Elementary Statistics

In a World of Applications

Second Edition

Ramakant Khazanie

ELEMENTARY STATISTICS

In a World of Applications

SECOND EDITION

RAMAKANT KHAZANIE
Humboldt State University

SCOTT, FORESMAN AND COMPANY
Glenview, Illinois London, England

To the memory of Dada, Aai, and my brother, Suresh

AVAILABLE SUPPLEMENTS

Study Guide: includes chapter summaries and listing of key terms, additional examples and exercises with corresponding solutions, a selection of answers or solutions to chapter tests in the book, and assorted MINITAB applications.

Statistical Toolkit (by Tony Patricelli of Northeastern Illinois University): two-sided 48K Apple II or compatible diskette (color monochrome option) which provides both computational experience and graphical displays for key introductory concepts.

Library of Congress Cataloging-in-Publication Data

Khazanie, Ramakant.
Elementary statistics in a world of applications.

1. Statistics. I. Title. II. Title: Elementary statistics.
QA276.12.K5 1986 519.5 85-27654
ISBN 0-673-16618-X

456-RRC-91 90 898887

PREFACE

This book is a substantial revision of the first edition. Considerable portions of the text have been entirely rewritten with greater attention to clarity of exposition. Some new material has been added, such as stem-and-leaf diagrams, Chebyshev's Theorem, significance probability, paired t-tests, and prediction intervals. Certain topics have been reorganized to streamline the material and make it more teachable. A large number of drill exercises have been added in appropriate places to prepare the student to tackle word problems. Also, there are many new applied exercises, and the chapter tests have been expanded. An additional feature of this edition is the use of color to highlight important statements and results.

More and more academic disciplines are requiring a course in introductory statistics. Statistical methodology has become an important component of scientific reasoning and the areas of its application are numerous. Statistical techniques are now employed in many fields such as engineering, education, agriculture, business, biology, medicine, fisheries, natural resources, geology, communications, psychology, and ecology. For this reason, the treatment in this book is not slanted toward any particular discipline but touches on diverse areas of interest in all walks of life. The student's acquaintance with the fundamentals of statistics developed here should provide a basis for studying specialized methods within his or her field.

In preparing the second edition, the basic format of the book is left untouched, and the order of coverage is much the same as in the earlier edition. The following changes have shaped this revised edition:

Chapter 1 is substantially reorganized, having fewer sections now. Also, stem-and-leaf diagrams have been added in the graphing section.

Chapter 2 on measures of central tendency and dispersion is a result of considerable rearrangement of topics in Chapters 2 and 3 of the first edition. In particular, mean and variance for grouped data are now discussed in a separate section after having been introduced for ungrouped data in the earlier sections. Also, the definition of sample variance is now in conformity with the more traditional usage where the divisor $n - 1$ is used instead of n. (In the first edition, understandably, the divisor n was an irritant for some users.) Finally, a small discussion on Chebyshev's Theorem is added.

Section 3–2 has been almost entirely rewritten in order to give the student a better feeling for the concepts of sample space and events. Also, more illustrative examples with detailed explanations have been added in the section on counting techniques. In the final analysis, it is my considered opinion that counting techniques *per se* are extraneous to understanding the basic concepts of probability. An instructor who wishes to stay clear of Section 3-4 can do so without impairing a systematic development of later topics. The only place where these techniques will come up in the text is in the treatment of the binomial distribution and its approximation to the normal distribution in Chapter 6. The parts that might be skipped without losing continuity are clearly indicated at the appropriate places.

Chapter 7 on estimation (single population) and Chapter 9 on testing of hypotheses have undergone major changes. One change is that, whereas in the first edition the t distribution, the chi-square distribution, and the F distribution were discussed in one section of the chapter on sampling distributions, in the current edition they are discussed as and when they are needed. Also added, under testing of hypotheses, are significance probability (P-value) and the paired t-test.

Substantial changes have been effected in Section 11–2 on linear regression and linear correlation. The prediction interval for y is an addition to this section.

Chapter 12 on analysis of variance and Chapter 13 on nonparametric tests have also been revised somewhat, though not in a major way.

This textbook covers topics in elementary statistics that, in recent years, have been offered at the freshman-sophomore level in various colleges and universities. The treatment presented here employs mathematics at the most elementary level, with the express purpose of stimulating interest among students who have no special interest in or taste for mathematical formalism. Instead of addressing such formalism, constant appeal is made to the reader's intuition. A background in high school algebra would suffice, although a course in finite mathematics could help the student attain a certain degree of maturity and self-confidence.

In developing the ideas, the overriding concern has been to present the material in as clear a way as possible so that any student who wishes to read it on his or her own can do so with a minimum of difficulty. Each new concept is introduced through an example often drawn from a real-life situation to which the student can easily relate. This introduction is followed by numerous detailed examples. Many of these examples are based on hypothetical data. The book is replete with exercises at the end of each section giving students ample opportunities to test their skills and to fortify their understanding of the subject matter. These exercises serve an additional purpose—to show how pervasively statistical applications are used in everyday life. As a final measure of the competency of the student, a test is provided at the end of each chapter. To derive maximum benefit, the chapter test should be treated as such—a test—and should be attempted only after acquiring a certain degree of proficiency in the topics covered in the chapter.

The contents of this book could be divided into three parts. The first part, consisting of Chapters 1–2, gives an account of descriptive statistics and is intended to familiarize the student with field data summary. The second part, Chapters 3–6, involves elementary probability. In this part the goal is to build the tools essential to understanding the concepts of statistical inference. In the third part, which deals with the inferential aspect of statistics, Chapters 7–9 should be considered by far the most important. Of the remaining chapters, I feel the order of importance is as follows: Chapters 11, 10, 13, 12. However, this will vary from instructor to instructor.

The topics treated in the book should prove adequate for an introductory one-semester course meeting three hours a week or a one-quarter course meeting four hours a week. In my teaching of a one-semester course meeting three hours a week or during a one-quarter course meeting four hours a week, I have invariably covered the first nine chapters. There is enough optional material from which to fashion a two-quarter course meeting three hours a week. The chapters on linear regression and linear correlation, goodness of fit, analysis of variance, and nonparametric methods are included to enable flexibility in structuring such an offering.

In a first course such as this, one can hardly be expected to attain mastery of the vast subject of statistics. This book sets for itself the very modest goal of introducing the student to some aspects of statistical methodology. It is hoped that a student who has acquired a sound understanding of the topics developed here will be in a position to realize the importance of statistical reasoning in work and in life, to discern the basic statistical assumptions underlying a given situation, and to pick an appropriate test if the problem falls within the framework of his or her introduction to the subject. But more than anything, it is hoped that students will be able to interpret the results of a statistical investigation and decipher the barrage of statistical information aimed at them constantly.

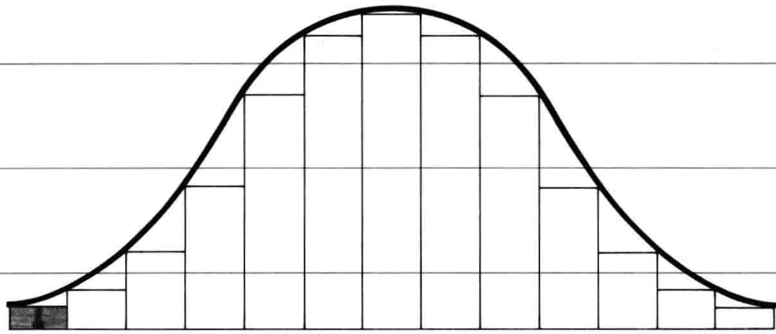
In the course of developing this edition, I have availed myself of the many valuable suggestions made by my colleagues and users of the first edition. I have also benefited from the many helpful comments of the reviewers. I express my thanks and appreciation to the following individuals who contributed in many ways to the first edition and this one: Jerald Ball, Chabot College; Leonard Deaton, California State University, Los Angeles; Robert Goosey, Eastern Michigan University; John Skillings, Miami University; Frank C. Hammons, Sinclair Community College; Wilbur Waggoner, Central Michigan University; Shu-ping Hodgson, Central Michigan University; Ian W. McKeague, Florida State University; Jan F. Bjørnstad, University of Texas at Austin; Donald L. Meyer, University of Pittsburgh; Robert J. Lacher, South Dakota State University; Arthur Dull, Diablo Valley College; Donald Fridshal, California State University, Chico; August Zarcone, College of DuPage; Andre Lehre and Victor Tang, Humboldt State University. I extend my special appreciation to my colleague, Professor Charles Biles, who took a keen interest in the project, read the entire manuscript, and made many helpful suggestions. My thanks are also due to Professor Virgil Anderson of Purdue University who inculcated in me a taste and liking for statistical applications. I also express my appreciation to Kathleen McCutcheon, who typed parts of the manuscript for this edition. Finally, I thank my family for enduring patiently during the course of this project.

I am grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S. and to Longman Group, Ltd., London, for permission to reprint Table III from their book, Statistical Tables for Biological, Agricultural and Medical Research (6th edition, 1974).

Ramakant Khazanie

CONTENTS

1



ORGANIZATION OF DATA

1-1 Population, Sample, and Variables

1-2 Frequency Distributions

1-3 Graphs and Charts

INTRODUCTION

What is statistics? Allusions to statistics are abundant and a penchant to quote statistics is almost universal. Strangely enough, however, skepticism about statistics is widespread. Much of this reaction stems from the fact that misleading conclusions are often drawn either from poor planning of an experiment or because some pertinent information is withheld. In this context the following observation is appropriate:

> The secret language of statistics, so appealing in a fact-minded culture, is employed to sensationalize, inflate, confuse and oversimplify. Statistical methods and statistical terms are necessary in reporting the mass data of social and economic trends, business conditions, "opinion" polls, and census. But without writers who use the words with honesty and understanding and readers who know what they mean, the result can only be semantic nonsense.*

Statistics is a science that deals with the methods of collecting, organizing, and summarizing data in such a way that valid conclusions can be drawn from them. Statistical investigations and analyses of data fall into two broad categories—*descriptive statistics* and *inductive statistics.*

Descriptive statistics deals with processing data without attempting to draw any inferences from it. It refers to the presentation of data in the form of tables as in Table 1-1 and graphs as in Figure 1-1 and to the description of some of its features, such as averages, which we will consider in Chapter 2. To most people it is this notion that is conveyed by the word *statistics:* a compilation of a huge mass of data and charts dealing with taxes, accidents, epidemics, farm outputs, incomes, university enrollments, brands of televisions in use, sports statistics, and so on.

TABLE 1-1

Statistical information presented as a table
Newspaper Advertising—Expenditures for 64 Cities: (in millions of dollars)

Type of Advertising	*1957*	*1970*	*1974*	*1975*	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*
Total	1,611	3,120	3,845	4,117	5,352	5,696	6,666	7,641	8,186	9,575
Automotive	63	93	109	93	127	145	151	196	182	226
Classified	313	724	967	982	1,342	1,522	1,892	2,179	2,196	2,515
Financial	42	117	135	131	148	147	203	244	297	387
General	326	426	514	547	731	752	827	982	1,122	1,380
Retail	866	1,759	2,120	2,364	3,005	3,129	3,593	4,040	4,389	5,068

Source: Compiled by Media Records, Inc. Current data in U.S. Bureau of Economic Analysis, *Survey of Current Business*, monthly.

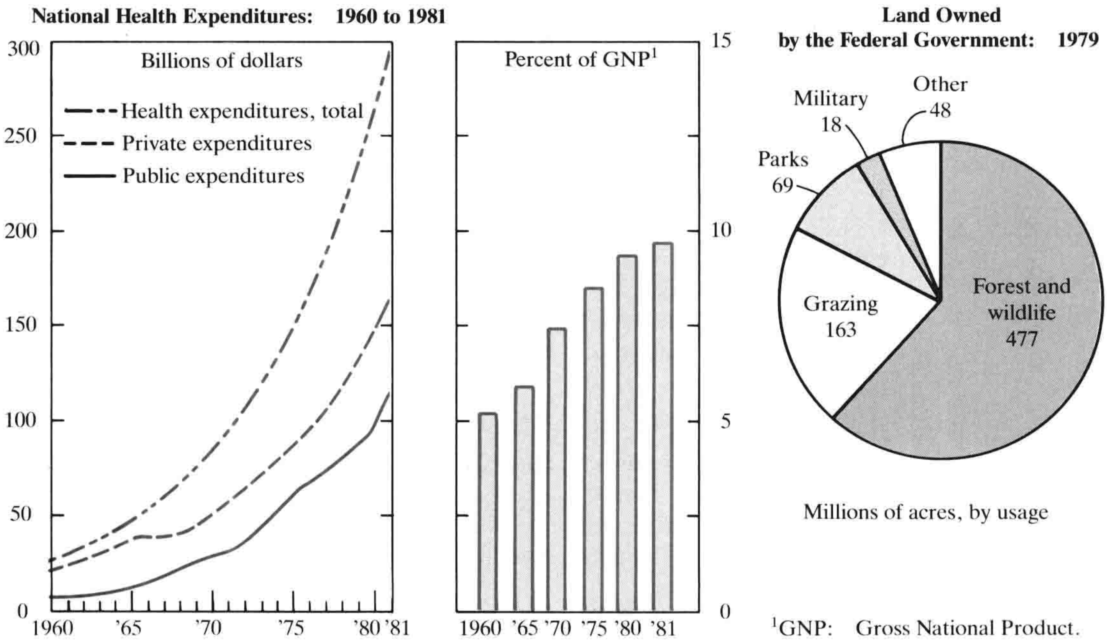
*From Darrell Huff, *How to Lie With Statistics* (New York: W. W. Norton & Co.), p. 8.

As distinguished from descriptive statistics we have **inductive statistics,** also called **inferential statistics,** which is a scientific discipline concerned with developing and using mathematical tools to make forecasts and inferences. In this role of devising procedures for carrying out analyses of data, statistics caters to the needs of research scientists in such diverse disciplines as industrial engineering, biological sciences, social sciences, marketing research, and economics, to mention just a few. Basic to the development and understanding of inductive statistics are the concepts of probability theory.

In this and the next chapter we shall concentrate specifically on some aspects of descriptive statistics. In subsequent chapters we shall be concerned mainly with inductive statistics, developing concepts in Chapters 3 through 6 and then, in the remainder of the book, applying these to decision-making procedures.

FIGURE 1-1

Statistical information presented through charts.



Source: Charts prepared by U.S. Bureau of Census.

1-1 POPULATION, SAMPLE, AND VARIABLES

One of the goals of a statistical investigation is to explore the characteristics of a large group of items on the basis of a few. Sometimes it is physically, economically, or for some other reason almost impossible to examine each item in a group under study. In such a situation the only recourse is to examine a subcollection of items from this group.

For example, if in a primary election we are interested in finding the disposition of the voters toward a certain candidate, it would not be feasible in terms of time and money to interview every potential voter. Instead we might interview only a few of the voters, the selection being carried out in such a way as to provide a good cross section of the voters. Since we are interviewing only a few and drawing conclusions about the entire set of voters, our inferences are subject to an element of uncertainty and so the statements should be couched in the language of probability.

All the conceivable members of a group under study constitute a **population,** or a **universe.** The words *population* and *universe* do not carry the usual dictionary meanings but refer simply to the totality of observations relevant to a given discussion.

What group of items make up a population will depend upon what the researcher is interested in investigating. For example, suppose a retailer is interested in knowing whether the bolts in a large shipment are within specifications. In this case the shipment of bolts constitutes the population relevant to the retailer's investigation.

EXAMPLE 1 Suppose an ornithologist is interested in investigating migration patterns of birds in the Northern Hemisphere. Then all the birds in the Northern Hemisphere will represent the population of interest to him. His choice of the universe restricts him, for it does not include birds that are native to Australia and do not migrate to the Northern Hemisphere. ■

EXAMPLE 2 Every ten years the U.S. Bureau of Census conducts a census of the entire population of the United States accounting for every person regarding sex, age, and other characteristics. The last such census was carried out in 1980. Only the federal government with its vast resources can undertake a task of this magnitude. In this case the entire population of the United States is the population in the statistical sense. ■

A population can be finite or infinite. For a *finite population,* if we start counting the members, the counting process will ultimately come to an end. This is not the case for an *infinite population.* All the examples given above are examples of finite populations. The following are examples of infinite populations: the possible weights of people; the possible temperature readings on a day (before the day starts one can conceive of any possible reading over a range); the possible heights that a balloon can attain.

A **sample** is a subcollection of items drawn from the population under study.

EXAMPLE 3 A fisheries researcher is interested in the behavior pattern of Dungeness crab *(Cancer Magister Dana)* along the Pacific Coast of North America. It would be inconceivable and impossible to investigate every crab individually. The only way to make any kind of educated guess about their behavior would be by examining a small subcollection, that is, a sample.

EXAMPLE 4 Suppose a machine has produced 10,000 electric bulbs and we are interested in getting some idea about how long the bulbs will last. It would not be practical to test all the bulbs because the bulbs that are tested for their life will never reach the market. So we might pick 50 of these bulbs to test. Our interest is in learning about the 10,000 bulbs and we study 50. The 10,000 bulbs constitute the population, and the 50 bulbs the sample.

Any quantitative measure that describes a characteristic of a population is called a **parameter.** Parameters are the constants that are peculiar to a given population. A quantitative measure that describes a characteristic of a sample is called a **statistic.**

For example, the proportion of Democrats among all the voters in the United States is a parameter. If we interview 1000 people among all the voters in the United States, then the proportion of Democrats in this sample is a statistic.

As another example, suppose we are interested in the heights of the people in the United States at a given time. Then the tallest person at that instant in the United States is a parameter. The height of the tallest person in a sample of 500 people, for example, is a statistic.

Usually access to the entire population is limited by such factors as the time and costs involved, the physical size of the population, and, quite often, the very nature of the investigation, as in Example 4. An understanding of the population is acquired via a sample. When we draw a sample generally we are interested not in studying the sample per se but in extrapolating the nature of the population from which the sample is drawn. On the basis of the findings from the sample, if it is representative of the population, we acquire a better understanding of the population.

Professional pollsters such as Gallup, Lou Harris, and Roper, various advertising agencies interested in consumer preferences, and research scientists invariably use samples to make projections regarding population characteristics, that is, parameters based on their findings from the sample. Thus the aspect of statistics they are dealing with is inferential.

Statistical data or information that we gather is obtained by interviewing people, by inspecting items, or by many other different ways. The characteristic that is being studied is called a *variable*. Heights of people, grades on a high school test, the time it takes for the bus to arrive at the bus stop, and hair color are examples of variables. There are two kinds of variables—*qualitative* and *quantitative*.

A **qualitative variable** can be identified simply by noting its presence. The color of an object is an example of a qualitative variable. In the same way, the outcome on tossing a coin is an example of a qualitative variable. The outcome is either heads or tails and has no numerical value. Although a qualitative variable has no numerical value, it is possible to assign numerical values to a qualitative variable by giving values to each quality. For example, we could assign the value of 1 to heads and the value of 0 to tails and in this way quantify the qualitative information.

A **quantitative variable** consists of numerical values. For example, the height of an individual when expressed in feet or inches is a quantitative variable. Diameters of bolts (in inches) produced by a machine, waiting time at a bus stop (in minutes), the price of a stock (in dollars), the annual income of a family (in dollars), the number of bacteria in a culture, the volume of sales in a day (in dollars) are other examples of quantitative variables.

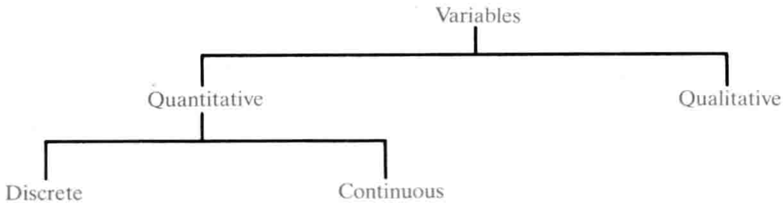
A familiar example of a variable which is given qualitatively as well as quantitatively is a score on an examination. If the score is given numerically as 80 points, 65 points, and so on, then the variable is quantitative; but if the score is given in letter grades A, B, C, D, F, then the same variable becomes qualitative.

As another example, the height of an individual measured in feet is a quantitative variable; the same variable recorded as tall, medium, or short is a qualitative variable.

We can classify quantitative variables further as continuous or discrete. If the variable can assume any numerical value over an interval or intervals, it is a **continuous variable.** Weight, height, and time are examples of continuous variables. The observations can be measured to any degree of accuracy on a numerical scale. For example, if we choose, we can report the time of travel as 2 hours, 2.5 hours, or 2.5378 hours, although, as a rule, we are not interested in such precision. Loosely speaking, if a variable is continuous, then there are no ''breaks'' in the possible values the variable may assume.

In contrast with a continuous variable, a **discrete variable** is one whose possible values consist of breaks between successive values. For our purpose this definition will suffice. The number of Democrats is an example of a discrete variable, since the only values that the variable can take are integral values such as 0, 1, 2, and so forth. For instance, we would not report that there were 2.5 Democrats in a group. Other examples of discrete variables are the number of bacteria in a culture, the number of defective bulbs, the number of telephone calls received during a one-hour period, and the number of unfilled orders at the end of a day.

A classification of variables can be displayed schematically as follows:



EXAMPLE 5 The following is an excerpt from an article that appeared in Time magazine in the Economy and Business section (Dec. 12, 1983).

> The "little guy" is back in the stock market in a big way and, as it turns out, is not so likely to be a guy. Women have rushed to buy stocks during the past two years, more than men. They constitute 57% of all new shareholders.

The information in the statement "Women have rushed to buy stocks during the past two years, more than men" is qualitative in nature where the variable of interest is the sex of the shareholder—male or female.

Presumably the study is conducted to describe the new shareholders during the period mentioned. Naturally, then, "all new shareholders during the two-year period" would constitute the population.

The statement that women "constitute 57% of all new shareholders" gives a quantitative measure of the percent of women shareholders in the population. Hence 57 percent is the value of the parameter.

EXAMPLE 6 The following table based on data from U.S. International Trade Commission, *Synthetic Organic Chemicals,* annual, is reproduced from the Statistical Abstract of the United States 1982–83.

There are several variables involved in describing the data in Table 1-2.

TABLE 1-2

Synthetic Organic Pesticides—Production and Sales

Item	*Unit*	*1960*	*1965*	*1970*	*1975*	*1976*	*1977*	*1978*	*1979*	*1980*
Production, total	Mil. lb	648	877	1,034	1,603	1,364	1,388	1,416	1,429	1,468
Herbicides	Mil. lb	102	263	404	788	656	674	664	657	806
Insecticides	Mil. lb	366	490	490	660	566	570	605	617	506
Fungicides	Mil. lb	179	124	140	155	142	143	147	155	156
Production, value[1]	Mil. dol	307	577	1,058	2,900	2,880	3,116	3,342	3,685	4,269
Sales, total	Mil. lb	570	764	881	1,317	1,193	1,263	1,300	1,369	1,406
Sales value[1]	Mil. dol	262	497	870	2,359	2,410	2,808	3,041	3,631	4,078

[1]Manufacturers unit value multipled by production.

Source: U.S. Dept. of Agriculture, Agriculture Stabilization and Conservation Service, *The Pesticide Review*, 1979. Based on data from U.S. International Trade Commission, *Synthetic Organic Chemicals*, annual