TESTING STATISTICAL HYPOTHESES

E. L. LEHMANN

Testing Statistical Hypotheses

E. L. LEHMANN
Professor of Statistics
University of California, Berkeley

New York . John Wiley & Sons, Inc.
London . Chapman & Hall, Limited

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 59–11803

PRINTED IN THE UNITED STATES OF AMERICA

Preface

A mathematical theory of hypothesis testing in which tests are derived as solutions of clearly stated optimum problems was developed by Neyman and Pearson in the 1930's and since then has been considerably extended. The purpose of the present book is to give a systematic account of this theory and of the closely related theory of confidence sets, together with their principal applications. These include the standard one- and two-sample problems concerning normal, binomial, and Poisson distributions; some aspects of the analysis of variance and of regression analysis (linear hypothesis); certain multivariate and sequential problems. There is also an introduction to nonparametric tests, although here the theoretical approach has not yet been fully developed. One large area of methodology, the class of methods based on large-sample considerations, in particular χ^2 and likelihood ratio tests, essentially has been omitted because the approach and the mathematical tools used are so different that an adequate treatmen would require a separate volume. The theory of these tests is only briefly indicated at the end of Chapter 7.

At present the theory of hypothesis testing is undergoing important changes in at least two directions. One of these stems from the realization that the standard formulation constitutes a serious oversimplification of the problem. The theory is therefore being re-examined from the point of view of Wald's statistical decision functions. Although these investigations throw new light on the classical theory, they essentially confirm its findings. I have retained the Neyman-Pearson formulation in the main part of this book but have included a discussion of the concepts of general decision theory in Chapter 1 to provide a basis for giving a broader justification of some of the results. It also serves as a background for the development of the theories of hypothesis testing and confidence sets.

Of much greater importance is the fact that many of the problems, which traditionally have been formulated in terms of hypothesis testing, are in reality multiple decision problems involving a choice between several decisions when the hypothesis is rejected. The development of suitable procedures for such problems is at present one of the most important tasks of statistics and is finding much attention in the current literature. However, since most of the work so far has been tentative, I have preferred to present the traditional tests even in cases in which the majority of the applications appear to call for a more elaborate procedure, adding only a warning regarding the limitations of this approach. Actually, it seems likely that the tests will remain useful because of their simplicity even when a more complete theory of multiple decision methods is available.

The natural mathematical framework for a systematic treatment of hypothesis testing is the theory of measure in abstract spaces. Since introductory courses in real variables or measure theory frequently present only Lebesgue measure, a brief orientation with regard to the abstract theory is given in Sections 1 and 2 of Chapter 2. Actually, much of the book can be read without knowledge of measure theory if the symbol $\int p(x)\, d\mu(x)$ is interpreted as meaning either $\int p(x)\, dx$ or $\Sigma p(x)$, and if the measure theoretic aspects of certain proofs together with all occurrences of the letters a.e. (almost everywhere) are ignored. With respect to statistics, no specific requirements are made, all statistical concepts being developed from the beginning. On the other hand, since readers will usually have had previous experience with statistical methods, applications of each method are indicated in general terms but concrete examples with data are not included. These are available in many of the standard textbooks.

The problems at the end of each chapter, many of them with outlines of solutions, provide exercises, further examples, and introductions to some additional topics. There is also given at the end of each chapter an annotated list of references regarding sources, both of ideas and of specific results. The notes are not intended to summarize the principal results of each paper cited but merely to indicate its significance for the chapter in question. In presenting these references I have not aimed for completeness but rather have tried to give a usable guide to the literature.

An outline of this book appeared in 1949 in the form of lecture notes taken by Colin Blyth during a summer course at the University of California. Since then, I have presented parts of the material in courses at Columbia, Princeton, and Stanford Universities and several times at the University of California. During these years I greatly

benefited from comments of students and I regret that I cannot here thank them individually. At different stages of the writing I received many helpful suggestions from W. Gautschi, A. Høyland, and L. J. Savage, and particularly from Mrs. C. Striebel, whose critical reading of the next to final version of the manuscript resulted in many improvements. Also, I should like to mention gratefully the benefit I derived from many long discussions with Charles Stein.

It is a pleasure to acknowledge the generous support of this work by the Office of Naval Research; without it the book would probably not have been written. Finally, I should like to thank Mrs. J. Rubalcava, who typed and retyped the various drafts of the manuscript with unfailing patience, accuracy, and speed.

E. L. LEHMANN

Berkeley, Calif.
June, 1959

Contents

CHAPTER 1

The General Decision Problem

1. STATISTICAL INFERENCE AND STATISTICAL DECISIONS

The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables X whose distribution P_θ is at least partly unknown. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Ω, the *parameter space*. *Statistical inference* is concerned with methods of using this observational material to obtain information concerning the distribution of X or the parameter θ with which it is labeled. To arrive at a more precise formulation of the problem we shall consider the purpose of the inference.

The need for statistical analysis stems from the fact that the distribution of X, and hence some aspect of the situation underlying the mathematical model, is not known. The consequence of such a lack of knowledge is uncertainty as to the best mode of behavior. To formalize this, suppose that a choice has to be made between a number of alternative actions. The observations, by providing information about the distribution from which they came, also provide guidance as to the best decision. The problem is to determine a rule which, for each set of values of the observations, specifies what decision should be taken. Mathematically such a rule is a function δ, which to each possible value x of the random variables assigns a decision $d = \delta(x)$, that is, a function whose domain is the set of values of X and whose range is the set of possible decisions.

In order to see how δ should be chosen, one must compare the consequences of using different rules. To this end suppose that the consequence of taking decision d when the distribution of X is P_θ is a *loss*, which can be expressed as a nonnegative real number $L(\theta, d)$. Then the long-term average loss that would result from the use of δ in a number of repetitions of the experiment is the expectation $E[L(\theta, \delta(X))]$ evaluated

under the assumption that P_θ is the true distribution of X. This expectation, which depends on the decision rule δ and the distribution P_θ, is called the *risk function* of δ and will be denoted by $R(\theta, \delta)$. By basing the decision on the observations, the original problem of choosing a decision d with loss function $L(\theta, d)$ is thus replaced by that of choosing δ where the loss is now $R(\theta, \delta)$.*

The above discussion suggests that the aim of statistics is the selection of a decision function which minimizes the resulting risk. As will be seen later, this statement of aims is not sufficiently precise to be meaningful; its proper interpretation is in fact one of the basic problems of the theory.

2. SPECIFICATION OF A DECISION PROBLEM

The methods required for the solution of a specific statistical problem depend quite strongly on the three elements that define it: the class $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ to which the distribution of X is assumed to belong; the structure of the space D of possible decisions d; and the form of the loss function L. In order to obtain concrete results it is therefore necessary to make specific assumptions about these elements. On the other hand, if the theory is to be more than a collection of isolated results, the assumptions must be broad enough either to be of wide applicability or to define classes of problems for which a unified treatment is possible.

Consider first the specification of the class $\mathcal{P}$. Precise numerical assumptions concerning probabilities or probability distributions are usually not warranted. However, it is frequently possible to assume that certain events have equal probabilities and that certain others are statistically independent. Another type of assumption concerns the relative order of certain infinitesimal probabilities, for example the probability of occurrences in an interval of time or space as the length of the interval tends to zero. The following classes of distributions are derived on the basis of only such assumptions, and are therefore applicable in a great variety of situations.

The *binomial* distribution $b(p, n)$ with

$$(1) \quad P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \cdots, n; \quad 0 \leqq p \leqq 1.$$

This is the distribution of the total number of successes in n independent trials when the probability of success for each trial is p.

* Sometimes, aspects of a decision rule other than the expectation of its loss are also taken into account.

The *Poisson* distribution $P(\tau)$ with

$$(2) \qquad P(X = x) = \frac{\tau^x}{x!} e^{-\tau}, \qquad x = 0, 1, \cdots; \quad 0 < \tau.$$

This is the distribution of the number of events occurring in a fixed interval of time or space if the probability of more than one occurrence in a very short interval is of smaller order of magnitude than that of a single occurrence, and if the numbers of events in nonoverlapping intervals are statistically independent. Under these assumptions, the process generating the events is called a *Poisson process.**

The *normal* distribution $N(\xi, \sigma^2)$ with probability density

$$(3) \quad p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(x - \xi)^2\right], \quad -\infty < x, \xi < \infty; \; 0 < \sigma.$$

Under very general conditions, which are made precise by the central limit theorem, this is the approximate distribution of the sum of a large number of independent random variables when the relative contribution of each term to the sum is small.

We consider next the structure of the decision space D. The great variety of possibilities is indicated by the following examples.

Example 1. Let $X_1, \cdots, X_n$ be a *sample* from one of the distributions (1)–(3), that is, let the X's be distributed independently and identically according to one of these distributions. Let θ be p, τ, or the pair (ξ, σ) respectively, and let $\gamma = \gamma(\theta)$ be a real-valued function of θ.

(i) If one wishes to decide whether or not γ exceeds some specified value γ_0, the choice lies between the two decisions $d_0: \gamma > \gamma_0$ and $d_1: \gamma \leqq \gamma_0$. In specific applications these decisions might correspond to the acceptance or rejection of a lot of manufactured goods, of an experimental airplane as ready for flight testing, of a new treatment as an improvement over a standard one, etc. The loss function of course depends on the application to be made. Typically, the loss is 0 if the correct decision is chosen, while for an incorrect decision the losses $L(\gamma, d_0)$ and $L(\gamma, d_1)$ are increasing functions of $|\gamma - \gamma_0|$.

(ii) At the other end of the scale is the much more detailed problem of obtaining a numerical estimate of γ. Here a decision d of the statistician is a real number, the estimate of γ, and the losses might be $L(\gamma, d) = v(\gamma)w(|d - \gamma|)$ where w is a strictly increasing function of the error $|d - \gamma|$.

(iii) An intermediate case is the choice between the three alternatives $d_0: \gamma < \gamma_0$, $d_1: \gamma > \gamma_1$, $d_2: \gamma_0 \leqq \gamma \leqq \gamma_1$, for example accepting a new treatment, rejecting it, or recommending it for further study.

* Such processes are discussed in the books by Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, New York, John Wiley & Sons, 2nd ed., 1957, and by Doob, *Stochastic Processes*, New York, John Wiley & Sons, 1953.

The distinction illustrated by this example is the basis for one of the principal classifications of statistical methods. Two-decision problems such as (i) are usually formulated in terms of *testing a hypothesis* which is to be accepted or rejected (see Chapter 3). It is the theory of this class of problems with which we shall be mainly concerned. The other principal branch of statistics is the theory of *point estimation* dealing with such problems as (ii). The investigation of *multiple-decision procedures* illustrated by (iii) has only begun in recent years.

Example 2. Suppose that the data consist of samples X_{ij}, $j = 1, \cdots, n_i$, from normal populations $N(\xi_i, \sigma^2)$, $i = 1, \cdots, s$.

(i) Consider first the case $s = 2$ and the question of whether or not there is a material difference between the two populations. This has the same structure as problem (iii) of the previous example. Here the choice lies between the three decisions d_0: $|\xi_2 - \xi_1| \leqq \Delta$, d_1: $\xi_2 > \xi_1 + \Delta$, d_2: $\xi_2 < \xi_1 - \Delta$ where Δ is preassigned. An analogous problem, involving $k + 1$ possible decisions, occurs in the general case of k populations. In this case one must choose between the decision that the k distributions do not differ materially, d_0: $\max |\xi_j - \xi_i| \leqq \Delta$, and the decisions d_k: $\max |\xi_j - \xi_i| > \Delta$ and ξ_k is the largest of the means.

(ii) A related problem is that of ranking the distributions in increasing order of their mean ξ.

(iii) Alternatively, a standard ξ_0 may be given and the problem is to decide which, if any, of the population means exceed that standard.

Example 3. Consider two distributions—to be specific, two Poisson distributions $P(\tau_1)$, $P(\tau_2)$—and suppose that τ_1 is known to be less that τ_2 but that otherwise the τ's are unknown. Let $Z_1, \cdots, Z_n$ be independently distributed, each according to either $P(\tau_1)$ or $P(\tau_2)$. Then each Z is to be classified as to which of the two distributions it comes from. Here the loss might be the number of Z's that are incorrectly classified, multiplied by a suitable function of τ_1 and τ_2. An example of the complexity that such problems can attain and the conceptual as well as mathematical difficulties that they may involve is provided by the efforts of anthropologists to classify the human population into a number of homogeneous races by studying the frequencies of the various blood groups and of other genetic characters.

All the problems considered so far could be termed *action problems.* It was assumed in all of them that if θ were known a unique correct decision would be available, that is, given any θ there exists a unique d for which $L(\theta, d) = 0$. However, not all statistical problems are so clear-cut. Frequently it is a question of providing a convenient summary of the data or indicating what information is available concerning the unknown parameter or distribution. This information will be used for guidance in various considerations but will not provide the sole basis for any specific decisions. In such cases the emphasis is on the inference rather than on the decision aspect of the problem, although formally it can still be considered a decision problem if the inferential statement

itself is interpreted as the decision to be taken. An important class of such problems, estimation by interval,* is illustrated by the following example.

Example 4. Let $X = (X_1, \cdots, X_n)$ be a sample from $N(\xi, \sigma^2)$ and let a decision consist in selecting an interval $[\underline{L}, \bar{L}]$ and stating that it contains ξ. Suppose that decision procedures are restricted to intervals $[\underline{L}(X), \bar{L}(X)]$ whose expected length for all ξ and σ does not exceed $k\sigma$ where k is some preassigned constant. An appropriate loss function would be 0 if the decision is correct and would otherwise depend on the relative position of the interval to the true value of ξ. In this case there are many correct decisions corresponding to a given distribution $N(\xi, \sigma^2)$.

It remains to discuss the choice of loss function, and of the three elements defining the problem this is perhaps the most difficult to specify. Even in the simplest case, where all losses eventually reduce to financial ones, it can hardly be expected that one will be able to evaluate all the short- and long-term consequences of an action. Frequently it is possible to simplify the formulation by taking into account only certain aspects of the loss function. As an illustration consider Example 1(i) and let $L(\theta, d_0) = a$ for $\gamma(\theta) \leqq \gamma_0$ and $L(\theta, d_1) = b$ for $\gamma(\theta) > \gamma_0$. The risk function becomes

$$R(\theta, \delta) = \begin{cases} aP_\theta\{\delta(X) = d_0\} & \text{if} \quad \gamma \leqq \gamma_0 \\ bP_\theta\{\delta(X) = d_1\} & \text{if} \quad \gamma > \gamma_0, \end{cases} \tag{4}$$

and is seen to involve only the two probabilities of error with weights which can be adjusted according to the relative importance of these errors. Similarly, in Example 3 one may wish to restrict attention to the number of misclassifications.

Unfortunately, such a natural simplification is not always available, and in the absence of specific knowledge it becomes necessary to select the loss function in some conventional way, with mathematical simplicity usually an important consideration. In point estimation problems such as that considered in Example 1(ii), if one is interested in estimating a real-valued function $\gamma = \gamma(\theta)$ it is customary to take the square of the error, or somewhat more generally to put

$$L(\theta, d) = v(\theta)(d - \gamma)^2. \tag{5}$$

Besides being particularly simple mathematically, this can be considered as an approximation to the true loss function L provided that for each fixed θ, $L(\theta, d)$ is twice differentiable in d, that $L(\theta, \gamma(\theta)) = 0$ for all θ, and that the error is not large.

* For the more usual formulation in terms of confidence intervals, see Chapter 3, Section 5, and Chapter 5, Sections 4 and 5.

It is frequently found that, within one problem, quite different types of losses may occur, which are difficult to measure on a common scale. Consider once more Example 1(i) and suppose that γ_0 is the value of γ when a standard treatment is applied to a situation in medicine, agriculture, or industry. The problem is that of comparing some new process with unknown γ to the standard one. Turning down the new method when it is actually superior, or adopting it when it is not, clearly entails quite different consequences. In such cases it is sometimes convenient to treat the various components, say $L_1, L_2, \cdots, L_r$, separately. Suppose in particular that $r = 2$ and that L_1 represents the more serious possibility. One can then assign a bound to this risk component, that is, impose the condition

$$E_{L_1}(\theta, \delta(X)) \leqq \alpha, \tag{6}$$

and subject to this condition minimize the other component of the risk. Example 4 provides an illustration of this procedure. The length of the interval $[\underline{L}, \bar{L}]$ (measured in σ-units) is one component of the loss function, the other being the loss that results if the interval does not cover the true ξ.

3. RANDOMIZATION; CHOICE OF EXPERIMENT

The description of the general decision problem given so far is still too narrow in certain respects. It has been assumed that for each possible value of the random variables a definite decision must be chosen. Instead, it is convenient to permit the selection of one out of a number of decisions according to stated probabilities, or more generally the selection of a decision according to a probability distribution defined over the decision space; which distribution depends of course on what x is observed. One way to describe such a randomized procedure is in terms of a nonrandomized procedure depending on X and a random variable Y whose values lie in the decision space and whose conditional distribution given x is independent of θ.

Although it may run counter to one's intuition that such extra randomization should have any value, there is no harm in permitting this greater freedom of choice. If the intuitive misgivings are correct it will turn out that the optimum procedures always are of the simple nonrandomized kind. Actually, the introduction of randomized procedures leads to an important mathematical simplification by enlarging the class of risk functions so that it becomes convex. In addition, there are problems in which some features of the risk function such as its maximum can be improved by using a randomized procedure.

Another assumption that tacitly has been made so far is that a definite experiment has already been decided upon so that it is known what observations will be taken. However, the statistical considerations involved in designing an experiment are no less important than those concerning its analysis. One question in particular that must be decided before an investigation is undertaken is how many observations should be taken so that the risk resulting from wrong decisions will not be excessive. Frequently it turns out that the required sample size depends on the unknown distribution and therefore cannot be determined in advance as a fixed number. Instead it is then specified as a function of the observations and the decision whether or not to continue experimentation is made *sequentially* at each stage of the experiment on the basis of the observations taken up to that point.

Example 5. On the basis of a sample $X_1, \cdots, X_n$ from a normal distribution $N(\xi, \sigma^2)$ one wishes to estimate ξ. Here the risk function of an estimate, for example its expected squared error, depends on σ. For large σ the sample contains only little information in the sense that two distributions $N(\xi_1, \sigma^2)$ and $N(\xi_2, \sigma^2)$ with fixed difference $\xi_2 - \xi_1$ become indistinguishable as $\sigma \to \infty$, with the result that the risk tends to infinity. Conversely, the risk approaches zero as $\sigma \to 0$ since then effectively the mean becomes known. Thus the number of observations needed to control the risk at a given level is unknown. However, as soon as some observations have been taken, it is possible to estimate σ^2 and hence to determine the additional number of observations required.

Example 6. In a sequence of trials with constant probability p of success, one wishes to decide whether $p \leqq 1/2$ or $p > 1/2$. It will usually be possible to reach a decision at an early stage if p is close to 0 or 1 so that practically all observations are of one kind, while a larger sample will be needed for intermediate values of p. This difference may be partially balanced by the fact that for intermediate values a loss resulting from a wrong decision is presumably less serious than for the more extreme values.

Example 7. The possibility of determining the sample size sequentially is important not only because the distributions P_θ can be more or less informative but also because the same is true of the observations themselves. Consider, for example, observations from the uniform distribution* over the interval $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ and the problem of estimating θ. Here there is no difference in the amount of information provided by the different distributions P_θ. However, a sample $X_1, X_2, \cdots, X_n$ can practically pinpoint θ if $\max |X_j - X_i|$ is sufficiently close to 1, or it can give essentially no more information than a single observation if $\max |X_j - X_i|$ is close to 0. Again the required sample size should be determined sequentially.

Except in the simplest situations, the determination of the appropriate sample size is only one aspect of the design problem. In general, one

* This distribution is defined in Problem 1 at the end of the chapter.

must decide not only how many but also what kind of observations to take. Formally all these questions can be subsumed under the general decision problem described at the beginning of the section, by interpreting X as the set of all available variables, by introducing the decisions of whether or not to stop experimentation at the various stages, by specifying in case of continuance which type of variable to observe next, and by including the cost of observation in the loss function. However, in spite of this formal possibility, the determination of optimum designs in specific situations is typically of a higher order of difficulty than finding the optimum decision rule for a given experiment, and it has been carried out in only a few cases. Here, we shall be concerned primarily with the problem as it presents itself once the experiment has been set up, and only in a few special cases attempt a comparison of different designs.

4. OPTIMUM PROCEDURES

At the end of Section 1 the aim of statistical theory was stated to be the determination of a decision function δ which minimizes the risk function

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]. \tag{7}$$

Unfortunately, in general the minimizing δ depends on θ, which is unknown. Consider, for example, some particular decision d_0, and the decision procedure $\delta(x) \equiv d_0$ according to which decision d_0 is taken regardless of the outcome of the experiment. Suppose that d_0 is the correct decision for some θ_0 so that $L(\theta_0, d_0) = 0$. Then δ minimizes the risk at θ_0 since $R(\theta_0, \delta) = 0$, but presumably at the cost of a high risk for other values of θ.

In the absence of a decision function that minimizes the risk for all θ, the mathematical problem is still not defined since it is not clear what is meant by a best procedure. Although it does not seem possible to give a definition of optimality which will be appropriate in all situations, the following two methods of approach frequently are satisfactory.

The nonexistence of an optimum decision rule is a consequence of the possibility that a procedure devotes too much of its attention to a single parameter value at the cost of neglecting the various other values that might arise. This suggests the restriction to decision procedures which possess a certain degree of impartiality, and the possibility that within such a restricted class there may exist a procedure with uniformly smallest risk. Two conditions of this kind, invariance and unbiasedness, will be discussed in the next section.

Instead of restricting the class of procedures, one can approach the