The Annals of Mathematical Statistics
1971, Vol. 42, No. 6, 1957–1961

A NEW PROOF OF THE BAHADUR REPRESENTATION OF QUANTILES AND AN APPLICATION

BY J. K. GHOSH

Indian Statistical Institute

1. Introduction and summary. Let $\{X_i\}$ be a sequence of independent random variables with the same distribution function $F(x) = \Pr\{X_i \leqq x\}$. Let $F(M_p) = p, 0 < p < 1$. Suppose F has two derivatives in a neighborhood of M_p, $F'(x)$ is bounded there and $F'(M_p)$ is positive. Let $Y_{p,n}$ be a sample p-quantile based on $X_1, \cdots, X_n$. Let $nG_n(x)$ be the number of X_i's among $(X_1, \cdots, X_n)$ which are $> x$. Bahadur (1966) has proved

$$Y_{p,n} = M_p + [G_n(M_p) - (1-p)]/F'(M_p) + R_n \tag{1}$$

where the remainder term R_n becomes negligible as $n \to \infty$. More precisely, he has shown $R_n = O(n^{-\frac{1}{4}} \log n)$ a.s. as $n \to \infty$. The best result of this type is due to Kiefer (1967) who has calculated the exact order of R_n. Sen (1968) has extended Bahadur's result to random variables which are neither independent nor identically distributed.

We shall give a new and much simpler proof of a weaker version of Bahadur's result which suffices for many statistical applications. Our proof involves fewer assumptions than Bahadur's.

For arbitrary p_n let M_{p_n} be defined as $M_p + (p_n - p)/F'(M_p)$. Consider

$$Y_{p_n,n} = M_{p_n} + [G_n(M_p) - (1-p)]/F'(M_p) + R_n \tag{2}$$

where $Y_{p_n,n}$ is a sample p_n-quantile.

In Section 2 we have proved the following result about R_n.

THEOREM 1. *Suppose $F'(M_p)$ exists and is strictly positive and $p_n - p = O(1/n^{\frac{1}{2}})$. Then R_n as defined in* (2) (*and, a fortiori, R_n as defined in* (1) *satisfies*

$$n^{\frac{1}{2}} R_n \to 0 \quad \textit{in probability.} \tag{3}$$

(After writing this paper the author discovered that the result for $p_n = p$ is stated without proof in Chernoff et al (1967).)

It is easy to extend this result as in Sen (1968). An outline is sketched in one of the remarks. Once again it is possible to achieve some economy in assumptions.

The representation (1) is not new. Its use in deriving the asymptotic moments of $Y_{p,n}$ goes back to Karl Pearson. See, for example, (1) in Hojo (1931). But the formulation therein is very imprecise and lacks a rigorous justification.

We next consider an application of Theorem 1. Let $\bar{X}_n = (\sum_1^n X_i)/n$ and $P_n =$ proportion of X_i's above $\bar{X}_n$. David (1962) proved the asymptotic normality of P_n when F is a normal distribution function. Using the same elegant trick, Mustafi

Received June 12, 1969; revised February 17, 1971.

(1968) has proved a similar result for bivariate normal distributions. We shall extend these results considerably by providing alternative proofs based on Theorem 1, which dispense with the normality assumption on F. Moreover, in our proof we may consider—though we shall not do so for purposes of simplicity—instead of the sample mean $\bar{X}_n$ an U-statistic to which the central limit theorem of Hoeffding (1948) applies.

2. Proof of Theorem 1. Let Ω, A, P be the probability space on which all the random variables are defined.

LEMMA 1. *Let $\{V_n\}$, $\{W_n\}$ be two sequences of random variables satisfying the following conditions.*

(4) *For all* $\delta > 0$ *there exists* λ *(depending on* δ*)* s.t. $P(|W_n| > \lambda) < \delta$.

(5) *For all* k *and all* $\varepsilon > 0$

$$\lim_{n\to\infty} P(V_n \leqq k, W_n \geqq k+\varepsilon) = 0$$
$$\lim_{n\to\infty} P(V_n \geqq k+\varepsilon, W_n \leqq k) = 0.$$

Then $V_n - W_n \to 0$ *in probability as* $n \to \infty$.

The lemma is quite easy to prove but for the sake of completeness we include a proof.

PROOF OF LEMMA 1. It follows from (5) that

(6) $\lim P(a \leqq W_n \leqq b, V_n \geqq b+\varepsilon \text{ or } V_n \leqq a-\varepsilon) = 0$ for all $a < b$, $\varepsilon > 0$. Choose λ as in (4). Let $I_j = [a_j, b_j]$, $j = 1, \cdots, m$ be m intervals such that $a_1 = -\lambda$, $b_m = \lambda$ and $b_j - a_j < \varepsilon$. Then

$$P(|V_n - W_n| > 2\varepsilon) \leqq \delta + \sum P(a_j \leqq W_n \leqq b_j, |V_n - W_n| > 2\varepsilon).$$

The lemma is an immediate consequence of this inequality and (6).

Let $F_n(x) = 1 - G_n(x)$ be the sample distribution function and $G(x) = 1 - F(x)$.

We first prove the theorem when $Y_{p_n,n}$ is the smallest sample p_n-quantile, i.e. $Y_{p_n,n}$ is the smallest y for which $F_n(y-0) \leqq p_n \leqq F_n(y)$. Then for any real number t,

$$\{w; n^{\frac{1}{2}}(Y_{p_n,n}(w) - M_{p_n}) \leqq t\} = \{w; p_n \leqq F_n(M_{p_n} + t/n^{\frac{1}{2}})\}$$

$$= \left\{w; n^{\frac{1}{2}} \frac{\{G_n(M_{p_n} + t/n^{\frac{1}{2}}) - G(M_{p_n} + t/n^{\frac{1}{2}})\}}{F'(M_p)} \leqq t_n\right\} \tag{7}$$

where $t_n = n^{\frac{1}{2}}(F(M_{p_n} + t/n^{\frac{1}{2}}) - p_n)/F'(M_p)$.

Clearly $t_n = n^{\frac{1}{2}}\{F(M_p) + \{t/n^{\frac{1}{2}} + (p_n - p)/F'(M_p)\}(F'(M_p) + o(1)) - p_n\}/F'(M_p)$ and so tends to t as $n \to \infty$.

Let $Z_{t,n} = n^{\frac{1}{2}}\{G_n(M_{p_n} + t/n^{\frac{1}{2}}) - G(M_{p_n} + t/n^{\frac{1}{2}})\}/F'(M_p)$.

Let $W_n = n^{\frac{1}{2}}\{G_n(M_p) - G(M_p)\}/F'(M_p)$.

Easy calculation shows $E(Z_{t,n}-W_n)^2 = p_n(1-p_n)/\{F'(M_p)\}^2$ where $p_n = |F(M_p) - F(M_{p_n}+t/n^{\frac{1}{2}}| \to 0$.

So,

$$E(Z_{t,n}-W_n)^2 \to 0. \tag{8}$$

It follows that

$$Z_{t,n}-W_n \to 0 \quad \text{in probability.} \tag{9}$$

Also W_n has an asymptotic normal distribution. So if we let $V_n = n^{\frac{1}{2}}(Y_{p_n,n}-M_{p_n})$, then by (7) and (9) V_n and W_n satisfy the conditions of the lemma. Thus $V_n - W_n \to 0$ in probability and the theorem is proved when $Y_{p_n,n}$ is the smallest p_n-quantile.

Let $Y'_{p_n,n}$ denote an arbitrary p_n-quantile and $Y_{p_n,n}$ the smallest, as above. Then $Y_{p_n,n} \leqq Y'_{p_n,n} \leqq Y_{p_n',n}$ where $p_n' = p_n + 1/n$ also satisfies the hypothesis of the theorem. Then

$$\begin{aligned} n^{\frac{1}{2}}(Y_{p_n,n}-M_{p_n})-W_n &\leqq n^{\frac{1}{2}}(Y'_{p_n}-M_{p_n})-W_n \\ &\leqq n^{\frac{1}{2}}(Y_{p_n'}-M_{p_n})-W_n. \end{aligned}$$

The theorem follows from the special case proved since $n^{\frac{1}{2}}(M_{p_n}-M_{p_n'}) \to 0$.

REMARKS. (i) If $F'(M_p) > 0$ and F has a continuous first derivative in a neighborhood around M_p then the theorem is true for all $p_n \to p$. The proof is exactly similar but M_{p_n} is redefined so that $F(M_{p_n}) = p_n$.

(ii) To see how we may extend as in Sen (1968) and weaken his assumptions, suppose X_i's are independent but not necessarily identically distributed random variables. Let $\bar{F}_n(x) = 1/n \sum_{i=1}^n P\{X_i \leqq x\}$ and $\bar{G}_n(x) = 1-\bar{F}_n(x)$. Suppose $\bar{F}_n(M_{p,n}) = p \; \forall \; n$ and the following two conditions hold.

CONDITION 1. $\dfrac{\bar{F}_n(M_{p,n}+t/n^{\frac{1}{2}})-\bar{F}_n(M_{p,n})}{t/n^{\frac{1}{2}}} \to_{\forall t} c > 0$ (independent of t).

CONDITION 2. $T_n = n^{\frac{1}{2}}\{G_n(M_{p,n})-(1-p)\}$ is asymptotically normal.

Then, taking $Y_{p,n}$ to be the smallest p-quantile, we have $n^{\frac{1}{2}}(Y_{p,n}-M_{p,n})-T_n/c \to 0$ in probability. The proof is quite similar. (For the step corresponding to (8) note that $1/n \sum p_i(1-p_i) \leqq \bar{p}(1-\bar{p})$ where $\bar{p} = 1/n \sum_1^n p_i$). If we further assume $\bar{F}_n$ is continuous in $(M_{p,n}-\varepsilon,\ M_{p,n}+\varepsilon)$, $\varepsilon > 0$ then the previous assertion holds for all p-quantiles. To see this consider the biggest p-quantile $Y'_{p,n}$, the set $\{w;\ n^{\frac{1}{2}}(Y'_{p,n}-M_{p,n}) < t\}$ and construct the proof as before through steps similar to (7), (8) and (9).

3. An application of Theorem 1. Let $\{X_i\}$ be a sequence of i.i.d. random variables with common distribution function $F(x)$. We make the following assumptions on F. The first two moments of F exist and without loss of generality $E(X_i) = 0$. $F'(0)$ exists and is positive. We shall say F satisfies Assumption A if all the preceding conditions hold.

Let $p = F(0)$, $q = 1-p$. Let

$$\begin{aligned} u_i &= 1 \quad \text{if} \quad X_i \leqq 0; \\ &= 0 \quad \text{if} \quad X_i > 0. \end{aligned}$$

Then $F_n(0) = 1/n \sum u_i$. Let $T_i = (p - U_i) - X_i F'(0)$. Then the T_i's are i.i.d. with mean 0 and finite variance. Let $\bar{T}_n = (\sum_1^n T_i)/n$. We recall that $P_n = 1 - F_n(\bar{X}_n)$.

THEOREM 2. *Let F satisfy Assumption A. Then $n^{\frac{1}{2}}(P_n - q) - n^{\frac{1}{2}}\bar{T}_n$ converges to zero in probability and hence $n^{\frac{1}{2}}(P_n - q)$ is asymptotically normal with zero mean and variance equal to $E(T_i^2)$.*

Before proving Theorem 2 we observe that since $n^{\frac{1}{2}}\bar{X}_n$ is asymptotically normal, plim $n^{\frac{1}{2}}[P_n - q - T_n] = 0$ is equivalent to plim $[(F_n(\bar{X}_n) - F_n(0))/\bar{X}_n - F'(0)] = 0$. This should make clear the intuitive content of Theorem 2.

PROOF OF THEOREM 2. Now

$$\{w; n^{\frac{1}{2}}(P_n(W) - q) \leqq t\} = \{w; p - t/n^{\frac{1}{2}} \leqq F_n(\bar{X}_n)\}$$
$$= \{w;\ Y_{p_n,n} \leqq \bar{X}_n\}$$

where $p_n = p - t/n^{\frac{1}{2}}$ and $Y_{p_n,n}$ is the smallest p_n-quantile,

$$= \{w; n^{\frac{1}{2}}F'(0)(Y_{p_n,n} - M_{p_n}) - \bar{X}_n) \leqq t\} \tag{10}$$

since $M_{p_n} = M_p + (p_n - p)/F'(0) = t/(n^{\frac{1}{2}}F'(0))$.
But by Theorem 1, $n^{\frac{1}{2}}\{F'(0)(Y_{p,n} - M_{p_n}) - (G_n(0) - q)\}$ converges in probability to zero. Using (10) and applying the lemma (with $V_n = n^{\frac{1}{2}}(P_n - q)$ and $W_n = n^{\frac{1}{2}}T_n$) we get Theorem 2.

If $(X_i,\ Y_i)$ are i.i.d. random vectors and their marginal distribution functions $F_1(x)$, $F_2(y)$ satisfy Assumption A then Theorem 1 may be applied to both P_n and Q_n = proportion of Y_i's above $\bar{Y}_n$. This shows the asymptotic joint distribution of P_n, Q_n is bivariate normal. It is easy to give a sufficient condition for the non-singularity of this normal distribution. If we define V_i similarly to U_i in terms of Y_i then it is easy to show that if there exists a linear relation between $u_i - X_i F_1'(0)$ and $V_i - Y_i F_2'(0)$ then the joint distribution function $F(x, y)$ of $(X_i,\ Y_i)$ is concentrated on at most four lines. Thus if in particular, $F(x, y)$ is not singular with respect to the two dimensional Lebesgue measure then the asymptotic joint distribution of P_n and Q_n is a non-singular normal distribution. Mustafi's theorem (1968) is a special case

Acknowledgment. I wish to thank Dr. C. K. Mustafi for telling me of his result before publication. The referee's suggestions improved the presentation.

REFERENCES

BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

CHERNOFF, H., GASTWIRTH, J. L. and JOHNS, M. V. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.* **38** 52–72.

DAVID, H. T. (1962). The sample mean among the moderate order statistics. *Ann. Math. Statist.* **33** 1160–1166.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

HOJO, J. (1931). Distribution of the median, quartiles and interquartile distance in sample from a normal population. *Biometrika* **23** 315–360.

KIEFER, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* 38 1323–1342.

MUSTAFI, C. K. (1968). On the proportion of observations above sample means in a bivariate normal distribution. *Ann. Math. Statist.* 39 1350–1353.

SEN, P. K. (1968). Asymptotic normality of sample, quantiles for m-dependent processes. *Ann. Math. Statist.* 39 1724–1730.