ANNALS OF MATHEMATICS
Vol. 46, No. 2, April, 1945

STATISTICAL DECISION FUNCTIONS WHICH MINIMIZE THE MAXIMUM RISK

BY ABRAHAM WALD

(Received November 7, 1944)

1. Introduction

In some previous publications (see [1] and the last chapter in [2]) the author outlined a theory of statistical inference which deals with the following general problem: Let $X = (X_1, \cdots, X_n)$ be a set of random variables and suppose that the joint cumulative distribution function $F(t_1, \cdots, t_n)$ of the random variables $X_1, \cdots, X_n$ is not known. However it is known that $F(t_1, \cdots, t_n)$ is an element of a given class Ω of distribution functions. Consider a system S of subsets of Ω and for each element ω of S let H_ω denote the hypothesis that the joint distribution function of $X_1, \cdots, X_n$ is an element of ω. Furthermore, denote by H_S the system of all hypotheses H_ω corresponding to all elements ω of S. Let $E = (x_1, \cdots, x_n)$ denote an observation on X, i.e., x_i denotes an observed value of X_i $(i = 1, 2, \cdots, n)$. The totality of all possible observations E on X is the n-dimensional Cartesian space and is called the sample space. Any point of the sample space is called a sample point. The problem of statistical inference is to decide on the basis of the observed sample point E which hypothesis H_ω of the system H_S should be accepted. In other words, the problem of statistical inference can be formulated as follows: Given a class Ω of distribution functions and given a system S of subsets of Ω, the problem is to construct a function $\omega(E)$, called *statistical decision function*, which associates with each sample point E an element $\omega(E)$ of S so that the hypothesis $H_{\omega(E)}$ is accepted when the sample point E is observed. Thus, a statistical decision function is defined over all points of the sample space and for any sample point E the value of the function is an element of S.

The problem of statistical inference, as formulated above, is very general. It contains the problems of testing hypotheses and of statistical estimation treated in the literature. For example, if the problem is to test a hypothesis H_ω where ω is a given subset of Ω, then the system S consists only of two elements ω and $\bar{\omega}$ where $\bar{\omega}$ is the complement of ω in Ω. If we want to estimate the unknown distribution function of X by a single element of Ω, then S is the system of all points of Ω.

For simplicity we shall assume that Ω is a k-parameter family of distribution functions. Then each element of Ω can be represented by a point $\theta = (\theta_1, \cdots, \theta_k)$, called parameter point, in the k-dimensional Cartesian space. The class Ω of distribution functions is represented by a subset of the k-dimensional Cartesian space, called parameter space. In what follows we shall refer to Ω as the parameter space and the elements of Ω will be the parameter points.

To formulate principles for the proper choice of the statistical decision func-

tion the notion of a weight function for the possible erroneous decisions was introduced in [1]. The weight function $W(\theta, \omega)$ is a real valued non-negative function defined for all points θ of Ω and for all elements ω of S. For any pair (θ, ω) the value of the weight function $W(\theta, \omega)$ expresses the relative importance of the error committed by accepting H_ω when θ is true. We shall also say that $W(\theta, \omega)$ is the loss caused by the acceptance of H_ω when θ is true. If θ is contained in ω, $W(\theta, \omega)$ is, of course, equal to zero. The determination of the weight function $W(\theta, \omega)$ is not a statistical question and is considered here as given. The choice of the statistical decision function will, of course, be affected by the weight function $W(\theta, \omega)$. If the decision function $\omega(E)$ is used and if θ is the true parameter point, the expected value of the loss is given by the Stieltjes integral

$$\int_M W[\theta, \omega(E)]\, dF(E) \tag{1.1}$$

where the integration is to be taken over the whole sample space M and $F(E) = F(x_1, \cdots, x_n)$ is the joint cumulative distribution function of $X_1, \cdots, X_n$ corresponding to the parameter point θ.

The expression (1.1) is also called the risk. Clearly, the risk depends on the true parameter point and since this is unknown, it will be necessary to study the risk as a function of the parameter point θ. We shall denote the expression (1.1) by $r(\theta)$ and we shall call it risk function. Thus, the risk function is a real valued non-negative function defined for all points θ of the parameter space Ω. For any point θ the value of $r(\theta)$ is the expected value of the loss we would suffer if θ were the true parameter point. The shape of the risk function $r(\theta)$ depends, as can be seen from (1.1), on the decision function $\omega(E)$ used for making decisions. To put this in evidence, we shall occasionally use the notation $r[\theta \mid \omega(E)]$ instead of $r(\theta)$ indicating that we refer to the risk function which corresponds to the decision function $\omega(E)$. We shall also refer to $r[\theta \mid \omega(E)]$ as the risk function generated by the decision function $\omega(E)$. The goodness of a statistical decision function $\omega(E)$ is judged entirely on the basis of the risk function $r(\theta)$ generated by $\omega(E)$. A statistical decision function $\omega(E)$ is said to be *uniformly better* than the statistical decision function $\omega^*(E)$ if the risk functions $r(\theta)$ and $r^*(\theta)$ generated by $\omega(E)$ and $\omega^*(E)$, respectively, are not identically equal and if $r(\theta) \leqq r^*(\theta)$ for all points θ in Ω. A statistical decision function is said to be *admissible* if no uniformly better decision function exists.

In general, there will be infinitely many admissible decision functions. As to the question which of them should be chosen, the following considerations may be advanced: One possibility would be to choose an admissible decision function for which some weighted average of the risk function becomes a minimum. In other words, we choose a decision function $\omega(E)$ for which the Stieltjes integral

$$\int_\Omega r[\theta \mid \omega(E)]\, df(\theta) \tag{1.2}$$

becomes a minimum where $f(\theta)$ is some cumulative distribution function of θ. A decision function $\omega(E)$ for which (1.2) is a minimum, will be said to minimize the average risk relative to the distribution function $f(\theta)$. The difficulty with this approach is that the decision function $\omega(E)$ for which (1.2) is a minimum will, in general, depend on the distribution function $f(\theta)$ and one can hardly justify any particular choice of $f(\theta)$. If there would exist an a priori probability distribution $g(\theta)$ of the parameter θ and if this distribution were known, one could put $f(\theta)$ equal to $g(\theta)$. However, in most of the applications not even the existence of such an a priori probability distribution of θ can be postulated, and in those few cases where the existence of an a priori distribution of θ may be assumed this distribution is usually unknown. Under such circumstances it seems of interest to consider a decision function which minimizes the maximum (instead of some weighted average) of the risk function. We shall say that a decision function $\omega_0(E)$ minimizes the maximum of the risk function, or more briefly, that it minimizes the maximum risk, if the maximum of $r[\theta \mid \omega(E)]$ with respect to θ becomes a minimum for $\omega(E) = \omega_0(E)$. We shall also say that a decision function $\omega(E)$ is an "optimum" decision function if $\omega(E)$ is admissible and minimizes the maximum risk.

The above definition of an optimum decision function was already given in [1] where several results concerning optimum decision functions were obtained. The results derived in [1] were based on six assumptions made concerning Ω, the distribution functions represented by the points of Ω and the weight function $W(\theta, \omega)$. In this paper only the first three assumptions and a weakened form of the fourth assumption will be retained and several theorems concerning optimum decision functions will be derived which go considerably beyond the results obtained in [1]. The main results obtained in this paper can be briefly stated as follows: (1) there exists an optimum decision function; (2) if $\omega(E)$ and $\omega^*(E)$ are optimum decision functions, then the risk function $r(\theta)$ generated by $\omega(E)$ is identically equal to the risk function $r^*(\theta)$ generated by $\omega^*(E)$; (3) a decision function which minimizes the average risk relative to a particular distribution function $f(\theta)$, defined in Section 3 and designated there as the least favorable distribution of θ, is an optimum decision function; (4) the risk function generated by an optimum decision function is constant over a subset of Ω containing all points θ for which the probability measure of any open subset of Ω containing θ, calculated on the basis of a least favorable distribution of θ, is positive. Frequently the least favorable distribution of θ will be such that any open subset of Ω has a positive probability measure. In such cases the risk function $r(\theta)$ corresponding to an optimum decision function will be constant over the whole parameter space Ω.

An important and unsolved problem is to find a general method which would permit the actual calculation of an optimum statistical decision function. The fact that a decision function which minimizes the average risk relative to a least favorable distribution of θ is an optimum decision function may be very helpful in finding an optimum decision function, since the calculation of a least favorable distribution of θ seems to be a considerably simpler problem.

2. Reduction of the general problem of statistical inference to the problem of point estimation

It was pointed out in [1] (p. 305) that the general problem of statistical inference may be reduced to the problem of point estimation, i.e., to the case where S is the system of all points of Ω, provided that the power of the system S is not greater than the power of the set Ω. This can be seen as follows: Consider the problem of finding an optimum decision function $\omega(E)$ for a given system H_S of hypotheses and a given weight function $W(\theta, \omega)$. Assume that the power of S is not greater than that of Ω. Then there exists a single valued function $\gamma(\theta)$ defined for all points θ of Ω such that the following condition holds: for each θ the value of $\gamma(\theta)$ is an element of S and for each element ω of S there exists at least one point θ in Ω such that $\gamma(\theta) = \omega$. Let S^* be the system of all points of Ω and let $W^*(\theta, \bar{\theta})$ be a weight function defined for all pairs $(\theta, \bar{\theta})$ of parameter points as follows:

$$W^*(\theta, \bar{\theta}) = W[\theta, \gamma(\bar{\theta})]. \tag{2.1}$$

Any decision function for the system H_{S^*} of hypotheses is a single valued function $\bar{\theta}(E)$ defined for all sample points E such that for each E the value $\bar{\theta}(E)$ is an element of Ω. Such a function $\bar{\theta}(E)$ is also called a point estimate, or more briefly, an estimate of θ. Thus, any decision function for the system H_{S^*} of hypotheses is an estimate of θ. If the system S is equal to S^* we shall use the word "estimate" synonymously with "statistical decision function." Let $\bar{\theta}(E)$ be an optimum decision function for the problem where the system of hypotheses is given by H_{S^*} and the weight function of errors is given by $W^*(\theta, \bar{\theta})$. Then the decision function

$$\omega(E) = \gamma[\bar{\theta}(E)] \tag{2.2}$$

can easily be seen to be an optimum decision function for the original problem given by the system H_S of hypotheses and the weight function $W(\theta, \omega)$. Thus, we see that to any problem of statistical inference defined by a system S of subsets of Ω and a weight function $W(\theta, \omega)$, there corresponds another problem defined by the system S^* of all points of Ω and by the weight function $W^*(\theta, \bar{\theta})$ given in (2.1) such that an optimum decision function $\omega(E)$ for the first problem can be obtained from an optimum decision function $\bar{\theta}(E)$ for the second problem by the relationship (2.2). Hence, it is sufficient to consider only problems of estimation, i.e., problems where S is the system of all points of Ω. In what follows we shall assume that S is the system of all points of Ω and we shall use the phrase "estimate of θ," or more briefly, "estimate" to mean a statistical decision function.

3. Assumptions concerning Ω, the distribution functions represented by the points of Ω and the weight function

Throughout this paper the following assumptions will be made:

ASSUMPTION 1. *The parameter space Ω is a bounded and closed subset of a finite dimensional, say k-dimensional, Cartesian space.*

ASSUMPTION 2. *The weight function $W(\theta, \bar{\theta})$ defined for all pairs $(\theta, \bar{\theta})$ of parameter points is continuous in θ and $\bar{\theta}$ jointly.*

ASSUMPTION 3. *For any point θ of Ω the joint distribution function of $X_1, \cdots, X_n$ represented by θ admits a density function $p(E, \theta)$ for all points E of the n-dimensional Cartesian space M(sample space). The density function $p(E, \theta)$ is assumed to be continuous in E and θ jointly.*

In what follows we shall mean by a distribution function $f(\theta)$ of θ a cumulative distribution function defined for all points of the k-dimensional Cartesian space for which

$$\int_\Omega df(\theta) = 1.$$

For any distribution function $f(\theta)$ of θ and for any sample point E we shall denote by $\omega_f(E)$ the set of all parameter points $\bar{\theta}$ for which

$$\int_\Omega W(\theta, \bar{\theta}) p(E, \theta)\, df(\theta) \tag{3.1}$$

takes its minimum value with respect to $\bar{\theta}$. An element of $\omega_f(E)$ will be denoted by $\theta_f(E)$. If several elements of $\omega_f(E)$ are considered, they will be distinguished by superscripts, such as $\theta_f'(E)$, $\theta_f''(E)$, etc. We shall say that two parameter points θ' and θ'' are interchangeable if $W(\theta, \theta') = W(\theta, \theta'')$ identically in θ. Obviously, if $\theta(E)$ and $\theta^*(E)$ are two estimates such that for any sample point E the parameter points $\theta(E)$ and $\theta^*(E)$ are interchangeable, then the risk function generated by $\theta(E)$ is identically equal to the risk function generated by $\theta^*(E)$.

ASSUMPTION 4. *For any distribution function $f(\theta)$ of θ and for any sample point E the set $\omega_f(E)$ consists only of parameter points which are interchangeable with each other.*

It is clear that if a parameter point θ' is interchangeable with a point of $\omega_f(E)$, then θ' is a point of $\omega_f(E)$. Assumption 4 is weaker than Assumption 4 in [1] which states that $\omega_f(E)$ contains at most one parameter point. If for the original problem the system S of subsets is not the system of all points of Ω then in the reduced problem, as defined by the system S^* of all points of Ω and the weight function given in (2.1), there will be, in general, points of Ω which are interchangeable and, therefore, the set $\omega_f(E)$ will frequently contain more than one point, while the present weaker assumption 4 still may be satisfied. Thus there seems to be a considerable advantage in formulating Assumption 4 in its present weaker form.

It should be mentioned that Assumption 4 is not as restrictive as it would appear. For example, if Ω is a one dimensional closed interval, the set $\omega_f(E)$ can easily be shown to contain exactly one point if the following conditions are fulfilled: (1) $p(E, \theta) > 0$ for all points E and θ. (2) $W(\theta, \bar{\theta})$ is continuous in θ and is a polynomial of the second degree in $\bar{\theta}$. (3) The coefficient of $\bar{\theta}^2$ in the expression $W(\theta, \bar{\theta})$ is positive for every θ.

The assumptions 1–4 are sufficient but by no means necessary for the validity

of the results obtained in Sections 4 and 5. Although these assumptions could be weakened in various ways, the author did not endeavor to do so for the sake of simplicity.

4. Estimates of θ which minimize the average risk relative to a given distribution $f(\theta)$ of θ. Least favorable distribution of θ

In this section we shall define the notion of a least favorable distribution of θ and we shall prove the existence and several properties of such a distribution. These properties will then be used in Section 5 to establish several theorems about estimates of θ which minimize the maximum risk. First we shall derive several lemmas.

LEMMA 4.1. *For any distribution function $f(\theta)$ of θ and for any sample point E the set $\omega_f(E)$ is not empty.*

PROOF: From Assumption 1 it follows that Ω is compact. From this and Assumption 2 it follows that $W(\theta, \bar{\theta})$ is uniformly continuous. According to Assumption 3 $p(E, \theta)$ is continuous. Hence for any sample point E, $p(E, \theta)$ is a bounded function of θ. From these facts it follows easily that the expression (3.1) is a continuous function of $\bar{\theta}$ for any fixed sample point E. Hence, $\omega_f(E)$ is not empty and Lemma 4.1 is proved.

LEMMA 4.2. *Let $\{\varphi_n(\theta)\}$ $(n = 1, 2, \cdots, ad\ inf.)$ be a sequence of functions of θ such that $\lim \varphi_n(\theta) = \varphi(\theta)$ uniformly in θ where $\varphi(\theta)$ is a continuous function of θ. Furthermore, let $\{f_n(\theta)\}$ be a sequence of distribution functions of θ which converges to a distribution function $f(\theta)$ for all continuity points of $f(\theta)$. Then*

$$\lim_{n=\infty} \left\{ \int_\Omega \varphi_n(\theta)\, df_n(\theta) - \int_\Omega \varphi_n(\theta)\, df(\theta) \right\} = 0.$$

PROOF: Since $\lim_{n=\infty} \varphi_n(\theta) = \varphi(\theta)$ uniformly in θ, and since Ω is compact, we have

$$\lim_{n=\infty} \left\{ \int_\Omega \varphi_n(\theta)\, df_n(\theta) - \int_\Omega \varphi(\theta)\, df_n(\theta) \right\} = 0 \tag{4.1}$$

$$\lim_{n=\infty} \left\{ \int_\Omega \varphi_n(\theta)\, df(\theta) - \int_\Omega \varphi(\theta)\, df(\theta) \right\} = 0 \tag{4.2}$$

and

$$\lim_{n=\infty} \left\{ \int_\Omega \varphi(\theta)\, df_n(\theta) - \int_\Omega \varphi(\theta)\, df(\theta) \right\} = 0. \tag{4.3}$$

Lemma 4.2 is an immediate consequence of (4.1), (4.2) and (4.3).

For any point θ of Ω and for any subset ω of Ω we shall denote by $\delta(\theta, \omega)$ the greatest lower bound of the Euclidean distance $\delta(\theta, \theta')$ of the points θ and θ' where θ' may be any point of ω. For any pair (ω_1, ω_2) of subsets of Ω we shall denote by $\delta(\omega_1, \omega_2)$ the least upper bound of $\delta(\theta, \omega_2)$ with respect to θ where θ is restricted to points of ω_1.

LEMMA 4.3. *Let $f(\theta)$ be a distribution function of θ and let $\{f_n(\theta)\}$ $(n = 1, 2, \cdots, ad\ inf.)$ be a sequence of distribution functions of θ such that $\lim_{n=\infty} f_n(\theta) =$*

$f(\theta)$ at all continuity points of $f(\theta)$. Furthermore, let $\{E_n\}$ $(n = 1, 2, \cdots, ad\ inf.)$ be a sequence of sample points such that $\lim E_n = E$. *Then*

$$\lim_{n=\infty} \delta[\omega_{f_n}(E_n), \omega_f(E)] = 0.$$

PROOF: We shall assume that Lemma 4.3 is not true and we shall derive a contradiction. If Lemma 4.3 does not hold, there exists a positive δ, a subsequence $\{n'\}$ $(n = 1, 2, \cdots,$ ad inf.) of the sequence of positive integers and for each n' a point $\theta_{f_{n'}}(E_{n'})$ of $\omega_{f_{n'}}(E_{n'})$ such that

$$\lim_{n=\infty} \delta[\theta_{f_{n'}}(E_{n'}), \omega_f(E)] = \delta. \tag{4.4}$$

Since the expression (3.1) is a continuous function of $\bar{\theta}$, and since Ω is compact, it follows from (4.4) that

$$\liminf_{n=\infty} \left\{ \int_\Omega W[\theta, \theta_{f_{n'}}(E_{n'})] p(E, \theta)\, df(\theta) - \int_\Omega W[\theta, \theta_f(E)] p(E, \theta)\, df(\theta) \right\} = \delta^* > 0 \tag{4.5}$$

where $\theta_f(E)$ is an element of $\omega_f(E)$. From the compactness of Ω and Assumption 3 it follows that

$$\lim_{n=\infty} p(E_{n'}, \theta) = p(E, \theta) \tag{4.6}$$

uniformly in θ. Since $W(\theta, \bar{\theta})$ is bounded, we obtain from (4.5) and (4.6)

$$\liminf_{n=\infty} \left\{ \int_\Omega W[\theta, \theta_{f_{n'}}(E_{n'})] p(E_{n'}, \theta)\, df(\theta) - \int_\Omega W[\theta, \theta_f(E)] p(E_{n'}, \theta)\, df(\theta) \right\} = \delta^* > 0. \tag{4.7}$$

Let $\{n''\}$ be a subsequence of the sequence $\{n'\}$ such that the sequence $\{\theta_{f_{n''}}(E_{n''})\}$ $(n = 1, 2, \cdots,$ ad inf.) of parameter points converges and denote the limit point by θ^*. Then it follows from (4.6), the continuity of $W(\theta, \bar{\theta})$ and the compactness of Ω that

$$\lim_{n=\infty} W[\theta, \theta_{f_{n''}}(E_{n''})] p(E_{n''}, \theta) = W(\theta, \theta^*) p(E, \theta) \tag{4.8}$$

and

$$\lim_{n=\infty} W[\theta, \theta_f(E)] p(E_{n''}, \theta) = W[\theta, \theta_f(E)] p(E, \theta) \tag{4.9}$$

uniformly in θ. Since the right hand side expressions in (4.8) and (4.9) are continuous functions of θ, it follows from (4.8), (4.9) and Lemma 4.2 that

$$\lim_{n=\infty} \left\{ \int_\Omega W[\theta, \theta_{f_{n''}}(E_{n''})] p(E_{n''}, \theta)\, df_{n''}(\theta) - \int_\Omega W[\theta, \theta_{f_{n''}}(E_{n''})] p(E_{n''}, \theta)\, df(\theta) \right\} = 0 \tag{4.10}$$

and

$$\lim_{n=\infty}\left\{\int_\Omega W[\theta, \theta_f(E)]p(E_{n''}, \theta)\, df_{n''}(\theta) - \int_\Omega W[\theta, \theta_f(E)]p(E_{n''}, \theta)\, df(\theta)\right\} = 0. \tag{4.11}$$

From (4.7), (4.10) and (4.11) we obtain

$$\liminf_{n=\infty}\left\{\int_\Omega W[\theta, \theta_{f_{n''}}(E_{n''})]p(E_{n''}, \theta)\, df_{n''}(\theta) - \int_\Omega W[\theta, \theta_f(E)]p(E_{n''}, \theta)\, df_{n''}(\theta)\right\} \geqq \delta^* > 0. \tag{4.12}$$

But this is a contradiction, since $\theta_{f_{n''}}(E_{n''})$ is an element of $\omega_{f_{n''}}(E_{n''})$ and, therefore, the expression in the braces cannot be positive. Hence, Lemma 4.3 is proved.

LEMMA 4.4. *For each positive ϵ a bounded and closed subset M_ϵ of the sample space (n-dimensional Cartesian space) M can be given such that*

$$\int_{M_\epsilon} p(E, \theta)\, dE \geqq 1 - \epsilon$$

for all points θ of the parameter space Ω.

This lemma is the same as Lemma 4 in [1], p. 309. It was proved there using only Assumptions 1 and 3.

LEMMA 4.5. *For any positive η a positive δ can be given such that for any estimate $\theta(E)$ and for any pair (θ, θ') of parameter points whose Euclidian distance is less than δ the inequality*

$$|r(\theta) - r(\theta')| = \left|\int_M W[\theta, \theta(E)]p(E, \theta)\, dE - \int_M W[\theta', \theta(E)]p(E, \theta')\, dE\right| < \eta$$

holds.

This lemma is the same as Lemma 5 in [1], p. 310. It was proved there using only Assumptions 1, 2 and 3.

With the help of Lemmas 1–5 we shall be able to establish several theorems.

THEOREM 4.1. *For any estimate $\theta(E)$ the risk function $r(\theta)$ generated by $\theta(E)$ is a uniformly continuous function of θ.*

This theorem is an immediate consequence of Lemma 4.5.

THEOREM 4.2. *For any distribution function $f(\theta)$ of θ there exists an estimate $\theta(E)$ which minimizes the average risk relative to the distribution $f(\theta)$. If both estimates $\theta^*(E)$ and $\theta^{**}(E)$ minimize the average risk relative to the distribution $f(\theta)$, then the risk function $r^*(\theta)$ generated by $\theta^*(E)$ is identically equal to the risk function $r^{**}(\theta)$ generated by $\theta^{**}(E)$.*

PROOF: Since according to Lemma 4.1 the set $\omega_f(E)$ is not empty, we may put $\theta(E)$ equal to an element $\theta_f(E)$ of $\omega_f(E)$ for each sample point E. The esti-

mate $\theta_f(E)$ minimizes the average risk relative to the distribution $f(\theta)$ since the average risk is given by

$$\int_\Omega r[\theta \mid \theta_f(E)]\, df(\theta) = \int_\Omega \int_M W[\theta, \theta_f(E)]p(E, \theta)\, dE\, df(\theta) = \int_M \int_\Omega W[\theta, \theta_f(E)]p(E, \theta)\, df(\theta)\, dE \tag{4.13}$$

where M denotes the sample space. Hence, the first half of Theorem 4.2 is proved. Suppose that $\theta(E)$ is an estimate which minimizes the average risk relative to $f(\theta)$ and denote by R the set of all sample points E for which $\theta(E)$ is not an element of $\omega_f(E)$. The second half of Theorem 4.2 is proved if we show that the Lebesgue measure of the set R is equal to zero. Suppose that R has a positive Lebesgue measure. Since

$$\int_\Omega W[\theta, \theta(E)]p(E, \theta)\, df(\theta) > \int_\Omega W[\theta, \theta_f(E)]p(E, \theta)\, df(\theta) \tag{4.14}$$

for all points E in R, there exists a subset R' of R and a positive δ such that R' has a positive Lebesgue measure and

$$\int_\Omega W[\theta, \theta(E)]p(E, \theta)\, df(\theta) \geqq \int_\Omega W[\theta, \theta_f(E)]p(E, \theta)\, df(\theta) + \delta \tag{4.15}$$

for all points E in R'. From (4.15) it follows that

$$\int_M \int_\Omega W[\theta, \theta(E)]p(E, \theta)\, df(\theta)\, dE > \int_M \int_\Omega W[\theta, \theta_f(E)]p(E, \theta)\, df(\theta)\, dE$$

which is impossible, since $\theta(E)$ is an estimate which minimizes the average risk. Hence R must be of measure zero and Theorem 4.2 is proved.

In what follows for any distribution $f(\theta)$ of θ we shall denote by $r_f(\theta)$ the risk function generated by an estimate which minimizes the average risk relative to $f(\theta)$.

THEOREM 4.3. *If the sequence* $\{f_n(\theta)\}$ $(n = 1, 2, \cdots, ad\ inf.)$ *of distribution functions of* θ *converges to a distribution function* $f(\theta)$ *at any continuity point of* $f(\theta)$, *then*

$$\lim_{n=\infty} r_{f_n}(\theta) = r_f(\theta)$$

uniformly in θ.

PROOF: For each sample point E let $\theta_{f_n}(E)$ be an element of $\omega_{f_n}(E)$ and $\theta_f(E)$ an element of $\omega_f(E)$. First we show that

$$\lim_{n=\infty} W[\theta, \theta_{f_n}(E)] = W[\theta, \theta_f(E)] \tag{4.16}$$

uniformly in E over any bounded subset M' of the sample space M. Suppose that this is not true. Then there exists a sequence $\{E_n\}$ of points of M' and a

subsequence $\{n'\}$ $(n = 1, \cdots, \text{ad inf.})$ of the sequence of positive integers such that $\{E_{n'}\}$ converges to a point E_0 of M and

$$\lim_{n=\infty} \{W[\theta, \theta_{f_{n'}}(E_{n'})] - W[\theta, \theta_f(E_{n'})]\} = \rho \neq 0. \tag{4.17}$$

From Lemma 4.3 it follows that

$$\lim_{n=\infty} \delta[\omega_{f_{n'}}(E_{n'}), \omega_f(E_0)] = 0 \tag{4.18}$$

and

$$\lim_{n=\infty} \delta[\omega_f(E_{n'}), \omega_f(E_0)] = 0. \tag{4.19}$$

From (4.18) and (4.19) we obtain

$$\lim_{n=\infty} W[\theta, \theta_{f_{n'}}(E_{n'})] = W[\theta, \theta_f(E_0)] \tag{4.20}$$

and

$$\lim_{n=\infty} W[\theta, \theta_f(E_{n'})] = W[\theta, \theta_f(E_0)]. \tag{4.21}$$

But this is in contradiction to (4.17). Hence, the convergence in (4.16) is shown to be uniform in E over any bounded subset of M. From this it follows that

$$\lim_{n=\infty} \int_{M'} W[\theta, \theta_{f_n}(E)]p(E, \theta)\, dE = \int_{M'} W[\theta, \theta_f(E)]p(E, \theta)\, dE \tag{4.22}$$

for any bounded subset M' of M.

According to Lemma 4.4 for any positive ϵ a bounded and closed subset M_ϵ of M can be given such that

$$\int_{M_\epsilon} p(E, \theta)\, dE \geqq 1 - \epsilon \tag{4.23}$$

for every θ. Since $W(\theta, \bar{\theta})$ is bounded and since ϵ can be chosen arbitrarily small, we obtain from (4.22) and (4.23).

$$\begin{aligned} \lim_{n=\infty} r_{f_n}(\theta) &= \lim_{n=\infty} \int_M W[\theta, \theta_{f_n}(E)]p(E, \theta)\, dE \\ &= \int_M W[\theta, \theta_f(E)]p(E, \theta)\, dE = r_f(\theta). \end{aligned} \tag{4.24}$$

The uniformity of the convergence follows easily from Lemma 4.5 and the compactness of Ω. Hence Theorem 4.3 is proved.

For any distribution function $f(\theta)$ of θ we shall denote $\int_\Omega r_f(\theta)\, df(\theta)$ by r_f. A distribution function $g(\theta)$ of θ will be said to be a least favorable distribution if $r_g \geqq r_f$ for all distribution functions $f(\theta)$.

If an a priori distribution $f(\theta)$ of θ exists and is known, the best we can do is to use an estimate $\theta_f(E)$ for which the expected value of the loss is equal to its minimum value r_f. Thus, whenever an *a priori* distribution of θ exists and is known, the expected value of the loss will be greatest if the *a priori* distribution of θ happens to be equal to a least favorable distribution $g(\theta)$.

THEOREM 4.4. *There exists a least favorable distribution.*

PROOF: Since $W(\theta, \bar{\theta})$ is bounded, r_f also has a finite upper bound. Let r be the least upper bound of r_f with respect to all distribution functions $f(\theta)$. Then there exists a sequence $\{g_n(\theta)\}$ $(n = 1, 2, \cdots, \text{ad inf.})$ of distribution functions such that

$$\lim_{n=\infty} r_{g_n} = r. \tag{4.25}$$

Let $\{n'\}$ $(n = 1, 2, \cdots, \text{ad inf.})$ be a subsequence of the sequence $\{n\}$ $(n = 1, 2, \cdots, \text{ad inf.})$ such that $\{g_{n'}(\theta)\}$ converges to a distribution function $g(\theta)$ at all continuity points of $g(\theta)$. Then according to Theorem 4.3

$$\lim_{n=\infty} r_{g_{n'}}(\theta) = r_g(\theta) \quad \text{uniformly in } \theta. \tag{4.26}$$

Hence

$$\lim_{n=\infty} \left\{ \int_\Omega r_{g_{n'}}(\theta)\, dg(\theta) - \int_\Omega r_g(\theta)\, dg(\theta) \right\} = 0. \tag{4.27}$$

From (4.26) and Lemma 4.2 we obtain

$$\lim_{n=\infty} \left\{ \int_\Omega r_{g_{n'}}(\theta)\, dg_{n'}(\theta) - \int_\Omega r_{g_{n'}}(\theta)\, dg(\theta) \right\} = 0. \tag{4.28}$$

From (4.27) and (4.28) it follows that

$$r = \lim_{n=\infty} r_{g_{n'}} = r_g.$$

Hence $g(\theta)$ is a least favorable distribution and Theorem 4.4 is proved.

THEOREM 4.5. *If $g(\theta)$ is a least favorable distribution of θ, then $r_g(\theta) \leqq r_g$ for all points θ in Ω.*

PROOF: Suppose that Theorem 4.5 is not true. Then there exists a point θ_0 such that $r_g(\theta_0) > r_g$. Hence there also exists a distribution function $h(\theta)$ of θ such that

$$\int_\Omega r_g(\theta)\, dh(\theta) > r_g. \tag{4.29}$$

Let $k(\theta)$ be a distribution function defined by

$$k(\theta) = \frac{g(\theta) + \lambda h(\theta)}{1 + \lambda} \qquad (\lambda \geqq 0) \tag{4.30}$$

where λ is some non-negative value. Then

$$\lim_{\lambda=0} k(\theta) = g(\theta) \tag{4.31}$$

and therefore on account of Theorem 4.3

$$\lim_{\lambda=0} r_k(\theta) = r_g(\theta) \tag{4.32}$$

uniformly in θ. From (4.29) and (4.32) it follows that for sufficiently small values of λ

$$\int_\Omega r_k(\theta)\, dh(\theta) > r_g\,. \tag{4.33}$$

Clearly,

$$\int_\Omega r_k(\theta)\, dg(\theta) \geqq r_g\,. \tag{4.34}$$

From (4.30), (4.33) and (4.34) we obtain

$$\int_\Omega r_k(\theta)\, dk(\theta) > r_g \tag{4.35}$$

for sufficiently small values of λ. But this is in contradiction to our assumption that $g(\theta)$ is a least favorable distribution. Hence, Theorem 4.5 is proved.

For any distribution function $f(\theta)$ of θ we shall denote by Ω_f the set of all points θ for which the following condition is satisfied: for any open subset ω_θ of Ω which contains θ the inequality

$$\int_{\omega_\theta} df(\theta) > 0$$

holds. We shall denote by $\bar{\Omega}_f$ the complement of Ω_f in Ω.

THEOREM 4.6. *If $g(\theta)$ is a least favorable distribution of θ, then $r_g(\theta) = r_g$ for all points θ in Ω_g.*

This theorem is an immediate consequence of Theorem 4.5 and the continuity of the risk function $r_g(\theta)$.

THEOREM 4.7. *If both $g(\theta)$ and $h(\theta)$ are least favorable distributions of θ, then $r_g(\theta)$ is identically equal to $r_h(\theta)$.*

PROOF: Since $g(\theta)$ and $h(\theta)$ are least favorable distributions, we must have

$$r_g = r_h = r \text{ (say)}. \tag{4.36}$$

From this and Theorem 4.5 it follows that

$$\int_\Omega r_g(\theta)\, dh(\theta) = \int_\Omega r_h(\theta)\, dh(\theta) = r. \tag{4.37}$$

Let $\theta_g(E)$ be an element of $\omega_g(E)$ and $\theta_h(E)$ an element of $\omega_h(E)$. Then the estimate $\theta_g(E)$ generates the risk function $r_g(\theta)$ and the estimate $\theta_h(E)$ generates the risk function $r_h(\theta)$. Equation (4.37) shows that both estimates $\theta_g(p)$ and $\theta_h(E)$ minimize the average risk relative to $h(\theta)$. But then according to the second half of Theorem 4.2 $r_g(\theta)$ is identically equal to $r_h(\theta)$. This proves Theorem 4.7.

THEOREM 4.8. *If $g(\theta)$ is a distribution function of θ such that the maximum of $r_g(\theta)$ is equal to r_g, then $g(\theta)$ is a least favorable distribution of θ.*

PROOF: Let $h(\theta)$ be any distribution function of θ. Since the maximum of $r_g(\theta)$ is equal to r_g, we have

$$\int_\Omega r_g(\theta)\, dh(\theta) \leqq r_g = \int_\Omega r_g(\theta)\, dg(\theta). \tag{4.38}$$

Clearly,

$$\int_\Omega r_h(\theta)\, dh(\theta) \leqq \int_\Omega r_g(\theta)\, dh(\theta). \tag{4.39}$$

Hence

$$\int_\Omega r_h(\theta)\, dh(\theta) \leqq \int_\Omega r_g(\theta)\, dg(\theta)$$

for any distribution function $h(\theta)$. This proves Theorem 4.8.

5. Estimates which minimize the maximum risk

On the basis of the results obtained in Section 4 we shall be able to derive several theorems concerning estimates which minimize the maximum risk.

THEOREM 5.1. *If $\theta(E)$ is an estimate which minimizes the average risk relative to a least favorable distribution $g(\theta)$ of θ, then $\theta(E)$ also minimizes the maximum risk.*

PROOF: According to Theorem 4.5 the risk function $r_g(\theta)$ generated by $\theta(E)$ satisfies the inequality $r_g(\theta) \leqq r_g = \int_\Omega r_g(\theta) dg(\theta)$. Hence the maximum of $r_g(\theta)$ with respect to θ is equal to r_g. Suppose that there exists an estimate $\theta^*(E)$ which generates a risk function $r^*(\theta)$ the maximum of which is less than r_g. Then

$$\int_\Omega r^*(\theta)\, dg(\theta) < r_g$$

in contradiction to the assumption that $\theta(E)$ minimizes the average risk relative to the distribution $g(\theta)$. Hence Theorem 5.1 is proved.

THEOREM 5.2. *There exists an estimate $\theta(E)$ which minimizes the maximum risk.*

This theorem is an immediate consequence of Theorems 4.2, 4.4 and 5.1.

Now we shall prove the converse of Theorem 5.1.

THEOREM 5.3. *Let $\theta(E)$ be an estimate which minimizes the maximum risk and let $g(\theta)$ be a least favorable distribution of θ. Then $\theta(E)$ minimizes the average risk relative to $g(\theta)$.*

PROOF: Let $r(\theta)$ be the risk function generated by $\theta(E)$. From Theorem 5.1 it follows that the maximum of $r(\theta)$ is equal to the maximum of $r_g(\theta)$ which in turn is equal to

$$\int_\Omega r_g(\theta)\, dg(\theta) = r_g .$$

Hence

$$\int_{\Omega} r(\theta)\, dg(\theta) \leqq r_g \,. \tag{5.1}$$

Since $r_g(\theta)$ is generated by an estimate which minimizes the average risk relative to $g(\theta)$, the equality sign must hold in (5.1). Thus, we have

$$\int_{\Omega} r(\theta)\, dg(\theta) = r_g \,. \tag{5.2}$$

Hence, $\theta(E)$ minimizes the average risk relative to $g(\theta)$. This proves Theorem 5.3.

THEOREM 5.4. *If both estimates $\theta^*(E)$ and $\theta^{**}(E)$ minimize the maximum risk, the risk function $r^*(\theta)$ generated by $\theta^*(E)$ is identically equal to the risk function $r^{**}(\theta)$ generated by $\theta^{**}(E)$.*

PROOF: Let $g(\theta)$ be a least favorable distribution of θ and let $r_g(\theta)$ be the risk function generated by an estimate which minimizes the average risk relative to $g(\theta)$. According to the second half of Theorem 4.2 $r_g(\theta)$ is uniquely determined. Then it follows from Theorem 5.3 that $r^*(\theta) = r_g(\theta)$ and $r^{**}(\theta) = r_g(\theta)$ for all points θ in Ω. Hence Theorem 5.4 is proved.

THEOREM 5.5. *If $\theta(E)$ is an estimate which minimizes the maximum risk, then $\theta(E)$ is admissible.*

PROOF: Suppose that $\theta(E)$ minimizes the maximum risk but is not admissible. Then there exists another estimate $\theta^*(E)$ such that the risk function $r^*(\theta)$ generated by $\theta^*(E)$ is less than or equal to the risk function $r(\theta)$ generated by $\theta(E)$ for all parameter points θ and $r^*(\theta) < r(\theta)$ for at least one point θ. Since $\theta(E)$ minimizes the maximum risk, also $\theta^*(E)$ minimizes the maximum risk. But then, according to Theorem 5.4, $r^*(\theta)$ must be identically equal to $r(\theta)$. Thus, we arrive at a contradiction and Theorem 5.5 is proved.

For any distribution $f(\theta)$ the symbol Ω_f was used in Section 4 to denote the set of all parameter points θ with the following property: For any open subset ω_θ of Ω which contains θ the inequality $\int_{\omega_\theta} df(\theta) > 0$ holds. The complement of Ω_f in Ω is denoted by $\bar{\Omega}_f$.

THEOREM 5.6. *If $g(\theta)$ is a least favorable distribution of θ and if $\theta(E)$ is an estimate which minimizes the maximum risk, then the risk function $r(\theta)$ generated by $\theta(E)$ has a constant value over the set Ω_g.*

This theorem is a simple consequence of the Theorems 4.6, 5.1 and 5.4.

From Theorem 5.6 we obtain the following corollary.

COROLLARY 5.1. *If a least favorable distribution $g(\theta)$ exists with the property that for any open subset ω of Ω $\int_{\omega} dg(\theta) > 0$, then the risk function $r(\theta)$ generated by an estimate which minimizes the maximum risk will be constant over the whole parameter space Ω.*

6. Relationship to von Neumann's theory of games

The theory of statistical decision functions which minimize the maximum risk is very closely related to a theory of games developed by John von Neumann [3], [4]. In fact, the problem of statistical inference as formulated here can be interpreted as a zero sum two person game in v. Neumann's theory. The normalized form of a zero sum two person game is defined as follows (see section 14.1 in [4]): There are two players and there is a function $K(\tau_1, \tau_2)$ of two variables τ_1 and τ_2 given where τ_1 and τ_2 can take only a finite number of values. Player 1 chooses a value of τ_1 and player 2 chooses a value of τ_2, each choice being made in complete ignorance of the other, and then the players 1 and 2 get the amounts $K(\tau_1, \tau_2)$ and $-K(\tau_1, \tau_2)$, respectively. Clearly, player 1 wishes to maximize $K(\tau_1, \tau_2)$ and player 2 wishes to minimize $K(\tau_1, \tau_2)$.

A problem of statistical inference may be interpreted as a zero sum two person game as follows: Player 1 is Nature and player 2 is the statistician. The variable τ_1 is the parameter point θ the value of which is chosen by Nature. The variable τ_2 is the statistical decision function $\omega(E)$ which is chosen by the statistician. The outcome $K[\theta, \omega(E)]$ of the game is the risk $r[\theta \mid \omega(E)]$ of the statistician. Clearly, the statistician wishes to minimize $r[\theta \mid \omega(E)]$. Of course, we cannot say that Nature wants to maximize $r[\theta \mid \omega(E)]$. However, if the statistician is in complete ignorance as to Nature's choice, it is perhaps not unreasonable to base the theory of a proper choice of $\omega(E)$ on the assumption that Nature wants to maximize $r[\theta \mid \omega(E)]$. Under this assumption a problem of statistical inference becomes identical with a zero sum two person game.[1]

The choice of τ_1 by player 1 and the choice of τ_2 by player 2 can be rationalized if the game is strictly determined (see section 14.5.1 in [4]) i.e., if

$$\text{(6.1)} \qquad \text{Max}_{\tau_1}\text{Min}_{\tau_2}K(\tau_1, \tau_2) = \text{Min}_{\tau_2}\text{Max}_{\tau_1}K(\tau_1, \tau_2).$$

If (6.1) is fulfilled, a good way for 1 to play the game is to choose a value τ_1 for which $\text{Min}_{\tau_2}K(\tau_1, \tau_2)$ assumes its maximum value, and a good way for 2 to play the game is to choose a value τ_2 for which $\text{Max}_{\tau_1}K(\tau_1, \tau_2)$ assumes its minimum value.[2]

To overcome the difficulty caused by the fact that for many games (6.1) does not hold, the problem is reformulated as follows (see section 17 in [4]): Instead of choosing a particular value of τ_i, player i considers all possible values of τ_i and chooses only the probabilities with which he is going to use them, respectively. In other words, if the possible values of τ_i are $1, 2, \cdots, \beta_i$, player i does not choose any particular value in this set, but chooses a set of probabilities $\rho_1, \rho_2, \cdots, \rho_{\beta_i}$ and the value of τ_i is then determined by a chance mechanism

[1] The only difference is that in the theory of games τ_1 and τ_2 take only a finite number of values, while in the theory of statistical inference the corresponding variables θ and $\omega(E)$ are not subject to such a restriction. They take a finite number of values only if both the parameter space and the sample space are finite.

[2] In the terminology of the present paper this means that a good way for the statistician to play is to minimize the maximum risk.

constructed in such a way that the probability that $\tau_i = j$ is equal to ρ_j. Thus, the choice of player 1 is now characterized by a vector $\xi = (\xi_1, \xi_2, \cdots, \xi_{\beta_1})$ and the choice of 2 by a vector $\eta = (\eta_1, \eta_2, \cdots, \eta_{\beta_2})$. The mathematical expectation of the outcome is then given by

$$K^*(\xi, \eta) = \sum_{\tau_1=1}^{\beta_1} \sum_{\tau_2=1}^{\beta_2} K(\tau_1, \tau_2)\xi_{\tau_1}\eta_{\tau_2}. \tag{6.2}$$

The main theorem proved by v. Neumann (see section 17.6 in [4]) states that for any arbitrary function $K(\tau_1, \tau_2)$ the game corresponding to $K^*(\xi, \eta)$ is always strictly determined, i.e.,

$$\text{Max}_\xi \text{Min}_\eta K^*(\xi, \eta) = \text{Min}_\eta \text{Max}_\xi K^*(\xi, \eta). \tag{6.3}$$

If player i does not choose a particular value of τ_i but a probability distribution of the possible values of τ_i, we shall say that player i uses a mixed strategy. On the other hand, if player i chooses a particular value of τ_i, we shall say that player i has a pure strategy. Then v. Neumann's result can be stated as follows: if both players are permitted to use mixed strategies, the game is always strictly determined. This result of v. Neymann is closely related to some of the principal results of the present paper. In fact, Theorems 4.5 and 5.1 of the present paper clearly imply that the problem of statistical inference, viewed as a zero sum two person game, is strictly determined. However, the strict determinateness is derived in the present paper under conditions which differ in several respects from those in v. Neumann's theory. In v. Neumann's theory both players are permitted to use a mixed strategy, while in the present paper only Nature is permitted to use a mixed strategy (the statistician chooses a definite decision function $\omega(E)$ and not a probability distribution in the space of all possible decision functions). Furthermore, the variables τ_1 and τ_2 can take only a finite number of values in v. Neumann's theory, while in the present paper the corresponding variables θ and $\omega(E)$ are not subject to this restriction. Because of the omission of this restriction and because of the requirement that the strategy of the statistician be pure, we were not able to prove strict determinateness without postulating the validity of Assumptions 1–4. No such assumptions are needed in v. Neumann's theory.

COLUMBIA UNIVERSITY

REFERENCES

[1] A. WALD, *Contributions to the theory of statistical estimation and testing hypotheses*, Annals of Mathematical Statistics, Vol. 10, December, 1939.

[2] A. WALD, *On the principles of statistical inference*, Notre Dame Mathematical Lectures, Number 1, 1942, Notre Dame, Indiana.

[3] JOHN V. NEUMANN, *Zur theorie der Gesellschaftspiele*, Math. Annalen, vol. 100, (1928), pp. 295–320.

[4] JOHN V. NEUMANN AND OSKAR MORGANSTERN, *Theory of games and economic behavior*, Princeton, Princeton University Press, 1944.