PROBABILITY AND STATISTICS

PROBABILITY AND STATISTICS

Morris H. DeGroot
Carnegie-Mellon University

Menlo Park, California · London · Don Mills, Ontario · Sydney

 Printed in the United States of America. Published simultaneously in Canada. Library of Congress Catalog Card No. 74-19691.

ISBN 0-201-01503-X
BCDEFGHIJ-MA-798765

Preface

This book contains enough material for a one-year course in probability and statistics. The mathematical requirements for the course are a knowledge of the elements of calculus and a familiarity with the concepts and elementary properties of vectors and matrices. No previous knowledge of probability or statistics is assumed.

This book has been written with both the student and the teacher in mind. Special care has been taken to make sure that the text can be read and understood with few obscure passages or other stumbling blocks. Theorems and proofs are presented where appropriate, and illustrative examples are given at almost every step of the way. More than 800 exercises are included in the book. Some of these exercises provide numerical applications of results presented in the text, and others are intended to stimulate further thought about these results.

The first five chapters are devoted to probability and can serve as the text for a one-semester course on that topic. The elementary concepts of probability are illustrated by such famous examples as the birthday problem, the tennis tournament problem, the matching problem, the collector's problem, and the game of craps. Standard material on random variables and probability distributions is highlighted by discussions of the use of a table of random digits, the elementary notions of life testing, a comparison of the relative advantages of the mean and the median as predictors, and the importance of the central limit theorem. Also included as special features of these chapters are sections on Markov chains, the Gambler's Ruin problem, and utility and preferences among gambles. These topics are treated in a completely elementary fashion, but they can be omitted without loss of continuity if time is limited.

The last five chapters of the book are devoted to statistical inference. The coverage here is modern in outlook. Both classical and Bayesian statistical methods are developed in an integrated presentation. No single school of thought is treated in a dogmatic fashion. My goal is to equip the student with the theory and methodology that have proved to be useful in the past and promise to be useful in the future.

These chapters contain a comprehensive but elementary survey of estimation, testing hypotheses, nonparametric methods, multiple regression, and the analysis of variance. The strengths and weaknesses and the advantages and disadvantages of such basic concepts as maximum likelihood estimation, Bayesian decision procedures, unbiased estimation, confidence intervals, and levels of significance are discussed from a contemporary viewpoint. Special features of these chapters include discussions of prior and posterior distributions, sufficient statistics, Fisher information, the Bayesian analysis of samples from a normal distribution, multidecision problems, tests of goodness-of-fit, contingency tables, inferences about the median and other quantiles, and the important and current topic of robust estimation and trimmed means. If time does not permit complete coverage of the contents of these chapters, any of the following sections can be omitted without loss of continuity: 7.6, 7.8, 8.3, 9.5, 9.6, 9.7, 9.8, and 9.9.

Although a computer can be a valuable adjunct in a course in probability and statistics such as this one, none of the exercises in this book requires access to a large-scale computer or a knowledge of programming. For this reason, the use of this book is not tied to a computer in any way. Instructors are urged, however, to utilize computers in the course as much as is feasible. A small calculator is a helpful aid for solving some of the numerical exercises in the second half of the book.

One further point about the style in which the book is written should be emphasized. The pronoun "he" is used throughout the book in reference to a person who is confronted with a statistical problem. This usage certainly does not mean that only males calculate probabilities and make decisions, or that only males can be statisticians. The word "he" is used quite literally as defined in Webster's Third New International Dictionary to mean "that one whose sex is unknown or immaterial." The field of statistics should certainly be as accessible to women as it is to men. It should certainly be as accessible to members of minority groups as it is to the majority. It is my sincere hope that this book will help create among all groups an awareness and appreciation of probability and statistics as an interesting, lively, and important branch of science.

A preliminary version of this material was used in a course at Carnegie-Mellon University taught by my colleague John P. Lehoczky. The book has been significantly improved as a result of Professor Lehoczky's sage and sensible advice. I have also benefited from the helpful comments of Joseph B. Kadane and Paul Shaman. Prem K. Goel and Richard S. Luckew were of great assistance in the early stages of the manuscript. Frederick Mosteller (Harvard University), consulting editor, and David Bedworth (Arizona State University, Tempe), Robert

Berk (Rutgers University), Maurice Bryson (Colorado State University, Fort Collins), H. C. Fryer (Kansas State University, Manhattan), Merrill Hume (University of Missouri, Rolla), Eugene Lukacs (Bowling Green State University), Loren Meeker (University of New Hampshire), and W. R. Schucany (Southern Methodist University) read the manuscript for Addison-Wesley and made many useful suggestions. At the final stage, Chandra Gulati's comments were valuable.

The editorial work of my father, Archibald DeGroot, has improved the manuscript substantially. The book would probably not have been completed without the friendship and support of Richard M. Cyert. It would probably never have been started without Dolores, who brought so much that is worthwhile into my life.

Dolores died on September 18, 1974. During her life she fought hard for the dignity and rights of all human beings. She was brilliant and talented, gentle and beautiful, and very brave. She will be missed.

Pittsburgh, Pennsylvania M.H.D.
November, 1974

CONTENTS

Introduction to Probability

1.1. THE HISTORY OF PROBABILITY

The concepts of chance and uncertainty are as old as civilization itself. People have always had to cope with uncertainty about the weather, their food supply, and other aspects of their environment, and have strived to reduce this uncertainty and its effects. Even the idea of gambling has a long history. Games of chance played with bone objects that could be considered precursors of dice were apparently highly developed in Egypt and elsewhere by about the year 3500 B.C. Cubical dice with markings virtually identical to those on modern dice have been found in Egyptian tombs dating from 2000 B.C. We know that gambling with dice has been popular ever since that time and played an important part in the early development of probability theory.

It is generally believed that the mathematical theory of probability was started by the French mathematicians Blaise Pascal (1623–1662) and Pierre Fermat (1601–1665) when they succeeded in deriving exact probabilities for certain gambling problems involving dice. Some of the problems that they solved had been outstanding for about 300 years. However, numerical probabilities of various dice combinations had been calculated by Girolamo Cardano (1501–1576) more than a century before Pascal and Fermat, and by Galileo Galilei (1564–1642) more than 50 years before them.

The theory of probability has been developed steadily since the seventeenth century and has been widely applied in diverse fields of study. Today, probability theory is an important tool in most areas of engineering, science, and management. Many research workers are actively engaged in the discovery and establishment of new applications of probability in fields such as medicine, the control of spaceships, and marketing.

References

The ancient history of gambling and the origins of the mathematical theory of probability are discussed by David (1962), Ore (1960), and Todhunter (1865).

Some introductory books on probability theory, which discuss many of the same topics that will be studied in this book, are Feller (1968); Hoel, Port, and Stone (1971); and Parzen (1960). Other introductory books, which discuss both probability theory and statistics at about the same level as they will be discussed in this book, are Blum and Rosenblatt (1972), Hogg and Craig (1970), Kempthorne and Folks (1971), Lindgren (1968), and Mood and Graybill (1963).

1.2. INTERPRETATIONS OF PROBABILITY

In addition to the many formal applications of probability theory, the concept of probability enters our everyday life and conversation. We often hear and use such expressions as: "It probably will rain tomorrow afternoon"; "It is very likely that the plane will arrive late"; or "The chances are good that he will be able to join us for dinner this evening." Each of these expressions is based on the concept of the probability, or the likelihood, that some specific event will occur.

Despite the fact that the concept of probability is such a common and natural part of our experience, no single scientific interpretation of the term probability is accepted by all statisticians, philosophers, and other authorities in this field. Through the years, each interpretation of probability that has been proposed by some authorities has been criticized by others. Indeed, the true meaning of probability is still a highly controversial subject and is involved in many current philosophical discussions pertaining to the foundations of statistics. Three different interpretations of probability will be described here. Each of these interpretations can be very useful in applying probability theory to practical problems.

The Frequency Interpretation of Probability

In many problems, the probability that some specific outcome of a process will be obtained can be interpreted to mean the *relative frequency* with which that outcome would be obtained if the process were repeated a large number of times under similar conditions. For example, the probability of obtaining a head when a coin is tossed is considered to be $\frac{1}{2}$ because the relative frequency of heads should be approximately $\frac{1}{2}$ when the coin is tossed a large number of times under similar conditions. In other words, it is assumed that the proportion of tosses on which a head is obtained would be approximately $\frac{1}{2}$.

Of course, the conditions mentioned in this example are too vague to serve as the basis for a scientific definition of probability. First, a "large number" of tosses of the coin is specified, but there is no definite indication of an actual number that would be considered large enough. Second, it is stated that the coin should be tossed each time "under similar conditions," but these conditions are not described precisely. The conditions under which the coin is tossed must not be completely identical for each toss because the outcomes would then be the same, and there would be either all heads or all tails. In fact, a skilled person can toss a coin into the air repeatedly and catch it in such a way that a head is obtained on almost every toss. Hence, the tosses must not be completely controlled, but must have some "random" features.

Furthermore, it is stated that the relative frequency of heads should be "approximately $\frac{1}{2}$," but no limit is specified for the permissible variation from $\frac{1}{2}$. If a coin were tossed 1,000,000 times, we would not expect to obtain exactly 500,000 heads. Indeed, we would be extremely surprised if we obtained exactly 500,000 heads. On the other hand, neither would we expect the number of heads to be very far from 500,000. It would be desirable to be able to make a precise statement of the likelihoods of the different possible numbers of heads, but these likelihoods would of necessity depend on the very concept of probability that we are trying to define.

Another shortcoming of the frequency interpretation of probability is that it applies only to a problem in which there can be, at least in principle, a large number of similar repetitions of a certain process. Many important problems are not of this type. For example, the frequency interpretation of probability cannot be applied directly to the probability that a specific acquaintance will get married within the next two years or to the probability that a particular medical research project will lead to the development of a new treatment for a certain disease within a specified period of time.

The Classical Interpretation of Probability

The classical interpretation of probability is based on the concept of *equally likely outcomes*. For example, when a coin is tossed, there are two possible outcomes: either a head or a tail. If it may be assumed that these outcomes are equally likely to occur, then they must have the same probability. Since the sum of the probabilities must be 1, both the probability of a head and the probability of a tail must be $\frac{1}{2}$. More generally, if the outcome of some process must be one of n different outcomes, and if these n outcomes are equally likely to occur, then the probability of each outcome is $1/n$.

Two basic difficulties arise when an attempt is made to develop a formal definition of probability from the classical interpretation. First, the concept of equally likely outcomes is essentially based on the concept of probability that we are trying to define. The statement that two possible outcomes are equally likely to occur is the same as the statement that two outcomes have the same probability. Second, no systematic method is given for assigning probabilities to outcomes that are not assumed to be equally likely. When a coin is tossed, or a well balanced die is rolled, or a card is chosen from a well shuffled deck of cards, the different possible outcomes can usually be regarded as equally likely because of the nature of the process. However, when the problem is to guess whether an acquaintance will get married or whether a research project will be successful, the possible outcomes would not typically be considered to be equally likely, and a different method is needed for assigning probabilities to these outcomes.

The Subjective Interpretation of Probability

According to the subjective, or personal, interpretation of probability, the probability that a person assigns to a possible outcome of some process represents his own judgment of the likelihood that the outcome will be obtained. This judgment will be based on that person's beliefs and information about the process. Another person, who may have different beliefs or different information, may assign a different probability to the same outcome. For this reason, it is appropriate to speak of a certain person's *subjective probability* of an outcome, rather than to speak of the *true probability* of that outcome.

As an illustration of this interpretation, suppose that a coin is to be tossed once. A person with no special information about the coin or the way in which it is tossed might regard a head and a tail to be equally likely outcomes. That person would then assign subjective probability $\frac{1}{2}$ to the possibility of obtaining a head. The person who is actually tossing the coin, however, might feel that a head is much more likely to be obtained than a tail. In order for this person to be able to assign subjective probabilities to the outcomes, he must express the strength of his belief in numerical terms. Suppose, for example, that he regards the likelihood of obtaining a head to be the same as the likelihood of obtaining a red card when one card is chosen from a well shuffled deck containing four red cards and one black card. Since the person would assign probability $\frac{4}{5}$ to the possibility of obtaining a red card, he should also assign probability $\frac{4}{5}$ to the possibility of obtaining a head when the coin is tossed.

This subjective interpretation of probability can be formalized. In general, if a person's judgments of the relative likelihoods of various combinations of outcomes satisfy certain conditions of consistency, then it can be shown that his subjective probabilities of the different possible events can be uniquely determined. However, there are two difficulties with the subjective interpretation. First, the requirement that a person's judgments of the relative likelihoods of an infinite number of events be completely consistent and free from contradictions does not seem to be humanly attainable. Second, the subjective interpretation provides no "objective" basis for two or more scientists working together to reach a common evaluation of the state of knowledge in some scientific area of common interest.

On the other hand, recognition of the subjective interpretation of probability has the salutary effect of emphasizing some of the subjective aspects of science. A particular scientist's evaluation of the probability of some uncertain outcome must ultimately be his own evaluation based on all the evidence available to him. This evaluation may well be based in part on the frequency interpretation of probability, since the scientist may take into account the relative frequencv of occurrence of this outcome or similar outcomes in the past. It may also be based in part on the classical interpretation of probability, since the scientist may take into account the total number of possible outcomes that he considers equally likely to occur. Nevertheless, the final assignment of numerical probabilities is the responsibility of the scientist himself.

The subjective nature of science is also revealed in the actual problems that the scientist chooses to study from the class of problems that might have been chosen, in the experiments that he decides to perform in carrying out this study, and in the conclusions that he draws from his experimental data. The mathematical theory of probability and statistics can play an important part in these choices, decisions, and conclusions. Moreover, this theory of probability and statistics can be developed, and will be presented in this book, without regard to the controversy surrounding the different interpretations of the term probability. This theory is correct and can be usefully applied, regardless of which interpretation of probability is used in a particular problem. The theories and techniques that will be presented in this book have served as valuable guides and tools in almost all aspects of the design and analysis of effective experimentation.

1.3. EXPERIMENTS AND EVENTS

Types of Experiments

The theory of probability pertains to the various possible outcomes that might be obtained and the possible events that might occur when an experiment is performed. The term "experiment" is used in probability theory to describe virtually any process whose outcome is not known in advance with certainty. Some examples of experiments will now be given.

1. In an experiment in which a coin is to be tossed 10 times, the experimenter might want to determine the probability that at least four heads will be obtained.
2. In an experiment in which a sample of 1000 transistors is to be selected from a large shipment of similar items and each selected item is to be inspected, a person might want to determine the probability that not more than one of the selected transistors will be defective.
3. In an experiment in which the air temperature at a certain location is to be observed every day at noon for 90 successive days, a person might want to determine the probability that the average temperature during this period will be less than some specified value.
4. From information relating to the year of birth of Thomas Jefferson, a certain person might want to determine the probability that Jefferson was born in 1741.

5. In evaluating an industrial research and development project, a person might want to determine, at a certain time, the probability that the project will result in the successful development of a new product within a specified number of months.

It can be seen from these examples that the possible outcomes of an experiment may be either random or nonrandom, in accordance with the usual meanings of those terms. The interesting feature of an experiment is that each of its possible outcomes can be specified before the experiment is performed, and probabilities can be assigned to various combinations of outcomes that are of interest.

The Mathematical Theory of Probability

As was explained in Section 1.2, there is controversy in regard to the proper meaning and interpretation of some of the probabilities that are assigned to the outcomes of many experiments. However, once probabilities have been assigned to some simple outcomes in an experiment, there is complete agreement among all authorities that the mathematical theory of probability provides the appropriate methodology for the further study of these probabilities. Almost all work in the mathematical theory of probability, from the most elementary textbooks to the most advanced research, has been related to the following two problems: (i) methods for determining the probabilities of certain events from the specified probabilities of each possible outcome of an experiment and (ii) methods for revising the probabilities of events when additional relevant information is obtained.

These methods are based on standard mathematical techniques. The purpose of the first five chapters of this book is to present these techniques which, together, form the mathematical theory of probability.

1.4. SET THEORY

The Sample Space

The collection of all possible outcomes of an experiment is called the *sample space* of the experiment. In other words, the sample space of an experiment can be thought of as a *set*, or collection, of different possible outcomes, and each outcome can be thought of as a *point*, or an *element*, in the sample space. Because of this interpretation, the language and concepts of set theory provide a natural context for the development of probability theory. The basic ideas and notation of set theory will now be reviewed.

Relations of Set Theory

Let S denote the sample space of some experiment. Then any possible outcome s of the experiment is said to be a member of the space S, or to belong to the space S. The statement that s is a member of S is denoted symbolically by the relation $s \in S$.

When an experiment has been performed and we say that some event occurred, we mean that the outcome of the experiment satisfied certain conditions which specified that event. In other words, some outcomes in the space S signify that the event occurred, and all other outcomes in S signify that the event did not occur. In accordance with this interpretation, any event can be regarded as a certain subset of possible outcomes in the space S.

For example, when a six-sided die is rolled, the sample space can be regarded as containing the six numbers 1, 2, 3, 4, 5, 6. Symbolically, we write

$$S = \{1, 2, 3, 4, 5, 6\}.$$

The event A that an even number is obtained is defined by the subset $A = \{2, 4, 6\}$. The event B that a number greater than 2 is obtained is defined by the subset $B = \{3, 4, 5, 6\}$.

It is said that an event A *is contained in* another event B if every outcome that belongs to the subset defining the event A also belongs to the subset defining the event B. This relation between two events is expressed symbolically by the relation $A \subset B$. The relation $A \subset B$ is also expressed by saying that A is a subset of B. Equivalently, if $A \subset B$, we may say that B contains A and may write $B \supset A$.

In the example pertaining to the die, suppose that A is the event that an even number is obtained and C is the event that a number greater than 1 is obtained. Since $A = \{2, 4, 6\}$ and $C = \{2, 3, 4, 5, 6\}$, it follows that $A \subset C$. It should be noted that $A \subset S$ for any event A.

If two events A and B are so related that $A \subset B$ and $B \subset A$, it follows that A and B must contain exactly the same points. In other words, $A = B$.

If A, B, and C are three events such that $A \subset B$ and $B \subset C$, then it follows that $A \subset C$. The proof of this fact is left as an exercise.

The Empty Set

Some events are impossible. For example, when a die is rolled, it is impossible to obtain a negative number. Hence, the event that a negative number will be obtained is defined by the subset of S that contains no outcomes. This subset of S is called the *empty set*, or *null set*, and it is denoted by the symbol $\emptyset$.

Now consider any arbitrary event A. Since the empty set $\emptyset$ contains no points, it is logically correct to say that any point belonging to $\emptyset$ also belongs to A, or $\emptyset \subset A$. In other words, for any event A, it is true that $\emptyset \subset A \subset S$.

Operations of Set Theory

Unions. If A and B are any two events, the *union* of A and B is defined to be the event containing all outcomes that belong to A alone, to B alone, or to both A and B. The notation for the union of A and B is $A \cup B$. The event $A \cup B$ is sketched in Fig. 1.1. A sketch of this type is called a *Venn diagram*.