WILEY PUBLICATIONS IN STATISTICS

Elementary Statistics

PAUL G. HOEL
Professor of Mathematics
University of California
Los Angeles

John Wiley & Sons, Inc.
New York · London

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 60–9671

PRINTED IN THE UNITED STATES OF AMERICA

Preface

This book is designed for a one-semester course for the student whose background in mathematics is limited to high-school algebra. A number of mathematics departments offer a service course of this kind; however, in many schools a department with strong interests in the applications of statistics gives the course.

My interest in writing a book at this level arose from having taught such a service course and being somewhat dissatisfied with the available texts. I felt that a mathematical statistician could write a descriptive book just as well as someone primarily interested in the applications and at the same time present the theory in a more satisfactory manner. Blandishments on the part of the publisher of my *Introduction To Mathematical Statistics* were undoubtedly as influential in causing the book to be written as these less mundane reasons.

I have some rather positive views on how elementary statistics should be taught. In particular, I believe in teaching students the simple classical ideas of sampling very thoroughly rather than attempting to introduce them to some of the more sophisticated modern notions. As in mathematics, there is a considerable body of material that needs to be understood before one can proceed to the study of more recent material. In writing this book I have attempted to include what I consider to be basic material, and I have tried to write for the student's benefit rather than to impress my colleagues.

The material in the first nine chapters, exclusive of the sections marked by a boldface arrowhead (▶), should suffice for the ordinary one-semester course. The arrowhead sections of these chapters contain material for amplifying the basic course if time permits. The last four chapters have been included as additional optional material to enable the instructor with particular interests to include topics related to those interests. None of these chapters is strictly elementary, although the nonparametric methods are relatively easy to explain.

I feel compelled to justify the existence of the last chapter. Time series is not a subject that I include in my own course, but several referees urged me to include a brief introduction to the topic as an optional chapter. I feel that many of the statistical concepts involved are too advanced and too delicate for beginning undergraduates. My approach has been to select some of the more elementary topics that are both sound and useful. As a consequence, though the treatment may not satisfy some economists and statisticians, it should represent an honest and elementary introduction from the viewpoint of students.

A number of individuals whose names I do not know read an early version of the manuscript and made many helpful suggestions for its improvement. If the book is slightly more difficult than I first intended it to be, the blame is partly theirs because most of the suggestions involved the addition of somewhat more difficult material. The revised manuscript was given a final reading by Professor W. K. Smith, who was most helpful in improving its exposition by his exceptionally careful reading of it. One of my graduate students, Robert Jennrich, was very helpful in eliminating errors in the answers to the exercises. Answers to the even-numbered problems can be obtained in pamphlet form from the publisher.

PAUL G. HOEL

Los Angeles, California
January 1960

Contents

CHAPTER 1

The Nature of Statistical Methods

1. INTRODUCTION

Statistical methods are often described as methods for treating numerical data. Such a definition, however, is much too broad in scope. It is necessary to restrict both the nature of the data and the reasons for studying them before such methods can rightfully be called statistical.

Statisticians are concerned with data that have been obtained from taking observations, in the form of measurements or counts, from a source of such observations. For example, in studying the quality of bricks in a certain brickyard, a number of bricks would be selected and tested for quality; or, in studying public opinion on a controversial measure being considered by a city council, a small percentage of the inhabitants of the city would be selected and asked whether they favored the measure.

Statisticians take observations of the type just described for the purpose of drawing conclusions about the source of the observations. Thus, bricks are selected and tested for quality for the purpose of trying to determine the quality of the entire supply of bricks in the brickyard. Similarly, the purpose of questioning only a small percentage of the inhabitants of a city concerning a controversial measure is to determine to a satisfactory approximation the opinions of all the inhabitants on this measure.

The set of observations that is taken from some source of observations for the purpose of obtaining information about the source is called a *sample*, whereas the source of those observations is called a *population*. In view of the preceding discussion, *statistical methods may be described as methods for drawing conclusions about populations by means of samples.* The single word "statistics" is often used in place of statistical methods. Thus a student who is taking a course in statistics is taking a course in statistical methods.

At first glance, the foregoing definition may seem to be rather technical and contrary to the popular notion about statistics. For example, many business people look upon statistical methods as methods for collecting and summarizing business facts. The Federal government employs a number of statisticians whose principal duty is to design efficient ways of collecting and summarizing various kinds of information. According to the preceding definition of statistical methods, these statisticians do not appear to be using statistical methods because they do not apply the information they have collected for drawing conclusions about the sources of the information. This viewpoint, however, does not take into account the fact that such information is gathered for the consumption of others who will use it to reach conclusions. Business concerns do not collect and summarize business facts just to admire the information obtained. They expect to use the information to make decisions, and whether or not they openly arrive at conclusions concerning the sources of the information the fact remains that they do make decisions on the basis of samples.

That part of statistical methods concerned with the collecting and summarizing of data is usually called *descriptive statistics*. The part concerned with drawing conclusions about the source of the data is called *statistical inference*. Since the ultimate objective is to make inferences, that is, draw conclusions, the descriptive part of statistics should be looked upon as a sort of preliminary to the main bout.

The use of statistical methods has increased remarkably in the last few decades, particularly in the biological and social sciences. Such methods have also proved very useful in various branches of the physical sciences and engineering. Because of this varied and strong interest, these methods have developed rapidly and have increased in complexity and diversity; nevertheless, many of the most important techniques are quite simple and are the same for all branches of application. Some of these universal methods are studied in this book.

2. ILLUSTRATIONS

This section describes a few problems of the type that statistical methods were designed to solve. It does not begin to cover the broad class of problems capable of being solved by statistical methods but rather illustrates a few of the simpler ones that can be solved by using only the methods developed in this book. One problem is of academic interest, whereas the others are typical real-life problems.

(*a*) A television program sponsor wishes to know how popular his program is, compared to others at the same hour. In particular, he wishes to know what percentage of the television audience is viewing his program rather than some other. To satisfy him, an organization engaged in determining program popularity agrees to take a poll of the television audience at that hour to evaluate program preferences. By using statistical methods, such an organization can decide how large a poll will be necessary in order to estimate, within any desired degree of accuracy, the percentage of the audience viewing this program.

(*b*) A medical research team has developed a new serum it hopes will help to prevent a common children's disease. It wishes to test the serum. In order to assist the researchers in carrying out such a test, a school system in a large city has agreed to inoculate half of the children in certain grades with the serum. Records of all children in those grades are kept during the following year. On the basis of the percentages of those children who contract the disease during that year, both for the inoculated group and for the remaining half, it is possible by statistical methods to determine whether the serum is really beneficial.

(*c*) An industrial firm is concerned about the large number of accidents occurring in its plant. In the process of trying to discover the various causes of such accidents, an investigator considers factors related to the time of day. He collects information on the number of accidents occurring during the various working hours of the day, and by using statistical methods he is able to show that the accident rate increases during the morning and also during the afternoon. Further statistical studies then reveal some of the major contributing factors involved in these accidents.

One might be tempted to say that statistical methods are not needed in a problem such as this, and that all one needs to do is to calculate percentages and look at them to decide what is happening. If one has a large amount of properly selected data, such decisions will often be correct; however the high cost of collecting data usually forces one to work with only small amounts and it is precisely in such situations that statistical methods are needed to yield valid conclusions.

(*d*) A rabbit farmer, interested in experimenting with different rations for his rabbits, wishes to determine the effect on gain in weight of giving rabbits increased amounts of a certain food. In order to study the effect, he uses a standard ration and six new rations obtained by adding from one to six ounces, respectively, of this food to the standard ration. Seventy rabbits are divided into 7 groups of 10 each, with each group receiving one of the rations. After the rabbits have been on those rations for a certain period of time, their gains in weight are determined. By using statistical methods, it is then possible to estimate the increased gains due to the increased amounts of the supplementary food, together with the accuracy of those estimates.

(*e*) An instructor of an elementary statistics course is having difficulty convincing some of his students that the chances of winning from a slot machine are just as good immediately after someone has won some money as after a run of losses. For the purpose of convincing them, he, together with a few students of sterling character, performs the following experiment on a slot machine located in a private golf club. The machine is played for one hour, or until the combined resources of instructor and students are exhausted, whichever occurs first. A record is kept of the number of wins and losses that occur immediately after a win, together with the amounts won, and also of the number of wins and losses, and amounts won, immediately after a run of, say, five losses. With data of this type available, the instructor should be able to apply statistical methods to convince the skeptics of his wisdom in this matter. Since a run of bad luck might make it difficult to demonstrate this wisdom, unless the machine were played a long while, the instructor would be well advised to come amply supplied with cash. An experiment of this type should also convince the students that slot machines are designed to extract money from naïve individuals.

3. ESTIMATION AND HYPOTHESIS TESTING

An analysis of the preceding illustrations will show that they properly belong to the field of statistics because all are concerned with drawing conclusions about some population of objects or individuals and propose to do so on the basis of a sample from that population.

It may also be observed that these problems fall into two general categories. They are concerned either with estimating some property of the population or with testing some hypothesis about the population. The first illustration, for example, is concerned with estimating the percentage of the television audience that is watching a particular program at a certain hour. The second illustration is one of testing the hypothesis that the percentage of children contracting a disease is the same for inoculated children as for children receiving no inoculation. The third illustration considers the problem of testing the hypothesis that the accident rate for a population of workers is constant over the day. The fourth illustration is concerned with estimating the average gain in weight of rabbits as a function of increasing the amounts of a food supplementing a standard diet. The fifth illustration is one of testing the hypothesis that the average amount of money won from a slot machine after a run of losses is the same as after a win.

Most of the statistical methods to be explained in this book are those for treating problems of these two types, namely, estimating properties of or testing hypotheses about populations. Although there are other types of conclusions or decisions that can be related to populations on the basis of samples, the bulk of those made by statisticians falls into one of the two aforementioned categories, and therefore they alone are studied in this book.

4. PROBABILITY

In the problem of estimating the percentage of a certain kind of television audience the solution will consist of a percentage based on the sample and a statement of the accuracy of the estimate, usually in the form of "the probability is .95 that the estimate will be in error by less than 3 per cent." Similarly, in problems involving the testing of some hypothesis the decision to accept or reject the hypothesis will be based on certain probabilities.

It is necessary to use probability in such conclusions because a conclusion based on a sample involves incomplete information about the population, and therefore it cannot be made with certainty. The magnitude of the probability attached to a conclusion represents the degree of confidence one should have in the truth of the conclusion. The basic ideas and rules of probability are studied in a later chapter; meanwhile it should be treated from an intuitive point of view. Thus the statement that the probability is .95 that an estimate will be in error by less than 3 per cent should be interpreted as meaning that about 95 per cent of such statements made by a statistician are valid and about 5 per cent are not. In the process of studying statistical methods one will soon discover that probability is the basic tool of those methods.

Probability is an exceedingly interesting subject, even for those who have little liking for mathematics or quantitative methods. Many people enjoy some of the events associated with probability, if not the study itself; otherwise, how can one account for the large number of people who love to gamble at horse races, lotteries, cards, etc.? It may well be that it is their lack of probability sense that encourages them to gamble as they do. In any case, probability is used consciously or unconsciously by everyone in making all sorts of decisions based on uncertainty, and any student who wishes to be well educated, or to behave rationally, should have some knowledge of probability.

5. ORGANIZATION

The study of the statistical methods discussed in the preceding sections will proceed by first considering properties of samples and then properties of populations. As indicated in section 1, such studies constitute the descriptive part of statistics. It will then be possible to consider the two basic problems of statistical inference, namely the problems of estimation and hypothesis testing. This means that for any given type of problem the sample data will always be studied first before any attempt is made to introduce a theoretical population from which the sample might have come. In Chapter 2 a beginning is made in the study of properties of sample data, after which some basic theoretical populations are introduced. It is at this theoretical stage that probability will appear.

CHAPTER 2

The Description of Sample Data

1. INTRODUCTION

Since the purpose of this chapter is to study properties of samples taken from populations, it would seem necessary to agree first on how samples are to be taken because the desired properties may well depend upon the method employed. Suppose, for example, that a student newspaper reporter has been assigned the task of determining the percentage of students having part time jobs. He might attempt to get an estimate of this percentage by polling the first 100 students he encountered in front of the student union. This method of sampling, however, is not likely to give a valid estimate of the population percentage because students found lolling in front of the union are often the campus loafers and social butterflies, and they are seldom the working type. The reporter would undoubtedly do much better if he were to select 100 cards blindly from the student enrollment card file and poll the selected students.

The problem of how to select a sample from a population so that valid conclusions about the population can be drawn from the sample is quite complicated. This problem is discussed rather extensively in Chapter 5. In that discussion a method of sampling called random sampling is advocated and justified. Anticipating that material somewhat, it will be assumed hereafter that whenever a sample is to be taken it will be obtained by a random sampling method. For the present, it suffices to depend on one's intuition and to think of random sampling as a method of sampling in which individuals are selected "blindly" or "haphazardly." Later discussions clarify this language; meanwhile one should let his natural understanding of the meaning of words guide him here.

Now turn to the problem of studying properties of samples taken from populations. Only some of the simpler properties are discussed here.

In this connection, consider the problem of what a physical education department at a university would do if it were interested in determining whether its male dormitory students were typical university students with respect to physical characteristics. In such a study it would undoubtedly wish to compare, as one source of information, the weight distribution of the dormitory students with that of nondormitory students. Now, weighing every male student on campus would certainly yield the desired information on weight distribution; however, this would become quite an undertaking in a large school at which such information is not required at registration time. The desired information, to sufficient accuracy, could be obtained much more easily by studying the weight distributions of samples of dormitory and nondormitory students.

Suppose then that a random sample of, say, 120 students has been obtained from the dormitory population. Since the only concern here is what to do with samples, the nondormitory sample can be ignored in this discussion—it would be treated in the same manner as the dormitory sample. Suppose, furthermore, that the weights of these 120 students have been recorded to the nearest pound and that they range from the lightest at 110 pounds to the heaviest at 218 pounds.

It is very difficult to look at 120 measurements and obtain any reasonably accurate idea of how those measurements are distributed. For the purpose of obtaining a better idea of the weight distribution of the 120 students, it is therefore convenient to condense the data somewhat by classifying the measurements into groups. It will then be possible to graph the modified distribution and learn more about the original set of 120 measurements. This condensation will also be useful for simplifying the computations of various averages that need to be evaluated, particularly if fast computing facilities are not available. These averages will supply additional information about the distribution. Thus the purpose of classifying data is to assist in the extraction of certain kinds of useful information from the data.

The weight measurements considered here comprise an example of observations made on what is called a *continuous variable.* This name is applied to variables, such as length, weight, temperature, and time, that can be thought of as capable of assuming any value in some interval of values. Thus the weight of a student in the 140–150 pound range can be deemed capable of assuming any value in this range. Variables such as the number of automobile accidents during a day, the number of beetles dying when they are sprayed with an insecticide, or the number of children in a family are examples of what is called a *discrete variable.* For the purposes of this book, discrete variables can be considered as variables whose possible values are integers; hence they involve counting rather than measuring.

Since any measuring device is of limited accuracy, measurements in real life are actually discrete in nature rather than continuous; however, this should not deter one from thinking about such variables as being continuous. Although the dormitory weights have been recorded to the nearest pound, they should be regarded as the values of a continuous variable, the values having been rounded off to the nearest integer. When a weight is recorded as, say, 152 pounds, it is assumed that the actual weight is somewhere between 151.5 and 152.5 pounds.

2. CLASSIFICATION OF DATA

The problem of classifying the data of a sample usually arises only for continuous variables because discrete variables by their very nature are naturally classified; therefore, consider the problem for the 120 dormitory weight measurements. What needs to be done is to place each weight in its proper weight class, for instance, between 130 and 140 pounds. Experience and theory indicate that for most types of data it is desirable to use from 10 to 20 classes, with the smaller number of classes for smaller quantities of data. With less than about 10 classes, too much sample detail is lost, whereas with more than about 20 classes computations become unnecessarily tedious. In order to determine boundaries between the various class intervals, it is necessary merely to know the smallest and largest measurements of the set. For the weight data, these are 110 and 218 pounds, respectively. Since the range of values, which is 108 pounds here, is to be divided into 10 to 20 equal intervals, the length of the class interval is first determined for those two extreme cases. If 10 intervals were chosen, the class interval length would be $108/10 = 10.8$ pounds, whereas if 20 intervals were chosen, it would be $108/20 = 5.4$ pounds. Any convenient number between 5.4 and 10.8 may therefore be chosen. A class interval length of 10 pounds will evidently be very convenient. Other class interval lengths such as 6, 7, 8, or 9 would have been satisfactory also, although preference should be given to one of the larger intervals because 120 is not considered to be a large number of measurements. Since the first class interval should contain the smallest measurement of the set, it must begin at least as low as 110. Furthermore, in order to avoid having measurements fall on the boundary of two adjacent class intervals, it is customary to choose class boundaries to $\frac{1}{2}$ a unit beyond the accuracy of the measurements. Thus in this problem, with weights recorded to the nearest pound, it is satisfactory to choose the first class interval as 109.5–119.5, since 109.5 is $\frac{1}{2}$ a unit below the smallest measurement of 110, and it was agreed to use 10 pounds as the length of the class interval. This interval is certain to contain the smallest measurement, in view of the fact that a recorded weight of 110 pounds represents an actual weight between 109.5 and 110.5 pounds. The remaining class boundaries are determined by merely adding the class interval length 10 repeatedly until the largest measurement, namely 218, is enclosed in the final interval. If 109.5–119.5 is chosen as the first class interval, there will be 11 class intervals, and the last class interval will turn out to be 209.5–219.5. When the class boundaries have been determined, it is a simple matter to list each measurement of the set in its proper class interval by merely recording a short vertical bar to represent it, as shown in Table 1(a). When the number of bars corresponding to each class interval has been listed, the data are said to have been classified in a *frequency table.*

It is assumed in such a classification that all measurements in a given class interval have been assigned the value at the mid-point of the interval. This mid-point value is called the *class mark* for that interval. Thus, for the first interval, 109.5–119.5, the class mark is 114.5, and any weight within this interval is assigned the value 114.5. This means, for example, that the smallest measurement, 110, will be replaced by the measurement 114.5. The process of classification thus replaces a set of measurements by a new, more convenient, set, whose values are approximately equal to the original values; the approximation, however, is usually very good.

Table 1 illustrates the tabulation (a) and resulting frequency table (b) for a set of 120 weights of the type under consideration. The class marks are usually listed in such a table because they are the new values assigned to the measurements. The letter x is used to denote a class mark, and the letter f, to denote the corresponding frequency. Sub-

TABLE **1**

(a)		(b)	
Class Boundaries	Frequencies	Class Marks: x	Frequencies: f
109.5–119.5	/	114.5	1
119.5–129.5	////	124.5	4
129.5–139.5	𝍸 𝍸 𝍸 //	134.5	17
139.5–149.5	𝍸 𝍸 𝍸 𝍸 𝍸 ///	144.5	28
149.5–159.5	𝍸 𝍸 𝍸 𝍸 𝍸	154.5	25
159.5–169.5	𝍸 𝍸 𝍸 ///	164.5	18
169.5–179.5	𝍸 𝍸 ///	174.5	13
179.5–189.5	𝍸 /	184.5	6
189.5–199.5	𝍸	194.5	5
199.5–209.5	//	204.5	2
209.5–219.5	/	214.5	1

scripts on x and f designate the class interval. Thus $x_1, x_2, x_3, \ldots, x_{11}$ denote the class marks for the 11 class intervals in Table 1, and $f_1, f_2, f_3, \ldots, f_{11}$ denote the corresponding frequencies. For example, $x_2 = 124.5$, and $f_2 = 4$. The letter n is used to denote the total number of measurements. Since the sum of the frequencies for the various intervals must equal the total number of measurements, it follows that

$$n = f_1 + f_2 + \cdots + f_h,$$

in which h denotes the number of class intervals in the frequency table.

Magazines and newspapers often indicate class intervals in a slightly different manner from that suggested here. They do not record actual class interval boundaries but rather noncontiguous boundaries. Thus they would indicate the first three class intervals in the preceding problem by 110–119, 120–129, and 130–139. When intervals are so indicated, the boundaries, as defined earlier, are ordinarily half way between the upper and lower recorded boundaries of adjacent intervals. Another method used by them employs common boundaries but agrees that an interval includes measurements up to but not including the upper boundary. With this method, the first three class intervals would be indicated by 110–120, 120–130, and 130–140. A measurement that falls on a boundary is placed in the higher of the two intervals. These alternative methods are undoubtedly used because the reading public finds them easier to follow. If one knows the accuracy of measurement of the variable involved, there will be little difficulty in determining the correct class marks for the two methods of classification. It is important to use the correct class marks; otherwise a systematic error will be introduced in many of the computations to follow.

3. GRAPHICAL REPRESENTATION

A frequency distribution, such as that in Table 1, is easier to visualize if it is represented graphically. A particularly useful type of graph for this kind of classified data is a graph called a *histogram*. The histogram for the frequency distribution of Table 1 is shown in Fig. 1. The class boundaries of Table 1 are marked off on the x axis starting and finishing at any convenient points. The frequency corresponding to any class interval is represented by the height of the rectangle whose base is the interval in question. The vertical axis is therefore the frequency, or f, axis. Histograms are particularly useful graphs for later work when frequency distributions of populations are introduced.

The histogram of Fig. 1 is typical of many frequency distributions obtained from data found in nature and industry. They usually range from a rough bell-shaped distribution, such as that in Fig. 2, to something