A NOTE ON QUANTILES IN LARGE SAMPLES[1]

By R. R. Bahadur

University of Chicago

1. Introduction. Let $F(x)$ be a probability distribution function on the real line. Let ξ be a fixed point and let

$$F(\xi) = p. \tag{1}$$

It is assumed that F has at least two derivatives in some neighborhood of ξ, that $F''(x)$ is bounded in the neighborhood, and that $F'(\xi) = f(\xi) > 0$. These assumptions imply, in particular, that $0 < p < 1$ and that ξ is the unique p-quantile of F.

Let $\omega = (X_1, X_2, \cdots \text{ ad inf})$ be a sequence of independent random variables X_i with each X_i distributed according to F. For each $n = 1, 2, \cdots$, let $Y_n = Y_n(\omega)$ be the sample p-quantile when the sample is $(X_1, \cdots, X_n)$. Let $Z_n = Z_n(\omega)$ be the number of observations X_i in the sample $(X_1, \cdots, X_n)$ such that $X_i > \xi$. This note points out that, with $q = 1 - p$,

$$Y_n(\omega) = \xi + [(Z_n(\omega) - nq)/n \cdot f(\xi)] + R_n(\omega) \tag{2}$$

where R_n becomes negligible as $n \to \infty$. It is shown here that

$$R_n(\omega) = O(n^{-3/4} \log n) \quad \text{as} \quad n \to \infty \tag{3}$$

with probability one, but the exact order of R_n is not known at present.

The above representation of Y_n gives new insight into the well known result that $n^{\frac{1}{2}}(Y_n - \xi)$ is asymptotically normally distributed with mean 0 and variance $v = pq/f^2(\xi)$. It gives an easy access, via the multivariate central limit theorem for zero-one variables, to the asymptotic joint distribution of several quantiles in samples from a multivariate distribution [2]. The representation also shows that the law of the iterated logarithm holds for quantiles, i.e.,

$$\begin{aligned} &\limsup_{n\to\infty} [n^{\frac{1}{2}}(Y_n - \xi)/(2 \log \log n)^{\frac{1}{2}}] = v^{\frac{1}{2}}, \\ &\liminf_{n\to\infty} [n^{\frac{1}{2}}(Y_n - \xi)/(2 \log \log n)^{\frac{1}{2}}] = -v^{\frac{1}{2}} \end{aligned} \tag{4}$$

with probability one.

The proof in the following section may be outlined as follows. Let $F_n(x, \omega)$ be the sample distribution function when the sample is $(X_1, \cdots, X_n)$, i.e., $F_n(x, \omega) = (\text{The number of } X_i \leqq x \text{ in the sample})/n$. It is shown that, with I_n a suitable neighborhood of ξ, $F_n(x, \omega) \doteq F_n(\xi, \omega) + F(x) - F(\xi)$ uniformly

Received 3 January 1966.

[1] This research was supported in part by Research Grant No. NSF GP 3707 from the Division of Mathematical Physical and Engineering Sciences of the National Science Foundation, and in part by the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

for x in I_n, and that Y_n is in I_n for all sufficiently large n. Hence $p \doteq F_n(Y_n, \omega) \doteq F_n(\xi, \omega) + F(Y_n) - F(\xi) \doteq F_n(\xi, \omega) + (Y_n - \xi)f(\xi)$, so $Y_n \doteq \xi + (Z_n - nq)/nf(\xi)$.

2. Proof. Let

$$G_n(x, \omega) = [F_n(x, \omega) - F_n(\xi, \omega)] - [F(x) - F(\xi)]. \tag{5}$$

Let $\{a_n : n = 1, 2, \cdots\}$ be a sequence of positive constants such that

$$a_n \sim (\log n)/n^{\frac{1}{2}} \quad \text{as} \quad n \to \infty. \tag{6}$$

Let $I_n = (\xi - a_n, \xi + a_n)$, and let

$$H_n(\omega) = \sup \{|G_n(x, \omega)| : x \text{ in } I_n\}. \tag{7}$$

Lemma 1. *With probability one, $H_n(\omega) = O(n^{-3/4} \log n)$ as $n \to \infty$.*

Proof. Let $\{b_n : n = 1, 2, \cdots\}$ be a sequence of positive integers such that

$$b_n \sim n^{\frac{1}{4}} \quad \text{as} \quad n \to \infty. \tag{8}$$

Consider a particular n. For any integer r, let $\eta_{r,n} = \xi + a_n b_n^{-1} r$, let $J_{r,n}$ denote the interval $[\eta_{r,n}, \eta_{r+1,n}]$, and let $\alpha_{r,n} = F(\eta_{r+1,n}) - F(\eta_{r,n})$. Since F_n and F are non-decreasing in x, it is plain from (5) that, for x in $J_{r,n}$,

$$\begin{aligned} G_n(x, \omega) &\leqq F_n(\eta_{r+1,n}, \omega) - F_n(\xi, \omega) - F(\eta_{r,n}) + F(\xi) \\ &= G_n(\eta_{r+1,n}, \omega) + \alpha_{r,n}. \end{aligned}$$

Similarly, for x in $J_{r,n}$, $G_n(x, \omega) \geqq G_n(\eta_{r,n}, \omega) - \alpha_{r,n}$. It follows hence from (7) that

$$\begin{aligned} H_n(\omega) &\leqq \max \{|G_n(\eta_{r,n}, \omega)| : -b_n \leqq r \leqq b_n\} \\ &\qquad + \max \{\alpha_{r,n} : -b_n \leqq r \leqq b_n - 1\} \\ &= K_n(\omega) + \beta_n \qquad \text{say.} \end{aligned} \tag{9}$$

Since $\eta_{r+1,n} - \eta_{r,n} = a_n b_n^{-1}$ for each r, since $|\eta_{r,n} - \xi| \leqq a_n$ for $|r| \leqq b_n$, and since F is sufficiently smooth in a fixed neighborhood of ξ, it follows from (6) and (8) that $\beta_n = O(n^{-3/4} \log n)$. In view of (9), it will therefore suffice to show that if $c_1 > 0$ is sufficiently large, and if $\gamma_n = c_1 n^{-3/4} \log n$ for $n = 1, 2, \cdots$ then

$$\textstyle\sum_n P(K_n \geqq \gamma_n) < \infty. \tag{10}$$

To establish (10) we will use the following inequality due to S. N. Bernstein. For any n and any z, $0 \leqq z \leqq 1$, let $B(n, z)$ denote a random variable such that $P(B(n, z) = r) = \binom{n}{r} z^r (1 - z)^{n-r}$ for $r = 0, 1, \cdots, n$. Then

$$P(|B(n, z) - nz| \geqq t) \leqq 2 \exp(-h) \tag{11}$$

for all $t > 0$, where

$$h = h(n, z, t) = t^2/\{2[nz(1 - z) + (t/3) \max\{z, 1 - z\}]\}. \tag{12}$$

For a proof of this version of Bernstein's inequality see [3], pp. 204–205, where a generalization of (11)–(12) is given. See [1] for other generalizations, and for certain closer bounds.

Choose and fix $c_2 > F'(\xi)$. Let N be an integer so large that $F(\xi + a_n) - F(\xi) < c_2 a_n$ and $F(\xi) - F(\xi - a_n) < c_2 a_n$ for all $n > N$. We see from (5) that, for any n and r, the probability distribution of $|G_n(\eta_r, \omega)|$ is the same as that of $n^{-1}|B(n, z) - nz|$ with $z = |F(\eta_{r,n}) - F(\xi)| = z_{r,n}$ say. Consequently, $P(|G_n(\eta_r)| \geqq \gamma_n) \leqq 2 \exp(-h_n(r))$ by (11), where $h_n(r) = h(n, z_{r,n}, n\gamma_n)$ is given by (12). Since $h(n, z, t) \geqq t^2/2[nz + t]$, and since $n > N$ and $|r| \leqq b_n$ imply $z_{r,n} \leqq c_2 \cdot a_n$, it follows that

$$P(|G_n(\eta_r, \omega)| \geqq \gamma_n) \leqq 2 \exp(-\delta_n) \tag{13}$$

for $n > N$ and $|r| \leqq b_n$, where $\delta_n = n^2\gamma_n^2/2[c_2 \cdot na_n + n\gamma_n]$. Since δ_n does not depend on r, it follows from (9) and (13) that $P(K_n \geqq \gamma_n) \leqq 4b_n \exp(-\delta_n) = \lambda_n$ say, for $n > N$. It follows easily from (6) and (8) by the definitions of γ_n, δ_n, and λ_n that

$$\log \lambda_n/\log n \to \tfrac{1}{4} - (c_1^2/2c_2) \tag{14}$$

as $n \to \infty$. The limit in (14) is less than -1 if, given c_2, c_1 is chosen sufficiently large; then $\sum_n \lambda_n < \infty$ and (10) holds. This completes the proof.

Let $\{k_n : n = 1, 2, \cdots\}$ be a sequence of positive integers such that $1 \leqq k_n \leqq n$ for each n and

$$k_n = np + o(n^{\frac{1}{2}} \log n) \quad \text{as} \quad n \to \infty. \tag{15}$$

For each n let $U_{n1} \leqq \cdots \leqq U_{nn}$ be the sample values $X_1, \cdots, X_n$ arranged in ascending order, and let

$$V_n(\omega) = U_{nk_n}. \tag{16}$$

In other words, V_n is the k_nth order statistic in the sample $(X_1, \cdots, X_n)$.

Lemma 2. *With probability one, V_n is in I_n for all sufficiently large n.*

Proof. For each n, $P(V_n \leqq \xi - a_n) = P(B(n, z_n) \geqq k_n)$ where $z_n = F(\xi - a_n)$. An upper bound for $P(V_n \leqq \xi - a_n)$ may therefore be obtained by putting $z = z_n$ and $t = t_n = k_n - nz_n$ in (11) and (12), provided $t_n > 0$. Since $z_n = F(\xi) - a_n f(\xi) + o(a_n)$, and $f(\xi) > 0$, it follows from (1), (6) and (15) that $t_n \sim f(\xi)n^{\frac{1}{2}} \log n$ as $n \to \infty$. Consequently, $h_n = h(n, z_n, t_n) \sim c_3 (\log n)^2$ by (12), where $c_3 = f^2(\xi)/2pq > 0$, so that $\sum_n \exp(-h_n) < \infty$. Thus $\sum_n P(V_n \leqq \xi - a_n) < \infty$. A similar argument shows that $\sum_n P(V_n \geqq \xi + a_n) < \infty$, and this completes the proof.

Lemma 3. *With probability one,*

$$V_n(\omega) = \xi + \{[k_n - nF_n(\xi, \omega)]/nf(\xi)\} + O(n^{-3/4} \log n) \tag{17}$$

as $n \to \infty$.

Proof. Choose and fix an ω such that V_n is in I_n for all sufficiently large n.

Let $N = N(\omega)$ and c_4 be such that, for all $n > N$, V_n is in I_n, and $F''(x)$ exists and $\frac{1}{2}|F''(x)| \leqq c_4$ for all x in I_n.

We may suppose that, for $n > N$, $F_n(V_n, \omega) = k_n/n$. It follows hence from (5) and (7) that, for $n > N$,

$$(18) \qquad k_n/n = F_n(\xi, \omega) + F(V_n) - F(\xi) + \theta_n(\omega)\cdot H_n(\omega)$$

where $|\theta_n| \leqq 1$. We observe next that, for $n > N$, $F(V_n) = F(\xi) + (V_n - \xi)f(\xi) + c_4\cdot\varphi_n(\omega)\cdot a_n^2$ where $|\varphi_n| \leqq 1$. It follows hence from (18) that $k_n/n = F_n(\xi, \omega) + (V_n - \xi)f(\xi) + \zeta_n$ where $\zeta_n(\omega) = O(\max\{a_n^2, H_n(\omega)\})$. It is thus plain from (6) that (17) holds with probability one.

Let $[np]$ be the integral part of np, and let $\psi_n = np - [np]$, $0 \leqq \psi_n < 1$. For $n > 1/p$ let $k_n^{(1)} = [np]$ and $k_n^{(2)} = k_n^{(1)} + 1$, and let $V_n^{(i)}$ be determined by $k_n^{(i)}$ according to (16), $i = 1, 2$. Then (17) holds for $V_n^{(i)}$ and $k_n^{(i)}$, $i = 1, 2$. Since $Y_n = (1 - \psi_n)V_n^{(1)} + \psi_n V_n^{(2)}$ for $n > 1/p$, and since $k_n^{(i)} = np + O(1)$ for $i = 1, 2$, it follows that (3) holds for R_n defined by (2).

As noted in Section 1, (2) and (3) imply (4). It follows from (4) that the best choice of $I_n = (\xi - a_n, \xi + a_n)$ in the preceding proof is not given by (6) but by $a_n \sim c_5(2n^{-1}\log\log n)^{\frac{1}{2}}$ with $c_5 > v^{\frac{1}{2}}$. By repeating the arguments of this section for the revised I_n (but omitting the now redundant Lemma 2) it is easily seen that in fact $R_n = O(n^{-3/4}l_n)$ where $l_n = (\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}}$. This however is not a substantial improvement or clarification of (3).

REFERENCES

[1] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *JASA* **58** 13–30.

[2] Siddiqui, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *J. Res. Nat. Bur. Standards Sect. B* **64** 145–150.

[3] Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.