ESTIMATION OF NON-UNIQUE QUANTILES

By Dorian Feldman[1] and Howard G. Tucker[2]

Michigan State University and University of California, Riverside

1. Introduction and summary. This paper is concerned with consistent estimates of a quantile of a distribution function when the quantile is not unique. To be more precise, since the quantile is assumed not to be unique, we are concerned with obtaining a consistent estimate of the smallest pth quantile for a fixed $p (0 < p < 1)$, and from this procedure we can estimate the largest pth quantile. In Section 2 we consider the oscillating character and limit distribution of the sample pth quantile. Also included in this section is a precise statement of the problem to be solved. In Section three the problem of medians only is considered. Here we treat the sample median of the set of averages of all $\binom{n+1}{2}$ pairs of observations $X_1, \cdots, X_n$, which is briefly mentioned in [1]. We give a proof that this sample median converges almost surely to the center median of the original population, provided that the original distribution function is symmetric about a median. If this symmetry condition is relaxed, it is shown that this sample median of averages of pairs need not converge; and even if it did converge, it might converge to a number which is not a median of the parent distribution. In Section 4, strongly consistent estimates of the smallest pth quantile are obtained (for fixed p, $0 < p < 1$) which do not depend on the functional form of the parent distribution function, and a characterization of weakly consistent estimates is given.

2. Oscillatory effect of the pth sample quantile. Let $\{X_n\}$ be a sequence of independent identically distributed random variables with common distribution function F. For fixed $p \in (0, 1)$ we assume there exist numbers $a < b$ such that $a = \inf \{x \mid F(x) = p\}$ ann $b = \sup \{x \mid F(x) = p\}$. Let $X_{n,1} \leqq X_{n,2} \leqq \cdots \leqq X_{n,n}$ a.s. denote the order statistics of $X_1, \cdots, X_n$, and let $\hat{m}_n = X_{n,[np]}$, where $[x]$ denotes the largest integer which does not exceed x. In this section the oscillatory property and the limit distribution of $\hat{m}_n$ are obtained.

Theorem 1. *The sequence $\{\hat{m}_n\}$ obeys the oscillatory effect with respect to the interval $[a, b]$*, i.e., $P[\hat{m}_n \leqq a \text{ i.o.}] = P[\hat{m}_n \geqq b \text{ i.o.}] = 1$, *whatever be F and p.* (i.o. *means "infinitely often."*)

Proof. We prove that $P[\hat{m}_n \geqq b \text{ i.o.}] = 1$ and show that this does not depend on p or F. Thus the other conclusion will also hold. We define $\{U_n\}$ by $U_n = I_{[X_n \geqq b]} - I_{[X_n \leqq a]}$. The random variables $\{U_n\}$ are independent, and $P[U_n = 1] = 1 - p$ and $P[U_n = -1] = p$. We first establish that

$$[\textstyle\sum_{k=1}^{n} U_k \geqq n(1 - 2p)] \subset [\hat{m}_n \geqq b].$$

Received 9 August 1965; revised 2 November 1965.

[1] This research was supported in part by the National Science Foundation grant number GP-92494.

[2] This research was supported in part by the Air Force Office of Scientific Research Grant No. AF-AFOSR-62-328 and AF-AFOSR-851-65.

Indeed,

$$[\hat{m}_n \geqq b] = [\text{at least } n - [np] + 1 \text{ of the } X_i,\ 1 \leqq i \leqq n, \text{ are } \geqq b]$$
$$= [\textstyle\sum_{k=1}^{n} U_k \geqq n - 2[np] + 2] \supset [\sum_{k=1}^{n} U_k \geqq n(1 - 2p)].$$

One easily verifies that $EU_n = 1 - 2p$ and $\operatorname{Var} U_n = 4p(1 - p)$. By the law of the iterated logarithm (see Loève [2], p. 260)

$$P\left[\limsup_{n\to\infty} \frac{\sum_{k=1}^{n} U_k - n(1 - 2p)}{[8np(1 - p) \log\log 4np(1 - p)]^{\frac{1}{2}}} = 1\right] = 1.$$

Hence for any $\epsilon > 0$,

$$P[\textstyle\sum_{k=1}^{n} U_k - n(1 - 2p) > [8np(1 - p) \log\log 4np(1 - p)]^{\frac{1}{2}}(1 - \epsilon) \text{ i.o.}] = 1,$$

or

$$P[\textstyle\sum_{k=1}^{n} U_k \geqq n(1 - 2p) \text{ i.o.}] = 1.$$

Hence $P[\hat{m}_n \geqq b \text{ i.o.}] = 1$. Similarly,

$$[\textstyle\sum_{k=1}^{n} U_k \leqq n(1 - 2p) - 1] \subset [\hat{m}_n \leqq a].$$

Also, by the law of the iterated logarithm,

$$P\left[\liminf_{n\to\infty} \frac{\sum_{k=1}^{n} U_k - n(1 - 2p)}{[8np(1 - p) \log\log 4np(1 - p)]^{\frac{1}{2}}} = -1\right] = 1,$$

and we obtain, as in the above case, $P[\hat{m}_n \leqq a \text{ i.o.}] = 1$, which concludes the proof of the theorem.

Not only does $\{\hat{m}_n\}$ oscillate across $[a, b]$ infinitely often with probability one but actually $\{\hat{m}_n\}$ does not converge in probability to any constant. We can, however, show that whatever be F and p that $X_{n,[np]}$ has equal chance of being below a and of being above b. More precisely, we have

THEOREM 2. *The limiting distribution of* $X_{n,[np]}$ *is*

$$\lim_{n\to\infty} P[X_{n,[np]} \leqq x] = 0 \quad \text{if} \quad x < a$$
$$= \tfrac{1}{2} \quad \text{if} \quad a \leqq x < b$$
$$= 1 \quad \text{if} \quad x \geqq b.$$

PROOF. Let Y_n be a random variable whose distribution is binomial $B(n, F(x))$. Then

$$P[X_{n,[np]} \leqq x] = \textstyle\sum_{k=[np]}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}$$
$$= P[Y_n \geqq [np]].$$

Now

$$\{[np] - nF(x)\}/\{nF(x)(1 - F(x))\}^{\frac{1}{2}} \sim \{n^{\frac{1}{2}}(p - F(x))\}/\{F(x)(1 - F(x))\}^{\frac{1}{2}}.$$

By the Laplace-DeMoivre theorem,

$$P[X_{n,[np]} \leqq x] = \lim_{n\to\infty} P\left[\frac{Y_n - nF(x)}{[nF(x)(1-F(x))]^{\frac{1}{2}}} \geqq \frac{[np] - nF(x)}{[nF(x)(1-F(x))]^{\frac{1}{2}}}\right]$$
$$= 0, 1/2 \text{ or } 1 \text{ according as } F(x) < p, F(x) = p \text{ or } F(x) > p,$$

i.e., according as $x < a$, $a \leqq x < b$ or $x \geqq b$, which proves the theorem.

We would still get the oscillatory effect of Theorem 1 if we were to try to estimate a by $X_{n,L(n)}$ where $L(n) < [np]$ but $[np] - L(n) < \theta(8np(1-p) \log\log 4np(1-p))^{\frac{1}{2}}$ for sufficiently large n, and where $0 < \theta < 1$. Thus, even the dropping below the pth sample quantile by a relatively slowly increasing number of ordered observations does not assure us of a consistent estimate of a.

The problem then is: given p, can one estimate a and b, and what are such estimates? It is true that by the Glivenko-Cantelli theorem the empirical distribution function $\hat{F}_n$ converges uniformly to F with probability one. Hence eventually $\hat{F}_n(x)$ will be between the horizontal lines $p \pm \epsilon$, but when this will happen we do not know. Even if we knew, the curve of $F(x)$ might drop off ever so slowly to the left of a and might increase ever so slowly to the right of b, so that if n is too large, the intersections of $F_n(x)$ with $p \pm \epsilon$ would be far removed from a and b. The problem is thus not easily solvable by use of the Glivenko-Cantelli theorem and hence is non-trivial.

3. Sample medians of averages of pairs. In case the quantile that one wishes to estimate is a median, and if the distribution function F is symmetric about a median, then it is possible to find a consistent sequence of estimates of the center median $(a + b)/2$ (where now $p = \frac{1}{2}$). In this section we prove that under the hypothesis on F given above, $F * F(2x)$ has only one median μ which is the center median of F. A very easily obtained consistent estimator of μ is exhibited, and the inadequacy of this approach to the more general case is demonstrated.

LEMMA 1. *If F is symmetric about a median μ, then $F * F(2x)$ has only one median, which is equal to μ.*

PROOF. We may assume $\mu = 0$. Let X and X' be independent and identically distributed random variables with F as their common distribution function. Then $(X + X')/2$ and $(-X - X')/2$ have the same distribution which is also symmetric about 0. It is sufficient to show that for every $\eta > 0$ the event $[-2\eta \leqq X + X' \leqq 2\eta]$ has positive probability. But this event is implied by the event

$$[-b - \eta \leqq X \leqq -b + \eta] \cap [b - \eta \leqq X' \leqq b + \eta],$$

whose probability is positive because of the independence of X and X' and the definition of b. Q.E.D.

Thus there is a unique median of $F * F(2x)$, and it is the center median of F. We now show how this center median of F can be estimated. Let $X_1, X_2, \cdots, X_n, \cdots$ be a sequence of independent observations on a random variable X whose distribution function is F. Let $\hat{\mu}_n$ denote the sample median of the following $\binom{n+1}{2}$ random variables: $\{(X_i + X_j)/2, 1 \leqq i \leqq j \leqq n\}$.

THEOREM 3. *Under the hypotheses of Lemma 1, $\mu_n \to \mu$ a.s..*

PROOF. Let $\hat{F}_n$ be the empirical distribution function of $X_1, \cdots, X_n$, and

let $\hat{G}_n$ be the empirical distribution function of $\{(X_i + X_j)/2, 1 \leqq i \leqq j \leqq n\}$. It is easy to see that $\hat{F}_n * \hat{F}_n(2x)$ is the empirical distribution function of $\{(X_i + X_j)/2, 1 \leqq i \leqq n, 1 \leqq j \leqq n\}$. Let $\hat{f}_n$ be the (empirical) characteristic function of $\hat{F}_n$. The Glivenko-Cantelli theorem states that

$$P[\sup_x |\hat{F}_n(x) - F(x)| \to 0] = 1.$$

Hence $P[\hat{f}_n(u) \to f(u) \text{ at all } u] = 1$. (Note where "at all u" occurs.) Thus,

$$P[\hat{f}_n^2(u/2) \to f^2(u/2) \text{ at all } u] = 1,$$

where f is the characteristic function of F, which in turn implies (because of the continuity theorem) that

$$P[\hat{F}_n * \hat{F}_n(2x) \to F * F(2x) \text{ at all } x \text{ at which } F * F(2x) \text{ is continuous}] = 1.$$

An easy computation shows that

$$\begin{aligned} \hat{F}_n * \hat{F}_n(2x) &= n^{-2} \sum_{1 \leqq i \leqq n, 1 \leqq j \leqq n} I_{[(X_i + X_j)/2 \leqq x]} \\ &= [(n^2 + n)/n^2]\hat{G}_n(x) - (1/n)\hat{F}_n(x). \end{aligned}$$

But $(1/n)\hat{F}_n(x) \to 0$, so

$$P[\hat{G}_n(x) \to F * F(2x) \text{ at all } x \text{ at which } F * F(2x) \text{ is continuous}] = 1.$$

Since $\hat{\mu}_n$ is the sample median of $\hat{G}_n(x)$, and since $F * F(2x)$ has exactly one median, namely μ, it follows that $\hat{\mu}_n \to \mu$ with probability one. Q.E.D.

This procedure is inadequate for estimating the center median μ of F if F is not symmetric about μ. This statement is justified because: (i) if F is not symmetric about a median, then it can occur that $F * F(2x)$ has a median outside the interval of medians of F, and (ii) unlike the symmetric case, $F * F(2x)$ can have a non-degenerate interval of medians. As an example of (i) we consider $0 < a < b$, and let $\epsilon > 0$ be small, i.e., $0 < \epsilon < a$, let $c > 2b + \epsilon$ and let $0 < \delta < \frac{1}{8}$. Let F have a density f defined by

$$\begin{aligned} f(x) &= 1/2\epsilon && \text{if } x \in (a - \epsilon, a) \\ &= \delta/(c - b) && \text{if } x \in (b, c) \\ &= (\tfrac{1}{2} - \delta)/\epsilon && \text{if } x \in (c, c + \epsilon) \\ &= 0 && \text{otherwise.} \end{aligned}$$

Let X, Y be independent, identically distributed random variables with density f just given. The medians of X are all the points in $[a, b]$. We shall show that the smallest median of $(X + Y)/2$ is greater than b, and shall do this by proving $P[(X + Y)/2 > b] > \frac{1}{2}$. Since $c > 2b$ we need only show that $P[X + Y > c] > \frac{1}{2}$, or $P[X + Y \leqq c] < \frac{1}{2}$. Now it is easy to see that

$$P[X + Y \leqq c] < \tfrac{1}{4} + 2(\delta/2) + \delta^2 = (\tfrac{1}{2} + \delta)^2,$$

and since $\delta < \frac{1}{8}$, we have

$$P[X + Y \leqq c] < (\tfrac{5}{8})^2 < \tfrac{1}{2}.$$

We now construct an example to prove assertion (ii) given above. Let F have a density

$$\begin{aligned} f(x) &= 1/4\epsilon \quad \text{if} \quad x \,\varepsilon\, [0, \epsilon] \\ &= 1/4\epsilon \quad \text{if} \quad x \,\varepsilon\, [2, 2 + \epsilon] \\ &= 1/2\epsilon \quad \text{if} \quad x \,\varepsilon\, [4, 4 + \epsilon] \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

where $0 < \epsilon < \frac{1}{4}$. Then it is easily seen that $F * F$ has a triangular density over the intervals $[0, 2\epsilon]$, $[2, 2 + 2\epsilon]$, $[4, 4 + 2\epsilon]$, $[6, 6 + 2\epsilon]$ and $[8, 8 + 2\epsilon]$ and spreads over these intervals the probabilities $\frac{1}{16}$, $\frac{1}{8}$, $\frac{5}{16}$, $\frac{1}{4}$, $\frac{1}{4}$ respectively. Hence $F * F(2x)$ has triangular densities over $[0, \epsilon]$, $[1, 1 + \epsilon]$, $[2, 2 + \epsilon]$, $[3, 3 + \epsilon]$, $[4, 4 + \epsilon]$ with probabilities $\frac{1}{16}$, $\frac{1}{8}$, $\frac{5}{16}$, $\frac{1}{4}$, $\frac{1}{4}$ associated with the corresponding intervals. Thus, in this non-symmetric case $F * F(2x)$ has an interval of medians $[2 + \epsilon, 3]$ of positive length.

4. Consistent estimates of the smallest pth quantile. We have seen in Section 2 that the $[np]$th order statistic $X_{n,[np]}$ is not a consistent estimate of the smallest pth quantile, a. In this section we show that in the notation of Section 2 there does exist an order statistic $X_{n,L(n)}$, where $L(n) < [np]$, which does converge with probability 1 to the smallest pth quantile, a.

THEOREM 4. *If for some selected $K > 0$, $\delta > 0$ the sequence of integers $\{L(n)\}$ satisfies* (i) $0 < [np] - L(n) \leqq Kn^{\frac{1}{2}+\epsilon}$ *for some* ϵ, $0 < \epsilon < \frac{1}{2}$, *and* (ii) $[np] - L(n) \geqq (1 + \delta)(2n \log \log n/2)^{\frac{1}{2}}$, *then* $X_{n,L(n)} \to a$ a.s.

PROOF. The theorem will be proved when we have proved

(i) $P[X_{n,L(n)} \leqq a - \delta' \text{ i.o.}] = 0$ in Case (i) for any $\delta' > 0$, and

(ii) $P[X_{n,L(n)} > a \text{ i.o}] = 0$ in Case (ii).

We first prove (i). Let $\delta' > 0$ be arbitrary, and let $p' = P[X_1 \leqq a - \delta']$. By the definition of a and p, $p' < p$. Let us denote $V_n = I_{[X_n \leqq a-\delta']} - I_{[X_n > a-\delta']}$. Since the event $[X_{n,L(n)} \leqq a - \delta']$ is the event that at least $L(n)$ of the random variables $\{X_1, \cdots, X_n\}$ are $\leqq a - \delta'$, we have the identity

$$[X_{n,L(n)} \leqq a - \delta'] = [\textstyle\sum_{k=1}^{n} V_k \geqq 2L(n) - n].$$

Now $EV_n = 2p' - 1$ and $\operatorname{Var} V_n = 4p'(1 - p')$. By the law of the iterated logarithm,

$$P\left[\limsup_{n\to\infty} \frac{\sum_{j=1}^{n} V_j - n(2p' - 1)}{\psi(n, p')} = 1\right] = 1,$$

where $\psi(n, p') = [8np'(1 - p') \log \log 4np'(1 - p')]^{\frac{1}{2}}$. This implies that

$$P[\textstyle\sum_{k=1}^{n} V_k > n(2p' - 1) + (1 + \eta)\psi(n, p') \text{ i.o.}] = 0$$

for any $\eta > 0$. From the inequality $[np] - L(n) \leqq Kn^{\frac{1}{2}+\epsilon}$, we obtain

$$2L(n) - n \geqq n(2p' - 1) + 2n(p - p') - 2Kn^{\frac{1}{2}+\epsilon} - 2.$$

But $p - p' > 0$, so $2n(p - p') - 2Kn^{\frac{1}{2}+\epsilon} - 2$ is greater than $(1 + \eta)\psi(n, p')$ for sufficiently large n. Hence $P[\sum_{j=1}^{n} V_j \geqq 2L(n) - n \text{ i.o.}] = 0$, which proves (i). In order to prove (ii), let U_n be as in the proof of Theorem 1. Then

$$[X_{n,L(n)} > a] = [\sum_{j=1}^{n} U_j \geqq n - 2L(n) + 2],$$

and again by the law of the iterated logarithm, for any $\eta > 0$,

$$P[\sum_{j=1}^{n} U_j \geqq n(1 - 2p) + (1 + \eta)\psi(n, p) \text{ i.o.}] = 0.$$

Since for $x \in [0, 1]$, $0 \leqq (1 - x)x \leqq \frac{1}{4}$, then $(2n \log \log n)^{\frac{1}{2}} \geqq \psi(n, p)$ for all $p \in [0, 1]$. Hence, since $[np] - L(n) \geqq (1 + \delta)(2n \log \log n/2)^{\frac{1}{2}}$,

$$\begin{aligned} P[X_{n,L(n)} > a \text{ i.o.}] &\leqq [\sum_{j=1}^{n} U_j \geqq n(1 - 2p) + 2([pn] - L(n)) \text{ i.o.}] \\ &\leqq P[\sum_{j=1}^{n} U_j \geqq n(1 - 2p) + (1 + \delta)(2n \log \log n)^{\frac{1}{2}} \text{ i.o.}] \\ &\leqq P[\sum_{j=1}^{n} U_j \geqq n(1 - 2p) + (1 + \delta)\psi(n, p) \text{ i.o.}] = 0, \end{aligned}$$

which concludes the proof.

In Theorem 4 sufficient conditions on $L(n)$ were obtained in order that $X_{n,L(n)} \to a$ with probability one. For convergence in probability slightly broader conditions are both necessary and sufficient, as is shown in the following theorem.

THEOREM 5. *A necessary and sufficient condition that $X_{n,L(n)} \to a$ in probability is that* (i) $L(n)/n \to p$ *as* $n \to \infty$ *and* (ii) $n^{\frac{1}{2}}(p - L(n)/n) \to +\infty$ *as* $n \to \infty$.

PROOF: Let $b - a > \epsilon > 0$. Then $P[|X_{n,L(n)} - a| \geqq \epsilon] = P[X_{n,L(n)} \leqq a - \epsilon] + P[X_{n,L(n)} \geqq a + \epsilon] = A_n + B_n$, where

$$A_n = \sum_{k=L(n)}^{n} \binom{n}{k} F^k(a - \epsilon)(1 - F(a - \epsilon))^{n-k},$$

and $B_n = \sum_{k=0}^{L(n)-1} \binom{n}{k} p^k(1 - p)^{n-k}$. By the definition of a and b, $F(a - \epsilon) < p = F(a + \epsilon)$, so by the Laplace-De Moivre theorem,

$A_n = 1 - \Phi\{(L(n) - nF(a - \epsilon))/[nF(a - \epsilon)(1 - F(a - \epsilon))]^{\frac{1}{2}}\} + o(1)$, and

$B_n = \Phi\{(L(n) - 1 - np)/[np(1 - p)]^{\frac{1}{2}}\} + o(1)$, where Φ is the normal distribution function with mean zero and variance one. We first prove that Conditions (i) and (ii) are sufficient. Condition (ii) immediately implies that $B_n \to 0$ as $n \to \infty$. Both conditions imply, for sufficiently large n, that $L(n)/n \in (F(a - \epsilon), p]$, and thus (i) implies that $A_n \to 0$. Thus $X_{n,L(n)} \to a$ in probability. In order to prove that the conditions are necessary we assume $A_n \to 0$ and $B_n \to 0$ as $n \to \infty$. But $B_n \to 0$ immediately implies that $(L(n) - np)/n^{\frac{1}{2}} = n^{\frac{1}{2}}(L(n)/n - p) \to -\infty$, which proves that (ii) is necessary. Since $A_n \to 0$ as $n \to \infty$, we have, for every $p' < p$, that $(L(n) - np')/n^{\frac{1}{2}} = n^{\frac{1}{2}}(L(n)/n - p') \to \infty$. Hence $L(n)/n > p - \epsilon$ for every $\epsilon > 0$ and all large n. Also, for large n, $L(n) < np$ because of Condition (ii) which was already proved necessary. From these two observations, Condition (i) follows. Q.E.D.

It should be noticed that Conditions (i) and (ii) of Theorem 5 are equivalent to requiring that $p - L(n)/n = g(n)/n^{\frac{1}{2}}$, where g is some function such that $g(n)/n^{\frac{1}{2}} \to 0$ and $g(n) \to \infty$ as $n \to \infty$. Hence for weak consistency there is a slight improvement over both Conditions (i) and (ii) of Theorem 4 in an obvious way.

5. Acknowledgment. The authors wish to express their gratitude to the referee for supplying a much simpler proof of a more general Lemma 1 and for a number of other very helpful comments.

REFERENCES

[1] HODGES, J. L. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

2] LOÈVE, M. (1963). *Probability Theory* (*3rd ed.*). Van Nostrand, Princeton, New Jersey.