APPLICATIONES MATHEMATICAE
39,2 (2012), pp. 137–142

Ryszard Zieliński (Warszawa)

OPTIMAL ESTIMATION OF HIGH QUANTILES IN A LARGE NONPARAMETRIC MODEL

Abstract. "A high quantile is a quantile of order q with q close to one." A precise constructive definition of high quantiles is given and optimal estimates are presented.

1. The problem. In practice one often has to assess the risk of large but possibly rare losses and indicate thresholds for parameters in economic, especially financial (e.g. value at risk), ecological (e.g. floods) or technological (e.g. atomic power stations) systems. Due to insufficient real data it is rather difficult to formulate an appropriate statistical model with a specified family of heavy tailed probability distributions and a natural solution is to adopt a suitable nonparametric approach. There are many nonparametric models known in statistics and its applications. A simple and widespread example is the statistical model with continuous distribution functions and finite first and second moments. Another popular nonparametric model considered in problems of estimating density functions is a model with distributions which have densities satisfying some special conditions (e.g. bounded). We call the statistical model with all continuous and strictly increasing (on their supports) distribution functions the *large nonparametric model.* We denote by $\mathcal{F}$ the family of all such distributions. Our problem consists in estimating the quantile $x_q = F^{-1}(q)$ of order q (with q close to one) of an unknown distribution $F \in \mathcal{F}$ on the basis of a sample $X_1, \ldots, X_n$ drawn from F.

The lack of information beyond the range of the sample makes it difficult to estimate high quantiles. For an empirical distribution function F_n we have $F_n(x) = 0$ for $x < X_{1:n}$ and $F_n(x) = 1$ for $x \geq X_{n:n}$ and it seems impossible to estimate sufficiently low or sufficiently high quantiles. Typically different tricks with extrapolation beyond the sample range are used, and different

2010 *Mathematics Subject Classification*: 62G30, 62G32.
Key words and phrases: tails, quantiles, high quantiles, order statistics.

DOI: 10.4064/am39-2-2

extrapolation ideas lead to a variety of estimates. See for example recent results by Wang et al. (2010), Li et al. (2010) or an excellent short review in Markovich (2007), chapt. 6, as well as the abundant references therein.

We shall restrict ourselves to estimating high quantiles in the large nonparametric model without restrictions on tails of distributions. It is obvious that what a high quantile is depends on the size n of the sample.

2. Basic facts on optimal estimates in the large statistical model. The minimal complete sufficient statistic is the vector of order statistics $(X_{1:n},\dots,X_{n:n})$ (Lehmann, 1983), hence we confine ourselves to estimates of the form $T(X_{1:n},\dots,X_{n:n})$.

A specific property of the model is that if X is a random variable with a distribution $F\in\mathcal{F}$ and g is any strictly monotone function then the distribution of the random variable $g(X)$ also belongs to $\mathcal{F}$, and if $x_q=x_q(F)$ is the quantile of order q of F then $g(x_q)$ is the quantile of order q of the distribution of $g(X)$. It follows that if $T(X_{1:n},\dots,X_{n:n})$ is an estimate of the quantile x_q of the distribution F of a random variable X, then, for every monotone transformation g, $T(g(X_{1:n}),\dots,g(X_{n:n}))$ should be a suitable estimate of the qth quantile of the distribution of $g(X)$. Otherwise an estimate that works well for quantiles of one distribution $F\in\mathcal{F}$ may be completely unacceptable for another one. Formally, for any fixed $x_1\le\dots\le x_n$, and for every monotone transformation g, the estimate T should satisfy

$$T(g(x_1),\dots,g(x_n))=g(T(x_1,\dots,x_n)).$$

The class $\mathcal{T}$ of estimates which satisfy this condition is identical with the class of randomized order statistics:

$$T\in\mathcal{T}\quad\text{iff}\quad T=X_{J:n}\text{ for a random variable }J\text{ on }\{1,\dots,n\}$$

(Uhlmann, 1963; Zieliński, 2009).

In the large nonparametric model $\mathcal{F}$, a natural counterpart of the minimum variance unbiased estimate is the maximum concentrated median-unbiased estimate.

For the estimate $T=X_{J:n}$ of the qth quantile $x_q(F)$, with

$$P\{J=j\}=\lambda_j,\quad \lambda_j\ge 0,\quad \sum_{j=1}^{n}\lambda_j=1,$$

we have

$$\begin{aligned}P_F\{T\le x_q(F)\}&=P_F\{X_{j:n}\le x_q(F)\}=P_F\{F(X_{J:n})\le q)\}\\&=P\{U_{J:n}\le q\}=\sum_{j=1}^{n}\lambda_jP\{U_{j:n}\le q\},\end{aligned}$$

where $U_{j:n}$ is the jth order statistic from a sample of size n from the uniform distribution on $(0,1)$.

Denote

$$\pi_j(q) = P\{U_{j:n} \le q\} = \sum_{k=j}^{n} \binom{n}{k} q^k (1-q)^{n-k}.$$

It follows that T is a median-unbiased estimate of the qth quantile $x_q(F)$, i.e. $P_F\{T \le x_q(F)\} = 1/2$ for all $F \in \mathcal{F}$, iff $\lambda_j,\ j = 1, \dots, n$, satisfy

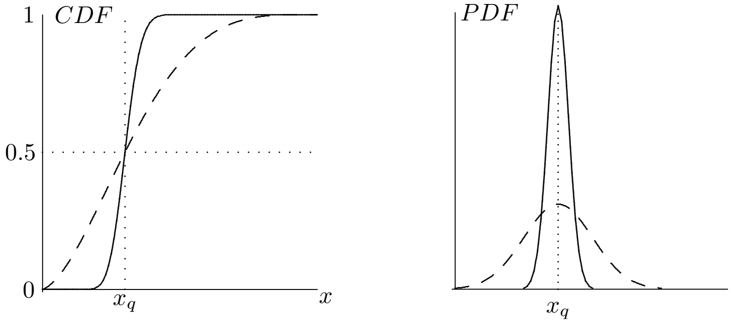
$$\sum_{j=1}^{n} \lambda_j \pi_j(q) = \frac{1}{2}, \quad \text{where} \quad \pi_j(q) = \sum_{k=j}^{n} \binom{n}{k} q^k (1-q)^{n-k}.$$

Note that a median-unbiased estimate exists iff $\pi_1(q) \ge 1/2 \ge \pi_n(q)$. Given q, the smallest $n = n(q)$ for which a median-unbiased estimate exists is $n(q) = \min\{n : n \ge -\log 2/\log(\max\{q, 1-q\})\}$. On the other hand, given n, the order q of a quantile to be estimated without bias should satisfy $1 - (1/2)^{1/n} \le q \le (1/2)^{1/n}$.

A median-unbiased estimate T^* of the qth quantile $x_q(F)$ is said to be the *most concentrated at* $x_q(F)$ if for all $F \in \mathcal{F}$,

$$P_F\{T^* \le t\} \begin{cases} \le P_F\{T \le t\} & \text{for } t \le x_q(F), \\ \ge P_F\{T \le t\} & \text{for } t \ge x_q(F), \end{cases}$$

for any other median-unbiased estimate T.



Estimate of x_q with solid cdf and pdf is more concentrated than those with dashed ones.

Given q and $n \ge n(q)$, let k be an integer such that $\pi_k(q) > 1/2 > \pi_{k+1}(q)$. Let

$$\lambda_k^* = \frac{1/2 - \pi_{k+1}(q)}{\pi_k(q) - \pi_{k+1}(q)}, \quad \lambda_{k+1}^* = 1 - \lambda_k^*, \quad \lambda_i^* = 0 \text{ for } i \notin \{k, k+1\}.$$

If $\pi_k(q) = 1/2$ for some $k = 1, \dots, n$, put $\lambda_k^* = 1$.

Theorem 1. *In the class of all median-unbiased estimates of the qth quantile, $X_{J^*:n}$, where J^* has the distribution $(\lambda_1^*,\ldots,\lambda_n^*)$, is the most concentrated one.*

A proof is given in Zieliński (1988).

3. High quantile: definitions. We shall restrict ourselves to high quantiles; it is easy to see that the theory may be easily applied to estimation of low quantiles, i.e. quantiles of order q with q close to zero.

Definition 1. Given n, we say that x_q is a *high quantile* if no median-unbiased estimate of x_q exists, i.e. if $q > q(n) = (1/2)^{1/n}$.

Definition 2. Given q, we say that x_q is a *high quantile* if no median-unbiased estimate of x_q exists, i.e. $n < n(q) = -\log 2/\log q$.

Observe that if a median-unbiased estimate of x_q exists then also the maximally concentrated median-unbiased estimate exists so that the quantile can be optimally estimated.

By the definitions above, for the order of a high quantile x_q we have $q \ge q(n) = (1/2)^{1/n}$ (see Table 1).

Table 1

n	5	10	20	50	100
$q(n)$	0.8706	0.9331	0.9660	0.9863	0.9931
n	200	500	1000	2000	5000
$q(n)$	0.9966	0.9987	0.9993	0.9997	0.9999

On the other hand, for a given q, to construct a median-unbiased estimate of the quantile x_q a sample of size $n \ge n(q)$ is needed (see Table 2).

Table 2

q	0.9	0.95	0.99	0.999	0.9999	0.99999
$n(q)$	7	14	69	693	6932	69315

If q is close to one but $n \ge n(q)$, then the maximum concentrated median-unbiased estimator for the quantile x_q can be easily constructed as above and x_q is not a high quantile.

4. Optimal estimation of high quantiles. F transformation. If T is an estimate of the qth quantile $x_q = x_q(F)$ of an unknown distribution $F \in \mathcal{F}$, then $F(T)$ may be considered as an estimate of the known value q. The distribution of $F(T)$ is concentrated in the interval $(0,1)$ so that for every $F \in \mathcal{F}$ the expectations $E_F F(T)$ as well as the F-mean-square errors $E_F(F(T)-q)^2$ exist.

We say that T is an F-*unbiased estimate* of the qth quantile $x_q(F)$ if $E_F F(T) = q$ for all $F \in \mathcal{F}$. If $T \in \mathcal{T}$ then $F(T)$ is a pivot. For $T \in \mathcal{T}$ we have

$$E_F F(X_{J:n}) = \sum_{j=1}^{n} \lambda_j E U_{j:n} = \frac{1}{n+1} \sum_{j=1}^{n} j\lambda_j.$$

It follows that an F-unbiased estimate of the quantile of order $q \in (0,1)$ exists in $\mathcal{T}$ iff for some $\lambda_1, \ldots, \lambda_n$,

$$\frac{1}{n+1} \sum_{j=1}^{n} j\lambda_j = q,$$

i.e. iff $1/(n+1) \le q \le n/(n+1)$. For the order q of high quantiles we have $q > q(n) > n/(n+1)$ so that no F-unbiased estimate exists.

THEOREM 2. *The high quantile estimate $X_{n:n}$ is an estimate with uniformly minimum F-mean-square error $E_F(F(T) - q)^2$.*

Proof. For the F-mean-square error of an estimate $T \in \mathcal{T}$ of the qth quantile we have

$$\begin{aligned} FMSE_n(q) &= E_F(F(X_{J:n}) - q)^2 = E(U_{J:n} - q)^2 = \sum_{j=1}^{n} \lambda_j E(U_{j:n} - q)^2 \\ &= \sum_{j=1}^{n} \frac{\lambda_j \Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} \int_0^1 (x-q)^2 x^{j-1}(1-x)^{n-j}\, dx \\ &= \frac{1}{(n+1)(n+2)} \sum_{j=1}^{n} j\big(j+1-2(n+2)q\big)\lambda_j + q^2. \end{aligned}$$

It follows that the optimal $(\lambda_1^*, \ldots, \lambda_n^*)$ is that for $\lambda_{j^*} = 1$, $\lambda_j = 0$, $j \ne j^*$, with j^* that minimizes $j(j+1-2(n+2)q)$. Observe that for $q \ge (n+1/2)/(n+2)$ we have $j^* = n$ so that for high quantiles with $q \ge q(n) = (1/2)^{1/n} > (n+1/2)/(n+2)$, $X_{n:n}$ is the optimal estimate. ■

The uniformly minimum mean square error estimate is $X_{n:n}$, with $FMSE_n(q)$ given by the formula

$$FMSE_n(q) = \frac{n(n+1-2(n+2)q)}{(n+1)(n+2)} + q^2, \quad q \ge q(n).$$

Observe that, given n, $FMSE_n \nearrow 1 - n(n+3)/[(n+1)(n+2)]$ as $q \nearrow 1$, and $FMSE_n \searrow 0$ as $n \nearrow \infty$, uniformly in $q \ge q(n)$.

5. A comment.

For high quantiles in the large nonparametric model the error of estimation is measured in terms of $F(T) - q$ rather than $T - x_q(F)$. To assess the error in terms of $T - x_q(F)$, for example the bias $E_F T - x_q(F)$

or the mean square error $E_F(T - x_q(F))^2$, precise assumptions concerning the tail of model distributions are needed; the problem is that for $F \in \mathcal{F}$ no moments may exist. It seems that instead of imposing artificial conditions on tails one could consider smaller nonparametric models, for example the first order nonparametric model $\mathcal{F}_1 = \mathcal{F} \cap \{F : \int_0^1 |F^{-1}(t)|\,dt < \infty\}$ or the second order nonparametric model $\mathcal{F}_2 = \mathcal{F} \cap \{F : \int_0^1 (F^{-1}(t))^2\,dt < \infty\}$, i.e. the models with all continuous and strictly increasing distribution functions for which the first or the first and second moments exist.

References

B. Wang, S. N. Mishra, M. S. Mulekar, N. Mishra, K. Huang (2010), *Comparison of bootstrap and generalized bootstrap methods for estimating high quantiles*, J. Statist. Plann. Inference 140, 2926–2935.

D. Y. Li, L. Peng, J. P. Yang (2010), *Bias reduction for high quantiles*, ibid. 140, 2433–2441.

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

N. Markovich (2007), *Nonparametric Analysis of Univariate Heavy-Tailed Data*, Wiley.

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

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

R. Zieliński (2009), *Optimal nonparametric quantile estimates. Towards a general theory. A review*, Comm. Statist. Theory Methods 38, 980–992.

Ryszard Zieliński
Institute of Mathematics
Polish Academy of Sciences
Śniadeckich 8
P.O. Box 21
00-956 Warszawa, Poland
E-mail: rziel@impan.pl

Received on 1.7.2011;
revised version on 5.9.2011 (2096)