PMC-optimal nonparametric quantile estimator

Ryszard Zieliński

Institute of Mathematics Polish Academy of Sciences

According to Pitman's Measure of Closeness, if T_1 and T_2 are two estimators of a real parameter θ, then T_1 is better than T_2 if $P_\theta\{|T_1-\theta|<|T_2-\theta|\}>1/2$ for all θ. It may however happen that while T_1 is better than T_2 and T_2 is better than T_3, T_3 is better than T_1. Given $q\in(0,1)$ and a sample $X_1,X_2,\ldots,X_n$ from an unknown $F\in\mathcal{F}$, an estimator $T^*=T^*(X_1,X_2,\ldots,X_n)$ of the q-th quantile of the distribution F is constructed such that $P_F\{|F(T^*)-q|\leq|F(T)-q|\}\geq 1/2$ for all $F\in\mathcal{F}$ and for all $T\in\mathcal{T}$, where $\mathcal{F}$ is a nonparametric family of distributions and $\mathcal{T}$ is a class of estimators. It is shown that $T^*=X_{j:n}$ for a suitably chosen jth order statistic.

AMS 1991 subject classification: Primary 62G05; secondary 62G30

Key words and phrases: quantile estimators, nonparametric model, optimal estimation, equivariant estimators, PMC (Pitman's Measure of Closeness)

Address: Institute of Mathematics PAN, P.O.Box 137, 00-950 Warsaw, Poland
E-mail: rziel@impan.gov.pl

Pitman's Measure of Closeness If T and S are two estimators of a real parameter θ we define T as *better than* S if $P_\theta\{|T-\theta| \leq |S-\theta|\} \geq 1/2$ for all θ (Keating et al. 1991, 1993). A rationale behind that criterion is that the absolute error of estimator T is more often smaller than that of S. A restricted applicability of the idea is a consequence of the fact that while T_1 is better than T_2 and T_2 is better than T_3 it may happen that T_3 is better than T_1. It may however happen that in a given statistical model and in a given class of estimators there exists one which is better than any other. We define such estimator as *PMC-optimal* . In what follows we construct a *PMC-optimal* estimator of a qth quantile of an unknown continuous and strictly increasing distribution function.

Statistical model. Let $\mathcal{F}$ be the family of all continuous and strictly increasing distribution functions on the real line: $F \in \mathcal{F}$ if and only if $F(a)=0$,$F(b)=1$, and F is strictly increasing on (a,b) for some a and b, $-\infty \leq a < b \leq +\infty$. Let $X_1, X_2, \ldots, X_n$ be a sample from an unknown $F \in \mathcal{F}$ and let $X_{1:n}, X_{2:n}, \ldots$ $\ldots, X_{n:n}$ $(X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n})$ be the order statistic from the sample. The sample size n is assumed to be fixed (nonasymptotic approach). Let $q \in (0,1)$ be a given number and let $x_q(F)$ denote the qth quantile (the quantile of order q) of the distribution $F \in \mathcal{F}$. The problem is to estimate $x_q(F)$.

Due to the fact that $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ is a minimal sufficient and complete statistic for $\mathcal{F}$ (Lehmann 1983) we confine ourselves to estimators $T = T(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$.

Observe that if X is a random variable with a distribution $F \in \mathcal{F}$ with the qth quantile equal to x then, for every strictly increasing fucntion φ, the random variable $\varphi(X)$ has a distribution from $\mathcal{F}$ with the qth quantile equal to $\varphi(x)$. According to that property we confine ourselves to the class $\mathcal{T}$ of equivariant estimators: $T \in \mathcal{T}$ iff

$$T\left(\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_n)\right) = \varphi\left(T(x_1, x_2, \ldots, x_n)\right)$$

for all strictly increasing functions φ and for all $x_1 \leq x_2 \leq \ldots \leq x_n$

It follows that $T(x_1, x_2, ..., x_n) = x_k$ for any fixed k (Uhlmann (1963)). Allowing randomization (Zieliński 1999) we conclude that the class $\mathcal{T}$ of equivariant

estimators (1) is identical with the class of estimators

$$T = X_{J(\lambda):n}$$

where $J(\lambda)$ is a random variable independent of the sample $X_1, X_2, \ldots, X_n$, such that

$$P\{J(\lambda) = j\} = \lambda_j, \quad \lambda_j \geq 0, \quad j = 1, 2, \ldots, n, \quad \sum_{j=1}^{n} \lambda_j = 1$$

This gives us an explicit and easily tractable characterization of the class $\mathcal{T}$ of estimators under consideration.

Observe that if T is to be a good estimator of the qth quantile $x_q(F)$ of an unknown distribution $F \in \mathcal{F}$, then $F(T)$ should be close to q. Hence we shall measure the error of estimation in terms of differences $|F(T(X_1, X_2, \ldots, X_n)) - q|$ rather than in terms of differences $|T(X_1, X_2, \ldots, X_n)) - x_q(F)|$. According to the Pitman's Measure of Closeness an estimator T is better than S if

$$P_F\{|F(T) - q| \leq |F(S) - q|\} \geq 1/2 \text{ for all } F \in \mathcal{F} \tag{1}$$

(for more fine definitions see Keating *et al.* 1993).

DEFINITION. *An estimator T^* which satisfies*

$$P_F\{|F(T^*) - q| \leq |F(S) - q|\} \geq 1/2 \text{ for all } F \in \mathcal{F} \text{ and for all } S \in \mathcal{T} \tag{2}$$

is said to be PMC-optimal .

We use $\leq$ in the first inequality in the above definition because for $T = S$ we prefer to have LHS of (1) to be equal to one rather than to zero; otherwise the part "for all $T \in \mathcal{T}$" in (2) would not be true. For example two different estimators $X_{[nq]:n}$ and $X_{[(n+1)q]:n}$ are identical for $n = 7$ when estimating qth quantile for $q = 0.2$.

One can easily conclude from the proof of the Theorem below that the second inequality $\geq 1/2$ may be strengthened in the following sense: if there are two optimal estimators T_1^* and T_2^* (we can see from the proof of the Theorem that it may happen), then $P_F\{|F(T_1^*) - q| \leq |F(T_2^*) - q|\} = \frac{1}{2}$ and $P_F\{|F(T_1^*) - q| \leq |F(T) - q|\} > \frac{1}{2}$ for all other estimators $T \in \mathcal{T}$.

Denote LHS of (1) by $p(T,S)$ and observe that to construct T^* it is enough to find T' such that

$$\min_{S\in\mathcal{T}} p(T',S) = \max_{T\in\mathcal{T}} \min_{S\in\mathcal{T}} p(T,S) \text{ for all } F\in\mathcal{F}$$

and take $T^* = T'$ if $\min_{S\in\mathcal{T}} p(T^*,S) \geq \frac{1}{2}$ for all $F\in\mathcal{F}$. If the inequality does not hold then the optimal estimator T^* does not exist. In what follows we construct the estimator T^*.

The optimal estimator T^*. Let $T = X_{J(\lambda):n}$ and $S = X_{J(\mu):n}$. If the sample $X_1, X_2, \ldots, X_n$ comes from a distribution function F then $F(T) = U_{J(\lambda):n}$ and $F(S) = U_{J(\mu):n}$, respectively, where $U_{1:n}, U_{2:n}, \ldots, U_{n:n}$ are the order statistics from a sample $U_1, U_2, \ldots, U_n$ drawn from the uniform distribution $U(0,1)$. Denote

$$w_q(i,j) = P\{|U_{i:n} - q| \leq |U_{j:n} - q|\}, \quad 1 \leq i,j \leq n$$

Then

$$p(T,S) = p(\lambda,\mu) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \mu_j w_q(i,j)$$

and $T^* = X_{J(\lambda^*:n)}$ is optimal if

$$\min_{\mu} p(\lambda^*,\mu) = \max_{\lambda}\min_{\mu} p(\lambda,\mu)$$

and

$$\min_{\mu} p(\lambda^*,\mu) \geq \frac{1}{2}$$

For a fixed i, the sum $\sum_{j=1}^{n} \mu_j w_q(i,j)$ is minimal for $\mu_{j^*} = 1$, $\mu_j = 0, j \neq j^*$, where $j^* = j^*(i)$ is such that $w_q(i,j^*) \leq w_q(i,j)$, $j = 1,2,\ldots,n$. Then the optimal λ^* satisfies $\lambda_{i^*} = 1$, $\lambda_i = 0, i \neq i^*$, where i^* maximizes $w_q(i,j^*(i))$. It follows that the optimal estimator T^* is of the form $X_{i^*:n}$ with a suitable i^* and the problem reduces to finding i^*.

Denote $v_q^-(i) = w_q(i,i-1)$, $v_q^+(i) = w_q(i,i+1)$ and define $v_q^-(1) = v_q^+(n) = 1$. Proofs of all Lemmas and the Theorem below are postponed to next Section.

LEMMA 1. *For a fixed* $i = 1, 2, \ldots, n$, *we have* $\min_j w_q(i,j) = \min\{v_q^-(i), v_q^+(i)\}$.

By Lemma 1, the problem reduces to finding i^* which maximizes $\min\{v_q^-(i), v_q^+(i)\}$.

LEMMA 2. *The sequence* $v_q^+(i)$, $i = 1, 2, \ldots, n$, *is increassing and the sequence* $v_q^-(i)$, $i = 1, 2, \ldots, n$ *is decreasing.*

By Lemma 2, to get i^* one should find $i' \in \{1, 2, \ldots, n-1\}$ such that

$$v_q^-(i') \geq v_q^+(i') \quad \text{and} \quad v_q^-(i'+1) < v_q^+(i'+1) \tag{3}$$

and then calculate

$$i^* = \begin{cases} i', & \text{if } v_q^+(i') \geq v_q^-(i'+1) \\ i'+1, & \text{otherwise} \end{cases} \tag{4}$$

Eventually we obtain the following theorem.

THEOREM. *Let* i^* *be defined by the formula*

$$i^* = \begin{cases} i', & \textit{if } v_q^+(i') \geq \dfrac{1}{2} \\ i'+1, & \textit{otherwise} \end{cases} \tag{5}$$

where

(6)

$$i' = \begin{cases} \textit{the smallest integer } i \in \{1, 2, \ldots, n-2\} & \textit{such that } Q(i+1; n, q) < \dfrac{1}{2} \\ n-1, & \textit{if } Q(n-1, n, q) \geq \dfrac{1}{2} \end{cases}$$

where

$$Q(k; n, q) = \sum_{j=k}^{n} \binom{n}{j} q^j (1-q)^{n-j} = I_q(k, n-k+1)$$

and

$$I_x(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt$$

For i^* *defined by (5) we have*

$$P_F\{|F(X_{i^*:n} - q| \leq |F(T) - q|\} \geq \frac{1}{2} \tag{7}$$

for all $F \in \mathcal{F}$ *and for all equivariant estimators* T *of the* q*th quantile, which means that* i^* *is optimal.*

Index i' can be easily found by tables or suitable computer programs for Bernoulli or Beta distributions. Checking the condition in (5) will be commented in Section *Practical applications*.

As a conclusion we obtain that $X_{i^*:n}$ is *PMC-optimal* in the class of all equivariant estimators of the qth quantile.

Proofs.

Proof of Lemma 1. Suppose first that $i < j$ and consider the following events

$$A_1 = \{U_{i:n} > q\}, \quad A_2 = \{U_{i:n} \leq q < U_{j:n}\}, \quad A_3 = \{U_{j:n} < q\} \tag{8}$$

The events are pairwise disjoint and $P(A_1 \cup A_2 \cup A_3) = 1$. Hence

$$w_q(i,j) = \sum_{j=1}^{3} P\{|U_{i:n} - q| \leq |U_{j:n} - q|, A_j\}$$

For the first summand we have

$$P\{|U_{i:n} - q| \leq |U_{j:n} - q|, A_1\} = P\{U_{i:n} > q\}$$

The second summand can be written in the form

$$\begin{aligned} P\{|U_{i:n} - q| \leq |U_{j:n} - q|, A_2\} &= P\{U_{i:n} + U_{j:n} \geq 2q, U_{i:n} \leq q < U_{j:n}\} \\ &= P\{U_{i:n} \leq q < U_{j:n}, U_{j:n} \geq 2q - U_{i:n}\} \end{aligned}$$

and the third one equals zero.

If $j' > j$ then $U_{j':n} \geq U_{j:n}$, the event $\{U_{i:n} \leq q < U_{j:n}, U_{j:n} \geq 2q - U_{i:n}\}$ implies the event $\{U_{i:n} \leq q < U_{j':n}, U_{j':n} \geq 2q - U_{i:n}\}$, and hence

$$w_q(i,j') \geq w_q(i,j)$$

In consequence

$$\min_{j>i} w_q(i,j) = w_q(i,i+1) = v_q^+(i)$$

Similarly $\min_{j<i} w_q(i,j) = v_q^-(i)$, which ends the proof of Lemma 1. □

PROOF OF LEMMA 2. Similarly as in the proof of Lemma 1, considering events (8) with $j = i + 1$, we obtain

$$v_q^+(i) = P\{U_{i:n} > q\} + P\{U_{i:n} + U_{i+1:n} \geq 2q, U_{i:n} \leq q < U_{i+1:n}\}$$

and by standard calculations

$$v_q^+(i) = \frac{n!}{(i-1)!(n-i)!} \left(\int_q^1 x^{i-1}(1-x)^{n-i}dx + \int_{(2q-1)^+}^q x^{i-1}(1-2q+x)^{n-i}dx \right)$$

where $x^+ = \max\{x, 0\}$. For $i = n - 1$ we obviously have $v_q^+(n-1) < v_q^+(n) = 1$. For $i \in \{1, 2, \ldots, n-1\}$ the inequality $v_q^+(i) < v_q^+(i+1)$ can be written in the form

$$i\Big(\int_q^1 x^{i-1}(1-x)^{n-i}dx + \int_{(2q-1)^+}^q x^{i-1}(1-2q+x)^{n-i}dx \Big) <$$

$$< (n-i)\Big(\int_q^1 x^{i}(1-x)^{n-i-1}dx + \int_{(2q-1)^+}^q x^{i}(1-2q+x)^{n-i-1}dx \Big)$$

Integrating LHS by parts we obtain an equivalent inequality

$$2(n-i) \int_{(2q-1)^+}^q x^i(1-2q+x)^{n-i-1}dx > 0$$

which is obviously always true.

In full analogy to the calculation of $v_q^+(i)$, for $i \in \{2, 3, \ldots, n\}$ we obtain

$$v_q^-(i) = \frac{n!}{(i-1)!(n-i)!} \left(\int_0^q x^{i-1}(1-x)^{n-i}dx + \int_q^{min\{1,2q\}} (2q-x)^{i-1}(1-x)^{n-i}dx \right)$$

and the inequality $v_q^-(i-1) > v_q^-(i)$ can be proved as above, which ends the proof of Lemma 2. □

PROOF OF THE THEOREM. We shall use following facts

$$v_q^+(i) + v_q^-(i+1) = 1 \tag{9}$$

which follows from the obvious equality $w_q(i,j) + w_q(j,i) = 1$, and

$$v_q^+(i) + v_q^+(i+1) = 2\left(1 - Q(i+1;n,q)\right), \quad i = 1,2,\ldots,n-1 \tag{10}$$

Equality (10) follows from integrating by parts both integrals in $v_q^+(i)$ and then calculating the sum $v_q^+(i) + v_q^+(i+1)$.

Let us consider condition (3) for $i = 1$, $i = n-1$, and $i \in \{2,3,\ldots,n-2\}$, separately.

For $i = 1$ we have $v_1^-(1) = 1 > v_q^+(1)$ hence $i' = 1$ iff $v_q^-(2) < v_q^+(2)$ which by (9) amounts to $1 - v_q^+(1) < v_q^+(2)$ and by (10) to $2\left(1 - Q(2,n,q)\right) > 1$ or $Q(2,n,q) < \dfrac{1}{2}$. Now $i^* = 1$ if $v_q^+(1) \geq v_q^-(2)$ or $v_q^+(1) \geq 1 - v_q^+(1)$ or $v_q^+(1) \geq \dfrac{1}{2}$, and $i^* = 2$ if $v_q^+(1) < \dfrac{1}{2}$.

Due to the equality $v_q^-(n) < v_q^+(n) = 1$, by (3) we have $i' = n-1$ iff $v_q^-(n-1) \geq v_q^+(n-1)$ which by (9) amounts to $v_q^+(n-2) + v_q^+(n-1) \leq 1$, and by (10) to $Q(n-1;n,q) \geq \dfrac{1}{2}$. Now $i^* = n-1$ if $v_q^+(n-1) \geq v_q^-(n)$ or $v_q^+(n-1) \geq \dfrac{1}{2}$; otherwise $i^* = n$.

For $i \in \{2,3,\ldots,n-2\}$, by (9), condition (3) can be written in the form

$$v_q^+(i-1) + v_q^+(i) \leq 1 \quad \text{and} \quad v_q^+(i) + v_q^+(i+1) > 1$$

and by (10) in the form

$$Q(i;n,q) \geq \frac{1}{2} \quad \text{and} \quad Q(i+1;n,q) < \frac{1}{2}$$

Now by (4) and (9)

$$i^* = \begin{cases} i', & \text{if } v_q^+(i') \geq \dfrac{1}{2} \\ i'+1, & \text{otherwise} \end{cases}$$

Summing up all above and taking into account that $Q(i;n,q)$ decreases in $i = 1, 2, \ldots, n-1$, we obtain

$$i' = \begin{cases} \text{first } i \in \{1, 2, \ldots, n-2\} \text{ such that } Q(i+1;n,q) < \dfrac{1}{2} \\ n-1, \text{if such } i \text{ does not exist} \end{cases}$$

Then $i^* = i'$ if $v_q^+(i') \geq \dfrac{1}{2}$ and $i^* = i' + 1$ otherwise, which gives us statement (5)-(6) of the Theorem.

To prove statement (7) of the Theorem observe that if $i^* = 1$ then $v_q^+(1) \geq \dfrac{1}{2}$ and if $i^* = n$ then $v_q^-(n) = 1 - v_q^+(n-1) \geq \dfrac{1}{2}$. For $i^* \in \{2, 3, \ldots, n-1\}$ we have: 1) if $i^* = i'$ then by (5) $v_q^+(i^*) \geq \dfrac{1}{2}$ and by the first inequality in (3) $v_q^-(i^*) \geq v_q^+(i^*) \geq \dfrac{1}{2}$, hence $\min\{v_q^-(i^*), v_q^+(i^*)\} \geq \dfrac{1}{2}$ and 2) if $i^* = i' + 1$ then by (5) $v_q^+(i^* - 1) < \dfrac{1}{2}$ which amounts to $1 - v_q^-(i^*) < \dfrac{1}{2}$ or $v_q^-(i^*) > \dfrac{1}{2}$. Then by the second inequality in (3) we have $v_q^+(i^*) > v_q^-(i^*) > \dfrac{1}{2}$, so that again $\min\{v_q^-(i^*), v_q^+(i^*)\} \geq \dfrac{1}{2}$, which ends the proof of the theorem. □

Practical applications. While calculating i' in the Theorem is easy, checking condition (5) needs a comment.

First of all observe that $v_0^+(i) = 1$, $v_1^+(i) = 0$, and the first derivative of $v_q^+(i)$ with respect to q is negative. It follows that $v_q^+(i) \geq \dfrac{1}{2}$ iff $q \leq q_n(i)$ where $q_n(i)$ is the unique solution (with respect to q) of the equation $v_q^+(i) = \dfrac{1}{2}$. For $q \in (0,1)$, $v_q^+(i)$ is a strictly decreasing function with known values at both ends of the interval so that $q_n(i)$ can be easily found by a standard numerical routine. Table 1 gives us the values of $q_n(i)$ for $n = 3, 4, \ldots, 20$. Due to the equality

$$v_q^+(i) + v_{1-q}^+(n-i) = 1$$

we have $q \leq q_n(i)$ iff $1 - q \geq q_n(n-i)$ so that in Table 1 only the values $q_n(i)$ for $i < [n/2]$ are presented. Sometimes the following fact may be useful: if i^* is optimal for estimating the qth quantile from sample of size n, then $n - i^* + 1$ is optimal for estimating the $(1-q)$th quantile from the same sample.

Table 1

n	i								
	1	2	3	4	5	6	7	8	9
3	.3612								
4	.2800								
5	.2283	.4086							
6	.1926	.3450							
7	.1666	.2984	.4326						
8	.1467	.2628	.3811						
9	.1311	.2348	.3406	.4468					
10	.1184	.2122	.3077	.4038					
11	.1080	.1936	.2807	.3683	.4561				
12	.0993	.1779	.2580	.3385	.4192				
13	.0919	.1646	.2387	.3132	.3879	.4626			
14	.0855	.1532	.2221	.2914	.3609	.4304			
15	.0799	.1432	.2076	.2724	.3374	.4024	.4675		
16	.0750	.1344	.1949	.2558	.3168	.3778	.4389		
17	.0707	.1267	.1837	.2411	.2985	.3560	.4136	.4712	
18	.0669	.1198	.1737	.2279	.2823	.3367	.3911	.4455	
19	.0634	.1136	.1647	.2162	.2677	.3193	.3709	.4225	.4742
20	.0603	.1080	.1567	.2055	.2545	.3036	.3527	.4018	.4509

Examples.

1. Suppose we want to estimate the qth quantile with $q = 0.3$ from a sample of size $n = 10$. For the Bernoulli distribution we have

$$B(4, 10; 0.3) = 0.3504 < \frac{1}{2} < B(3, 10; 0.3)$$

hence $i' = 3$. Now $q_{10}(3) = 0.3077$ so that $q < q_n(i')$, hence $i^* = 3$.

2. For $n = 8$ and $q = 0.75$ we have $B(7, 8; 0.75) = 0.3671 < \frac{1}{2} < B(6, 8; 0.75) = 0.6785$ and $i' = 6$. By Table 1 we have $q_8(6) = 1 - q_8(2) = 0.7372$. Now $q > q_8(6)$ so that $i^* = i' + 1 = 7$.

A comment.

It is interesting to observe that PMC-optimal estimator differs from that which minimizes Mean Absolute Deviation $E_F|F(T)-q|$; the latter has been constructed in Zieliński (1999). For example, to estimate the quantile of order $q = 0.225$, $X_{3:10}$ is *PMC-optimal* , while $X_{2:10}$ minimizes Mean Absolute Deviation.

Acknowledgment.

The research was supported by grant KBN 2 P03A 033 17.

References

Keating, J.P., Mason, R.L., and Sen, P.K. (1993) *Pitman's Measure of Closeness: A comparison of Statistical Estimators*. SIAM, Philadelphia.

Keating, J.P., Mason, R.L., Rao, C.R., Sen, P.K. ed's (1991) *Pitman's Measure of Closeness*. Special Issue of *Communications in Statistics - Theory and Methods*, Vol. 20, Number 11

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. (1999), *Best equivariant nonparametric estimator of a quantile*, Statistics and Probability Letters 45, 79–84.