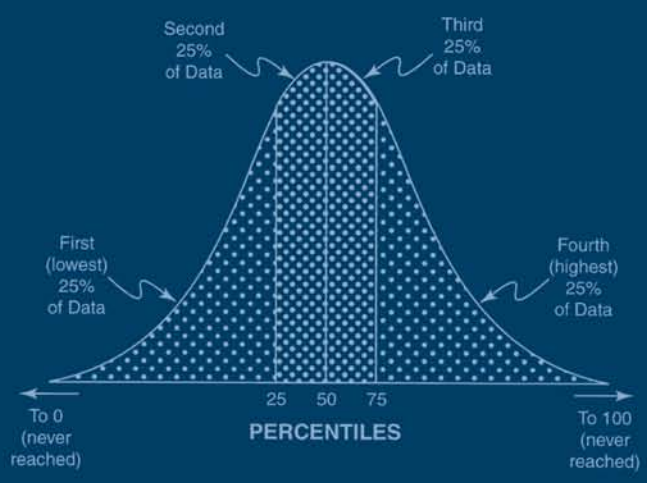
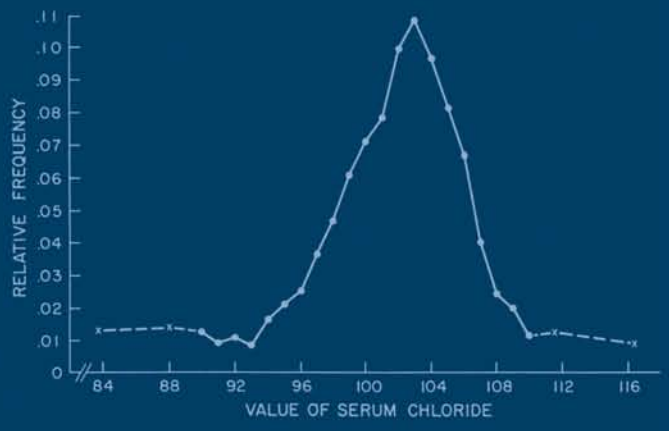
PRINCIPLES OF MEDICAL STATISTICS





Alvan R. Feinstein, M.D.

CHAPMAN & HALL/CRC

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

Part III Evaluating Associations

Part IV Additional Activities

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?"

If the mark is placed as follows,

the measured distance could represent 67 "pain units." Despite the dimensional expression, such scales are not truly dimensional because adjacent categories — such as **70, 71,** and **72** — in the arbitrary graphic ratings are not accompanied by criteria that demarcate equal intervals.

2.3.1.4 Binary (Dichotomous) Scales — A binary or dichotomous scale has two categories, such as **male** and **female** for the variable *sex*. Other common scales with two categories are **alive/dead, success/failure,** and **case/control.** In medical research, many binary scales refer to entities, such as chest pain, that can be classified as **present** or **absent**, or to a variable such as *previous pregnancy* that can be cited as **yes** or **no**. When used to represent existence or nonexistence, a binary scale is sometimes called an *existential scale*.

Although apparently binary, existential data can often be regarded as a subset of ordinal data. The ordinal ranking becomes apparent if gradations of probability are added to the scale. Thus, when *existence of acute myocardial infarction* is expressed as likelihood of existence, the available scale is ordinally expanded to include the categories of **definitely present, probably present, uncertain whether present or absent, probably absent**, and **definitely absent.**

2.3.1.5 Nominal Scales — In a nominal scale, the unranked categories have no magnitudes. Nominal scales would express such variables as a person's name, address, color of eyes, birthplace, religion, diagnosis, or type of therapy.

Nominal characteristics can sometimes receive implicit ranks based on cultural, social, political, or even clinical preferences. Thus, a short name might be preferred to a long one; certain addresses might be regarded as being in better neighborhoods than others; someone might like blue eyes better than brown (or vice versa); a person raised in Alabama might be more accustomed to social patterns in Mississippi than someone from Maine; and a diagnosis of acute pancreatitis might be more desirable than pancreatic carcinoma. When such rankings occur, however, the variable is no longer nominal and needs a different title for its expression. For example, if the variable is designated as *desirability of diagnosis* rather than *diagnosis*, the previously nominal categories **acute pancreatitis, gallstone obstructive jaundice,** and **pancreatic carcinoma** might become an ordinal scale.

For analytic purposes, the individual categories of a nominal scale are often decomposed and converted into a set of "dummy" binary variables, each rated as **present** or **absent.** For example, suppose *current occupation* is expressed in the nominal categories **doctor, merchant,** or **other**. The categories can be converted into three binary variables: *occupation as a doctor*, *occupation as a merchant*, and *occupation as neither a doctor nor a merchant*. (The three variables could be suitably condensed into only the first two, because someone whose values are **no** for both *occupation as a doctor* and *occupation as a merchant* would have to be rated as **yes** for *occupation as neither a doctor nor a merchant*.)

2.3.2 Other Forms of Nomenclature

Because of their fundamental roles in scientific information, the five types of scales and data have received considerable attention from statisticians and psychosocial scientists, who have labeled the scales with many different titles. For dimensional data, we have already met the terms *interval* and *ratio*, which are used by psychologists, and *continuous*, used by statisticians. To distinguish the arbitrary nondimensional categories, ordinal, binary, or nominal data are sometimes called *attribute data*, *categorical data*, or *discrete data* (although the last term is ambiguous, because integer dimensions are also discrete).

Being readily added, multiplied, and otherwise manipulated mathematically, dimensional data are sometimes called *quantitative data*, whereas categorical data, which can only be enumerated and cited as frequency counts, are sometimes called *qualitative* or *quantal data*.

For practical purposes as well as simplicity, we can escape all of the possible jargon by using five main terms: *dimensional*, *quasi-dimensional*, *ordinal*, *binary*, and *nominal*. To indicate the arbitrary categories of non-dimensional scales, they are sometimes here called *categorical*, although the term *non-dimensional* will usually convey the desired distinction.

2.4 Multi-Component Variables

Variables can also be classified according to their number of components and timing of component states. Regardless of the type of scale, a single variable can contain one component or more than one. A *simple* variable has only one main component; a *multi-component* or *composite* variable has more than one. A variable can also refer to only one stated condition, or to two or more states.

2.4.1 Composite Variables

A composite variable contains a combination of data from two or more component variables. Each component variable, expressed in its own scale, is "input" that is aggregated to form a separate "output" scale. For example, the *Apgar Score* for the condition of a newborn baby is a composite variable. Its output scale, ranging from **0** to **10**, is a sum of ratings for five simple variables — representing *heart rate, respiration, skin color*, *muscle tone*, and *reflex response* — each cited in its input scale of **0, 1,** or **2**. Although the aggregation can be an "additive score," some composite variables are formed as "Boolean clusters." A Boolean cluster contains a combination of categories joined in logical unions or intersections. For example, the TNM staging system for cancer contains categorical ratings for three component variables representing *Tumor, Nodes*, and *Metastases*. The individual ratings, which might be cited as **T5NlM0** or **T2N0M2**, are often combined into an ordinal set of clustered categories called *stages*. In one common pattern, **Stage III** represents patients who have distant metastases, regardless of the ratings for tumor and regional nodes. **Stage II** represents patients with involvement of regional nodes *and* no evidence of distant metastases, regardless of the ratings for tumor. **Stage I** represents patients with no evidence of either regional node or distant metastatic involvement.

A composite variable can also be formed from dimensional components. For example, for the variable called *anion gap*, the sum of serum chloride and bicarbonate concentrations is subtracted from the serum sodium concentration. The *Quetelet index* is the quotient of *weight* (in kg.) divided by the square of *height* (in cm.).

Composite variables have many different names. They are sometimes called *indexes, factors, stages, systems, classes, scales, ratings, criteria, multi-dimensional scales*, or *clinimetric indexes*. Although composite variables are usually expressed in dimensional or ordinal scales, some of the most complex constructions have binary existential citations. They occur for variables formed as diagnostic criteria for presence or absence of a particular disease. Thus, in the clinical diagnostic criteria for tuberculosis, rheumatic fever, rheumatoid arthritis, or myocardial infarction, the output scale is a simple binary expression such as **yes** or **no** (or **present** or **absent**). The input, however, consists of multiple variables, arranged in complex patterns, that often form intermediate axes or "subscales" before everything is ultimately aggregated into the binary output rating.

2.4.2 Component States

All of the variables described thus far refer to the *single-state* condition of a particular person at a single point in time. A single variable, however, can also express the comparison of results for two or more states. They can be the single-state condition of two persons, the change noted in a person's single state on several occasions, or descriptions of the same state by several independent observers.

In these situations, the comparative result is often cited as a separate variable, having its own scale for denoting an increment, decrement, or some other comparative distinction in the linked values. For example, if **70 kg** and **68 kg** are the values for two people "matched" in a particular study of weight, the difference between the two single-state values could be cited in a separate variable as **+2 kg** (or **–2 kg**). Alternatively, using a separate ordinal scale, one person could be regarded as **slightly** heavier than the other. In another study, a particular person's single-state values of serum cholesterol might have been **261** before treatment and **218** afterward. The change could be recorded in a separate variable as *fall in cholesterol*: **43**. In a study of observer variability, the S-T wave depression in the same electrocardiogram might be reported as **2.0 mm** by one electrocardiographer and as **2.4 mm** by another. The difference in the two measurements could be expressed in a separate variable as **0.4 mm**.

For the three foregoing examples, the value of the additional variable was calculated from values of the original single-state variables. In other circumstances, however, the change can be rated directly, without specific reference to single-state values. For example, a patient might express *response to treatment* in a rating scale such as **excellent, good, fair, poor** or in specifically comparative terms such as **much better, somewhat better, same, somewhat worse, much worse.** When a comparative change is cited without reference to single-state components, the expression is called a *transition variable*.

The total effect of more than two states in the same person often cannot be expressed in simple comparative terms, such as **larger, smaller,** or **same.** The expressions may therefore describe the *trend* in the entire series of values. Thus, the pattern of successive blood pressure measurements for a particular person might be cited in a scale such as **rising, stable,** or **falling.** Sometimes the general pattern of trend is categorized for its *average value, desirability,* or *diagnostic features*. For example, a set of successive daily values for post-therapeutic pain might be cited for their mean value, or in ratings of the total *response* as **excellent, good, …, poor.** The set of successive dimensional values in a glucose tolerance curve might be classified as **normal, diabetic,** or **hypoglycemic**.

2.4.3 Scales for Multi-Component Variables

Regardless of how the components are expressed and assembled, multi-component variables are cited in the same types of dimensional, ordinal, binary, or other scales used for single-component variables. Thus, an *Apgar Score* of **10**, a *TNM Stage* of **III**, an *anion gap* of **9**, a *diagnosis* of **acute myocardial infarction**, or a *post-treatment response* of **much improved** could all come from multi-component variables, but could all be analyzed as values in individual variables.

To evaluate the scientific quality of the multi-component expression, however, we would need to consider not only the quality of each component, but also the suitability and effectiveness of the way they have been put together. For example, suppose a composite scale to indicate *satisfaction with life* contained certain weighted ratings for age, occupation, and income. Each of the components might be scientifically measured, but the combination would be an unsatisfactory expression for *satisfaction with life*. As another example, each of the components of the Apgar score might be suitably chosen and appraised, but their aggregate would be unsuitable if reflex responses were rated on a scale of **0, 5, 10** while the other four variables were rated as **0, 1, 2.**

2.5 Problems and Customs in "Precision"

The idea of *precision* is a fundamental concept in both science and statistics, but the concept is used with different meanings in the two domains. Furthermore, the attribute of *numerical precision* is different from ordinary *precision* and is usually a matter of arbitrary custom.

2.5.1 Concepts of Precision

Scientists use *precision* to denote "the quality of being exactly or sharply defined or stated."[2] Thus, the phrase "cachectic, dyspneic, anemic old man" is more precise than the phrase "sick patient." The

numerical expression **3.14159** is a more precise value for π than **3.1**. Statisticians, however, use *precision* to refer to "the way in which repeated observations" conform to themselves and … to the dispersion of the observations."[3] In this idea, a series of repeated measurements of the same entity, or individual measurements of multiple entities, is statistically precise if the spread of values is small relative to the average value. To a scientist, therefore, precision refers to increased detail and is an attribute of a *single* measurement. To a statistician, precision refers to a small spread or range in a *group* of measurements.

This fundamental difference in concepts may cause confusion when scientists and statisticians communicate with one another, using the same word for two different ideas. Furthermore, the difference in concepts leads to different goals in measurement. A prominent aim of scientific measurement is to increase precision by increasing details. Statistics does not have a word to indicate "detail," however, and many statistical activities are deliberately aimed at "data reduction,"[4] which helps eliminate details.

Doing studies of quality control or observer variability in a measurement process, regardless of its precision in detail, the scientist may be surprised to discover that excellent *reproducibility* for the process is regarded by the statistician as *precision*. The scientist may also be chagrined to find that the hard work of developing instruments to measure with high precision may be regarded statistically as producing excessive details to be reduced.

The two sets of concepts can readily be reconciled by recognizing that scientific *precision* refers to individual items of data, whereas statistical *precision* refers to a group of data. In trying to understand or summarize meaning for a group of data, both the scientist and the statistician will reduce the amount of detail conveyed by the *collection* of individual values. This reduction occurs with the formation of statistical indexes that denote the average value, spread, or other attributes of the collective group. Before or while the reduction occurs, however, the individual items of data will require certain policies or customs about the management of numerical precision.

2.5.2 Strategies in Numerical Precision

The idea of numerical precision involves concepts and customs for "significant figures" or "significant digits." In the decimal system of notation, any number can be converted to a standard numerical expression by moving the decimal point so that the number is cited as a value between 0 and 10, multiplied by a power of 10. Thus, .01072 becomes 1.072×10^{-2} and 1072 becomes 1.072×10^{3}. With this convention, a significant figure occurs at the left of the decimal point and the other significant figures occur at the right. The number 15,000 becomes 1.5000×10^{4} and 14,999 becomes 1.4999×10^{4}. Each number has five significant figures.

A prominent challenge in data reduction — at the level of individual values of data — is the decision about how many significant figures (or digits) to retain in each number. The decision will depend on how the number was obtained and how it will be used.

2.5.2.1 Production of Numbers — Numbers can be produced by measurement or by calculation, but there are two forms of measurement: mensuration and enumeration.

2.5.2.1.1 Mensuration. This unfamiliar term refers to the most familiar form of scientific measurement: identifying the magnitude of an entity on a calibrated scale. *Mensuration* is used to distinguish this type of measurement from the other type, *enumeration*, which is discussed in the next subsection. In the customary form of mensuration, the scale contains dimensional values, and the number is a measured amount, such as **137** meq/dl for serum sodium concentration or **15.2** units of hemoglobin. With more precise systems of measurement and calibration, sodium might be determined as **137.419** or hemoglobin as **15.23**, but such systems are rarely used because the extra precision is seldom necessary or helpful.

Because precision will depend on the "refinement" of the dimensional scale, decisions must be made about how to report results (i.e., how many significant figures to include) that lie between the two finest units of calibration. For example, if an ordinary ruler is marked at intervals of 1 mm, magnitudes that lie between two marks are often cited at the midpoint and expressed as 1.5 mm., 2.5 mm., etc.

Note that mensurations do not always produce dimensional numbers. With an ordinal, binary, or nominal scale, the result of the mensuration process may be values such as **3+, yes**, or **hepatitis**. The foregoing discussion of mensuration, however, was confined to dimensional numbers.

2.5.2.1.2 Enumeration. This fancy word for counting is often used because it refers to both the process and result and, therefore, offers a single term for the two words, *frequency count*. An enumeration has unlimited precision because the number of significant digits continues to increase as more entities are counted. For example, the census bureau can report counts of 1873 or 2,194,876 people in different regions.

2.5.2.1.3 Calculation. Most of the main numbers used in statistics are calculated from the basic results obtained with mensuration and/or enumeration. For example, as noted later, a *mean* is constructed when a sum of mensurations or enumerations is divided by another enumeration.

During the calculations, numerical precision can be altered in ways that no longer reflect the original precision of measurement. Addition and multiplication may increase the original number of significant digits; subtraction can decrease them; and division can produce either increases or decreases. For example, if one hematocrit is measured with high precision as **23.091** and another as **46,** the sum of **69.091** suggests that both measurements have 5 significant digits. Multiplication of the two numbers would produce the 7 significant digits of **1062.186.** For the three-digit serum calcium values of **11.3** and **10.9,** subtraction loses two digits, yielding **.4.** With division, the ratio of the two-digit ages **84** and **42** produces the one-digit **2,** but the ratio of **56/32** produces the three-digit **1.75.** In some instances, calculation can produce a "repeating" or "unending" decimal having as many significant digits as desired. Thus, 2/3 = .66666666… and 1/22 = .0454545… .

2.5.2.2 Transformations — Some variables are regularly transformed either as raw data or for analytic purposes. The raw-data transformations, often done for scientific custom, can express weight as **kg** rather than **lb**, height in **cm** rather than **in.**, and temperature in **°C** rather than **°F**. Many chemical concentrations today are transformed into standard international units from their original measurements in **mgm** or **Gm.** Other raw-data transformations, done for scientific convenience, will express hydrogen ion concentration as its negative logarithm, called **pH**.

In yet other transformations, used for analytic purposes discussed later, values of *survival time in months* or *bacterial counts* might be converted to logarithms. A series of dimensional values for *hemoglobin* might be compressed into an ordinal array such as **≤8, 9–11, 12–13, 14–15, 16–17,** and **≥18.** A series of values for *serum calcium* might be compressed into the categorical zones **low, normal,** and **high**. These transformations, however, usually occur during the statistical analyses, not in the original expression of the data.

2.5.2.3 Customs in Management — Certain customs have been established to achieve consistency in managing the diverse possibilities for computational alteration of numerical precision. Each custom depends on whether the number is being understood, reported, or formally calculated.

2.5.2.3.1 Understanding. For understanding what the numbers imply, and for doing certain mental or in-the-head calculations that can facilitate comprehension, most people are accustomed to using two significant digits. For example, if two percentages are reported as 49.63% and 24.87%, you may not immediately perceive their comparative relationship. The relationship is promptly apparent, however, when the numbers are "rounded" to 50% and 25%. Because most people have become accustomed to dealing with percentages, which range from 0 to 100 while usually having two digits, the process of comprehension usually requires no more than *two* significant digits.

Furthermore, the digits are usually most rapidly understood when presented in a range of 0 to 100. Thus, the proportions .50 and .25 are easier to compare when expressed in percentages as 50% and 25%. This custom is responsible for the frequency with which proportions are multiplied by 100 and expressed as percentages. The custom is also responsible for the epidemiologic tactic, discussed later, of expressing

populational rates in unit values per 10,000 or per 100,000. Thus, an occurrence rate of 217/86,034 might be cited as 25.2×10^{-4} or 252 per hundred thousand rather than .00252.

(Certain tactics in prompt understanding will work with "convenient" numbers rather than significant digits. For example, if you see a number such as $\sqrt{157}$, you can almost immediately recognize that the result lies between 12 and 13 because $12^2 = 144$ and $13^2 = 169$.)

2.5.2.3.2 Reporting. In reporting numbers for readers, three significant digits are usually listed for measured amounts, and two digits for most percentages. The "rules" may be altered, however, if they produce ambiguous discriminations or misleading impressions. For example, two baseball players with "rounded" averages of .333 after the last day of the season may be tied for the batting championship. The winner, determined with extra decimal places, will be the batter with 216/648 = .3333 rather than 217/652 = .3328. Among a series of percentages, one member of the group may be listed as 0% because the proportion 2/483 = .00414 was rounded to 0 when converted to integers between 0 and 100. To avoid confusion between this 0 and another that arises as 0/512, the first one can be cited as 0.4%.

An unresolved custom in reporting is whether to use a "leading" **0** before the decimal point in numbers such as **.183.** Frequently seen expressions such as **P < .05** need not be reported as **P < 0.05,** but some writers believe that the leading zero, as in **0.183**, helps improve clarity in both typing and reporting. Other writers believe the leading **0** is a waste of space. In computer or other printouts of integer digits, numbers such as 01 or 08 look peculiar, but are increasingly used and accepted. (In this text, the leading zeros will be omitted except when needed to avoid ambiguity.)

2.5.2.3.3 Calculating. In contrast to the truncated two or three digits used for understanding and reporting, calculations should always be done with all pertinent digits retained during the calculation. Any truncation (or "rounding") of "trailing" digits on the right should be reserved until everything is done, when the final result is reported. The rule to follow is "Round at the end." Keep all numbers as intact as possible throughout the calculation and save any rounding until all the work is completed.

The rule is particularly important if you use an electronic hand calculator that can store no more than 9 digits. Many personal digital computers work in "single precision" with 6 digits or in "double precision" with 12. In the latter situation, or when a mainframe computer works with 24 significant digits in "extended precision," the computer seldom has problems in rounding. The reader of the printout may then be suffused with an excess of trailing digits, however, when the computer displays results such as ".04545454545" for divisions such as 1/22.

2.5.3 Rounding

Many people in the world of medicine learned about mathematical *rounding* long before the same word was applied in patient care. In mathematics, rounding consists of eliminating trailing digits at the right end of a number. The excess digits are those beyond the stipulated boundary of two, three, or sometimes more "significant figures."

With simple truncation, the trailing digits are chopped off directly with no "adjustment" of previous digits. With rounding, the previous adjacent digits are adjusted; and the adjustment may involve more than one digit. Thus, **19.09** might be rounded to **19** for two digits and **19.1** for three, but **19.99** would become **20** or **20.0.**

The rounding process is simple and straightforward if the terminal digits are between 0–4 and 6–9. If <5, the candidate digit is dropped without further action. If the candidate is >5, the preceding digit is incremented one unit. Thus, 42.7 becomes 43; 42.4 becomes 42; and 5378 becomes 5380 (or 5400 with more drastic rounding).

The only pertinent problem is what to do when the terminal digit is exactly 5 or has the sequence of 5000… . In many branches of science (and in most hand-held calculators), the rule is simple: If the last digit is ≥5, round upward. With this rule, 42.5 would become 43. A statistical argument has been offered, however, that the strategy is biased. Of the ten candidate digits, four (1, 2, 3, 4) are rounded down; four (6, 7, 8, 9) are rounded up; and one (0) has no effect. Therefore, a bias toward higher values will occur if all terminal 5's are rounded upward. To avoid this problem, terminal 5's are rounded to the

preceding even digit. Thus, 17.5, having an odd digit before the 5, is rounded upward to 18; but 16.5, with an even digit before the 5, is rounded down to 16. With this tactic, the *preceding* digits of 0, 2, 4, 6, 8 are left intact, but 1, 3, 5, 7, 9 are rounded upward.

You may now want to argue that the procedure produces a bias toward getting *even* terminal digits. This bias might be a problem for betting in roulette, but should not be a difficulty in quantitative calculations. In fact, the rule would work just as well and the results would be just as "unbiased" if rounding were aimed toward a preceding odd rather than even digit. The *even* direction is probably maintained because rounding toward an even number is easier to remember than rounding "odd." If you (or your calculator) happen to neglect the rule, and if terminal 5's are routinely rounded upward, however, no great disasters are likely to occur.

Bear in mind that the round-even strategy applies only when the terminal nonzero digit is 5 (or 50000…). Otherwise, use the simple rule of rounding up for > 5 and down for < 5.

Like any other set of rules, the foregoing recommendations have exceptions that occur for discrimination, identification, and decrementation.

2.5.3.1 Discrimination — If the relative magnitude of two numbers must be discriminated, rounding can be counterproductive, as in the previous illustration of ranking baseball players. If the extra digits are needed to discriminate distinctions, do not round.

2.5.3.2 Identification — Among a series of percentages ranging from 1% to 99%, one member of the group may have the proportion 2/483 = 0.4%. If listed with no more than two integer digits, this result would be a misleading 0%. To avoid the problem, the rule can be "bent" for an exception listing the result as 0.4%.

2.5.3.3 Decrementation — Trailing digits are particularly important when two numbers are subtracted, because the size of the decrement may be determined entirely from distinctions in last few digits. A disastrous statistical syndrome, called *premature rounding*, occurs when trailing digits are "ejaculated" too soon. The malady can occur with hand-calculator determination of an entity called *group variance*, which precedes the determination of a standard deviation in Chapter 4.

2.6 Tallying

Despite the advances of electronic computation, you may sometimes (or often) find yourself doing enumeration by counting the frequency of data in a set of nominal, ordinal, or binary categories. Doing it manually can save a great deal of effort if you have a relatively small amount of data. By the time everything gets punched and verified for electronic processing, the manual tally can be long done, checked, and tabulated. For example, suppose a questionnaire returned by 150 people contains ratings of 1, 2, 3, 4, or 5 for each of 4 statements, A, B, C, and D. You want to know and summarize the distribution of frequency counts for the ratings. The rapid manual way to do this job is to prepare a "tally box" or "tally table" that has the skeleton shown in Figure 2.1.

The next step is to go through the questionnaires, one at a time, putting an appropriate tally in the columns for the rating given to A, B, C, and D by each person. With a convenient method of tallying, the frequency counts are evident as soon as you have completed the last questionnaire.

2.6.1 Conventional Methods

The conventional method of tallying uses a vertical slash or up-and-down mark for each of four tallies, then a diagonal mark through the fifth. The first two marks would be ||, the first four would be ||||; the fifth would be 𝍸. With this tactic, the tally 𝍸 𝍸 𝍸 ||| would promptly be identified as 18.

	1	2	3	4	5
A					
B					
C					
D					

FIGURE 2.1
Illustration of a "Tally box" with five possible ratings for four statements.

Complete ' cheng'	Writing Sequence (1)	(2)	(3)	(4)	(5)
正	一	丅	下	𠂇	正

The Chinese Character Cheng

Age in Years	Tally Mark	Frequency in Group
5	正 下	8
6	正 正 丅	12
7	正 正 正 一	16
8	正 正 正 丅	17
9	正 正 丅	12
10	正	5
Total		70

FIGURE 2.2
Alternative approach for tallying. [Taken from Chapter Reference 5.].

2.6.2 Alternative Methods

An alternative way of bunching groups of five, proposed by K. H. Hsieh,[5] is the Chinese tally count, which is based on the five strokes used in the Chinese character *cheng*. According to Hsieh, the result is "neater and easier to read than a series of slashes"; and because "an incomplete character can be identified at a glance," the method avoids having "to count carefully to see if four slashes have been recorded" before crossing the fifth. The upper part of Figure 2.2 shows the writing sequence for forming *cheng*, and the lower part illustrates its application in a tally table.

John Tukey[6] has proposed a system of dots and lines that shows 10, rather than 5, for each completed visual group. In the Tukey system, : : is 4, then lines are drawn around the periphery to form a box (so that ⊓ is 7). Diagonal lines are then drawn so that ⊠ is 10.

Any of the tally methods will work well if you become familiar and comfortable with it, and use it carefully. The Tukey method has the advantage of saving space if only a small area is available for the tally box.

References

1. Stevens, 1946; 2. Lapedes, 1974; 3. Kendall, 1971; 4. Ehrenberg, 1975; 5. Hsieh, 1981; 6. Tukey, 1977.

Exercises

2.1. The skeleton table in Section 2.2 contains variables for age, height, sex, diagnosis, treatment, and three values of blood pressure.

2.1.1. Identify the scale of each of those variables according to the classifications listed in Section 2.3.1.

2.1.2. You have been asked to prepare an analysis of the response of blood pressure to treatment. You can use the three cited values of blood pressure to form any additional variables you might wish to analyze. Indicate what those variables might be and how they would be formed.

2.2. The variable age is almost always expressed in a dimensional scale. Prepare transformations of the data that will express age in three different scales that are ordinal, binary, and nominal. Cite the contents of the categories that form the new scales.

2.3. The value of a variable expressing a change in state is usually determined by direct comparison of results in two single-state variables. Occasionally, however, a transition variable is created during clinical activities without specifically establishing two individual single-state values. Give an example of a question you might ask a patient (or about a patient) in which the answer would produce data that can be directly classified as a transition variable.

2.4. In Section 2.4.1, a composite variable was created by combining two or more separate variables. Can you give at least one example of a nonmedical phenomenon, occurring in daily life, that is regularly expressed as a composite variable?

2.5. In the TNM system for classifying patients with cancer, the T is expressed in 4 or 5 categories for rating certain attributes of the primary tumor. The N contains three ordinal categories for rating the spread of the tumor to regional lymph nodes. The M contains three ordinal categories for rating the spread of the tumor to distant metastatic sites. With numerical ratings assigned to each category, the results for a particular patient may be cited in such TNM index expressions as T3NlM0, T2N0Ml, TlN2M2, and so on. According to the classifications you have learned in this chapter, what kind of a scale is this? Give the reasons for your decision.

2.6. Using the five types of classification listed in Section 2.3.1, what kind of scales are formed by the Apgar Score, TNM Stage, anion gap, and Quetelet index?

2.7. In extracting data from medical records, an investigator decides to cite *age in years* according to decade. The scale of categories will be **0, 1, 2, 3, …, 8, 9** according to ages **0–9, 10–19, 20–29, 30–39, …, 80–89,** and ≥ **90.** What do you foresee as the advantages and disadvantages of this scale when the extracted data are later analyzed statistically?

2.8. Find a set of diagnostic criteria for any disease in which you are interested. *Briefly* outline the construction of the criteria and classify the scale used to express the result.

2.9. From any literature at your disposal, find a variable that has been reported with what you regard as unsatisfactory precision, or that seems to have been categorized or coded in an unsatisfactory manner. The categorization or coding is "unsatisfactory" if you do not like the chosen categories, cannot determine how they were identified or demarcated, or believe that the selected codes were inadequate. For example, you might complain about a component of the Apgar Score if you think that heart rate should have been divided into four categories rather than three, if you cannot decide where to assign a baby with a heart rate of 100, if you would have difficulty using the stated criteria for scoring *muscle tone*, or if you think that *muscle tone* is not a "sensible" or sufficiently important variable to be included in the total score.

All that is needed for the answer here is one complaint. You need not identify or include a copy of the source publication, but please say enough about the topic or circumstances for your complaint to be clearly understood and justified.

Part I

Evaluating a Single Group of Data

If individual items of data are the statistical counterpart of biologic "molecules," groups of data are the "cells." In fact, the word *cell* is regularly used in statistics to denote the particular group of data in each location of the basic entity called a *table*. Just as cells are the smallest viable structure in biology, a group of data for a single variable—containing information such as each person's diastolic blood pressure, sex, clinical stage, or principal diagnosis—is the smallest analyzable structure in statistics.

The next seven chapters describe the many things that might be done with such a "univariate" group of data. The activities extend all the way from simple descriptive summary indexes, such as medians and standard deviations, to more complicated inferential calculations, such as standard errors, confidence intervals, and a "one-sample" t-test. This long tour through a single group of data will also include the main conventional activities of statistical inference — probabilities and parametric sampling — and some attractive new procedures made possible by modern computers: jackknife analyses and bootstrap resampling.

If you become impatient with the amount of attention being given to a single group, please bear in mind that (1) as a counterpart of the cells of biology, the phenomena have fundamental importance in statistics; and (2) all the activities of "advanced" statistics are essentially an extension of what can happen for a single group. If you understand the univarate descriptions and inferences, the rest is relatively easy.

3

Central Index of a Group

CONTENTS

The items in a large collection of data can always be examined individually, but the collection cannot be interpreted until it receives a meaningful summary. To "make sense" of a collection of individual items of data, we begin by forming groups, inspecting the spectrum of data in each group, and choosing appropriate summaries for each spectrum. This chapter is concerned with the crucial first choice in that summary: a central index.

3.1 Formation of a Group

Most statistical discussions begin with the idea that a "sample" has been obtained and is about to be analyzed. In most medical activities, however, the idea of a sample is often misleading because a process of "sampling" did not occur.

Instead, the investigator usually collects data from conveniently available groups of people. Sometimes the group is "natural," consisting of everyone who came to the emergency room or received a

particular treatment during a specified interval of time. Often the group is more "artificial," containing the individual persons who decided to respond to a mailed questionnaire or who volunteered for a special test. Sometimes the group may contain not individual people but a set of observations obtained repeatedly for the same person.

The way that groups are chosen and formed is an important scientific issue that is commonly overlooked when statistical procedures are applied. As an arbitrarily defined collection of conveniently available people, most groups are usually biased either for representing a larger population or for being compared with one another. The prevention, evaluation, and management of these biases are prime challenges in scientific architecture, but the statistical discussion here will proceed with the assumption that the groups have been suitably formed. The statistical goals are to prepare a "univariate" summary of the data for each variable in a group.

3.2 Role of Summary Indexes

The summary of a group is used for both internal and external purposes. *Internally*, it is a reference point for locating the relative status of individual members within the group. Is a particular person much older or much younger than the other people? *Externally*, the summary index locates the group for comparison with the outside general world or with some other selected group. Thus, we might want to know whether a particular collection of people contains mainly children or octogenarians or whether the group is generally older than a compared group. In other external activities, we may examine the relationship of one variable, such as *age*, to other variables, such as *serum cholesterol* and *survival time*.

The formation of a summary often receives perfunctory statistical attention. After a description of the *means* that offer an "index of central tendency" and the *standard deviations* that offer an "index of spread," the discourse promptly advances to more "interesting" topics: the inferential statistical procedures used for analyzing the summarized data. If the descriptive data have prime scientific importance, however, the choice of summary expressions is a key decision, requiring substantial thought and attention.

At least three major problems can arise if we merely calculate means and standard deviations, while ignoring the main scientific goals and the other statistical options:

1. *Representation:* The mean or other value chosen as a central index becomes the one single number that repeatedly represents the group thereafter, in all of the internal and external comparisons. If this central index is unsatisfactory, the group will not be adequately represented.
2. *Distortion:* If the spread of the data in the group's spectrum is denoted by the standard deviation, and if the standard deviation is not an effective index of spread, the data will be inadequately represented and possibly distorted for internal placement of relative locations, for various external comparisons, and for analytic correlations with other variables.
3. *Omission:* Many important clinical phenomena are expressed in categorical data, which cannot be summarized with means and standard deviations. Consequently, a focus on only the latter two indexes will omit the challenge of getting suitable summaries for non-dimensional variables.

For all these reasons, the role and choice of summary indexes require careful attention. The rest of this chapter is devoted to indexes of central tendency. The next two chapters are concerned with indexes of internal location and spread.

3.3 Spectrum (Distribution) of a Group

To choose appropriate summary indexes, the first step is to see what is in the data. The collection of values in the group form what is usually called a *spectrum* by biologic scientists and a *distribution* by statisticians. The contents and "shape" of the spectrum will indicate how best to summarize the results.

3.3.1 Displaying the Spectrum

A spectrum is arranged, displayed, and summarized differently for non-dimensional and dimensional data.

3.3.1.1 One-Way Frequency Tabulations — Non-dimensional data are usually displayed in a "one-way table" showing frequency counts and the relative frequencies (or proportions) for each categorical value. The results are called "one-way tables" because only a single variable is tabulated. Figure 3.1 shows the contents of such a table for the ordinal variable, *TNM stage*, in a group of 200 patients with lung cancer. The distribution has its highest individual proportions (25.5 and 34.0 percent) at the two "ends" of the data. The lowest proportion (9.0 percent) is in the "middle." (The "cumulative" frequencies marked in Figure 3.1 will be discussed in Chapter 4.)

TNM STAGE

TNM STAGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
I	51	25.5	51	25.5
II	27	13.5	78	39.0
IIIA	18	9.0	96	48.0
IIIB	36	18.0	132	66.0
IV	68	34.0	200	100.0

FIGURE 3.1
Distribution of TNM stages in 200 patients with lung cancer.

Dimensional data can also be displayed in one-way tables, but a "pattern" may not become apparent if the spectrum contains too many individual values.

3.3.1.2 Stem-Leaf Plots — To show the pattern of dimensional data, John Tukey[1] introduced the display called a *stem-leaf plot*. For this plot, each numerical value of data is split into two parts: the left side becomes a *stem*, and the right side, a *leaf*. For example, the numbers 114, 120, 93, 107, 128, 99, and 121 might be split as 11|4, 12|0, 9|3, 10|7, 12|8, 9|9, and 12|1. In the tabulated plot, the leaves are listed consecutively next to each common stem. In the foregoing group of values, which has three 12 stems and two 9 stems, the plot would be

12|081
11|4
10|7
9|39

The number of digits in the available values will determine where to place the splits that separate stems and leaves. The subsequent plot can then be constructed simply and rapidly to show the spread, concentration, and shape of the data.

Figure 3.2 shows a stem-leaf plot for the values of hematocrit in 200 patients with lung cancer. The stems extend from **24** to **56**; and most of the leaves are **0** because hematocrit values are usually recorded as integers. The column on the right shows the number of leaves at each stem.

3.3.1.3 Histograms — Before stem-leaf plots were invented and became generally available via commercial computer programs, dimensional data were often displayed with histograms. (The term is a reduced combination of history-gram; it has nothing to do with histology.)

Instead of demarcations formed by the stems of leading digits, the data are divided into a series of intervals. Instead of showing the trailing digit(s) as leaves, the number of items are counted as frequencies for each interval. When placed on a graph, the X-axis for the results indicates the boundaries of each interval, and the Y-axis, the associated frequency counts or their relative proportional frequencies. The graph forms a *frequency polygon*, if shown with points, and a *histogram*, if shown with bars.

Stem	Leaf	#
56	0	1
55		
54	0	1
53	00	2
52	0	1
51	00	2
50	000000	6
49	000	3
48	000009	6
47	0001	4
46	0000000	8
45	000000000000006	15
44	000000000000	12
43	000000000000555	15
42	00000000000000000009	19
41	00000000047	11
40	0000000000000005	16
39	000000000000003399	18
38	00000000004	11
37	0000000000558	13
36	000033	6
35	0000000	7
34	000000255	9
33	0000003	7
32	07	2
31	00	2
30	0	1
29		
28	0	1
27		
26		
25		
24	0	1

FIGURE 3.2
Stem-leaf plot for values of hematocrit in 200 patients with lung cancer.

To illustrate the procedures, consider the data of Table 3.1. For the raw values of data in the left-hand column, the boundaries of intervals are chosen arbitrarily. In one common approach, at least seven intervals are used for the spread of the data; in another approach each interval spans 5 dimensional units. Both tactics can be used to divide the data of Table 3.1 into seven intervals: 1**0–14, 15–19, …,** and **40–44** units. The results for relative frequency are graphed in Figure 3.3 as both a frequency polygon and a histogram.

Table 3.2 shows a stem-leaf plot, prepared with an ordinary typewriter, for the data of Table 3.1. (Each digit must be given equal width if a stem-leaf plot is prepared by hand or by a typewriter that ordinarily shortens the width of the digit 1.) The horizontal pattern shown in Table 3.2 is essentially identical to the vertical pattern in Figure 3.3.

3.3.2 Inspection of Shape

The *shape* of a distribution is the pattern formed by the frequencies (or relative frequencies) displayed at each of the categorical values, stems, or demarcated intervals. For non-dimensional data, the shape is not particularly important because, as noted later, it does not affect the choice of summary indexes. For dimensional data, however, the patterns — when arranged vertically rather than in horizontal stem-leaf plots — can form the series of shapes shown in Figure 3.4.

TABLE 3.1

Raw Data, Interval Values, and Relative Frequencies

Raw Data	Demarcated Intervals	Frequency Count	Relative Frequency (%)
12,14,12,11,13	10–14	5	9
15,18,17,17,16,19,19,18,15	15–19	19	34
17,16,16,17,19,15,17,19,16,18			
21,24,20,22,22,21,24,20	20–24	13	23
23,21,22,24,21			
28,26,29,28,28,25	25–29	10	18
29,26,28,29			
33,30,34,33	30–34	4	7
36,39,37	35–39	3	5
41,43	40–44	2	4
TOTAL		56	100

TABLE 3.2

Stem-Leaf Plot for Data in Table 3.1

4*	13
3•	697
3*	3043
2•	8698859689
2*	1402214031241
1•	5877699857667957968
1*	24213

Note: The symbol * is used for leaves having values of 0–4; the • symbol is for values of 5–9.

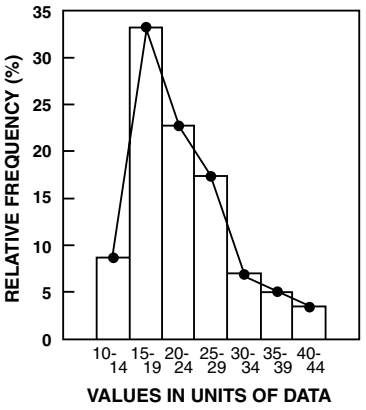


FIGURE 3.3
Histogram (*rectangles*) and frequency polygon (*dots and lines*) for data contained in Table 3.1.

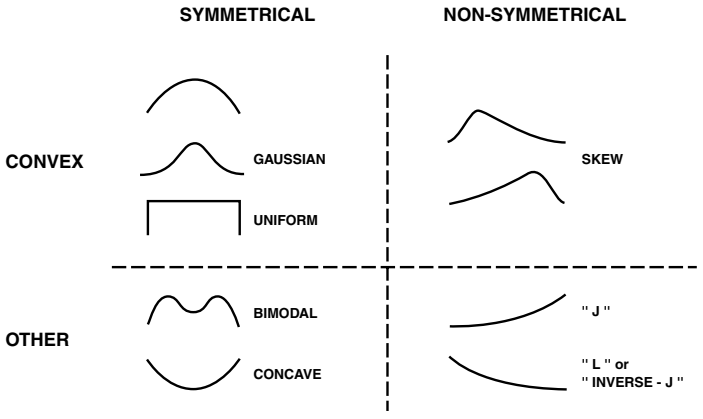


FIGURE 3.4
Shapes of different distributions of data. The vertical axes for each shape are relative frequencies of occurrence for the values (or intervals) in the horizontal axes.

These univariate patterns are sometimes mistakenly regarded as bivariate, because the graphs seem to have two variables, placed along an X- and Y-axis. The entity on the Y-axis, however, is not a variable. It is a statistic determined from counting the available values, not from acts of individual measurement. The Y-axis shows frequency counts or relative proportions for each category of the single variable on the X-axis.

3.3.2.1 Catalog of Shapes — The shapes in Figure 3.4 can be catalogued as convex or other, and as symmetrical and non-symmetrical.

Most patterns have one or more crests that form a basically *convex* shape, with small relative frequencies at one end of the distribution, rising toward one or more crests in the interior and falling toward the other end. With a single crest at the center of the curve, the convex shape is also *symmetrical*. The famous "bell" or "cocked-hat" shape of the Gaussian (or "normal") distribution has an essentially symmetrical convex pattern. (The distribution will be called Gaussian throughout this text to avoid the medical ambiguity produced when statistical jargon refers to this shape as "normal"). Another shape that can be regarded as symmetrically convex is a *uniform* (or rectangular) distribution. When the single crest is asymmetrically off-center, located toward one of the exterior ends, the distribution is called *skew*. The frequency polygon in Figure 3.3 has a right skew pattern.

A symmetrical distribution need not be convex. Symmetrical shapes can also be produced by a *bimodal* distribution with two distinctive crests and by a *concave* distribution with a U-shaped trough. (The latter type of shape would be produced by the data in Figure 3.1.) Bimodal and concave distributions can also be asymmetrical, according to the location of the crests and troughs.

A *sloping* distribution does not have the fall-rise-fall or rise-fall-rise pattern of any of the previous shapes. Instead, the curve of relative frequencies is "monotonic," continuously rising toward (or falling from) a crest that occurs at one end of the distribution. One type of sloping distribution is sometimes called *J-shaped*; another type is called *inverse-J-shaped*, or *L-shaped*. Statisticians often use the word *exponential* for these sloping shapes.

Shapes can also be described as *centripetal* if their crest (or crests) are somewhere near the center of the distribution away from either end. All basically convex shapes are centripetal. The shape is *centrifugal* if the crests are located at the ends, or if no specific crest occurs. The uniform, concave, and sloping distributions are all centrifugal.

The reason for all the attention to shape is that symmetrical distributions — whether convex or concave, centripetal or centrifugal — can be well represented by the means and standard deviations discussed later. All other distributions are eccentric and may require other types of summary indexes to avoid distortions.

3.3.2.2 Outliers — When arranged in ascending or descending order of magnitude, the items of a dimensional distribution are usually close to one another, forming a shape that seems well "connected" from its lowest to highest values. Sometimes, however, a large gap may occur between the main bulk of the data and one or more values at the extreme ends.

The extreme values, called *outliers*, create two types of problems. First, they may sometimes arise as errors in observing or transferring data. The error can be immediately detected if it represents an impossible value. For example, values of **–21** or **275** for *age in years* are obviously erroneous. A value of **275** for *weight in kg*. is unusual but not necessarily an error. (A complex *scientific* decision is whether to accept an outlier value as unusual or to discard it as erroneous.)

The second problem arises if non-symmetrical outlier values, which produce an eccentric shape, are accepted and included in the data. To avoid the possible distortion of summary indexes calculated as means and standard deviations, the raw data may be transformed (as noted in Section 3.8.1) into scales that reduce the impact of outliers, or the summary indexes can be chosen with methods that are unaffected by outliers.

3.4 Types of Summary Indexes

At least two types of *indexes of location* are used to summarize a set of data. For external comparisons, an *index of central tendency*, or *central index*, will represent the entire group of data when it is referred to the outside world. For this purpose, we want something that represents the group by being particularly typical, common, or central in the distribution. For internal comparisons, an *index of cumulative relative frequency* (or *percentile*) will locate the items of data with reference to one another. These (and other) indexes of inner location can also indicate the spread or dispersion of the data around the central index.

3.5 Indexes of Central Tendency

Although no single index can perfectly represent an array of values, we want something that will do the best possible job. The job will vary for non-dimensional or dimensional data.

3.5.1 Indexes for Non-Dimensional Data

Non-dimensional data are often summarized not with a single central index, but with an apportionment that shows the relative frequencies of each category. This type of expression, marked "percent," was used to summarize the five categories listed in Figure 3.1.

If a group has c categories, and if n_i represents the number of members in category i, the total size of the group will be $N = n_1 + n_2 + n_3 + ... + n_c$. The Greek symbol Σ is regularly used to show these summations. Thus, $N = \Sigma n_i = n_1 + n_2 + ... + n_c$. [In strict mathematical usage, the Σ is accompanied by upper and lower symbols that indicate exactly where the summation begins and ends. For the complete collection of data, the result here would be shown as $N = \sum_{i=1}^{c} n_i$. In pragmatic usage, however, the sum almost always extends from the first to the last items. The upper and lower symbols are therefore omitted from the Σ sign.]

The actual frequencies are converted to relative frequencies when cited as proportions of the total group. Thus, the relative proportion of the i^{th} category is $p_i = n_i/N$. In Figure 3.1, which shows 5 categories, the third category has $n_3 = 18$ and N is 200, so that $p_3 = 18/200 = .09$ or 9%.

3.5.1.1 Binary Data — For binary data, only two proportions can be cited. If 35 people have lived and 15 have died in a group of 50, the survival proportion is p = 35/50 = .70 or 70%. The fatality proportion is q = 15/50 = .30 or 30%. Similarly, if a group of 60 people contains 21 women, the proportions are p = .35 for women and q = .65 for men, or vice versa.

Because q = 1 − p as a proportion (and 100 − p as a percentage), each of the two proportions (or percentages) is called the *complement* of the other. The complete spectrum of binary data is effectively summarized with either proportion. The choice of whether to use p or q usually depends on the item of interest, rather than the item that is most common. Thus, the summary of a group might say that 14% of its members have coronary disease, rather than 86% do not.

3.5.1.2 Nominal Data — Suppose a group of patients on a medical service has the following spectrum of nominal diagnoses: **pneumonia**, 8; **myocardial infarction**, 5; **ulcerative colitis**, 1; **stroke**, 4; **cancer**, 6; and **renal failure**, 2. In these 26 people, the relative frequencies of each diagnosis are **pneumonia**, .31 (or 31%) calculated as 8/26; **myocardial infarction**, .19; **ulcerative colitis**, .04; **stroke**, .15; **cancer**, .23; and **renal failure**, .08. To summarize a nominal spectrum, the relatively uncommon items of data are often consolidated into a single category called **other** or **miscellaneous**. Thus, the **ulcerative colitis** and **renal failure** diagnoses might be combined into an **other** category, with frequency of 3 and a relative frequency of .12. For the total spectrum, the individual proportions will add up to 1 (or 100%).

The spread of the spectrum is shown in the array of proportions for each cited category, but a single central index is difficult to choose for nominal data. The best single choice is usually the proportion of the most common item, which is called the *mode*. For the spectrum of the preceding 26 diagnoses, *pneumonia = 31%* would be the best single central index.

Alternatively, certain categories may be consolidated and expressed accordingly. Thus, *myocardial infarction* and *stroke* might be combined, and the central index listed as *cardiovascular = 34%* [= (5 + 4)/26].

3.5.1.3 Ordinal Data — For grades of ordinal data, the central index can come from the proportion of the most common category or from two or more consolidated categories. Suppose *response to treatment* is rated as **poor** in 6 people, **fair** in 12, **good** in 11, and **excellent** in 10. A single summary index could be formed from the mode of **fair** in 31% (= 12/39) or from the consolidated category, **good or excellent,** in 54% (= 21/39). The data in Figure 3.1 could be summarized with the statement that 61% (= 9% + 18% + 34%) of the patients are in the **metastatic** stages of **IIIA, IIIB,** or **IV.**

Because ordinal data can be ranked, the results can be summarized with a *median*, which is the middle value in the ranked grades. In the preceding 39 people, the 20th rank has the middle position. Counting the 20th rank from either end of the data, **good** is the median value. For the 200 people in Figure 3.1, the middle ranks are 100 and 101. **Stage IIIB,** the value at both those ranks, is the median. (Medians are more thoroughly discussed in the next section.)

In some instances, ordinal grades are given arbitrary numerical values, which are then treated as though they were dimensional. Thus, if the foregoing responses were rated as **1** = poor, **2** = fair, **3** = good, and **4** = excellent, the individual ratings might be added and averaged in the same manner shown in the next section for dimensional data. Although the propriety is often disputed, this additive procedure is often applied to ordinal data.

3.5.2 Indexes for Dimensional Data

For dimensional data, three different contenders are immediately available as an index of central tendency: the mode, the median, and the mean. The *mode* is the most common value in the data. It can be determined for any type of data — dimensional or non-dimensional. The *median* is the value that occupies the middle rank when the data are arrayed monotonically in ascending or descending magnitudes. A median can be determined for ordinal as well as dimensional data. The *mean*, as an arithmetical average of the added values, is most properly used only for the equi-interval data of dimensional variables.

Consider the group of data {13, 7, 1, 15, 22, 7, 21}. For these 7 items, the mean is produced with the formula $\Sigma X_i/n = \bar{X}$, when their sum, 86, is divided by the number of items to yield 86/7 = 12.3. The mode is 7, which happens to be the only value that appears more than once in this group. To find the median, we rearrange the values according to their ranks as 1, 7, 7, 13, 15, 21, and 22. The middle-ranked number, 13, is the median.

The mode can be determined, without any calculations, from a simple count of frequencies in the data. The median can also be determined without any substantial calculations. We simply rank the data and count the ranks until reaching the middle one. For N items of data, the median is the value of the item of data that has rank (N + 1)/2 if N is odd. If N is even, (N + 1)/2 is not an integer; and the median occurs between the values in ranks N/2 and (N + 2)/2. By custom, the median is assumed to be midway between them. Thus, in the foregoing array of 7 items, the median was at the rank of (7 + 1)/2 = 4. If the data contained 8 items — arranged as 1, 7, 7, 13, 15, 21, 22, and 25 — the median would be midway between the 4th and 5th items, at the value whose rank is (N + l)/2 = (8 + 1)/2 = 4.5. There is a 2-unit distance between 13 (the 4th ranked value) and 15, the 5th ranked value. Half of this distance is 1 unit, and so the median would be 13 + 1 = 14.

Each of the three contenders has a reasonable claim for selection as the central index. The *mode* has the right of popularity: it is the most common occurrence. The *median* can argue that its central position

is literally in the exact middle (or center) of the array. Half of the items are on one side of the median; half are on the other. The *mean* has the centrality of a fulcrum. It is the point that divides the data by weight. Half of the summed amounts (or weights) of the values are on one side of the mean; half are on the other.

Given these three reasonable options, which should be chosen as the index of central tendency? For people with commercial marketing interests, the best choice is often the *mode*. It lets the vendor know where to find the largest audience of potential buyers for a product. In medical research, however, the mode is generally not a good central index because the data may contain either no naturally occurring mode or several widely separated modal values.

3.5.3 Choice of Median or Mean

With the mode eliminated as a contender, the choice is between the mean or median. For this decision, we need to examine the pattern (or "shape") of the dimensional distribution.

The shapes of the diverse spectrums in Figure 3.4 demonstrate the problems of making a suitable choice. A centripetal distribution can be reasonably represented by a central index, but a centrifugal distribution does not have a central tendency; and any efforts to create one will inevitably be undesirable compromises.

For a symmetrical convex distribution, the mean and median will occur at essentially the same crest of the curve, and so no major decision is necessary. Either choice will give the same value. For a symmetrical distribution that is concave, bimodal, or even multi-modal, no central index can be really satisfactory. Citing a mode would be inadequate, since two or more modes would be needed to represent the two "crests" on either side of the center. The mean and median will have essentially similar values, but neither one will be "typical" of the main bulk of the data, which occurs on both sides of the center. Nevertheless, for lack of anything better, either the mean or the median would seem to be a reasonable compromise in these circumstances.

The main problems and challenges occur for the many medical distributions that are eccentric. Their shape can be overtly centrifugal, or centripetal but non-symmetric, or basically symmetric with distinctive outlier values at the low or high ends of the spectrum. Figure 3.5 illustrates the different locations of the mode, median, and mean in a common type of right-skewed non-symmetrical distribution that creates difficult decisions in choosing a single central index. [A simple mnemonic for remembering the location of the three indexes in a skewed distribution is that they are arranged both statistically and alphabetically starting at the skewed end of the data. Mean precedes median, which precedes mode.]

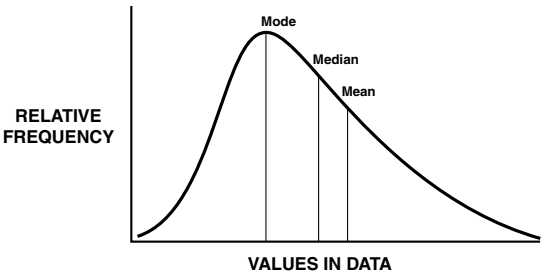


FIGURE 3.5
Customary location of mode, median, and mean in a right-skewed distribution.

3.6 Advantages of the Median

The median has four main descriptive advantages as an index of central tendency. It is better than the mean for summarizing eccentric distributions; it can also be properly used, unlike the mean, for ordinal data; it is often, unlike the mean, an actual member of the data set; and it can be applied, also unlike the mean, to sets containing incomplete longitudinal data.

3.6.1 Eccentric Distributions

The median is preferable to the mean for avoiding problems in summarizing eccentric distributions created by outliers or by asymmetrical (skew) patterns.

3.6.1.1 Avoidance of "Outlier" Distortions — Consider a group of eleven people aged 1, 2, 4, 4, 5, 5, 5, 6, 6, 8, and 97 years. Except for the first and last members of the group, all other values are arranged symmetrically around a central value of 5. If the last value were 9 rather than 97, the entire distribution would be symmetrical, with the mean and median each being 5. The outlier value of 97 substantially distorts the symmetry, however, and raises the mean to a value of 13.0 — an index that misrepresents the group's central location for age and that is not typical for any member of the group. The median, on the other hand, has the advantage of being unaffected by extreme outliers. For the eleven people just listed, the median value would be 5, regardless of what the highest value may be.

3.6.1.2 Use in Asymmetrical (Skew) Distributions — The virtues just cited for the median become more apparent when a distribution is overtly asymmetrical. For a convex skew distribution or for a sloping distribution, the crest of relative frequencies, the median, and the mean will each usually be located at different sites. For example, for the raw data listed in Table 3.1, the mode is 17, the median is 21, and the mean is 22.7. The shape of these data, as shown in Figure 3.3, is skewed to the right — a pattern that commonly occurs in medical data.

The more general pattern of a right-skewed distribution is shown in Figure 3.5. The most "typical" or common values of skew distributions will occur at their crests, which are the modal values that are unsatisfactory as indexes of central tendency. In choosing an alternative central index, the median seems preferable because the outlier values that create the skew will also move the mean far from the crest, leaving the median to fall between the crest and the mean.

The median thus has two advantages over the mean for summarizing an eccentric distribution. Being unaffected by the impact of outlier values, the median will always be consistently located in the central rank of the distribution; and the median will also be closer than the mean to the "popular" values in the crest.

3.6.2 Application to Ordinal Data

Another advantage of the median is that it can readily be applied to ordinal data because its proper calculation does not require dimensional values. In Section 3.5.1.3, the median produced **good** as a central index for 39 ordinal ratings of *response to therapy*, and **Stage IIIB** for 200 cases of lung cancer.

3.6.3 Membership in Group

Another advantage is that the median value is particularly likely to occur in an actual member of the group. For example, when a census report says that the average American family contains **2.3** children (calculated as a mean), you may wonder what the family does with the 0.3 child. If the average is reported as a median of **2**, however, the statement is immediately plausible.

If the group contains an even number of people, however, the median is calculated as an average of two middle values. If they are different, the median will not be a member of the group. Thus, for six people whose ages are 7, 9, 10, 28, 37, and 43, the two middle values would be 10 and 28. The median would be cited as halfway between them, or 19. Although not found in any member of the group, this median value seems reasonable as a central index. (The alternative choice would be a mean of 22.3.)

3.6.4 Application to Incomplete Longitudinal Data

Finally, the median is particularly valuable for expressing results of longitudinal data (such as *length of survival*) in groups where the *mean* value cannot be calculated because some of the people have not yet died (or otherwise ended the follow-up period). For example, suppose 17 patients followed for 3 years

after treatment of a particular cancer have survival times that are **1, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 7, 12, 17,** and **25** months, and **unknown** for two patients who are still alive after 36 months. The mean survival time cannot be calculated yet for this group; and the result will vary according to whether the two survivors die soon after 36 months or live much longer. The median survival, however, can be determined now. It is 5 months, and will remain at 5 months no matter how long the last two people live.

The median is also helpful as a *single* summary for survival of the group. The survival rates might be reported for individual time points such as 65% (=11/17) at three months, 35% (=6/17) at six months, and 24% (=4/17) at one year; and a survival curve could be drawn to show each temporal rate. To summarize survival with a single quantitative value, however, the best of the available options would be the median of 5 months.

3.7 Disadvantages of the Median

In view of all these advantages of the median, you may wonder why the mean is so commonly used in reports of medical research.

3.7.1 Symbolic

A trivial reason for preferring the mean is that it has a standard symbol, $\bar{X}$, but the median does not. It is sometimes labeled as X_m, M, or X_{med}. This problem can readily be solved, particularly with modern typewriter symbols, by using $\tilde{X}$ for a median.

3.7.2 Computational

A more substantive disadvantage of the median is the enumerative challenge of finding it. The data must be arranged in ranked order and the ranks must be counted to find the middle one. In the days before computers and programmable electronic hand calculators, this ranking procedure could be a nuisance, requiring considerable time to get all the numerical values ranked and then counted for a large group. A mean was easily determined, however, by the simple act of pushing buttons in a mechanical calculator, which would do the adding and dividing.

Today, this disadvantage has little pertinence. With modern computational devices, the median can often be determined just as easily as the mean; and, in fact, with a suitably programmed calculator, both values can emerge from the same act of button pushing.

3.7.3 "Elitist"

With the problem in counting eliminated, the only remaining objections to the median are conceptual. One conceptual complaint is that the median is somewhat "elitist." It is determined from only the middle or the two middlemost values in the data. The rest of the data are used as ranks and counts in finding the median, but do not participate as actual values. Since the choice of a central index is intended to provide intellectual communication rather than participatory populism, this objection does not seem cogent. Besides, a value that denotes "mediocrity" can hardly be regarded as "elitist."

3.7.4 Inferential

The most important conceptual objection to the median is that it seldom appears in the probabilistic reasoning and calculations of inferential statistics.

After being chosen for a group, the central index is often used for inferences about a "parent population," or for stochastic contrasts of "statistical significance" in two groups. The various strategies (confidence intervals, P values, etc.) used for these statistical inferences, as discussed later, have traditionally been developed with concepts that depend on means, rather than medians. Consequently, the

mean is usually preferred because it can be analyzed for both descriptive and inferential statistics. Even if a median were used for description, we might still need the mean for inference.

This point is true, but not germane to the main descriptive issues; and the point may eventually become irrelevant for the inferential issues. In description, we want a summary that will communicate as effectively as possible. The median usually offers more effective communication than the mean. In the inferential activities, if a data set is poorly represented by the mean, its use is scientifically unattractive, regardless of whatever traditional mathematical theory may be associated. Most importantly, however, modern computational devices have allowed inferential statistical work to be done with new strategies called *non-parametric*, *permutation*, or *resampling procedures*, which do not require a mean for the calculations. As these new procedures become more popular and eventually replace the older ones (such as t or Z tests), the communicative advantages of the median will no longer be inhibited by extraneous features of statistical inference.

Until that time arrives, however, you will often see the mean, rather than the median, in most summaries of dimensional data.

3.8 Alternative Approaches

The descriptive problems of using the mean as a central index have led to many proposed substitutions and adjustments, having diverse titles or eponyms. Among the many proposed alternatives, only two are likely to appear in current medical literature: transformations of data and geometric means.

3.8.1 Transformations of Data

Various mathematical transformations have been used to convert non-symmetrical distributions into more symmetrical shapes, from which the mean is then calculated. After the raw data are changed into their square roots, logarithms, or some other suitably chosen value that will produce a "better" pattern, an additive mean is calculated for the transformed data. The result is then converted back to the original units of expression.

For example, in a study of antibody levels to poliomyelitis vaccine, the following values were found in 17 persons: 4, 4, 8, 8, 16, 16, 32, 64, 64, 256, 256, 256, 512, 512, 512, 1024, and 2048. For these skewed data, the median is 64, but the arithmetic mean is the unrepresentative value of 328.9. The data would become more symmetrical (or "Gaussian") if transformed to their logarithmic values (base 10). The logarithms of the 17 antibody values are respectively .60, .60, .90, .90, 1.20, 1.20, 1.51, 1.81, 1.81, 2.41, 2.41, 2.41, 2.71, 2.71, 2.71, 3.01, and 3.31. This distribution seems reasonably symmetrical, and its mean is 1.89. To convert the "log mean" back to the original values, we calculate $10^{1.89}$ and get 78.47. (Note that the reconverted log mean is reasonably close to the median value of 64.)

The logarithmic transformation is particularly useful (and common) for data that have a right skew or high-value outliers. A square root transformation is sometimes applied to help "adjust" data that have a left skew or low-value outliers. A transformation having the delightful eponym of Box-Cox[2] is often mentioned in statistical literature, but rarely (if ever) receives medical usage.

Two other seldom-used transformations are applied for their logical rather than "Gaussianizing" virtues. The *harmonic mean*, calculated from the reciprocals ($1/X_i$) of the observed values, is often used for averaging rates of speed. The *quadratic mean* or *root mean square*, calculated from the squares of the observed values as $\sqrt{\Sigma X_i^2/n}$, appears later in the text here as a mechanism for summarizing a set of differences in two measurements of the same entities.

3.8.2 The Geometric Mean

The log-mean transformation was popular in the days before modern electronic calculators. Today, the same result can be calculated directly as an entity called the *geometric mean*. It is obtained by multiplying the N values of observed data and then taking the Nth root of their product. (The advantage of modern

calculators is the "y^x" function that allows an easy computation of the N^{th} root, where y = product of the values and x = 1/N.)

The direct product of the 17 previous antibody values is $4 \times 4 \times 8 \times \ldots \times 512 \times 1024 \times 2048 = 1.622592769 \times 10^{32}$. The 17th (or 1/17) root of this number is 78.47, which is the same as the reconverted log-mean value.

The logarithmic transformation was popular for many years because it could often "Gaussianize" skew data and provide the numerical precursor of the geometric mean. Today, with the ease and rapidity permitted by electronic calculators, the geometric mean can be determined directly with hardly more effort than what is needed to obtain an arithmetic mean.

For people who want a "populist" index of central tendency, the geometric mean has the advantage of including all the values of data in the calculation, but the result is often close to the "elitist" median. This advantage of the geometric mean is lost, however, in several important situations. If any of the data values is 0, the entire product is 0, and the geometric mean must be 0. If the product of the data values is negative, a logarithm cannot be determined and the Nth root of the product (even if calculatable) may be meaningless. A different kind of problem occurs if the data are skewed to the left, rather than rightward. For example, suppose the data set contains the values 1, 62, 67, 67, and 75. The arithmetic mean of this group is 54.4 and the median is 67, but the geometric mean is 29.1, which is farther from the median than the arithmetic mean. On the other hand, if the data set were skewed right, with values of 62, 67, 67, 75, and 200, the arithmetic mean would be 94.2 and the geometric mean of 83.97 would be closer to the median of 67. For these reasons, except perhaps in an all-positive skewed-right distribution, the median is a better general index of central tendency than the geometric mean.

The geometric mean is often preferred, however, if the basic data are expressed "exponentially" in powers of 10 (or some other number) for such entities as bacterial counts and antibody titers. In the earlier example (Section 3.8.1) of 17 antibody levels to poliomyelitis vaccine, all of the values were expressed in exponential powers of 2, and the geometric mean was much better than the arithmetic mean as an index of central tendency. The geometric mean has also been used to summarize the "skewed distribution" of plasma prolactin concentrations[3] and the "log normal distribution" of blood-lead values.[4]

3.8.3 The Midrange

A particularly simple index, the *midrange*, is half of the sum of the minimum and maximum items in the data set. The *midrange* is a good "screening clue" for discerning an asymmetrical distribution, which probably exists if the midrange differs substantially from the ordinary additive mean, $\overline{X}$. In Table 3.1, the midrange of (12 + 43)/2 = 27.5 immediately indicates the asymmetry of the data set, for which the mean is 22.7.

3.8.4 "Robust" Means

In modern new approaches, called Exploratory Data Analysis (EDA),[5] the mean is still determined by addition, but is made more "robust" by different methods of thwarting the effect of outliers. The calculations use various tactics to eliminate or "trim" the extremes of the data. The results have names such as trimmed mean, Tukey's tri-mean, Studentized mean, Windsorized mean, and Hempel adjusted mean; and new suggestions are constantly proposed in the literature of statistics and data processing.

The mathematical attraction of the "robust means" seldom overcomes both their unfamiliarity and their inapplicability for inferential (rather than descriptive) statistics. The median seems much simpler and more intuitively appealing as a descriptive mechanism for avoiding the effects of outliers.

References

1. Tukey, 1972; 2. Sakia, 1992; 3. Baron, 1986; 4. Fulton, 1987; 5. Tukey, 1977.

Exercises

3.1. Which of the three central indexes — mean, median, and mode — would you choose to summarize the central tendency of the data in Table 3.1 and Figure 3.3? Why?

3.2. The results of a clinical trial are presented in a two-way table as follows:

	Number of Patients Who Received		
Outcome	**Treatment A**	**Treatment B**	**Treatment C**
Success	40	70	50
Failure	320	280	295

3.2.1. Classify the types of variables under assessment in this table.

3.2.2. What proportion of patients received Treatment A?

3.2.3. How would you summarize the results for success with Treatment C?

3.2.4. If you had to give a single index for the success achieved in this entire trial, what would you choose?

3.3. A group of patients on a medical service shows the following values for fasting blood sugar: 62, 78, 79, 80, 82, 82, 83, 85, 87, 91, 96, 97, 97, 97, 101, 120, 135, 180, 270, and 400. Calculate or demonstrate the following indexes for these data: mean, median, mode, geometric mean.

3.4. Using any method you want to apply for estimating shape, classify the shape of the distribution of data in Exercise 3.3.

4

Indexes of Inner Location

CONTENTS

For the *external* role of showing an "average" general location, the central index lets us know whether the members of a group are mainly children or octogenarians, fat or thin, rich or poor, high or low achievers — or somewhere in between. The central index also represents each group when it is compared to decide whether some other group is older, fatter, richer, or more achieving.

An index of inner location has a different job. It shows the relative position of individual members internally within the group. If told that someone had the 10th highest score in a certification test, we might be impressed if the group contained 273 people, and unimpressed if it contained 11. In another group with a central index of **80,** a value of **75** seems relatively close to the center — only 5 units away. The actual closeness of the **75** item, however, depends on all the other items of data. In the data set {75, 77, 77, 79, 79, 79, 81, 81, 81, 83, 83, 85} the **75** is relatively far away, separated by five intermediate items.

4.1 Analytic Roles of Inner Locations

An index of inner location can have four important roles in description and subsequent analysis.

1. The most obvious role is to identify members of a group as being relatively near or far from the center.
2. In a subsequent role, indexes of inner location can demarcate boundaries for *inner zones* that contain selected proportions of the data.
3. The size of the inner zones can indicate the data set's relative compactness or dispersion.
4. The relative dispersion of the data can then be used to assess the adequacy and (later) the stability of the central index itself.

The last three roles of indexes of inner location will be discussed in subsequent chapters. The rest of this chapter is concerned with the indexes themselves. The two most commonly used expressions are *percentiles* and *standard deviations*. The *percentile* index is simple, obvious, and direct: it relies on counting the relative "depth" of ranked items in the observed distribution. The *standard deviation*, as a calculated entity, is more complex, but is currently more popular because it has become a traditional statistical custom.

4.2 Percentiles

Percentiles are points that divide a set of ranked (dimensional or ordinal) data into proportions. At the 25th percentile point, .25 or 25% of the data are below the cited value, and 75% are above. The 50th percentile point divides the data into two equal halves.

The proportions are easy to determine because each member of a set of n items occupies a 1/n proportion of the data. The ranking can be done in either direction — from low values toward high, or vice versa — but as the ranks ascend (or descend), the proportions accumulate so that the first r ranks will occupy r/n of the data. For example, in the 56-item data set of Table 3.l, each item occupies 1/56 = .018 of the data. At the 39th rank, the first 39 items will occupy 39/56 = .696 of the data.

4.2.1 Problems in Identification

The concept is easy to describe and understand, but its application produces three tricky problems: shared boundaries, intra-item percentiles, and shared ranks.

4.2.1.1 Shared Boundaries — The proportions occupied by two adjacent members of a data set have boundaries that "touch" or are shared by the upper end of the first and the lower end of the second. For example, in a five-item data set, the first item occupies the proportions 0–.20; the second item

occupies .20–.40; the fifth occupies .80–1.00. To do the demarcation properly, the corresponding percentile point is located between the two items that might claim it. Thus, in the five-item data set, the 20th percentile would be placed between the first and second items, and the 80th percentile would be between the fourth and fifth.

When adjacent items "touch" the pertinent proportions, the in-between location allows the percentile to be additively reciprocal. Thus, the percentile point, P, when counted in one direction, will be 100 – P when counted in the opposite direction. If we went from high to low ranks instead of from low to high in the five-item data set, the 20th percentile would become the 80th, and vice versa. The direction of counting ranks is unimportant as long as it is used consistently, because the percentile in one direction is always the additive reciprocal of the percentile in the other.

4.2.1.2 Intra-Item Percentiles —

In mathematical theories for "continuous" data, each value is a point of infinitesimal width on a curve. In pragmatic reality, however, each item of data occupies its own zonal proportion of the total collection. This attribute creates problems in the small data sets that are commonly analyzed in biomedical research, because certain percentiles may be contained *within* an item of data.

For example, in a five-item data set, the first item occupies the proportions 0–.20; and any of the percentiles from 1 to 19 would be contained within this item. Similarly, the fifth item of data would contain any of the percentiles from 81 to 99. The 50th percentile or median of the data set would pass through the third item, which occupies the proportions from .40 to .60.

This distinction of small data sets is responsible for the use of two different methods (discussed later) for determining percentiles. The pragmatic-reality method uses proportions of the discrete data; the mathematical-theory method relies on interval divisions for a continuous curve.

4.2.1.3 Shared (Tied) Ranks —

When two or more items have the same value in a data set, they are "tied" and will share the same ranks. To preserve the correct sums and proportions for the total of ranks, each tied value is assigned the same *average* rank. For example, in Table 3.1, the two values of **12** would occupy ranks 2 and 3. They are each assigned rank 2.5. (The next value, **13**, becomes rank 4.) The four values of **21** would occupy ranks 27, 28, 29, and 30. Their average rank can be determined as (27 + 28 + 29 + 30)/4 = 28.5, or, more simply, as the first plus last divided by 2. Thus, (27 + 30)/2 = 28.5. (The proper management of tied ranks is an important challenge that is further discussed for Rank Tests in Chapter 15.)

Tied ranks are seldom a substantial problem in determining percentiles, particularly if the same values "compete" for the location of a particular percentile point. For example, in the 56-item data set of Table 3.1, the first 28 items occupy the proportions 0–.50 and the last 28 items occupy .50 – 1. Because the .50 value is shared by ranks 28 and 29, the 50th percentile, or median, is placed between them. Since these ranks, as noted in the foregoing paragraph, are occupied by the same value, i.e., **21**, the value of the 50th percentile is **21**.

4.2.2 General Procedure

Each of the 56 items in the data set of Table 3.1 will have two ranks — one counted from each end. Thus, **11** is the first rank counting in one direction, and the 56th rank in the other; and **43** would correspondingly be either the 56th or the first rank.

As just noted, the middle rank or *median* for this distribution would be at the rank of (56 + 1)/2 = 28.5. Counting from low to high values, this rank occurs among the four values of **21** that occupy ranks 27 through 30. Going in the other direction, if **43** is counted as the first rank, the values of **24** occupy ranks 20–22; **23** has rank 23; values of **22** occupy ranks 24–26; and values of **21** occupy ranks 27–30.

Since the number of ranks will vary as n changes from one data set to another, the counted ranks are standardized by citation as proportions or relative frequencies of the data. The cumulative relative frequency or cumulative proportions will be 2/n after the second ranked item, 3/n after the third, and r/n after the rth ranked item.

In the ordinal data of Figure 3.1, the two rightmost columns show the cumulative count of frequencies as the categories ascend, and the associated cumulative relative frequency is marked cumulative percent. After the last category in any array, the cumulative count is always N (the total size of the group), and the cumulative percent is always 100%.

4.2.3 Points and Zones

The first percentile point demarcates a zone of cumulative relative frequency that includes 1/100 of the data. The second percentile point demarcates a zone for 2/100 of the data, and so on. Note that percentile points demarcate the boundaries of zones, whereas cumulative proportions refer to the amounts of data contained in each zone. Thus, the 50th percentile point demarcates two zones, each containing a .50 cumulative proportion, or half the data in the set. The 25th, 50th, and 75th percentile points demarcate four zones, each containing one fourth of the data.

The distinction is shown in Figure 4.1. The upper part of the figure indicates the percentile points for values of the data. The lower part shows the proportionate amount of the items of data located in each zone demarcated by the percentiles.

Figure 4.2 shows the location of the 25th, 50th, and 75th percentiles in a Gaussian distribution. Because the distribution is much denser near the center than at the edges, the zones containing the second and

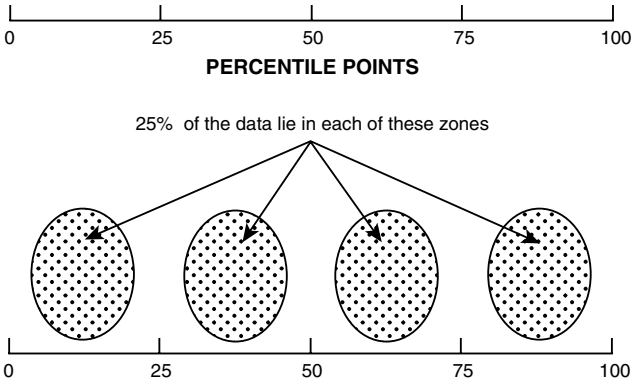


FIGURE 4.1
Distinction between percentile points (upper) and the corresponding zones of demarcated data (lower).

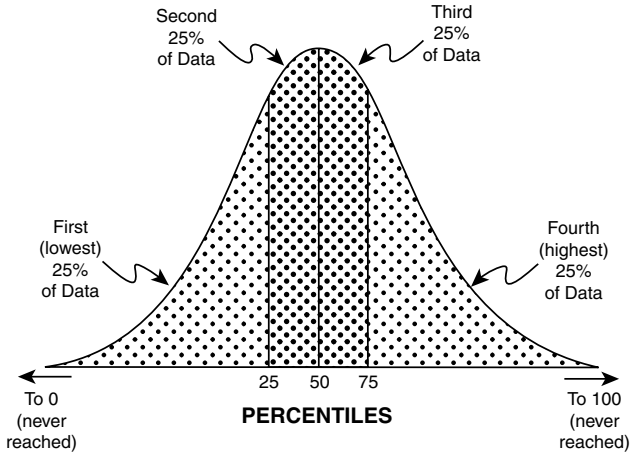


FIGURE 4.2
Location of 25th, 50th, and 75th percentiles, and associated zones in a Gaussian distribution.

third "quarters" of the data are taller and thinner than the zones for the first and fourth quarters. Despite the symmetrical shape of the Gaussian curve around its center, the distances between adjacent percentile points are not symmetrical. Thus, the distance from the 0th percentile (which is never quite reached) to the 25th percentile is much larger than the distance from the 25th to the 50th percentile. On the other hand, on the two sides of the curve, the *interpercentile* distances are symmetrical between the 25th to 50th and 50th to 75th percentiles, and between the 0th to 25th and 75th to 100th percentiles.

4.2.4 Tabular Illustration of Cumulative Frequencies

An empirical display of frequencies, relative frequencies, and cumulative relative frequencies is shown in Table 4.1, which contains the observed serum chloride values and the frequency of each value in 2039 consecutive patients tested at a university hospital's laboratory. The relative frequencies of each value, which are shown in the third column of the table, are plotted as a frequency polygon in Figure 4.3. The shape of the spectrum seems reasonably symmetrical, with the crest of the "curve" at the mode of 103.

TABLE 4.1

Serum Chloride Values for a Population of about 2,000 People (Courtesy of laboratory service, Yale-New Haven Hospital.)

Serum Chloride Concentration* mEq/dl)	Observed Frequency	Relative Frequency	Ascending Cumulative Relative Frequency	Descending Cumulative Relative Frequency	Z Scores
75–83	8	.0039	.0039	1.0000	—
84–85	12	.0059	.0098	.9961	—
86–87	12	.0059	.0157	.9902	—
88	10	.0049	.0206	.9843	–2.611
89	10	.0049	.0255	.9794	–2.418
90	28	.0137	.0392	.9745	–2.226
91	20	.0098	.0490	.9608	–2.034
92	22	.0108	.0598	.9510	–1.842
93	18	.0088	.0687	.9402	–1.649
94	32	.0157	.0844	.9313	–1.457
95	43	.0211	.1054	.9156	–1.264
96	51	.0250	.1305	.8946	–1.072
97	76	.0372	.1677	.8695	–0.879
98	96	.0471	.2148	.8322	–0.687
99	125	.0613	.2761	.7852	–0.495
100	146	.0716	.3477	.7239	–0.302
101	159	.0780	.4256	.6523	–0.110
102	204	.1000	.5257	.5743	0.082
103	222	.1089	.6346	.4743	–0.275
104	195	.0956	.7303	.3654	0.467
105	167	.0819	.8122	.2697	0.659
106	136	.0667	.8789	.1878	0.852
107	83	.0407	.9196	.1211	1.044
108	49	.0240	.9436	.0804	1.237
109	41	.0201	.9637	.0564	1.429
110	26	.0128	.9765	.0363	1.621
111	13	.0064	.9828	.0235	1.813
112	15	.0074	.9902	.0172	2.006
113–114	10	.0049	.9951	.0098	—
115–127	10	.0049	1.0000	.0049	
TOTAL	2,039	1.0000	—	—	

Note: Mean = μ = 101.572; standard deviation = σ = 5.199; the Z score values are described in Chapter 5 of the text.

* At the top and bottom of this table, certain extreme values have been compressed to save space. The individual values and their observed frequencies are as follows: 75–1; 76–1; 77–1; 80–2; 82–1; 83–2; 84–5; 85–7; 86–7; 87–5; 113–4; 114–6; 115–1; 116–3; 117–1; 118–1; 120–1; 123–2; 127–1.

The relative frequencies of the chloride values can be cumulated in either an ascending or descending direction, as shown in the fourth and fifth columns of Table 4.1. In the ascending direction, starting with the lowest chloride value, each successive chloride value (or interval) adds its own relative frequency to what was previously cumulated. The cumulative results continue to increase until they reach their final total of 1.000 after the highest chloride value is added. In the descending direction, starting with the highest chloride value, the process is similar but opposite.

Figure 4.4 shows the relative *cumulative* frequencies in an ascending direction for each chloride value. The S-shaped or sigmoidal curve is called an *ogive*. The shape is characteristic of cumulative frequencies in a Gaussian distribution.

The data in Table 4.1 (or in Figure 4.4) allow any cited chloride value to be immediately located in the spectrum. The value of 94 occurs near the lower end, as shown by its ascending cumulative relative frequency of .0844. The chloride value of 108, with an analogous proportion of .9436, is located near the upper end. The value of 102, with an ascending cumulative relative frequency of .5257, is close to the middle.

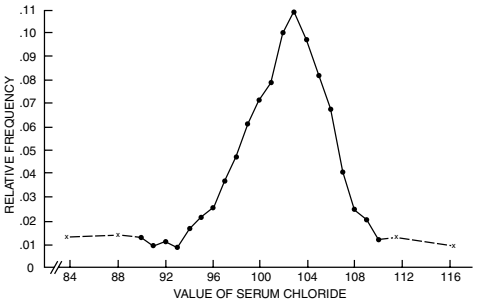


FIGURE 4.3
Frequency polygon showing the relative frequencies associated with the observed values of serum chloride listed in Table 4.1. The two extreme "tails" of the curve are shown with dotted lines because the points marked "×" are a composite of several values, as noted in Table 4.1.

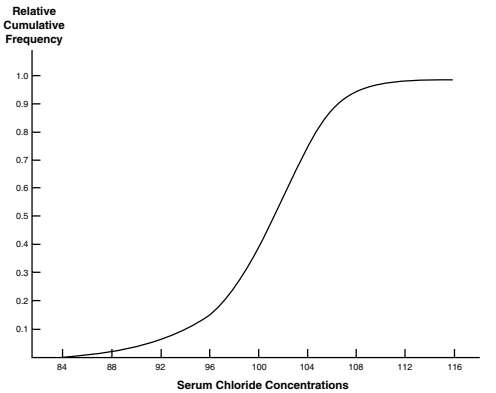


FIGURE 4.4
Ogive curve of ascending relative cumulative frequencies for the chloride concentrations shown in Table 4.1 and Figure 4.3.

With either Table 4.1 or Figure 4.4, we can immediately find the chloride value that corresponds to a selected percentile, and vice versa. In an ascending direction, the value of **102** occupies the cumulative proportions from .4256 to .5257. In a descending direction, **102** occupies the cumulative proportions from .4743 to .5743. In either direction, the cumulative proportion of .50 is passed within the value of 102, which is therefore the 50th percentile point or median. With a similar inspection, the chloride value of **99** contains the 25th percentile point in an ascending direction and the 75th percentile point in a descending direction.

4.2.5 Nomenclature

When a set is divided into hundredths, the demarcations can properly be called either *centiles* or *percentiles*. The former term is preferred by some writers, but the latter (employed throughout this text) is more commonly used and seems to be preferred in both ordinary and statistical dictionaries.

The percentile (or centile, if you wish) is regularly used to denote the inner location of an item of data in a dimensional spectrum. If a particular item occupies several percentiles, it is commonly cited with the "esthetic" or "conventional" demarcations that end in (or near) the digits 0 or 5. Thus, the

chloride value of 95, which occupies the ascending cumulative relative frequencies between .0844 and .1054 in Table 4.1, could be called the 9th percentile, but would usually be called the 10th.

4.2.5.1 Quantiles — Certain percentiles are given special names, called *quantiles*. The *tertiles* are the 33rd and 67th percentiles. The *quintiles* are the 20th, 40th, 60th, and 80th percentiles. The frequently used *quartiles* are the 25th, 50th, and 75th percentile points. These three quartiles are also respectively called the *lower quartile*, *median*, and *upper quartile*. The lower and upper quartiles are often symbolized as Q_1 and Q_3. (The median would be Q_2, but is usually shown with $\tilde{X}$ or some other special symbol.) The percentiles set at 2.5, 5.0, 7.5, …, 92.5, 95, 97.5, 100 each demarcate zones containing 1/40 of the distribution. Henrik Wulff[1] has proposed that they be called *quadragintiles*.

4.2.5.2 Quantile Zones — The names of certain percentiles are regularly applied erroneously to the demarcated zones as well as the boundary points. A set of k quantiles will demarcate $k + 1$ zones. Thus, the four quintile points demarcate five zones, each containing one fifth of the data. The three quartile points demarcate four zones, each containing one fourth of the data. These *zones* are really fifths and fourths, but they are often mislabeled as quintiles or quartiles.

4.2.5.3 Points, Ranks, and Values — An important distinction in nomenclature is the points, ranks, and values of percentiles.

A percentile *point* is a general location. Among 100 such points, it is regularly set (or found to be) at 20, 25, 50, 70, 75, or some other percentage of 100. In demarcating cumulative proportions of the data, these points would be denoted as .20, .25, .50, .70, and .75. When the term *percentile* is used without further specification, the reference is almost always to the point.

Percentile points will have different *ranks* within a set of data, according to the number of members in the distribution. The 50th percentile point (or median) occurs at rank 5 in a group of 9 items and at rank 14 in a group of 27.

A percentile *value* is the actual value of the particular item (or interpolated item) of data that is located at the percentile rank. This value will be 0.6, −12, 78.3, 1.9×10^6 or whatever occurs at that rank in the observed collection of data.

The following statement indicates all three uses of the terms: "He scored 786, which placed him in the 95th percentile, because he ranked 60th among the 1206 candidates who took the test." In this instance, the percentile *value* is 786, which is the actual score in the test. The percentile *rank* is cited directly as 60. The percentile *point* is determined from the proportion 60/1206 = .0498. Since high scores are usually reported with reciprocal percentiles, this proportion becomes converted to 1 − .0498 = .9502, or the 95th percentile.

4.3 Calculating Ranks, Zones, and Percentile Points

If r is the rank of an item in an n-item data set, the percentile point, P, occurs when the cumulative proportion, r/n, just exceeds P; i.e., (r/n) > P or r > Pn. Consequently, if the product Pn = r is exactly an integer, the selected rank must exceed Pn. It will have an intermediate rank that lies between r = Pn and the next rank, r + 1. For example, the formula Pn = r for the 25th percentile point of a 28-item data set yields (.25)(28) = 7; and the point will lie between the 7th and 8th ranks. Similarly, for the median or 50th percentile in that set, Pn = r is (.50)(28) = 14, and the point occurs between the 14th and 15th ranks.

On the other hand, if Pn = r is not an integer, the rank is *within* the next item. Thus, for the 5th percentile in a set of 28 items, Pn = r = (.05)(28) = 1.4. The 5th percentile is within the second rank.

This process will produce symmetrical results, working from either end of the data. With the proportion formula for the 95th percentile in a set of 28 items, we get r = (.95)(28) = 26.6, which becomes the 27th ranked item, counting from the lower end. This item will correspond to the 2nd ranked item (i.e., 5th percentile) counting from the upper end.

4.3.1 Calculating Ranks from Percentile Points

To calculate a rank from a percentile point, the simplest approach is to use the r = Pn formula. If the result is not an integer, the rank is the next higher item. If the result is an integer, the rank lies between the surrounding items. Thus, if Pn = 8, the rank is between 8 and 9. If Pn = 8.3, the rank is within the 9th item.

4.3.2 Calculating Percentile Points from Ranks

Going in the other direction is not as easy because you start with a ranked integer, r, and cannot promptly use the foregoing rule to calculate r = Pn. After getting the value of P = r/n, however, you can use r = Pn to see whether an integer emerges. Thus, for the 12th rank in a 37 item data set, P = 12/37 = .32 and r = (.32)(37) = 11.8, which confirms that the 32nd percentile is within rank 12. For the 14th item in a 28 item data set, P = 14/28 = .5, and r = (.5)(28) = 14, which is an integer. Therefore the 50th percentile must lie between ranks 14 and 15. In zones, the cumulative proportions end at .464 (=13/28) for the 13th rank, at .500 (=14/28) for the 14th rank, and at .536 (=15/28) for the 15th. The individual zones are from .464 to .500 for the 14th rank and .500 to .536 for the 15th. Because the 14th and 15th ranks share the "edges" of the .500 cumulative proportion, the percentile point must lie between them.

4.3.3 Alternative "Statistical Method"

The foregoing discussion is straightforward and direct, and the proposed method will always do the correct job. This method, however, is not what is offered in most conventional statistical discussions of percentiles. Accustomed to the theoretical collection of "continuous" data in Gaussian curves, the statistical approach does not deal with the problems of data sets that contain a finite number of individual items. Accordingly, in most statistical discussions, percentiles are used to demarcate intervals and interval boundaries, rather than the zones demarcated by those boundaries.

For large data sets (e.g., n > 100) or for the continuous curves that seldom occur in medicine but that are a common substrate in statistical reasoning, no real distinctions arise between the interval and the zonal approaches. In small data sets, however, where each item can occupy a substantial zone, the customary statistical approach will sometimes give slightly different (and erroneous) results. The next few subsections describe the statistical concepts and the reasons why the statistical formula usually appears as r = P(n + 1) rather than r = Pn. If you are content with what you already know about percentiles, skip to section 4.4. If you want to know about the customary statistical reasoning, read on.

4.3.3.1 Concept of Intervals — A data set containing n members will have n + 1 interval boundaries. Consider the alphabet as a 26-member data set symbolized as

a, b, c, d,…, w, x, y, z.

There are 25 or n − 1 interval boundaries between each pair of adjacent members; and there are 2 boundaries at each end, separating the first and last members from the outside world. Thus, the total number of interval boundaries is 27, or n − 1 + 2 = n + 1. The first of these boundaries is the 0th percentile, which never actually occurs. Each letter then occupies 1/26 or .038 of the zones of this data set, and so the sequence of cumulative proportions is .038, .076, .115,... until 1 is reached after the letter z.

4.3.3.2 Use of n + 1 Formula — To convert back and forth from ranks to percentile points, the interval-based formula is r = P(n + 1), where n is the number of items in the data set, P is the percentile point (expressed in hundredths), and r is the corresponding rank.

With this approach, if P(n + 1) is an integer, it is the rank. Thus, the median (or 50th percentile) rank in a set of 29 items is r = (.50)(30) = 15. If P(n + 1) is not an integer, the percentile lies between the two surrounding ranks. Thus, for the lower quartile in a set of 40 items, (.25)(41) = 10.25, and so Q_1 will lie between the 10th and 11th ranks.

The location of ranks for the two methods of calculations are shown in Table 4.2. For the two examples just cited, the results are the same with either formula. Thus, the proportion formula, r = Pn, yields (.50)(30) = 14.5, which becomes 15 for the median of a set of 29 items, and (.25)(40) = 10, which places Q_1 between ranks 10 and 11 for a set of 40 items.

TABLE 4.2

Location of Rank with Two Formula for Percentiles

Result of Calculation	Proportion: r = Pn	Interval: r = P(n + 1)
Integer (exactly r)	Between r and r + 1	r
Non-integer (r +…)	r + 1	Between r and r + 1

4.3.3.3 Problems in Small Data Sets — The r = P(n + 1) formula will work well for finding percentile ranks in large data sets, but can lead to difficulties in small data sets where the cumulative proportions and boundaries of intervals are not always the same.

For example, consider the five ranked items in the data set {a, b, c, d, e}, which has 6 interval boundaries, numbered as

1 2 3 4 5 6
| a | b | c | d | e |

The median item, c, contains the .50 cumulative proportion; and it also will split the six interval boundaries evenly so that three of them are on each side of c. With the interval P(n +l) formula the ranked location of c is (.50)(6) = 3. The rank of the 25th percentile point, or lower quartile, however, would be (.25)(6) = 1.5, which suggests a location between the first and second ranks, which in this instance are the items a and b. In reality, however, the cumulative portion of .25 in a five-item data set occurs *within* the second item of data, which occupies the proportions from .20 to .40. Thus, the P(n + 1) calculation yields something between a and b for the lower quartile, but the cumulative proportion shows that it actually occurs within b.

Correspondingly, the upper quartile, determined for the cumulative proportion of .75, will lie within d, which occupies .60 to .80 of the data. If the percentile is calculated as a split interval, however, the upper quartile will lie between d and e.

4.3.3.4 No Problem in Large Data Sets — For large data sets, the difference in interval vs. proportion methods of calculation becomes trivial because the results will usually be the same whether the data are apportioned as n items or split into n + 1 intervals. Consequently, the P(n + 1) method is often preferred because it seems easier to work with.

For example, the 5th percentile value for the 2039 items in Table 4.1 is in the 102nd rank whether calculated as a cumulative proportion, (.05)(2039) = 101.95, or as an interval marker, (.05)(2040) = 102. In the table itself, the serum chloride values from 75 to 91 occupy 100 observed frequencies (=8 + 12 + 12 + 12 + 10 + 10 + 28 + 20), which are the ranks from 1 to 100. The chloride level of **92** has 22 observed frequencies, which occupy the ranks from 101 to 122. Thus, the 102nd rank occurs at a chloride level of **92**.

Analogously, to get the 2.5 percentile value for the 2039 items in Table 4.1, P(n) would produce 50.975 and P(n + 1) would produce 51. With either calculation, the percentile value would occur in the 51st rank. As shown in the table, the chloride levels from **75** to **88** occupy the first 42 ranks (=8 + 12 + 12 + 10). The chloride level of **89** occupies the next 10 ranks, from 43 to 52, and so the 2.5 percentile is at the chloride level of **89**.

In small data sets, however, the results of the two methods may differ. Being immediately anchored to the data, the proportion technique is always preferable for small data sets.

4.3.4 Management of Values for Inter-Rank Points

If the calculated rank lies between two integer ranks, the easiest way to choose the appropriate percentile value is to split the difference in values at those two ranks. Thus, if the calculated rank lies between the two values of **a** and **b,** the percentile value can be set at (a + b)/2. This result is equivalent to calculating a + .5[(b − a)].

A preferable method of getting the inter-rank value, however, is to adjust the difference according to the "weight" of the percentile. With this technique, the inter-item value for the 20th percentile would be a + .20(b − a) = .80a + .20b. The calculation is a + P(b − a) where P is the percentile in hundredths. In the opposite direction, if the 80th percentile lies between a and b, the formula for its value would be b − .80(b − a) = .80a + .20b, which produces the same result as before. Thus, if the lower quartile lies between the values of 31 and 37, the quartile value will be (.75)(31) + (.25)(37) = 32.5. The alternative formula for the calculation would be 31 + (.25)(6) = 32.5.

This method of calculating the inter-rank values can be used regardless of whether the ranks are determined with the zonal formula r = Pn or with the interval formula r = P(n + 1). On the other hand, no great harm will be done if you ignore the distinction and simply split the value as (a + b)/2 rather than as a + P(b − a).

4.3.5 Choice of Methods

Because of the possible disparities, data analysts are sometimes advised not to use percentiles unless the data set contains at least 100 members. Because this advice would eliminate a valuable index for the many small data sets that constantly occur in medical research, a better approach is to recognize that the r = P(n) method is always more accurate for demarcating the *zones* that are usually a prime focus of analytic attention. The only disadvantages of the r = P(n) method are the need to remember (1) that an integer result implies an intermediate rank, (2) that a non-integer result implies the next higher rank and (3) that most statistical textbooks (ignoring the challenges of small data sets) cite r = P(n + 1) as a routine method of calculation. If in doubt about whether the data set is large enough, always use the r = P(n) method. It will be correct for small data sets, and should agree with the r = P(n + 1) result for large data sets.

Regardless of which method is used, the same percentile value should emerge for the rank calculated in both directions. Thus, the 2.5th, 5th, and 25th percentile values in one direction should be identical to the corresponding values that emerge, respectively, for the 97.5th, 95th, and 75th percentile ranks in the opposite direction.

4.4 Percentiles as Indexes of Inner Location

Despite the ambiguities of calculation for small data sets, the percentile technique is a simple method of immediately locating an item within a distribution. Low percentile points will denote a location at one end of the distribution; high percentiles will denote the other end; and the 40–60 region of percentile points will denote values near the middle.

4.4.1 Identifying Remoteness from Central Index

The percentile technique can be used to illustrate earlier comments (on the first page of this chapter) about two data sets, each containing 13 members, in which the value of **75** seemed close to or far away from the median of **80**.

Using the proportional technique to calculate percentiles for the small data sets, each item contains the proportions of 1/13 = .077 of the data. The first item occupies percentiles from 0 to 7.7. The fourth ranked item occupies the percentiles from 3/13 to 4/13, which contains the proportions from .231 to .308. The sixth ranked item occupies the percentiles from 5/13 to 6/13, containing .385 to .462 of the cumulative proportions. Using customary expressions, we could say that the first item is at the 5th, the

fourth item is at the 25th, and the sixth item is at the 40th percentile. Thus, in the first data set, **75** (as the sixth ranked item) is roughly at the 40th percentile, and in the second data set, **75** (as the first ranked item) is roughly at the 5th percentile.

[With the customary statistical (interval) method of calculation, the $r = P(n + 1)$ formula would produce the percentile point as $P = r/(n + 1)$. In the first data set, **75** is the sixth ranked member, and its percentile point would be $6/(13 + 1) = .43$. In the second data set, **75** is the first ranked member; and its percentile would be $P = 1/(13 + 1) = .07$. With either technique of calculation, the value of **75** is relatively close to the median of **80** in the first data set and relatively distant in the second.]

The percentile-point index shows where an item lies in the total distribution, but does not specifically designate distance from the center. Because the central percentile is at 50, the subtraction of $P - 50$ can promptly indicate remoteness from the center. The absolute size of the $P - 50$ difference, which can extend from 0 to 50, will indicate increasing remoteness, with negative values being below the center, and positive values above it.

4.4.2 Clinical Applications

Percentiles are commonly used by practicing pediatricians as a method for checking a child's growth. The patient's actual height and weight are located in a chart that shows the percentile distributions of these values in "standard" groups of normal children at different ages. If the child's percentile location for height and weight seems satisfactory for the age, the child is regarded as having normal development (at least in these two variables).

The height–weight percentiles can also be used for certain descriptive comparisons. If a child is at the 20th percentile for height and 80th percentile for weight, each percentile is within the limits of normal, but their relationship suggests that the child may be relatively too fat.

Percentiles are commonly used (as noted in a previous example) to rank candidates in a certifying examination, and can also be employed (as noted later) to compare the similarity or differences in data for two groups.

Some of the most valuable statistical applications of percentiles, however, are discussed in Chapter 5, where they are used to form zones of data.

4.5 Percentiles as Probabilities

Beyond their role in describing inner locations, percentiles can indicate probabilities for what might happen if members were randomly chosen from a distribution. For example, in the earlier data of Table 4.1, a chloride value of 99 has an ascending cumulative relative frequency of .2761; and 100 has a descending cumulative relative frequency of .7239. Therefore, if someone were randomly chosen from this distribution, the probability would be .2761, or about 28%, for getting a chloride value of ≤ 99 and about 72% for getting a chloride value of ≥ 100.

These probability values, as discussed later, have prime importance in the statistical applications of percentiles. In *descriptive* statistics, percentile boundaries are used to demarcate a range of "customary" or "normal" values. In *inferential* statistics, percentile zones of probability are used for decisions about "significant" distinctions.

4.6 Standard Deviation

The statistical entity called the "standard deviation" was originally devised and is best known for its role (discussed later in Chapter 5) as an index of dispersion. With certain mathematical principles, however, the standard deviation can become an index of inner location.

4.6.1 Definitions, Symbols, and Calculations

Usually symbolized as *s*, the standard deviation denotes the average deviation between the individual items of data and the mean. The ideas require some new symbols and terms. If X_i is a member of a data set having n members, with ΣX_i as the sum of values in the set, the mean is calculated as $\bar{X} = \Sigma X_i/n$. Each member of the data set deviates from the mean by the amount $X_i - \bar{X}$.

The obvious way to get the average of the deviations is to take their sum and divide it by n, the number of items in the data. Since some of the deviations in a set of data will lie above and others below the mean, however, the sum of the deviations, $\Sigma(X_i - \bar{X})$, will always be zero.

To avoid this problem, we can find the average deviation by eliminating the positive or negative signs, determining the sum of each absolute deviation, $\Sigma|X_i - \bar{X}|$, and then calculating the average absolute deviation as $\Sigma|X_i - \bar{X}|/n$. Although an excellent way to get average deviations, this method has not become statistically popular because the calculations were difficult in the days before electronic computation.

To escape from absolute values, and from positive and negative signs, the deviations are squared, as $(X_i - \bar{X})^2$. Their sum, which is always a positive number, is expressed symbolically as

$$S_{xx} = \Sigma(X_i - \bar{X})^2$$

[The reason for the xx subscript is that the calculation is really $\Sigma(X_i - \bar{X})(X_i - \bar{X})$. Later on, in Chapter 19, the calculation of $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$ for co-deviations between two variables is symbolized as S_{xy}.] S_{xx} has many synonymous names: *deviance, group variance, system variance*, and *sum of squares*. In strict statistical concepts, S_{xx} is the sum of squares in a "sample," and is used to estimate group variance in the parent population. This distinction is commonly neglected, however, in many discussions and computer printouts. Because both deviance and sum of squares can pertain to diverse computational arrangements, the calculation using the original mean of the data ($\bar{X}$) is easily and unambiguously cited as *group variance* — the term commonly used in this text.

In its role now, S_{xx} is merely an entity calculated enroute to a standard deviation. In many subsequent statistical activities, however, S_{xx} has an important job of its own.

4.6.2 Calculation of Group Variance: S_{xx}

The computation of S_{xx} can be simplified for hand calculators by using an algebraic expansion and reformulation of $\Sigma(X_i - \bar{X})^2$. It becomes

$$S_{xx} = \Sigma X_i^2 - n\bar{X}^2$$

or

$$S_{xx} = \Sigma X_i^2 - [(\Sigma X_i)^2/n]$$

The latter formula is preferable for hand calculation because "rounding" is avoided until the end of the calculations.

To illustrate the calculations of S_{xx}, consider three data sets, each containing three items, with **30** as the mean for each set. The calculations are as follows:

Data set	ΣX_i^2	$\bar{X}$	$n\bar{X}^2 = (\Sigma X_i)^2/n$	$S_{xx} = \Sigma X_i^2 - n\bar{X}^2 = \Sigma X_i^2 - [(\Sigma X_i)^2/n]$
29, 30, 31	$29^2 + 30^2 + 31^2 = 2702$	30	$(90)^2/3 = 2700$	2
25, 30, 35	$25^2 + 30^2 + 35^2 = 2750$	30	2700	50
1, 30, 59	$1^2 + 30^2 + 59^2 = 4382$	30	2700	1682

These results immediately show the important role of the squared original observations, ΣX_i^2, in contributing to the group variance. Each of the three cited data sets had the same mean, and hence the same values of $\overline{X}$, ΣX_i and $(\Sigma X_i)^2/n$. The source of the differences in the S_{xx} values was the value of ΣX_i^2.

Each of the original values, X_i, contributes X_i^2 to the sum ΣX_i^2, which is sometimes called the "original" or "uncorrected" sum of squares. If each value of X_i were replaced by the mean, $\overline{X}$, the new square for each value would be $\overline{X}^2$, and the sum of the n squared values would be $n\overline{X}^2$. The difference in the two sums of squares, i.e., $\Sigma X_i^2 - n\overline{X}^2$, represents the reduction or "correction" achieved in the original sum of squares when each of the original values is replaced by a "fitted model." In this instance, the fitted model is the mean.

An advantage of using the mean as an index of central tendency is that the sum of squared deviations, S_{xx}, is smaller with the mean than with any other "model" (such as the median) that might have been chosen as a central index. (The median, $\tilde{X}$, is best for minimizing the sum of absolute deviations as $\Sigma|X_i - \tilde{X}|$, but not for sums of squared deviations.) The idea of a smallest value for sums of squares may not seem particularly impressive now, but it becomes important later on, when we fit data with other models and evaluate the achievement by noting the reductions produced in sums of squared deviations from the model.

4.6.3 Formula for Repeated Frequencies

If each value of the data, X_i, occurs repeatedly with a frequency of f_i, the formula for calculating S_{xx} is

$$\Sigma f_i(X_i)^2 - (\Sigma f_i X_i)^2/N$$

where $N = \Sigma f_i$. For the data in Table 3.1, the group variance calculated in the usual manner is 3255. If each interval in Table 3.1 is replaced by its midpoint (so that **10–14** is represented by **12, 15 – 19** by **17**, etc.), the result could be approximated as $5(12)^2 + 19(17)^2 + 13(22)^2 + 4(32)^2 + 3(37)^2 + 2(42)^2 - [5(12) + 19(17) + 13(22) + 10(27) + 4(32) + 3(37) + 2(42)]^2/56$. The result is $31524 - (1262)^2/56 = 31524 - 28440 = 3084$, and is not too different from the more accurate calculation of S_{xx}.

4.6.4 Calculation of Variance

With S_{xx} as the group variance for the total of n members, we want an average value that is "standardized" for the size of the group. For this purpose, S_{xx} is divided either by n or by $n-1$ to produce a result called the *variance*, which is symbolized as s^2. The square root of the variance is s, the *standard deviation*.

In the three sets of data in Section 4.6.2, dividing each S_{xx} value by $n = 3$ produces the respective variances of .67, 16.67, and 560.67. The square roots of these variances are the respective standard deviations of .82, 4.08, and 23.68. If S_{xx} had been divided by $n - 1$, which is 2 in this instance, the respective standard deviations would be 1, 5, and 29.

4.6.5 Choice of Division by n or n – 1

A subtle issue in calculating the variance is whether to use n or $n - 1$ in the denominator. Because ordinary intuition suggests the use of n, you may wonder why $n - 1$ becomes a tenable or even desirable candidate for the divisor. The reason, as noted later when we consider the inferential strategies of statistics, is that division by $n - 1$ generally provides a better or allegedly "unbiased" estimate of variance in the hypothetical population from which the observed group is thought to have been "sampled."

For the moment, you can use the rule that if the data set is complete and is being summarized for purely descriptive purposes, S_{xx} is divided by n to yield the variance. On the other hand, if the standard deviation is going to be used for anything inferential — such as calculating the standard errors, confidence intervals, or P values discussed later in the text — S_{xx} is divided by $n - 1$.

If you are not sure about whether to use n or $n - 1$ in a particular situation, use $n - 1$. With a large data set, the results will be hardly affected. With a small data set, your statistical colleagues are more

likely to feel comfortable and to approve. For most purposes, therefore, the formula for calculating the standard deviation is

$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{\Sigma X_i^2 - [(\Sigma X_i)^2/n]}{n-1}}$$

If n is used as divisor, the variance becomes

$$S_{XX}/n = (\Sigma X_i^2 - n\bar{X}^2)/n = \Sigma(X_i^2/n) - \bar{X}^2$$

The latter (descriptive) formula is responsible for the statement that the variance is the "mean of the squares minus the square of the mean." With this formula, the standard deviation is

$$s = \sqrt{(\Sigma X_i^2/n) - \bar{X}^2}$$

4.6.6 Other Approaches

Calculations that use a central index can be avoided if an average deviation is determined for all possible pairs of values in the data set. If X_i and X_j are any two members in the data set, this average can be calculated from the squares of all differences, $(X_i - X_j)^2$, or from the absolute differences $|X_i - X_j|$. The latter approach is called *Gini's mean difference*. Although available and interesting, these "pairing" approaches are hardly ever used.

A different pairing tactic, called *Walsh averages*, can be applied for descriptions of non-Gaussian data sets. An array of means is calculated as $(X_i - X_j)/2$ for each pair of data, including each value paired with itself. For n items of data, there will be $n(n + 1)/2$ Walsh averages. The median of the array can be used as a central index, and the range of values above and below the median can help denote (as discussed later) the stability of the central index.

4.7 Standardized Z-Score

The inner location of any item in a distribution of data can be cited with a standardized Z-score, which is calculated as

$$Z_i = \frac{X_i - \bar{X}}{s}$$

and sometimes called a *standardized deviate*. The division by s produces a "standardization" that is unit-free, because Z_i is expressed in magnitudes of the standard deviation. The positive or negative value of each Z_i indicates whether the item lies above or below the mean.

To illustrate the standardizing effect, consider {9, 12, 17, 18, 21} as a data set for *age in years*. This 5-item set of data has $\bar{X} = 15.4$ and $s = 4.827$. If age were expressed in months, however, the same set of data would be {108, 144, 204, 216, 252}. For the latter expression, $\bar{X} = 184.8$ and $s = 57.924$. Thus, with a simple change in units of measurement, the same set of data can have strikingly different values for $\bar{X}$, s, and the deviations, $X_i - \bar{X}$. The values of Z_i, however, will be identical in the two sets. In the first set, Z_i for the first item will be $(9 - 15.4)/4.827 = -1.326$. In the second set, Z_1 for the first item will be $Z_1 = (108 - 184.8)/57.924 = -1.326$. Identical Z_i values will also occur for all other corresponding items in the two data sets.

4.7.1 Distribution of Z-Scores

In the two foregoing sets of data, the Z-scores for age are {−1.326, −.704, .331, .539, 1.160}. The scores have a mean of 0 and a standard deviation of 1. This attribute of any set of Z-scores is inherent in their calculation and has nothing to do with the shape of the distribution of data. Regardless of whether a distribution is symmetrical or skew, convex or concave, the Z-scores will have a mean of 0 and a standard deviation of 1. (A proof of this statement appears in Appendixes 4.1 and 4.2.) This remarkable feature gives Z-scores many statistical advantages.

4.7.2 Z-Scores as Indexes of Inner Location

The main advantage to be discussed now is that Z-scores can be used as indexes of inner location. An item's Z-score of −0.6, 1.2, 0, −2.3, etc. will immediately indicate how relatively far (in standard deviation units) the item lies above or below the mean.

The rightmost column of Table 4.1 shows the Z scores calculated as $Z_i = (X_i - 101.572)/5.199$ for each item of the chloride data, using their mean and standard deviation. These Z-scores, rather than percentiles, could cite the relative inner location of any item in the data. Note that most of the items are included in a zone of Z-scores that extends from –2 to +2. If values of $Z > |2|$ are excluded, the Z-score zone from −2 to +2 in Table 4.1 contains the chloride values from **92** to **111**. The zone will contain about 93% of the data, excluding .049 of the data at the low end and .0172 of the data at the high end.

4.7.3 Application of Z-Scores

The ability of a Z-score to denote inner location is exploited in diverse ways. A common procedure, which you have probably already encountered in previous scholastic adventures, is the use of Z-scores for educational tests, where the process is sometimes called "grading on a curve." The Z-scores can avoid calculation of percentiles for small data sets, and can immediately be used to rank the test takers for performance in tests that have different scoring systems.

Standard reference Z-scores[2] were used in a recent study of zinc vs. placebo supplements in breastfed infants[3] to compare each infant's status for length and weight before and after treatments. Z-scores have also been advocated[4] as a replacement for percentiles in determining the nutritional status of children, particularly in developing countries. The argument is that the rank of many malnourished children is difficult to establish as a percentile, because they are below the first percentile of the World Health Organization (WHO) reference population.[5] A "road to health" chart, shown in Figure 4.5, has been proposed as an illustration that can help health workers understand the Z-score procedure. Figure 4.6 illustrates the mean Z-scores of weight according to age in a population of children in a region of South Africa.[6]

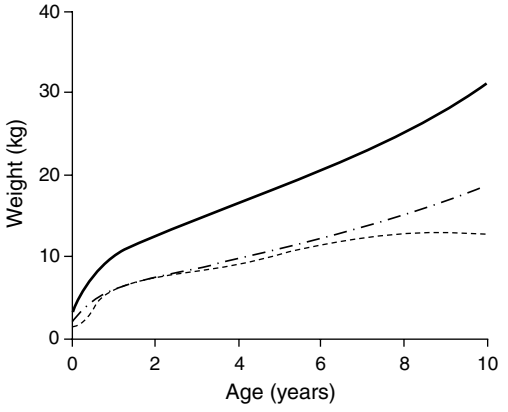


FIGURE 4.5
Z-scores for road-to-health charts. Comparison of 60% of median weight-for-age for males (–·–) with a 4 SD below the median (----); the bold line is the median. [Figure taken from Chapter Reference 4.]

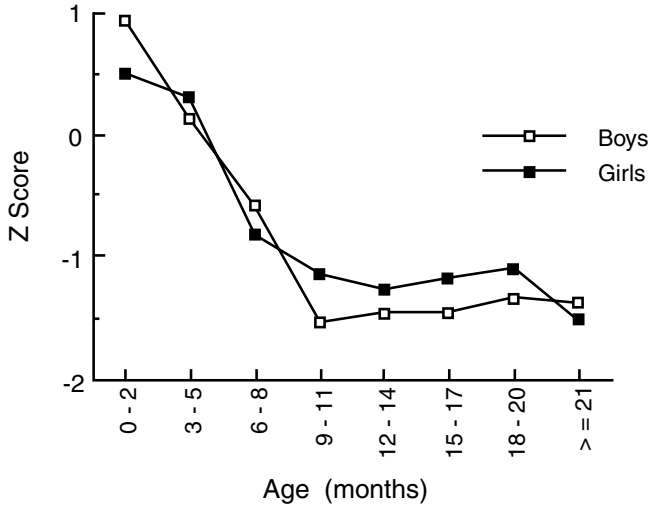


FIGURE 4.6
Mean weight-for-age Z-scores of 921 Basotho children by age and sex, Lesotho, 1985–1986. Z-scores were computed using the reference population of the National Center for Health Statistics. The child's weight was compared with the 50th percentile value of the reference populations and divided by the standard deviation of the references at that age. A Z-score of zero means that the child's weight is equal to the 50th percentile of the reference population. [Legend and figure taken from Chapter Reference 6.]

The use of Z-scores also avoids an undesirable technique, discussed in Chapter 5, which identifies inner locations according to the percentage by which each value is greater or less than the median value.

4.8 Conversion of Z-Scores to Probabilities

Compared with percentiles, Z-scores have one major disadvantage as indexes of inner location. By showing cumulative relative frequency, a percentile immediately denotes a specific location relative to other ranks; and the location can also be converted to a probability. Thus, a percentile of 95% tells us that only 5% of items have higher values and that the probability of getting one of those higher values is .05. A Z-score of 1.2, however, does not indicate either of these features. The array of Z-scores could be used to arrange a set of 5 items in ranked order as {−1.3, −.70, .33, .54, 1.2}, but we would not have known that 1.2 was the highest rank until we saw its location relative to the others.

This advantage can be overcome if the distributional shape of the data can be described with a conventional mathematical formula. When a suitable mathematical expression is available, each Z value can promptly be converted to a probability value for cumulative relative frequency. The eponymic mathematical distributions named after Gauss, Poisson, Bernouilli, etc. have been popular because of the correspondence between probability values and statistical indexes, such as Z.

4.8.1 Gaussian Distribution

The famous Gaussian distribution was promulgated in the 19th century by Carl Gauss, studying what today would be called "observer variability." As new technologic devices became available, the results found for repeated measurements of the same entity did not always agree. Confronted with the choice of which observation to accept, Gauss decided that their mean would be the correct value. He then noted that the deviations from the mean, called errors, were distributed around the mean in a symmetrical pattern that is today sometimes called *normal* by statisticians and *Gaussian* by people who want to avoid the medical connotations of *normal*.

Later in the 19th century, Adolphe Quetelet found that this same pattern occurred for the distribution of height in Belgian soldiers. This similarity of patterns helped give the Gaussian distribution its enthusiastic application in biostatistics of the 19th and early 20th centuries. The same pattern seemed to fit two different phenomena: multiple measurements of the same individual entity and individual measurements of multiple entities.

This dual accomplishment made the Gaussian distribution so mathematically attractive that Francis Galton, often regarded as the founder of modern biometry, believed it was a "Law of Frequency of Error." He expressed his admiration by saying, "the law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion."[7]

The diverse technologic measurements of the 20th century, however, have demonstrated that most medical data do not have Gaussian distributions. Consequently, clinical investigators have had to think about other distributions and to use statistical strategies that do not depend on theoretical mathematical patterns. Nevertheless, despite the *descriptive* inadequacies, the Gaussian distribution remains valuable for *inferential* activities that will be discussed later.

4.8.2 Gaussian Mathematics

Although popularized by Gauss, the mathematical formula for the distributional pattern was actually discovered by Abraham DeMoivre. In that formula, the height of the curve shows the relative frequency, y, at each Z_i score. The mathematical expression is essentially

$$y = .4e^{-\frac{1}{2}Z_i^2}$$

where e is the well-known mathematical constant, 2.7183…. This formula describes a *standard* Gaussian curve, with mean = 0 and standard deviation = 1.

In the standard Gaussian curve, the height of the curve, y, is the relative frequency for each possible value of Z. Thus, when Z = 0, i.e., at the mean, $y = .4e^0 = .4$. When Z = ±.5, $y = .4e^{-\frac{1}{2}(.25)} = .35$. When Z = ±2.0, y = .054. The corresponding values of y and z are as follows:

Z	0	±.25	±.50	±.75	±1.0	±1.25	±1.50	±1.75	±2.0	±2.25	±2.5	±3.0	±3.5
y	.4	.39	.35	.30	.24	.18	.13	.087	.054	.032	.018	.004	.0009

The pattern of these points is shown in Figure 4.7. Note that the height of the curve becomes infinitesimal for Z values that exceed ±3, but the height never reaches zero. For example, if Z = 6, $y = 6.092 \times 10^{-9}$.

When Gaussian curves are expressed in values of X_i, rather than Z_i, their mean will be at $\overline{X}$ and their relative width will be determined by the size of the standard deviation, s. The value of Z_i in the foregoing Gaussian formula will be replaced by $(X_i - \overline{X})/s$ and the .4 factor will be divided by s, thus making the curve wider and shorter as s enlarges, and narrower and taller as s gets smaller.

4.8.3 Citation of Cumulative Probabilities

With integral calculus, the formula for the Gaussian curve can be converted to show the cumulative rather than individual relative frequencies at each value of Z_i. With this conversion, a specific probability value of cumulative relative frequency is associated for each of the possible negative and positive Z values.

The cumulative relative frequencies are usually expressed as proportions (or percentiles) that begin with 0 at Z = 0, and enlarge as Z becomes larger in either a positive or negative direction. In one mode of expression, the cumulative proportions on either side of 0 can be cited as follows:

Z	0	±.25	±.5	±.75	±1.0	±1.5	±2.0	±2.5	±3.0	±3.5
Cumulative Proportion	0	.099	.191	.27	.34	.43	.477	.494	.498	.499

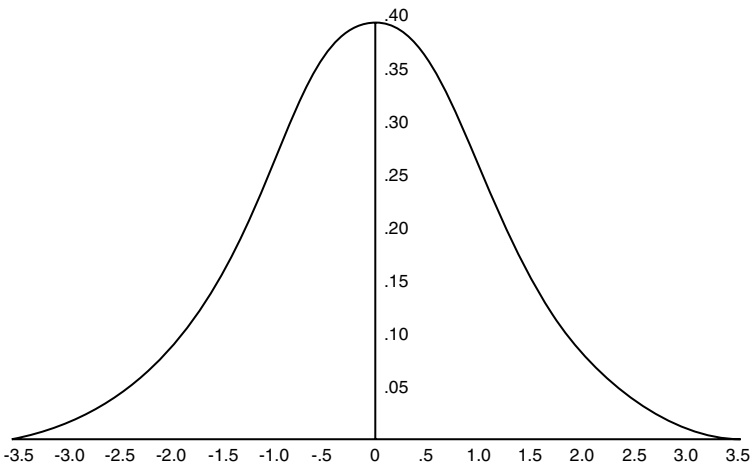


FIGURE 4.7
Values of y axis for Z-scores in Gaussian curve.

A more detailed account of the corresponding results is shown in Table 4.3. If you want to cumulate these proportions starting from the lower end of a Gaussian distribution and going upward, remember that the value of .5 should be added or subtracted appropriately for each of the foregoing values of the corresponding proportions. Thus, the Z value of –3.08, with a cumulative relative frequency of .4990 on the left side of the distribution, becomes .5 – .4990 = .001, which is the .1 percentile. The corresponding results for converting Z values to cumulative relative proportions are –.6 → .5 – .2257 = .2743, –.26 → .5 – .1026 = .3974, 0 → .5, .26 → .599, .6 → .7257, …, 2.0 → .9772, …, 3.6 → .999.

Although the cumulative proportions (and probabilities) can be used like percentiles merely to identify the inner locations of a distribution, their most frequent application occurs as P values in statistical inference. The P values are created when inner locations and probabilities for individual items in the data are demarcated into *zones* of location and probability. The demarcations and applications are discussed in the next chapter.

4.9 Inner Locations for Non-Dimensional Data

The percentiles and standard deviations that are pertinent for dimensional data cannot be used directly and must be modified to rank inner locations for the categories of non-dimensional data.

4.9.1 Nominal Data

Questions about the rank of an inner location do not arise for nominal data, which cannot be ranked.

4.9.2 Ordinal Data

Percentiles can easily be determined for ordinal data, although the same values will occupy many percentiles if the data are expressed in grades. For example, consider the 39 people whose therapeutic responses were rated as **poor**, 6; **fair**, 12; **good**, 11; and **excellent**, 10 in Section 3.5.1.3. Because each item occupies 1/39 = .026 of the data, the poor group would include the percentiles from 0 to 15 (since 6/39 = .154). The **fair** group occupies the subsequent percentiles up to 46 (since 18/39 = .462); the **good** group extends to the 74th percentile (since 29/39 = .744); and the **excellent** group occupies the rest.

Although not appropriate for ordinal data, standard deviations could be calculated if the four grades were regarded as dimensions (such as **1, 2, 3, 4**) and summarized with means.

TABLE 4.3

Cumulative Normal Frequency Distribution (Area under the standard normal curve from 0 to Z)

Z	0.00	0.02	0.04	0.06	0.08
0.0	0.0000	0.0080	0.0160	0.0239	0.0319
0.2	.0793	.0871	.0948	.1026	.1103
0.4	.1554	.1628	.1700	.1772	.1844
0.6	.2257	.2324	.2389	.2454	.2517
0.8	.2881	.2939	.2995	.3051	.3106
1.0	.3413	.3461	.3508	.3554	.3599
1.2	.3849	.3888	.3925	.3962	.3997
1.4	.4192	.4222	.4251	.4279	.4306
1.6	.4452	.4474	.4495	.4515	.4535
1.8	.4641	.4656	.4671	.4686	.4699
1.9	.4713	.4726	.4738	.4750	.4761
2.0	.4772	.4783	.4793	.4803	.4812
2.1	.4821	.4830	.4838	.4846	.4854
2.2	.4861	.4868	.4875	.4881	.4887
2.3	.4893	.4898	.4904	.4909	.4913
2.4	.4918	.4922	.4927	.4931	.4934
2.6	.4953	.4956	.4959	.4961	.4963
2.8	.4974	.4976	.4977	.4979	.4980
3.0	.4987	.4987	.4988	.4989	.4990
3.2	.4993	.4994	.4994	.4994	.4995
3.4	.4997	.4997	.4997	.4997	.4997
3.6	.4998	.4999	.4999	.4999	.4999
3.9	.5000				

Note: Abridged table showing correspondence of Z-scores (shown in first column, augmented by values in subsequent columns) and unilateral cumulative proportions (shown in cells) of Gaussian curve in zone from 0 to Z. Values not shown here can be obtained by interpolation.

4.9.3 Binary Data

A set of binary data, expressed as **yes/no**, **success/failure**, **alive/dead**, etc. can be coded as **1/0**. If the set contains n items, r items will be coded as **1** and n − r items will be coded as **0**. The binary proportion that summarizes the data will be either p = r/n or q = (n − r)/n.

The percentiles of such a data set are relatively trivial. The first r values of **1** in the set **1, 1, 1, …, 1, 0, 0, …, 0** will occupy r/n = p or the first P percentiles. The standard deviation of the data set, however, is not trivial. It becomes particularly important later on when we consider the "standard error" of a proportion.

The standard deviation of a set of binary data can be calculated with the same formula used for dimensional data. In r items of the data, the deviation is (1 − p) from the "mean" value, p; and n − r items will have the deviation (0 − p). Thus, the group variance will be $r(1 - p)^2 + (n - r) \times (0 - p)^2$. Letting 1 − p = q, r = pn, and n − r = qn, this expression becomes $pnq^2 + qnp^2 = npq(q + p) = npq$.

The group variance is customarily divided by n − 1 to form the variance of dimensional data, but by n for binary data. (The reason will be explained later.) Accordingly, the variance of a set of binary data is npq/n = pq, and the standard deviation is

$$s = \sqrt{pq}$$

Note that this remarkable formula refers to standard deviation of the data, not to variance of the binary proportion that is the central index. In other words, each of the **1** or **0** values in the data deviates from

the central index, p, by an average of $\sqrt{pq}$. This same average deviation occurs if we express the central index as q = 1 − p, rather than p. Thus, for the 12 values of **1** and 7 values of **0** that are summarized as p = 12/19 = .63 or as q = 7/19 = .37, the standard deviation is $\sqrt{(12/19)(7/19)} = \sqrt{.233} = .48$. This same standard deviation would occur if the data set were three times larger, with p expressed as 36/57 = .63.

Appendixes for Chapter 4

A.4.1 The Mean of a Set of Z-Scores is 0

Proof: If the original data are represented by $\{X_i\}$, $\bar{X}$, and s, each Z-score will be $Z_i = (X_i - \bar{X})/s$. Then $\Sigma Z_i = \Sigma(X_i - \bar{X})/s = (1/s)[\Sigma X_i - \Sigma\bar{X}] = (1/s)[N\bar{X} - N\bar{X}] = 0$.

A.4.2 The Standard Deviation of a Set of Z-Scores is 1

Proof: The group variance of the Z scores will be $\Sigma(Z_i - \bar{Z})^2 = \Sigma Z_i^2$, since $\bar{Z} = 0$. Since $Z_i = (X_i - \bar{X})/s$, each Z_i^2 will be $(X_i^2 - 2X_i\bar{X} + \bar{X}^2)/s^2$ and $\Sigma Z_i^2 = (\Sigma X_i^2 - n\bar{X}^2)/s^2$. If calculated with n − 1, the variance of the Z scores will be $\Sigma Z_i^2/(n-1) = (\Sigma X_i^2 - n\bar{X}^2)/[(n-1)s^2]$. Since $(n-1)s^2 = \Sigma X_i^2 - n\bar{X}^2$, the quotient for the variance is 1. Its square root, 1, will be the standard deviation.

References

1. Wulff, 1981; 2, Hall, 1993; 3. Walravens, 1992; 4. Shann, 1993; 5. WHO Working Group, 1986; 6. Ruel, 1992; 7. Galton, 1889.

Exercises

4.1. For the data in Table 3.1 (Chapter 3),
- 4.1.1. What are the lower and upper quartile values?
- 4.1.2. What values are at the 2.5 and 97.5 percentiles?
- 4.1.3. In what percentile is the value of **30**?

4.2. For the data set in Exercise 3.3 (Chapter 3),
- 4.2.1. What is the standard deviation calculated with n and with n − 1?
- 4.2.2. Calculate the Z-scores for the two lowest and two highest items of the data. What do these Z-scores tell you about whether the distribution is Gaussian?

4.3. What feature of the Z values in Table 4.1 indicates that the chloride data do *not* have a Gaussian distribution?

4.4. Students graduating in the 25th percentile of "performance" at Medical School X had higher MCAT scores on admission than students graduating in the 75th percentile of performance at Institution Y. Assuming that performance has been determined and graded in the same way at both institutions, does this result indicate that MCAT scores are poor predictors of performance in medical school? If your answer is No, give an alternative possible explanation for the observed results.

4.5. In the certifying examination of American specialty boards (in Medicine, Pediatrics, Surgery, etc.), the candidates receive a total "raw" numerical score based on right and wrong answers to the questions. For psychometric purposes in the theory of educational testing, the Boards transform these results so that each group of candidates will have a mean score of 500, with a standard deviation of 100. How are the raw scores altered to accomplish this goal?

4.6. As the Dean of Students for your Medical School, you must prepare each student's class ranking in clinical work. The ranking, which will be used for internship and other recommendations, comes from a combination of grades for clinical clerkships in five departments. Each grade is given an equal weight in the student's "class standing." The five clinical departments express their individuality by using different scales for giving grades. The scales are as follows:

Internal Medicine: A,B,C,D,E, (with A = highest and E = lowest)
Obstetrics-Gynecology: A+, A, A–, B+, B, B–, C+, C, C–, D, and E
Pediatrics: Numerical examination grade from 100 (perfect) to 0 (terrible)
Psychiatry: Superior, Satisfactory, Fail
Surgery: High honors, Honors, Pass, Fail

How would you combine these non-commensurate scaling systems to form a composite score for class standing?

4.7. Please mark each of the following statements as true or false (correct or incorrect). If you disagree, indicate why.

4.7.1. Harry had a percentile score of 70 in a board certification test. The result implies that he correctly answered 70% of the items.

4.7.2. At the beginning of the year, Mary had a percentile score of 90 in class standing. By the end of the year, she had moved to a percentile score of 99. During the same period, John moved from the 50th to 59th percentile. The two students, therefore, made about equal progress.

4.7.3. There is little difference between Joan's score at the 98 percentile and Bob's score at the 99.9 percentile, but a large difference between Bill's 84th percentile and Joan's 98th.

4.7.4. Since Nancy scored at the 58th percentile and Dick at the 64th, Dick obviously did much better in the test.

4.7.5. The Dean of Students at Almamammy medical school wants to evaluate secular trend in the students' performance on National Board examinations. He believes the trend can be correctly determined if he forms an average of each student's overall percentile score for members of last year's class and uses it to compare this year's performance.

4.8. A pediatrician who likes to be different decides to use Z-scores rather than percentiles for characterizing the status of patients. What would be the approximate Z-scores associated with the following percentiles: 2.5, 25, 40, 50, 60, 75, 97.5?

4.9. Joe has brought a lawsuit against the American Board of Omphalology because he failed the certifying examination, although he knows (from information sent by the Board) that he correctly answered 72% of the items. What would your argument be if you were Joe's lawyer? If you were the Board's lawyer, how would you defend it?

4.10. If you enjoy playing with algebra, and want to check your ability to manipulate statistical symbols (and also have the available time), here are two interesting but optional assignments.

4.10.1. Show that $\Sigma(X_i - \overline{X})^2 = \Sigma X_i^2 - N\overline{X}^2$

4.10.2. A sum of squares or group variance can be calculated as $\Sigma(X_i - M)^2$ with any selected central index, M. Prove that this group variance is a minimum when $M = \overline{X}$ (the mean). (Hint: Let $M = \overline{X} \pm c$, where $c > 0$. Then see what happens. Another approach, if you know the calculus, is to differentiate this expression in search of a minimum.)

5

Inner Zones and Spreads

CONTENTS

We now know how to indicate a group's external location and the relative internal location of individual members, but we do not yet have a way to denote a group's *spread*. This chapter is devoted to indexes of spread and to the demarcation of several valuable *inner zones* of spread. Their important roles are to denote the compactness of the group and the adequacy with which the group is represented by its central index.

5.1 The Range

A simple, obvious index of spread for a set of dimensional data is the range from the lowest to highest values. In Table 3.1, where the data extend from values of **11** to **43**, the range can be cited as **11–43**. For the chloride values in Table 4.1, the range (noted from the footnotes in that table) is **75–127**.

The two extreme boundaries of a range are usually listed individually and are seldom subtracted to produce such single values as **32** or **52**.

5.1.1 Non-Dimensional Ranges

With non-dimensional data, a formal expression is seldom needed, because the range is readily evident when the array of categories in each variable's distribution is summarized with frequency counts or relative frequencies.

For a binary variable, the range is obvious: it covers the two categories for **yes/no**, **present/absent, success/failure, 0/1**, etc. For ordinal data, the range extends from the smallest to largest of the available grades or ranks. The ordinal grades would have ranges of **I–III** for the three TNM Stages **I, II,** or **III,** and **1–10** for the 10 values of the composite Apgar Score. If not expressed as a set of counted ranks, an ordinal variable seldom has more than ten grades.

Not having ranked magnitudes, the categories of nominal variables cannot be shown as a range. A nominal variable, such as *occupation*, might contain twelve categories arbitrarily coded as **01, 02, …, 12**, but would not be summarized as a range of **01–12**.

5.1.2 Problems in Dimensional Ranges

The main problem in using range for the spread of dimensional data is potential distortion by outliers. This problem, which keeps the mean from being an optimal central index, may also make the range an unsatisfactory index of spread. For example, a range that extends 12 units from **19 to 31** for an array of 200 items would suddenly have its spread become five times larger (from 12 units to 60) if the data set contains a single outlier value of **79**. The range thus offers a useful "quick and dirty" way to appraise the spread of the data, but some other index will be preferred to avoid the potential adverse effect of outliers.

5.2 Concept of an Inner Zone

To eliminate the problem of outliers, the spread of dimensional data can be truncated and converted to a smaller inner zone. The goal is to exclude the extreme values but to have the inner zone contain either most of the "bulk" of the data or a selected proportion thereof. [The "bulk" of the data refers to individual items, not the "weight" or "magnitude" of the values of those items. Thus, the data sets {1, 3, 6, 7} and {95, 108, 234, and 776} each have a bulk of four items.]

The location and magnitude of inner zones are chosen arbitrarily. Some of them can readily be demarcated in a relatively "natural" manner. For example, because 25% of the items of data are contained in each zone demarcated by the 25th, 50th, and 75th percentiles, the inner zone between the lower and upper quartiles will comprise 50% of the data, with each half of that zone containing 25%. With quartile demarcations, the **bulk** of items in the inner 50% zone is symmetrically distributed around the median, but in other circumstances, the selected inner zone may be located eccentrically, or may hold a much larger "bulk" than 50%.

The decisions about symmetry in location and quantity of "bulk" will depend on the purpose to be served by the inner zone.

5.2.1 Symmetry in Location

In many circumstances, the low and high values of a data set are relatively well balanced at the two ends of the spectrum. This reasonably symmetrical pattern, which occurs in the chloride data of Table 4.1, could allow bidirectional decisions that certain values are "too low" and others are "too high." An inner zone for the "range of normal" would therefore be located symmetrically (or concentrically) around the center of the data.

In other instances, however, the data may be overtly unbalanced or skewed, with most of the items occurring eccentrically at one end of the distribution. For example, in a large group of people examined routinely at a hospital emergency department, most of the levels of blood alcohol would be **0**. On various occasions, the patients' values would be high enough to make the mean for the total group exceed **0**, but the "range of normal" for blood alcohol would be placed eccentrically, beginning at the **0** end of the data.

Even if the spectrum is relatively symmetrical, however, we may sometimes be interested in only one side of the distribution. For example, we might want to know the probability of finding a person with hyperchloridemia in Table 4.1. For this purpose, the "inner zone" would be located eccentrically. It would extend from the lowest chloride value in the table to an upper boundary chosen as demarcating a "too high" chloride. If the latter boundary is set at **109**, the inner zone would start at **75** (the lowest chloride value in Table 4.1) and extend through **108**. The zone would include .9436 of all values in the data. The "outer zone," at chloride levels of **109** onward, would include the remaining .0564 proportion of the group. The probability would be .0564 that a particular person has a value in the zone denoted by chloride levels $\geq$ **109**.

Inner zones are most commonly set symmetrically with two boundaries that form an external zone at each end of the distribution. The two external zones beyond the boundaries are often called "tails" when the results are cited as probability values. Sometimes, however, the inner zone has only one demarcated boundary, and the result (as in the example just cited) is called "one-tailed" for probability.

5.2.2 Decisions about Magnitude of "Bulk"

A separate decision refers to the quantity of the distribution that will be regarded as the "bulk," "great bulk," or "most" of all the items in the data. For some comparative purposes, the "bulk" may be set at 50%, 80%, or 90%. For conclusions about uncommon, unusual, or "abnormal" members of a distribution, the bulk is often set at 95%.

The choice of 95% is completely arbitrary. It has no inherent biologic or logical virtue and could just as well have been set at 83%, 94%, 96%, or some other value. The current custom arose when R.A. Fisher, a prominent leader in 20th century statistics, noted that 95% of the data in a Gaussian distribution was contained in a symmetrical inner zone bounded by 1.96 standard deviations on both sides of the mean. Because the "1.96" could easily be remembered as "2," a convenient mnemonic symbol was constructed. With $\overline{X}$ as mean and s as standard deviation, $\overline{X} \pm 2s$ became the symbol that designated the 95% inner zone of a Gaussian distribution. Fisher then referred to the remaining values of data, located in the two segments of the outer zone, as highly uncommon or unusual.

5.3 Demarcation of Inner Zones

After the decisions about symmetrical location and magnitude of bulk, inner zones can be demarcated with three tactics: percentiles, standard deviations, or Gaussian Z-scores.

5.3.1 Percentile Demarcations

An inner zone demarcated with percentiles is called an *inner-percentile range*, abbreviated as **ipr**. It is centered at the median, and its bulk is indicated with a subscript, such as ipr_{95} for a zone covering 95%