VI OUTLINE OF A GENERAL THEORY OF STATISTICAL INFERENCE

The theories of Fisher, Neyman and Pearson are restricted in two respects. First, they consider only the problem of testing a hypothesis and that of estimation by point or interval. The second restriction is that only the case in which Ω is a k-parameter family of distribution functions is investigated. Both restrictions are serious from the point of view of applications.

There are many important statistical problems which are neither problems of testing a hypothesis, nor problems of estimation. We have already given such an example in Section 1. As a further illustration, let us consider the following case: Let $X_1,\ldots,X_p$ be p independently and normally distributed random variables with unit variances and unknown means $\theta_1,\ldots\theta_p$. Furthermore, let $x_{11},\ldots,x_{in}$ be n independent observations on $X_i(i = 1,2,\ldots,p)$. Suppose we test the hypothesis that $\theta_1 = \ldots = \theta_p = 0$, and decide to reject this hypothesis on the basis of the pn observations $x_{i\alpha}(\alpha = 1,2,\ldots,n;\ i = 1,2,\ldots,p)$. In such cases we are usually interested in knowing which mean values are not zero, i.e., we wish to subdivide the set of p mean values $\theta_1,\ldots,\theta_p$ into two subsets, such that one of them contains the mean values which are zero and the other the mean values which are not zero. This subdivision has to be done, of course, on the basis of the pn observations $x_{i\alpha}$. More precisely, we have to deal with the following statistical problem: There exist 2^p different subsets of the set $(\theta_1,\ldots,\theta_p)$. Denote these subsets by $\omega_1,\ldots,\omega_{2^p}$, respectively. Let H_k

($k = 1,\ldots,2^p$) be the hypothesis that the mean values contained in the set ω_k are equal to zero and all other mean values are unequal to zero. On the basis of the pn observations we have to decide which hypothesis H_k from the set of the 2^p possible hypotheses should be accepted. This problem cannot be considered as a problem of testing a hypothesis nor a problem of estimation.

A similar problem arises if we wish to classify a set of regression coefficients into the class of non-zero and the class of zero regression coefficients. In problems of regression we often take it for granted that the regression in question is a polynomial and we have to determine on the basis of the observations the degree of the polynomial to be fitted. That is to say, we have to decide on the basis of the observations which hypothesis of the sequence of hypotheses $H_1, H_2, H_3, \ldots, H_n, \ldots$ should be accepted. The symbol H_n ($n = 1,2,\ldots$) denotes the hypothesis that the regression is a polynomial of n-th degree. These examples illustrate sufficiently the necessity of the extension of the theory of statistical inference to the general case as formulated in Section 1.

The case in which Ω cannot be represented as a k-parameter family of distribution functions is quite important. As an illustration, consider the following problem: Let $(x_1,y_1),\ldots (x_n,y_n)$ be n independent pairs of observations on a pair (X,Y) of random variables. Suppose we wish to test the hypothesis that X and Y are independently distributed and we do not have any a priori knowledge about the joint distribution of X and Y. In this case Ω consists of all distribution functions

$F(x_1,y_1,\ldots,x_n,y_n)$ which can be written in the form

$$F(x_1,y_1,\ldots) = \overline{\Phi}(x_1,y_1) \ldots \overline{\Phi}(x_n,y_n)$$

where $\overline{\Phi}$ may be an arbitrary function. The subclass ω consists of all distribution functions $F(x_1,y_1,\ldots,x_n,y_n)$ which can be written in the form

$$F(x_1,y_1,\ldots,x_n,y_n) = \varphi(x_1)\psi(y_1)\varphi(x_2)\psi(y_2)\ldots\varphi(x_n)\psi(y_n).$$

Hence, Ω cannot be represented as a k-parameter family of functions.

The problem given above as an illustration has been treated by H. Hotelling and Margaret Pabst (see reference 8). Another problem, where Ω is the class of all continuous distributions, has been considered in paper (see reference 21). We shall give here an outline of a theory of statistical inference dealing with the following general problem[11]:

Let $X_1,\ldots,X_n$ be a set of n random variables. It is known that the joint probability distribution function $F(x_1,\ldots,x_n)$ of $X_1,\ldots,X_n$ is an element of a certain class Ω of distribution functions. Let S be a system of subclasses of Ω. For each element ω of S denote by H_ω the hypothesis that the true distribution $F(x_1,\ldots,x_n)$ of $X_1,\ldots,X_n$ is an element of ω. Denote by H_S the system of all hypotheses corresponding to all elements of S. Let x_i be the observed value of X_i (i=1,...,n). We have to decide by means of the observed sample point $E_n = (x_1,\ldots,x_n)$ which hypothesis of the system H_S of hypotheses should be accepted. That is to say, for each hypothesis H_ω we have to determine a region of acceptance M_ω in the n-dimensional sample space. The hypothesis H_ω will be accepted

---

11) This theory has been developed in reference 16 for the case that Ω is a k-parameter family

if and only if the sample point falls in the region M_ω. The regions M_ω and $M_{\omega'}$ are, of course, disjoint for $\omega \neq \omega'$. Furthermore, $\sum_\omega M_\omega$ is equal to the whole sample space. The statistical problem is that of the proper choice of the system M_S of the regions of acceptance.

The choice of the system M_S of regions of acceptance is equivalent to the choice of a function $\omega(E_n)$ defined over all points E_n of the sample space. The value of the function $\omega(E_n)$ is an element of S determined as follows: Since the elements of M_S are disjoint and since $\sum_\omega M_\omega$ is equal to the whole sample space, for each point E_n there exists exactly one element ω of S such that E_n is contained in M_ω. The value of the function $\omega(E_n)$ is that element ω of S for which E_n is an element of M_ω. Hence, we can replace M_S by the function $\omega(E_n)$ and for each sample point E_n we decide to accept the hypothesis $H_\omega(E_n)$. We will call $\omega(E_n)$ the statistical decision function. Hence, the statistical problem is that of choosing the statistical decision function $\omega(E_n)$.

The choice of $\omega(E_n)$ will essentially be affected by the relative importance of the different possible errors we may commit. We commit an error whenever we accept a hypothesis H_ω and the true distribution is not an element of ω. We introduce a weight function for the possible errors. The weight function $w[F,\omega]$ is a real valued non-negative function defined for all elements F of Ω and all elements ω of S, expressing the relative importance of the error committed by accepting H_ω when F is true. If F is an element of ω then $w[F,\omega] = 0$, otherwise $w[F,\omega] > 0$. The question as to how the form of the weight function $w[F,\omega]$ should be chosen is not a mathematical nor statistical

one. The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors and this will depend on the special purposes of his investigation. If this is done, we shall in general be able to give a more satisfactory answer to the question as to how the statistical decision function should be chosen. In many cases, especially in statistical questions concerning industrial production, we are able to express the importance of an error in monetary terms, that is, we can express the loss caused by the error considered in terms of money. We shall also say that $w[F,\omega]$ is the loss caused by accepting H_ω when F is true.

Suppose that we make our decisions according to a statistical decision function $\omega(E_n)$, and that the true distribution is the element $F(x_1,\ldots,x_n)$ of Ω. Then the expected value of the loss is obviously given by the Stieltjes integral

$$(5) \qquad \int_{M_n} w[F,\omega(E_n)]\,dF(x_1,\ldots,x_n) = r[F],$$

where the integration is to be taken over the whole sample space M_n. We shall call the expression (5) the risk of accepting a false hypothesis when F is the true distribution function. Since we do not know the true distribution F we shall have to study the risk $r[F]$ as a function of F. We shall call this function the <u>risk function</u>. Hence, the risk function is defined over all elements F of Ω. The form of the risk function depends on the statistical decision function $\omega(E_n)$ and on the weight function $w[F,\omega]$. In order to express this fact, we shall denote the risk function associated with the statistical decision function $\omega(E_n)$ and the weight function $w[F,\omega]$ also by

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\}$$

We introduce the following definitions:

Definition 1. Denote by $\omega(E_n)$ and $\omega'(E_n)$ two statistical decision functions for the same system H_s of hypotheses. We shall say that $\omega(E_n)$ and $\omega'(E_n)$ are equivalent relative to the weight $w[F,\omega]$ if the risk function $r\left\{F|\omega(E_n),\ w[F,\omega]\right\}$ is identically equal to the risk function $r\left\{F|\omega'(E_n), w[F,\omega]\right\}$ i.e., for any element F of Ω we have

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\} = r\left\{F|\omega'(E_n),\ w[F,\omega]\right\}.$$

Definition 2. Denote by $\omega(E_n)$ and $\omega'(E_n)$ two statistical decision functions for the same system H_s of hypotheses. We shall say that $\omega(E_n)$ is uniformly better than $\omega'(E_n)$ relative to the weight function $w[F,\omega]$ if $\omega(E_n)$ and $\omega'(E_n)$ are not equivalent and for each element F of Ω we have

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\} \leq r\left\{F|\omega'(E_n),\ w[F,\omega]\right\}.$$

Definition 3. A statistical decision function $\omega(E_n)$ is said to be admissible relative to the weight function $w[F,\omega]$ if no uniformly better statistical decision function exists relative to the weight function considered.

First principle for the choice of the statistical decision function. We choose a statistical decision function which is admissible relative to the weight function considered.

There can scarcely be given any argument against the acceptance of the above principle for the selection of $\omega(E_n)$. However, this principle does not lead in general to a unique solution. There exist in general many admissible statistical decision functions. We need a second principle for the choice of a best admissible decision function.

The choice between two admissible decision functions $\omega(E_n)$ and $\omega'(E_n)$ may be affected by the degree of our a priori confidence in the truth of the different elements of Ω. Suppose, for instance, that for a certain element F_1 of Ω we have

$$r\left\{F_1|\omega(E_n), w[F,\omega]\right\} < r\left\{F_2|\omega'(E_n), w[F,\omega]\right\}$$

for another element F_2 of Ω we have

$$r\left\{F_2|\omega(E_n), w[F,\omega]\right\} > r\left\{F_2|\omega'(E_n), w[F,\omega]\right\}$$

and for any other element $F \neq F_1, \neq F_2$ we have

$$r\left\{F|\omega(E_n), w[F,\omega]\right\} = r\left\{F|\omega'(E_n), w[F,\omega]\right\}.$$

If we have much greater a priori confidence in the truth of F_1 than in that of F_2, we will probably prefer $\omega(E_n)$ to $\omega'(E_n)$. On the other hand, if we think a priori that F_2 is more likely to be true than F_1, we may prefer $\omega'(E_n)$ to $\omega(E_n)$.

Suppose we can express our a priori degree of confidence by a non-negative additive set function $\rho(\eta)$ defined over a certain system of subsets η of Ω, where $\rho(\Omega) = 1$. That is to say the value of $\rho(\eta)$ expresses the degree of our a priori belief that the true distribution is an element of the subset η. In such a case it seems very reasonable to consider a decision function $\omega^*(E_n)$ as "best" if the value of the integral

$$\int_{\Omega} r\left\{F|\omega(E_n), w[F,\omega]\right\} d\rho$$

becomes a minimum for $\omega(E_n) = \omega^*(E_n)$. That is, we consider a decision function $\omega^*(E_n)$ as "best" if it minimizes a certain weighted average of the risk function.

However, it is doubtful that a set function expressing our a priori degree of belief can meaningfully be constructed. Therefore, we prefer to formulate the notion of a "best" decision function independently of such considerations.

Denote by $r\left\{\omega(E_n),\ w[F,\omega]\right\}$ the least upper bound of $r\left\{F|\omega(E_n),\ w[F,\omega]\right\}$ with respect to F, where F may be any element of Ω.

Definition 4. A decision function $\omega^*(E_n)$ is said to be a "best" decision function if $r\left\{\omega(E_n),\ w[F,\omega]\right\}$ becomes a minimum for $\omega(E_n) = \omega^*(E_n)$. (The weight function $w[F,\omega]$ is considered fixed.)

This definition of a "best" decision function seems to be a very reasonable one, although it is not the only possible one. One could reasonably define a decision function as "best" if it minimizes a certain weighted average of the risk function. However, there are certain properties of the "best" decision function according to definition 4, which seem to justify the use of that definition. One of the most important properties of a "best" decision function in the sense of definition 4 is that the risk function is a constant, i.e., it has the same value for all elements F of Ω. This has been shown in the case that Ω is a k-parameter family of distributions, and the weight function $w[F,\omega]$ and the distribution functions F satisfy certain restrictive conditions. The constancy of the risk function seems to be very desirable from the point of view of applications since this property makes it possible to evaluate the exact magnitude of the risk associated with the statistical decision. In the theory of confidence intervals the confidence coefficient, α, i.e., the probability that the confidence interval will cover the unknown parameter, is independent of the value of the unknown parameter. This fact, which is considered to be of basic importance in the theory of interval-estimation,

is analogous to the constancy of the risk function in our general theory since 1-α can be considered in a certain sense as the risk associated with the interval estimation. (The quantity 1-α is exactly equal to the risk in the sense of our definition, if the weight function takes only the values 0 and 1.)

Finally, I should like to make some remarks about the relationship of the general theory as outlined here, to the particular theory of uniformly most powerful and asymptotically most powerful tests which were discussed before. In the case of testing the simple hypothesis that the unknown distribution $F(x_1,\ldots,x_n)$ is equal to a particular distribution $F_0(x_1,\ldots x_n)$, the system S of subsets of Ω consists only of two elements ω_1 and ω_2 where ω_1 contains the single element F_0 and ω_2 is the complement of ω_1 in Ω. Hence, the decision function $\omega(E_n)$ can assume merely the values ω_1 and ω_2. Let M_{ω_1} be the subset of the sample space consisting of the points E_n for which $\omega(E_n)=\omega_1$ and let M_{ω_2} be the set of points E_n for which $\omega(E_n)=\omega_2$. The set M_{ω_2} is the complement of M_{ω_1} in the sample space. Obviously the set M_{ω_2} is the critical region, in the sense of the Neyman-Pearson theory. It is easy to see that if for any $\alpha(0<\alpha<1)$ a uniformly best critical region of size α for testing $F = F_0$ exists, then for any arbitrary weight function and for any admissible (see definition 3) decision function $\omega(E_n)$, the set M_{ω_2} will be a uniformly best critical region. In particular, the set M_{ω_2} corresponding to the "best" decision function (see definition 4) will be a uniformly best critical region. Hence, the form of the weight function affects merely the size of the region M_{ω_2} associated with the "best" decision function $\omega(E_n)$,

but it will always be a uniformly best critical region in the sense of the Neyman-Pearson theory. Similar considerations hold concerning asymptotically most powerful tests. Let the sequence $\{W_n\}$ (n=1,2,...,ad inf.) of critical regions be an asymptotically most powerful test for testing the simple hypothesis $F = F_0$. Then for sufficiently large n the region W_n is practically a uniformly best critical region and, therefore, it will be an excellent approximation to the region which is "best" in the sense of definition 4 irrespective of the shape of the weight function of errors.

As we have seen, for building up a general theory of statistical inference, the following three steps have to be made:

1. Formulation of the general problem of statistical inference.
2. Definition of the "best" procedure for making statistical decisions, i.e., definition of the "best" statistical decision function.
3. Solution of the mathematical problem of calculating the "best" statistical decision function.

The problem of statistical inference, as we have formulated it here, seems to be sufficiently broad to cover the problems in practical applications. The second step will always be, to a certain extent, arbitrary. The definition of "best" decision function given here seems to be a satisfactory one. Moreover, under certain restrictive conditions it has the important property that the risk function associated with the "best" decision function is constant, i.e., it has the same value for all elements of Ω. However, there may be other definitions of a

"best" decision function worth investigating. Decision functions which minimize a certain average of the risk function may be of special interest. Concerning step 3, there are many mathematical problems as yet unsolved.