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 $F_1 = F_2 = \dots = F_c$ against alternatives of the form $F_i(x) = F(x - \theta_i)$ $(i = i,2,\dots,c)$, where the θ_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 $\{x_{ij}, j = 1,2,.,.,n_i\}$ be a random sample from the i th population with continuous c.d.f. F_i, $i = 1,2,\dots, c$, and suppose that these samples are independent. We consider a nonparametric test of the hypothesis

$$H_o: \; F_1 = F_2 = \dots = F_c$$

against alternatives of the form $F_i(x) = F(x-\theta_i)$ with the θ_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

(2.1) $$\phi_i\,(x_1,x_2,\ldots,x_c) = r - 1\ ,$$

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

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

It is then seen that

$$v_i = \sum_{j=1}^{n_i}\ \sum_{r=1}^{c}\ (r-1)\ n_{ij}^{(r)}\quad ,$$

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 ith sample, such that x_{ij} has rank r in each of these c-plets. Let $u_i = v_i/n_1n_2\ldots n_c$ and $N = \Sigma_i n_i$. Then the statistic now being proposed is

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

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

It can be seen that $\Sigma_i V_i = (n_1n_2\ldots n_c)c(c-1)/2$, 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 $|U-n_1n_2/2|$, where U is the number of pairs $(X_{1\alpha}, X_{2\beta})$ with, say, $X_{1\alpha} > X_{2\beta}$. It is also known that the Mann-Whitney U-test is equivalent to the Wilcoxon [10] test based on $\bar{R}_1$, 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 $\bar{R}_i$'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 ith 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 ith 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

$$(2.4) \qquad \phi_i(x_1,x_2,\ldots,x_c) = \sum_{j=1}^{c} \phi_{ij}(x_i,x_j) \quad ,$$

where

$$(2.5) \qquad \phi_{ij}(x_i,x_j) = 1 \quad \text{if } x_i > x_j$$
$$= \text{otherwise}$$

Then, from (2.2),

$$v_i = n_1n_2\ldots n_c \sum_{j=1}^{c} u_{ij} \quad ,$$

so that

(2.6) $$u_i = \sum_{j=1}^{c} u_{ij} \quad ,$$

where

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

and $u_{ii} = 0$. In the special case $n_1=n_2=\ldots=n_c=n$, say, we have $n^2 u_i = n[\bar{R}_i - (n+1)/2]$, where $\bar{R}_i$ is the mean rank of the ith 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.

3. The asymptotic distribution of W under H_o.

From (2.2) it is seen that U_i is a generalized U-statistic corresponding to ϕ_i . From the c-sample version (e.g. see [2]) of Hoeffding's theorem [5] on U-statistics, it then follows that $N^{\frac{1}{2}}[\underset{\sim}{U}_N - \underset{\sim}{\eta}]$ 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 $\Sigma_i p_i = 1$, normally distributed with zero mean and covariance matrix $\underset{\sim}{\Sigma} = (\sigma_{rs})$ given by

(3.1) $$\sigma_{rs} = \sum_{i=1}^{c} \frac{1}{p_i} \zeta^{(i)}(r,s) \quad , \qquad r,s = 1,2,\ldots,c,$$

where

$$\underset{\sim}{U}'_N = (U_{1N},\ldots, U_{cN}) \quad , \underset{\sim}{\eta}' = (\eta_1,\eta_2,\ldots,\eta_c) \ ,$$

$$\eta_i = \mathcal{E}[\phi_i(X_1,X_2,\ldots,X_c)] \ ,$$

(3.2) $$\zeta^{(i)}(r,s) = \mathcal{E}\left[\phi_r(X_1,X_2,\ldots,X_c)\phi_s(X'_1, \ldots ,X'_{i-1},X_i,X'_{i+1},\ldots,X'_c)\right] - \eta_r\eta_s \ ,$$

where X_j, X'_j are independent random variables with c.d.f. F_j $(j=1,2,\ldots,c)$.

Now, when H_o holds, $F_1 = F_2 = \cdots = F_c = F$, say. Then $\eta_i = \sum_{j=1}^{c} \mathcal{E} [\phi_{ij} (X_i, X_j)] = (c-1)/2$. Here, and hereafter in this section, X's are independent random variables each with c.d.f. F. Also

$$\zeta^{(i)}(i,i) = \mathcal{E} \{\sum_j \phi_{ij}(X_i, X_j)\} \{ \sum_k \phi_{ik}(X_i, X_k')\} - (c-1)^2/4$$

$$= \sum_j \sum_k \mathcal{E} \{\phi_{ij}(X_i, X_j) \phi_{ik}(X_i, X_k')\} - (c-1)^2/4$$

$$= \sum_j \sum_k (1/3) - (c-1)^2/4 = (c-1)^2/12 \quad , \tag{3.3}$$

$$\zeta^{(j)} (i,i) = \mathcal{E} \{\sum_r \phi_{ir}(X_i, X_r)\} \{ \sum_{s \neq j} \phi_{is}(X_i', X_s') + \phi_{ij}(X_i', X_j)\} - (c-1)^2/4 \tag{3.4}$$

$$= \sum_r \sum_{s \neq j} (1/4) + \sum_{r \neq j} (1/4) + 1/3 - (c-1)^2/4$$

$$= 1/12 \, , \qquad (i \neq j)$$

$$\zeta^{(j)}(i,j) = \mathcal{E} \{\sum_r \phi_{ir}(X_i, X_r)\} \{ \sum_s \phi_{js}(X_j, X_s')\} - (c-1)^2/4 \tag{3.5}$$

$$= \sum_{r \neq j} \sum_s (1/4) + \sum_s (1/6) - (c-1)^2/4 = -(c-1)/12 \; ,$$

and finally

$$\zeta^{(k)}(i,j) = \mathcal{E} \{ \sum_r \phi_{ir} (X_i, X_r)\} \{ \sum_{s \neq k} \phi_{js}(X_j', X_s') + \phi_{jk}(X_j', X_k)\} - (c-1)^2/4 \tag{3.6}$$

$$= \sum_r \sum_{s \neq k} (1/4) + \sum_{r \neq k} (1/4) + 1/3 - (c-1)^2/4$$

$$= 1/12 \; .$$

Thus, when H_o holds, we have

$$\sigma_{ii} = \frac{(c-1)^2}{12p_i} + \sum_{j \neq i} \frac{1}{12p_j} \tag{3.7}$$

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

Thus

$$(3.8) \quad 12\, \Sigma = c^2 P^{-1} - c\, q\, j' - c\, j\, q' + a\, J ,$$

where P = diagonal $(p_i,\ i=1,2,\dots,c)$, $a = \sum_{i=1}^{c} (1/p_i)$,

$J = (1)_{c\times c}$, $j' = (1)_{1\times c}$ and $q' = (1/p_1,\dots,1/p_c)$.

Since $\Sigma_i U_{iN} = c(c-1)/2$, the distribution of U_N is singular and, hence, the asymptotic normal distribution of $\sqrt{N}\,(U_N - \eta)$ 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_o. Thus suppressing N in the subscript of U, we have the statistic (2.3) proposed earlier.

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

Let $\eta_i = f^{(i)}(F_1,F_2,\dots,F_c)$, $i=1,2,\dots,g$, be real-valued functions such that $f^{(i)}(F,\dots,F) = \eta_{io}$ for all $(F,F,\dots,F)$ in a class $\mathcal{C}_o$. Let $T^{(i)}_{n_1,\dots,n_c} = t^{(i)}(X_{11},\dots,X_{1n_1};\dots;X_{c1},\dots,X_{cn_c})$, $i=1,2,\dots,g$, be sequences of real-valued statistics such that $T^{(i)}_{n_k,\dots,n_c}$ tends to η_i in probability as $\min(n_1,\dots,n_c) \to \infty$. Suppose that $f^{(i)}(F_1,F_2,\dots,F_c) \neq \eta_{io}$ for some i for all $(F_1,\dots,F_c)$ in a class $\mathcal{C}_1$. Further let

$$W_{n_1,\dots,n_c} = w\,(T^{(1)}_{n_1,\dots,n_c},\dots,T^{(g)}_{n_1,\dots,n_c})$$

be a nonnegative function which is zero if, and only if, $T^{(i)}_{n_1,\ldots,n_c} = \eta_{io}$ 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: $\mathcal{C}_o$ at every fixed level of significance against the alternatives $\mathcal{C}_1$.

If we take $\eta_i = \mathcal{E}\left[\phi_i(X_1,X_2,\ldots,X_c) \right]$, where the X_i's are independent random variables with continuous c.d.f. $F_1,\ldots,F_c$, respectively, and $T^{(i)}_{n_1,n_2,\ldots,n_c} = U_{iN}$, then the convergence in probability of U_{iN} to η_i follows from the asymptotic normality of $\sqrt{N}(\underset{\sim}{U}_N - \underset{\sim}{\eta})$. For the class $\mathcal{C}_1$ of translation-type alternatives $F_i(x) = F(x - \theta_i)$, with θ's not all equal, it may be easily seen that $\eta_r > (c-1)/2$, i.e. η_{ro}, where θ_r is the (or one of the) largest among θ'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 $\mathcal{E}\,[\phi_i(X_1,X_2,\ldots,X_c)] \neq (c-1)/2$ for at least one i.

5. 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 θ's equal, the power $\to 1$ as $\min(n_1,\ldots,n_c) \to \infty$. The asymptotic power is then defined as the limiting power under a sequence of alternatives H_n tending to H_o, as $n \to \infty$, provided that this limit is different from both 1 and the level of significance α. It can be seen that, for our purpose, the asymptotic power can be computed if we take the sequence of alternatives

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

with not all θ's equal and $n_i = ns_i$, 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

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

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 of 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

$$(5.1) \qquad \lambda_W = 12\lambda^2 \sum_i s_i\, (\theta_i - \bar{\theta})^2 ,$$

where $\bar{\theta} = \Sigma_i s_i\, \theta_i/\Sigma_i s_i$ and

$$(5.2) \qquad \lambda = \int_{-\infty}^{\infty} f^2(y)\, dy .$$

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

$$\eta_{in} = \sum_j \mathcal{E}\left[\phi_{ij}(X_i,X_j) \mid H_n\right]$$
$$= (c-1)/2 + n^{-\frac{1}{2}} \lambda \sum_j (\theta_i - \theta_j) + O(n^{-1}) .$$

Similarly it can be shown that

$$N \operatorname{cov} (\underset{\sim}{U}_N \mid H_n) \to \underset{\sim}{\Sigma} , \text{ as } n \to \infty ,$$

and

$\sqrt{N}\, [\, \underset{\sim}{U}_N - (c-1/2)\, \underset{\sim}{j}\,]$ is asymptotically normal with mean $(\Sigma_i s_i)^{\frac{1}{2}} \lambda\, \underset{\sim}{\delta}$ and covariance matrix $\underset{\sim}{\Sigma}$, where $\underset{\sim}{\delta}' = (\delta_1,\delta_2,\ldots,\delta_c)$,

$$\delta_i = c\, \theta_i - \sum_j \theta_j$$

and $\underset{\sim}{\Sigma}$ 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 χ^2 variable with c-1 degrees of freedom and the noncentrality parameter λ_W exceeds $\chi^2_{c-1,\alpha}$, the usual upper α-point of the central χ^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 ith 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.

R E F E R E N C E S

[1] FRED C. ANDREWS, "Asymptotic behavior of some rank tests for analysis of variance," Ann. Math. Stat., Vo. 25(1954), pp. 724-736.

[2] V. P. BHAPKAR, "A nonparametric test for the problem of several samples," Ann. Math. Stat., Vol. 32 (1961), pp. 1108-1117.

[3] JAYANT V. DESHPANDE, "A nonparametric test based on U-statistics for several samples," (Abstract), Ann. Math. Stat., Vol. 34 (1963) p. 1624.

[4] MEYER DWASS, " Some k-sample rank-order tests," Contributions to Probability and Statistics, Stanford University Press, Stanford, California (1960), pp. 198-202.

[5 WASSILY HOEFFDING, "A class of statistics with asymptotic normal distributions," Ann. Math. Stat. Vol. 19 (1948), pp. 293-325.

[6] J. KIEFER, "K-sample analogues of the Kolmogoror-Smirnov and Cramer-v. Mises tests," Ann. Math. Stat., Vol. 30(1959), pp. 420-447.

[7] WILLIAM H. KRUSKAL, "A nonparametric test for the several sample problem," Ann. Math. Stat. Vol. 23(1952), pp. 525-540.

[8] E. L. LEHMANN, "Consistency and unbiasedness of certain nonparametric tests," Ann. Math. Stat., Vol. 22(1951), pp. 165-179.

[9] H. B. MANN and D. R. WHITNEY, "On a test of whether one of two random variables is stochastically larger than the other," Ann. Math. Stat., Vol. 18(1947), pp. 50-60.

[10] F. WILCOXON, "Individual comparison by ranking methods," Biometrics Bull., Vol. (1945), pp. 80-83.