ON THE ESTIMATION OF THE POPULATION MEAN BASED ON ORDERED SAMPLES FROM AN EQUICORRELATED MULTIVARIATE DISTRIBUTION

KOITI TAKAHASI

(Received Feb. 7, 1969; revised June 26, 1969)

1. Introduction

In a previous paper [1] the author has studied some properties of an unbiased estimator of the population mean based on ordered samples. The estimator was defined as follows. Let $f(x)$ be a probability density function (pdf) with mean μ and variance σ^2 and let $\{X_{ij}^*,\ j=1,\cdots,n;\ i=1,\cdots,n\}$ be a random sample of size n^2 from the pdf $f(x)$ divided into n groups of size n. Let Y_{ni} denote the ith order statistic within the ith subgroup $\{X_{i1}^*,\cdots,X_{in}^*\}$. Then $\bar{Y}_{[n]}=(Y_{n1}+\cdots+Y_{nn})/n$ is an unbiased estimator of μ. A sampling and estimation procedure like this will be useful in the situation where the selection of the element of the ith least value from among n elements can be done without measuring the values of the n elements, for example, merely by taking a glance at the elements. If n is small and n elements are located not so distantly from each other, such situation will be of frequent occurrence. However, a trouble occurs in the practical application of our estimation: If the n elements are located closely to each other, they can not necessarily be considered as a random sample of size n from the population. They may be correlated positively or negatively. Thus it becomes necessary or at least desirable to construct a model concerned with the case of dependence and investigate properties of the estimator corresponding to $\bar{Y}_{[n]}$ in the case of independence. The purpose of this paper is to deal with this problem. A model and an estimator are presented in section 2. In section 3 several properties of the estimator are shown. The efficiencies of the estimator for some particular distributions are given in section 4.

2. Model and estimator

Let $F_m(x_1,\cdots,x_m)$ be a cdf which is symmetric in $x_1,x_2,\cdots,x_m$ and has the pdf $f_m(x_1, \cdots, x_m)$, and let $F_n(x_1, \cdots, x_n)$ and $f_n(x_1, \cdots, x_n)$ be the cdf and the pdf of the n-dimensional marginal distribution of the $F_m(x_1, \cdots, x_m)$, respectively, for $n=1, 2, \cdots, m-1$. Notice that every F_n is also symmetric in its arguments. Denote $F_1(x)$ and $f_1(x)$ merely by $F(x)$ and $f(x)$, respectively.

Typical examples are the normal distribution with the same intraclass correlation coefficients and the symmetric mixture of independently and identically distributed random variables. (See section 4.)

Let $X_n^* = (X_{n,1}^*, \cdots, X_{n,n}^*)$ be a random vector from the cdf $F_n(x_1, \cdots, x_n)$, and let $X_n = (X_{n,1}, \cdots, X_{n,n})$ be the one obtained by rearranging the components of X_n^* in increasing order of magnitude. We denote the marginal cdf and pdf of $X_{n,i}$ by $F_{n,i}(x)$ and $f_{n,i}(x)$ respectively, and a random vector with the cdf $F_{n,1}(x_1) \cdot F_{n,2}(x_2) \cdots F_{n,n}(x_n)$ by $(Y_{n,1}, Y_{n,2}, \cdots, Y_{n,n})$. Then the statistic

$$\bar{Y}_{[n]} = \sum_{i=1}^{n} Y_{n,i}/n$$

is an unbiased estimator of the mean μ of $F(x)$, since we have the relation

$$\sum_{k=1}^{n} f_{n,k}(x) = nf(x) . \tag{1}$$

The estimator $\bar{Y}_{[n]}$ was considered in [1] for the case $F_n(x_1, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$. Throughout this paper, this case will be called 'the case of independence'.

3. Properties of the estimator

Let $\mu_{n,k}$ and $\sigma_{n,k}^2$ denote the mean and variance of $X_{n,k}$. We use the symbol 'tilder' for representing 'the case of independence'; for example, $\tilde{\mu}_{n,k}$ and $\tilde{\sigma}_{n,k}^2$ denote the mean and variance of $X_{n,k}$ for the case of independence where $(X_{n,1}^*, \cdots, X_{n,n}^*)$ has the joint cdf $\prod_{i=1}^{n} F(x_i)$.

From the symmetry of $F_n(x_1, \cdots, x_n)$ and (1) we have the recurrence relation between the $F_{n,i}$'s ([2], [3])

$$\frac{n+1-i}{n+1} F_{n+1,i} + \frac{i}{n+1} F_{n+1,i+1} = F_{n,i} .$$

Thus, as in the proof of Theorem 2 in [1] we have

$$\sigma_{[n]}^2 > \sigma_{[n+1]}^2 \qquad \text{for} \quad 1 \leqq n \leqq m-1 , \tag{2}$$

where $\sigma_{[n]}^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_{n,i}^2$. The variance of $\bar{Y}_{[n]}$ is $\sigma_{[n]}^2/n$.

It follows from (2) that

$$\frac{\sigma^2}{n} > \sigma^2(\bar{Y}_{[n]}),$$

that is, the variance of estimator $\bar{Y}_{[n]}$ is smaller than that of the mean of a random sample of size n from the distribution with the cdf $F(x)$. As in the case of independence, the relative efficiency $\tau_{[n]}$ of $\bar{Y}_{[n]}$ with respect to the mean $\bar{X}_n$ of a random sample of size n, is defined by

$$(3) \qquad \tau_{[n]} = \frac{\dfrac{\sigma^2}{n} - \dfrac{\sigma^2_{[n]}}{n}}{\dfrac{\sigma^2}{n}} = \frac{\sigma^2 - \sigma^2_{[n]}}{\sigma^2}$$

(although $\sigma^2/\sigma^2_{[n]}$ might be a more familiar measure of efficiency).

In terms of the efficiency $\tau_{[n]}$ we can write the relation (3) as

$$\tau_{[n]} < \tau_{[n+1]} \qquad \text{for } n = 1, 2, \cdots, m-1.$$

In particular, we have, for $n \geqq 2$,

$$\tau_{[n]} > 0.$$

In short, the $\bar{Y}_{[n]}$ is more efficient than the $\bar{X}_n$, no matter whether ρ is positive or negative.

Now we consider the effect of dependency to the efficiency $\tau_{[n]}$. In many cases, it seems intuitively that the positive dependence gives rise to lower efficiency of $\bar{Y}_{[n]}$ and the negative dependence to higher efficiency.

From (1) we have

$$\sigma^2_{[n]} = \sigma^2 + \mu^2 - \frac{1}{n} \sum_{k=1}^{n} \mu^2_{n,k}.$$

Thus

$$(4) \qquad \sigma^2_{[n]} - \tilde{\sigma}^2_{[n]} = \frac{1}{n} \left(\sum_{k=1}^{n} \tilde{\mu}^2_{n,k} - \sum_{k=1}^{n} \mu^2_{n,k} \right).$$

Denote $F_l(x, \cdots, x)$ by $F_l(x)$, in short, for $l = 1, 2, \cdots, m$. We, then, have

$$(5) \qquad F_{n,k}(x) = \sum_{l=k}^{n} c_{n,k,l} F_l(x)$$

where

$$c_{n,k,l} = (-1)^l \sum_{\nu=k}^{l} (-1)^\nu \binom{n}{\nu} \binom{n-\nu}{l-\nu}.$$

If we denote $P(X_{n,i}^*\leqq x,\ i=1,2,\cdots,k$ and $X_{n,j}^*>x,\ j=k+1,\cdots,n)$ by $G_{n,k}(x)$, we immediately have

$$(6)\qquad F_{n,k}(x)=\sum_{l=k}^{n}\binom{n}{l}G_{n,l}(x)$$

and

$$(7)\qquad F_k(x)=\sum_{r=0}^{n-k}\binom{n-k}{r}G_{n,k+r}(x)\,.$$

From (6) and (7), (5) is obtained after some simple calculation. Noting that $\tilde{F}_l(x)=F^l(x)$ for the case of independence, from (4) and (5) we have

$$(8)\qquad \sigma^2_{[n]}-\tilde{\sigma}^2_{[n]}=\frac{1}{n}\sum_{l=1}^{n}\sum_{k=1}^{l}c_{n,k,l}(\mu_{n,k}+\tilde{\mu}_{n,k})\int_{-\infty}^{\infty}(F^l(x)-F_l(x))dx\,.$$

For $n=2$ the expression (8) reduces to the simple form

$$(9)\qquad \sigma^2_{[2]}-\tilde{\sigma}^2_{[2]}=\frac{1}{2}\{(\mu_{2,2}-\mu_{2,1})+(\tilde{\mu}_{2,2}-\tilde{\mu}_{2,1})\}\int_{-\infty}^{\infty}(F^2(x)-F_2(x))dx\,.$$

The expression in the bracket of the right-hand side of (9) is obviously positive. Hence $\sigma^2_{[2]}>,\ =,\ <\tilde{\sigma}^2_{[2]}$ correspond to $\int_{-\infty}^{\infty}(F_2(x,x)-F^2(x))dx <,\ =,\ >0$, respectively. A two dimensional distribution is called positively (or negatively) quadrant dependent ([4]), if

$$P(X\leqq x)P(Y\leqq y)\leqq(\text{or }\geqq)P(X\leqq x,Y\leqq y)\qquad\text{for all } x,\ y.$$

In terms of this concept, we can say as follows: if the cdf $F_2(x_1,x_2)$ is negatively (or positively) quadrant dependent, the efficiency of $\bar{Y}_{[2]}$ is larger (or smaller) than that in the case of independence. For $n\geqq 3$ no simple interpretation of (8) is obtained.

It may be useful to show an example that the correlation coefficient of $X_{n,i}^*$ and $X_{n,j}^*$ does not necessarily determine the efficiency.

Example 1. Here we limit ourselves to the case $m=2$.

(i) Suppose

$$(10)\qquad f_2(x_1,x_2)=\begin{cases} s & \text{if } (x_1,x_2)\in\bigcup_{j=1}^{s}\left\{\left[\frac{j-1}{s},\frac{j}{s}\right)\times\left[\frac{j-1}{s},\frac{j}{s}\right)\right\}\\ 0 & \text{otherwise,}\end{cases}$$

where s is a positive integer. We, then, have

$$\rho=1-\frac{1}{s^2}$$

and

$$\tau_{[2]}=\frac{1}{3s^2}=\frac{1}{3}(1-\rho)=(1-\rho)\tilde{\tau}_{[2]}\,.$$

(ii) Suppose

$$f_2(x_1, x_2)=\begin{cases} 1/\theta & \text{if} \quad (x_1, x_2)\in[0, \theta)\times[0, \theta) \\ 1/(1-\theta) & \text{if} \quad (x_1, x_2)\in[\theta, 1)\times[\theta, 1) \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $0<\theta<1$. We, then, have

$$\rho=3\theta(1-\theta)$$

and

$$\tau_{[2]}=\frac{1}{3}\{1-2\theta(1-\theta)\}^2=\frac{1}{3}\left(1-\frac{2}{3}\rho\right)^2\,.$$

4. Some examples

4.1 *Normal distribution*

Let $f_n(x_1,\cdots, x_n)$ be the pdf of the n-dimensional normal distribution with the mean vector $(\mu, \mu,\cdots, \mu)$ and the variance-covariance matrix (σ_{ij}) where $\sigma_{ii}=\sigma^2$, $\sigma_{ij}=\rho\sigma^2$ $(i\neq j)$. By the result of Owen and Steck [5] we can easily obtain

$$\tau_{[n]}(\rho)=(1-\rho)\tilde{\tau}_{[n]}\,,$$

for $n=2, 3,\cdots$ if $\rho\geqq 0$, and for $n\leqq 1-\dfrac{1}{\rho}$ if $\rho<0$.

4.2 *Mixture*

Let $f_n(x_1,\cdots, x_n)$ be an n-dimensional pdf given by

$$f_n(x_1,\cdots, x_n)=\int_0^\infty g(x_1|\omega)\cdots g(x_n|\omega)dP(\omega)\,, \tag{12}$$

for $n=1, 2,\cdots$, where $g(x|\omega)$ is a pdf and $P(\omega)$ is a cdf in $(0, \infty)$.

i) Let

$$g(x|\omega)=\begin{cases} \omega e^{-\omega x} & \text{if} \quad x>0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$dP(\omega)=\frac{\alpha^p}{\Gamma(p)}\omega^{p-1}e^{-\alpha\omega}d\omega\,.$$

The pdf $f_n(x_1,\cdots,x_n)$ corresponds to a multivariate Burr's distribution [6]. We have, in particular,

$$f(x)=f_1(x)=\frac{p\alpha^p}{(\alpha+x)^{p+1}}\,.$$

For $n=2$ we have ([7], [6], [1])

$$\tau_{[2]}=\frac{p-2}{4p}\,,$$

$$\tilde{\tau}_{[2]}=\frac{p(p-2)}{(2p-1)^2}\,,$$

and

$$\rho=\frac{1}{p}\qquad(p>2)\,.$$

Thus we have

$$\tau_{[2]}=\left(1-\frac{1}{2p}\right)^2\tilde{\tau}_{[2]}=\left(1-\frac{\rho}{2}\right)^2\tilde{\tau}_{[2]}=\frac{1}{4}(1-2\rho)\,.$$

ii) Let

$$g(x|\omega)=\begin{cases} e^{-(x-\omega)} & \text{if}\quad x>\omega>0 \\ 0 & \text{otherwise}\end{cases}$$

and

$$dP(\omega)=\alpha e^{-\alpha\omega}d\omega\,.$$

We, then, have

$$f(x)=\begin{cases} \dfrac{\alpha}{\alpha-1}(e^{-x}-e^{-\alpha x}) & \text{for}\quad \alpha\neq 1 \\ xe^{-x} & \text{for}\quad \alpha=1\,,\end{cases}$$

$$\tau_{[n]}=\frac{\alpha^2}{1+\alpha^2}\left\{\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}+\cdots+\frac{1}{n-(k-1)}+\frac{1}{\alpha}\right)^2-\left(1+\frac{1}{\alpha}\right)^2\right\}\,,$$

in particular,

$$\tau_{[2]}=\frac{\alpha^2}{4(1+\alpha^2)}\,,$$

$$\rho=\frac{1}{1+\alpha^2}$$

and

$$\tau_{[2]}=\frac{1}{4}(1-\rho)\,.$$

Acknowledgement

The author wishes to thank the referee for his suggestions which made the paper easier to follow.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] K. Takahasi and K. Wakimoto, "On unbiased estimates of the population mean based on the sample stratified by means of ordering," *Ann. Inst. Statist. Math.*, 20 (1968), 1-31.

[2] H. A. David and P. C. Joshi, "Recurrence relations between moments of order statistics for exchangeable variates," *Ann. Math. Statist.*, 39 (1968), 272-274.

[3] D. H. Young, "Recurrence relations between the P.D.F.'s of order statistics of dependent variables, and some applications," *Biometrika*, 54 (1967), 283-292.

[4] E. L. Lehmann, "Some concepts of dependence," *Ann. Math. Statist.*, 37 (1966), 1137-1153.

[5] D. B. Owen and G. P. Steck, "Moments of order statistics from the equicorrelated multivariate normal distribution," *Ann. Math. Statist.*, 33 (1962), 1286-1291.

[6] K. Takahasi, "Note on the multivariate Burr's distribution," *Ann. Inst. Statist. Math.*, 17 (1965), 257-260.

[7] K. Takahasi, "Estimation of population mean based on ordered sample and its applications," (in Japanese) *Proc. Inst. Statist. Math.* (to appear).