A NONPARAMETRIC TEST FOR THE SEVERAL
SAMPLE LOCATION PROBLEM

by

V. P. Bhapkar

University of North Carolina and
University of Poona

Institute of Statistics Mimeo Series No. 411

October 1964

This research was supported by Air Force Office
of Scientific Research Grant No. 84-63.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.
A NONPARAMETRIC TEST FOR THE SEVERAL
SAMPLE LOCATION PROBLEM

by

V. P. Bhapkar

University of North Carolina and
University of Poona

1. Summary. This paper offers a new nonparametric test of the null hypothesis \( F_1 = F_2 = \ldots = F_c \) against alternatives of the form
\[ F_i(x) = F(x - \theta_i) \quad (i = 1, 2, \ldots, c), \]
where the \( \theta_i \)'s are not all equal and \( F_i \) is the unknown (continuous) cumulative distribution function of the univariate population from which the \( i \)th random sample comes. It is based on \( c \)-plets that can be formed by choosing one observation from each sample. The asymptotic distribution of the new test statistic, \( W \), is shown to be the chi-square distribution with \( c - 1 \) degrees of freedom, under quite general conditions, when the null hypothesis holds. The asymptotic power of the test is computed for translation-type alternatives and it is shown that the test is asymptotically as efficient, in the Pitman-sense, as the Kruskal-Wallis \( H \)-test.

2. Introduction. Let \( \{x_{ij}, \ j = 1, 2, \ldots, n_i\} \) be a random sample from the \( i \)th population with continuous c.d.f. \( F_i, \ i = 1, 2, \ldots, c \), and suppose that these samples are independent. We consider a nonparametric test of the hypothesis
\[ H_0: \ F_1 = F_2 = \ldots = F_c \]
against alternatives of the form \( F_i(x) = F(x - \theta_i) \) with the \( \theta_i \)'s not all equal. Reference to prior work and some of the recent work may be found in [7], [6], [4], [2] and [3].
The observations, when regarded as random variables, will be represented by the corresponding capital letters. Let

\[ \phi_i(x_1, x_2, \ldots, x_c) = r - 1, \]

where \( r \) is the rank of \( x_i \) in \( x_1, x_2, \ldots, x_c \) arranged in increasing order, \( i = 1, 2, \ldots, c \). Since the distributions are assumed to be continuous, the probability that any two \( x \)'s are equal in zero. Let

\[ v_i = \sum_{t_1=1}^{n_1} \sum_{t_2=1}^{n_2} \sum_{t_c=1}^{n_c} \phi_i(x_{1t_1}, x_{2t_2}, \ldots, x_{ct_c}). \]

It is then seen that

\[ v_i = \sum_j \sum_{r=1}^{n(r)} n_{ij}, \]

where \( n_{ij}^{(r)} \) is the number of c-plets that can be formed by taking one observation from each sample, \( x_{ij} \) being the observation from the \( j \)th sample, such that \( x_{ij} \) has rank \( r \) in each of these c-plets. Let \( u_i = v_i/n_1 n_2 \ldots n_c \) and \( N = \Sigma_i n_i \). Then the statistic now being proposed is

\[ w = \frac{12}{c^2} \left[ \sum_{i=1}^{c} n_i u_i^2 - \frac{\left( \Sigma_i n_i u_i \right)^2}{N} \right]. \]

It is seen that \( w \) may be regarded as a suitable measure of deviation from the null hypothesis \( H_0 \) since \( w = \left( \frac{12}{c^2} \right) \Sigma_i n_i (u_i - \overline{u})^2 \), where \( \overline{u} = \Sigma_i n_i u_i / N \) and random variables \( U_i \)'s are expected to be equal when \( H_0 \) holds. The test consists in rejecting \( H_0 \) at a significance level \( \alpha \) if \( w \) exceeds some pre-determined number \( w_\alpha \). In the next section it is shown that, when \( H_0 \) is true, \( W \) is asymptotically distributed as a chi-square variable with \( c-1 \) degrees of freedom. Thus a large sample approximation for \( W_\alpha \) is provided by the upper \( \alpha \)-point of the \( \chi^2 \) distribution with \( c-1 \) degrees of freedom.

It can be seen that \( \Sigma_i V_i = (n_1 n_2 \ldots n_c)c(c-1)/2 \), so that \( \Sigma_i U_i = c(c-1)/2 \). Then for \( c = 2 \), i.e., for the two-sample problem the statistic
W is seen to be equivalent to the Mann-Whitney [9] statistic \( |U_n n_2^{1/2}| \), where \( U \) is the number of pairs \( (x_{\alpha_1}, x_{\beta_2}) \) with, say, \( x_{\alpha_1} > x_{\beta_2} \). It is also known that the Mann-Whitney \( U \)-test is equivalent to the Wilcoxon [10] test based on \( \overline{R}_l \), the mean rank of the first sample. The multisample analogue of the Wilcoxon statistic is provided by the Kruskal-Wallis [7] \( H \)-statistic based on \( \overline{R}_l \)'s. The motivation behind the tests based on \( c \)-plets is to use, for the case of \( c \) samples, Mann-Whitney-type test statistics. In [2] a test-statistic, \( V \), has been offered; it is based on the number of \( c \)-plets that can be formed by choosing one observation from each sample such that the observation from the \( i \)-th sample is the least \((i = 1, 2, \ldots, c)\). It was shown to be consistent for the class of translation alternatives and asymptotically more efficient, in the Pitman sense, than the \( H \)-statistic for some distributions. But it was asymptotically less efficient for normal distribution. Deshpande [3] proposed a statistic based on the numbers of \( c \)-plets such that the observation from the \( i \)-th sample is (i) the least or (ii) the largest. That statistic also suffers from a similar drawback. The statistic being proposed now extracts, presumably, more information with the result that it is asymptotically as efficient, as will be shown later, as the \( H \)-statistic. In fact, it can be seen that

\[
(2.4) \quad \Phi_i (x_1, x_2, \ldots, x_c) = \sum_{j=1}^{c} \phi_{ij}(x_i, x_j),
\]

where

\[
(2.5) \quad \phi_{ij}(x_i, x_j) = 1 \quad \text{if} \quad x_i > x_j
\]

\[
\text{otherwise}
\]

Then, from (2.2),

\[
v_i = n_1 n_2 \cdots n_c \sum_{j=1}^{c} u_{ij},
\]

so that
(2.6) \[ u_i = \sum_{j=1}^{c} u_{ij}, \]

where

(2.7) \[ u_{ij} = \sum_{t_i=1}^{n_i} \sum_{t_j=1}^{n_j} \phi_j(x_{it_i}, x_{jt_j})/n_i n_j, \quad i \neq j \]

and \( u_{ii} = 0 \). In the special case \( n_1=n_2=\ldots=n_c=n \), say, we have \( n^2 u_1 = n \left[ \bar{R}_i - (n+1)/2 \right] \), where \( \bar{R}_i \) is the mean rank of the \( i \)th sample; thus, in this case, \( W \) statistic is equivalent to the \( H \)-statistic. Such a simple relation does not exist if the \( n_i \)'s are not all equal.

The asymptotic distribution of \( W \) under \( H_0 \).

From (2.2) it is seen that \( U_1 \) is a generalized \( U \)-statistic corresponding to \( \phi_i \). From the \( c \)-sample version (e.g. see [2]) of Hoeffding's theorem [5] on \( U \)-statistics, it then follows that \( N^{\frac{1}{2}} [U_N - \mu] \) is, in the limit as \( n_i \to \infty \) in such a way that \( n_i = Np_i \), the \( p \)'s being fixed positive numbers such that \( \sum_p p_i = 1 \), normally distributed with zero mean and covariance matrix \( \Sigma = (\sigma_{rs}) \) given by

(3.1) \[ \sigma_{rs} = \sum_{i=1}^{c} \frac{1}{p_i} \xi^{(i)}(r,s), \quad r, s = 1, 2, \ldots, c, \]

where

(3.2) \[ \xi^{(i)}(r,s) = \xi \left[ \phi_i(X_1, X_2, \ldots, X_c) \phi_s(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_c) \right] - \eta_r \eta_s, \]

where \( X_j, X'_j \) are independent random variables with c.d.f. \( F_j \) (\( j=1, 2, \ldots, c \)).
Now, when $H_0$ holds, $F_1 = F_2 = \cdots = F_c = F$, say. Then

$$\eta_i = \sum_{j=1}^c \mathbb{E} \left[ \phi_{ij}(X_i, X_j) \right] = (c-1)/2.$$ Here, and hereafter in this section, $X$'s are independent random variables each with c.d.f. $F$. Also

$$\xi^{(1)}(i,i) = \mathbb{E} \left[ \phi_{ij}(X_i, X_j) \right] \left[ \mathbb{E} \phi_{ik}(X_i, X_k) \right] - (c-1)^2/4$$

$$= \sum_{j \neq i} \sum_{k} \mathbb{E} \left[ \phi_{ij}(X_i, X_j) \phi_{ik}(X_i, X_k) \right] - (c-1)^2/4$$

$$\xi^{(3)} = \sum_{j \neq i} \left( (1/3) - (c-1)^2/4 = (c-1)^2/12 \right),$$

$$\xi^{(2)}(i,i) = \mathbb{E} \left[ \phi_{ir}(X_i, X_r) \right] \left[ \mathbb{E} \phi_{is}(X_i, X_s) \right] + \phi_{ij}(X_i, X_j) - (c-1)^2/4$$

$$\xi^{(4)} = \sum_{r \neq i} \left( \sum_{s \neq i} \left( (1/4) + \sum_{r \neq j} \mathbb{E} \phi_{is}(X_i, X_s) \right) \right) - (c-1)^2/4$$

and finally

$$\xi^{(5)} = \sum_{r \neq i} \left( \sum_{s \neq k} \left( (1/4) + \sum_{r \neq k} \mathbb{E} \phi_{is}(X_i, X_s) \right) \right) - (c-1)^2/4$$

Thus, when $H_0$ holds, we have

$$\sigma_{ii} = \frac{(c-1)^2}{12p_i} + \sum_{j \neq i} \frac{1}{12p_j}$$

$$\sigma_{ij} = \frac{1}{12} \sum_{k \neq i} \frac{1}{p_k} - \frac{c-1}{12p_i} - \frac{c-1}{12p_j}.$$
Thus

\[(3.8) \quad 12 \Sigma = c^2 \Sigma - c \Sigma j' - c j q' + a j, \]

where \( \Sigma = \text{diagonal } (p_i, i=1,2,\ldots,c) \), \( a = \sum_{i=1}^{c} (1/p_i) \),

\( \Sigma j = (1)_{c \times c}, \quad j' = (1)_{1 \times c} \) and \( q' = (1/p_1,\ldots,1/p_c) \).

Since \( \Sigma U_{IN} = c(c-1)/2 \), the distribution of \( U_{IN} \) is singular and, hence, the asymptotic normal distribution of \( \sqrt{N} (U_{IN} - \eta) \) is also singular. In fact, it can be verified that \( \Sigma j = 0 \). Then arguing exactly as in [2] it follows that

\[(3.9) \quad W = \frac{12N}{c^2} \left[ \sum_{i=1}^{c} p_i (U_{IN} - \frac{c-1}{2})^2 - \left\{ \sum_{i=1}^{c} p_i (U_{IN} - \frac{c-1}{2}) \right\}^2 \right] \]

\[= \frac{12N}{c^2} \left[ \sum_{i=1}^{c} p_i U_{IN}^2 - \left\{ \sum_{i=1}^{c} p_i U_{IN} \right\}^2 \right] \]

has asymptotically chi-square distribution with \( c-1 \) degrees of freedom under \( H_0 \). Thus suppressing \( N \) in the subscript of \( U \), we have the statistic (2.3) proposed earlier.

\[\text{\underline{4. Consistency of the W-test:}} \text{ We quote here the following extension [2] of a lemma due to Lehmann [8]:} \]

Let \( \eta_i = \sum(i)(F_1,F_2,\ldots,F_c) \), \( i=1,2,\ldots,g \), be real-valued functions such that \( f(i)(F,\ldots,F) = \eta_{i0} \) for all \( (F,F,\ldots,F) \) in a class \( C_0 \). Let

\[T(i)_{n_1,\ldots,n_c} = t(i)(X_{11},\ldots,X_{1n_1};\ldots;X_{c1},\ldots,X_{cn_c}) \], \( i=1,2,\ldots,g \), be sequences of real-valued statistics such that \( T(i)_{n_1,\ldots,n_c} \) tends to \( \eta_i \) in probability as \( \min(n_1,\ldots,n_c) \rightarrow \infty \). Suppose that \( f(i)(F_1,F_2,\ldots,F_c) \neq \eta_{i0} \) for some \( i \) for all \( (F_1,\ldots,F_c) \) in a class \( C_1 \). Further let

\[W_{n_1,\ldots,n_c} = W(T(i)_{n_1,\ldots,n_c},\ldots,T(g)_{n_1,\ldots,n_c})\]
be a nonnegative function which is zero if, and only if, $\tau^{(i)}_{n_1, \ldots, n_c} = \eta_{i0}$ for all $i=1,2,\ldots,g$. Then the sequence of tests which reject when $W_{n_1, \ldots, n_c} > d_{n_1, \ldots, n_c}$ is consistent for testing $H_0 : C_o$ at every fixed level of significance against the alternatives $C_1$.

If we take $\eta_i = \mathbb{E} \left[ \phi_i(X_1, X_2, \ldots, X_c) \right]$, where the $X_i$'s are independent random variables with continuous c.d.f. $F_1, \ldots, F_c$ respectively, and $r_{n_1, n_2, \ldots, n_c} = U_{1N}$, then the convergence in probability of $U_{1N}$ to $\eta_i$ follows from the asymptotic normality of $\sqrt{N}(U_{1N} - \eta_i)$. For the class $C_1$ of translation-type alternatives $F_i(x) = F(x - \theta_i)$, with $\theta$'s not all equal, it may be easily seen that $\eta > (c-1)/2$, i.e. $\eta_{i0}$, where $\eta_{i0}$ is the (or one of the) largest among $\theta$'s. The $W$-test, thus, is seen to be consistent against the class of translation-type alternatives.

More generally, the $W$-test is consistent against the wider class of alternatives for which $\mathbb{E} \left[ \phi_i(X_1, X_2, \ldots, X_c) \right] \neq (c-1)/2$ for at least one $i$.

2. The asymptotic power of $W$ under a sequence of translation-type alternatives: As the $W$-test is consistent for a fixed translation-type alternative $F_i(x) = F(x - \theta_i)$, with not all $\theta$'s equal, the power $\rightarrow 1$ as $\min (n_1, \ldots, n_c) \rightarrow \infty$. The asymptotic power is then defined as the limiting power under a sequence of alternatives $H_n$ tending to $H_0$, as $n \rightarrow \infty$, provided that this limit is different from both 1 and the level of significance $\alpha$. It can be seen that, for our purpose, the asymptotic power can be computed if we take the sequence of alternatives

$$H_n : F_{in}(x) = F(x - n^{-\frac{1}{2}} \theta_i)$$

with not all $\theta$'s equal and $n_i = n s_i$, with $s_i$ a positive integer. The asymptotic power can then be computed in a manner similar to the one employed
in [2].

**Theorem 5.1.** If \( F \) possesses a continuous derivative \( f \) and there exists a function \( g \) such that

\[
||f(y+h) - f(y)||_1 \leq g(y) \quad \text{for all } y \text{ and } h,
\]

and

\[
\int_{-\infty}^{\infty} g(y) f(y) dy < \infty,
\]

then with \( n_i = n s_i \), with \( s_i \) a fixed positive integer, and under the sequence/distributions \( F_{in}, i=1,2,\ldots,c \), as \( n \to \infty \) the statistic \( W \) has a limiting noncentral chi-square distribution with \( c-1 \) degrees of freedom and the noncentrality parameter

\[
\lambda_N = 12\lambda^2 \sum_i s_i (\theta_i - \bar{\theta})^2,
\]

where \( \bar{\theta} = \sum_i s_i \theta_i / \sum_i s_i \) and

\[
\lambda = \int_{-\infty}^{\infty} f^2(y) dy.
\]

**Proof:** Let \( \eta_{in} = \mathcal{E} \left[ \phi_i(X_1,X_2,\ldots,X_c) / H_n \right]. \)

Then it can be easily shown that

\[
\eta_{in} = \sum_j \mathcal{E} \left[ \phi_{ij}(X_1,X_j) \mid H_n \right]
\]

\[
= (c-1)/2 + n^{-1/2} \lambda \sum_j (\theta_i - \theta_j) + O(n^{-1}).
\]

Similarly it can be shown that

\[
N \text{ cov } (U_{in} \mid H_n) \to \Sigma_X \quad \text{as } n \to \infty,
\]

and

\[
\sqrt{n} \left[ U_{in} - (c-1/2) \bar{\theta} \right] \text{ is asymptotically normal with mean}
\]

\[
(\sum_i s_i)^{1/2} \lambda \Sigma_X \quad \text{and covariance matrix } \Sigma_X, \text{ where } \Sigma_X = (\xi_1, \xi_2, \ldots, \xi_c),
\]

\[
\xi_i = c \theta_i - \frac{1}{j} \theta_j
\]

8
and $\xi$ is given by (3.8). The theorem then follows as in [2].

The asymptotic power is thus seen to be equal to the probability that a noncentral $\chi^2$ variable with $c-1$ degrees of freedom and the noncentrality parameter $\lambda_W$ exceeds $\chi^2_{c-1, \alpha}$, the usual upper $\alpha$-point of the central $\chi^2$ variable with $c-1$ degrees of freedom.

6. Remarks. Andrews [1] has obtained the asymptotic distribution of the $H$-statistic. Under the same sequence of alternatives it is the same as the asymptotic distribution of $W$, so that the asymptotic efficiency of $W$ relative to $H$ is one and, hence, relative to the $F$-statistic is $3/\pi$ if the underlying distributions are normal.

Comparing the efficiency figures in [2] it appears that the $V$-statistic, i.e., the test based on the number of c-plets such that the observations from the $i$th population are the least, is much more efficient for populations bounded below (e.g. exponential distribution $f(y, \alpha) = e^{-(y-\alpha)}$, $y \geq \alpha$); the statistic based on the number of c-plets with respect to the largest observation is similarly much more efficient for populations bounded above (e.g. reversed exponential distribution $f(y, \alpha) = e^{(y-\alpha)}$, $y \leq \alpha$). Both of them are fairly efficient (and the statistic based on c-plets with respect to both the smallest and the largest observations is even much more so) for distributions bounded on both sides (e.g. uniform distribution $f(x, \alpha, \beta) = 1/(\beta-\alpha)$, $\alpha \leq x \leq \beta$) while the $W$-statistic (based on c-plets with respect to all the positions) appears to be more efficient for unbounded distributions.
REFERENCES


