A NONPARAMETRIC TEST FOR THE SEVERAL SAMPLE PROBLEM[1]

By William H. Kruskal

University of Chicago

1. Summary. Suppose that C independent random samples of sizes $n_1, \cdots, n_C$ are to be drawn from C univariate populations with unknown cumulative distribution functions $F_1, \cdots, F_C$. This paper discusses a test of the null hypothesis $F_1 = F_2 = \cdots = F_C$ against alternatives of the form

$$F_i(x) = F(x-\theta_i) \qquad \text{(all } x,\ i = 1, \cdots, C\text{)}$$

with the θ_i's not all equal, or against alternatives of a much more general sort to be specified in Section 5. The test to be discussed has as its critical region large values of the ordinary F-ratio for one-way analysis of variance, computed after the observations have been replaced by their ranks in the $\sum n_i$-fold over-all sample. This use of ranks simplifies the distribution theory, and permits application of the test to cases where the ranks are available but the numerical values of the observations are difficult to obtain. Briefly, then, we shall consider a nonparametric analogue, based on ranks, of one-way analysis of variance.

It is shown in Section 4 that, under quite general conditions, the proposed test statistic, H, is asymptotically chi-square with $C - 1$ degrees of freedom when the null hypothesis holds. Section 5 derives a necessary and sufficient condition that the natural family of sequences of tests based on large values of H all be consistent against a given alternative. Section 6 derives the variance of H under the null hypothesis, Section 7 derives the maximum value of H, and Section 8 gives a difference equation which may be used to obtain exact small-sample distributions under the null hypothesis. These derivations are made on the assumption of continuity for the cumulative distribution functions; Section 9 considers extensions to the possibly discontinuous case.

2. Introduction. Until Section 9 all cumulative distribution functions will be supposed continuous. The over-all sample consists of the $\sum n_i = N$ (say) independent random variables $\xi_j^{(i)}$ (i $= 1, \cdots, C; j = 1, \cdots, n_i$), where the superscript refers to the (sub)sample and the subscript indexes observations within a (sub)sample. Under the null hypothesis all the ξ's have the same continuous but unknown cdf (cumulative distribution function): $F(x)$. Each $\xi_j^{(i)}$ is immediately replaced by $X_j^{(i)}$, its rank in the over-all sample. Then, under the null hypothesis, the N-tuple $(X_1^{(1)}, \cdots, X_{n_1}^{(1)}, X_1^{(2)}, \cdots, X_{n_2}^{(2)}, \cdots\cdots, X_1^{(C)}, \cdots, X_{n_C}^{(C)})$ takes as values with equal probability the $N!$ permutations of $(1, 2, \cdots, N)$.

Next let $R_i = \sum_{j=1}^{n_i} X_j^{(i)}$ be the sum of ranks of sample from the ith population and let $\bar{R}_i = R_i/n_i$. Of course $\sum R_i = \frac{1}{2}N(N+1)$. The standard one-way

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

analysis of variance test based on the X's has for its critical region large values of

$$\sum_{i=1}^{C} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2 \Big/ \left[\sum_{i=1}^{C}\sum_{j=1}^{n_i} (X_j^{(i)} - \bar{R}_i)^2\right].$$

But this is a monotone increasing function of

$$\sum_{i=1}^{C} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2 \Big/ \left[\sum_{i=1}^{C}\sum_{j=1}^{n_i} \left(X_j^{(i)} - \frac{N+1}{2}\right)^2\right] \tag{2.1}$$

and because of the use of ranks the denominator of the above expression is a constant. Hence a critical region consisting of large values of the numerator of (2.1) is suggested. The corresponding test is the one to be discussed in this paper. Actually this test will be discussed in terms of the random variable

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} n_i \left(\bar{R}_i - \frac{N+1}{2}\right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1).$$

Since the variance of the uniform distribution on the integers $1, 2, \cdots, N$ is $(N^2-1)/12$, it is natural to expect that the numerator of the F-ratio in terms of ranks divided by this variance is asymptotically chi-square with $C-1$ degrees of freedom. But this normalized numerator is just $H\ N/(N-1)$. The minor advantage of H over this and other asymptotically equivalent random variables upon which the test might be based is that under the null hypothesis $EH = C - 1$ $= E$(chi-square with $C - 1$ degrees of freedom).

3. Relationship to other tests. When $C = 2$ the test discussed in this paper is the same as the symmetrical two-tail version of a test considered by Wilcoxon [11] and by Mann and Whitney [2]; for when $C = 2$

$$H = \frac{12}{(N+1)n_1(N-n_1)} \left(R_1 - n_1 \frac{N+1}{2}\right)^2.$$

For $C = 2$, a test against *any* alternative (subject to existence of and weak conditions on a density function) is provided by the work of Wald and Wolfowitz [9] who propose and discuss a test based on runs in the over-all sample. A generalization of the Wald-Wolfowitz test for any C is available (e.g., Wallis [12]). For any C a test based on the median of the over-all sample and reducing to a conventional chi-square test has been proposed by Brown and Mood in Chapter 16 of [4]. A generalization of this test using several previously determined order statistics of the over-all sample is described by Massey [3]. Other tests are discussed in [1] and [13]. For $C = 2$ a recent addition to the list of tests has been made by Marshall [14].

Whitney [10] in the case of $C = 3$ considers two tests designed to have particular power against the following two types of alternatives, respectively:

(1) $F_1(x) > F_2(x)$ and $F_1(x) > F_3(x)$ (all x),

(2) $F_1(x) > F_2(x) > F_3(x)$ (all x).

His tests appear to be generalizable for any C.

The test discussed in this paper, that is, that based on large values of H, is closely akin to tests considered by Pitman, Friedman and others for two-way analysis of variance problems (see the expository paper by Scheffé [7] for a general discussion and references). However, this particular application of the randomization method has not to my knowledge been discussed in the statistical literature. Further discussion of related tests will be found in [15].

4. Asymptotic distributions. The term "asymptotic distribution" will be taken in the sense of convergence in distribution as $N \to \infty$. We shall assume at present that, for all i, $\lim_{N\to\infty} n_i/N = \nu_i$ exists and is positive. We proceed to show that under the null hypothesis the R_i's, properly normed, have asymptotically a singular multivariate normal distribution, and that from this the asymptotic chi-square distribution of H readily follows. The proof will be a direct application of a powerful general theorem of Wald and Wolfowitz [8], and I shall suppose that the reader is familiar with this reference. A consequence of the Wald-Wolfowitz theorem, in a form appropriate for our purpose, may be stated as follows:

THEOREM 4W (*Wald-Wolfowitz*). *Let* $\{A_N\}$ *be a sequence of ordered* N*-tuples,* $A_N = (a_{N1}, a_{N2}, \cdots, a_{NN})$, $(N = 1, 2, 3, \cdots)$ *satisfying condition* W *of* [8]. *Let* $(Z_{N1}, \cdots, Z_{NN})$ *for each* N *be a random ordered* N*-tuple taking as values with equal probability the permutations of* A_N. *Let* $\{n_i^{(N)}\}$ $(i = 1, \cdots, C; N = 1, 2, 3, \cdots)$ *be* C *sequences of non-negative integers such that*

$$\sum_{i=1}^{C} n_i^{(N)} = N, \qquad \lim_{N\to\infty} n_i^{(N)}/N = \nu_i$$

exists. Let $L_N^{(i)} = \sum Z_{N\alpha}$ *for* $i = 1, \cdots, C$ *where the summation is from* $\alpha = \sum_{j=1}^{i-1} n_j^{(N)} + 1$ *to* $\alpha = \sum_{j=1}^{i} n_j^{(N)}$. *Let* v_N *be the common variance of any* $Z_{N\alpha}$. *Then, asymptotically, the random variables* $[L_N^{(i)} - EL_N^{(i)}]/\sqrt{Nv_N}$ *have the singular* C*-variate normal distribution with mean zero and covariance matrix whose* i, i' *term is*

$$\delta_{ii'}\nu_i - \nu_i\nu_{i'}. \tag{4.1}$$

The proof of Theorem 4W follows from Corollary 1 of [8] via the technique used in Section 7 of [8], that is, via the consideration of arbitrary linear combinations of the random variables $[L_N^{(i)} - EL_N^{(i)}]/\sqrt{Nv_N}$. In order to save space the details are omitted. Note that for Theorem 4W itself it is not necessary to assume that the ν_i's are positive.

To apply Theorem 4W to our case, set $a_{N\alpha} = \alpha$ and observe that the resulting $\{A_N\}$ satisfies condition W of [8] (see, e.g., Section 3 of [8]). $L_N^{(i)}$ is called R_i, and it may be readily computed that $ER_i = \frac{1}{2} n_i^{(N)} (N + 1)$ and $v_N = (N^2 - 1)/12$. Hence the variables

$$\sqrt{12}\,\frac{R_i - n_i \dfrac{N+1}{2}}{N^{3/2}}$$

(dropping the superscript "N" for convenience) have asymptotically the singular multivariate normal distribution with zero mean and covariance matrix

given by (4.1). We next use the assumption that all the ν_i's are positive, from which it follows immediately that the variables

$$T_i = \sqrt{12}\,\frac{R_i - ER_i}{N^{3/2}\sqrt{n_i/N}}$$

have a joint asymptotic normal distribution with zero mean vector and with the covariance matrix whose i, i' term is $\delta_{ii'} - \sqrt{\nu_i \nu_{i'}}$. We now make the standard analysis of variance transformation

$$S_0 = \sum_{i'=1}^{C} \sqrt{\nu_{i'}}\, T_{i'}, \qquad S_i = \sum_{i'=1}^{C} e_{ii'} T_{i'} \qquad (i = 1, 2, \cdots, C-1)$$

with the e's chosen to make the transformation orthogonal. It follows that $\sum_{i=1}^{C} T_i^2$ is asymptotically chi-square with $C-1$ degrees of freedom. But

$$\sum_{i=1}^{C} T_i^2 = \frac{12}{N^2} \sum_{i=1}^{C} \frac{1}{n_i}\left(R_i - n_i \frac{N+1}{2}\right)^2 = \frac{N+1}{N} H.$$

Hence H is asymptotically chi-square with $C-1$ degrees of freedom.

It seems desirable to make a few comments regarding possible weakening of the conditions for an asymptotic chi-square distribution. In the first place, no great difficulty arises if some ν_i's should be zero—for example, suppose that $\nu_1 = 0$ and the other ν_i's are positive. Then $(R_1 - ER_1)/N^{3/2}$ approaches zero stochastically and $[N/(N+1)] \sum_{i=2}^{C} T_i^2$, that is, H computed from the sample *without* including R_1, is asymptotically chi-square with C−2 degrees of freedom. It is not however true in general that $\sum_{i=1}^{C} T_i^2$ is asymptotically chi-square; for example, consider the case of $n_1 = 1$ for all N. Analogous remarks apply if more than one $\nu_i = 0$.

If we use chi-square with $C-1$ degrees of freedom to approximate the critical region, then it may be wise to drop from the total sample any (sub)samples with n_i's very small. We would do this in order to obtain a better approximation to the critical region, at the expense of losing power against certain alternatives involving the populations from which the omitted observations arose. On the other hand, because of the smallness of the n_i's in question, even the exact critical region would probably have had little power against these alternatives.

Whitney in [10] uses a kind of limit requirement which might be thought applicable here and weaker than the existence of the ν_i's; that is to suppose the existence, for $i \neq j$, of

$$\tau_{ij} = \lim_{N\to\infty} \frac{n_i n_j}{(N - n_i)(N - n_j)}.$$

(Assume all $n_i < N$, so that the τ's are defined.) That this requirement is little weaker than ours except effectively for the case $C = 2$ is shown by the following lemma.

LEMMA 4.1. *If the ν_i's exist and no $\nu_i = 1$, then the τ_{ij}'s exist; and if $\nu_i = 0$, then every $\tau_{ij} = 0$ ($j = 1, 2, \cdots, C, j \neq i$). When $C \geqq 3$ if the τ_{ij}'s exist and at*

least two different τ_{ij}*'s are* $\neq 0$*, then the* ν_i*'s exist; and if* $\tau_{ij} = 0$*, then either* ν_i *or* $\nu_j = 0$.

PROOF. The first part is obvious. To prove the second choose an i, say $i = 1$ for convenience. Suppose that we can then find a j and k ($j \neq k; j, k \neq 1$) such that $\tau_{jk} \neq 0$. Then

$$\tau_{1j}\tau_{1k}/\tau_{jk} = \lim_{N\to\infty}\left(\frac{n_1}{N - n_1}\right)^2.$$

Hence ν_1 exists, and if $\tau_{1j} = 0$ then $\nu_1 = 0$. Next suppose that no such τ_{jk} exists, that is, that the only possible nonzero τ's are $\tau_{12}, \tau_{13}, \cdots, \tau_{1C}$. By hypothesis at least two of these must be nonzero, say τ_{12} and τ_{13}. Since $\tau_{23} = 0$, it follows as above that ν_2 and ν_3 exist and are zero. The same comment holds for any other ν_j for which $\tau_{1j} \neq 0$. Finally suppose a $\tau_{1j} = 0$ ($j \neq 2, 3$). Since $\tau_{12} \neq 0$ we have

$$\lim_{N\to\infty} \frac{n_j/(N - n_j)}{n_2/(N - n_2)} = 0.$$

But the denominator here approaches zero itself. Hence the limit of $n_j/(N-n_j)$ is zero, ν_j exists, and it is zero. Of course $\nu_1 = 1$.

If only one τ, say τ_{12}, is $\neq 0$, then $\nu_3, \nu_4, \cdots, \nu_C$ must all be 0. It can be shown that $\tau_{12} = 1$, as well, so that we are effectively in the $C = 2$ case. It is impossible for all the τ's to be zero.

The material of this section is summarized in the following theorem:

THEOREM 4. *If for all* i, *lim* $n_i/N = \nu_i$ *exists and is positive, then under the null hypothesis* H *is asymptotically distributed as chi-square with* $C - 1$ *degrees of freedom. If* p ν_i*'s are zero* ($p = 1, 2, \cdots, C - 1$), *then* H *computed with only the* R_i*'s corresponding to nonzero* ν_i*'s is asymptotically distributed under the null hypothesis as chi-square with* $C - p - 1$ *degrees of freedom.*

5. Consistency of the test based on large values of H. Suppose that the n_i's are functions of N, and consider the family of sequences of critical regions $H \geqq t_\alpha(N)$, where the level of significance $\alpha \in (0, 1)$ indexes the sequences of the family, N indexes the members of a sequence, and $t_\alpha(N)$ is the least number with the property $Pr\{H \geqq t_\alpha(N)\} \leqq \alpha$ under the null hypothesis. Let us say that this family of sequences is consistent against a given alternative if every member sequence is consistent in the usual sense against the given alternative, that is, if for all $\alpha \in (0, 1)$

$$\lim_{N\to\infty} \Pr\{H \geqq t_\alpha(N)\} = 1,$$

where the probabilities are taken under the alternative. For brevity we may simply say that the test based on large values of H is consistent. (Note that failure of consistency for the family of sequences against an alternative implies only that there is some α_0 such that for all $\alpha \leqq \alpha_0$ the sequence of tests $H \geqq t_\alpha(N)$ fails to be consistent in the usual sense.) This use of the word "consistent" will permit more compact statements and will not, I think, cause any confusion.

Under what circumstances is the test based on large values of H consistent? We consider alternatives of the following form: all the ξ's are independent, and $\xi_j^{(i)}$ has a continuous cdf F_i. Assume that as $N \to \infty$, $n_i/N = \nu_i + o(N^{-\frac{1}{2}})$ with $\nu_i > 0$; and note that this assumption subsumes the most natural case: $n_i = [\nu_i N]$ or $[\nu_i N] + 1$. Since $t_\alpha(N)$ for given α has as its limit the upper 100α-percent point of the chi-square distribution with C-1 degrees of freedom, it is equivalent to ask under what circumstances $\lim_{N\to\infty} \Pr\{H \geqq t\} = 1$ for all positive t. We may also replace H by $\sum T_i^2$ since the two differ by a factor of $N/(N+1)$.

First, we ask under what circumstances, for all positive t, $\lim_{N\to\infty} \Pr\{|\ T_1\ | \geqq t\} = 1$. Following the useful procedure of Mann and Whitney, set V = the number of couples $(X_j^{(1)}, X_{j'}^{(i)})$ where $i = 2$ or 3 or $\cdots C$ and for which $X_j^{(i)} > X_{j'}^{(i)}$. Then

$$R_1 = \tfrac{1}{2} n_1(n_1 + 1) + V.$$

This relationship holds for the special case $X_j^{(1)} \leqq n_1, j = 1, 2, \cdots, n_1$; for then $V = 0$. It holds in general, since an interchange of the superscripts of two adjacent X's, one from sample 1 and the other from sample $i \neq 1$, increases or decreases R_1 and V together by unity. Then

$$T_1 = \sqrt{\frac{12}{n_1 N^2}} \left\{ \frac{n_1(n_1+1)}{2} + V - n_1 \frac{N+1}{2} \right\}$$

$$= \sqrt{\frac{12}{n_1 N^2}} \{ V - \tfrac{1}{2} n_1 (N - n_1) \}.$$

Next define the following $n_1(N-n_1)$ counter-variables

$$Y_{jj'}^{(i)} = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \text{ when } X_j^{(1)} \begin{Bmatrix} \leqq \\ > \end{Bmatrix} X_{j'}^{(i)}$$

so that

$$V = \sum_{i=2}^{C} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_i} Y_{jj'}^{(i)}.$$

From now on we deal with a specific alternative F_i which will be described in slightly different terms further on.

LEMMA 5.1. *The set of values of Var T_1, as N runs through the positive integers, is bounded.*

PROOF. Set Var $Y_{j_1 j_2}^{(i)} = v_i$ and

$$\text{Cov}\,(Y_{j_1 j_2}^{(i)}, Y_{j_3 j_4}^{(i')}) = \begin{cases} c_{ii'} \text{ for } j_1 = j_3, \text{ and either } i \neq i' \text{ or } j_2 \neq j_4, \\ d_i \text{ for } i = i', j_2 = j_4, \text{ and } j_1 \neq j_3, \\ 0 \text{ otherwise (since } \xi\text{'s independent).} \end{cases}$$

Clearly the v's, c's, and d's are all less than 1 absolutely. So

$$\begin{aligned}\text{Var } T_1 &= \frac{12}{N^2 n_1} \text{ Var } V \\ &\leqq \{\text{number of } v_i \text{ terms} + \text{number of } c_{ii'} \text{ terms} + \text{number of } d_i \text{ terms}\} \\ &\leqq \frac{12}{N^2 n_1}\{n_1(N-n_1) + n_1(N-n_1)^2 + n_1^2(N-n_1)\} = \frac{12(N-n_1)(N+1)}{N^2}\end{aligned}$$

which is finite and has the limit $12(1-\nu_1) < \infty$. We next introduce the numbers

$$g_{i,i'} = \Pr\{X_j^{(i)} > X_{j'}^{(i')}\}$$

(under the alternative) for $i \neq i'$, and $g_{i,i'} = \frac{1}{2}$. Hence for $i > 1$ $g_{1,i} = EY_{jj'}^{(i)}$ and

$$ET_1 = \sqrt{\frac{12}{n_1 N^2}}\left[n_1 \sum_{i=2}^{c} n_i g_{1,i} - \tfrac{1}{2} n_1(N-n_1)\right].$$

Hence the limit of $ET_1/\sqrt{n_1}$ is

$$\sqrt{12}\left[\sum_{i=2}^{c} \nu_i g_{1,i} - \tfrac{1}{2}(1-\nu_1)\right] = \sqrt{12}\left[\sum_{i=1}^{c} \nu_i g_{1,i} - \tfrac{1}{2}\right],$$

whence the limit of ET_1 is

$$\left.\begin{matrix}\infty \\ 0 \\ -\infty\end{matrix}\right\} \text{as } \sum_{i=1}^{c} \nu_i g_{1,i} \left\{\begin{matrix}> \\ = \\ <\end{matrix}\right\} \tfrac{1}{2},$$

and we have

LEMMA 5.2. *If* $\sum_{i=1}^{c} \nu_i g_{1,i} \neq \frac{1}{2}$, *then* $\lim_{N\to\infty} \Pr\{|T_1| \geqq t\} = 1$. This follows immediately from Tchebycheff's inequality. Consequently we may state

LEMMA 5.3. *If for some* i, $\sum_{i'=1}^{c} \nu_{i'} g_{i,i'} \neq \frac{1}{2}$, *then the test based on large values of* H *is consistent.* We now turn to implications in the other direction.

LEMMA 5.4. *If* $\sum_{i=1}^{c} \nu_i g_{1,i} = \frac{1}{2}$, *then there exists a* t_0, *a function* $N_0(t)$, *and a decreasing function* $G(t)$ *such that*

(1) $\lim_{t\to\infty} G(t) = 0$,

(2) *For* $t > t_0$ *and* $N > N_0(t)$, $\Pr\{|T_1| \geqq t\} \leqq G(t) < 1$.

PROOF. Let K be an upper bound for Var T_1. Then by Tchebycheff's inequality, for $t > 0$

$$\Pr\{|T_1| \geqq t\} = \Pr\{T_1 \geqq t\} + \Pr\{T_1 \leqq -t\} \leqq \frac{K}{[t - ET_1]^2} + \frac{K}{[t + ET_1]^2}$$

which has the limit $2K/t^2$, since $ET_1 \to 0$. Putting $\epsilon(t) = [\max(1, t)]^{-1}/4$ and $t_0 = 2\sqrt{K}$, it follows that for any $t > t_0$, there is an $N_0(t)$ such that for $N > N_0(t)$

$$\Pr\{|T_1| \geqq t\} \leqq \frac{2K}{t^2} + \epsilon(t) < \tfrac{1}{2} + \tfrac{1}{4} < 1.$$

$G(t)$ is the function in the middle of the above double inequality. Next this is generalized as follows:

LEMMA 5.5. *If for all* i, $\sum_{i'=1}^{C} \nu_{i'} g_{i,i'} = \frac{1}{2}$, *then the test based on large values of* H *is not consistent.*

PROOF. By the previous lemma there are (for each i) numbers, $t_0^{(i)}$, and functions $N_0^{(i)}(t)$ and $G^{(i)}(t)$ such that $G^{(i)}(t) \to 0$ from above monotonically as $t \to \infty$, and such that for $t > t_0^{(i)}$ and $N > N_0^{(i)}(t)$, $Pr\{|\mathrm{T}_i| \geqq t\} \leqq G^{(i)}(t) < 1$. Let $t_0^* = \max_i t_0^{(i)}$, $N_0^*(t) = \max_i N_0^{(i)}(t)$, and $G^*(t) = \max_i G^{(i)}(t)$. Then $G^*(t)$ is a monotone decreasing function with limit 0, and for all i, $t > t_0^*$, and $N > N_0^*(t)$, $\Pr\{|T_i| \geqq t\} \leqq G^*(t) < 1$. For $t > t_0^*$ and $N > N_0^*(t)$, $\Pr\{\text{some } |T_i| \geqq t\} \leqq C \cdot G^*(t)$. But $\sum T_i^2 \geqq s > 0$ implies that some $|T_i| \geqq \sqrt{s/C}$. Hence for $s > C \cdot t_0^{*2}$, and $N > N_0^*(\sqrt{s/C})$ we have

$$\Pr\{\sum T_i^2 \geqq s\} \leqq C \cdot G^*(\sqrt{s/C})$$

To complete the proof take s large enough so that $C \cdot G^*(\sqrt{s/C}) < 1$.

It is natural to ask for a simple probabilistic interpretation of the necessary and sufficient condition for consistency which has been proven. This may be done as follows. Recall that we are still discussing a fixed alternative $\{F_i\}$ and that all probabilities are taken with respect to this alternative. Now let $\eta^{(1)}, \cdots, \eta^{(C)}$ be C independent random variables independent of all the ξ's and with cdf's F_i. Then

$$g_{i,i'} = \Pr\{\eta^{(i)} > \xi_j^{(i')}\}.$$

Next choose a $\xi_J^{(I)}$ at random from among the N possibilities (i.e., take an observation in the space of N ordered couples (I, J) where each has the same probability $1/N$.) Then

$$\Pr\{\eta^{(i)} > \xi_J^{(I)}\} = \frac{1}{N} \sum_{i'=1}^{C} \sum_{j=1}^{n_{i'}} g_{i,i'} = \sum_{i'=1}^{C} \frac{n_{i'}}{N} g_{i,i'}$$

so that the test based on large values of H is *in*consistent if and only if for all i, $\lim_{N\to\infty} \Pr\{\eta^{(i)} > \xi_J^{(I)}\} = \frac{1}{2}$. Roughly speaking this means that the test is consistent if and only if the variables from at least one population tend in the limit to be either larger or smaller than the other variables.

In particular we have consistency under the following circumstances which generalize to the C-population case the sufficient conditions for $C = 2$ given in [2] by Mann and Whitney[2]

$$F_1(x) < F_2(x), F_1(x) \leqq F_i(x) \qquad (i = 3, 4, \cdots, C)$$

for all x. (Of course the choice of subscripts 1 and 2 here is just for convenience.) To show that the consistency condition is satisfied, note that for $i = 3, 4, \cdots, C$

[2] The unnecessary specialization of the Mann and Whitney consistency condition when $C = 2$ was noted (separately) by Lehmann and van Dantzig; see p. 166 of [1] and [16]. In the latter both sufficiency and necessity are considered by a method similar to that of this paper, and further results are obtained. In 1948 E. J. G. Pitman gave the same necessary and sufficient condition for $C = 2$ during lectures at Columbia University.

$$g_{1,i} = \int_{-\infty}^{\infty} F_i(x)\, dF_1(x) \geqq \int_{-\infty}^{\infty} F_1(x)\, dF_1(x) = \tfrac{1}{2}$$

because of symmetry. However $g_{1,2} > \frac{1}{2}$; for let m run through the positive integers and set B_m equal to the set of all x satisfying $F_1(x) + 1/m \geqq F_2(x) > F_1(x) + 1/(m+1)$. Then

$$\begin{aligned} g_{1,2} &= \int_{-\infty}^{\infty} F_2(x)\, dF_1(x) = \sum_{m=1}^{\infty} \int_{B_m} F_2(x)\, dF_1(x) \\ &\geqq \sum_{m=1}^{\infty} \left\{ \int_{B_m} F_1(x)\, dF_1(x) + \frac{1}{m+1} \int_{B_m} dF_1(x) \right\} \\ &= \frac{1}{2} + \sum_{m=1}^{\infty} \frac{1}{m+1} \int_{B_m} dF_1(x), \end{aligned}$$

and clearly at least one B_m must have positive measure with respect to F_1. Hence $\sum_{i=1}^{C} \nu_i g_{1,i} > \frac{1}{2}$, and we have consistency. The circumstances just discussed include the translational sort of alternative described in the introductory paragraph of this paper.

A simple class of cases for which consistency fails and yet the null hypothesis need not hold is given by the following characteristic: that all the C distributions be symmetrical about the same point f in the following sense:

$$F_i(f-x) = 1 - F_i(f+x)$$

for all i and x. For, setting $f = 0$ without loss of generality, this means that the distribution of every ξ is the same as that of its negative. Hence for all i, i', $g_{i,i'} = g_{i',i} = \frac{1}{2}$ and consistency fails.

The material of this section may be summarized as follows.

THEOREM 5. *Suppose that the n_i's are functions of N and that for all i, $n_i/N = \nu_i + o(N^{-\frac{1}{2}})$ and $\nu_i > 0$. For each level of significance $\alpha (0 < \alpha < 1)$ consider the sequence of tests: reject the null hypothesis if $H \geqq t_\alpha(N)$ where $t_\alpha(N)$ is the least number giving rise to level of significance α at the Nth step. Then these sequences of tests are all consistent against a given continuous alternative $\{F_i\}$ if and only if for some i, with probabilities taken under the alternative*

$$\sum_{i'=1}^{C} \nu_{i'} [\Pr\{\eta^{(i)} > \eta^{(i')}\} + \tfrac{1}{2} \Pr\{\eta^{(i)} = \eta^{(i')}\}] \neq \tfrac{1}{2},$$

where the $\eta^{(i)}$'s are C independent random variables having respectively the cdf's F_i. The sufficiency of the above condition holds regardless of the order of $(n_i/N) - \nu_i$. When $C = 2$ the denial of the above condition implies $g_{12} = g_{21} = \frac{1}{2}$.

6. The variance of H under the null hypothesis. As an aid in approximating the distribution of H when the null hypothesis is true, we seek the variance of H under the null hypothesis. This seems to be a tedious computation by any method; we shall outline a direct method, omitting most of the routine algebra. Directly from the definition of H we have

$$\text{(6.1)} \quad \operatorname{Var} H = \frac{144}{N^2(N+1)^2} \left\{ \sum_i \frac{1}{n_i^2} ER_i^4 + \sum_{i \neq j} \frac{1}{n_i n_j} E(R_i^2 R_j^2) - \left[\sum_i \frac{1}{n_i} ER_i^2 \right]^2 \right\}.$$

$E\,R_i^4$ is readily found from formula (8) of [2], which when translated into our notation says

$$E\left(R_i - n_i\frac{N+1}{2}\right)^4 = \frac{n_i(N-n_i)(N+1)}{240}$$
$$[5Nn_i(N-n_i) - 2n_i^2 - 2(N-n_i)^2 + 3n_i(N-n_i) - 2N].$$

From this

$$\frac{1}{n_i^2}ER_i^4 = \frac{N+1}{240}\Big\{n_i^2[15N^3+15N^2-10N-8] + n_i\,[30N^3+50N^2+16N]$$
$$+ [5N^3+9N^2+2N] - \frac{1}{n_i}[2N^3+2N^2]\Big\},$$

and, summing over i

$$\sum_i \frac{1}{n_i^2}ER_i^4 = \frac{N+1}{240}[15N^3+15N^2-10N-8]\sum_i n_i^2$$
$$(6.2)\quad + \frac{N^2(N+1)}{240}[30N^2+50N+16] + \frac{C(N+1)N}{240}[5N^2+9N+2]$$
$$- \frac{N+1}{240}[2N^3+2N^2]\sum_i \frac{1}{n_i}.$$

Next we find $E(R_1^2R_2^2)$ as follows:

$$E(R_1^2R_2^2) = \sum_i\sum_{i'}\sum_j\sum_{j'} E[X_i^{(1)}\,X_{i'}^{(1)}\,X_j^{(2)}\,X_{j'}^{(2)}],$$

where $i, i' = 1, 2, \cdots, n_1$ and $j, j' = 1, 2, \cdots, n_2$. This quantity is

$$n_1n_2(n_1-1)(n_2-1)E[X_1^{(1)}\,X_2^{(1)}\,X_1^{(2)}\,X_2^{(2)}] + n_1n_2(n_1-1)E[X_1^{(1)}\,X_2^{(1)}\,X_1^{(2)^2}]$$
$$+ n_1n_2(n_2-1)\,E\,[X_1^{(1)^2}X_1^{(2)}\,X_2^{(2)}] + n_1n_2E[X_1^{(1)^2}\,X_1^{(2)^2}]$$
$$= \frac{n_1n_2(n_1-1)\,(n_2-1)}{N(N-1)(N-2(N-3)}\sum pp'\,qq' + \frac{n_1n_2(n_1+n_2-2)}{N(N-1)(N-2)}\sum p^2qq'$$
$$+ \frac{n_1n_2}{N(N-1)}\sum p^2q^2,$$

where the p's and q's run from 1 to N and within any term of a summation no two are equal. This simplifies after some algebraic labor; we divide by n_1n_2 and make the obvious generalization from (1, 2) to (i, j) to find

$$\frac{1}{n_in_j}E(R_i^2R_j^2) = (n_i-1)(n_j-1)\frac{N+1}{240}\,[15N^3+15N^2-10N-8]$$
$$+ (n_i+n_j-2)\frac{N+1}{360}\,[30N^3+35N^2-11N-12]$$
$$+ \frac{N+1}{180}\,[20N^3+24N^2-5N-6].$$

Sum this over $i \neq j$ to obtain

$$\begin{aligned}(6.3)\qquad &\frac{N+1}{240}[(N-C)^2+2N-C-\sum_i n_i^2][15N^3+15N^2-10N-8]\\ &+\frac{N+1}{360}[2(C-1)(N-C)][30N^3+35N^2-11N-12]\\ &+\frac{N+1}{180}C(C-1)[20N^3+24N^2-5N-6].\end{aligned}$$

Next, we note that

$$\frac{1}{n_i}ER_i^2=\frac{N+1}{12}(N-n_i)+n_i\frac{(N+1)^2}{4}=\frac{N+1}{12}[n_i(3N+2)+N]$$

which, summed over i and squared, gives

$$(6.4)\qquad \frac{N^2(N+1)^2}{144}[9N^2+6N(C+2)+(C+2)^2].$$

Finally, substituting (6.2), (6.3), and (6.4) in (6.1), and simplifying, we obtain

$$(6.5)\qquad \begin{aligned}&\operatorname{Var} H\\ &=2(C-1)-\frac{2}{5N(N+1)}[3C^2-6C+N(2C^2-6C+1)]-\frac{6}{5}\sum_i\frac{1}{n_i}.\end{aligned}$$

Note that as all the n_i's $\to \infty$, Var $H \to 2(C-1)$ = Var (chi-square with C$-$1 degrees of freedom).

7. The maximum value of H. It is an aid in approximating the distribution of H to know its maximum value. This may be obtained from the well-known analysis of variance algebraic identity

$$(7.1)\qquad \begin{aligned}\frac{N(N+1)}{12}H+\sum_{i=1}^{C}\sum_{j=1}^{n_i}(X_j^{(i)}-\bar{R}_i)^2&=\sum_{i=1}^{C}\sum_{j=1}^{n_i}\left(X_j^{(i)}-\frac{N+1}{2}\right)^2\\ &=\frac{1}{12}N(N^2-1).\end{aligned}$$

A sample point maximizing H is a sample point minimizing the second term on the left side of the above identity, that is the within sum of squares. Clearly this sum of squares is minimized when the ranks within each (sub) sample form consecutive integers, that is $X_j^{(i)}-\bar{R}_i = j-\frac{1}{2}(n_i-1)$, so that

$$\begin{aligned}\sum_{i=1}^{C}\sum_{j=1}^{n_i}(X_j^{(i)}-\bar{R}_i)^2&=\frac{1}{12}\sum_{i=1}^{C}n_i(n_i^2-1)\\ &=\frac{1}{12}(\sum n_i^3-N).\end{aligned}$$

Substituting back in (7.1) it follows that the maximum value of H is

$$(7.2)\qquad (N-1)+\frac{N-\sum n_i^3}{N(N+1)}=\frac{N^3-\sum n_i^3}{N(N+1)}.$$

8. On the distribution of the R_i's and H. If we set

$$\Gamma(r_1, r_2, \cdots, r_C; n_1, n_2, \cdots, n_C) = \frac{\left(\sum_{i=1}^{C} n_j\right)!}{n_1! \cdots n_C!} Pr\{R_i = r_i \quad (i = 1, 2, \cdots, C) \text{ with stated } n_i\text{'s}\},$$

then, under the null hypothesis, the following difference equation holds for Γ.

$$(8.1)\quad \Gamma(r_1, \cdots, r_C; n_1, \cdots, n_C) = \sum_{i=1}^{C} \Gamma\left(r_1, \cdots, r_{i-1}, r_i - \sum_{j=1}^{C} n_j, r_{i+1}, \cdots, r_C; n_1, \cdots, n_{i-1}, n_i - 1, n_{i+1}, \cdots, n_C\right)$$

with the following boundary conditions:

1. If any argument fails to be a non-negative integer, $\Gamma = 0$.
2. If $r_i n_i = 0$, but $r_i + n_i \neq 0$, then $\Gamma = 0$.
3. When $n_1 = n_2 = \cdots = n_C = 0$, $\Gamma = 1$ when all r_i's are 0, and otherwise $\Gamma = 0$.

The above equation and conditions follow readily from the partition of the chance event $\{R_1 = r_1, \cdots, R_C = r_C\}$ into the C chance events $\{R_1 = r_1, \cdots, R_C = r_C$, and $\max_{i',j} X_j^{(i')}$ is an $X_j^{(i)}\}$ for $i = 1, 2, \cdots, C$.

I have been unable to find a closed solution for this equation. For $C = 2$ and small n_i's, values of $[n_1!n_2!/(n_1+n_2)!]\ \Gamma(r_1,r_2; n_1, n_2)$ are given in [2]. (Comment on notation: m, n, and U of [2] correspond to our n_2, n_1, and $R_1 - \frac{1}{2}\ n_1(n_1+1)$ respectively. The tables of [2] actually give the cumulative probabilities.) For $C = 3$ and for small values of the n_i's, recursive computations based on (8.1) are being carried out in order to obtain exact distributions of H. It is hoped that from these exact distributions some idea of the accuracy of various approximations may be obtained.

In Section 6 formula (6.5) for the variance of H under the null hypothesis was obtained as a function of N, C, and $\sum 1/n_i$. It seems reasonable to attempt to better the chi-square approximation by fitting a Type III distribution with density function

$$\frac{a^\nu}{\Gamma(\nu)} t^{\nu-1} e^{-at}$$

for $t \geqq 0$, and 0 for $t < 0$. Equating first and second moments

$$a = (C-1)/\text{Var } H, \qquad \nu = (C-1)^2/\text{Var } H.$$

Equivalently, one may approximate $Pr\{H \leqq x\}$ under the null hypothesis by the use of K. Pearson's incomplete Γ function tables [5], setting u and p of those tables equal respectively to

$$x/\sqrt{\text{Var } H}, \qquad (C-1)^2/\text{Var } H - 1.$$

Again equivalently, one can make the same approximation by interpolation in a chi-square table using $2(C-1)^2/\text{Var } H$ degrees of freedom and argument $2x(C-1)/\text{Var } H$.

In Section 7 formula (7.2) for the maximum value of H, say H_M, was obtained in terms of N and $\sum n_i^3$. It seems reasonable to attempt to better the chi-square and Type III approximations by fitting an incomplete B distribution and using [6] to obtain approximate probabilities. Thus, equating moments again, we may approximate $\Pr\{H \leqq x\}$ under the null hypothesis by $I_{x/H_M}(p, q)$, in the notation of [6], where

$$p = \frac{C-1}{H_M} \cdot \frac{(C-1)(H_M - C + 1) - \text{Var } H}{\text{Var } H}, \qquad q = \frac{H_M - C + 1}{C-1} p.$$

The above formulas are given for convenient reference. The relative merits of these approximations will be discussed in [15].

9. The possibly discontinuous case. Much of the preceding material carries almost directly over to the general case in which the cdf's need not be continuous, providing that the following randomization convention is followed: when two or more ξ's are equal, define the corresponding X's at random. More precisely, suppose that $\xi_{j_1}^{(i_1)} = \xi_{j_2}^{(i_2)} = \cdots = \xi_{j_\omega}^{(i_\omega)}$ for a given sample point, and that all other ξ's are unequal to the common value of the above ω ξ's with (say) λ ξ's less than the common value. Then assign ranks $\lambda+1, \lambda+2, \cdots, \lambda+\omega$ to the tied ξ's by performing a random experiment in which each of the $\omega!$ possible assignments is an equally likely outcome. With this convention the joint distribution of the X's under the null hypothesis is the same as that stated in Section 2, so that the asymptotic chi-square distribution (Section 4) holds.

The following minor changes would be made in Section 5:

(1) In the discussion of the intuitive interpretation for the consistency condition, replace the given expression for $g_{i,i'}$ by $g_{i,i'} = \Pr\{\eta^{(i)} > \xi_j^{(i')}\} + \frac{1}{2}\Pr\{\eta^{(i)} = \xi_j^{(i')}\}$, and replace the necessary and sufficient condition for inconsistency by

$$\lim_{N\to\infty} [\Pr\{\eta^{(i)} > \xi_J^{(I)}\} + \tfrac{1}{2} \Pr\{\eta^{(i)} = \xi_J^{(I)}\}] = \tfrac{1}{2}, \text{ for all } i.$$

(2) In the discussion of consistency when $F_1 < F_2$, and $F_1 \leqq F_i$ $(i = 3, 4, \cdots, C)$, insert the remark that the result continues to hold if we consider the cdf's not in one of the usual senses (i.e., continuous to the left or to the right), but rather in the sense of Lévy: $\frac{1}{2}F(x^-) + \frac{1}{2}F(x^+)$. The same interpretation of the cdf notation should be made in the discussion of a class of cases for which consistency fails.

(3) Delete the word "continuous" in the statement of Theorem 5.

Another way to treat ties, much discussed in connection with the rank correlation coefficient, is to give tied ξ's equal fractional ranks so as to keep the sum of ranks at its usual value; i.e., in the notation of the first paragraph of this section, assign the fractional rank $\lambda + \frac{1}{2}(\omega + 1)$ to all the ω tied ξ's. We proceed to show

that if we do this, and also change the norming constants appropriately, the altered H still is asymptotically chi-square with $C - 1$ degrees of freedom.

Suppose that there are K groups of ties with, respectively, $\omega_1, \omega_2, \cdots, \omega_K$ members. We agree to use mean ranks in the tied groups and to work in the conditional distribution wherein just K tied groups exist of sizes $\omega_1, \cdots, \omega_K$ and covering fixed rank intervals, but permitting the numbers of observations from the C subsamples falling in any tied group to vary. In other words, instead of the finite population $(1, 2, \cdots, N)$, we deal with

$$
\begin{aligned}
&1, 2, \cdots, \lambda_1, \\
&\lambda_1 + \tfrac{1}{2}(\omega_1 + 1), \cdots, \lambda_1 + \tfrac{1}{2}(\omega_1 + 1), \qquad \lambda_1 + \omega_1 + 1, \cdots, \lambda_2, \\
&\lambda_2 + \tfrac{1}{2}(\omega_2 + 1), \cdots, \lambda_2 + \tfrac{1}{2}(\omega_2 + 1), \qquad \lambda_2 + \omega_2 + 1, \cdots, \lambda_3, \\
&\quad\vdots \\
&\lambda_K + \tfrac{1}{2}(\omega_K + 1), \cdots, \lambda_K + \tfrac{1}{2}(\omega_K + 1), \qquad \lambda_K + \omega_K + 1, \cdots, N,
\end{aligned} \tag{9.1}
$$

where $\lambda_k + \frac{1}{2}(\omega_k + 1)$ occurs ω_k times. Under the null hypothesis the ordered N-tuple of $X_j^{(i)}$'s takes as its values the permutations of the above finite population, all with equal probability. We compute that $ER_i = \frac{1}{2}n_i(N + 1)$, as before, and that

$$
\operatorname{Var} R_i - \frac{n_i(N - n_i)(N + 1)}{12} = -\frac{n_i(N - n_i)}{N(N - 1)} \sum_{k=1}^{K} \frac{1}{12} \omega_k(\omega_k - 1)(\omega_k + 1),
$$

$$
\operatorname{Cov}(R_i, R_{i'}) + \frac{n_i n_{i'}(N + 1)}{12} = \frac{n_i n_{i'}}{N(N - 1)} \sum_{k=1}^{K} \frac{1}{12} \omega_k(\omega_k - 1)(\omega_k + 1),
$$

so that, setting

$$
\gamma = \sum_{k=1}^{K} \omega_k(\omega_k - 1)(\omega_k + 1),
$$

we have

$$
E[(R_i - ER_i)(R_{i'} - ER_{i'})] = \frac{1}{12}[n_i N \delta_{ii'} - n_i n_{i'}] \frac{N^3 - N - \gamma}{N(N - 1)}
$$

or the corresponding second moment in the untied case times

$$
[N^3 - N - \gamma]/[N^3 - N].
$$

Now let the λ_k's, the ω_k's and the n_i's all be functions of N, and assume that $\lim_{N\to\infty} n_i/N = \nu_i > 0$ exists, $\lim_{N\to\infty} \gamma/N^3 = \gamma^*$ exists and $\gamma^* \neq 1$. To say that $\gamma^* \neq 1$ is to say that $\operatorname{Max}_k \omega_k/N$ does not approach 1, and one can readily show then that the sequence of finite populations (9.1) satisfies condition W of Theorem 4W. It follows from Theorem 4W that the variables

$$
\sqrt{12}\, \frac{R_i - n_i \dfrac{N + 1}{2}}{\sqrt{N^3 - N - \gamma}}
$$

are asymptotically multivariate normal with zero mean and covariance matrix given by (4.1). Hence, just as in Section 4

$$(9.2)\qquad \frac{12}{N(N+1)\left[1-\frac{\gamma^*}{N(N^2-1)}\right]}\sum_{i=1}^{c}\frac{1}{n_i}\left(R_i - n_i\frac{N+1}{2}\right)^2 = H^*$$

(say) is asymptotically chi-square with C − 1 degrees of freedom and has expected value C − 1. Note that no limit condition on the λ_k's is needed.

10. Further work. It would be interesting to investigate further the power function of the test described in this paper, perhaps along the lines of [1], or by considering its asymptotic relative efficiency to ordinary one-way analysis of variance in the normal case. Again in the spirit of [1], it would seem desirable to propose and investigate related tests specifically designed to be powerful against more restricted classes of alternatives, e.g., $F_1 \geqq F_2 \geqq \cdots \geqq F_c$, with at least one inequality strong.[3] Another extension is to consider the use of H-like tests in two-way analyses of variance or more general linear hypothesis situations, in a manner analogous to that of [4].

11. Acknowledgments. The test discussed here was suggested to me in a slightly variant form by W. A. Wallis. I wish to acknowledge with gratitude Professor Wallis' encouragement and helpful suggestions in carrying through the work reported here. In particular Professor Wallis suggested the applicability of the mean rank approach in the case of ties and obtained the proper norming constant. I wish also to thank D. A. S. Fraser for his derivation, replacing a longer version, of the maximum value of H.

REFERENCES

[1] E. L. Lehmann, "Consistency and unbiasedness of certain nonparametric tests" *Annals of Math. Stat.*, Vol. 22 (1951), pp. 165–179.

[2] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[3] F. J. Massey, Jr., "A note on a two sample test," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 304–306.

[4] A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill Book Co., New York, 1950.

[5] K. Pearson, *Tables of the Incomplete* Γ*-Function*, His Majesty's Stationery Office, London, 1922.

[6] K. Pearson, *Tables of the Incomplete* B*-function*, Cambridge University Press, 1934.

[7] H. Scheffé, "Statistical inference in the non-parametric case," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 305–332.

[8] A. Wald and J. Wolfowitz, "Statistical tests based on permutations of the observations," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 358–372.

[9] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 147–162.

[10] D. R. Whitney, "A bivariate extension of the U statistic," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 274–282.

[3] A recent investigation in this direction is that of T. J. Terpstra [17].

[11] F. WILCOXON, "Individual comparison by ranking methods," *Biometrics Bull.*, Vol. 1 (1945), p. 80–83.

[12] W. A. WALLIS, "Rough and ready statistical tests," *Industrial Quality Control*, Vol. 8 (1952), p. 35–40.

[13] J. WOLFOWITZ, "Non-parametric statistical inference," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949.

[14] A. W. MARSHALL, "A large sample test of the hypothesis that one of two random variables is stochastically larger than the other," *Jour. Am. Stat. Assn.*, Vol. 46 (1951), pp. 366–374.

[15] W. H. KRUSKAL AND W. A. WALLIS, "Use of ranks in one-criterion analysis of variance," *Jour. Am. Stat. Assn.*, Vol. 47 (1952), pp. 583–621.

[16] D. VAN DANTZIG, "On the consistency and power of Wilcoxon's two sample test," *Indagationes Mathematicae*, Vol. 13 (1951), pp. 1–8.

[17] T. J. TERPSTRA, "The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking," *Indagationes Mathematicae*, Vol. 14 (1952), pp. 327–333.