# ON THE DISTRIBUTION OF THE EXPECTED VALUES OF THE ORDER STATISTICS[1]

By Wassily Hoeffding

*University of North Carolina*

**Summary.** Let $X_1, X_2, \cdots, X_n$ be independent with a common distribution function $F(x)$ which has a finite mean, and let $Z_{n1} \leqq Z_{n2} \leqq \cdots \leqq Z_{nn}$ be the ordered values $X_1, \cdots, X_n$. The distribution of the $n$ values $EZ_{n1}, \cdots, EZ_{nn}$ on the real line is studied for large $n$. In particular, it is shown that as $n \to \infty$, the corresponding distribution function converges to $F(x)$ and any moment of that distribution converges to the corresponding moment of $F(x)$ if the latter exists. The distribution of the values $Ef(Z_{nm})$ for certain functions $f(x)$ is also considered.

**1. Introduction and statement of results.** Let $X_1, X_2, \cdots, X_n, \cdots$ be mutually independent random variables with a common (cumulative) distribution function $F(x)$. Let $Z_{n1} \leqq Z_{n2} \leqq \cdots \leqq Z_{nn}$ be the ordered values $X_1, X_2, \cdots, X_n$. It will be assumed that

$$\int_{-\infty}^{\infty} |x| \, dF(x) < \infty, \tag{1}$$

which implies that the expected values $EZ_{n1}, EZ_{n2}, \cdots, EZ_{nn}$ exist. (Throughout this paper the statement that an expected value exists will imply that it is finite.) The distribution which assigns equal weights to the $n$ values $EZ_{n1}, \cdots, EZ_{nn}$ will be referred to as the distribution of the $EZ_{nm}$, and its distribution function will be denoted by $F_n(x)$. The primary object of this paper is to show that this distribution approximates the distribution represented by $F(x)$ when $n$ is large. More precisely, the following will be proved.

Theorem 1. *Suppose that* (1) *is satisfied and let $g(x)$ be a real-valued, continuous function such that*

$$|g(x)| \leqq h(x), \tag{2}$$

*where the function $h(x)$ is convex and*

$$\int_{-\infty}^{\infty} h(x) \, dF(x) < \infty. \tag{3}$$

*Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} g(EZ_{nj}) = \int_{-\infty}^{\infty} g(x) \, dF(x). \tag{4}$$

Received 8/12/52.

[1] Work done under the sponsorship of the Office of Naval Research.

The assumption that $h(x)$ is convex is understood in the sense that for any two real numbers $x$, $y$

$$h(ax + (1 - a)y) \leqq ah(x) + (1 - a)h(y) \qquad \text{if } 0 < a < 1.$$

With $g(x) = \cos tx$ and $\sin tx$, Theorem 1 implies that the characteristic function of the distribution of the $EZ_{nm}$ converges to that of $X_j$ as $n \to \infty$, and hence $F_n(x) \to F(x)$ for all points of continuity of $F(x)$. With $g(x) = x^k$, $k > 0$, we obtain that the moment of order $k$ of the distribution of the $EZ_{nm}$ converges to the corresponding moment of $F(x)$ if the latter exists.

If $f(x)$ is a function such that $Ef(X_j)$ exists, we can, more generally, consider the distribution of $Ef(Z_{n1}), \cdots, Ef(Z_{nn})$. If $f(x)$ is a strictly monotone function, Theorem 1 can be applied in an obvious way. The general case will not be considered, but the following special result will be obtained as a simple consequence of Theorem 1.

THEOREM 2. *Let $f(x)$ be convex, $g(x)$ convex and nondecreasing (for $x \geqq A$ if $f(y) \geqq A$* for all *$y$), and suppose that*

$$\int_{-\infty}^{\infty} x \, dF(x), \qquad \int_{-\infty}^{\infty} f(x) \, dF(x) \qquad \textit{and} \qquad \int_{-\infty}^{\infty} g(f(x)) \, dF(x)$$

*exist. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} g(Ef(Z_{nj})) = \int_{-\infty}^{\infty} g(f(x)) \, dF(x).$$

Theorem 2 and the indicated modification of Theorem 1 apply, in particular, to the case where $f(x)$ and $g(x)$ are powers of $x$.

The behavior of the distributions of the $EZ_{nm}$ and the $Ef(Z_{nm})$ is of interest in connection with certain rank order tests. It has been shown by Hoeffding [4] and Terry [6] that rank order tests for testing a hypothesis of randomness which are most powerful against certain alternatives are based on statistics of the form $c(R) = \sum_{j=1}^{n} a_j Ef(Z_{nR_j})$, where $R = (R_1, \cdots, R_n)$ is the vector of the ranks of the observations and $f(x)$ is a given function. If all permutations of the ranks are equally probable, the moments of $c(R)$ are functions of the power sums $\sum_{j=1}^{n} [Ef(Z_{nj})]^k$. Theorems 1 and 2 give asymptotic expressions for these power sums. Tests of this type were already considered by Fisher and Yates [2] whose tables XX and XXI give the values of $EZ_{nj}$ and the (approximate) values of $\sum_{j=1}^{n} (EZ_{nj})^2$ for $n \leqq 50$ when $F(x)$ is normal with mean 0 and variance 1. Dwass [1] and Terry [6] use results implied by Theorems 1 and 2 to study the asymptotic distributions of statistics of the form $c(R)$.

**2. Preliminaries.** The distribution function of $Z_{nm}$ will be denoted by $F_{nm}(x)$. Since $Z_{nm} \leqq x$ if and only if at least $m$ of the values $X_1, \cdots, X_n$ are $\leqq x$, we have

$$F_{nm}(x) = \sum_{j=m}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}$$

$$= \frac{n!}{(m-1)!(n-m)!} \int_0^{F(x)} t^{m-1}(1-t)^{n-m} \, dt. \tag{5}$$

The following three facts, which are known or easily verified, will be used in the sequel.

I. If $Ef(X_1)$ exists, so does $Ef(Z_{nm})$ for all $n$, $m$.

II. $\sum_{m=1}^{n} Ef(Z_{nm}) = nEf(X_1)$.

III. (Cf. Jensen [5].) If $h(x)$ is convex and $U$ is a random variable such that $EU$ and $Eh(U)$ exist, we have $h(EU) \leqq Eh(U)$.

Repeated use will be made of the following Lemma 1, which is an immediate consequence of an extension by Fréchet and Shohat [3] of a theorem of Helly.

LEMMA 1. *Let $V(x)$, $V_n(x)$, $n = 1, 2, \cdots$, be a sequence of functions which are uniformly bounded and of uniformly bounded variation on any finite interval, such that $\lim_{n\to\infty} V_n(x) = V(x)$ for all $x$, with the possible exception of a countable set. Let $f(x)$ be a continuous function such that*

$$\int_{-\infty}^{\infty} f(x)\, dV(x) \qquad \text{and} \qquad \int_{-\infty}^{\infty} f(x)\, dV_n(x), \qquad n = 1, 2, \cdots$$

*exist and*

$$\lim_{A\to\infty} \int_{|x|>A} f(x)\, dV_n(x) = 0$$

*uniformly with respect to $n$. Then*

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)\, dV_n(x) = \int_{-\infty}^{\infty} f(x)\, dV(x).$$

**3. Proofs.** Theorem 1 will be proved with the help of several lemmas.

LEMMA 2. *Given $\epsilon > 0$, there exist two numbers $C$ and $a$, where $0 < a < 1$, such that for every $n \geqq 2$*

$$F_{nm}(x) \leqq Ca^n F(x) \qquad \text{if} \qquad F(x) + \epsilon \leqq \frac{m-1}{n-1} \leqq 1, \tag{6}$$

$$1 - F_{nm}(x) \leqq Ca^n[1 - F(x)] \qquad \text{if} \qquad 0 \leqq \frac{m-1}{n-1} \leqq F(x) - \epsilon. \tag{7}$$

PROOF. Let $s = (m-1)/(n-1)$, $v = F(x)$,

$$H(s, v) = \frac{\int_0^v [t^s(1-t)^{1-s}]^{n-1}\, dt}{\int_0^1 [t^s(1-t)^{1-s}]^{n-1}\, dt}.$$

Then inequalities (6) and (7) can be written as

$$H(s, v) \leqq Ca^n v \qquad \text{if } v + \epsilon \leqq s \leqq 1, \tag{8}$$

$$1 - H(s, v) \leqq Ca^n(1 - v) \qquad \text{if } 0 \leqq s \leqq v - \epsilon. \tag{9}$$

For $s$ arbitrarily fixed, $0 \leqq s \leqq 1$, the function $t^s(1-t)^{1-s}$ increases for $0 < t < s$ and decreases for $s < t < 1$. Hence the quantity

$$2b = \min_{\epsilon \leqq s \leqq 1} [s^s(1 - s)^{1-s} - (s - \epsilon)^s(1 - s + \epsilon)^{1-s}],$$

where $s^s(1 - s)^{1-s} = 1$ if $s = 0$ or 1, is positive. We have for $v \leqq s - \epsilon$

$$(10) \qquad \int_0^v [t^s(1 - t)^{1-s}]^{n-1}\, dt \leqq [(s - \epsilon)^s(1 - s + \epsilon)^{1-s}]^{-n-1} v$$
$$\leqq [s^s(1 - s)^{1-s} - 2b]^{n-1}\, v.$$

On the other hand, we can choose a positive number $d$ so that for every $s$, $0 \leqq s \leqq 1$,

$$s^s(1 - s)^{1-s} - t^s(1 - t)^{1-s} \leqq b \qquad \text{if } |t - s| \leqq d.$$

Then we have

$$(11) \qquad \int_0^1 [t^s(1 - t)^{1-s}]^{n-1}\, dt \geqq \int_{\substack{|t-s| \leqq d \\ 0 \leqq t \leqq 1}} [t^s(1 - t)^{1-s}]^{n-1}\, dt$$
$$\geqq d[s^s(1 - s)^{1-s} - b]^{n-1}.$$

From (10) and (11) we have for $v + \epsilon \leqq s \leqq 1$

$$(12) \qquad H(s, v) \leqq d^{-1}[K(s)]^{n-1} v,$$

where

$$(13) \qquad K(s) = \frac{s^s(1 - s)^{1-s} - 2b}{s^s(1 - s)^{1-s} - b} \leqq \frac{1 - 2b}{1 - b}.$$

If we put $a = (1 - 2b)/(1 - b)$ and $C = d^{-1}a^{-1}$, inequality (8) follows from (12) and (13).

Inequality (9) is obtained from (8) by observing that $1 - H(s, v) = H(1 - s, 1 - v)$. This completes the proof.

The following Lemmas 3 and 4 are immediately obtained from Lemma 2.

LEMMA 3. *If $m/n \to c$ as $n \to \infty$, then*

$$\lim_{n\to\infty} F_{nm}(x) = \begin{cases} 0 & \text{if } F(x) < c \\ 1 & \text{if } F(x) > c. \end{cases}$$

LEMMA 4. *If $m/n \to c$ as $n \to \infty$, where $0 < c < 1$, there exist two numbers $N$ and $d > 0$ such that for $n > N$*

$$F_{nm}(x) \lneqq F(x) \qquad \text{if } F(x) < d,$$
$$1 - F_{nm}(x) \leqq 1 - F(x) \qquad \text{if } 1 - F(x) < d.$$

Let $S$ be the set on the real line which consists of all points of discontinuity of $F(x)$ and all points $x$ such that $F(x - h) < F(x) < F(x + h)$ for every $h > 0$.

LEMMA 5. *Let $y \in S$, $0 < a < 1$. If $m/n \to aF(y - 0) + (1 - a)F(y + 0)$ as $n \to \infty$, then*

$$(14) \qquad \lim_{n\to\infty} EZ_{nm} = y.$$

PROOF. By Lemma 1 it suffices to show that

$$\lim_{n\to\infty} F_{nm}(x) = \begin{cases} 0 & \text{if } x < y \\ 1 & \text{if } x > y, \end{cases} \tag{15}$$

and that

$$\lim_{A\to\infty} \int_{|x|>A} x\, dF_{nm}(x) = 0 \text{ uniformly with respect to } n. \tag{16}$$

Let $c = aF(y - 0) + (1 - a)F(y + 0)$. Since $y \in S$, the inequalities $x < y < z$ imply $F(x) < c < F(z)$. Hence (15) follows from Lemma 3.

The assumptions $y \in S$, $0 < a < 1$ imply that $0 < c < 1$. Let $d$ and $N$ be defined as in Lemma 4. Given $\epsilon > 0$, choose $B > 0$ so that $F(-B) < d$, $1 - F(B) < d$,

$$-\int_{-\infty}^{-B} x\, dF(x) < \frac{\epsilon}{2}, \qquad \int_{B}^{\infty} x\, dF(x) < \frac{\epsilon}{2},$$

and $F(x)$ and $F_{nm}(x)$ are continuous at $x = \pm B$. Then

$$-\int_{-\infty}^{-B} x\, dF_{nm}(x) = BF_{nm}(-B) + \int_{-\infty}^{-B} F_{nm}(x)\, dx. \tag{17}$$

Applying Lemma 4, we have that for $n > N$ the right-hand side of (17) does not exceed

$$BF(-B) + \int_{-\infty}^{-B} F(x)\, dx = -\int_{-\infty}^{-B} x\, dF(x).$$

Hence if $n > N$, $-\int_{-\infty}^{-B} x\, dF_{nm}(x) < \epsilon/2$ and, similarly $\int_{-\infty}^{-B} x\, dF_{nm}(x) < \epsilon/2$. This implies (16). The proof is complete.

Let

$$G_{nm}(x) = \frac{1}{n} \sum_{j=1}^{m} F_{nj}(x). \tag{18}$$

LEMMA 6. *If $m/n \to c$ as $n \to \infty$, then*

$$\lim_{n\to\infty} G_{nm}(x) = \begin{cases} F(x) & \textit{if } F(x) < c \\ c & \textit{if } F(x) > c. \end{cases}$$

PROOF. By (5) and (18),

$$\begin{aligned} nG_{nm}(x) &= \sum_{j=1}^{m} \sum_{k=j}^{n} \binom{n}{k} F(x)^k [1 - F(x)]^{n-k} \\ &= \sum_{k=1}^{m} k \binom{n}{k} F(x)^k [1 - F(x)]^{n-k} + m \sum_{k=m+1}^{n} \binom{n}{k} F(x)^k [1 - F(x)]^{n-k}, \end{aligned}$$

whence

$$G_{nm}(x) = F(x)[1 - F_{n-1,m}(x)] + \frac{m}{n} F_{n,m+1}(x) \quad \text{if} \quad m < n \tag{19}$$

and $G_{nn}(x) = F(x)$. Lemma 6 now follows from Lemma 3.

From (19) and Lemma 4 we easily obtain

LEMMA 7. *If $m/n \to c$ as $n \to \infty$, where $0 < c < 1$, there exist two numbers $N$ and $d > 0$ such that for $n > N$*

$$G_{nm}(x) \leqq 2F(x) \quad \textit{if} \quad F(x) < d,$$

$$\frac{m}{n} - G_{nm}(x) \leqq 1 - F(x) \quad \textit{if} \quad 1 - F(x) < d.$$

LEMMA 8. *If $g(x)$ satisfies the conditions of Theorem 1 and $m/n \to F(y)$ as $n \to \infty$, where $y$ is a point of continuity of $F(x)$, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{m} Eg(Z_{nj}) = \int_{-\infty}^{y} g(x)\, dF(x). \tag{20}$$

PROOF. Equation (20) can be written in the form

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g(x)\, dG_{nm}(x) = \int_{-\infty}^{y} g(x)\, dF(x). \tag{21}$$

By Lemma 1 it suffices to show that

$$\lim_{n\to\infty} G_{nm}(x) = \begin{cases} F(x) & \text{if} \quad x < y \\ F(y) & \text{if} \quad x > y \end{cases} \tag{22}$$

for every $x$ at which $F(x)$ is continuous and that

$$\lim_{A\to\infty} \int_{|x|>A} g(x)\, dG_{nm}(x) = 0 \quad \text{uniformly with respect to } n. \tag{23}$$

For every $y$ which is a point of continuity of $F(x)$ we can choose two numbers $y_1$, $y_2$ in $S$ and two numbers $a_1$, $a_2$ in (0, 1) such that if we let

$$c_i = a_i F(y_i - 0) + (1 - a_i) F(y_i + 0), \qquad i = 1, 2,$$

we have $c_1 \leqq F(y) \leqq c_2$ and $c_2 - c_1$ is arbitrarily small. Now choose $m_1 \leqq m$ and $m_2 \geqq m$ in such a way that $m_1/n \to c_1$ and $m_2/n \to c_2$ as $n \to \infty$. Since $G_{nm_1}(x) \leqq G_{nm}(x) \leqq G_{nm_2}(x)$, (22) now follows from Lemma 6.

To prove (23), we may assume without loss of generality that the function $h(x)$ of Theorem 1 is nonincreasing for $-x$ sufficiently large and nondecreasing for $x$ sufficiently large. Then (23) follows from

$$\left| \int_{|x|>A} g(x)\, dG_{nm}(x) \right| \leqq \int_{|x|>A} h(x)\, dG_{nm}(x)$$

and Lemma 7 in a similar way as in the proof of (16). This completes the proof of Lemma 8.

Let

$$H_n(y) = \frac{1}{n} \sum_{EZ_{nj} \leqq y} EZ_{nj},$$

$$H(y) = \int_{-\infty}^{y} x \, dF(x).$$

LEMMA 9. *If $y$ is a point of continuity of $F(x)$, $\lim_{n\to\infty} H_n(y) = H(y)$.*

PROOF. We can write $H_n(y) = n^{-1} \sum_{j=1}^{m} EZ_{nj}$, where $m = m(y)$ is determined by

$$EZ_{nm} \leqq y < EZ_{n,m+1}. \tag{24}$$

This implies $m/n \to F(y)$. For otherwise a subsequence $\{m'/n'\}$ of $\{m/n\}$ must converge to a number $v \neq F(y)$. If $v < F(y)$, we can choose $x \in S$ and $a$ in (0, 1) so that $v \leqq c < F(y)$, where $c = aF(x - 0) + (1 - a)F(x + 0)$. To every $(m', n')$ we can choose an integer $m'' \geqq m'$ so that $m''/n' \to c$. By Lemma 5 this implies $x = \lim_{n'\to\infty} EZ_{n',m''+1}$, hence $\limsup EZ_{n',m'+1} \leqq x < y$, which contradicts (24). In a similar way the assumption $v > F(y)$ leads to a contradiction.

Lemma 9 now follows from Lemma 8 with $g(x) = x$.

LEMMA 10. *If $g(x)$ satisfies the conditions of Theorem 1, we have*

$$\lim_{A\to\infty} \int_{|x|>A} g(x) \, dF_n(x) = 0$$

*uniformly with respect to $n$.*

PROOF. If $A$ is a point of continuity of $F(x)$,

$$\left| \int_{A}^{\infty} g(x) \, dF_n(x) \right| \leqq \int_{A}^{\infty} h(x) \, dF_n(x) = \frac{1}{n} \sum_{j=m}^{n} h(EZ_{nj}),$$

where $EZ_{n,m-1} \leqq A < EZ_{nm}$. As shown in the proof of Lemma 9, $m/n \to F(A)$ as $n \to \infty$. Since $h(x)$ is convex, $n^{-1} \sum_{j=m}^{n} h(EZ_{nj}) \leqq n^{-1} \sum_{j=m}^{n} Eh(Z_{nj})$. By Lemma 8 the right-hand side converges to $\int_{A}^{\infty} h(x) \, dF(x)$. Thus we obtain an upper bound which can be made arbitrarily small and is independent of $n$. The remainder of the proof is obvious.

PROOF OF THEOREM 1. Equation (4), which is to be proved, can be written in the form

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g(x) \, dF_n(x) = \int_{-\infty}^{\infty} g(x) \, dF(x), \tag{25}$$

and this is equivalent to

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{g(x) - g(0)}{x} \, dH_n(x) = \int_{-\infty}^{\infty} \frac{g(x) - g(0)}{x} \, dH(x). \tag{26}$$

First, suppose that the function $(g(x) - g(0))/x$ is continuous everywhere. Then (26), and hence (25), follows from Lemmas 9 and 10 by using Lemma 1. In particular, (25) is now proved for $g(x) = \cos tx$ and $\sin tx$. By the continuity theorem for characteristic functions this implies that

$$\lim_{n \to \infty} F_n(x) = F(x) \tag{27}$$

for all points of continuity of $F(x)$. Equation (25) now follows for every $g(x)$ which satisfies the conditions of Theorem 1 by applying Lemma 1, (27) and Lemma 10.

Proof of Theorem 2. Since $f(x)$ and $g(x)$ are convex, we have $f(EZ_{nj}) \leqq Ef(Z_{nj})$ and $g(Ef(Z_{nj})) \leqq Eg(f(Z_{nj}))$. Since $g(x)$ is nondecreasing, $g(f(EZ_{nj})) \leqq g(Ef(Z_{nj}))$. Hence

$$\frac{1}{n}\sum_{j=1}^{n} g(f(EZ_{nj})) \leqq \frac{1}{n}\sum_{j=1}^{n} g(Ef(Z_{nj})) \leqq \frac{1}{n}\sum_{j=1}^{n} Eg(f(Z_{nj})) = \int_{-\infty}^{\infty} g(f(x))\, dF(x). \tag{28}$$

The first member of (28) converges to the last member if the function $\bar{g}(x) = g(f(x))$ satisfies the conditions for $g(x)$ in Theorem 1. That these conditions are satisfied, follows from the fact that $\bar{g}(x)$ is convex.

## REFERENCES

[1] M. Dwass, "On the asymptotic normality of certain rank order statistics," *Ann. Math. Stat.*, to be published.

[2] R. A. Fisher and F. Yates, *Statistical Tables*. Hafner Publishing Co., New York, 1949.

[3] M. Fréchet and J. Shohat, "A proof of the generalized second limit theorem in the theory of probability," *Trans. Amer. Math. Soc.*, Vol. 33 (1931), pp. 533–543.

[4] W. Hoeffding, " 'Optimum' nonparametric tests." *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 83–92.

[5] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Math.*, Vol. 30 (1906), pp. 175–193.

[6] M. E. Terry, "Some rank order tests which are most powerful against specific parametric alternatives," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 346–366.