ON THE ASYMPTOTIC NORMALITY OF SOME STATISTICS USED IN NON-PARAMETRIC TESTS

By Meyer Dwass

Northwestern University

1. Summary. Let $(R_1, \cdots, R_n)$ be a random vector which takes on each of the $N!$ permutations of $(1, 2, \cdots, N)$ with equal probability, $1/N!$. Let $(a_{N1}, \cdots, a_{NN})$ and $(b_{N1}, \cdots, b_{NN})$ be two sets of real numbers given for every N. We will assume throughout that for no N are the a_{Ni} all equal or the b_{Ni} all equal. We also assume that the a_{Ni} and b_{Ni} have been so normalized that

$$\sum a_{Ni} = \sum b_{Ni} = 0; \qquad \sum a_{Ni}^2 = N^{-1} \sum b_{Ni}^2 = 1.$$

Unless otherwise stated, $\sum$ will mean $\sum_{i=1}^{N}$. Define

$$S_N = \sum a_{Ni} b_{NR_i}. \tag{1.1}$$

Let $\Phi(x)$ be the unit normal c.d.f. In Section 2 sufficient conditions are given for $\Pr\{S_N < x\}$ to approach $\Phi(x)$ as $N \to \infty$. (The first two moments of S_N are 0 and $N/(N-1)$, respectively.)

For every N, let $Y = (Y_{11}, \cdots, Y_{1N_1}, \cdots, Y_{m1}, \cdots, Y_{mN_m})$ be $N = N_1 + \cdots + N_m$ random variables which are mutually independent and independent of the R_i. We assume that all Y_{ij} with the same first subscript are identically distributed. Define

$$\bar{Y} = N^{-1} \sum_{i,j} Y_{ij}, \qquad S^2 = N^{-1} \sum_{i,j} (Y_{ij} - \bar{Y})^2, \qquad SY'_{ij} = Y_{ij} - \bar{Y},$$

$$Y'_{ij} = 0 \text{ if } S = 0.$$

Let Y' denote the vector of the Y'_{ij}. Let $y = (y_1, \cdots, y_N)$ denote a point in N-space. By $F_N(x, y)$ we mean the c.d.f. of the random variable $S_N = \sum a_{Ni} y_{NR_i}$. In Section 3 are considered sufficient conditions for convergence with probability one of the random c.d.f. $F_N(x, Y')$ to $\Phi(x)$.

2. Asymptotic distribution of S_N when the b_{Ni} are nonrandom. Let $G_N(x)$ denote the c.d.f. of the b_{Ni} (continuous to the left). We assume that the b_{Ni} have been so indexed that $b_{N1} \leqq b_{N2} \leqq \cdots \leqq b_{NN}$.

Theorem 2.1. *Suppose*

A) *there is a* c.d.f. $G(x)$ *such that* $\lim_{N\to\infty} G_N(x) = G(x)$ *at every point of continuity of* $G(x)$ *and*

$$\int_{-\infty}^{\infty} x \, dG(x) = 0, \qquad \int_{-\infty}^{\infty} x^2 \, dG(x) = 1;$$

and either

B) $\lim_{N\to\infty} \max_{1\leqq i\leqq N} |a_{Ni}| = 0$ or B′) $G(x) = \Phi(x)$.

Received July 13, 1953, revised July 21, 1954.

Then S_N (1.1) *is asymptotically normally distributed. That is,* $\Pr\{S_N < x\} \to \Phi(x)$ *as* $N \to \infty$.

This theorem can be compared with that of a previous paper [1]. There it was assumed that the b_{Ni} were essentially powers of expected values of order statistics. In this paper we use the fact that the b_{Ni} behave asymptotically like expected values of order statistics. The idea of the earlier theorem [1] is then used in proving this theorem. In [1] the counterpart of $G(x)$ was assumed continuous; in this paper we make no such assumption. Theorem 3.1, below, deals with the case where the b_{Ni} are random variables. From now on, the function $G(x)$ will be assumed to have the properties stated in Theorem 2.1.

The proof of this theorem is aided by Lemmas 2.1 to 2.7 below. The main lines of the proof are as follows. Let $X_1, \cdots, X_N$ be independent, identically distributed random variables, each with the c.d.f. $G(x)$; let $Z_{N1} \leqq \cdots \leqq Z_{NN}$ be the ordered values of the X's. We will define below a vector function of the X's, $(R_1, \cdots, R_N)$, which has the property that it assumes each of the permutations of $(1, \cdots, N)$ with equal probability, $1/N!$. Then $\sum a_{Ni} b_{NR_i} = S_N$ is the random variable whose asymptotic normality we seek to prove.

Either condition B) or B′) will assure us of the asymptotic normality of $\sum a_{Ni}X_i$. (See [1]; however, existence of third moment is not required. This follows from the counterpart of Lindeberg's condition for a sequence of sequences of random variables. I am obliged to Prof. W. Hoeffding for pointing this out to me.)

Hence it is sufficient to show that $\sum a_{Ni}X_i - S_N$ converges in probability to zero. Sufficient for this is to show that $\lim_{N\to\infty} E(\sum a_{Ni}X_i - S_N)^2 = 0$. We have that

$$(2.1) \quad E(\sum a_{Ni}X_i - S_N)^2 = EX_1^2 - 2E(\sum a_{Ni}X_i)\,(S_N) + N/(N-1).$$

The purpose of Lemma 2.1 is to show that our particular definition of $(R_1, \cdots, R_N)$ provides us with the fact that $E(\sum a_{Ni}X_i)\,(S_N) = (N-1)^{-1} \cdot \sum EZ_{Ni}b_{Ni}$. Since $EX_1^2 = 1$, then if $\lim_{N\to\infty} (N-1)^{-1}\sum EZ_{Ni}b_{Ni} = 1$, this will imply that (2.1) approaches zero as $N \to \infty$. Making use of the fact [3] that $\lim_{N\to\infty} N^{-1}\sum(EZ_{Ni})^2 = 1$, it is easy to see that

$$\lim_{N\to\infty} N^{-1}\sum EZ_{Ni}b_{Ni} = 1 \quad \text{is equivalent to} \quad \lim_{N\to\infty} N^{-1}\sum(b_{Ni} - EZ_{Ni})^2 = 0.$$

The truth of this last limit is shown in the sequence of Lemmas 2.2 through 2.7, where important use is made of some results of Hoeffding [3].

A random vector $R = (R_1, \cdots, R_N)$ is defined by considering the ordered arrangement of the X's,

$$(2.2) \qquad X_{i_1} \leqq X_{i_2} \leqq \cdots \leqq X_{i_N}.$$

If no two of the X's are equal to each other, then let $R_{i_1} = 1$, $R_{i_2} = 2, \cdots, R_{i_N} = N$. If there are ties among the X's then let

$$(X_{i_1} \leqq X_{i_2} \leqq \cdots \leqq X_{i_N}), \quad \cdots, \quad (X_{j_1} \leqq X_{j_2} \leqq \cdots \leqq X_{j_N}),$$

be the set of all possible distinct ordered arrangements of the X's. Say there are p such arrangements. In such a case, let

$$P\{(R_{i_1} = 1), \quad (R_{i_2} = 2), \quad \cdots, \quad (R_{i_N} = N) \mid (X_1, \cdots, X_N)\} = 1/p,$$
$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots$$
$$P\{(R_{j_1} = 1), \quad (R_{j_2} = 2), \quad \cdots, \quad (R_{j_N} = N) \mid (X_1, \cdots, X_N)\} = 1/p.$$

It is easy to see that R equals each of the $N!$ permutations of $(1, \cdots, N)$ with probability $1/N!$

LEMMA 2.1. $E(\sum a_{Ni}X_i)\,(\sum a_{Ni}b_{NR_i}) = (N-1)^{-1}\sum EZ_{Ni}b_{Ni}$.

PROOF. The left side can be written as

$$(N!)^{-1}\sum{}' \{\sum a_{Ni}b_{Nr_i}E(\sum a_{Ni}X_i \mid (R_1 = r_1), \cdots, (R_N = r_N)\}$$

where $(r_1, \cdots, r_N)$ is one of the permutations of $(1, \cdots, N)$ and $\sum'$ is summation over all $N!$ permutations. We also have that

$$E[X_1 \mid (R_1 = r_1), \cdots, (R_N = r_N)] = E[Z_{Nr_i} \mid (R_1 = r_1), \cdots, (R_N = r_N)].$$

Since the distribution of $X_1, \cdots, X_N$ is invariant under all permutations, this last conditional expectation does not depend on the values of the R's and hence equals EZ_{Nr_i}. This shows that the left side equals

$$(N!)^{-1}\sum{}' (\sum a_{Ni}b_{Nr_i})\,(\sum a_{Ni}EZ_{Nr_i}).$$

By an elementary calculation, this in turn is equal to the right side.

Since $Z_{N1} \leqq Z_{N2} \leqq \cdots \leqq Z_{NN}$, then $EZ_{N1} \leqq EZ_{N2} \leqq \cdots \leqq EZ_{NN}$. Denote by $H_N(x)$ the c.d.f. of the EZ_{Ni} (continuous to the left).

LEMMA 2.2 (Hoeffding). $\lim_{N\to\infty} H_N(x) = G(x)$ *for every* x *which is a point of continuity of* $G(x)$.

This is proved by Hoeffding ([3], Theorems 1 and 2).

Let S be that set of points on the real line where $G(x)$ is discontinuous, together with all points x where $G(x-h) < G(x) < G(x+h)$ for every $h > 0$.

LEMMA 2.3. *Given* $\epsilon_1 > 0$ and $\epsilon_2 > 0$, *there is an even integer* $2n$ *and there is a sequence of closed intervals which are mutually disjoint,*

$$[t_1, t_2], [t_3, t_4], \cdots, [t_{2n-1}, t_{2n}],$$

having the following properties.

a) *With each* t_i *we can associate an* a_i *with* $0 < a_i < 1$ *and a* $y_i \in S$ *such that*

$$t_i = a_i\,G(y_i - 0) + (1 - a_i)\,G(y_i + 0);$$

b) $y_1^2(t_2 - t_1) + y_3^2(t_4 - t_3) + \cdots + y_{2n-1}^2(t_{2n} - t_{2n-1}) > 1 - \epsilon_1$;

c) $\max[(y_2 - y_1), (y_4 - y_3), \cdots, (y_{2n} - y_{2n-1})] < \epsilon_2$.

PROOF. We will denote the right side of the equality in part a) by $G_{a_i}(y_i)$. Since $\int_A^B y^2\,dG(y)$ exists in Riemann-Stieltjes sense for any A, B, we can find finite A', B' and a partitioning,

$$A' = y_1' < y_2' < \cdots < y_m' = B',$$

such that

$$y_1'^2[G(y_2') - G(y_1')] + y_2'^2[G(y_3') - G(y_2')] + \cdots + y_{m-1}'^2[G(y_m') - G(y_{m-1}')] > 1 - \tfrac{1}{2}\epsilon_1,$$

$$\max_{1 \leqq i \leqq m-1} (y_{i+1}' - y_i') < \tfrac{1}{2}\epsilon_2 .$$

We can replace this partitioning by one

$$A = y_1 < y_2 < \cdots < y_{2n} = B,$$

and find constants $0 < a_i < 1$ where each $y_i \varepsilon S$ and for which

$$y_1^2[G_{a_2}(y_2) - G_{a_1}(y_1)] + y_2^3[G_{a_4}(y_4) - G_{a_3}(y_3)] + \cdots + y_{2n-1}^2[G_{a_{2n}}(y_{2n}) - G_{a_{2n-1}}(y_{2n-1})] > 1 - \epsilon_1 ,$$

and where c) is satisfied. The clumsy but elementary details are omitted. We need now only set $t_i = G_{a_i}(y_i)$ for $i = 1, 2, \cdots, 2n$; adjacent t_i's can be equal to each other.

LEMMA 2.4. *For every N, let there be given an $m = m(N)$. Let $y \varepsilon S$. If $m/N \to aG(y - 0) + (1 - a)\, G(y + 0)$, for some a where $0 < a < 1$, as $N \to \infty$, then $b_{Nm} \to y$ and $EZ_{Nm} \to y$.*

PROOF. Suppose $\overline{\lim}_{N\to\infty} b_{Nm} = y + \delta$, where $\delta > 0$. Then there are δ', where $0 < \delta' < \delta$, such that $y + \delta'$ is a continuity point of G, and a subsequence N_i, m_i, such that

$$\lim_{i\to\infty} m_i/N_i \geqq G(y + \delta') > aG(y - 0) + (1 - a)\, G(y + 0),$$

which gives a contradiction. We treat $\underline{\lim}\, b_{Nm}$ similarly. Making use of Lemma 2.2, we can prove this for the EZ_{Ni} in the same way. This was proved by Hoeffding ([3], Lemma 5) without using the result quoted in our Lemma 2.2.

Let $T = [t_1, t_2] \cup [t_3, t_4] \cup \cdots \cup [t_{2n-1}, t_{2n}]$, a finite union of disjoint closed intervals. Let $\sum_T a_i$ mean summation over all indices i such that $(i/N) \varepsilon T$.

LEMMA 2.5 *For every $\epsilon > 0$, there exists a finite union of disjoint closed intervals, $[t_1, t_2] \cup \cdots \cup [t_{2n-1}, t_{2n}] = T$, such that*

$$\underline{\lim}\, N^{-1} \sum_T b_{Ni}^2 > 1 - \epsilon, \qquad \underline{\lim}\, N^{-1} \sum_T (EZ_{Ni})^2 > 1 - \epsilon.$$

PROOF. Choose $\epsilon > 0$. We can find a sequence of closed intervals which are mutually disjoint, $[t_1, t_2]$, $[t_3, t_4]$, $\cdots$, $[t_{2n-1}, t_{2n}]$, satisfying the requirements of Lemma 2.3 for $\epsilon_1 = \frac{1}{2}\epsilon$; the value of ϵ_2 of Lemma 2.3 is not material here. Then

$$N^{-1} \sum_T b_{Ni}^2 = N^{-1} \sum_{Nt_1 \leqq i \leqq Nt_2} b_{Ni}^2 + N^{-1} \sum_{Nt_3 \leqq i \leqq Nt_4} b_{Ni}^2 + \cdots + N^{-1} \sum_{Nt_{2n-1} \leqq i \leqq Nt_{2n}} b_{Ni}^2 .$$

For simplicity, look at the first term on the right side of the above equality. Let p denote the number of integers contained between Nt_1 and Nt_2 inclusive. Let $[r]$ denote the least integer greater than or equal to r. Then

$$N^{-1} \sum_{Nt_1 \leqq i \leqq Nt_2} b_{Ni}^2 \geqq b_{N[Nt_1]} N^{-1} p.$$

By Lemma 2.4 the first factor of the right side has the limit y_1^2. The second factor has the limit $t_2 - t_1$. Hence the limit inferior of the left side is greater than $y_1^2(t_2 - t_1) - \frac{1}{2}\epsilon/n$. Therefore

$$\varliminf_{N\to\infty} N^{-1} \sum_T b_{Ni}^2 > \sum y_{2j-1}^2 (t_{2j} - t_{2j-1}) - \tfrac{1}{2}\epsilon > 1 - \epsilon.$$

The last inequality of Lemma 2.5 is proved in the same way.

LEMMA 2.6. *Given* $\epsilon > 0$, *we can find a set* $T = T(\epsilon)$, *such that Lemma* 2.5 *holds, and also* $\varlimsup_{N\to\infty} N^{-1} \sum_T (b_{Ni} - EZ_{Ni})^2 < \epsilon$.

We omit the proof, since it is very similar to that of Lemma 2.5. It uses Lemmas 2.2–2.4; in particular, Lemma 2.3 is essential.

LEMMA 2.7. $\lim_{N\to\infty} N^{-1} \sum (b_{Ni} - EZ_{Ni})^2 = 0$.

PROOF. For $\epsilon > 0$, find $T = T(\epsilon)$ satisfying Lemma 2.6. Write

$$N^{-1} \sum (b_{Ni} - EZ_{Ni})^2 = N^{-1} \sum_T (b_{Ni} - EZ_{Ni})^2 + N^{-1} \sum{}' (b_{Ni} - EZ_{Ni})^2,$$

where $\sum'$ means summation over those indices not summed in $\sum_T$. By Lemma 2.6, for all N sufficiently large, $N^{-1}\sum_T (b_{Ni} - EZ_{Ni})^2 < 2\epsilon$. It is true that

$$N^{-1} \sum{}' (b_{Ni} - EZ_{Ni})^2 \leqq (\sqrt{N^{-1} \sum{}' b_{Ni}^2} + \sqrt{N^{-1} \sum{}' (EZ_{Ni})^2})^2.$$

For all N sufficiently large, each of the terms under the square root sign is, by Lemma 2.5, less than 2ϵ. Thus for all N sufficiently large, $N^{-1} \sum (b_{Ni} - EZ_{Ni})^2$ is arbitrarily small, which proves the lemma.

3. Asymptotic distribution of S_N when the b_{Ni} are random variables. We will now deal with the generalization of Theorem 2.1 that was described in the last paragraph of Section 1. We refer to the definitions made there. Let $G_N(x, Y)$ be the proportion of Y_{ij} which are less than x, and let $G'_N(x, Y')$ be the proportion of Y'_{ij} which are less than x. Let $G_1(x), \cdots, G_m(x)$ be the c.d.f.'s of $Y_{11}, Y_{21}, \cdots, Y_{m1}$ respectively. Let

$$\int_{-\infty}^{\infty} x\, dG_i(x) = \mu_i, \qquad \int_{-\infty}^{\infty} (x - \mu)^2\, dG_i(x) = \sigma_i^2, \qquad G(x) = \sum_{i=1}^{m} k_i G_i(x),$$

where $k_1, \cdots, k_m$ are constants defined in Theorem 3.1 below, with $k_i \geqq 0$, and $\sum_1^m k_i = 1$.

THEOREM 3.1. *Let* $G'(x)$ *be the* c.d.f. *obtained from* $G(x)$ *by a linear transformation of the independent variable, so that the first two moments of* $G'(x)$ *are* 0 *and* 1. *Suppose*

A) $\lim_{N\to\infty} N_i/N = k_i$ *exists for* $i = 1, \cdots, m$;

B) $0 < \sigma_i^2 < \infty$;

C) *condition* B) *or* B′) *of Theorem* 2.1 *is satisfied for* $G'(x)$.

Then $F_N(x, Y')$, *as defined in Section* 1, *converges with probability one to* $\Phi(x)$.

PROOF. According to the Cantelli-Glivenko theorem, and some well-known facts about strong convergence, we have that $G_N(t, Y)$ converges with probability one to $G(t)$, uniformly in t ([2], pp. 257–279, 280). Also we have that, with probability one, $\bar{Y}$ and S^2 converge respectively to the mean and variance of a random variable distributed with c.d.f. $G(x)$. Then, with probability one, $G'_N(t, Y') \to G'(t)$ at the continuity points of $G'(t)$, as $N \to \infty$. Hence the required result follows from Theorem 2.1.

4. Applications. Some applications of a theorem like Theorem 2.1 have been pointed out previously [1]. We wish here to point out that Theorem 3.1 can be useful in evaluating the "large sample power" of tests of the sort that were studied by Hoeffding [4], specifically his Theorem 6.2. Analogous results can be obtained where the X's of that theorem are essentially like the Y's of our Theorem 3.1. Also, the results of this paper can be used to study analysis of variance tests like those of Hoeffding's Section 5. We plan a future paper on the large sample power of analysis of variance tests of this type.

REFERENCES

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

[2] M. FRECHÉT, *Généralités sur les Probabilités*, premier livre, Gauthier-Villars, Paris, 1950.

[3] WASSILY HOEFFDING, "On the distribution of the expected values of the order statistics," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 93–100.

[4] WASSILY HOEFFDING, "The large-sample power of tests based on permutations of observations," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 169–192.