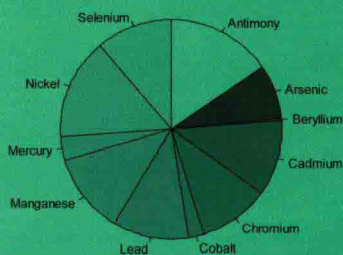


JOSEPH OFUNGWU

Statistical Applications for Environmental Analysis and Risk Assessment

STATISTICS IN PRACTICE



WILEY

STATISTICAL APPLICATIONS FOR ENVIRONMENTAL ANALYSIS AND RISK ASSESSMENT

JOSEPH OFUNGWU
Hackettstown, NJ



WILEY

Copyright © 2014 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Ofungwu, Joseph.

Statistical applications for environmental analysis and risk assessment /
Joseph Ofungwu. – First edition.

pages cm – (Statistics in practice)

Includes bibliographical references and index.

ISBN 978-1-118-63453-0 (hardback)

1. Environmental risk assessment—Statistical methods. I. Title.

GE145.O38 2014

363.7'02—dc23

2013047836

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

STATISTICAL
APPLICATIONS
FOR ENVIRONMENTAL
ANALYSIS AND RISK
ASSESSMENT

WILEY SERIES IN STATISTICS IN PRACTICE

Advisory Editor, MARIAN SCOTT, *University of Glasgow, Scotland, UK*

Founding Editor, VIC BARNETT, *Nottingham Trent University, UK*

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods, and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceuticals; industry, finance, and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

To Divine Providence

PREFACE

Although the subject of environmental statistics has been around for decades, the average environmental professional is far from comfortable with statistics, in my experience. This should be concerning because the protection of public health and ecological well-being falls largely on these professionals, and statistics should be a prominent part of their arsenal. It is fair to say that the environmental profession revolves around data. The environmental engineer, geologist, or scientist routinely collects data from soil, sediment, water, ambient air or other environmental media, for analysis and interpretation to determine the potential presence and concentrations of environmental contaminants. Based on the results of the data analysis and professional judgment, the practitioner makes a recommendation to the appropriate environmental protection authority that remedial action is necessary to reduce contaminant concentrations and minimize exposure risks, or that the exposure risks are minimal, warranting no further action. More often than not, the environmental regulatory authority concurs.

As it turns out, environmental data samples are rarely well-behaved, with nondetects, outliers, skewness, sustained and/or cyclical trend as habitual offenders in many data samples. Without functional familiarity with at least the basic statistical analysis principles and practices, how can we make sense of data such as these? Even more worrying, risk management decisions are often legalistic, based on numbers produced by statistical and associated analyses, where even a single mishandled outlier could possibly result in serious consequences for public and environmental health.

One reason for the lukewarm attitude toward statistics is the lack of regular access to competent software, as manual computation of most statistical procedures is now considered “ancient.” An insufficient statistics or math background is another reason. Not surprisingly, cost is mainly to blame for the lack of software access. The freely available, high-quality software system, R, along with others such as ProUCL and VSP used in this book, has come to the rescue in this regard. Although these systems have been in existence for over a decade, many in the environmental profession have still not heard the “good news.” The allure of a zero-cost, high-powered software package should be irresistible. No excuse to wait any longer!

ACKNOWLEDGMENTS

I owe a debt of gratitude to Steve Quigley, the Associate Publisher at Wiley, for his enthusiastic support and encouragement through the numerous twists and turns that finally brought this book to closure. Many thanks to Sari Friedman at Wiley, unfailingly courteous and professional. The book production was superbly orchestrated, thanks to Danielle LaCourciere and Faraz Sharique Ali for their expertise. I also wish to express my sincere appreciation to the many reviewers of the book proposal whose comments and criticisms alike helped shape the final product.

CONTENTS

PREFACE	xvii
ACKNOWLEDGMENTS	xix
1 INTRODUCTION	1
1.1 Introduction and Overview / 1	
1.2 The Aim of the Book: Get Involved! / 2	
1.3 The Approach and Style: Clarity, Clarity, Clarity / 3	
PART I BASIC STATISTICAL MEASURES AND CONCEPTS	5
2 INTRODUCTION TO SOFTWARE PACKAGES USED IN THIS BOOK	7
2.1 R / 8	
2.1.1 Helpful R Tips / 9	
2.1.2 Disadvantages of R / 10	
2.2 ProUCL / 10	
2.2.1 Helpful ProUCL Tips / 11	
2.2.2 Potential Deficiencies of ProUCL / 12	
2.3 Visual Sample Plan / 12	
2.4 DATAPLOT / 13	
2.4.1 Helpful Tips for Running DATAPLOT in Batch Mode / 13	
2.5 Kendall–Thiel Robust Line / 14	
2.6 Minitab® / 14	
2.7 Microsoft Excel / 15	

3	LABORATORY DETECTION LIMITS, NONDETECTS, AND DATA ANALYSIS	17
3.1	Introduction and Overview / 17	
3.2	Types of Laboratory Data Detection Limits / 18	
3.3	Problems with Nondetects in Statistical Data Samples / 19	
3.4	Options for Addressing Nondetects in Data Analysis / 20	
3.4.1	Kaplan–Meier Estimation / 21	
3.4.2	Robust Regression on Order Statistics / 22	
3.4.3	Maximum Likelihood Estimation / 23	
4	DATA SAMPLE, DATA POPULATION, AND DATA DISTRIBUTION	25
4.1	Introduction and Overview / 25	
4.2	Data Sample Versus Data Population or Universe / 26	
4.3	The Concept of a Distribution / 27	
4.3.1	The Concept of a Probability Distribution Function / 28	
4.3.2	Cumulative Probability Distribution and Empirical Cumulative Distribution Functions / 31	
4.4	Types of Distributions / 34	
4.4.1	Normal Distribution / 34	
4.4.1.1	<i>Goodness-of-Fit (GOF) Tests for the Normal Distribution / 40</i>	
4.4.1.2	<i>Central Limit Theorem / 48</i>	
4.4.2	Lognormal, Gamma, and Other Continuous Distributions / 49	
4.4.2.1	<i>Gamma Distribution / 51</i>	
4.4.2.2	<i>Logistic Distribution / 51</i>	
4.4.2.3	<i>Other Continuous Distributions / 52</i>	
4.4.3	Distributions Used in Inferential Statistics (Student's <i>t</i> , Chi-Square, <i>F</i>) / 53	
4.4.3.1	<i>Student's t Distribution / 53</i>	
4.4.3.2	<i>Chi-Square Distribution / 55</i>	
4.4.3.3	<i>F Distribution / 57</i>	
4.4.4	Discrete Distributions / 57	
4.4.4.1	<i>Binomial Distribution / 57</i>	
4.4.4.2	<i>Poisson Distribution / 61</i>	
	Exercises / 64	
5	GRAPHICS FOR DATA ANALYSIS AND PRESENTATION	67
5.1	Introduction and Overview / 67	
5.2	Graphics for Single Univariate Data Samples / 68	
5.2.1	Box and Whiskers Plot / 68	
5.2.2	Probability Plots (i.e., Quantile–Quantile Plots for Comparing a Data Sample to a Theoretical Distribution) / 72	

5.2.3	Quantile Plots / 79	
5.2.4	Histograms and Kernel Density Plots / 82	
5.3	Graphics for Two or More Univariate Data Samples / 86	
5.3.1	Quantile–Quantile Plots for Comparing Two Univariate Data Samples / 86	
5.3.2	Side-by-Side Box Plots / 89	
5.4	Graphics for Bivariate and Multivariate Data Samples / 91	
5.4.1	Graphical Data Analysis for Bivariate Data Samples / 91	
5.4.2	Graphical Data Analysis for Multivariate Data Samples / 95	
5.5	Graphics for Data Presentation / 98	
5.6	Data Smoothing / 105	
5.6.1	Moving Average and Moving Median Smoothing / 105	
5.6.2	Locally Weighted Scatterplot Smoothing (LOWESS or LOESS) / 108	
5.6.2.1	<i>Smoothness Factor and the Degree of the Local Regression / 109</i>	
5.6.2.2	<i>Basic and Robust LOWESS Weighting Functions / 109</i>	
5.6.2.3	<i>LOESS Scatterplot Smoothing for Data with Multiple Variables / 112</i>	
	Exercises / 113	

6 BASIC STATISTICAL MEASURES: DESCRIPTIVE OR SUMMARY STATISTICS

115

6.1	Introduction and Overview / 115	
6.2	Arithmetic Mean and Weighted Mean / 116	
6.3	Median and Other Robust Measures of Central Tendency / 117	
6.4	Standard Deviation, Variance, and Other Measures of Dispersion or Spread / 119	
6.4.1	Quantiles (Including Percentiles) / 121	
6.4.2	Robust Measures of Spread: Interquartile Range and Median Absolute Deviation / 124	
6.5	Skewness and Other Measures of Shape / 124	
6.6	Outliers / 134	
6.6.1	Tests for Outliers / 135	
6.7	Data Transformations / 139	
	Exercises / 141	

PART II STATISTICAL PROCEDURES FOR MOSTLY UNIVARIATE DATA

143

7 STATISTICAL INTERVALS: CONFIDENCE, TOLERANCE, AND PREDICTION INTERVALS

145

7.1	Introduction and Overview / 145	
-----	---------------------------------	--

- 7.2 Confidence Intervals / 146
 - 7.2.1 Parametric Confidence Intervals / 151
 - 7.2.1.1 *Parametric Confidence Interval around the Aritihmetic Mean or Median for Normally Distributed Data* / 151
 - 7.2.1.2 *Lognormal and Other Parametric Confidence Intervals* / 153
 - 7.2.2 Nonparametric Confidence Intervals Around the Mean, Median, and Other Percentiles / 154
 - 7.2.3 Parametric Confidence Band Around a Trend Line / 164
 - 7.2.4 Nonparametric Confidence Band Around a Trend Line / 166
- 7.3 Tolerance Intervals / 168
 - 7.3.1 Parametric Tolerance Intervals / 169
 - 7.3.2 Nonparametric Tolerance Intervals / 170
- 7.4 Prediction Intervals / 173
 - 7.4.1 Parametric Prediction Intervals for Future Individual Values and Future Means / 175
 - 7.4.2 Nonparametric Prediction Intervals for Future Individual Values and Future Medians / 176
- 7.5 Control Charts / 178
 - Exercises / 178

8 TESTS OF HYPOTHESIS AND DECISION MAKING **181**

- 8.1 Introduction and Overview / 181
- 8.2 Basic Terminology and Procedures for Tests of Hypothesis / 182
- 8.3 Type I and Type II Decision Errors, Statistical Power, and Interrelationships / 190
- 8.4 The Problem with Multiple Tests or Comparisons: Site-Wide False Positive Error Rates / 193
- 8.5 Tests for Equality of Variance / 195
 - Exercises / 199

9 APPLICATIONS OF HYPOTHESIS TESTS: COMPARING POPULATIONS, ANALYSIS OF VARIANCE **201**

- 9.1 Introduction and Overview / 201
- 9.2 Single Sample Tests / 202
 - 9.2.1 Parametric Single-Sample Tests: One-Sample *t*-Test and One-Sample Proportion Test / 203
 - 9.2.2 Nonparametric Single-Sample Tests: One-Sample Sign Test and One-Sample Wilcoxon Signed Rank Test / 205
 - 9.2.2.1 *Nonparametric One-Sample Sign Test* / 206
 - 9.2.2.2 *Nonparametric One-Sample Wilcoxon Signed Rank Test* / 208

9.3	Two-Sample Tests / 208	
9.3.1	Parametric Two-Sample Tests / 210	
9.3.1.1	<i>Parametric Two-Sample t-Test for Independent Populations / 210</i>	
9.3.1.2	<i>Parametric Two-Sample t-Test for Paired Populations / 214</i>	
9.3.2	Nonparametric Two-Sample Tests / 216	
9.3.2.1	<i>Nonparametric Wilcoxon Rank Sum Test for Two Independent Populations / 216</i>	
9.3.2.2	<i>Nonparametric Gehan Test for Two Independent Populations / 220</i>	
9.3.2.3	<i>Nonparametric Quantile Test for Two Independent Populations / 221</i>	
9.3.2.4	<i>Nonparametric Two-Sample Paired Sign Test and Paired Wilcoxon Signed Rank Test / 222</i>	
9.4	Comparing Three or More Populations: Parametric ANOVA and Nonparametric Kruskal–Wallis Tests / 227	
9.4.1	Parametric One-Way ANOVA / 228	
9.4.1.1	<i>Computation of Parametric One-Way ANOVA / 230</i>	
9.4.2	Nonparametric One-Way ANOVA (Kruskal–Wallis Test) / 235	
9.4.3	Follow-Up or Post Hoc Comparisons After Parametric and Nonparametric One-Way ANOVA / 238	
9.4.4	Parametric and Nonparametric Two-Way and Multifactor ANOVA / 244	
	Exercises / 255	
10	TRENDS, AUTOCORRELATION, AND TEMPORAL DEPENDENCE	257
10.1	Introduction and Overview / 257	
10.2	Tests for Autocorrelation and Temporal Effects / 258	
10.2.1	Test for Autocorrelation Using the Sample Autocorrelation Function / 259	
10.2.2	Test for Autocorrelation Using the Rank Von Neumann Ratio Method / 261	
10.2.3	An Example on Site-Wide Temporal Effects / 264	
10.3	Tests for Trend / 265	
10.3.1	Parametric Test for Trends—Simple Linear Regression / 266	
10.3.2	Nonparametric Test for Trends—Mann–Kendall Test and Seasonal Mann–Kendall Test / 271	
10.3.3	Nonparametric Test for Trends—Theil–Sen Trend Test / 273	
10.4	Correcting Seasonality and Temporal Effects in the Data / 279	
10.4.1	Correcting Seasonality for a Single Data Series / 280	
10.4.2	Simultaneously Correcting Temporal Dependence for Multiple Data Sets / 281	
10.5	Effects of Exogenous Variables on Trend Tests / 282	
	Exercises / 285	

PART III STATISTICAL PROCEDURES FOR MOSTLY MULTIVARIATE DATA	287
11 CORRELATION, COVARIANCE, GEOSTATISTICS	289
11.1 Introduction and Overview / 289	
11.2 Correlation and Covariance / 290	
11.2.1 Pearson's Correlation Coefficient / 292	
11.2.2 Spearman's and Kendall's Correlation Coefficients / 294	
11.3 Introduction to Geostatistics / 300	
11.3.1 The Variogram or Covariogram / 300	
11.3.2 Kriging / 302	
11.3.3 A Note on Data Sample Size and Lag Distance Requirements / 311	
Exercises / 312	
12 SIMPLE LINEAR REGRESSION	315
12.1 Introduction and Overview / 315	
12.2 The Simple Linear Regression Model / 316	
12.2.1 The True or Population X - Y Relationship / 317	
12.2.2 The Estimated X - Y Relationship Based on a Data Sample / 320	
12.3 Basic Applications of Simple Linear Regression / 324	
12.3.1 Description and Graphical Review of the Data Sample for Regression / 324	
12.3.1.1 <i>Computing the Regression</i> / 325	
12.3.1.2 <i>Interpreting the Regression Results</i> / 326	
12.4 Verify Compliance with the Assumptions of Conventional Linear Regression / 332	
12.4.1 Assumptions of Linearity and Homoscedasticity / 332	
12.4.2 Assumption of Independence / 334	
12.4.3 Exogeneity Assumption, Normality of the Y Errors, and Absence of Outliers / 337	
12.5 Check the Regression Diagnostics for the Presence of Influential Data Points / 339	
12.6 Confidence Intervals for the Predicted Y Values / 343	
12.7 Regression for Left-Censored Data (Non-detects) / 344	
Exercises / 349	
13 DATA TRANSFORMATION VERSUS GENERALIZED LINEAR MODEL	351
13.1 Introduction and Overview / 351	
13.2 Data Transformation / 352	
13.2.1 General Approach for Data Transformations / 355	

- 13.2.2 The Ladder of Powers / 357
- 13.2.3 The Bulging Rule and Data Transformations for Regression Analysis / 359
- 13.2.4 Facilitating Data Transformations Using Box–Cox Methods / 366
- 13.2.5 Back-Transformation Bias and Other Issues with Data Transformation / 367
 - 13.2.5.1 *Logarithmic Transformations* / 369
 - 13.2.5.2 *Other Transformations* / 370
- 13.2.6 Transformation Bias Correction / 371
- 13.3 The Generalized Linear Model (GLM) and Applications for Regression / 374
 - 13.3.1 Components of the Generalized Linear Model and Inherent Limitations / 374
 - 13.3.2 Estimation and Hypothesis Tests of Significance for GLM Parameters / 376
 - 13.3.3 Deviance, Null Deviance, Residual Deviance, and Goodness of Fit / 377
 - 13.3.4 Diagnostics for GLM / 379
 - 13.3.5 Procedural Steps for Regression with GLM in R / 380
- 13.4 Extension of Data Transformation and Generalized Linear Model to Multiple Regression / 385
 - 13.4.1 Data Transformation for Multiple Regression / 385
 - 13.4.2 Generalized Linear Models for Multiple Regression / 387
- Exercises / 387

14 ROBUST REGRESSION

391

- 14.1 Introduction and Overview / 391
- 14.2 Kendall–Theil Robust Line / 393
 - 14.2.1 Computation of the Kendall–Theil Robust Line Regression / 393
 - 14.2.2 Test of Significance for the Kendall–Theil Robust Line / 396
 - 14.2.3 Bias Correction for Y Predictions by the Kendall–Theil Robust Line / 397
- 14.3 Weighted Least Squares Regression / 398
 - 14.3.1 Procedure for Weighted Least Squares Regression for Known Variances of the Observations / 399
- 14.4 Iteratively Reweighted Least Squares Regression / 405
 - 14.4.1 The Iteratively Reweighted Least Squares Procedure / 409
- 14.5 Other Robust Regression Alternatives: Bounded Influence Methods / 412
 - 14.5.1 Least Absolute Deviation or Least Absolute Values / 412
 - 14.5.2 Quantile Regression / 413
 - 14.5.3 Least Median of Squares / 413
 - 14.5.4 Least Trimmed Squares / 414

- 14.6 Robust Regression Methods for Multiple-Variable Data / 416
Exercises / 417

15 MULTIPLE LINEAR REGRESSION

419

- 15.1 Introduction and Overview / 419
- 15.2 The Need for Multiple Regression / 420
- 15.3 The Multiple Linear Regression (MLR) Model / 421
- 15.4 The Estimated Multivariable X - Y Relationship Based on a Data Sample / 422
- 15.5 Assumptions of Multiple Linear Regression / 430
- 15.5.1 Linearity of the Relationship Between the Dependent and Explanatory Variables / 431
- 15.5.2 Absence of Multicollinearity Among the Explanatory Variables / 433
15.5.2.1 *Potential Remedies for Multicollinearity* / 436
- 15.5.3 Homoscedasticity or Constancy of Variance of the Y Population Errors / 439
- 15.5.4 Statistical Independence of the Y Population Errors / 441
- 15.5.5 Exogeneity Assumption, Normality of the Y Errors, and Absence of Outliers / 445
- 15.5.6 Absence of Variability or Errors in the Explanatory Variables / 446
- 15.6 Hypothesis Tests for Reliability of the MLR Model / 447
- 15.6.1 ANOVA F Test for Overall Significance of the Regression / 447
15.6.1.1 *A Note on ANOVA Tables* / 448
- 15.6.2 Partial t and Partial F Tests for Individual Regression Coefficients / 452
- 15.6.3 Complete and Reduced Models / 452
- 15.7 Confidence Intervals for the Regression Coefficients and Predicted Y Values / 457
- 15.8 Coefficient of Multiple Correlation (R), Multiple Determination (R^2), Adjusted R^2 , and Partial Correlation Coefficients / 458
- 15.8.1 Coefficient of Multiple Correlation (R) / 458
- 15.8.2 Coefficient of Multiple Determination (R^2) and Adjusted R^2 / 459
- 15.8.3 Partial Correlations and Squared Partial Correlations / 460
- 15.9 Regression Diagnostics / 462
- 15.10 Model Interactions and Multiplicative Effects / 467
- 15.10.1 The Multiple Linear Regression Interaction Model / 467
- 15.10.2 Hypothesis Tests of the Interaction Terms for Significance / 468
Exercises / 474

16 CATEGORICAL DATA ANALYSIS

477

- 16.1 Introduction and Overview / 477