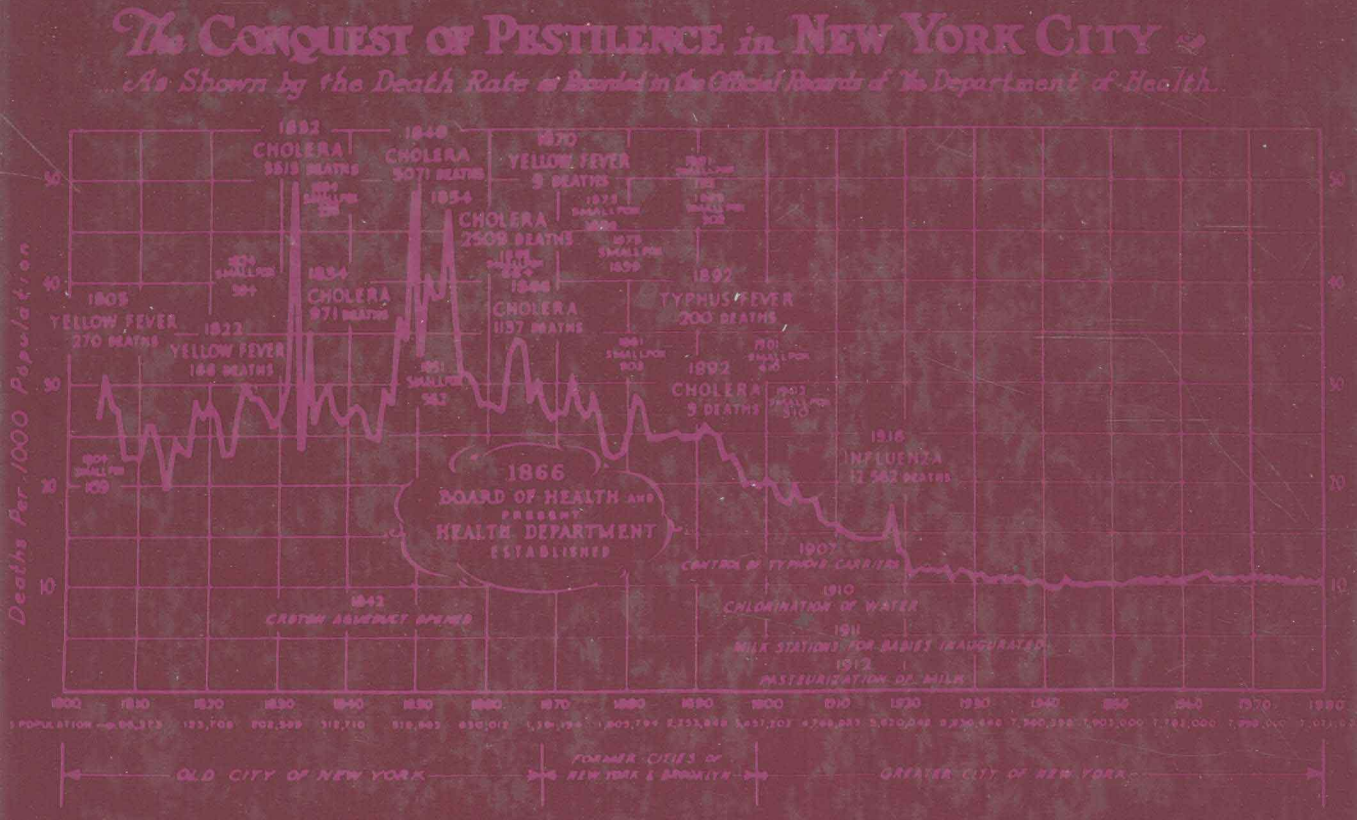
PRINCIPLES OF BIOSTATISTICS



MARCELLO PAGANO

KIMBERLEE GAUVREAU

3½" IBM DISK ENCLOSED

Principles of Biostatistics

Marcello Pagano

Harvard School of Public Health

Kimberlee Gauvreau

Harvard School of Public Health

Duxbury Press
An Imprint of Wadsworth Publishing Company
Belmont, California

Sponsoring Editors: *Michael J. Sugarman* and *Alex Kugushev*
Marketing Representative: *John Moroney*
Editorial Assistant: *Carol Ann Benedict*
Production Coordinator: *Marlene Thom*
Manuscript Editor: *Stacey C. Sawyer*
Interior Design: *Katherine Minerva*
Cover Design: *Susan Haberkorn*
Art Coordinator: *Lisa Torri*
Interior Illustration: *Gloria Langer*
Typesetting: *SuperScript Typography*
Printing and Binding: *R. R. Donnelley & Sons*

This book is printed on acid-free recycled paper.

4 5 6 7 8 9 10–97 96

Library of Congress Cataloging in Publication Data

Pagano, Marcello, [date–]
Principles of biostatistics / Marcello Pagano, Kimberlee Gauvreau.
p. cm.
Includes bibliographical references and index.
ISBN: 0-534-14069-6
1. Biometry I. Gauvreau, Kimberlee, [date–]. II. Title.
QH323.5.P34 1992
574'.01'5195–dc20 92-32841
CIP

This book is dedicated with love to
Phyllis, John-Paul, Marisa, Loris, Alice, and Lilian.
David, Lois, Karen, and Tami.

Preface

This book was written for students of the health sciences and serves as an introduction to the study of biostatistics—the use of numerical techniques to extract information from data and facts. Because numbers are more precise than words, they are particularly well suited for communicating scientific results.

However, just as one can lie with words, one can also lie with numbers. Indeed, numbers and lies have been linked for quite some time; there is even a book entitled *How to Lie with Statistics.* This association may owe its origin, or its affirmation at the very least, to the British prime minister Benjamin Disraeli. Disraeli is credited by Mark Twain as having said "There are three kinds of lies: lies, damned lies, and statistics." One has only to observe any modern political campaign to be convinced of the abuse of statistics. But enough about lies; this book adopts the position of Frederick Mosteller, who said "It is easy to lie with statistics, but it is easier to lie without them."

Principles of Biostatistics is aimed at the student who wishes to learn modern research methods. It is based on a required course offered at the Harvard School of Public Health. A large number of health professionals in the rest of the Harvard medical area attend as well. The course is as old as the School itself, which attests to its importance. It spans 16 weeks of lectures and laboratory sessions. Each week includes two 50-minute lectures and one 2-hour lab. The entire class is together for the lectures but divides into smaller groups, headed by teaching assistants, for the lab sessions. These labs reinforce the material covered in the lectures, review the homework assignments, and introduce the computer into the course. We have included the lab materials—except those dealing with the homework assignments and specific computer commands—in the sections called Further Applications. These sections present additional examples or a different perspective of the material already covered. The labs are designed to provoke discussion, although they are sufficiently complete for the reader who is not using the book as a course text to benefit from reading them.

This book has evolved over the years to include topics that we believe can be covered at some depth in one American semester. Clearly, some choices had to be made; we hope that we have chosen well. In our course, we have sufficient time to cover most of the material in the first 20 chapters. However, there is enough material to allow the instructor some flexibility in the choice of topics. For example, some instructors may choose to omit the sections covering grouped data (Section 3.3), Chebychev's inequality (Section 3.4), and the Poisson distribution (Section 7.3) or the chapter on the analysis of variance (Chapter 12), if they consider these concepts to be less important than others.

Some say that statistics is the study of variability and uncertainty. We believe there is some truth to this adage and have used it as a guide by dividing the book into three parts. The first five chapters deal with collections of numbers and ways in which to summarize, explore, and explain them; these methods are easily comprehensible. The next two chapters consider probability and serve as an introduction to the tools needed for the subsequent investigation of uncertainty. It is only in the eighth chapter and thereafter that we distinguish between populations and samples and begin to investigate the inherent variability introduced by sampling, thus progressing to inference. We think that this modular introduction to the quantification of uncertainty is justified by the success achieved by our students. Postponing the slightly more difficult parts until a solid foundation has been established makes it easier for the reader to comprehend them.

Throughout the text we have used data drawn from published studies to exemplify biostatistical concepts. Not only is real data more meaningful, it is usually more interesting as well. Of course, we must take care not to use examples in which the subject matter is too esoteric or too complex. To this end, we have been guided by the backgrounds and interests of our students—mostly public health and clinical research—to choose examples that best illustrate the concepts at hand.

There is some risk involved in using published data. We cannot guarantee that all the examples are honest and that the data were properly collected; for this we must rely on the reputations of our sources. We do not belittle the importance of this consideration. The value of our inference depends critically on the worth of the data, and we strongly recommend that a good deal of effort be expended on evaluating their quality. We assume that this is understood by the reader.

More than once we have used examples in which the population of the United States is broken down along racial lines. In reporting these official statistics we follow the lead of the government agencies that release them. We do not wish to reify this racial categorization, since in fact the observed differences may well be due to socioeconomic factors rather than the implied racial ones. One option would be to ignore these statistics; however, that would hide inequities that exist in our health system—inequities that need to be eliminated. We focus attention on the problem in the hope of stimulating interest in promoting solutions.

We have minimized the use of mathematical notation because of its well-deserved reputation of being the ultimate jargon. On one hand, it can be over used and intimidate even the most ardent scholar. On the other hand, we should not eliminate its use entirely; it has been developed over the ages to be helpful in communicating results. We hope that in this respect we have written a succinct and understandable text.

Over and above their precision, there is something more to numbers—maybe a little magic—that makes them fun to study. The fun is in the conceptualization more so than the calculations, and we are fortunate to have the computer to do the drudge work. This allows students to concentrate on the ideas. In other words, the computer allows the instructor to teach the poetry of statistics and not the plumbing.

To take advantage of the computer, one needs a good statistical package. We use Stata, which is available from the Computing Resource Center in Santa Monica, California. We find this statistical package to be one of the best on the market today; it is user friendly, accurate, powerful, reasonably priced, and works on a number of different platforms, including DOS, Unix, and Macintosh. Furthermore, the output from this package is acceptable to the Federal Drug Administration in New Drug Approval submissions. Other packages are available, and this book can be supplemented by any one of them. We strongly recommend that some statistical package be used.

Some of the exercises in the text require the use of the computer. To help the reader, we have included the data sets used in the review exercises both in Appendix B and on the disk that accompanies the book. The disk contains each data set in two different formats: an ASCII file (the "raw" suffix) and a Stata file (the "dta" suffix). There are also many exercises that do not require the computer. Answers to selected exercises are supplied in Appendix C. As always, active learning yields better results than passive observation does. To this end, we cannot stress enough the importance of the exercises and urge the reader to attempt as many as time permits.

A debt of gratitude is owed a number of people: President Derek Bok for providing the support that got this book off the ground, Dr. Michael K. Martin for calculating all the statistical tables in Appendix A, and John-Paul Pagano for assisting in the editing. We thank the individuals who reviewed the manuscript: Stewart Anderson, University of Pittsburgh; Ronald Cody, Robert Wood Johnson Medical School; Charles Davis, University of Iowa; Janice Derr, Pennsylvania State University; Christiana Drake, University of California, Davis; David Gjertson, University of California, Los Angeles; James Godbold, Mt. Sinai School of Medicine; Maureen Lahiff, University of California, Berkeley; Nicholas Lange, Brown University; John Lefante, Tulane Medical Center; and Donald Slymen, San Diego State University. Our thanks also go to Karen U. Abbett and Judith Nahas for assistance in typing an early version of the text and to Stacey Sawyer, whose editing suggestions have resulted in an improved book. The teaching assistants who have helped us teach the course and who have made many valuable suggestions, we thank. Probably the most deserving of thanks are the students who have taken the course over the years and who have tolerated us as we learned how to teach it. We are still learning.

Marcello Pagano
Kimberlee Gauvreau

Boston, Massachusetts
August 1992

Contents

1 Introduction 1

1.1 Review Exercises 5

References 6

2 Data Presentation 7

2.1 Types of Numerical Data 7

2.1.1 Nominal Data 7
2.1.2 Ordinal Data 9
2.1.3 Ranked Data 9
2.1.4 Discrete Data 10
2.1.5 Continuous Data 11

2.2 Tables 11

2.2.1 Frequency Distributions 12
2.2.2 Relative Frequency 13

2.3 Graphs 15

2.3.1 Bar Charts 15
2.3.2 Histograms 16
2.3.3 Frequency Polygons 18
2.3.4 One-Way Scatter Plots 20
2.3.5 Box Plots 21
2.3.6 Two-Way Scatter Plots 21
2.3.7 Line Graphs 22

2.4 Further Applications 25

2.5 Review Exercises 30
References 33

3 Numerical Summary Measures 35

3.1 Measures of Central Tendency 35
3.1.1 Mean 35
3.1.2 Median 38
3.1.3 Mode 38
3.2 Measures of Dispersion 40
3.2.1 Range 40
3.2.2 Interquartile Range 41
3.2.3 Variance and Standard Deviation 42
3.2.4 Coefficient of Variation 44
3.3 Grouped Data 44
3.4 Chebychev's Inequality 48
3.5 Further Applications 50
3.6 Review Exercises 54
References 57

4 Rates and Standardization 59

4.1 Rates 59
4.2 Standardization of Rates 61
4.2.1 Direct Method of Standardization 66
4.2.2 Indirect Method of Standardization 67
4.2.3 Use of Standardized Rates 69
4.3 Further Applications 78
4.4 Review Exercises 82
References 84

5 Life Tables 85

5.1 Computation of the Life Table 85
5.1.1 Column 1 85
5.1.2 Column 2 87
5.1.3 Columns 3 and 4 89

5.1.4 Column 5 90
5.1.5 Column 6 91
5.1.6 Column 7 91
5.2 Applications 92
5.3 Years of Potential Life Lost 95
5.4 Further Applications 100
5.5 Review Exercises 104
References 112

6 Probability 115
6.1 Operations on Events 115
6.2 Conditional Probability 117
6.3 Bayes' Theorem 119
6.4 Diagnostic Tests 124
6.4.1 Sensitivity and Specificity 124
6.4.2 Applications of Bayes' Theorem 125
6.4.3 The ROC Curve 128
6.5 Calculation of Prevalence 130
6.6 The Relative Risk and the Odds Ratio 132
6.7 Further Applications 136
6.8 Review Exercises 141
References 143

7 Theoretical Probability Distributions 145
7.1 Probability Distributions 145
7.2 The Binomial Distribution 147
7.3 The Poisson Distribution 153
7.4 The Normal Distribution 157
7.5 Further Applications 168
7.6 Review Exercises 173
References 176

8 Sampling Distribution of the Mean 179
8.1 Sampling Distributions 179
8.2 The Central Limit Theorem 180
8.3 Applications 181
8.4 Further Applications 187
8.5 Review Exercises 193
References 194

9 Confidence Intervals 195
9.1 Two-Sided Confidence Intervals 195
9.2 One-Sided Confidence Intervals 200
9.3 Student's t Distribution 201
9.4 Further Applications 206
9.5 Review Exercises 207
References 208

10 Hypothesis Testing 211
10.1 General Concepts 211
10.2 Two-Sided Tests of Hypotheses 213
10.3 One-Sided Tests of Hypotheses 216
10.4 Types of Error 218
10.5 Power 222
10.6 Sample Size 224
10.7 Further Applications 227
10.8 Review Exercises 230
References 232

11 Comparison of Two Means 235
11.1 Paired Samples 236
11.2 Independent Samples 241
11.2.1 Equal Variances 242

11.2.2 Unequal Variances 245
11.3 Further Applications 248
11.4 Review Exercises 252
References 255

12 Analysis of Variance 257
12.1 One-Way Analysis of Variance 257
12.1.1 The Problem 257
12.1.2 Sources of Variation 260
12.2 Multiple Comparisons Procedures 263
12.3 Further Applications 266
12.4 Review Exercises 269
References 271

13 Nonparametric Methods 273
13.1 The Sign Test 273
13.2 The Wilcoxon Signed-Rank Test 276
13.3 The Wilcoxon Rank Sum Test 279
13.4 Advantages and Disadvantages of Nonparametric Methods 283
13.5 Further Applications 284
13.6 Review Exercises 288
References 290

14 Inference on Proportions 293
14.1 Normal Approximation to the Binomial Distribution 294
14.2 Sampling Distribution of a Proportion 296
14.3 Confidence Intervals 297
14.4 Hypothesis Testing 299
14.5 Sample Size Estimation 300

14.6 Comparison of Two Proportions 301
14.7 Further Applications 305
14.8 Review Exercises 308
References 310

15 Contingency Tables 311
15.1 The Chi-Square Test 311
15.1.1 2×2 Tables 311
15.1.2 $r \times c$ Tables 316
15.2 McNemar's Test 318
15.3 The Odds Ratio 320
15.4 Berkson's Fallacy 325
15.5 Further Applications 328
15.6 Review Exercises 332
References 336

16 Multiple 2×2 Tables 339
16.1 Simpson's Paradox 339
16.2 The Mantel-Haenszel Method 341
16.2.1 Test of Homogeneity 343
16.2.2 Summary Odds Ratio 346
16.2.3 Test of Association 350
16.3 Further Applications 353
16.4 Review Exercises 359
References 360

17 Correlation 363
17.1 The Two-Way Scatter Plot 363
17.2 Pearson's Correlation Coefficient 365
17.3 Spearman's Rank Correlation Coefficient 369
17.4 Further Applications 372
17.5 Review Exercises 376

References 378

18 Simple Linear Regression 379

18.1 Regression Concepts 379

18.2 The Model 384

18.2.1 The Population Regression Line 384
18.2.2 The Method of Least Squares 386
18.2.3 Inference for Regression Coefficients 388
18.2.4 Inference for Predicted Values 392

18.3 Evaluation of the Model 395

18.3.1 The Coefficient of Determination 395
18.3.2 Residual Plots 395
18.3.3 Transformations 397

18.4 Further Applications 400

18.5 Review Exercises 404

References 408

19 Multiple Regression 409

19.1 The Model 409

19.1.1 The Least Squares Regression Equation 410
19.1.2 Inference for Regression Coefficients 411
19.1.3 Evaluation of the Model 412
19.1.4 Indicator Variables 413
19.1.5 Interaction Terms 414

19.2 Model Selection 415

19.3 Further Applications 418

19.4 Review Exercises 422

References 424

20 Logistic Regression 427

20.1 The Model 427

20.1.1 The Logistic Function 429
20.1.2 The Fitted Equation 430

20.2 Multiple Logistic Regression 432

20.3 Indicator Variables 435
20.4 Further Applications 438
20.5 Review Exercises 441
References 443

21 Survival Analysis 445
21.1 The Life Table Method 446
21.2 The Product-Limit Method 451
21.3 The Log-Rank Test 455
21.4 Further Applications 459
21.5 Review Exercises 466
References 467

22 Sampling Theory 469
22.1 Sampling Schemes 469
22.1.1 Simple Random Sampling 470
22.1.2 Systematic Sampling 471
22.1.3 Stratified Sampling 471
22.1.4 Cluster Sampling 472
22.2 Sources of Bias 472
22.3 Further Applications 475
22.4 Review Exercises 479
References 480

Appendix A Tables 483

Appendix B Data Sets 503

Appendix C Solutions to Selected Exercises 513

Index 521

CHAPTER 1

Introduction

H. G. Wells is reputed to have said that statistical reasoning will one day be as important to good citizenship as the ability to read and write. Indeed, statistics play an important role in many decision-making processes. Before a new drug can be marketed, for instance, the U.S. Food and Drug Administration requires that it be subjected to a clinical trial—an experimental study involving human subjects. The data from this study must be compiled and analyzed to determine whether the drug is effective. In addition, the U. S. government's decisions regarding Social Security and public health programs rely in part on predictions about the longevity of the nation's population; as a result, it must be able to predict the number of years that a particular individual will live. There are many other questions that we would like to answer as well. Where should a government invest its resources if it wishes to reduce infant mortality? Does the use of a seat belt decrease the chance of death in a motor vehicle accident? Should a mastectomy always be recommended to a patient with breast cancer? What factors increase the risk that an individual will develop coronary heart disease? To address these issues and others, we rely on the methods of biostatistics.

The study of statistics investigates the collection, organization, analysis, and interpretation of numerical data. The concepts of statistics may be applied to a number of fields that include business, psychology, and agriculture. When focus is on the biological and health sciences, we use the term *biostatistics*.

Historically, statistics have been used to tell a story with numbers. Numbers often communicate ideas more succinctly than do words. For example, for the following data, the picture is quite clear. In 1979, 48 persons in Japan, 34 in Switzerland, 52 in Canada, 58 in Israel, 21 in Sweden, 42 in Germany, 8 in England, and 10,728 in the United States were killed by handguns [1]. The power of these numbers is obvious; the message would remain clear even if we were to correct for differences in population size.

As a second example, consider the following quotation taken from an editorial in *The Boston Globe* [2]:

> Lack of contraception is linked to an exceptionally high abortion rate in the Soviet Union—120 abortions for every 100 births, compared with 20 per 100 births in Great Britain, where access to contraception is guaranteed. Inadequate support for family planning in the United States has resulted in 40 abortions for every 100 births—a lower rate than the Soviet Union, but twice as high as most industrialized nations.

In this case, a great deal of information is contained in only three numbers: 120, 20, and 40. The statistics provide some insight into the consequences of differing attitudes toward family planning.

In both these examples, the numbers provide a concise summary of certain aspects of the situation being studied. Surely the numerical explanation of the handgun data is more illuminating than if we had been told that some people got killed in Japan, fewer in Switzerland, more in Canada, still more in Israel, but far fewer in Sweden, and so forth. Both examples deal with very complex situations, yet the numbers convey the essential information. Of course, no matter how powerful, no statistic will convince everyone that a given conclusion is true. The handgun data are often brushed away with the aphorism, "Guns don't kill people, people do." This should not be surprising; after all, there are still members in the Flat Earth Society. The aim of a biostatistical study is to provide the numbers that contain information about a certain situation and to present them in such a way that valid interpretations are possible.

If we wish to study the effects of a new diet, we might begin by measuring the changes in body mass for all individuals who have been placed on the diet. Similarly, if we want to investigate the success of a certain therapy for treating breast cancer, we would look at the lengths of time that women treated with this therapy survive beyond diagnosis of the disease. However, the collections of numbers are not very informative until we start combining them in some way. Descriptive statistics are methods for organizing and summarizing data that help us to describe the attributes of a group or population. In Chapter 2, we examine both tabular and graphical descriptive techniques. The graphical capabilities of computers have made this type of summarization more feasible than in the past, and a whole new mode of presentation is now available for even the most modest analyses.

Chapter 3 goes beyond the graphical techniques presented in Chapter 2 and introduces numerical summary measures. By definition, a summary captures only a particular aspect of the data being studied; consequently, it is also important to have an idea of how well the summary represents the data as a whole. For example, we might wish to know how long AIDS patients survive after diagnosis with one of the characteristic opportunistic infections. If we calculate an average survival time, is this average representative of all patients? Furthermore, how