SOME NONPARAMETRIC CONSISTENT ESTIMATES FROM CENSORED SAMPLES

KOITI TAKAHASI

(Received June 19, 1970)

1. Introduction

In some cases it happens that only the smallest observations in each sample can be obtained (cf. David [2]). When considering the cost or the time of an experiment it also happens in some cases including life testing that an observation of the minimum in a small sample can be obtained more easily than, or, at least, as easily as, an observation of a sample of size one (cf. Takahasi and Wakimoto [7]).

David [2] considered the estimation of means of normal populations from n observations each of which is the minimum of m independent $N(\mu, \sigma^2)$ variates.

In this paper we shall give a method of obtaining nonparametric consistent estimates for some problems of estimation from n observations each of which is the kth least order statistic of a sample of size m from a population. Some properties of the estimates obtained by this method are discussed in Sections 3, 4 and 5 for the problems of estimation of the cumulative distribution, mean and quantile of populations, respectively.

2. Derivation of consistent estimates

Let $F(x)$ be a continuous cumulative distribution function (cdf). We denote by $F_{m,k}(x)$ the cdf of the kth least order statistic of a sample of size m from the distribution with cdf $F(x)$. We have (cf. Hoeffding [3])

$$(1)\qquad \begin{aligned} F_{m,k}(x) &= \sum_{j=k}^{m} \binom{m}{j} F(x)^j [1-F(x)]^{m-j} \\ &= k\binom{m}{k} \int_0^{F(x)} t^{k-1}(1-t)^{m-k} dt \,. \end{aligned}$$

It is easily seen that for any given continuous cdf $F_{m,k}(x)$ there is a unique cdf $F(x)$ which satisfies the relation (1). The key to our method is this fact. Let

$$(2)\qquad \gamma_{m,k}^{-1}(u)=\sum_{j=k}^{m}\binom{m}{j}u^j(1-u)^{m-j}$$
$$=k\binom{m}{k}\int_0^u t^{k-1}(1-t)^{m-k}dt\ .$$

Denote the inverse function of $\gamma_{m,k}^{-1}(u)$ by $\gamma_{m,k}(u)$. Using these notations we can write

$$(3)\qquad F_{m,k}(x)=\gamma_{m,k}^{-1}(F(x))$$

and

$$(4)\qquad F(x)=\gamma_{m,k}(F_{m,k}(x))\ .$$

Denote the empirical cdf of a sample of size n from the $F_{m,k}(x)$ by $F_{m,k,n}^*(x)$. By the central statistical theorem (cf. Loève [4]) $F_{m,k,n}^*(x)$ converges to $F_{m,k}(x)$ uniformly in x with probability 1. From this fact and the uniform continuity of the function $\gamma_{m,k}(u)$, ($u\in[0,1]$), the cdf $\gamma_{m,k}(F_{m,k,n}^*(x))$ converges to $\gamma_{m,k}(F_{m,k}(x))$ with probability 1 as n tends to infinity, that is,

$$(5)\qquad \gamma_{m,k}(F_{m,k,n}^*(x))\xrightarrow{\text{a.s.}}F(x)\ .$$

Thus we have for every bounded continuous function $h(x)$ (cf. [4], p.182)

$$(6)\qquad \int_{-\infty}^{\infty}h(x)d\gamma_{m,k}(F_{m,k,n}^*(x))\to\int_{-\infty}^{\infty}h(x)dF(x)\ ,$$

with probability 1.

THEOREM 1. *Let $h(x)$ be a bounded continuous function, let X_i ($i=1,2,\cdots,n$) be independent and identically distributed with the cdf $F_{m,k}(x)$ which is the cdf of the k-th least order statistic in a sample of size m from the cdf $F(x)$ and let $X_{(j)}$ be the j-th least order statistic of $\{X_i\}$. Define an estimator of $\int_{-\infty}^{\infty}h(x)dF(x)$ based on $(X_1, X_2,\cdots,X_n)$ by*

$$(7)\qquad T_n=\sum_{i=1}^{n}\left\{\gamma_{m,k}\left(\frac{i}{n}\right)-\gamma_{m,k}\left(\frac{i-1}{n}\right)\right\}h(X_{(i)})\ .$$

Then the estimator T_n is (strongly) consistent (cf. Rao [5], p. 281) for $\int_{-\infty}^{\infty}h(x)dF(x)$.

It should be noted that the quantity $\int_{-\infty}^{\infty}h(x)dF(x)$ to be estimated is associated with the distribution with cdf $F(x)$ and on the other hand the sample on which the estimation is based comes from the distribution with cdf $F_{m,k}(x)$.

To calculate the value of T_n for a given sample we require the numerical values of $\gamma_{m,k}(i/n)-\gamma_{m,k}((i-1)/n)$, $i=1, 2, \cdots, n$. Recall that the function $\gamma_{m,k}(u)$ is the inverse function of the polynomial $\gamma_{m,k}^{-1}(u)$ of degree m which is defined by (2). Hence we may use a table of incomplete beta function to calculate $\gamma_{m,k}(i/n)-\gamma_{m,k}((i-1)/n)$. On the other hand it is not so difficult to solve the equation $\gamma_{m,k}^{-1}(x)=i/n$. For the the case $m=2$ we have

$$\gamma_{2,1}(u)=1-\sqrt{1-u} \qquad \text{and} \qquad \gamma_{2,2}(u)=\sqrt{u}\ . \tag{8}$$

3. Estimation of values of cdf

In this section we consider the problem of estimating $F(x_0)$ for any fixed value x_0. Let $J(x; x_0)$ be the indicator function of the set $(-\infty, x_0]$. Substituting $J(x; x_0)$ for $h(x)$ in (7) we have

$$A_n=\sum_{i=1}^{n}\left\{\gamma_{m,k}\left(\frac{i}{n}\right)-\gamma_{m,k}\left(\frac{i-1}{n}\right)\right\}J(X_{(i)}; x_0)=\gamma_{m,k}\left(\frac{s}{n}\right), \tag{9}$$

where s is the number of X_i which does not exceed x_0. Since $\int_{-\infty}^{\infty} J(x; x_0)dF(x)=F(x_0)$ we obtain the following corollary.

COROLLARY 1. *The estimator A_n given by* (9) *is* (*strongly*) *consistent for $F(x_0)$.*

Since $F(x)$ is assumed to be continuous the discontinuity of $J(x; x_0)$ at $x=x_0$ makes no problem in using Theorem 1 to prove this corollary.

The random variable s in (9) has the binomial distribution $B(n, F_{m,k}(x_0))$. Therefore we have the expressions for the moments of A_n

$$\mathrm{E}(A_n^\nu)=\sum_{i=0}^{n}\binom{n}{i}(F_{m,k}(x_0))^i(1-F_{m,k}(x_0))^{n-i}\gamma_{m,k}^\nu\left(\frac{i}{n}\right), \qquad (\nu=1, 2, \cdots)\ . \tag{10}$$

Since the function $\gamma_{m,k}$ is bounded and continuous on [0, 1] we have from a theorem on Bernstein polynomials (cf. Rivlin [6])

$$\begin{aligned}\lim_{n\to\infty} \mathrm{E}(A_n)&=\gamma_{m,k}(F_{m,k}(x_0))\\ &=F(x_0)=p\ , \qquad \text{say.}\end{aligned} \tag{11}$$

This implies that the bias of A_n tends to 0 as n increases. Using the Tailor expansion about the point $u=F_{m,k}(x_0)$ and noting that $\gamma_{m,k}'(\gamma_{m,k}^{-1}(p))=1/\gamma_{m,k}^{-1\prime}(p)$ we have, for large n, approximately

$$\begin{aligned}\mathrm{E}(A_n)-p&\fallingdotseq F_{m,k}(x_0)(1-F_{m,k}(x_0))\gamma_{m,k}''(F_{m,k}(x_0))/(2n)\\ &=\gamma_{m,k}^{-1}(p)(1-\gamma_{m,k}^{-1}(p))\gamma_{m,k}''(\gamma_{m,k}^{-1}(p))/(2n)\end{aligned} \tag{12}$$

and

$$\begin{aligned}(13)\qquad \mathrm{E}(A_n-F(x_0))^2 &\doteqdot F_{m,k}(x_0)(1-F_{m,k}(x_0))\gamma'^2_{m,k}(F_{m,k}(x_0))/n \\ &=\gamma^{-1}_{m,k}(p)(1-\gamma^{-1}_{m,k}(p))\gamma'^2_{m,k}(\gamma^{-1}_{m,k}(p))/n \\ &=\gamma^{-1}_{m,k}(p)(1-\gamma^{-1}_{m,k}(p))/\{n(\gamma^{-1\prime}_{m,k}(p))^2\}\,.\end{aligned}$$

From (13) the condition for which the mean square error of A_n would be smaller than that of the usual estimate based on a sample of size n from $F(x)$ is approximately

$$(14)\qquad \gamma^{-1}_{m,k}(p)(1-\gamma^{-1}_{m,k}(p))/\{p(1-p)(\gamma^{-1\prime}_{m,k}(p))^2\}<1\,.$$

For example, in the case $k=1$, the condition (14) reduces to

$$(15)\qquad m^2q^{m-1}-(m^2-1)q^m-1>0\,,$$

where $q=1-p$.

It may be expected that the mean square error of A_n would be smaller than $p(1-p)/n$ provided that $F(x_0)$ is close to $k/(m+1)$. From (2) we have

$$\gamma^{-1}_{m,k}(p)=\sum_{j=k}^{m}\binom{m}{j}p^j(1-p)^{m-j}\,,$$

$$1-\gamma^{-1}_{m,k}(p)=\sum_{j=0}^{k-1}\binom{m}{j}p^j(1-p)^{m-j}$$

and

$$\gamma^{-1\prime}_{m,k}(p)=k\binom{m}{k}p^{k-1}(1-p)^{m-k}\,.$$

The denominator of the left-hand side of (14) can be written as

$$\begin{aligned}(\gamma^{-1\prime}_{m,k}(p))^2p(1-p)=&\left\{k\binom{m}{k}p^k(1-p)^{m-k}\right\}\\ &\cdot\left\{(m-k+1)\binom{m}{m-k+1}p^{k-1}(1-p)^{m-k+1}\right\}.\end{aligned}$$

Assume that $p=k/(m+1)$. Then it is easily seen that

$$\max_{0\leqq j\leqq k-1}\binom{m}{j}p^j(1-p)^{m-j}\leqq\binom{m}{k}p^k(1-p)^{m-k}$$

and

$$\max_{k\leqq j\leqq m}\binom{m}{j}p^j(1-p)^{m-j}\leqq\binom{m}{m-k+1}p^{k-1}(1-p)^{m-k+1}\,.$$

Therefore we have

$$\sum_{j=k}^{m}\binom{m}{j}p^j(1-p)^{m-j}\leqq(m-k+1)\binom{m}{m-k+1}p^{k-1}(1-p)^{m-k+1}$$

and

$$\sum_{j=0}^{k-1}\binom{m}{j}p^j(1-p)^{m-j}\leqq k\binom{m}{k}p^k(1-p)^{m-k}\,.$$

Thus we have proved that the inequality (14) holds if $k/(m+1)=F(x_0)$.

4. Estimation of means

In this section we shall treat the problem of estimation of means. From Theorem 1 we have the following corollary.

COROLLARY 2. *Assume that the support of $F(x)$ is bounded. Then,*

$$B_n=\sum_{i=1}^{n}\left\{\gamma_{m,k}\left(\frac{i}{n}\right)-\gamma_{m,k}\left(\frac{i-1}{n}\right)\right\}X_{(i)} \tag{16}$$

is a (strong) consistent estimate of the mean of $F(x)$.

In order to apply Theorem 1 to the estimation of means we assumed that the cdf $F(x)$ has a bounded support. We can, however, show the consistency of B_n for some special cases where the support of $F(x)$ is not bounded. Assume that $F(x)=1-\exp(-x/\theta)$ and $k=1$. In this case $F_{m,1}(x)$ is also the exponential distribution with the scale parameter θ/m. Therefore, we can express the mean and the variance of B_n explicitly. It is well known that

$$\mathrm{E}(X_{(i)})=\frac{\theta}{m}\sum_{j=1}^{i}\frac{1}{n+1-j} \tag{17}$$

and

$$\underset{i\leq j}{\mathrm{Cov}}\,(X_{(i)},\,X_{(j)})=\frac{\theta^2}{m^2}\sum_{j=1}^{i}\frac{1}{(n+1-j)^2}\,. \tag{18}$$

From these expressions we have

$$\mathrm{E}(B_n)=\frac{\theta}{m}n^{-1}\sum_{h=1}^{n}\left(1-\frac{h-1}{n}\right)^{-1+(1/m)} \tag{19}$$

and

$$\begin{aligned}\mathrm{Var}\,B_n&=\left(\frac{\theta}{m}\right)^2 n^{-2/m}\sum_{j=1}^{n}(n+1-j)^{(2/m)-2}\\&=\left(\frac{\theta}{m}\right)^2 n^{-2/m}\sum_{j=1}^{n}j^{(2/m)-2}\,.\end{aligned} \tag{20}$$

It follows that

$$\lim_{n\to\infty}\mathrm{E}(B_n)=\theta \tag{21}$$

and

$$\lim_{n\to\infty} \operatorname{Var} B_n = 0 . \tag{22}$$

Thus the estimator B_n is (weakly) consistent in this case. Further assume that $m=2$. Then, from Theorem 1 of [1] it follows that the cdf of $n^{1/2}(B_n - EB_n)/\sqrt{\operatorname{Var} B_n}$ converges to that of the standard normal distribution.

5. Estimation of quantiles

In this section we assume that the support of $F(x)$ is an interval (it may be infinite) and $F(x)$ is strictly increasing on the interval. Then, the pth quantile of $F(x)$ is unique for each p, $0<p<1$. It follows from (1) that the cdf $F_{m,k}(x)$ is also strictly increasing on the same interval and has therefore the unique pth quantile for p, $0<p<1$. The order statistic $X_{([np])}$, where $[np]$ is the greatest integer that does not exceed np, is a consistent estimate of the pth quantile of $F_{m,k}(x)$ under the condition of the uniqueness of the pth quantile (Wilks [8]). Thus we have the following corollary.

COROLLARY 3. *The order statistic*

$$C_n = X_{([n\gamma_{m,k}^{-1}(p)])} \tag{23}$$

is a consistent estimate of the p-th quantile of $F(x)$.

Let us denote the pth quantile of $F(x)$ by $\xi_p(F)$. Note that

$$\xi_{\gamma_{m,k}^{-1}(p)}(F_{m,k}) = \xi_p(F) . \tag{24}$$

We assume further that $F(x)$ has the derivative $f(x)$ which is positive on its support. The probability density function $f_{m,k}(x)$ of $F_{m,k}(x)$ is

$$f_{m,k}(x) = k\binom{m}{k} F(x)^{k-1}(1-F(x))^{m-k} f(x) . \tag{25}$$

This implies that the cdf $F_{m,k}(x)$ also satisfies the condition that the cdf $F_{m,k}(x)$ has the derivative which is positive on its support. For large m, $C_n = X_{([n\gamma_{m,k}^{-1}(p)])}$ is asymptotically distributed according to

$$N(\xi_{\gamma_{m,k}^{-1}(p)}(F), \gamma_{m,k}^{-1}(p)(1-\gamma_{m,k}^{-1}(p))/(nf_{m,k}^2(\xi_{\gamma_{m,k}^{-1}(p)}(F_{m,k}))))$$

(cf. Wilks [8]). That is, the estimator C_n is asymptotically distributed according to

$$N(\xi_p(F), \gamma_{m,k}^{-1}(p)(1-\gamma_{m,k}^{-1}(p))/(nf_{m,k}^2(\xi_p(F)))) . \tag{26}$$

The usual estimator of $\xi_p(F)$ based on a sample of size n from $F(x)$ is the $[np]$th least order statistic of the sample. The ratio of the asymptotic variance of C_n to that of the usual estimator is

$$\frac{f^2(\xi_p(F))\gamma_{m,k}^{-1}(p)(1-\gamma_{m,k}^{-1}(p))}{p(1-p)f_{m,k}^2(\xi_p(F))}. \tag{27}$$

Since $f_{m,k}(x)=F'_{m,k}(x)=\gamma_{m,k}^{-1\prime}(F(x))f(x)$, the ratio (27) can be written as

$$\frac{\gamma_{m,k}^{-1}(p)(1-\gamma_{m,k}^{-1}(p))}{p(1-p)(\gamma_{m,k}^{-1\prime}(p))^2}. \tag{28}$$

This coincides with the left-hand side of (14). Thus, we have a similar result to the one mentioned in the last paragraph of Section 3.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Chernoff, H., Gastwirth, J. L. and Johns, M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation, *Ann. Math. Statist.*, **38**, 52-72.
[2] David, H. A. (1957). "Estimation of means of normal populations from observed minima," *Biometrika*, **44**, 282-286.
[3] Hoeffding, W. (1953). On the distribution of the expected values of the order statistics, *Ann. Math. Statist.*, **24**, 93-100.
[4] Loève, M. (1963). *Probability Theory*, Van Nostrand, Princeton, N.J.
[5] Rao, C. R. (1969). *Linear Statistical Inference and Its Applications*, John Wiley, New York.
[6] Rivlin, J. R. (1969). *An Introduction to the Approximation of Functions*, Blaisdell, Waltham, Mass.
[7] Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of population means based on the sample stratified by means of ordering, *Ann. Inst. Statist.*, **20**, 1-31.
[8] Wilks, S. S. (1950). *Mathematical Statistics*, Princeton University Press, Princeton, N.J.