Hoel
Port
Stone

Introduction to Statistical Theory

Introduction to Statistical Theory

Paul G. Hoel
Sidney C. Port
Charles J. Stone
University of California, Los Angeles

HOUGHTON MIFFLIN COMPANY BOSTON
Atlanta Dallas Geneva, Illinois Hopewell, New Jersey
Palo Alto London

PRINTED IN THE U.S.A.

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 70-136172

ISBN: 0-395-04637-8

The Houghton Mifflin Series in Statistics
under the Editorship of Herman Chernoff

LEO BREIMAN
Probability and Stochastic Processes: With a View Toward Applications
Statistics: With a View Toward Applications

Y. S. CHOW, HERBERT ROBBINS, AND DAVID SIEGMUND
Great Expectations: The Theory of Optimal Stopping

PAUL G. HOEL, SIDNEY C. PORT, AND CHARLES J. STONE
Introduction to Probability Theory
Introduction to Statistical Theory
Introduction to Stochastic Processes

PAUL F. LAZARSFELD AND NEIL W. HENRY
Latent Structure Analysis

GOTTFRIED E. NOETHER
Introduction to Statistics—A Fresh Approach

I. RICHARD SAVAGE
Statistics: Uncertainty and Behavior

General Preface

This three-volume series grew out of a three-quarter course in probability, statistics, and stochastic processes taught for a number of years at UCLA. We felt a need for a series of books that would treat these subjects in a way that is well coordinated, but which would also give adequate emphasis to each subject as being interesting and useful on its own merits.

The first volume, *Introduction to Probability Theory*, presents the fundamental ideas of probability theory and also prepares the student both for courses in statistics and for further study in probability theory, including stochastic processes.

The second volume, *Introduction to Statistical Theory*, develops the basic theory of mathematical statistics in a systematic, unified manner. Together, the first two volumes contain the material that is often covered in a two-semester course in mathematical statistics.

The third volume, *Introduction to Stochastic Processes*, treats Markov chains, Poisson processes, birth and death plocesses, Gaussian processes, Brownian motion, and processes defined in terms of Brownian motion by means of elementary stochastic differential equations.

Preface

This book is designed for a one-semester course in mathematical statistics. It was written in close conjunction with *Introduction to Probability Theory*, the first volume of our three-volume series, and assumes that the student is acquainted with the material covered in a one-semester course in probability for which elementary calculus is a prerequisite.

The objective of this book is to present an elementary systematic treatment of mathematical statistics from a theoretical point of view. An attempt has been made to restrict consideration to important fundamental ideas and to describe these in some detail, so that the student will appreciate the motivation as well as the mathematics of the theory. Too often students who have finished a course in statistics come away with only a vague notion of the central ideas and methods of the subject. It is hoped that this text will uncover the unity and logical structure of statistical methods.

The theoretical development has been based on a few of the elementary notions of decision theory. This permits the treatment of Bayesian methods in addition to the more traditional methods; however, space did not permit the introduction of more than the most basic of these methods. The Bayesian techniques occur at the end of each chapter; therefore they can be omitted if time does not permit their inclusion.

One of the most important theorems and most useful techniques in statistics is concerned with testing the general linear hypothesis. This theorem is the foundation of numerous special tests and it can be applied to a host of important problems. A proof of the theorem is seldom presented at this elementary level; however, because of the importance of the theorem and because elementary calculus students are now receiving some training in matrix algebra, a proof based on simple algebraic and geometric techniques is presented. This material, which occurs in Chapter 5, is undoubtedly the most difficult part of the book, but its mastery is well worth the effort. For students who do not possess the necessary algebraic background, it is best to skip the proofs in this chapter and pass on to the applications.

Although elementary calculus suffices as a prerequisite for *Introduction to Probability Theory*, the present volume also assumes some elementary knowledge of matrix algebra. This knowledge will be needed in Chapter 4 as well as in Chapter 5, because the least squares theory in that chapter is presented by means of matrix notation and techniques. However it is only in Chapter 5 that a student needs to know anything more than the simplest notions of matrix algebra. A review of the matrix methods that are needed for these two chapters is presented in an appendix.

Some instructors may be surprised to discover that the concept of sufficiency is not introduced in this book. Sufficiency is very useful in developing statistical theory at an advanced level, but it would serve no useful purpose at this introductory stage. There are several other topics that are often found in introductory texts which are not included here. The justification for such omissions is that classroom time spent on such topics would leave insufficient time for an adequate discussion of the basic material.

The exercises at the end of each chapter are arranged according to the order in which the material of that chapter was presented. Problems of a computational nature occur first. Answers are given in an appendix.

Although this book was designed for a one-semester course meeting three times a week, it is sufficiently flexible in arrangement of material to adjust itself to a shorter course. This is easily accomplished either by omitting the chapter on nonparametric methods, or by omitting the material on Bayesian methods, or by omitting some of the proofs, or by a combination of such omissions. Sections that may be omitted are indicated by an asterisk. A somewhat longer course can be accommodated by including all the theory and spending more time on the exercises.

We would like to thank Frank Samaniego for obtaining answers to many of the exercises and Mrs. Gerry Formanack for her excellent typing.

Table of Contents

1 Basic Principles

Many phenomena in the various sciences are governed by laws or relations of a stochastic nature. This implies that a probability model may be appropriate for representing the occurrence of the phenomenon. Such models were introduced and applied to games of chance and other physical experiments in Volume I, *Introduction to Probability Theory*. Although the physical sciences yield experiments that are likely to be more stable than those in the nonphysical sciences, and hence for which a probability model might seem to be more appropriate, such models can be just as appropriate in these other fields. The fundamental difference is that there may be more uncontrolled variables interfering with the variable being studied and leading to greater variability of it. The application of probability models to phenomena in these various sciences led to the development of methods that are now commonly called the methods of statistics.

Some simple typical problems that the methods of statistics are designed to solve are:

deciding on the basis of testing a few samples from a shipment of a certain drug whether the quality of the shipment is satisfactory,

predicting on the basis of a small poll what the voters' preferences are on a vital issue,

calculating on the basis of a high school student's record and the records of students who have gone to college what the chances are that he will be successful in college,

deciding on the basis of analyzing sonar signals whether a submarine is approaching.

Problems of this type can be formulated mathematically by considering the data that are to be used for making a decision as the observed values of a random variable X. The distribution of X is assumed to belong to a certain family of distributions, a particular member of which is specified when the value of a parameter θ is specified. The problem is to decide on the basis of the data which member, or members, of the family could represent the distribution of X. This is called *the problem of statistical inference.* For example, in the problem of determining voter preferences on an issue by means of a poll, the variable X may be treated as a discrete random variable with X assuming the value 1 or 0 corresponding to an individual's favoring or not favoring the issue, and with the parameter θ representing the unknown proportion of voters favoring the issue. A typical problem of statistical inference is to decide on the basis of a set of responses $(x_1, \ldots, x_n)$ obtained by a pollster whether the value of θ exceeds .60.

The methods of statistics are much broader in scope than the statistical inference problems illustrated here would suggest. They concern themselves also with such problems as how experiments should be conducted, what models are appropriate, and how information should be utilized. However, we shall be concerned almost exclusively with finding the best methods for making inferences about distributions of random variables. This means that we assume our random variable X possesses a given type of distribution depending upon an unknown parameter θ and that our objective is to draw some inference concerning θ. In most problems θ will be an ex licit parameter of a probability density function $f(x \mid \theta)$; however, θ may be merely an index to distinguish different members of a family of such functions. The random variable X and the parameter θ may be vector variables with several components each; but in our discussion of basic principles they will be treated as one dimensional to simplify the exposition. The extension to vector variables will be considered in the next chapter. Typical probability models of this type are the binomial distribution with θ representing the probability of success in a single trial of an experiment, the Poisson distribution with θ representing the mean of the distribution, and the normal distribution with known variance and with θ representing the mean of the distribution. If both the mean and variance of a normal distribution were unknown, θ would represent the vector parameter (μ, σ).

In an inference problem such as the one where θ is the proportion of voters favoring an issue, the parameter θ is considered to be a fixed but unknown constant at the time the poll is taken. However, in some problems the parameter θ may be treated as a random variable with a known probability distribution. If so, the distribution will be assumed to be given by a density function, $\pi(\theta)$. The function $f(x \mid \theta)$ will then represent a conditional distribution with the variable θ fixed, and the joint distribution of X and θ will be given by the density $f(x, \theta) = \pi(\theta) f(x \mid \theta)$. Here the word density is used for both discrete and continuous random variables. This convention was used in Volume I and will be used throughout this book without further reminders. There is a slight inconsistency in notation here because a capital letter normally represents a random variable and the corresponding small letter its numerical value, whereas θ is being used here to represent both the random variable and its numerical value. When θ is treated as a random variable its probability distribution is called a *prior distribution*. This name arises because $\pi(\theta)$ is known prior to the experimentation that is carried out to make an inference concerning θ. An illustration of a problem for which θ might be treated as a random variable is that of deciding by means of testing a sample taken from a shipment whether the proportion θ of defectives in the shipment exceeds some tolerance proportion θ_0. Assume the purchasing firm receives such shipments regularly and has determined a distribution for θ from past shipment θ's. The θ for this shipment may be treated as the value, although unknown, of a random variable possessing this distribution. Most of the classical methods of statistical theory treat θ as a constant and rely only on a set of observed values of the random variable X for drawing inferences. This is partly because for many of the problems in the social and life sciences, from which much of statistical theory evolved, no $\pi(\theta)$ is available or appropriate. It would, of course, be a mistake to ignore prior information on θ if such information were available, even if it is not expressible in the form of a precise probability distribution.

To recapitulate, we shall assume that we are given a probability distribution of a random variable X that depends upon a parameter θ, and that we wish to make some inference concerning θ on the basis of some observed values of the random variable X and of a prior distribution for θ (if available).

1.1. Types of problems

Since an inference is to be made by means of a set of observed values $x_1, \ldots, x_n$ of the random variable X, it is necessary to introduce a function $d = d(x_1, \ldots, x_n)$ of those values for making the inference. Such a function is called a *decision function.* The nature of this function will depend upon the kind of inference concerning θ that is to be made. In the simplest problems we merely wish to know whether a certain proposition is true or false. For example, we might wish to know whether a shipment of drugs is up to quality specifications, whether a radar scanning has picked up a missile, or whether the number of children in a school district who suffer from malnutrition exceeds ten percent. We shall take a positive point of view in decision making by associating any decision with an action. Thus, in the preceding problems there will be two possible actions available, which will be denoted by a_1 and a_2, with a_1 corresponding to the decision of accepting the truth of the proposition and a_2 corresponding to its rejection. For a set of n observational or sample values, we will have an n-dimensional sample space. Since a decision function $d(x_1, \ldots, x_n)$ must determine for each point of this sample space whether action a_1 or a_2 is to be taken, such a function must separate the sample space into exactly two parts, one part consisting of those sample points for which a_1 will be taken and the other part consisting of points for which action a_2 will be taken. For example, if small values of X correspond to the truth of a proposition, a possible division of an n dimensional sample space might be to assign all points inside the sphere $x_1^2 + \cdots + x_n^2 = r^2$, where r is a suitably chosen constant, to a_1 and all other points to a_2. Problems of the preceding type in which there are only two possible actions are called *hypothesis testing problems.*

A more complicated decision making problem arises when there are more than two possible actions available. For example, suppose a certain region in Europe is known to have been inhabited by five different races of people. An archaeologist might wish to decide on the basis of bone measurements taken of a group of skeletons found in that region to which of the five races they belonged. Because of the simplicity of hypothesis testing procedures, problems involving more than two possible actions are sometimes incorrectly treated as hypothesis testing problems. For example, if a new drug is introduced as a cure for a disease, it is important to decide whether the drug is superior, inferior, or about equally effective in curing the disease; therefore it would be improper to treat it as a two action problem, in which one tests, for example, only whether the drug is superior or is equally effective in curing the disease. Problems in which there are a finite number, $k > 2$, possible actions available are called *multiple decision problems.* A decision function for such problems must divide the sample space into k parts, the ith part consisting of those sample points that are associated with taking action a_i, $i = 1, \ldots, k$.

A third class of problems arises when interest centers on trying to predict the value of the parameter θ and there are an infinite number of possibilities for θ. Thus if θ represents the proportion of voters who will vote for candidate A, it may be important to have a precise estimate of that proportion, rather than merely to decide whether it exceeds 1/2. The decision function $d(x_1, \ldots, x_n)$ will then be a real-valued function whose range of values theoretically may be taken to be the interval $[0, 1]$. Problems of this type are called *estimation problems.*

As an illustration of how these three types of problems could arise in the same experimental situation, suppose that two new drugs for lowering blood pressure are to be compared for effectiveness. Let X denote the ratio of a patient's blood pressure after treatment to his blood pressure before treatment, and let θ represent the mean value of this random variable with respect to a class of patients. A typical problem of testing a hypothesis is to decide on the basis of experimentation with both drugs whether $\theta_1 \leq \theta_2$ or $\theta_1 > \theta_2$, where θ_1 and θ_2 correspond to the two drugs. From a practical point of view there is little point in preferring one drug to the other unless it shows a meaningful advantage. Thus, it might be more realistic to treat the problem as a multiple decision problem by considering the three possibilities $\theta_1 - \theta_2 \leq -\delta$, $-\delta < \theta_1 - \theta_2 < \delta$, $\theta_1 - \theta_2 \geq \delta$, where δ is the smallest difference that is considered practically useful. If the experiment yielded the decision, for whichever formulation was chosen, that the first drug is superior, then additional experimentation with this drug would be desirable so that an accurate estimate of θ_1 could be obtained.

1.2. The risk function

The success of a given decision function in accomplishing its objective needs to be measured in a numerical manner. If an experimenter is able to assign weights to the seriousness of making various incorrect decisions, these weights can serve to define a *loss function* $\mathscr{L}(\theta, a)$. This function is designed to numerically measure the penalty that arises from taking action a when θ is the true value of the parameter. For example, if we were given the observational values $x_1, \ldots, x_n$ of a normal variable with unknown mean θ, and we wished to use them to estimate θ, the action a would consist in stating that the value of θ is $d(x_1, \ldots, x_n)$. The magnitude of the error in this decision would be given by $|\theta - d(x_1, \ldots, x_n)|$. A typical loss function might then be $\mathscr{L}(\theta, a) = |\theta - a|$. To indicate the dependence of the loss function upon the decision function and the observational values, we will express it in the form $\mathscr{L}(\theta, d(x_1, \ldots, x_n))$. The name loss is attached to this function to indicate that the objective is to minimize $\mathscr{L}$. Since we wish to minimize our decision errors, it is clear that we would like $\mathscr{L}$ to be a function that decreases as the magnitude of the error decreases. The problem of how to choose $\mathscr{L}$ for the three types of problems discussed before will be considered in Section 1.4.

Problems arise in which the available decisions are qualitative in nature and for which it would be difficult to assign a numerical value to incorrect decisions. Thus, an individual might be faced with a choice of five color schemes for redecorating his house. If aesthetic considerations are as important as monetary ones in such a choice, it would be inconvenient to assign a loss function here. Problems of this type can be treated satisfactorily if one is able to assign a preference ordering to the possible choices by employing a numerical valued function, the utility function, that is based on this ordering. Even for problems of a quantitative nature, it is often necessary to introduce a utility function to express in quantitative form one's preferences among the various possibilities. We shall assume hereafter that if the problem is one in which it is necessary or desirable to introduce a utility function, then $\mathscr{L}$ represents that function.

Thus far the discussion has been on the basis of having available a set of observed values $x_1, \ldots, x_n$ of some random variable X and then trying to select a good function of those values. In measuring the effectiveness of any decision function, we must look at its overall performance and not just at how well it does for a single experiment. We therefore consider an experiment in which a set of n observations is to be taken of some random variable X. These potential observational values will be denoted by $X_1, \ldots, X_n$. If the observations are to be obtained by random sampling, then we know from the definition of random sampling in Volume I that these n random variables will be independently and identically distributed with the same distribution as that of X. A decision function $d = d(X_1, \ldots, X_n)$ is then a random variable, and therefore the loss $\mathscr{L} = \mathscr{L}(\theta, d(X_1, \ldots, X_n))$ is also a random variable. To measure the overall effectiveness of a decision function we calculate the expected value of the loss function and use it as our measure. This defines a new function called the risk function $\mathscr{R}$. Thus,

Definition 1 *The risk function $\mathscr{R}$ is given by the formula*

$$\mathscr{R}(\theta, d) = E_\theta \mathscr{L}(\theta, d(X_1, \ldots, X_n)),$$

where the expectation is taken with respect to the distribution of the random variables $X_1, \ldots, X_n$ with θ fixed.

Under random sampling and for the situation in which X possesses the density $f(x \mid \theta)$, the joint distribution of these n random variables is given by $\prod_{i=1}^{n} f(x_i \mid \theta)$. If X is a continuous random variable the risk function will then be given by

$$(1) \qquad \mathscr{R}(\theta, d) = \int \cdots \int \mathscr{L}(\theta, d(x_1, \ldots, x_n)) \prod_{i=1}^{n} f(x_i \mid \theta)\, dx_1 \cdots dx_n.$$

For a discrete random variable these integrals must be replaced by corresponding sums over all possible values of the x's. Since it is inconvenient to write out multiple integrals, an abbreviated notation will be used in which X will represent the basic random variable if a sample of size one is to be taken but will represent the vector random variable $X = (X_1, \ldots, X_n)$ if a sample of size $n > 1$ is to be taken. Then we may replace the multiple integration notation of (1) by the more compact representation

$$(2) \qquad \mathscr{R}(\theta, d) = \int \mathscr{L}(\theta, d(x)) f(x \mid \theta)\, dx$$

where now $f(x \mid \theta)$ denotes the density of the vector variable X and dx represents $dx_1 \cdots dx_n$.

Now suppose we wish to compare two decision functions d_1 and d_2 by means of their risk functions $\mathscr{R}(\theta, d_1)$ and $\mathscr{R}(\theta, d_2)$. This comparison is most easily made by means of their graphs. Consider the two sets of graphs shown in Figures 1 and 2 which represent two possible occurrences. It is clear in Figure 1 that decision function d_1 is better than d_2 because its risk function value is less than that of d_2 for each value of θ and our objective is to minimize the risk function. In Figure 2, however, neither function is superior to the other because for some values of θ the function d_1 is better than d_2 but for other values the advantage is reversed.

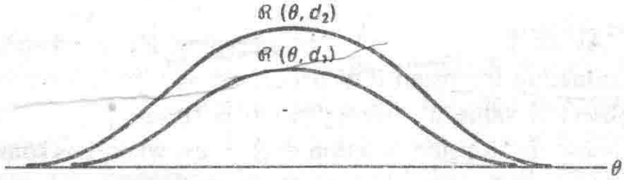

Figure 1

Unfortunately, the type of situation illustrated in Figure 1 rarely occurs, at least when the comparison is being made between decision functions that have been selected with intelligence. It would be difficult to make a choice between d_1 and d_2 for the more typical situation illustrated in Figure 2. In such cases, it is necessary to introduce some additional principle or criterion in order to arrive at a choice. One such principle that is quite popular is based on using the maximum value of the risk function as a criterion for comparison. If one risk function has a smaller maximum value than another, the decision function yielding the smaller maximum value is considered the better. If there is a decision function d whose risk function possesses a maximum value that is a minimum for all competing risk functions, d is called a best decision function in the minimax sense. Thus,

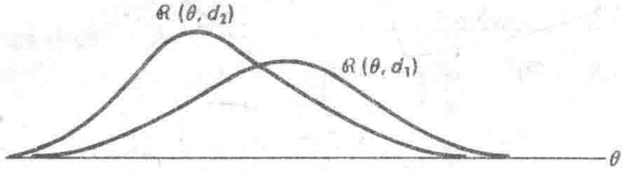

Figure 2

Definition 2 *The function d_0 is called a minimax decision function in the class D of decision functions if it satisfies*

$$\max_{\theta} \mathcal{R}(\theta, d_0) = \min_{d \in D} \max_{\theta} \mathcal{R}(\theta, d).$$

From Figure 2 it will be seen that d_1 is a better decision function than d_2 in the minimax sense, because its risk function clearly has a smaller maximum than the risk function for d_2. If the class D of decision functions consisted only of these two functions, then d_1 would be a minimax decision function relative to this class.

The advantage of introducing an additional principle, such as that of minimax, is that it reduces the comparison of decision functions to the comparison of real numbers. There are other principles that could be introduced here, but they will be discussed later when they are needed.

As an illustration of the preceding ideas, consider the problem of estimating the mean θ of a Poisson distribution on the basis of a single observed value X. Here $f(x \mid \theta) = (e^{-\theta}\theta^x)/x!$, $x = 0, 1, \ldots$. We shall choose the decision function $d(x) = cx$, where c is some positive constant, and we shall assume that the loss function is $\mathscr{L}(\theta, d) = (d - \theta)^2/\theta$. According to formula (1) the risk function is given by the sum

$$\mathscr{R}(\theta, d) = \sum_{x=0}^{\infty} \frac{(cx - \theta)^2}{\theta} \frac{e^{-\theta}\theta^x}{x!}.$$

The evaluation of $\mathscr{R}(\theta, d)$ is simpler, however, if the basic definition of $\mathscr{R}$ as the expected value of $\mathscr{L}$ is used and the properties of E that were derived in Volume I are employed; hence we calculate $E(cX - \theta)^2/\theta$. This expected value is most easily carried out by first writing $(cX - \theta)^2/\theta$ in the following form.

$$\frac{(cX - \theta)^2}{\theta} = \frac{c^2}{\theta}\left(X - \frac{\theta}{c}\right)^2 = \frac{c^2}{\theta}\left(X - \theta + \theta\left(1 - \frac{1}{c}\right)\right)^2$$

$$= \frac{c^2}{\theta}\left\{(X - \theta)^2 + 2\theta\left(1 - \frac{1}{c}\right)(X - \theta) + \theta^2\left(1 - \frac{1}{c}\right)^2\right\}.$$

Since X is a Poisson variable, we know from Volume I that the mean and variance of X are both equal to θ, and therefore that $E(X - \theta) = 0$ and $E(X - \theta)^2 = \theta$. Application of these facts to the preceding sum will yield the result

$$\text{(3)} \qquad \mathscr{R}(\theta, d) = \frac{c^2}{\theta}\left\{\theta + \theta^2\left(1 - \frac{1}{c}\right)^2\right\} = c^2 + \theta(c - 1)^2.$$

Now consider what value of c should be chosen to produce a good estimate of θ. If $c = 1$, the risk function has the constant value of 1. If $c \neq 1$, the risk function is a linear function of θ with a positive derivative, and therefore it will assume increasingly large values as θ becomes increasingly large. If there is no restriction on the value of θ, except that it must be positive, then $c = 1$ produces a minimax estimator in the class of estimators $d(X) = cX$. It is the only estimator in that class which yields a finite maximum for the risk function. The word estimator is used to denote a function of the random variables, whereas the word estimate denotes its numerical value.

1.3. Mean risk

If, in addition to the sample $X = (X_1, \ldots, X_n)$, a prior distribution for θ is available, then X and θ are both considered to be random variables.