
Basic Biostatistics

Statistics for Public Health Practice

B. Burt Gerstman

Professor

Department of Health Science

San Jose State University

San Jose, California



JONES AND BARTLETT PUBLISHERS

Sudbury, Massachusetts

BOSTON TORONTO LONDON SINGAPORE

World Headquarters

Jones and Bartlett Publishers
40 Tall Pine Drive
Sudbury, MA 01776
978-443-5000
info@jbpub.com
www.jbpub.com

Jones and Bartlett Publishers
International
Barb House, Barb Mews
London W6 7PA
United Kingdom

Jones and Bartlett Publishers
Canada
6339 Ormindale Way
Mississauga, Ontario L5V 1J2
Canada

Jones and Bartlett's books and products are available through most bookstores and online booksellers. To contact Jones and Bartlett Publishers directly, call 800-832-0034, fax 978-443-8000, or visit our website www.jbpub.com.

Substantial discounts on bulk quantities of Jones and Bartlett's publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones and Bartlett via the above contact information or send an email to specialsales@jbpub.com.

Copyright © 2008 by Jones and Bartlett Publishers, Inc.

All rights reserved. No part of the material protected by this copyright may be reproduced or utilized in any form, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the copyright owner.

This publication is designed to provide accurate and authoritative information in regard to the Subject Matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the service of a competent professional person should be sought.

Production Credits

Publisher: Michael Brown
Associate Editor: Katey Birtcher
Production Director: Amy Rose
Production Editor: Tracey Chapman
Associate Production Editor: Rachel Rossi
Marketing Manager: Sophie Fleck
Manufacturing Buyer: Therese Connell

Composition: Graphic World, Inc.
Cover Design: Kristin E. Ohlin
Cover Image: © Sebastian Kaulitzki/
Shutterstock, Inc.; © Li Wa/Shutterstock, Inc.
Printing and Binding: Malloy, Inc
Cover Printing: Malloy, Inc

Library of Congress Cataloging-in-Publication Data

Gerstman, B. Burt.

Basic biostatistics : statistics for public health practice / B. Burt Gerstman.

p. ; cm.

Includes index.

ISBN-13: 978-0-7637-3580-7 (alk. paper)

ISBN-10: 0-7637-3580-9 (alk. paper)

1. Medical statistics. 2. Biometry. 3. Public health—Statistical methods. I. Title.

[DNLM: 1. Biometry—methods. 2. Public Health Practice. WA 950 G383b 2008]

RA409.G47 2008

362.1072'7—dc22

2007003334

6048

Printed in the United States of America

11 10 09 08 07 10 9 8 7 6 5 4 3 2 1

To my mother, Bernadine, and in memory of my father, Joseph.

Table of Contents

	Preface	xi
	Acknowledgments	xv
	About the Author	xvii
Part I	<i>General Concept and Techniques</i>	
Chapter 1	Measurement	1
	1.1 What Is Biostatistics?	1
	1.2 Organization of Data	2
	1.3 Types of Measurements	5
	1.4 Data Quality	7
Chapter 2	Types of Studies	15
	2.1 Surveys	15
	2.2 Comparative Studies	21
Chapter 3	Frequency Distributions	35
	3.1 Stemplots	35
	3.2 Frequency Tables	51
	3.3 Additional Frequency Charts	55
Chapter 4	Summary Statistics	63
	4.1 Central Location: Mean	63
	4.2 Central Location: Median	67
	4.3 Central Location: Mode	70
	4.4 Comparison of the Mean, Median, and Mode	70
	4.5 Spread: Quartiles	71

	4.6 Boxplots.....	75
	4.7 Spread: Variance and Standard Deviation	78
	4.8 Selecting Summary Statistics	84
Chapter 5	Probability Concepts.....	89
	5.1 What Is Probability?.....	89
	5.2 Types of Random Variables	92
	5.3 Discrete Random Variables.....	93
	5.4 Continuous Random Variables.....	100
	5.5 More Rules and Properties of Probability	105
Chapter 6	Binomial Probability Distributions.....	115
	6.1 Binomial Random Variables	115
	6.2 Calculating Binomial Probabilities	116
	6.3 Cumulative Probabilities	119
	6.4 Probability Calculators	120
	6.5 Expected Value and Variance of a Binomial	123
	Random Variable	
	6.6 Using the Binomial Distribution to Help	125
	Make Judgments	
Chapter 7	Normal Probability Distributions	129
	7.1 Normal Distributions.....	129
	7.2 Determining Normal Probabilities	139
	7.3 Finding Values That Correspond to Normal	145
	Probabilities	
	7.4 Assessing Departures from Normality.....	147
Chapter 8	Introduction to Statistical Inference.....	155
	8.1 Concepts.....	155
	8.2 Sampling Behavior of a Mean.....	158
	8.3 Sampling Behavior of a Count and Proportion.....	167
Chapter 9	Basics of Hypothesis Testing	175
	9.1 The Null and Alternative Hypotheses.....	175
	9.2 Test Statistic	178
	9.3 <i>P</i> -Value	181
	9.4 Significance Level.....	182
	9.5 One-Sample <i>z</i> Test	184
	9.6 Power and Sample Size	188

Chapter 10	Basics of Confidence Intervals.....	197
	10.1 Introduction to Estimation	197
	10.2 Confidence Interval for μ When σ Known	199
	10.3 Sample Size Requirements.....	203
	10.4 Relationship Between Hypothesis Testing	205
	and Confidence Intervals	
Part II	<i>Quantitative Response Variable</i>	
Chapter 11	Inference About a Mean	209
	11.1 Estimated Standard Error of the Mean	209
	11.2 Student's t Distributions	210
	11.3 One-Sample t Test.....	214
	11.4 Confidence Interval for μ	217
	11.5 Paired Samples	218
	11.6 Conditions for Inference	224
	11.7 Sample Size and Power	226
Chapter 12	Comparing Independent Means.....	235
	12.1 Paired and Independent Samples.....	235
	12.2 Exploratory and Descriptive Statistics.....	239
	12.3 Inference About the Mean Difference.....	243
	12.4 Equal Variance t Procedure (Optional)	247
	12.5 Conditions for Inference	248
	12.6 Sample Size and Power	250
Chapter 13	Comparing Several Means (One-Way ANOVA)	259
	13.1 Descriptive Statistics.....	260
	13.2 The Problem of Multiple Comparisons	265
	13.3 Analysis of Variance (ANOVA)	266
	13.4 Post Hoc Comparisons.....	276
	13.5 The Equal Variance Assumption.....	282
	13.6 Introduction to Non-Parametric Tests	287
Chapter 14	Correlation and Regression	295
	14.1 Data.....	295
	14.2 Scatterplots	296
	14.3 Correlation.....	299
	14.4 Regression	311

Chapter 15	Multiple Linear Regression	333
	15.1 The General Idea.....	333
	15.2 The Multiple Linear Regression Model.....	334
	15.3 Categorical Explanatory Variables in	337
	Regression Models	
	15.4 Regression Coefficients.....	340
	15.5 ANOVA for Multiple Linear Regression.....	342
	15.6 Examining Multiple Regression Conditions	346
Part III	<i>Categorical Response Variable</i>	
Chapter 16	Inference About a Proportion	349
	16.1 Proportions	349
	16.2 The Sampling Distribution of a Proportion.....	352
	16.3 Hypothesis Test, Normal Approximation	354
	16.4 Hypothesis Test, Exact Binomial Method.....	357
	16.5 Confidence Interval for a Population Proportion.....	363
	16.6 Sample Size and Power	366
Chapter 17	Comparing Two Proportions.....	373
	17.1 Data.....	373
	17.2 Proportion Difference (Risk Difference).....	375
	17.3 Hypothesis Test	380
	17.4 Proportion Ratio (Relative Risk)	389
	17.5 Systematic Sources of Error	393
	17.6 Power and Sample Size	396
Chapter 18	Cross-Tabulated Counts	407
	18.1 Types of Samples	407
	18.2 Describing Naturalistic and Cohort Samples.....	409
	18.3 Chi-Square Test of Association	421
	18.4 Test for Trend.....	431
	18.5 Case-Control Samples	436
	18.6 Matched Pairs	446
Chapter 19	Stratified 2-by-2 Tables	465
	19.1 Preventing Confounding.....	465
	19.2 Simpson's Paradox	466
	19.3 Mantel-Haenszel Methods	468
	19.4 Interaction	474

Appendix A Table of 2000 Random Digits.....	483
Appendix B z Table. Cummulative Probablities for a Standard Normal Random Variable	485
Appendix C t Table	487
Appendix D F Table	489
Appendix E X^2 Table	493
Appendix F Two-Tails of z	495
Answers to Odd Numbered Exercises	497
Index	547

Preface

Basic Biostatistics is an introductory text that presents statistical ideas and techniques for students and workers in public health and biomedical practice and research. The book is designed to be accessible to students with modest mathematical backgrounds; no more than high school algebra is needed to understand this book. With this said, I hope to get past the notion that biostatistics is just an extension of math. Biostatistics is much more than that; it is a combination of mathematics and careful reasoning. Do not let the former interfere with the latter.

Biostatistical analysis is more than just number crunching; it considers how research questions are generated, studies are designed, data are collected, and results are interpreted.

Analysis of data, with a more or less statistical flavor, should play many roles.^a

Basic Biostatistics pays particular attention to exploratory and descriptive analyses. Whereas many introductory biostatistics texts give this topic intermittent attention, this text gives it ongoing consideration.

Both exploratory and confirmatory data analysis deserves our attention.^b

Biostatistics entails formulating research questions and designing processes for exploring and testing theories. I hope students who come to the study of biostatistics asking “What’s the right answer?” leave asking questions like “Was that the right question?” and “Has the question been answered adequately?”

Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.^c

^a Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, 34(1), 23–25.

^b Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24, 83.

^c Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33(1), 13–14.

Several additional points bear emphasis:

Point 1: Practice, practice, practice. In studying biostatistics, you are developing a new set of reasoning skills. What is true of developing other skills is true of developing biostatistical skills—the only way to get better is to practice with the proper awareness and attention. To this end, illustrative examples and exercises are incorporated throughout the book. I've tried to make illustrations and exercises relevant. Many are contemporary, and many have historical importance. Carefully following the reasoning of illustrations and exercises is an important opportunity to learn. Answers to odd-numbered exercises are provided toward the back of the book. Qualified instructors may request answers to even-numbered exercises from the publisher.

Point 2: Structure of the book. The structure of this book may differ from that of other texts. Chapters are intentionally brief. They allow for flexibility in the order of coverage. The book is organized into three main parts. Part I (Chapters 1–10) addresses basic concepts and techniques. Students should complete these chapters (or a comparable introductory course) before moving on to Parts II and III.

Part II (Chapters 11–15) covers analytic techniques for quantitative responses. Part III (Chapters 16–19) covers techniques for categorical responses. Chapters in these sections can be covered in many different orders at the discretion of the instructor. One instructor may choose to cover these chapters in sequence, while another may cover Chapter 11 and Chapter 16 simultaneously (as an example), because these chapters both address one-sample problems. (Chapter 11 covers one-sample problems for quantitative responses; Chapter 16 covers one-sample problems for binary responses.) As another example, one could cover the chapters on categorical responses (Chapters 16–19) before covering the chapters on quantitative responses (Chapter 11–15).

Point 3: Hand calculations and computational support. While I believe there is still benefit in learning how to calculate statistics by hand, students are encouraged to use statistical software to supplement and check calculations. Use of the proper software tools can free us from some of the tedium of numerical manipulations, leaving more time to step back and think about practical implications of results.

The only way humans can do BETTER than computers is to take a chance of doing WORSE. So we have got to take seriously the need for steady progress toward

teaching routine procedures to computers rather than to people. That will leave the teachers of people with only things hard to teach, but this is our proper fate.^d

The book is not tied to any particular software package, but does make frequent use of these three programs: *StaTable*, *SPSS*, and *WinPepi*.

- *StaTable*^e is a freeware program that provides access to 25 commonly used statistical distributions. It runs on Windows, Palm, and Web-browser (Java) platforms. This utility eliminates the need to look up probabilities in hard-copy tables. It also allows for more exact interpolations for probabilities, especially for continuous random variables. The website for this book includes a link to the *StaTable* website.
- *SPSS*^f is a commercial software package with versions that run on Windows and Macintosh computers. A student version of the program can be purchased at most campus bookstores. It can also be purchased online at www.journeyed.com. An economical alternative to purchasing the product is to lease it for short-term use through the Web site www.e-academy.com.
- *WinPepi*^g stands for **W**INDOWS **P**ROGRAMS for **E**PIdemiologists. This is a series of computer programs written by Joe Abramson of the Hebrew University–Hadassah School of Public Health and Community Medicine, (Jerusalem, Israel) and Paul Gahlinger (University of Utah in Salt Lake City). The programs are designed for use in practice, but are also excellent learning aids. *WinPepi* is free and can be downloaded from the website for this book: <http://publichealth.jbpub.com/book/gerstman>.

^d Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, 34, 23–25.

^e www.cytel.com/Products/StaTable/, Cytel Inc., 675 Massachusetts Ave., Cambridge, Massachusetts 02139.

^f SPSS, Inc., Chicago, IL.

^g Abramson, J. H. (2004). *WINPEPI* (PEPI-for-Windows): Computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations*, 1(1), 6.

Acknowledgments

I wish to express my appreciation to San Jose State University for affording me the leave to work on this book. I would especially like to thank the chair of my department, Kathleen Roe, and dean of my college, Inger Sagatun-Edwards, for administrative support in this regard. I am grateful to the colleagues in my department who taught many of my classes during my absence, especially Jane Pham, Dan Perales, Jenny McNeill, and to those who covered other duties, including Ramani Rangavajhula, Nancy Hikoyeda, Polly Bith-Melander, and Edward Mamary. I greatly appreciate the artistic and technical support of Jean Shiota of the Center for Faculty Development for her work in preparing illustrations for the text. Thanks, Jean. Finally, I wish to express my thanks to those many students in my classes over the years who have provided me with helpful comments, encouragement, and camaraderie.

While writing this book, I had many constructive discussions with Joe Abramson of the Department of Social Medicine, Hebrew University–Hadassah School of Public Health and Community Medicine. I thank Joe for sharing his insights generously. I also greatly appreciate his careful work in developing WINDows Programs for EPIdemiologists.ⁱ This is really an exceptional set of programs for public health workers. Along these same lines, Paul Gahlinger (University of Utah) deserves credit for conceiving and creating the progenitor of *WinPepi*, PEPI (Programs for EPIdemiologists).^j I also wish to express my thanks to Mads Haahr (University of Dublin, Trinity College, Ireland) for creating his true random number generator at www.random.org and to John C. Pezzullo

ⁱ Abramson, J. H. (2004). WINPEPI (PEPI-for-Windows): Computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations*, 1(1), 6.

^j Abramson, J. H., & Gahlinger, P. M. (2001). *Computer Programs for Epidemiologic Analyses: PEPI v. 4.0*. Salt Lake City, UT: Sagebrush Press.

(Georgetown University) for his helpful compilation of web pages that perform statistical calculations at www.statpages.org.

Finally, I would like to acknowledge the contributions of my wife, who has been patient, understanding, supportive, and encouraging throughout the work on this marathon project. As Ralph Kramden (Jackie Gleason) used to tell his wife Alice (Audrey Meadows), “[Honey], you’re the greatest!”

About the Author

Dr. Gerstman did his undergraduate work at Harpur College (State University of New York, Binghamton). He later received a doctor of veterinary medicine (Cornell University), a masters of public health (University of California at Berkeley), and a doctor of philosophy degree (University of California, Davis). He has been a U.S. Public Health Service Epidemiology Fellow and epidemiologist at the U.S. Food and Drug Administration and was an instructor at the National Institutes of Health Foundation Graduate School. Since 1990, Dr. Gerstman has been a professor in the Department of Health Science at San Jose State University where he teaches epidemiology, biostatistics, and general education courses. Dr. Gerstman's research interests are in the areas of epidemiologic methods, the history of public health, drug safety, and medical and public health record linkage.

