INTRODUCTION TO PROBABILITY and STATISTICS

SIXTH EDITION

ALDER / ROESSLER

INTRODUCTION TO
Probability and Statistics

SIXTH EDITION

Henry L. Alder Edward B. Roessler
UNIVERSITY OF CALIFORNIA, DAVIS

W. H. FREEMAN AND COMPANY
San Francisco

Cover illustration: A computer-generated figure composed of bivariate normal distribution functions with positions, widths, and amplitudes chosen randomly. From *Computer Graphics: 118 Computer-Generated Designs* by Melvin L. Prueitt. Dover Publications, Inc., 1975.

Library of Congress Cataloging in Publication Data

Alder, Henry L
Introduction to probability and statistics.
Bibliography: p.
Includes index.
1. Probabilities. 2. Statistics.
I. Roessler, Edward Biffer, joint author.
II. Title.
QA273.A43 1976 519.2 76-13643
ISBN 0-7167-0467-6

Printed in the United States of America

AMS 1970 subject classifications: 60-01, 62-01

1 2 3 4 5 6 7 8 9

Preface to the Sixth Edition

This book has been written to serve as a textbook for a one-semester introductory course in probability and statistics. It is also suitable for either a quarter course meeting four hours a week for ten weeks or a two-quarter course meeting three hours a week. The first edition of the text was a revision of the mimeographed and lithoprinted forms used for several years by instructors on the Davis campus of the University of California and at some other institutions. The many hundreds of students in these classes came from almost all fields of specialization in the natural and social sciences, and a few from the humanities.

In the typical American college or university, introductory statistics courses are offered in many different departments, are directed toward a great number of different specializations, and require widely differing levels of mathematical preparation. This phenomenon might lead to the mistaken impression that basic statistics is essentially different in the various fields of application. In this respect, instruction in statistics seems to occupy a unique position; in almost every other science there exists one introductory course with a fairly well-defined amount of subject material and anyone requiring some background in that science takes that course. Thus, almost every college or university offers an introductory chemistry course in which all students who need some knowledge of chemistry enroll. There seem to be no reasons except purely administrative or traditional ones why statistics should not be treated similarly. Indeed, much benefit could be derived from centralizing the teaching of basic statistics in one course in one department, which, in

most schools, would be either the mathematics or the statistics department. If only one such course is offered, much needless duplication can be avoided and a single standard of achievement can be maintained. It is for such a course that this book is designed.

Mathematical knowledge equivalent to two years of high school algebra is all that is required as background for this textbook. At one point in Chapter 12 trigonometry is used, but readers without training in that subject may omit this passage without loss of continuity. The limited level of mathematical preparation required for this book makes it necessary to state some theorems without proof. However, almost all theorems whose proofs depend only on mathematics as commonly taught in two years of high school algebra are proved. The reader may be surprised to find that the number of theorems not proved here, because of the limited mathematical preparation assumed, is relatively small, and he may find consolation in the fact that his understanding of the omitted proofs would not be increased appreciably even after a course or two in beginning calculus. To avoid the mathematical difficulties associated with limiting processes, with which the reader ordinarily will not become acquainted until he studies calculus, discussions have been restricted, wherever possible, to the finite case; in particular, all populations are assumed to be finite.

The students who enroll in the statistics course at Davis come from all the agricultural, biological, and medical sciences, business administration, economics, home economics, psychology, sociology, and geology. To allow for these differing fields of specialization, the examples and exercises have been chosen from all of these subject areas.

Exercises are an integral part of the text. It has been our experience that an introductory course in statistics is taught most effectively in the same way as are other introductory courses in mathematics, a number of homework problems being assigned after every lecture. Special care has been taken to insure that exercises involve a minimum of computation, and that none (except for a few in Chapter 15) requires machine calculation. This has often necessitated the inclusion of fictitious data, although those in many exercises represent the results of actual experiments.

Although most students taking the beginning course in statistics on the Davis campus enroll in it during their second year in college, some are freshmen To take this course as early as possible appears the best advice. There seems no valid reason why it could not be taken in high school, since it is certainly no more difficult than a course in trigonometry and, judging from the reaction of many students, may even be simpler. An early course in statistics enables the student to use his knowledge in the many courses in

which some form of statistics is presently which involve almost all other obvious benefits gained from an early exposure, a course in statistics is a convenient and effective means of reviewing high school algebra before taking other mathematics courses that demand a thorough knowledge of algebra. Because of this, we have included exercises which involve almost all topics ordinarily encountered in two years of high school algebra—students seem to find solution of practical, interesting problems a much less painful way of reviewing their algebra than traditional methods.

Chapters 1 through 13 constitute the basic core of information required in all fields where statistics is used. Chapters 14 and 15 are designed primarily for students in business administration and economics. Consequently it is recommended that instructors in general one-semester statistics courses, particularly those oriented towards the biological sciences, omit Chapters 14 and 15 and consider two possible alternatives: one, complete coverage of Chapters 1 through 13; the other, coverage of Chapters 1 through 10, 12, 13, 16, and 17. Instructors in one-semester courses oriented towards the social sciences—in particular, economics—may find it preferable to omit Chapters 16 and 17, but to include Chapters 11, 14, and 15. If the complete coverage of all chapters is desired, the text would be suitable for a one-semester course meeting four hours a week or a two-quarter course of ten weeks each meeting three hours a week.

With an increasing number of universities on the quarter system, more and more institutions now desire to cover the basic material of probability and statistics in a one-quarter course meeting four hours a week for ten weeks. Experience has shown that, for such a course, Chapters 1 through 13 of this text can be covered, with the following exclusions: Section 5.4; Section 7.2; Chapter 11; Sections 12.6, 12.7, 12.8, 12.9, 12,10, and 12.11; and Section 13.5. For such a course, it is recommended that Chapter 11 be given as a reading assignment after Chapter 10 has been covered.

The present edition differs from the Fifth Edition not only in a large number of minor improvements but also in four major changes.

1. Since the numbers used in the examples and exercises of this text usually represent actual data, wherever possible they have been brought up to date with latest available information.

2. Exercises have been added in a few chapters, and review exercises covering the material of Chapters 3 to 13 appear in a new Appendix consisting of a set of exercises, all of which are different from those in Chapters 3 to 13 and are designed to give the reader experience in deciding which statistical technique learned in Chapters 3 to 13 is the appropriate one to use in each of these exercises.

3. Small additions have been made in a few chapters to include material which, experience has indicated, is frequently needed or useful. In particular, discussion of the binomial distribution (Chapter 6) has been expanded by treating this distribution as a special case of a probability distribution of a discrete variable and similarly the normal distribution (Chapter 7) is treated as a special case of a distribution of a continuous variable. Also, in the chapter on regression and correlation (Chapter 12), we have added an elementary proof, which does not require calculus, for obtaining the line of regression, and we have added a section on exponential and power curves. In the chapter on chi-square distribution (Chapter 13), a discussion of the proper way of drawing conclusions from a contingency table has been added and, in the chapter on analysis of variance (Chapter 17), material has been included to clarify the conditions under which Duncan's new multiple range test should be applied.

4. A few changes have been made which delete, add, or alter terminology in order to conform to current practices and usage in statistics.

We are grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., and to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint, in abridged form, Tables III, IV, and VII from their book *Statistical Tables for Biological, Agricultural, and Medical Research* (6th edition, 1974).

We gratefully acknowledge the many valuable suggestions and comments made by Professors Hubert A. Arnold, George A. Baker, and Curtis M. Fulton in the course of their use of an early draft of this book in their classes. We are indebted to Professor Jerry Foytik for his advice and helpful suggestions concerning the original draft and, in particular, for his comments on the material from economics covered in Chapter 15. We express special appreciation to Professor Gordon V. Shute of the City College of Chicago for his careful reading of the Fifth Edition and his many valuable suggestions. We are also grateful to Professors Dorothy L. Bernstein of Goucher College and Corwin L. Atwood of the University of California, Davis for their advice on certain parts of this new edition.

Henry L. Alder
Edward B. Roessler

August 1976

Contents

1
Introduction

Statistics is the science dealing with the *collection*, *organization*, *analysis*, and *interpretation* of numerical data.

"Collection of data" is the process of obtaining measurements or counts. Valid conclusions can result only from properly collected or from representative data. Although this is a very important part of statistical procedure, in the interest of conciseness we shall not discuss it but shall consider the treatment of data already available.

"Organization of data" is the task of presenting the collected measurements or counts in a form suitable for deriving logical conclusions. Representative methods of organizing and presenting data by means of tables and graphs are discussed in Chapter 2.

"Analysis of data" is the process of extracting from the given measurements or counts relevant information, from which a summarized and comprehensible numerical description can be formulated. The most important measures used for this purpose—the mean, the median, the range, the standard deviation, and others—are discussed in Chapter 4.

"Interpretation of data" is the task of drawing conclusions from the

analysis of the data and usually involves the formulation of predictions concerning a large collection of objects from information available for a small collection of similar objects. The interpretation of data forms the main portion of the text.

Statistics, then, is a science that deals with problems capable of being answered to some degree by numerical information, that, is information obtained by counting or measuring. It matters little whether we are making insect counts for a biological study or surveying the number of workers or work-hours in an industrial plant. The duties of the statistician are first to select the kind of information needed, then to direct the proper and efficient collection and processing of that information, and, finally, to interpret the results. In interpreting results, especially where they are based on incomplete data, the statistician must apply principles and techniques that yield valid findings. He is often expected to make wise decisions in the face of uncertainty.

The word statistics has two greatly differing meanings. When used as indicated in the preceding paragraph, it is a scientific procedure used in the study and evaluation of numerical data. When used as the plural of statistic, it is synonymous with the term "numerical data." Thus if we say there are statistics in the *World Almanac* or in the *Statistical Abstract of the United States*, we mean there are numerical data in them. This is the older and more common meaning of the word. Originally, statistics were gathered for the purpose of providing governmental heads with data for managing the affairs of state. Such information expressed in numbers dates back to Aristotle and his treatises on the "matters of state." In fact, there is evidence that the words "statistics" and "state" are derived from the same root. From earliest times most civilized countries have compiled large scale "statistics" in order to ascertain, for military and fiscal reasons, the manpower and material strength of the nation. We read in the Bible of such censuses, and compilations for purposes of taxation were common practices in all parts of the Roman Empire.

The study of probability began during the Italian Renaissance when gamblers seeking to develop systems of winning at dice consulted such scholars as Girolamo Cardano (1501–1576) and the famous mathematician-astronomer Galileo Galilei (1564–1642); Galileo wrote a short essay in which he set forth the fundamental laws of probability that form the basis for the whole science of statistics.

In the sixteenth and seventeenth centuries, games of chance were especially popular with people of wealth and as more complicated games were introduced and larger sums of money were involved, the need for a rational method of calculating the chances in various games became increasingly important. A French intellectual who was also a passionate gambler, Chevalier de Méré,

consulted the famous mathematician and philosopher Blaise Pascal (1623–1662), whose interest led Pascal to correspond with some of his mathematical friends, especially Pierre de Fermat (1601–1665). That correspondence forms the origin of modern probability theory and combinatorial analysis.

Other well-known mathematicians active in the study of the laws of chance were Gottfried Wilhelm Leibniz (1646–1716) and Jakob Bernoulli (1654–1705), who was the first of nine mathematicians in the famous Bernoulli family. All won measures of distinction, and Jakob, his brother Johann Bernoulli (1667–1748), and his nephews Nikolaus Bernoulli (1687–1759) and Daniel Bernoulli (1700–1782) achieved worldwide renown. The first extensive treatise on the theory of probability as a whole was written by Jakob Bernoulli, who expounded the principle of the Law of Large Numbers. Nikolaus Bernoulli applied the concept of probability to problems in law. Daniel Bernoulli applied the calculus of probability to epidemiology and the study of insurance.

During the same period, important advances were made in the collection of demographic data and the development of the body of knowledge now known as statistics. In England, John Graunt (1620–1674) made a semi-mathematical study of vital statistics and the statistics of insurance and economics. His work was extended by Sir William Petty (1623–1687), who studied the vital statistics of the population of the city of London, the first work of this kind ever undertaken, and by Edmund Halley (1656–1742), who developed mortality tables and is credited with originating the science of life statistics.

Abraham De Moivre (1667–1754) enunciated procedures for the probabilities of compound events, derived the theory of permutations and combinations from the principles of probability, and founded the science of life contingencies. In 1733 he discovered the equation of the normal curve upon which much of the theory of inductive statistics is based. The same bell-shaped curve is often referred to as the "Laplacian curve," the "Gaussian curve," or the "Gauss-Laplace curve," in honor of Marquis de Laplace (1749–1827) and Karl Friedrich Gauss (1777–1855), who independently rediscovered the equation. Gauss derived it from a study of errors in repeated measurements of the same quantity. He also originated the method of least squares and developed the theory of observational error. Laplace made great contributions to the application of statistics to astronomy and with Adrien Marie Legendre (1752–1833) introduced the use of partial differential equations into the study of probability. In 1815 the term "probable error" appeared for the first time in the writings of Friedrich Wilhelm Bessel (1784–1846), who also developed the theory of instrumental errors.

Other contributors to the theory were James Stirling (1692–1770), who de-

veloped an approximation to $n!$; Marquis de Condorcet (1743–1794), who applied probability and statistics to social problems; Thomas Bayes (1702–1761), who first used probability inductively; Leonhard Euler (1707–1783), the originator of the use of the Greek letter sigma as a symbol to denote summation; and Thomas Simpson (1710–1761), who introduced the principle of continuity into the theory of mathematical probability. In his study of probability Jean Le Rond d' Alembert (1717–1783) used meteorological data; Joseph Louis Lagrange (1736–1813) applied the differential calculus; and Pierre Rémond de Montmort (1678–1719) introduced the calculus of finite differences. The Comte de Buffon (1707–1788) anticipated some aspects of modern genetics and the calculus of probabilities, and Siméon Denis Poisson (1781–1840) developed the distribution that bears his name.

Between 1835 and 1870 the Belgian scientist Lambert A. J. Quetelet (1796–1874) made great contributions to the development and use of probability and statistics. He showed that biological and anthropological measurements closely follow the normal curve. Quetelet applied statistical methods not only in biology but also in education and sociology. He displayed a tremendous breadth of interest and is credited with being the first to recognize the constancy of large numbers and one of the first to demonstrate that statistical techniques developed in one area of research are applicable in most other areas.

In Germany, Georg Friedrich Knapp (1842–1926), following up Quetelet's principles, investigated extensively the statistics of mortality, and Wilhelm Lexis (1837–1914) developed a procedure that today is called one-way analysis of variance.

During the last quarter of the nineteenth century, Sir Francis Galton (1822–1911), the founder of the School of Eugenics in England, displayed unbounded enthusiasm as he verified the principle of systematic variation in every biological variable for which he was able to accumulate adequate data. Revelation of the principle of orderliness in biological variation formed the beginning of a new era in biological research. Galton and his great successor Karl Pearson (1857–1936), using problems from genetics, developed the ideas of regression and correlation. Later, Pearson and Charles Edward Spearman (1863–1945) extended this theory and applied it to studies in the social sciences. Pearson also studied extensively the effects of errors of sampling, developed the chi-square test and introduced the terms "mean deviation" and "standard deviation" into the literature.

Early in the present century, William Sealy Gosset (1876–1937), a statistician for Guinness, an Irish brewery, writing under the pseudonym of "Student," published many papers on the interpretation of data obtained

by sampling. He was the first to recognize the importance of developing methods of extracting reliable information from small samples. His methods were later popularized in England by Sir Ronald A. Fisher (1890–1962) and his colleagues, who made many contributions to science, especially to population genetics, and greatly extended the theory of experimentation, increasing the interest in statistical methods as well as their use in all areas of scientific investigation. It was Fisher who introduced the now widely-used term "null hypothesis" and developed statistical techniques for the analysis of variance.

In the twentieth century, many noteworthy statisticians, too numerous to mention, have been active in developing new theories and applications. The availability of electronic computers has greatly helped in these developments. Today the research worker considers statistics one of his most useful tools.

Everyday life is influenced more and more by decisions based on quantitative information. The scientific sequence of hypothesis, experiment, and test of hypothesis is now a familiar approach to problems in every area of activity. Today, modern statistical methods, founded on probability theory, are proving indispensable as aids in the physical and biological sciences, in economics and sociology, in psychology and education, in medicine and agriculture, and in government and industry. The astronomer predicts future positions of heavenly bodies on the basis of statistical methods; conformity to genetic segregation is ascertained statistically; life insurance premiums and annuity payments are determined from mortality tables based on statistical records; power companies cannot supply electricity efficiently without statistical data of load requirements; research workers determine significance in agricultural field trials from statistical considerations; engineers find sampling theory invaluable in controlling quality of manufactured products; and business executives and governmental analysts use statistical procedures in decision-making. Although these are widely differing fields of application, most of the statistical methods employed are the same. One aspect of statistical analysis may be stressed more in one field of application than in another, but, in general, the same statistical procedures are used in all fields.

In approaching the study of statistics, a word of caution is in order. It is important to realize that no statistical procedure can, in itself, insure against mistakes, inaccuracies, faulty reasoning, or incorrect conclusions. The original data must be accurate; the methods must be properly applied; and the results must be interpreted by one who understands not only the methods themselves but also the field to which they are applied. The statistical methods discussed in this book are to be considered as tools that, in proper hands and applied to the situation for which they are designed, can produce useful results, but that, by themselves, have no power to work miracles.

2

Organization of Data

2.1 INTRODUCTION

Frequently, the collection of information leads to large masses of data that, if they are to be understood or to be presented effectively, must be in some manner summarized. Clear and forceful presentation is an important aid to the understanding and correct interpretation of such data. Two methods of presenting quantitative data are in common use. One method involves a summarized presentation of the numbers themselves, usually in tabular form; the other consists in presenting the quantitative data in pictorial form—graphs, diagrams, or other similar representations. Representation of a mass of data by either of these methods is the part of statistical analysis that should lead to a better over-all comprehension of the data.

2.2 TABULAR AND GRAPHICAL METHODS OF PRESENTING DATA

In nearly all scientific and business publications, in government reports, in magazines and in newspapers, data of all sorts are presented by means of tables, diagrams, or pictures.