PRINCIPLES OF MEDICAL STATISTICS

Alvan R. Feinstein, M.D.

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Feinstein, Alvan R., 1925–
Principles of medical statistics / Alvan R. Feinstein.
p. ; cm.
Includes bibliographical references and index.
ISBN 1-58488-216-6 (alk. paper)
1. Medicine—Statistical methods.
[DNLM: 1. Statistics—methods. 2. Data Interpretation, Statistical. WA 950 F299p 2001] I. Title.
R853.S7 F45 2001
610'.7'27—dc21 2001001794

International Standard Book Number 1-58488-216-6
Library of Congress Card Number 2001001794
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Preface

What! Yet another book on medical biostatistics! Why? What for?

The purpose of this preface is to answer those questions and to add a few other pertinent remarks. The sections that follow describe a series of distinctions, some of them unique, that make this book different from other texts.

Goals and Objectives

The goal of the text is to get biomedical readers to think about data and statistical procedures, rather than learn a set of "cook-book recipes." In many statistics books aimed at medical students or biomedical researchers, the readers are believed to have either little interest or limited attention. They are then offered a simple, superficial account of the most common doctrines and applications of statistical theory. The "get-it-over-with-quickly" approach has been encouraged and often necessitated by the short time given to statistics in modern biomedical education. The curriculum is supposed to provide fundamental background for the later careers of medical and other graduate students, but the heavily stressed "basic science" topics are usually cellular and molecular biology. If included at all, statistics is usually presented briefly, as a drudgery to be endured mainly because pertinent questions may appear in subsequent examinations for licensure or other certifications.

Nevertheless, in later professional activities, practicing clinicians and biomedical researchers will constantly be confronted with reports containing statistical expressions and analyses. The practitioners will regularly see and use statistical results when making clinical decisions in patient care; and the researchers will regularly be challenged by statistical methods when planning investigations and appraising data. For these activities, readers who respect their own intellects, and who want to understand and interpret the statistical procedures, cannot be merely passive learners and compliant appliers of doctrinaire customs. The readers should think about what they want, need, and receive. They should also recognize that their knowledge of the substantive biomedical phenomena is a major strength and dominant factor in determining how to get, organize, and evaluate the data. This book is aimed at stimulating and contributing to those thoughts.

Another distinction of the text is that the author is a physician with intimate and extensive experience in both patient care and biomedical investigation. I had obtained a master's degree in mathematics before entering medical school, but thereafter my roots were firmly and irrevocably grounded in clinical medicine. When I later began doing clinical research and encountering statistical strategies, my old mathematical background saved me from being intimidated by established theories and dogmas. Although not all statisticians will approve the temerity of an "unauthorized" writer who dares to compose a text in which the fundamental basis of old statistical traditions is sometimes questioned,

other statisticians may be happy to know more about the substantive issues contained in biomedical research, to learn what their clients are (or should be) thinking about, and to lead or collaborate in developing the new methods that are sometimes needed.

New Methods and Approaches

The text contains many new methods and approaches that have been made possible by advances in statistical strategy for both analytic description and inferential decisions.

Statistical description has traditionally relied on certain mathematical models, such as the Gaussian distribution of a "normal" curve, that summarize data with means, standard deviations, and arbitrarily constructed histograms. Readers who begin to think about what they really want, however, may no longer happily accept what is offered by those old models. For example, because biomedical data seldom have a Guassian distribution, the *median* is usually a much better summary value than the *mean*; and new forms of data display — the stem-and-leaf plot and the box plot — not only are superior to histograms, but are more natural forms of expression.

Another descriptive distinction, which is omitted or blurred in many text books, is the difference between a trend (for citing correlation or regression) and a concordance (for citing agreement). Investigators who study variability in observers or in laboratory procedures have usually been taught to express results with the conventional indexes of "association" that denote trend, but not concordance. This text emphasizes the difference between correlation and agreement; and separate chapters are devoted to both "nondirectional" concordance (for observer variability) and "directional" concordance (for accuracy of marker tests).

In statistical inference for decisions about probability, the customary approach has used hard-to-understand mathematical theories and hypothetical assumptions that were developed, established, and entrenched (for topics such as t tests and chi-square tests), because they led to standard formulas for relatively simple calculations. During the past few decades, however, the elaborate mathematical theories and assumptions have been augmented, and sometimes replaced, by easy-to-understand new methods, which use rearrangements or resamplings of the observed data. The new methods often require formidable calculations that were not practical in the pre-computer era; but today, the "computer-intensive" work can be done quickly and easily, requiring no more effort than pushing the right "button" for an appropriate program. The new methods, which may eventually replace the old ones, are discussed here as additional procedures that involve no complicated mathematical backgrounds or unrealistic assumptions about "parametric" sampling from a theoretical population. In the new methods — which have such names as *Fisher exact test*, *bootstrap*, and *jackknife* — all of the rearrangements, resamplings, and statistical decisions about probability come directly from the empirical real-world data. Another departure from tradition is a reappraisal of the use of probability itself, with discussions of what a reader really wants to know, which is *stability* of the numbers, not just probabilistic assessments.

The text also has sections that encourage methods of "physical diagnosis" to examine the data with procedures using only common sense and in-the-head-without-a-calculator appraisals. From appropriate summary statistics and such graphic tactics as box-plot displays, a reader can promptly see what is in the data and can then make some simple,

effective, mental calculations. The results will often offer a crude but powerful check on more complex mathematical computations.

A particularly novel and valuable approach is the careful dissection (and proposed elimination) of the term *statistical significance*, which has been a source of major confusion and intellectual pathogenicity throughout 20th-century science. *Statistical significance* is an ambiguous term, because it does not distinguish between the theoretical stochastic significance of calculated probabilities (expressed as P values and confidence intervals) and the pragmatic quantitative significance or clinical importance of the "effect sizes" found in the observed results. Not only is the crucial difference between stochastic and quantitative significance emphasized and thoroughly discussed, but also a special chapter, absent from conventional texts, is devoted to the indexes of contrast used for expressing and evaluating the "effect size" of quantitative distinctions.

Two other unique features of this text are the following:

- Two chapters on the display of statistical data in tables, charts, and graphs contain good and bad examples that can be helpful to readers, investigators, and the artists who prepare medical illustrations.
- A chapter that discusses the challenges of evaluating "equivalence" rather than "superiority" also considers the management of problems that arise when discordance arises in what the investigator wants, what the results show, and what the statistical tests produce.

Sequence, Scope, Rigor, and Orientation

The text is arranged in a logical sequence of basic principles that advance from simple to more elaborate activities. It moves from evaluating one group of data to comparing two groups and then associating two variables. Thereafter, the scope extends into more complex but important topics that frequently appear as challenges in biomedical literature: controversies about stochastic issues in choosing one- or two-tailed tests, the graphic patterns of survival analysis, and the problems of appraising "power," determining "equivalence," and adjudicating "multiple hypotheses."

Nevertheless, despite some of the cited deviations from customary biostatistical discourse, the text describes all the conventional statistical procedures and offers reasonably rigorous accounts of many of their mathematical justifications. Whether retaining or rejecting the conventional procedures, a reader should know what they do, how they do it, and why they have been chosen to do it. Besides, the conventional procedures will continue to appear in biomedical literature for many years. Learning the mechanisms (and limitations) of the traditional tactics will be an enlightened act of self-defense.

Finally, although the conventional mathematical principles are given a respectful account, the book has a distinctly clinical orientation. The literary style is aimed at biomedical readers; and the examples and teaching exercises all come from the real-world medical phenomena. The readers are not expected to become statisticians, although appropriate historical events are sometimes cited and occasional mathematical challenges are sometimes offered. Clinical and biomedical investigators have made many contributions to other "basic" domains, such as cell and molecular biology, and should not be discouraged from helping the development of another "basic" domain, particularly the *bio*-portion

of biostatistics. As preparation for a future medical career, such basic tools as the methods of history taking, auscultation, imaging, catheterization, and laboratory tests are almost always taught with a clinical orientation. As another important basic tool, statistics receives that same orientation here.

Containing much more than most "elementary" books, this text can help repair the current curricular imbalance that gives so little attention to the role of statistics as a prime component of "basic" biomedical education. Statistical procedures are a vital, integral part of the "basic" background for clinical or biomedical careers, and are essential for readers and investigators who want to be at least as thoughtful in analyzing results as in planning and doing the research. The biomedical readers, however, are asked to read the text rather than race through it. What they learn will help them think for themselves when evaluating various statistical claims in the future. They can then use their own minds rather than depending on editorial decisions, authoritarian pronouncements, or the blandishments of various medical, commercial, or political entrepreneurs.

Before concluding, I want to thank various faculty colleagues — (alphabetically) Domenic Cicchetti, John Concato, Theodore Holford, Ralph Horwitz, James Jekel, Harlan Krumholz, Robert Makuch, Peter Peduzzi, and Carolyn Wells—who have contributed to my own statistical education. I also want to acknowledge the late Donald Mainland, whose writings made me realize that statistics could be profound but comprehensible while also being fun, and who launched my career in statistics when he invited me to succeed him in writing a bimonthly set of journal essays on biostatistical topics. I am immensely grateful to the post-residency physicians who have been research fellows in the Yale Clinical Scholar Program, sponsored by the Robert Wood Johnson Foundation and also supported by the U.S. Department of Veterans Affairs. The Clinical Scholars are the people who inspired the writing of this text, who have received and worked through its many drafts, whose comments and suggestions have produced worthwhile improvements, and who helped create an intellectual atmosphere that I hope will be reflected and preserved. While I was trying to prod the group into learning and thinking, their responses gave me the stimuli and pleasures of additional learning and thinking.

My last set of acknowledgments contains thanks to people whose contributions were essential for the text itself. Donna Cavaliere and many other persons — now too numerous to all be named — did the hard, heroic work of preparing everything on a word processor. Robert Stern, of Chapman and Hall publishers, has been an excellent and constructive editor. Carole Gustafson, also of Chapman and Hall, has done a magnificent job of checking everything in the text for logic, consistency, and even grammar. I am grateful to Sam Feinstein for esthetic advice and to Yale's Biomedical Communications department for preparing many of the illustrations. And finally, my wife Lilli, has been a constant source of patience, encouragement, and joy.

Alvan R. Feinstein
New Haven
July, 2001

Biographical Sketch

Alvan R. Feinstein was born in Philadelphia and went to schools in that city before attending the University of Chicago, from which he received a bachelor's degree, a master's degree in mathematics, and his doctor of medicine degree. After residency training in internal medicine at Yale and at Columbia-Presbyterian Hospital in New York, and after a research fellowship at Rockefeller Institute, he became medical director at Irvington House, just outside New York City, where he studied a large population of patients with rheumatic fever. In this research, he began developing new clinical investigative techniques that were eventually expanded beyond rheumatic fever into many other activities, particularly work on the prognosis and therapy of cancer.

His new clinical epidemiologic approaches and methods have been reported in three books, *Clinical Judgment*, *Clinical Epidemiology*, and *Clinimetrics*, which describe the goals and methods of clinical reasoning, the structure and contents of clinical research with groups, and the strategy used to form clinical indexes and rating scales for important human clinical phenomena — such as pain, distress, and disability — that have not received adequate attention in an age of technologic data. His clinical orientation to quantitative data has been presented in two previous books, *Clinical Biostatistics* and *Multivariable Analysis*, and now in the current text.

To supplement the current biomedical forms of "basic science" that are used for explanatory decisions about pathophysiologic mechanisms of disease, Feinstein has vigorously advocated that clinical epidemiology and clinimetrics be developed as an additional humanistic "basic science" for the managerial decisions of clinical practice.

He is Sterling Professor of Medicine and Epidemiology at the Yale University School of Medicine, where he is also Director of the Clinical Epidemiology Unit and Director Emeritus of the Robert Wood Johnson Clinical Scholars Program. For many years he also directed the Clinical Examination Course (for second-year students).

Table of Contents

1

Introduction

CONTENTS

Suppose you have just read a report in your favorite medical journal. The success rates were said to be 50% with a new treatment for Disease D, and 33% in the control group, receiving the customary old treatment. You now have a clinical challenge: Should you beginning the new treatment instead of the old one for patients with Disease D?

Decisions of this type are constantly provoked by claims that appear in the medical literature or in other media, such as teaching rounds, professional meetings, conferences, newspapers, magazines, and television. In the example just cited, the relative merits were compared for two treatments, but many other medical decisions involve appraisals of therapeutic hazards or comparisons of new technologic procedures for diagnosis. In other instances, the questions are issues in public and personal health, rather than the clinical decisions in diagnosis or treatment. These additional questions usually require evaluations for the medical risks or benefits of phenomena that occur in everyday life: the food we eat; the water we drink; the air we breathe; the chemicals we encounter at work or elsewhere; the exercise we take (or avoid); and the diverse patterns of behavior that are often called "life style."

If not provoked by statistics about therapeutic agents or the hazards of daily life, the questions may arise from claims made when investigators report the results of laboratory research. A set of points that looks like scattered buckshot on a graph may have been fitted with a straight line and accompanied by a statement that they show a "significant" relationship. The mean values for results in two compared groups may not seem far apart, but may be presented with the claim that they are distinctively different.

Either these contentions can be accepted in the assumption that they were verified by the wisdom of the journal's referees and editors, or — mindful of the long history of erroneous doctrines that were accepted and promulgated by the medical "establishment" in different eras — we can try to evaluate things ourselves. To do these evaluations, we need some type of rational mechanism. What kinds of things shall we think about? What should we look for? How should we analyze what we find? How do we interpret the results of the analysis?

The final *conclusions* drawn from the evaluations are seldom expressed in statistical terms. We conclude that treatment A is preferable to treatment B, that diagnostic procedure C is better than

diagnostic procedure E, that treatment F is too hazardous to use, or that G is a risk factor for disease H. Before we reach these nonstatistical conclusions, however, the things that begin the thought process are often statistical expressions, such as success rates of 50% vs. 33%.

The statistical citation of results has become one of the most common, striking phenomena of modern medical literature. No matter what topic is under investigation, and no matter how the data have been collected, the results are constantly presented in statistical "wrappings." To evaluate the results scientifically, we need to look beneath the wrapping to determine the scientific quality of the contents. This look inside may not occur, however, if a reader is too flustered or uncomfortable with the exterior statistical covering. Someone familiar with medical science might easily understand the interior contents, but can seldom reach them if the statistical material becomes an obscure or intimidating barrier.

The frequent need to appraise numerical information creates an intriguing irony in the professional lives of workers in the field of medicine or public health. Many clinicians and public-health personnel entered those fields because they liked people and liked science, but hated mathematics. After the basic professional education is completed, the subsequent careers may bring the anticipated pleasure of working in a humanistic science, but the pleasure is often mitigated by the oppression of having to think about statistics

This book is intended to reduce or eliminate that oppression, and even perhaps to show that statistical thinking can be intellectually attractive. The main point to recognize is that the primary base of statistical thinking is not statistical. It requires no particular knowledge or talent in mathematics; and it involves only the use of enlightened common sense — acquired as ordinary common sense plus professional knowledge and experience. Somewhat like a "review of systems" in examining patients, a statistical evaluation can be divided into several distinctive components. Some of the components involve arithmetic or mathematics, but most of them require only the ability to think effectively about what we already know.

1.1 Components of a Statistical Evaluation

The six main components of a statistical "review of systems" can be illustrated with examples of the way they might occur for the decision described at the start of this chapter.

1.1.1 Summary Expressions

The first component we usually meet in statistical data is a summary expression, such as a *50% success rate*. The statistical appraisal begins with the adequacy of this expression. Is it a satisfactory way of summarizing results for the observed group? Suppose the goal of treatment was to lower blood pressure. Are you satisfied with a summary in which the results are expressed as success rates of 50% or 33%? Would you have wanted, instead, to know the average amount by which blood pressure was lowered in each group? Would some other quantitative expression, such as the average weekly change in blood pressure, be a preferable way of summarizing each set of data?

1.1.2 Quantitative Contrasts

Assuming that you are satisfied with whatever quantitative summary was used to express the individual results for each group, the second component of evaluation is a contrast of the two summaries. Are you impressed with the comparative distinction noted in the two groups? Does a 17% difference in success rates of 50% and 33% seem big enough to be important? Suppose the difference of 17% occurred as a contrast of 95% vs. 78%, or as 20% vs. 3%. Would these values make you more impressed or less impressed by the distinction? If you were impressed not by the 17% difference in the two numbers, but by their ratio of 1.5 (50/33), would you still be impressed if the same ratio were obtained from the contrast of 6% vs. 4% or from .0039 vs. 0026? What, in fact, is the strategy you use for deciding that a distinction in two contrasted numbers has an "impressive" magnitude?

1.1.3 Stochastic Contrasts

If you decided that the 50% vs. 33% distinction was impressive, the next step is to look at the numerical sources of the compared percentages. This component of evaluation contains the type of statistics that may be particularly distressing for medical people. It often involves appraising the stochastic (or probabilistic) role of random chance in the observed numerical results. Although the quantitative contrast of 50% vs. 33% may have seemed impressive, suppose the results came from only five patients. The 50% and 33% values may have emerged, respectively, as one success in two patients and one success in three patients. With constituent numbers as small as 1/2 and 1/3, you would sense intuitively that results such as 50% vs. 33% could easily occur by chance alone, even if the two treatments were identical.

Suppose, however, that the two percentages (50% vs. 33%) were based on such numbers as 150/300 vs. 100/300? Would you now worry about chance possibilities? Probably not, because the contrast in these large numbers seems distinctive by what Joseph Berkson has called the traumatic interocular test. (The difference hits you between the eyes.) Now suppose that the difference of 50% and 33% came from numbers lying somewhere between the two extremes of 1/2 vs. 1/3 and 150/300 vs. 100/300. If the results were 8/16 vs. 6/18, the decision would not be so obvious. These numbers are neither small enough to be dismissed immediately as "chancy" nor large enough to be accepted promptly as "intuitively evident."

The main role of the third component of statistical evaluation is to deal with this type of problem. The process uses mathematical methods to evaluate the "stability" of numerical results. The methods produce the P values, confidence intervals, and other probabilistic expressions for which statistics has become famous (or infamous).

1.1.4 Architectural Structure

Suppose you felt satisfied after all the numerical thinking in the first three steps. You accepted the expression of "success"; you were impressed by the quantitative distinction of 50% vs. 33%; and the numbers of patients were large enough to convince you that the differences were not likely to arise by chance. Are you now ready to start using the new treatment instead of the old one? If you respect your own common sense, the answer to this question should be a resounding NO. At this point in the evaluation, you have thought only about the statistics, but you have not yet given any attention to the science that lies behind the statistics. You have no idea of how the research was done, and what kind of architectural structure was used to produce the compared results of 50% and 33%.

The architectural structure of a research project refers to the scientific arrangement of persons and circumstances in which the research was carried out. The ability to evaluate the scientific architecture requires no knowledge of statistics and is the most powerful analytic skill at your disposal. You can use this skill to answer the following kinds of architectural questions: Under what clinical conditions were the two treatments compared? Were the patients' conditions reasonably similar in the two groups, and are they the kind of conditions in which you would want to use the treatment? Were the treatments administered in an appropriate dosage and in a similar manner for the patients in the two groups. Was "success" observed and determined in the same way for both groups?

If you are not happy with the answers to these questions, all of the preceding numerical appraisals may become unimportant. No matter how statistically impressive, the results may be unacceptable because of their architectural flaws. The comparison may have been done with biases that destroy the scientific credibility of the results; or the results, even if scientifically credible, may not be pertinent for the particular kinds of patients you treat and the way you give the treatment.

1.1.5 Data Acquisition

The process of acquiring data involves two crucial activities: observation and classification. For clinical work, the observation process involves listening, looking, touching, smelling, and sometimes tasting. The observations are then described in various informal or formal ways. For example, the observer

might see a 5 mm. cutaneous red zone that blanches with pressure, surrounding a smaller darker-red zone that does not blanch. For classification, the observer chooses a category from an available taxonomy of cutaneous lesions. In this instance, the entity might be called a *petechia*. If the detailed description is not fully recorded, the entry of "petechia" may become the basic item of data. Analogously, a specimen of serum may be "observed" with a technologic process, and then "classified" as sodium, 120 meq/dl. Sometimes, the classification process may go a step further, to report the foregoing sodium value as *hyponatremia*.

Because the available data will have been acquired with diverse methods of observation and classification, these methods will need separate scientific attention beyond the basic plan of the research itself. What procedures (history taking, self-administered questionnaire, blood pressure measurements, laboratory tests, biopsy specimens, etc.) were used to make and record the basic observations that produced the raw data? How was each patient's original condition identified, and how was each post-therapeutic response observed and classified as success or no success? Was "success" defined according to achievement of normotension, or an arbitrary magnitude of reduction in blood pressure? What kind of "quality control" or criteria for classification were used to make the basic raw data trustworthy?

The answers to these questions may reveal that the basic data are too fallible or unsatisfactory to be accepted, even if all other elements of the research architecture seem satisfactory.

1.1.6 Data Processing

To be analyzed statistically, the categories of classification must be transformed into coded digits that become the entities receiving data processing. This last step in the activities is particularly vital in an era of electronic analysis. Because the transformed data become the basic entities that are processed, we need to know how well the transformation was done. What arrangements were used to convert the raw data into designated categories, to convert the categories into coded digits, and to convert those digits into magnetized disks or diverse other media that became the analyzed information? What mechanisms were used to check the accuracy of the conversions?

The transformation from raw data into processed data must be suitably evaluated to demonstrate that the collected basic information was correctly converted into the analyzed information.

1.2 Statistical and Nonstatistical Judgments

Of the six activities just cited, the last three involve no knowledge of mathematics, and they also have prime importance in the scientific evaluation. During the actual research, these three activities all occur before any statistical expressions are produced. The architectural structure of the research, the quality of the basic data, and the quality of the processed data are the fundamental scientific issues that underlie the statistical results. If the basic scientific structure and data are inadequate, the numbers that emerge as results will be unsatisfactory no matter how "significant" they may seem statistically. Because no mathematical talent is needed to judge those three fundamental components, an intelligent reader who recognizes their primacy will have won at least half the battle of statistical evaluation before it begins.

Many readers of the medical literature, however, may not recognize this crucial role of nonstatistical judgment, because they do not get past the first three statistical components. If the summary expressions and quantitative contrasts are presented in unfamiliar terms, such as an odds ratio or a multivariable coefficient of association, the reader may not understand what is being said. Even if the summaries and contrasts are readily understood, the reader may be baffled by the P values or confidence intervals used in the stochastic evaluations. Flustered or awed by these uncertainties, a medically oriented reader may not penetrate beyond the outside statistics to reach the inside place where enlightened common sense and scientific judgment are powerful and paramount.

Because this book is about statistics, it will emphasize the three specifically statistical aspects of evaluation. The three sets of scientific issues in architecture and data will be outlined only briefly; and readers who want to know more about them can find extensive discussions elsewhere.[1–3] Despite the statistical focus, however, most of the text relies on enlightened judgment rather than mathematical

reasoning. Once you have learned the strategy of the descriptive expressions used in the first two parts of the statistical evaluation, you will discover that their appraisal is usually an act of common sense. Except for some of the complicated multivariable descriptions that appear much later in the text, no mathematical prowess is needed to understand the descriptive statistical expressions used for summaries, contrasts, and simple associations. Only the third statistical activity — concerned with probabilities and other stochastic expressions — involves distinctively mathematical ideas; and many of them are presented here with modern approaches (such as permutation tests) that are much easier to understand than the traditional (parametric) theories used in most elementary instruction.

1.3 Arrangement of the Text

This book has been prepared for readers who have some form of "medical" interest. The interest can be clinical medicine, nursing, medical biology, epidemiology, dentistry, personal or public health, or health-care administration. All of the illustrations and examples are drawn from those fields, and the text has been written with the assumption that the reader is becoming (or has already become) knowledgeable about activities in those fields.

A major challenge for any writer on statistical topics is to keep things simple, without oversimplifying and without becoming too superficial. The achievement of these goals is particularly difficult if the writer wants to respect the basic intellectual traditions of both science and mathematics. In both fields, the traditions involve processes of assertion and documentation. In science, the assertion is called a *hypothesis*, and the documentation is called *supporting evidence*. In mathematics, the assertion is called a *theorem* or *operating principle*, and the documentation is called a *proof*.

To make things attractive or more palatable, however, many of the assertions in conventional statistical textbooks are presented without documentation. For example, a reader may be told, without explanation or justification, to divide something by $n - 1$, although intuition suggests that the division should be done with n. Many medical readers are delighted to accept the simple "recipes" and to avoid details of their justification. Other readers, however, may be distressed when the documentary traditions of both science and mathematics are violated by the absence of justifying evidence for assertions. If documentary details are included in an effort to avoid such distress, however, readers who want only "the beef" may become bored, confused by the complexity, or harassed by the struggle to understand the details.

The compromise solution for this dilemma is to use both approaches. The main body of the text has been kept relatively simple. Many of the sustaining mathematical explanations and proofs have been included, but they are relegated to the Appendixes in the back of the pertinent chapters. The additional details are thus available for readers who want them and readily skipped for those who do not.

1.4 A Note about References

The references for each chapter are numbered sequentially as they appear. At the end of each chapter, they are cited by first author and year of publication (and by other features that may be needed to avoid ambiguity). The complete references for all chapters are listed alphabetically at the end of the text; and each reference is accompanied by an indication of the chapter(s) in which it appeared.

For the first chapter, the references are as follows:

1. Feinstein, 1985; 2. Sackett, 1991; 3. Hulley, 2000.

1.5 A Note about "Exercises"

The list of references for each chapter is followed by a set of exercises that can be used as "homework." The end of the text contains a list of the official answers to many of the exercises, usually those with

odd numbers. The other answers are available in an "Instructor's Manual." You may not always agree that the answers are right or wrong, but they are "official."

1.6 A Note about Computation

Statistics books today may contain instructions and illustrations for exercises to be managed with the computer programs of a particular commercial system, such as BMDP, Excel, Minitab, SAS, or SPSS. Such exercises have been omitted here, for two reasons. First, the specific system chosen for the illustrations may differ from what is available to, or preferred by, individual readers. Second, and more importantly, your understanding and familiarity with the procedures will be greatly increased if you "get into their guts" and see exactly how the computations are done with an electronic hand calculator. You may want to use a computer program in the future, but learning to do the calculations yourself is a valuable introduction. Besides, unlike a computer, a hand calculator is easily portable and always accessible. Furthermore, the results obtained with the calculator can be used to check the results of the computer programs, which sometimes do the procedures erroneously because of wrong formulas or strategies.

Nevertheless, a few illustrations in the text show printouts from the SAS system, which happens to be the one most often used at my location.

Exercises

1.1 Six types of "evaluation" were described in this chapter. In which of those categories would you classify appraisals of the following statements? [All that is needed for each answer is a number from 1–6, corresponding to the six parts of Section 1.1]

1.1.1. "The controls are inadequate."

1.1.2. "The data were punched and verified."

1.1.3. "Statistical analyses of categorical data were performed with a chi-square test."

1.1.4. "Compared with non-potato eaters, potato eaters had a risk ratio of 2.4 for developing omphalosis."

1.1.5. "The patient had a normal glucose tolerance test."

1.1.6. "The trial was performed with double-blind procedures."

1.1.7. "Newborn babies were regarded as sick if the Apgar Score was ≤ 6."

1.1.8. "The rate of recurrent myocardial infarction was significantly lower in patients treated with excellitol than in the placebo group."

1.1.9. "The reports obtained in the interviews had an 85% agreement with what was noted in the medical records."

1.1.10. "The small confidence interval suggests that the difference cannot be large, despite the relatively small sample sizes."

1.1.11. "The compared treatments were assigned by randomization."

1.1.12. "We are distressed by the apparent slowing of the annual decline in infant mortality rates."

2

Formation, Expression, and Coding of Data

CONTENTS

During the 17th and 18th centuries, nations began assembling quantitative information, called *Political Arithmetic*, about their wealth. It was counted with economic data for imports, exports, and agriculture, and with demographic data for population census, births, and deaths. The people who collected and tabulated these descriptions for the state were called *statists*; and the items of information were called *statistics*.

Later in the 18th century, the royalty who amused themselves in gambling gave "grants" to develop ideas that could help guide the betting. The research produced a "calculus of probabilities" that became eventually applied beyond the world of gambling. The application occurred when smaller "samples" rather than the entire large population were studied to answer descriptive questions about regional statistics. The theory that had been developed for the probabilities of bets in gambling became an effective mechanism to make inferential decisions from the descriptive attributes found in the samples. Those theories and decisions thus brought together the two statistical worlds of description and probability, while also bringing P values, confidence intervals, and other mathematical inferences into modern "statistical analysis."

The descriptive origin of statistics is still retained as a job title, however, when "statisticians" collect and analyze data about sports, economics, and demography. The descriptive statistics can be examined by sports fans for the performances of teams or individual players; by economists for stock market indexes, gross national product, and trade imbalances; and by demographers for changes in geographic distribution of population and mortality rates. The descriptive origin of statistics is also the fundamental basis for all the numerical expressions and quantitative tabulations that appear as evidence in modern

biologic science. The evidence may often be analyzed with inferential "tests of significance," but the inferences are a secondary activity. The primary information is the descriptive numerical evidence.

The numbers come from even more basic elements, which have the same role in statistics that molecules have in biology. The basic molecular elements of statistics are items of data. To understand fundamental biologic structures, we need to know about molecules; to understand the fundamentals of statistics, we need to know about data.

In biology, most of the studied molecules exist in nature, but some of them are made by humans. No data, however, exist in nature. All items of data are artifacts produced when something has been observed and described. The observed entity can be a landscape, a person, a conversation, a set of noises, a specimen of tissue, a graphic tracing, or the events that occur in a person's life. The medical observations can be done by simple clinical examination or with technologic procedures. The description can be expressed in letters, symbols, numbers, or words; and the words can occupy a phrase, a sentence, a paragraph, or an entire book.

The scientific quality of basic observations and descriptions depends on whether the process is suitable, reproducible, and accurate. Is the score on a set of multiple-choice examination questions a suitable description of a person's intelligence? Would several clinicians, each taking a history from the same patient, emerge with the same collection of information? Would several histopathologists, reviewing the same specimen of tissue, all give the same reading? If serum cholesterol is measured in several different laboratories, would the results — even if similar — agree with a measurement performed by the National Bureau of Standards?

2.1 Scientific Quality of Data

Suitability, reproducibility, and accuracy are the attributes of scientific quality. For *reproducibility*, the same result should be obtained consistently when the measurement process is repeated. For *accuracy*, the result should be similar to what is obtained with a "gold-standard" measurement. For *suitability*, which is sometimes called *sensibility* or *face validity*, the measurement process and its result should be appropriate according to both scientific and ordinary standards of "common sense."

These three attributes determine whether the raw data are trustworthy enough to receive serious attention when converted into statistics, statistical analyses, and subsequent conclusions. Of the three attributes, *accuracy* may often be difficult or impossible to check, because a "gold standard" may not exist for the definitive measurement of such entities as pain, discomfort, or gratification. *Reproducibility* can always be checked, however. Even if it was not specifically tested in the original work, a reader can get an excellent idea about reproducibility by noting the guidelines or criteria used for pertinent decisions. *Suitability* can also be checked if the reader thinks about it and knows enough about the subject matter to apply enlightened common sense.

The foregoing comments should demonstrate that the production of trustworthy data is a scientific rather than statistical challenge. The challenge requires scientific attention to the purpose of the observations, the setting in which they occurred, the way they were made, the process that transformed observed phenomena into descriptive expressions, the people who were included or excluded in the observed groups, and many other considerations that are issues in science rather than mathematics.

These issues in scientific architecture are the basic "molecular" elements that lie behind the assembled statistics. The issues have paramount importance whenever statistical work is done or evaluated — but the issues themselves are not an inherent part of the statistical activities. The data constitute the basic scientific evidence available as "news"; statistical procedures help provide summaries of the news and help lead to the "editorials" and other conclusions.

To give adequate attention to what makes the basic data scientifically trustworthy and credible, however, would require too many digressions from the statistical procedures. A reader who wants to learn mainly about the statistics would become distressed by the constant diversions into scientific priorities. Therefore, to allow the statistical discourse to proceed, many of the fundamental scientific issues receive little or no attention in this text. This neglect does not alter their primacy, but relies on

the assumption that they will be known, recognized, and suitably considered by readers whose minds are liberated from mathematical confusion.

For the statistical discussion, the raw descriptive information and groups will be accepted here as satisfactory scientific evidence. This acceptance, an *ad hoc* literary convenience, is never warranted in the real world, where scientific credibility leads to the most important, but commonly overlooked, problems in statistical analyses. In fundamental scientific thinking, the focus is on what the information represents, how it got there, and how good it is. Accepting the underlying scientific process as adequate, however, we can turn to the statistical process discussed in this and the subsequent 27 chapters. The next level of thinking begins with conversion of the raw data to statistics.

2.2 Formation of Variables

The available raw information can be called *data*, but the data can seldom be analyzed statistically. For example, statements such as "John Doe is a 37-year-old man" and "Mary Smith is a 31-year-old woman" contain specific data about specific persons, but the information is not arranged in a manner suitable for statistical analysis. To get the data into an appropriate format, the first statement can be converted to "*Name*: **John Doe**; *Age in years*: **37;** *Sex*: **male**." The second statement would become "*Name*: **Mary Smith**; *Age in years*: **31;** *Sex*: **female**.

In the format just cited, specific attributes such as *name*, *age*, and *sex* were chosen as labels for each citation. Analogous attributes and formats would be needed to express details of the medical information contained in a patient's history, an account of a surgical operation, or a description of a pathologic specimen. Many other types of raw information, however, receive expressions that are immediately suitable for analysis. Examples are the laboratory report, *hematocrit*: **47%**, or the clinical description of *systolic blood pressure* as **125 mm Hg**.

Each of these formats refers to a selected attribute, cited in an available set of expressions. In the parlance of mathematics or statistics, the attributes are called *variables*, and the expressions available for citation are called a *scale*.

2.2.1 Definition of a Variable

A *variable* is a class of data in which different distinctions can be expressed. A person's *age*, *height*, *sex*, *diagnosis*, and *therapy* are all variables. Their distinctions might be cited respectively for a particular person as **52 years, 69 inches, male, peptic ulcer,** and **antacid tablets**. Although not particularly desirable as a scientific term, the word *variable* has become firmly entrenched in its mathematical role. You can think of it not as something that denotes diversity or change — in the sense of *variation* or *variability* — but simply as the name used for a class of data that can be cited differently for different people.

The expression of eight variables for six individual persons could be shown in a "matrix" table that has the following skeleton structure:

	Names of Variables								
Identity of Person	**Age**	**Height**	**Sex**	**Diagnosis**	**Therapy**	**Systolic Blood Pressure Before Treatment**	**Systolic Blood Pressure During Treatment**	**Systolic Blood Pressure After Treatment**	
A.B.									
C.D.									
E.F.									
G.H.									
I.J.									
K.L.									

The interior "cells" of this table would contain the assembled citations of each variable for each person. (Exercises will use different ways and the additional categories for analyzing these data.)

2.2.2 Scales, Categories, and Values

The *scale* of a variable contains the available *categories* for its expression. The scale for the variable *sex* usually has two categories: **male** and **female**. The scale for *age in years* has the categories **1, 2, 3, 4, …, 99, 100, … .** (In statistics, as in literature, the symbol "…" indicates that certain items in a sequence have been omitted.)

For a particular person, the pertinent category of a scale is called the *value* of the variable. Thus, a 52-year-old man with peptic ulcer has 52 as the value of *age in years*, **male** as the value of *sex*, and **peptic ulcer** as the value of *diagnosis*. The word *value* is another entrenched mathematical term that has nothing to do with judgments or beliefs about such "values" as importance, worth, or merit. The value of a variable is the *result* of an observational process that assigns to a particular person the appropriate category of the variable's scale.

Any descriptive account of a person can be converted into an organized array of data using variables, scales, categories, and values.

2.3 Classification of Scales and Variables

Scales and variables can be classified in various ways. The most common and useful classification depends on the precision of ranking for the constituent categories.

2.3.1 Precision of Rankings

A 31-year-old person is 14 years younger than someone who is 45. A person with severe dyspnea is more short of breath than someone with mild dyspnea, but we cannot measure the exact difference. In both these examples, definite ranks of magnitude were present in the values of **31** and **45** for *age*, and in the values of **severe** and **mild** for *severity* of dyspnea. The ranks were distinct for both variables, but the magnitudes were more precise for *age* than for *severity of dyspnea.*

Certain other variables, however, are expressed in categories that have no magnitudes and cannot be ranked. Thus, no obvious rankings seem possible if *history of myocardial infarction* is **present** in one patient and **absent** in another; or if one patient has an **anterior** and another has a **posterior** *location of myocardial infarction*. We might want to regard **present** as being "more" than absent, but we cannot rank the magnitude of such locations as **anterior** or **posterior**.

The four examples just cited illustrate patterns of precision in ranking for the *dimensional*, *ordinal*, *binary*, and *nominal* variables that were used, respectively, to denote age, severity, existence, and location. These four patterns, together with *quasi-dimensional* scales, are the basic arrangements for categories in the scales of variables. The patterns are further discussed in the sections that follow.

2.3.1.1 Dimensional Scales — In a dimensional scale, the successive categories are monotonic and equi-interval. In the directional sequence of a monotonic ranking, each category is progressively either greater or smaller than the preceding adjacent category. For equi-interval ranks, a measurably equal interval can be demarcated between any two adjacent monotonic categories.

Thus, in the scale for *age in years*, a measurably equal interval of 1 year separates each of the successive categories **1, 2, 3, 4, …, 99, 100**, **… .** Similarly, for the variable *height in inches*, each of the successive categories **…, 59, 60, 61, 62, …, 74, 75, …** has an incremental interval of 1 inch.

Many alternative terms have been used as names for a dimensional scale, which is now the traditional form of scientific measurement. Psychologists and sociologists often refer to *interval scales*, but the

word *interval* is often used medically for a period of time. The term *metric* might be a satisfactory name, but it regularly connotes a particular system of measurement (in meters, liters, etc.).

Mathematicians sometimes talk about *continuous* scales, but many dimensional categories cannot be divided into the smaller and smaller units that occur in continuous variables. For example, *age* and *height* are continuous variables. We could express age in finer and finer units such as years, months, days, hours, seconds, and fractions of seconds since birth. Similarly, with a suitably precise measuring system, we could express height not merely in inches but also in tenths, hundredths, thousandths, or millionths of an inch. On the other hand, *number of children* or *highest grade completed in school* are dimensional variables that are discrete rather than continuous. Their scale of successive integers has equi-interval characteristics, but the integers cannot be reduced to smaller units.

Psychologists sometimes use *ratio scale* for a dimensional scale that has an absolute zero point, allowing ratio comparisons of the categories. Thus, *age in years* has a ratio scale: a 24-year-old person is twice as old as someone who is 12. *Fahrenheit temperature* does not have a ratio scale: 68°F is not twice as warm as 34°F.

Although these distinctions are sometimes regarded as important,[1] they can generally be ignored. Any type of scale that has equi-interval monotonic categories can be called *dimensional*.

2.3.1.2 Ordinal Scales —

In an ordinal scale, the successive categories can be ranked monotonically, but the ranks have arbitrary magnitudes, without measurably equal intervals between every two adjacent categories.

Clinicians constantly use ordinal scales to express such variables as *briskness of reflexes* in the graded categories of **0, 1+, 2+, 3+, 4+.** *Severity of pain* is a variable often cited as **none, mild, moderate,** or **severe**. Although *age* can be expressed in dimensional data, it is sometimes converted to an ordinal scale with citations such as **neonatal, infant, child, adolescent, young adult, middle-aged adult, … .**

An ordinal scale can have either *unlimited ranks* or a *limited* number of *grades*. Most ordinal scales in medical activities have a finite group of grades, such as **0, 1+, …, 4+** or **none, mild, …, severe.** If we wanted to rank the people who have applied for admission to a medical school, however, we could use an unlimited-rank scale to arrange the applicants with ratings such as **1, 2, 3, 4, 5, …, 147, 148, … .** In this limitless scale, the lowest ranked person might be rated as **238** or **964**, according to the number of applicants. Scales with unlimited ranks seldom appear in medical research, but have been used (as discussed much later) for the mathematical reasoning with which certain types of statistical tests were developed.

2.3.1.3 Quasi-Dimensional Scales —

A quasi-dimensional scale seems to be dimensional, but does not really have measurably equal intervals between categories. Quasi-dimensional scales can be formed in two ways. In one technique, the scale is the sum of arbitrary ratings from several ordinal scales. For example, a licensure examination might contain 50 questions, for which each answer is regarded as a separate variable and scored as **0** for *wrong*, **1** for *partially correct*, and **2** for *completely correct*. Despite the arbitrary ordinal values, which have none of the equi-interval characteristics of dimensional data, the candidate's scores on each question can be added to form a total score, such as **46, 78, 85,** or **100**. The arbitrary result looks dimensional and is often manipulated mathematically as though it were truly dimensional.

A second source of quasi-dimensional data is a graphic rating technique called a *visual analog scale*. The respondent rates the magnitude of a feeling, opinion, or attitude by placing a mark on a line that is usually 100 mm. long. The measured location of the mark then becomes converted to an apparently dimensional rating. For example, someone might be asked to mark the following line in response to a question such as "How bad is your pain?"