ON THE FUNDAMENTAL THEOREM FOR THE DECISION RULE BASED ON DISTANCE || ||

By KAMEO MATUSITA AND MINORU MOTOO

(Received Oct. 25, 1955)

1. In the theory of decision rule based on distance || ||, which has been developed in the papers [1, 2], the probabilistic inequalities concerning the distance play a fundamental role. We have given in [1, 2] the following inequalities :

For a given positive number η *we have*

$$\text{(I)} \qquad P_r(||F-S_n||>\eta)\leqq\frac{k-1}{n\eta^2}$$

and

$$\text{(II)} \qquad P_r(||F-S_n||>\eta)\leqq 2ke^{-\frac{n\eta^4}{2k^2}}$$

where F *denotes the discrete distribution of the random variable under observation,* k *the number of events the variable takes on and* S_n *the empirical distribution for* n *observations.*

In this note we intend to give another inequality, which will serve better for a wide class of application than the above two. The comparison of it with the others and its applications will also be given.

2. THEOREM. *For any positive number* η *we have*

$$\text{(III)} \qquad P_r(||F-S_n||>\eta)\leqq\frac{k^2+k-1}{(n\eta^2)^2}\leqq\frac{1.25k^2}{(n\eta^2)^2}$$

provided that $k\geqq 2$ *and* $n>k$.

PROOF. Let $F=\{p_1, p_2, \cdots, p_k\}$ and $S_n=\left\{\frac{n_1}{n}, \frac{n_2}{n}, \cdots, \frac{n_k}{n}\right\}$.

Further, let

$$p_i>\frac{1}{n} \qquad (i=1, 2, \cdots, r)$$

$$p_j\leqq\frac{1}{n} \qquad (j=r+1, r+2, \cdots, k)$$

and put $r+s=k$, $\sum_{i=1}^{r} p_i=p$, $\sum_{j=r+1}^{k} p_j=q$. Then, we have

$$r \geqq 1 .$$

For, if $r=0$, we should have

$$s=k$$

and

$$1=\sum_{j=1}^{k} p_j \leqq \frac{k}{n} < 1$$

which is a contradiction.

Now we have

$$E(||F-S_n||^4)=E\left(\sum_{i=1}^{k}\left(\sqrt{\frac{n_i}{n}}-\sqrt{p_i}\right)^2\right)^2$$

$$=E\left(\sum_{i=1}^{r}\left(\sqrt{\frac{n_i}{n}}-\sqrt{p_i}\right)^2+\sum_{j=r+1}^{k}\left(\sqrt{\frac{n_j}{n}}-\sqrt{p_j}\right)^2\right)^2$$

$$\leqq E\left(\sum_{i=1}^{r}\frac{\left(\frac{n_i}{n}-p_i\right)^2}{p_i}+\sum_{j=r+1}^{n}\left(\sqrt{\frac{n_j}{n}}\right)^2\right)^2$$

$$=E\left(\sum_{i=1}^{r}\frac{\left(\frac{n_i}{n}-p_i\right)^2}{p_i}+\frac{m}{n}\right)^2 \qquad \text{where } m=\sum_{j=r+1}^{n} n_j$$

$$=\sum_{i=1}^{r}\frac{E\left(\frac{n_i}{n}-p_i\right)^4}{p_i^2}+\sum_{i\neq k}\frac{E\left(\frac{n_i}{n}-p_i\right)^2\left(\frac{n_k}{n}-p_k\right)^2}{p_ip_k}$$

$$+2\sum_{i=1}^{r}\frac{E\left(\sum\left(\frac{n_i}{n}-p_i\right)^2\left(\frac{m}{n}-q\right)\right)}{p_i}+2q\sum_{i=1}^{r}\frac{E\left(\frac{n_i}{n}-p_i\right)^2}{p_i}$$

$$+E\left(\frac{m}{n}-q\right)^2+q^2 .$$

Since

$$E\left(\left(\frac{n_i}{n}-p_i\right)^2\right)=\frac{1}{n}p_i(1-p_i) ,$$

$$E\left(\left(\frac{n_i}{n}-p_i\right)\left(\frac{n_j}{n}-p_j\right)^2\right)=-\frac{1}{n^2}p_ip_j(1-2p_j) ,$$

$$E\left(\left(\frac{n_i}{n}-p_i\right)^2\left(\frac{n_j}{n}-p_j\right)^2\right)=\frac{1}{n^2}\left[p_ip_j\{(1-p_i)(1-p_j)+2p_ip_j\}\right.$$
$$\left.+\frac{1}{n}p_ip_j\{-1+2p_i+2p_j-6p_ip_j\}\right],$$

$$E\left(\left(\frac{n_i}{n}-p_i\right)^4\right)=\frac{1}{n^2}\left\{3\left(1-\frac{1}{n}\right)p_i^2(1-p_i)^2\right.$$
$$\left.+\frac{1}{n}p_i(1-p_i)(1-3p_i+3p_i^2)\right\}$$
$$=\frac{1}{n^2}\left\{3p_i^2(1-p_i)^2+\frac{1}{n}p_i(1-p_i)(1-6p_i+6p_i^2)\right\},$$

we obtain

$$n^2\left(\sum_{i=1}^{r}\frac{\left(\frac{n_i}{n}-p_i\right)^4}{p_i^2}+\sum_{i\neq j}\frac{E\left(\frac{n_i}{n}-p_i\right)^2\left(\frac{n_j}{n}-p_j\right)^2}{p_ip_j}\right)$$
$$=\sum_{i=1}^{r}3(1-p_i)^2+\frac{1}{np_i}+\frac{1}{n}\{-7+12p_i-6p_i^2\}$$
$$+\sum_{i\neq j}\left\{(1-p_i)(1-p_j)+2p_ip_j+\frac{1}{n}-1+2p_i+2p_j-6p_ip_j\right\}$$
$$\leqq(r-p)^2+2r-4p+2p^2+r+\frac{1}{n}\{-r^2-7r+12p-6p^2+4p(r-1)\}$$
$$=r^2+r(3-2p)-4p+3p^2+\frac{1}{n}\{-r^2-7r+4pr+8p-6p^2\}$$
$$=r^2+r-1+2qr-2g+3q^2+\frac{1}{n}\{-r^2-7r+4pr+8p-6p^2\}$$
$$\leqq r^2+r-1+2qr-2q+3q^2$$

$$n^2\left(2\sum_{i=1}^{r}\frac{E\left(\frac{n_i}{n}-p_i\right)^2\left(\frac{m}{n}-q\right)}{p_i}\right)=-2q\sum_{i=1}^{r}(1-2p_i)$$
$$=-2qr+2pq=-2qr+2q-2q^2,$$

$$n^2\left(2q\sum_{i=1}^{r}\frac{E\left(\frac{n_i}{n}-p_i\right)^2}{p_i}\right)=2nq\sum_{i=1}^{r}(1-p_i)\leqq 2rs,$$

$$n^2\left(E\left(\frac{m}{n}-q\right)^2+q^2\right)=nq(1-q)+n^2q^2\leqq nq+(n^2-n)q^2$$
$$=s+s^2-\frac{s^2}{n}.$$

Therefore, we have

$$E(||F-S_n||^4)\leqq\frac{1}{n^2}\left\{r^2+2rs+s^2+r+s-1+q^2-\frac{s^2}{n}\right\}$$

$$\leqq\frac{1}{n^2}\{(r+s)^2+(r+s)-1\}$$

$$=\frac{1}{n^2}(k^2+k-1).$$

As $k\geqq 2$, the last term is less than or equal to $1.25\,k^2$. We thus obtain

$$Pr\{||F-S_n||>\eta\}\leqq\frac{k^2+k-1}{(n\eta^2)^2}\leqq\frac{1.25k^2}{(n\eta^2)^2}$$

3. *Comparison with the previous inequalities.* Now, we compare the inequality (III) with (I) and (II).

Put

$$A=\frac{k^2+k-1}{(n\eta^2)^2},\quad B=\frac{k-1}{n\eta^2},\quad C=2ke^{-n\eta^4/2k^2}$$

and denote by α the upper bound which we want to set on $Pr\{||F-S_n||>\eta\}$. Actually we evaluate an upper bound of $Pr\{||F-S_n||>\eta$ by A, B or C. Therefore, for a given α we take n such that at least one of A, B and C becomes less than α.

First we have :

$$A\geqq B\rightleftarrows n\eta^2\leqq\frac{k^2+k-1}{k-1}$$

$$\rightleftarrows B\geqq\frac{(k-1)^2}{k^2+k-1}$$

and

$$\frac{(k-1)^2}{k^2+k-1}\begin{cases}=\frac{1}{5} & \text{when } k=2,\\ =\frac{4}{11} & \text{when } k=3,\\ \rightarrow 1 & \text{when } k\rightarrow\infty.\end{cases}$$

Therefore, when we take α less than or equal to 0.2, or 0.36 if $k\geqq 3$, and when we make A or B less than α, it always holds that

$$A\leqq B,$$

which means (III) is preferable to (I). When $\alpha > 0.2$, $A \geqq B$ does not necessarily hold. We have $A \leqq B$ almost always for sufficiently large k. For example, we have $A \leqq B$ for $k \geqq 10$ and $\alpha \leqq 0.73$.

Secondly, we have :

$$A \geqq C \rightleftarrows A \geqq 2ke^{-\frac{k^2+k-1}{2nAk^2}} \rightleftarrows A \geqq 2ke^{-\frac{\sqrt{k^2+k-1}}{2k^2\sqrt{A}}\eta^2}$$

From the second relation it follows that $A \leqq \dfrac{1}{n \log 2k}$, and from the last relation it can be seen that

$$(1) \qquad 1 \geqq \frac{4}{A} e^{-\frac{\sqrt{5}}{4\sqrt{A}}} \quad \text{for } k \geqq 2 \text{ and } \eta^2 < 2\,,$$

$$(2) \qquad 1 \geqq \frac{4}{A} e^{-\frac{\sqrt{5}}{8\sqrt{A}}} \quad \text{for } \eta^2 \leqq 1\,,$$

$$(3) \qquad 1 \geqq \frac{6}{A} e^{-\frac{\sqrt{11}}{9\sqrt{A}}} \quad \text{for } k \geqq 3\,.$$

Now, when $1 \geqq A \geqq 1/100$, (1) does not hold, when $1 \geqq A \geqq 1/500$ (2) does not hold and when $1 \geqq A \geqq 1/400$ (B) does not hold. Therefore, (III) is always preferable to (II) for $\alpha \geqq 0.01$, that is, when A or C can be greater than or equal to 0.01, although at least one of them remains less than α. (III) is preferable to (II) for $\eta^2 \leqq 1$, $\alpha \geqq 0.002$ or for $\alpha \geqq 0.0025$ when $k \geqq 3$. Even when $\alpha < 0.01$, $A \geqq C$ does not necessarily hold. As k becomes larger and η smaller, the case $A \leqq C$ happens more frequently. For example, when $\eta^2 < 1/2$, we have $A \leqq C$ for $\alpha > 1/5000$. On the other hand, when we fix k and η, and make n large, (accordingly α small), we have $A \geqq C$. For example, when $k \leqq 10$ and $\eta \geqq 0.2$, we have always $A \geqq C$ for $\alpha \leqq 2.5 \times 10^{-5}$.

4. *Application.* Let ω be a set of distributions which are defined on the same k events. Further, let F_0 be the distribution of the random variable defined on the same events under observation, and δ_n the empirical distribution on n observations of the variable. The problem then is to decide whether F_0 is contained in ω or F_0 lies apart by ε (>0) from ω. This problem has been treated in various forms in [1, 2]. The decision is, however, made more precisely in a wide class of cases by

employing (III) than by employing (I) or (II). This will be illustrated by the examples below.*

Let d denote $\inf_{F\in\omega} ||F-S_n||$. Then we have:

	n	k	d^2	$\frac{k-1}{nd^2}$	$\frac{k^2+k-1}{n^2d^4}$
1	253	9	0.0214	1.47	3.03
2	497	9	0.0699	0.2302	0.0737
3	300	9	0.410	0.656	0.00588
4	300	9	0.238	0.2738	0.0788
5	300	9	0.426	0.0626	0.00545
6	300	6	0.294	0.057	0.00398
7	300	6	0.032	0.521	0.0336
8	300	6	0.416	0.040	0.00199
9	270	9	0.296	0.100	0.0114
10	270	9	0.178	0.166	0.0316
11	270	9	0.238	0.124	0.0177
12	270	9	0.402	0.074	0.0062

Thus, we can decide with risk 0.05 that F_0 is contained in ω in examples 1, 2, 4, and F_0 is not contained in ω, that is, lies by $2\sqrt[4]{\frac{k^2+k-1}{n^2(0.05)^2}}$ apart from ω in examples 3, 5, 6, 7, 8, 9, 10, 11 and 12. For the above examples it can be seen at a glance that (III) is more precise than (I). It can also be seen easily that (III) is more precise than (II) for the above examples.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCE

[1] Matsusita, Kameo, Decision rules, based on the distance, for the problems of fit, two samples, and estimation, *Ann. Math. Stat.*, Vol, 26 (1955), pp. 631-640.

[2] Matsusita, Kameo and Hirotugu AKAIKE: Decision rules, based on the distance, for the problems of independence, invariance and two samples, *Ann. Inst. Stat. Math.*, Vol. VII (1956), pp. 67-80.

* As to these examples see the examples in [2].