

DATA ANALYSIS and STATISTICS

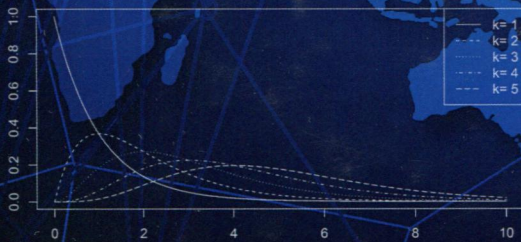
for Geography, Environmental
Science, and Engineering

MIGUEL F. ACEVEDO

$$\mu_X = b\Gamma(1+1/c) = (b/c)\Gamma(1/c)$$

$$\sigma_X^2 = b^2 \left(\Gamma(1+2/c) - (\Gamma(1+1/c))^2 \right)$$

$$\begin{bmatrix} z_{i1} \\ z_{i2} \end{bmatrix} = \begin{bmatrix} 0.33 & 0.62 & 0.74 \\ -0.84 & 0.58 & 0.13 \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{bmatrix}$$



$$E[Z(\mathbf{x}_0)] = E \left[\sum_{i=1}^k \lambda_i Z(\mathbf{x}_i) \right] = \sum_{i=1}^k \lambda_i E[Z(\mathbf{x}_i)] = \mu_Z \sum_{i=1}^k \lambda_i$$



CRC Press
Taylor & Francis Group

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

© 2013 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

International Standard Book Number: 978-1-4398-8501-7 (Hardback)

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

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

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

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

Library of Congress Cataloging-in-Publication Data

Acevedo, Miguel F.
Data analysis and statistics for geography, environmental science, and engineering / Miguel F. Acevedo.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4398-8501-7 (hardcover : alk. paper)
1. Geography--Data processing. 2. Geography--Statistical methods. 3. Environmental sciences--Data processing. 4. Environmental sciences--Statistical methods. 5. Engineering--Data processing.
6. Engineering--Statistical methods. I. Title.

G70.2.A26 2012
519.5--dc23

2012032357

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

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

Contents

Preface.....	xv
Acknowledgments.....	xix
Author	xxi

PART I Introduction to Probability, Statistics, Time Series, and Spatial Analysis

Chapter 1 Introduction.....	3
1.1 Brief History of Statistical and Probabilistic Analysis	3
1.2 Computers.....	4
1.3 Applications.....	4
1.4 Types of Variables	4
1.4.1 Discrete.....	5
1.4.2 Continuous.....	5
1.4.3 Discretization	5
1.4.4 Independent vs. Dependent Variables	6
1.5 Probability Theory and Random Variables.....	6
1.6 Methodology.....	6
1.7 Descriptive Statistics	7
1.8 Inferential Statistics.....	7
1.9 Predictors, Models, and Regression	7
1.10 Time Series.....	8
1.11 Spatial Data Analysis	8
1.12 Matrices and Multiple Dimensions	8
1.13 Other Approaches: Process-Based Models	9
1.14 Baby Steps: Calculations and Graphs.....	9
1.14.1 Mean, Variance, and Standard Deviation of a Sample	9
1.14.2 Simple Graphs as Text: Stem-and-Leaf Plots.....	10
1.14.3 Histograms	11
1.15 Exercises.....	11
1.16 Computer Session: Introduction to R	11
1.16.1 Working Directory	11
1.16.2 Installing R.....	11
1.16.3 Personalize the R GUI Shortcut	11
1.16.4 Running R.....	13
1.16.5 Basic R Skills	13
1.16.6 R Console	15
1.16.7 Scripts.....	15
1.16.8 Graphics Device	16
1.16.9 Downloading Data Files	17
1.16.10 Read a Simple Text Data File	17
1.16.11 Simple Statistics	19
1.16.12 Simple Graphs as Text: Stem-and-Leaf Plots.....	20

1.16.13	Simple Graphs to a Graphics Window	20
1.16.14	Addressing Entries of an Array	20
1.16.15	Example: Salinity	22
1.16.16	CSV Text Files	23
1.16.17	Store Your Data Files and Objects	24
1.16.18	Command History and Long Sequences of Commands	25
1.16.19	Editing Data in Objects	25
1.16.20	Cleanup and Close R Session	26
1.16.21	Computer Exercises	26
	Supplementary Reading	27
Chapter 2	Probability Theory	29
2.1	Events and Probabilities	29
2.2	Algebra of Events	29
2.3	Combinations	31
2.4	Probability Trees	32
2.5	Conditional Probability	33
2.6	Testing Water Quality: False Negative and False Positive	34
2.7	Bayes' Theorem	35
2.8	Generalization of Bayes' Rule to Many Events	36
2.9	Bio-Sensing	36
2.10	Decision Making	37
2.11	Exercises	39
2.12	Computer Session: Introduction to Rcmdr, Programming, and Multiple Plots	40
2.12.1	R Commander	40
2.12.2	Package Installation and Loading	40
2.12.3	R GUI SDI Option: Best for R Commander	43
2.12.4	How to Import a Text Data File Using Rcmdr	43
2.12.5	Simple Graphs on a Text Window	45
2.12.6	Simple Graphs on a Graphics Window: Histograms	46
2.12.7	More than One Variable: Reading Files and Plot Variables	47
2.12.7.1	Using the R Console	48
2.12.7.2	Using the R Commander	51
2.12.8	Programming Loops	53
2.12.9	Application: Bayes' Theorem	54
2.12.10	Application: Decision Making	55
2.12.11	More on Graphics Windows	55
2.12.12	Editing Data in Objects	56
2.12.13	Clean Up and Exit	56
2.12.14	Additional GUIs to Use R	57
2.12.15	Modifying the R Commander	57
2.12.16	Other Packages to Be Used in the Book	57
2.12.17	Computer Exercises	58
	Supplementary Reading	58
Chapter 3	Random Variables, Distributions, Moments, and Statistics	59
3.1	Random Variables	59
3.2	Distributions	59

3.2.1	Probability Mass and Density Functions (pmf and pdf)	59
3.2.2	Cumulative Functions (cmf and cdf)	62
3.2.3	Histograms	62
3.3	Moments	63
3.3.1	First Moment or Mean	63
3.3.2	Second Central Moment or Variance	64
3.3.3	Population and Sample	66
3.3.4	Other Statistics and Ways of Characterizing a Sample	67
3.4	Some Important RV and Distributions	68
3.5	Application Examples: Species Diversity	72
3.6	Central Limit Theorem	72
3.7	Random Number Generation	73
3.8	Exercises	74
3.9	Computer Session: Probability and Descriptive Statistics	75
3.9.1	Descriptive Statistics: Categorical Data, Table, and Pie Chart	75
3.9.2	Using a Previously Generated Object or a Dataset	78
3.9.3	Summary of Descriptive Statistics and Histogram	78
3.9.4	Density Approximation	81
3.9.5	Theoretical Distribution: Example Binomial Distribution	82
3.9.6	Application Example: Species Diversity	86
3.9.7	Random Number Generation	86
3.9.8	Comparing Sample and Theoretical Distributions: Example Binomial	89
3.9.9	Programming Application: Central Limit Theorem	90
3.9.10	Sampling: Function Sample	92
3.9.11	Cleanup and Close R Session	92
3.9.12	Computer Exercises	93
	Supplementary Reading	93
Chapter 4	Exploratory Analysis and Introduction to Inferential Statistics	95
4.1	Exploratory Data Analysis (EDA)	95
4.1.1	Index Plot	95
4.1.2	Boxplot	95
4.1.3	Empirical Cumulative Distribution Function (ecdf)	96
4.1.4	Quantile–Quantile (q–q) Plots	98
4.1.5	Combining Plots for Exploratory Data Analysis (EDA)	98
4.2	Relationships: Covariance and Correlation	98
4.2.1	Serial Data: Time Series and Autocorrelation	101
4.3	Statistical Inference	102
4.3.1	Hypothesis Testing	103
4.3.2	p -Value	105
4.3.3	Power	105
4.3.4	Confidence Intervals	107
4.4	Statistical Methods	109
4.5	Parametric Methods	110
4.5.1	Z Test or Standard Normal	110
4.5.2	The t -Test	110
4.5.3	The F Test	111
4.5.4	Correlation	112

4.6	Nonparametric Methods.....	112
4.6.1	Mann–Whitney or Wilcoxon Rank Sum Test.....	112
4.6.2	Wilcoxon Signed Rank Test.....	112
4.6.3	Spearman Correlation.....	112
4.7	Exercises.....	113
4.8	Computer Session: Exploratory Analysis and Inferential Statistics.....	113
4.8.1	Create an Example Dataset.....	113
4.8.2	Index Plot.....	113
4.8.3	Boxplot.....	114
4.8.4	Empirical Cumulative Plot.....	114
4.8.5	Functions.....	115
4.8.6	Building a Function: Example.....	115
4.8.7	More on the Standard Normal.....	116
4.8.8	Quantile–Quantile (q–q) Plots.....	118
4.8.9	Function to Plot Exploratory Data Analysis (EDA) Graphs.....	119
4.8.10	Time Series and Autocorrelation Plots.....	120
4.8.11	Additional Functions for the Rconsole and the R Commander.....	121
4.8.12	Parametric: One Sample <i>t</i> -Test or Means Test.....	122
4.8.13	Power Analysis: One Sample <i>t</i> -Test.....	124
4.8.14	Parametric: Two-Sample <i>t</i> -Test.....	126
4.8.15	Power Analysis: Two Sample <i>t</i> -Test.....	128
4.8.16	Using Data Sets from Packages.....	129
4.8.17	Nonparametric: Wilcoxon Test.....	130
4.8.18	Bivariate Data and Correlation Test.....	132
4.8.19	Computer Exercises.....	135
	Supplementary Reading.....	136
Chapter 5	More on Inferential Statistics: Goodness of Fit, Contingency Analysis, and Analysis of Variance.....	137
5.1	Goodness of Fit (GOF).....	137
5.1.1	Qualitative: Exploratory Analysis.....	137
5.1.2	χ^2 (Chi-Square) Test.....	137
5.1.3	Kolmogorov–Smirnov (K–S).....	140
5.1.4	Shapiro–Wilk Test.....	140
5.2	Counts and Proportions.....	141
5.3	Contingency Tables and Cross-Tabulation.....	141
5.4	Analysis of Variance.....	144
5.4.1	ANOVA One-Way.....	145
5.4.2	ANOVA Two-Way.....	148
5.4.3	Factor Interaction in ANOVA Two-Way.....	149
5.4.4	Nonparametric Analysis of Variance.....	150
5.5	Exercises.....	151
5.6	Computer Session: More on Inferential Statistics.....	153
5.6.1	GOF: Exploratory Analysis.....	153
5.6.2	GOF: Chi-Square Test.....	154
5.6.3	GOF: Kolmogorov–Smirnov Test.....	155
5.6.4	GOF: Shapiro–Wilk.....	156
5.6.5	Count Tests and the Binomial.....	156
5.6.6	Obtaining a Single Element of a Test Result.....	157

5.6.7	Comparing Proportions: <code>prop.test</code>	158
5.6.8	Contingency Tables: Direct Input.....	159
5.6.9	Contingency Tables: Cross-Tabulation.....	160
5.6.10	ANOVA One-Way.....	162
5.6.11	ANOVA Two-Way.....	166
5.6.12	ANOVA Nonparametric: Kruskal–Wallis.....	169
5.6.13	ANOVA Nonparametric: Friedman.....	172
5.6.14	ANOVA: Generating Fictional Data for Further Learning.....	172
5.6.15	Computer Exercises.....	175
	Supplementary Reading.....	176
Chapter 6	Regression.....	177
6.1	Simple Linear Least Squares Regression.....	177
6.1.1	Derivatives and Optimization.....	178
6.1.2	Calculating Regression Coefficients.....	180
6.1.3	Interpreting the Coefficients Using Sample Means, Variances, and Covariance.....	183
6.1.4	Regression Coefficients from Expected Values.....	184
6.1.5	Interpretation of the Error Terms.....	185
6.1.6	Evaluating Regression Models.....	188
6.1.7	Regression through the Origin.....	192
6.2	ANOVA as Predictive Tool.....	195
6.3	Nonlinear Regression.....	196
6.3.1	Log Transform.....	197
6.3.2	Nonlinear Optimization.....	197
6.3.3	Polynomial Regression.....	198
6.3.4	Predicted vs. Observed Plots.....	198
6.4	Computer Session: Simple Regression.....	200
6.4.1	Scatter Plots.....	200
6.4.2	Simple Linear Regression.....	202
6.4.3	Nonintercept Model or Regression through the Origin.....	206
6.4.4	ANOVA One Way: As Linear Model.....	208
6.4.5	Linear Regression: Lack-of-Fit to Nonlinear Data.....	211
6.4.6	Nonlinear Regression by Transformation.....	214
6.4.7	Nonlinear Regression by Optimization.....	216
6.4.8	Polynomial Regression.....	219
6.4.9	Predicted vs. Observed Plots.....	221
6.4.10	Computer Exercises.....	221
	Supplementary Reading.....	223
Chapter 7	Stochastic or Random Processes and Time Series.....	225
7.1	Stochastic Processes and Time Series: Basics.....	225
7.2	Gaussian.....	225
7.3	Autocovariance and Autocorrelation.....	227
7.4	Periodic Series, Filtering, and Spectral Analysis.....	231
7.5	Poisson Process.....	238
7.6	Marked Poisson Process.....	241

7.7	Simulation.....	247
7.8	Exercises.....	249
7.9	Computer Session: Random Processes and Time Series.....	250
7.9.1	Gaussian Random Processes.....	250
7.9.2	Autocorrelation.....	252
7.9.3	Periodic Process.....	252
7.9.4	Filtering and Spectrum.....	253
7.9.5	Sunspots Example.....	254
7.9.6	Poisson Process.....	255
7.9.7	Poisson Process Simulation.....	255
7.9.8	Marked Poisson Process Simulation: Rainfall.....	256
7.9.9	Computer Exercises.....	257
	Supplementary Reading.....	258
Chapter 8	Spatial Point Patterns.....	259
8.1	Types of Spatially Explicit Data.....	259
8.2	Types of Spatial Point Patterns.....	259
8.3	Spatial Distribution.....	259
8.4	Testing Spatial Patterns: Cell Count Methods.....	260
8.4.1	Testing Uniform Patterns.....	260
8.4.2	Testing for Spatial Randomness.....	261
8.4.3	Clustered Patterns.....	263
8.5	Nearest-Neighbor Analysis.....	264
8.5.1	First-Order Analysis.....	264
8.5.2	Second-Order Analysis.....	266
8.6	Marked Point Patterns.....	268
8.7	Geostatistics: Regionalized Variables.....	269
8.8	Variograms: Covariance and Semivariance.....	270
8.8.1	Covariance.....	271
8.8.2	Semivariance.....	272
8.9	Directions.....	274
8.10	Variogram Models.....	276
8.10.1	Exponential Model.....	276
8.10.2	Spherical Model.....	278
8.10.3	Gaussian Model.....	278
8.10.4	Linear and Power Models.....	279
8.10.5	Modeling the Empirical Variogram.....	280
8.11	Exercises.....	281
8.12	Computer Session: Spatial Analysis.....	284
8.12.1	Packages and Functions.....	284
8.12.2	File Format.....	284
8.12.3	Creating a Pattern: Location-Only.....	285
8.12.4	Generating Patterns with Random Numbers.....	286
8.12.5	Grid or Quadrat Analysis: Chi-Square Test for Uniformity.....	288
8.12.6	Grid or Quadrat Analysis: Randomness, Poisson Model.....	289
8.12.7	Nearest-Neighbor Analysis: <i>G</i> and <i>K</i> Functions.....	290
8.12.8	Monte Carlo: Nearest-Neighbor Analysis of Uniformity.....	293
8.12.9	Marked Spatial Patterns: Categorical Marks.....	294
8.12.10	Marked Spatial Patterns: Continuous Values.....	298

8.12.11 Marked Patterns: Use Sample Data from sgeostat 301
 8.12.12 Computer Exercises 305
 Supplementary Reading 306

PART II Matrices, Temporal and Spatial Autoregressive Processes, and Multivariate Analysis

Chapter 9 Matrices and Linear Algebra 309

9.1 Matrices 309
 9.2 Dimension of a Matrix 309
 9.3 Vectors 310
 9.4 Square Matrices 310
 9.4.1 Trace 311
 9.4.2 Symmetric Matrices: Covariance Matrix 311
 9.4.3 Identity 312
 9.5 Matrix Operations 312
 9.5.1 Addition and Subtraction 312
 9.5.2 Scalar Multiplication 313
 9.5.3 Linear Combination 313
 9.5.4 Matrix Multiplication 313
 9.5.5 Determinant of a Matrix 315
 9.5.6 Matrix Transposition 316
 9.5.7 Major Product 316
 9.5.8 Matrix Inversion 317
 9.6 Solving Systems of Linear Equations 319
 9.7 Linear Algebra Solution of the Regression Problem 321
 9.8 Alternative Matrix Approach to Linear Regression 323
 9.9 Exercises 325
 9.10 Computer Session: Matrices and Linear Algebra 326
 9.10.1 Creating Matrices 326
 9.10.2 Operations 327
 9.10.3 Other Operations 330
 9.10.4 Solving System of Linear Equations 331
 9.10.5 Inverse 331
 9.10.6 Computer Exercises 332
 Supplementary Reading 332

Chapter 10 Multivariate Models 333

10.1 Multiple Linear Regression 333
 10.1.1 Matrix Approach 333
 10.1.2 Population Concepts and Expected Values 338
 10.1.3 Evaluation and Diagnostics 339
 10.1.4 Variable Selection 340
 10.2 Multivariate Regression 342
 10.3 Two-Group Discriminant Analysis 344
 10.4 Multiple Analysis of Variance (MANOVA) 349
 10.5 Exercises 353

10.6	Computer Session: Multivariate Models	355
10.6.1	Multiple Linear Regression	355
10.6.2	Multivariate Regression	359
10.6.3	Two-Group Linear Discriminant Analysis	361
10.6.4	MANOVA	363
10.6.5	Computer Exercises.....	365
10.6.6	Functions	365
	Supplementary Reading	367
Chapter 11	Dependent Stochastic Processes and Time Series	369
11.1	Markov.....	369
11.1.1	Dependent Models: Markov Chain	369
11.1.2	Two-Step Rainfall Generation: First Step Markov Sequence	371
11.1.3	Combining Dry/Wet Days with Amount on Wet Days	371
11.1.4	Forest Succession	374
11.2	Semi-Markov Processes	378
11.3	Autoregressive (AR) Process	381
11.4	ARMA and ARIMA Models	387
11.5	Exercises.....	389
11.6	Computer Session: Markov Processes and Autoregressive Time Series	389
11.6.1	Weather Generation: Rainfall Models.....	389
11.6.2	Semi-Markov.....	391
11.6.3	AR(p) Modeling and Forecast.....	392
11.6.4	ARIMA(p, d, q) Modeling and Forecast.....	395
11.6.5	Computer Exercises.....	398
11.6.6	SEEG Functions	400
	Supplementary Reading	403
Chapter 12	Geostatistics: Kriging.....	405
12.1	Kriging	405
12.2	Ordinary Kriging.....	405
12.3	Universal Kriging.....	413
12.4	Data Transformations	414
12.5	Exercises.....	414
12.6	Computer Session: Geostatistics, Kriging.....	415
12.6.1	Ordinary Kriging	415
12.6.2	Universal Kriging.....	417
12.6.3	Regular Grid Data Files	422
12.6.4	Functions	425
12.6.5	Computer Exercises.....	428
	Supplementary Reading	428
Chapter 13	Spatial Auto-Correlation and Auto-Regression	429
13.1	Lattice Data: Spatial Auto-Correlation and Auto-Regression.....	429
13.2	Spatial Structure and Variance Inflation	429
13.3	Neighborhood Structure	429
13.4	Spatial Auto-Correlation	432

13.4.1	Moran's I	432
13.4.2	Transformations.....	433
13.4.3	Geary's c	434
13.5	Spatial Auto-Regression.....	434
13.6	Exercises.....	436
13.7	Computer Session: Spatial Correlation and Regression.....	437
13.7.1	Packages.....	437
13.7.2	Mapping Regions.....	438
13.7.3	Neighborhood Structure.....	440
13.7.4	Structure Using Distance.....	441
13.7.5	Structure Based on Borders.....	445
13.7.6	Spatial Auto-Correlation.....	446
13.7.7	Spatial Auto-Regression Models.....	448
13.7.8	Neighborhood Structure Using Tripack.....	451
13.7.9	Neighborhood Structure for Grid Data.....	452
13.7.10	Computer Exercises.....	453
	Supplementary Reading.....	454
Chapter 14	Multivariate Analysis I: Reducing Dimensionality.....	455
14.1	Multivariate Analysis: Eigen-Decomposition.....	455
14.2	Vectors and Linear Transformation.....	455
14.3	Eigenvalues and Eigenvectors.....	455
14.3.1	Finding Eigenvalues.....	457
14.3.2	Finding Eigenvectors.....	458
14.4	Eigen-Decomposition of a Covariance Matrix.....	459
14.4.1	Covariance Matrix.....	459
14.4.2	Bivariate Case.....	461
14.5	Principal Components Analysis (PCA).....	465
14.6	Singular Value Decomposition and Biplots.....	469
14.7	Factor Analysis.....	472
14.8	Correspondence Analysis.....	475
14.9	Exercises.....	479
14.10	Computer Session: Multivariate Analysis, PCA.....	480
14.10.1	Eigenvalues and Eigenvectors of Covariance Matrices.....	480
14.10.2	PCA: A Simple 2×2 Example Using Eigenvalues and Eigenvectors.....	481
14.10.3	PCA: A 2×2 Example.....	483
14.10.4	PCA Higher-Dimensional Example.....	485
14.10.5	PCA Using the Rcmdr.....	486
14.10.6	Factor Analysis.....	490
14.10.7	Factor Analysis Using Rcmdr.....	493
14.10.8	Correspondence Analysis.....	495
14.10.9	Computer Exercises.....	499
	Supplementary Reading.....	500
Chapter 15	Multivariate Analysis II: Identifying and Developing Relationships among Observations and Variables.....	501
15.1	Introduction.....	501
15.2	Multigroup Discriminant Analysis (MDA).....	501
15.3	Canonical Correlation.....	502

15.4	Constrained (or Canonical) Correspondence Analysis (CCA).....	505
15.5	Cluster Analysis.....	506
15.6	Multidimensional Scaling (MDS)	508
15.7	Exercises	509
15.8	Computer Session: Multivariate Analysis II	509
15.8.1	Multigroup Linear Discriminant Analysis.....	509
15.8.2	Canonical Correlation	514
15.8.3	Canonical Correspondence Analysis	515
15.8.4	Cluster Analysis	516
15.8.5	Multidimensional Scaling (MDS).....	518
15.8.6	Computer Exercises.....	520
	Supplementary Reading	520
	Bibliography	521
	Index.....	525