On the Estimation by the Minimum Distance Method

By Kameo MATUSITA

(Received January 15, 1954)

1. As to the problem of estimation Wolfowitz ([1], [2]) showed that the minimum distance method provides a (super-) consistent estimate. He employed there

$$\delta(F_1, F_2)=\sup_x |F_1(x)-F_2(x)|$$

as the distance between two distribution functions F_1, F_2, but stressed that his method can be applied with very many definitions of distance and is in no way tied to any paticular definition of distance.

In this paper we shall give a probabilistic inequality which holds for a wide class of distributions, and show that the minimum distance method with the definition of distance given in [3] a'so provides a super-consistent estimate, i.e., an estimate which converges with probability one to the parameter concerned. Confidence intervall will also be referred to.

Our distance between two distributions F_1, F_2 is given by

$$|| F_1-F_2 || = \left\{\int_R (\sqrt{p_1(x)}-\sqrt{p_2(x)})^2 dm \right\}^{\frac{1}{2}}$$

when the probabilities of a set E by F_1 and F_2 are expressed by the aid of a suitable measure m as

$$F_1(E)=\int_E p_1(x)dm$$

$$F_2(E)=\int_E p_2(x)dm$$

respectively, and R denotes the whole sample space. This distance is invariant under any transformations of random variables which has the continuous derivative in the continuous case, and also under any orthogonal transformation of the square root system of probabilities in the discrete case, so it can be used for non-parametric problems. Further, in the discrete case, the distance between the theoretical and the empirical distribution is closely connected with the chi-square, and with our results it is established that the minimum chi-square method provides a super-consistent estimate. The same can be said concerning another related distance, for example, a distance like the Euclidean (see below).

First, we shall consider a parameter of one distribution, and second, a "structural" parameter involved in many distributions. Finally, in connection with the convergence of estimate, we shall refer to metric in the parameter space. Throughout this paper we are concerned with finite discrete distributions, since the other cases can be reduced to that case by, for instance, subdividing the whole region of definition of distribution functions into a finite number of regions appropriately.

2. Let F be a discrete distribution with probabilities $p_1, p_2, \cdots, p_k$ for the events (1), (2), $\cdots$, (k), respectively. Let n_i be a number of occurences of event (i) in n observations. We denote the empirical distribution $\left(\frac{n_1}{n}, \cdots, \frac{n_k}{n}\right)$ by S_n. Then our distance between F and S_n is given by

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

THEOREM I. *When the random variable concerned has a distribution F and each p_i of F is positive, then we have*

$$P_r(||F-S_n||^2>\eta)\leqq\frac{1}{n^2\eta^2}\left\{k^2-1+\frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{p_i}-k^2-2k+2\right)\right\}$$

for any positive number η.

Proof. Put

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$$

and we have clearly

$$||F-S_n||^2\leqq\frac{\chi^2}{n}$$

consequently

$$P_r\{||F-S_n||^2>\eta\}\leqq P_r\left\{\frac{1}{n}\chi^2>\eta\right\}\leqq\frac{1}{n^2\eta^2}E(\chi^2)^2$$
$$=\frac{1}{n^2\eta^2}\left\{k^2-1+\frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{p_i}-k^2-2k+2\right)\right\}$$

When we apply an inequality of Wald (see [4]) to χ^2 in the above proof, we obtain

THEOREM I'. *When* $n\eta>(k+1)+\frac{1}{(k-1)n}\left\{\sum_{i=1}^{k}\frac{1}{p_i}-k^2-2k+2\right\}$, *we have*

$$P_r\{||F-S_n||^2>\eta\}\leqq\left(1+\frac{(n\eta-k+1)^2}{2(k-1)+\frac{1}{n}\left(\frac{k}{p_0}-k^2-2k+2\right)}\right)^{-1}$$

Applying these theorems, we can obtain confidence intervals for the parameters involved in distribution F or test a hypothesis concerning the parameters. In this paper, however, we shall confine ourselves to the problem of estimation, especially that of point estimation, and for this purpose theorem 1 alone will do.

THEOREM II. *Under the same condition as in THEOREM I we have*

$$P_r\{\lim_{n\to\infty} \| F-S_n \| = 0\} = 1$$

Let θ be a parameter involved in the distribution F and R a set within which θ is known to lie. θ may be a vector, but for the sake of brevity we assume that θ is a scalar.

In the following use will be made of

Condition (B): For any θ in R, we have

$$p_i(\theta) \geqq p_0 > 0 \qquad (i=1, \cdots, k)$$

When this condition is satisfied, we set

$$G = k^2 + 1 + \frac{k}{p_0}$$

Then we have clearly

THEOREM III. *Under condition (B) we have*

$$P_r(\| F(\theta) - S_n \|^2 > \eta) \leqq \frac{1}{n^2 \eta^2} G$$

for any θ in R, and

$$P_r\{\min_\theta \| F(\theta) - S_n \|^2 < \eta\} \geqq 1 - \frac{1}{n^2\eta^2} G$$

From this theorem and the Borel-Cantelli lemma follows

THEOREM IV. *We have*

$$P_r\{\lim_{n\to\infty} \min_\theta \| F(\theta) - S_n \| = 0\} = 1$$

Therefore, the minimum distance-method with respect to our distance gives a super-consistent estimate of θ, i. e., $\min_\theta \| F-S_n \|$ estimate of θ converges to θ with probability one, when θ depends continuously on F. Of course, we consider here the convergence of estimates in the ordinary sense, i. e., the convergence as real numbers. The convergence in the sense of another metric will be referred to in 4.

From our result now established it is easily seen that the minimum chi-square method also provides a super-consistent estimate, when condition (B) holds. For, concerning χ^2 defined in the proof of theorem 1, we have

$$\frac{p_0}{4n}\chi^2 \leqq \| F-S_n \|^2 \leqq \frac{1}{n}\chi^2$$

As to the distance given by

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

we have

$$\frac{\delta_e}{2} \leqq \| F - S_n \| \leqq \frac{\delta_e}{\sqrt{p_0}}$$

under condition (B). Therefore, the δ_e minimum method gives a super-consistent estimate, too.

3. Now, let us turn our attention to the problem of estimating the structural parameter which was treated in [1], [2], [5].

Suppose, we make a set of n observations at l steps and n_i denotes a number of observations at the i-th step. Let $F_i(\theta)$ be a discrete distribution with probabilities $p_{i1}(\theta)$, $p_{i2}(\theta)$, $p_{ik}(\theta)$ for the events (1), (2), $\cdots$, (k). Further, for each i let n_{ij} be a number of occurences of event (j) in the i-th step, and represent the empirical distribution $\left(\frac{n_{i1}}{n_i}, \cdots, \frac{n_{ik}}{n_i}\right)$ by $S_{ni}^{(i)}$.

As generalizations of distance $\| \quad \|$ we introduce two quantities. First, denote the systems $(F_1(\theta), F_2(\theta), \cdots, F_l(\theta))$ and $(S_{n1}^{(1)}, \cdots, S_{nl}^{(l)})$ by $F^{*(l)}$ and $S^{*(l)}$, respectively, and put

$$\| F^{*(l)} - S^{*(l)} \| = \left\{\frac{1}{n}\sum_{i=1}^{l} n_i \sum_{j=1}^{k}\left(\sqrt{p_{ij}} - \sqrt{\frac{n_{ij}}{n_i}}\right)^2\right\}^{\frac{1}{2}}$$

i. e.,

$$\| F^{*(l)} - S^{*(l)} \|^2 = \frac{1}{n}\sum_{i=1}^{l} n_i \| F_i - S_{ni}^{(i)} \|^2$$

Second, define

$$F^{(n)} = \left\{\frac{\sum_{i=1}^{l} n_i p_{i1}}{n}, \frac{\sum_{i=1}^{l} n_i p_{i2}}{n}, \cdots, \frac{\sum_{i=1}^{l} n_i p_{ik}}{n}\right\}$$

$$S^{(n)} = \left\{\frac{\sum_{i=1}^{l} n_i \frac{n_{i1}}{n_i}}{n}, \frac{\sum_{i=1}^{l} n_i \frac{n_{i2}}{n_i}}{n}, \cdots, \frac{\sum_{i=1}^{l} n_i \frac{n_{ik}}{n_i}}{n}\right\}$$

(See [2]). Then we obtain

$$\| F^{(n)} - S^{(n)} \|^2 = \sum_{j=1}^{k}\left\{\sqrt{\sum_{i=1}^{l} n_i p_{ij}\Big/n} - \sqrt{\sum_{i=1}^{l} n_i \frac{n_{ij}}{n_i}\Big/n}\right\}^2$$

$$\leqq \frac{1}{n}\sum_{i=1}^{l} n_i \sum_{j=1}^{k}\left\{\sqrt{p_{ij}} - \sqrt{\frac{n_{ij}}{n_i}}\right\}^2$$

$$=\frac{1}{n}\sum_{i=1}^{l} n_i \| F_i - S_{ni}^{(i)} \|^2 = \| F^{*(l)} - S^{*(l)} \|$$

Now, assume that condition (B) is satisfied for all F_i. It holds then

$$E(\| F_i - S_{ni}^{(i)} \|^4) \leqq \frac{1}{n_i^2} E\left\{\sum_{j=1}^{k} \frac{(n_i p_{ij} - n_{ij})^2}{n_i p_{ij}}\right\}^2$$

$$= \frac{1}{n_i^2}\left\{(k^2-1) + \frac{1}{n_i}\left(\sum_{j=1}^{k}\frac{1}{p_{ij}} - k^2 - 2k + 2\right)\right\}$$

$$\leqq k^2 + 1 + \frac{1}{n_i}\sum_{i=1}^{k}\frac{1}{p_{ij}}$$

$$\leqq k^2 + 1 + \frac{k}{p_0} = G$$

$$(i=1,\ 2,\ \cdots,\ l)$$

On the other hand, we have

$$\| F^{*(l)} - S^{*(l)} \|^8 \leqq \frac{8}{n^4}\sum_{i=1}^{l} n_i^4 \| F_i - S_{ni}^{(i)} \|^8$$

$$\leqq \frac{32}{n^4}\sum_{i=1}^{l} n_i^4 \| F_i - S_{ni}^{(i)} \|^4$$

(according to $\| F_i - S_{ni}^{(i)} \|^2 \leqq 2$)

hence,

$$E(\| F^{*(l)} - S^{*(l)} \|^8) \leqq \frac{32}{n^4}\sum_{i=1}^{l} n_i^4 E(\| F_i - S_{ni}^{(i)} \|^4)$$

$$\leqq \frac{32G}{n^4}\sum_{i=1}^{l} n_i^2 \leqq \frac{32G}{n^2}$$

Since for any positive number η it holds

$$P\{\| F^{*(l)} - S^{*(l)} \|^2 > \eta\} \leqq \frac{1}{\eta^4} E(\| F^{*(l)} - S^{*(l)} \|^8)$$

we have the following

THEOREM V. *Under condition (B) we have*

$$P_r\{\| F^{*(l)} - S^{*(l)} \|^2 > \eta\} \leqq \frac{32G}{n^2\eta^2}$$

for any positive number η.

From this theorem and relation $\| F^{(n)} - S^{(n)} \| \leqq \| F^{*(l)} - S^{*(l)} \|$ we obtain

THEOREM VI. *Under condition (B) we have*

$$P_r\{\| F^{(n)} - S^{(n)} \| > \eta\} \leqq \frac{32G}{n^2\eta^2}$$

for any positive number η.

These theorems can be made more precise for large n (cf. theorem I′), and all these enable us to obtain a confidence interval for the structural parameter θ. Of course, we assume that θ depends continuously on each F_i when we consider the convergence in the ordinary sense. We also obtain by the aid of the Borel-Cantelli lemma

THEOREM VII. *Under condition* (B) *we have*

$$Pr\{\lim_{n\to\infty} \| F^{*(l)} - S^{*(l)} \| = 0\} = 1$$

THEOREM VIII. *Under condition* (B) *we have*

$$Pr\{\lim_{n\to\infty} \| F^{(n)} - S^{(n)} \| = 0\} = 1$$

Therefore, the $\min \| F^{*(l)} - S^{*(l)} \|$ estimate and the $\min \| F^{(n)} - S^{(n)} \|$ estimate of structural parameter θ are super-consistent in the ordinary sense, respectively, when θ has some continuity property with respect to $\{F_i\}$, for example, when $\{F_i, n_i\}$ has the asymptotic positive distance property as defined in [2].

Now, as related quantities with $\| F^{*(l)} - S^{*(l)} \|^2$ or $\| F^{(n)} - S^{(n)} \|^2$, define

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(n_i p_{ij} - n_{ij})^2}{n p_{ij}}$$

and

$$\delta_e^2(F^{*(n)}, S^{*(n)}) = \sum_{i=1}^{l} \frac{n_i}{n} \sum_{j=1}^{k} \left(p_{ij} - \frac{n_{ij}}{n_i} \right)^2$$

Then we have clearly

$$\frac{p_0}{4n} \chi^2 \leqq \| F^{*(l)} - S^{*(l)} \|^2 \leqq \frac{1}{n} \chi^2$$

$$\frac{1}{4} \delta_e^2 (F^{*(n)}, S^{*(n)}) \leqq \| F^{*(l)} - S^{*(l)} \| \leqq \frac{1}{p_0} \delta_e^2(F^{*(n)}, S^{*(n)})$$

provided that condition (B) is satisfied. From these inequalities and theorem VII follows that under condition (B) the minimum χ^2 method and the minimum δ_e^2 method give super-consistent estimates of the structural parameter θ, respectively. (As regards only the consistency of the minimum χ^2 estimate, it was shown by Neyman [6] without condition (B).)

4. Finally we want to make a remark about metric in the parameter space. The distance between parameters thus far used in most literatures is the Euclidean, i.e., the parameter space is considered to be or be embedded in a Euclidean space. If we, however, consider the parameter always with the distribution and its rôle in the distribution, it would be natural to introduce a metric defined by the distribution into the parameter space, for example, a metric

in the space of distributions. In our case, then, a distance between parameters θ_1, θ_2 involved in distribution F is defined as $||F(\theta_1)-F(\theta_2)||$. With this definition of distance in the parameter space, we do not need any continuity assumption of the parameter on the distribution such as mentioned above, when we want to get confidence intervals or (super-) consistent estimates. Further, the efficiency of estimate by the minimum $||\ ||$ method can be considered by $E(||F-S_n||^2)\left(\leqq\frac{k-1}{n}\right)$.

INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Wolfowitz, J.: Consistent estimators of the parameters of a linear structural relation, *Skand. Aktuarietidskrift*, 1952.

[2] Wolfowitz, J.: Estimation by the minimum distance method, *Annals of the Institute of Statistical Mathematics*, Vol. 5, 1953.

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

[4] Wald, A.: A generalization of Markoff's inequality, *Annals of Mathematical Statistics*, Vol. IX, 1938.

[5] Neyman, J. and E. L. Scott: Consistent estimates based on partially consistent observations, *Econometrika*, Vol. 16, 1948.

[6] Neyman, J.: Contribution to the theory of the χ^2 test, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, 1949.

A REMARK TO "ON THE ESTIMATION BY THE MINIMUM DISTANCE METHOD"

BY KAMEO MATUSITA

(Received June 15, 1954)

In the above-named paper (*this Annals* Vol. V), we can prove theorems IV, VII, VIII and the statements concerning δ_e without condition (B). The proofs are very easy.

As to theorem IV, for instance, we have clearly

$$\min_{\theta} \| F(\theta)-S_n \| \leqq \| F(\theta_0)-S_n \|$$

for any θ_0. Now, suppose, θ_0 is the value of the parameter which we want to estimate. Then, by theorem II we have

$$Pr\{\lim_{n\to\infty} \| F(\theta_0)-S_n \| =0\} =1$$

consequently

$$Pr\{\lim_{n\to\infty} \min_{\theta} \| F(\theta)-S_n \| =0\} =1$$

I owe this remark to Professor J. Wolfowitz and wish to thank him.

THE INSTITUTE OF STATISTICAL MATHEMATICS

ERRATA

This Annals Vol. V, No. 2

Page 60, line 9, read "the number of ..." instead of "a number of ..."

Page 63, line 5, read $\frac{1}{n_i^2}\left\{k^2+1+\frac{1}{n_i}\sum_{j=1}^{k}\frac{1}{p_{ij}}\right\}$

instead of $k^2+1+\frac{1}{n_i}\sum_{i=1}^{k}\frac{1}{p_{ij}}$

Page 63, line 6, read $\frac{1}{n_i^2}\left(k^2+1+\frac{k}{p_0}\right)=\frac{G}{n_i^2}$

instead of $k^2+1+\frac{k}{y_0}=G$

Page 80, line 20, read $\int_x^y e^{-\frac{t^2}{2}}\,dt-\left(\int_x^y te^{-\frac{t^2}{2}}\,dt\right)^2\Big\}\,/$

instead of $\int_x^y e^{-\frac{t^2}{2}}\,dt\left(\int_x^y te^{-\frac{t^2}{2}}\right)^2\Big\}\,/$

Page 86, line 12, read $2\ e(y)\,/\left(\int_x^y\right)^2$

instead of $2\left(e(y)\,/\int_x^y\right)^2$

Page 87, line 2, read $y<0$ instead of $y>0$

Page 99, line 1, read $+\frac{1}{y}\int_y^z (y-t)^2\int_y^z t^2-e(z)$

instead of $+\frac{1}{y}\int_y^x (y-t)^2\int_y^z -e(z$

Page 90, line 18, read $\int_y^z (y-t)\int_y^z t^2-(z-y)^2e(z)$

instead of $\int_y^z (y-t)-(z-y)^2\,e(z)$

ERRATA

These Annals Vol. V, No. 2

Page 63, line 9,

read "$\leqq \frac{27}{n^4} \sum_{i=1}^{l} n \cdots$" instead of "$\leqq \frac{8}{n^4} \sum_{i=1}^{l} \cdots$"

Page 63, line 10, 13, 14, 19, 23,

read "108" instead of "32"

Vol. VII, No. 2

Page 117, line 10,

read "$X_i^{(2)} = X_i^{(1)} \left(1 + \frac{X_i' - \frac{Y_i'}{b_{ii}}}{X_i'} \cdot \cdots \right.$"

instead of "$X_i^{(2)} = X_i^{(1)} \left(1 + \frac{X_i' - \frac{Y_i'}{b_{ii}}}{X_i} \cdot \cdots \right.$"

Page 121, line 2 from bottom,

read "12.713", instead of "2.713"

Vol. VII, No. 3

Page 147, line 2 & 3,

read "$\eta_u = \frac{\sinh^{-1}\left[\frac{1}{2} \cdots\right.}{\alpha'(n+1/\sigma^2)}$" instead of "$-\eta_L = \frac{\sin h^{-1}\left[\frac{1}{2}(\cdots\right.}{\alpha'(n+1/\sigma^2)}$"

Page 151, line 6 from bottom,

read "$P_{\theta'}\{S_n \text{ covers } \theta\}$ is$\cdots$"

instead of "$P_{\theta}\{S_n \text{ covers } \theta'\}$ is$\cdots$"

Vol. VIII, No. 1

Page 59, line 9 from bottom,

read "0.01234, 0.04344, 0.12803, 0.28807, 0.52812"

instead of "0.52812, 0.28807, 0.12803, 0.04344, 0.01234"