Order Statistics as Quantile Estimators in a Large Nonparametric Model

Ryszard Zieliński

Institute of Mathematics, Polish Acad.Sc.

POBox 21, Warszawa 10, Poland

R.Zielinski@impan.gov.pl, www.impan.gov.pl/~rziel

Summary and introduction

The large nonparametric model in this note is a statistical model with the family $\mathcal{F}$ of all continuous and strictly increasing distribution functions. In abundant literature of the subject, there are many proposals for nonparametric estimators of quantiles, for example simple order statistics or convex combination of two consecutive order statistics [David and Steinberg (1986): *Quantile estimation, In Encyclopedia of Statistical Sciences, Vol. 7, Wiley*], some more sophisticated L-statistics such as Harrell and Davis (1982) [*A new distribution-free quantile estimator, Biometrika*], or Kaigh and Cheng(1991) [*Subsampling quantile estimators and uniformity criteria. Commun. Statist.-Theory Meth.*], etc. Asymptotically the estimators do not differ substantially but if the sample size n is fixed, which is the case of our concern, differences may be serious. It appears that **in the nonparametric statistical model with the family $\mathcal{F}$ of possible distributions nontrivial L-statistics (the L-statistics which do not degenerate to a single order statistic) are highly unsatisfactory.** For example Zieliński (1995)[*Estimating Median and Other Quantiles in Nonparametric Models. Applicationes Math.*] take the well known estimator of the median $m(F)$ of an unknown distribution $F \in \mathcal{F}$ from a sample of size $2n$, defined as the arithmetic mean of two central observations $M_{2n} = (X_{n:2n} + X_{n+1:2n})/2$. Let $Med(F, M_{2n})$ denote a median of the distribution of the statistic M_{2n} if the sample comes from the distribution F. Then for every $C > 0$ there exists $F \in \mathcal{F}$ such that $Med(F, M_{2n}) - m(F) > C$.

A numerical study (simulations)

To demonstrate that L-statistics are useless for estimating quantiles in the nonparametric model $\mathcal{F}$ with all continuous and strictly increasing distribution functions we decided to present the problem of estimating median with the following well known estimators:

Davis and Steinberg

$$X_{(n+1)/2:n}, \text{if } n \text{ is odd}; \quad \left(X_{n/2:n} + X_{n/2+1:n}\right)/2, \text{if } n \text{ is even}$$

Harrell and Davis

$$\frac{n!}{[(\frac{n-1}{2})!]^2} \sum_{j=1}^{n} \left[\int_{(j-1)/2}^{j/n} [u(1-u)]^{(n-1)/2} du \right] X_{j:n}$$

Kaigh and Cheng

$$\frac{1}{\binom{2n-1}{n}} \sum_{j=1}^{n} \binom{\frac{n-3}{2}+j}{\frac{n-1}{2}} \binom{\frac{3n-1}{2}-j}{\frac{n-1}{2}} X_{j:n}$$

with n odd

Reference distributions $\Big($ for $\alpha = \frac{1}{2},\ \frac{1}{4},\ \frac{1}{8}\Big)$:

Pareto with cdf

$$1 - \frac{1}{x^{\alpha}}, \quad x > 1, \quad \text{heavy tails, no moments of order } k \geq \alpha$$

Power (special case of Beta) with cdf

$$x^{\alpha}, \quad x \in (0, 1), \quad \text{no tails, all moments}$$

Exponential with cdf

$$1 - Exp\{-\alpha x\}, \quad x > 0, \quad \text{very regular}$$

If T is an estimator of the quantile $x_q(F)$ of order $q \in (0,1)$ of an unknown distribution $F \in \mathcal{F}$ then assessing the quality of the estimator in terms of

$$bias: \quad E_F T - x_q(F)$$

$$Mean\ Square\ Error: \quad E_F\big(T - x_q(F)\big)^2$$

etc, is impossible because the moments of $F \in \mathcal{F}$ may not exist.
We decided to study the differences

$$Med(F,T) - x_q(F)$$

which always exists and are finite. Here $Med(F,T)$ is a median of estimator T if the sample comes from the parent distribution F.

We have done our simulation for estimating the population median from samples of size $n = 9$; then $X_{5:9}$ is a natural estimator; for a comparison we have performed simulations for the estimator $(X_{5:10} + X_{6:10})/2$ which is also very "natural" and popular but unfortunately rather unsatisfactory.

The Harrell-Davis estimator takes on the form

$$HD = 0.00145X_{1:9} + 0.0289X_{2:9} + 0.1145X_{3:9} + 0.2207X_{4:9} + 0.2690X_{5:9} + 0.2207X_{6:9} + 0.1145X_{7:9} + 0.0289X_{8:9} + 0.00145X_{9:9}$$

and the Kaigh-Cheng subsampling estimator takes on the form

$$KC = 0.0204X_{1:9} + 0.0679X_{2:9} + 0.1296X_{3:9} + 0.1814X_{4:9} + 0.2016X_{5:9} + 0.1814X_{6:9} + 0.1296X_{7:9} + 0.0679X_{8:9} + 0.0204X_{9:9}$$

Results of our numerical investigations are presented in the following Table:

Table 1. Medians of estimators (simulated)

Distribution	Median	HD	KC	$X_{5:9}$	$\frac{X_{5:10}+X_{6:10}}{2}$
Pareto					
$\alpha = 1/2$	4	7.72	13.71	4.03	4.13
$\alpha = 1/4$	16	255	1107	15.93	18.45
$\alpha = 1/8$	256	3.3×10^6	2.8×10^7	265	383
Power					
$\alpha = 1/2$	0.25	0.2780	0.2919	0.2508	0.2535
$\alpha = 1/4$	0.0625	0.1055	0.1286	0.0629	0.0692
$\alpha = 1/8$	0.0039	0.0241	0.0432	0.0039	0.0053
Exponential					
$\alpha = 1/2$	1.3863	1.5138	1.6235	1.3805	1.4079
$\alpha = 1/4$	2.7726	3.0571	3.2731	2.7718	2.8036
$\alpha = 1/8$	5.5452	6.0595	6.4897	5.5426	5.6143

To assess the exactness of the simulation we may compare columns "Median" and "$X_{5:9}$"; the latter is an unbiased estimator of the median so that the entries of both columns should be approximately equal.

It seems however that absolute differences $b_F(T) = Med(F,T) - x_q(F)$ are not suitable measures of quality of an estimator (is the bias of HD really smaller when estimating median of the Power distribution than that for Exponential distribution?)

To "normalize" the bias we may argue as follows.

If T is an estimator of the qth quantile $x_q(F)$ then $F(T)$ may be considered as an estimator of the (known!) value q.

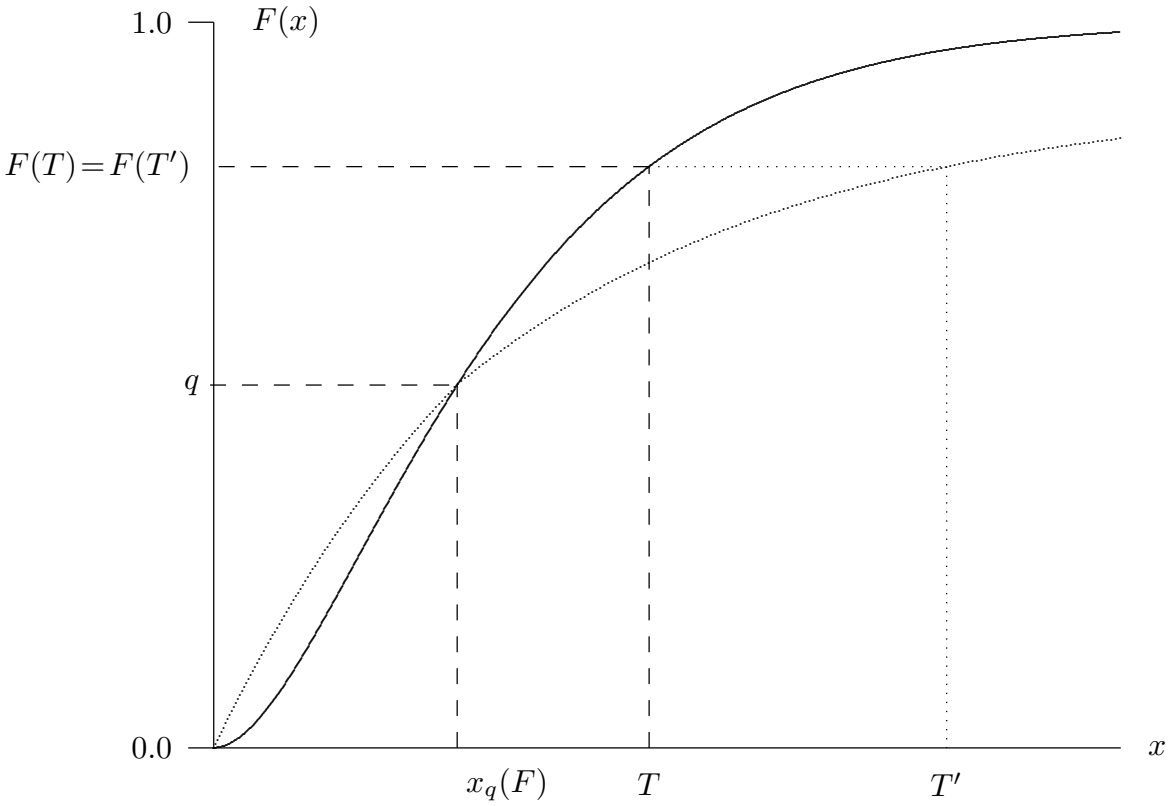

Figure 1

Measuring quality of estimators in terms of differences $F(T) - q$:

Table 2. F–medians of estimators (simulated)

Distribution	Median	HD	KC	$X_{5:9}$	$\frac{X_{5:10} + X_{6:10}}{2}$
Pareto					
$\alpha = 1/2$	0.5	0.6401	0.7299	0.5016	0.5132
$\alpha = 1/4$	0.5	0.7498	0.8265	0.4995	0.5175
$\alpha = 1/8$	0.5	0.8471	0.8830	0.5022	0.5245
Power					
$\alpha = 1/2$	0.5	0.5272	0.5403	0.5008	0.5035
$\alpha = 1/4$	0.5	0.5700	0.5988	0.5008	0.5128
$\alpha = 1/8$	0.5	0.6276	0.6752	0.5004	0.5197
Exponential					
$\alpha = 1/2$	0.5	0.5308	0.5559	0.4986	0.5054
$\alpha = 1/4$	0.5	0.5343	0.5588	0.4999	0.5039
$\alpha = 1/8$	0.5	0.5319	0.5557	0.4998	0.5043

Theoretical results

Theorem 1. *Let T be the Harrell-Davis, or Kaigh-Cheng, or any L-estimator $\sum_{j=1}^{n} \lambda_j X_{j:n}$ such that $\lambda_n > 0$. Then for every $C > 0$ there exists a distribution $F \in \mathcal{F}$ such that*

$$Med(F, T) - m(F) > C.$$

Theorem 2. *If $T = \sum_{j=k}^{m} \lambda_j X_{j:n}$ is an L-statistic such that $\lambda_k > 0$, $\lambda_m > 0$, and $\lambda_k + \lambda_{k+1} + \ldots + \lambda_m = 1$, then*

$$m(U_{k:n}) \leq Med\big(F, F(T)\big) \leq m(U_{m:n})$$

where $m(U_{k:n})$ and $m(U_{m:n})$ are the medians of order statistics $U_{k:n}$ and $U_{m:n}$ from a sample of size n from the uniform $U(0, 1)$ parent distribution. The bounds are sharp in the sense that for every $\varepsilon > 0$ there exists $F \in \mathcal{F}$ such that $Med(F, T) > m(U_{m:n}) - \varepsilon$ and for every $\eta > 0$ there exists $G \in \mathcal{F}$ such that $Med(G, T) < m(U_{k:n}) + \eta$.

Proof of Theorem 1.

Observe that $T \geq \lambda_n X_{n:n}$ *a.s.* and in consequence $Med(F,T) \geq \lambda_n Med(F, X_{n:n})$. Consider the family

$$F_{M,\alpha}(x) = \Big(\frac{x-1}{M-1}\Big)^{1/\alpha}, \quad 1 < x < M, \quad M > 1, \quad \alpha > 0.$$

The median of the distribution is

$$m(F_{M,\alpha}) = 1 + (M-1)2^{-\alpha}$$

The distribution function of $X_{n:n}$ is $F^n_{M,\alpha}(x)$ and the median of that distribution is

$$Med(F_{M,\alpha}, X_{n:n}) = 1 + (M-1)2^{-\alpha/n}$$

Now

$$\begin{aligned} Med(F_{M,\alpha}, T) - m(F_{M,\alpha}) &\geq \lambda_n Med(F_{M,\alpha}, X_{n:n}) - m(F_{M,\alpha}) \\ &= (M-1)\Big[\lambda_n 2^{-\alpha/n} - 2^{-\alpha}\Big] - (1-\lambda_n) \end{aligned}$$

Choosing any $\alpha > -\dfrac{n}{n-1} Log_2 \lambda_n$ (then $\lambda_n 2^{-\alpha/n} - 2^{-\alpha}$ is positive) and any M satisfying

$$M > 1 + \frac{C + (1-\lambda_n)}{\lambda_n 2^{-\alpha/n} - 2^{-\lambda}}$$

we obtain $Med(F_{M,\alpha}, T) - m(F_{M,\alpha}) > C$. □

Proof of Theorem 2.

The first statement follows easily from the fact that $X_{k:n} < T < X_{m:n}$ and hence for every $F \in \mathcal{F}$ we have $U_{k:n} = F(X_{k:n}) < F(T) < F(X_{m:n}) = U_{m:n}$. To prove the second part of the theorem it is enough to construct families of distributions $F_\alpha, \alpha > 0$, and $G_\alpha, \alpha > 0$, such that $Med(F_\alpha, F_\alpha(T)) \to m(U_{m:n})$ and $Med(G_\alpha, G_\alpha(T)) \to m(U_{k:n})$, as $\alpha \to 0$.
Consider the family of power distributions $F_\alpha(x) = x^\alpha$, $0 < x < 1$, $\alpha > 0$. Then $X_{j:n} = F_\alpha^{-1}(U_{j:n}) = U_{j:n}^{1/\alpha}$ and

$$\begin{aligned} F_\alpha(T) &= \Big(\lambda_k U_{k:n}^{1/\alpha} + \lambda_{k+1} U_{k+1:n}^{1/\alpha} + \ldots + \lambda_{m-1} U_{m-1:n}^{1/\alpha} + \lambda_m U_{m:n}^{1/\alpha}\Big)^\alpha \\ &= U_{m:n}\Big[\lambda_k \Big(\frac{U_{k:n}}{U_{m:n}}\Big)^{1/\alpha} + \lambda_{k+1}\Big(\frac{U_{k+1:n}}{U_{m:n}}\Big)^{1/\alpha} + \ldots + \lambda_{m-1}\Big(\frac{U_{m-1:n}}{U_{m:n}}\Big)^{1/\alpha} + \lambda_m\Big]^\alpha \end{aligned}$$

If $\alpha \to 0$ then $F_\alpha(T) \to U_{m:n}$ and $Med(F_\alpha, F_\alpha(T)) \to m(U_{m:n})$.
Now consider the family G_α with $G_\alpha(x) = 1 - (1-x)^\alpha$; in full analogy to the above we conclude that then $G_\alpha(T) \to U_{k:n}$ and $Med(G_\alpha, G_\alpha(T)) \to m(U_{k:n})$ as $\alpha \to 0$. □

Example. For $n = 9$ and L-estimators with $\lambda_9 > 0$ (Harrell-Davis Kaigh-Cheng) we have

$$0.074 \leq Med(F, F(HD)) \leq 0.926$$

The bounds do not depend of the order q of the quantile to be estimated. It follows that the normalized bias $Med(F, F(HD)) - q$ of the estimator, when estimating a quantile of order close to zero, may be close to 0.926. By Theorem 1 the absolute bias may be arbitrarily large.

Conclusions

A reason for the strange behavior of nontrivial L-statistics as quantile estimators is that they are not equivariant under monotonic transformation of data while the class $\mathcal{F}$ of all continuous and strictly increasing distribution functions is closed under such transformations: if X is a random variable with a distribution $F \in \mathcal{F}$ and g is any strictly monotonic transformation then the distribution of $g(X)$ also belongs to $\mathcal{F}$. The class of all statistics which are equivariant with respect to monotonic transformations of data is identical with the class of all order statistics $X_{J:n}$, where J is a random index: $P\{J = j\} = p_j$, $p_j \geq 0$, $\sum_{j=1}^{n} p_j = 1$. Observe that if the sample comes from a distribution $F \in \mathcal{F}$ then $F(X_{J:n}) = U_{J:n}$ and the distribution of $F(X_{J:n})$ does not depend of a specific $F \in \mathcal{F}$. It follows that in the large nonparametric statistical model with the class $\mathcal{F}$ of all continuous and strictly increasing distribution functions the only reasonable estimators of quantiles are single order statistics $X_{J:n}$ with suitably chosen random index J. The random index may be chosen in such a way that

The random index may bo chosen in such a way that $F(X_{J:n})$ is an estimator of q which

- is unbiased $\Big(E_F F(X_{J:n}) = q \text{ for all } F \in \mathcal{F}\Big)$, or

- minimizes Mean Square Error $\Big(E_F(F(X_{J:n}) - q)^2 \text{ for all } F \in \mathcal{F}\Big)$, or

- minimizes Mean Absolute Error $\Big(E_F|F(X_{J:n}) - q| \text{ for all } F \in \mathcal{F}\Big)$, or etc.

Median-unbiasedness of an estimator T is a *sine qua non* condition. To see that, for a distribution $F \in \mathcal{F}$ define the distribution $F_\theta(x) = F(x/\theta)$; then $F_\theta \in \mathcal{F}$ and $x_q(F_\theta) = \theta x_q(F)$. If T is a scale-equivariant estimator (L-statistic are) and if $Med(F,T) - x_q(F) > 0$ for a distribution $F \in \mathcal{F}$ then

$$Med(F_\theta, T) - x_q(F_\theta) = \theta\Big(Med(F,T) - x_q(F)\Big)$$

may be arbitrarily large.

Estimator

$$X_{J:n} \quad \text{with} \quad P\{J = j\} = p_j$$

is median-unbiased if $p_j, j = 1, 2, \ldots, n$, satisfy the condition

$$\sum_{j=1}^{n} p_j \pi_j(q) = \frac{1}{2}, \quad \text{where } \pi_j(q) = \sum_{k=j}^{n} \binom{n}{k} q^k (1-q)^{n-k}$$

In the class of all equivariant median-unbiased estimators of x_q one can find the estimator that is the most concentrated around x_q.

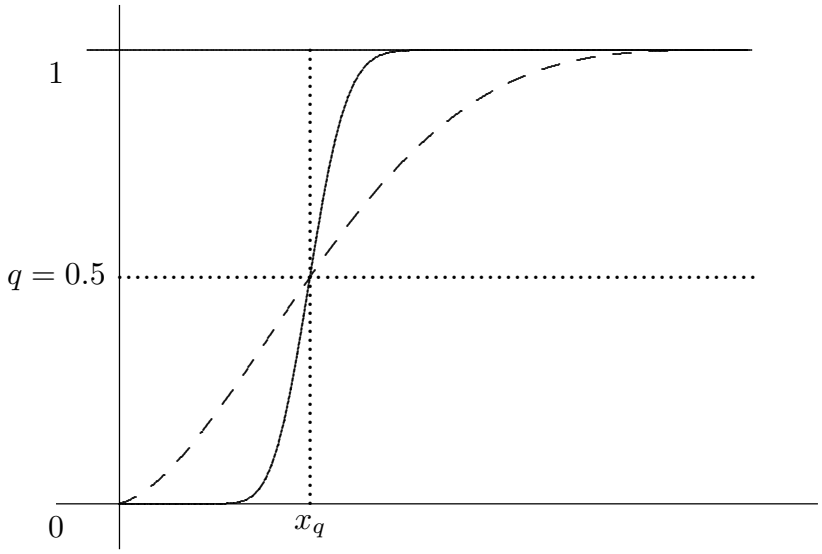

Figure 2. Estimator of x_q with solid cdf
is more concentrated median-unbiased estimator of x_q
than that with dashed cdf

A construction of the estimator, as well as estimators optimal under some other criteria, one can find in my Technical Report at www.impan.gov.pl/˜rziel/Preprint653.pdf

References

Davis, C.E. and Steinberg, S.M. (1986), Quantile estimation, In *Encyclopedia of Statistical Sciences*, Vol. 7, Wiley, New York

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

Kaigh, W.D. and Cheng, C. (1991): Subsampling quantile estimators and uniformity criteria. *Commun. Statist. Theor. Meth.* 20, 539-560

Zieliński, R.(1995), Estimating Median and Other Quantiles in Nonparametric Models. Applicationes Math. 23.3, 363-370. Correction: Applicationes Math. 23.4 (1996) p. 475

Zieliński, R. (2004), Optimal quantile estimators. Small sample approach. *IMPAN, Preprint 653, November 2004*. Available at www.impan.gov.pl/~rziel/Preprint653.pdf