ON THE DISTRIBUTION OF RANKS AND OF CERTAIN RANK ORDER STATISTICS[1]

By Meyer Dwass

Northwestern University and Stanford University

1. Introduction. Suppose $X_1, \cdots, X_m$ and $X_{m+1}, \cdots, X_N$ are two independent samples from two possibly different populations, and $R_1, \cdots, R_m$ are the ranks of the first m observations in the combined sample and $R_{m+1}, \cdots, R_N$ the ranks of the remaining observations. In the first part of the paper, various moment generating functions connected with these ranks are derived. Of particular interest may be the moment generating function of the Wilcoxon statistic. The asymptotic distribution of a finite number of ranks is derived as $N \to \infty$. The remainder of the paper studies certain aspects of the distribution theory of rank order statistics of the form $\sum_{i=1}^{m} f_N(R_i/N)$. The Wilcoxon statistic and the Hoeffding c_1-statistic are special cases of such a statistic. Many previous studies have been devoted to showing its asymptotic normality. The main purpose of the last half of this paper is to show that for certain combinations of sample sizes m, n, and parent populations, the limiting distribution is non-normal as $m \to \infty$, $n \to \infty$, and $m/N \to 0$.

2. Generating functions for ranks. Throughout this paper we suppose that $X_1, \cdots, X_m, X_{m+1}, \cdots, X_{m+n}$ are $N = m + n$ independent random variables, the first m identically distributed, each with c.d.f. F_1 and the last n identically distributed, each with c.d.f. F_2. We suppose these c.d.f.'s are continuous. By the random variable R_i, the rank of X_i, we mean the number of X_j's less than or equal to X_i. The main object of this section is to write an expression for a generating function for ranks, and the following notation is intended to be useful toward that end. Let $u_0 = -\infty$, $u_{r+1} = \infty$, and

$$u_0 < u_1 < \cdots < u_r < u_{r+1}.$$

Then we denote

$$G_{i,j+1} = G_{i,j+1}(u_j, u_{j+1}) = F_i(u_{j+1}) - F_i(u_j) \qquad (i = 1, 2; j = 0, \cdots, r).$$

Let $i_1, i_2, \cdots, i_r$ be a permutation of the $r = p + q (p \leqq m, q \leqq n)$ integers $1, 2, \cdots, p, m + 1, \cdots, m + q$, and let $e_{i_1}, \cdots, e_{i_r}$ be defined by

$$e_{i_1} = \begin{cases} 1 & \text{if } i_1 \text{ is one of } 1, \cdots, p, \\ 2 & \text{if } i_1 \text{ is one of } m + 1, \cdots, m + q, \end{cases}$$

with similar definitions for $e_{i_2}, \cdots, e_{i_r}$. If $d_1 < d_2 < \cdots < d_r$ is a set of positive integers, they uniquely determine a set of non-negative integers w_1,

Received June 18, 1926; revised October 22, 1956.

[1] Work performed under Office of Naval Research Contract Nonr-225(21).

$w_2, \cdots, w_r$ such that

$$
\begin{aligned}
d_1 &= w_1 + 1, \\
d_2 &= w_1 + w_2 + 2, \\
&\vdots \\
d_r &= w_1 + w_2 + \cdots + w_r + r.
\end{aligned} \tag{1}
$$

Conversely, (1) determines for non-negative integers $w_1, \cdots, w_r$ a corresponding set $d_1 < \cdots < d_r$.

We first want to evaluate

$$P\{R_{i_1} = d_1, \cdots, R_{i_r} = d_r ; \quad d_1 < \cdots < d_r \leqq N\}$$

for positive integers d_i. By an elementary computation we can write this probability as

$$
\sum \frac{(m-p)!}{s_1! \cdots s_{r+1}!} \frac{(n-q)!}{t_1! \cdots t_{r+1}!} \int_{u_1 < \cdots < u_r} G_{11}^{s_1} \cdots G_{1,r+1}^{s_{r+1}} \cdot G_{21}^{t_1} \cdots G_{2,r+1}^{t_{r+1}} \, dF_{e_{i1}}(u_1) \cdots dF_{e_{ir}}(u_r). \tag{2}
$$

where, to facilitate printing, e_{i_1} is written as e_{i1} when it occurs as a subscript etc., and where $\sum$ is summation over all non-negative integers s_i, t_i such that

$$
\begin{aligned}
s_1 + \cdots + s_{r+1} &= m - p, \\
t_1 + \cdots + t_{r+1} &= n - q, \\
s_1 + t_1 &= w_1, \\
s_2 + t_2 &= w_2, \\
&\vdots \\
s_r + t_r &= w_r,
\end{aligned}
$$

the w_i being determined by the d_i as in (1). Next we recognize that (2) equals the coefficients of $T_{i_1}^{w_1} \cdots T_{i_r}^{w_r}$ in

$$
\int_{u_1 < \cdots < u_r} (T_{i_1} G_{11} + \cdots T_{i_r} G_{1,r} + G_{1,r+1})^{m-p} \cdot (T_{i_1} G_{21} + \cdots + T_{1_r} G_{2,r} + G_{2,r+1})^{n-q} \, dF_{e_{i1}}(u_1) \cdots dF_{e_{ir}}(u_r). \tag{3}
$$

Since we can write

$$
T_{i_1}^{w_1} T_{i_2}^{w_2} \cdots T_{i_r}^{w_r} = \left(\frac{T_{i_1}}{T_{i_2}}\right)^{d_1} \left(\frac{T_{i_2}}{T_{i_3}}\right)^{d_2} \cdots \left(\frac{T_{i_{r-1}}}{T_{i_r}}\right)^{d_{r-1}} (T_{i_r})^{d_r} \cdot \frac{1}{T_{i_1} \cdots T_{i_r}}
$$

this suggests that we make the relabelling

$$\frac{T_{i_1}}{T_{i_2}} = \theta_{i_1}, \cdots, \frac{T_{i_{r-1}}}{T_{i_r}} = \theta_{i_{r-1}}, \qquad T_{i_r} = \theta_{i_r} \tag{4}$$

or

$$\begin{aligned} T_{i_1} &= \theta_{i_1} \cdots \theta_{i_r}, \\ T_{i_2} &= \theta_{i_2} \cdots \theta_{i_r}, \\ &\vdots \\ T_{i_r} &= \theta_{i_r}. \end{aligned} \tag{5}$$

Substituting from (5) into (3) and denoting the resulting function of $\theta_1, \cdots, \theta_r$ by $\varphi(i_1, \cdots, i_r)$, we have

$$\sum P(R_{i_1} = d_1, \cdots, R_{i_r} = d_r)\theta_{i_1}^{d_1} \cdots \theta_{i_r}^{d_r} = \theta_{i_1}\theta_{i_2}^2 \cdots \theta_{i_r}^r \varphi(i_1, \cdots, i_r),$$

where $\sum$ is summation over all integers d_i such that $1 \leqq d_1 < d_2 < \cdots < d_r \leqq N$. We can now state the following.

THEOREM 1. *The generating function of* $R_1, \cdots, R_p, R_{m+1}, \cdots, R_{m+q}$ equals

$$\sum P(R_1 = d_1, \cdots, R_{m+q} = d_r)\theta_1^{d_1} \cdots \theta_r^{d_r} = \sum{}' \theta_{i_1}\theta_{i_2}^2 \cdots \theta_{i_r}^r \varphi(i_1, \cdots, i_r), \tag{6}$$

where $\sum$ *is over all possible integers* $d_1, \cdots, d_r$ *between* 1 *and* N *(no two equals to each other) and* $\sum'$ *is over all permutations* $i_1, \cdots, i_r$ *of the integers* $1, \cdots, p, m+1, \cdots, m+q$.

REMARK. Equality among any two d_i's is equivalent to tied ranks which is excluded with probability one by the assumption of continuity of F_1, F_2.

3. Several special cases. In this section we look at three special cases. They will be referred to again later.

A. *The generating function for a single rank.* To find the generating function for a single rank, say R_1 to be specific, set

$$p = 1, \qquad q = 0, \qquad r = 1, \qquad \theta_1 = \theta$$

in (6). We then obtain that

$$\begin{aligned} E\theta^{R_1} &= \theta \int_{-\infty}^{\infty} (\theta G_{11} + G_{12})^{m-1} (\theta G_{21} + G_{22})^n \, dF_1(u) \\ &= \theta \int_{-\infty}^{\infty} ((\theta - 1) F_1(u) + 1)^{m-1} ((\theta - 1) F_2(u) + 1)^n \, dF_1(u). \end{aligned} \tag{7}$$

B. *The generating function for* $R_1, R_2, \cdots, R_m$. For this case we set

$$p = r = m, \qquad q = 0,$$

in (6) and obtain

$$\begin{aligned} &E\theta_1^{R_1}\theta_2^{R_2} \cdots \theta_m^{R_m} \\ &\quad = \sum \theta_{i_1}\theta_{i_2}^2 \cdots \theta_{i_m}^m \int_{u_1 < \cdots < u_m} [\theta_{i_2} \cdots \theta_{i_m}(\theta_{i_1} - 1) F_2(u_1) \\ &\qquad + \cdots + (\theta_{i_m} - 1) F_2(u_m) + 1]^n \, dF_1(u_1) \cdots dF_1(u_m), \end{aligned}$$

where $\sum$ is over all permutations $i_1, i_2, \cdots, i_m$ of $1, 2, \cdots, m$.

C. *The Wilcoxon statistic.* This statistic is $R_1 + \cdots + R_m$. In case B above, set $\theta_1 = \theta_2 = \cdots = \theta_m = \theta$ and we find that

$$E\theta^{R_1+\cdots+R_m} = m!\theta^{m(m+1)/2} \int_{u_1<\cdots<u_m} [\theta^{m-1}(\theta - 1) F_2(u_1) + \cdots + (\theta - 1) F_2(u_m) + 1]^n \, dF_1(u_1) \cdots dF_1(u_m).$$

4. Limiting distributions involving a fixed number of ranks.

A. *A single rank.* From (7) we have that

$$Ee^{i\theta R/N} = e^{i\theta/N} \int_{-\infty}^{\infty} ((e^{i\theta/N} - 1) F_1(u) + 1)^{m-1} ((e^{i\theta/N} - 1) F_2(u) + 1)^n \, dF_1(u).$$

Suppose $m \to \infty$, $n \to \infty$, $m/N \to \rho$. Since

$$|(e^{i\theta/N} - 1)F_j(u) + 1|^2 = |F_j(u)^2 + (1 - F_j(u))^2 + 2(1 - F_j(u))F_j(u) \cos(\theta/N)| \leqq 1,$$

and since, as $N \to \infty$,

$$((e^{i\theta/N} - 1)F_j(u) + 1)^N \to \exp[i\theta F_j(u)] \qquad j = 1, 2,$$

we can apply the Lebesgue bounded convergence theorem to conclude that

$$E \exp(i\theta R_1/N) \to \int_{-\infty}^{\infty} \exp[i\theta(\rho F_1(u) + (1 - \rho) F_2(u))] \, dF_1(u), \tag{8}$$

as $N \to \infty$. Hence R_1/N is asymptotically distributed as a random variable

$$\rho F_1(X) + (1 - \rho)F_2(X),$$

where X has c.d.f. $F_1(X)$.

REMARKS. (a) Notice that the extremes, $\rho = 0$ and $\rho = 1$ are included in this result.

(b) If we do the above computation for R_j/N, its limiting characteristic function is given by the right side of (8) for $j = 1, \cdots, m$, and by the right side of (8) with dF_1 replaced by dF_2 for $j = m + 1, \cdots, m + n$.

(c) A similar analysis shows that $R_{j_1}/N, \cdots, R_{j_k}/N$ are asymptotically independent as $N \to \infty$ if $j_1 < j_2 < \cdots < j_k$ are fixed indices which do not depend on N.

B. $R_1, \cdots, R_m$. We hold m fixed and let $n \to \infty$. Then by the above remarks,

$$E \exp[i(\theta_1 R_1/N + \cdots + \theta_m R_m/N)] \to \prod_{j=1}^{m} \int_{-\infty}^{\infty} \exp[i\theta_j F_2(u)]$$

as $n \to \infty$. Thus, $R_1/N, \cdots, R_m/N$ are asymptotically independent and each is distributed as a random variable $F_2(X)$, where X has c.d.f. F_1. Let us denote the limiting c.d.f. of R_1/N by $S(t)$. That is, there is a c.d.f. S, such that at any

continuity point t of S,

$$P(R_1/N \leqq t) \to S(t), \tag{9}$$

as $N \to \infty$, for fixed m. Notice that in case F_2 has an inverse F_2^{-1}, then $S(t) = F_1(F_2^{-1}(t))$.

5. Limiting distributions of $S = \sum_{i=1}^{m} f_N(R_i/N)$.

In this section we study the asymptotic distribution of rank order statistics of the form

$$S_N = \sum_{i=1}^{m} f_N(R_i/N), \tag{10}$$

where $f_N(i/N)$ is a real number defined for $i = 1, \cdots, N$. We give below a short discussion on why S_N is of interest and on some of the known results regarding its asymptotic distribution.

For convenience, suppose that

$$f_N(1/N) \leqq f_N(2/N) \leqq \cdots \leqq f_N(N/N).$$

Let $H_N(t)$, $(0 < t < \infty)$ be the c.d.f. of the N numbers $f_N(i/N)$. That is, $H_N(t) =$ proportion of $f_N(i/N)$ less than t. Perhaps the most notable example of a statistic of the form (10) is the Wilcoxon statistic, in which case $f_N(t) = t$, $(0 \leqq t \leqq 1)$, for all N. In [3] it was shown that in case F_1, F_2 depend on a single parameter θ and $F_1 = F_2$ when $\theta = 0$, then often a test of $H_0{:}\theta = 0$ against $H{:}\theta > 0$ based on (10) for suitably chosen f_N is a locally most powerful rank order test (local in the sense that θ is close to zero). Studies relevant to the asymptotic normality[2] of (10) can be found in [1], [2], [3], [4], [6], [8]. In particular, it may be worth while to mention some specific conditions which insure the asymptotic normality of S_n. In each case we assume that $m/N \to \rho$ as $N \to \infty$ and $0 < \rho < 1$.

(1) $F_1 = F_2$. $f_N(i/N) = EZ_{Ni}^k$ for some positive integer k, where $Z_{N1} \leqq \cdots \leqq Z_{NN}$ are the ordered values of N independent identically distributed random variables [1].

(2) $F_1 = F_2$. The assumption preceding Theorem 2 below holds and the c.d.f. H has its first two moments [3].

(3) $F_1 \neq F_2$ or $F_1 = F_2$. $f_N(t) = f(t) =$ a polynomial in t, which does not depend on N [3].

We shall now construct the examples referred to in the introduction. The main tools are Theorems 2 and 3 which follow. In addition to the basic assumptions made at the beginning of Section 2, we assume also through the remainder of the paper that there is a c.d.f. $H(t)$ such that at every continuity point of

[2] Whenever we refer to the asymptotic distribution we mean a limiting distribution of $(S_N - a_N)/b_N$, as $m \to \infty$, $n \to \infty$ for a proper choice of $\{a_N\}$, $\{b_N\}$. It may be that m depends on n.

$H(t)$,

$$H_N(t) \to H(t), \qquad \text{as } N \to \infty.$$

THEOREM 2. *Let $t_1, \cdots, t_m$ be such that they are continuity points of $H(t)$ and such that $H(t_1), \cdots, H(t_m)$ are continuity points of $S(u)$. Then*

$$P\{f_N(R_1/N) < t_1, \cdots, f_N(R_m/N) < t_m\} \to S(H(t_1)) \cdots S(H(t_m))$$

as $N \to \infty$, provided m is fixed.

PROOF. We can write

$$P\{f_N(R_1/N) < t_1, \cdots, f_N(R_m/N) < t_m\}$$
$$= P\{R_1/N \leqq H_N(t_1), \cdots, R_m/N \leqq H_N(t_m)\}.$$

The result follows from (9) and the remarks preceding it.

COROLLARY. *Let $\varphi(u)$ be the characteristic function of a random variable with* c.d.f. $R(t) = S(H(t))$. *Then*

$$E \exp i(u_1 f_N(R_1/N) + \cdots + u_m f_N(R_m/N)) \to \varphi(u_1) \cdots \varphi(u_m)$$

as $N \to \infty$, m fixed.

LEMMA 1. *Let $X_{11}, X_{12}, \cdots; X_{21}, X_{22}, \cdots$, be two infinite sequences of random variables, and let the random variable*

$$t_{m,n} = t_{m,n}(X_{11}, \cdots, X_{1m}; X_{21}, \cdots, X_{2n})$$

be a function of the $m + n$ random variables in parentheses. Let $\varphi_{m,n}(u) =$ $E \exp(iut_{m,n})$. Suppose

(a) *There is a characteristic function φ such that for every positive integer m and every real u,*

$$\varphi_{m,n}(u) \to [\varphi(u)]^m$$

as $n \to \infty$ and m is fixed.

(b) *There are norming constants a_m, b_m and a characteristic function Ψ, such that for every real u*

$$\exp(-a_m u/b_m)[\varphi(u/b_m)]^m \to \Psi(u) \tag{11}$$

as $m \to \infty$. Then there is a sequence of pairs of positive integers $(1, n(1)), (2, n(2)),$ $\cdots, (m, n(m)), \cdots$ such that

$$\exp(-ia_m u/b_m)[\varphi_{m,n}(u/b_m)] \to \Psi(u) \tag{12}$$

as $m \to \infty$, $n \to \infty$, provided $n \geqq n(m)$.

The proof is elementary and we omit it. We point out that (12) says that the distribution of $(t_{m,n} - a_m)/b_m$ converges to that distribution whose characteristic function is Ψ.

LEMMA 2. Let $R(t) = S(H(t))$ be as defined above. Suppose $0 < F_2(t) < 1$ if and only if $0 < F_1(t) < 1$. Suppose also that $t = F^{-1}_2(u)$, the inverse of $F_2(t)$,

is defined for all $0 < u < 1$. Then

(a) if $F_1 = F_2$, $R(t) = H(t)$,

(b) if $F_2 = H$, $R(t) = F_1(t)$.

PROOF. The proof follows from the fact that $S(u) = F_1(F_2^{-1}(u))$.

THEOREM 3. *Suppose that if* $Y_1, Y_2, \cdots$ *is an infinite sequence of independent identically distributed random variables, each with* c.d.f. $R(t) = S(H(t))$, *then then there are norming constants* $\{a_m\}, \{b_m\}$, *such that the* c.d.f. *of* $(\sum_1^m Y_i - a_m)/b_m$ *converges to a* c.d.f. $L(t)$, *as* $m \to \infty$. *Then there is a sequence* $(1, n(1)), \cdots, (m, n(m)), \cdots$, *such that the* c.d.f. of

$$\frac{\sum_{i=1}^{m} f_N(R_i/N) - a_m}{b_m} \tag{13}$$

converges to $L(t)$ *as* $m \to \infty$, $n \to \infty$, *provided that* $n \geqq n(m)$.

PROOF. Let $t_{m,n} = \sum_{i=1}^m f_N(R_i/N)$. This statistic satisfies Condition (a) of Lemma 1 by the corollary to Theorem 2. (Let $u_1 = \cdots = u_m = u$ in that corollary.) Condition (b) holds by assumption and this completes the proof.

REMARKS. (a) An unsatisfactory feature of this result is that it tells nothing about the relative orders of m and n. It is clear that we can find sequences $\{m_i\}$, $\{n_i\}$, such that if $m = m_i$, $n = n_i$, then the asymptotic distribution of the proposition holds and

$$\lim_i m_i/(m_i + n_i) = 0. \tag{14}$$

Though our methods here are not sensitive enough to yield this information, the sense of the derivation is such to make reasonable the conjecture that this asymptotic distribution holds for *all* sample size sequences for which (14) holds.

(b) By Lemma 2, if $F_2 = H$ then $R(t) = F_1(t)$. By the proper choice of F_1 we can determine the limiting distribution L to be *any stable* distribution. For example, suppose (10) is the Hoeffding c_1-statistic [7]. That is, $f_N(i/N) = EZ_{Ni}$, where $Z_{N1} \leqq \cdots \leqq Z_{NN}$ are the ordered values of N independent $N(0, 1)$ random variables. According to [5], H is the unit normal c.d.f. Now suppose that the alternative to the usual null hypothesis that $F_1 = F_2$ is that F_2 is the unit normal c.d.f. and that F_1 is the Cauchy c.d.f., centered at θ. Then there are sequences $\{m, n(m)\}$ such that $[\sum_{i=1}^m f_N(R_i/N) - m\theta]/m$ has asymptotically the Cauchy distribution centered at zero. This is so because Lemma 2 (Case (a)) insures that R is the Cauchy c.d.f. and because an average of independent and identically distributed Cauchy variables is distributed like any one of its components.

(c) In case $H(t)$ concentrates all its mass on a bounded interval, then so does $R(t)$ and excluding the one point limiting distribution, the limiting distribution of this theorem must be normal. This will happen if $f_N(t)(0 \leqq t \leqq 1)$ is a polynomial in t which does not depend on N. This is not surprising since for this case [3] shows that if $\lim_{N\to\infty} m/N = \rho$ exists, then S_N is asymptotically normal for all $0 < \rho < 1$. As a matter of fact, these results would seem to imply that one should be able to include the extreme cases $\rho = 0$ and $\rho = 1$. Similarly, if $F_1 = F_2$

and H has its first two moments, then S_N is asymptotically normal. This is also not surprising since for this case [2] shows asymptotic normality for $0 < \rho < 1$.

(d) We can construct further examples of non-normal limiting distributions by supposing $F_1 = F_2$ and by choosing H properly, since by Lemma 2(b), $R = H$. This is presumably of lesser interest than the construction given in Remark (b) above.

REFERENCES

[1] M. Dwass, "On the asymptotic normality of certain rank order statistics," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 303-306.

[2] M. Dwass, "On the asymptotic normality of some statistics used in non-parametric tests," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 334-339.

[3] M. Dwass, "The large sample power of rank order tests in the two-sample problem," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 352-373.

[4] W. Hoeffding, "A combinatorial central limit theorem," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 558-66.

[5] W. Hoeffding, "On the distribution of the expected values of the order statistics," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 93-100.

[6] G. E. Noether, "On a theorem by Wald and Wolfowitz," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 455-58.

[7] M. E. Terry, "Some rank order tests which are most powerful against specific parametric alternatives," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 346-66.

[8] Wald, A., and J. Wolfowitz, "Statistical tests based on permutations of the observations," *Ann. Math. Stat.* Vol. 15 (1944), pp. 358-72.