ON THE MINIMAX APPROACH TO THE PROBLEM OF ESTIMATION.

By D. BASU, *Research Fellow, N.I.S.I., Statistical Laboratory, Calcutta.*

(Communicated by Prof. S. N. Bose, F.N.I.)

(*Received July 23; read October 5, 1951.*)

INTRODUCTION.

The purpose of this paper is to lay emphasis on a few aspects of Professor Wald's approach to the problems of Statistical Inference with particular reference to Estimation. It has been pointed out that a minimax estimator may not exist in many important cases unless we suitably modify our loss function or truncate the parameter space.

The problem of point estimation is as follows. The form of the distribution function $F(x|\theta)$ of a certain population being known we have to estimate the unknown population parameter θ by means of a random sample $\mathbf{x} = (x_1, x_2, \ldots . x_n)$ from the population. The problem is clearly a problem of decision functions. We have to define a single valued function $d(\mathbf{x}) = d(x_1, x_2, \ldots . x_n)$ defined over the entire n-dimensional sample space M such that $d(\mathbf{x})$ takes values in the parameter space Ω. When we get a sample $\mathbf{x} = (x_1, \ldots . x_n)$ we estimate the unknown θ by $d(\mathbf{x})$.

Let $W(\theta, d)$ be the loss or weight function. It stands for the loss that we suffer if we estimate the true θ by d. In other words $W(\theta, d)$ stands for the different weights that the statistician wants to give to the different possible wrong decisions. It will not be unrealistic to assume that $W(\theta, d)$ is a monotonic non-decreasing function of $|\theta - d|$. As a matter of fact we shall generally work with the following two simple types of weight functions namely

$$W(\theta, d) = (\theta - d)^2 \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (1.1)$$

$$W(\theta, d) = \begin{cases} 0 \text{ if } |\theta - d| \leqq l \\ 1 \text{ if } |\theta - d| > l \end{cases} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (1.2)$$

The risk function or the expected loss function $r(\theta/d)$ corresponding to a particular decision function $d(\mathbf{x})$ is defined as follows:

$$r(\theta/d) = \int_M W(\theta, d(\mathbf{x}))\, dF(\mathbf{x}/\theta) \quad \ldots \quad \ldots \quad \ldots \quad (1.3)$$

where $$F(\mathbf{x}|\theta) = F(x_1|\theta) \ldots . F(x_n|\theta)$$

and M is the n-dimensional sample space. It will be noted that $r(\theta/d)$ is a function of θ and the form of the decision function $d(\mathbf{x})$ but is independent of the sample point.

It is very natural to try to discover a decision function for which the associated risk function is small over the whole parameter space Ω. This immediately leads to the following definitions.

Definition 1.1: The decision function $d_0(\mathbf{x})$ is said to be uniformly more powerful than $d_1(\mathbf{x})$ if $r(\theta|d_0) \leqq r(\theta|d_1)$ for all θ with the sign of inequality holding for at least one θ.

It is almost obvious that there cannot exist a decision function $d_0(\mathbf{x})$ that is uniformly more powerful than all other decision functions. For if possible let $d_0(\mathbf{x})$ be uniformly the most powerful and let $d_1(\mathbf{x}) \equiv \theta_1$. Then it is clear that $r(\theta_1|d_1) = 0$ (assuming that $W(\theta, d) = 0$ when $\theta = d$) and since $r(\theta_1|d_0) \leqq r(\theta_1|d_1)$ it follows that $r(\theta|d_0) = 0$ when $\theta = \theta_1$. Since $d_0(\mathbf{x})$ is more powerful than any alternative $d_1(\mathbf{x})$ it follows that $r(\theta|d_0) \equiv 0$, and this can be true only if $W(\theta_1 d)$ is of a trivial nature. The above consideration leads us to the very important definition of admissible decision functions.

Definition 1.2:—The decision function $d_0(\mathbf{x})$ is said to be admissible if there exists no other decision function that is uniformly more powerful than $d_0(\mathbf{x})$.

Obviously no decision function that is inadmissible can be accepted. The main problem that arises in the Waldian approach to the classical problem of estimation is to show that certain estimators commonly in use are admissible decision functions. The following theorem first obtained by Rao (1945) is of great importance in the above connection. The theorem was independently obtained by Blackwell (1947) and was extended to convex loss functions by Hodges and Lehmann (1950) and by Barankin (1950).

Theorem 1.1:—If $t = t(\mathbf{x})$ be a sufficient statistic for θ then the class of admissible estimators of θ is a sub-class of all functions of t provided $W(\theta, d) = (\theta - d)$.[2]

The extension of Rao's theorem to the case where $W(\theta_1 d)$ for every θ is a convex (downwards) function of d is almost immediate. For from convexity of $W(\theta, d)$ it follows that

$$E\{W(\theta, d)|t\} \geqq W(\theta, \psi) \text{ where } \psi = \psi(t) = E(d|t)$$

and hence

$$\begin{aligned} r(\theta|d) &= E[W(\theta, d)] = E_t[E\{W(\theta, d)|t\}] \\ &\geqq E\, W(\theta, \psi) \\ &= r(\theta|\psi) \end{aligned}$$

Wald has shown that under certain conditions the class of all Bayes' solutions forms a complete class of decision functions and it will be seen that when a sufficient statistic exists every Bayes' solution is a function of the sufficient statistic. Wald's theorem however is proved under some restrictive conditions on the parameter space Ω and the weight function $W(\theta, d)$. We take up the problem in some greater details in the next section

2. Minimax Decision Rule.

If the population parameter θ be itself a random variable with $\mu(\theta)$ as the distribution function defined over the space Ω then it may be considered desirable to minimise the average risk function

$$\bar{r}(\mu|d) = \int_{\Omega} r(\theta|d)\, d\mu(\theta) \qquad \ldots \quad \ldots \quad \ldots \quad (2.1)$$

Definition 2.1:—A decision function that minimises the average risk function $\bar{r}(\mu/d)$ is called a Bayes' solution of the decision problem.

Definition 2.2:—A class of decision functions is said to be a complete class of decision functions if no decision function outside that class is admissible.

Definition 2.3:—A decision function $d_0(\mathbf{x})$ is said to be a minimax decision function if it is admissible and further if it minimises the maximum risk associated with any decision function. In other words d_0 is a minimax decision function if it is admissible and

$$\operatorname*{Sup}_{\theta} r(\theta|d_0) \leqq \operatorname*{Sup}_{\theta} r(\theta|d) \qquad \ldots \quad \ldots \quad \ldots \quad (2.2)$$

where d is any other decision function. By sup (supremum) of $r(\theta/d)$ we mean the least upper bound of $r(\theta/d)$ in Ω.

The criterion of admissibility and the further criterion of minimax are the two new criteria set up by Wald. The following theorems proved by Wald are of fundamental importance.

Theorem 2.1:—The class of Bayes' solutions is a complete class of decision functions, i.e. no decision function that does not minimise the average risk with respect to some a priori distribution $\mu(\theta)$ is admissible.

Theorem 2.2:—There exists a minimax decision function.

Theorem 2.3:—The minimax decision function is a Bayes' solution with respect to a least favourable a priori distribution and it generates a risk function that is constant over the parameter space Ω expecting possibly in a set of probability measure zero (with reference to the least favourable a priori distribution).

The decision problem considered by Wald covers almost all the problems of Statistical Inference and it is to be expected that such general theorems can be proved only under a set of restrictive assumptions (vide Wald (1950), Chap. 3). We shall presently study the real nature of some of these assumptions. It should be noted that a Bayes' solution is nothing but a decision function that minimises some average risk function. No a priori arguments are really brought into the picture and the Bayes' solutions are studied because of their important properties.

3. Non-admissibility of uniform weight function over an infinite range:— Wald (1939) attempted to solve the problem of finding the minimax estimate of a location parameter. He demonstrated that under certain condition there exists a minimax estimate of the location parameter and that the maximum likelihood estimate under certain further conditions is the minimax estimate. It is then deduced that $\bar{x}$ is the minimax estimate of the Normal mean when the variance is known. The proofs of Theorems 5 and 6 in the above paper are however not valid because the multiple integrals considered in (30), (31) and (32) of that paper are all divergent and as such the proofs through the change of the order of integration in (30) and (31) is inadmissible. The Bayes' solution with respect to the uniform a priori distribution over the infinite range $-\infty < \theta < \infty$ was found to generate a constant risk function and so it was deduced that the particular Bayes' solution is the minimax estimate. But the a priori density function $d\theta(-\infty < \theta < \infty)$ is not a true probability distribution and may very well lead to a decision function that is not admissible. The average risk function $\bar{r}(\mu/d)$, where $d\mu(\theta) = d\theta(-\infty < \theta < \infty)$, will be infinite and the question of minimising $\bar{r}(\mu/d)$ with respect to d does not arise at all. Take for instance the following example where the problem is to make an estimate of the Poisson mean θ the loss function being $(\theta-d)^2$ the space Ω being $0 < \theta < \infty$.

Let $$X = x_1+x_2+\ldots+x_n.$$

Since X is a sufficient statistic for θ it follows from Rao's theorem that we need restrict ourselves only to functions of X. Since X is a Poisson variable with mean $n\theta$ we have

$$p(X|\theta) = e^{-n\theta}\frac{(n\theta)^X}{X!} \quad (X = 0, 1, 2, \ldots.) \quad \ldots \quad \ldots \quad (3.1)$$

and

$$r(\theta|d) = \sum_0^\infty (\theta-d(X))^2 p(X/\theta) \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (3.2)$$

Now if $d\mu(\theta) = d\theta \quad (0<\theta<\infty)$ we have

$$\bar{r}(\mu|d) = \int_0^\infty d\theta \sum_0^\infty (\theta-d(X))^2 e^{-n\theta}\frac{(n\theta)^X}{X!} \quad \ldots \quad \ldots \quad (3.3)$$

Proceeding in the same way as in Theorem 5 of Wald (1939) we may write (by formally integrating term by term)

$$\bar{r}(\mu|d) = \sum_{0}^{\infty} \frac{1}{X!} \int_{0}^{\infty} (\theta - d)^2 \, e^{-n\theta} (n\theta)^X \, d\theta \,. \qquad .. \qquad .. \qquad (3.4)$$

Now it is easily seen that the integral

$$\int_{0}^{\infty} (\theta - d)^2 \, e^{-n\theta} (n\theta)^X \, d\theta \qquad .. \qquad .. \qquad .. \qquad (3.5)$$

is a minimum when

$$d = \frac{\int_{0}^{\infty} \theta \, e^{-n\theta} (n\theta)^X \, d\theta}{\int_{0}^{\infty} e^{-n\theta} (n\theta)^X \, d\theta} \qquad .. \qquad .. \qquad .. \qquad .. \qquad (3.6)$$

$$= \frac{1}{n} \cdot \frac{\Gamma(X+2)}{\Gamma(X+1)} = \frac{X}{n} + \frac{1}{n}.$$

The risk function generated by the above decision function is

$$r\left(\theta \,\middle|\, \frac{X}{n} + \frac{1}{n}\right) = E\left(\theta - \frac{X}{n} - \frac{1}{n}\right)^2 = \frac{\theta}{n} + \frac{1}{n^2} \qquad .. \qquad .. \qquad (3.7)$$

whereas the risk function generated by $\frac{X}{n}$ is

$$r\left(\theta \,\middle|\, \frac{X}{n}\right) = E\left(\theta - \frac{X}{n}\right)^2 = \frac{\theta}{n} \qquad .. \qquad .. \qquad .. \qquad .. \qquad (3.8)$$

Thus the decision function (3.6) is not admissible.

It is therefore clear that the type of argument employed in Theorem 5 of Wald (1939) is not valid. We should be very careful while dealing with a priori weight functions in space Ω that are not true probability distributions.

Suppose the problem is to find the minimax estimate of the normal mean the variance being known. If we take our loss function as $(\theta - d)^2$ then assumption 3.3 in Wald (1950) is violated as the loss (or weight) function is not bounded. This however is not a very serious difficulty. Even if the loss function is unbounded it can be shown that the minimax estimate will exist in many cases.

In estimation the real difficulty arises with assumption 3.4 where it is assumed that the space of terminal decisions D^t (since for the present we restrict ourselves to non-sequential decision functions only the space D^t is the whole decision space) is compact in the sense of a suitably defined metric in D^t space. Now if the parameter space Ω be $-\infty < \theta < \infty$ then the space of all possible decisions also is from $-\infty$ to ∞ and as such assumption 3.4 is violated. Of course we can make the decision space compact by taking the parameter space Ω as $a \leqq \theta \leqq b$ but in that case further difficulties arise. With this truncated decision space the theorems of Wald hold true but the resulting analysis becomes extremely difficult. For instance the decision function $\bar{x}$ ceases to be admissible, for clearly the decision function

$$d_0(\mathbf{x}) = \begin{cases} \bar{x} \text{ if } a \leqq \bar{x} \leqq b \\ a \text{ if } \bar{x} < a \\ b \text{ if } \bar{x} > b \end{cases} \qquad .. \qquad .. \qquad .. \qquad (3.9)$$

is uniformly more powerful than $\bar{x}$. Wald (1950) has shown that $d_0(\mathbf{x})$ also is not admissible. Although a minimax decision function exists in theory it is not known what it is or how to find it out. Thus for the sake of simplicity at least we have to take Ω as $-\infty < \theta < \infty$, in which case $\bar{x}$ is the minimax estimator.

4. Non-existence of a minimax solution when the decision space is not compact. Consider a rectangular population with range $(0, \theta)$ the problem being to find a minimax estimator of θ on the basis of a random sample $\mathbf{x} = (x_1, \ldots . x_n)$ the parameter space Ω being $0 < \theta < \infty$ and the loss function being of the type (1.2). The risk function is

$$r(\theta/d) = \int_0^\theta \ldots . \int_0^\theta W(\theta, d) \frac{1}{\theta^n} dx_1 \ldots dx_n.$$

Let $\xi = \max (x_1, \ldots . x_n)$. Then it is easily verified that the density function of ξ is $p(\xi|\theta)d\xi = \frac{1}{\theta^n} n\xi^{n-1} d\xi$ and hence it follows that the probability density of $(x_1, x_2, \ldots . x_n)$ on the surface S_ξ where ξ lies between ξ and $\xi + d\xi$ is simply

$$p(x_1, \ldots , x_n/\xi)\, dx_1 \ldots dx_n = \frac{1}{n\xi^{n-1} d\xi} dx_1 \ldots dx_n.$$

Thus

$$r(\theta/d) = \int_0^\theta \frac{n\xi^{n-1}}{\theta^n} d\xi \int_{S_\xi} W(\theta, d) \frac{dx_1 \ldots dx_n}{n\xi^{n-1} d\xi}. \qquad \ldots \quad \ldots \quad (4.1)$$

Consider the following a priori distribution for θ namely

$$g(\theta|\alpha)\, d\theta = \frac{1}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta}\, d\theta \; (0 < \theta < \infty, \alpha > 0).$$

Then

$$\bar{r}(g(\theta|\alpha)/d) = \bar{r}(\alpha/d) = \int_0^\infty r(\theta|d)\, g(\theta|\alpha)\, d\theta$$

$$= \text{the average risk function.}$$

By an easy change of the order of integration we have

$$\bar{r}(\alpha/d) = \int_0^\infty n\xi^{n-1} d\xi \int_{S_\xi} \frac{dx_1 \ldots dx_n}{n\xi^{n-1} d\xi} \int_\xi^\infty \frac{W(\theta, d)}{\theta^n} g(\theta/\alpha)\, d\theta.$$

Thus in order to minimise $\bar{r}(\alpha/d)$ all that we have to do is to choose $d = d(x_1, \ldots . x_n)$ in such a way that for every $(x_1, x_2, \ldots . x_n)$ the integral

$$\int_\xi^\infty \frac{W(\theta, d)}{\theta^n} g(\theta/\alpha)\, d\theta \qquad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (4.2)$$

$$= \int_\xi^\infty W(\theta, d) \frac{1}{\Gamma(\alpha)} \theta^{\alpha-n-1} e^{-\theta}\, d\theta$$

is a minimum.

Now since $W(\theta, d) = 0$ when $d-l \leqq \theta \leqq d+l$ and $W(\theta, d) = 1$ elsewhere we should clearly choose $d \geqq \xi+l$ and should choose d in such a way that

$$\int_{d-l}^{d+l} \frac{1}{\Gamma(\alpha)} \theta^{\alpha-n-1} e^{-\theta} d\theta \text{ is maximum.} \quad .. \quad .. \quad .. \quad (4.3)$$

Let $\alpha > n+1$ and let $\phi(y)$ be defined as

$$\phi(y) = \int_{y-l}^{y+l} \theta^{\alpha-n-1} e^{-\theta} d\theta \qquad (l \leqq y < \infty)$$

Then

$$\phi^1(y) = (y+l)^{\alpha-n-1} e^{-(y+l)} - (y-l)^{\alpha-n-1} e^{-(y-l)}$$

It is easily verified that $\phi^1(y)$ remains > 0 up to a stage and then changes sign and remains always < 0. That is $\phi(y)$ increases for some time and then goes on decreasing.

Let $y = f(\alpha)$ be the point where $\phi(y)$ is a maximum.

Thus in order to minimise (4.2) or to maximise (4.3) we should choose

$$d(\mathfrak{x}) = d\alpha(\xi) = \begin{cases} f(\alpha) & \text{so long as} \quad \xi \leqq f(\alpha)-l \\ \xi+l & \text{as soon as} \quad \xi > f(\alpha)-l \end{cases} \quad .. \quad .. \quad (4.4)$$

Thus corresponding to the a priori density function $g(\theta/\alpha)d\theta$ for θ the Bayes' solution is as defined in (4.4) where it is easily verifiable that $f(\alpha) = \alpha+0(1)$ (by $0(1)$ we mean a function of α that remains bounded as $\alpha \to \infty$).

Now consider the risk function $r(\theta/d\alpha)$ generated by the decision function $d\alpha(\xi)$. Clearly so long as $\theta < f(\alpha)-l$, $d\alpha(\xi) = f(\alpha)$ for every ξ and hence from the definition of our loss function we have $r(\theta/d\alpha) = 1$ for every $\theta < f(\alpha)-l$. (Since we shall ultimately make $\alpha \to \infty$ therefore we can assume α to be so large that $f(\alpha)-l > 0$.). When $f(\alpha)-l \leqq \theta \leqq f(\alpha)+l$ then either $\xi \leqq f(\alpha)-l$ or ξ lies between $f(\alpha)-l$ and θ and in both the cases $d\alpha(\xi)$ (which is either $f(\alpha)$ or $\xi+l$) must lie between $\theta-l$ and $\theta+l$ and as such $r(\theta/d\alpha) = 0$ in this case.

When $\theta > f(\alpha)+l$ then $W(\theta, d\alpha) = 0$ only when $\theta-2l \leqq \xi \leqq \theta$ and hence in this case

$$r(\theta|d\alpha) = \int_0^{\theta-2l} \frac{n\xi^{n-1}}{\theta^n} d\xi = \left(1-\frac{2l}{\theta}\right)^n.$$

Thus we have

$$r(\theta|d\alpha) = \begin{cases} 1 & \text{for} \quad 0 < \theta < f(\alpha)-l \\ 0 & \text{for} \quad f(\alpha)-l \leqq \theta \leqq f(\alpha)+l \\ \left(1-\frac{2l}{\theta}\right)^n & \text{for} \quad f(\alpha)+l < \theta < \infty. \end{cases} \quad .. \quad .. \quad (4.5)$$

Hence the average risk function

$$\bar{r}(\alpha|d_\alpha) = \int_0^\infty r(\theta/d_\alpha)\, g(\theta/\alpha) d\theta \quad .. \quad .. \quad .. \quad .. \quad (4.6)$$

$$= \int_0^{f(\alpha)-l} g(\theta|\alpha) d\theta + \int_{f(\alpha)+l}^\infty \left(1-\frac{2l}{\theta}\right)^n g(\theta|\alpha)\, d\theta.$$

We now show that $\bar{r}(\alpha|d\alpha)\to 1$ as $\alpha\to\infty$. As noted before $f(\alpha)=\alpha+0(1)\to\infty$ as $\alpha\to\infty$ Hence for every $\epsilon>0$, no matter how small, we can find an A such that for every $\alpha>A$ the integral

$$\int_{f(\alpha)+l}^{\infty}\left(1-\frac{2l}{\theta}\right)^{n}g(\theta|\alpha)\,d\theta>(1-\epsilon)\int_{f(\alpha)+l}^{\infty}g(\theta|\alpha)\,d\theta>\int_{f(\alpha)+l}^{\infty}g(\theta|\alpha)\,d\theta-\epsilon.$$

$\therefore$ for every $\alpha>A$ we have

$$\bar{r}(\alpha|d\alpha)>1-\int_{f(\alpha)-l}^{f(\alpha)+l}g(\theta|\alpha)\,d\theta-\epsilon.$$

It is easily verifiable that the maximum of $g(\theta|\alpha)$ is attained at $\theta=\alpha-1$ and that $g(\alpha-1\,|\,\alpha)\to 0$ as $\alpha\to\infty$

Hence
$$\int_{f(\alpha)-l}^{f(\alpha)+l}g(\theta|\alpha)\,d\theta\to 0 \text{ as } \alpha\to\infty.$$

Thus it is proved that $\bar{r}(\alpha|d\alpha)\to 1$ as $\alpha\to\infty$. It at once follows that $\sup_{\theta} r(\theta|d)=1$ for every decision function $d(\mathbf{x})$. For if for a particular $d(\mathbf{x})$, $\sup_{\theta} r(\theta|d)=1-\delta$ then for that d the average risk function

$$\bar{r}(\alpha|d)=\int_{0}^{\infty}r(\theta|d)\,g\,(\theta/\alpha)\,d\theta\leqq 1-\delta \text{ for all } \alpha.$$

But as $\bar{r}(\alpha|d\alpha)\to 1$ there exists an α for which

$$\bar{r}(\alpha|d\alpha)>\bar{r}(\alpha|d) \quad .. \quad .. \quad .. \quad .. \quad (4.7)$$

Which is a contradiction because $d\alpha$ is the Bayes' solution corresponding to the a priori distribution $g(\theta|\alpha)$.

Thus we find that if the range of θ be from 0 to ∞ then for every decision function the maximum risk must be unity and hence the question of minimising the maximum risk does not arise. The reason why Wald's existence theorems do not hold in this case is that the decision space is not compact although all the other conditions are satisfied. It is conjectured that if θ be a location parameter then under certain very mild conditions (as for example $W(\theta, d)$ is a function of $|\theta-d|$ etc.) the minimax estimator for θ will exist even though the space Ω be unbounded. A general proof is yet to be given. It is also conjectured that in every case the minimax estimator will exist if we define our loss function suitably. As for example for the upper bound of the rectangular distribution the minimax estimator for θ will exist if we take our loss function $W(\theta, d)$ as say $(d-\log\theta)^2$.

5. Some comments on the estimation of Poisson mean when the parameter space is restricted:—The problem of estimating the parameters of a statistical distribution function when it is known, before the sample is drawn, that the parameter certainly belongs to a given set of numbers has been considered recently by Hammersley (1950) and it seems to present some points of peculiar interest. We now consider the problem of estimating the Poisson mean θ when it is a priori known that $0\leqq\theta\leqq 1$. The more general case of $0\leqq\theta\leqq a$ presents no new difficulties.

Let $\mathbf{x}=(x_1, x_2, \ldots, x_n)$ be a random sample drawn from a Poisson population with mean θ where it is known that $0\leqq\theta\leqq 1$ and let $X=x_1+\ldots+x_n$. Let $d(\mathbf{x})$ be an estimator of θ and let

$$r(\theta|d)=E(\theta-d(\mathbf{x}))^2$$

be the corresponding risk function.

Since the loss function is $(\theta-d)^2$ it follows from Rao's Theorem that we need consider only such estimators $d(\mathbf{x})$ as are functions of the sufficient statistics X. Again since X is a Poisson variable with mean $n\theta$ therefore there is no real loss of generality if we somewhat simplify our problem assuming that we are estimating on the basis of a single sample x. Thus if $d(x)$ be any estimator based on a single sample x then the corresponding risk function is

$$r(\theta\,|\,d) = \sum_0^\infty (\theta-d(x))^2\, e^{-\theta}\, \frac{\theta^x}{x!} \quad (0 \leqq \theta \leqq 1) \qquad .. \qquad .. \qquad (5.1)$$

Now if $\xi(\theta)$ be any a priori distribution function over the parameter space $0 \leqq \theta \leqq 1$ then the corresponding average risk function is

$$\bar{r}(\xi\,|\,d) = \int_0^1 r(\theta\,|\,d)\, d\xi(\theta) \qquad .. \qquad .. \qquad .. \qquad (5.2)$$

$$= \sum_0^\infty \frac{1}{x!} \int_0^1 (\theta-d)^2\, e^{-\theta}\, \theta^x d\xi(\theta)$$

The term by term integration in (5.2) is permissible as the series in (5.1) is uniformly convergent in $0 \leqq \theta \leqq 1$ for all estimators $d(x)$ that take values only in the range 0 to 1. Clearly we need not consider any estimator $d(x)$ that takes values outside the interval (0, 1) for corresponding to any such estimator we can always find another estimator always lying in (0, 1) but generating a risk function that is uniformly less than $r(\theta|d)$. Now $\bar{r}(\xi|d)$ will be a minimum if corresponding to any x we choose $d(x)$ in such a way that the integral

$$\int_0^1 (\theta-d(x))^2\, e^{-\theta}\, \theta^x\, d\xi(\theta) \text{ is a minimum.}$$

In other words the Bayes' solution corresponding to the a priori distribution $\xi(\theta)$ is

$$d_\xi^B(x) = \frac{\int_0^1 \theta^{x+1}\, e^{-\theta}\, d\xi(\theta)}{\int_0^1 \theta^x\, e^{-\theta}\, d\xi(\theta)} \qquad .. \qquad .. \qquad .. \qquad (5.3)$$

If we define the distribution function $\xi_1(\theta)$ as

$$\xi_1(\theta) = \frac{\int_0^\theta e^{-\eta}\, d\xi(\eta)}{\int_0^1 e^{-\eta}\, d\xi(\eta)} \qquad .. \qquad .. \qquad .. \qquad .. \qquad (5.4)$$

then it follows that

$$d_\xi(x) = \frac{\int_0^1 \theta^{x+1}\, d\xi_1(\theta)}{\int_0^1 \theta^x\, d\xi_1(\theta)} = \frac{\mu_{x+1}}{\mu_x} \qquad .. \qquad .. \qquad (5.5)$$

where μ_x is the x^{th} moment of θ with respect to the distribution $\xi_1(\theta)$. Wald has proved that under certain conditions (which are satisfied here) every Bayes' solution

is an admissible decision function and that the class of all Bayes' solutions is a complete class of decision functions. Thus we have the following:

Theorem 5.1:—A necessary and sufficient condition in order that an estimator $d(x)$ of the Poisson parameter θ $(0 \leqq \theta \leqq 1)$ be admissible is that

$$d(x) = \frac{\mu_{x+1}}{\mu_x}$$

where μ_x is the x^{th} moment of a random variable distributed over the interval $(0, 1)$.

That the condition is sufficient is proved as follows:

If
$$d(x) = \frac{\mu_{x+1}}{\mu_x} = \frac{\int_0^1 \theta^{x+1}\, d\xi_1(\theta)}{\int_0^1 \theta^x\, d\xi_1(\theta)}$$

then defining

$$\xi(\theta) = \int_0^\theta e^\eta\, d\xi_1(\eta) \Big/ \int_0^1 e^\eta\, d\xi_1(\eta)$$

we have

$$d(x) = \int_0^1 \theta^{x+1} e^{-\theta}\, d\xi(\theta) \Big/ \int_0^1 \theta^x e^{-\theta}\, d\xi(\theta) = d_\xi(x)$$

i.e. $d(x)$ is the Bayes' solution corresponding to the a priori distribution $\xi(\theta)$ and as such is admissible. Now since the quadratic form in u, v, namely

$$\int_0^1 \left(u\theta^{\frac{x-1}{2}} + v\theta^{\frac{x+1}{2}}\right)^2 d\xi_1(\theta)$$

$$= u^2 \mu_{x-1} + 2uv\, \mu_x + v^2 \mu_{x+1}$$

is positive definite therefore it follows that

$$\mu_{x-1}\, \mu_{x+1} \geqq \mu_x^2$$

or
$$d_\xi(x) = \mu_{x+1} \,|\, \mu_x \geqq \mu_x \,|\, \mu_{x-1} = d_\xi(x-1).$$

If the sign of equality holds for any x then it follows that there exists two constants u_0 and v_0 such that

$$u_0\theta^{\frac{x-1}{2}} + v_0\theta^{\frac{x+1}{2}} = 0$$

for all θ excepting possibly in a set of probability measure zero under the d.f. $\xi_1(\theta)$.

But the above equation is satisfied only when

$$\theta = 0 \quad \text{or} \quad \theta = -\frac{u_0}{v_0}$$

Thus we have proved the following:

Theorem 5.2:—The Bayes' solution $d_\xi(x)$ is a strictly increasing function of x excepting when the d.f. $\xi(\theta)$ is such that θ can take only one or two values (one of which is zero) under $\xi(\theta)$.

If under $\xi(\theta)$ θ takes the value α $(0 \leqq \alpha \leqq 1)$ with unit probability then the corresponding Bayes' solution is obviously $d_\xi(x) = \alpha$ for all x.

If under $\xi(\theta)$ θ takes only the two values 0 and α $(0 < \alpha \leqq 1)$ then the corresponding Bayes' solution is

$$d_\xi(x) = \begin{cases} C \text{ when } x = 0 \\ \alpha \text{ when } x > 0 \end{cases} \quad (0 < C < \alpha \leqq 1)$$

If $d\xi(\theta) = \text{Const. } e^\theta \theta^{l-1}(1-\theta)^{m-1} d\theta$ $(0 \leqq \theta \leqq 1, l > 0, m > 0)$ then it is easily verified that

$$d_\xi(x) = \frac{x+l}{x+l+m}$$

Thus corresponding to every a priori distribution function $\xi(\theta)$ there exists a unique Bayes' solution $d\xi(x)$ that minimises the average risk $\bar{r}(\xi/d)$ the minimum value being denoted by $\bar{r}_\xi = \bar{r}(\xi/d_\xi)$

Wald has shown that there exists a d.f. $g(\theta)$ called the least favourable a priori d.f., such that $\bar{r}_g \geqq \bar{r}_\xi$ for any alternative d.f. $\xi(\theta)$.

Further the minimax decision function is the Bayes' solution corresponding to the least favourable a priori d.f. $g(\theta)$ and the risk function generated by $d_g(x)$ is $\leqq \bar{r}_g$ for all θ $(0 \leqq \theta \leqq 1)$. It at once follows that $r(\theta/d_g) = r_g$ for all θ excepting possibly in a set if probability measure zero under the d.f. $g(\theta)$.

Now since the series in (5.1) is uniformly convergent in $0 \leqq \theta \leqq 1$ for all admissible decision function $d(x)$ we have that $r(\theta/d_g)$ is a continuous function of θ. Hence if $g(\theta)$ be a continuous distribution function then it follows that the risk function $r(\theta|d_g)$ generated by the minimax decision function $d_g(x)$ is a constant over the range $0 \leqq \theta \leqq 1$.

But if
$$r(\theta|d) = \sum_0^\infty (\theta - d(x))^2 e^{-\theta} \frac{\theta^x}{x!} = C$$

then by multiplying both sides by e^θ and equating for like powers of θ we have the difference equation

$$d^2(0) = C \quad .. \quad .. \quad .. \quad .. \quad (5.6)$$

$$d^2(n) - 2n\, d(n-1) + n(n-1) = C \text{ for } n > 0.$$

But since $d(n)$ lies in the range (0, 1), $d(x)$ being admissible, therefore the l.h.s. of (6) is of order n^2 whereas the r.h.s. is a constant. Thus we have the following:

Theorem 5.3. There exists no admissible decision function that generates a constant risk function.

In an appendix we give an interesting proof of the fact that the difference equation (6) cannot be solved even when there is no restriction on $d(x)$ and the r.h.s. is a function of n that does not increase too rapidly.

An immediate consequence of Theorem 5.3 is

Theorem 5.4 :—The least favourable a priori d.f. $g(\theta)$ is discrete.

Let ξ^π be the simple discrete distribution of θ under which θ takes the values 0 and 1 with probabilities $(1-\pi)/1+(e-1)\pi$ and $e\pi/1+(e-1)\pi$ where $0 < \pi < 1$. From (5.4) we have the associated d.f. ξ_1^π as one under which θ takes the values 0 and 1 with probabilities $1-\pi$ and π respectively.

From (5.5) we have the corresponding Bayes' solution

$$d^\pi = d_{\xi^\pi}(x) = \begin{cases} \pi \text{ when } x = 0 \\ 1 \text{ when } x > 0 \end{cases} \quad .. \quad .. \quad .. \quad (5.7)$$

The risk function generated by $d^{\pi}(x)$ is

$$r(\theta|d^{\pi}) = (\theta-\pi)^2 e^{-\theta}+(\theta-1)^2(1-e^{-\theta}) \quad \ldots \quad \ldots \quad \ldots \quad (5.8)$$

and the corresponding average risk is

$$\bar{r}_{\pi} = \bar{r}(\xi^{\pi}|d^{\pi}) = \int_0^1 r(\theta/d^{\pi})\, d\xi^{\pi}(\theta) \quad \ldots \quad \ldots \quad \ldots \quad (5.9)$$

$$= \frac{\pi(1-\pi)}{1+(e-1)\pi}$$

It can be easily verified that $\bar{r}_{\pi}$ is a maximum when

$$\pi = (1+\sqrt{e})^{-1} = \pi_0 \text{ (say)}$$

Thus in the class of all a priori distributions of the form $\xi^{\pi}(0 \leqq \pi \leqq 1)$ the least favourable d.f. is the one for which $\pi = \pi_0$. We shall see afterwards that this is not the least favourable d.f. in the whole class. Now it is easily verified that

$$r(0/d^{\pi_0}) = r(1/d^{\pi_0}) = \bar{r}_{\pi_0} = (1+\sqrt{e})^{-2} \quad \ldots \quad \ldots \quad (5.10)$$

Again since d^{π_0} is the Bayes' solution corresponding to the a priori d.f. ξ^{π_0} under which θ can take only the two values 0 and 1, therefore we have the following.

Theorem 5.5.:—For every $d(x)$ $\operatorname{Sup}_{\theta} r(\theta/d) > (1+\sqrt{e})^{-2}$.

For if there exists a $d(x)$ other than $d^{\pi_0}(x)$ for which $r(\theta/d) \leqq (1+\sqrt{e})^{-2}$ for both $\theta = 0$ and 1 then that will contradict the fact that $d^{\pi_0}(x)$ is the unique Bayes' solution corresponding to ξ^{π_0}. Again it is easily verified that $\frac{d}{d\theta} r(\theta|d^{\pi}) > 0$ at $\theta = 0$ so long as $\pi < (1+\sqrt{2})^{-1}$ and since $\pi_0 < (1+\sqrt{2})^{-1}$ we have that $r(\theta|d^{\pi_0})$ is an increasing function at $\theta = 0$ which proves the theorem.

Since $r(\theta|d^{\pi_0})$ is an increasing function at $\theta = 0$ it follows that $r(\theta|d^{\pi_0}) > \bar{r}_{\pi_0}$ for some θ and so ξ^{π_0} is not the least favourable a priori d.f. When $\pi = (1+\sqrt{2})^{-1}$ it is easily seen that $\operatorname{Sup} r(\theta|d^{\pi})$ is attained at $\theta = 0$ and $\operatorname{Sup} r(\theta|d^{\pi}) = (1+\sqrt{2})^{-2}$ Thus we have finally:

Theorem 5.6. $$(1+\sqrt{e})^{-2} < \operatorname{Inf}_{d} \operatorname{Sup}_{\theta} r(\theta|d) < (1+\sqrt{2})^{-2}.$$

Appendix:—

Here we prove that the difference equation (a particular case of which we considered in (5.6))

$$u_0^2 = K(0)$$

$$u_m^2 - 2m\, u_{m-1} + m(m-1) = K(m) \text{ for } m > 0$$

has no real solution if

$$\overline{\lim} \frac{K(m)}{m} < 1$$

$$\left(\text{If } K(m) \text{ is a constant then } \overline{\lim} \frac{K(m)}{m} = 0\right).$$

Proof:—From $$u_m^2 - 2m\, u_{m-1} + m(m-1) = K(m)$$

$$2m\, u_{m-1} \geqq m(m-1) - K(m)$$

or

$$\frac{u_{m-1}}{m-1} \geqq \tfrac{1}{2} - \frac{K(m)}{2m(m-1)}$$

$$\therefore \underline{\lim} \frac{u_{m-1}}{m-1} \geqq \tfrac{1}{2} - \overline{\lim} \frac{K(m)}{2m(m-1)}$$

Since by supposition $\overline{\lim} \dfrac{K(m)}{m} < 1$

$$\therefore \overline{\lim} \frac{K(m)}{2m(m-1)} \leqq 0 \text{ and therefore we have}$$

$$\underline{\lim} \frac{u_m}{m} \geqq \tfrac{1}{2}. \qquad \ldots\ldots\ldots\ldots\ldots\ldots \text{(a.1)}$$

Now if $\underline{\lim} \dfrac{u_m}{m} \geqq \alpha_0$

then $u_m > (\alpha - \delta)\, m$ for all $m \geqq N = N(\delta)$

and then $2m\, u_{m-1} = u_m^2 + m(m-1) - K(m)$

$$> (\alpha_0 - \delta)^2 m^2 + m(m-1) - K(m)$$

$$\therefore \frac{u_{m-1}}{m-1} > (\alpha_0 - \delta)^2 \frac{m^2}{2m(m-1)} + \tfrac{1}{2} - \frac{K(m)}{2m(m-1)} \text{ for all } m \geqq N.$$

$\therefore \underline{\lim} \dfrac{u_{m-1}}{m-1} \geqq \tfrac{1}{2}\{1 + (\alpha_0 - \delta)^2\}$ and since δ is arbitrary we have the following result

$$\underline{\lim} \frac{u_m}{m} \geqq \alpha_0 \text{ implies that } \underline{\lim} \frac{u_m}{m} \geqq \tfrac{1}{2}(1 + \alpha_0^2). \qquad \ldots\ldots \text{(a.2)}$$

Now let $\alpha_0 = \frac{1}{2}$ and $\alpha_n = \frac{1}{2}\left(1 + \alpha_{n-1}^2\right)$

It is easily seen that the sequence $\{\alpha_n\}$ is a monotonic increasing bounded sequence and as such $\lim \alpha_n = \alpha$ exists.

$$\because\ \alpha_n = \tfrac{1}{2}\left(1 + \alpha_{n-1}^2\right)$$

$\therefore$ making $n \to \infty$ we have $\alpha = \frac{1}{2}(1 + \alpha^2)$ and hence $\alpha = 1$.

$$\text{But } \underline{\lim} \frac{u_m}{m} \geqq \alpha_n \text{ for every } n \text{ and therefore } \ldots\ldots\ldots \text{(a.3)}$$

$$\underline{\lim} \frac{u_m}{m} \geqq 1.$$

Again since the difference equation can be written in the form $(u_m - m)^2 + 2m(u_m - u_{m-1}) = m + K(m)$ we at once have

$$u_m - u_{m-1} \leqq \frac{1}{2} + \frac{K(m)}{2m}$$

and since $\overline{\lim} \dfrac{K(m)}{m} < 1$

$$\therefore \overline{\lim}\, (u_m - u_{m-1}) < 1. \qquad \ldots\ldots\ldots\ldots \text{(a.4)}$$

Hence there exists a $\delta > 0$ and an N such that

$$u_{N+i} - u_{N+i-1} < 1-\delta \quad \text{for } i = 1, 2, -\ldots \text{ ad inf.}$$

$$\therefore u_{N+n} - u_N = \sum_{1}^{n} (u_{N+i} - u_{N+i-1})$$

$$< (1-\delta)n$$

$$\therefore \frac{u_{n+N}}{N+n} < (1-\delta)\frac{n}{N+n} + \frac{u_N}{N+n}$$

and hence making $n \to \infty$ we have

$$\overline{\lim} \frac{u_m}{m} \leqq 1-\delta \quad \text{which contradicts (a.3)} \ldots \quad \ldots \quad \ldots \quad \text{(a.5)}$$

Thus the required result is proved.

SUMMARY.

It is shown that minimax estimators need not exist in many familiar cases unless the parameter space be truncated or the loss function suitably modified. For the rectangular distribution with range $(0, \theta)$ there do not exist any minimax estimator for θ if the loss function be of the simple zero-one type. The case of the Poisson mean has been considered in some details and a few interesting results obtained. The mean θ is assumed to be between 0 and 1 and the loss function is taken to be the square of the error. A complete characterization of the class of admissible estimators is given. It is proved that the least favourable a priori distribution must be discrete, for there exists no estimator that generates a constant risk function. Bounds for the minimum of the maximum risk have also been given.

REFERENCES.

Barankin, E. W. (1950). Extension of a theorem of Blackwell. *Annals of Math. Stat.*, **21**, 280.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ibid.*, **18**, 105.

Hammersley, J. M. (1950). On estimating restricted parameters. *Jour. of the Roy. Stat. Soc.*, **12**, No. 2.

Hodges, J. L. and Lehmann, E. H. (1950). Some problems of Minimax point estimation. *Annals of Math. Stat.*, **21**, 182.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Cal. Math. Soc.*, **37**, 81.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Math. Stat.*, **10**, 299.

———— (1945). Statistical Decision Functions which minimise the maximum risk. *Annals of Math.*, **46**, 265.

———— (1947). An essentially complete class of admissible decision functions. *Annals of Math. Stat.*, **18**, 549.

———— (1950). Statistical Decision Functions. Wiley Publications.

Walfowitz, J. (1950). Minimax estimates of the mean of a Normal variable with known variance. *Annals of Math. Stat.*, **21**, 218.