WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

Introduction to Mathematical Statistics

FOURTH EDITION

PAUL G. HOEL

Professor of Mathematics
University of California
Los Angeles

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Catalog Card Number: 70-139277
ISBN 0-471-40365-2

Printed in the United States of America

20 19 18 17 16 15 14

Preface

This edition differs from the third edition principally in the organization of material. The first six chapters have been rewritten to enable the instructor to postpone statistical inference until after the material on probability has been covered. This will permit a two-quarter (or two-semester) course to be given that will satisfy both the students who desire only a one-quarter course in probability and those who want a two-quarter course in probability and statistics. In order to organize a course in this manner it is necessary to postpone Chapter 4 and section 9 of Chapter 5 until after Chapter 6 has been completed.

Although Chapter 4, which is a gentle introduction to statistical inference, may be postponed until after the basic material on probability has been studied, I feel that so little time is required to cover it and so much insight into statistical reasoning is obtained by studying it that it is a mistake to omit it even for students who wish to learn only probability. It should be part of the knowledge acquired by any student who is interested in probability.

The reorganization of material required considerable rewriting of the text. This was particularly true in the early chapters on probability in which it was necessary to formalize the notation and treatment. The later chapters differ very little from those of the earlier edition except in the choice of topics. The exercises at the end of each chapter have been changed considerably by changing the data in many of the numerical problems (so that old student files on solutions will be worthless) and by adding new problems. The exercises have been labeled with the number of the section to which they belong, beginning with routine numerical problems and followed by theoretical ones. The more difficult problems occur at the end of each set.

Since this book is written for the student who has only an elementary calculus background in mathematics, much of the statistical theory must be accepted on faith. Some of the more important proofs are outlined in the appendix for the benefit of those students who possess more mathematical maturity than that normally obtained from an elementary calculus course.

From time to time I have received letters and reviews from some of the users of the earlier editions of this book with suggestions for its improvement.

I have always appreciated such letters and reviews even though I may not have included all those suggestions in the revision. I wish to thank those who have used my book over the years and especially those who have submitted ideas for its improvement.

Paul G. Hoel

Los Angeles, California
January 1971

Contents

CHAPTER 1

Introduction

Mathematical statistics is the study of how to deal with data by means of probability models. It grew out of methods for treating data that were obtained by some repetitive operation such as those encountered in games of chance and in industrial processes. These methods soon found application in such diverse fields as medical research, insurance, marketing, agriculture, chemistry, and industrial experimentation.

In its broadest sense, statistical methods are often described as methods for making decisions in the face of uncertainty. The outcome of an experiment is usually uncertain but, hopefully, if it is repeated a number of times one may be able to construct a probability model for it and make decisions concerning the experimental process by means of it. Although probabilty can be applied to experiments or situations in which it is difficult to conceive of them as being of the repetitive type, the emphasis in this book will be on the repetitive type situation.

Experience indicates that many repetitive operations or experiments behave as though they occurred under essentially stable circumstances. Games of chance, such as coin tossing or dice rolling, usually exhibit this property. Many experiments and operations in the various branches of science and industry do likewise. Under such circumstances, it is often possible to construct a satisfactory mathematical model of the repetitive operation. This model can then be employed to study properties of the operation and to draw conclusions concerning it. Although mathematical models are especially useful devices for studying real-life problems when the model is realistic of the actual operation involved, it often happens that such models prove useful even though the operation is not highly stable.

The mathematical model that a statistician selects for a repetitive operation is usually one that enables him to make predictions about the frequency with which certain results can be expected to occur when the operation is repeated a number of times. For example, the model for studying the inheritance of color in the propagation of certain flowers might be one that predicted three times as many flowers of one color as of another color. In the investigation of the quality of manufactured parts the model might be one that predicts the percentage of defective parts that can be expected in the manufacturing process.

Because of the nature of statistical data and models, it is only natural that probability should be the fundamental tool in statistical theory. The statistician looks on probability as an idealization of the proportion of times that a certain result will occur in repeated trials of an experiment; consequently, a probability model is the type of mathematical model selected by him. Because probability is so important in the theory and applications of statistical methods, an introduction to probability is given before the study of statistical methods as such is taken up.

Although probability is being interpreted herein as an idealized relative frequency, it is treated as a measure of an individual's betting odds by those individuals who apply probability to a broader class of problems than those considered in this book. In applying probability techniques to such problems, however, it is necessary to realize that the reliability of a decision is heavily dependent on the realism of the individual's betting odds.

The idea of a mathematical model for assisting in the solution of real-life problems is a familiar one in the various sciences. For example, a physicist studying projectile motion often assumes that the simple laws of mechanics yield a satisfactory model, in spite of the complexity of the actual problem. For more refined work, he introduces a more complicated model. Since a model is only an idealization of the actual situation, the conclusions derived from it can be relied on only to the extent that the model chosen is a sufficiently good approximation to the actual situation being studied. In any given problem, therefore, it is essential to be well acquainted with the field of application in order to know what models are likely to be realistic. This is just as true for statistical models as for models in the various branches of science.

The science student will soon discover the similarity between certain of the statistical methods and certain scientific methods in which the scientist sets up a hypothesis, conducts an experiment, and then tests the hypothesis by means of his experimental data. Although statistical methods are applicable to all branches of science, they have been applied most actively in the biological and social sciences because the laboratory methods of the physical sciences have not been sufficiently broad to treat many of the problems of those other sciences. Problems in the biological and social sciences often involve undesired variables that cannot be controlled, as contrasted to the physical sciences in which such variables can often be controlled satisfactorily in the laboratory. Statistical theory is concerned not only with how to solve certain problems of the various sciences but also with how experiments in those sciences should be designed. Thus the science student should expect to learn statistical techniques to assist him in treating his experimental data and in designing his experiments in a more efficient manner.

The theory of statistics can be treated as a branch of mathematics in which probability is the basic tool; however, since the theory developed from an attempt to solve real-life problems, much of it would not be fully appreciated if it were removed from such applications. Therefore the theory and the applications are considered simultaneously throughout this book, although the emphasis is on the theory.

In the process of solving a real-life problem in statistics three steps may be recognized. First, a mathematical model is selected. Second, a check is made of the reasonableness of the model. Third, the proper conclusions are drawn from this model to solve the proposed problem. In this book the emphasis is on the first and third steps. In order to do justice to the second step, it would be necessary to be well acquainted with the field of application. It would also be necessary to know how the conclusions are affected by changes in the assumptions necessary for the model.

Students who have not had experience with applied science are sometimes disturbed by the readiness with which a statistician will accept certain of his model assumptions as being sufficiently well satisfied in a given problem to justify confidence in the validity of the conclusions. One of the striking features of much of statistical theory is that its field of application is much broader than the assumptions involved would seem to justify. The rapid development of, and interest in, statistical methods during the last few decades can be attributed in part to the highly successful application of statistical techniques to so many different branches of science and industry.

CHAPTER 2

Probability

1 INTRODUCTION

An individual's approach to probability depends on the nature of his interest in the subject. The pure mathematician usually prefers to treat probability from an axiomatic point of view, just as he does, say, the study of geometry. The applied statistician usually prefers to think of probability as the proportion of times that a certain event will occur if the experiment related to the event is repeated indefinitely. The approach to probability here is based on a blending of these two points of view.

The statistician is usually interested in probability only as it pertains to the possible outcomes of experiments. Furthermore, most statisticians are interested in only those experiments that are repetitive in nature or that can be conceived of as being so. Experiments such as tossing a coin, counting the number of defective parts in a box of parts, or reading the daily temperature on a thermometer are examples of simple repetitive experiments. An experiment in which several experimental animals are fed different rations may be performed only once with those same animals; nevertheless, the experiment may be thought of as the first in an unlimited number of similar experiments and therefore may be conceived of as being repetitive.

2 SAMPLE SPACE

Consider a simple experiment such as tossing a coin. In this experiment there are but two possible outcomes, a head and a tail. It is convenient to represent the possible outcomes of such an experiment, and experiments in general, by points on a line or by points in higher dimensions. Here it would be convenient to represent a head by the point 1 on the x axis and a tail by the point 0. This choice is convenient because the number corresponds to the number of heads obtained in the toss. If the experiment had consisted of

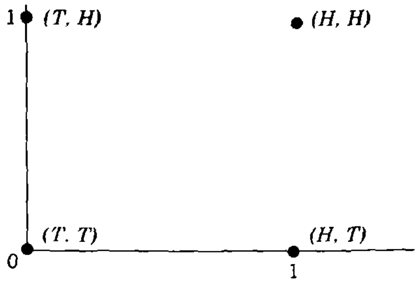

Fig. 1. A simple sample space.

tossing the coin twice, there would have been four possible outcomes, namely *HH*, *HT*, *TH*, and *TT*. For reasons of symmetry, it would be desirable to represent these outcomes by the points (1, 1), (1, 0), (0, 1), and (0, 0) in the x, y plane. Figure 1 illustrates this choice of points to represent the possible outcomes of the experiment.

If the coin were tossed three times, it would be convenient to use three dimensions to represent the possible experimental outcomes. This representation, of course, is merely a convenience, and if desired one could just as well mark off any eight points on the x axis to represent the eight possible outcomes.

In the experiment of rolling two dice, there are 36 possible outcomes, which have been listed in Table 1. The first number of each pair denotes

TABLE 1

11	21	31	41	51	61
12	22	32	42	52	62
13	23	33	43	53	63
14	24	34	44	54	64
15	25	35	45	55	65
16	26	36	46	56	66

the number that came up on one of the dice and the second number denotes the number that came up on the other die. It is assumed that the two dice are distinguishable or are rolled in order. For this experiment a natural set of points to represent the possible outcomes are the 36 points in the x, y plane whose coordinates are the corresponding number pairs of Table 1. This choice is shown in Fig. 2.

An experiment that consists of reading the temperature of a patient in a hospital has a very large number of possible outcomes depending on the degree of accuracy with which the thermometer is read. For such an experiment it is convenient to assume that the patient's temperature can assume

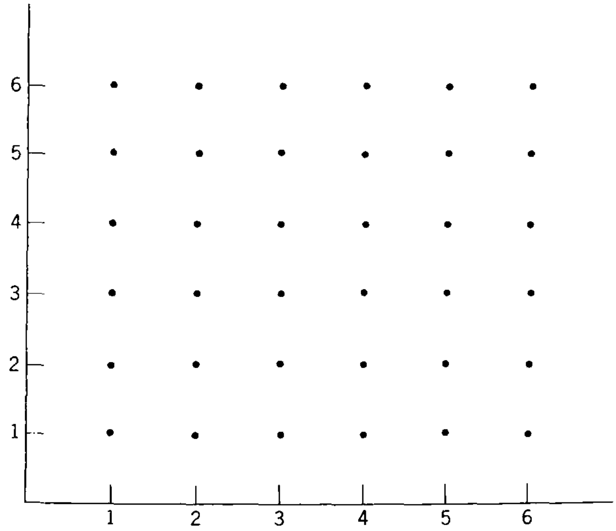

Fig. 2. A sample space for rolling two dice.

any value between, say, 95° and 110°; therefore the possible outcomes would be represented in a natural way by the points inside the interval from 95 to 110 on the x axis. This, of course, is a convenient idealization and ignores the impossibility of reading a thermometer to unlimited accuracy.

Definition: *The set of points representing the possible outcomes of an experiment is called the sample space of the experiment.*

The idea of a sample space is introduced because it is a convenient mathematical device for developing the theory of probability as it pertains to the outcomes of experiments.

3 EVENTS

Consider an experiment such that whatever the outcome of the experiment it can be decided whether an event A has occurred. This means that each sample point can be classified as one for which A will occur or as one for which A will not occur. Thus, if A is the event of getting exactly one head and one tail in tossing a coin twice, the two sample points (H, T) and (T, H) of Fig. 1 correspond to the occurrence of A. If A is the event of getting a total of seven points in rolling two dice, then A is associated with the six sample points (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1) of Fig. 2. If A is the event that a patient's temperature will be at least as high as 102, then A will consist of the interval of points from 102 to 110 on the x axis.

Definition: *An event is a subset of a sample space.*

Since a subset of a set of points is understood to include the possibility that the subset is the entire set of points or that it contains none of the points of the set, this definition includes an event that is certain to occur or one that cannot possibly occur when the experiment is performed.

In view of the correspondence between events and sets of points the study of the relationship between various events is reduced to the study of the relationship between the corresponding sets. For this purpose it is convenient to represent the sample space, whatever its dimension or whatever the number of points in it, by a set of points inside a rectangle in a plane. An event A, which is therefore a subset of the points in this rectangle, is represented by the points lying inside a closed curve contained in the rectangle. If B is some other event of interest, it will be represented by the points inside some other closed curve in the rectangle. This representation is shown in Fig. 3. No attempt has been made to indicate whether the number of points is finite or infinite because a representation for both types of sample spaces is desired.

If A and B are two events associated with an experiment, one may be interested in knowing whether at least one of the events A and B will occur when the experiment is performed. Now the set of points that consists of all points that belong to A, or B, or both A and B is called the union of A and B and is denoted by the symbol $A \cup B$. This set of points, which is shown as the shaded region in Fig. 3, therefore represents the event that at least one of the events A and B will occur.

As an illustration, if A is the event of getting a six on the first die when rolling two dice and B is the event of getting a six on the second die, then $A \cup B$ is the event of getting at least one six in rolling two dice. The event A consists of the six points found in the last column of points in the sample space shown in Fig. 2 and the event B consists of the six points found in the last row of points in that sample space. The event $A \cup B$ is then the set of eleven points found in the union of the last column and last row of points.

Another event of possible interest is that of knowing whether both events A and B will occur when the experiment is performed. The set of points that consists of all points that belong to both A and B is called the intersection of

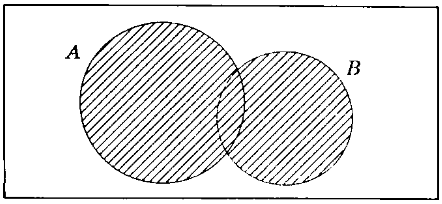

Fig. 3. Representation of $A \cup B$.

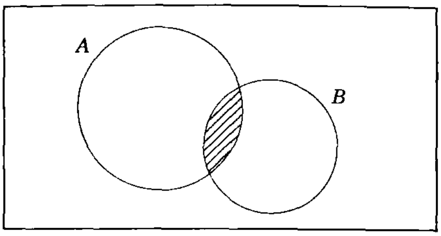

Fig. 4. Representation of $A \cap B$.

A and B and is denoted by the symbol $A \cap B$. This set of points, which is shown as the shaded region in Fig. 4, therefore represents the event that both A and B will occur when the experiment is performed.

In the preceding illustration concerning the two dice, $A \cap B$ is the event that both dice will show a six. It is represented by the single point (6, 6), which is the intersection of the last column, and last row, sets of points.

Corresponding to any event A there is an associated event, denoted by $\bar{A}$, which states that A will not occur when the experiment is performed. It is represented by all the points of the rectangle not found in A and is shown as the shaded region in Fig. 5. The set $\bar{A}$ is called the complement of the set A relative to the sample space.

If two sets, A and B, have no points in common they are said to be disjoint sets. In the language of events, such events are called *disjoint events*, but they are also called *mutually exclusive events*, because the occurrence of one of those events excludes the possible occurrence of the other.

4 PROBABILITY

The familiar functions of calculus are what are known as point functions. The function defined by the formula $f(x) = x^2$ is an example of a point function, for to each point on the x axis this formula assigns the value of the function. The notion of function is much broader than this, however, and permits the elements of the domain of the function to be sets of points rather than individual points. In this case the function is called a *set function*. As

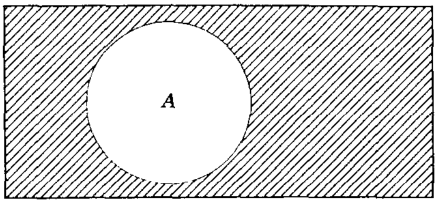

Fig. 5. Representation of $\bar{A}$.