The Annals of Statistics
1973, Vol. 1, No. 3, 446-453

THE MEAN-SQUARE ERROR OF BAHADUR'S ORDER-STATISTIC APPROXIMATION

By D. L. Duttweiler

Bell Laboratories

Under the assumption of a uniform parent distribution, we find an exact expression for the mean-square error of an order-statistic approximation suggested by R. R. Bahadur. From this result we obtain an asymptotic formula for the mean-square error when the parent distribution is not necessarily uniform by transforming with the inverse of the cumulative distribution function of the parent.

1. Introduction. Let $X_1, X_2, \cdots$ be a sequence of independent, finite-mean-square, random variables on the real line each with the same cumulative distribution function $F(x) = \Pr\{X_m \leqq x\}$. For all integers m and n with $m \leqq n$ define $X_{(m):n}$ (or simply $X_{(m)}$) as the mth smallest member of the set $\{X_1, \cdots, X_n\}$ or, in other words, as the mth order statistic in a sample of size n. Bahadur (1966) has suggested approximating $X_{(m):n}$ by (results will be stated informally in this introductory section)

$$\hat{X}_{(m):n} = \xi + (Z_n - nq)/(nf(\xi)) \tag{1}$$

where

$$f(x) = F'(x) = \frac{d}{dx} F(x)\,,$$

$$p = m/(n+1)\,, \qquad q = 1 - p\,,$$

ξ is such that $F(\xi) = p$, and Z_n is the number of observations X_i in the set $\{X_1, \cdots, X_n\}$ that are greater than ξ. He has shown that if m and n increase together maintaining the relationship

$$p \doteq m/(n+1)\,,$$

then

$$R_n = X_{(m):n} - \hat{X}_{(m):n}$$

is $O(n^{-\frac{3}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})$ almost surely. Since Bahadur's initial work, Kiefer (1967) (see also, Kiefer (1970)) has found the exact order of R_n.

For many applications of Bahadur's approximation, however, knowing the almost sure order of R_n is not sufficient and an estimate of the size of R_n for finite n is needed. The engineering problem that led to our interest in Bahadur's work provides an example. To simulate a proposed system for concentrating digitally encoded speech channels, we needed to know the statistics of a discrete

Received April 1972; revised August 1972.
AMS 1970 *subject classifications*. Primary 62G30; Secondary 33A15, 60F99.
Key words and phrases. Order statistics, approximation, incomplete Beta function.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

random process $T(k)$, defined at each time k as the mth smallest of the values at time k of n independent, stationary, discrete random processes $\{X_1(\cdot), \cdots, X_n(\cdot)\}$. In other words, we needed the statistics of

$$T(k) = X_{(m):n}(k) . \tag{2}$$

Results of Mood (1941) and Siddiqui (1960) showed that $T(k)$ was asymptotically (m and n increasing together and maintaining the relationship $p = m/(n+1)$) Gaussian with mean ξ and covariance

$$C(k) = p(F_k(\xi \mid \xi) - p)/(nf^2(\xi)) \tag{3}$$

with $F_k(\xi \mid \xi)$ equal to the probability $X_m(l+k) \leqq \xi$ given that $X_m(l) \leqq \xi$ (independent of l and m). Unfortunately, neither Mood's nor Siddiqui's methods of derivation suggested a way of estimating the accuracy of (3) for finite n. Bahadur's approximation did. The covariance (3) is easily shown to be the covariance of

$$\hat{T}(k) = \hat{X}_{(m):n}(k) . \tag{4}$$

Thus an estimate of $E[R_n^2]$ provides an estimate of the accuracy of (3) for finite n.

In this paper we show that under certain smoothness conditions on $F(x)$

$$E[R_n^2] \doteq n^{-\frac{3}{2}} f^{-2}(\xi)(2pq/\pi)^{\frac{1}{2}} . \tag{5}$$

This result is consistent with a result of Kiefer (1967, Equation 1.6), who showed that $n^{\frac{3}{2}} f(\xi) R_n$ converged in distribution to a distribution with mean zero and variance $(2pq/\pi)^{\frac{1}{2}}$. The procedure we shall follow in deriving (5) is to first find an exact formula for $E[R_n^2]$ for the case $F(x) = x$, $x \in (0, 1)$, and then extend this result for arbitrary $F(x)$ by transforming with F^{-1}.

2. Exact results for a uniform parent.

THEOREM 1. *Let $U_1, U_2, \cdots, U_n$ be independent random variables each distributed uniformly on $(0, 1)$ and let $U_{(1)}, \cdots, U_{(n)}$ denote their order statistics. For any $m \leqq n$ define*

$$\begin{aligned} p &= m/(n+1) , \qquad q = 1 - p , \\ Z &= \text{the number of} \quad U_i \geqq p \\ \hat{U}_{(m)} &= p + (Z - nq)/n \end{aligned} \tag{6}$$

and

$$R = U_{(m)} - \hat{U}_{(m)} .$$

Then

$$\begin{aligned} E[R^2] &= 2n^{-1}p(I_p(m, n+1-m) - I_p(m+1, n+1-m)) \\ &\quad - 2pqn^{-1}(n+2)^{-1} \end{aligned} \tag{7}$$

$$\leqq (2/n)(pq/(n+2))^{\frac{1}{2}} \tag{8}$$

where $I_x(a, b)$ is the incomplete Beta function and is defined by

$$I_x(a, b) = \frac{(a + b - 1)!}{(a - 1)!\,(b - 1)!} \int_0^x \lambda^{a-1}(1 - \lambda)^{b-1}\, d\lambda .$$

PROOF. We have

$$\begin{aligned} E[R^2] &= E[U_{(m)} - (p + (Z - nq)/n)]^2 \\ &= E[(U_{(m)} - p) - (Z - nq)/n]^2 \\ &= E[U_{(m)} - p]^2 + E[Z - nq]^2/n^2 - 2E[(U_{(m)} - p)(Z - nq)]/n . \end{aligned} \tag{9}$$

It is well known (see, for example, David (1970), page 28) that $U_{(m)}$ has mean p and variance $pq/(n + 2)$. Since Z is a binomial (n, q) variable it has mean nq and variance npq. Using these facts, Equation (9) can be rewritten as

$$E[R^2] = pq/(n + 2) + pq/n - 2E[U_{(m)}Z]/n + 2pq . \tag{10}$$

The difficult step is evaluating $E[U_{(m)}Z]$. The random variable Z is the number of $U_i \geqq p$ or, equivalently, the number of order statistics $U_{(i)} \geqq p$. Since the conditional distribution of the order statistics $\{U_{(1)}, \cdots, U_{(m-1)}\}$ given $U_{(m)}$ is that of the order statistics in a sample of size $m - 1$ from the uniform distribution over $(0, U_{(m)})$ and the conditional distribution of the order statistics $\{U_{(m+1)}, \cdots, U_{(n)}\}$ given $U_{(m)}$ is that of the order statistics in a sample of size $n - m$ from the uniform distribution over $(U_{(m)}, 1)$ we have[1]

$$\begin{aligned} E[Z \mid U_{(m)}] &= \begin{cases} (n - m)q/(1 - U_{(m)}) , & U_{(m)} \leqq p \\ (n - m + 1) + (m - 1)(U_{(m)} - p)/U_{(m)} , & U_{(m)} > p \end{cases} \\ &= \begin{cases} (n - m)q/(1 - U_{(m)}) , & U_{(m)} \leqq p \\ n - (m - 1)p/U_{(m)} , & U_{(m)} > p . \end{cases} \end{aligned} \tag{11}$$

Letting $1(\cdot)$ denote the function that is one whenever its argument is nonnegative and zero otherwise, we can rewrite (11) as

$$\begin{aligned} E[Z \mid U_{(m)}] &= \frac{(n - m)q}{1 - U_{(m)}}\, 1(p - U_{(m)}) + \left(n - \frac{(m - 1)p}{U_{(m)}}\right) 1(U_{(m)} - p) \\ &= n - \frac{(m - 1)p}{U_{(m)}} + \left(\frac{(n - m)q}{1 - U_{(m)}} + \frac{(m - 1)p}{U_{(m)}} - n\right) 1(p - U_{(m)}) . \end{aligned} \tag{12}$$

We have

$$\begin{aligned} E[U_{(m)}Z] &= E[U_{(m)}E[Z \mid U_{(m)}]] \\ &= E[nU_{(m)} - (m - 1)p] \\ &\quad + E\left[\left(\frac{(n - m)qU_{(m)}}{1 - U_{(m)}} + (m - 1)p - nU_{(m)}\right) 1(p - U_{(m)})\right] \\ &= p(n - m + 1) + E[(U_{(m)} - p)1(p - U_{(m)})] \\ &\quad + E\left[\left(\frac{(n - m)qU_{(m)}}{1 - U_{(m)}} + mp - (n + 1)U_{(m)}\right) 1(p - U_{(m)})\right] . \end{aligned} \tag{13}$$

[1] The idea of evaluating $E[U_{(m)}Z]$ by first finding $E[Z \mid U_{(m)}]$ was suggested to us by a referee and substantially shortened our original proof.

Remembering that the probability density of $U_{(m)}$ is

$$\frac{n!}{(m-1)!\,(n-m)!}\,u_{(m)}^{m-1}(1-u_{(m)})^{n-m} \tag{14}$$

and integrating, it is straightforward to show that the last expectation in (13) is equal to

$$-m(I_p(m+1, n-m+1) - pI_p(m, n-m+1) - qI_p(m+1, n-m)),$$

which in turn by Equation 6.6.5 of Abramowitz and Segun (1964) equals zero. Therefore,

$$E[U_{(m)}Z] = p(n-m+1) - E[(p-U_{(m)})1(p-U_{(m)})]. \tag{15}$$

Substituting (15) in (10) and algebraically simplifying the resulting expression gives

$$E[R^2] = 2n^{-1}E[(p-U_{(m)})1(p-U_{(m)})] - 2pqn^{-1}(n+2)^{-1}. \tag{16}$$

Integrating the density (14) of $U_{(m)}$ with $(p-U_{(m)})$ over $(0, p)$, we obtain

$$\begin{aligned} E[(p-U_{(m)})1(p-U_{(m)})] & \\ &= p(I_p(m, n+1-m) - I_p(m+1, n+1-m)), \end{aligned} \tag{17}$$

which substituted into (16) gives (7). The inequality (8) is obtained from the relationship

$$\begin{aligned} E[(p-U_{(m)})1(p-U_{(m)})] &\leqq E\,|U_{(m)}-p| \\ &\leqq (E(U_{(m)}-p)^2)^{\frac{1}{2}} \\ &= (pq/(n+2))^{\frac{1}{2}}. \end{aligned} \tag{18}$$

3. Arbitrary parent.

THEOREM 2. *Let $X_1, X_2, \cdots$ be independent, finite-mean-square, random variables on the real line each with the same cumulative distribution function $F(x) = \Pr\{X_l \leqq x\}$, and for all positive integers m and n with $m \leqq n$ let $X_{(m):n}$ denote the* m*th smallest member of the set $\{X_1, \cdots, X_n\}$. Let ξ be an arbitrary constant on the real line and assume*

(i) *$f(x) = F'(x) = (d/dx)F(x)$ exists and is strictly positive at $x = \xi$, and*
(ii) *$F''(x)$ exists in a neighborhood of ξ.*

Then, for any $\delta > 0$, if $m = m(n)$ is such that the limit, as n increases, of $m/(n+1)$ exists and equals $F(\xi)$,

$$(E[R_n^2])^{\frac{1}{2}} = n^{-\frac{3}{4}}f^{-1}(\xi)(2pq/\pi)^{\frac{1}{4}} + o(n^{-1+\delta/2}) \tag{19}$$

where

$$\begin{aligned} &p = F(\xi), \quad q = 1-p, \quad R_n = X_{(m):n} - \hat{X}_{(m):n}, \\ &\hat{X}_{(m):n} = \xi_n + (Z_n - nq_n)/(nf(\xi_n)), \quad Z_n = \textstyle\sum_{l=1}^{n} 1(X_l - \xi_n), \\ &p_n = m(n)/(n+1), \quad q_n = 1-p_n, \quad \xi_n = Q(p_n), \end{aligned}$$

and the function $Q(\cdot)$ is defined by

$$Q(u) = \sup\{x : F(x) \leqq u\}$$

($Q(\cdot) = F^{-1}(\cdot)$ *if this inverse exists*).

REMARK 1. The theorem is still true if ξ_n and q_n in the definitions of $\hat{X}_{(m):n}$ and Z_n are replaced by ξ and q. The theorem is stated in the form we feel will be most useful for applications.

REMARK 2. If it is also assumed that $F'''(x)$ exists in a neighborhood of ξ, then it can be shown that (24) is true with $o(n^{-1+\delta/2})$ replaced by $O(n^{-1})$. The proof of this stronger result is essentially the same as the following proof. Using the notation of the following proof, the function $H_n(U_{(m):n})$ must be expanded as $(\frac{1}{2})H_n''(p_n)(U_{(m):n} - p_n)^2 + G_n(U_{(m):n})$ and $E[G_n^2(U_{(m):n})]$ bounded by a procedure similar to that used here to bound $E[H_n^2(U_{(m):n})]$.

PROOF. Let $U_1, U_2, \cdots$ be independent random variables distributed uniformly on (0, 1). It is not difficult to show that for any l, $Q(U_l)$ has the cumulative distribution function $F(\cdot)$ and, moreover, since $Q(\cdot)$ is nondecreasing, that for all n the collection $\{X_1, \cdots, X_n, X_{(1):n}, \cdots, X_{(n):n}\}$ has the same joint statistics as the collection $\{Q(U_1), \cdots, Q(U_n), Q(U_{(1):n}), \cdots, Q(U_{(n):n})\}$. Thus

$$R_n \approx Q(U_{(m):n}) - \xi_n - n^{-1}(f(\xi_n))^{-1} \sum_{l=1}^{n} (1(Q(U_l) - \xi_n) - q_n) \tag{20}$$

where $\approx$ denotes identical distributions.

Conditions (i) and (ii) imply the existence of an open interval A containing ξ in which $F''(x)$ exists and $f(x) = F'(x)$ exists and is strictly positive. Let $B = \{u : u = F(x), x \in A\}$. Then $F(\cdot)$ restricted to A is invertible and $Q(\cdot)$ restricted to B is its inverse.

Since $p = \lim p_n$ and p is in the open interval B, there must exist N such that for all $n \geqq N$, $p_n \in B$. For $n \geqq N$,

(i) $Q'(p_n)$ exists and equals $(f(\xi_n))^{-1}$,
(ii) $Q''(p_n)$ exists,
(iii) $Q(U_l) \geqq \xi_n$ if and only if $U_l \geqq p_n$.

Define

$$H_n(U_{(m):n}) = Q(U_{(m):n}) - Q(p_n) - Q'(p_n)(U_{(m):n} - p_n)\,. \tag{21}$$

Using (21) and (iii) in (20), we have

$$\begin{aligned} R_n &\approx Q(p_n) + Q'(p_n)(U_{(m):n} - p_n) + H_n(U_{(m):n}) \\ &\quad - \xi_n - \frac{1}{nf(\xi_n)} \sum_{l=1}^{n} (1(U_l - p_n) - q_n) \\ &= H_n(U_{(m):n}) + (f(\xi_n))^{-1}(U_{(m):n} - \hat{U}_{(m):n}) \end{aligned} \tag{22}$$

where $\hat{U}_{(m):n}$ is as in Theorem 1.

From Theorem 1, we have

$$E(U_{(m):n} - \hat{U}_{(m):n})^2 = 2n^{-1}p_n(I_{p_n}(m, n+1-m) - I_{p_n}(m+1, n+1-m)) + O(n^{-2}) . \tag{23}$$

Since $\lim f(\xi_n) = f(\xi)$, the proof will be complete if we can show

$$2n^{-1}p_n(I_{p_n}(m, n+1-m) - I_{p_n}(m+1, n+1-m)) = n^{-\frac{3}{2}}(2pq/\pi)^{\frac{1}{2}} + o(n^{-2+\delta}) \tag{24}$$

and

$$E[H_n^2(U_{(m):n})] = o(n^{-2+\delta}) . \tag{25}$$

Let L denote the left-hand side of (24). We have from Abramowitz and Segun (1964, Equation 26.5.16)

$$L = \frac{2}{n} p_n \frac{n!}{m!\,(n-m)!} p_n^m (1-p_n)^{n+1-m} = \frac{2}{n} \frac{m}{n+1} \frac{n!}{m(m-1)!\,(n-m)!} \left(\frac{m}{n+1}\right)^m \left(\frac{n+1-m}{n+1}\right)^{n+1-m} . \tag{26}$$

Using Stirling's rule (see, for example, Abramowitz and Segun (1964), Equation 6.1.37) to approximate the factorials in (26), it is straightforward to show that

$$L = n^{-\frac{3}{2}}(2pq/\pi)^{\frac{1}{2}} + O(n^{-5/2}) , \tag{27}$$

which is stronger than (24).

The problem of bounding $E[H_n^2(U_{(m):n})]$ is considered in Section 5.4 of Blom (1958) and Section 3.2 of Van Zwet (1964). Unfortunately, our conditions and desired conclusions are such that we cannot use the results in either of these references directly.

Let

$$\varepsilon_n = n^{-1/2+\delta/5}$$

and $I_n = (p_n - \varepsilon_n, p_n + \varepsilon_n)$. Denoting the probability density function of $U_{(m):n}$ by $g_n(\cdot)$, we have

$$E[H_n^2(U_{(m):n})] = \int_0^1 H_n^2(u) g_n(u)\, du = \left(\int_{u \in I_n} + \int_{u \notin I_n}\right) H_n^2(u) g_n(u)\, du . \tag{28}$$

Let

$$H_{n,\max} = \sup_{u \in I_n} |H_n(u)| \tag{29}$$

and

$$g_{n,\max} = \sup_{u \notin I_n} g_n(u) . \tag{30}$$

Then

$$E[H_n^2(U_{(m):n})] \leqq H_{n,\max}^2 \int_{u \in I_n} g_n(u)\, du + g_{n,\max} \int_{u \notin I_n} H_n^2(u)\, du \leqq H_{n,\max}^2 + g_{n,\max} \int_0^1 H_n^2(u)\, du . \tag{31}$$

Since

$$H_n(u) = Q(u) - Q(p_n) - Q'(p_n)(u - p_n)\,, \tag{32}$$

we have, using the fact that the square of the sum of three numbers is less than three times the sum of the squares of the numbers,

$$\begin{aligned}(\tfrac{1}{3})\int_0^1 H_n^2(u)\,du &\leqq \int_0^1 Q^2(u)\,du + \int_0^1 Q^2(p_n)\,du + \int_0^1 (Q'(p_n))^2(u - p_n)^2\,du \\ &\leqq E[Z_l^2] + Q^2(p_n) + (Q'(p_n))^2\,.\end{aligned} \tag{33}$$

Since the Z_l are of finite mean-square, since $Q(\cdot)$ and $Q'(\cdot)$ are continuous over B, and since $p_n \in B$ for $n \geqq N$, there exists a constant C_1 such that

$$\int_0^1 H_n^2(u)\,du \leqq C_1\,, \qquad n \geqq N\,. \tag{34}$$

Therefore, for $n > N$

$$E[H_n^2(U_{(m):n})] \leqq H_{n,\max}^2 + C_1 g_{n,\max}\,. \tag{35}$$

Let M be an integer greater than N such that I_n is properly contained in B for $n \geqq M$. Letting C_2 denote the maximum of $|Q''(u)|$ in I_M, we have by Taylor's theorem

$$H_{n,\max} \leqq C_2\varepsilon_n^2/2\,, \qquad n \geqq M\,. \tag{36}$$

The density $g_n(u_{(m):n})$ has its mode p_n^* at $(m-1)/(n-1)$ and decreases monotonically on both sides. Since

$$\begin{aligned}|p_n - p_n^*| &= \left|\frac{m}{n+1} - \frac{m-1}{n-1}\right| \\ &= |q_n - p_n|/(n-1) \leqq \varepsilon_n\,,\end{aligned}$$

p_n^* will exist in I_n for all n. Let k be an arbitrary positive integer. We have

$$\begin{aligned}E(U_{(m):n} - p_n)^{2k} &= \int_0^1 (u - p_n)^{2k} g_n(u)\,du \\ &\geqq \int_{p_n^*}^{p_n+\varepsilon_n} (u - p_n)^{2k} g_n(u)\,du \\ &\geqq g_n(p_n + \varepsilon_n)\int_{p_n^*}^{p_n+\varepsilon_n} (u - p_n)^{2k}\,du \\ &= g_n(p_n + \varepsilon_n)(2k+1)^{-1}(\varepsilon_n^{2k+1} - (p_n^* - p_n)^{2k+1})\,.\end{aligned} \tag{37}$$

Since $\varepsilon_n = n^{-\frac{1}{2}+\delta/5}$ and $p_n^* - p_n = O(n^{-1})$, there must exist a constant $C_3 > 0$ such that for all n

$$E(U_{(m):n} - p_n)^{2k} \geqq C_3 g_n(p_n + \varepsilon_n)\varepsilon_n^{2k+1}\,,$$

or equivalently,

$$g_n(p_n + \varepsilon_n) \leqq E(U_{(m):n} - p_n)^{2k} C_3^{-1}\varepsilon_n^{-(2k+1)}\,.$$

Similarly, it can be shown that there exists a constant $C_4 > 0$ such that

$$g_n(p_n - \varepsilon_n) \leqq E(U_{(m):n} - p_n)^{2k} C_4^{-1}\varepsilon_n^{-(2k+1)}\,.$$

Letting C_5 denote the minimum of C_3 and C_4, we have

$$g_{n,\max} \leqq E(U_{(m):n} - p_n)^{2k} C_5^{-1}\varepsilon_n^{-(2k+1)}\,. \tag{38}$$

Since (see Blom (1958) or Van Zwet (1964)) there exists C_6 such that

$$E(U_{(m):n} - p_n)^{2k} \leqq C_6 n^{-k}, \tag{39}$$

we have

$$g_{n,\max} \leqq C_6 C_5^{-1} n^{-k} \varepsilon_n^{-(2k+1)}. \tag{40}$$

Combining (35), (36), and (40), we obtain

$$\begin{aligned} E[H_n^2(U_{(m):n})] &\leqq 4^{-1} C_2^2 \varepsilon_n^4 + C_1 C_6^{-1} \varepsilon_n^{-(2k+1)} n^{-k} \\ &= 4^{-1} C_2^2 n^{-2+4\delta/5} + C_1 C_6 C_5^{-1} n^{\frac{1}{2}-2\delta k/5-\delta/5}. \end{aligned} \tag{41}$$

Since k is arbitrary, we can choose it to be such that

$$2\delta k/5 + \delta/5 - \tfrac{1}{2} \geqq 2,$$

proving the theorem.

REFERENCES

[1] ABRAMOWITZ, M. and SEGUN, I. A. (1964). *Handbook of Mathematical Functions*. Dover, New York.

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

[3] BLOM, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Wiley, New York.

[4] DAVID, H. A. (1970). *Order Statistics*. Wiley, New York.

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

[6] KIEFER, J. (1970). Deviations between the sample quantile process and the sample df. *Nonparametric Techniques in Statistical Interference* (M. L. Puri, ed.). Cambridge Univ. Press. 299–319.

[7] MOOD, A. M. (1941). On the joint distribution of the medians in samples from a multivariate population. *Ann. Math. Statist.* **12** 268–278.

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

[9] VAN ZWET, W. R. (1964). *Convex Transformations of Random Variables*. Mathematical Center Tracts 7, Mathematisch Centrum, Amsterdam.

BELL LABORATORIES
HOLMDEL, NEW JERSEY 07733