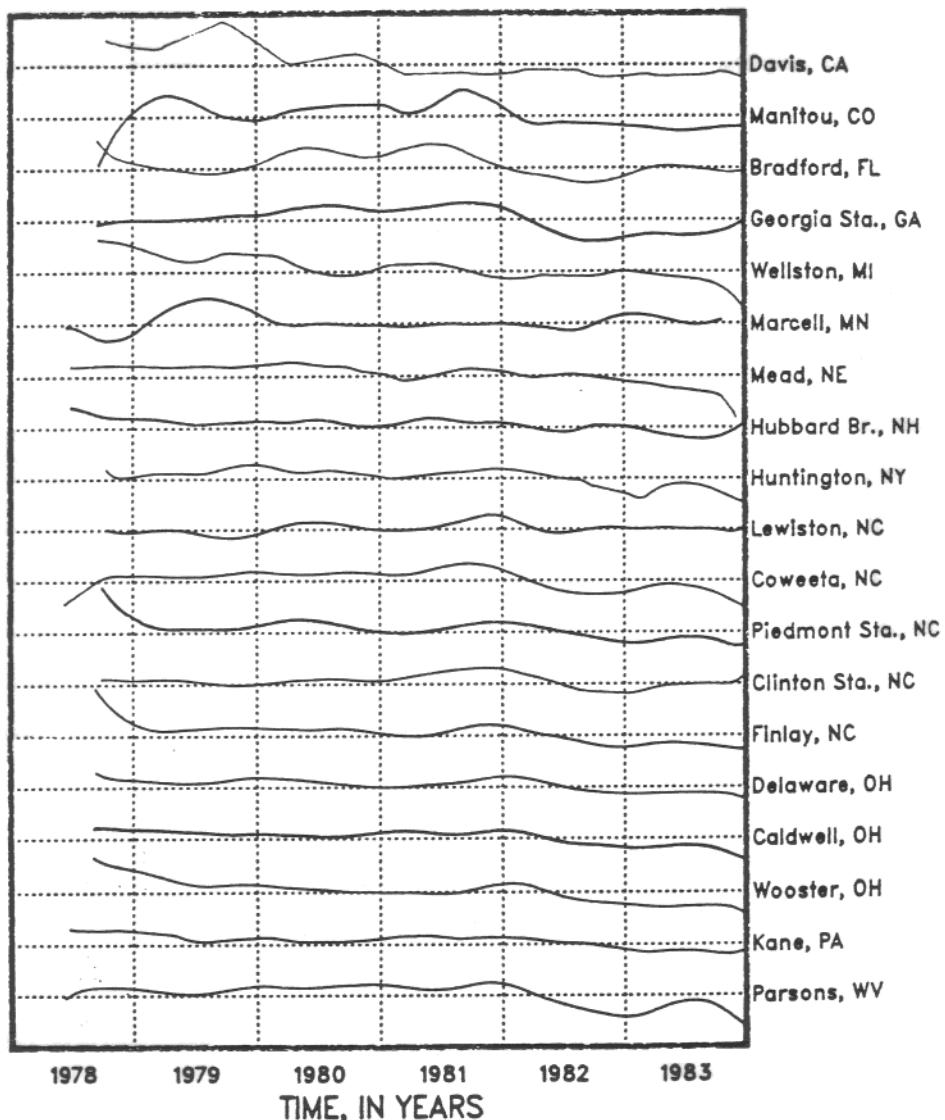


Figure 10.16. Plot of lowess smooths of sulfate concentrations at 19 stations, 1979–83 (Schertz and Hirsch, 1985).

### 10.3.5 Variations on Smoothing Algorithms

There are many appropriate types of variations on the general idea of smoothing that can be considered. Data analysts should be careful to describe the variations that they may use in their applications. One possible variation is to base  $d_{max}$ , not on some type of span parameter (a proportion of the dataset), but rather to set it based on distance in the units in which the  $x$  variable is measured. For example, in measuring some type of biogeochemical process for which rates vary with temperature, one might want to do a loess where  $d_{max}$  is set by the distance as measured by temperature, rather than simply using a standard fraction of the dataset. The analyst would state that the smoothing is done with weighted regressions with a half width of  $2^\circ \text{C}$  (for example). One would have to write a specialized function to do this, but the loess concept is simple enough that this can be done by repeated use of the `lm` function in R.

Another example of a departure from the standard loess model is where the  $x$  variable is cyclic, such as hour of the day or day of the year. A standard loess approach would not be appropriate if one of these were used as the  $x$  value. This is because the loess model would see values, say from late December, as being very distant from those in early January, or 11:59 PM being very distant from 12:01 AM, when in fact they are quite close together. Specialized code can be written that measures time distances in a circular manner rather than linear in order to achieve the desired continuity through the entire year. The weighted regression on time, discharge, and season (WRTDS) method (Hirsch and others, 2010) introduced in chapter 12 uses both of these types of variations on the basic principles of loess.

## Exercises

1. For the data below,
  - A. Compute the Kendall slope estimator.
  - B. Compute Kendall's  $\tau$ .
  - C. Compute the Theil-Sen equation.
  - D. Compute the significance level of the test.

$y$	10	40	30	55	62	56
$x$	1	2	3	4	5	6

2. One value has been altered from the first exercise. Again, compute the slope estimate, intercept,  $\tau$ , and significance level. By how much have these changed in response to the one (large) change in  $y$ ? Also compute a 95-percent confidence interval on the slope estimate.

$y$	10	40	30	55	200	56
$x$	1	2	3	4	5	6

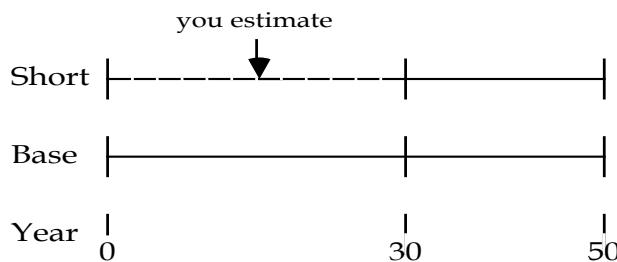
3. Williams and Wolman (1984) relate the lowering of streambed elevation downstream of a major dam to the number of years following its installation. Calculate a linear least-squares regression of bed lowering ( $L$ ) as the response variable, versus years (Yrs) as the explanatory variable, and compute its  $R^2$ .

Yrs	Lowering (L, in meters)	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

Now compute either a lowess or loess smooth of the data. Plot the smooth and regression line along with a scatterplot of the data. Describe how well each represents the data.

4. Record Extension

Monthly discharges in million cubic meters per month for September at two rivers are given in `Joint10_4.RData`. Data are for years 31 to 50 for "Short" and "Base". The two sites are close enough that the data are reasonably well correlated with each other. Using the 20 years of joint record, compute a regression line and an LOC, saving the slopes and intercepts (regression parameters are stored for you within the regression object in R). Using these two linear models, estimate the early 30 years of record at "Short" from the 30 years of early record at "Base" found in `Early10_4.RData`.



Take the estimated 30-year record at “Short” produced by both methods 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 that actually occurred (the Actual.Short.streamflow column in `Early10_4.RData`). Which technique, regression or LOC, most closely matches the observed streamflow characteristics at “Short”?

5. The pulp liquor waste contamination of shallow groundwater (see exercise 7.1) is revisited. Of interest is now the relation between pH and chemical oxygen demand (COD) in samples taken from all three piezometers. Calculate a straight line which best describes the innate relation between these two chemical constituents (either pH or COD could be predicted from the other).

<b>pH</b>	<b>COD</b>	<b>pH</b>	<b>COD</b>	<b>pH</b>	<b>COD</b>
7.0	51	6.3	21	8.4	283
7.2	60	6.9	17	7.6	2,170
7.5	51	7.0	34	7.5	6,580
7.7	3,600	6.4	43	7.4	3,340
8.7	6,900	6.8	34	9.3	7,080
7.8	7,700	6.7	43	9.0	10,800



# Chapter 11

## Multiple Linear Regression

---

*The flood with a 1-percent annual exceedance probability (AEP) is to be estimated for locations without streamflow gages using basin characteristics at those locations. A regression equation is first developed relating the 1-percent AEP flood to several basin characteristics at sites that have a streamgage. Each characteristic used is known to influence the magnitude of the 1-percent AEP flood, has already been used in adjoining states, and thus will be included in the equation regardless of whether it is significant for any individual dataset. Values for the basin characteristics at each ungaged site are then input to the multiple regression equation to produce the 1-percent AEP flood estimate for that site.*

*Residuals from an ordinary least squares 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 mechanisms 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. All-subsets regression or similar procedures are performed to select the most important variables, and the subsequent regression equation is then used to predict concentrations.*

Multiple linear regression (MLR) is the extension of simple linear regression (SLR) to the case of multiple explanatory variables. The goal of this relation 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 MLR model are explained, as are the common pitfalls, such as multicollinearity and relying on  $R^2$ . The mathematics of MLR—best handled by matrix notation—will not be extensively covered here, see Kutner and others (2004) or Davis (2004) for that. In this chapter we refer to SLR and MLR to distinguish between regressions with only one explanatory variable versus those with two or more explanatory variables. In either case these methods both use an ordinary least squares (OLS) approach to estimation.

### 11.1 Why Use Multiple Linear Regression?

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, such as application rates of fertilizers or pesticides, and subsurface conditions, such as soil permeability or depth to groundwater.

The use of MLR might also be indicated by the residuals from an 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 the season the data point was collected in). Analysis of a residuals plot (described in chap. 9) may also show that patterns of residuals occur as a function of some categorical grouping representing a special condition, for example, on the rising limb of a hydrograph, at cultivating time, during or after frontal storms, in wells with PVC casing, or measurements taken before 10:00 a.m. 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 relations should lead to definition of an appropriate explanatory variable and its inclusion in the model if it significantly improves the fit.

## 11.2 Multiple Linear Regression Model

The MLR model will be denoted

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon , \quad (11.1)$$

where

- $y$  is the response variable,
- $\beta_0$  is the intercept,
- $x_i$  is the  $i$ th explanatory variable,
- $\beta_1$  is the slope coefficient for the first explanatory variable,
- $\beta_2$  is the slope coefficient for the second explanatory variable,
- $\beta_k$  is the slope coefficient for the  $k$ th explanatory variable, and
- $\varepsilon$  is the remaining unexplained noise in the data (the error).

There are  $k$  explanatory variables, some of which may be related or correlated to each other (such as the previous 5-day's total rainfall and the previous 1-day rainfall). Sometimes the explanatory variables are referred to as independent variables; however, it best to avoid calling them independent because they may not be independent of each other. Calling them explanatory variables describes their purpose: to explain the variation in the response variable.

In SLR, the regression function is a line. In MLR, the regression function describes a plane (when there are two explanatory variables) or a hyperplane (when there are more than two explanatory variables) and is sometimes called a regression surface.

### 11.2.1 Assumptions and Computation

The assumptions necessary for MLR are the same as those for SLR described in table 9.2 of chapter 9, except that there are now multiple explanatory variables. As described in chapter 9 for SLR, estimation of the parameters in MLR is a minimization problem. MLR uses maximum likelihood methods to estimate parameters describing a regression surface. The estimates that meet the least squares criterion minimize the squared distances between the observed response data and the regression surface. The solution has the same properties listed in chapter 9 for the least squares solution.

## 11.3 Hypothesis Tests for Multiple Regression

### 11.3.1 Nested F-Tests

The most important hypothesis test for MLR is the  $F$ -test for comparing any two nested models. Let model “ $s$ ” (eq. 11.2) be the simpler MLR model:

$$y_s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_s . \quad (11.2)$$

It has  $k+1$  parameters including the intercept, with degrees of freedom ( $df_s$ ) of  $n-(k+1)$ , where  $k$  is the number of explanatory variables and  $n$  is the sample size. As in SLR, the degrees of freedom equals the number of observations minus the number of parameters estimated. The error sum of squares (defined in table 9.1) is  $SSE_s$ .

Let model “ $c$ ” (eq. 11.3) be the more complex regression model (meaning it has more explanatory variables than model  $s$ ):

$$y_c = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_m x_m + \varepsilon_c . \quad (11.3)$$

It has  $m+1$  parameters and residual degrees of freedom ( $df_c$ ) of  $n-(m+1)$ . Its error sum of squares 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

$$H_0 : \beta_{k+1} = \beta_{k+2} = \dots = \beta_m = 0 ,$$

and the alternative hypothesis is

$$H_1 : \text{at least one of these } m-k \text{ coefficients is not equal to zero.}$$

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

$$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{SSE_c / df_c} , \quad (11.4)$$

where  $(df_s - df_c) = m - k$  (Neter and others, 1985).

If  $F$  exceeds the tabulated value of the  $F$ -distribution with  $(df_s - df_c)$  and  $df_c$  degrees of freedom for the selected significance level (say  $\alpha=0.05$ ), 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 one (or more) of the additional coefficients in the more complex model is 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. See also section 11.6 on choosing the best linear model.

### 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

$$y_s = \beta_0 + \varepsilon_s , \quad (11.5)$$

where

$$\beta_0 = \bar{y} .$$

The rules for a nested  $F$ -test still apply: the  $df_s = n-1$  and  $SSE_s = (n-1) \times$  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 and does not indicate whether useful predictions can be made. 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 (SAS Institute Inc., 2014) and is performed in R by fitting the complex and simpler model separately and comparing them using the `anova` function. Here the complex model has only one additional explanatory variable over the simpler model, so that  $m=k+1$ . The partial  $F$ -test evaluates whether the  $m$ th 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. The  $F$  statistics on a coefficient will change depending on what other variables are in the model, thus we cannot determine if the variable  $m$  belongs in the model. What can be determined is whether  $m$  belongs in the model in the presence of the other variables. Note the comparison is valid only if the models are fitted to the same dataset; this can be a problem if values are missing for some potential explanatory 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 (described in chap. 9). 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 Student's  $t$ -distribution with  $n-(m+1)$  degrees of freedom.  $H_0$  is rejected if  $|t| > t_{1-\alpha/2;n-(m+1)}$ . For a two-sided test with  $\alpha=0.05$  and sample sizes  $n$  of 20 or more, the critical value of  $t$  is  $|t| \geq 2$ . 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  $t$ -statistics 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 MLR. These procedures do not guarantee that the best model will be obtained, as discussed later, better procedures are available for doing so.

## 11.4 Confidence and Prediction Intervals

Confidence intervals can be computed for the regression slope coefficients,  $\beta_k$ , 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 SLR in chapter 9 but require matrix manipulations (linear algebra) for computation. A brief discussion of them follows. More discussion can be found in chapter 9 and more complete treatment can be found in many statistics textbooks, such as Draper and Smith (1981), Walpole and Myers (1985), Kutner and others (2004), and Montgomery and others (2012).

### 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 ones for the intercept term, can be combined into a matrix  $X$  (eq. 11.6):

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}. \quad (11.6)$$

$X$  is used in MLR to compute the variance-covariance matrix  $\sigma^2(X'X)^{-1}$ , where  $(X'X)^{-1}$  is often called the  $X$  prime  $X$  inverse matrix, where  $X'$  ( $X$  prime) is the transpose of the  $X$  matrix and  $(\cdot)^{-1}$  denotes the inverse of the quantity in parentheses. Elements of  $(X'X)^{-1}$  for three explanatory variables are as follows:

$$(X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} & C_{03} \\ C_{10} & C_{11} & C_{12} & C_{13} \\ C_{20} & C_{21} & C_{22} & C_{23} \\ C_{30} & C_{31} & C_{32} & C_{33} \end{bmatrix}. \quad (11.7)$$

When multiplied by the error variance,  $\sigma^2$  (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, and 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  $\beta_0$  through  $\beta_k$  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,  $\sigma^2$ . A  $100 \cdot (1 - \alpha)$ -percent confidence interval on  $\beta_j$  is

$$\hat{\beta}_j - t_{(1-\alpha/2;n-p)} \sqrt{s^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{(1-\alpha/2;n-p)} \sqrt{s^2 C_{jj}} , \quad (11.8)$$

where  $C_{jj}$  is the diagonal element of  $(X'X)^{-1}$  corresponding to the  $j$ th explanatory variable. Often printed is the standard error of the regression coefficient:

$$se(\hat{\beta}_j) = \sqrt{s^2 C_{jj}} . \quad (11.9)$$

Note that  $C_{jj}$  is a function of the other explanatory variables as well as the  $j$ th. Therefore, the interval estimate, like  $\hat{\beta}_j$  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  $100 \cdot (1 - \alpha)$ -percent 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.

$$\hat{y}_0 - t_{(1-\alpha/2;n-p)} \sqrt{s^2 x_0' (X'X)^{-1} x_0} \leq \mu(y_0) \leq \hat{y}_0 + t_{(1-\alpha/2;n-p)} \sqrt{s^2 x_0' (X'X)^{-1} x_0} \quad (11.10)$$

The variance of the mean is the term under the square root symbol. 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,  $h_i$ , which expresses the distance that  $x_0$  is from the center of the data.

### 11.4.4 Prediction Intervals for an Individual $y$

A  $100 \cdot (1 - \alpha)$  percent 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.

$$\hat{y}_0 - t_{(1-\alpha/2;n-p)} \sqrt{s^2 x_0' (X'X)^{-1} x_0} \leq y_0 \leq \hat{y}_0 + t_{(1-\alpha/2;n-p)} \sqrt{s^2 (1 + x_0' (X'X)^{-1} x_0)} \quad (11.11)$$

As in SLR, the prediction interval for a new observation is wider than the confidence interval for a mean response because it takes into account the uncertainty in parameter estimates and in the unexplained variability in  $y$  (see chap. 9 for more discussion).

## 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 Diagnostic 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 relation 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 relation between  $e$  and  $x_2$  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-regression plots (also called added-variable plots) should be constructed to view the partial relation between the response and an explanatory variable, adjusted for the other explanatory variables (Fox and Weisberg, 2011).

The partial residual is

$$e_j^* = y - \hat{y}_j , \quad (11.12)$$

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. The residual is the part of  $y$  that is not explained by all the regressors except  $x_j$ . This partial residual is then plotted versus another residual

$$x_j^* = x_j - \hat{x}_j , \quad (11.13)$$

where  $\hat{x}_j$  is the  $x_j$  predicted from a regression against all other explanatory variables. Therefore,  $x_j$  is treated as a response variable in order to compute its adjusted value. The partial plot ( $e_j^*$  versus  $x_j^*$ ) describes the relation between  $y$  and the  $j$ th 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 (see fig. 9.4 and associated text in chap. 9 for more on transformations). The `avPlots` function in the `car` package (Fox and Weisberg, 2011) for R will generate partial-regression/added-variable plots.

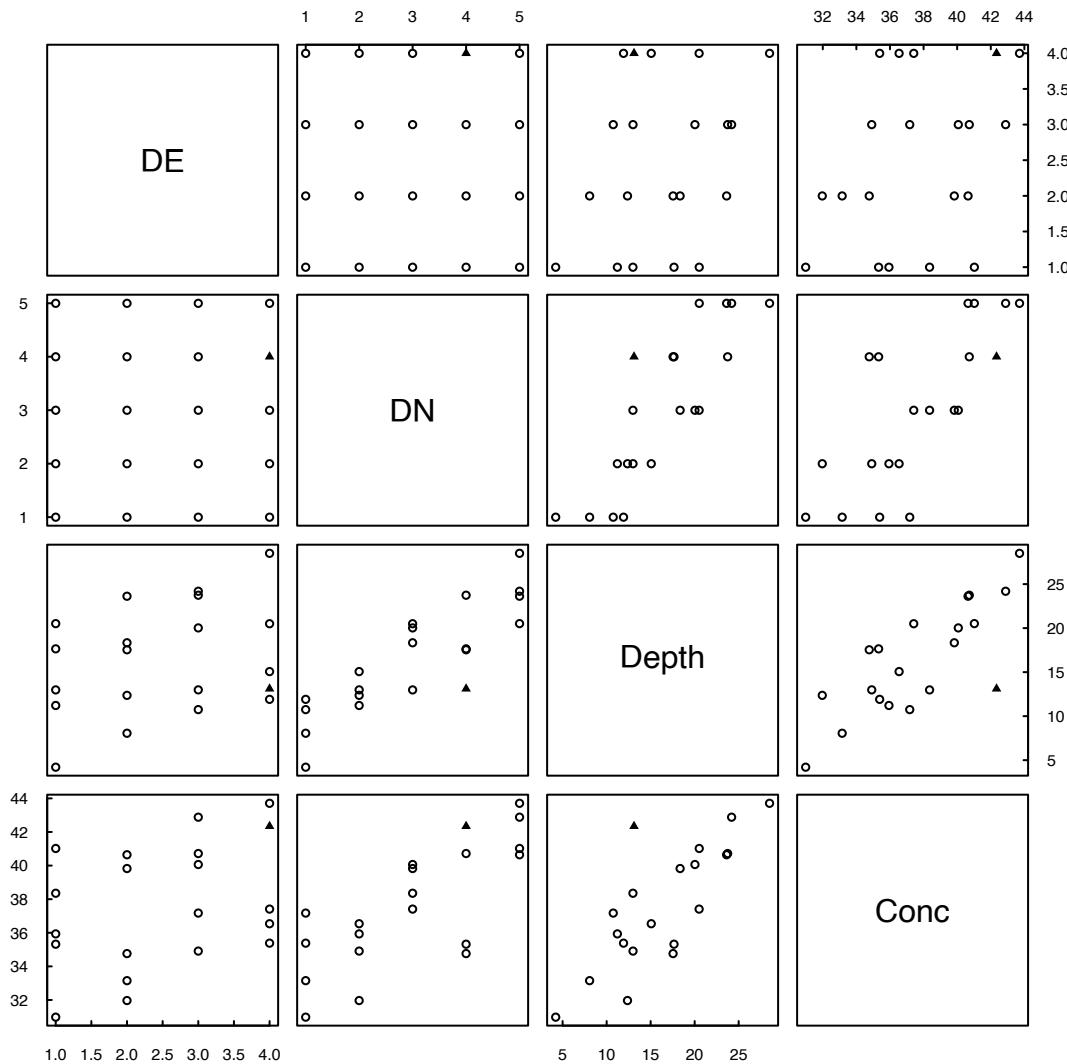
Another type of plot that is helpful in determining whether to transform one or more of the  $x$  variables is a component plus residual (component + residual) plot. These plots examine whether the explanatory variables have linear relations with the response variable by adding the linear component of the relation between  $y$  and  $x_j$  back to the residual from the full model and plotting this component plus residual against  $x_j$ . Component + residual plots are better at highlighting nonlinearities than the partial-regression plots. The function to generate these plots is available in the `car` package for R.

### 11.5.2 Leverage and Influence

The regression diagnostics of chapter 9 are much more important in MLR than in SLR. It is difficult when performing MLR 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 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 unusual explanatory variable values. Such points warrant further checking as possible errors, or they may indicate a poor model (transformation required, relations not linear, and so forth).

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 dataset, 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.



**Figure 11.1.** Scatterplot matrix for the variables listed in table 11.1 (observation 16 is shown as a solid triangle). DE, distance east; DN, distance north; Depth, well depth; Conc, concentration.

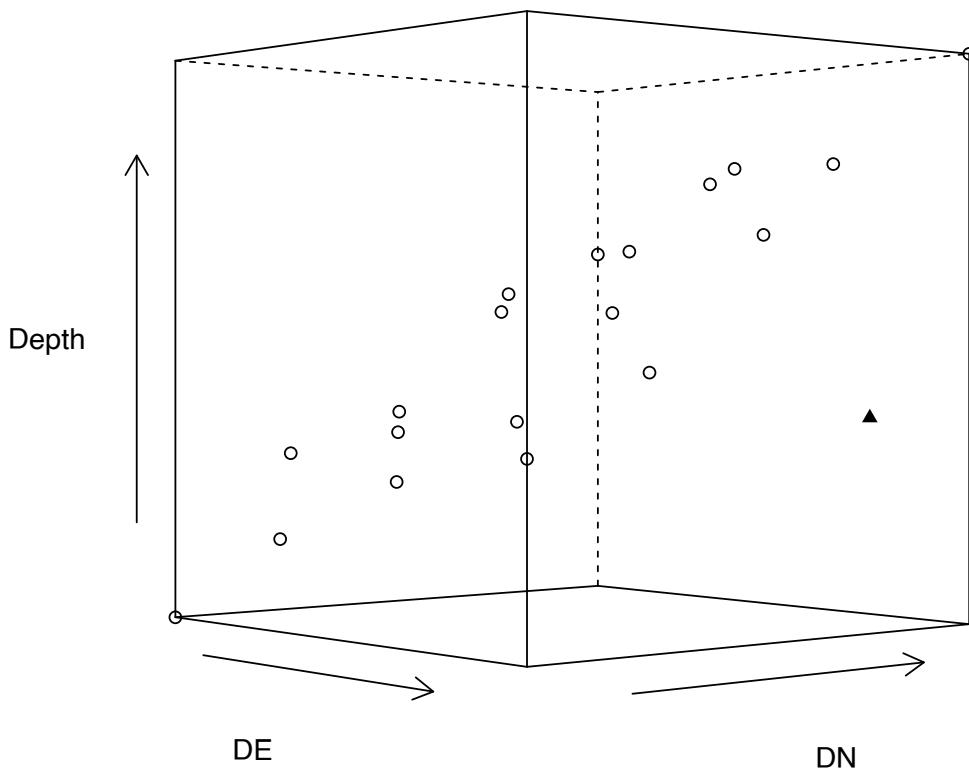
### Example 11.1. Chemical concentrations in an aquifer.

Variations in chemical concentrations within a steeply dipping aquifer are to be described by location and depth. The data are concentrations (Conc) plus three coordinates: distance east (DE), distance north (DN), and well depth (Depth). Data are artificial and were generated using  $\text{Conc} = 30 + 0.5 \cdot \text{Depth} + \varepsilon$ . Any acceptable regression model should closely reproduce this true model, and should find Conc to be independent of DE and DN. Pairs plots of the variables (fig. 11.1) do not reveal any extreme outliers in the dataset, yet compared to the critical leverage statistic  $h_i = 3(p/n) = 0.6$ , and critical influence statistic  $DFFITS = 2\sqrt{p/n} = 0.9$ , where  $n$  is the sample size, 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 potential explanatory variables, even though its individual values for the three explanatory variables are not unusual.

**Table 11.1.** Data and diagnostics for chemical concentrations used in example 11.1.

[DE, distance east; DN, distance north; Depth, well depth; Conc, concentration;  $h_i$ , critical leverage statistic; DFFITS, critical influence statistic]

<b>DE</b>	<b>DN</b>	<b>Depth</b>	<b>Conc</b>	<b><math>h_i</math></b>	<b>DFFITS</b>
1	1	4.2122	30.9812	0.289433	-0.30866
2	1	8.0671	33.1540	0.160670	-0.01365
3	1	10.7503	37.1772	0.164776	0.63801
4	1	11.9187	35.3864	0.241083	-0.04715
1	2	11.2197	35.9388	0.170226	0.42264
2	2	12.3710	31.9702	0.086198	-0.51043
3	2	12.9976	34.9144	0.087354	-0.19810
4	2	15.0709	36.5436	0.165040	-0.19591
1	3	12.9886	38.3574	0.147528	0.53418
2	3	18.3469	39.8291	0.117550	0.45879
3	3	20.0328	40.0678	0.121758	0.28961
4	3	20.5083	37.4143	0.163195	-0.47616
1	4	17.6537	35.3238	0.165025	-0.59508
2	4	17.5484	34.7647	0.105025	-0.77690
3	4	23.7468	40.7207	0.151517	0.06278
4	4	13.1110	42.3420	0.805951	4.58558
1	5	20.5215	41.0219	0.243468	0.38314
2	5	23.6314	40.6483	0.165337	-0.08027
3	5	24.1979	42.8845	0.160233	0.17958
4	5	28.5071	43.7115	0.288632	0.09397



The depth value for observation 16 was a typographical error, and should be 23.1110 instead of 13.1110. What does this error and resulting high leverage point do to a regression of concentration versus the three explanatory variables? R code and output for a MLR model of Conc explained by DE, DN, and Depth is shown below. From the *t*-ratios it can be seen that DN and perhaps DE appear to be significantly related to Conc, but that Depth is not. This is exactly opposite of what is known to be true.

```
> mod1 <- lm(Conc ~ DE + DN + Depth, data=Chap11Ex1)
> summary(mod1)
```

Call:

```
lm(formula = Conc ~ DE + DN + Depth, data = Chap11Ex1)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.101	-1.006	0.106	1.645	2.726

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.90857	1.58151	18.279	3.81e-12 ***
DE	0.99058	0.52033	1.904	0.0751 .
DN	1.59599	0.75055	2.126	0.0494 *
Depth	0.09069	0.18572	0.488	0.6319
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’			1

Residual standard error: 2.144 on 16 degrees of freedom

Multiple R-squared: 0.7112, Adjusted R-squared: 0.657

F-statistic: 13.13 on 3 and 16 DF, p-value: 0.0001393

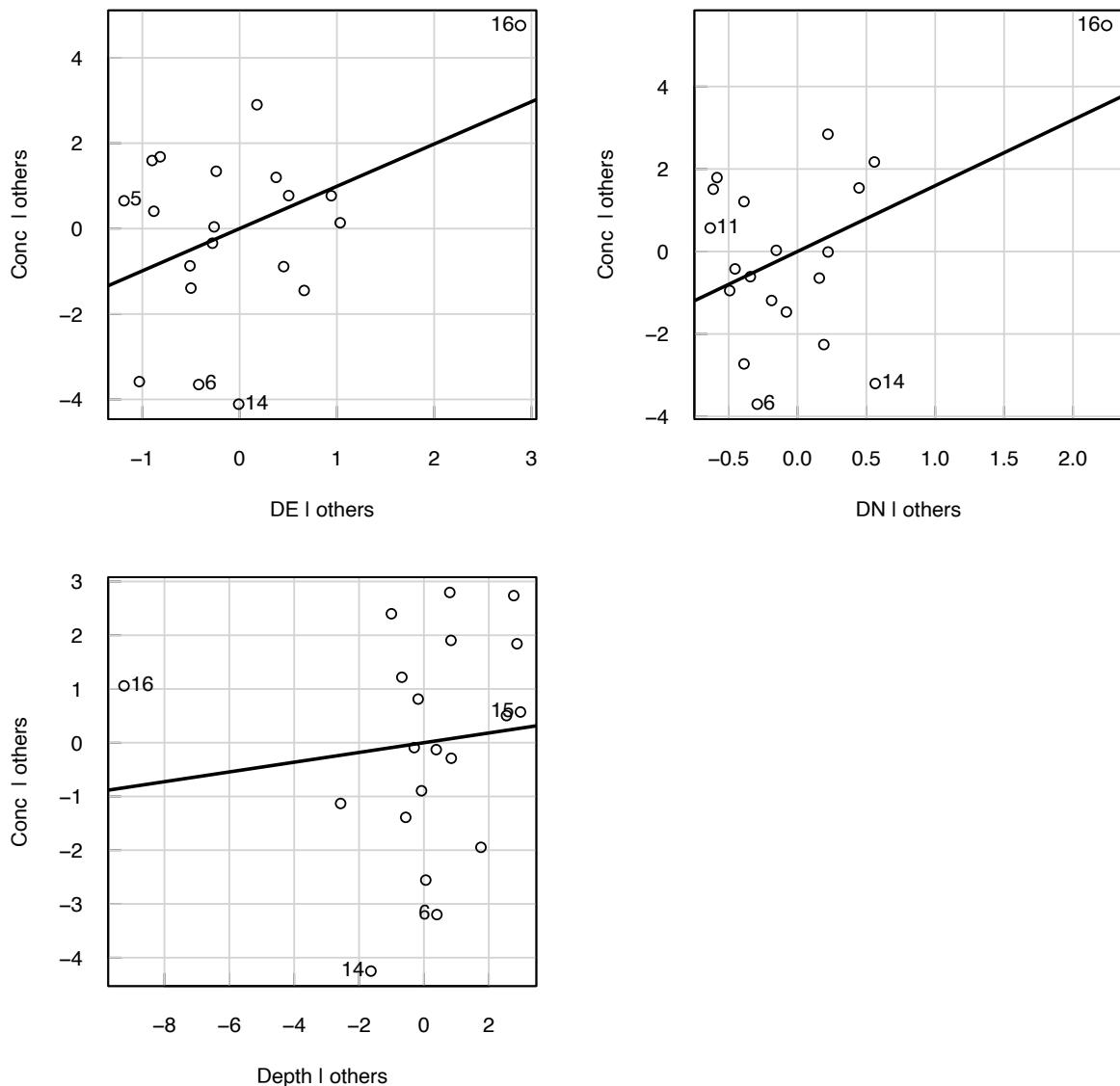
The above code and output result in the regression equation

$$\text{Conc} = 28.9 + 0.991\text{DE} + 1.60\text{DN} + 0.091\text{Depth} ,$$

where

$n$  = 20,  
 $s$  = 2.14, and  
 $R^2$  = 0.71.

**Figure 11.2 (facing page).** Rotated scatterplot showing the position of the high leverage point (observation 16, shown as a triangle). DE, distance east; DN, distance north; Depth, well depth.



**Figure 11.3.** Partial-regression plots for concentration (Conc) as a function of distance east (DE), distance north (DN), and well depth.

The outlier had a severe, detrimental effect on the regression coefficients and model structure and is visible in the partial-regression plots (fig. 11.3). 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. The R code and output below show that the resulting regression relation is drastically changed.

```
> Chap11Ex1.correct <- Chap11Ex1
> Chap11Ex1.correct[16,"Depth"] <- 23.111
> mod2 <- lm(Conc ~ DE + DN + D Depth, data = Chap11Ex1.correct)
> summary(mod2)
```

Call:

```
lm(formula = Conc ~ DE + DN + Depth, data = Chap11Ex1.correct)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.5166	-0.6483	0.0513	1.1086	2.8290

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.1680	1.3872	21.026	4.42e-13 ***
DE	-0.4186	0.8331	-0.502	0.6222
DN	-0.8157	1.3396	-0.609	0.5512
Depth	0.7103	0.3385	2.098	0.0521 .
---				
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 1.913 on 16 degrees of freedom

Multiple R-squared: 0.7701, Adjusted R-squared: 0.7271

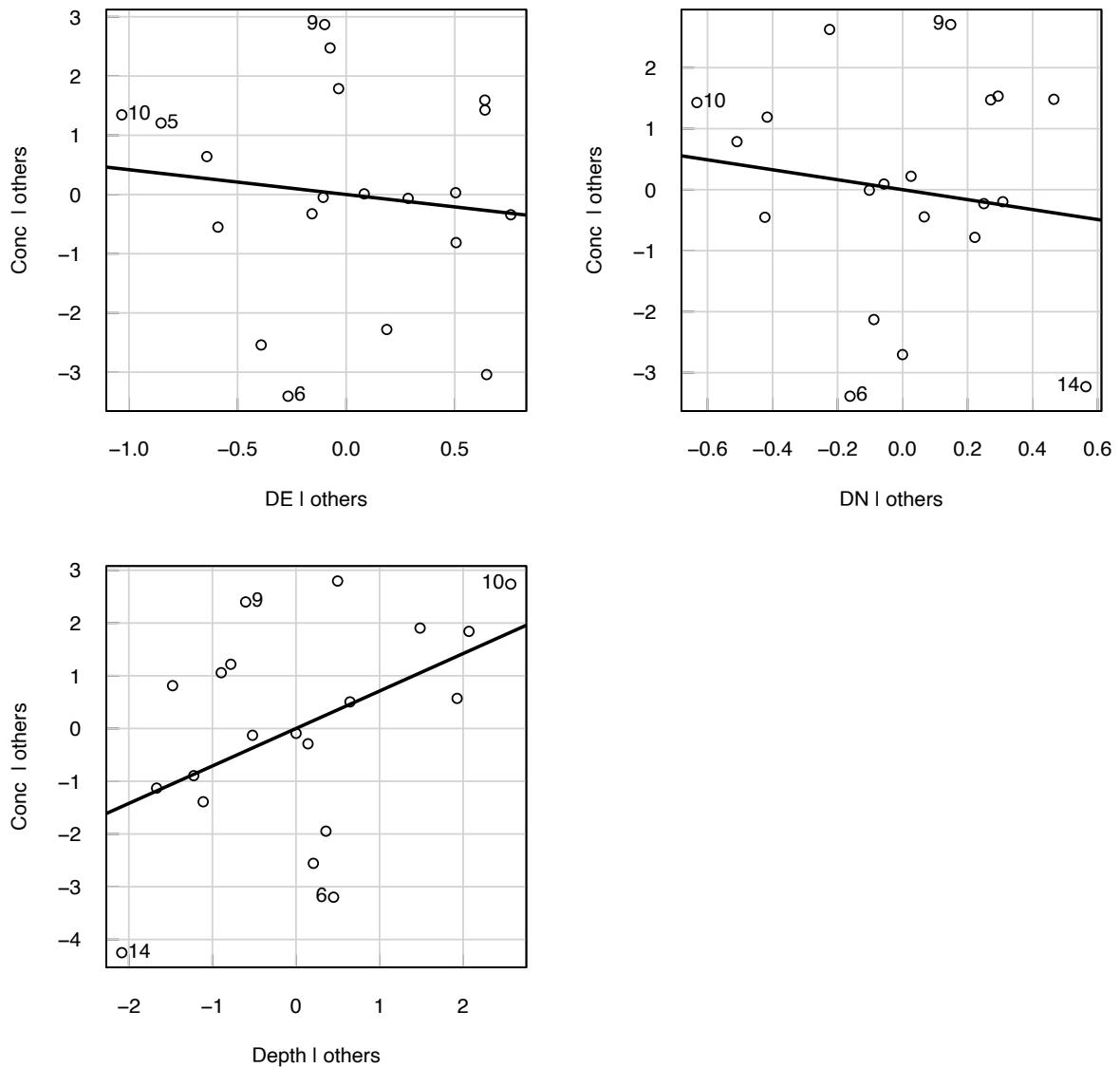
F-statistic: 17.87 on 3 and 16 DF, p-value: 2.32e-05

The above code and output, result in the regression equation

$$\text{Conc} = 29.2 - 0.419\text{DE} - 0.816\text{DN} + 0.710\text{Depth} ,$$

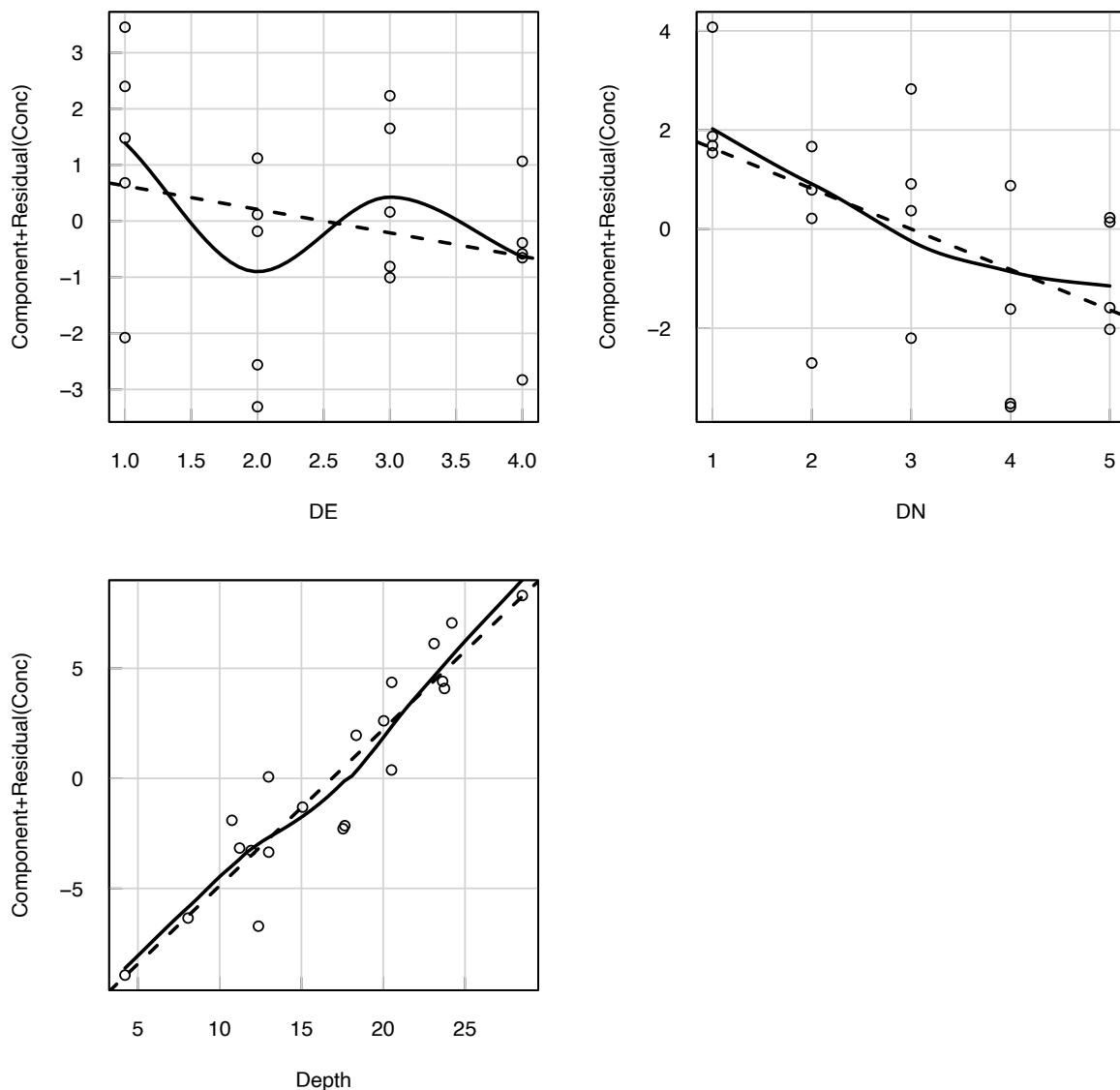
where

$n$  = 20,  
 $s$  = 1.91, and  
 $R^2$  = 0.77.



**Figure 11.4.** Partial-regression plots for concentration (Conc) as a function of distance east (DE), distance north (DN), and well depth (Depth), with outlier corrected.

Based on the *t*-statistics, DE and DN are not significantly related to Conc, but Depth is related. One can draw the same conclusion from the partial-residual plots (fig. 11.4) that show DE and DN having negative slopes, but little relation to the response variable when adjusted for the other explanatory variables. 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,  $h_i = 0.19$  and  $DFFITS = 0.48$ , both well below their critical values (see chap. 9). Thus, no observations have undue influence on the regression equation and no outlier stands out in the partial-regression plots (fig. 11.4).



**Figure 11.5.** Component + residual plots for concentration (Conc) as a function of distance east (DE), distance north (DN), and well depth (Depth), with outlier corrected.

The component + residual plots are shown in figure 11.5. The dashed line is the line of best fit (if the explanatory variables have a linear relation to the response). The solid line is a representation of where the component + residuals actually fall. Deviations from the dashed line indicate non-normalities; figure 15 indicates a nonlinear relation with DE.

Based on the numerical and graphical results, DE and DN do not appear to belong in the regression model, dropping them produces the model below, 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.

```
> mod3 <- lm(Conc ~ Depth, data=Chap11Ex1.correct)
> summary(mod3)
```

Call:

```
lm(formula = Conc ~ Depth, data = Chap11Ex1.correct)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.3882	-0.5395	0.0510	1.4331	2.6834

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.0363	1.1985	24.23	3.44e-15 ***
Depth	0.5110	0.0668	7.65	4.60e-07 ***
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’			1

Residual standard error: 1.825 on 18 degrees of freedom

Multiple R-squared: 0.7648, Adjusted R-squared: 0.7517

F-statistic: 58.53 on 1 and 18 DF, p-value: 4.604e-07

The above code and output result in the regression equation

$$\text{Conc} = 29.0 + 0.511\text{Depth} ,$$

where

$$\begin{aligned} n &= 20, \\ s &= 1.82, \text{ and} \\ R^2 &= 0.77. \end{aligned}$$

Note, the *F*-test for the overall model was significant in all three models, indicating that they were all better than no model (intercept only). This highlights the fact that the *F*-test is of little value when evaluating the quality of a model. The final, correct model had the lowest residual standard error, *s*. The final model also had the highest adjusted *R*<sup>2</sup> (a measure used when comparing models with differing numbers of explanatory variables, discussed later in this chapter).

### 11.5.3 Multicollinearity

Multicollinearity is the condition where at least one explanatory variable is closely related to one or more other explanatory variables. It is important that practitioners of MLR understand the causes and consequences of multicollinearity and can diagnose its presence, because 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 that are unrealistically large, and whose partial  $F$ - or  $t$ -tests are found to be insignificant.
2. Coefficients may be unrealistic in sign (for example, a negative slope for a regression of streamflow versus precipitation). Usually this occurs when two variables describing approximately the same thing counter-balance each other in the equation by 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.

Monte Carlo simulation has found that these undesirable consequences are magnified when sample sizes are smaller, correlations between variables are higher, and model error variances are higher (Kroll and Song, 2013). Concern over multicollinearity should be strongest when the purpose is to make inferences about coefficients, in which case advanced techniques could be used such as principal component analysis or partial least squares regression (Kroll and Song, 2013). Concern can be somewhat less when only predictions are of interest, if these predictions are for cases within the observed range of the  $x$  data.

A widely used diagnostic for measuring multicollinearity is the variance inflation factor ( $VIF$ ) presented by Marquardt (1970). For variable  $j$  the  $VIF$  is

$$VIF_j = \frac{1}{(1 - R_j^2)}, \quad (11.14)$$

where  $R_j^2$  is the  $R^2$  from a regression of the  $j$ th explanatory variable on all of the other explanatory variables—the equation used for adjustment of  $x_j$  in partial plots. The ideal is  $VIF_j \geq 1$ , corresponding to  $R_j^2 \geq 0$ , meaning that there is no correlation between the  $j$ th predictor and the other predictors, and the variance of the  $\beta_j$  is not inflated. Serious problems are indicated when  $VIF_j$  is large. There is no universal definition of a large  $VIF$ , but  $VIF_j$  is commonly considered large when it is greater than 10; however, some use a value as low as 4 (Kutner and others, 2004; O'Brien, 2007; Vatcheva and others, 2016). When  $VIF$  is greater than 10,  $R_j^2$  is greater than 0.9, indicating a high degree of correlation between variables ( $r_j > 0.95$ ).

A useful interpretation of  $VIF$  is that multicollinearity inflates the width of the confidence interval for the  $j$ th regression coefficient by the amount  $\sqrt{VIF_j}$  compared to what it would be with a perfectly independent set of explanatory variables. When using the  $VIF$ , it is important to know its limitations, including that the  $VIF$  is sensitive to sample size (O'Brien, 2007) and it cannot distinguish between multiple simultaneous multicollinearities (Neter and others, 1996).

If the analyst determines that multicollinearity is a problem for a candidate regression model, there are four options for reducing multicollinearity.

1. Center the data. A simple solution that works in some specific cases is to center the data. Multicollinearity 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 relation 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 that produces about as many positive values as negative values. The constant could be the mean or median, or it could be a round number roughly in the middle of the data (for example, the year 2000 for a dataset that runs from 1990 to 2014). When all of the derived explanatory variables are recomputed as functions (squares, products, and so forth) of these centered variables, their multicollinearity will be reduced.

Centering is a mathematical solution to a mathematical problem, it will not reduce multicollinearity between two variables that 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 multicollinearity 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 (fitting a high order polynomial of distances north of the equator and west of Greenwich, as an example, or trend analysis values are a polynomial of years). This will be demonstrated in example 11.2.

When an explanatory variable is centered, the slope coefficient of the explanatory variable does not change, but the interpretation of the model intercept does change. In the SLR case, the intercept,  $\beta_0$ , is the value of the response variable, when the predictor variable equals 0. If the predictor variable is centered,  $\beta_0$  is the value of the response variable when the original predictor variable minus the value used to center it equals 0. When using a regression equation for prediction, one must take care to use the centered values in prediction, not the original values.

2. Eliminate variables. In some cases, prior judgment suggests the use of several different variables that describe related, but not identical, attributes. Examples of this might be air temperature and dew point temperature, the maximum 1-hour rainfall, 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 be eliminated on the basis of the analyst's knowledge of the system being studied or on the basis of comparisons of model fits with one eliminated variable versus the other eliminated variable, in order to lower the *VIF*.
3. Collect additional data. Multicollinearity problems can sometimes be solved with only a few additional but strategically selected observations. Suppose some attributes of river basins are being studied, where small basins tend to be heavily forested and 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 multicollinearity will result from including both variables in the regression equation. To solve this and allow the effects of each variable to be judged separately, the sampling design should include, if possible, the collection of additional samples from a few small but less forested basins and a few large but heavily forested basins. This should produce a much more reliable model. Similar problems arise in groundwater quality studies, where rural wells are shallow and urban wells are deep. Depth and population density may show strong multicollinearity, requiring some shallow urban and deep rural wells to be sampled.
4. Perform ridge regression (beyond the scope of this text). Ridge regression was proposed by Hoerl and Kennard (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 MLR.

### Example 11.2—Centering.

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 (simulated data). The city center was arbitrarily chosen as a geographic datum. The data are presented in table 11.2.

The following code and output show an MLR model for the log of concentration explained by DE, DN,  $DE^2$ ,  $DN^2$ , and an interaction term  $DE \cdot DN$ .

```
> m1 <- lm(lnConc ~ DE + DN + DESQ + DNSQ + DEDN, data = Ex2)
> summary(m1)
```

Call:

```
lm(formula = lnConc ~ DE + DN + DESQ + DNSQ + DEDN, data = Ex2)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.33468	-0.16896	-0.01947	0.15423	0.61474

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.790e+02	9.166e+01	-5.226	0.000128 ***
DE	1.055e+01	1.121e+00	9.405	1.99e-07 ***
DN	1.514e+01	3.602e+00	4.202	0.000887 ***
DESQ	-2.642e-01	1.498e-02	-17.635	5.88e-11 ***
DNSQ	-1.514e-01	3.581e-02	-4.229	0.000842 ***
DEDN	1.383e-03	1.895e-02	0.073	0.942864
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Residual standard error: 0.268 on 14 degrees of freedom

Multiple R-squared: 0.9596, Adjusted R-squared: 0.9451

F-statistic: 66.46 on 5 and 14 DF, p-value: 2.978e-09

The regression equation is:

$$\ln(\text{Conc}) = -479 + 10.5\text{DE} + 15.1\text{DN} - 0.262\text{DESQ} - 0.151\text{DNSQ} + 0.0014\text{DEDN} .$$

**Table 11.2.** Data for example 11.2.

[Obs. #, observation number, Conc, concentration, DE, distance east; DN, distance north; DESQ, distance east squared, DNSQ, distance north squared]

Obs. #	Conc	ln(Conc)	DE	DN	DESQ	DNSQ	DE•DN
1	14	2.63906	17	48	289	2,304	816
2	88	4.47734	19	48	361	2,304	912
3	249	5.51745	21	48	441	2,304	1,008
4	14	2.63906	23	48	529	2,304	1,104
5	29	3.36730	17	49	289	2,401	833
6	147	4.99043	19	49	361	2,401	931
7	195	5.27300	21	49	441	2,401	1,029
8	28	3.33220	23	49	529	2,401	1,127
9	21	3.04452	17	50	289	2,500	850
10	276	5.62040	19	50	361	2,500	950
11	219	5.38907	21	50	441	2,500	1,050
12	40	3.68888	23	50	529	2,500	1,150
13	22	3.09104	17	51	289	2,601	867
14	234	5.45532	19	51	361	2,601	969
15	203	5.31320	21	51	441	2,601	1,071
16	35	3.55535	23	51	529	2,601	1,173
17	15	2.70805	17	52	289	2,704	884
18	115	4.74493	19	52	361	2,704	988
19	180	5.19296	21	52	441	2,704	1,092
20	16	2.77259	23	52	529	2,704	1,196

Since the square of distance east (DESQ) must be strongly related to DE, and similarly DNSQ must be strongly related to DN, and  $DE \cdot DN$  with both DE and DN, multicollinearity between these variables will be detected by their *VIFs*. One may write a function as outlined above to calculate *VIFs* in the software they are using. SAS has collinearity diagnostics as part of its regression procedure, PROC REG (SAS Institute Inc., 2014). A number of R packages also have a function that calculates the *VIFs*, including `car` (Fox and Weisberg, 2011), `fmsb` (Nakazawa, 2017), and `usdm` (Naimi and others, 2014). Using the `car` package in R, the *VIFs* for the above coefficients follow.

```
> library(car)
> vif(m1)

      DE        DN      DESQ      DNSQ      DEDN
1751.000 7223.857 501.000 7143.857 1331.000
```

Using the rule that any *VIF* above 10 indicates a strong dependence between variables, the result of the *VIF* analysis above show that all variables have high *VIFs*. Therefore, all of the slope coefficients are unstable, and no conclusions can be drawn from the slope coefficients. This cannot be considered a good regression model, even though the  $R^2$  is large.

To solve the problem, DE and DN are centered by subtracting their medians. Following this, the three derived variables DESQ, DNSQ, and  $DE \cdot DN$  are recomputed, and the regression rerun. The following results show that all multicollinearity is completely removed. The coefficients for DE and DN are now more reasonable in size, and the coefficients for the derived variables are exactly the same as in the original regression. 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 unchanged. In fact, this is exactly the same model as the uncentered equation, but only in a different and centered coordinate system.

```
> cent <- Ex2
> cent$DE <- cent$DE - median(cent$DE)
> cent$DN <- cent$DN - median(cent$DN)
> cent$DESQ <- cent$DE^2
> cent$DNSQ <- cent$DN^2
> cent$DEDN <- cent$DE * cent$DN
> m2 <- lm(lnConc ~ DE + DN + DESQ + DNSQ + DEDN,
+           data = cent)
> summary(m2)

Call:
lm(formula = lnConc ~ DE + DN + DESQ + DNSQ + DEDN, data = cent)
```

#### Residuals:

Min	1Q	Median	3Q	Max
-0.33468	-0.16896	-0.01947	0.15423	0.61474

#### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.764495	0.119721	48.149	< 2e-16 ***
DE	0.048116	0.026800	1.795	0.094208 .

```

DN          0.018581  0.042375  0.438  0.667726
DESQ        -0.264201  0.014982 -17.635 5.88e-11 ***
DNSQ        -0.151442  0.035813 -4.229  0.000842 ***
DEDN        0.001383  0.018951  0.073  0.942864
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

Residual standard error: 0.268 on 14 degrees of freedom

Multiple R-squared: 0.9596, Adjusted R-squared: 0.9451

F-statistic: 66.46 on 5 and 14 DF, p-value: 2.978e-09

Resulting in the equation

$$\ln(\text{Conc}) = 5.76 + 0.048(\text{DE} - \text{median}(\text{DE})) + 0.019(\text{DN} - \text{median}(\text{DN})) - 0.264\text{DESQ} - 0.151\text{DNSQ} + 0.001(\text{DE} - \text{median}(\text{DE})) \cdot (\text{DN} - \text{median}(\text{DN}))$$

Multicollinearity is now no longer a concern.

```

> vif(m2)
DE    DN DESQ DNSQ DEDN
1     1    1    1    1

```

What we see from this result is that there are three explanatory variables we might now want to consider removing from the model (because they have *t*-statistics that are relatively low). These are DE, DN, and DE · DN. The next section describes approaches for this kind of problem, selecting the best model from among a number of possible models.

## 11.6 Choosing the Best Multiple Linear Regression Model

One of the major issues in MLR is finding the appropriate approach to variable selection. The benefit of adding additional variables to an MLR 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,  $n-k-1$ , decreases (the number of independent pieces of information in the sample), 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 that 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 (error sum of squares, defined in chap. 9) 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, because there is a loss of statistical power as more variables are added. 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, all subsets regression, uses one or more overall measure of model quality.

### 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 forward, backward, 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 multicollinearity is present; they also do not evaluate all possible models, and so cannot guarantee that the best model is even tested. The procedures were developed before 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 that are not significantly different from zero; they must only be significant when they enter. A second drawback is that two variables that each individually provide little explanation of  $y$  may never enter, but together the variables might 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.

These stepwise procedures are influenced by the sequence in which variables are added and removed and can move in a direction of suboptimal model space. The procedures can also result in an overfit model, that is one that is excessively complex and mathematically describes random error rather than the underlying hydrologic process. Overfit models may not perform well to another sample from the same population.

#### Example 11.3. Regression for mean annual runoff.

Haan (1977) attempted to relate the mean annual runoff of several streams ( $ROFF$ ) with nine other variables: the precipitation falling at the gage ( $PRECIP$ ), 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 ( $Rr$ ). Haan chose to select a three-variable model (using  $PRECIP$ ,  $PERIM$ , and  $Rr$ ) based on a leveling off of the incremental increase in  $R^2$  as more variables were added to the equation.

What models would be selected if the stepwise or overall methods were applied to this data? If a forward 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  $\alpha$ ,  $AREA$  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 in table 11.3).

The backward model begins with all variables in the model. It checks all partial  $t$ - or  $F$ -statistics, throwing away the variable that is least significant. Here the least significant single variable is  $AREA$ . So whereas the forward model made  $AREA$  the first variable to bring in, the backward model discarded  $AREA$  first. Then other variables were also removed, resulting in a model with  $Rr$ ,  $PCIP$ ,  $PERIM$ ,  $DI$ , and  $FREQ$  remaining in the model. Multicollinearity between measures of drainage basin size, as well as between other variables, has produced models from backward and forward procedures that 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.

The stepwise model first enters  $AREA$ ,  $Rr$ ,  $PRECIP$ , 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 five-variable model as did the backward model (table 11.4).

**Table 11.3.** Results of forward selection procedure for example 11.3.

[*AREA*, drainage area of the basin; *Rr*, relief ratio for the basin; *PRECIP*, precipitation falling at the gage; *PERIM*, perimeter of the basin; -, not applicable]

Forward	Coefficient or t-statistic	Model 1	Model 2	Model 3	Model 4
<i>AREA</i>	$\beta$	0.43	0.81	0.83	-0.62
	<i>t</i>	1.77	4.36	4.97	-1.68
<i>Rr</i>	$\beta$	-	0.013	0.011	0.009
	<i>t</i>	-	3.95	3.49	4.89
<i>PRECIP</i>	$\beta$	-	-	0.26	0.54
	<i>t</i>	-	-	1.91	5.05
<i>PERIM</i>	$\beta$	-	-	-	1.02
	<i>t</i>	-	-	-	4.09

**Table 11.4.** Results of stepwise selection procedure for example 11.3.

[*AREA*, drainage area of the basin; *Rr*, relief ratio for the basin; *PRECIP*, precipitation falling at the gage; *PERIM*, perimeter of the basin; *DI*, diameter of the largest circle which could be inscribed within the drainage basin; *FREQ*, stream frequency—the ratio of the number of streams in the basin to the basin area; -, not applicable]

Stepwise	Coefficient or t-statistic	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
<i>AREA</i>	$\beta$	0.43	0.81	0.83	-0.62	-	-	-
	<i>t</i>	1.77	4.36	4.97	-1.68	-	-	-
<i>Rr</i>	$\beta$	-	0.013	0.011	0.009	0.010	0.010	0.011
	<i>t</i>	-	3.95	3.49	4.89	5.19	5.02	6.40
<i>PRECIP</i>	$\beta$	-	-	0.26	0.54	0.430	0.495	0.516
	<i>t</i>	-	-	1.91	5.05	4.62	5.39	6.71
<i>PERIM</i>	$\beta$	-	-	-	1.02	0.617	0.770	0.878
	<i>t</i>	-	-	-	4.09	8.24	6.98	8.38
<i>DI</i>	$\beta$	-	-	-	-	-	-1.18	-1.30
	<i>t</i>	-	-	-	-	-	-1.75	-2.32
<i>FREQ</i>	$\beta$	-	-	-	-	-	-	0.36
	<i>t</i>	-	-	-	-	-	-	2.14

## 11.6.2 Overall Measures of Quality

Four statistics can be used to evaluate each of the  $2^k$  regression equations possible from  $k$  candidate explanatory variables. These are the adjusted  $R^2$  ( $R_a^2$ ); Mallow's  $C_p$  ( $C_p$ ); Schwartz's information criterion (BIC); and the prediction error sum of squares (PRESS<sub>p</sub>) statistic, usually shortened to PRESS.

Adjusted  $R^2$  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  $R_a^2$  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:

$$R^2 = 1 - \frac{SSE}{SS_y} . \quad (11.15)$$

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. Adjusted  $R^2$  is adjusted to offset the loss in degrees of freedom by including as a weight the ratio of total degrees of freedom to error degrees of freedom:

$$R_a^2 = 1 - \frac{(n-1)SSE}{(n-p)SS_y} = 1 - \frac{MSE}{SS_y / (n-1)} . \quad (11.16)$$

As  $SS_y / (n-1)$  is constant for a given dataset,  $R_a^2$  increases as  $MSE$  decreases; therefore, an analyst can either maximize  $R_a^2$  or minimize  $MSE$  as an overall measure of quality. However, when  $p$ , the number of coefficients, is considerably smaller than  $n$ ,  $R_a^2$  is a less sensitive measure than either PRESS or  $C_p$ . PRESS has the additional advantage of being a validation criterion.

The statistic  $C_p$  (Mallows, 1973) 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  $C_p$  statistic is

$$C_p = p + \frac{(n-p)(MSE_p - \hat{\sigma}^2)}{\hat{\sigma}^2} , \quad (11.17)$$

where  $n$  is the number of observations,  $p$  is the number of coefficients (number of explanatory variables plus 1),  $MSE_p$  is the mean square error of the  $p$ -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  $C_p$  value. When several models have nearly equal  $C_p$  values, they may be compared in terms of reasonableness, multicollinearity, importance of high influence points, and cost in order to select the model with the best overall properties.

The statistic BIC, sometimes known as Bayesian information criterion or Schwarz criterion (Schwarz, 1978), takes into account goodness of fit of the model and applies a penalty for increasing the number of parameters in a model. It approximates Bayes factor, which is a “summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another” (Kass and Raftery, 1995). In model selection using BIC, the goal is to minimize the BIC, while making a tradeoff between model fit and number of parameters in the model. For large  $n$ , the BIC criterion is

$$BIC = -2 \cdot \ln(\hat{L}) + \ln(n) \cdot p , \quad (11.18)$$

where  $\hat{L}$  is the maximized value of the likelihood function (the likelihood of a set of parameter values of a statistical model, given the observed data),  $n$  is the number of observations, and  $p$  is the number of coefficients in the fitted model (R Core Team, 2016). It can also be expressed as

$$BIC = n + n \cdot \ln(2\pi) + n \cdot \ln\left(\frac{SSE}{n}\right) + \ln(n)(p+1) . \quad (11.19)$$

This is very similar to another criterion, Akaike's information criterion (Akaike, 1974), AIC, which is calculated in the same manner as BIC by replacing  $\ln(n)$  with 2 (R Core Team, 2016). When  $n \geq 8$ , the BIC penalty for each additional parameter is larger than it is for AIC. BIC can be useful when an analyst has many potential explanatory variables, but wants a parsimonious model.

The statistic PRESS was defined in chapter 9 as a validation-type estimator of error that uses the deleted residuals to provide an estimate of the prediction error (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.

The  $R_a^2$  and BIC can disagree as to which is the best model when comparing models of different sizes because of the larger penalty BIC applies to additional coefficients. PRESS and  $C_p$  generally agree as to which model is best, even though their criteria for selection are not identical. Among the several metrics introduced here ( $R_a^2$ , BIC,  $C_p$ , and PRESS) there is no universally accepted best metric. They are all reasonable approaches to model selection and all of them are better metrics than simply using  $R^2$ .

### 11.6.3 All-subsets Regression

The all-subsets regression method, an exhaustive search method, uses the computer to perform a large number of computations, letting the scientist judge which model to use. This allows flexibility in choosing between models—some may make more sense from an economic or scientific perspective, rather than from a purely statistical perspective. For example, two best models may be nearly identical in terms of their  $R_a^2$ ,  $C_p$ , BIC, 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. Likewise, several best models may have similar measures of model quality, but one better matches the science of the process being modeled. 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 and they may even fail to test the best model.

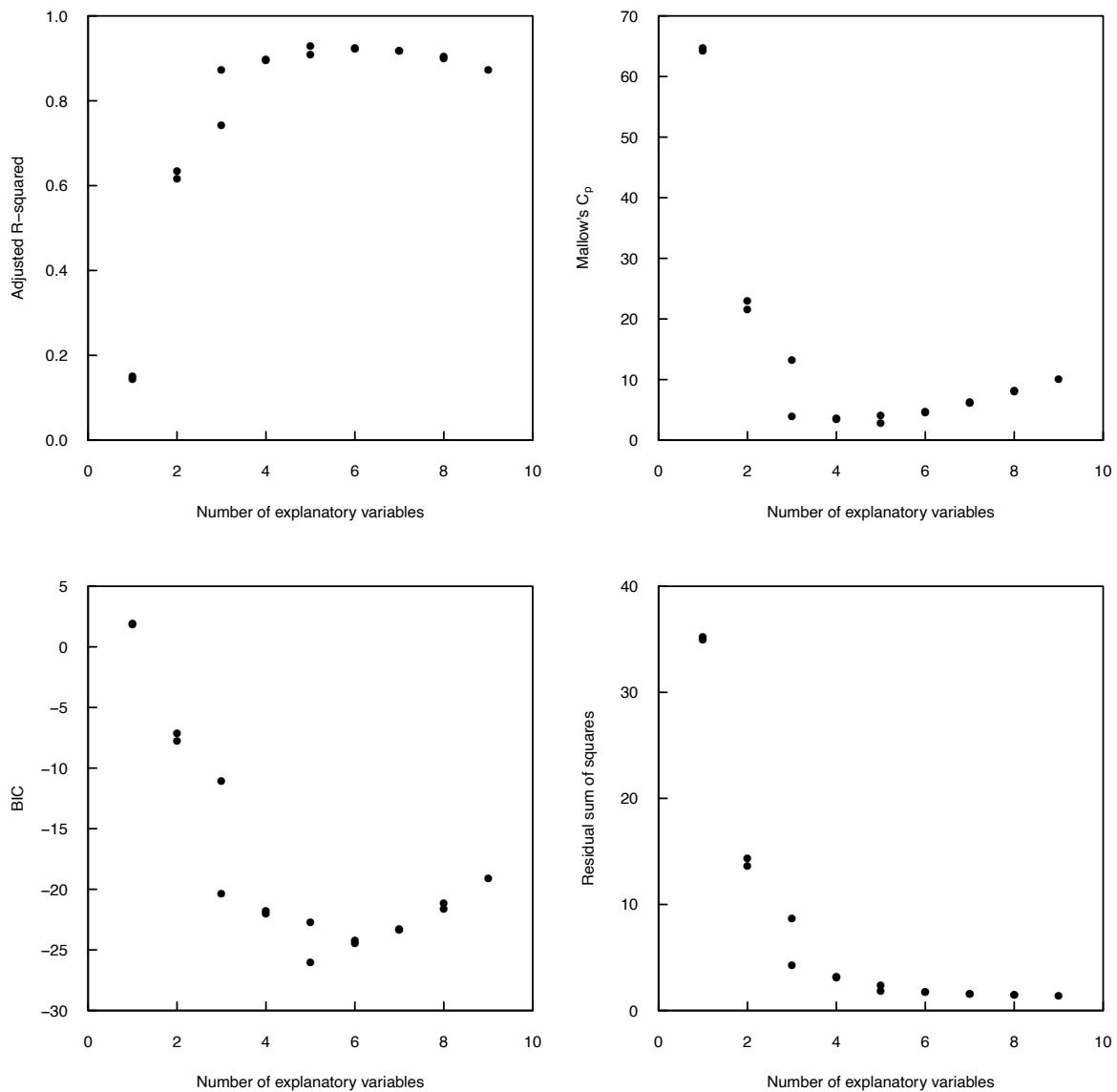
### **Example 11.3. Regression for mean annual runoff—Continued.**

Instead of the stepwise procedures run on Haan's data, models are evaluated using the overall statistics  $C_p$  and PRESS. Smaller values of  $C_p$  and PRESS are associated with better models. Computing PRESS and  $C_p$  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 11.5.

Based on  $C_p$ , the best model would be the five variable model having *PCIP*, *PERIM*, *DI*, *FREQ*, and *Rr* as explanatory variables ( $C_p = 2.9$ ). It is the same model as selected by the stepwise and forward methods. Remember that there is no guarantee that stepwise procedures regularly select the lowest  $C_p$  or PRESS models. The advantage of using an overall statistic like  $C_p$  is that options are given to the scientist to select what is best. If the modest multicollinearity ( $VIF = 5.1$ ) between *PERIM* and *DI* is of concern (with its resultant negative slope for *DI*) the model with the next lowest  $C_p$  that does not contain both these variables, a four-variable model with  $C_p = 3.6$  could be selected. If the scientist decided *AREA* must be in the model, the lowest  $C_p$  model containing *AREA* (the same four-variable model) could be selected.

**Table 11.5.** Statistics for several multiple regression models of Haan's (1977) data.

[PRESS, prediction error sum of squares; *VIF*, variance inflation factor; *PRECIP*, precipitation; *AREA*, the drainage area of the basin; *SLOPE*, the average slope of the basin; *LEN*, the length of the drainage basin; *PERIM*, the perimeter of the basin; *DI*, the diameter of the largest circle which could be inscribed within the drainage basin; *Rs*, the shape factor of the basin; *FREQ*, the stream frequency—the ratio of the number of streams in the basin to the basin area; *Rr*, the relief ratio for the basin; -, not applicable]



**Figure 11.6.** Plots of the magnitude of adjusted R-squared, Mallow's  $C_p$ , BIC (Bayesian information criterion), and residual sum of squares for the two best explanatory variable models as a function of the number of explanatory variables.

Using the `regsubsets` function of the `leaps` package (Lumley, 2017) in R, selecting the best subsets from all-subsets regression can be automated. The following code performs an exhaustive all-subsets regression and returns the two best models of each size from one to nine explanatory variables (returning two models of size one to eight and the model of size nine is specified by the user and the user could specify a different best subset). The best models of a particular size are the best based on  $R_a^2$ ,  $C_p$ , or BIC, as the differences among these measures of model quality occur when comparing models of different sizes. The resulting indicators of which variables are included with each potential model is merged with  $R_a^2$ ,  $C_p$ , BIC, and the residual sum of squares for each model. The results of selecting the two best models of each size from one to nine explanatory variables are graphically displayed in figure 11.6, the plots show that the two best models of each size often have similar measures of model quality, but not always. The plots also show that a model with five explanatory variables is likely best.

The user can then examine all of this information and select the preferred model based on the user's definition of best. The user may also want to select a subset of the models and calculate PRESS for them as the `regsubsets` function does not calculate this function.

```

> load("Chapter11.RData")
> head(AC14)
> library(leaps)
> ex3.all <- regsubsets(AC14[, -1], AC14[, 1], method = "exhaustive", nmax = 9, nbest = 2)
> ex3.step.res <- as.data.frame(cbind(as.data.frame(summary(ex3.all)$which, row.names = c(1:17)),
+ summary(ex3.all)$adjr2, summary(ex3.all)$cp, summary(ex3.all)$bic, summary(ex3.all)$rss))
> dimnames(ex3.step.res)[[2]][11:14] <- c("adjr2", "cp", "bic", "rss")
> ex3.step.res

      (Intercept) PRECIP AREA SLOPE LEN PERIM DI RS FREQ Rr adjr2 cp bic rss
1      TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE 0.15009488 64.183501 1.871486 34.977527
2      TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE 0.1446269 64.728412 1.967923 35.237963
3      TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE 0.63351892 21.586058 -7.784226 13.662500
4      TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE 0.6166390 23.039626 -7.139449 14.357223
5      TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE 0.87355822 3.915331 -20.366184 4.261018
6      TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE 0.7420792 13.189278 -11.096405 8.693434
7      TRUE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE 0.8972964 3.438170 -22.033100 3.077077
8      TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE 0.8949710 3.583937 -21.742049 3.146745
9      TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE FALSE TRUE 0.9289799 2.895526 -25.999553 1.861839
10     TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE 0.9085162 4.017979 -22.707973 2.398307
11     TRUE TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE 0.9235907 4.592394 -24.487732 1.716959
12     TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE FALSE TRUE 0.9221146 4.661795 -24.238984 1.750128
13     TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE 0.9179914 6.213039 -23.373604 1.535649
14     TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE 0.9172172 6.243371 -23.251455 1.550146
15     TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE 0.9037325 8.017354 -21.625535 1.442123
16     TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE 0.8999080 8.137226 -21.119070 1.499415
17     TRUE 0.8723816 10.000000 -19.135570 1.433828

```

## 11.7 Summary of Model Selection Criteria

Rules for selection of linear regression models are summarized in the five 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:
  - A. constant variance across the range of  $\hat{y}$ ,
  - B. normality of residuals, and
  - C. a linear pattern, not curvature.
- The statistics  $R^2$ ,  $R_a^2$ , SSE,  $C_p$ , BIC, and PRESS are not appropriate for comparison of models having different transformations of  $y$ .
2. Should  $x$  (or several  $x$ 's) be transformed? Transformation of an  $x$  variable should be made using component + residual plots. Check for the same three patterns of constant variance, normality, and linearity. Considerable help can be obtained by optimizing statistics such as  $R_a^2$  (maximize), BIC (minimize), SSE (minimize),  $C_p$  (minimize), or PRESS (minimize). Many transformations can be rapidly checked with such statistics, but a residual plot should always be inspected before 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  $R_a^2$ , SSE,  $C_p$ , 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 the best one. The analyst may also select the model based on minimum  $C_p$  or minimum PRESS.
5. Which of several models is preferable when each uses the same  $y$  variable but is not necessarily nested?  $C_p$  or PRESS should be used in this situation.

## 11.8 Analysis of Covariance

Often there are factors that 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 and night; winter and summer; before and 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 an MLR context. They can be incorporated by the use of binary variables, also called indicator or dummy variables. This method is essentially a blending of regression and analysis of variance into an analysis of covariance, or ANCOVA.

### 11.8.1 Use of One Binary Variable

Starting with the simple one-variable regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon , \quad (11.20)$$

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

$$Z_i = \begin{cases} 0 & \text{if } i \text{ is from cold season} \\ 1 & \text{if } i \text{ is from warm season} \end{cases} \quad (11.21)$$

to produce the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \quad (11.22)$$

When the slope coefficient,  $\beta_2$ , is significant, the model in equation 11.22 would be preferred to the SLR model equation 11.20. A significant result also says that the relation between  $Y$  and  $X$  is affected by season.

Consider the hypothesis test where  $H_0: \beta_2 = 0$  and  $H_A: \beta_2 \neq 0$ . 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  $\beta$ 's being estimated. If the partial  $|t| \geq t_{1-\alpha/2}$ ,  $H_0$  is rejected. Thus, we should infer that there are two models:

$$\hat{Y} = b_0 + b_1 X \text{ for the cold season } (Z=0), \text{ and}$$

$$\hat{Y} = b_0 + b_1 X + b_2 \text{ for the warm season } (Z=1), \text{ or}$$

$$\hat{Y} = (b_0 + b_2) + b_1 X.$$

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 (fig. 11.7).

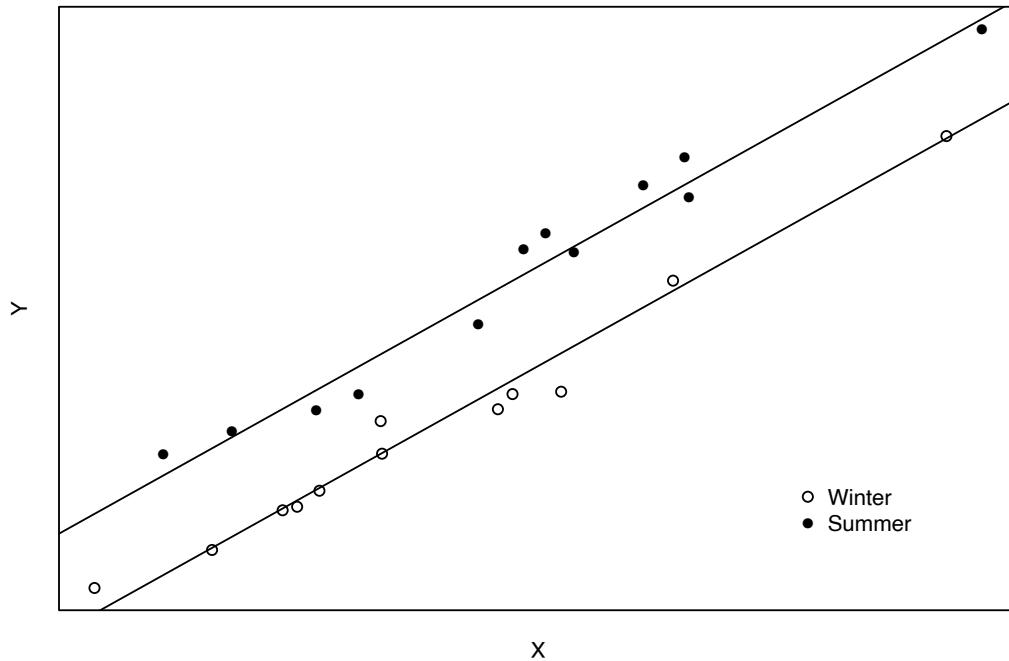
The binary variables always have a value of 0 or 1 because their coefficient is used to change the intercept or slope for some of the values by a constant. In the above example, the warm season values have an intercept that is increased by  $b_2$  when  $Z=1$ . An example of using binary variables to change both the intercept and slope for some values follows.

Suppose that the relation 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

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX + \varepsilon, \quad (11.23)$$

which is equivalent to

$$Y = (\beta_0 + \beta_2 Z) + (\beta_1 + \beta_3 Z) \cdot X + \varepsilon. \quad (11.24)$$



**Figure 11.7.** Plot of regression lines for data differing in intercept between two seasons.

The intercept equals  $\beta_0$  for the cold season and  $\beta_0 + \beta_2$  for the warm season; the slope equals  $\beta_1$  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  $ZX$ , 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 the model in equation 11.23, the following hypotheses are tested:

$$H_0: \beta_2 = 0 \text{ versus } H_A: \beta_2 \text{ and/or } \beta_3 \neq 0.$$

A nested  $F$ -statistic is computed

$$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{SSE_c / df_c}, \quad (11.25)$$

where  $s$  refers to the simpler (no  $Z$  terms) model, and  $c$  refers to the more complex model. For the two nested models shown in equations 11.20 and 11.23 this becomes

$$F = \frac{(SSE_{20} - SSE_{23}) / 2}{MSE_{23}},$$

where  $MSE_{23} = SSE_{23} / (n - 4)$ , rejecting  $H_0$  if  $F > F_{\alpha/2, n-4}$ .

If  $H_0$  is rejected, the model in equation 11.23 should also be compared to the model in equation 11.22 (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 11.20 in favor of 11.23 was only the result of a shift in intercept. The null hypothesis  $H'_0: \beta_3 = 0$  is compared to  $H'_A: \beta_3 \neq 0$  using the test statistic

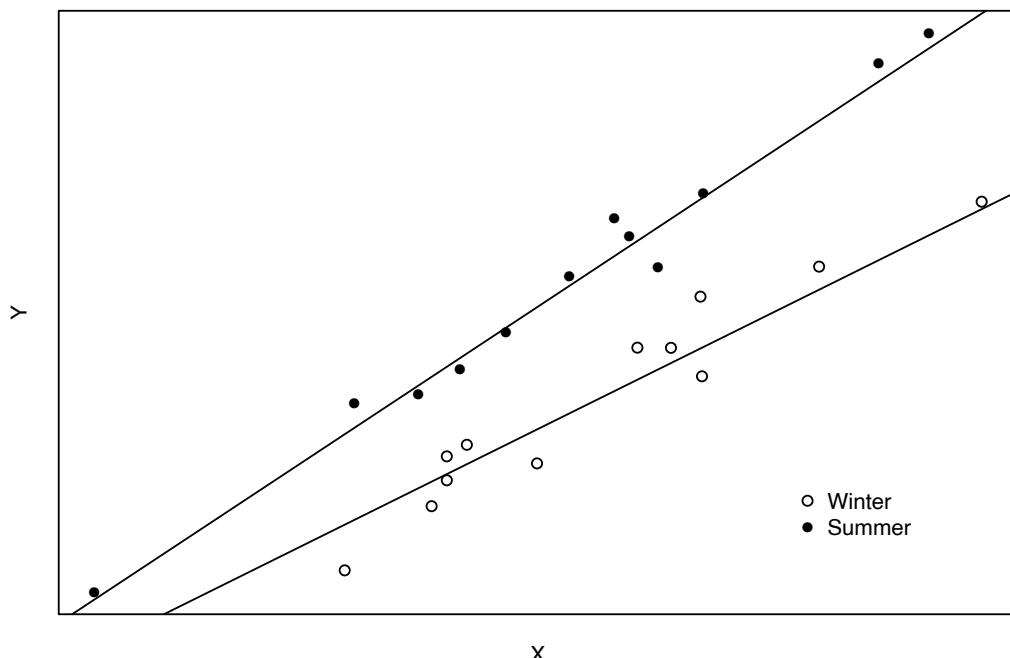
$$F = \frac{(SSE_{22} - SSE_{23}) / 1}{MSE_{23}}$$

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 fig. 11.8):

$$\hat{Y} = b_0 + b_1 X \quad \text{cold season}$$

$$\hat{Y} = (b_0 + b_2) + (b_1 + b_3) X \quad \text{warm season}$$



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.

### 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), and so forth. To illustrate, assume there are precise definitions of three 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.

$$\text{Let } R_i = \begin{cases} 1 & \text{if } i \text{ is a rising limb observation} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Let } D_i = \begin{cases} 1 & \text{if } i \text{ is a falling limb observation} \\ 0 & \text{otherwise} \end{cases}$$

So that

Category	Value of R	Value of D
Rising	1	0
Falling	0	1
Base flow	0	0

The following model results

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \varepsilon . \quad (11.26)$$

To test  $H_0: \beta_2 = \beta_3 = 0$  versus  $H_A: \beta_2 \neq 0$  and (or)  $\beta_3 \neq 0$ , F-tests comparing simpler and more complex models are again performed. To compare the model in equation 11.26 versus the SLR model in equation 11.20 with no rising or falling terms,

$$F = \frac{(SSE_{20} - SSE_{26}) / 2}{MSE_{26}} ,$$

where  $MSE_{26} = SSE_{26} / (n - 4)$ , rejecting  $H_0$  if  $F > F_{\alpha/2, n-4}$ .

To test for differences between each pair of categories consider the following questions.

1. Is the rising limb different from base flow? This is tested using the t-statistic on the coefficient  $\beta_2$ . If  $|t| > t_{\alpha/2}$  with  $n - 4$  degrees of freedom, reject  $H_0$  where  $H_0: \beta_2 = 0$ .
2. Is the falling limb different from base flow? This is tested using the t-statistic on the coefficient  $\beta_3$ . If  $|t| > t_{1-\alpha/2}$  with  $n - 4$  degrees of freedom, reject  $H_0$  where  $H_0: \beta_3 = 0$ .

**Figure 11.8 (facing page).** Plot of regression lines differing in slope and intercept for data from two seasons.

3. Is the rising limb different from the falling limb? 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  $b_2 - b_3$ ,  $\widehat{Var}(b_2 - b_3) = \widehat{Var}(b_2) + \widehat{Var}(b_3) - 2\widehat{Cov}(b_2, b_3)$ , where  $\widehat{Cov}$  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

$$\widehat{Var}(b_2) = C_{22}s^2 ,$$

$$\widehat{Var}(b_3) = C_{33}s^2 , \text{ and}$$

$$\widehat{Cov}(b_2, b_3) = C_{23}s^2 .$$

The test statistic is  $t = (b_2 - b_3) / \sqrt{\widehat{Var}(b_2 - b_3)}$ . If  $|t| > t_{1-\alpha/2}$  with  $n-4$  degrees of freedom, reject  $H_0$ .

- B. The binary variables can be redefined so that a direct contrast between rising and falling limbs 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

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \beta_4 RX + \beta_5 DX + \varepsilon . \quad (11.27)$$

The procedures for selecting models follow the pattern described above. The significance of an individual  $\beta$  coefficient, given all the other  $\beta$ s, 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 redefining 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

1. Mustard and others (1987) presented data from 42 small urban drainage basins located in several cities around the United States. 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. The data are in the data frame AC15 in the Chapter11.RData workspace.

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 nonurban
PCOMM	percentage of area commercial
PIND	percentage of area industrial

Do not bother with transformations of the  $x$  variables—use these variables as they are. Pay special attention to multicollinearity. Try one or more of the approaches described in this chapter to select the best model for predicting LOGTN from these explanatory variables.

2. Analysis of covariance. The following 10 possible models describe the variation in sand-size particles (0.125–0.250 millimeters) in the Colorado River at Lees Ferry, Arizona. (There is no dataset to load, the regression results are shown in tabular form below). Select the best model from this set of 10 and interpret its meaning. The basic model describes a quadratic relation between concentration and discharge ( $X$ ). Do the intercept and (or) slope vary with the three seasons (S, summer; W, winter; or other)? Use  $\alpha=0.05$  for all hypothesis tests.

Basic model

$$Y = b_0 + b_1 X + b_2 X^2$$

where

$$\begin{aligned} Y &= \ln(\text{concentration of suspended sands}) \\ X &= \ln(\text{discharge}) \end{aligned}$$

	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

xModel #	Explanatory variables	SSE	df(error)
1	X, X2	69.89	124
2	X, X2, S	65.80	123
3	X, X2, S, SX	65.18	122
4	X, X2, S, SX, SX2	64.84	121
5	X, X2, W	63.75	123
6	X, X2, W, WX	63.53	122
7	X, X2, W, WX, WX2	63.46	121
8	X, X2, S, W	63.03	122
9	X, X2, S, W, SX, WX	62.54	120
10	X, X2, S, W, SX, WX, SX2, WX2	61.45	118

3. The Ogallala aquifer was investigated to determine relations 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 the data frame `AC16` of the workspace `Chapter11.RData`. What is the significance of these predictor variables?

# Chapter 12

## Trend Analysis

---

*Concentrations and loads of phosphorus have been observed at numerous tributaries to an important estuary over a 20-year period. How much has the central tendency of those concentrations and (or) loads changed over this 20-year time period? How confident are we about the direction of those changes? Are the changes we observe simply a reflection of the fact that the early part of the record may have been particularly dry and the later part wetter, or is the change more than we might expect after we consider these variations in weather? Is there an observable effect associated with a ban on phosphorus compounds in detergents that was implemented in the middle of the observed record?*

*Groundwater levels were recorded for many wells in a study area over 14 years. During the ninth year of observations, installation of new irrigation systems in the area increased groundwater withdrawals dramatically. Do the data show decreasing water levels in the region's wells as a result of increased groundwater pumping? How large do we think that decrease is and how confident are we about that estimated change?*

Trend analysis is a process of building a statistical model of the behavior of some hydrologic variable over time. Examples of the kind of variable we might consider include discharge, concentration of solutes or suspended matter, water levels, or water temperature. The data may take the form of measurements made at a regular interval such as hourly, daily, weekly, monthly, or yearly; it may also take the form of measurements that are made on an irregular basis such as concentrations of some solute in water quality samples. The statistical model we build generally consists of one or more of the following (1) components related to regular cycles such as seasonal, diurnal or tidal; (2) patterns driven by some exogenous variable (for example, discharge as a driver of solute concentration or precipitation as a driver of discharge); (3) a long-term trend in the central tendency of the process; and (4) some random variability which is composed of natural variability, measurement error, and some type of serial dependence. The focus of trend analysis is on the long-term trend component, we wish to know the direction and rate of change in the central tendency over some multi-year period of interest. If we could measure the variable of interest at an infinitely high sampling rate and with perfect accuracy it is virtually certain that we would conclude that there is some trend over time. But we don't have the luxury of infinitely high sampling rates or perfect measurements, thus statistical trend analysis methods are needed in order to estimate of the rate of change in the central tendency and to estimate our uncertainty about that rate of change. In particular we want our methods to avoid coming to the wrong conclusion about the sign of the trend slope (for example, we conclude that it is increasing when, in fact, it is decreasing), see McBride and others (2014).

Methods for trend analysis build on the basic concepts of regression analysis (including ordinary least squares [OLS] and nonparametric alternatives to OLS) in which the explanatory variable of interest is time. A variation on trend analysis is to use a variable that may represent a driver of hydrologic change, which has itself been changing monotonically (or nearly monotonically) over time, as an explanatory variable. Examples of these types of variables include percent of watershed that is impervious, percent of watershed that is in row-crop agriculture, pounds of nitrogen fertilizer applied, or measures of atmospheric properties such as global average carbon dioxide concentration. This alternative approach can be particularly valuable when the time history of the hypothesized driver of change has a highly nonlinear time trend. If we are correct about the role of this driver of change, then we may be better able to characterize the temporal pattern of the trend than would be the case if we simply used time as the explanatory variable. Examples of such situations might include measures of urbanization that may be very low for a long period, followed by a period of rapid increase, followed by more stable values once the watershed is developed. In this chapter, most of our examples will evaluate trends over time, but these other more causally linked examples are possible approaches that shouldn't be ignored.

Trend analysis considers both sudden and gradual trends over time. It also considers trends with, or without, consideration of the effects of time-varying natural drivers of the trends of interest. In this chapter, various common types of tests are classified, and their strengths and weaknesses compared. Spatial trends are also a type of trend (for example, does chloride concentration increase with decreasing distance from the ocean). Many of the methods described here have extensions to spatial trend analysis, but the details of spatial trend analysis are not considered in this book.

## 12.1 General Structure of Trend Tests

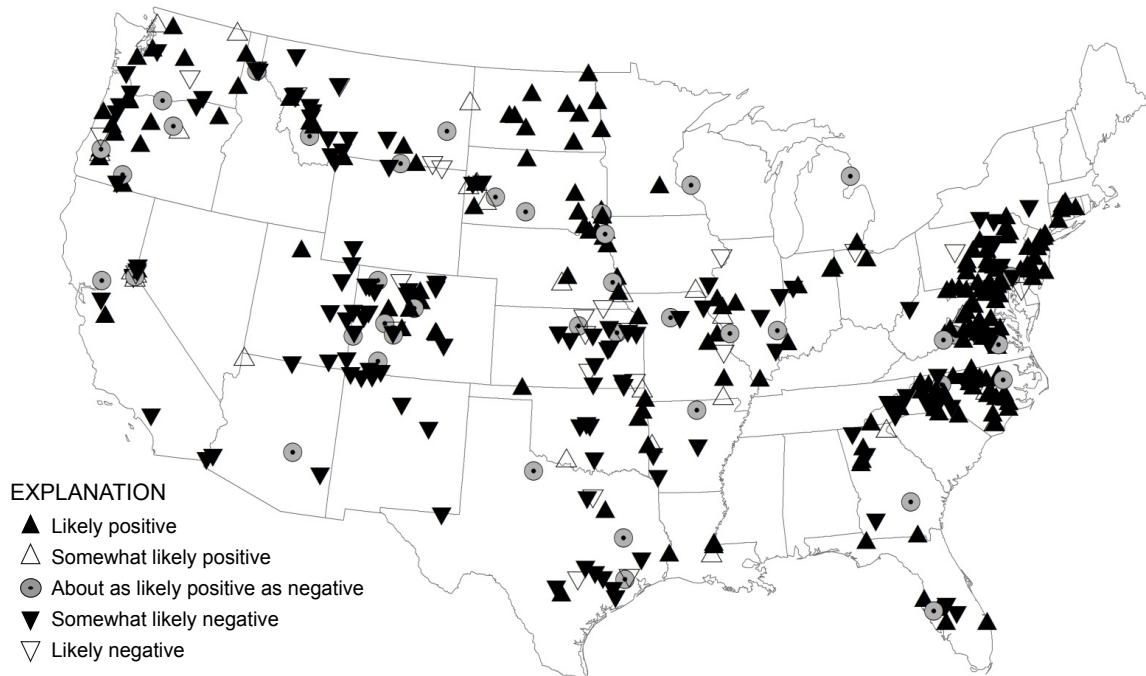
### 12.1.1 Purpose of Trend Testing

There are several possible motivations for conducting trend tests. In some cases, we are aware of some type of change on the landscape or in the climate that we expect might influence the central tendency of some hydrologic variable (for example, did the mean annual minimum daily discharge change after the removal of an upstream dam, or did the mean riverine load of nitrate change after farmers started planting winter cover crops upstream). In other cases, we may not have a specific hypothesis about a change driver, but recognize that many things are happening upstream or upgradient of the monitoring location and we want to know if the variable we have measured has changed its central tendency over a period of years. We might do a trend analysis to guide future investigation of the possible cause of the change or to update statistics used in engineering designs (for example, for flood control or water supply).

Another motivation for trend analysis comes about when there is a network of many sites at which long-term datasets have been collected (water levels, discharge, or water quality) and the goal is to categorize them by the trends that have taken place over some defined period of time, such as the most recent three decades. A trend test might be used to categorize the datasets into three groups: those where the variable of interest has generally increased, those where it has decreased, and those where the evidence for change is equivocal. These results might then be explored to see if the groupings correspond to specific types of trends in human activities that are likely drivers of such trends. Other purposes of trend tests can include estimating the magnitude of the trend over the period. The trend test can also be designed to help provide insights about the nature of the change that has taken place. For example, determining if a change in water quality is more pronounced at high discharges or at low discharges. Changes at high discharges may be driven by changes in nonpoint sources. In contrast, changes at low discharges may be driven by changes in point source loading or in groundwater baseflow contributions. Sometimes, because intersite comparability is crucial, it is vital that the same method be applied across all of the sites examined. In other cases, the procedure might be tailored to the specific site and dataset, with the goal of maximum understanding of the changes at that specific site. The choice of approaches should be substantially influenced by the purpose of the test as well as the statistical properties of the dataset.

There may be a desire by some practitioners to seek a unique, right way to apply trend tests for any given type of dataset. In reality, there can be several good approaches, each with its own advantages and disadvantages. When the results are of major consequence, using two or three methods may be appropriate. If all results from different tests are in approximate agreement, then the analyst should have high confidence in the results of any of them. But if there is strong disagreement, the analyst should try to understand the reason for those differences and, from that, draw conclusions as to which test might be the best way of characterizing the trend. Trend analysis can be viewed as a process of model building in which the data analyst considers many factors that may affect the central tendency of the data and then evaluates if, after accounting for those factors, there is a significant change in the central tendency in the variable of interest over time. There is no unique correct way to make such a model and do the tests but there are some principles and a set of good practices that need to be applied to trend testing, and this chapter is designed to lay out those principles and examples.

In the general formulation of a trend test, a series of observations of a random variable (concentration, unit well yield, biologic diversity, and so forth) has been collected over some period of time. We would like to determine if the values generally increase or decrease. We would also like to describe the amount or rate of that change, usually in percentage terms relative to the mean or median, or in original units relative to some important environmental benchmark such as a drinking water or aquatic life standard. Interest may be in data at one location or across a large geographical area. An example of a product



**Figure 12.1.** Map showing trend analysis results for specific conductance for the time period 1992–2002 (based on Oelsner and others, 2017). Definitions of the symbol categories are as follows. Likely positive the likelihood that the trend is actually positive is between 0.85 and 1.0; Somewhat likely positive, the likelihood that the trend is actually positive is between 0.7 and 0.85; About as likely positive as negative, the likelihood that the trend is actually positive is between 0.3 and 0.7 and also that the likelihood that it is actually negative is between 0.3 and 0.7; Somewhat likely negative, the likelihood that the trend is actually negative is between 0.7 and 0.85; Likely negative, the likelihood that the trend is actually negative is between 0.85 and 1.0.

designed to depict results across a broad geographical area is shown in figure 12.1—a set of trend results of specific conductance for the period 1992–2012 (a measure of the ionic strength of water) for rivers across the United States. An important feature of such a presentation is that it shows all of the results across the United States, and not just those that are considered to have attained a high degree of statistical significance. A common problem with the display of multiple trend results is that the analyst's judgment about what is significant enough to be shown may have the effect of eliminating potentially important information about the tendencies for broad areas to show similar results, even if many of them are not individually statistically significant using a typical two-sided significance test with a low  $\alpha$  value such as 0.05 or 0.02. In this example, a likelihood criterion is used so that every site analyzed has an indication of trend direction unless the statistical evidence for trend direction was poor. The circles are used here to indicate where the likelihood of a positive trend is in the range of 0.3 to 0.7, which is called "about as likely as not" to be a positive trend. Another way to think about those sites denoted by the circles is that positive and negative trends are about equally likely. The triangles indicate locations where the evidence about trend direction is relatively strong, the strongest being shown by the filled triangles. These results are developed using the weighted regressions on time, discharge, and season (WRTDS) method of trend analysis in conjunction with the WRTDS bootstrap test (Hirsch and others, 2015). It is always important to remember that indicators of likelihood or significance are not indicators that the trend is of practical significance. This can only be determined by evaluating the kinds of impacts that the estimated trends have on things like ecosystem conditions, water supply reliability, flood damages, or human health.

In most applications of trend analysis, the null hypothesis,  $H_0$ , is that there is no trend in the central tendency of the random variable being tested. The precise mathematical definition of  $H_0$  depends on the test that is being applied. The null hypothesis typically includes a set of assumptions related to the distribution of the data (normal versus non-normal), the type of trend (linear, monotonic, or step), and the degree of serial correlation. As discussed in chapter 2, the outcome of the test is a decision—either  $H_0$  is rejected or not.

Failing to reject  $H_0$  does not prove there is no trend. Rather, it is a statement that given the available data and the assumptions of the particular test, there is not sufficient evidence to conclude that there is a trend. The possible outcomes of a statistical test in the context of trend analysis are summarized in table 12.1.

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. It is a function of the true magnitude of the trend and other characteristics of the data (for example, distribution type, variance, serial correlation). In practice, these are never known. The analyst should try to select a test that has high power for the characteristics of the data they expect to encounter, such as the nature and magnitude of the trend and the variability of the data. Because this is often not known, the test selected should be robust—it should have relatively high power in all situations and types of data that might reasonably be expected to occur in the study being conducted. Thus, it is crucial that we give some thought to the characteristics that our datasets might have that would influence the suitability and power of a particular trend test. Some of the common characteristics to consider are (1) skewed distributions, often with a finite lower bound; (2) presence of outliers; (3) natural cycles such as annual, weekly, diurnal, or tidal; (4) missing values or irregularly spaced samples; (5) the presence of censoring; and (6) serial dependence.

In addition to having high power, another criterion for selecting a test is the type of information provided by the test. It is commonly the case that we are interested in much more than simply the decision to reject or fail to reject the null hypothesis. We may be interested in quantifying changes over particular periods at some site (for example, concentrations declined from 1980 to 1990 and then increased substantially from 1990 to 2010), or changes in different seasons (for example, summer trends are large but winter trends are much smaller), or changes in different discharge conditions (for example, concentrations have declined substantially under low flow conditions but are nearly unchanged at high flow).

Before delving into the details of various approaches to trend analysis, it should be mentioned that there are emerging alternative views about how the results of trend assessments should be reported. The emphasis on these approaches, as described by McBride and others (2014), Hirsch and others (2015), and McBride (2019) is to summarize the analysis with information about the best estimate of the trend magnitude, accompanied by a statement of the probability that the sign of the trend magnitude might be wrong. This is a different perspective than the null hypothesis significance test approach described in table 12.1, but computationally they are closely related. These approaches assume that if enough data were collected, we could demonstrate conclusively that there is either a positive or negative trend. That is, there is never a trend of zero magnitude, it is always positive or negative. In these approaches, the role of the data analyst is to inform a decision maker to help them reach one of three conclusions (see Tukey [1960] and Jones and Tukey [2000]): (1) act as if the trend is positive, (2) act as if the trend is negative, or (3) decide that there is insufficient evidence to decide between a positive versus a negative trend. As such, the key information about uncertainty is a statement of the likelihood that the true trend is of a different sign than the trend they have decided on. These outcomes might be best expressed in words (like “positive trend is highly likely” or “positive trend and negative trend are about equally likely”) using a lexicon that is quantitatively defined by the analyst in consultation with decision makers.

This chapter will not delve into the mechanics of how these likelihoods are computed. In McBride’s analysis, they involve the use of one-sided confidence intervals, in Hirsch’s analysis, they are the outcome of a bootstrap procedure. The analyses both adhere to the general ideas emphasized in the American Statistical Association (ASA) principles (presented in chap. 4) regarding statistical significance and  $p$ -values (Wasserstein and Lazar, 2016) in particular, “scientific conclusions and business or policy decision should not be based only on whether a  $p$ -value passes a specific threshold” and also “by itself, a

**Table 12.1.** Probabilities associated with possible outcomes of a trend test.

Decision	True situation (unknown, in reality)	
	No trend	Trend exists
No trend	$1-\alpha$ Probability that $H_0$ is not rejected	(Type II error) $\beta$ Probability that $H_0$ is not rejected
Trend	(Type I error) Significance level $\alpha$ Probability that $H_0$ is rejected	(Power) $1-\beta$ Probability that $H_0$ is rejected

*p*-value does not provide a good measure of evidence regarding a model or hypothesis.” These alternative approaches honor the idea that each decision maker has a unique level of risk-tolerance, and the data analyst’s job is to provide them with information that will help them make that decision. Prespecifying a particular  $\alpha$ -value takes away that opportunity for the decision maker to select their preferred risk level. These new approaches are now being used in applications related to water quality trend assessments for rivers in the Chesapeake Bay watershed in the United States (<https://cbrim.er.usgs.gov/>) and nationwide assessments in New Zealand (Larned and others, 2016) and the United States (Oelsner and others, 2017). The remainder of the chapter will use a traditional null hypothesis significance testing vocabulary and mathematics, but readers should know that there are emerging alternative approaches to conveying this uncertainty information to decision makers. Much of the underlying probabilistic quantification of the classical approaches still applies in these alternative approaches.

## 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 parametric and nonparametric. The second factor, shown in the columns, is whether there is some attempt to remove variation caused by other associated variables. See the headnote to table 12.2 for the definitions of the types of variables used here.

Examples of exogenous variables might be precipitation amount when the  $Y$  variable is streamflow or water level change, or it might be river discharge when the  $Y$  variable is concentration of some solute. For our purposes in this chapter, an exogenous variable is an explanatory variable that is also a random variable. Time may also be an explanatory variable in a trend analysis, but it is not a random variable. The reason for using exogenous variables is that they may explain a substantial part of the variance of the response variable (for example, precipitation explains a great deal of the variance of runoff) and by accounting for these kinds of relations the trend signal may be much more easily detected, which increases the power and accuracy of the trend analysis method. This doesn’t negate the value of doing the simpler trend test without an exogenous variable, it simply adds to our ability to discern and potentially explain the trend that is taking place. Simple trend tests (not adjusted for  $X$ ) are discussed in section 12.2. Tests adjusted for  $X$  are discussed in section 12.3. The approaches shown in sections 12.2 and 12.3 assume that the data have no seasonal or other regular periodic component. Section 12.4 expands on the previous sections by adjusting the methods for the presence of a regular periodic component.

If the trend is spatial rather than temporal,  $T$  could be downstream order, distance downdip, distance from a dam or a pollution source, and so forth. Examples of  $X$  and  $Y$  include the following

- For trends in surface water quality,  $Y$  could be concentration,  $X$  could be streamflow, and  $R$  could be called the flow-adjusted concentration.
- For trends in flood magnitudes,  $Y$  could be streamflow,  $X$  could be the precipitation amount, and  $R$  could be called the precipitation-adjusted flow. In this type of problem, the duration of precipitation variable used must be appropriate to the flow variable being tested. For example, if  $Y$  is the annual flood peak from a 25 square kilometer ( $\text{km}^2$ ) basin, then  $X$  might be the 1-hour maximum rainfall, whereas if  $Y$  is the annual flood peak for a 25,000  $\text{km}^2$  basin, then  $X$  might be the 24-hour maximum rainfall.

**Table 12.2.** Classification of five types of tests for trend.

[ $Y$ , 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 loess of  $Y$  versus  $X$ ;  $T$ , time (typically expressed in years); -, not applicable]

Type of trend test	Not adjusted for the exogenous variable, $X$	Adjusted for the exogenous variable, $X$
Nonparametric	Mann-Kendall trend test of $Y$	Mann-Kendall trend test on residuals $R$ from loess 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 $T$

## 12.2 Trend Tests with No Exogenous Variable

### 12.2.1 Nonparametric Mann-Kendall Test

A simple way to construct a trend test is to determine if the central tendency of the variable of interest,  $Y$ , changes, in a monotonic fashion, with the time variable,  $T$ . Mann (1945) first suggested using the test for significance of the Kendall's  $\tau$  correlation value as a test for trend, whereby the two variables being related are  $Y$  and the time variable,  $T$ . The Mann-Kendall test can be stated most generally as whether  $Y$  values tend to increase or decrease as  $T$  increases (monotonic change).

$$H_0 : \text{Prob}[Y_j > Y_i] = 0.5, \text{ where time } T_j > T_i.$$

$$H_1 : \text{Prob}[Y_j > Y_i] \neq 0.5 \text{ (two-sided test).}$$

In the two-sided test we are only interested in determining, for any pair of observations in the record, if the probability that the later observation is greater than the earlier one is different from a value of 0.5. If the probability were  $>0.5$  that means that there is a tendency for  $Y$  to increase over time and if the probability were  $<0.5$  that means that there is a tendency for  $Y$  to decrease over time.

No assumption of normality is required, but there must be no serial correlation in the  $Y$  values (after detrending the data) for the resulting  $\alpha$  level of the test to be correct. Typically, the test is used to determine whether the central value or median changes over time. As discussed in chapters 8 and 10, the results of the test do not change if the  $Y$  data are transformed by any monotonic transformation (such as logarithm or square root transformation). The test is performed by computing the Kendall's  $S$  statistic from the  $Y$  and  $T$  data pairs (see chaps. 8 and 10 for details).  $H_0$  is rejected if the value of  $S$  is statistically significantly different from zero. For large samples, the distribution of  $S$  given  $H_0$  is approximately normal, with variance determined only by the sample size.

There is also an estimate of the slope of such a temporal trend that is closely related to the  $S$  statistic and the hypothesis test. This is the Theil-Sen estimator (also known as the Kendall-Theil robust line; see section 10.1). If  $S$  is positive the Theil-Sen slope will be positive and if  $S$  is negative the Theil-Sen slope will be negative. The units of the slope estimate are the units of the  $Y$  variable divided by the units of time. For example, if the trend test is concerned with annual mean discharge, then the  $Y$  units might be cubic meters per second ( $\text{m}^3/\text{s}$ ) and the  $T$  units would be year (yr), so the slope units would be  $\text{m}^3/\text{s}/\text{yr}$ .

The example in figure 12.2 illustrates a dataset with a fairly strong long-term trend (the annual mean discharge of the Mississippi River at Keokuk, Iowa, 1931–2013). Let's say the selected  $\alpha$  level for the test was 0.05. The two-sided  $p$ -value for the Mann-Kendall test is 0.000006 with  $\tau=0.34$ . Thus, we can reject  $H_0$  and conclude that there is a trend and that the trend is positive. The units of discharge are  $\text{m}^3/\text{s}$  and the units of time are years. The Theil-Sen robust line slope is 13.0  $\text{m}^3/\text{s}/\text{yr}$ . Using the median of the discharge values and the median of the time values as suggested in section 10.1, the equation for the Theil-Sen robust line is

$$\hat{Q} = \beta_0 + \beta_1 \cdot T \quad (12.1)$$

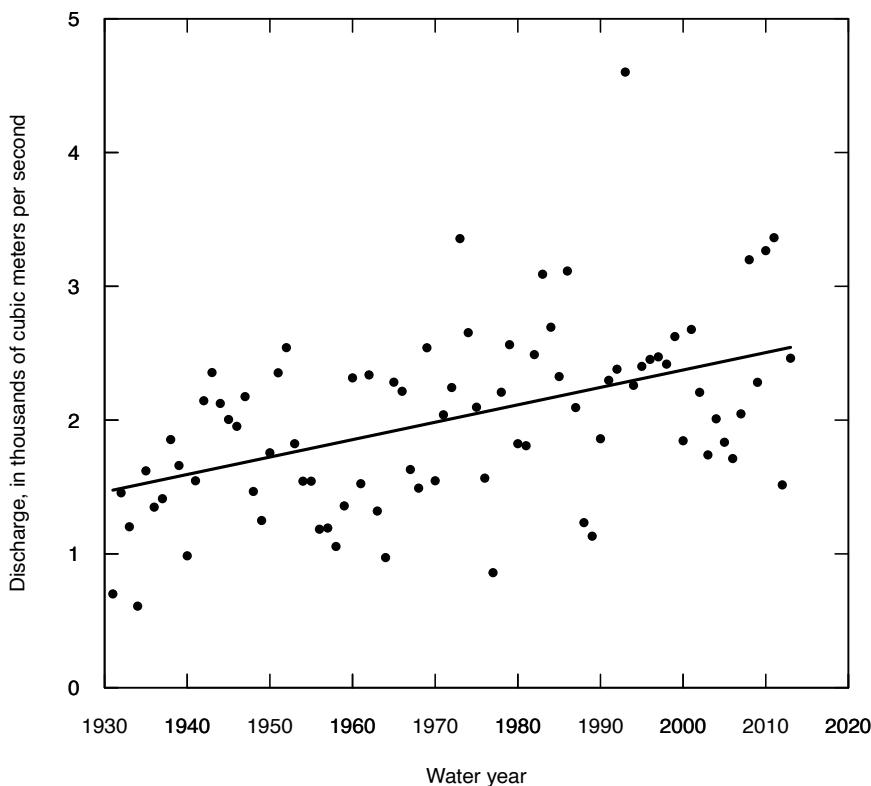
where

- $\hat{Q}$  = estimate of discharge in  $\text{m}^3/\text{s}$ ;
- $T$  = time in years;
- $\beta_1$  = Theil-Sen slope; and
- $\beta_0$  =  $\text{median}(Q) - \beta_1 \cdot \text{median}(T)$ .

In this particular example  $\beta_0 = -23,659$  and  $\beta_1 = 13.0$ . The Theil-Sen line shown in figure 12.2 is

$$\hat{Q} = -23,659 + 13.0 \cdot T$$

One could consider doing this analysis on log-transformed discharge data. The Mann-Kendall hypothesis test for trend will have exactly the same result ( $p$ -value is still 0.000006) as it did without the log transformation; however, the slope will be different because of the transformation. If we use the natural logarithm for our transformation (as opposed to base 10) the slope is 0.006696 per year. The  $p$ -value will



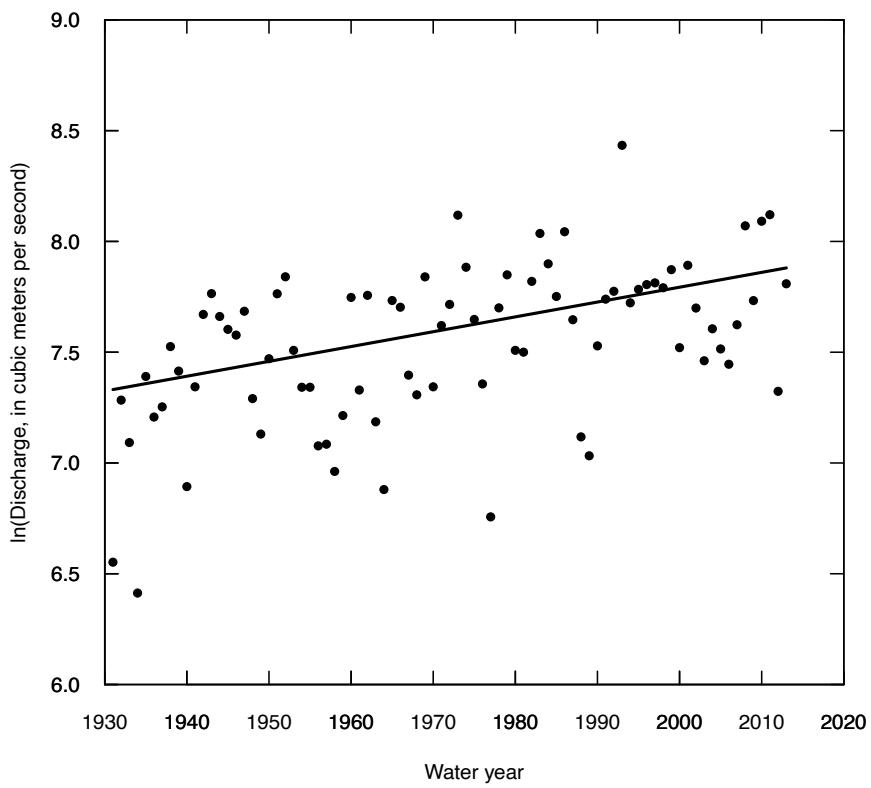
**Figure 12.2.** Plot of annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013, shown with the Theil-Sen robust line.

remain unchanged no matter what monotonic transformation is used (such as any one of the ladder of powers transformations). If the change is assumed to be linear with time, in this case linear in  $\ln(Q)$ , then that means that the trend in the original discharge units will be exponential and the ratio of the expected value from one year to the expected value for the previous year will be  $\exp(0.006696)$ , which is 1.006718. This result can be interpreted as an increase of 0.6718 percent from one year to the next. We will return to this transformation issue in more detail in section 12.5. The graphical representation of the Mann-Kendall test for trend and Theil-Sen robust line on the natural log of discharge is shown in figure 12.3.

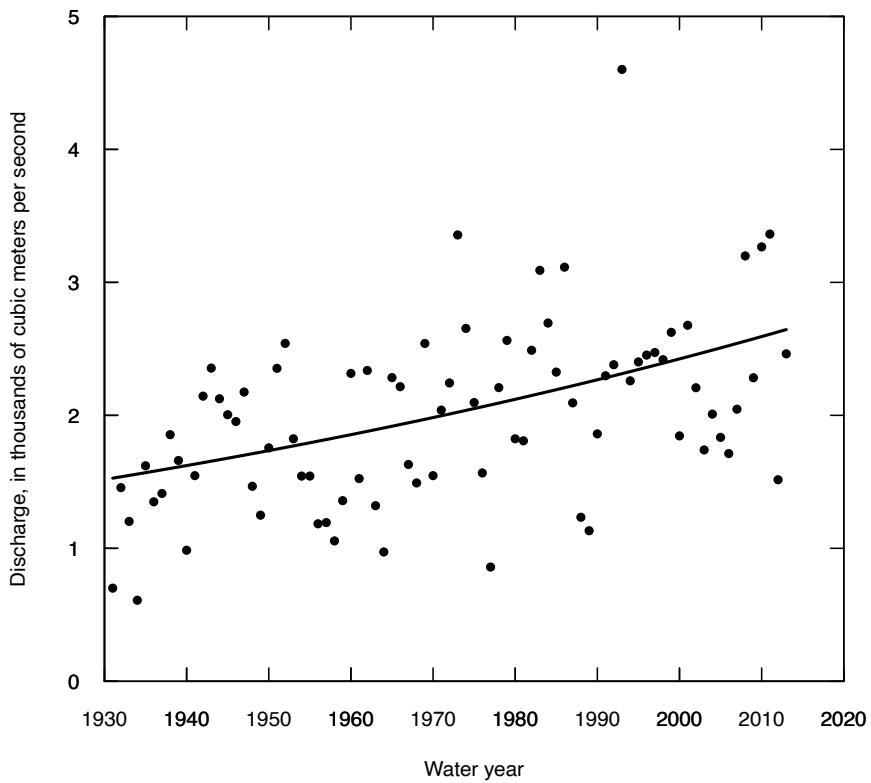
The original data are shown in figure 12.4, but the solid line is the Theil-Sen estimate transformed back into the original units (by exponentiating the estimates from the log space model). In this case, there is relatively little difference between the two estimates (compare figs. 12.2 and 12.4), but in some cases there can be a substantial difference.

There is no definitive test that will indicate which is a better estimate, it is really a matter of judgment based on apparent curvature. In this case, either approach could be acceptable. One final comment on the Theil-Sen robust line: it should not be considered an estimate of the conditional mean of the distribution for any given year. It is more suitable to think of it as an estimate of the conditional median. For example, in figure 12.3, we may say that for any given year the probability of the true value being above the Theil-Sen line is 0.5 and being below is 0.5.

Because the Mann-Kendall test for trend provides the same results no matter what transformation is used on the  $Y$  values, it can be suitable for studies of a large number of similar datasets (for example, 50-year records of mean discharge) across a study area. Each dataset may suggest the need for a different ladder-of-powers transformation to make the relation more nearly linear, but because the Mann-Kendall test results do not vary across all the possible ladder-of-powers transformations, this step of transformations becomes unnecessary for applying the test to multiple sites. This is one reason that a nonparametric test such as the Mann-Kendall is well suited to a study that encompasses tests of many datasets.



**Figure 12.3.** Plot of the natural log of annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013, shown with the Theil-Sen robust line.



**Figure 12.4.** Plot of the annual mean discharge, Mississippi River at Keokuk, Iowa, 1931–2013, shown with the transformed Theil-Sen robust line based on slope of the natural log discharge values.

## 12.2.2 Ordinary Least Squares Regression of $Y$ on Time, $T$

Ordinary least squares (OLS) regression of  $Y$  on  $T$  is a test for trend

$$Y = \beta_0 + \beta_1 \cdot T + \varepsilon \quad (12.2)$$

The null hypothesis is that the slope coefficient,  $\beta_1$ , is zero. OLS regression (eq. 12.2) makes stronger assumptions about the behavior of  $Y$  over time than does Mann-Kendall. The relation must be checked for normality of residuals, constant variance, and linearity of the relation (best done with residuals plots—see chap. 9). If  $Y$  is not linear with time, a transformation will likely be necessary. If all of the conditions of OLS regression are met, then the slope estimate is  $b_1$ , and the  $t$ -statistic on  $b_1$  can be used to determine if the slope is significantly different from zero. Further, this test for trend has slightly more power than the nonparametric Mann-Kendall test if the conditions of OLS regression are met. But, with modest departures from normality of residuals, the Mann-Kendall test can be a good deal more powerful than regression. If the  $t$ -statistic is large (in absolute value), typically  $|t| > 2$ , we can reject the null hypothesis and conclude that there is a time trend (upwards if the estimate is positive and downward if it is negative). Unlike Mann-Kendall, the test results for regression (specifically the  $p$ -value on the slope coefficient) will not be the same before and after a transformation of  $Y$ .

## 12.2.3 Comparison of Simple Tests for Trend

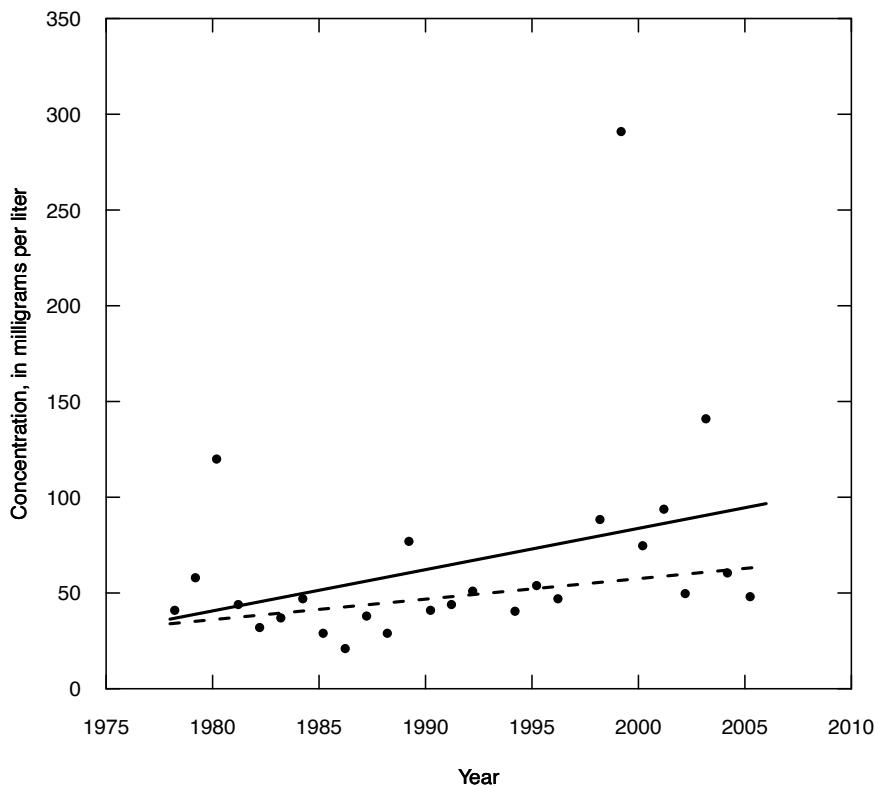
If the model form specified in a regression equation were known to be correct ( $Y$  is linear with  $T$ ) and the residuals were truly normal, then OLS regression would be optimal (most powerful and lowest error variance for the slope). Of course, usually we cannot know this in any actual situation. 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 chap. 10, and Hirsch and others [1991], p. 805–806).

There are practical cases where the OLS regression approach is preferable, particularly in the multiple regression context (see section 12.3). A good deal of care needs to be taken to ensure the regression is correctly applied and enough information is provided such that the audience is able to verify that the assumptions have been met. 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. Nonparametric methods are almost always nearly as powerful as OLS regression, and failure to edit out a small percentage of bad data or correctly transform the data will not have a substantial effect on the results.

### Example 12.1. Milwaukee River chloride trends.

Chloride concentrations sampled in the month of March in the Milwaukee River at Milwaukee, Wisconsin for the years 1978–2005 are shown in figure 12.5. Two trend tests were conducted, the Mann-Kendall test and OLS regression on time. The Theil-Sen and OLS regression lines are plotted along with the data. Using  $\alpha=0.05$  the OLS regression line is not significantly different from a slope of zero, and thus we would not reject the null hypothesis of no trend ( $p$ -value is 0.091), but the Mann-Kendall test statistic ( $S$ ) is significantly different from zero ( $p$ -value is 0.017).

It is interesting to note that the linear regression line is a good deal steeper than the Theil-Sen line even though it is not significant, but the Mann-Kendall test is significant. The linear regression line is heavily influenced by the high outlier value in 1999. This high value and the highly skewed distribution of the residuals are the reasons that the linear regression approach fails to provide confirmation of a trend, whereas the Theil-Sen line is unaffected by the magnitude of this high value. Later in this chapter we will return to this dataset and consider ways that the parametric approach could be improved upon to result in a more sensitive and meaningful description of the trend, which the Mann-Kendall test strongly suggests is present.



**Figure 12.5.** Graph of chloride concentration in the month of March for the Milwaukee River, at Milwaukee, Wisconsin. Solid line is the estimated trend using linear regression. The dashed line is the estimated trend using the Theil-Sen robust line. Note how the high value in 1999 has a strong influence on the linear regression but not the Theil-Sen robust line.

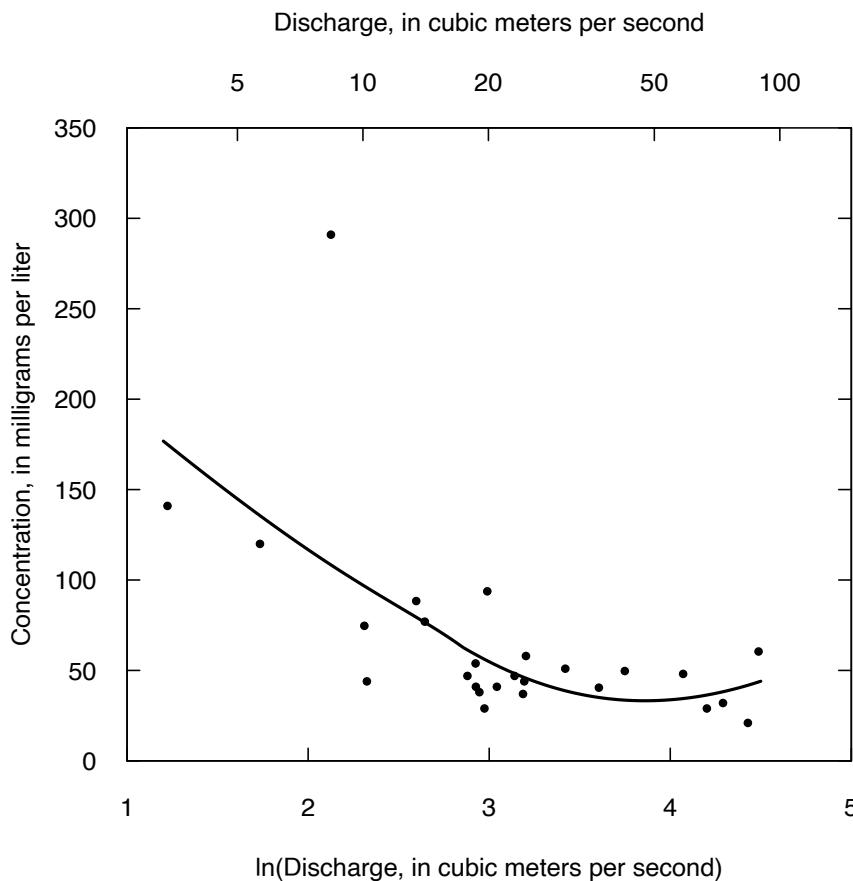
## 12.3 Accounting for Exogenous Variables

Variables other than time 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 the trend signal that may exist can be seen, and the power (ability) of a trend test to discern changes in  $Y$  with  $T$  is increased. Two important types of situations are relevant here. The first is exploration of trends in some measure of water quantity. In these cases, the  $Y$  variable may be something like annual mean discharge, annual minimum discharge, annual maximum discharge, or some measure of change in storage in a lake or aquifer. The obvious candidate for an exogenous ( $X$ ) variable would be some measure of precipitation at an appropriate spatial and temporal scale (for example, annual total precipitation averaged over several precipitation gages that span the watershed of interest). The second type of situation is the exploration of trends in some measure of water quality such as concentrations of solutes or sediment, a biological measure such as chlorophyll, or water temperature. In these cases, the obvious candidate for an exogenous ( $X$ ) variable might be discharge at the time of water quality measurement or perhaps discharge during some period before the time of measurement. The discussion presented below applies to the latter case, but the principles discussed here could apply to either one. The process of removing the variation resulting from  $X$  involves modelling the effects of exogenous variables with OLS regression or loess (for computation of loess, see chap. 10). Using the same dataset used in the example above, we will consider all three of the options mentioned in the “Adjusted for the exogenous variable,  $X$ ” column in table 12.2.

Consider a regression of  $Y$  versus  $X$ . The residuals ( $R$ ) from this regression express the variation in  $Y$  not explained by  $X$ . A trend in  $R$  implies a trend in the relation between  $X$  and  $Y$ . This in turn implies a trend in the distribution of  $Y$ , but this conclusion may not hold if there is a trend in  $X$ . What kind of variable is appropriate to select as an exogenous ( $X$ ) variable? We use the term exogenous here to indicate that it is a particular kind of explanatory variable that is itself a random variable which is externally driven rather than being driven by some human activity that may also be driving variations in  $Y$ , the variable for which we are doing the trend test. This exogenous variable should be a measure of a driving force behind the process of interest, but it must be relatively free of changes owing to human manipulation. For a water-quality trend study, the streamflow record at (or near) the site where the water quality data were collected is an obvious choice for an exogenous ( $X$ ) variable. However, if the streamflow record being used includes a time span that covers a period both prior to and after major upstream water management changes, then the streamflow record would be unacceptable as a choice of an exogenous random variable because the probability distribution of  $X$  has likely changed substantially during the period of interest. Examples of such changes include the completion of a major dam, removal of a major dam, initiation of a major new diversion in or out of the watershed, or a major change in operating policy of a water resource system. A streamflow record which reflects some subtle human influence is acceptable, provided that the effect 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 streamgage 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. Of course, as landscape manipulations (such as artificial drainage), regional groundwater depletion, or global greenhouse gas concentrations increase over time, it becomes impossible to say that any hydrologic or climatic variable is free of human manipulation. Decisions to use climate records or streamflow records as exogenous variables in a trend analysis involve a trade-off. The use of exogenous variables is very helpful in reducing the unexplained variance in the variable of interest ( $Y$ ) and thereby increasing our ability to discern and describe the trend. However, as time periods get longer and climate, landscape, or groundwater storage changes get stronger, the use of such exogenous variables becomes problematic. To do a trend study of a random variable,  $Y$ , we need to be confident that the observed trend in  $Y$  is a function of a shift in the  $X$ - $Y$  relation and not simply a function of a trend in  $X$ . Resolution of this trade-off will be a challenge to water-related trend studies for the foreseeable future.

Where  $Y$  is a concentration (of a solute or particulate matter), 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 processes. One process is dilution: a solute may be delivered to the stream at a reasonably constant rate (for example, effluents from a point source or groundwater discharge to the stream) as discharge changes over time. The result of this situation is a decrease in concentration with increasing flow; 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. 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. The resulting pattern is an initial dilution at the low end of the discharge range, followed by an increase with discharge at higher values of discharge. The Milwaukee River chloride record exhibits this kind of nonmonotonic behavior in the  $X$ - $Y$  relation, as illustrated in figure 12.6.

Subsections 12.3.1., 12.3.2., and 12.3.3. consider three types of approaches to trend testing using an exogenous variable. All three are applied to the dataset presented in the previous example (Milwaukee River chloride data). We will consider a dataset that consists of a single chloride concentration value for



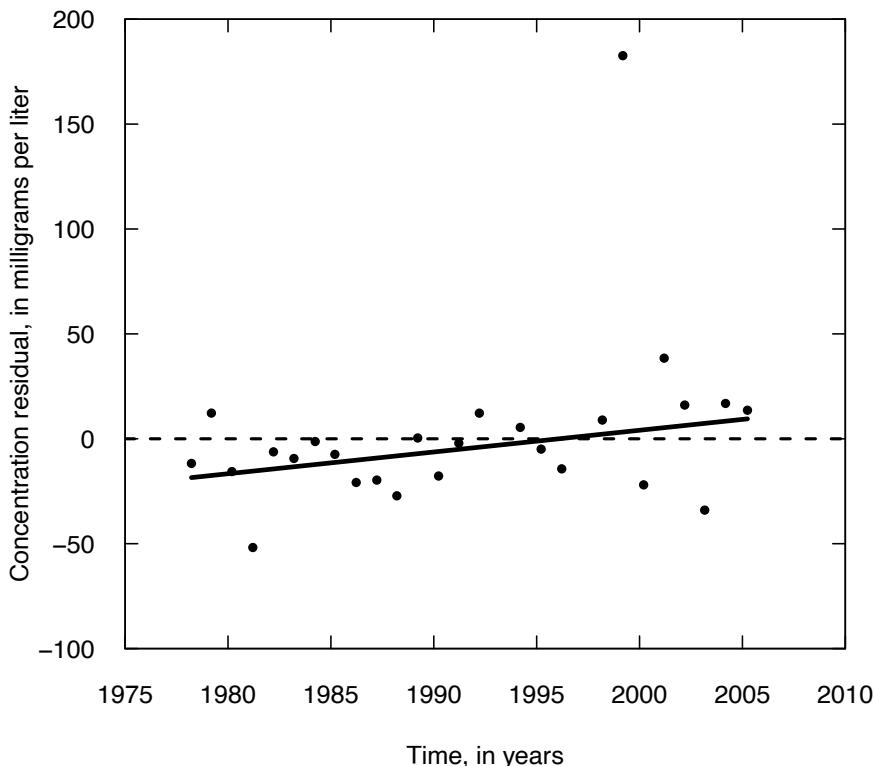
**Figure 12.6.** Graph of the relation between chloride concentration and the natural log of discharge, Milwaukee River at Milwaukee, Wisconsin, for samples collected in March 1978–2005. Solid line is the loess smooth. Residuals are the vertical differences between the data points and the line.

the month of March from each year, and the associated daily discharge value on that sampling date for the Milwaukee River for the 26 years from 1978 through 2005.

### 12.3.1 Mann-Kendall Trend Test on Residuals, $R$ , from Loess of $Y$ on $X$

The first of the three approaches to trend testing is to remove the influence of discharge on the chloride data and then do a Mann-Kendall test for trend in the residuals. The relation between chloride concentration and discharge is clearly evident in figure 12.6. In this case the exogenous variable is the natural log of discharge, the smoothing method is the `loess` function, and the `span` value is 1.0 (the default value for `span` is 0.75). See chapter 10 for a discussion of `loess` and setting the `span`. The resulting scatterplot with the superimposed loess curve indicates that these are reasonable choices (the curve follows the bulk of the data and doesn't have any jagged oscillations). The decision to transform the explanatory variable and the choice of the `span` value for the loess are judgement calls on the part of the analyst. One could also try an OLS regression fit for the purpose of computing residuals, but an examination of figure 12.6 provides ample evidence that OLS would be problematic because of the curvature in the relation at higher discharges. With OLS, the residuals would not be independent of the explanatory variable.

Based on this loess fit, residuals are computed and they can be plotted against time (fig. 12.7). One thing that is clear at a glance is that the distribution of the residuals contains one very extreme outlier. Based on this fact, the use of a nonparametric approach to trend testing is probably a good approach. For the null hypothesis that the residuals are trend free, we apply the Mann-Kendall test using a two sided  $\alpha=0.1$  and the result is that we should reject the null hypothesis (the attained  $p$ -value is 0.052).

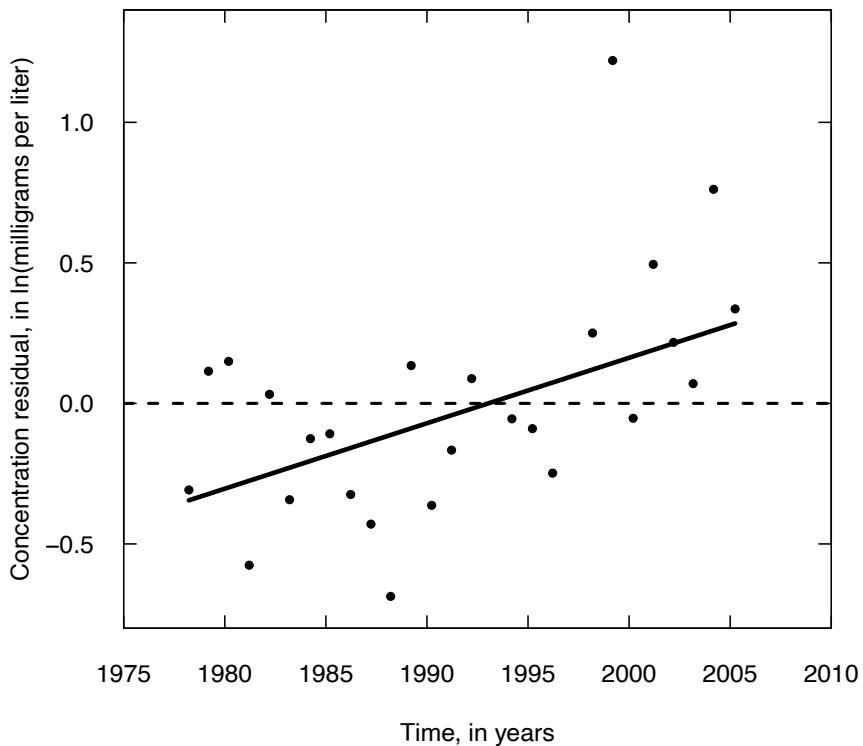


**Figure 12.7.** Graph of concentration residuals (from a loess fit of concentration as a function of the log of discharge) versus time for chloride concentrations in the Milwaukee River at Milwaukee, Wisconsin, for samples collected in March, from 1978 through 2005. Solid line shows the Theil-Sen line, with a slope of 1.03 milligrams per liter per year. Dashed line is residual = 0 for all years.

We can quantify the slope of the trend using the Theil-Sen slope estimator and obtain a result of a slope of 1.03 milligrams per liter (mg/L) per year. Using this slope and the median value of  $X$  (the natural log of discharge) and the median value of  $R$  (the residuals from the loess fit) we can depict the trend with a straight line (also shown in fig. 12.7) using the method defined in section 10.1.1. Note that if we were to test for trend using OLS regression of residuals versus time (or through the use of the product-moment correlation coefficient or Pearson correlation) the result would still indicate a positive trend but the  $p$ -value would be 0.090. The reason for the difference in  $p$ -values is that in the case of the OLS regression, the one large outlier causes the residual variance of the relation to be high and thus the strength of the trend is diminished by the indication of high variability. In the case of the Mann-Kendall test the large outlier is not consequential. The resulting  $p$ -value would have been the same if, instead of a residual of 177.7 mg/L it had been 50 mg/L or 500 mg/L. As discussed before, the test is resistant to the magnitude of outliers.

### 12.3.2 Mann-Kendall Trend Test on Residuals, $R$ , from Regression of $Y$ on $X$

Consider a mixed parametric and nonparametric approach to test for trends. In this approach, we use OLS regression of  $Y$  on  $X$  to obtain the residuals, which we then test for trend using a Mann-Kendall trend test. In this particular case, exploration of various forms of the regression model of concentration as a function of discharge suggests that an appropriate model would take the form  $\ln(C) = \beta_0 + \beta_1 \cdot \ln(Q) + R$ , where  $C$  is concentration,  $Q$  is discharge, and  $R$  is the residual. Using this type of model, with the dependent variable being a log transformation of the variable of interest, means that the residuals are no longer in the original units (mg/L) but are now residuals expressed in log space. We can then proceed to do a test for trend on these residuals using the Mann-Kendall test. The results are shown in figure 12.8, the trend slope is 0.0232/yr and the  $p$ -value for the test is 0.0082, which is a stronger indicator of trend than was attained from the previous test where the trend was computed on loess residuals. With this particular



**Figure 12.8.** Graph of log concentration residuals versus time for chloride concentrations in the Milwaukee River at Milwaukee, Wisconsin, for samples collected in March, from 1978 through 2005. Solid line shows the Theil-Sen line, with a slope of 0.023 log units/year. Dashed line is residual = 0 for all years.

dataset, the use of linear regression to compute the residuals is problematic because the small size of the dataset makes it difficult to verify the soundness of the approach. In this case, the loess approach to computing the residuals should probably be given more credence. The fact that in this case the trend is expressed in log units will be discussed later in section 12.5.

Alley (1988) showed that this type of 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  $e^*$  rather than  $R$  versus  $T$ , where  $e^*$  are the residuals from an OLS regression of  $T$  versus  $X$ :

$$T = b_0 + b_1 \cdot X + e^* \quad (12.3)$$

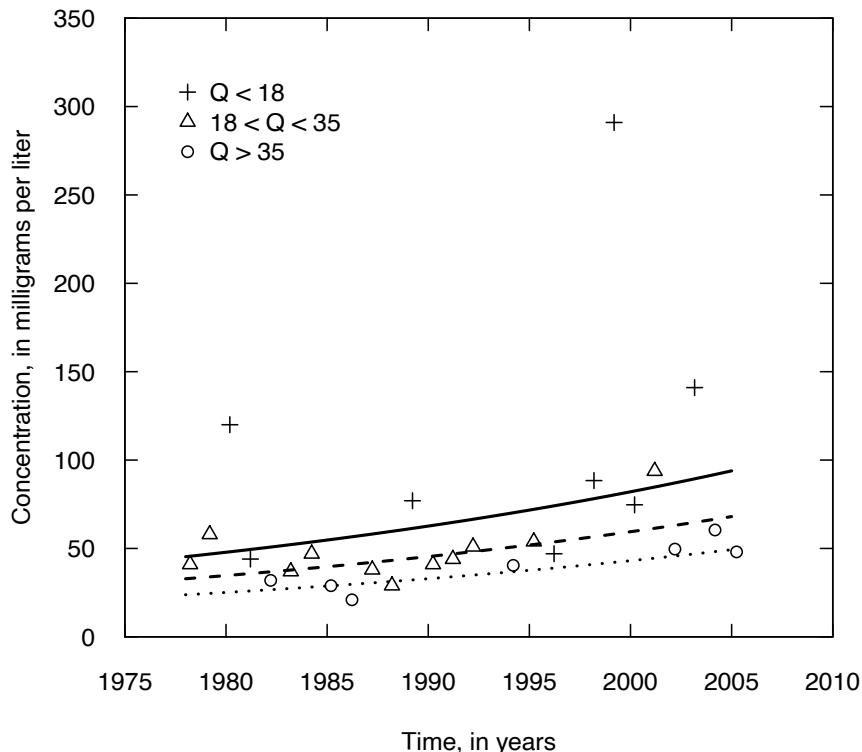
In this way the effect of a drift in  $X$  over time is removed, so that the  $R$  versus  $e^*$  relation is totally free of the influence of  $X$ . This test is a Mann-Kendall test on the partial residuals of  $Y$  versus  $T$ , having removed the effect of all other exogenous variable(s)  $X$  from both  $Y$  and  $T$  by regression. For more discussion of the partial Mann-Kendall test see Libiseller and Grimvall (2002).

### 12.3.3 Regression of $Y$ on $X$ and $T$

The regression of  $Y$  on  $T$  is an entirely parametric approach to evaluating trends in  $Y$  adjusted for some exogenous random variable,  $X$ . This approach uses multiple linear regression to do in a single step what the preceding two approaches did sequentially. If we knew that the dataset had the right sort of characteristics for multiple regression, then this would be the most powerful and least biased approach to the problem. The characteristics that would be most important are (1) that the relations be linear, and (2) that the errors

**Table 12.3.** Model coefficients and their *t*-statistics and *p*-values.

Regression results	$b_0$	$b_1$	$b_2$
Coefficient value	-48.27	0.0270	-0.465
<i>t</i> -statistic	-2.90	3.23	-5.31
<i>p</i> -value	0.008	0.003	<0.001

**Figure 12.9.** Graph of curves that represent median estimates of chloride concentration as a function of time, from the Milwaukee River at Milwaukee, Wisconsin. Solid line is at discharge ( $Q$ ) = 15 cubic meters per second ( $\text{m}^3/\text{s}$ ), dashed line is at  $Q$  = 30  $\text{m}^3/\text{s}$ , and dotted line is at  $Q$  = 60  $\text{m}^3/\text{s}$ . Symbols represent data points based on discharge in  $\text{m}^3/\text{s}$ .

be approximately normal. Of course, we can't determine these things for certain, but effective use of model building and checking can help to identify if the data at least approximate these characteristics. The building of a multiple regression model for this data needs to be done with care considering the many issues discussed in previous chapters on regression, particularly those related to issues of transformation of explanatory variables and dependent variables and concerns about the behavior of the residuals (chap. 11). The appropriate regression model for this dataset is

$$\ln(C) = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot \ln(Q) + \varepsilon \quad (12.4)$$

where  $C$  is concentration,  $T$  is time in years, and  $Q$  is discharge in  $\text{m}^3/\text{s}$ . The  $R^2$  for the fitted model is 0.63 and the overall *F*-test for the model shows it to be highly significant ( $p < 0.0001$ ). We can summarize the fitted coefficients, their *t*-statistics, and their *p*-values in table 12.3.

Viewing this as a trend test, we can focus directly on the  $\beta_1$  coefficient. We can see that the coefficient value is very nearly the same as the one we found in the previous test (Mann-Kendall on residuals from the  $\ln(C)$  versus  $\ln(Q)$  linear regression). We also see that it is statistically significant ( $p = 0.003$ ). This is exactly what we would hope to be the case, that this simultaneous approach gives similar results to those that come from the sequential parametric approach. Our conclusion for this example then is, for any particular value

of discharge, the expected value of the natural log of concentration is increasing over time and is doing so at a rate of 0.027 per year. This is equivalent to a 2.74 percent per year increase ( $\exp(0.027)=1.0274$ ). The statement is accurate, but not particularly easy to picture or understand; However, figure 12.9 provides a simple way to illustrate the fitted model. The data are categorized into three groups (low, medium, and high discharges) using different symbols for each group. Then the fitted regression is evaluated at three example discharges (15, 30, and 60 m<sup>3</sup>/s). The trends are linear in the natural log of concentration, but the graph is presented in concentration units so these linear trends in the logs become exponential trends in concentration. The curves represent a median estimate of concentration for each of the three example discharges because the model is built on the assumption that the errors around the regression line are normal with zero mean and constant variance. When the regression lines are transformed to concentration units, they still represent the conditional median of concentration for the given discharge and year, but they do not represent the conditional mean (see the discussion in chap. 9 regarding retransformation of regression results).

## 12.4 Dealing with Seasonality or Multiple Sites

There are many instances where changes between different seasons of the year are a major source of variation in the response ( $Y$ ) variable. As with other exogenous effects, seasonal variation must be compensated for or removed in order to discern the underlying trend in  $Y$  over time. If not, there may be only little power to detect trends which are truly present. A statistical test like the Mann-Kendall or OLS regression does not detect that there are seasonal patterns; rather the pattern registers as random noise in the process. This is because those tests are designed to detect monotonic or linear changes. Because trend tests (parametric or nonparametric) are fundamentally about being able to see a trend signal stand out above the noise, this seasonality will hinder our ability to truly observe the trend.

This issue can arise in many contexts. Concentrations of a water quality constituent in surface waters typically show strong seasonal patterns. Water temperatures also have a strong seasonal component and streamflow itself almost always varies across seasons. These seasonal patterns arise from seasonal variations in precipitation volume and air temperature. Air temperature affects the precipitation type (rain versus snow) and the rate of evapotranspiration. Some of the observed seasonal variation in water quality concentration may be explained by accounting for this seasonal variation in discharge. However, seasonality often remains even after discharge effects have been removed (Hirsch and others, 1982). Possible additional causes of seasonal patterns include biological processes and managed activities such as agriculture. For example, nutrient concentrations commonly vary with the seasonal application of fertilizers and the natural pattern of uptake and release by plants. Other effects are the result of different sources of water contributing to the water feature (stream, lake, or pond) at different times of the year, such as snow melt or intense rainfall. The seasonal rise and fall of ground water can also influence water quality. For example, a given discharge magnitude in one season may derive mostly from groundwater whereas the same discharge magnitude during 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. Thinking in terms of trends in water quantity (streamflow or groundwater level change) it is generally the case that the response of the hydrologic variable to a given amount of precipitation can differ considerably across seasons. This can be true because of the changes in interception by plant canopy, changes in runoff characteristics between bare fields and fields with active vegetation, and also differences in the amount of evapotranspiration across the seasons. In general, a rainfall-runoff relation is likely to be different at different times of the year, and failure to consider it will greatly diminish the power of any trend analysis.

Techniques for dealing with seasonality fall into three major categories, as shown in table 12.4; one is fully nonparametric, one is a mixed procedure, and the last is fully parametric. For the upper four cells of table 12.4 it is necessary to define a season. In general, seasons should be just long enough so that there

**Table 12.4.** General categories of options for dealing with seasonality in conducting trend tests.  $X$  is an exogenous variable such as precipitation or streamflow that may influence the variable of interest, in addition to the influence of season.

Type of trend test	Not adjusted for $X$	Adjusted for $X$
Nonparametric	Seasonal Kendall test for trend on $Y$	Seasonal Kendall trend test on residuals from loess of $Y$ on $X$
Mixed	OLS regression of deseasonalized $Y$ on $T$	Seasonal Kendall trend test on residuals from OLS regression of $Y$ on $X$
Parametric	Multiple regression of $Y$ on $T$ and seasonal terms	Multiple regression of $Y$ on $X$ , $T$ , and seasonal terms

is some data available for most of the seasons in most of the years of record but no shorter than monthly blocks. 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 four seasons, and so forth.

### 12.4.1 The Seasonal Kendall Test

The Seasonal Kendall test (Hirsch and others, 1982) accounts for seasonality by computing the Mann-Kendall test on each of  $m$  seasons separately, and then combining the results. In this test, a season can be monthly, quarterly, or some other definition of time. The advantage of this approach over an ordinary Mann-Kendall approach is that in the Seasonal Kendall test no comparisons are made across season boundaries. For example, for monthly seasons, January data are compared only with January, February only with February, and so on. The idea behind the test is that making comparisons across the seasons (say comparing the January 2012 value to the June 2008 value) is really not informative about trend and is more likely to be an expression of the differences between seasons rather than differences between years. The information we seek is found by only comparing data from the same season over different years. To perform the test, Kendall's  $S$  statistics, denoted  $S_i$ , are computed for each season and these are then summed to form the overall statistic  $S_k$  (eq. 12.5).

$$S_k = \sum_{i=1}^m S_i \quad (12.5)$$

When the product of the 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 (eq. 12.6):

$$\sigma_{S_k}^2 = \sum_{i=1}^m \sigma_{S_i}^2 = \sum_{i=1}^m \frac{n_i \cdot (n_i - 1) \cdot (2n_i + 5)}{18}, \quad (12.6)$$

where  $n_i$  = number of years of data in season  $i$ . Note that the formula for the variance of  $S_i$  is exactly the same as the formula for the ordinary  $S$  statistic in the original Mann-Kendall test for trend.

$S_k$  is standardized (eq. 12.7) by subtracting its expectation and adding in a continuity correction and dividing by its standard deviation,  $\sigma_{S_k}$ , all under the assumption that there is no correlation among the seasons, and no ties and no cases of multiple values in a given season. Also note that the value of  $\sigma_{S_k}^2$  depends only on the number of seasons and the number of years of observations in each season, and not on the data values themselves. The resulting standardized value,  $Z_{S_k}$ , of the Seasonal Kendall test statistic  $S_k$  is evaluated against a table of the standard normal distribution.

$$Z_{S_k} = \begin{cases} \frac{S_k - 1}{\sigma_{S_k}} & \text{if } S_k > 0 \\ 0 & \text{if } S_k = 0 \\ \frac{S_k + 1}{\sigma_{S_k}} & \text{if } S_k < 0 \end{cases} \quad (12.7)$$

The null hypothesis (no trend) is rejected at significance level  $\alpha$  if  $|Z_{S_k}| > Z_{crit}$  where  $Z_{crit}$  is the value of the standard normal distribution with a probability of exceedance of  $\alpha/2$ . When some of the  $Y$  or  $T$  values are tied, the formula for  $\sigma_{S_k}$  must be modified as discussed in chapter 8. For datasets of about 10 years or longer, the variance ( $\sigma_{S_k}^2$ ) can be modified to account for serial correlation (Hirsch and Slack, 1984). The Seasonal Kendall Test, with and without this adjustment for serial correlation, is implemented in the `rkt` (Marchetto, 2017) package in R, in the function `rkt`, and also in the `EnvStats` (Millard, 2013) package in the function `kendallSeasonalTrendTest`.

If there is variation in sampling frequency during the years of interest (meaning the sample sizes differ across the seasons), the dataset used in the trend test may need to be modified. If the variations 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 there happen to be no values in a particular season of a particular year, then there would be no value used for that season of that year. If, however, there is a systematic variation in the sampling frequency (for example, monthly for 7 years followed by quarterly for 5 years) then a different approach is necessary. When there is a systematic variation in sampling frequency, define the seasons on the basis of the lowest sampling frequency. For the 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. If the sampling frequency is reasonably consistent and there are generally two or more samples taken in each season, then the median of the available data should be used for each season of each year. The `rkt` function allows the user to select this approach.

The trend slope can be computed in a manner that is compatible with the approach used in the test. It is based on the Theil-Sen slope estimator, but only uses the pairwise comparisons within a given season. The slope estimate is the median of all of these within-season slopes. This statistic is also computed by the `rkt` function.

When there is an important exogenous variable ( $X$ , as discussed in section 12.3) the Seasonal Kendall test can also be used. One option is to perform the Seasonal Kendall test on residuals from a loess model of  $Y$  as a function of  $X$ . Another option is to perform the Seasonal Kendall test on residuals from an OLS regression of  $Y$  on  $X$ . The choice between the two should be based on how well the underlying assumptions

of linear regression are met. If they are met reasonably well, this mixed approach of the Seasonal Kendall Test on residuals from linear regression would be an appropriate method.

### 12.4.2 Mixed Method—OLS Regression on Deseasonalized Data

Another possible approach is to deseasonalize the data by subtracting seasonal medians from all data within the season, and then doing OLS regression of 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 preferred over the alternatives.

When seasonal medians are subtracted, this is equivalent to using dummy variables for  $m-1$  seasons in a fully parametric regression. This approach causes the loss of  $m-1$  degrees of freedom in computing the seasonal statistics, a disadvantage which can be avoided by using the fully parametric method introduced in the next section. However, one drawback to the fully parametric approach is that it makes a fairly restrictive assumption about the shape of the seasonal pattern (a sine wave). In some cases, hydrologic data may strongly depart from that characteristic, for example having abrupt changes between the growing season and the nongrowing season, or having distinct regular shifts from dry season to wet season. Where these more abrupt changes exist, the mixed approach may have merit.

### 12.4.3 Fully Parametric Model—Multiple Regression with Periodic Functions

The third option for analysis of trends in seasonal data is to use periodic functions to describe seasonal variation. The simplest approach, and one that is sufficient for many purposes, takes the form of equation 12.8.

$$Y = \beta_0 + \beta_1 \cdot \sin(2\pi T) + \beta_2 \cdot \cos(2\pi T) + \beta_3 \cdot T + \beta_4 \cdot X + \varepsilon \quad (12.8)$$

where  $T$  is time in years and  $X$  is an exogenous explanatory variable such as discharge, precipitation, or level of some human activity (for example, waste discharge, basin population, or production). The  $X$  variable may be continuous or binary dummy variables as in analysis of covariance (for example, before or after the dam was removed, or before or after the treatment plant upgrade). The trend test is conducted by determining if the slope coefficient on  $T(\beta_3)$  is significantly different from zero. Other terms in the equation should be significant and appropriately modeled (the standard assumptions for multiple linear regression described in chap. 11). The residuals,  $\varepsilon$ , must be approximately normal.

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,  $M$ , at which the peak occurs. The sum of the sine and cosine terms can be re-expressed this way.

$$\beta_1 \cdot \sin(2\pi t) + \beta_2 \cdot \cos(2\pi t) = A \cdot \sin[2\pi(t + t_0)] \quad (12.9)$$

where  $A = \sqrt{\beta_1^2 + \beta_2^2}$ , and  $t_0$  is the phase shift in years (the point in time when the sine wave crosses zero and has a positive slope is at time of  $-t_0$  years).

Let  $M$  denote the day of the year when the function reaches its maximum. It can be determined as follows

$$\text{if } \beta_1 > 0 \quad M = \frac{365.25 \cdot \left[ \frac{\pi}{2} - \tan^{-1}(\beta_2 / \beta_1) \right]}{2\pi} \quad (12.10)$$

$$\text{if } \beta_1 < 0 \quad M = \frac{365.25 \cdot \left[ -\frac{\pi}{2} - \tan^{-1}(\beta_2 / \beta_1) \right]}{2\pi} + 365.25 \quad (12.11)$$

There are three special cases,

$$\text{if } \beta_1 = 0 \text{ and } \beta_2 > 0 \text{ then } M = 365.25 \text{ or } 0$$

$$\text{if } \beta_1 = 0 \text{ and } \beta_2 < 0 \text{ then } M = 182.625$$

$$\text{if } \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ then } M = \text{undefined}$$

(when  $M$  is undefined that means there is no annual sine wave). An R function, called **MaxDay** is included in the supplementary material (SM.12) for this chapter. It determines, given a pair of  $\beta_1$  and  $\beta_2$  values, the day of the year at which this sine wave attains its maximum value. Note that the same sort of trigonometric formulation can be used for other possible periodicities that might arise in water resources. The most common of these being a diurnal cycle (24 hours) and also possibly a weekly cycle (where the variable of interest is affected by human activities focused around the 7-day week).

After including sine and cosine terms in a multiple regression to account for seasonality, the residuals may still show a seasonal pattern. If this occurs, additional periodic functions with periods of half a year or some 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 in building more complex functions. For example, to use two waves, one with a period of a year and the other with a period of a half a year the following equation would be appropriate:

$$Y = \beta_0 + \beta_1 \cdot \sin(2\pi t) + \beta_2 \cdot \cos(2\pi t) + \beta_3 \cdot \sin(4\pi t) + \beta_4 \cdot \cos(4\pi t) + \varepsilon \quad (12.12)$$

One way to determine how many periodic seasonal 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 legitimately settle on a model in which the  $t$ -statistics for one of the two coefficients in the pair is not significant. What matters is that 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. There are also cases where the seasonal pattern may be described by a functional form that is not a simple trigonometric function. For example, see Vecchia and others (2008) for an approach used with pesticide data, which has a very specific temporal pattern throughout the year based on the timing of pesticide application.

#### 12.4.4 Comparison of Methods for Dealing with Seasonality

The Seasonal Kendall test and mixed approaches have the disadvantages of only being applicable to univariate data (either the original data or residuals from a previous analysis) and are not amenable to simultaneous analysis of multiple sources of variation. For this reason, these methods take at least two steps to compute. Multiple regression allows many variables to be considered easily and simultaneously by a single model.

The Seasonal Kendall test has the usual advantage of nonparametrics: robustness against departures from normality. The OLS regression of deseasonalized  $Y$  on  $T$  is perhaps the least robust because the individual seasonal datasets 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.

**Table 12.5.** Methods for characterizing seasonal patterns.

Rating	Graphical methods	Tabular methods
Best	Boxplot by season, or loess of data versus time of year	List the amplitude and peak day of cycle
Next best	None	List the 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 the seasonal means, standard deviations, or standard errors

For example, it is unlikely that April will be very different from May, even though the sample statistics may suggest that this is so.

Multiple regression with periodic functions involves very few parameters. However, the functional form (sine and cosine terms) can become a straightjacket, constraining the seasonal pattern to a single form. Perhaps the annual cycle really does have abrupt breaks associated with freezing and thawing, or the growing season. Multiple regression can always use binary variables to distinguish the 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 have abrupt changes in slope, but doing this would require a good, physically based definition of the timing of the influential factors.

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 dataset (table 12.5). Any of the methods can be applied to the raw data or to residuals from a loess or OLS 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.

## 12.4.6 Seasonal Differences in Trend Magnitude

The approaches described in previous sections assume a single pattern of trend across all seasons. For example, if we are testing for trends in the mean monthly discharge, these approaches assume that the trend slope or magnitude is identical for every month. The reality is that trend magnitudes can differ greatly across seasons (for example, winter season discharge might have increased substantially but summer might not have changed or might even have decreased). None of the test statistics described above will provide any clue of these differences. This may be a gross over-simplification and can fail to reveal large differences in the trends across the different seasons. In fact, it is entirely possible that the  $Y$  variable exhibits a strong positive trend in the spring and summer, but a strong negative trend in the fall and winter. These changes might cancel each other out, resulting in an overall Seasonal-Kendall test statistic indicating little or no trend or slope for the time term in a multiple regression approach that suggests that there is no trend. This is not to suggest that single test statistics are never useful with seasonal data; many times we desire a single number to characterize what is happening at a site over the course of an entire year. Yet, when a more detailed examination of trends at an individual site is needed, it is often useful to perform and present the full, within-season analysis on each season. A good approach to graphically presenting the results of such multi-season analyses is seen in figure 12.10. The graph is the result of individual Mann-Kendall trend tests on the monthly discharges for the Sugar River near Brodhead, Wisconsin, for a period of 62 years. The trends in all months (expressed as  $m^3/s/yr$ ) are all positive but vary over a range from a low of 0.04 in

March to a high of 0.16 in June. The significance level of these trends also varies greatly. Several of the months have  $p$ -values substantially less than 0.01, but for the month of March it is around 0.5 (not even close to being significant).

In the approaches using the Seasonal-Kendall test, one can also examine contrasts between the different seasonal statistics. Contrasting these results provides a single statistic that 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  $i$  ( $i=1, 1, 2, \dots, m$ ) 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:

$$\chi^2_{(total)} = \sum_{i=1}^m Z_i^2 \quad (12.13)$$

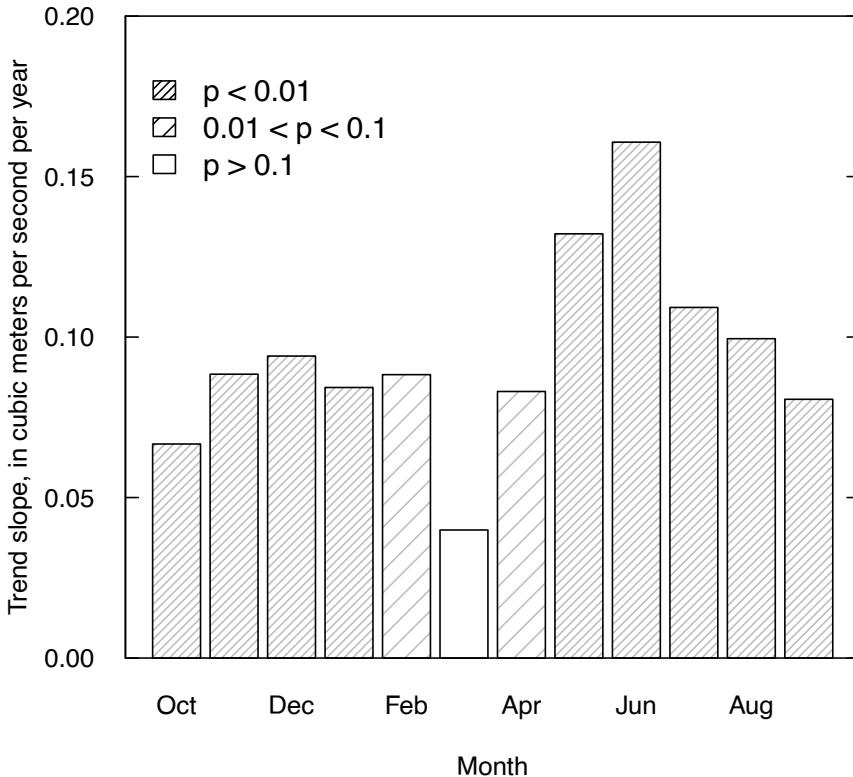
$$\chi^2_{(trend)} = m \cdot \bar{Z}^2 \quad (12.14)$$

where

$$\bar{Z} = \frac{\sum_{i=1}^m Z_i}{m}$$

$$\chi^2_{(homogeneous)} = \chi^2_{(total)} - \chi^2_{(trend)} \quad (12.15)$$

The null hypothesis, that the seasons are homogeneous with respect to trend ( $\tau_1 = \tau_2 = \dots = \tau_m$ ), is tested for homogeneity of trend 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  $\alpha$ , reject the null hypothesis and conclude that different seasons exhibit different trends. For the dataset considered in figure 12.10, the



**Figure 12.10.** Graph of monthly trends in discharge, Sugar River near Brodhead, Wisconsin, for water years 1952–2016. Bar heights show the Theil-Sen slope, by month, for each of the 12 months. Shading indicates the  $p$ -value for the Mann-Kendall trend test for the month.

Seasonal Kendall test, which considers all 12 months, gives an overall slope of  $0.093 \text{ m}^3/\text{s}/\text{yr}$  and a  $p$ -value of  $<0.001$ . The `EnvStats` (Millard, 2013) function `kendallSeasonalTrendTest` also considers the contrasts of the trends in the 12 months and does not show the trends to be heterogeneous, which means that we should not reject the null hypothesis that the trend is the same in all 12 months. The test itself will not indicate which months have different trends and which are similar, but the graphics can help sort out these differences.

### 12.4.7 The Regional Kendall Test

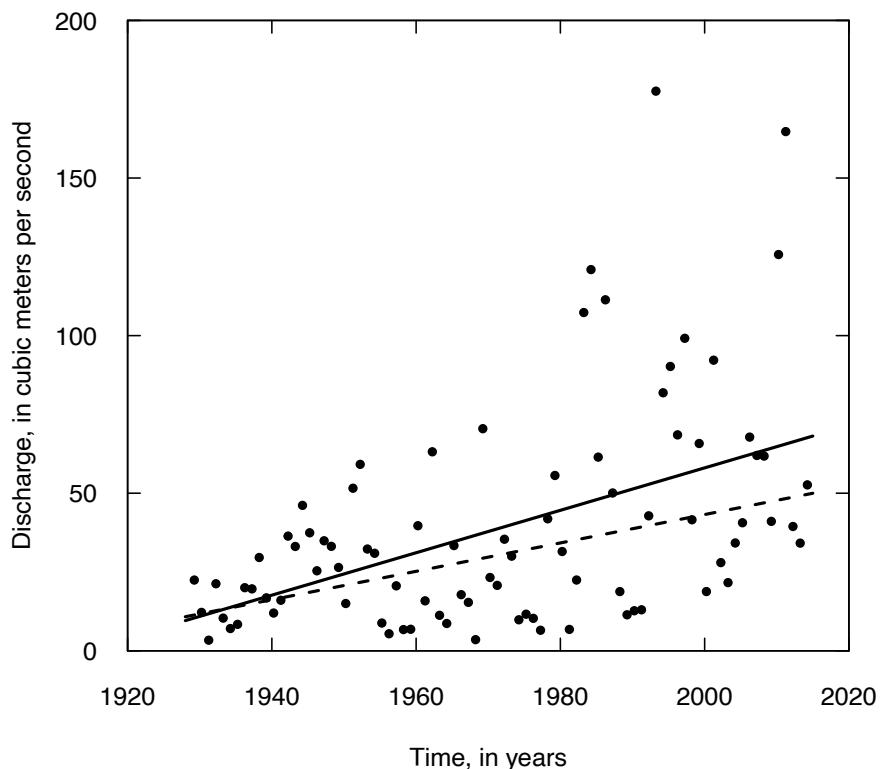
Another variation on the Mann-Kendall test for trend is the Regional Kendall test, which was introduced by Helsel and Frans (2006). This test is intended for use when a set of Mann-Kendall trend tests are applied to data from a set of monitoring locations that are near enough to each other that one may expect their data to be correlated across the sites. It operates in the same manner as the Seasonal Kendall test adjusted for serial correlation (mentioned in section 12.4.1.). It computes the  $S$  statistic for each individual site and sums them to form an overall test statistic for the network of sites, with the variance of this test statistic modified to account for the cross correlations between all possible pairs of sites. These cross correlations are each computed using the Kendall  $\tau$  correlation.

Suppose we are considering the question of trends in annual mean discharges for a set of streamgages that are located very close together, so close that they generally respond to the same precipitation events. In such a situation, the cross correlation between the pairs of sites is very high (close to 1). Given this situation we can expect that the Mann-Kendall trend test  $S$  statistic for each site will be very similar to the others. If the overall test statistic were not adjusted for this cross correlation, we might conclude that the regional trend is highly significant because so many of the sites in the region show significance. However, the reality is that the trend tests results are highly dependent on each other, so the overall strength of the evidence for a trend for the region may be no higher than that provided by the evidence at any individual site. Examples of use are Clow (2010), Garmo and others (2014), and Archfield and others (2016). The Regional Kendall test is implemented in R in the `rkt` package (Marchetto, 2017) by the `rkt` function. In the `EnvStats` package (Millard, 2013) it can be implemented using the `kendallSeasonalTrendTest`, but instead of indexing the data by season the data must be indexed by site.

## 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 several 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 such as the Mann-Kendall or Seasonal Kendall 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 original 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 the significance levels of a parametric test, 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 that slope is estimated by OLS regression or as a Theil-Sen slope estimate. It is quite possible that negative predictions may result for some values of time or  $X$ . One can transform the data so that the trend becomes approximately linear, and a trend slope can be estimated by OLS regression or Kendall-Theil slope. This slope can then be re-expressed in the original units. The resulting nonlinear trend will better fit the data than the linear expression. When doing a trend analysis that includes an exogenous variable it may be appropriate to run analyses using transformations of the  $X$  values. For example, in a test of trend in a solute concentration ( $Y$ ) we might want to use the logarithm of discharge as our  $X$  value.



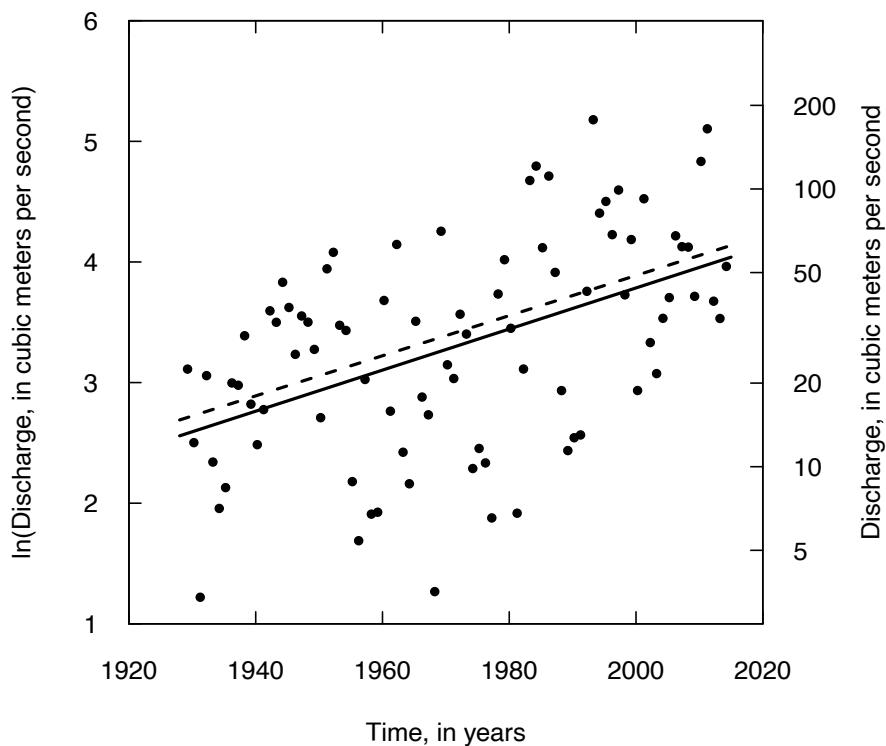
**Figure 12.11.** Graph showing trend in annual mean discharge, Big Sioux River at Akron, Iowa. The solid line is from an ordinary least squares regression trend analysis and the dashed line is the Theil-Sen robust line (associated with the Mann-Kendall test).

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. If we use a natural log transformation, then the linear trend we estimated translates to an exponential trend in original units. This exponential trend 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  $(e^{b_1} - 1) \cdot 100$ . If a slope value in the 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 in original units.

#### Example 12.2. Trends in annual mean discharge, Big Sioux River, South Dakota and Iowa.

The Big Sioux River watershed is located in eastern South Dakota and northwestern Iowa has shown remarkable trends in streamflow over the 87-year period from 1928 to 2014. If an analysis is conducted on the untransformed data, the resulting estimated trends are shown in figure 12.11. The trend is highly significant according to both OLS regression and the Mann-Kendall test. The slope of the linear regression estimate is substantially steeper than the Theil-Sen line ( $0.67 \text{ m}^3/\text{s}/\text{yr}$  versus  $0.45 \text{ m}^3/\text{s}/\text{yr}$ ), because the regression is strongly influenced by some of the highest values in the later part of the record. Note that the data do not conform well to the assumption of constant variance (homoscedasticity) that is required for either OLS regression or the Mann-Kendall test. It is safe to say that because of the strong departure from homoscedasticity, the estimates of uncertainty inherent in each approach will not be accurate and as a consequence  $p$ -values will not be accurate.

The alternative formulation of this problem, evaluating the natural log of the discharge, is shown in figure 12.12. The figure shows that both the OLS and Mann-Kendall methods produce virtually identical slopes (about 0.017 per year) and that the data now conform much more closely to the assumption of



**Figure 12.12.** Graph of trend in the natural log of annual mean discharge, Big Sioux River at Akron, Iowa. The solid line is from an ordinary least squares regression trend analysis and the dashed line is from a Mann-Kendall trend analysis (line represents the Theil-Sen robust line).

homoscedastic errors and the residuals also appear to be more nearly normal. The significance levels for the tests are again both highly significant, noting that the test statistic for the Mann-Kendall test is exactly the same as in the previous test on the untransformed data.

Regardless of the slope estimation method used with the log-transformed data, we can express these results as indicating a rate of increase of about 1.7 percent per year. The percent per year increase is calculated as  $100 \cdot (e^B - 1)$ , where  $B$  is the slope computed for natural log discharge as a function of year. It is also possible to express the trend in terms of longer time periods. For example, it could be described as about 18.5 percent per decade, which is  $100 \cdot (e^{10B} - 1)$ , or as 331 percent over entire record (computed using the last year minus the first year, in this case 86 years), which is  $100 \cdot (e^{86B} - 1)$ . The exponent in this formula is always the product of the slope and the duration of the trend analysis.

In general, when using parametric approaches, 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 to stabilize the variance and avoid having estimates that are negative when negative values are actually impossible. 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 flux (concentration multiplied by discharge) for virtually any constituent. One potential issue with transformations is the case where some of the observations are zero values. Many transformations, including the logarithm, cannot be calculated in these cases. This may rule out the use of transformations (especially in low flow studies) but the nonparametric tests will work in situations where the variable of interest goes to zero. Censored estimation techniques can be used if one

treats the zero values as less than some very small value. However, setting this censoring threshold can be very arbitrary and its choice might have considerable influence on the result.

Some argue that data should always be transformed to ensure normality, and then parametric procedures computed on the transformed data. Transformations to enforce normality are not always possible, as some data are non-normal not as a result of skewness, but of the heavy tails of the distribution (Schertz and Hirsch, 1985). Transformations can make a dataset suitable for OLS regression methods (because they may cause the trends to be more nearly linear and have more nearly constant variance) and regression methods allow for 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) that at least allow removal of multiple exogenous effects in one step. The `loess` function in R allows for such multivariate smoothing.

A difficult situation arises when one is undertaking the analysis of multiple datasets (for example, the same variable evaluated for trend at multiple sites). 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. The transformation appropriate to one dataset may not be appropriate to another. If different transformations are used on different datasets then comparisons among results are difficult, if not impossible. Also, there is an element of subjectivity in the choice of transformation. The argument of the skeptic—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 datasets in a study, and this is next to impossible to accomplish with the several judgements of model adequacy required for parametric methods. Nonparametric procedures are therefore well suited to multirecord trend analysis studies because they generally don't require transformations in order for them to work well. In contrast, for stand-alone analyses of individual records, the use of multiple regression methods in which the dependent variable is transformed can be very appropriate and may offer a richer description of the nature of the trend.

## 12.6 Monotonic Trend Versus Step Trend

The previous sections of this chapter discuss tests for monotonic trends, trends that are assumed to be gradual and don't change direction over time. The Mann-Kendall test and OLS regression are the two basic tests for this purpose. Another class of tests compares two nonoverlapping 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 to those described in other chapters, including the rank-sum test, two-sample  $t$ -tests, and analysis of covariance. Each of these tests also can be modified to account for seasonality.

### 12.6.1 When to Use a Step-trend Approach

Step-trend procedures should be used in two situations. The first is when the record being analyzed is 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 over monotonic tests for trend. A general rule is 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. In general, if the within-period trends are small in comparison to the between-period differences, then step-trend procedures should be used.

The second situation to test for step trend is when a known event has occurred at a specific time during the record that is likely to have led to a change in the distribution of the variable of interest. 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 owing to completion of treatment plant improvements, or the closing of some facility. It is also possible that some natural event, such as a large flood flow that changed the channel and floodplain configuration, could be considered in a step trend analysis. The question posed might be related to some aspect of how the river behaved before the extreme event versus how it behaved after the extreme event. In some cases, the known event is not instantaneous but its implementation period is brief in comparison

**Table 12.6.** Step-trend tests (two sample) that do not consider seasonality. In the case of analysis of covariance, group is represented by a dummy variable with value of 0 in the before group and 1 in the after group.

[-, not applicable]

Type of trend test	Not adjusted for $X$	Adjusted for $X$
Nonparametric	Rank-sum test on $Y$	Rank-sum test on residuals from loess of $Y$ on $X$
Mixed	-	Rank-sum test on residuals from regression of $Y$ on $X$
Parametric	$t$ -test	Analysis of covariance of $Y$ on $X$ and group

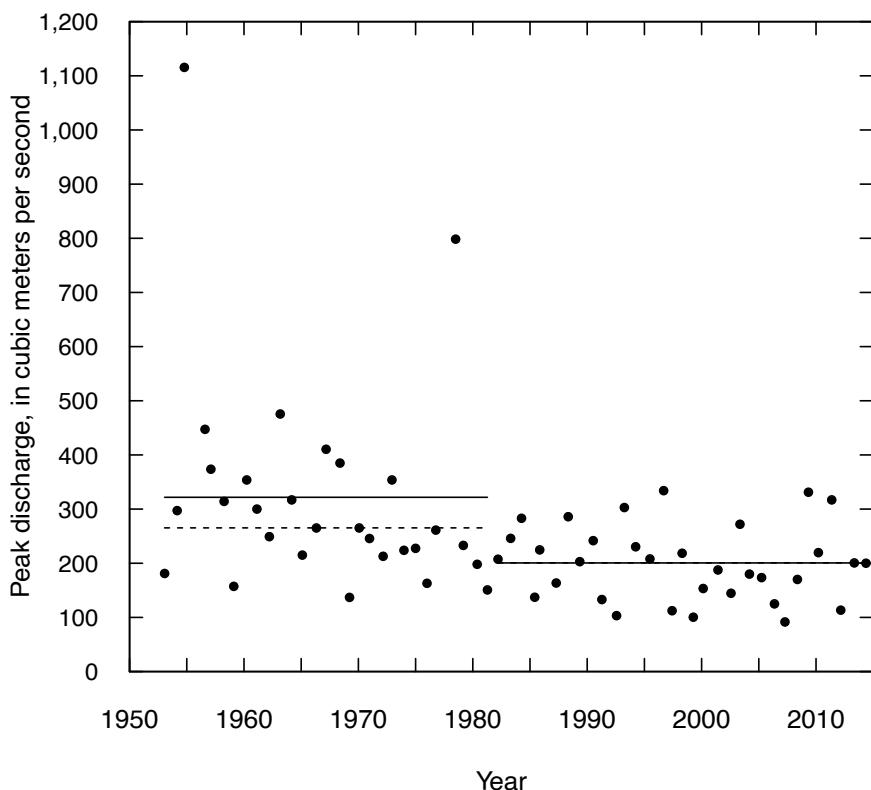
**Table 12.7.** Step-trend tests (two sample) that consider seasonality. In the case of analysis of covariance, group is represented by a dummy variable with value of 0 in the before group and 1 in the after group.

Type of trend test	Not adjusted for $X$	Adjusted for $X$
Nonparametric	Seasonal rank-sum test on $Y$	Seasonal rank-sum test on residuals from loess 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

to the total length of the record being analyzed. In such cases there may be a buffer period between the before and after periods that might be simply left out of the analysis. It is imperative that the decision to use step-trend procedures not be based on examination of the data (in other words, 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. Step-trend procedures require a highly specific situation, and the decision to use them should be made prior to any examination of the data (except for determining where large data gaps exist). If there is no prior hypothesis of a step 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 to use because comparable periods of time are more easily examined among all the records.

## 12.6.2 Step-trend Computation Methods

Step-trend approaches that do not consider seasonality are summarized in table 12.6 and those that consider seasonality are summarized in table 12.7. 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 that are non-normal. Those disadvantages include 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 (Hirsch, 1988). The H-L estimator is the median of all possible differences between data in the before and after periods. The rank-sum test can be implemented in a seasonal manner just like



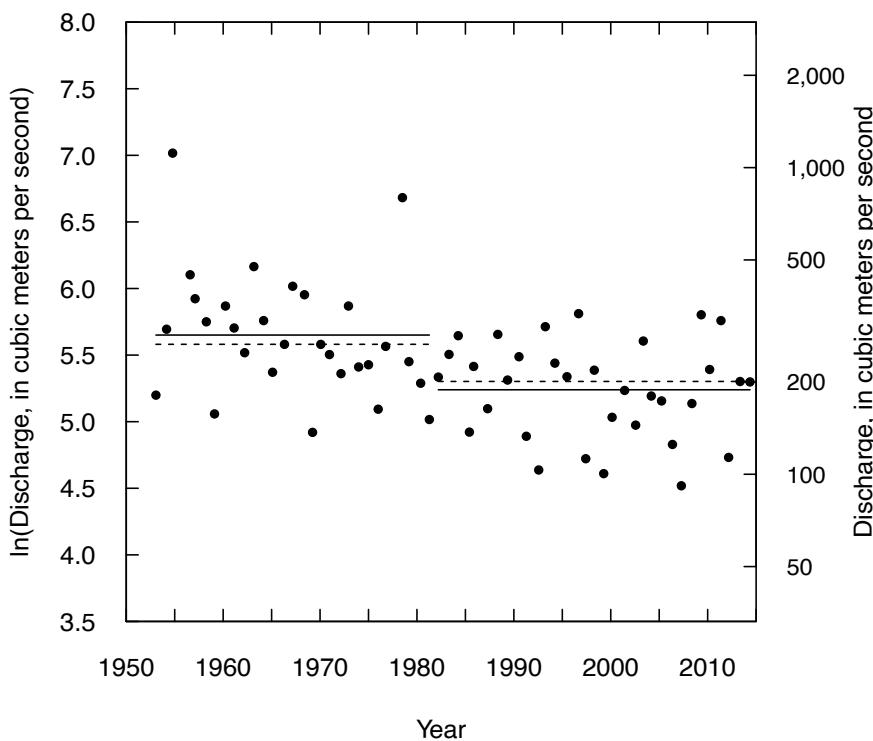
**Figure 12.13.** Graph of annual peak discharge, North Branch Potomac River at Luke, Maryland. The solid lines represent the mean value for the periods before and after the completion of Jennings Randolph reservoir (9 miles upstream). The dashed lines represent the median value for the periods before and after. Note that for the after period the mean and median are virtually identical so that the two lines overprint each other.

the Mann-Kendall test and is called the seasonal rank-sum test. The test 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.

**Example 12.3. Step trend in annual peak discharge, North Branch Potomac River at Luke, Maryland.**

This example dataset consists of the annual peak discharge values for the North Branch Potomac River, at Luke, Maryland. The streamgage is located 9 miles downstream from the Jennings Randolph reservoir (U.S. Army Corps of Engineers), which was completed in 1981. The question is, has the distribution of annual peak discharge changed from the period before the dam was completed to the period since it was completed? The dataset is shown in figure 12.13.

The two-sample  $t$ -test (run by using the `lm` function in R, with a dummy variable set to 0 for 1952–81 and set to 1 for 1982–2014) indicates a step trend of  $-121.2 \text{ m}^3/\text{s}$  and  $p=0.0019$ . Note that the pre-1982 data are highly skewed, which has a strong influence on the pre-1982 mean. It also results in a high variance for the pre-1982 mean value, but in spite of that, the  $p$ -value conveys a very strong indication of a trend. The Mann-Whitney-Wilcoxon Rank Sum (MWWSRS) test also indicates a trend,  $p=0.00072$ . The estimated size of the shift is a decrease of  $79.3 \text{ m}^3/\text{s}$ . It is interesting that the nonparametric test results in a lower  $p$ -value, but the magnitude of the estimated trend is smaller than for the parametric approach. This is



**Figure 12.14.** Graph of the natural log of annual peak discharge, North Branch Potomac River at Luke, Maryland. The solid lines represent the mean value for the periods before and after the completion of Jennings Randolph reservoir (9 miles upstream). The dashed lines represent the median value for the periods before and after.

because in the MWWR test, the difference is determined by the difference in medians and these are closer together because in the pre-dam period the median is quite a bit below the mean (as a result of the two high outliers), whereas in the post-dam period the median and mean are virtually identical.

We can take natural logs of the data and re-run the analyses. The transformed data are shown in figure 12.14. The  $p$ -value for the  $t$ -test has changed from the previous test on the real data. It suggests an even higher level of evidence that the two periods are different. The  $p$ -value is 0.00026, compared to the previous value of 0.0019. The  $p$ -value for the MWWR test is unchanged by the transformation and remains a value of 0.00072. The estimated magnitude of the shift is slightly larger for the  $t$ -test approach (a change of  $-0.411$ ) than for the nonparametric approach (a change of  $-0.363$ ). These step changes can be turned into a more meaningful set of numbers. In the case of the  $t$ -test the step size is  $-33.7$  percent,  $100 \cdot (\exp(-0.411) - 1)$ . Using the nonparametric approach the step size is  $-30.5$  percent. The similarity of these two results is explained by the fact that the log data, after removing the step, are approximately normally distributed.

### 12.6.3 Identification of the Timing of a Step Change

A common error in the use of step-trend approaches is that the analyst looks at a dataset and notes what appears to be a large shift that happens in a particular year, then based on this observed shift, divides the data into pre- and post-shift groups and reports a significance level for the shift. A  $p$ -value computed this way is always biased because it was determined based on selecting the point in time that will result in a highly significant trend. There is a nonparametric test that determines the change point and computes a significance of the change, it is called the Pettitt test (Pettitt, 1979). In the Pettitt test the  $p$ -value is computed in a manner that adjusts for the fact that the method is designed to find the most advantageous point in the record to consider as the change point.

## 12.7 Applicability of Trend Tests with Censored Data

Recall that censored samples are records in which some of the data are known only to be less than or greater than some threshold. The existence of censored values complicates the use of parametric trend tests and trend procedures that involve removing the effects of an exogenous variable. They also complicate the use of nonparametric tests when there is more than one reporting limit. For the parametric methods of trend testing one option that is often used, but which is highly inappropriate, is to replace the censored value with some arbitrary value such as zero, the reporting limit, or half the reporting limit. These approaches will give inaccurate results for hypothesis tests and biased estimates of trend slopes (see Helsel [1990, 2012]). A better 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). In Tobit estimation, the user specifies a model that looks like a conventional multiple regression model, but some of the data used to fit the model are specified to be censored. Because the Tobit method assumes a linear model with normally distributed errors, transformations (such as logarithms) of  $Y$  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 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 whereas a more recent part of the record (when floodplain development became more intense and record keeping more complete) may contain estimates of floods exceeding a more moderate threshold. See England and others (2018) for an extensive discussion of the use of multiple thresholds in flood frequency studies.

The Mann-Kendall test can be used without any difficulty when only one censoring threshold exists, although the power of the test declines as the fraction of the data that is censored rises. Comparisons between all pairs of observations are possible. All the less than values are smaller than the other values and are considered to be tied with each other. Thus the  $S$  statistic and  $\tau$  are easily computed using the tie correction for the standard deviation in the large-sample approximation. Equation 12.16 is the formula for the standard deviation of the  $S$  statistic in the face of ties.

$$\sigma_s = \sqrt{\frac{[n(n-1)(2n+5) - \sum_{i=1}^n t_i(i)(i-1)(2i+5)]}{18}} \quad (12.16)$$

where

- $n$  is sample size,
- $i$  is the extent of a tie (for example  $i=2$  represents a tie of just two values and  $i=3$  is a tie that consists of three values), and
- $t_i$  is the number of ties of extent  $i$ .

For example, if we have a dataset of 50 samples and 7 of them are reported as “less than 1” and all others are some value greater than 1 (and none are tied with each other), then equation 12.16 would become

$$\sigma_s = \sqrt{\frac{50 \cdot 49 \cdot 105 - 1 \cdot 7 \cdot 6 \cdot 19}{18}}$$

$$\sigma_s = \sqrt{\frac{257250 - 798}{18}} = 119.362 .$$

**Table 12.8.** Classification of monotonic trend tests for censored data.

Type of trend test	Not adjusted for $X$	Adjusted for $X$
Nonparametric	ATS or Mann-Kendall test for trend on $Y$	No test available
Fully parametric	Tobit regression of $Y$ on $T$	Tobit regression of $Y$ on $X$ and $T$

When more than one detection limit exists, the Akritas-Theil-Sen (ATS) method will perform the equivalent nonparametric trend test and Theil-Sen slope (Helsel, 2012), but description of the test is beyond the scope of this book. A simpler procedure, but not as powerful as ATS, is to perform the Mann-Kendall test after recensoring the data as follows. Consider the dataset <1, <1, 3, <5, 7. A <1 and <5 must be considered tied. Also, a 3 must be considered tied with a <5. The Mann-Kendall test can be computed after recoding to set all data below the highest detection limit as tied with one another. 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. But if a trend is found after recoding, it can be believed. This is not true if substitution is used. Using values such as one-half the detection limit for each less-than value can place a false trend into the dataset, as well as decreasing the power to find a trend that is actually occurring. Recoding data as described above is far preferable to using substituted values such as half of the reporting limit. Better yet, use methods such as ATS that are designed for data with multiple detection limits.

Although the sign of the estimated Theil-Sen trend slope is accurate for highly censored data records, the magnitude of the slope estimate is likely to be in error. Substitution of an arbitrarily chosen value below the reporting limit for all censored values will give biased estimates of the trend slope. Although the amount of bias cannot be stated precisely, the presence of only a few less-than values in a record (less than about five percent) is not likely to affect the accuracy of the trend slope magnitude substantially and an arbitrary substitution of a value such as half the detection limit can be used for the censored data when computing a Theil-Sen robust line. In situations where we want to adjust for an exogenous variable ( $X$ ) and we have censoring, there is no suitable nonparametric procedure.

Monotonic trend tests for datasets with censored data are classified in table 12.8.

## 12.8 More Flexible Approaches to Water Quality Trend Analysis

One feature of all trend tests in this chapter is that they depend on a number of fairly restrictive assumptions about the nature of the trend and relations between the water quality variables and factors such as season or discharge. In this section we will introduce new methods that are aimed at achieving greater flexibility in the underlying assumptions.

### 12.8.1 Methods Using Antecedent Discharge Information

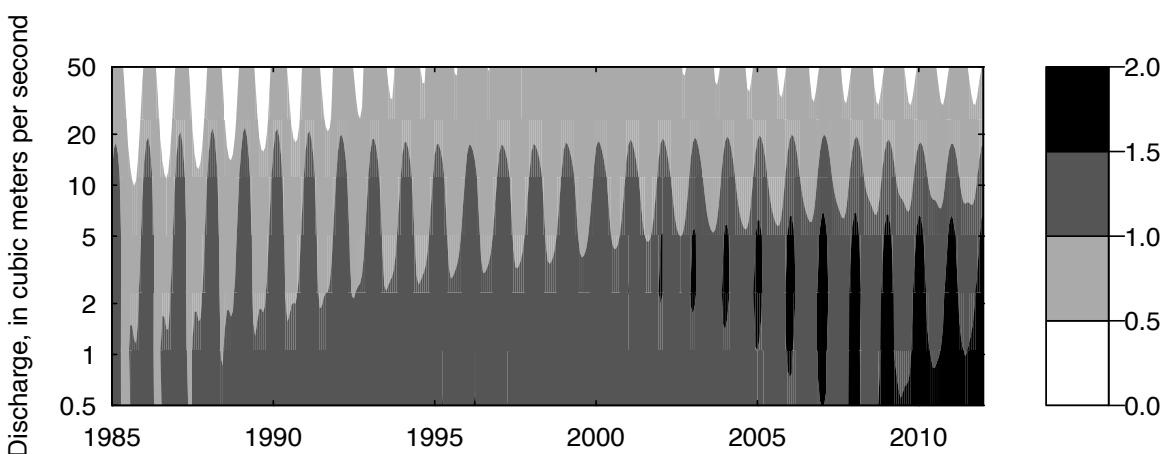
One commonality among most methods for removing or adjusting for the effect of discharge on water quality is that they are all predicated on the assumption that the discharge on the day of sampling is the best metric of discharge to be used. Yet, it is well known that antecedent hydrologic conditions can have a strong influence on water quality. Murphy and others (2014) discuss these phenomena in some detail and their paper provides an extensive set of references to these antecedent effects on water quality. For example, in the case of a nutrient such as nitrogen, if there has been a dry period of many months in duration, then there is likely to be a large amount of nitrogen available because the opportunities for leaching to groundwater, surface runoff, or loss processes such as denitrification have been limited during the dry period. When the watershed does begin to receive normal or above normal inputs of precipitation one can expect that concentrations of nitrogen in rivers will be higher than would be the case for that same amount of discharge following a wet or normal period.

Vecchia (2005) developed and applied a method of trend analysis that explicitly considers these antecedent conditions through the use of multiple regression models with explanatory variables referred to as annual and seasonal discharge anomalies and high frequency variability. These methods have been shown to provide substantially better explanation of concentration variability than is possible in models

that only depend on the discharge on the day of sampling. The idea of using long-term, mid-term, and short-term flow anomalies has been combined with the use of specific seasonal wave functions that describe the seasonal pattern of pesticide transport (a function that is specific to the particular pesticide in a given agricultural region) to create a water quality trend analysis method called SEAWAVE-Q (Ryberg and others, 2014). In general, when the water quality dataset is large enough to support estimation of a large number of explanatory variables, there can be a considerable advantage (in terms of trend detection or estimation of a full daily water quality record) to adding explanatory variables that capture aspects of the history of discharge over some preceding time period such as a year, a season, or a few days. As continuous water quality sensors become more common, it will become possible to use much more complex statistical models.

### 12.8.2 Smoothing Approaches

Another development in water quality trend analysis is the use of smoothing algorithms (similar to loess) to describe the evolving nature of the relation of analyte concentrations to discharge, season, and trend components. Using a smoothing approach requires a fairly large dataset (more than 100 observations) and allows the relation of concentration to the explanatory variables to be free from assumptions about functional forms, allowing a more data-driven description of the changes in water quality than a typical multiple regression approach would require. The WRTDS (weighted regressions on time, discharge, and season) method (Hirsch and others, 2010) is one such method. It has been documented and is available within the EGRET (Explorations and Graphics for RivEr Trends) R package (Hirsch and De Cicco, 2015). The WRTDS method was designed as an exploratory data analysis tool that relaxes many of the assumptions that are common to the trend methods that are described in this chapter. In particular, it does not prescribe one shape of the concentration versus discharge relation, and it does not prescribe that the trends in these relations be the same in different seasons of the year. As a smoothing method, it will cause a very abrupt change (such as an upgrade of a point source that is a major contributor to pollution at the monitoring location) to appear more gradual than it really was, but it is highly appropriate to the case where the changes in pollution are the aggregate of many small changes taking place over the entire watershed. WRTDS does not require an assumption of constant variance over the entire range of years, seasons, and discharge values. An example of one of the many types of graphical outputs that are created by the WRTDS method is shown in figure 12.15. An additional feature of the WRTDS method is that it can describe annual or seasonal time series of flow-normalized concentrations and flow-normalized fluxes. Flow-normalization is a method that removes the influence of year to year variation in discharge to get at the underlying trends as seen in the contour plot, by integrating the relation shown there over the seasonally specific probability density function of discharge. WRTDS also explicitly considers the possibility that trends may not follow a prescribed functional form (such as linear or quadratic in time). Thus, it can describe trends that are



**Figure 12.15.** Graph of contoured surface describing the relation between the expected value of filtered nitrate plus nitrite concentration as a function of time and discharge for the Choptank River near Greensboro, Maryland. Scale at right of figure is concentration in milligrams per liter.

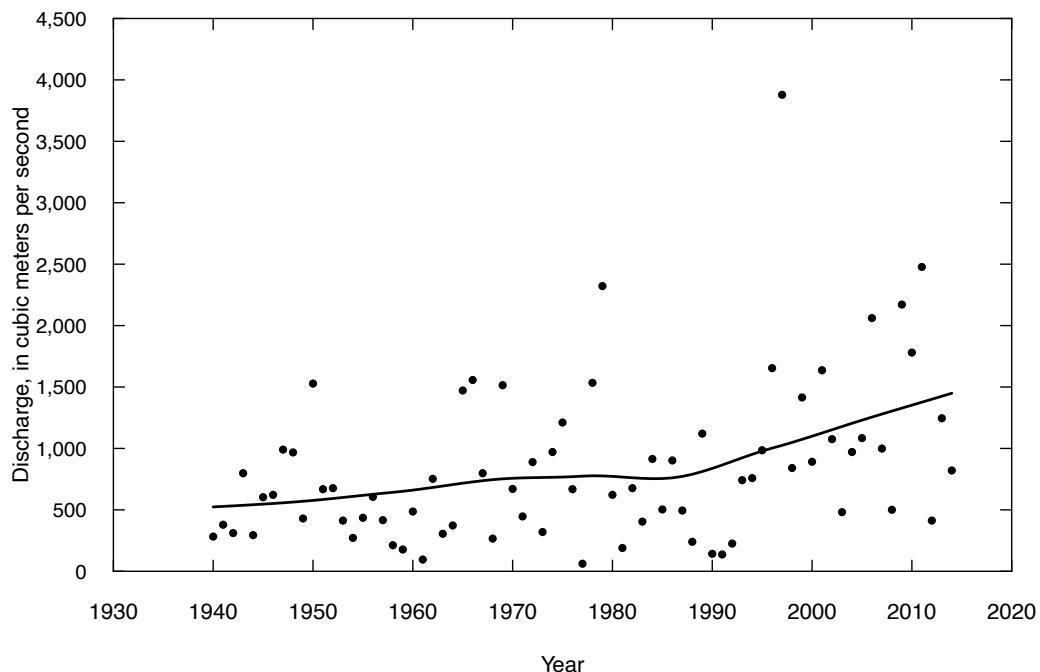
nonmonotonic without having to resort to the arbitrary choice of a functional form for that trend, such as quadratic or cubic. In addition, in WRTDS results for some specified time interval can have percentage changes in concentration that are quite different from the percentage changes in flux over the same time period. For example, in the case shown in figure 12.15, the change in flow-normalized concentration from 1985 to 2012 is estimated to be +40 percent, but the change in flow-normalized flux over this same period is only +27 percent. The discrepancy arises because the strongest trends are happening at the low to moderate discharges (which reflects changes in groundwater quality) but the trends are smaller in the higher discharges (reflecting little change in wash-off at high flow conditions).

More complex approaches to trend analysis have been developed based on a combination of the ideas presented by Vecchia (for inclusion of antecedent conditions) and by Hirsch (for the use of a flexible smoothing model). An example of such a hybrid approach is described by Zhang and Ball (2017).

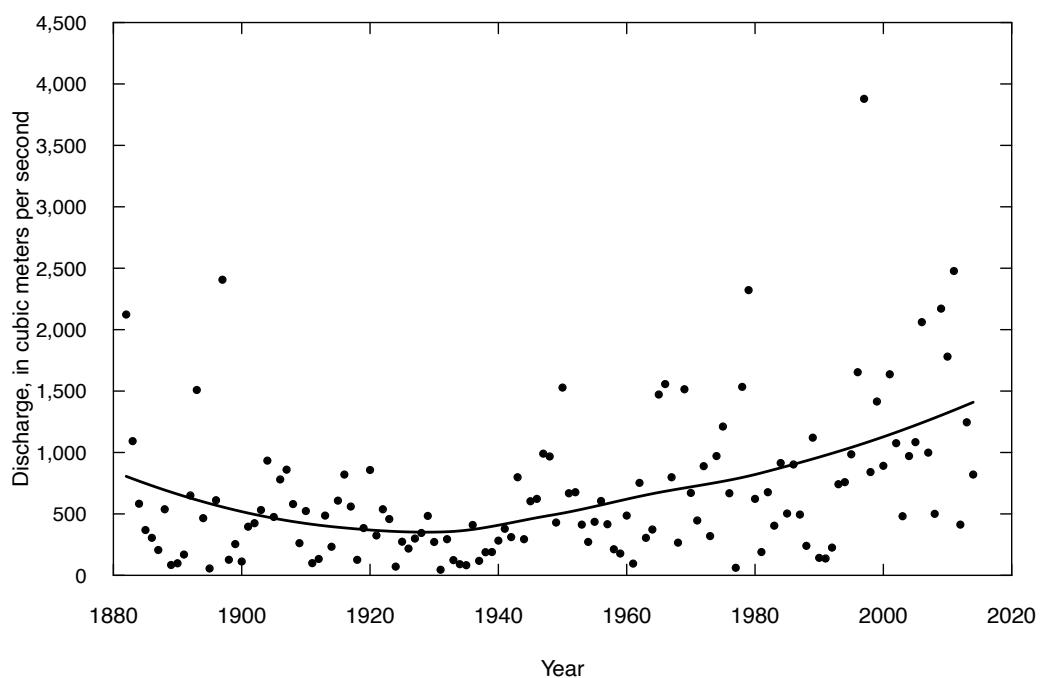
## 12.9 Discriminating Between Long-term Trends and Long-term Persistence

One of the most vexing problems in the analysis of trends in streamflow, groundwater levels, or water quality is that it can be very difficult to distinguish between deterministic trends and long-term persistence. A deterministic trend in hydrology is a change in the central tendency of a process that comes about because of human activity in the watershed (for example, increasing water consumption, increasing urbanization, decreasing point source loadings owing to improved waste treatment). Long term persistence is a change in the central tendency of a hydrologic variable that is a result of the chaotic behavior of the land-atmosphere-ocean system, that often manifests itself as quasi-periodic oscillation (an oscillation that may have a characteristic length but does not have a regular and predictable periodicity) that exist across a wide range of time scales from years, to centuries, or millennia. Long-term persistence (LTP) has been recognized for many decades (see Hurst, 1951; Mandelbrot and Wallis, 1968; Klemeš, 1974; Koutsoyiannis, 2002); the literature shows that hydrologic time series exhibit persistence at all time scales and that this persistence goes well beyond what can be modeled by simple auto-regressive moving-average (ARMA) processes (Box and others, 2015). The annual mean discharge data for the Big Sioux River shown in figure 12.11 provides an excellent example of a highly persistent hydrologic time series. There is no obvious cause for this nearly century-long increase in streamflow that can be related to water or land management actions. Thus, we cannot consider it to be a deterministic trend. Rather, it appears to be driven by a particular combination of climate variations at large temporal and spatial scale, along with the properties of the regional groundwater system. Long-term persistence (as well as short term serial correlation) will inflate the variance of a trend-test statistic. This means that for a stationary, but highly persistent process, the variance of the trend-test statistic will be larger than what it would be under the standard assumptions that the random variable is independent and identically distributed. Thus, type I errors have a higher probability than the nominal type I error rate for the trend test being used. Although this text does not address these issues in any detail, it is nonetheless an important concept when thinking about the statistics of water resources and has been the subject of extensive discussion in the literature. Cohn and Lins (2005) provide an excellent overview of the difficulties encountered in attempting to distinguish between long-term persistence and trend. They reference many important papers that deal with these issues and offer some potential solutions in terms of hypothesis testing. Being able to distinguish between deterministic trend and long-term persistence is important because it provides guidance about how to understand what the future will be like. If an apparent trend is largely a deterministic one, then if we can successfully build a simulation model of how the drivers of that deterministic trend influence the variable of interest, then we can base our forecast of the future on forecasts of that driving variable. But if the apparent trend is simply a manifestation of long-term persistence then we have no real ability to forecast the variable of interest because the drivers of the phenomena could just as easily continue at their current levels or reverse themselves tomorrow and take the system in a new direction. Most hydrologic variables are responding to a mix of these two types of drivers, and assessing that mixture is critical to projecting future conditions.

One way that long term persistence affects trend analysis is in the selection of the time period over which a trend is assessed. The following example data illustrates the problem. Annual peak discharges of the Red River of the North at Grand Forks, North Dakota from 1940–2014, and a loess smooth of



**Figure 12.16.** Graph of annual peak discharge, Red River of the North, at Grand Forks, North Dakota, 1940–2014, and a loess smooth of the data (using the default span value of 0.75).



**Figure 12.17.** Graph of annual peak discharge, Red River of the North, at Grand Forks, North Dakota, 1882–2014, and a loess smooth of the data (using the default span value of 0.75).

the dataset are shown in figure 12.16. It is obvious from visual inspection, and confirmed by trend test results, that the null hypothesis of no trend is easily rejected. However, if we look further back in the historical record to the first recorded observation in 1882, we come away with a very different impression (fig. 12.17). From the 133-year time scale perspective we may conclude that we have some type of quasi-periodic oscillation and our observations constitute something less than one full cycle of this oscillation. This conclusion should inform us that very large swings in flood magnitude over multidecadal periods are something we should expect in this watershed. Results of statistical tests (on either the shorter or longer version of this dataset) that include adjustment for serial correlation, including possible LTP, will indicate the presence of a statistically significant trend, but with an attained significance level (*p*-value) that is much larger than would arise from a standard method that assumed independence (such as Mann-Kendall or OLS regression approaches). Regardless of the specific *p*-value reported the records shown here tell us that over the next few decades flood magnitudes are unlikely to be similar to those observed in the mid-20th century. As a consequence, management actions should consider the possibility of the recent trend continuing for some years into the future, while recognizing that a reversal of this trend could certainly happen at any time. Admittedly this case is a rather extreme example, however, quasi-periodic oscillations are well known to exist in climate and hydrology time series (typically associated with phenomena such as El Niño or the Atlantic Multidecadal Oscillation). The analyst should always consider the possibility that what is seen as a deterministic trend is merely one limb of such an oscillation. Where strong persistence is evident in the data, planners must be prepared for a future that is quite different from the recent past.

In the limit, using a null hypothesis that includes oscillatory or long-term persistent processes, it becomes impossible to ever reject the hypothesis that the process is highly persistent but stationary. Vogel and others (2013) stated the problem succinctly, “Our ability to distinguish stochastic persistence from deterministic trends is in its infancy (Cohn and Lins, 2005)... Earth systems evolve over space and time, thus new theory and practical algorithms are needed to address long term social and physical drivers and feed-backs. New exploratory and statistical tools are needed to sharpen our insights into the emergent properties of such systems, and to guide modeling and prediction.” Loftis and others (1991) also offer a useful perspective on this problem. They note that “... a process which is stationary over long times ... may contain short-term runs which would be important from a management standpoint.”

There are methods of adjusting parametric and nonparametric trend tests to partly account for the effects of serial correlation. These methods are designed to achieve the nominal type I error rate (for example  $\alpha=0.05$ ) when the data arise from a stationary process that follows a specified temporally dependent process. These adjustments are helpful, but in practice the process parameters are generally estimated from the data, and the presence of a deterministic trend in the data will influence these adjustments. For example, in the case of the Mann-Kendall test, an adjustment for autocorrelation was developed by Yue and others (2002) and has been implemented in the zyp package in R (Bronaugh and Werner, 2013). The adjustment is referred to as “prewhitened nonlinear trend analysis” and variations and power analysis of the method have been discussed and debated in the literature. Other recent literature relevant to this problem include Zhang and others (2000), Matalas and Sankarasubramanian (2003), Yue and others (2003), Bayazit and Önöz (2007), Hamed (2008), Önöz and Bayazit (2012), and Wang and others (2015). Fortunately, as was demonstrated by Cohn and Lins (2005), even if we are dealing with a process that had no serial correlation or LTP there is only a very small loss of power associated with using a trend test designed for a serially correlated or persistent process versus using a test designed for white noise.

The WRTDS method, mentioned above, uses a block bootstrap method for computing statistical significance of trends (Hirsch and others, 2015). This method accounts for the influence of serial correlation at time scales of about 200 days or less, but does not account for longer term persistence such as decadal or multidecadal persistence or quasi-periodic oscillations. The problem of distinguishing deterministic trends from persistence is a serious issue in the analysis of water quality, where records are generally much shorter and have much lower sampling frequencies than those for discharge. An added difficulty regarding water quality trends is that there are so few datasets of long duration (say greater than 30 years) that also have a high sampling frequency (at least a few dozen samples per year). Without such long-term high-frequency datasets it will continue to be difficult to sort out natural long-term variations, shorter-term serial correlation, and trends owing to changes in human activity. The advent of multiyear records from monitors that measure water quality at time steps such as 15 minutes or 1 hour could be very helpful in sorting out some of these issues (see for example, Godsey and others [2010]).

These issues of distinguishing deterministic trends from persistent or quasi-periodic oscillations of hydrologic processes remains a challenge for the water resources community. We know that trends are

ubiquitous in hydrologic time series and it is difficult to sort out the relative roles of human activities on the landscape, human driven changes in the global atmosphere, and chaotic behavior of the land-atmosphere-ocean system that would exist in the absence of human activity. Awareness of all of these drivers of change cause us to question the long-standing statistical foundations of hydrology (in other words, time series that are independent and identically distributed). Milly and others (2008) stated the problem this way: "...we assert that stationarity is dead and should no longer serve as a central default assumption in water-resource risk assessment and planning. Finding a suitable successor [to stationarity] is crucial..." Describing trends is an important scientific goal in support of hazard mitigation, water resources planning, and evaluation of water quality improvement strategies. Statistical tools will need to continue to evolve and improve (see Salas and others, 2018). Such tools need to be cognizant of the atmospheric and watershed processes that drive the changes that would exist even in the absence of human interventions. There is no simple solution to this problem of distinguishing persistence from deterministic trends. The best advice to the analyst is to be cognizant of these issues in designing their analysis and explaining the meaning of the results.

## 12.10 Final Thoughts About Trend Assessments

Conveying trend results is often done best through the use of graphics. Using data-driven methods to capture the major patterns of change should be the primary goal of the assessment. The task for the analyst is to try to maximize the signal-to-noise ratio by removing as much variation as possible that is driven by other exogenous factors such as seasonality so that the signal can be better recognized and better estimated. Thoughtful trend analysis protocols can be crucial to identifying unexpected behaviors and should be used to check that actual outcomes are congruent with outcomes estimated with deterministic models operated in hindcast mode. The results can help to build confidence in deterministic modeling tools or reveal serious weaknesses in those models that must be resolved. We should expect surprises in trend analyses and those surprises can help to build improved understanding and predictive capabilities needed for effective management.

As time series get longer, it is critically important that we allow the analysis to reveal nonmonotonic trends and identify the timing and magnitude of the reversals. Identifying the congruence between the observed trends and natural or human driven forces (for example, regulations, controls, land use changes, atmospheric changes) is a major goal of trend assessments. This is often best accomplished by looking across many datasets (many sites, many analytes at a single site, or many parts of the probability distributions). The problems 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 1975–85 trend at one station to a 1960–80 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.

Presentation of trend results should not be filtered by the analyst with rules such as showing only those trends significant at  $\alpha=0.05$ . It may be that there are many sites where there is fairly strong evidence of

an increasing or decreasing trend but it tends to fall short of qualifying as a significant trend. Those results could be of great value and should not be censored by the analyst. It is the analyst's job to try to quantify how sure we should be about the direction or magnitude of the trend, rather than eliminating them from the results because they fall short of a prespecified significance level. Ultimately, the concerned citizen or decision maker needs to decide if they should act as if the trends are positive, act as if the trends are negative, or conclude that the evidence for trend in either direction is insufficient to act on (McBride and others, 2014). One duty of the analyst is to give them an evaluation of the likelihood that their decisions might be wrong so that they can decide if they should act on the information that is being presented. The other duty of the analyst is to characterize the nature of the trend (for example, is it trending up or down, what is the rate of change, is there a distinct change in slope or even a change in the sign of the slope). This can be important to verification of predictive models, to the evaluation of the success of water management strategies, and to provide early indicators of emerging water issues. Hydrologic systems are very noisy, and the job of the data analyst is to try to make trend signals, if they exist, rise above that noise level and become known and understood by the public and by decision makers. They should also use that understanding to improve the models and tools used for future planning and management of water resources.

## Exercises

### 1. Green River sediment load trends

During the period 1962–69 the Green River Dam was constructed about 35 miles upstream from a streamgaging station, which also collected suspended sediment data, on the Green River at Munfordville, Kentucky. It is a flood control dam, thus it regulates flow (changes the flow duration curve) but has very little effect on total annual runoff from its 4,300 square kilometer basin. Over the period of record for the sediment and streamflow record, 1952 through 1981, (which spans the pre-dam, dam construction, and regulated periods) has there been a monotonic trend in sediment transport? The data are in the file `Green.RData` located in SM.12. They consist of the water year, the annual mean discharge in  $\text{m}^3/\text{s}$ , and mean sediment load for the year in  $10^6$  kilograms per day ( $\text{kg}/\text{day}$ ). Using the various approaches to trend analysis presented in this chapter, what would you conclude about sediment load trends at this site?

### 2. Seasonal Kendall test with extensive ties and censoring

The following data are intended to represent samples of a dissolved metal collected for a period of 9 years with one sample in the winter and one sample in the summer. Data are in micrograms per liter ( $\mu\text{g}/\text{L}$ ).

Using these hypothetical data do separate Mann-Kendall tests for winter concentrations and summer concentrations and then a Seasonal Kendall test. Compute the  $S$  statistics by hand and then use the computer if you like to compute the necessary variances and final test statistics. Doing it by hand may seem tedious but it helps to get a feel for what is happening in the Mann-Kendall test when there is extensive censoring and ties.

Year	Winter concentration, in $\mu\text{g}/\text{L}$	Summer concentration, in $\mu\text{g}/\text{L}$
2001	<2	3
2002	3	<2
2003	<2	3
2004	4	<2
2005	<2	7
2006	3	<2
2007	8	9
2008	4	<2
2009	9	9

### 3. Water quality trend analysis, Potomac River at Chain Bridge, Washington, D.C., dissolved nitrate data

The dataset we will use is `PotomacNitrate.RData`, located in SM.12. It is a subset of the nitrate data collected by the USGS at Chain Bridge, Washington, D.C., from January 1999 through September 2015. It is a subset in the sense that it only contains the first sample taken in each month (that was done to keep the analysis here simpler, in a real application we would likely use all of the data). It consists of 197 values of filtered nitrate concentration (expressed as N in milligrams per liter [ $\text{mg}/\text{L}$ ], denoted as the variable `C`). For each observation the dataset also contains the daily mean discharge for the day the sample was collected (in cubic meters per second, denoted as `Q`) and also contains the variable time (which expresses the date in decimal years), and also the month (the digits 1 through 12). The data are stored in the data frame called `SampleS` and the columns are `time`, `month`, `Q`, and `C`. Can we detect and describe a time trend in concentration, using some of the methods described in the chapter? Try multiple approaches and describe what you think is the best way to approach this problem.

# Chapter 13

## How Many Observations Do I Need?

---

*During project planning, a hydrologist calculated the number of observations of sediment and metals concentrations required in order to determine if the mean concentrations exceeded those of the means for data collected 10 years prior. The hydrologist used equations for computing sample size for a two-group t-test, even though the older data were shaped quite differently from a normal distribution. Presumably the newer data would be non-normal as well. For many of the constituents, the equation produced an estimate of several hundred new observations required to distinguish their mean from the mean of the older data. The hydrologist knew that this number was unaffordable, so ignored the calculations and collected quarterly samples for three years. No difference in the mean of the twelve new observations could be determined from that for the older data using a t-test. What better process could the hydrologist have used?*

One of the most common questions asked when using statistical methods is, “How many observations (sample size) do I need to achieve the desired precision in interval estimates or power in hypothesis tests?” The purpose of this chapter is to introduce some examples of power analysis for required sample size estimation and show that methods of estimating required numbers of observations exist for nonparametric methods as well as for parametric methods such as the *t*-test. It is not intended to be a complete description of all methods for sample size estimation. The relation between sample size and statistical interval length was discussed in chapter 3; in this chapter, we discuss the relation between power of a hypothesis test and sample size in more detail.

Recall from chapter 4 that the power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false - the probability of seeing a signal that occurs. The power of a hypothesis test is a function of several parameters. To determine the required sample size you will need to specify

1.  $\alpha$ , the significance level of the test;
2. The desired power,  $1-\beta$ ;
3. The minimum signal you wish to detect; and
4. The statistical test you are using.

### 13.1 Power Calculation for Parametric Tests

For parametric tests such as the Student’s *t*-test, the minimum signal is  $\Delta / \sigma$ , where

1. The minimum distance between the mean difference for the null hypothesis (usually zero) and a specific mean difference for the alternative hypothesis,  $H_A$ , is  $\Delta$ . In other words, it is the minimum signal (difference in group means) that you want to be able to detect.
2. The standard deviation,  $\sigma$  is, in other words, the noise of the data.

The required sample size,  $n$ , can be written symbolically as

$$n = \text{Function}(\alpha, 1-\beta, \Delta / \sigma) \quad (13.1)$$

where  $\alpha$  is the significance level of the test and  $(1-\beta)$  is the desired power of the test. For example, the approximate sample size formula for the two-sample *t*-test found in most elementary textbooks is given by

$$n \geq \left[ \frac{Z_{1-\alpha} + Z_{1-\beta}}{\Delta/\sigma} \right]^2 \quad (13.2)$$

where  $Z$  is a quantile of the standard normal distribution (the normal distribution with mean of zero and standard deviation of 1). This equation requires either an assumption of normality of the underlying data or a large enough sample size for the Central Limit Theorem to hold, which we suggest is around 70 or more observations for the severity of skewness common to field data. It is an approximation because  $t$ -distribution quantiles should be used and not those for the normal distribution. However, quantiles for the  $t$ -distribution depend on the sample size,  $n$ , and  $n$  is not known. Kupper and Hafner (1989) used simulations to determine that the sample size,  $n$ , in the above equation should be increased by three to adjust for the approximate nature of the equation above. A second alternative is to use  $t$  instead of  $Z$  in the above equation, beginning with an estimate of a reasonable  $n$ , solving for  $n$ , and then iteratively using the computed  $n$  to obtain  $t$  in a subsequent computation, until the resulting estimates of  $n$  become stable.

The sample size,  $n$ , depends on the input parameters  $(\alpha, 1-\beta, \Delta / \sigma)$  in the following ways:

1. As  $\Delta$  increases (bigger minimum desired signal), the required  $n$  decreases.
2. As  $\sigma$  increases (more noise), the required  $n$  increases.
3. As  $(1-\beta)$  increases (more power desired), the required  $n$  increases.
4. As  $\alpha$  increases (less evidence of a signal needed), the required  $n$  decreases.

The analyst must specify the minimum signal size,  $\Delta$ , to be detectable by the test. Often the process is begun with a somewhat idealized specification of  $\Delta$ , say to measure a difference of 0.1 milligrams per liter (mg/L) of nitrate nitrogen, and the resulting large sample size to achieve this encourages the analyst to redo the analysis with a  $\Delta$  that is more realistic. Consider laboratory precision and estimates of variability from quality assurance studies when determining a (hopefully) realistic minimum desired signal.

The noise,  $\sigma$ , is often estimated using the sample standard deviations of available data. For determining the power of a two-sample  $t$ -test, use the pooled standard deviation (eq. 13.3) when the standard deviation is not the same in both groups (but data in both groups follow a normal distribution).

$$\sigma_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \quad (13.3)$$

The pooled standard deviation can be computed using the `sdpool.R` script in the supplemental material (SM.13) for chapter 13.

If the magnitude of the underlying noise is unknown, a rough estimate of  $\sigma$  can be obtained using  $\sigma = (\text{Max} - \text{Min}) / 4$ , where the *Max* and *Min* are the maximum and minimum of the data (or transformed data if using logarithms or another transformation when performing the test). This estimate is based on the fact that 95 percent of a normal population lies in the interval  $(\mu - 2\sigma, \mu + 2\sigma)$ .

For either method of estimating  $\sigma_p$ , if the original data is not normally distributed then to use these formulae, units that transform data into approximate normality should be used, recognizing that the eventual test will not be testing differences in means in the original units but instead means in the transformed units.

The library `stats` in R has scripts to compute power for  $t$ -tests, analysis of variance (ANOVA), and the test of difference in proportions. Typing `?power` lists available power functions. Typing `?power.t.test` lists the arguments for the `power.t.test` command:

```
> power.t.test(n, delta, sd, power, sig.level = 0.05,
  > type = c("two.sample", "one.sample", "paired"),
  > alternative = c("two.sided", "one.sided"))
```

Note `sig.level` is  $\alpha$  with a default set to 0.05. It can be changed by specifying `sig.level=.01` or other values. Exactly one of (`n`, `delta`, `sd`, `power`, `sig.level`) must not be specified, it will then be calculated from the values of the others.

**Example 13.1. Molybdenum—Estimating sample size for a *t*-test.**

We repeat the molybdenum concentration data in micrograms per liter ( $\mu\text{g/L}$ ) from exercise 2 in chapter 5 here and compute the pooled standard deviation of the two groups.

```
> downgrad <- c(0.85, 0.39, 0.32, 0.3, 0.3, 0.205, 0.2,
+      0.2, 0.14, 0.14, 0.09, 0.046, 0.035)
> upgrad <- c(6.9, 3.2, 1.7)
```

Then load and run the `sdpool` script:

```
> source("sdpool.R")
> sdpool(downgrad, upgrad)
sdpool = 1.030282
```

The R command for computing the required sample size,  $n$ , for each group in a two-sample *t*-test is:

```
> power.t.test(delta = 2, sd = 1.03, power = 0.9,
+      type = "two.sample", alternative = "one.sided")
```

Two-sample t test power calculation

```
n = 5.382099
delta = 2
sd = 1.03
sig.level = 0.05
power = 0.9
alternative = one.sided
```

**NOTE:** `n` is number in \*each\* group

In the function statement, the minimum difference `delta = 2` was chosen by the subject matter specialist as the required minimum. The estimated standard deviation, `sd = 1.03`, is the pooled standard deviation of the observed data. The specified power of 0.9 is something of a default when computing sample size requirements, a statement that signals should be detected approximately 90 percent of the times that they occur.

If the two groups followed a normal distribution and had the same standard deviation of 1.03, six observations per group (after rounding up) would be required to find that the mean of the `upgrad` data was 2  $\mu\text{g/L}$  of molybdenum or more greater than the `downgrad` data, with a probability of 90 percent when that level of difference is present. However, from the Shapiro-Wilk test, the residuals from the group means do not follow a normal distribution, and there are too few data to invoke the Central Limit Theorem.

```
> resids <- c(upgrad - mean (upgrad), downgrad - mean (downgrad))
> shapiro.test(resids)
```

```
Shapiro-Wilk normality test
data: resids
W = 0.716, p-value = 0.0002546
```

Owing to non-normality and unequal variance (and so unequal standard deviation), we expect that the computed sample size requirement of six observations per group may not be sufficient to find an actual difference with 90 percent power. The fact that no significant difference was found between the two groups in their original units may be partly the result of the findings that neither group appears to be close to a normal distribution, they have differing standard deviations, and that the upgradient group consists of only three observations.

The natural logarithms of the molybdenum concentrations appear closer to a normal distribution according to the Shapiro-Wilk test, so a more realistic computation of sample size would be to use the logarithms of the data.

```
> lndown <- log(dowgrad)
> lnup <- log(upgrad)
> ln.resid <- c( lnup - mean(lnup), lndown - mean(lndown))
> shapiro.test (ln.resid)
```

```
Shapiro-Wilk normality test
data: ln.resid
W = 0.96621, p-value = 0.7741
```

This transformation changes the units for  $\Delta$  and  $s$ , and the type of comparison between groups. A  $t$ -test on the difference in mean logarithms is equivalent to a test for whether the ratio of geometric means of the two groups equals 1. Setting the ratio of geometric means to two, the  $t$ -test should determine whether the difference in mean logarithms of the upgradient and downgradient groups equals 0.69.

```
> ln.delta <- log(2)
> ln.delta
[1] 0.6931472
> sdpool(lndown, lnup)
sdpool = 0.8465613
> power.t.test(delta = 0.69, sd = 0.85, power = 0.9,
+     type = "two.sample", alternative = "one.sided")
```

```
Two-sample t test power calculation
```

```
n = 26.69657
delta = 0.69
sd = 0.85
sig.level = 0.05
power = 0.9
alternative = one.sided
```

NOTE: n is number in \*each\* group

Approximately 27 observations per group are required when using a  $t$ -test to determine whether the geometric mean of the upgradient group is twice the geometric mean of the downgradient group. Note that after using a logarithm transformation, means of the original data are no longer being compared. As the

logarithms are closer in shape to a normal distribution, whereas data in original units are strongly skewed, this estimate of sample size is more likely to be accurate, at the cost of changing the parameters being compared to a ratio of geometric means.

The power achieved for a fixed sample size can also be calculated by specifying  $n$ ,  $\alpha$ , and  $\Delta$ . For example, specifying 30 observations per group, standard deviation of 0.90, a difference in mean logarithms of 0.69 units is expected to be detected with a probability of 92.8 percent:

```
> power.t.test(n = 30, delta = 0.69, sd = 0.85,
+     type = "two.sample", alternative = "one.sided")
```

Two-sample t test power calculation

```
n = 30
delta = 0.69
sd = 0.85
sig.level = 0.05
power = 0.9281424
alternative = one.sided
```

NOTE: n is number in \*each\* group

## 13.2 Why Estimate Power or Sample Size?

There are two reasons for estimating power or sample size.

1. The prospective study. A determination of power answers the question, “What power will I achieve with the number of observations that I plan to collect?” This is done before collecting data, to determine whether there is sufficient power to detect the smallest desired difference of importance. Power can be increased by increasing sample size, increasing  $\Delta$ , or by using a better test.
2. The retrospective study. The determination of power answers the question, “Since I did not reject the null hypothesis, did I not have sufficient power to reject it though the groups differ, or did I just not have enough data?” If you failed to reject the null hypothesis and the power was low, say 0.2, then there would not be sufficient power to dependably detect differences even if they were present. You are left not knowing whether the two groups are different by your specified  $\Delta$  or not. Given that, you could
  - A. Modify your requirements (increase  $\Delta$ );
  - B. Collect more data;
  - C. Stop the study, as it is not worth the money when the likelihood of finding the desired signal is so low; or
  - D. Use a different test, perhaps a nonparametric test.

If you failed to reject the null hypothesis and the power is high, say 0.9, then you have shown there is likely no statistically meaningful difference between groups ( $\Delta$  is not significantly different from zero).

Note that the parametric power functions discussed so far are based on a normality assumption, and all require estimates of the standard deviation,  $\Delta$ , and a specification of  $\alpha$ . Note also that it is often useful to compute power for a range of plausible true values of delta and standard deviation—called power functions. Graphical presentations of the power functions for plausible ranges of the values of the parameters will provide a basis for discussion about project planning, such as whether it would be a small or large cost to collect sufficient numbers of samples to increase the power appreciably.

### 13.3 Power Calculation for Nonparametric Tests

Power function calculations are also available for nonparametric tests. These do not require that data follow a normal distribution, but they do require estimates of a minimum detectable signal, called *PPlus* (in contrast to  $\Delta/s$  for parametric tests). *PPlus* is the probability that for any random  $x$  in one group and  $y$  in the other,  $x$  is greater than  $y$ . *PPlus* equals 0.5 for the null hypothesis and increases as the  $x$  data more frequently are larger than the  $y$  data. *PPlus* decreases below 0.5 when the  $x$  data more frequently are smaller than the  $y$  data. When *PPlus* is sufficiently large or small and so differs from 0.5, the null hypothesis is rejected in favor of the alternate hypothesis.

The procedure to compute power or required sample size is described in a brief paper by Noether (1987). The sample size,  $n$ , depends on three input parameters ( $\alpha, 1-\beta, PPlus$ ) in the following ways:

1. As *PPlus* increases (bigger minimum desired signal), the required  $n$  decreases.
2. As  $1-\beta$  increases (more power desired), the required  $n$  increases.
3. As  $\alpha$  increases (less evidence of a signal needed), the required  $n$  decreases.

Noether (1987) gives the formula describing the relation between sample sizes in each group, the significance level  $\alpha$ , the power  $1-\beta$ , and *PPlus* as

$$n \geq \frac{\left[ \frac{Z_{1-\alpha} + Z_{1-\beta}}{PPlus - 0.5} \right]^2}{[12 \cdot c \cdot (1-c)]} \quad (13.4)$$

where

- $c = m_1/n$ , the proportion of observations in the first group;
- $m_i =$  the number of observations in group  $i$ ,  $i=1,2$ ; and
- $n = m_1 + m_2$  is the total number of observations.

Note the similarity between this formula and the one for the two-sample  $t$ -test. The script `power.WMW` uses this formula to solve for  $n$  given the other variables, or to solve for the power  $1-\beta$  given the other variables.

#### 13.3.1 Estimating the Minimum Difference *PPlus*

How then does the user estimate the desired minimum difference, *PPlus*? With the sign test, it is easy to specify *PPlus* directly. Only the algebraic sign of the difference in paired data is used for the sign test (chap. 6), and the null hypothesis,  $H_0$ , is that *PPlus*=0.5. The analyst specifies the probability of paired  $(x, y)$  values where  $x > y$ , choosing a *PPlus* greater than 0.5.

For the two-group Wilcoxon rank-sum test (chap. 5), *PPlus* can be computed from an expression easier to conceptualize by the analyst, the ratio of geometric means (*GMratio*) of the two groups. The first method is to specify a numeric estimate of *GMratio*. The null hypothesis is that the *GMratio* equals 1. A *GMratio* of 1.5 states that the data in the first group are typically about 50 percent higher than data in the second group.

Using the ratio of medians or geometric means to express a minimum difference is a consequence of taking logarithms or another power transformation of the data. If the data are strictly positive (no zeros or negative values), logarithms can be computed. If this is not true, then another monotonic transformation to normality such as a cube-root transform could be used instead. The probability of  $x$  being greater than  $y$  is also the probability that the natural logarithm (or other monotonic power transformation) of  $x$  is greater than the same transformation of  $y$ , based on the fact that a monotonic transformation does not alter relative probabilities. Using a natural logarithm,

$$\text{Prob}[x > y] = \text{Prob}[\ln(x) > \ln(y)]. \quad (13.5)$$

After a monotonic transformation to achieve symmetry (or normality), the mean and median of the transformed data are identical. The rank-sum test of difference in medians of transformed data is also a test for difference in the means. The difference in the mean of the logs equals the log of the ratio of geometric means, it is the ratio of medians when using other monotonic transformations. Using a natural log transformation, the null hypothesis for the rank-sum test is therefore

$$H_0 : \text{mean}(\ln x) - \text{mean}(\ln y) = \ln(\text{geomean } x / \text{geomean } y) = 0 ,$$

where *geomean* is the geometric mean. In original units this is

$$H_0 : \text{geomean}(x) / \text{geomean}(y) = 1 .$$

The standard deviation of the difference in logarithms is

$$\text{StdDev}[\ln x - \ln y] = \sqrt{\text{variance}(\ln x) + \text{variance}(\ln y)} ,$$

assuming that *x* and *y* are uncorrelated. Given this we can calculate the value of *PPlus* from the ratio of geometric means:

$$PPlus = \text{Prob}[(\ln x - \ln y) > 0] \text{ or } \text{Prob}[x > y] . \quad (13.6)$$

The probability is computed using the probability function for a normal distribution (*pnorm*) in R. For example, using a difference in group means of logarithms  $\text{mean}(\ln x) - \text{mean}(\ln y) = 0.69$ , and  $\text{StdDev}[\ln x - \ln y] = 2$ , using the *pnorm* function in R:

```
> pnorm(0, 0.69, 2, lower.tail = FALSE)
[1] 0.6349528
```

This is *PPlus*, the probability of  $[\ln x - \ln y] > 0$  given the observed data, which is the probability of a random *x* value exceeding a random *y* value. Computing *PPlus* in this way requires a minimum desired *GMratio* and an estimate of the standard deviation of the logarithms.

A second method to come up with a desired value for *PPlus* is to look at graphs as shown in example 13.2. *GMratio* can be seen as a measure of how separated the two boxplots of data are for the null hypothesis of a geometric mean ratio of one, identical boxplots are positioned side by side at the identical position on the *y*-axis. The minimum desired difference, *PPlus*, of the alternate hypothesis can be estimated by computing the estimated ratio of geometric means, either from observed data or a desire to see a specified ratio: for a 30 percent difference in medians the  $GMratio = \text{geomean } x / \text{geomean } y = 1.30$ . The *power.WMW* script then computes the corresponding *PPlus*, which then is used to return the required sample size to achieve that power.

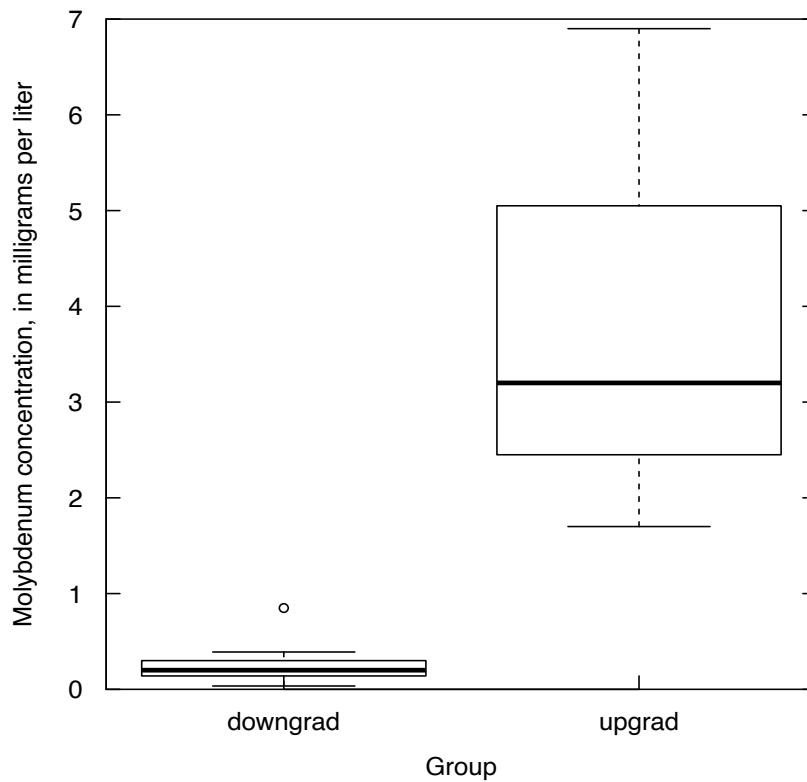
### Example 13.2. Molybdenum—Choosing the *GMratio* from plots.

For the molybdenum data (fig. 13.1), the two boxplots are widely separated, the ratio of geometric means is large, and so *PPlus* is close to 1, it equals 0.995. This results from all of the data in the upgradient (upgrad) group being larger than all of the observations in the downgradient (downgrad) group.

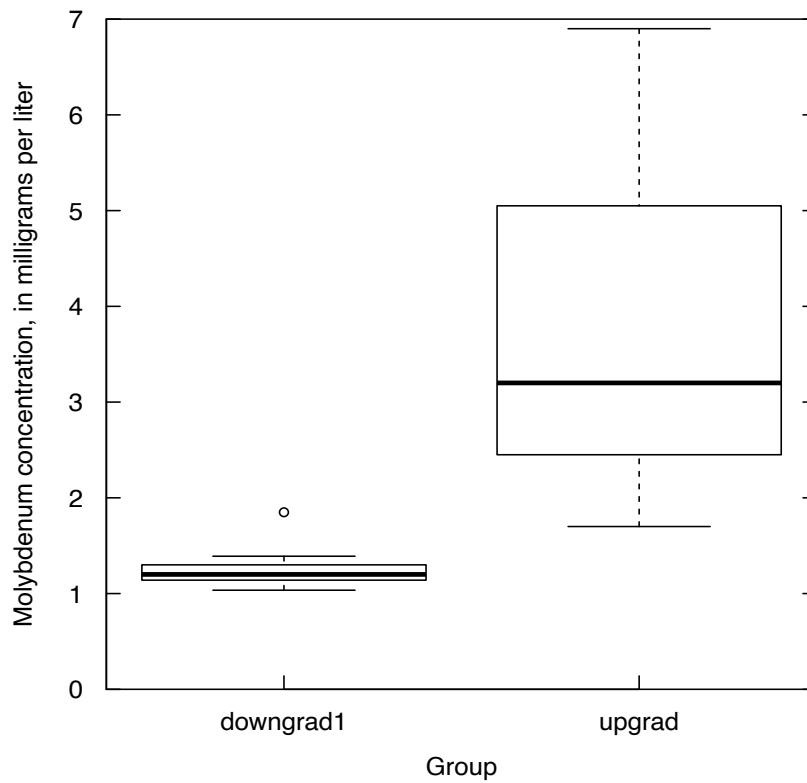
```
> boxplot(downgrad, upgrad, names = c("downgrad", "upgrad"), ylab =
  "molybdenum")
```

By adding the constants 1, 2, or 3 successively to the downgrad group we can visualize how *PPlus* changes as the groups overlap more and the *GMratio* decreases (figs. 13.2–13.4).

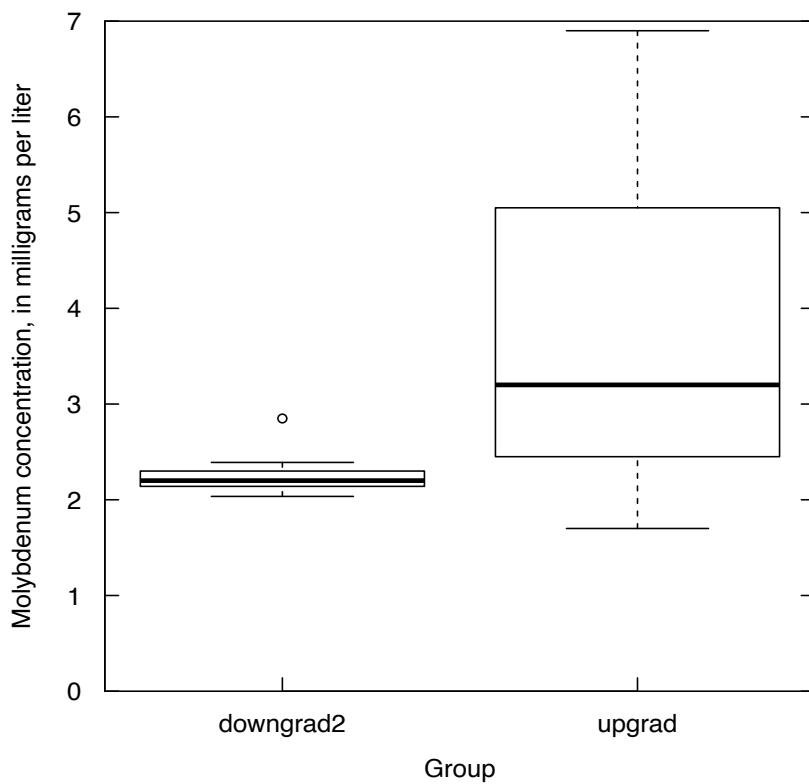
Using the selected *GMratio*, the *power.WMW* script (SM.13) will compute *PPlus* and the power obtained for specific sample sizes.



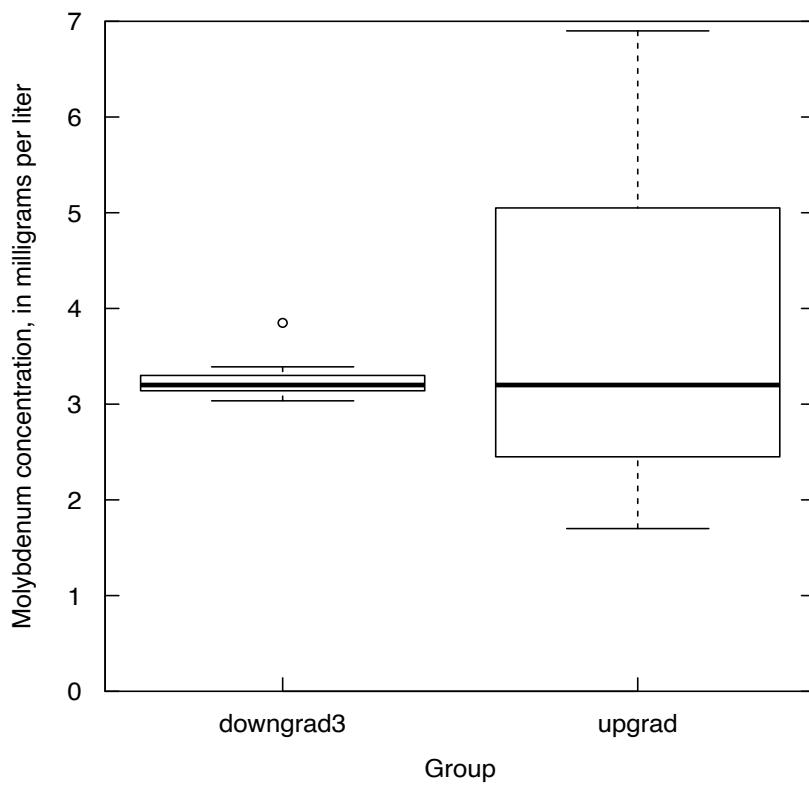
**Figure 13.1.** Boxplots of the two groups of molybdenum data from exercise 2 of chapter 5.  $GMratio = 19.3$ ,  $PPlus = 0.995$ .



**Figure 13.2.** Graph showing the effect on  $PPlus$  of adding 1.0 to the downgradient observations.  $GMratio = 2.7$ ,  $PPlus = 0.917$ .



**Figure 13.3.** Graph showing the effect on *PPlus* of adding 2.0 to the downgradient observations.  $GMratio = 1.5$ ,  $PPlus = 0.715$ .



**Figure 13.4.** Graph showing the effect on *PPlus* of adding 3.0 to the downgradient observations.  $GMratio = 1.03$ ,  $PPlus = 0.518$ .

### 13.3.2 Using the power.WMW Script

The power.WMW script computes power as a function of sample size for the Wilcoxon rank-sum test. The required arguments are shown below, along with their default values.

```
> power.WMW <- function(y1, y2, gmratio, conf = 95, stacked = TRUE)
> # Computes power function for Wilcoxon rank-sum test,
> # one-sided alternative
> # y1 = data (for both groups if stacked = TRUE; 1st group if
> # stacked = FALSE)
> # y2 = grouping variable if stacked = TRUE; 2nd data column if
> # stacked = FALSE
> # gmratio is the ratio of geometric means
```

For a prospective study, gmratio is the ratio of geometric means chosen from past data or from an educated guess of requirements by the analyst. For a retrospective study, gmratio is the observed ratio of geometric means for the data. For example, a retrospective study on the molybdenum data would use the observed ratio of geometric means of 18.5 to input to the power.WMW script:

```
> lnup <- log(upgrad)
> lndown <- log(dowograd)
> obsvdGMratio <- exp(mean(lnup) - mean(lndown))
> obsvdGMratio
[1] 18.50774
> source("power.WMW.R")
> power.WMW(upgrad, dowograd, gmratio = 18.5)

DATA ANALYZED: upgrad dowograd
-----
Results for Wilcoxon rank-sum test (one-tailed)
      with specified gmratio.
SampleSize is the required number of obs in both groups together.
Nxratio is the proportion of SampleSize for 1st variable entered.
      SampleSize Nxratio GMratio PPlus ObsrvPower
1       16  0.1875    18.5  0.996     84.97
-----
SampleSize is integer for closest Power
      not less than specified Power
Sample sizes are rounded up to smallest integer
      not less than the computed sample size
      SampleSize Power
1       7   55.1
2       7   55.1
```

3	9	64.2
4	10	68.2
5	11	71.8
6	12	75.0
7	14	80.6
8	17	86.8
9	20	91.2
10	25	95.6

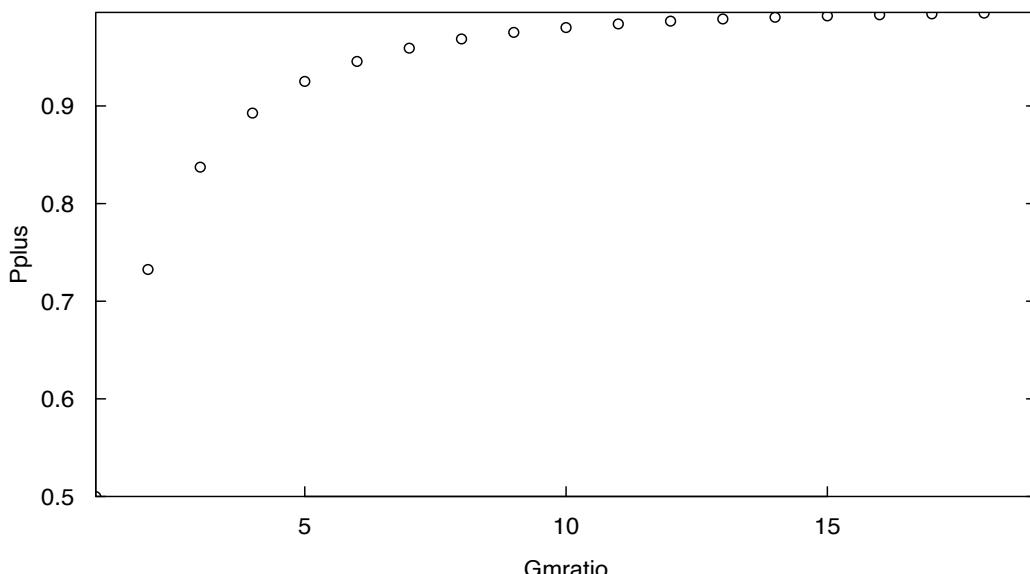
The observed power for a sample size of 16 with 19 percent observations in the upgradient group results in a power of 85 percent. The lower `SampleSize` column lists the total number of observations in both groups together needed to achieve the power in the same row. The distribution of samples among the two groups for a retrospective result will be the observed proportion for the input data—81 percent of the molybdenum data are in the downgrad group and 19 percent in the upgrad group. Therefore, with a total sample size of 25, an arrangement of  $0.81 \cdot 25 = 20$  observations in the downgrad group and 5 observations in the upgrad group would result in a power of 95.6 percent to see differences in group medians by the rank-sum test. To achieve 90 percent power, 19 total observations would be required.

The relation between `PPlus` and `GMratio` for the molybdenum data is shown in figure 13.5. Figure 13.5 was generated using the `pplusplot` function in SM.13. It requires the standard deviations of the logarithms for each group separately:

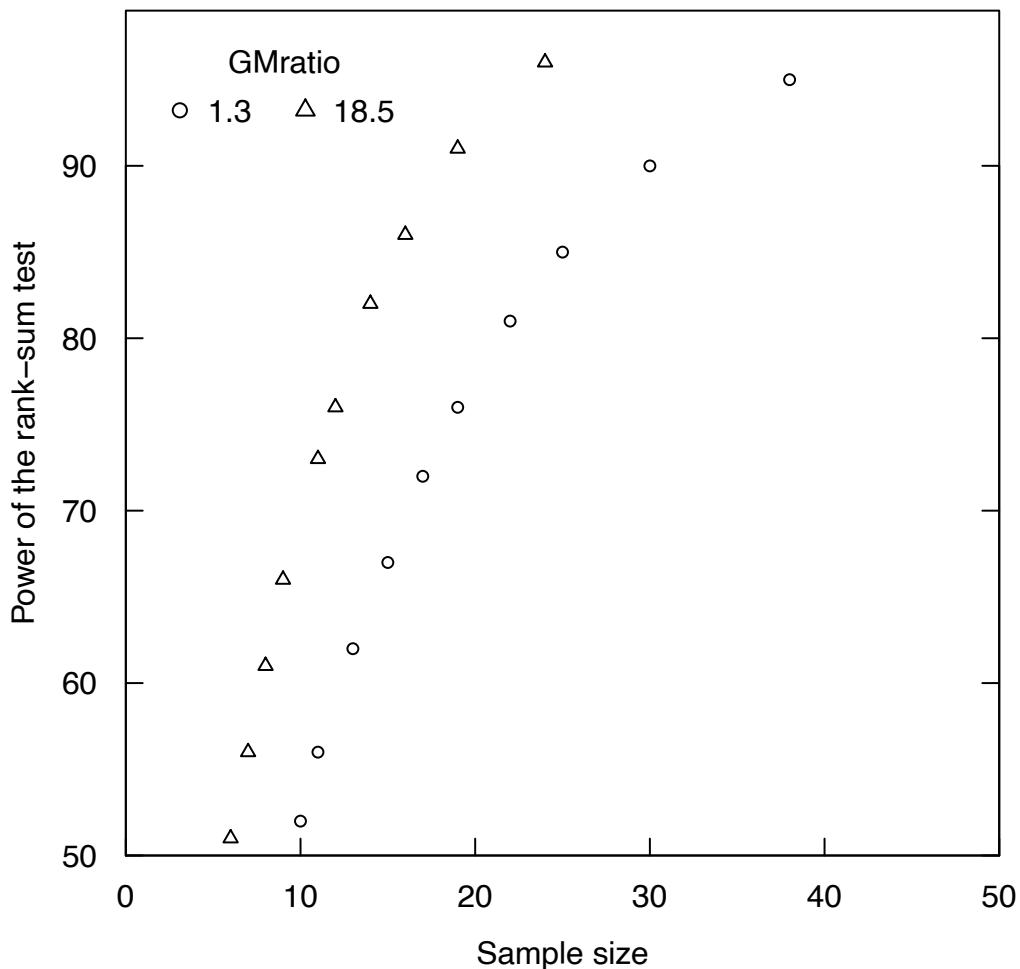
```
> sd(lnup)
[1] 0.7015435
> sd(lndown)
[1] 0.8683797
```

Run the script with the command

```
> pplusplot(seq(1, 19, by = 1), 0.70, 0.87)
```



**Figure 13.5.** Graph showing `GMratio` versus `PPlus` using the observed standard deviation of logarithms of 0.70 and 0.87 from the molybdenum dataset of chapter 5.



**Figure 13.6.** Graph showing power to differentiate downgradient from upgradient molybdenum concentrations for various samples sizes for dissimilar data ( $GMratio = 18.5$ ) versus quite similar data ( $GMratio = 1.3$ ). More observations are needed to discern the smaller  $GMratio$  group differences.

Taking low (1.3, just above the null hypothesis value of 1) and high (18.5) values of  $GMratio$ , the obtained power is plotted versus sample size for the molybdenum data in figure 13.6.

The two power curves plotted together show that when the geometric means differ by a larger amount (18.5 as compared to 1.3), many fewer samples are required to obtain the same power. All of these power calculations are obtained by using the script `power.WMW`, located in SM.13 accompanying this report.

## 13.4 Comparison of Power for Parametric and Nonparametric Tests

Which tests, parametric or nonparametric, have greater power to detect differences for a given sample size? This depends, of course, on the distribution of the data to be used. Generally, parametric tests have slightly more power when data are exactly normal, are equal in power when data are approximately normal

(the best-case scenario for parametric tests with field data), and far lower power for non-normal data that are asymmetric and have outliers (chap. 4).

Using a dataset where both variables have the same number of observations, we can directly compare the number of observations needed to detect with 90 percent power a mean (*t*-test) or median (rank-sum test) difference between groups. Two of the four rock types from the specific capacity dataset of chapter 7 provide a good example. The dolomite and siliciclastic rock types differ in their means using the *t*-test, and in their quantiles using the rank-sum test, at a significance level of 0.05:

```
> load("specapic.RData")
> specapic2 <- specapic[rock == "Siliciclastic" | rock == "Dolomite",]
> t.test(spcap ~ rock, data = specapic2)
```

Welch Two Sample t-test

```
data: spcap by rock
t = 2.5913, df = 49.117, p-value = 0.01255
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
3.238134 25.604258
sample estimates:
mean in group Dolomite mean in group Siliciclastic
15.492200          1.071004
```

```
> wilcox.test(spcap ~ rock, data = specapic2)
```

Wilcoxon rank sum test with continuity correction

```
data: spcap by rock
W = 1666, p-value = 0.004174
alternative hypothesis: true location shift is not equal to 0
```

The parametric power analysis is done after computing the pooled standard deviation of the two groups.

```
> source("sdpool.R")
> sdpool(spcap[rock == "Dolomite"], spcap[rock == "Siliciclastic"])
sdpool = 27.82611
> power.t.test(delta = 2, sd = 27.8, power = 0.9,
+     type = "two.sample", alternative = "one.sided")
```

Two-sample t test power calculation

```
n = 3309.919
```

```

delta = 2
sd = 27.8
sig.level = 0.05
power = 0.9
alternative = one.sided

```

NOTE: n is number in \*each\* group

The non-normality and unequal variances of the data result in a statement that 3,310 observations are required for each group to detect a 2 unit difference in specific capacity with 90 percent power. In comparison, the nonparametric power.WMW script will determine the rank-sum test's data requirement to achieve 90 percent power. The *GMratio* for the required minimum difference of 2 is first computed:

```

> source("power.WMW.R")
> geomean.silic <- exp(mean(log(spcap[rock == "Siliciclastic"])))
> gmratio.input <- (geomean.silic + 2) / geomean.silic
> gmratio.input
[1] 5.268879
> power.WMW(spcap[rock == "Dolomite"],
+           spcap[rock == "Siliciclastic"], gmratio = 5.27)

```

DATA ANALYZED: spcap[rock == "Dolomite"] spcap[rock ==  
"Siliciclastic"]

-----  
Results for Wilcoxon rank-sum test (one-tailed)

with specified gmratio.

SampleSize is the required number of obs in both groups together.

Nxratio is the proportion of SampleSize for 1st variable entered.

SampleSize	Nxratio	GMratio	PPlus	ObsrvPower
1	100	0.5	5.27	0.715

-----  
SampleSize is integer for closest Power

not less than specified Power

Sample sizes are rounded up to smallest integer

not less than the computed sample size

SampleSize	Power
1	20 51.0
2	23 55.8
3	26 60.2
4	30 65.5

```

5      34  70.2
6      39  75.4
7      45  80.5
8      52  85.2
9      62  90.2
10     78  95.1

```

A total sample size of 62 (31 observations in each group) would produce 90 percent power using the rank-sum test. This is far fewer than the recommended number for the *t*-test. Given that the *t*-test requires the data to follow a normal distribution for its power computations to be correct, and that these data are non-normally distributed with unequal standard deviations, the result of the `power.t.test` command is suspect.

An alternate approach would be to use a logarithm transformation and use `power.t.test` on the logarithms, because they are similar in shape to a normal distribution according to the Shapiro-Wilk test:

```
> shapiro.test(log(spcap[rock == "Siliciclastic"]))
```

```
Shapiro-Wilk normality test
```

```

data: log(spcap[rock == "Siliciclastic"])
W = 0.97456, p-value = 0.351

```

```
> shapiro.test(log(spcap[rock == "Dolomite"]))
```

```
Shapiro-Wilk normality test
```

```

data: log(spcap[rock == "Dolomite"])
W = 0.97211, p-value = 0.2815

```

```

> sdpool(log(spcap[rock == "Dolomite"]),
+         log(spcap[rock == "Siliciclastic"]))
sdpool = 2.063606
> power.t.test(delta = log(5.27), sd = 2.06, power = 0.9,
+                type = "two.sample", alternative ="one.sided")

```

```
Two-sample t test power calculation
```

```

n = 27.01639
delta = 1.66203
sd = 2.06
sig.level = 0.05

```

```
power = 0.9
alternative = one.sided
```

NOTE: n is number in \*each\* group

The parametric power calculation on the logarithms estimated a recommended sample size of 27 observations in each group or 54 total. This is a more realistic calculation, given the similarity to a normal distribution of the logarithms. It is a similar result to the sample size required by the rank-sum test (62), as both the *t*-test on the logarithms and the rank-sum test are testing differences in medians (the geometric mean estimates a median when the logarithms are symmetric). A summary of power obtained with three different sample sizes for the three tests are given in table 13.1, following the R code necessary to calculate them.

First the *t*-test:

```
> power.t.test(n = c(10, 20, 35), delta = 2.0, sd = 27.8,
+               type = "two.sample", alternative = "one.sided")
```

Two-sample t test power calculation

```
n = 10, 20, 35
delta = 2
sd = 27.8
sig.level = 0.05
power = 0.06812103, 0.07760416, 0.08900946
alternative = one.sided
```

NOTE: n is number in \*each\* group

Next the *t*-test on logarithms:

```
> power.t.test(n = c(10, 20, 35), delta = log(5.27), sd = 2.06,
+               type = "two.sample", alternative = "one.sided")
```

Two-sample t test power calculation

```
n = 10, 20, 35
delta = 1.66203
sd = 2.06
sig.level = 0.05
power = 0.5362065, 0.8052854, 0.9550959
alternative = one.sided
```

NOTE: n is number in \*each\* group

Last, the rank-sum test:

```
> power.WMW(spcap[rock == "Dolomite"], spcap[rock == "Siliciclastic"],
+           gmratio = 5.27, power = seq(0.5, 0.99, by = 0.02))

DATA ANALYZED: spcap[rock == "Dolomite"]
spcap[rock == "Siliciclastic"]
-----
Results for Wilcoxon rank-sum test (one-tailed)
with specified gmratio.

SampleSize is the required number of obs in both groups together.
Nxratio is the proportion of SampleSize for 1st variable entered.

  SampleSize Nxratio GMratio PPlus ObsrvPower
1       100     0.5    5.27  0.715      98.16
-----
SampleSize is integer for closest Power
not less than specified Power
Sample sizes are rounded up to smallest integer
not less than the computed sample size

  SampleSize Power
1       20   51.0
2       21   52.6
3       22   54.2
4       24   57.3
5       25   58.8
6       26   60.2
7       28   62.9
8       29   64.2
9       31   66.8
10      33   69.1
11      34   70.2
12      36   72.4
13      38   74.4
14      40   76.3
15      42   78.1
16      45   80.5
17      48   82.7
18      51   84.6
19      54   86.4
```

20	58	88.4
21	62	90.2
22	67	92.1
23	74	94.1
24	83	96.0
25	99	98.1

When there is asymmetry and (or) outliers on the original scale, the nonparametric test will have a great advantage in power over the parametric test (table 13.1). If data are transformed to symmetry with a monotonic power transformation or logarithms, the results on transformed data will be similar to those for the nonparametric test. Both tests will be testing differences in medians or (in the case of logarithms) geometric means, not means on the original scale. Permutation tests for differences in means (chap. 4 and demonstrated in chaps. 5–7) will have power characteristics similar to the equivalent nonparametric test, as their *p*-values are computed in a similar fashion. Permutation tests will have a strong power advantage over their equivalent parametric tests for testing means of non-normal data without transformations of scale. For additional reading on power and sample size computations, see Hsieh and others (1998), Millard and Neerchal (2000), and Bacchetti (2010).

**Table 13.1.** Power for three samples sizes for the specific capacity data using the *t*-test, *t*-test on the logarithms, and the rank-sum test. All values except number of observations given in percent.

Total number of observations (half in each group for <i>t</i> -test)	<i>t</i> -test on mean (assumes normal distribution and equal variance)	<i>t</i> -test on logarithms (more realistic for skewed data)	Rank-sum test
20	7	54	51
40	8	80	76
70	9	96	93

## Exercise

1. Nitrogen concentrations in precipitation were collected in both residential and industrial areas. The data are found in the dataset `precipn` from chapter 5. Nitrogen in precipitation is in the `NH4orgN` column and the location variable is named `where`. No difference was found with the two-sample *t*-test.
  - A. Compute the pooled standard deviation and use the `power.t.test` command to estimate the number of observations that would have been needed to find the observed two-sided difference  $\Delta$  of 0.03 mg/L, with a power of 0.9.
    - I. What was the power achieved with 10 observations in each group?
    - II. How many observations would be needed if the calculations were done in natural logs?
    - III. What was the power achieved with 10 observations in each group?
  - B. Use the `power.WMW` function to estimate the number of observations needed to find a difference in medians between the groups with a power of 0.9 using the rank-sum test. Hint, you will need to compute the ratio of the two groups' geometric means. Also find the power obtained with 20 samples total.
  - C. Did the rank-sum test or *t*-test on geometric means achieve greater power with the observed amount of data?



# Chapter 14

## Discrete Relations

---

Investigators sampled three aquifers to determine whether they differ in their concentrations of an organic contaminant. In all three, more than 60 percent of the samples were listed as below the detection limit. What method will test whether the distribution of the chemical is similar in the three aquifers while effectively incorporating data categorized as below versus above the detection limit?

Ecosystem health is evaluated at three stream locations by using counts of three macroinvertebrate species. The three species cover a 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 relation between two discrete (also called categorical) variables.

### 14.1 Recording Categorical Data

Categorical variables are those whose values are not along a continuous scale (such as concentration) but take on only one of a finite number of discrete values separated into one of several categories. These may be nominal data, data without a scale from low to high. Examples of nominal categorical data used in water resources include soil type, land use group, and location variables such as aquifer unit, site number, or name. Categorical data may also be ordinal, data that can be placed in an order from low to high. Examples of ordinal categorical variables used in water resources studies include presence or absence of a benthic invertebrate, whether an organic compound is above or below the detection limit, and ratings such as an impairment rating in five categories from not impaired to highly impaired.

To investigate the relation between two categorical variables, data are recorded as a matrix of counts (fig. 14.1). The matrix is composed of two categorical variables, one assigned to the columns and one to the rows. The entry in any given cell of the matrix is the number of observations,  $O_{ij}$ , that fall into the  $i$ th row and  $j$ th column of the matrix.

		Variable 2		
Variable 1		Group 1	Group 2	Group 3
Response 1	Group 1	$O_{11}$	$O_{12}$	$O_{13}$
	Group 2	$O_{21}$	$O_{22}$	$O_{23}$
		$C_1 = \sum O_{1j}$	$C_2 = \sum O_{2j}$	$C_3 = \sum O_{3j}$
				$A_1 = \sum O_{1j}$
				$A_2 = \sum O_{2j}$
				$N = A_1 + A_2$
				$= \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{C}_3$

Figure 14.1. Structure of a two variable, 2x3 matrix of counts.

## 14.2 Contingency Tables (Both Variables Nominal)

Contingency table analysis evaluates the association between two nominal categorical variables, the row and column variables. Nominal variables have no natural ordering, so their categories may be re-ordered without any loss in meaning. The null hypothesis,  $H_0$ , is that the row and column variables are not associated—the proportion of data residing in each row is the same for each of the columns (given some small variation due to noise). For the 2x3 contingency table of figure 14.1, with  $p_{21}$  being the proportion of data of column 1 that are in row 2,  $p_{22}$  being the proportion in column 2 in row 2, and so forth, the null hypothesis is

$$H_0: p_{21} = p_{22} = p_{23}$$

The alternate hypothesis,  $H_A$ , is that at least one proportion differs from the others.

If there is sufficient evidence to reject  $H_0$  in favor of  $H_A$ , the variables are associated or correlated. This correlation manifests as a difference in the proportion of observations in the rows, going from column to column. The statement that the row or column variable causes the observed values for the other variable is not necessarily implied, analogous to the implications of correlation. Causation must be determined by knowledge of the relevant processes, not only the mathematical association. For example, a third underlying variable could cause the observed patterns in both variables.

### 14.2.1 Performing the Test for Association

To test for association, the observed counts,  $O_{ij}$ , in each cell are summed across each row  $i$  to produce the row totals,  $A_i$ , and down each column  $j$  to produce column totals,  $C_j$ . There are  $m$  rows ( $i=1, 2, \dots, m$ ) and  $k$  columns ( $j=1, 2, \dots, k$ ). The total sample size,  $N$ , is the sum of all counts in every cell, or  $N = \sum A_i = \sum C_j = \sum O_{ij}$ . 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  $C_j$  totals by  $N$ :

$$a_i = A_i / N \text{ and } c_j = C_j / N \quad (14.1)$$

If  $H_0$  is true, the probability of a new sample falling into row 1 is estimated by the marginal probability,  $a_1$ , regardless of which column the sample is taken from. Thus, the marginal probability for the  $i$ th 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. Therefore, with  $H_0$  true, the best estimate of the joint probability,  $e_{ij}$ , of being in a single cell in the table equals  $a_i$ , the marginal probability of being in row  $i$ , multiplied by  $c_j$ , the marginal probability of being in column  $j$ :

$$e_{ij} = a_i \cdot C_j . \quad (14.2)$$

Finally, for a sample size of  $N$ , the expected number of observations in each cell given  $H_0$  is true is computed by multiplying each joint probability,  $e_{ij}$ , by  $N$ :

$$E_{ij} = N \cdot e_{ij} = N \cdot a_i \cdot c_j \quad (14.3)$$

or

$$E_{ij} = \frac{A_i C_j}{N} .$$

To test  $H_0$ , first compute cell residuals as observed counts,  $O_{ij}$ , minus the counts  $E_{ij}$  expected when  $H_0$  is true, divided by expected counts, or  $\frac{(O_{ij} - E_{ij})}{E_{ij}}$ . The test statistic,  $X_{ct}$ , is the sum of the squared residuals, summed over all  $i \cdot j$  cells:

$$X_{ct} = \sum_{i=1}^m \sum_{j=1}^k \left[ \frac{(O_{ij} - E_{ij})}{E_{ij}} \right]^2. \quad (14.4)$$

If the observed cell counts,  $O_{ij}$ , are sufficiently different from the  $E_{ij}$  for at least one cell,  $X_{ct}$  will be large and  $H_0$  will be rejected. If the  $O_{ij} \cong E_{ij}$  for all  $i \cdot j$  cells,  $X_{ct}$  will be small and  $H_0$  will not be rejected. To evaluate whether  $X_{ct}$  is sufficiently large to reject  $H_0$ , software will compute the probability of equaling or exceeding the observed test statistic for a chi-square distribution having  $(m-1) \cdot (k-1)$  degrees of freedom. This chi-square approximation for the  $p$ -value does not depend on the shape of individual variables but does require 5 or more expected counts in 80 percent or more of the cells (see next section). If this sample-size requirement is not fulfilled, a warning is usually output by the software. Permutation tests are available in R and other software that do not depend on this sample-size requirement and so are more valid for small samples.

To understand why there are  $(m-1) \cdot (k-1)$  degrees of freedom, when the marginal sums  $A_{ij}$  and  $C_{ij}$  are known, once  $(m-1) \cdot (k-1)$  of the cell counts,  $O_{ij}$ , are specified, the remaining cell counts can be computed. Therefore, only  $(m-1) \cdot (k-1)$  cell counts can be freely specified.

#### Example 14.1. Computing the contingency table test.

Three streams are sampled to determine if they differ in their macrobiological community structure. In particular, the counts 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, we infer that they differ in their pollution sources as well. Here we test whether the streams are identical in their proportion of pollution-tolerant organisms ( $H_0$ : row and column variables are not associated) or whether the three streams differ in this proportion ( $H_A$ : row and column proportions are associated).

$H_0$ : the proportions,  $a_1$ , of pollution-tolerant species and pollution-intolerant species,  $a_2$ , are the same for all three streams (proportions of data in rows are not associated with the column classification).

$H_A$ : the proportions,  $a_i$ , differ between (are associated with) streams.

The observed counts,  $O_{ij}$ , and marginal totals,  $A_i$  and  $C_j$ , are included in figure 14.2.

The expected counts,  $E_{ij}$ , assuming  $H_0$  is true, are computed using equation 14.3 and shown in figure 14.3.

		Stream 1	Stream 2	Stream 3	
Tolerant	4	8	12		$A_1 = 24$
	18	12	6		$A_2 = 36$
	$C_1 = 22$	$C_2 = 20$	$C_3 = 18$	$N = 60$	

Figure 14.2. The 2x3 matrix of observed counts for the data from example 14.1.

	Stream 1	Stream 2	Stream 3	
Tolerant	8.8	8.0	7.2	$A_1 = 24$
Intolerant	13.2	12.0	10.8	$A_2 = 36$
	$C_1 = 22$	$C_2 = 20$	$C_3 = 18$	$N = 60$

Figure 14.3. The 2x3 matrix of expected counts for the data from example 14.1.

Compare observed counts to expected counts to test whether the proportion of pollution-tolerant species significantly differs for the three streams. The test statistic,  $X_{ct}$  (eq. 14.4), is the sum of the squared cell residuals:

$$X_{ct} = \frac{(4.0 - 8.8)^2}{8.8} + \frac{(8 - 8.0)^2}{8.0} + \frac{(12 - 7.2)^2}{7.2} + \frac{(18 - 13.2)^2}{13.2} + \frac{(12 - 12)^2}{12} + \frac{(6 - 10.8)^2}{10.8} = 9.70 .$$

In R, perform the test using the `chisq.test` command. The `rep.int` function below repeats the first entry by the number of the second entry. For example, `rep.int(1, 22)` places the data value of 1 in the first 22 entries for the column `STREAM`, declaring that the first 22 observations are in column 1. The second and third uses of `rep.int` complete the column assignment with 20 observations for column 2 and 18 observations for column 3. In defining the `SPECIES` row variable, the text for the row name is repeated by the cell counts,  $O_{ij}$ . The `STREAM` and `SPECIES` variables then define the table of observed counts after being formatted as a table in R using the `xtabs` function. Print the formatted table with the `ftable` function.

```

> STREAM <- c(rep.int(1, 22), rep.int(2, 20), rep.int(3, 18))
> SPECIES <- c(rep.int("Tolerant", 4), rep.int("Intolerant", 18),
+   rep.int("Tolerant", 8), rep.int("Intolerant", 12),
+   rep.int("Tolerant", 12), rep.int("Intolerant", 6))
> SPECIES <- factor(SPECIES, levels = c("Tolerant", "Intolerant"))
> # rows now not in alphabetical order
>
> tab1 <- xtabs(~ SPECIES + STREAM)
> ftable(tab1)

          STREAM  1  2  3
SPECIES
Tolerant          4  8 12
Intolerant        18 12  6
> chisq.test(tab1)

```

Pearson's Chi-squared test

```

data: tab1
X-squared = 9.697, df = 2, p-value = 0.00784

```

The `chisq.test` function computes the test statistic  $X_{ct}$ , named “X-squared” in the output. The degrees of freedom  $(m-1) \cdot (k-1) = (2-1) \cdot (3-1) = 2$ . The  $p$ -value for  $X_{ct}$  is 0.00784, a small probability that the observed association between row and column variables is the result of only chance. Therefore, reject  $H_0$ . Conclude that the proportion of pollution-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 large-sample approximation  $p$ -value (see chap. 4) produced by the `chisq.test` function is accurate as long as all  $E_{ij} > 1$  and at least 80 percent of cells have  $E_{ij} \geq 5$  (Conover, 1999). When the second condition is not met, the following error message is produced by the `chisq.test` command: `expected value in a cell is too low`. When either condition is not met, there are two options.

1. Compute Fisher’s exact test. Use the `fisher.test` command in R. This is possible for smaller datasets. For the example data, the command and corresponding output are

```
> fisher.test(tab1)

Fisher's Exact Test for Count Data

data: tab1
p-value = 0.007201
alternative hypothesis: two.sided
```

2. Compute a permutation test for the contingency table setup. This is performed using the `simulate.p.value=TRUE` option of the `chisq.test` function. The constant,  $B$ , is the number of rearrangements performed for the test. We suggest using  $B=10,000$ .

```
> chisq.test(tab1, simulate.p.value = TRUE, B=10000)
```

Pearson’s Chi-squared test with simulated p-value  
(based on 10000 replicates)

```
data: tab1
X-squared = 9.697, df = NA, p-value = 0.006799
```

Although the expected counts,  $E_{ij}$ , were sufficiently large in this example, the permutation test can be used as a good approximation to the exact test for any sample size.

### 14.2.3 Location of the Differences

When a contingency table test finds an association between the two variables, we should determine how the two are related. Which cells are higher or lower in proportion to that expected when  $H_0$  is true, causing the rejection of  $H_0$ ? Use the individual cell residuals (prior to being squared) to answer this question.

Cells having large absolute values of the residual,  $\frac{(O_{ij} - E_{ij})}{E_{ij}}$ , are the cells contributing most to the rejection of the null hypothesis. The sign of the residual indicates the direction of the departure. For example, the individual cell residuals for the species data of the previous example are printed in R as follows:

```

> Xct <- chisq.test(tab1)
> Xct$residuals
      STREAM
SPECIES          1          2          3
Tolerant     -1.618080  0.000000  1.788854
Intolerant    1.321157  0.000000 -1.460593

```

Stream 3 has higher counts of the pollution-tolerant species  $O_{13}$  than the number expected if all three streams were alike ( $E_{13}$ ). This is evident because the residual for that cell equals 1.78. Similarly, stream 1 has many fewer of the tolerant species because the residual is a large negative. Therefore, stream 3 appears to be the most affected by pollution (large proportion of pollution-tolerant species), whereas stream 1 is the least affected (large proportion of pollution-intolerant species).

Contingency tables tests are designed for nominal data whose categories are not ordered. As noted at the start of the chapter, a contingency table test is not capable of recognizing the information contained in ordinal scales for rows or columns. The following two sections detail the appropriate methods when one or both variables have ordered values (such as low < medium < high).

## 14.3 Kruskal-Wallis Test for Ordered Categorical Responses

Chapter 7 introduced the Kruskal-Wallis test as a nonparametric test for differences in medians among three or more groups. The response variable in that case was continuous. Here the test is applied to ordinal data, where the response variable is 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

$H_0$ : the proportion of data in each response category (row) is the same for each group (column).

$H_A$ : the proportions in the rows differ among (is associated with) the groups.

### 14.3.1 Computing the Test

Organize the data as a matrix identical in format to that for a contingency table. The computations at the margins will differ (fig. 14.4). Compute the row sums,  $A_i$ , assigning ranks,  $R_i$ , to each observation in the table in accordance with the levels of the response variable. Ranks for all observations in the category with the lowest responses (response row 1 in fig. 14.4) will be tied at the average rank for that row, or  $\bar{R}_1 = (A_1 + 1) / 2$ . 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 figure 14.4, the equation for the average rank  $\bar{R}_2$  is  $\bar{R}_2 = A_2 + (A_2 + 1) / 2$ . For any of the  $i=1$  to  $m$  rows the average rank will equal

$$\bar{R}_i = \sum_{i=1}^m A_{i-1} + \frac{(A_i + 1)}{2} \quad (14.5)$$

where

$$A_0 = 0.$$

When the response variable categories express relative magnitude on a numeric scale, compute the Kruskal-Wallis test on the numerical values. The actual values are unimportant; all that matters is their relative order. For example, a rating of 10 for responses in the first row, 15 in the second row, and 28 in the third row will produce ranks identical to those had the responses been 2, 5, and 8. Only their relative ordering is preserved in the ranks.

	Group 1	Group 2	Group 3	
Response 1	$O_{11}$	$O_{12}$	$O_{13}$	$A_1 = \sum(O_{11} + O_{12} + O_{13})$
Response 2	$O_{21}$	$O_{22}$	$O_{23}$	$A_2 = \sum(O_{21} + O_{22} + O_{23})$
	$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	N

where

$$\bar{D}_1 = \frac{(O_{11}\bar{R}_1 + O_{21}\bar{R}_2)}{O_{11} + O_{21}} \quad \bar{D}_2 = \frac{(O_{12}\bar{R}_1 + O_{22}\bar{R}_2)}{O_{12} + O_{22}} \quad \bar{D}_3 = \frac{(O_{13}\bar{R}_1 + O_{23}\bar{R}_2)}{O_{13} + O_{23}}$$

**Figure 14.4.** A 2x3 matrix for a Kruskal-Wallis analysis of an ordered response variable.

Determine whether the distribution of proportions differs among the  $k$  groups (the  $k$  columns), by first averaging the column ranks to create  $\bar{D}_j$  (see fig. 14.4):

$$\bar{D}_j = \frac{\left( \sum_{i=1}^m O_{ij} \bar{R}_i \right)}{C_j} \quad (14.6)$$

where

$$C_j = \sum_{i=1}^m O_{ij} .$$

Now compute the Kruskal-Wallis test statistic  $K$  from these average group ranks (eq. 14.7). If  $H_0$  is true, the average ranks,  $\bar{D}_j$ , will all be about the same, and similar to the overall average rank of  $(N+1)/2$ . If  $H_0$  is not true, the average rank for at least one of the columns will differ.

$$K = (N-1) \frac{\sum_{j=1}^k (C_j \bar{D}_j^2) - N \left[ \frac{N+1}{2} \right]^2}{\sum_{i=1}^m (A_i \bar{R}_i^2) - N \left[ \frac{N+1}{2} \right]^2} \quad (14.7)$$

where

$C_j$  is the number of observations in column  $j$ ,

$\bar{D}_j$  is the average rank of observations in column  $j$ ,

$A_i$  is the number of observations in row  $i$ , and

$\bar{R}_i$  is the average rank of observations in row  $i$ .

To evaluate its significance, compare  $K$  to a table of the chi-square distribution with  $k-1$  degrees of freedom. After getting the data into a suitable format, all of these computations can be done using the R function `kruskal.test`.

#### Example 14.2. Kruskal-Wallis test for ordered responses.

An organic chemical is measured in 60 wells, which are each screened in one of three aquifers. The concentration is recorded only as being either above or below the reporting limit (RL). This is illustrated in figure 14.5. Note that we are repurposing the data from the first example, declaring the values in the first row to be nondetects designated as  $<\text{RL}$ , and counts in the second row to be detected values designated as  $\geq\text{RL}$ .

		$O_{ij}$		
		Aquifer 1	Aquifer 2	Aquifer 3
<RL		4	8	12
$\geq RL$		18	12	6
$C_1 = 22$		$C_2 = 20$	$C_3 = 18$	$N = 60$

**Figure 14.5.** The 2x3 matrix of observed counts for example 14.2. RL, reporting limit.

Does the distribution of the chemical differ among the three aquifers? To perform the Kruskal-Wallis test for these alphanumeric categories, we first assign a 0 to all first-row data ( $<RL$ ), and a 1 to all second-row data ( $\geq RL$ ). This creates numerical values that can be ordered and used by the Kruskal-Wallis test.

```

> conc.category <- c(rep.int(0,24), rep.int(1,36))
> aquifer <- as.factor(c(rep.int(1,4), rep.int(2,8), rep.int(3,12),
+ rep.int(1,18), rep.int(2,12), rep.int(3,6)))
> kruskal.test(conc.category~aquifer)

```

```

Kruskal-Wallis rank-sum test
data: conc.category by aquifer
Kruskal-Wallis chi-squared = 9.5354, df = 2,
p-value = 0.0085

```

The Kruskal-Wallis test statistic  $K$  (labeled *Kruskal-Wallis chi-squared* in the output) equals 9.53. The corresponding  $p$ -value, which is adjusted by the command for all data tied at the same rank within each row, equals 0.0085. Using a significance level,  $\alpha$ , of 0.05 (our default), reject  $H_0$ , finding different percentages between the aquifers of data above and below 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 can be easily done using the `p.adjust.method="BH"` option of the R function `posthoc.kruskal.dunn.test` in the `PMCMR` package (Pohlert, 2014) to compute Dunn's multiple comparison test, as presented in chapter 7, using an overall or family error rate equal to  $\alpha$ .

```

> library(PMCMR)
> posthoc.kruskal.dunn.test(conc.cat, aquifer, p.adjust.method="BH")

```

```

Pairwise comparisons using Dunn's-test for multiple
comparisons of independent samples

```

```
data: conc.cat and aquifer
```

1	2
2	0.153 -
3	0.006 0.145

P value adjustment method: BH

The triangular output of estimated  $p$ -values shows that the proportions of detects and nondetects for aquifers 1 and 3 significantly differ ( $p=0.006$ ), but neither aquifer's data is significantly different from the proportions in aquifer 2 at the 0.05 significance level.

## 14.4 Kendall's Tau ( $\tau$ ) for Categorical Data (Both Variables Ordinal)

If both row and column variables are ordinal, a contingency table would test only for differences in distribution of the row categories among the columns, but would ignore the ordering of row and column categories. The Kendall's  $\tau$  rank-correlation test will incorporate the directional structure of the categories, answering questions such as, "Do increases in the column variable coincide with increases or decreases in the row variable?"

### 14.4.1 Kendall's $\tau_b$ for Tied Data

Recall from chapter 8 that Kendall's correlation coefficient,  $\tau$ , is modified in the presence of ties and called  $\tau_b$ . With categorical data, there are generally many ties among the table cells, so  $\tau_b$  should be used. Kendall (1975) presented the formula

$$\tau_b = \frac{S}{\sqrt{\frac{1}{2} \left( N^2 - SS_a \right) \left( N^2 - SS_c \right)}} \quad (14.8)$$

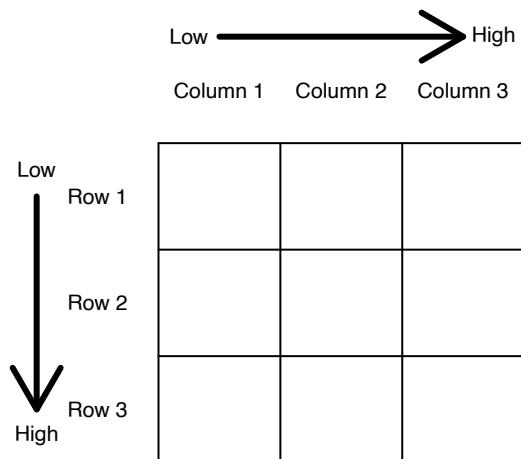
The numerator,  $S$ , for  $\tau_b$  is the number of pluses (P) minus the number of minuses (M)—see chapter 8. R computes and reports  $\tau_b$  when there are ties.

Consider a contingency table structure with the lowest values on the upper left, where 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 (fig. 14.6). With this format, compute the number of pluses (P) as the number of data in cells to the right and below each cell (fig. 14.7). Compute the number of minuses (M) as the number of data in cells to the left and below each cell (fig. 14.8). Do not count the numbers of data in cells of the same row or column, as these are tied in either the row or column variable and do not contribute to  $S=P-M$ .

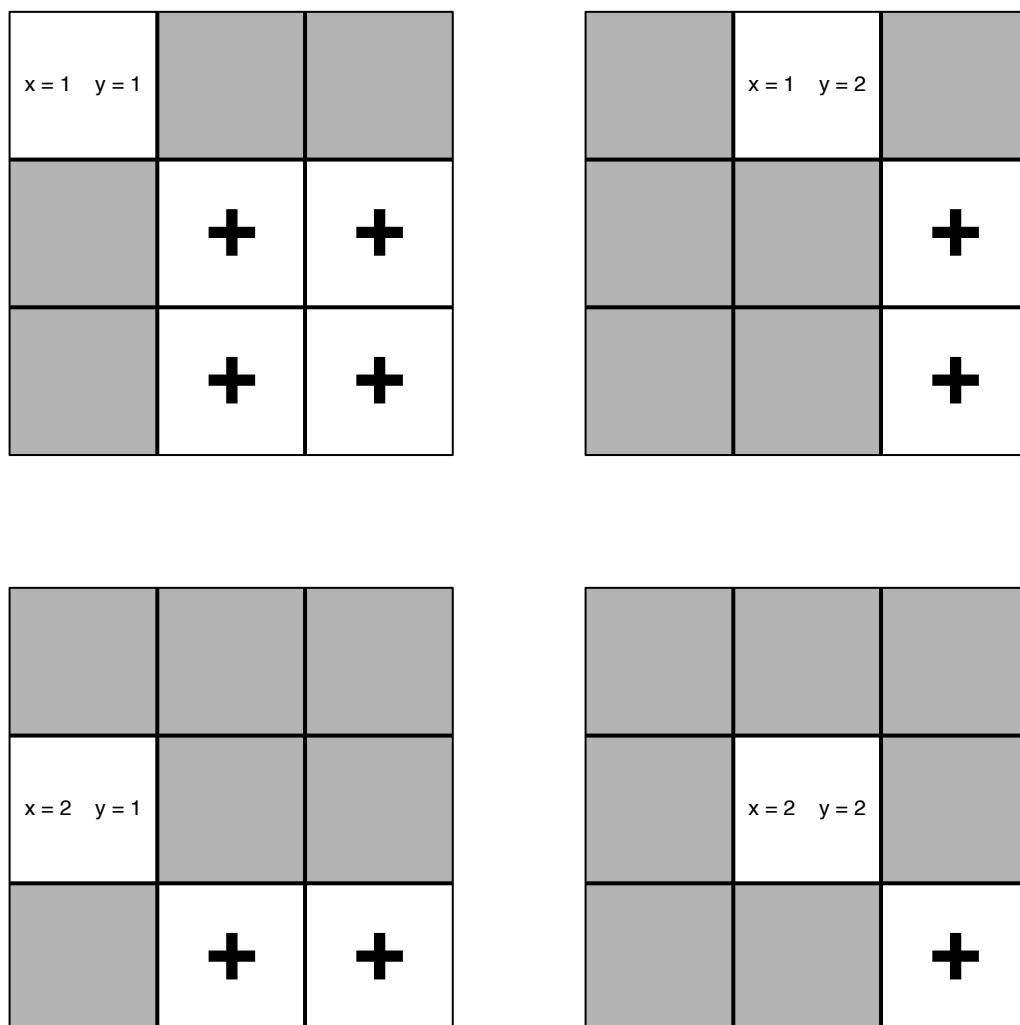
The denominator for  $\tau_b$  is not  $n \cdot (n+1)/2$  as it was for  $\tau$ . Instead, divide  $S$  by the number of untied comparisons. To compute this efficiently for a contingency table, use equation 14.9 to compute  $SS_a$  and equation 14.10 to compute  $SS_c$  (the sums of squares of the row and column marginal totals, respectively). Then insert the result into equation 14.8 to compute  $\tau_b$ .

$$SS_a = \sum_{i=1}^m A_i^2 \quad (14.9)$$

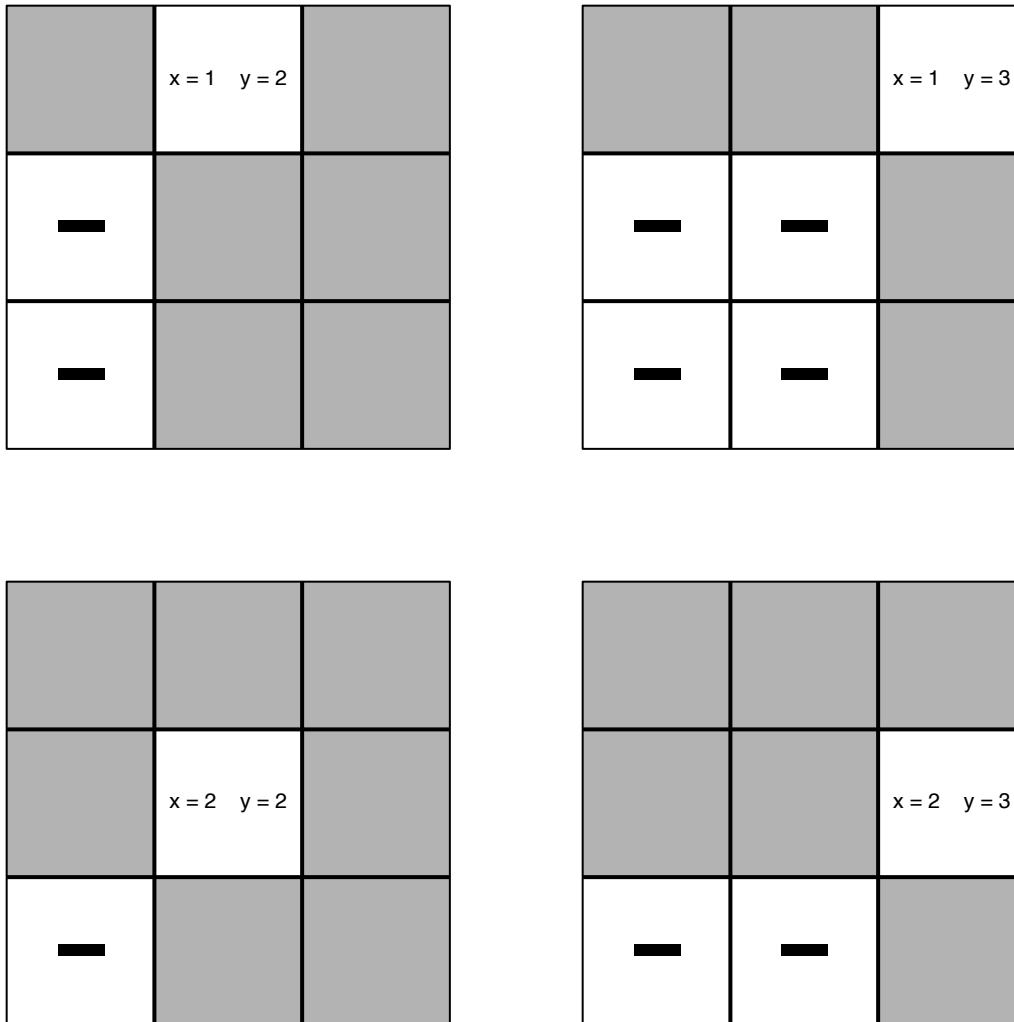
$$SS_c = \sum_{j=1}^k C_j^2 \quad (14.10)$$



**Figure 14.6.** Diagram showing suggested ordering of rows and columns for computing  $\tau_b$ .



**Figure 14.7.** Diagrams of 3x3 matrix cells contributing to  $P$  (column  $i > x$ , and row  $j > y$ ).



**Figure 14.8.** Diagrams of 3x3 matrix cells contributing to M (column i >x, and row j >y).

#### 14.4.2 Test of Significance for $\tau_b$

Determine whether  $\tau_b$  is significantly different from zero by dividing  $S$  by its standard error,  $\sigma_S$ , and compare the result to a table of the normal distribution. Agresti (2002) provides the following approximation to  $\sigma_S$ , which is valid for the number of plusses (P) and number of minuses (M) each larger than 100:

$$\sigma_S \cong \sqrt{\frac{1}{9} \cdot \left(1 - \sum_{i=1}^m a_i^3\right) \left(1 - \sum_{j=1}^k c_j^3\right) \cdot N^3} . \quad (14.11)$$

where  $a_i$  and  $c_j$  are the marginal probabilities of each row and column.

The exact formula for  $\sigma_S$  (Kendall, 1975) is more complicated. It is the square root of equation 14.9:

$$\sigma_s^2 = \frac{\left( n(n-1)(2n+5) - \sum_{i=1}^m A_i (A_i - 1)(2A_i + 5) - \sum_{j=1}^k C_j (C_j - 1)(2C_j + 5) \right)}{18} + \frac{\left( \sum_{i=1}^m A_i (A_i - 1)(A_i - 2) \right) \left( \sum_{j=1}^k C_j (C_j - 1)(C_j - 2) \right)}{9 \cdot N(N-1)(N-2)} + \frac{\left( \sum_{i=1}^m A_i (A_i - 1) \right) \left( \sum_{j=1}^k C_j (C_j - 1) \right)}{2 \cdot N(N-1)}. \quad (14.12)$$

If one variable were continuous and contained no ties, equation 14.12 would simplify to the square of the formula for the standard deviation of  $S$ , presented in section 8.4.2. Compute the test statistic  $Z_s$  as in chapter 8:

$$Z_s = \begin{cases} \frac{S-1}{\sigma_s} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sigma_s} & \text{if } S < 0 \end{cases}. \quad (14.13)$$

Compare  $Z_s$  to the  $\alpha/2$  quantile of the normal distribution to obtain the two-sided  $p$ -value for the test of significance on  $\tau_b$ . This will be computed by the `cor.test` function in R using the `method=kendall` and `continuity=TRUE` options.

#### Example 14.3. Kendall's tau for ordered row and column categories.

Pesticide concentrations in shallow aquifers were investigated to test whether their distribution was the same across wells located in three different soil-drainage classes. The null hypothesis,  $H_0$ , is that pesticide concentrations do not differ by soil-drainage class. The laboratory reported concentrations for the pesticide when levels were above the reporting limit, here categorized as  $>\text{RL}$ . For concentrations less than the reporting limit but above the instrument detection limit (DL), the compound was reported as present, and as  $<\text{DL}$  if concentrations were below the instrument detection limit. We now compute Kendall's  $\tau_b$  and its significance test. The data are organized as a contingency table in figure 14.9.

Soil Drainage					
Concentration	Poor	Moderate	High	$A_i$	$a_i$
Present	18	12	7	37	0.47
	5	10	8	23	0.29
	2	6	11	19	0.24
$C_j$	25	28	26	<b>79</b>	
$c_j$	0.32	0.35	0.33		<b>1.0</b>

Figure 14.9. The 3x3 matrix of observed counts for example 14.3. DL, detection limit; RL, reporting limit.

To simplify manual computation, organize the table so that variables increase in grouping level to the right and below. The number of pluses equals the number of comparisons between cells to the right and below. These are concordant comparisons where both row and column variables increase in grouping level. Comparisons within the same row and column are ties in one of the variables, so do not count them when computing  $\tau_b$ . Thus, the number of pluses (P) is  $P=18 \cdot (10+8+6+11)+12 \cdot (8+11)+5 \cdot (6+11)+10 \cdot (11)=1,053$ . Similarly, the number of minuses (M) equals the number of comparisons between cells to the left and below. These are discordant observations, with one variable increasing in grouping level and the second variable decreasing. The number of minuses is therefore  $M=12 \cdot (5+2)+7 \cdot (5+10+2+6)+10 \cdot (2)+8 \cdot (2+6)=329$ . So  $S=1,053-329=724$ .

To compute the denominator of  $\tau_b$ ,

$$SS_a = 37^2 + 23^2 + 19^2 = 2,259$$

$$SS_c = 25^2 + 28^2 + 26^2 = 2,085$$

and

$$\tau_b = \frac{724}{\sqrt{\frac{(79^2 - 2,259)(79^2 - 2,085)}{2}}} = \frac{724}{2034} = 0.3559$$

Is the value of  $\tau_b=0.3559$  significantly different from zero? From equation 14.11 the large-sample approximation for  $\sigma_s$  is

$$\begin{aligned}\sigma_s &\cong \sqrt{\frac{1}{9} \left( 1 - \left( 0.47^3 + 0.29^3 + 0.24^3 \right) \right) \left( 1 - \left( 0.32^3 + 0.35^3 + 0.33^3 \right) \right) \cdot 79^3} \cong \\ &\sqrt{\frac{(0.86)(0.89) \cdot 79^3}{9}} = \sqrt{42,329} = 205.74\end{aligned}$$

This results in a test statistic

$$Z_s \cong \frac{724-1}{205.74} = 3.51$$

and from a table of the normal distribution the two-sided  $p$ -value is  $p=0.0004$ . Therefore  $H_0: \tau_b=0$  is rejected and the positive value indicates that pesticide concentrations increase (the distribution shifts to a greater proportion of higher classes) as soil drainage increases.

It is much simpler to use software for these calculations. In R, the row and column variables `pest`, `category` and `soilclass` are defined with categories increasing from low to high in the order used by the `levels`= list below, though the ordering used for hand calculation convenience is not required for machine computations. Print the table to show what was accomplished using the `ftable` command.

```
> pest.category <- c(rep.int("<dl",37), rep.int("present",23),
+      rep.int(">dl",19))
> soilclass <- c(rep.int("Poor",18), rep.int("Moderate",12),
+      rep.int("High",7), rep.int("Poor",5),
+      rep.int("Moderate",10), rep.int("High",8),
```

```

+      rep.int("Poor",2), rep.int("Moderate",6),
+      rep.int("High",11))
> pest.category <- factor(pest.category,
+      levels = c("<dl", "present", ">dl"))
> # rows now not in alphabetical order
> soilclass <- factor(soilclass,
+      levels = c("Poor", "Moderate", "High"))
> # columns now not in alphabetical order
> Ex3.table <- xtabs(~pest.category + soilclass)
> # in table format for printing
> ftable(Ex3.table)

soilclass Poor Moderate High
pest.category
<dl          18       12      7
present       5        10      8
>dl          2        6     11

```

Convert the ordered categories to numeric values using the `as.numeric` function. For `soilclass`, this results in Poor = 1, Moderate = 2 and High = 3. Similarly for the three rows of `pest.category`, the `as.numeric` function results in <DL = 1, present = 2, and >DL = 3. Then compute Kendall's  $\tau_b$  (because there are ties) from the numeric values using the `cor.test` command.  $\tau_b$  equals 0.3559 with  $p$ -value 0.00044, agreeing with the hand computations above.

```

> soilclass.num <- as.numeric(factor(soilclass,
+      levels=c("Poor", "Moderate", "High")))
> pest.num <- as.numeric(factor(pest.category,
+      levels = c("<dl", "present", ">dl")))
> cor.test(pest.num, soilclass.num,
+      method="kendall", continuity=TRUE)

```

Kendall's rank correlation tau

```

data: pest.num and soilclass.num
z = 3.5141, p-value = 0.0004412
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.3559428

```

## 14.5 Other Methods for Analysis of Categorical Data

Categorical data may also be analyzed by loglinear modeling (Agresti, 2002). Loglinear models are prominent in the statistical literature for analysis of all three situations discussed in this chapter. Loglinear models transform the expected cell probabilities  $e_{ij} = a_i \cdot c_j$  by taking logarithms to produce a linear equation

$$\ln(e_{ij}) = \mu + \ln(a_i) + \ln(c_j), \quad (14.14)$$

where  $\mu$  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, and so on are possible using the loglinear model. Tests for higher-dimensioned matrices (such as a three-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 (2002) provides ample detail.

## Exercises

1. Chloride ion concentrations collected at the U.S. Geological Survey National Stream Quality Accounting Network (NASQAN) stations from 1974 to 1981 show more frequent increases than decreases. Smith and others (1987) classified 245 stations by their trend analysis results at  $\alpha=0.1$ . One reasonable cause for observed trends is road salt application. Estimates of tons of road salt applied to the 245 basins in 1975 and 1980 were used to place the stations into 3 road salt application groups: Down (1980 was more than 20 percent less than 1975), Up (1980 was more than 20 percent greater than 1975), and (little or) No change. The two variables are summarized in this 3x3 table:

Trend in Cl <sup>-</sup> (1974–81, $\alpha=0.1$ )				
Change in road salt application	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	<b>245</b>

Test  $H_0$ : A basin's trend in chloride ion concentration is not associated with its change in road salt application versus the alternative that they are associated.

- A. Use a contingency table. Interpret the test result.
  - B. Use Kendall's  $\tau_b$ . Interpret the test result.
  - C. State which test is more appropriate and explain why.
2. Fusillo and others (1985) sampled 294 wells in New Jersey for volatile organic compounds (VOCs). The wells were classified by whether they were located in an outcrop location near the surface, or whether they were located farther down dip 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. If a difference is found, which location has the higher probability of detected VOCs?

Location	Non-detects	Detected VOCs	Totals
Down dip	106	9	115
Outcrop	129	50	179
Totals	235	59	<b>249</b>

3. Switzerland began regulating organo-tin antifouling paints for boats in 1988. Concentrations in nanograms per liter (ng/L) of tributyltin (TBT, an organo-tin compound) 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?

Year	Number of samples		
	TBT $\leq 200$	TBT >200	Totals
1988	2	7	9
1989	9	13	22
1990	10	10	20
Totals	21	30	<b>51</b>

# Chapter 15

## Regression for Discrete Responses

---

*Concentrations of a volatile organic chemical are measured in wells across a large study area. About 75 percent 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 a result being above the reporting limit using a regression-like relation?*

*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 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 the use of a statistical model that is similar to ordinary least squares (OLS) in that the explanatory variables are continuous or discrete, but different from OLS in that the response variable is discrete (see fig. 4.1). Discrete (or categorical) response variables are often encountered when the measurement process is not sufficiently precise to provide a continuous scale. For example, instead of an estimate of concentration, only whether or not a sample exceeds some threshold, such as a reporting limit, health standard, or assessment criteria is recorded. Logistic regression is the most commonly used procedure for the situation where there are only two possible responses (for example, yes versus no). The equation predicts the probability of being in one of the two possible response groups.

Discrete response variables are commonly binary (two categories). For example, species of organisms or attributes of an organism are listed as either present or absent. Sections 15.1–15.3 deal with binary responses. Analysis of multiple response categories is discussed briefly in section 15.4.

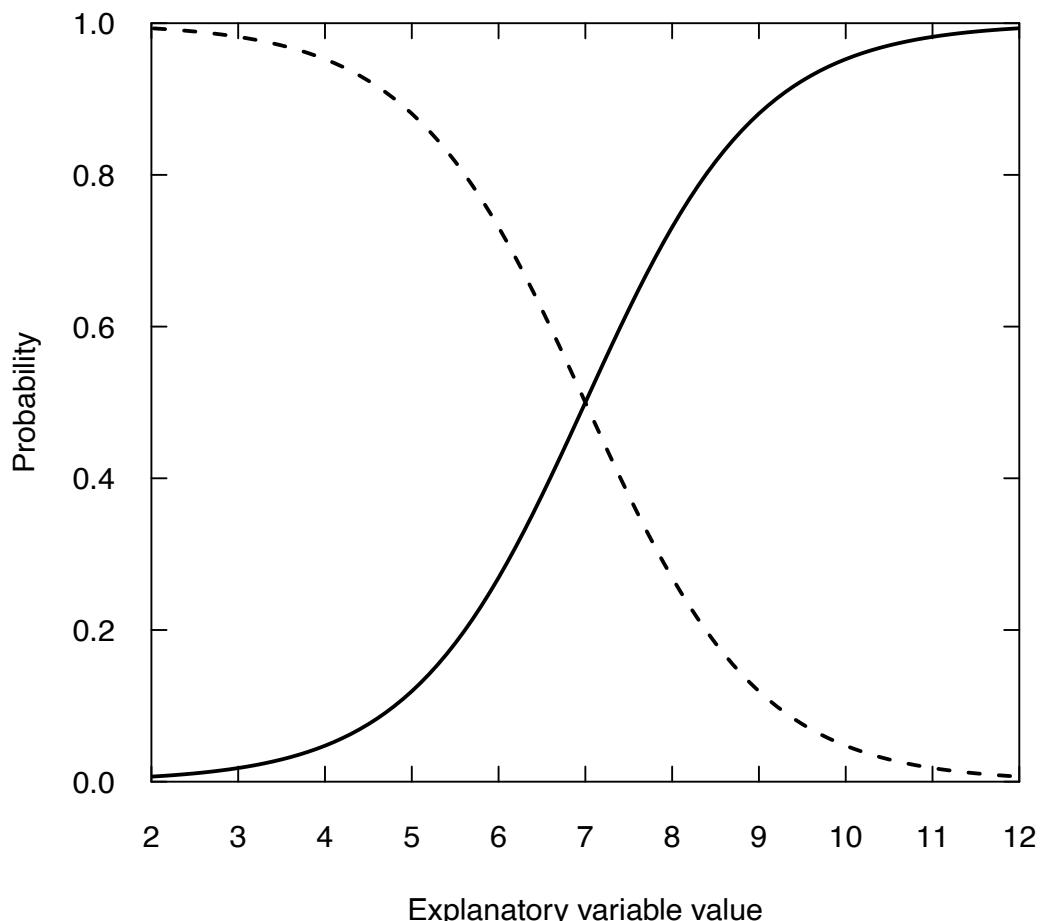
### 15.1 Regression for Binary Response Variables

With OLS regression, the magnitude of a response variable is modeled as a function of the magnitudes of one or more continuous explanatory variables. When the response is a binary categorical variable, however, the probability of being in one of the two response groups is modeled. The response variable is coded as either a 0 or a 1. For convenience we might code a lower group (for example, below a standard or below a detection limit) as 0 and code the higher group (for example, above a standard or a detection limit) as 1, but the choice is arbitrary and has no influence on the conclusions. The probability of the response having a value of 1 is  $p$ , and  $1-p$  is the probability of the response being a 0. In the various methods, what is predicted is typically the odds ( $p/(1-p)$ ) or the log of the odds. The explanatory variables may be either continuous, as in OLS regression, or a mixture of continuous and discrete variables similar to analysis of covariance.

An obvious approach to consider is to use OLS to estimate  $p$ , the probability of a response of 1. This would be a simple but inappropriate approach. Judge and others (1985), discuss several reasons why OLS should not be used for this estimation problem. The simplest reason is that the OLS regression model may result in predictions,  $\hat{p}$ , that fall outside of the range of 0 to 1. Probabilities outside of this range are meaningless.

### 15.1.1 The Logistic Regression Model

Logistic regression, also called logit regression, transforms this probability,  $p$ , into a response variable with values possible from  $-$  to  $+$  infinity ( $\infty$ ). Logistic regression models a function of the probability, the log odds, which in turns permits predictions to vary on the real line from  $-\infty$  to  $+\infty$ . The log odds is defined as  $\ln\left(\frac{p}{1-p}\right)$ , and the log odds can be transformed to  $p$ -values between 0 and 1. If  $y$  is the log odds, then  $p=e^y/(1+e^y)$ . A plot of  $p$  as a function of some explanatory variable typically has an S shape (fig. 15.1). 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). Ayotte and others (2016) provide an example of the use of logistic regression, and contrasts the results with other methods such as random forest classification and random forest regressions for estimating the probability of high arsenic levels in groundwater in the Central Valley of California based on geologic and hydrologic variables.



**Figure 15.1.** Graph of logistic regression equations; solid curve has a positive relation between the explanatory variable and  $p$ , the dashed curve has a negative relation. Note that estimates change more rapidly in the center than at the extremes.

### 15.1.2 Important Formulae for Logistic Regression

The odds are defined as the ratio of the probability of the specified event occurring (denoted as a 1) divided by the probability of it not occurring:

$$\text{odds ratio} = \left( \frac{p}{1-p} \right) \quad (15.1)$$

where  $p$  is the probability of the event occurring.

The log of the odds, or logit, transforms a variable constrained between 0 and 1, such as a proportion, into a continuous variable that can take on any value between  $-\infty$  and  $\infty$ . The logit can then be modeled as a linear function of one or more explanatory variables. The model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (15.2)$$

where the  $x$  values are the explanatory variables, there are  $k$  explanatory variables,  $\varepsilon$  is the error term, and the  $\beta$ 's are the true coefficients (there are  $k+1$  of them).

The estimated model is

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k \quad (15.3)$$

where  $b_0$  is the intercept and  $b_i$  is the slope coefficient for the  $i$ th explanatory variable. Equation 15.2 may be rearranged to view the odds as a function of an exponentiation of the linear predictors.

$$\left( \frac{p}{1-p} \right) = \exp(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k) \quad (15.4)$$

To return the predicted values of the response variable to original units (probabilities ranging from 0 to 1), the logistic transformation (the inverse of the logit transformation) is used:

$$p = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)} \quad (15.5)$$

For example, the multiple logistic regression equation with two explanatory variables would look like

$$p = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2)} \quad (15.6)$$

For a single  $x$  variable, the odds of obtaining a response of 1 increase multiplicatively by  $e^{b_1}$  for every unit increase in  $x$ . The inflection point of the curve is at  $-b_0/b_1$ ; the slope of the estimated probability is greatest at this point. Inflection points can also be computed for the case of multiple explanatory variables, but they need to be conditioned on specific values of the other explanatory variables. Toxicologists evaluating effects of toxicants on the mortality of some species call the inflection point the median lethal dose ( $LD_{50}$ ), which is the dose that results in a 50 percent chance of survival.

### 15.1.3 Estimation of the Logistic Regression Model

The estimation of a logistic model can be accomplished in R using the `glm` (general linear model) function in the base `stats` package (R Core Team, 2017). It is very similar in structure to the `lm` function used in regression models (also in the `stats` package). In the output, each coefficient estimate is given and a z-value for that estimate (the ratio of the estimate to its standard error) and from that a  $p$ -value indicates the probability that the estimate could be this much different from zero (in absolute value) under the null hypothesis that the true coefficient is zero. Unlike the `lm` function, it does not directly provide a statement

of the residual squared error ( $R^2$ ) value, but the critical information about the explanatory power of the model is contained in the null deviance, which is a measure of error with the null model (where there is only an intercept term and all slopes equal 0). The residual deviance is a measure of error using the full model. The only difference between the arguments used in the `glm` function compared to the `lm` function is that for logistic regression the `glm` function requires the specification of a family, which is binomial because the results for each realization can take on one of two possible values (0 or 1). So, the command in R would take the form:

```
> modLogistic <- glm(y ~ x1 + x2 + x3, data = mydata, family = binomial)
```

where `y` is the vector of results (0 for the case where the event did not occur and 1 for the case where it did occur); `x1`, `x2`, and `x3` would be three possible explanatory variables; and `mydata` is simply a data frame that contains the `y`, `x1`, `x2`, and `x3` values for the set of cases in the dataset, followed by `family = binomial`.

### 15.1.4 Hypothesis Tests for Nested Models

To compare nested logistic regression models, similar to the partial  $F$ -tests in regression, we can use the `anova` function (analysis of variance) in R to compare the residual deviance of the simpler model to that of the more complex model, and determine if the decline in residual deviance is more than we would expect by chance alone if the coefficients on the additional explanatory variables were actually zero. If there were two models fitted, where `model1` had `nf` fewer explanatory variables than `model2`, then we can conduct the test as follows:

```
> anovaResult <- anova(model1, model2)
> pvalue <- 1 - pchisq(anovaResult$Deviance[2], df = nf)
```

If `pvalue` is less than our specified  $\alpha$  value, then we should reject the null hypothesis that the additional explanatory variables have coefficients equal to zero.

In the case where `nf` = 1 (there is only one additional explanatory variable in the more complex model) then the  $p$ -value from this  $F$ -test will be equivalent to using the  $p$ -value on the specific coefficient.

A special case of testing nested models is the comparison of any given model to the null model (one where only the intercept term is nonzero). This can be done by specifying the null model formula as `y~1` (this is how an intercept-only model is designated in either the `glm` or `lm` functions). Then the same ANOVA test can be used to see if the model with one or more explanatory variables is significantly better than the null model.

### 15.1.5 Amount of Uncertainty Explained, $R^2$

A measure of the amount of uncertainty explained by the model is McFadden's  $R^2$ , or the likelihood- $R^2$ ,

$$R^2 = 1 - \frac{D_k}{D_0} \quad (15.7)$$

where  $D_k$  is the deviance of the  $k$  variable model and  $D_0$  is the deviance of the null model (intercept-only model). The likelihood- $R^2$  is uncorrected for the number of coefficients in the model, much like  $R^2$  in OLS regression.

McFadden's  $R^2$  is one of a class of pseudo- $R^2$ 's, considered pseudo because they are not equivalent to the  $R^2$  of OLS regression. They are pseudo- $R^2$ 's because they are calculated based on maximum likelihood estimates and are not designed to minimize variance. Pseudo- $R^2$  values are similar to  $R^2$  values in that they have a scale of 0 to 1, with higher values indicating a better model fit; however, they do not have the same interpretation as  $R^2$  in OLS (Smith and McKenna, 2013). Low pseudo- $R^2$ 's are typical in logistic regression and may be helpful in evaluating competing models but are not an effective measure of final model quality. The `glm` function of base R (used in the example below) does not report an  $R^2$ , but it is easily calculated from the results. The `lrm` function of the R package `rms` (Harrell, 2016) reports the Nagelkerke  $R^2$  index (Nagelkerke, 1991). Other software programs report other pseudo- $R^2$ 's and these pseudo- $R^2$ 's can vary greatly for the same model.

### 15.1.6 Comparing Non-nested Models

To compare two or more non-nested logistic regression models, these ANOVA procedures are not appropriate. For non-nested models a statistic related to Mallow's  $C_p$  is Akaike's information criterion (AIC). The AIC is the deviance plus two times the number of fitted coefficients. By this criterion the superior model is the one with the lowest AIC. There are several other approaches to select the best model, including an adjusted pseudo- $R^2$  approach and also the BIC (Bayesian information criterion). We will use the AIC here, but all are reasonable approaches.

Hosmer and Lemeshow (2000) indicate that a true measure of model fit is the comparison of observed to predicted values from the fitted model. One measure of the ability of a model to predict observations is the Brier score ( $BS$ ), modified for binary forecasts (Brier, 1950; Ferro, 2007):

$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2 \quad (15.8)$$

where  $n$  is the number of observations,  $p_i$  is the probability of an outcome of 1 for observation  $i$ , and  $o_i$  is the actual outcome (0 or 1) for observation  $i$ . The Brier score ranges from 0 to 1 and a perfect model would have a Brier score of 0, whereas a model that gets every prediction wrong would have a Brier score of 1.

There are numerous additional methods for assessing models (Harrell, 2016; Hosmer and Lemeshow, 2000), including receiver operating characteristic curves that plot the true positive rate against the false positive rate, where the area under the curve (AUC) is scaled from 0 to 1 and areas close to 1 indicate that model has good predictive ability (Green and Swets, 1966; Brown and Davis, 2006).

#### Example 15.1. Trichloroethylene (TCE) detections in groundwaters on Long Island, New York.

Eckhardt and others (1989) reported a pattern of occurrence for several volatile organic compounds in shallow groundwaters on Long Island, New York, and related them to the population density in the area near the well. Trichloroethylene (TCE) detections for 643 samples are listed in table 15.1, which shows, for each population density class, the number of measured concentrations above the reporting limit of 3 parts per billion (ppb) and the number below the reporting limit. The logistic regression model to be considered here is that the probability of a measured concentration above 3 ppb is a function of population density.

**Table 15.1.** Trichloroethylene data in the Upper Glacial Aquifer, Long Island, New York.

Population density, persons per acre	Number > 3	Number $\leq 3$	Total number of samples	Mean frequency
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
<b>Totals</b>	<b>117</b>	<b>526</b>	<b>643</b>	<b>18.2</b>

The single explanatory variable, population density, is called `popDens` and the response variable is called `detect` (defined as 0 for a concentration  $\leq 3$  ppb and a 1 for concentrations  $> 3$  ppb). The model is fitted using the following R command:

```
> glm(formula = detect ~ popDens, family = binomial)
```

The output is shown here

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.80093	0.20827	-13.448	<2e-16 ***
popDens	0.22559	0.02706	8.337	<2e-16 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 610.02 on 642 degrees of freedom

Residual deviance: 533.00 on 641 degrees of freedom

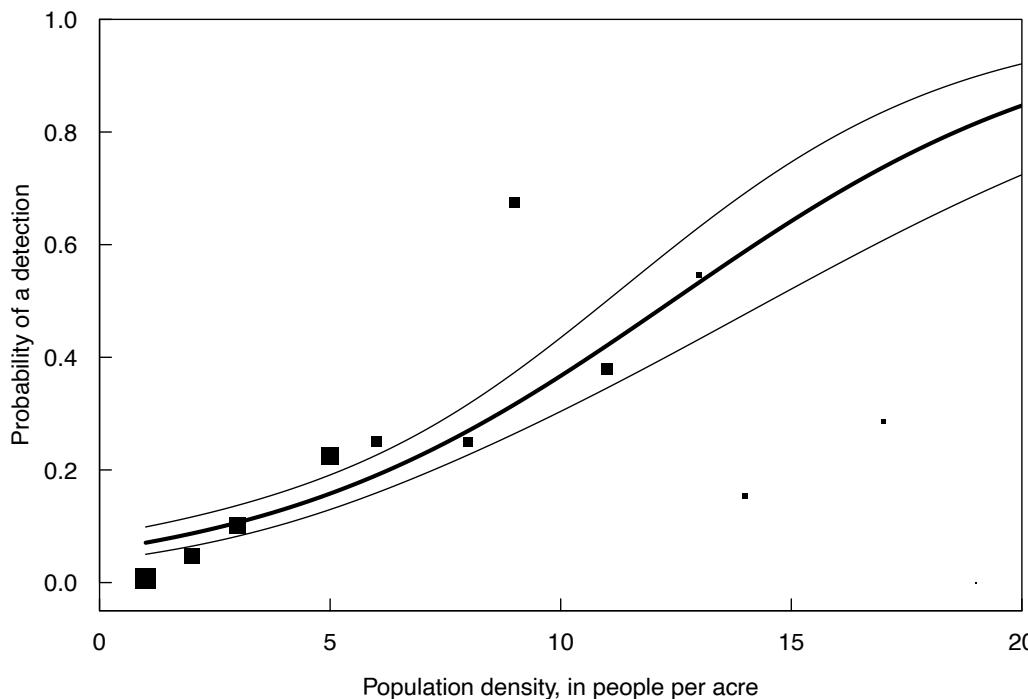
AIC: 537

Using equation 15.7, we see that this has a McFadden's  $R^2$  of  $0.126 = 1 - (533/610)$ . Stated in words, we can say that the model is not highly effective at predicting if the concentrations are above the reporting limit, but we can also test whether or not the `popDens` variable improves the model beyond what could be attained from the intercept-only model. The test statistic is the difference between the null deviance and the residual deviance and that value is 77. We can compare that value to the  $\chi^2$  distribution with 1 degree of freedom. The critical value of the test statistic at  $\alpha=0.05$  is 3.84, so we can reject the  $H_0$  that the coefficient on `popDens` = 0. In fact, the  $p$ -value for this result is  $<2 \cdot 10^{-16}$ . Thus, even though we find the model to be rather imprecise we can say, with very high confidence, that population density is related to the probability of detection. A plot of predicted probabilities from the logistic regression model with a 95-percent confidence band is shown in figure 15.2. See the code in the supplementary material (SM.15) to determine how these confidence intervals are computed.

The positive slope coefficient for `popDens` means that, under the proposed model, 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 or at 14, 17, or 19. But in the range of densities where most of the wells are located, the fit is reasonably strong.

### Example 15.2. Redox conditions in the Chesapeake Bay watershed.

Tesoriero and others (2015) developed a logistic regression model to evaluate the probability of low dissolved oxygen (DO) concentrations (<2 milligrams per liter [mg/L]) in groundwater samples collected in the Chesapeake Bay watershed. The existence of low dissolved oxygen in groundwater is important in this watershed because these conditions are conducive to denitrification, which is the microbially facilitated conversion of nitrate (an important driver of eutrophication in the Bay) to nitrogen gas (which will move to the atmosphere and not contribute to eutrophication). If areas of potentially high DO can be identified based on hydrogeologic characteristics, then this information can be used to identify areas where more effort to limit nitrogen inputs should be targeted to help mitigate eutrophication. The data used here are only a few of the explanatory variables in the model developed by Tesoriero and others (2015) and include only the 1,805 observations that contained no missing values, the so-called training dataset. See Tesoriero and others (2015) for examples of how a training dataset can be used along with an independent dataset to obtain estimates of the predictive power of the model and also learn about other diagnostics suited to logistic regression, such as the Hosmer-Lemeshow Test and AUC metric of performance.



**Figure 15.2.** Graph of estimated trichloroethylene (TCE) detection probability as a function of population density, showing 95-percent confidence intervals. Mean observed probability, by density category, is shown as squares. Area of square is proportional to the sample size for each density class.

The dataset used here contains the following explanatory variables. Recharge (Rech) in millimeters per year (mm/year), porosity (Por) in percent, and four categorical variables to represent five surficial geology types: fine coastal plain, coarse coastal plain, crystalline, carbonate, and siliciclastic. The four categorical variables used to denote these five types are “fine,” “coarse,” “cryst,” and “carb.” In each case, they are coded as a value of 1 for the specified geologic types, -1 for siliciclastic, and 0 for all other types. The use of the -1 value makes it possible to designate the five geologic types using only four explanatory variables. If five variables were used (one for each rock type) then the explanatory variables would be

$$\text{colinear. The dependent variable is coded as } \text{lowDO} = \begin{cases} 1 & \text{if } DO \leq 2 \text{ mg/L} \\ 0 & \text{if } DO > 2 \text{ mg/L} \end{cases}$$

We can compare a model that only considers the geologic types as explanatory variables (called modGeo) and then consider a model that contains those variables plus recharge and porosity (called modAll). In each case a positive coefficient in the model indicates a higher probability of low DO.

The R code to fit the modGeo would be

```
> modGeo <- glm(lowDO ~ carb + coarse + cryst + fine,
+     family = binomial, data = new2)
> summary(modGeo)
```

Call:

```
glm(formula = lowDO ~ carb + coarse + cryst + fine, family = binomial,
    data = new2)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.4288	-1.1912	-0.6029	1.1623	1.8946

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.47144	0.05654	-8.337	< 2e-16 ***
carb	-0.91486	0.13490	-6.782	1.19e-11 ***
coarse	0.50700	0.09237	5.489	4.05e-08 ***
cryst	-1.14157	0.12668	-9.012	< 2e-16 ***
fine	1.04546	0.09063	11.536	< 2e-16 ***
---				
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2471.7 on 1801 degrees of freedom  
 Residual deviance: 2196.4 on 1797 degrees of freedom  
 AIC: 2206.4

Number of Fisher Scoring iterations: 4

What we see here is that the coefficients in the model are all highly significant although the amount of deviance explained is rather modest. It is the equivalent of a McFadden  $R^2$  value of 11 percent. Now, if we add the two additional explanatory variables (por and Rech) we get these results.

Call:

```
glm(formula = lowDO ~ carb + coarse + cryst + fine + Por + Rech,
family = binomial, data = new2)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.9878	-0.9278	-0.5291	0.9445	2.2570

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.947532	1.194036	-4.144	3.42e-05 ***
carb	-1.122650	0.147558	-7.608	2.78e-14 ***
coarse	1.063700	0.139319	7.635	2.26e-14 ***
cryst	-1.462441	0.139562	-10.479	< 2e-16 ***

```

fine      1.425439   0.109048  13.072 < 2e-16 ***
Por       0.136731   0.024409   5.602 2.12e-08 ***
Rech     -0.007813   0.001110  -7.038 1.95e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 2471.7 on 1801 degrees of freedom
Residual deviance: 2063.8 on 1795 degrees of freedom
AIC: 2077.8

```

All of the explanatory variables have coefficients that are significant at the 0.001 significance level and the McFadden  $R^2$  value increases to 17 percent. The AIC value is lower for the more complex model, which suggests that it is a better model. We could do a formal likelihood ratio test, but simply considering the  $p$ -values on the additional coefficients and the lower AIC value is sufficient grounds to select the latter model.

## 15.2 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.

### 15.2.1 Discriminant Function Analysis

Discriminant function analysis is used as a multivariate classification tool to decide which of several groups a response is most likely to belong to (Johnson and Wichern, 2002). The probabilities of being in each of the groups are 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. Each group is assigned an integer value. The goal of this approach is identical to the goal 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 to departures from these assumptions (Press and Wilson, 1978). Therefore, logistic regression is preferred when multivariate normality and equality of variances cannot be assumed, as is the case for most of the data found in water resources.

### 15.2.2 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 (Dietz, 1985). The rank-sum test is equivalent to the significance test for Kendall's  $\tau$  between the binary  $y$  variable and a continuous  $x$ .

The rank-sum test can be used instead of the likelihood ratio test 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.

## 15.3 Logistic Regression for More Than Two Response Categories

In water resources applications, response variables may 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 may be computed, with multiple thresholds differing by offset intercepts in logit units. When responses are nominal (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 dependent. An example in economics would be a consumer's probability of selecting an apple versus an orange versus a banana. In this case, independent logit models may be fit for evaluating the probabilities of being in each group. The approach here is called nominal response models. They have not seen much use in water resources data analysis and are not discussed further in this text.

Categorical response variables sometimes represent an underlying continuous variable that cannot be measured with precision sufficient to provide a continuous scale. For example, concentration data may be discretized into three categories based on two thresholds (see below). Biologic activity may be categorized as not affected, slightly, or severely affected by pollution. The resulting multiple responses  $y_i$ ,  $i=1, 2, \dots, m$  are ordinal, so that  $y_1 < y_2 < \dots < y_m$ . For example, suppose three responses are possible:

- 0: Concentrations are below the reporting limit (a reporting limit is a concentration value below which the laboratory lacks confidence to state a specific concentration value).
- 1: Concentrations are above the reporting limit but below a health standard.
- 2: Concentrations are above the health standard.

This type of problem may be addressed using ordinal logistic regression. It is designed to estimate the probabilities of the random variable being less than each of two or more thresholds, and to evaluate the relation between those probabilities and one or more explanatory variables. This text will not address the theory behind this method. An R function that is appropriate in this situation is the function `polr` in the `MASS` package (Venables and Ripley, 2002). The name `polr` comes from proportional odds logistic regression. These methods are discussed in the texts Agresti (2002) and Venables and Ripley (2002). One hydrologic application is Ayotte and others (2012), which deals with the probability of arsenic in New Hampshire groundwater exceeding various thresholds based on geologic and other information about the aquifer from which the water is withdrawn.

## Exercises

- 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 best predicts contamination. The factors were

Factor	Possible scores (discrete levels)
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 dataset **Impoundment.RData**, in SM.15, 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?

- The Des Plaines River in the suburbs of Chicago, Illinois, has been experiencing increasing chloride concentrations over recent decades and the cause for this is almost certainly the increasing use of salt as a road-deicing compound in the winter. The Environmental Protection Agency's chronic water quality criteria for chloride is 230 milligrams per liter (mg/L) and the data available for the Des Plaines River generally indicate that this threshold seems to be exceeded with increasing frequency over time. The dataset (the object called **Sample**, stored in the file **DesPlainesChloride.Rdata** in SM.15) contains 194 observations spanning the years 1987 through 2009. The dataset contains the following variables: **Date**, **DecYear** (decimal year), **Q** (discharge in cubic meters per second [ $m^3/s$ ]), **Conc** (chloride concentration in mg/L), and **Month** (month of the year). The file also contains a metadata object called **INFO** providing information about the watershed, site, and the variable of interest.

Using these data, build a logistic regression model that estimates the probability of concentrations exceeding 230 mg/L as a function of year and time of year. Then build a model using year, time of year, and discharge. In both cases, you may want to consider multiple ways of expressing some of these explanatory variables. Which of these do you think is the best model of the probability of exceeding the criterion of 230 mg/L? Can you find a graphical way of expressing the results?



# 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 can be worth a thousand table entries.*

For oral presentations, only the simplest tables are likely to be effective in presenting information. Listeners are not familiar with the data and have not pored over them for many hours as has the presenter. Numbers are often not readable to those seated beyond 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 before the presentation. This 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 discusses. Figures should be a visual abstract of the report and provide one of the best ways to convince someone to take enough time to read your work. Such graphics give evidence that the data support the conclusions you have reached. In fact, many scientific journals now provide the option of publishing a graphical abstract for precisely this reason.

However, not all graphs are created equal. Some present quantitative information clearly and precisely. Others are not as effective and may even be misleading. Not only do researchers need to be aware of the effectiveness of their graphics, but they are often beholden to the graphical standards of their organization or of a publication outlet. These standards do not always keep up with current tools and research; however, they do help to standardize the presentation of results. Fortunately, since the 1992 publication of Statistical Methods in Water Resources, advancements have been made in the tools available to create effective graphics and in related cognitive science research. Because of these advancements, many excellent references specific to graphics are now available.

Creating effective graphics can be very time consuming. Increasing one's skill in generating graphs can save time in the end. Some references that are effective for learning graphics commands in R include, R Graphics (Murrell, 2011), Lattice: Multivariate Data Visualization with R (Sarkar, 2008), R Graphics Cookbook (Chang, 2013), Graphical Data Analysis with R (Unwin, 2015), and ggplot2: Elegant Graphics for Data Analysis (Wickham, 2016).

Understanding the strengths and weaknesses of various types of graphs is important when choosing the most appropriate way to present data. Three authors stand out in their evaluation of graphs for quantitative data. Cleveland (1985) discusses the ability of the human eye-brain system to process information. Schmid (1983) wrote a handbook listing numerous examples of both good and bad graphics. Finally, Tufte (1983, 1990, 1997, 2006) has a body of work that contributes to thinking about effective graphics and that describes the artistry involved in creating graphics. This chapter draws on ideas from these three authors and others.

### 16.1 The Value of Presentation Graphics

Graphs were introduced with great enthusiasm and effect by early developers. William Playfair invented the line graph, bar chart (Playfair, 1801a), pie chart, and circle graph (Playfair, 1801b). Charles Joseph Minard created what has been called one of the best statistical graphics ever (Tufte, 2006), titled "Map of the successive losses in men of the French Army in the Russian Campaign 1812–1813." This map

is used to highlight “fundamental principles of analytical design” in Tufte (2006). Florence Nightingale’s 1858 coxcomb diagram (Nightingale, 1858) contributed to improved sanitary conditions for the British Army, and she “was arguably the first to use this and other statistical graphs for political persuasion and popular impact” (Friendly, 2008). Early makers of statistical graphics seemed to understand the need for clarity of context and purpose better than many modern practitioners do. Their goal was to distill and simplify information, while we, with ever more sophisticated tools, keep adding complexity (color, perspective, numbers, graphics within graphics, and so on).

In 1915, attempts were being made by “engineers, economists, statisticians, psychologists, mathematicians, accountants, educators, and others” to standardize graphics (Goldenweiser, 1916). They divided graphics into three use categories that are still relevant today—printed publication, lectures, and exhibits. Each set of graphics fills a different need and, although a project may result in a print publication, a presentation, and an exhibit such as a poster, the same graphics are not always appropriate for each method of presentation. How many times have scientists heard during an oral presentation, “I know you can’t read this, but...” or “Excuse this graphic, I know you can’t read it in the back, but...”? One needs to first carefully consider the purpose of his or her graphic and the presentation venue, and then incorporate the information into creating that graphic.

Graphs can clarify complex interrelations between variables. They can picture the signal over and above the noise, letting the data tell its story. In chapter 2, graphs for exploratory data analysis were discussed. These same methods that provide insight to an analyst will also illustrate important patterns and contrasts to the audience. As Alexander von Humboldt said in 1811, “Whatever relates to extent and quantity may be represented by geometrical figures. Statistical projections which speak to the senses without fatiguing the mind possess the advantage of fixing the attention on a great number of important facts” (von Humboldt, 1811).

Complex tables, although useful for referencing specific numbers, rarely allow easy extraction of a data signal. In general, a good graph will contribute much more to audience understanding than a table, regardless of audience backgrounds and interests. Appropriate graphs often lead to new understanding or to better decisions.

## 16.2 General Guidelines for Graphics

As described in the preface, all of the authors have been involved in teaching a course titled “Statistical Methods for Environmental Data Analysis.” The course includes a section on presentation graphics and the following general guidelines have developed over the history of the course taught at the U.S. Geological Survey National Training Center.

- Maximize the information-to-ink ratio (Tufte, 1983). That is, do not add solid blocks of color, images, or additional features to the graphic unless the additions support the conveyance of information; otherwise these additions are considered chart junk.
- Avoid perspective. Just because one’s software can add perspective does not mean one should do it.
- Axes should start at zero when absolute magnitude is important (Robbins, 2005; Strange, 2007; Kelleher and Wagener, 2011).
- Consider the precision of visual perception (more discussion follows in the next section).
- When using a logarithm transformation, the axis scale should be in original units, not logarithms.
- In scatterplots, the errors should be shown in the vertical direction.
- Connect points in a plot only if the measurements are continuous. For example, annual peak-streamflow values are not continuous, and peaks should not be connected with a line. A loess smooth could be added to show a general pattern in the observations.
- Graphics for a presentation should generally be different from the graphics for a report, as the viewer has less time to study each graphic in a presentation and the text in explanations and captions can be more difficult to read on a screen.

To evaluate whether the guidelines you work under are sufficient (or excessive), be sure that your graphs are reviewed by colleagues and by nontechnical persons if the graph will find its way to the public. Request a summary of conclusions drawn from the graph(s) and revise when the conclusions do not match your intentions for the graph(s). Review of graphs is equally as important as review of text.

## 16.3 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 could perform various tasks of perceiving and processing graphical information. The purpose was to rank tasks necessary for interpreting common graphs, such as bar and pie charts, in order to understand which types of graphs are most effective at presenting information. Before 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 (p. 535–6).

Cleveland and McGill then ranked perceptual tasks based on accuracy, as determined by the number of correct judgments of identical data displayed by different graphs (table 16.1). 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 less precision is required to illustrate the main points and the audience is less technically inclined, graphs that make use of tasks found lower in the table may convey the information more effectively.

### 16.3.1 Color

Color can enhance the ability to read graphs precisely and accurately, as well as enhance the interest of the reader. However, it can interfere in judgments of size between areas of different colors and by biasing the interpretation of results (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). Areas shaded a bright red on a map, as is commonly done in maps related to pollution studies, will appear larger than they would if shaded another color or with a pastel such as light pink.

Pastels can 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, puts all areas on an equal footing. Of course, this minimizes the newspaper graphics effect of attracting attention to the graph, but enhances the graph's ability to portray information.

**Table 16.1.** Precision of perceptual tasks summarized from figure 1 of Cleveland and McGill (1984a).

Degree of precision	Elementary perceptual tasks
More precise	Position along a common scale
	Positions along nonaligned scales
	Length, direction, angle
	Area
	Volume, curvature
	Shading, color saturation

Color can also be quite helpful at 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, colored lines allow better perception than solid versus patterned lines.

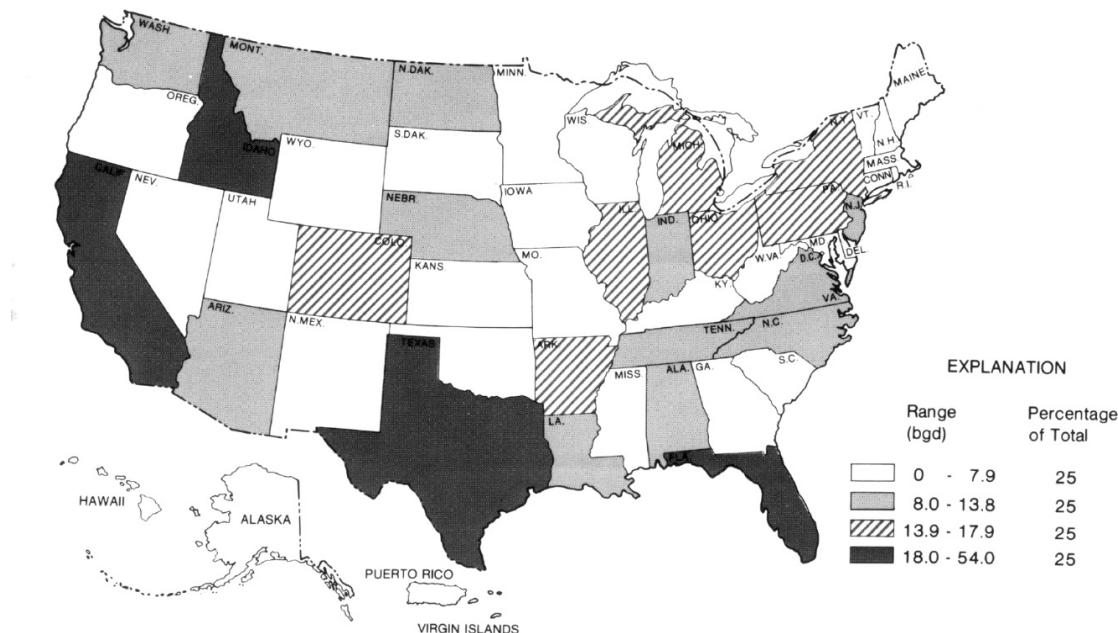
Caution is warranted when employing many different colors in a graphic because the maximum number of colors humans can differentiate is from 8 to 12 (Streit and Gehlenborg, 2014). Additionally, when creating graphics one should consider the possibility of having color-blind audience members. About 5 percent of men (as much as 8 percent of men and 0.5 percent of women of Northern European descent) have the most common form of color blindness—red-green color blindness (National Eye Institute, 2015; Bernhard and Kelso, 2018). These individuals may have a difficult time distinguishing items with red and green colors without other indicators, such as shape or position. Less common forms of color blindness include the inability to distinguish blue and yellow and complete inability to see color.

Software can simulate what images might look like to someone with color blindness. Adobe Illustrator has a method of “soft-proofing for color blindness.” Color Oracle is a software program one can use to view any image on the screen under differing forms of color blindness (Bernhard and Kelso, 2018). Coblis is an online simulator that an analyst can use by uploading a graphic (<http://www.color-blindness.com/coblis-color-blindness-simulator/>). The simulator creates examples of different types of color blindness with the graphic provided.

### 16.3.2 Shading

A common use of shading—shaded maps where the color saturation and lightness indicates the magnitude of a single variable—is shown in figure 16.1. Such maps may be of the entire country, a state, or a study area. These chloropleth maps are inherently difficult to correctly interpret; however, with increased access to geographic information systems, chloropleth maps, despite their shortcomings, may even be more popular now than when this map was published in 1983.

The first challenge with this map 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 16.1, Texas stands out not only because it is dark, but



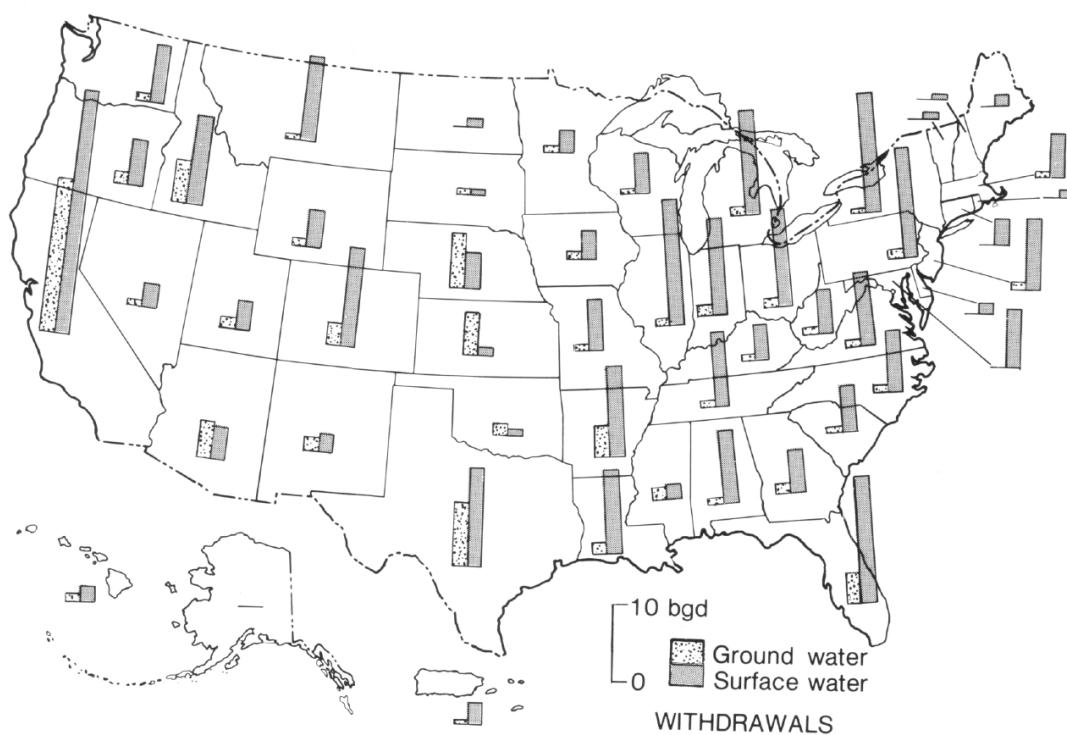
**Figure 16.1.** Map showing total offstream water withdrawals by state, 1980, from Solley and others (1983). Bgd, billion gallons per day.

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 (N.J.). 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 16.1, a state with a higher population like New Jersey may be far more important than is Montana, a state with a much smaller population. The weighting given to larger areas on a shaded map is often inappropriate to the data being illustrated.

Another challenge is that all variability within areas is 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 United States will produce greater impressions on the viewer than do smaller eastern counties.

Lastly, 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 also degrade as graphs are reproduced on a printer or 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 13.9–17.9 class in figure 16.1. Such patterns quickly become very confusing, and actually reduce the eye's ability to distinguish classes of data. One must also be careful to use a series of patterns whose color saturation and lightness changes with the data. Figure 16.1 seems to violate this rule, as the shade of the second class (8.0–13.8) appears darker than the third striped pattern.

An alternative to figure 16.1, figure 16.2 also uses a map to display the geographic distribution of the data, but adds bars depicting data classes within each state. With bars, the perceptual task is a judgment of length without a common datum resulting in the ability to distinguish more than five levels. However, it is often difficult to place the bars within state boundaries. Another alternative is to abandon the map background and construct bars or other graphics for each state. When abandoning maps, however, much of the regional perspective is sacrificed for state-by-state precision.



**Figure 16.2.** Map of withdrawals for offstream water use by source and state, from Solley and others (1983). Bgd, billion gallons per day.

### 16.3.3 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 (fig. 16.3). 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 16.3, it is only possible to see that the flow in the spring season is much larger than the other seasons. Determining whether the fall fraction is larger than the winter fraction is difficult, and getting a relative sense of how much bigger the summer fraction is than either the fall or the winter fraction is also difficult.

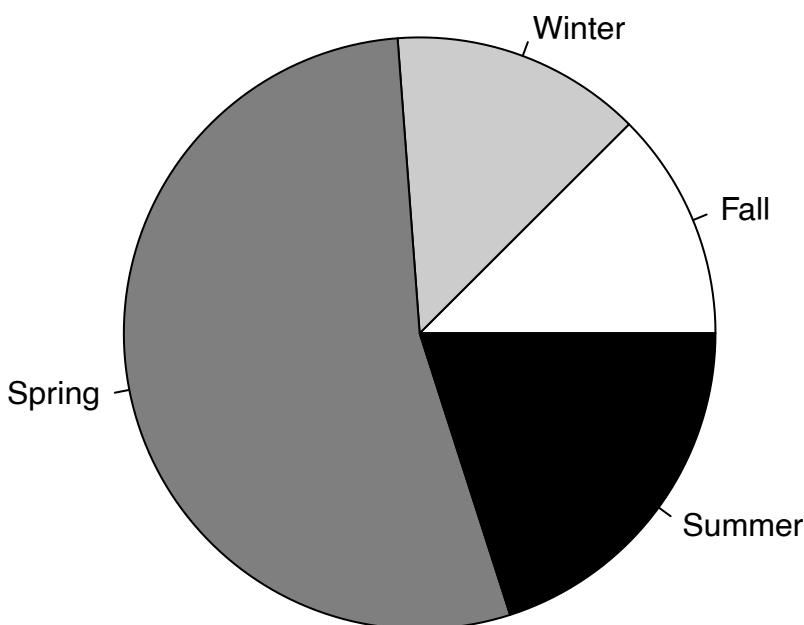
Pie charts have been used for more than 200 years (Spence, 2005) and, therefore, have an enduring appeal. However, it is always possible to replace a pie chart with a figure using one of the higher perceptive tasks in order to improve graphical precision. For the same data in figure 16.3, figure 16.4 presents a bar chart. Now the perceptive task is of location along a common scale (the *y*-axis), and all differences are more clearly seen. The four seasons 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, as length is a much more accurate visual cue than the angle-based approach of pie charts (Streit and Gehlenborg, 2014).

Pie charts have little utility for scientific publications because of their lack of precision. In addition, areas or volumes beyond pie charts, such as squares or cubes, are also more likely to be misinterpreted than linear bar charts (Joint Committee on Standards for Graphic Presentation, 1915). Perspective should not be added to turn bars into rectangular prisms (see section 16.4.1. on perspective).

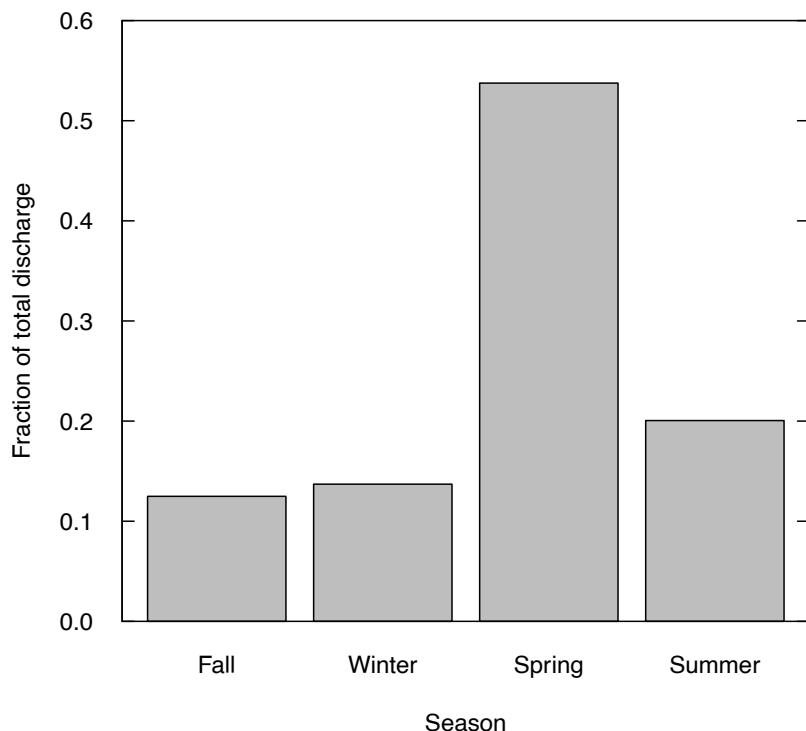
### 16.3.4 Angle and Slope

Judgments of angle and slope occur when comparing two curves (Cleveland and McGill, 1984a). 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. A good rule of thumb is that if differences are of interest, plot the differences directly.

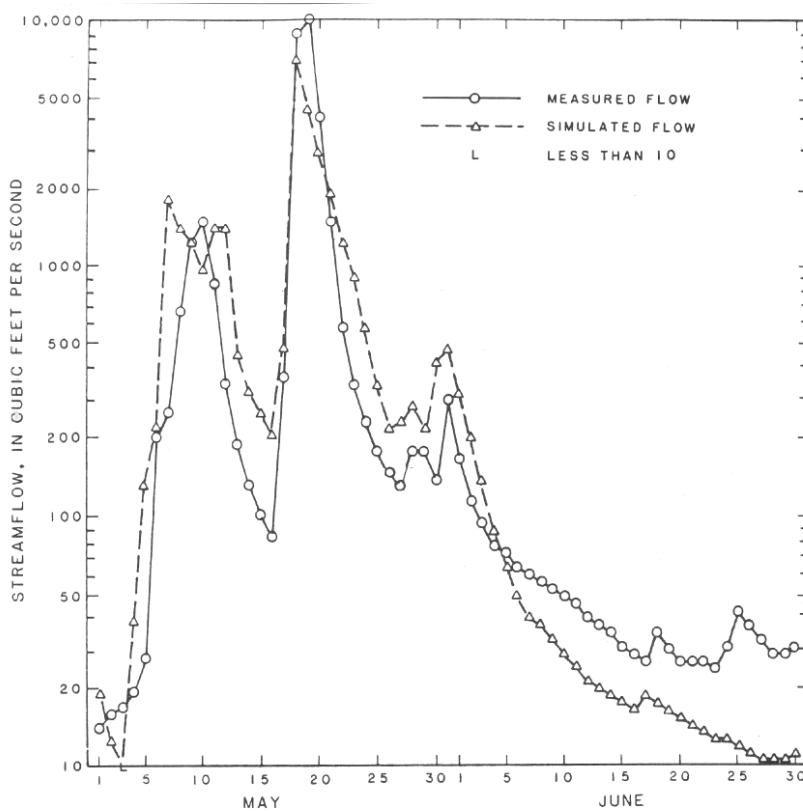
A comparison of measured and modeled logarithms of streamflow is shown in figure 16.5. Which days show the poorest predictions? Though it appears that the largest difference in log streamflow occurs on May 16 and in later June, the mismatch is actually much greater on and near May 6. If the purpose of the graph is to portray daily differences, the differences themselves should be plotted. If room is available in a report or article, one may plot the data as shown in figure 16.5 and plot the differences to provide as much information as possible to the reader.



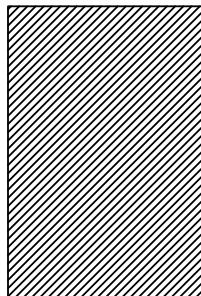
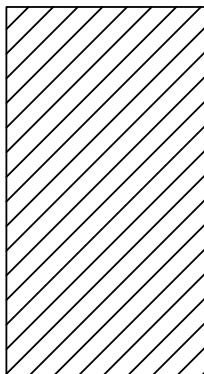
**Figure 16.3.** Graph of seasonal flow distribution for the Red River of the North at Grand Forks, North Dakota, for water years 1994–2015.



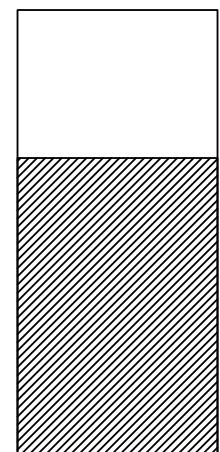
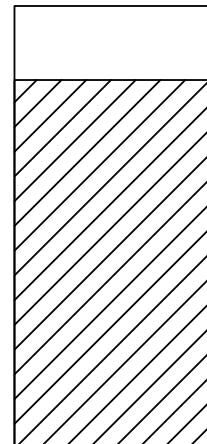
**Figure 16.4.** Bar chart of seasonal flow distribution for the Red River of the North at Grand Forks, North Dakota, for water years 1994–2015, using the same data as in figure 16.3.



**Figure 16.5.** Graph of measured and simulated streamflow. From Bloyd and others (1986).



**Figure 16.6.** Graph demonstrating that judgment of length is more difficult without a common scale.



**Figure 16.7.** Graph showing how framed rectangles improve figure 16.6 by adding a common scale.

### 16.3.5 Length

Judgments of length are required when symbols or bars are to be measured, but do not have a common datum or common scale available. The simplest such case, determination of the length of two offset bars, is shown in figure 16.6. It is difficult to visually determine which is longer. Another example requiring the use of length judgments are the bars displayed on the map of figure 16.2.

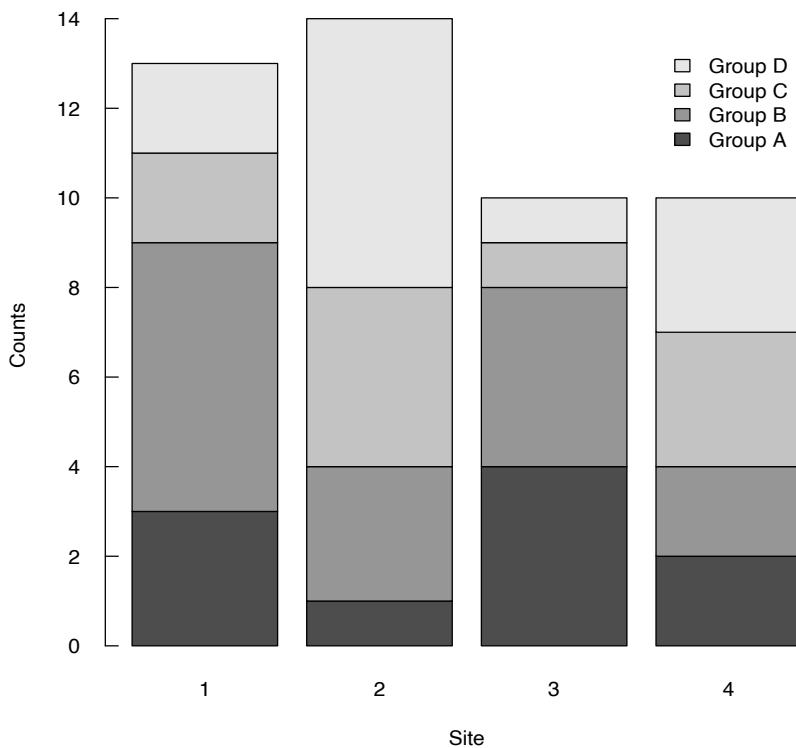
To aid in the judgment process and add precision, a common scale can be added to each bar, this is done in figure 16.7 as a framed rectangle. The rectangle surrounding each bar is of exactly the same length, a common reference frame. It is now easier to see that the first bar is indeed longer than the second. This is because the judgment is based on the positions of the white areas within the common scale. Their relative differences are greater than the shaded bars, and so are 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 ability to judge differences in length.

### 16.3.6 Position Along Nonaligned Scales

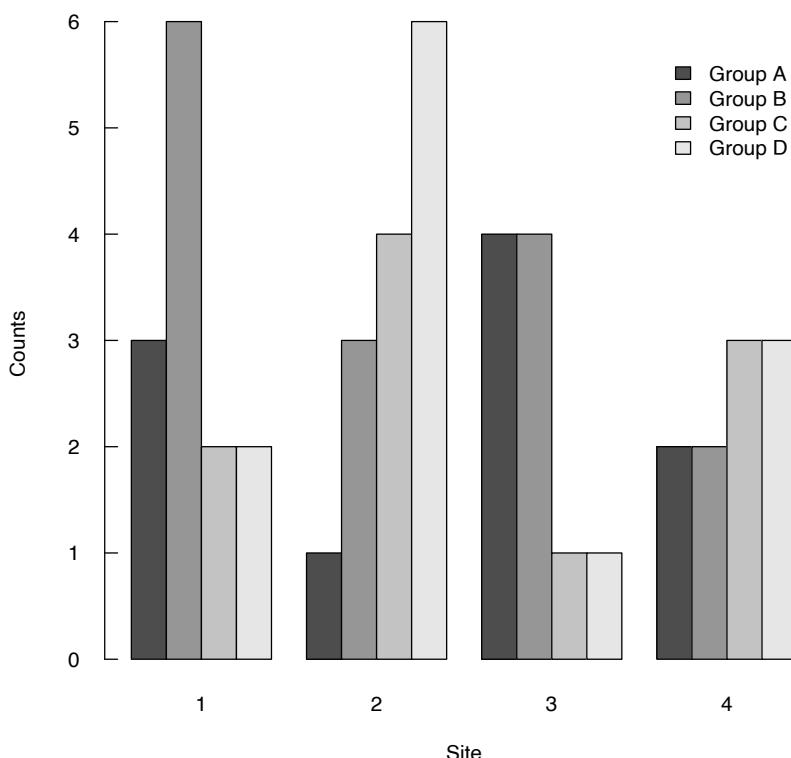
Framed rectangles are examples of graphs with a common but nonaligned scale, which is without a common datum. Another graph in this category is a stacked bar chart (fig. 16.8). These graphs of segmented bars require judgments of position and (or) length. Only the lowest segments of each bar possess a 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 16.8 it is difficult to determine which of the top two squares of bar 1 is larger.

To make comparisons more precise, stacked bars can always be unstacked and placed side-by-side, producing grouped bar charts (fig. 16.9). These are equivalent to multiple pie charts, but grouped bar charts belong in the highest precision category—position along a common scale (common datum)—and provide more accurate discrimination of values and comparisons within and between groups (Streit and Gehlenborg, 2014). By using a common datum, smaller differences are more easily seen. For example, in bar 2 it is now easy to see that C is larger than B. In bar 3, the group A square is larger than D, 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.

Bars are often 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.



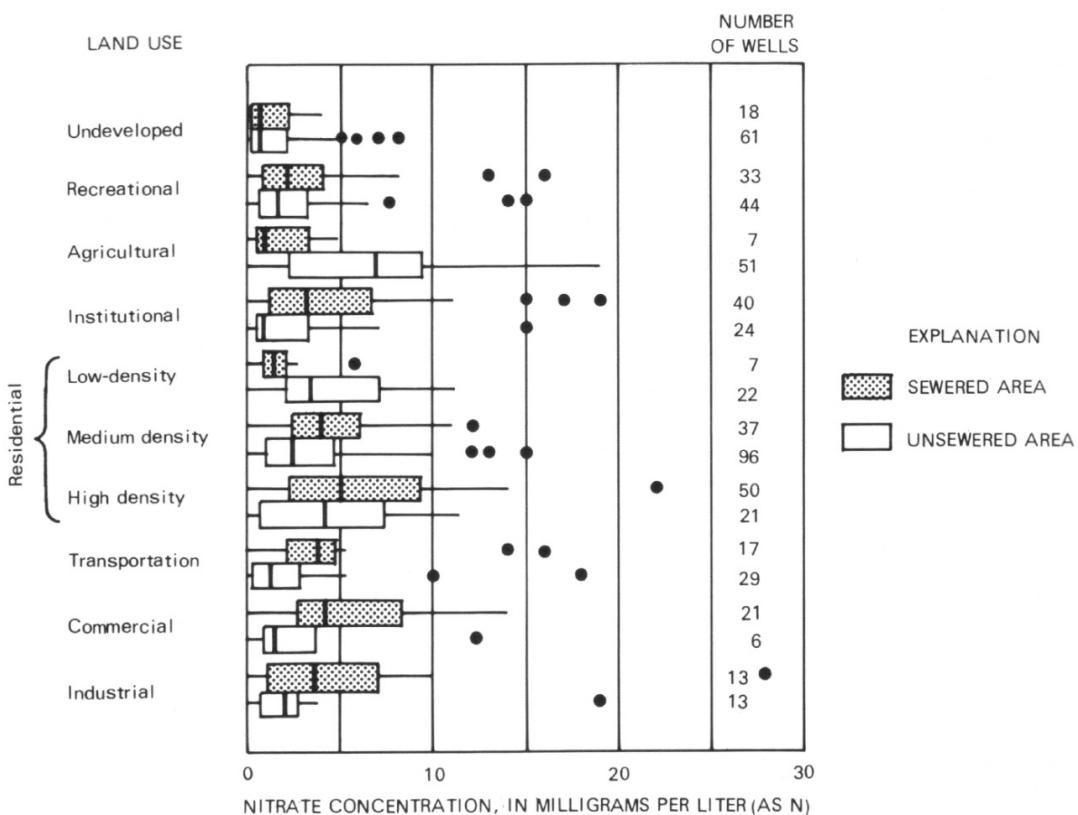
**Figure 16.8.** Stacked bar charts demonstrating the difficulty of comparing data not aligned with the y-axis.



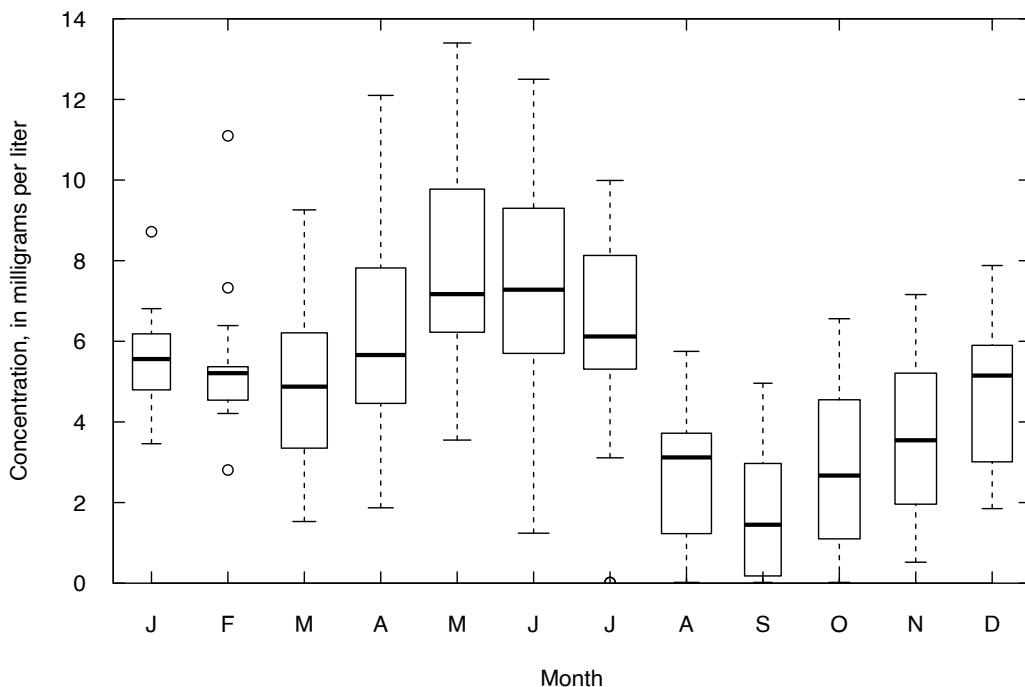
**Figure 16.9.** Grouped bar charts display the same data as in figure 16.8.

### 16.3.7 Position Along an Aligned Scale

Grouped bar charts are one example of graphs where data are positioned along an aligned (common datum) scale. Scatterplots and boxplots also belong in this category. Though discussed and used fully throughout this book, the strengths of boxplots bear repeating. Boxplots represent the characteristics of one or more distributions, but do not require assumptions about those distributions and are, therefore, robust and informative. Many boxes can be placed on a page, allowing precise summaries and comparisons of a large amount of information. Boxplots for a two-way analysis of variance are shown in figure 16.10. Differences in concentration owing to both land-use category and to sewerage are easily seen, as are skewness and outliers. Sample size should be indicated on boxplots because large sample sizes are generally more representative of the population distribution than very small sample sizes. In addition, there are a number of ways to define the whiskers (such as the whisker extending to the maximum and minimum values, spear style; or extending to data points within 1.5 times the interquartile range above and below the third and first quartiles, Tukey style; Krzywinski and Altman, 2014); therefore, the whiskers should be defined in the explanation or the caption (which was not done in fig. 16.10 from Eckhardt and others, 1989). Sample size also can be represented with varying box width that is proportional to the sample size (fig. 16.11). The user can make some comparisons between groups or seasons, for example, to see if a season does not have representative sampling; however, the user is unable to estimate the actual sample sizes from such a graph.



**Figure 16.10.** Boxplots of nitrate concentrations by land use and sewerage. From Eckhardt and others (1989).



**Figure 16.11.** Boxplots of nitrate concentrations in milligrams per liter as N, Iowa River at Wapello, Iowa, water years 2000–16. Box width is proportional to the number of samples in the month, based on 272 total samples. Whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range. The open circles represent the extreme data points more than 1.5 times the interquartile range.

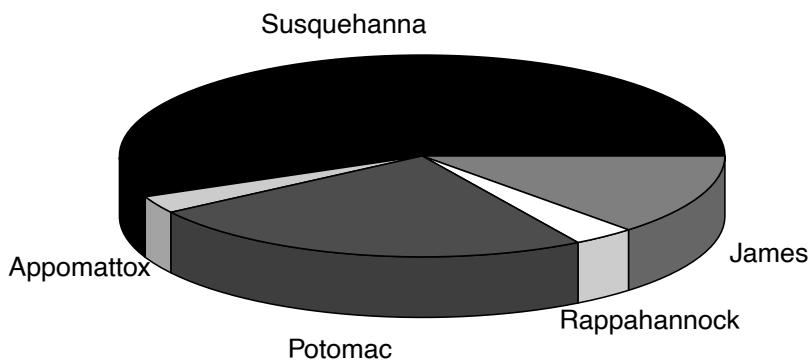
## 16.4 Misleading Graphics to Be Avoided

### 16.4.1 Perspective

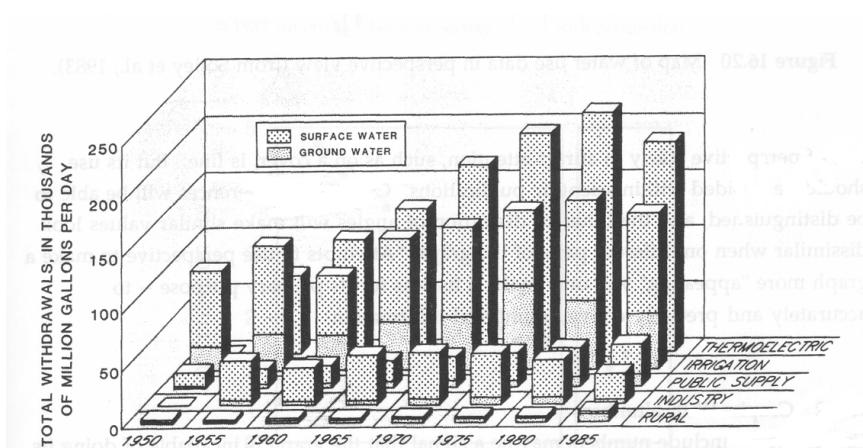
Figures are often put into a tilted perspective to give an impression of three dimensions. The intent is to make the figure look more multidimensional and sophisticated. Unfortunately, by doing so, judgments of area, length, and angle used by the viewer to extract information become impaired. This happens because the brain compensates for the perspective by determining that objects in the distance are larger (United Nations Economic Commission for Europe, 2009).

An example is given in figure 16.12. A pie chart is presented with five slices. The slices represent the relative sizes of the five largest monitored tributaries to the Chesapeake Bay watershed. Judgments of angle are impossible to get correct once the angles are altered by perspective. For example, it is very difficult to tell if the Susquehanna is more than half the total or less than half. It is also very difficult to determine the relative sizes of the Appomattox and Rappahannock watersheds. A bar chart would be a much more effective way to present this information.

A second example of perspective is shown in figure 16.13 where bar charts are placed into perspective so that many bars can be depicted in 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?



**Figure 16.12.** Pie chart in perspective view, showing the drainage area of the five largest monitored tributaries in the Chesapeake Bay watershed.

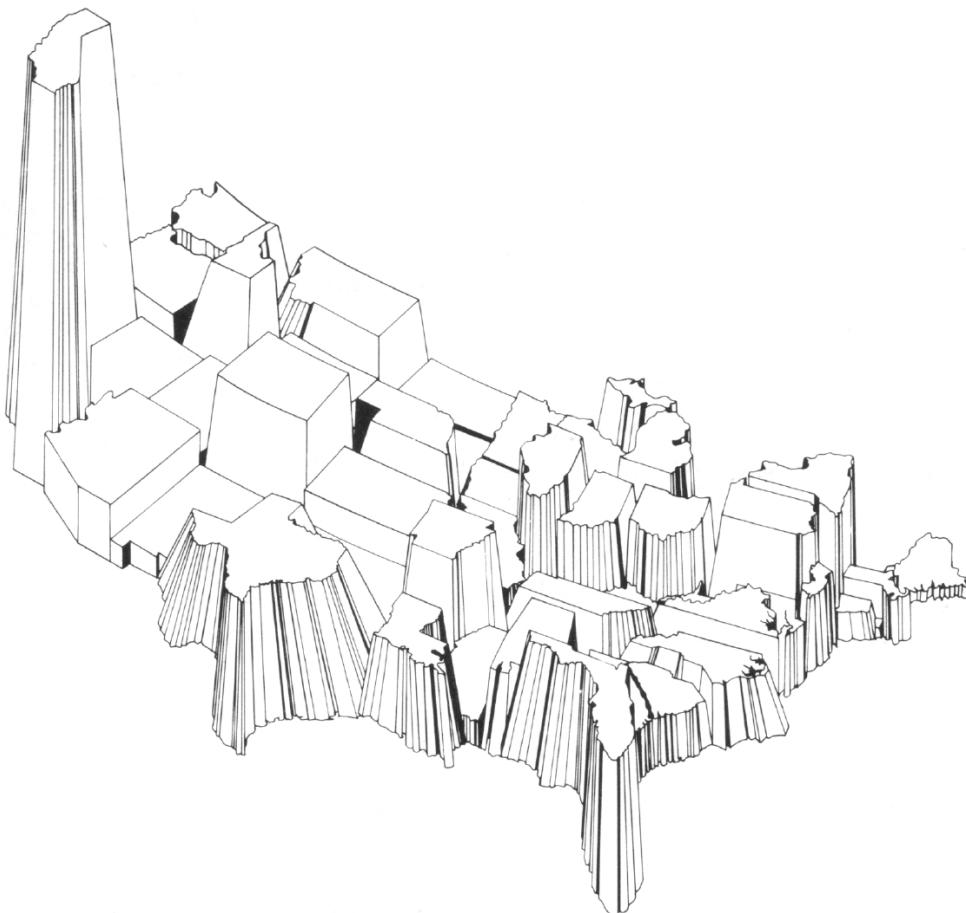


**Figure 16.13.** Bar chart of water use data, in perspective view (Solley and others, 1988).

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 assess accurately the data portrayed by bar heights. Comparisons of heights across nonadjacent rows are 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.

Perspective should also be avoided when presenting maps. A perspective map of water use in the United States (Solley and others, 1988) is shown in figure 16.14. 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 we can determine that Louisiana has a larger value, but it does not appear that way on the map. Note also that several states are again partly or totally hidden.

Use of perspective solely to attract attention, such as on a cover, is fine. However, 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 less useful for its primary purpose—to convey numerical information accurately and precisely.



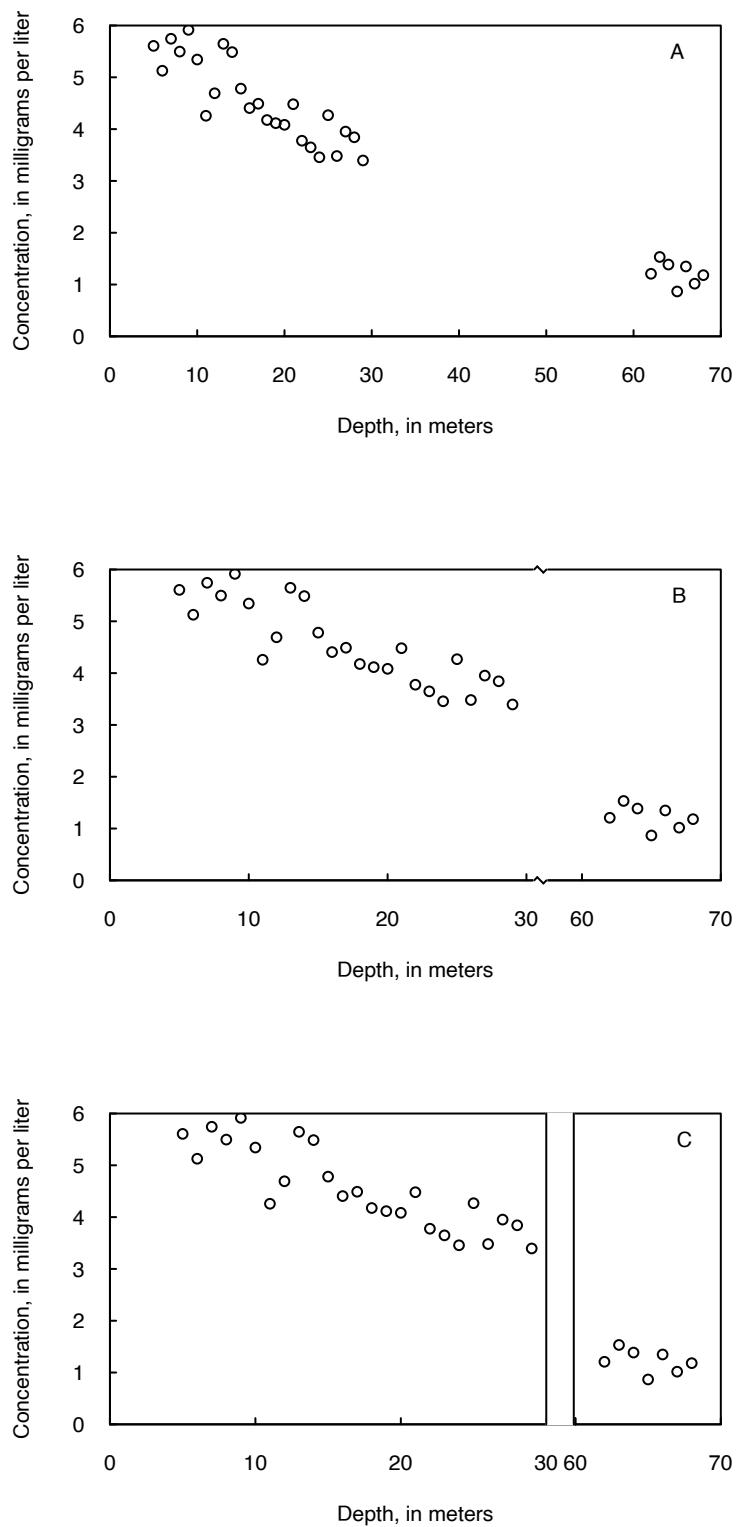
**Figure 16.14.** Map of water use data in perspective view (Solley and others, 1983).

### 16.4.2 Graphs with Numbers

Graphs that include numbers (other than on the axes) may be a signal that the graph is incapable of doing its job. The graph needs to be made more precise. Tables providing the necessary detail for computations can be placed elsewhere in the report if required. However, they do not provide the insight needed to quickly comprehend primary patterns of the data. Adding numbers to graphs that also do not portray those patterns does not add up to an effective graph.

### 16.4.3 Hidden Scale Breaks

Breaks in the scale of measurement on a graph can be misleading to the viewer. If scale breaks are used, it is the job of the presenter to make them as clear as possible. To make a scale break more obvious, Cleveland (1984) suggested the use of a “full scale break” as in figure 16.15C, 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.

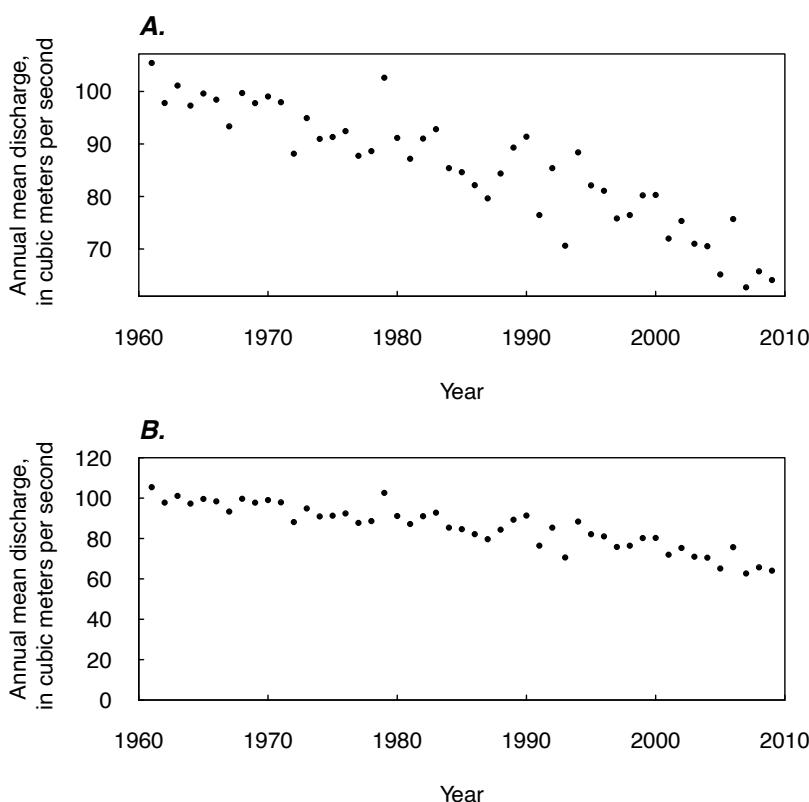


**Figure 16.15.** Graph of simulated concentration and depth data plotted with (A) no scale breaks, (B) a scale break indicated by a zigzag line at the break, and (C) a full-scale break.

### 16.4.4 Self-scaled Graphs

When producing a scatter plot or a line plot representing a time series of data, it is tempting to let the graphing function choose the axis scales for the user. This may be entirely appropriate for quick data exploration, but it may not be appropriate for a final graphical product. As an example, figure 16.16 shows two graphs of the same, randomly generated data intended to represent annual mean discharge. In figure 16.16A, self-scaling is used for the vertical axis. This makes it appear that the trend is quite significant and suggests at a glance that the river must be nearly going dry. The same dataset is shown in figure 16.16B, but here the scaling is set by the user, so that the  $y$ -axis starts at zero. The trend is also evident here, but it leads the reader toward a more appropriate interpretation, namely, that there is a trend in discharge, but it is a modest change over this 49-year period.

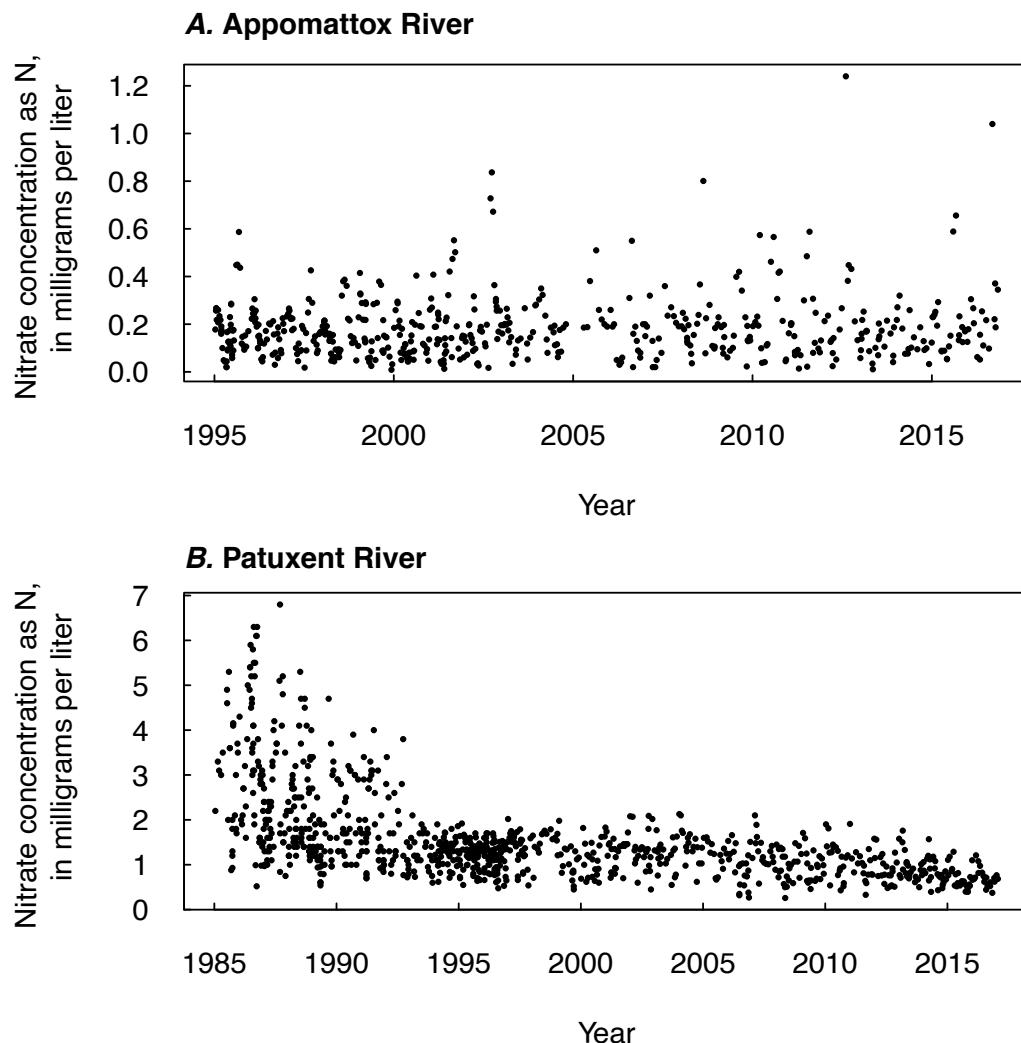
We propose a set of guidelines to govern these kinds of choices. These guidelines are suggested only when the quantity being graphed is one that always remains positive. Thus, concentrations of a solute or certain types of aquifer properties (such as porosity or permeability) would fall into this class. In addition, river discharge or downstream flux of some dissolved or suspended matter would also fall in this class, except in those cases where flow reversals occur (for example, backwater or tidal situations where flow may be in the upstream direction at certain times). Some data that would not fall into this class would be temperature data (unless one uses degrees Kelvin, but that is not helpful to the reader's understanding of the situation), or pH data (which is the negative logarithm of hydrogen ion concentration). In general, it is a good practice to graph physical quantities on a scale with a minimum value of zero. However, an important distinction for the scientist to consider is that when exploring a dataset to understand relations (as is typically done from scatterplots) it can often be very useful to graph the logarithms or some other transformation on the ladder of powers for one or both variables. It is often the case that such plots will facilitate development of statistical models and help resolve issues such as heteroscedasticity. Keep in mind that graphs done with log scales can be excellent tools for the data analyst to understand their dataset, but they may not be excellent tools for the less-technically oriented audience to whom the report is being presented.



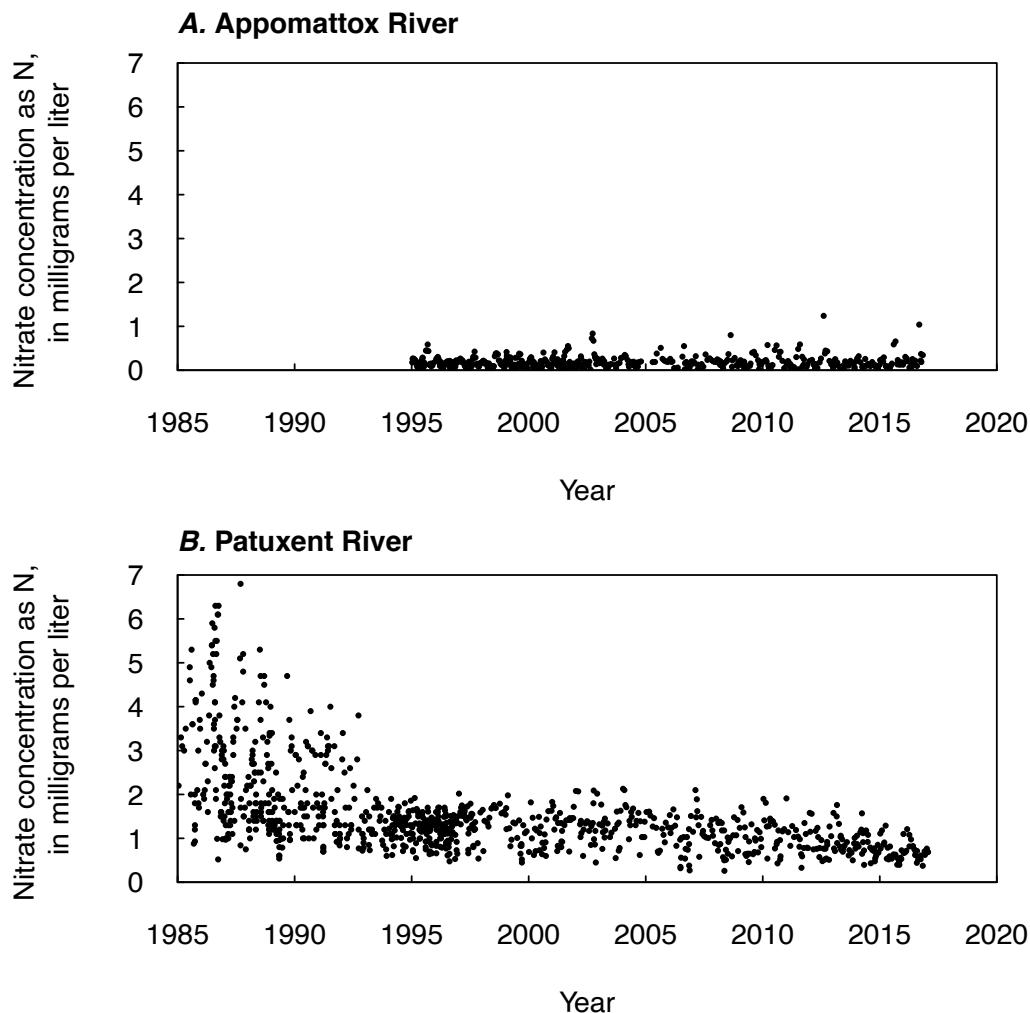
**Figure 16.16.** Two graphical presentations of the same hypothetical dataset. In (A) the vertical axis is self-scaled and does not extend to zero; in (B) the user set the scale on the vertical axis to go to zero.

Another issue that should be considered is the scaling of multiple graphs when each one is presenting similar sets of data and comparisons across datasets are of interest. Consider the following example based on two datasets of nitrate concentrations over a period of several decades. One is the Appomattox River in Virginia, which is a highly rural and mostly forested watershed. The other is the Patuxent River in Maryland, a watershed with substantial suburban land use. For purposes of this example, the full datasets have been shortened in length so that the Appomattox record appears to start in 1995 and the Patuxent starts in 1985. These two records are shown in figure 16.17 with self-scaling used in the presentation of both datasets. Thus, the time scales and concentration scales are different for each dataset. The Appomattox record appears to have a fairly constant central tendency but does show a few very high values towards the later part of the record. The Patuxent record shows a substantial downward trend particularly from 1985 through 1995. The strong downward trend is known to be related to major upgrades of sewage treatment plants in those early years.

A much better way of presenting the information, for a report that is intended to provide comparisons across multiple sites, is for the analyst to impose identical scales on both graphs (fig. 16.18). Viewing the information this way gives a greatly different impression of what has happened in the stream. We see immediately that the period of record is different in each case, so we do not know if perhaps there was a strong decline at the Appomattox site between 1985 and 1995. Much more importantly, we see at a glance that from 1995 to 2016 the Patuxent concentration values are almost all higher than the Appomattox concentration values. Even the few high values for the Appomattox appear to be no higher than the highest



**Figure 16.17.** Time series of nitrate concentration data for the (A) Appomattox River at Matoaca, Virginia, and (B) Patuxent River near Bowie, Maryland. In both cases, the graphs are self-scaled and thus their horizontal and vertical axes are both different.



**Figure 16.18.** Graphs showing the same datasets as in figure 16.17, but with the scaling of both axes identical across the two graphs.

values of the Patuxent record, and the lowest values for the Patuxent are higher than almost all of the Appomattox values. This presentation gives a much more meaningful comparison of the sites; however, the data are now compressed for the Appomattox River. Although we can see concentrations are considerably lower than in the Patuxent River, we lose the depiction of the within-site variable. One could add boxplots to the analysis or plot on the data on logarithmic scales to better depict the within-site variability and the between site differences.

One additional thought regarding the scaling of graphs is that for graphs of river discharge or river fluxes (of sediment or some chemical species) it is beneficial to the reader to standardize the variable of interest by dividing by the drainage area. Thus, discharge numbers, rather than being in cubic meters per second ( $m^3/s$ ), may be shown as runoff values in units such as millimeters per day (mm/day) or cubic meters per second per hectare ( $m^3/s/ha$ ). Similarly, for fluxes, instead of units such as kilograms per year (kg/year) we can present the results as yields in units such as kilograms per hectare per year (kg/ha/year).

## 16.5 Conclusion

“Good statistical graphics are hard to do, much harder than running regressions and making tables” (Gelman, 2011). Despite the degree of difficulty and the many caveats provided above, creating effective graphics can greatly contribute to satisfaction in one’s publications or presentations and can contribute to their usefulness. We encourage you to put special effort into presentation graphics and approach them

with enthusiasm, such as that expressed by Henry Gannet in the preface to the Statistical Atlas of the Tenth Census in 1880, “Let these facts be expressed not alone in figures, but graphically, by means of maps and diagrams, appealing to a quick sense of form and color and ‘clothing the dry bones of statistics in flesh and blood,’ and their study becomes a delight rather than a task” (Friendly, 2008).

# References Cited

---

- Adler, D., Murdoch, D., Nenadic, O., Urbanek, S., Chen, M., Gebhardt, A., Bolker, B., Csardi, G., Strzelecki, A., Senger, A., The R Core Team, and Eddelbuettel, D., 2018, rgl—3D visualization using OpenGL (ver. 0.99.16): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=rgl>.
- Agresti, A., 2002, Categorical data analysis: Hoboken, N.J., John Wiley & Sons, 721 p., accessed October 2019 at <https://doi.org/10.1002/0471249688>.
- Aho, K.A., 2016, Foundational and applied statistics for biologists using R: Boca Raton, Fla., Chapman & Hall/CRC Press, 596 p.
- Aho, K., 2019, asbio—A collection of statistical tools for biologists (ver. 1.5-5): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=asbio>.
- Aitchison, J., and Brown, J.A.C., 1981, The lognormal distribution: Cambridge, England, Cambridge University Press, 176 p.
- Akaike, H., 1974, A new look at the statistical model identification: Institute of Electrical and Electronics Engineer Transactions on Automatic Control, v. 19, no. 6, p. 716–723, accessed October 2019 at <https://doi.org/10.1109/TAC.1974.1100705>.
- Alley, W.M., 1988, Using exogenous variables in testing for monotonic trends in hydrologic time series: Water Resources Research, v. 24, no. 11, p. 1955–1961, accessed October 2019 at <https://doi.org/10.1029/WR024i011p01955>.
- Amemiya, T., 1981, Qualitative response models—A survey: Journal of Economic Literature, v. 19, p. 1483–1536.
- Anscombe, F.A., 1973, Graphs in statistical analysis: The American Statistician, v. 27, p. 17–21, accessed February 2020 at <https://doi.org/10.1080/00031305.1973.10478966>.
- Archfield, S.A., Hirsch, R.M., Viglione, A., and Blöschl, G., 2016, Fragmented patterns of flood change across the United States: Geophysical Research Letters, v. 43, no. 19, p. 10232–10239, accessed October 2019 at <https://doi.org/10.1002/2016GL070590>.
- Ayotte, J.D., Cahillane, M., Hayes, L., and Robinson, K.W., 2012, Estimated probability of arsenic in groundwater from bedrock aquifers in New Hampshire, 2011: U.S. Geological Survey Scientific Investigations Report 2012–5156, 25 p., accessed August 10, 2016, at <http://pubs.usgs.gov/sir/2012/5156/>.
- Ayotte, J.D., Nolan, B.T., and Gronberg, J.A., 2016, Predicting arsenic in drinking water wells of the Central Valley, California: Environmental Science & Technology, v. 50, no. 14, p. 7555–7563, accessed October 2019 at <https://doi.org/10.1021/acs.est.6b01914>.
- Bacchetti, P., 2010, Current sample size conventions—Flaws, harms, and alternatives: BMC Medicine, v. 8, no. 1, p. 17, accessed October 2019 at <https://doi.org/10.1186/1741-7015-8-17>.
- Bajgier, S.M., Atkinson, M., and Prybutok, V.R., 1989, Visual fits in the teaching of regression concepts: The American Statistician, v. 43, no. 4, p. 229–234, accessed February 2020 at <https://doi.org/10.1080/00031305.1989.10475664>.

- Barefield, E., and Mansouri, H., 2001, An empirical study of nonparametric multiple comparison procedures in randomized blocks: *Journal of Nonparametric Statistics*, v. 13, no. 4, p. 591–604, accessed October 2019 at <https://doi.org/10.1080/10485250108832867>.
- Bayazit, M., and Önöz, B., 2007, To prewhiten or not to prewhiten in trend analysis?: *Hydrological Sciences Journal*, v. 52, no. 4, p. 611–624, accessed October 2019 at <https://doi.org/10.1623/hysj.52.4.611>.
- Beckman, R.J., and Cook, R.D., 1983, Outlier ... .... s: *Technometrics*, v. 25, p. 119–149, accessed March 2020 at <https://doi.org/10.1080/00401706.1983.10487840>.
- Belsley, D.A., Kuh, E., and Welsch, R.E., 1980, *Regression diagnostics*: New York, John Wiley, 292 p., accessed October 2019 at <https://doi.org/10.1002/0471725153>.
- Benjamini, Y., and Hochberg, Y., 1995, Controlling the false discovery rate—A practical and powerful approach to multiple testing: *Journal of the Royal Statistical Society, Series B, Methodological*, v. 57, no. 1, p. 289–300.
- Benson, M.A., 1965, Spurious correlation in hydraulics and hydrology: *Journal of the Hydraulics Division*, v. 91, no. 4, p. 35–42.
- Bernhard, J., and Kelso, N.V., 2018, Color Oracle—Design for the color impaired (ver. 1.3): Color Oracle web page, accessed October 2019 at <http://colororacle.org/>.
- Blair, R.C., and Higgins, J.J., 1980, A comparison of the power of Wilcoxon's rank-sum statistic to that of student's t statistic under various nonnormal distributions: *Journal of Statistics Education*, v. 5, p. 309–335.
- Blom, G., 1958, *Statistical estimates and transformed beta variables*: New York, John Wiley, 176 p.
- Bloyd, R.M., Daddow, P.B., Jordan, P.R., and Lowham, H.W., 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.
- Boos, D.D., and Hughes-Oliver, J.M., 2000, How large does n have to be for Z and t intervals?: *The American Statistician*, v. 54, no. 2, p. 121–128, accessed February 2020 at <https://doi.org/10.1080/00031305.2000.10474524>.
- Borcard, D., Gillet, F., and Legendre, P., 2011, *Numerical ecology with R*: New York, Springer, 306 p., accessed October 2019 at <https://doi.org/10.1007/978-1-4419-7976-6>.
- Boudette, E.L., Canney, F.C., Cotton, J.E., Davis, R.I., Ficklin, W.H., and Matooka, J.M., 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., Jenkins, G.M., Reinsel, G.C., and Ljung, G.M., 2015, *Time series analysis—Forecasting and control*: Hoboken, N.J., John Wiley & Sons, 712 p.
- Bradley, J.V., 1968, *Distribution-free statistical tests*: Englewood Cliffs, N.J., Prentice-Hall, 388 p.
- Bradu, D., and Mundlak, Y., 1970, Estimation in lognormal linear models: *Journal of the American Statistical Association*, v. 65, no. 329, p. 198–211, accessed October 2019 at <https://doi.org/10.1080/01621459.1970.10481074>.
- Breusch, T.S., and Pagan, A.R., 1979, A simple test for heteroscedasticity and random coefficient variation: *Econometrica*, v. 47, no. 5, p. 1287–1294, accessed October 2019 at <https://doi.org/10.2307/1911963>.
- Brier, G.W., 1950, Verification of forecasts express in terms of probability: *Monthly Weather Review*, v. 78, no. 1, p. 1–3, accessed October 2019 at [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).

- Bronaugh, D., and Werner, A., 2013, zyp: Zhang + Yue-Pilon trends package for R (ver. 0.10-1): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=zyp>.
- Brown, C.D., and Davis, H.T., 2006, Receiver operating characteristic curves and related decision measures—A tutorial: *Chemometrics and Intelligent Laboratory Systems*, v. 80, no. 1, p. 24–38, accessed October 2019 at <https://doi.org/10.1016/j.chemolab.2005.05.004>.
- Brown, D., and Rothery, P., 1993, Models in biology—Mathematics, statistics and computing: Hoboken N.J., John Wiley & Sons, 708 p.
- Brunner, E., Dette, H., and Munk, A., 1997, Box-type approximations in nonparametric factorial designs: *Journal of the American Statistical Association*, v. 92, no. 440, p. 1494–1502, accessed October 2019 at <https://doi.org/10.1080/01621459.1997.10473671>.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A., 1983, Graphical methods for data analysis: Boston, PWS-Kent Publishing Co., 395 p.
- Chang, W., 2013, R graphics cookbook: Sebastopol, Calif., O'Reilly Media, Inc., 416 p.
- Charlton, B.G., 2004, Why a journal of negative results?: *Journal of Negative Results - Ecology & Evolutionary Biology*, v. 1, no. 1, p. 6–7.
- Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots: *Journal of the American Statistical Association*, v. 74, no. 368, p. 829–836, accessed October 2019 at <https://doi.org/10.1080/01621459.1979.10481038>.
- Cleveland, W.S., 1984, Graphical methods for data presentation—Full scale breaks, dot charts, and multibased logging: *The American Statistician*, v. 38, no. 4, p. 270–280, accessed March 2020, at <https://doi.org/10.1080/00031305.1984.10483224>.
- Cleveland, W.S., 1985, The elements of graphing data: Monterey, Calif., Wadsworth Books, 323 p.
- Cleveland, W.S., and Devlin, S.J., 1988, Locally weighted regression—An approach to regression analysis by local fitting: *Journal of the American Statistical Association*, v. 83, no. 403, p. 596–610, accessed October 2019 at <https://doi.org/10.1080/01621459.1988.10478639>.
- Cleveland, W.S., Grosse, E., and Shyu, W.M., 1992, Local regression models, chap. 8 of Chambers, J.M., and Hastie, T.J., eds., *Statistical models in S*: Pacific Grove, Calif., Wadsworth & Brooks/Cole Advanced Books & Software, p. 309–376.
- Cleveland, W.S., and McGill, R., 1983, A color-caused optical illusion on a statistical graph: *The American Statistician*, v. 37, no. 2, p. 101, accessed February 2020 at <https://doi.org/10.1080/00031305.1983.10482720>.
- Cleveland, W.S., and McGill, R., 1984a, Graphical perception—Theory, experimentation, and application to the development of graphical methods: *Journal of the American Statistical Association*, v. 79, no. 387, p. 531–554, accessed October 2019 at <https://doi.org/10.1080/01621459.1984.10478080>.
- Cleveland, W.S., and McGill, R., 1984b, The many faces of a scatterplot: *Journal of the American Statistical Association*, v. 79, no. 388, p. 807–822, accessed October 2019 at <https://doi.org/10.1080/01621459.1984.10477098>.
- Clow, D.W., 2010, Changes in the timing of snowmelt and streamflow in Colorado—A response to recent warming: *Journal of Climate*, v. 23, no. 9, p. 2293–2306, accessed October 2019 at <https://doi.org/10.1175/2009JCLI2951.1>.
- Cohen, A.C., 1950, Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples: *Annals of Mathematical Statistics*, v. 21, no. 4, p. 557–569, accessed October 2019 at <https://doi.org/10.1214/aoms/1177729751>.

- Cohen, A.C., 1976, Progressively censored sampling in the three parameter log-normal distribution: *Technometrics*, v. 18, no. 1, p. 99, accessed October 2019 at <https://doi.org/10.2307/1267922>.
- 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.
- Cohn, T.A., 2005, Estimating contaminant loads in rivers—An application of adjusted maximum likelihood to type 1 censored data: *Water Resources Research*, v. 41, no. 7, accessed October 2019 at <https://doi.org/10.1029/2004WR003833>.
- Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R., and Lamontagne, J.R., 2013, A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series: *Water Resources Research*, v. 49, no. 8, p. 5047–5058, accessed October 2019 at <https://doi.org/10.1002/wrcr.20392>.
- Cohn, T.A., and Lins, H.F., 2005, Nature's style—Naturally trendy: *Geophysical Research Letters*, v. 32, no. 23, L23402, accessed October 2019 at <https://doi.org/10.1029/2005GL024476>.
- Conover, W.J., 1999, Practical nonparametric statistics (3d ed.): New York, John Wiley and Sons, 493 p.
- Conover, W.L., and Iman, R.L., 1981, Rank transformations as a bridge between parametric and nonparametric statistics: *The American Statistician*, v. 35, no. 3, p. 124–129, accessed February 2020 at <https://doi.org/10.1080/00031305.1981.10479327>.
- Conover, W.J., Johnson, M.E., and Johnson, M.M., 1981, A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data: *Technometrics*, v. 23, no. 4, p. 351–361, accessed October 2019 at <https://doi.org/10.1080/00401706.1981.10487680>.
- Crabtree, R.W., Cluckie, I.D., and Forster, C.F., 1987, Percentile estimation for water quality data: *Water Research*, v. 21, no. 5, p. 583–590, accessed October 2019 at [https://doi.org/10.1016/0043-1354\(87\)90067-4](https://doi.org/10.1016/0043-1354(87)90067-4).
- Cribbie, R.A., Wilcox, R.R., Bewell, C., and Keselman, H.J., 2007, Tests for treatment group equality when data are nonnormal and heteroscedastic: *Journal of Modern Applied Statistical Methods*, v. 6, no. 1, p. 117–132, accessed October 2019 at <https://doi.org/10.22237/jmasm/1177992660>.
- Critchlow, D.E., and Fligner, M.A., 1991, On distribution-free multiple comparisons in the one-way analysis of variance: *Communications in Statistics. Theory and Methods*, v. 20, no. 1, p. 127–139, accessed October 2019 at <https://doi.org/10.1080/03610929108830487>.
- Cunnane, C., 1978, Unbiased plotting positions—A review: *Journal of Hydrology*, v. 37, no. 3–4, p. 205–222, accessed October 2019 at [https://doi.org/10.1016/0022-1694\(78\)90017-3](https://doi.org/10.1016/0022-1694(78)90017-3).
- Dang, X., Peng, H., Wang, X., and Zhang, H., 2008, Theil-Sen estimators in a multiple linear regression model: *Proceedings of the Joint Statistical Meetings*, American Statistical Association, August 3–7, 2008, Denver, Colo., accessed October 2019 at <http://home.olemiss.edu/~xdang/papers/MTSE.pdf>.
- Davis, J.C., 2004, Statistics and data analysis in geology: New York, John Wiley & Sons, 638 p.
- De Cicco, L.A., 2016, dataRetrieval tutorial: U.S. Geological Survey USGS-R web page, accessed October 2019 at <https://owi.usgs.gov/R/dataRetrieval.html>.
- Dietz, E.J., 1985, The rank sum test in the linear logistic model: *The American Statistician*, v. 39, no. 4, p. 322, accessed February 2020 at <https://doi.org/10.1080/00031305.1985.10479460>.
- Dietz, E.J., 1987, A comparison of robust estimators in simple linear regression: *Communications in Statistics—Simulation and Computation*, v. 16, p. 1209–1227.
- Dietz, E.J., 1989, Teaching regression in a nonparametric statistics course: *The American Statistician*, v. 43, no. 1, p. 35–40, accessed March 2020 at <https://doi.org/10.1080/00031305.1989.10475606>.

- Doornkamp, J.C., and King, C.A.M., 1971, Numerical analysis in geomorphology, An introduction: New York, St. Martins Press, 372 p.
- Draper, N.R., and Smith, H., 1981, Applied regression analysis (2d ed.): New York, John Wiley and Sons, 709 p.
- Draper, N.R., and Yang, Y.F., 1997, Generalization of the geometric mean functional relationship: Computational Statistics & Data Analysis, v. 23, no. 3, p. 355–372, accessed October 2019 at [https://doi.org/10.1016/S0167-9473\(96\)00037-0](https://doi.org/10.1016/S0167-9473(96)00037-0).
- Duan, N., 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610, accessed October 2019 at <https://doi.org/10.1080/01621459.1983.10478017>.
- Dunn, O.J., 1964, Multiple comparisons using rank sums: Technometrics, v. 6, no. 3, p. 241–252, accessed October 2019 at <https://doi.org/10.1080/00401706.1964.10490181>.
- Durbin, J., and Watson, G.S., 1950, Testing for serial correlation in least squares regression, I and II: Biometrika, v. 37, no. 3 and 4, p. 409–428.
- Eckhardt, D.A., Flipse, W.J., Jr., and Oaksford, E.T., 1989, Relation between land use and groundwater quality in the upper glacial aquifer in Nassau and Suffolk Counties, Long Island, New York: U.S. Geological Survey Water-Resources Investigations Report 86-4142, 26 p.
- Efron, B., and Tibshirani, R.J., 1994, An introduction to the bootstrap: Boca Raton, Fla., Chapman & Hall/CRC Press, 456 p.
- England, J.F., Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p., accessed October 2019 at <https://doi.org/10.3133/tm4B5>.
- Everitt, B., 2007, An R and S-plus companion to multivariate analysis: London, Springer-Verlag, 221 p.
- Everitt, B., and Hothorn, T., 2011, An introduction to applied multivariate analysis with R: New York, Springer, 273 p., accessed October 2019 at <https://doi.org/10.1007/978-1-4419-9650-3>.
- Faria, J.C. Jelihovschi, E.G., and Allaman, I.B., 2019, Conventional Tukey test: The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=TukeyC>.
- 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. Simulation and Computation, v. 13, no. 2, p. 213–225, accessed October 2019 at <https://doi.org/10.1080/03610918408812368>.
- Fay, M.P., and Shaw, P.A., 2010, Exact and asymptotic weighted logrank tests for interval censored data—The interval R package: Journal of Statistical Software, v. 36, no. 2, accessed October 2019 at <https://doi.org/10.18637/jss.v036.i02>.
- Fent, K., and Hunn, J., 1991, Phenyltins in water, sediment, and biota of freshwater marinas: Environmental Science & Technology, v. 25, no. 5, p. 956–963, accessed October 2019 at <https://doi.org/10.1021/es00017a020>.
- Ferguson, R.I., 1986, River loads underestimated by rating curves: Water Resources Research, v. 22, no. 1, p. 74–76, accessed October 2019 at <https://doi.org/10.1029/WR022i001p00074>.
- Ferro, C.A.T., 2007, Comparing probabilistic forecasting systems with the Brier Score: Weather and Forecasting, v. 22, no. 5, p. 1076–1088, accessed October 2019 at <https://doi.org/10.1175/WAF1034.1>.
- Feth, J.H.F., Robertson, C.E., and Polzer, W.L., 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.

- Filliben, J.J., 1975, The probability plot correlation coefficient test for normality: *Technometrics*, v. 17, no. 1, p. 111–117, accessed October 2019 at <https://doi.org/10.1080/00401706.1975.10489279>.
- Fisher, R.A., 1922, On the mathematical foundations of theoretical statistics: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, v. 222, no. 594–604, p. 309–368, accessed October 2019 at <https://doi.org/10.1098/rsta.1922.0009>.
- Fox, J., and Weisberg, S., 2011, An R companion to applied regression: Thousand Oaks, Calif., Sage Publications, 445 p.
- Franklin, L.A., 1988, The complete exact null distribution of Spearman’s Rho for n = 12(1)18: *Journal of Statistical Computation and Simulation*, v. 29, no. 3, p. 578–580, accessed October 2019 at <https://doi.org/10.1080/00949658808811066>.
- Friendly, M., 2008, The golden age of statistical graphics: *Statistical Science*, v. 23, no. 4, p. 502–535, accessed October 2019 at <https://doi.org/10.1214/08-STS268>.
- Frigge, M., Hoaglin, D.C., and Iglewicz, B., 1989, Some implementations of the boxplot: *The American Statistician*, v. 43, no. 1, p. 50–54, accessed February 2020 at <https://doi.org/10.1080/00031305.1989.10475612>.
- Frost, J., 2016, Did Welch’s ANOVA make Fisher’s classic one-way ANOVA obsolete?: *Minitab Blog*, April 13, 2014, accessed October 2019 at <https://blog.minitab.com/blog/adventures-in-statistics-2/did-welchs-anova-make-fishers-classic-one-way-anova-obsolete>.
- Fusillo, T.V., Hochreiter, J.J., and Lord, D.G., 1985, Distribution of volatile organic compounds in a New Jersey coastal plain aquifer system: *Groundwater*, v. 23, no. 3, p. 354–360, accessed October 2019 at <https://doi.org/10.1111/j.1745-6584.1985.tb00780.x>.
- Garmo, Ø.A., Skjelkvåle, B.L., de Wit, H.A., Colombo, L., Curtis, C., Fölster, J., Hoffmann, A., Hruška, J., Høgåsen, T., Jeffries, D.S., Keller, W.B., Krám, P., Majer, V., Monteith, D.T., Paterson, A.M., Rogora, M., Rzychon, D., Steingruber, S., Stoddard, J.L., Vuorenmaa, J., and Worsztynowicz, A., 2014, Trends in surface water chemistry in acidified areas in Europe and North America from 1990 to 2008: *Water, Air, and Soil Pollution*, v. 225, no. 3, p. 1880, accessed October 2019 at <https://doi.org/10.1007/s11270-014-1880-6>.
- Gelman, A., 2011, Rejoinder: *Journal of Computational and Graphical Statistics*, v. 20, no. 1, p. 36–40, accessed October 2019 at <https://doi.org/10.1198/jcgs.2011.09166rej>.
- Godsey, S.E., Aas, W., Clair, T.A., De Wit, H.A., Fernandez, I.J., Kahl, J.S., Malcolm, I.A., Neal, C., Neal, M., Nelson, S.J., Norton, S.A., Palucis, M.C., Skjelkvåle, B.L., Soulsby, C., Tetzlaff, D., and Kirchner, J.W., 2010, Generality of fractal 1/f scaling in catchment tracer time series, and its implications for catchment travel time distributions: *Hydrological Processes*, v. 24, no. 12, p. 1660–1671, accessed October 2019 at <https://doi.org/10.1002/hyp.7677>.
- Goldenweiser, E.A., 1916, Classification and limitations of statistical graphics: *Publications of the American Statistical Association*, v. 15, no. 114, p. 205–209, accessed October 2019 at <https://doi.org/10.2307/2965112>.
- Good, P.I., 2001, Resampling methods—A practical guide to data analysis (2d ed.): New York, Springer Science+Business Media, 238 p.
- Good, P.I., 2005, Permutation tests—A practical guide to resampling methods for testing hypotheses: New York, Springer-Verlag, 316 p.
- Goodchild van Hilten, L., 2015, Why it’s time to publish research “failures”: Elsevier Connect, accessed October 2019 at <https://www.elsevier.com/connect/scientists-we-want-your-negative-results-too>.

- Granato, G.E., 2006, Kendall-Theil Robust Line (KTRLLine--version 1.0)-A Visual Basic program for calculating and graphing robust nonparametric estimates of linear-regression coefficients between two continuous variables: U.S. Geological Survey Techniques and Methods, book 4, chap. A7, 31 p., accessed October 2019 at <https://doi.org/10.3133/tm4A7>.
- Green, D.M., and Swets, J.A., 1966, Signal detection theory and psychophysics: New York, John Wiley & Sons, Inc., 455 p.
- Gringorten, I.I., 1963, A plotting rule for extreme probability paper: *Journal of Geophysical Research*, v. 68, no. 3, p. 813–814, accessed October 2019 at <https://doi.org/10.1029/JZ068i003p00813>.
- Groggel, D.J., 1987, A Monte Carlo study of rank tests for block designs: *Communications in Statistics—Simulation and Computation*, v. 16, no. 3, p. 601–620, accessed October 2019 at <https://doi.org/10.1080/03610918708812607>.
- Grygier, J.C., Stedinger, J.R., and Yin, H.-B., 1989, A generalized maintenance of variance extension procedure for extending correlated series: *Water Resources Research*, v. 25, no. 3, p. 345–349, accessed October 2019 at <https://doi.org/10.1029/WR025i003p00345>.
- Gumbel, E., 1958, Statistics of extremes: New York, Columbia University Press, 375 p.
- Haan, C.T., 1977, Statistical methods in hydrology: Ames, Iowa, Iowa State University Press, 378 p.
- Hahn, G.J., and Meeker, W.Q., 1991, Statistical intervals, a guide for practitioners: New York, John Wiley & Sons, Inc., 392 p., accessed October 2019 at <https://doi.org/10.1002/9780470316771>.
- Håkanson, L., 1984, Sediment sampling in different aquatic environments—Statistical aspects: *Water Resources Research*, v. 20, no. 1, p. 41–46, accessed October 2019 at <https://doi.org/10.1029/WR020i001p00041>.
- Hald, A., 1949, Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point: *Skandinavisk Aktuarietidskrift*, no. 32, p. 119–134.
- Halfon, E., 1985, Regression method in ecotoxicology—A better formulation using the geometric mean functional regression: *Environmental Science & Technology*, v. 19, no. 8, p. 747–749, accessed October 2019 at <https://doi.org/10.1021/es00138a016>.
- Hamed, K.H., 2008, Trend detection in hydrologic data—The Mann–Kendall trend test under the scaling hypothesis: *Journal of Hydrology (Amsterdam)*, v. 349, no. 3–4, p. 350–363, accessed October 2019 at <https://doi.org/10.1016/j.jhydrol.2007.11.009>.
- Harrell, F.E., Jr., 2016, Rms—Regression modeling strategies (ver. 5.0-1): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=rms>.
- Hazen, A., 1914, Storage to be provided in the impounding reservoirs for municipal water supply: *Transactions of the American Society of Civil Engineers*, v. 77, p. 1547–1550.
- Helsel, D.R., 1983, Mine drainage and rock type influences on Eastern Ohio streamwater quality: *Water Resources Bulletin*, v. 19, no. 6, p. 881–888, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.1983.tb05936.x>.
- Helsel, D.R., 1990, Less than obvious—Statistical treatment of data below the detection limit: *Environmental Science & Technology*, v. 24, no. 12, p. 1766–1774, accessed October 2019 at <https://doi.org/10.1021/es00082a001>.
- Helsel, D.R., 2012, Statistics for censored environmental data using Minitab and R: New York, John Wiley & Sons, 344 p.
- Helsel, D.R., and Frans, L.M., 2006, Regional Kendall Test for trend: *Environmental Science & Technology*, v. 40, no. 13, p. 4066–4073, accessed October 2019 at <https://doi.org/10.1021/es051650b>.

- Helsel, D.R., and Griffith, L.M., 2003, Assess and interpret data: Water Resources Impact, v. 5, no. 5, p. 25–29.
- Helsel, D.R., and Hirsch, R.M., 1988, Discussion of applicability of the t-test for detecting trends in water quality variables: Water Resources Bulletin, v. 24, p. 201–204, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.1988.tb00896.x>.
- Helsel, D.R., and Ryker, S.J., 2002, Defining surfaces for skewed, highly variable data: Environmetrics, v. 13, no. 5–6, p. 445–452, accessed October 2019 at <https://doi.org/10.1002/env.531>.
- 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, T., 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., accessed October 2019 at <https://doi.org/10.3133/pp1402C>.
- Higgins, J.J., 2003, Introduction to modern nonparametric statistics: Pacific Grove, Calif., Brooks/Cole, 366 p.
- Hirsch, R.M., 1982, A comparison of four streamflow record extension techniques: Water Resources Research, v. 18, no. 4, p. 1081–1088, accessed October 2019 at <https://doi.org/10.1029/WR018i004p01081>.
- Hirsch, R.M., 1988, Statistical methods and sampling design for estimating step trends in surface-water quality: Water Resources Bulletin, v. 24, no. 3, p. 493–503, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.1988.tb00899.x>.
- Hirsch, R.M., Alexander, R.B., and Smith, R.A., 1991, Selection of methods for the detection and estimation of trends in water quality: Water Resources Research, v. 27, no. 5, p. 803–813, accessed October 2019 at <https://doi.org/10.1029/91WR00259>.
- Hirsch, R.M., Archfield, S.A., and De Cicco, L.A., 2015, A bootstrap method for estimating uncertainty of water quality trends: Environmental Modelling & Software, v. 73, p. 148–166, accessed October 2019 at <https://doi.org/10.1016/j.envsoft.2015.07.017>.
- Hirsch, R.M., and De Cicco, L.A., 2015, User guide to Exploration and Graphics for RivEr Trends (EGRET) and data Retrieval—R packages for hydrologic data (ver. 2.0, February 2015): U.S. Geological Survey Techniques and Methods, book 4, chap. A10, 93 p., accessed October 2019 at <http://dx.doi.org/10.3133/tm4A10>.
- Hirsch, R.M., and Gilroy, E.J., 1984, Methods of fitting a straight line to data—Examples in water resources: Water Resources Bulletin, v. 20, no. 5, p. 705–711, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.1984.tb04753.x>.
- Hirsch, R.M., Slack, J.R., and Smith, R.A., 1982, Techniques of trend analysis for monthly water quality data: Water Resources Research, v. 18, no. 1, p. 107–121, accessed October 2019 at <https://doi.org/10.1029/WR018i001p00107>.
- Hirsch, R.M., Moyer, D.L., and Archfield, S.A., 2010, Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs: Journal of the American Water Resources Association, v. 46, no. 5, p. 857–880, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.2010.00482.x>.
- Hirsch, R.M., and Ryberg, K.R., 2012, Has the magnitude of floods across the USA changed with global CO<sub>2</sub> levels?: Hydrological Sciences Journal, v. 57, no. 1, p. 1–9, accessed October 2019 at <https://doi.org/10.1080/02626667.2011.621895>.
- Hirsch, R.M., and Slack, J.R., 1984, A nonparametric trend test for seasonal data with serial dependence: Water Resources Research, v. 20, no. 6, p. 727–732, accessed October 2019 at <https://doi.org/10.1029/WR020i006p00727>.

- Hoaglin, D.C., 1983, Letter values—A set of order statistics, chap. 2 of Hoaglin, D.C., Mosteller, F., and Tukey, J.W., eds., *Understanding robust and exploratory data analysis*: New York, John Wiley, p. 33–57.
- Hoaglin, D.C., 1988, Transformations in everyday experience: *Chance*, v. 1, p. 40–45.
- Hoaglin, D.C., Mosteller, F., and Tukey, J.W., 1983, *Understanding robust and exploratory data analysis*: New York, John Wiley, 447 p.
- Hochberg, Y., and Tamhane, A.C., 1987, *Multiple comparison procedures*: New York, Wiley, 480 p., accessed October 2019 at <https://doi.org/10.1002/9780470316672>.
- Hodges, J.L., Jr., and Lehmann, E.L., 1963, Estimates of location based on rank tests: *Annals of Mathematical Statistics*, v. 34, no. 2, p. 598–611, accessed October 2019 at <https://doi.org/10.1214/aoms/1177704172>.
- Hoerl, A.E., and Kennard, R.W., 1970, Ridge regression—Biased estimation for nonorthogonal problems: *Technometrics*, v. 12, no. 1, p. 55–67, accessed October 2019 at <https://doi.org/10.1080/00401706.1970.10488634>.
- Hollander, M., and Wolfe, D.A., 1999, *Nonparametric statistical methods*: New York, John Wiley and Sons, 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 Report 87–4035, 120 p., accessed October 2019 at <https://doi.org/10.3133/wri874035>.
- Hosking, J.R.M., 1990, L-Moments—Analysis and estimation of distributions using linear combinations of order statistics: *Journal of the Royal Statistical Society. Series B. Methodological*, v. 52, no. 2, p. 105–124.
- Hosmer, D., and Lemeshow, S., 2000, *Applied logistic regression*: New York, Wiley, 373 p.
- Hothorn, T., Hornik, K., Van De Wiel, M.A., and Zeileis, A., 2008, Implementing a class of permutation tests—The coin package: *Journal of Statistical Software*, v. 28, no. 8, p. 1–23, accessed October 2019 at <https://doi.org/10.18637/jss.v028.i08>.
- Hsieh, F.Y., Bloch, D.A., and Larsen, M.D., 1998, A simple method of sample size calculation for linear and logistic regression: *Statistics in Medicine*, v. 17, no. 14, p. 1623–1634, accessed October 2019 at [https://doi.org/10.1002/\(SICI\)1097-0258\(19980730\)17:14<1623::AID-SIM871>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S).
- Hurst, H.E., 1951, Long-term storage capacity of reservoirs: *Transactions of the American Society of Civil Engineers*, v. 116, p. 770–808.
- Hyndman, R.J., and Fan, Y., 1996, Sample quantiles in statistical packages: *The American Statistician*, v. 50, no. 4, p. 361–365, accessed March, 2020 at <https://doi.org/10.1080/00031305.1996.10473566>.
- Iman, R.L., and Conover, W.J., 1983, *A modern approach to statistics*: New York, John Wiley and Sons, 497 p.
- Iman, R.L., and Davenport, J.M., 1980, Approximations of the critical region of the Friedman statistic: *Communications in Statistics A*, no. 9, p. 571–595.
- Inman, D.L., 1952, Measures for describing the size distribution of sediments: *Journal of Sedimentary Petrology*, v. 22, p. 125–145.
- Jensen, A.L., 1973, Statistical analysis of biological data from preoperational-postoperational industrial water quality monitoring: *Water Research*, v. 7, no. 9, p. 1331–1347, accessed October 2019 at [https://doi.org/10.1016/0043-1354\(73\)90010-9](https://doi.org/10.1016/0043-1354(73)90010-9).
- Johnson, D.H., 1995, Statistical sirens—The allure of nonparametrics: *Ecology*, v. 76, no. 6, p. 1998–2000, accessed October 2019 at <https://doi.org/10.2307/1940733>.

- Johnson, R.A., and Wichern, D.W., 2002, Applied multivariate statistical analysis: Upper Saddle River, N.J., Prentice Hall, 800 p.
- Johnston, J., 1984, Econometric methods: New York, McGraw-Hill, 480 p.
- Joint Committee on Standards for Graphic Presentation, 1915, Preliminary report published for the purpose of inviting suggestions for the benefit of the committee: Publications of the American Statistical Association, v. 14, no. 112, p. 790–797.
- Jones, L.V., and Tukey, J.W., 2000, A sensible formulation of the significance test: Psychological Methods, v. 5, no. 4, p. 411–414, accessed October 2019 at <https://doi.org/10.1037/1082-989X.5.4.411>.
- Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., and Lee, T.C., 1985, Introduction to the theory and practice of econometrics: New York, John Wiley and Sons, 839 p.
- Junk, G.A., Spalding, R.F., and Richard, J.J., 1980, Areal, vertical, and temporal differences in ground water chemistry—II. Organic constituents 1: Journal of Environmental Quality, v. 9, no. 3, p. 479, accessed October 2019 at <https://doi.org/10.2134/jeq1980.0047245000900030031x>.
- Kass, R.E., and Raftery, A.E., 1995, Bayes factors: Journal of the American Statistical Association, v. 90, no. 430, p. 773–795, accessed October 2019 at <https://doi.org/10.1080/01621459.1995.10476572>.
- Kay, M., and Wobbrock, J.O., 2016, ARTTool—Aligned rank transform for nonparametric factorial ANOVAs (ver. 0.10.6): The Comprehensive R Archive Network web page, accessed October 2019 at <https://cran.r-project.org/package=ARTTool>.
- Kelleher, C., and Wagener, T., 2011, Ten guidelines for effective data visualization in scientific publications: Environmental Modelling & Software, v. 26, no. 6, p. 822–827, accessed October 2019 at <https://doi.org/10.1016/j.envsoft.2010.12.006>.
- Kendall, M.G., 1938, A new measure of rank correlation: Biometrika, v. 30, no. 1-2, p. 81–93, accessed October 2019 at <https://doi.org/10.1093/biomet/30.1-2.81>.
- Kendall, M.G., 1975, Rank correlation methods: London, Charles Griffin, 202 p.
- Kenney, J.F., and Keeping, E.S., 1954, Mathematics of statistics, Part One: New York, D. Van Nostrand, 102 p.
- Kermack, K.A., and Haldane, J.B.S., 1950, Organic correlation and allometry: Biometrika, v. 37, no. 1-2, p. 30–41, accessed October 2019 at <https://doi.org/10.1093/biomet/37.1-2.30>.
- Khalil, B., and Adamowski, J., 2012, Record extension for short-gauged water quality parameters using a newly proposed robust version of the line of organic correlation technique: Hydrology and Earth System Sciences, v. 16, no. 7, p. 2253–2266, accessed October 2019 at <https://doi.org/10.5194/hess-16-2253-2012>.
- Khalil, B., Ouarda, T.B., and St-Hilaire, A., 2012, Comparison of record-extension techniques for water quality variables: Water Resources Management, v. 26, no. 14, p. 4259–4280, accessed October 2019 at <https://doi.org/10.1007/s11269-012-0143-9>.
- Kirby, W., 1974a, Algebraic boundedness of sample statistics: Water Resources Research, v. 10, no. 2, p. 220–222, accessed October 2019 at <https://doi.org/10.1029/WR010i002p00220>.
- Kirby, W., 1974b, Straight line fitting of an observation path by least normal squares: U.S. Geological Survey Open-File Report 74-197, 11 p.
- Klemeš, V., 1974, The Hurst phenomenon—A puzzle?: Water Resources Research, v. 10, no. 4, p. 675–688, accessed October 2019 at <https://doi.org/10.1029/WR010i004p00675>.
- Knopman, D.S., 1990, Factors relating 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 90-4147, 52 p.

- Kotze, D.J., Johnson, C.A., O'Hara, R.B., Vepsäläinen, K., and Fowler, M.S., 2004, Editorial: Journal of Negative Results - Ecology & Evolutionary Biology, v. 1, no. 1, p. 1–5.
- Koutsoyiannis, D., 2002, The Hurst phenomenon and fractional Gaussian noise made easy: Hydrological Sciences Journal, v. 47, no. 4, p. 573–595, accessed October 2019 at <https://doi.org/10.1080/0262660209492961>.
- Kritskiy, S.N., and Menkel, J.F., 1968, Some statistical methods in the analysis of hydrologic data: Soviet Hydrology Selected Papers, v. 1, p. 80–98.
- Kroll, C.N., and Song, P., 2013, Impact of multicollinearity on small sample hydrologic regression models: Water Resources Research, v. 49, no. 6, p. 3756–3769, accessed October 2019 at <https://doi.org/10.1002/wrcr.20315>.
- Kruskal, W.H., 1953, On the uniqueness of the line of organic correlation: Biometrics, v. 9, no. 1, p. 47–58, accessed October 2019 at <https://doi.org/10.2307/3001632>.
- Kruskal, J.B., 1964, Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis: Psychometrika, v. 29, no. 1, p. 1–27, accessed October 2019 at <https://doi.org/10.1007/BF02289565>.
- Krzywinski, M., and Altman, N., 2014, Points of significance—Visualizing samples with box plots: Nature Methods, v. 11, no. 2, p. 119–120, accessed October 2019 at <https://doi.org/10.1038/nmeth.2813>.
- Kupper, L.L., and Hafner, K.B., 1989, How appropriate are popular sample size formulas?: The American Statistician, v. 43, p. 101–105, accessed February 2020 at <https://doi.org/10.1080/00031305.1989.10475628>.
- Kutner, M., Nachtsheim, C., Neter, J., and Li, W., 2004, Applied linear statistical models (5th ed.): New York, McGraw-Hill/Irwin, 1,396 p.
- Land, C.E., 1971, Confidence intervals for linear functions of the normal mean and variance: Annals of Mathematical Statistics, v. 42, no. 4, p. 1187–1205, accessed October 2019 at <https://doi.org/10.1214/aoms/1177693235>.
- Land, C.E., 1972, An evaluation of approximate confidence interval estimation methods for lognormal means: Technometrics, v. 14, no. 1, p. 145–158, accessed October 2019 at <https://doi.org/10.1080/0040706.1972.10488891>.
- Langbein, W.B., 1960, Plotting positions in frequency analysis, in Dalrymple, T., ed., Flood frequency analysis: U.S. Geological Survey Water-Supply Paper 1543-A, p. 48–51.
- Larned, S., Snelder, T., Unwin, M., and McBride, G., 2016, Water quality in New Zealand rivers—Current state and trends: New Zealand Journal of Marine and Freshwater Research, v. 50, no. 3, p. 389–417, accessed October 2019 at <https://doi.org/10.1080/00288330.2016.1150309>.
- Lasi, M.A., Helsel, D.R., and Tweedale, W.A., 2011, Developing optical water quality models using geometric mean functional regression for the Indian River Lagoon system, Florida, in 21st Biennial Conference of the Coastal and Estuarine Research Federation, Daytona Beach, Fla.
- Lavagnini, I., Badocco, D., Pastore, P., and Magno, F., 2011, Theil–Sen nonparametric regression technique on univariate calibration, inverse regression and detection limits: Talanta, v. 87, p. 180–188, accessed October 2019 at <https://doi.org/10.1016/j.talanta.2011.09.059>.
- Lederman, N.G., and Lederman, J.S., 2016, Publishing findings that are not significant—Can non-significant findings be significant?: Journal of Science Teacher Education, v. 27, no. 4, p. 349–355, accessed October 2019 at <https://doi.org/10.1007/s10972-016-9475-2>.
- Lehmann, E.L., 1975, Nonparametrics, statistical methods based on ranks: Oakland, Calif., Holden-Day, 457 p.

- Lettenmaier, D.P., 1976, Detection of trends in water quality data from records with dependent observations: *Water Resources Research*, v. 12, no. 5, p. 1037–1046, accessed October 2019 at <https://doi.org/10.1029/WR012i005p01037>.
- Levine, T.R., 2013, A defense of publishing nonsignificant (ns) results: *Communication Research Reports*, v. 30, no. 3, p. 270–274, accessed October 2019 at <https://doi.org/10.1080/08824096.2013.806261>.
- Lewandowsky, S., and Spence, I., 1989, Discriminating strata in scatterplots: *Journal of the American Statistical Association*, v. 84, p. 682–688, accessed October 2019 at <https://doi.org/10.1080/01621459.1989.10478821>
- Libiseller, C., and Grimvall, A., 2002, Performance of partial Mann–Kendall tests for trend detection in the presence of covariates: *Environmetrics*, v. 13, no. 1, p. 71–84, accessed October 2019 at <https://doi.org/10.1002/env.507>.
- Lin, S.D., and Evans, R.L., 1980, Coliforms and fecal streptococcus in the Illinois River at Peoria, 1971–1976: *Illinois State Water Survey Report of Investigation* 93, 28 p.
- Loftis, J.C., McBride, G.B., and Ellis, J.C., 1991, Considerations of scale in water quality monitoring and data analysis: *Journal of the American Water Resources Association*, v. 27, no. 2, p. 255–264, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.1991.tb03130.x>.
- Looney, S.W., and Gulledge, T.R., 1985, Probability plotting positions and goodness of fit for the normal distribution: *The Statistician*, v. 34, no. 3, p. 297–303, accessed October 2019 at <https://doi.org/10.2307/2987656>.
- Lumley, T., 2017, leaps—Regression subset selection (ver. 3.0): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=leaps>.
- Maciak, W., 2009, Exact null distribution for  $n \leq 25$  and probability approximations for Spearman’s score in an absence of ties: *Journal of Nonparametric Statistics*, v. 21, no. 1, p. 113–133, accessed October 2019 at <https://doi.org/10.1080/10485250802401038>.
- Mallows, C.L., 1973, Some comments on  $C_p$ : *Technometrics*, v. 15, no. 4, p. 661–675, accessed February 2020 at <https://doi.org/10.1080/00401706.1973.10489103>.
- Mandelbrot, B.B., and Wallis, J.R., 1968, Noah, Joseph, and operational hydrology: *Water Resources Research*, v. 4, no. 5, p. 909–918, accessed October 2019 at <https://doi.org/10.1029/WR004i005p00909>.
- Manly, B.F.J., 2007, Randomization, bootstrap and Monte Carlo methods in biology (3d ed.): Boca Raton, Fla., Chapman & Hall/CRC, 455 p.
- Mann, H.B., 1945, Nonparametric test against trend: *Econometrica*, v. 13, no. 3, p. 245–259, accessed October 2019 at <https://doi.org/10.2307/1907187>.
- Mann, H.B., and Whitney, D.R., 1947, On a test of whether one of two random variables is stochastically larger than the other: *Annals of Mathematical Statistics*, v. 18, no. 1, p. 50–60, accessed October 2019 at <https://doi.org/10.1214/aoms/1177730491>.
- Mansouri, H., Paige, R.L., and Surles, J.G., 2004, Aligned rank transform techniques for analysis of variance and multiple comparisons: *Communications in Statistics. Theory and Methods*, v. 33, no. 9, p. 2217–2232, accessed October 2019 at <https://doi.org/10.1081/STA-200026599>.
- Marchetto, A., 2017, rkt—Mann-Kendall test, seasonal and regional Kendall tests (ver. 1.5): The Comprehensive R Archive Network web page, accessed October 2019 at <https://cran.r-project.org/package=rkt>.
- Marquardt, D.W., 1970, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation: *Technometrics*, v. 12, no. 3, p. 591–612, accessed October 2019 at <https://doi.org/10.2307/1267205>.

- Marquet, P.A., 2000, Invariants, scaling laws, and ecological complexity: *Science*, v. 289, no. 5484, p. 1487–1488, accessed October 2019 at <https://doi.org/10.1126/science.289.5484.1487>.
- Matalas, N.C., and Langbein, W.B., 1962, Information content of the mean: *Journal of Geophysical Research*, v. 67, no. 9, p. 3441–3448, accessed October 2019 at <https://doi.org/10.1029/JZ067i009p03441>.
- Matalas, N.C., and Sankarasubramanian, A., 2003, Effect of persistence on trend detection via regression: *Water Resources Research*, v. 39, no. 12, accessed October 2019 at <https://doi.org/10.1029/2003WR002292>.
- McBride, G., 2019, Has water quality improved or been maintained? A quantitative assessment procedure: *Journal of Environmental Quality*, v. 48, no. 2, p. 412–420, accessed October 2019 at <http://doi.org/10.2134/jeq2018.03.0101>.
- McBride, G., Cole, R.G., Westbrooke, I., and Jowett, I., 2014, Assessing environmentally significant effects—A better strength-of-evidence than a single P value?: *Environmental Monitoring and Assessment*, v. 186, no. 5, p. 2729–2740, accessed October 2019 at <https://doi.org/10.1007/s10661-013-3574-8>.
- McCuen, R.H., 2016, Assessment of hydrological and statistical significance: *Journal of Hydrologic Engineering*, v. 21, no. 4, 2 p., accessed October 2019 at [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001340](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001340).
- McGill, R., Tukey, J.W., and Larsen, W.A., 1978, Variations of box plots: *The American Statistician*, v. 32, p. 12–16, accessed February 2020 at <https://doi.org/10.1080/00031305.1978.10479236>.
- McLeod, A.I., 2011, Kendall—Kendall rank correlation and Mann-Kendall trend test (ver. 2.2): The Comprehensive R Archive Network web page, accessed October 2019 at <https://cran.r-project.org/package=Kendall>.
- Millard, S.P., 2013, EnvStats—An R package for environmental statistics: New York, Springer, 291 p., accessed October 2019 at <https://cran.r-project.org/package=EnvStats>.
- Millard, S.P., and Neerchal, N.K., 2000, Environmental statistics with S-Plus: Boca Raton, Fla., CRC Press, 848 p., accessed October 2019 at <https://doi.org/10.1201/9781420037173>.
- Miller, J.B., and Tans, P.P., 2003, Calculating isotopic fractionation from atmospheric measurements at various scales: *Tellus. Series B, Chemical and Physical Meteorology*, v. 55, no. 2, p. 207–214, accessed October 2019 at <https://doi.org/10.3402/tellusb.v55i2.16697>.
- Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., and Stouffer, R.J., 2008, Stationarity is dead—Whither water management?: *Science*, v. 319, no. 5863, p. 573–574, accessed October 2019 at <https://doi.org/10.1126/science.1151915>.
- Montgomery, D.C., 1991, Introduction to statistical quality control: New York, John Wiley, 674 p.
- Montgomery, D.C., and Peck, E.A., 1982, Introduction to linear regression analysis: New York, John Wiley, 504 p.
- Montgomery, D.C., Peck, E.A., and Vining, G.G., 2012, Introduction to linear regression analysis—Wiley Series in Probability and Statistics: Hoboken, N.J., John Wiley & Sons, Inc., 672 p.
- Mosteller, F., and Tukey, J.W., 1977, Data analysis and regression: Menlo Park, Calif., Addison-Wesley Publishers, 588 p.
- Mosteller, F., Siegel, A.F., Trapido, E., and Youtz, C., 1981, Eye fitting straight lines: *The American Statistician*, v. 35, no. 3, p. 150–152, accessed February 2020 at <https://doi.org/10.1080/00031305.1981.10479335>.
- Munroe, R., 2016, Significant: xkcd web page, accessed October 2019 at <https://xkcd.com/882/>.

- Murphy, J.C., Hirsch, R.M., and Sprague, L.A., 2014, Antecedent flow conditions and nitrate concentrations in the Mississippi River basin: *Hydrology and Earth System Sciences*, v. 18, no. 3, p. 967–979, accessed October 2019 at <https://doi.org/10.5194/hess-18-967-2014>.
- Murrell, P., 2011, *R graphics*: New York, CRC Press, The R Series, 546 p.
- Mustard, M.H., Driver, N.E., Chyr, J., and Hansen, B.G., 1987, U.S. Geological Survey urban-stormwater data base of constituent storm loads; characteristics of rainfall, runoff, and antecedent conditions; and basin characteristics: U.S. Geological Survey Water-Resources Investigations Report 87-4036, 328 p.
- Nagelkerke, N.J.D., 1991, A note on a general definition of the coefficient of determination: *Biometrika*, v. 78, no. 3, p. 691–692, accessed October 2019 at <https://doi.org/10.1093/biomet/78.3.691>.
- Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., and Toxopeus, A.G., 2014, Where is positional uncertainty a problem for species distribution modelling?: *Ecography*, v. 37, no. 2, p. 191–203, accessed October 2019 at <https://doi.org/10.1111/j.1600-0587.2013.00205.x>.
- Nakazawa, M., 2017, *fmsb*—Functions for medical statistics book with some demographic data: The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=fmsb>.
- National Eye Institute, 2015, Color Blindness: National Eye Institute web page, accessed October 2019 at [https://nei.nih.gov/health/color\\_blindness/facts\\_about](https://nei.nih.gov/health/color_blindness/facts_about).
- Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., 1996, *Applied linear statistical models*: New York, WCB McGraw-Hill, 1,408 p.
- Neter, J., Wasserman, W., and Kutner, M.H., 1985, *Applied linear statistical models*: Homewood, Ill., Irwin Publishers, 1,127 p.
- Nevitt, J., and Tam, H.P., 1997, A comparison of robust and nonparametric estimators under the simple linear regression model: Annual meeting of the American Educational Research Association, Chicago, Ill., March 24–28, 1997, 32 p., accessed October 2019 at <https://eric.ed.gov/?id=ED410289>.
- Nightingale, F., 1858, Notes on matters affecting the health, efficiency, and hospital administration of the British Army—Founded chiefly on the experience of the late war: London, Harrison and Sons, accessed January 30, 2019, at <https://omeka.lehigh.edu/items/show/3221>.
- Noether, G.E., 1987, Sample size determination for some common nonparametric tests: *Journal of the American Statistical Association*, v. 82, no. 398, p. 645–647, accessed October 2019 at <https://doi.org/10.1080/01621459.1987.10478478>.
- Nuzzo, R., 2014, Scientific method—Statistical errors: *Nature*, v. 506, no. 7487, p. 150–152, accessed October 2019 at <https://doi.org/10.1038/506150a>.
- Nyblom, J., 1992, Note on interpolated order statistics: *Statistics & Probability Letters*, v. 14, no. 2, p. 129–131.
- O'Brien, R.M., 2007, A caution regarding rules of thumb of variance inflation factors: *Quality & Quantity*, v. 41, no. 5, p. 673–690, accessed October 2019 at <https://doi.org/10.1007/s11135-006-9018-6>.
- Oelsner, G.P., Sprague, L.A., Murphy, J.C., Zuellig, R.E., Johnson, H.M., Ryberg, K.R., Falcone, J.A., Stets, E.G., Vecchia, A.V., and Riskin, M.L., 2017, Water-quality trends in the nation's rivers and streams, 1972–2012—Data preparation, statistical methods, and trend results: U.S. Geological Survey Scientific Investigations Report 2017-5006, 136 p., accessed January 2019 at <https://doi.org/10.3133/sir20175006>.
- Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., and Wagner, H., 2016, *vegan*—Community ecology package (ver. 2.4-1): The Comprehensive R Archive Network web page, accessed October 2019 at <https://cran.r-project.org/package=vegan>.

- Oltmann, R.N., and Shulters, M.V., 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., accessed October 2019 at <https://doi.org/10.3133/wsp2335>.
- Önöz, B., and Bayazit, M., 2012, Block bootstrap for Mann–Kendall trend test of serially dependent data: *Hydrological Processes*, v. 26, no. 23, p. 3552–3560, accessed October 2019 at <https://doi.org/10.1002/hyp.8438>.
- Parzen, E., 1979, Nonparametric statistical data modeling: *Journal of the American Statistical Association*, v. 74, no. 365, p. 105–121, accessed October 2019 at <https://doi.org/10.1080/01621459.1979.10481621>.
- Person, M., Antle, R., and Stephens, D.B., 1983, Evaluation of surface impoundment assessment in New Mexico: *Groundwater*, v. 21, no. 6, p. 679–688, accessed October 2019 at <https://doi.org/10.1111/j.1745-6584.1983.tb01939.x>.
- Pettitt, A.N., 1979, A non-parametric approach to the change-point problem: *Journal of the Royal Statistical Society, Series C, Applied Statistics*, v. 28, no. 2, p. 126–135.
- Piper, A.M., 1944, A graphic procedure in the geochemical interpretation of water-analyses: *Transactions—American Geophysical Union*, v. 25, no. 6, p. 914–928, accessed October 2019 at <https://doi.org/10.1029/TR025i006p00914>.
- Playfair, W., 1801a, The commercial and political atlas, representing, by means of copper-plate charts, the progress of the commerce, revenues, expenditure, and debts of England, during the whole of the eighteenth century—The third edition, corrected and brought down to the end of last year: London, Wallis. [Reprinted as a facsimile in 2005 by Cambridge University Press with an introduction by Howard Wainer and Ian Spence in a volume also containing “The Statistical Breviary” and titled “The Commercial and Political Atlas and Statistical Breviary.”]
- Playfair, W., 1801b, The statistical breviary—Shewing, on a principle entirely new, the resources of every and state and kingdom in Europe; illustrated with stained copper-plate charts ...: London, Wallis. [Reprinted as a facsimile in 2005 by Cambridge University Press with an introduction by Howard Wainer and Ian Spence in a volume also containing “The Commercial and Political Atlas” and titled “The Commercial and Political Atlas and Statistical Breviary.”]
- Pocock, S.J., 1982, When not to rely on the central limit theorem—An example from absenteeism data: *Communications in Statistics—Theory and Methods*, v. 11, no. 19, p. 2169–2179, accessed October 2019 at <https://doi.org/10.1080/03610928208828379>.
- Pohlert, T., 2014, The pairwise multiple comparison of mean ranks package (PMCMR): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=PMCMR>.
- Pohlert, T., 2017, ppcc—Probability plot correlation coefficient test (ver. 1.0): The Comprehensive R Archive Network web page, accessed October 2019 at <https://CRAN.R-project.org/package=ppcc>.
- Ponce, V.M., 1989, Engineering hydrology—Principles and practices: Englewood Cliffs, N.J., Prentice-Hall, 640 p.
- Press, S.J., and Wilson, S., 1978, Choosing between logistic regression and discriminant analysis: *Journal of the American Statistical Association*, v. 73, no. 364, p. 699–705, accessed October 2019 at <https://doi.org/10.1080/01621459.1978.10480080>.
- R Core Team, 2016, R—A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria, accessed October 2019 at <http://www.R-project.org/>.
- R Core Team, 2017, R—A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria, accessed October 2019 at <http://www.R-project.org/>.
- Reiss, R.-D., 1989, Approximate distributions of order statistics—With applications to nonparametric statistics—Springer Series in Statistics: New York, Springer Science+ Business Media, 354 p.

- Richter, S.J., and Payton, M.E., 2005, An improvement to the aligned-rank statistic for two-factor analysis of variance: *Journal of Applied Statistical Science*, v. 14, p. 225–235.
- Robbins, N.B., 2005, Creating more effective graphs: Hoboken, N.J., Wiley-Interscience, 402 p.
- Robertson, W.D., Barker, J.F., LeBeau, Y., and Marcoux, S., 1984, Contamination of an unconfined sand aquifer by waste pulp liquor—A case study: *Groundwater*, v. 22, no. 2, p. 191–197, accessed October 2019 at <https://doi.org/10.1111/j.1745-6584.1984.tb01488.x>.
- Roser, L.G., Ferreyra, L.I., Saidman, B.O., and Vilardi, J.C., 2017, EcoGenetics—An R package for the management and exploratory analysis of spatial data in landscape genetics: *Molecular Ecology Resources*, v. 17, no. 6, p. e241–e250, accessed October 2019 at <https://doi.org/10.1111/1755-0998.12697>.
- Rozeboom, W.W., 1960, The fallacy of the null-hypothesis significance test: *Psychological Bulletin*, v. 57, no. 5, p. 416–428, accessed October 2019 at <https://doi.org/10.1037/h0042040>.
- Runkel, R.L., Crawford, C.G., and Cohn, T.A., 2004, Load estimator (LOADEST)—A FORTRAN program for estimating constituent loads in streams and rivers: U.S. Geological Survey Techniques and Methods, book 4, chap. A5, 69 p.
- Ryberg, K.R., Vecchia, A.V., Gilliom, R.J., and Martin, J.D., 2014, Pesticide trends in major rivers of the United States, 1992–2010: U.S. Geological Survey Scientific Investigations Report 2014–5135, 63 p., accessed October 2019 at <https://doi.org/10.3133/sir20145135>.
- Salas, J.D., Obeysekera, J., and Vogel, R.M., 2018, Techniques for assessing water infrastructure for nonstationary extreme events—A review: *Hydrological Sciences Journal*, v. 63, no. 3, p. 325–352.
- Salsburg, D., 2001, The lady tasting tea—How statistics revolutionized science in the twentieth century: New York, W.H. Freeman and Company, 340 p.
- Samuelson, P.A., 1942, A note on alternative regressions: *Econometrica*, v. 10, no. 1, p. 80.
- Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D., and Yevjevich, V., 1983, Design of networks for monitoring water quality: Littleton, Colo., Water Resources Publications, 328 p.
- Sarkar, D., 2008, Lattice—Multivariate data visualization with R: New York, Springer Science+Business Media, 268 p.
- SAS Institute Inc, 2014, SAS/STAT(R) 13.2 User's guide: Cary, N.C., SAS Institute Inc., 9,429 p., accessed October 2019 at <https://support.sas.com/en/software/sas-stat-support.html#documentation>.
- Schertz, T.L., and Hirsch, R.M., 1985, Trend analysis of weekly acid rain data, 1978–83: U.S. Geological Survey Water-Resources Investigations Report 85–4211, 64 p., accessed October 2019 at <https://doi.org/10.3133/wri854211>.
- Schmid, C.F., 1983, Statistical graphics: New York, John Wiley and Sons, 212 p.
- Schroder, L.J., Brooks, M.H., and Willoughby, T.C., 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., accessed October 2019 at <https://doi.org/10.3133/wri864363>.
- Schwarz, G., 1978, Estimating the dimension of a model: *Annals of Statistics*, v. 6, no. 2, p. 461–464, accessed October 2019 at <https://doi.org/10.1214/aos/1176344136>.
- Sen, P.K., 1968, Estimates of the regression coefficient based on Kendall's tau: *Journal of the American Statistical Association*, v. 63, no. 324, p. 1379–1389, accessed October 2019 at <https://doi.org/10.1080/01621459.1968.10480934>.

- Shapiro, S.S., and Francia, R.S., 1972, An approximate analysis of variance test for normality: *Journal of the American Statistical Association*, v. 67, no. 337, p. 215–216, accessed October 2019 at <https://doi.org/10.1080/01621459.1972.10481232>.
- Shapiro, S.S., Wilk, M.B., and Chen, H.J., 1968, A comparative study of various tests for normality: *Journal of the American Statistical Association*, v. 63, no. 324, p. 1343–1372, accessed October 2019 at <https://doi.org/10.1080/01621459.1968.10480932>.
- Shelton, J.L., Fram, M.S., Munday, C.M., and Belitz, K., 2010, Groundwater-quality data for the Sierra Nevada study unit, 2008—Results from the California GAMA Program: U.S. Geological Survey Data Series 534, 106 p., accessed October 2019 at <https://doi.org/10.3133/ds534>.
- Sheskin, D.J., 2011, *Handbook of parametric and nonparametric statistical procedures*: Boca Raton, Fla., Chapman and Hall/CRC Press, 1,926 p.
- Singh, A.K., Singh, A., and Engelhardt, M., 1999, Some practical aspects of sample size and power computations for estimating the mean of positively skewed distributions in environmental applications. U.S. Environmental Protection Agency, 37 p.
- Smith, R.A., Alexander, R.B., and Wolman, M.G., 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., accessed October 2019 at <https://doi.org/10.3133/wsp2307>.
- Smith, T.J., and McKenna, C.M., 2013, A comparison of logistic regression pseudo  $R^2$  indices: *Multiple Linear Regression Viewpoints*, v. 39, no. 2, p. 17–26, accessed October 2019 at [http://www.glmj.org/archives/articles/Smith\\_v39n2.pdf](http://www.glmj.org/archives/articles/Smith_v39n2.pdf).
- Solley, W.B., Chase, E.B., and Mann, W.B., 1983, Estimated use of water in the United States in 1980: U.S. Geological Survey Circular 1001, 56 p., accessed August 6, 2014, at <https://doi.org/10.3133/cir1001>.
- Solley, W.B., Merk, C.F., and Pierce, R.R., 1988, Estimated use of water in the United States in 1985: U.S. Geological Survey Circular 1004, 94 p., accessed August 8, 2014, at <http://doi.org/10.3133/cir1004>.
- Spence, I., 2005, No humble pie—The origins and usage of a statistical chart: *Journal of Educational and Behavioral Statistics*, v. 30, no. 4, p. 353–368, accessed October 2019 at <https://doi.org/10.3102/10769986030004353>.
- Stedinger, J.R., 1983, Confidence intervals for design events: *Journal of Hydraulic Engineering*, v. 109, no. 1, p. 13–27, accessed October 2019 at [https://doi.org/10.1061/\(ASCE\)0733-9429\(1983\)109:1\(13\)](https://doi.org/10.1061/(ASCE)0733-9429(1983)109:1(13)).
- Stedinger, J., Vogel, R., and Foufoula-Georgiou, E., 1993, Frequency analysis of extreme events, chap. 18 of Maidment, D.R., ed., *Handbook of hydrology*: New York, McGraw-Hill, p. 18-1–18-66.
- Stelzer, E.A., Strickler, K.M., and Schill, W.B., 2012, Interlaboratory comparison of three microbial source tracking quantitative polymerase chain reaction (qPCR) assays from fecal-source and environmental samples: U.S. Geological Survey Scientific Investigations Report 2012–5087, 10 p., accessed October 2019 at <https://doi.org/10.3133/sir20125087>.
- Strange, N., 2007, *Smoke and mirrors—How to bend facts and figures to your advantage*: London, A&C Black Publishers, 2,008 p.
- Streit, M., and Gehlenborg, N., 2014, Points of view—Bar charts and box plots: *Nature Methods*, v. 11, no. 2, p. 117, accessed October 2019 at <https://www.nature.com/articles/nmeth.2807>.
- Strömberg, G., 1940, Accidental and systematic errors in spectroscopic absolute magnitudes for dwarf  $G_{\{0\}}K_{\{2\}}$  stars: *The Astrophysical Journal*, v. 92, p. 156, accessed October 2019 at <https://doi.org/10.1086/144209>.
- Sutton, C.D., 1993, Computer-intensive methods for tests about the mean of an asymmetrical distribution: *Journal of the American Statistical Association*, v. 88, no. 423, p. 802–810, accessed February 2020 at <https://doi.org/10.1080/01621459.1993.10476345>.

- Teissier, G., 1948, La relation d'allometrie sa signification statistique et biologique: Biometrics, v. 4, no. 1, p. 14–53, accessed October 2019 at <https://doi.org/10.2307/3001695>.
- Tesoriero, A.J., Terziotti, S., and Abrams, D.B., 2015, Predicting redox conditions in groundwater at a regional scale: Environmental Science & Technology, v. 49, no. 16, p. 9657–9664, accessed October 2019 at <https://doi.org/10.1021/acs.est.5b01869>.
- Theil, H., 1950, A rank-invariant method of linear and polynomial regression analysis, I, II, and III: Proceedings of the Royal Netherlands Academy of Sciences, v. 53, p. 386–392, 521–525, and 1397–1412, accessed October 2019 at [https://link.springer.com/chapter/10.1007/978-94-011-2546-8\\_20](https://link.springer.com/chapter/10.1007/978-94-011-2546-8_20).
- Thode, H.C., 2002, Testing for normality: New York, Marcel Dekker, Inc., 479 p., accessed October 2019 at <https://doi.org/10.1201/9780203910894>.
- Tofallis, C., 2002, Model fitting for multiple variables by minimising the geometric mean deviation, in Van Huffel, S., and Lemmerling, P., eds., Total least squares and errors-in-variables modeling: Dordrecht, Springer, p. 261–267, accessed October 2019 at [https://doi.org/10.1007/978-94-017-3552-0\\_23](https://doi.org/10.1007/978-94-017-3552-0_23).
- Tufte, E.R., 1983, The visual display of quantitative information: Cheshire, Conn., Graphics Press LLC, 197 p.
- Tufte, E.R., 1990, Envisioning information: Cheshire, Conn., Graphics Press LLC, 126 p.
- Tufte, E.R., 1997, Visual explanations—Images and quantities, evidence and narrative: Cheshire, Conn., Graphics Press LLC, 156 p.
- Tufte, E.R., 2006, Beautiful evidence: Cheshire, Conn., Graphics Press LLC, 213 p.
- Tukey, J.W., 1960, Conclusions vs decisions: Technometrics, v. 2, no. 4, p. 423–433, accessed October 2019 at <https://doi.org/10.1080/00401706.1960.10489909>.
- Tukey, J.W., 1977, Exploratory data analysis: Reading, Mass., Addison-Wesley, 712 p.
- United Nations Economic Commission for Europe, 2009, Making data meaningful, part 2—A guide to presenting statistics: Geneva, United Nations, 52 p.
- Unwin, A., 2015, Graphical data analysis with R—The R Series: Boca Raton, Fla., Chapman and Hall/CRC Press, 310 p.
- U.S. Environmental Protection Agency, 2002, Calculating upper confidence limits for exposure point concentrations at hazardous waste sites: OSWER 9285.6-10, 32 p., accessed March 10, 2017, at [nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100CYCE.TXT](http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100CYCE.TXT).
- U.S. Environmental Protection Agency, 2009, Statistical analysis of groundwater monitoring data at RCRA facilities: EPA 530-R-09-007, 884 p.
- U.S. Geological Survey, 2016, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed October 2019 at <http://dx.doi.org/10.5066/F7P55KJN>.
- van Belle, G., and Hughes, J.P., 1984, Nonparametric tests for trend in water quality: Water Resources Research, v. 20, no. 1, p. 127–136, accessed October 2019 at <https://doi.org/10.1029/WR020i001p00127>.
- Vannest, K.J., Parker, R.I., Davis, J.L., Soares, D.A., and Smith, S.L., 2012, The Theil–Sen slope for high-stakes decisions from progress monitoring: Behavioral Disorders, v. 37, no. 4, p. 271–280.
- Vatcheva, K.P., Lee, M., McCormick, J.B., and Rahbar, M.H., 2016, Multicollinearity in regression analyses conducted in epidemiologic studies: Epidemiology, v. 6, no. 2, p. 9, accessed October 2019 at <https://doi.org/10.4172/2161-1165.1000227>.
- Vecchia, A.V., 2005, Water-quality trend analysis and sampling design for streams in the Red River of the North basin, Minnesota, North Dakota, and South Dakota, 1970–2001: U.S. Geological Survey Scientific Investigations Report 2005–5224, 60 p.

- Vecchia, A.V., Martin, J.D., and Gilliom, R.J., 2008, Modeling variability and trends in pesticide concentrations in streams: *Journal of the American Water Resources Association*, v. 44, no. 5, p. 1308–1324, accessed October 2019 at <https://doi.org/10.1111/j.1752-1688.2008.00225.x>.
- Velleman, P.F., and Hoaglin, D.C., 1981, Applications, basics, and computing of exploratory data analysis: Boston, Mass., Duxbury Press, 354 p.
- Venables, W.N., and Ripley, B.D., 2002, Modern applied statistics with S: New York, Springer, 498 p., accessed October 2019 at <https://doi.org/10.1007/978-0-387-21706-2>.
- Vogel, R.M., 1986, The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distributional hypotheses: *Water Resources Research*, v. 22, no. 4, p. 587–590, accessed October 2019 at <https://doi.org/10.1029/WR022i004p00587>.
- Vogel, R.M., and Kroll, C.N., 1989, Low-flow frequency analysis using probability plot correlation coefficients: *Journal of Water Resources Planning and Management*, v. 115, no. 3, p. 338–357, accessed October 2019 at [https://doi.org/10.1061/\(ASCE\)0733-9496\(1989\)115:3\(338\)](https://doi.org/10.1061/(ASCE)0733-9496(1989)115:3(338)).
- Vogel, R.M., Rosner, A., and Kirshen, P.H., 2013, Brief communication, Likelihood of societal preparedness for global change—Trend detection: *Natural Hazards and Earth System Sciences*, v. 13, no. 7, p. 1773–1778, accessed October 2019 at <https://doi.org/10.5194/nhess-13-1773-2013>.
- Vogel, R.M., and Stedinger, J.R., 1985, Minimum variance streamflow record augmentation procedures: *Water Resources Research*, v. 21, no. 5, p. 715–723, accessed October 2019 at <https://doi.org/10.1029/WR021i005p00715>.
- von Humboldt, A., 1811, *Essai politique sur le royaume de la nouvelle-espagne*. (Political essay on the kingdom of New Spain—Founded on astronomical observations, and trigonometrical and barometrical measurements), vol. 1, translated to English by John Black: New York, Riley, 310 p.
- Wallis, J.R., Matalas, N.C., and Slack, J.R., 1974, Just a moment: *Water Resources Research*, v. 10, no. 2, p. 211–219, accessed October 2019 at <https://doi.org/10.1029/WR010i002p00211>.
- Walpole, R.E., and Myers, R.H., 1985, Probability and statistics for engineers and scientists: New York, MacMillan Publishing, 639 p.
- Wang, W., Chen, Y., Becker, S., and Liu, B., 2015, Variance correction prewhitening method for trend detection in autocorrelated data: *Journal of Hydrologic Engineering*, v. 20, no. 12, p. 04015033, accessed October 2019 at [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001234](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001234).
- Warton, D.I., and Weber, N.C., 2002, Common slope tests for bivariate errors-in-variables models: *Biometrical Journal*, v. 44, no. 2, p. 161–174, accessed October 2019 at [https://doi.org/10.1002/1521-4036\(200203\)44:2<161::AID-BIMJ161>3.0.CO;2-N](https://doi.org/10.1002/1521-4036(200203)44:2<161::AID-BIMJ161>3.0.CO;2-N).
- Warwick, R.M., 1971, Nematode associations in the Exe estuary: *Journal of the Marine Biological Association of the United Kingdom*, v. 51, no. 2, p. 439–454, accessed October 2019 at <https://doi.org/10.1017/S0025315400031908>.
- Wasserstein, R.L., and Lazar, N.A., 2016, The ASA's statement on p-values—Context, process, and purpose: *The American Statistician*, v. 70, no. 2, p. 129–133, accessed October 2019 at <https://doi.org/10.1080/00031305.2016.1154108>.
- Weibull, W., 1939, The phenomenon of rupture in solids: *Ingenjörsvetenskaps Akademien Handlingar*, no. 153, p. 17.
- Welch, B.L., 1951, On the comparison of several mean values: *Biometrika*, v. 38, no. 3–4, p. 330–336, accessed October 2019 at <https://doi.org/10.1093/biomet/38.3-4.330>.
- Welch, A.H., Lico, M.S., and Hughes, J.L., 1988, Arsenic in ground water of the western United States: *Groundwater*, v. 26, no. 3, p. 333–347, accessed October 2019 at <https://doi.org/10.1111/j.1745-6584.1988.tb00397.x>.

- Wickham, H., 2016, *ggplot2—Elegant graphics for data analysis*: New York, Springer-Verlag, 213 p.
- Wilcox, R., 1998, A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic: *Biometrical Journal*, v. 40, no. 3, p. 261–268, accessed October 2019 at [https://doi.org/10.1002/\(SICI\)1521-4036\(199807\)40:3<261::AID-BIMJ261>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1521-4036(199807)40:3<261::AID-BIMJ261>3.0.CO;2-V).
- Wilcoxon, F., 1945, Individual comparisons by ranking methods: *Biometrics Bulletin*, v. 1, no. 6, p. 80–83, accessed October 2019 at <https://doi.org/10.2307/3001968>.
- Williams, G.P., and Wolman, M.G., 1984, Downstream effects of dams on alluvial rivers: U.S. Geological Survey Professional Paper 1286, 83 p., accessed August 8, 2014, at <https://doi.org/10.3133/pp1286>.
- Wobbrock, J.O., Findlater, L., Gergle, D., and Higgins, J.J., 2011, The aligned rank transform for nonparametric factorial analyses using only anova procedures: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – ACM*, May 7–12, 2011, Vancouver, B.C., p. 143–146, accessed October 2019 at <https://doi.org/10.1145/1978942.1978963>.
- Wood, S.N., 2017, *Generalized additive models—An introduction with R* (2d ed.): Boca Raton, Fla., CRC Press, 476 p.
- World Health Organization, 2015, WHO Public disclosure of clinical trial results: World Health Organization International Clinical Trials Registry Platform web page, accessed January 2019 at <http://www.who.int/ictrp/results/en/>.
- Wright, W.G., 1985, Effects of fracturing on well yields in the coalfield areas of Wise and Dickenson Counties, southwestern Virginia: U.S. Geological Survey Water-Resources Investigations Report 85–4061, 21 p., accessed October 2019 at <https://doi.org/10.3133/wri854061>.
- Yue, S., Pilon, P., and Phinney, B., 2003, Canadian streamflow trend detection—Impacts of serial and cross-correlation: *Hydrological Sciences Journal*, v. 48, no. 1, p. 51–63, accessed October 2019 at <https://doi.org/10.1623/hysj.48.1.51.43478>.
- Yue, S., Pilon, P., Phinney, B., and Cavadias, G., 2002, The influence of autocorrelation on the ability to detect trend in hydrological series: *Hydrological Processes*, v. 16, no. 9, p. 1807–1829, accessed October 2019 at <https://doi.org/10.1002/hyp.1095>.
- Zeileis, A., and Grothendieck, G., 2005, zoo—S3 infrastructure for regular and irregular time series: *Journal of Statistical Software*, v. 14, no. 6, p. 1–27, accessed October 2019 at <https://doi.org/10.18637/jss.v014.i06>.
- Zhang, Q., and Ball, W.P., 2017, Improving riverine constituent concentration and flux estimation by accounting for antecedent discharge conditions: *Journal of Hydrology*, v. 547, p. 387–402, accessed October 2019 at <https://doi.org/10.1016/j.jhydrol.2016.12.052>.
- Zhang, X., Vincent, L.A., Hogg, W., and Niitsoo, A., 2000, Temperature and precipitation trends in Canada during the 20th century: *Atmosphere-Ocean*, v. 38, no. 3, p. 395–429, accessed October 2019 at <https://doi.org/10.1080/07055900.2000.9649654>.

# Index

## A

additive relation 119, 135, 151, 156, 159  
adjusted  $R^2$  308, 315  
Akaike's information criterion 316, 405  
 $\alpha$ -level 70, 73, 101, 102, 103, 107, 233, 234, 277, 332  
aligned-ranks test 97, 167, 198, 200, 201, 202  
all-subsets regression 317, 318  
alternate hypothesis 86, 100, 167, 169, 212, 215, 370, 371, 386  
analysis of covariance 93, 94, 95, 320, 323, 401  
analysis of variance 13, 42, 95, 96, 97, 141, 165, 166, 167, 171, 172, 174, 176, 177, 198, 202, 205, 320, 366, 399, 404, 409, 422  
factorial ANOVA 101, 177, 178  
one-way ANOVA 174, 178, 195, 201  
two-factor ANOVA 166, 177, 178, 179, 181, 182, 184, 192, 193, 202, 203, 204, 205  
ART 200, 201, 202, 204, 208  
attained significance level 102, 361  
autocorrelation 361

## B

bar chart 55, 413, 418, 420, 421, 422, 423  
BDM test 166, 170, 171, 185, 193, 201, 202, 203  
bias correction 13, 256, 260, 263  
bi-square weight function 288  
blocking 145, 193, 194, 195, 201, 205, 268  
Blom plotting position 9, 26, 108  
bootstrap 61, 63, 64, 65, 68, 69, 70, 78, 81, 92, 97, 134, 329, 330, 361  
boxplot 12, 14, 17, 20, 22, 24, 26, 29, 30, 31, 33, 34, 35, 36, 37, 45, 56, 58, 60, 62, 89, 95, 119, 120, 121, 130, 135, 148, 150, 154, 159, 169, 170, 178, 179, 181, 183, 184, 187, 195, 198, 201, 228, 245, 246, 273, 282, 293, 295, 371, 372, 422, 423, 429  
bulging rule 229, 230, 254

## C

categorical variable 95, 165, 246, 385, 386, 401, 407  
censored data 2, 126, 130, 138, 146, 256, 257, 356, 357  
centering 309, 310  
coefficient of determination 227  
coefficient of skewness 11  
compliance with water quality standards 75  
component + residual plots 300, 307, 320  
confidence interval 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 73, 75, 76, 77, 78, 80, 81, 88, 90, 91, 92, 98, 112, 122, 125, 129, 131, 132, 133, 134, 135, 137, 149, 150, 156, 157, 161, 162, 163, 188, 189, 200, 214, 234, 235, 236, 244, 246, 247, 248, 249, 250, 251, 252, 267, 268, 270, 277, 278, 292, 298, 299, 309, 313, 330, 377, 406, 407  
for percentiles 75, 76, 77  
for the mean 67, 68, 248, 249, 298, 299  
for the median 61, 65, 68, 91, 268

confidence level 58, 62, 67, 76, 82, 83, 84, 87, 88, 112, 187, 192, 202  
 constant variance assumptions 173, 178, 185  
 contingency table 95, 96, 386, 387, 389, 390, 393, 396, 400  
 continuity correction 78, 122, 123, 131, 138, 149, 150, 153, 218, 344  
 control chart 88, 90, 91  
 Cook's  $D$  238, 241, 242  
 correlation 5, 17, 18, 22, 30, 51, 52, 83, 93, 94, 95, 96, 97, 107, 108, 110, 114, 143, 209, 210, 211, 212, 214, 215, 216, 217, 218, 220, 221, 227, 233, 234, 236, 246, 247, 248, 255, 261, 268, 270, 275, 276, 277, 278, 279, 280, 281, 282, 284, 285, 309, 329, 330, 332, 339, 344, 349, 359, 361, 386, 393, 398  
 autocorrelation 5, 361  
 serial correlation 5, 246, 247, 248, 329, 330, 332, 344, 349, 359, 361  
 correlation coefficient 17, 22, 30, 83, 107, 108, 110, 143, 209, 210, 212, 214, 215, 218, 227, 233, 236, 246, 247, 255, 276, 279, 281, 282, 339, 393  
 cumulative distribution function 2, 3, 4, 9, 20, 26, 60, 61, 185, 203, 263, 281

**D**

degrees of freedom 60, 66, 80, 127, 134, 168, 169, 173, 174, 175, 178, 186, 192, 199, 201, 202, 204, 205, 212, 215, 216, 233, 235, 241, 245, 249, 250, 258, 262, 275, 296, 297, 298, 303, 305, 308, 311, 313, 315, 316, 321, 323, 324, 345, 348, 387, 389, 391, 406, 408, 409  
 detection limit 130, 143, 164, 206, 207, 215, 353, 357, 385, 396, 401  
 DFFITS 238, 241, 242, 302, 306  
 discriminant function analysis 409  
 dot-and-line plot 33, 36, 135  
 Dunn's multiple comparison test 392  
 Durbin-Watson statistic 247

**E**

efficiency 268, 270, 273, 274  
 equal variance 98, 118, 127, 140, 141, 173, 174, 175, 177, 181, 187, 192, 198, 208, 248, 382  
 error rate 97, 101, 103, 186, 187, 188, 190, 191, 193, 199, 200, 205, 359, 361, 392  
   family 186, 188, 190, 191, 193, 199, 200, 392  
   overall 392  
   pairwise 186, 187, 188  
 error sum of squares 173, 227, 248, 296, 313, 315, 317  
 exact test 93, 102, 103, 121, 122, 123, 132, 150, 151, 152, 153, 169, 176, 199, 216, 217, 220, 277, 389  
 expected value 11, 104, 107, 225, 333, 341, 358, 389  
 explanatory variable 93, 95, 165, 223, 224, 229, 233, 236, 237, 238, 241, 242, 243, 244, 246, 247, 248, 261, 264, 271, 273, 279, 284, 285, 286, 287, 289, 292, 295, 296, 297, 298, 299, 300, 301, 303, 306, 307, 308, 309, 310, 313, 314, 315, 316, 317, 318, 320, 322, 324, 325, 327, 331, 337, 338, 341, 345, 357, 358, 401, 402, 403, 404, 406, 407, 408, 409, 410, 411

**F**

factorial ANOVA. *See* analysis of variance  
 false discovery rate 186, 190, 191  
 Fligner-Killeen test 128, 140, 141, 175, 180, 181  
 flood frequency 9, 22, 89, 356  
 flow duration 21, 22, 75, 364  
 Friedman test 97, 167, 198, 199, 200, 201, 207  
*F*-test 138, 174, 176, 177, 178, 183, 184, 185, 201, 202, 204, 205, 296, 297, 298, 308, 309, 320, 323, 324, 341, 346, 404  
 nested *F*-test 296, 297, 320, 322, 324, 404, 405  
 partial *F*-test 297, 298, 320

**G**

geometric mean 7, 8, 15, 65, 66, 68, 69, 91, 96, 99, 129, 133, 137, 155, 156, 157, 182, 183, 184, 192, 280, 368, 369, 370, 371, 374, 376, 380, 382, 383  
 graphics 17, 18, 32, 36, 114, 283, 349, 358, 362, 413, 414, 415, 416, 417, 418, 423, 429  
 color 414, 415, 416, 417, 430  
 guidelines 245, 264, 414, 415, 427  
 length 420  
 misleading 11, 413, 423, 425  
 perspective 317, 330, 361, 414, 417, 418, 423, 424, 425  
 scale break 425, 426  
 shading 415, 416, 417

**H**

harmonic mean 186  
 Hazen plotting position 9  
 heavy tails 29, 30, 352  
 heteroscedasticity 138, 141, 173, 174, 185, 229, 230, 243, 254, 258, 263, 273, 299, 427  
 hinge 22, 24, 30  
 histograms 18, 20, 26, 30, 32, 36, 135  
 Hodges-Lehmann estimator 99, 112, 131, 132, 133, 135, 161, 163, 270  
 hypothesis tests 11, 38, 70, 74, 93, 94, 95, 96, 97, 100, 111, 112, 113, 114, 117, 135, 137, 166, 167, 212, 215, 218, 231, 233, 244, 245, 246, 247, 296, 313, 325, 356, 365, 404

**I**

influence 5, 6, 8, 9, 11, 12, 33, 39, 40, 107, 145, 154, 165, 177, 178, 179, 181, 182, 192, 237, 238, 239, 240, 241, 267, 273, 275, 285, 286, 287, 295, 300, 301, 302, 304, 306, 313, 314, 316, 320, 328, 330, 335, 336, 337, 338, 340, 342, 343, 350, 351, 354, 357, 358, 359, 361, 386, 401, 410  
 interaction 177, 178, 179, 181, 182, 184, 185, 203, 205, 310, 322, 324, 399  
 intercept 18, 135, 159, 224, 227, 231, 232, 233, 234, 235, 236, 258, 261, 262, 267, 268, 270, 275, 278, 281, 283, 284, 285, 292, 296, 298, 303, 305, 306, 308, 310, 311, 312, 314, 319, 321, 322, 323, 325, 403, 404, 406, 408, 410  
 interquartile range 2, 9, 15, 22, 36, 91, 118, 347, 422, 423  
 interval estimate 57, 58, 61, 62, 65, 67, 68, 69, 78, 79, 80, 81, 88, 91, 92, 132, 299, 365  
 invariance to rotation 283  
 IQR 9, 10, 12, 34, 118, 119, 135

**J**

joint probability 386

**K**

Kendall's  $S$  statistic 332, 343  
 Kendall's  $\tau$  97, 210, 217, 267, 275, 276, 277, 332, 349, 393, 396, 409  
 Kruskal-Wallis test 95, 97, 166, 167, 168, 169, 170, 173, 175, 186, 188, 199, 390, 391, 392

**L**

ladder of powers 14, 30, 182, 254, 333, 427  
 large-sample approximation 81, 85, 102, 120, 121, 122, 123, 132, 147, 149, 150, 151, 152, 153, 169, 199, 218, 220, 277, 356, 389, 397  
 least normal squares 278, 281, 283, 284, 285, 418  
 least significant range 186

least squares 45, 225, 227, 228, 238, 267, 274, 276, 280, 281, 284, 285, 295, 296, 309, 327, 335, 350, 351, 401  
 Levene's test 128, 141  
 leverage 237, 238, 239, 240, 241, 299, 300, 301, 302, 303, 304  
 likelihood ratio 409, 410  
 linear regression 13, 39, 95, 97, 212, 223, 224, 225, 226, 227, 228, 229, 231, 233, 234, 237, 238, 262, 285, 295, 296, 313, 320, 335, 336, 339, 340, 341, 344, 345, 350, 356  
 line of organic correlation 268, 278, 280, 281, 282, 284, 285  
 loess 40, 41, 42, 44, 45, 243, 286, 287, 288, 289, 290, 291, 292, 331, 336, 338, 339, 343, 344, 347, 352, 353, 358, 359, 360, 414  
 logistic regression 94, 95, 401, 402, 403, 404, 405, 406, 409, 410, 411  
 logistic transformation 403  
 logit 402, 403, 410  
 loglinear models 399  
 lognormal distribution 2, 3, 11, 60, 66, 67, 74, 75, 80, 81, 84, 88, 89, 90, 118, 119, 129, 138, 139, 140, 256  
 long-term persistence 359  
 LOWESS 40, 286, 288, 289, 290, 291, 292

## M

MAD 10, 289  
 major axis 280, 283  
 Mann-Kendall trend test 331, 338, 339, 347, 348, 349  
 Mann-Whitney test 97, 118  
 marginal probability 386, 389  
 matched-pair tests 145, 158, 159, 193  
 maximum-likelihood  
     Tobit regression 356, 357  
 measures of location 12  
 median difference 99, 122, 131, 132, 145, 150, 151, 159, 160, 161, 162  
 median polish 194, 195, 197, 198, 201  
 mixed effects 184  
 MLE 256, 260, 263, 356  
 mode 6, 7, 8, 362  
 monotonic trend 352, 353, 357, 364  
 MOVE 280, 282, 283  
 moving average 40, 286  
 multicollinearity 295, 308, 309, 310, 312, 313, 314, 316, 317, 325  
 multiple comparison test 166, 167, 185, 187, 199, 202, 204, 208, 392  
 multiple linear regression 39, 223, 238, 248, 295, 296, 313  
 multiplicative relation 36, 38, 135, 137, 142, 151, 154, 155, 160  
 multiply censored data 356  
 multivariate graphical methods 45

## N

nondetect 137, 138, 206, 207, 391, 393  
 non-normality 13, 74, 88, 90, 91, 111, 126, 137, 139, 140, 156, 158, 159, 161, 174, 175, 180, 182, 184, 185, 207, 267, 271, 368, 378  
 normal distribution 2, 4, 11, 21, 22, 24, 26, 27, 28, 29, 30, 58, 61, 63, 66, 67, 68, 70, 73, 74, 78, 80, 89, 91, 95, 96, 98, 99, 100, 102, 107, 108, 109, 115, 117, 122, 123, 124, 125, 128, 129, 134, 138, 139, 140, 141, 145, 146, 149, 150, 152, 153, 155, 156, 157, 158, 159, 162, 165, 166, 167, 179, 204, 206, 218, 220, 228, 244, 245, 255, 256, 260, 261, 263, 264, 268, 270,

271, 272, 277, 279, 283, 287, 343, 344, 349, 365, 366, 367, 368, 369, 370, 371, 379, 380, 382, 395, 396, 397  
 normality assumption 356, 369  
 normal probability plot 21, 22, 26, 29, 30, 31, 108, 109, 178, 245, 299  
 normal quantiles 26, 108, 132  
 normal scores 108  
 null hypothesis 86, 96, 97, 100, 101, 102, 103, 106, 107, 111, 112, 118, 124, 127, 129, 131, 140, 141, 146, 147, 148, 150, 154, 155, 156, 157, 158, 166, 168, 169, 170, 174, 175, 176, 177, 185, 186, 190, 199, 204, 205, 212, 214, 215, 218, 220, 233, 234, 242, 244, 297, 321, 322, 324, 329, 330, 331, 335, 338, 343, 344, 348, 361, 365, 369, 370, 371, 376, 386, 389, 396, 403, 404

## O

one-sided *p*-value 105, 106, 147, 149, 155, 157, 159  
 one-sided test 100, 101, 102, 103, 105, 107, 114, 118, 147, 149, 150, 157, 158, 214  
 ordinal variable 399  
 ordinary least squares 45, 225, 227, 228, 267, 274, 276, 280, 281, 284, 285, 295, 327, 335, 350, 351, 401  
 outliers 2, 6, 7, 8, 10, 11, 12, 13, 15, 22, 24, 28, 29, 30, 33, 36, 40, 45, 46, 49, 52, 53, 65, 67, 88, 89, 90, 91, 98, 99, 100, 102, 112, 130, 132, 135, 137, 146, 148, 154, 161, 169, 179, 194, 210, 212, 213, 214, 217, 228, 240, 243, 258, 267, 268, 270, 273, 283, 286, 289, 301, 330, 339, 354, 377, 382, 422  
 tests for outliers 89  
 outside values 22, 24, 252

## P

paired observations 145, 146, 150, 159  
 paired *t*-test 97, 145, 146, 155, 156, 157, 158, 161, 163, 167, 202, 205, 206  
 pairwise comparisons 186, 190, 191, 193, 200, 203, 206, 218, 268, 344, 392  
 pairwise error rate. *See* error rate  
 parametric prediction interval 73, 75, 252  
 partial-regression plots 300, 304, 306  
 Pearson's *r* 97, 210, 212, 214, 215, 216, 247  
 Pearson Type III distribution 81  
 percentile 9, 11, 12, 13, 20, 58, 64, 65, 68, 70, 74, 75, 76, 77, 78, 80, 81, 83, 84, 85, 87, 96, 98, 118, 130, 134, 169, 170, 184, 185, 263, 264, 267, 278, 281, 284  
 permutation test 30, 93, 95, 96, 97, 98, 99, 100, 103, 107, 123, 124, 125, 126, 129, 130, 134, 145, 146, 155, 157, 158, 159, 161, 165, 166, 173, 175, 176, 177, 178, 181, 182, 184, 185, 208, 382, 387, 389  
 Pettitt test 355  
*p*-hacking 112, 113  
 pie chart 55, 413, 415, 418, 420, 423, 424  
 Piper diagram 45, 48, 50  
 plotting position 9, 20, 22, 23, 24, 26, 36, 56, 64, 70, 108  
 point estimate 57, 67, 80  
 polar smooth 42, 43  
 population 1, 2, 3, 5, 11, 12, 14, 18, 22, 57, 58, 60, 61, 62, 63, 64, 67, 70, 76, 77, 79, 81, 88, 89, 93, 95, 96, 103, 117, 118, 124, 131, 138, 146, 148, 150, 172, 175, 256, 271, 283, 295, 310, 314, 345, 366, 401, 405, 406, 407, 417, 422  
 positive skewness 2, 11, 28, 29, 30, 36, 64, 74, 158, 182  
 power 11, 14, 30, 91, 92, 95, 96, 98, 99, 100, 101, 102, 111, 119, 124, 126, 128, 130, 135, 137, 141, 150, 158, 163, 166, 173, 174, 175, 178, 181, 182, 184, 185, 187, 188, 191, 192, 198, 199, 200, 201, 202, 203, 206, 210, 217, 229, 237, 244, 263, 296, 297, 313, 316, 330, 331

335, 336, 340, 342, 345, 349, 353, 356, 357, 361, 365, 366, 367, 368, 369, 370, 371, 374, 375, 376, 377, 378, 379, 380, 382, 383, 404, 406, 410  
 power transformation 119, 135, 182, 217, 229, 349, 370, 382  
 prediction interval 58, 70, 71, 72, 73, 74, 75, 89, 112, 228, 244, 245, 246, 248, 250, 251, 252, 253, 254, 298, 299  
 prediction residual 240, 241  
 PRESS statistic 248, 317  
 principal components analysis 51  
 probability density function 2, 3, 4, 18, 20, 24, 26, 272, 273, 358  
 probability plot 17, 21, 22, 26, 27, 28, 29, 30, 31, 36, 37, 56, 83, 107, 108, 109, 110, 111, 135, 143, 178, 228, 245, 254, 255, 293, 299  
 probability plot correlation coefficient 22, 30, 83, 107, 108, 110, 143, 255  
 profile plot 46  
 $p$ -value 93, 95, 96, 97, 98, 102, 103, 104, 105, 106, 107, 108, 109, 111, 112, 114, 119, 120, 122, 123, 124, 125, 126, 127, 128, 129, 131, 132, 133, 137, 138, 139, 140, 141, 147, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 162, 163, 167, 168, 169, 170, 174, 175, 176, 177, 178, 180, 181, 184, 186, 187, 190, 191, 192, 199, 200, 202, 204, 205, 206, 207, 214, 216, 217, 218, 219, 220, 233, 234, 258, 262, 270, 275, 276, 277, 298, 303, 305, 308, 311, 313, 330, 331, 332, 335, 338, 339, 341, 347, 348, 350, 354, 355, 361, 367, 368, 382, 387, 388, 389, 392, 393, 396, 397, 398, 402, 403, 404, 406, 409

## Q

Q-Q plot 14, 21, 26, 28, 36, 38, 56, 135, 136, 137, 142, 179, 180, 181  
 quality control 76, 88, 90  
 quantile plot 20, 21, 22, 24, 26, 27, 30, 56, 135, 170, 191  
 quantiles 1, 20, 22, 23, 26, 28, 30, 36, 42, 75, 78, 87, 103, 107, 108, 111, 132, 133, 135, 137, 149, 153, 162, 174, 204, 277, 366, 377  
 quartile 9, 12, 15, 22, 24, 29, 34, 36, 289, 422  
 quasi-periodic oscillation 359, 361

## R

$R^2$  108, 228, 229, 234, 236, 237, 255, 261, 263, 292, 295, 303, 305, 308, 309, 312, 313, 314, 315, 316, 317, 320, 341, 404, 405, 406, 408, 409  
 ranks 52, 61, 62, 63, 77, 78, 96, 103, 104, 119, 120, 121, 122, 123, 132, 133, 140, 151, 162, 167, 168, 169, 170, 185, 188, 194, 198, 199, 200, 201, 202, 203, 206, 207, 210, 214, 215, 216, 217, 253, 277, 390, 391, 409  
 rank-sum test 13, 93, 95, 97, 99, 100, 103, 104, 105, 106, 117, 118, 119, 120, 121, 122, 123, 129, 130, 131, 132, 135, 137, 138, 142, 143, 148, 166, 167, 188, 190, 191, 193, 203, 218, 352, 353, 370, 371, 374, 375, 377, 378, 379, 380, 381, 382, 383, 392, 409, 410  
 rating curve 254  
 record extension 268, 282, 292  
 reduced major axis 280  
 regional Kendall test 349  
 regression 13, 18, 39, 40, 42, 45, 86, 93, 94, 95, 96, 97, 159, 165, 177, 212, 223, 224, 225, 226, 227, 228, 229, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 260, 261, 262, 263, 264, 267, 268, 270, 271, 273, 275, 276, 279, 280, 282, 283, 285, 286, 287, 288, 289, 291, 292, 293, 295, 296, 297, 298, 299, 300, 301, 303, 304, 305, 306, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 320, 321, 322, 323, 325, 326, 327, 329, 331, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 349, 350, 351, 352, 353, 356, 357, 358, 361, 401, 402, 403, 404, 405, 406, 409, 410, 411, 429  
 assumptions 1, 11, 61, 67, 70, 86, 95, 96, 97, 99, 100, 111, 118, 123, 124, 126, 127, 129, 130, 135, 137, 154, 156, 159, 165, 167, 172, 173, 174, 175, 177, 178, 181, 182, 184, 185, 187, 192, 194, 208, 212, 223, 228, 229, 231, 234, 237, 238, 242, 245, 247, 248, 252, 255, 270, 273, 275, 296, 329, 330, 335, 344, 345, 356, 357, 358, 359, 409, 422

bias correction 13, 256, 260, 263  
 diagnostics 236, 237, 238, 243, 271, 299, 300, 302, 308, 312, 406  
 residual 45, 52, 140, 141, 171, 172, 173, 174, 175, 177, 178, 179, 180, 182, 183, 184, 186, 187, 194, 195, 197, 198, 201, 202, 204, 205, 206, 223, 224, 227, 228, 232, 233, 235, 236, 237, 238, 240, 241, 242, 243, 244, 245, 246, 247, 248, 251, 252, 253, 254, 255, 256, 257, 258, 260, 261, 262, 263, 267, 268, 270, 271, 273, 275, 279, 285, 286, 288, 289, 295, 296, 298, 299, 300, 303, 305, 306, 307, 308, 310, 311, 312, 313, 316, 318, 320, 331, 335, 337, 338, 339, 340, 341, 343, 344, 345, 346, 347, 350, 353, 367, 387, 388, 389, 390, 404, 406, 408, 409  
 robust 1, 12, 30, 40, 63, 75, 92, 99, 140, 161, 165, 184, 237, 238, 242, 247, 252, 257, 267, 268, 270, 273, 283, 288, 330, 332, 333, 334, 336, 346, 350, 351, 357, 409, 422  
 robust regression 238

## S

sample size 5, 6, 9, 11, 18, 20, 22, 23, 24, 58, 60, 61, 62, 63, 67, 68, 74, 77, 78, 80, 83, 84, 87, 88, 90, 91, 92, 98, 99, 101, 102, 103, 104, 105, 106, 107, 108, 111, 112, 119, 120, 121, 122, 123, 126, 130, 132, 135, 139, 146, 147, 149, 150, 151, 152, 153, 162, 165, 168, 169, 171, 173, 175, 178, 185, 186, 187, 199, 200, 215, 218, 219, 224, 246, 247, 251, 268, 271, 277, 283, 287, 288, 296, 298, 301, 309, 332, 344, 356, 365, 366, 367, 368, 369, 370, 371, 374, 375, 376, 379, 380, 382, 386, 389, 407, 409, 422  
 sample sizes 251  
 sampling design 310  
 Satterthwaite's approximation 127  
 scatterplot 231, 232, 276, 285, 286, 290, 292, 295, 301, 303, 338, 414, 422, 427  
 scatterplot matrix 50, 51, 301  
 scatterplots 17, 18, 38, 39, 40, 41, 42, 43, 44, 45, 50, 51, 52, 135, 155, 159, 160, 209  
 seasonality 92, 248, 295, 342, 343, 345, 346, 347, 352, 353, 362  
 seasonal Kendall test 268, 343, 344, 346, 348, 349, 364  
 seasonal rank-sum test 353  
 seasonal variation 295, 342, 345  
 use of periodic functions 345  
 serial correlation 5, 246, 247, 248, 329, 330, 332, 344, 349, 359, 361  
 Shapiro-Wilk test 107, 108, 109, 111, 128, 135, 156, 178, 181, 205, 367, 368, 379  
 signed-rank test 97, 145, 146, 150, 151, 152, 153, 154, 155, 156, 159, 161, 162, 163, 200, 215  
 sign test 97, 145, 146, 147, 148, 149, 150, 155, 159, 161, 198, 200, 217, 370  
 simple linear regression 223, 224, 231, 237, 295  
 skewness 2, 11, 12, 14, 20, 22, 24, 28, 29, 30, 32, 33, 34, 36, 58, 64, 65, 67, 74, 91, 98, 100, 102, 109, 130, 135, 137, 150, 158, 167, 170, 175, 179, 182, 212, 270, 273, 352, 366, 422  
 smooth 17, 38, 39, 40, 41, 42, 43, 44, 45, 244, 285, 286, 287, 288, 289, 290, 292, 338, 359, 360, 414  
 smoothing 40, 41, 44, 285, 286, 287, 289, 291, 338, 352, 358, 359  
 spatial trend 295, 328  
 Spearman's  $\rho$  97, 210, 214, 215, 217, 220  
 spread 2, 8, 10, 12, 20, 22, 32, 33, 38, 39, 42, 44, 91, 97, 111, 135, 167, 176, 278, 289  
 stacked bar chart 55, 420, 421  
 standard deviation 2, 8, 10, 11, 15, 26, 29, 33, 57, 59, 66, 67, 73, 80, 89, 90, 91, 96, 99, 102, 123, 127, 130, 134, 135, 137, 138, 140, 152, 153, 156, 159, 163, 175, 212, 213, 224, 227, 235, 236, 240, 241, 261, 271, 272, 279, 281, 344, 347, 356, 365, 366, 367, 368, 369, 371, 375, 377, 379, 383, 396  
 standard error 33, 227, 233, 234, 235, 236, 240, 254, 258, 261, 262, 263, 267, 275, 299, 303, 305, 308, 311, 313, 315, 316, 324, 347, 403  
 in percent 263  
 standardized residual 240, 242, 243, 245, 246  
 star diagram 50  
 step trend 352, 353, 354

stepwise 298, 314, 317  
 Stiff diagram 46, 48  
 studentized range 186  
 sum of squares 171, 172, 173, 174, 178, 204, 205, 227, 248, 279, 296, 313, 315, 317, 318  
 symmetry 1, 11, 12, 13, 18, 20, 33, 34, 41, 74, 98, 129, 145, 150, 151, 154, 155, 159, 167, 198,  
 201, 245, 271, 289, 371, 382

**T**

target population 1  
 $t$ -distribution 60, 66, 67, 80, 212, 214, 215, 216, 233, 235, 245, 249, 250, 298, 366  
 noncentral  $t$  80  
 Theil-Sen slope estimate 268, 270, 349  
 line 332  
 slope 267, 269, 270, 271, 274, 275, 276, 277, 278, 332, 338, 344, 348, 357  
 tolerance interval 58, 75, 76, 77  
 transformation bias 256, 257, 263, 273  
 $t$ -ratio 233, 234, 303  
 trilinear diagram 48, 49, 50  
 trimmed mean 7, 8, 12, 15  
 $t$ -statistic 156, 158, 186, 298, 314, 315, 323, 324, 335, 341  
 $t$ -test 13, 95, 97, 98, 99, 100, 103, 117, 119, 124, 125, 126, 127, 128, 129, 130, 133, 134, 135,  
 137, 138, 141, 143, 145, 146, 154, 155, 156, 157, 158, 159, 161, 163, 164, 165, 167, 170,  
 172, 173, 174, 186, 198, 202, 205, 206, 298, 309, 312, 313, 314, 321, 324, 352, 353, 354,  
 355, 365, 366, 367, 368, 370, 377, 379, 380, 382, 383, 409  
 Tukey's multiple comparison test 187  
 two-factor ANOVA 182. *See* (analysis of variance)  
 two-sided test 100, 106, 107, 118, 147, 150, 158, 169, 172, 209, 214, 218, 298, 332  
 type I error 99, 101, 102, 103, 123, 330, 356, 359, 361  
 type II error 98, 101, 102

**U**

unequal variance 126, 127, 128, 130, 138, 140, 173, 174, 175, 177, 182, 184, 368, 378

**V**

variance-covariance matrix 298

**W**

Weibull plotting position 9, 20, 26, 56, 70  
 weighted least squares 238  
 weight function 287, 288  
 Welch correction 128  
 Welch's  $t$ -test 127  
 whisker 22, 24, 30, 422  
 Wilcoxon rank-sum test 104  
 Wilcoxon signed-rank test 150  
 WRTDS 42, 45, 292, 329, 358, 359, 361

**X****Y****Z**

Manuscript was approved on December 26, 2018.

For more information about this publication, contact  
Chief, Analysis and Prediction Branch  
Integrated Modeling and Prediction Division  
Water Mission Area  
U.S. Geological Survey  
12201 Sunrise Valley Dr., Mail Stop 415  
Reston, VA 20192  
<https://water.usgs.gov/>

Publishing support provided by the USGS Science  
Publishing Network, Reston Publishing Service  
Center  
Editing by Katherine Jacques  
Layout by Cathy Y. Knutson and Jeannette Foltz  
Web support by Molly Newbrough

