DECISION RULE BY PROBABILITY RATIO

BY KAMEO MATUSITA

(Received Aug. 20, 1954)

1. Introduction

As is well known, Neyman and Pearson employed the probability ratio test to obtain the most powerful test. In the sample space of a fixed dimension, even by the most powerful test the Neyman-Pearson theory can not control both the errors, that is, it can control the error of first kind, but cannot control that of second kind. (This can at most be evaluated.) However, when we adopt the sequential test which Wald developped we can control both the errors. Wald employed there also the probability ratio test, admitting to take a sample of any size (see [1]). This probability ratio test was further shown by Wald and Wolfowitz [2] to have an optimum character in the sense that it has the smallest average sample number among the sequential tests with the errors equal to or smaller than those of it. On the other hand, we have shown in [3] how the smallest risk (the larger of the two errors) is bounded in the sample space of a finite dimension. This provides, and assures at the same time, a method of determining the sample size for obtaining a decision rule with the risk (due to the errors) smaller than a preassigned value. According to our result we can do always with samples of a bounded size as long as we are concerned with a given risk. In adopting the sequential test, therefore, we can also restrict ourselves to samples of a bounded size. Namely, we can always confine ourselves to the truncated sequential rule. This can be observed also from [1], but our treatment is more exact. To state about this together with Bayes and minimax solutions is the purpose of this paper. In the following section we shall give a simple proof to the fact that in a finite dimensional space a probability ratio rule gives a Bayes solution with respect to any a priori distribution, especially to a least favorable distribution, and in some case a minimax solution. In section 3, we shall consider a sequential probability ratio rule in a finite dimensional space. The truncated sequential rule is excellent, as is just oboious, in the point that the sample number is always bounded, compared with the non-truncated one, for the sample number in the non-truncated one may happen to be large.

2. Bayes and minimax solutions

Let R be any space, and let $F_1(E)=\int_E f_1(x)\,dm$, $F_2(E)=\int_E f_2(x)\,dm$ be two (distinct) distributions in R, where m is a (Lebesgue) measure in R. When it is known that the random variable X we observe has distribution F_1 or F_2, suppose we want to decide on the basis of the sample value which distribution X has. For this purpose, we consider a decision function or decision rule, that is, a mapping from R onto $D=\{d_1, d_2\}$, where d_i denotes the decision that X has F_i for $i=1, 2$. There are, of course, very many decision rules concerning a problem. We should, therefore, adopt an efficient one for the problem. To make clear the meaning of "efficient", we introduce a weight function $W(F_i, d_j)$. In the sequel, for the sake of simplicity we put

$$W(F_i, d_j)=\begin{cases}1 & \text{when} \quad i=j\\ 0 & \text{when} \quad i\neq j\end{cases} \qquad (i, j=1, 2)$$

This means that, in terms of test, we consider the error of first kind and that of second kind with the equal weight. Other cases are treated quite similarly as this case. Then, for an a priori probability $\mu=\{p, q\}$ on $\{F_1, F_2\}$ and a decision function $d=\varphi(x)$ we have

$$r(\mu, \varphi)=p\int_S f_1(x)\,dm+q\int_{S^c} f_2(x)\,dm$$

as the risk, where $S=\{x\,;\, d_1=\varphi(x)\}$ and $S^c=R-S$. This $r(\mu, \varphi)$ is rewritten as

$$r(\mu, \varphi)=p-\int_S (pf_1(x)-qf_2(x))\,dm$$

Therefore, for μ, $r(\mu, \varphi)$ takes on the minimum value when

$$S=\{x\,;\; pf_1(x)\geqq qf_2(x)\}.$$

THEOREM I. *A Bayes solution with respect to* $\mu=\{p, q\}$ *is given by*

$$S_\mu=\{x\,;\; pf_1(x)\geqq qf_2(x)\}\ ^{*)}$$

Now, let

*) This form is superior to the probability ratio form, since it prevails when $f_1(x)$ or $f_2(x)$, p or q attain zero.

$$F_1(S_{\mu_0})=F_2(S^c_{\mu_0})$$

where $\mu_0=\{p_0, q_0\}$ is an a priori distribution on $\{F_1, F_2\}$ and $S_{\mu_0}=\{x;\ p_0 f_1(x) \geqq q_0 f_2(x)\}$. Then we have

THEOREM II. *For any a priori distribution μ it holds that*

$$r(\mu, \varphi_\mu) \leqq r(\mu_0, \varphi_{\mu_0})$$

where φ_μ, φ_{μ_0} denote Bayes solutions with respect to μ and μ_0, respectively. Therefore, μ_0 is a least favorable distribution.

PROOF: Let $\mu=\{p, q\}$ and

$$p=p_0+\delta, \qquad \delta>0$$

Then we have clearly

$$S_\mu \supseteqq S_{\mu_0}$$

So, putting $S_\mu=S_{\mu_0}+\Delta$, we have

$$\begin{aligned}
r(\mu, \varphi_\mu)&=p\int_{S^c_\mu} f_1(x)\,dm+q\int_{S_\mu} f_0(x)\,dm\\
&=p\int_{S^c_{\mu_0}-\Delta} f_1(x)\,dm+q\int_{S_{\mu_0}+\Delta} f_2(x)\,dm\\
&=p\int_{S^c_{\mu_0}} f_1(x)\,dm-p\int_\Delta f_1(x)\,dm\\
&\quad+q\int_{S_{\mu_0}} f_2(x)\,dm+q\int_\Delta f_2(x)\,dm\\
&=p_0\int_{S^c_{\mu_0}} f_1(x)\,dm+q_0\int_{S_{\mu_0}} f_2(x)\,dm\\
&\quad+\delta\Big(\int_{S^c_{\mu_0}} f_1(x)\,dm-\int_{S_{\mu_0}} f_2(x)\,dm\Big)\\
&\quad-\int_\Delta (pf_1(x)-qf_2(x)\,dm\\
&=r(\mu_0, \varphi_{\mu_0})-\int_\Delta (pf_1(x)-qf_2(x))\,dm
\end{aligned}$$

(according to $F_1(S_{\mu_0})=F_2(S^c_{\mu_0})$)

Now, $\Delta \subseteqq S$, therefore

$$\int_\Delta (pf_1(x)-qf_2(x))\,dm \geqq 0$$

consequently

$$r(\mu,\ \varphi_{\mu}) \leqq r(\mu_0,\ \varphi_{\mu_0})$$

For $\mu=\{p,\ q\}$ with $p=p_0-\delta,\ \delta>0$, we have similarly

$$r(\mu,\ \varphi_{\mu}) \leqq r(\mu_0,\ \varphi_{\mu_0})$$

Thus we get

$$\max_{\mu}\ \min_{\varphi}\ r(\mu,\ \varphi)=r(\mu_0,\ \varphi_{\mu_0})$$

THEOREM III. *Let S be any set and $S_0=\{x;\ f_1(x)\geqq \alpha f_2(x)\}$ with $\alpha>0$. Then, if*

(*) $$F_2(S) \leqq F_2(S_0)$$

we have

$$F_1(S) \leqq F_1(S_0)$$

Similarly, if

($^*_*$) $$F_1(S^c) \leqq F_1(S_0^c)$$

we have

$$F_2(S^c) \leqq F_2(S_0^c)$$

This is a slight generalization of the Neyman-Pearson's fundamental lemma.

PROOF: For any S we have

$$\begin{aligned} F_2(S_0) \geqq F_2(S) &= F_2(S\frown S_0)+F_2(S\frown S_0^c) \\ &\geqq F_2(S\frown S_0)+\alpha F_1(S\frown S_0^c) \end{aligned}$$

therefore

$$\begin{aligned} F_1(S\frown S_0^c) &\leqq \alpha(F_2(S_0)-F_2(S\frown S_0)) \\ &= \alpha F_2(S_0\frown S^c) \\ &\leqq F_1(S_0\frown S^c) \end{aligned}$$

which implies that

$$\begin{aligned} F_1(S) &= F_1(S\frown S_0)+F_1(S\frown S_0^c) \\ &\leqq F_1(S\frown S_0)+F_1(S_0\frown S^c) \\ &= F_1(S_0) \end{aligned}$$

The remaining part is proved similarly.

THEOREM IV. *φ_{μ_0} is a minimax solution.*

PROOF: For a decision rule φ, let

$$F_1(S^c) < F_2(S)$$

where $S=\{x\,;\ d_1=\varphi(x)\}$. Then we have

$$\max_{\mu} r(\mu,\ \varphi)=F_2(S)$$

and

$$F_2(S_{\mu_0}) < F_2(S)$$

For, according to the preceding theorem, from $F_2(S) \leqq F_2(S_{\mu_0})$ follows $F_1(S) \leqq F_1(S_{\mu_0})$, hence $F_2(S) \leqq F_2(S_{\mu_0})=F_1(S^c_{\mu_0}) \leqq F_1(S^c)$, which contradicts

$$F_1(S^c) < F_2(S)$$

Therefore, we have

$$r(\mu_0,\ \varphi_{\mu_0}) < \max_{\mu} r(\mu,\ \varphi)$$

When $F_1(S^c) > F_2(S)$, we have the same relation, too.

Now, let

$$F_1(S^c)=F_2(S).$$

If $F_1(S^c) < F_1(S^c_{\mu_0})$, then $F_2(S^c) \leqq F_2(S^c_{\mu_0})$, hence $F_1(S^c) < F_1(S^b_{\mu_0})=F_2(S_{r_0}) \leqq F_2(S)$, which is a contradiction.

Therefore, it must hold:

$$F_1(S^c)=F_2(S)=F_1(S^c_{\mu_0})=F_2(S_{\mu_0})$$

consequently

$$r(\mu_0,\ \varphi_{\mu_0}) \leqq \max\ r(\mu,\ \varphi)$$

which shows that φ_{μ_0} is a minimax solution.

If F_1 and F_2 are continuous distributions, there exists a set S_0 such that

$$S_0=\{x\,;\ f_1(x) \geqq \alpha f_2(x)\}$$

and

$$F_1(S^c_0)=F_2(S_0)$$

This S_0 defines a minimax solution. Further, if F_1 and F_2 are unimodal distributions of one dimension, a set S_0 with $F_1(S^c_0)=F_2(S_0)$ is uniquely determined except for a set of m-measure zero, therefore, a minimax solution is almost uniquely determined. In case there does not exist a set S_0 such that $S_0=\{x\,;\ f_1(x) \geqq \alpha f_2(x)\}$ and $F_1(S^c_0)=F_2(S_0)$, a probability ratio rule does not always provide a minimax solution, as can be seen from the following example.[*]

[*] I am indebted to my colleague H. Akaike for this part.

Let F_1, F_2 be $\{0.2, 0.2, 0.6\}$, $\{0.1, 0.4, 0.5\}$ on events (1), (2), (3), respectively. Then we have

$$\min_{\alpha} \max \{F_1(S_\alpha^c),\ F_2(S_\alpha)\} = 0.6$$

where $S_\alpha = \{x:\ f_1(x) \geqq \alpha f_2(x)\}$, and

$$\min \max \{F_1(S^c),\ F(S)\} = 0.5$$

where S runs over any subset of $\{(1), (2), (3)\}$.

3. Sequential rule

Let $F_1(E) = \int_E f_1(x)\, dm$, $F_2(E) = \int_E f_2(x)\, dm$ be two distributions in R and ρ the affinity between F_1 and F_2, i.e.,

$$\rho = \int_R \sqrt{f_1(x)}\ \sqrt{f_2(x)}\ dm.$$

Further, let d_i denote decision that the random variable X we observe has F_i for $i=1, 2$. Then, if we want to make decision d_1 or d_2 with the risk smaller than a preassigned positive number ε, we have only to take a sample of size n with $\rho^n < \varepsilon$ and adopt a minimax solution. For we have then

$$F_1^{(n)}(S^c) < \varepsilon$$
$$F_2^{(n)}(S) < \varepsilon$$

where $S = \{(x_1 \ldots, x_n);\ f_1(x_1) \ldots f_1(x_n) \geqq f_2(x_1) \ldots f_2(x_2)\}$ and $F_1^{(n)}$, $F_2^{(n)}$ are the extentions of F_1, F_2 into $R_n = \{(x_1, \ldots x_n)\}$. In case F_1 and F_2 are continuous, a minimax solution is given by

$$S_n = \{(x_1, \ldots, x_n);\ f_1(x_1) \ldots f_1(x_n) \geqq \alpha f_2(x) \ldots f_2(x_n)\}$$

with $F_1^{(n)}(S_n^c) = F_2^{(n)}(S_n)$.

$$F_1^{(n)}(S_n^c) = F_2^{(n)}(S_n) = r_n < \rho^n < \varepsilon$$

Let S_k be generally a set in the k-dimensional space which gives a minimax solution, and denote the risk of the minimax solution by r_k. Then we have

$$r_1 \geqq r_2 \geqq \cdots \geqq r_k \geqq r_{k+1} \geqq \cdots$$

Now, with a sequential rule φ in R_n, there is associated a system of sets $\{A_k, B_k\}$ $(k=1, 2, \ldots, n)$ such that $A_k = \{(x_1, \ldots x_k);\ \varphi(x_1, \ldots, x_k) = d_1\}$ and $B_k = \{(x_1, \ldots x_n);\ \varphi(x_1, \ldots x_k) = d_2\}$. This system

$\{A_k, B_k\}$ has the property that $A_k, B_k \subseteqq R_k$, $A_k \frown B_k = O$ $(k=1, 2, \ldots, n)$ and $A_n + B_n = R_n$. Conversely, any system $\{A_k, B_k\}$ with this property defines a sequential rule in R_n. The average risk of decision rule $\{A_k, B_k\}$ with respect to an a priori probability $\mu = \{p, q\}$, which takes into account the sample number, is then given by

$$\begin{aligned} r(\mu, \varphi) = p\Big\{ & \sum_{k=1}^{n} F_1 (X_1 \in A_1^c \frown B_1^c, \ldots, \quad (X_1, \ldots, X_{k-1}) \in A_{k-1}^c \frown B_{k-1}^c, \\ & \qquad\qquad (X_1, \ldots, X_k) \in B_k) \\ & + c \sum_{k=1}^{n} k F_1 (X_1 \in A_1^c \frown B_1^c, \ldots, \quad (X_1, \ldots, X_{k-1}) \in A_{k-1}^c \frown B_{k-1}^c, \\ & \qquad\qquad (X_1, \ldots, X_k) \in A_k + B_k) \Big\} \\ + q\Big\{ & \sum_{k=1}^{n} F_2 (X_1 \in A_1^c \frown B_1^c, \ldots, \quad (X_1, \ldots, X_{k-1}) \in A_{k-1}^c \frown B_{k-1}^c, \\ & \qquad\qquad (X_1, \ldots, X_k) \in A_k) \\ & + c \sum_{k=1}^{n} k F_2 (X_1 \in A_1^c \frown B_1^c, \ldots, \quad (X_1, \ldots, X_{k-1}) \in A_{k-1}^c \frown B_{k-1}^c, \\ & \qquad\qquad (X_1, \ldots, X_k) \in A_k + B_k) \Big\} \end{aligned}$$

where c denotes the cost of one observation. We should, therefore, adopt an optimum decision rule concerning this $r(\mu, \varphi)$. Further, $\{A_k, B_k\}$ is desirable to be easily defined. As to this point Wald's definition of the sequential probability ratio test for a given strength is eminent, although it does not exactly have the given strength (see [1]). With this consideration in mind, for an arbitrarily chosen positive number ξ which is smaller than 1 and an integer n, we put

$$\begin{aligned} A_k &= \{(x_1, \ldots, x_k);\ f_1(x_1) \ldots f_1(x_k) \geqq \frac{1}{\xi} f_2(x_1) \ldots f_2(x_k)\} \\ B_k &= \{(x_1, \ldots, x_k);\ f_1(x_1) \ldots f_1(x_k) \leqq \xi f_2(x_1) \ldots f_2(x_2)\} \\ & \qquad\qquad (k=1, 2, \ldots, n-1) \\ A_n &= S_n \\ B_n &= S_n^c \end{aligned}$$

Then we have

THEOREM V. *The system* $\{A_k, B_k\}$ $(k=1, 2, \ldots, n)$ *defines a truncated sequentical rule with the risk due to the errors in inference smaller than* $\xi + \rho^n$.*)

The average sample number according to this rule is not so different from that of the sequential probability ratio test with $A_k' = \{(x_1, \ldots, x_k);$

*) Concerning this see [1].

$f_1(x_1)\ldots f_1(x_k) \geqq \frac{1}{\xi} f_2(x_1)\ldots f_2(x_k)\}$ and $B'_k = \{(x_1, \ldots, x_k);\ f_1(x_1)\ldots f_1(x_k) \leqq f_2(x_1)\ldots f_2(x_k)\}$ or $A''_k = \{(x_1, \ldots, x_k);\ f_1(x_1)\ldots f_1(x_k) \geqq \frac{1-\xi}{\xi} f_2(x_1)\ldots f_2(x_k)\}$ and $B''_k = \{(x_1, \ldots, x_k);\ f_1(x_i)\ldots f_1(x_k) \leqq \frac{\xi}{1-\xi} f_2(x_1)\ldots f_2(x_k)\}$ $(k=1, 2, \ldots$ ad. inf.) when ξ is not so large. Due to the above theorem, if we want to make decision with the risk due to the errors smaller than ε, we have only to take ξ and n such that $\xi + \rho^n < \varepsilon$. Therefore, if we want to make n small, we must make ξ also small, and if we want to make ξ large, we must make n also large. However, since it holds that $\varepsilon + \rho^n \fallingdotseq \varepsilon$ for n not so small we can actually employ

$$A'''_k = \left\{(x_1, \ldots, x_k);\ f_1(x_1)\ldots f_1(x_k) \geqq \frac{1}{\varepsilon} f_2(x_1)\ldots f_2(x_k)\right\}$$
$$B'''_k = \{(x_1, \ldots, x_k);\ f_1(x_1)\ldots f_1(x_k\} \leqq \varepsilon f_2(x_1)\ldots f_2(x_k)\}$$
$$(k=1, 2, \ldots, n-1)$$
$$A'''_n = \{(x_1, \ldots, x_n);\ f_1(x_1)\ldots f_1(x_n) \geqq f_2(x_1)\ldots f_2(x_n)\}$$
$$B'''_n = \{(x_1, \ldots, x_n);\ f_1(x_1)\ldots f_1(x_n) < f_2(x_1)\ldots f_2(x_n)\}$$

for that purpose.

For example, let $f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2}}$, $f_2(=x)\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m_2)^2}{2}}$ and $\varepsilon = 0.05$. Then we have in case $|m_1 - m_2| \geqq 1$,

$$\varepsilon + \rho^n \leqq 0.051 \quad \text{for} \quad n \geqq 56$$
$$\varepsilon + \rho^n \leqq 0.054 \quad \text{for} \quad n \geqq 45$$

and, in case $|m_1 - m_2| \geqq 2$,

$$\varepsilon + \rho^n \leqq 0.051 \quad \text{for} \quad n \geqq 14$$
$$\varepsilon + \rho^n \leqq 0.054 \quad \text{for} \quad n \geqq 12$$

as the bounds of the risk due to the errors in adopting our rule.

Thus far, we have explained sequential rules in which decision is made after each observation. But our rules apply in these forms to a case where at each stage a group of observatians are made at the same time. Moreover, it is not necessary for our rules that all observations are independent of each other, if the affinity in R_k concerning F_1 and F_2 tends to the zero as $k \to \infty$.

We shall now refer to the case when a sequential rule gives advantage. Let c_1 denote the cost of a thing to be inspected when observation (inspection) destroys the thing inspected, or zero otherwise, and c_2k+c_3 the labor of observation (inspection) for a sample of size k. Then, if we adopt the rule defined by $\{A_k, B_k\}$, the expected value of the cost and labor is

$$(C_1) \qquad (c_1+c_2+c_3)\sum_k kF_1(X_1 \in A_1^c \frown B_1^c, \dots, (X_1, \dots, X_{k-1}) \in A_{k-1}^c \frown B_{k-1}^c, (X_1, \dots, X_k) \in A_k+B_k)$$

or

$$(C_1) \qquad (c_1+c_2+c_3)\sum_k kF_2(X_1 \in A_1^c \frown B_1^c, \dots, (X_1, \dots, X_{-k1}) \in A_{k-1}^c \frown B_{k-1}^c, (X_1, \dots, X_k) \in A_k+B_k)$$

On the other hand, if we adopt a non-sequential rule, with the same risk due to errors in inference the value of the cost and labor is

$$(C_3) \qquad c_1n+c_2n+c_3$$

where n is the sample size in the non-sequential rule. Therefore, we obtain a profit by adopting the sequential rule when (C_1) and (C_2) is smaller than (C_3). In the case where the average sample number of the sequential rule is about $n/2$, the sequential rule is advantageous when $c_1+c_2 > c_3$. However, when adopting a non-truncated sequential rule, care must be taken in managing only with the average sample number, for we may have a large variance of the sample number.

The rules above mentioned apply, of course, to the case where d_1, d_2 are decisions concerning two classes of distributions, each of which has a representative distribution in the sense of lemma 3 in [4]. Further, the case where several (more than two) distributisns are essentially concerned, are similarly treated as above (see [3]).

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Wald, A, *Sequential analysis*, John Wiley, 1947.

[2] Wald, A. and J. Wolfowitz, Optimum character of the sequential probability test, *Annals of Mathematical Statistics*, Vol. XIX, 1948.

[3] Matusita, Kameo, On the theory of statistical decision functions, *Annals of the Institute of Statistical Mathematics*, Vol. III, 1951.

[4] Matusita, Kameo, Yukio Suzuki and Hirosi Hudimoto, On testing statistical hypotheses, the same number of the *Annals of the Institute of Statistical Mathematics*.