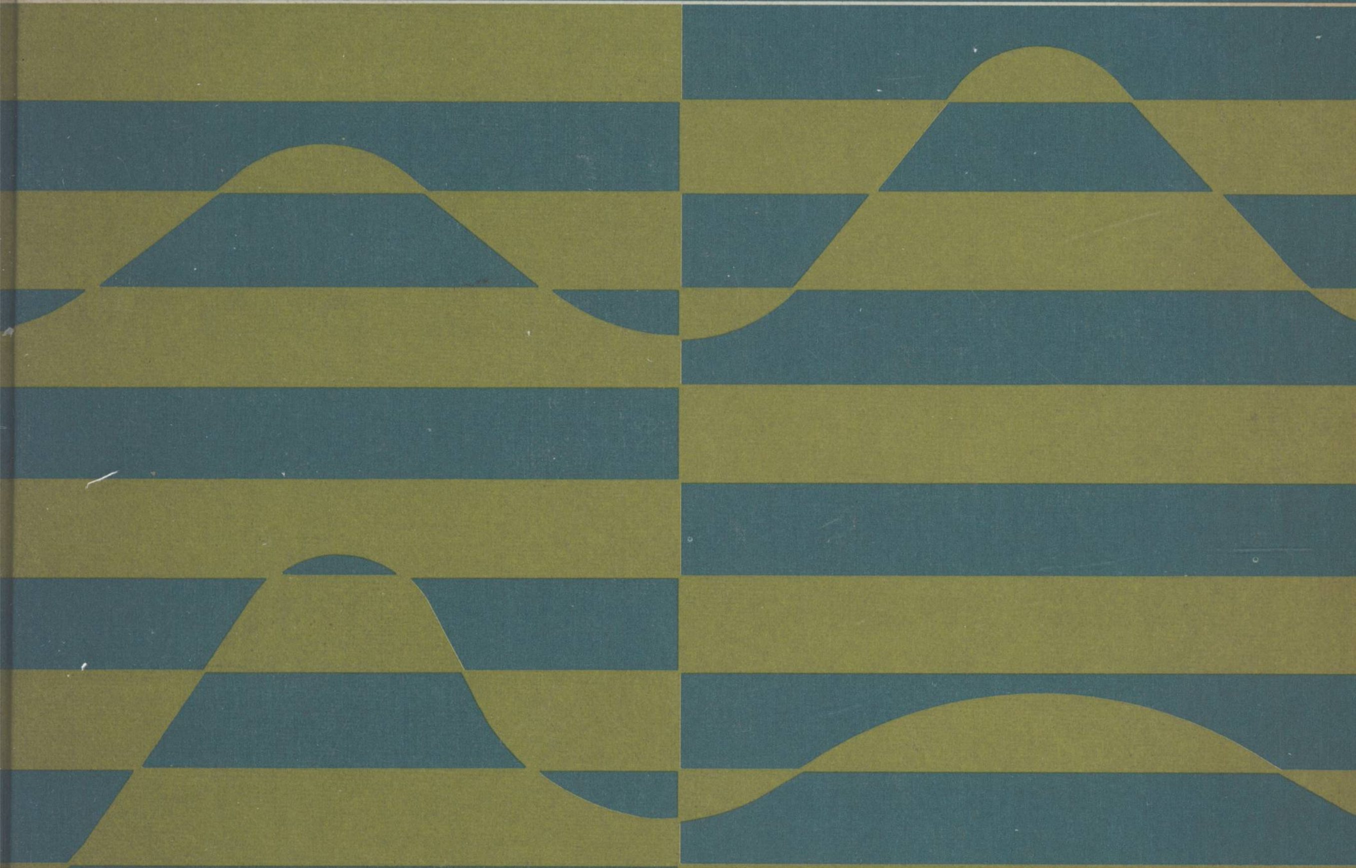
statistics: an intuitive approach



Lincoln L. Chao

STATISTICS:

AN INTUITIVE APPROACH

Lincoln L. Chao

California State University
Long Beach

SRA

SCIENCE RESEARCH ASSOCIATES, INC.
Chicago, Palo Alto, Toronto
Henley-on-Thames, Sydney, Paris

A Subsidiary of IBM

In Memory of
My Mother

Printed in the United States of America

Library of Congress Catalog Number: 73-89380

ISBN 0-574-19010-4

preface

At the time this book was being contemplated, the following announcement was made by a large state university to the statistics instructors of the colleges that provide most of its transfer students: their statistics courses would be accepted by the university as long as they put emphasis on the theory of probability and statistical inference rather than on traditional descriptive methods. This announcement illustrates two recent trends in the teaching of elementary statistics. One of these is a change in emphasis from descriptive statistics to probability theory and statistical inference. The other relates to the level at which beginning statistics courses are being taught. Elementary statistics, which has traditionally been an upper-division course for most students, is now being offered as a lower-division course in an increasing number of institutions.

This book was written with these two trends in mind. Its primary objective is to introduce the basic concepts of probability theory and the methods of statistical inference to students who have no mathematical background beyond one or two years of high school algebra.

The approach is essentially nonmathematical. The student is expected to know only how to apply statistical techniques in a decision-making process; the mathematical derivation of those techniques is not introduced. In most instances, the student is urged to follow the theoretical developments intuitively; in others, he is asked to take the derivations on faith. I hope that this approach will make statistics, which is often considered a great hurdle in a student's college career, an easier and more enjoyable course.

Each chapter starts with an introductory section explaining its scope, basic concepts, and objectives. Each theory or method is amply illustrated with examples taken from a variety of fields, including business, the social and natural sciences, and educational psychology. Similarly, the exercise problems provided after each major section of the text are interdisciplinary in nature. Short answers to most of the problems are listed at the end of the book. A solutions manual is provided for the instructor.

I wish to express my deep appreciation to my colleagues in quantitative methods at California State University, Long Beach, for their generous help and suggestions. I am especially grateful to Dr. Harry G. Romig, professor emeritus, who made many constructive criticisms, and to Dr. Perri Stinson for her invaluable comments. Heartfelt thanks are due also to Professor James Curl, of Modesto College, whose critical review and illuminating suggestions are in great part responsible for the present form of the text. I am also indebted to Professor Jerald T. Ball, of Chabot College; Kenneth Goldstein, of Miami-Dade College; Sidney Katoni, of New York City Community College; Barron G. Knechtel, of Orange Coast College; and Joseph Deken, of Stanford University, for their comments and assistance. Last but not least, I wish to express my gratitude to my wife and sons for their encouragement and sacrifices; without them this project could not have been undertaken. I wish to extend my special thanks to my son, Martin, for preparing appendix D, the table of square roots.

Lincoln L. Chao
Long Beach

contents

1 introduction

1-1 what is statistics?

Most words have multiple meanings; the word "statistics" is no exception. For example, "statistics" may be used to denote a collection of figures or numbers ("Here are the statistics for the first half." "Have you seen the latest statistics on unemployment?"), or the term may be used to designate a discipline, or branch of learning. These two meanings of statistics are actually closely related because statistics as numerical data is an essential part of statistics as a discipline.

The origin of statistics as a discipline can be traced to two areas of interest: *government records* and *games of chance*. When national states began to emerge at the end of the Middle Ages, it became necessary to collect information about the territories under their jurisdiction. This need for numerical information about citizens and resources led to the development of techniques for collecting and organizing numerical data. By the seventeenth century, surveys similar to our modern census were already in existence. At the same time, insurance companies were beginning to compile mortality tables for determining life insurance rates.

In the early stages of its development, statistics involved little more than the collection, classification, and summarization of numerical data. Even today, these activities are still an important part of statistics. (Think of all the tables, charts, and graphs you have studied.) Because the objective of this kind of data handling is to describe the important features and characteristics of collected information, we generally refer to it as *descriptive statistics*. The following is an example of descriptive statistics.

A school psychologist wants to know the average IQ of the five foreign students in his district. An intelligence test is administered to the students, and the IQ's are found to be 101, 103, 105, 107, and 109. The statistical method the psychologist uses to find the average IQ is the calculation of the arithmetic average. (The computation of the arithmetic average, simple as it is, is an important part of descriptive statistics.) The process is limited to the data collected in this particular case, and does not involve any inference or generalization about the IQ's of other foreign students. The method is descriptive in nature, because the average, which is 105, summarizes and describes the collected information.

The other source of statistics is found in gambling. In the seventeenth century, Europeans were intensely interested in games of chance, an interest that led to the development of the *theory of probability*. The study of probability, in turn, led to the development of a new field of statistics called *inferential statistics,* or *statistical inference*. Statistical inference is a technique by which conclusions or generalizations are drawn about the characteristics of a totality on the basis of partial or incomplete information. The following brief examples illustrate the meaning of statistical inference and its importance.

A drug manufacturer claims that a new cold vaccine developed by his company is 90 percent effective; that is, 90 out of every 100 persons who use the vaccine will survive the winter without catching cold. Let us assume that 30 persons have received the vaccine; of the 30, 25 survived the winter without catching a cold. Should we accept the manufacturer's claim?

Consider the problem of deciding whether to accept or reject a shipment of purchased goods in a manufacturing plant. A portion of the shipment is inspected to determine whether the shipment should be accepted. If 20 units are inspected, and two units are found to be defective, should the plant reject the shipment and return it to the supplier?

Let us assume that two different teaching methods are used to instruct two groups of students in a given subject. The two groups are believed to be comparable in ability. At the end of the instructional period, a standard test is administered to both groups. On the basis of the average test score for each group, can we evaluate the relative effectiveness of the two methods?

These are typical of the questions handled by inferential statistics. They involve a decision or choice between alternative courses of action in the light of empirical evidence. This is why statistics has been often referred to as a body of techniques dealing with decision making on the basis of observed data.

In summary, statistics as a discipline, or branch of learning, refers both to the treatment of numerical data and to the methods and theories that are used to handle numerical data for inferential purposes. More specifically,

> Modern statistics refers to a body of methods and principles that have been developed to handle the collection, description, summarization, and analysis of numerical data. Its primary objective is to assist the researcher in making decisions or generalizations about the nature and characteristics of all the potential observations under consideration of which the collected data form only a small part.

Thus, descriptive statistics and statistical inference are two areas that make up the discipline of modern statistics. Of the two, the latter has become increasingly more important.

1-1 exercises

1. Most newspapers contain a "vital statistics" section. What does the word "statistics" mean in this particular context?
2. In what context is the word "statistics" generally used on radio and television?
3. Why was statistics sometimes referred to as "political arithmetic" in the past?
4. In an economics examination, three freshmen received grades of 90, 85, and 80; and three sophomores received grades of 89, 86, and 92. From the following statements made on the basis of these figures, identify those that are derived from descriptive methods and those that are derived from statistical inference.
    a) The average grade of the three freshmen is 85, and the average grade of the three sophomores is 89.
    b) The three freshmen's grades fluctuate more than the three sophomores' grades.
    c) In economics, the average grade received by sophomores is generally higher than that received by freshmen.
    d) In the next economics test, freshmen will probably receive lower grades than sophomores.
    e) Freshmen's grades in economics usually fluctuate more than sophomores' grades.
5. Four brand A automobile tires and three brand B tires are tested to determine their service lives. The service lives for brand A are 29,000, 33,000, 37,000, and 41,000 miles; for brand B, they are 30,000, 32,000, and 34,000 miles. From the following statements made on the basis of these figures, identify those that are derived from descriptive methods and those that are derived from statistical inference.
    a) The average service life of brand A tires is 35,000 miles, whereas that of brand B tires is 32,000 miles.
    b) The service life of brand A tires varies more than that of brand B.
    c) If the price of brand A tires is the same as that of brand B tires, you should purchase brand A tires.
    d) The four tires of brand A vary more in service life than the three tires of brand B.
    e) The average service life of the four brand A tires is longer than that of the three brand B tires.
6. Use your own words and examples to illustrate the difference between descriptive statistics and statistical inference.

1-2 basic concepts: population and sample

population

The totality of "all potential observations" mentioned in the previous section is generally referred to as the "population." (Some statisticians refer to this concept as the "universe.") The term "population," like the word "statistics," has various meanings. In common usage, it refers to the number of people in a certain region or locality. (The population of China is 800 million; the population of the United States is 220 million; the population of New York City is more than 10 million.) In statistics,

> "Population" is defined as the totality of all potential measurements or observations under consideration in a given problem situation.

As shall be seen shortly, the term does not necessarily refer to the number of human beings in a given locality.

Each different problem situation involves a different population. If the problem is to ascertain a particular characteristic of all elementary school age children in California, then the measurements of that characteristic in that group of children constitute the population. If the purpose of investigation is to determine what proportion of all machine parts produced by a certain manufacturing process are defective, then the population consists of the measurements of the quality of all the machine parts produced by that process. If the problem is to determine the probability that heads will occur when a coin is tossed, the population consists of the outcomes of a presumably infinite number of tosses; similarly, in determining whether each of the six sides of a die has an equal chance to turn up when a die is rolled, the population consists of the outcomes of a presumably infinite number of rolls.

Populations are generally classified into two categories: finite and infinite. Finite means "capable of being reached or surpassed by counting"; infinite means "lacking limits or bounds" or "extending beyond measure or comprehension." A *finite population is one that includes a limited number of measurements or observations.* For instance, the total number of students presently attending colleges and universities in the United States, all calves born alive in the state of Wisconsin in 1973, and all the institutions of higher education in North America are finite populations. Some finite populations consist of only a few units, whereas others consist of millions. But as long as the total number of all potential measurements is capable of being reached, the population is considered finite.

A population is said to be infinite if it includes a large number of measurements or observations that cannot be reached by counting. For instance, the population of all live births of the human race is infinite because there is no limit to its number. Similarly, the population of digit numbers taken *with replacement* from an urn containing ten balls, each of which is marked with a different one of the ten digits, is infinite, because the number of balls that can be drawn is beyond limit or measure.

Similarly, the populations in the three examples mentioned earlier (the manufacturing process, the coin, and the die) are also infinite. They are infinite populations because, at least hypothetically, there is no limit to the number of observations each of them can include.

Population characteristics are generally called *parameters,* and population parameters are usually considered *true values.* For instance, the average IQ of all first graders in the country is a characteristic, and hence a parameter, of the population of the IQ's of first graders. It is the true average of the IQ's. Similarly, the proportion of all TV viewers who watch a particular program at a given time is a parameter of the population of the TV viewers; it is the *true proportion,* or *population proportion.* It is absolutely impossible to compute the true value of any parameter of an infinite population. In most cases, it is also impracticable to compute the true value of any parameter of a finite population. As a result, it is necessary to make inferences about population parameters from information contained in a small portion—or sample—of that population.

sample

A sample is a collection of measurements or observations taken from a given population. The number of observations in a sample is, of course, smaller than the number of possible observations in the population; otherwise the sample would be the population itself. Samples are taken because it is not economically feasible, though possible in some cases, to make all the potential observations in the population.

If, for example, we wanted to estimate the average annual expenditure of the American college student, we would probably draw a sample of, say, 1000 students, find out the annual expenditure for each of them, add the figures, and then find the average. We do this because we simply do not have the time and resources to contact all college students, even though it is possible to do so. On the basis of the sample average thus obtained, we make an inference about the average expenditure of all college students. Similarly, we may draw a sample of, say, 100 calves born in Wisconsin in a given year as a basis for estimating the average weight of all calves born in that state during the same year.

To estimate the value of a parameter of an infinite population, it is absolutely necessary to use sample information. Thus, to determine the proportion of defective parts produced by a production process, quality-control technicians examine a batch of parts to find out the number of defective parts contained in it. Such a batch, which constitutes a sample, is usually taken at regular periodical intervals. Tossing a coin 300 times to determine whether it is fair, and rolling a die 600 times to determine whether it is balanced, are other examples of sample taking from infinite populations.

A sample characteristic is generally called its *statistic*. For example, the average IQ of 1000 first graders selected at random from among all first graders in the country is a statistic; it is a characteristic of the sample of the IQ's of 1000 first graders. The proportion of a sample of 300 TV viewers who watch a particular program is also a statistic. Such a proportion is called a *sample proportion*. Statisticians use sample statistics to make inferences about population parameters.

1-2 exercises

1. Suppose that the freshman class of your college consists of 5000 students, all of whom have taken a standard aptitude test that was administered to all entering freshmen in the country. Explain the circumstances under which the 5000 scores received by these students can be considered (a) a sample and (b) a population.

2. During a certain week 2000 customers were served by a restaurant. Explain the circumstances under which these 2000 customers can be considered (a) a sample and (b) a population.

3. Suppose that a balanced coin is tossed 100 times, and 60 heads are obtained. Answer the following.
   a) What is the sample proportion?
   b) What is the population proportion?
   c) What is the size of the sample?
   d) What is the size of the population?

4. Suppose that 60 percent of all registered voters of a given country are members of a certain party and 40 percent are not. From a sample of 50 voters, it is found that 25 belong to the party. Answer the following.
   a) What is the sample proportion of voters who belong to the party?
   b) What is the population proportion of voters who belong to the party?
   c) What is the population? Is it finite or infinite?

5. A poll is conducted to determine voting preferences in a presidential election. Answer the following.
   a) What constitutes the sample?
   b) What constitutes the population?
   c) Is the population finite or infinite? Explain.

6. A survey is conducted to determine whether American housewives prefer one brand of detergent to another. Answer the following.
   a) What constitutes the sample?
   b) What constitutes the population?
   c) Is the population finite or infinite? Explain.
7. For each of the following statements, define the population that is being sampled, and tell whether it is finite or infinite.
   a) One thousand college students are interviewed about their opinions on the legalization of marijuana.
   b) Five hundred nonunion workers are interviewed to determine their attitudes toward unionization.
   c) We select a sample of screws produced by an automatic machine and test them to determine the proportion that are defective.
   d) We select a sample of hogs and obtain their birth weights.
8. Give an example to illustrate each of the following: (a) a finite population, and (b) an infinite population.

1-3 areas of application

Statistical methods and principles have found applications in many fields: business, the social sciences, engineering, and the natural and physical sciences.

The growing complexity of the economy has created a tremendous degree of uncertainty about the future operations of any business enterprise. As a result, more and more companies are using statistical analysis as a decision-making tool, especially in such areas as market research, forecasting, and analysis of economic trends.

Statistics also plays an important role in education and psychology. For instance, an educator may want to find out if, for a certain group of students, there is a correlation between achievement-test scores and grade-point averages. If there is a correlation between the two, he can make predictions about grade-point averages on the basis of achievement-test scores.

The need to analyze and interpret numerical data has made it necessary for educators and psychologists to have at least some basic understanding of statistical methods. Almost without exception, courses in statistics are required for education and psychology majors. In fact, the psychologist's need for special statistical tools has led to the development of new statistical techniques in recent decades. (Chapter 13 deals with some of these techniques.)

Because statistical techniques have been applied to a wide variety of research projects involving the study of individuals and groups, statistics has become, in most schools, a required course in sociology, anthropology, and related behavioral sciences.

Statistical methods have also been used extensively in the areas related to the biological sciences. In agriculture, they are used to determine the effects of seed strains, insecticides, and fertilizers on yields. In medicine, they are used to determine the possible side effects or effectiveness of drugs and to provide better methods to control the spread of contagious diseases. Statistics is a highly recommended, if not yet required, course for students majoring in these areas.

In recent years, statistics has found increasing applications in the physical sciences, where it has been used for handling the collection of data and the testing of hypotheses. The research needs of physical scientists have expanded the arena of the statistical technique of experimental design. It is hardly necessary to mention that in engineering the application of statistical principles to quality control has been an accepted practice for several decades.

One reason for the rapid growth in the application of statistics in recent decades is the increasing ease with which large quantities of numerical data can be handled. Electronic computers have made it possible to analyze, in a relatively short time, large quantities of data. The contribution of the desk calculator should not be ignored; it, too, has made the work of the statistician easier.

1-3 exercises

1. In what way is the electronic computer responsible for the increasingly widespread application of statistical techniques?
2. Explain why statistical techniques have found increasing application in business and economics.
3. Illustrate the use of descriptive statistics with an example from any field in the social sciences.
4. Illustrate the use of descriptive statistics with an example from the natural or physical sciences.
5. Illustrate the use of inferential statistics with an example from the social sciences.
6. Illustrate the use of inferential statistics with any example from the natural or physical sciences.

1-4 scope of the text

As mentioned in the preface, this is a textbook for beginning students who have never had a course in college mathematics. The approach is intuitive rather than mathematical; theoretical discussions are reduced to a minimum.

Most students have been exposed to the basic mathematical notations in high school mathematics courses; nevertheless, a brief review of the summation sign and various useful notations will help them understand better the material they will encounter in the text. These notations are reviewed in chapter 2. Chapter 3 deals with the concepts of events, the counting of events, and probability; chapter 4 discusses the axioms and basic rules of probability, the idea of a probability function, and discrete probability functions.

Descriptive statistics is treated in chapters 5 and 6, with emphasis on the arithmetic mean and the dispersion of observations for the sample as well as the population.

The rest of the text is devoted to a study of inferential statistics. Chapter 7 presents the normal distribution and the central limit theorem. These concepts are of paramount importance because they are invariably involved or implied in the techniques and principles covered in later chapters. Chapters 8 through 13 cover the most basic and important topics of statistical inference, including the methods of inference about a single mean, the difference between two means, and the equality among three or more means. They also deal with tests about goodness of fit, statistical independence, and simple regression and correlation coefficients. The text ends with chapter 14, which deals with the rudiments of modern decision theory.