ESTIMATING MEDIAN AND OTHER QUANTILES IN NONPARAMETRIC MODELS

Ryszard Zieliński

Summary. 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.

1991 AMS Classification: 62G05, 62G30

Key words: Median, quantiles, estimation

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 strictly increasing continuous 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 such nonparametric families as 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 $X_{1:2n} \leq X_{2:2n} \leq \ldots \leq X_{2n:2n}$ are the order statistics from the sample $X_1, \ldots, X_{2n}$. Let $Med(F, T)$ 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).

The aim of the 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*

$$Med(F, M_n) - m_F > C.$$

Proof (*Construction of F for a given $C > 0$*).

Let $\mathcal{F}_0$ be the class of all strictly increasing continuous functions G on $(0, 1)$ satisfying $G(0) = 0$, $G(1) = 1$. 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 $t \in (\frac{1}{4}, \frac{1}{2})$ and a fixed $\varepsilon \in (0, \frac{1}{4})$, let $F_{t,\varepsilon} \in \mathcal{F}_0$ be a distribution function such that

$$F_{t,\varepsilon}\left(\frac{1}{2}\right) = \frac{1}{2}, \quad F_{t,\varepsilon}(t) = \frac{1}{2} - \varepsilon,$$

$$F_{t,\varepsilon}(t - \frac{1}{4}) = \frac{1}{2} - 2\varepsilon, \quad F_{t,\varepsilon}(t + \frac{1}{4}) = 1 - 2\varepsilon$$

Let $Y_1, Y_2, \ldots, Y_{2n}$ be a sample from $F_{t,\varepsilon}$. We shall prove that for every $t \in (\frac{1}{4}, \frac{1}{2})$ there exists $\varepsilon > 0$ such that

$$Med\left(F_{t,\varepsilon}, \frac{1}{2}(Y_{n:2n} + Y_{n+1:2n})\right) \leq t \tag{1}$$

Consider two random events:

$$A_1 = \{0 \leq Y_{n:2n} \leq t, 0 \leq Y_{n+1:2n} \leq t\}$$

$$A_2 = \{0 \leq Y_{n:2n} \leq t - \frac{1}{4}, \frac{1}{2} \leq Y_{n+1:2n} \leq t + \frac{1}{4}\}$$

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

$$A_1 \cup A_2 \subseteq \{\frac{1}{2}(Y_{n:2n} + Y_{n+1:2n}) \leq t\} \tag{2}$$

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

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

and the probability of A_1 equals to

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

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

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

For $P_G(A_2)$ we obtain

$$P_G(A_2) = \int_0^{t-\frac{1}{4}} dx \int_{\frac{1}{2}}^{t+\frac{1}{4}} dy\, h(x,y)$$
$$= \binom{2n}{n} G^n(t-\frac{1}{4}) \left[\left(1 - G\left(\frac{1}{2}\right)\right)^n - \left(1 - G(t+\frac{1}{4})\right)^n \right]$$

Denote

$$C_1(\varepsilon) = P_{F_{t,\varepsilon}}(A_1), \quad C_2(\varepsilon) = P_{F_{t,\varepsilon}}(A_2)$$

Then

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

$$C_2(\varepsilon) = \binom{2n}{n} (\frac{1}{2} - 2\varepsilon)^n \left[\left(\frac{1}{2}\right)^n - (2\varepsilon)^n \right]$$

Observe that

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

and

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

Let $\varepsilon_1 > 0$ be such that

$$(\forall \varepsilon < \varepsilon_1) \qquad C_1(\varepsilon) > \frac{1}{2} - \frac{1}{2} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}$$

and let ε_2 be such that

$$(\forall \varepsilon < \varepsilon_2) \qquad C_2(\varepsilon) > \frac{1}{2} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}$$

Then for every $\varepsilon < \bar{\varepsilon} = min\{\varepsilon_1, \varepsilon_2\}$ we have

$$C_1(\varepsilon) + C_2(\varepsilon) > \frac{1}{2}$$

and by (2) for every $\varepsilon < \bar{\varepsilon}$

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

which proves (1).

For a fixed $t \in (\frac{1}{4}, \frac{1}{2})$ and $\varepsilon < \bar{\varepsilon}$, let $Y, Y_1, Y_2, \ldots, Y_{2n}$ be i.i.d. random variables distributed as $F_{t,\varepsilon}$, and for a given $C > 0$ define

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

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

Let F denote the distribution function of X. Then

$$P\{X \le 0\} = P\{Y \ge \frac{1}{2}\} = \frac{1}{2}$$

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

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

hence $Med\left(F, \frac{1}{2}(X_{n:2n} + X_{n+1:2n})\right) > C$, 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 $\xi_i' = g(\xi_i)$, $i = 1, 2, \ldots, N$, 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 continuous and strictly increasing transformations or $\mathcal{G}$-equivariant if

$$T(g(\xi_1), g(\xi_2), \ldots, g(\xi_N)) = g\left(T(\xi_1, \ldots, \xi_N)\right) \quad \text{for all } g \in \mathcal{G} \tag{3}$$

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 on values in the set $\{1, 2, \ldots, N\}$ (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 the distribution

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

which is constructed in such a way that

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

2) the sample median

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

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

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

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

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

$$R_1(F, T) = E_F\left(F(T) - \frac{1}{2}\right)^2$$

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 $X_1, \ldots, X_{2n}$ has the same risk as the nonrandomized estimator $M_n^{(2)}$ from the smaller sample $X_1, \ldots, X_{2n-1}$.

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 Zielinski (1988))

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

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

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

and denote

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

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

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

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 $S_\lambda = \sum_{i=1}^n \lambda_i X_{i:n}$, $\lambda_i \geq 0$, $\sum_{i=1}^n \lambda_i = 1$, 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 coefficients λ_i are strictly positive then S_λ is not an equivariant estimator. As a consequence, when estimating 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 $Med(F, S_\lambda) - x_F(q) > C$. The proof is similar to that of the Theorem above so we omit it and we confine ourselves to some simulation results.

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

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

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 $Beta(20, 1)$), and in Fig. 2b (for parent distribution H). In the figures the value of the quantile to be estimated is also exhibited.

In the following Table 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.

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

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

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

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

Gross, S.T. (1985), *Median estimation, inverse*, In *Encyclopedia of Statistical Sciences*, Vol. 5, John Wiley and Sons

Harrell, F.E., Davis, C.E. (1982), *A new distribution–free quantile estimator,* Biometrika, 669, 635–640

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

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

Uhlmann, W. (1963), *Ranggrössen als Schätzfuntionen*, Metrika 7, 1, 23 - 40

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