APPLICATIONES MATHEMATICAE
23,3 (1995), pp. 363–370

R. ZIELIŃSKI (Warszawa)

ESTIMATING MEDIAN AND OTHER QUANTILES IN NONPARAMETRIC MODELS

Abstract. Though widely accepted, in nonparametric models admitting asymmetric distributions the sample median, if $n = 2k$, may be a poor estimator of the population median. Shortcomings of estimators which are not equivariant are presented.

1. Results. Let $\mathcal{F}$ be the class of all distribution functions such that if $F \in \mathcal{F}$ then there exist a and b $(-\infty < a < b < \infty)$ such that $F(a) = 0$, $F(b) = 1$, and F is a strictly increasing differentiable function on (a, b). We consider $\mathcal{F}$ as a group family obtained by subjecting a random variable with a fixed distribution $F \in \mathcal{F}$ to the family of all strictly increasing continuous transformations (see Lehmann (1983), Sec. 1.3, Example 3.4).

In applications $\mathcal{F}$ can be considered as *a basic nonparametric family* which is contained in various nonparametric families including the family of all continuous distributions, the family of all distribution functions which have a density, the family of distributions which have first moments, and so on.

Let $X_1, \ldots, X_{2n}$, for a fixed n, be a sample from an $F \in \mathcal{F}$ and let $M_n = \frac{1}{2}(X_{n:2n} + X_{n+1:2n})$ be the sample estimator of the population median m_F. Here $X_{1:2n} \le X_{2:2n} \le \ldots \le X_{2n:2n}$ are the order statistics from the sample $X_1, \ldots, X_{2n}$. Let $\mathrm{Med}(F, T)$ denote the median of the distribution of the statistic T from a sample which comes from the distribution F.

The statistic M_n is a widely used estimator of the population median (see e.g. Gross (1985), Brown (1985), Bickel and Doksum (1977), Lehmann (1983), to mention only a few most important references in estimation theory).

1991 *Mathematics Subject Classification*: 62G05, 62G30.
Key words and phrases: median, quantiles, estimation.

The aim of this note is to show that M_n is a rather poor estimator of m_F for $F \in \mathcal{F}$. It appears that using M_n as a population median estimator requires some more restrictions on the nonparametric family $\mathcal{F}$.

THEOREM. *For every $C > 0$ there exists $F \in \mathcal{F}$ such that*

$$\mathrm{Med}(F, M_n) - m_F > C.$$

Proof (Construction of F for a given $C > 0$). Let $\mathcal{F}_0$ be the class of all strictly increasing differentiable functions G on $(0,1)$ satisfying $G(0) = 0$ and $G(1) = 1$. Then $\mathcal{F}$ is the class of all functions F satisfying $F(x) = G((x-a)/(b-a))$ for some a and b ($-\infty < a < b < \infty$), and for some $G \in \mathcal{F}_0$.

For a fixed $t \in \left(\frac{1}{4}, \frac{1}{2}\right)$ and a fixed $\varepsilon \in \left(0, \frac{1}{4}\right)$, let $F_{t,\varepsilon} \in \mathcal{F}_0$ be a distribution function such that

$$F_{t,\varepsilon}\left(\tfrac{1}{2}\right) = \tfrac{1}{2}, \qquad F_{t,\varepsilon}(t) = \tfrac{1}{2} - \varepsilon,$$
$$F_{t,\varepsilon}\left(t - \tfrac{1}{4}\right) = \tfrac{1}{2} - 2\varepsilon, \qquad F_{t,\varepsilon}\left(t + \tfrac{1}{4}\right) = 1 - 2\varepsilon.$$

Let $Y_1, \ldots, Y_{2n}$ be a sample from $F_{t,\varepsilon}$. We shall prove that for every $t \in \left(\frac{1}{4}, \frac{1}{2}\right)$ there exists $\varepsilon > 0$ such that

$$\mathrm{Med}\left(F_{t,\varepsilon}, \tfrac{1}{2}(Y_{n:2n} + Y_{n+1:2n})\right) \le t. \tag{1}$$

Consider two random events:

$$A_1 = \{0 \le Y_{n:2n} \le t, 0 \le Y_{n+1:2n} \le t\},$$
$$A_2 = \left\{0 \le Y_{n:2n} \le t - \tfrac{1}{4}, \tfrac{1}{2} \le Y_{n+1:2n} \le t + \tfrac{1}{4}\right\},$$

and observe that $A_1 \cap A_2 = \emptyset$ and

$$A_1 \cup A_2 \subseteq \left\{\tfrac{1}{2}(Y_{n:2n} + Y_{n+1:2n}) \le t\right\}. \tag{2}$$

If the sample comes from a distribution G with a probability density function g, then the joint probability density function $h(x,y)$ of $Y_{n:2n}, Y_{n+1:2n}$ is given by the formula

$$h(x,y) = \frac{\Gamma(2n+1)}{\Gamma(n)\Gamma(n)} G^{n-1}(x)\,[1 - G(y)]^{n-1}\, g(x)g(y), \qquad 0 \le x \le y \le 1,$$

and the probability of A_1 equals

$$P_G(A_1) = \int_0^t dx \int_x^t dy\, h(x,y).$$

Using the formula

$$\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x t^{p-1}(1-t)^{q-1}\, dt = \sum_{j=p}^{p+q-1} \binom{p+q-1}{j} x^j (1-x)^{p+q-1-j}$$

we obtain

$$P_G(A_1) = \sum_{j=n+1}^{2n} \binom{2n}{j} G^j(t)(1-G(t))^{2n-j}.$$

For $P_G(A_2)$ we obtain

$$\begin{aligned} P_G(A_2) &= \int_0^{t-1/4} dx \int_{1/2}^{t+1/4} dy\, h(x,y) \\ &= \binom{2n}{n} G^n\Big(t-\frac{1}{4}\Big)\bigg[\Big(1-G\Big(\frac{1}{2}\Big)\Big)^n - \Big(1-G\Big(t+\frac{1}{4}\Big)\Big)^n\bigg]. \end{aligned}$$

Define $C_1(\varepsilon) = P_{F_{t,\varepsilon}}(A_1)$ and $C_2(\varepsilon) = P_{F_{t,\varepsilon}}(A_2)$. Then

$$C_1(\varepsilon) = \sum_{j=n+1}^{2n} \binom{2n}{j} \Big(\frac{1}{2}-\varepsilon\Big)^j \Big(\frac{1}{2}+\varepsilon\Big)^{2n-j},$$

$$C_2(\varepsilon) = \binom{2n}{n} \Big(\frac{1}{2}-2\varepsilon\Big)^n \bigg[\Big(\frac{1}{2}\Big)^n - (2\varepsilon)^n\bigg].$$

Observe that

$$C_1(\varepsilon) \nearrow \frac{1}{2} \quad \text{as} \quad \varepsilon \searrow 0$$

and

$$C_2(\varepsilon) \nearrow \binom{2n}{n}\Big(\frac{1}{2}\Big)^{2n} \quad \text{as} \quad \varepsilon \searrow 0.$$

Let $\varepsilon_1 > 0$ be such that

$$(\forall \varepsilon < \varepsilon_1) \quad C_1(\varepsilon) > \frac{1}{2} - \frac{1}{2}\binom{2n}{n}\Big(\frac{1}{2}\Big)^{2n}$$

and let ε_2 be such that

$$(\forall \varepsilon < \varepsilon_2) \quad C_2(\varepsilon) > \frac{1}{2}\binom{2n}{n}\Big(\frac{1}{2}\Big)^{2n}.$$

Then for every $\varepsilon < \overline{\varepsilon} = \min\{\varepsilon_1, \varepsilon_2\}$ we have $C_1(\varepsilon) + C_2(\varepsilon) > \frac{1}{2}$ and by (2) for every $\varepsilon < \overline{\varepsilon}$,

$$P_{F_{t,\varepsilon}}\big\{\tfrac{1}{2}(Y_{n:2n} + Y_{n+1:2n}) \le t\big\} > C_1(\varepsilon) + C_2(\varepsilon) > \tfrac{1}{2},$$

which proves (1).

For a fixed $t \in \big(\frac{1}{4}, \frac{1}{2}\big)$ and $\varepsilon < \overline{\varepsilon}$, let $Y, Y_1, \ldots, Y_{2n}$ be i.i.d. random variables distributed as $F_{t,\varepsilon}$, and for a given $C > 0$ define

$$X = C \cdot \frac{\frac{1}{2} - Y}{\frac{1}{2} - t},$$

$$X_{i:2n} = C \cdot \frac{\frac{1}{2} - Y_{2n+1-i:2n}}{\frac{1}{2} - t}, \qquad i = 1, \ldots, 2n.$$

Let F denote the distribution function of X. Then

$$P\{X \le 0\} = P\big\{Y \ge \tfrac{1}{2}\big\} = \tfrac{1}{2},$$

hence $F^{-1}\big(\frac{1}{2}\big) = 0$ and

$$P\big\{\tfrac{1}{2}(X_{n:2n} + X_{n+1:2n}) \le C\big\} = P\big\{\tfrac{1}{2}(Y_{n:2n} + Y_{n+1:2n}) \ge t\big\} \le \tfrac{1}{2}.$$

Thus $\mathrm{Med}\big(F, \frac{1}{2}(X_{n:2n} + X_{n+1:2n})\big) > C$, which proves the Theorem. ∎

2. A comment. It is true that the sample median M_n is asymptotically normal with mean equal to m_F. The problem is that the convergence is not uniform in $\mathcal{F}$ and for every n the Theorem holds.

3. Two remedies. Let $\xi_1, \ldots, \xi_N$ be a sample and let $\mathcal{G}$ be the totality of transformations $\xi_i' = g(\xi_i)$, $i = 1, \ldots, N$, such that g is continuous and strictly increasing. A statistic $T = T(\xi_1, \ldots, \xi_N)$ is said to be equivariant with respect to strictly increasing continuous transformations or $\mathcal{G}$-*equivariant* if

$$T(g(\xi_1), \ldots, g(\xi_N)) = g(T(\xi_1, \ldots, \xi_N)) \quad \text{for all } g \in \mathcal{G}. \tag{3}$$

A reason for the above behaviour of M_n is that M_n is not $\mathcal{G}$-equivariant. Actually, the only $\mathcal{G}$-equivariant statistics are those of the form

$$T(\xi_1, \ldots, \xi_N) = \xi_{J:N}, \tag{4}$$

where J is a random variable taking values in the set $\{1, \ldots, N\}$ (see e.g. Uhlmann (1963)).

Having a sample $X_1, \ldots, X_{2n}$, two natural $\mathcal{G}$-equivariant estimators of the population median are available:

1) a randomized estimator

$$M_n^{(p)} = X_{J:2n},$$

where J is a random variable with distribution

$$p_j = \mathrm{Prob}\{J = j\}, \qquad j = 1, \ldots, 2n,$$

which is constructed in such a way that

$$\mathrm{Med}(F, M_n^{(p)}) = m_F \quad \text{for all } F \in \mathcal{F};$$

2) the sample median

$$M_n^{(2)} = X_{n:2n-1}$$

from the sample $X_1, \dots, X_{2n-1}$ obtained by removing one of the observations $X_1, \dots, X_{2n}$, say X_{2n}. Here again

$$\mathrm{Med}(F, M_n^{(2)}) = m_F \quad \text{for all } F \in \mathcal{F}.$$

The choice between $M_n^{(p)}$ and $M_n^{(2)}$, and if $M_n^{(p)}$ is chosen, the choice of the distribution $p = (p_1, \dots, p_{2n})$ depends of course on a "loss function" or a "criterion" adopted.

Mean Square Error criterion. If T is an estimator of the population median m_F then $F(T)$ should be close to $\frac{1}{2}$ whatever $F \in \mathcal{F}$. Uhlmann (1963) considered the risk of T defined as

$$R_1(F, T) = E_F\big(F(T) - \tfrac{1}{2}\big)^2.$$

He has proved that $M_n^{(p)}$ minimizing the risk in the class of all T satisfying (3), i.e. in the class of T of the form (4), is $M_n^{(p)}$ with $p_n = p_{n+1} = \frac{1}{2}$, $p_j = 0$ if $j \notin \{n, n+1\}$. This estimator will be denoted by $M_n^{(1)}$. He has also shown that

$$R_1(F, M_n^{(1)}) = R_1(F, M_n^{(2)}) = \frac{1}{4(2n+1)} \quad \text{for all } F \in \mathcal{F}.$$

It is interesting to observe that the optimal randomized estimator $M_n^{(1)}$ in the sample $X_1, \dots, X_{2n}$ has the same risk as the nonrandomized estimator $M_n^{(2)}$ from the smaller sample $X_1, \dots, X_{2n-1}$.

Interquartile criterion. Let $Q_p(F, T)$ denote the pth quantile of the distribution of the statistic $F(T)$ if the sample comes from the distribution F. Take

$$R_2(F, T) = Q_{3/4}(F, T) - Q_{1/4}(F, T)$$

as a criterion. Now again (see Zieliński (1988))

$$R_2(F, M_n^{(1)}) \le R_2(F, T) \quad \text{for all } F \in \mathcal{F}$$

for all T satisfying (3). Also

$$R_2(F, M_n^{(1)}) = R_2(F, M_n^{(2)}) \quad \text{for all } F \in \mathcal{F}. \tag{5}$$

To see this define the function

$$C_T(q) = P_F\{F(T) \le q\}$$

and write

$$C_1(q) = C_{M_n^{(1)}}(q), \qquad C_2(q) = C_{M_n^{(2)}}(q).$$

Then (5) is a consequence of the equality

$$C_1(q) = C_2(q) \quad \text{for all } q \in (0, 1). \tag{6}$$

To prove (6) observe that

$$\begin{aligned}
C_1(q) &= \frac{1}{2}P_F\{F(X_{n:2n}) \le q\} + \frac{1}{2}P_F\{F(X_{n+1:2n}) \le q\} \\
&= \frac{1}{2}\sum_{j=n}^{2n}\binom{2n}{j}q^j(1-q)^{2n-j} + \frac{1}{2}\sum_{j=n+1}^{2n}\binom{2n}{j}q^j(1-q)^{2n-j} \\
&= \frac{1}{2}\frac{\Gamma(2n+1)}{\Gamma(n)\Gamma(n+1)}\int_0^q (t^{n-1}(1-t)^n + t^n(1-t)^{n-1})\,dt
\end{aligned}$$

and similarly

$$C_2(q) = \frac{\Gamma(2n)}{\Gamma(n)\Gamma(n)}\int_0^q t^{n-1}(1-t)^{n-1}\,dt,$$

and hence $C_1(q) - C_2(q) = 0$ for all $q \in (0,1)$. Now again the optimal randomized estimator $M_n^{(1)}$ in the sample $X_1, \ldots, X_{2n}$ has the same risk as the nonrandomized estimator $M_n^{(2)}$ from the smaller sample $X_1, \ldots, X_{2n-1}$.

4. A generalization. Statistics of the form $S_\lambda = \sum_{i=1}^n \lambda_i X_{i:n}$, $\lambda_i \ge 0$, $\sum_{i=1}^n \lambda_i = 1$, are frequently used as quantile estimators in nonparametric models (e.g. Harrell and Davis (1982), and Kaigh and Lachenbruch (1982)). However, if two or more of the coefficients λ_i are strictly positive then S_λ is not an equivariant estimator. As a consequence, when estimating the qth quantile, for every $C > 0$ there exists a distribution $F \in \mathcal{F}$ with the qth quantile equal to $x_F(q)$, such that $\mathrm{Med}(F, S_\lambda) - x_F(q) > C$. The proof is similar to that of the Theorem above so we omit it and we confine ourselves to some simulation results.

Consider estimating the qth quantile for $q = 0.25$ of two distributions from $\mathcal{F}_0$: $\mathrm{Beta}(\alpha, 1)$ with $\alpha = 20$ (Fig. 1a) and

$$H(x) = \begin{cases} q\left(\dfrac{x}{q}\right)^\alpha & \text{if } 0 < x \le q, \\ q + (1-q)\left(\dfrac{x-q}{1-q}\right)^\alpha & \text{if } q < x < 1, \end{cases}$$

for $\alpha = 20$ (Fig. 1b).

Distributions of four estimators from samples of size $n = 10$ have been simulated: WU – Uhlmann (1963), RZ – Zieliński (1988), HD – Harrell–Davis (1982), and KL – Kaigh–Lachenbruch (1982) with the subsample size $m = 3$. The empirical distribution functions are given in Fig. 2a (for parent distribution $\mathrm{Beta}(20, 1)$), and in Fig. 2b (for parent distribution H). In the figures the value of the quantile to be estimated is also indicated.

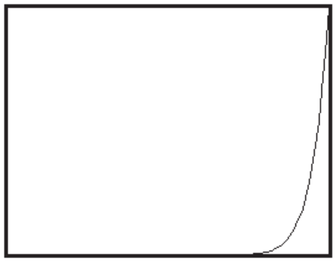

Fig. 1a. Cdf of Beta(20,1)

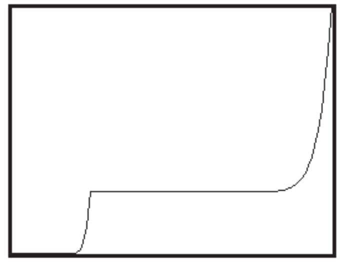

Fig. 1b. Cdf of $H(x)$

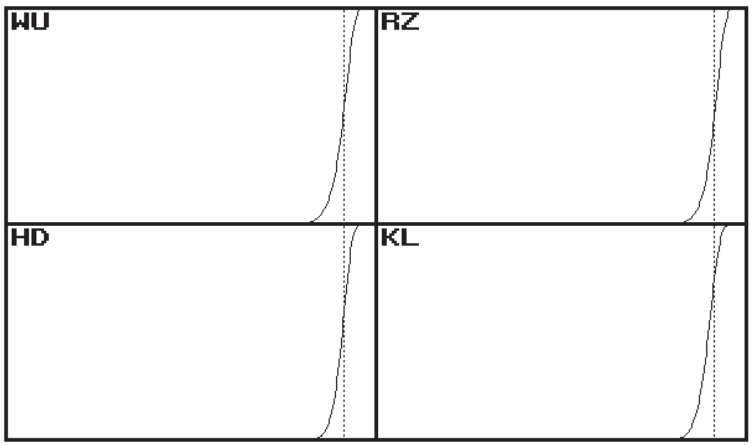

Fig. 2a. Simulated distributions of four estimators for the parent distribution from Fig. 1a

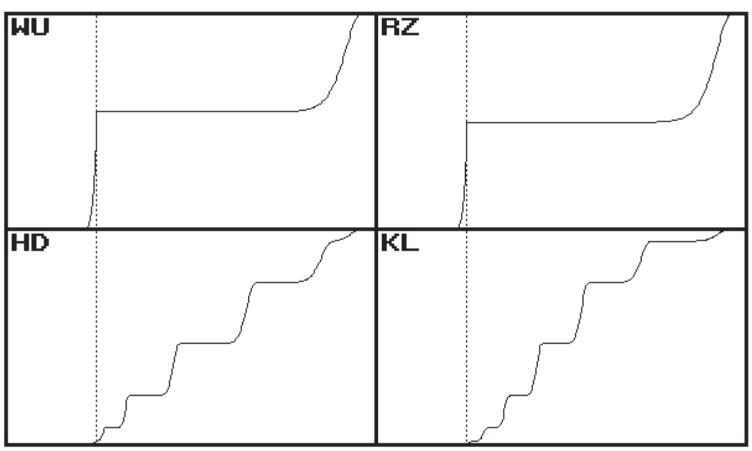

Fig. 2b. Simulated distributions of four estimators for the parent distribution from Fig. 1b

In Table 1 the simulated probabilities of taking on a value not greater than the estimated qth quantile ($q = 0.25$) for all four estimators and for both parent distributions are given; the probability is equal to 0.5 for every median-unbiased estimator.

TABLE 1

Parent distributions	Estimators			
	WU	RZ	HD	KL
Beta(20, 1)	0.5416	0.4985	0.6001	0.7486
H	0.5442	0.4953	0.0185	0.0065

All graphical and numerical results presented are based on 10,000 simulations.

References

P. J. Bickel and K. A. Doksum (1977), *Mathematical Statistics. Basic Ideas and Selected Topics*, Holden-Day.

B. M. Brown (1985), *Median estimates and sign tests*, in: *Encyclopedia of Statistical Sciences*, Vol. 5, Wiley.

C. E. Davis and S. M. Steinberg (1985), *Quantile estimation*, in: *Encyclopedia of Statistical Sciences*, Vol. 7, Wiley.

S. T. Gross (1985), *Median estimation, inverse*, in: *Encyclopedia of Statistical Sciences*, Vol. 5, Wiley.

F. E. Harrell and C. E. Davis (1982), *A new distribution-free quantile estimator*, Biometrika 69, 635–640.

W. D. Kaigh and P. A. Lachenbruch (1982), *A generalized quantile estimator*, Comm. Statist. Theory Methods 11, 2217–2238.

E. L. Lehmann (1983), *Theory of Point Estimation*, Wiley.

W. Uhlmann (1963), *Ranggrössen als Schätzfunktionen*, Metrika 7 (1), 23–40.

R. Zieliński (1988), *A distribution-free median-unbiased quantile estimator*, Statistics 19 (2), 223–227.

RYSZARD ZIELIŃSKI
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
P.O. BOX 137
00-950 WARSZAWA, POLAND

Received on 18.5.1995