SOME NONPARAMETRIC TESTS FOR THE MULTIVARIATE SEVERAL SAMPLE LOCATION PROBLEM

by

V. P. Bhapkar
University of North Carolina
and
University of Poona

Institute of Statistics Mimeo Series No. 415

December 1964

(Dedicated to the memory of
Professor S. N. Roy)

This research was supported by the Mathematics Division of the Air Force Office of Scientific Research Contract No. AF-AFOSR-760-65.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
Chapel Hill, N. C.

1. Summary. This paper offers nonparametric tests of the null hypothesis $F_1=F_2=\ldots=F_c$ against alternatives of the form $F_i(\underset{\sim}{x}) = F(\underset{\sim}{x} - \underset{\sim}{\theta}_i)$ $(i = 1,2,\ldots,c)$, where the $\underset{\sim}{\theta}_i$'s are not all equal and F_i is the unknown continuous cumulative distribution function of the p-variate population from which the $\underline{i}$ th random sample comes. The statistics offered are multivariate analogues of some univariate rank-order test-statistics and all these are shown to be asymptotically distributed as χ^2 with $p(c-1)$ degrees of freedom when the null hypothesis holds.

2. Introduction and Notation.

Quite a few nonparametric tests are available for the several-sample, (say c-sample), location problem in the univariate case. Among them are the H-test of Kruskal-Wallis [12], the M-test of Mood [13], and the tests based on c-plets, viz. the V and W tests offered by the author ([1], [2]) and the L-test offered by Deshpande [5]. Some other tests and references to earlier ones may be found in Dwasss [6], Kiefer [10] and Kruskal-Wallis [12].

Not much work though has been done in the corresponding multivariate case. Hodges [8], Vineze [14], Chatterjee and Sen [4] have developed non-parametric tests for the two-sample problem in the bivariate case. The author [3] had offered a test for the several-sample bivariate problem using a "step-down procedure" in which the regression of one variable on the other is assumed to be linear. But there the roles of the two variables do not appear to be symmetrical. The present paper offers symmetric nonparametric tests for the general several-sample multivariate location problem. These will be seen to be multivariate analogues of the univariate V, L, W and H tests.

Let $\{\underset{\sim}{X}_{ij},\ j = 1,2,\dots, n_i\}$ be n_i independent observations from a population with continuous nonsingular c.d.f. F_i, $i = 1,2,\dots, c$. Let the p components of observation $\underset{\sim}{X}_{ij}$ be denoted by $X_{ij}^{(\alpha)}$, $\alpha = 1,2,\dots, p$. $\underset{\sim}{X}'_{ij}$ then denotes the row-vector $(X_{ij}^{(1)},\dots, X_{ij}^{(p)})$ while $\underset{\sim}{X}_{ij}$ denotes the corresponding column-vector. The samples are assumed to be independent. We consider nonparametric tests of the hypothesis

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

against alternatives of the form $F_i(\underset{\sim}{x}) = F(\underset{\sim}{x}-\underset{\sim}{\theta}_i)$ with the vectors $\underset{\sim}{\theta}_i$'s not all equal.

Let $v_i^{(\alpha)}$ be the number of c-plets $(\underset{\sim}{X}_{1t_1},\underset{\sim}{X}_{2t_2},\dots,\underset{\sim}{X}_{ct_c})$ that can be formed by choosing one observation from each sample such that $X_{it_i}^{(\alpha)}$

is the smallest among $\{X^{(\alpha)}_{kt_k}, k = 1, 2, \ldots, c\}$. Similarly let $b^{(\alpha)}_i$ be the number of c-plets that can be formed such that $X^{(\alpha)}_{it_i}$ is the largest among $\{X^{(\alpha)}_{kt_k}, k = 1, 2, \ldots, c\}$; denote $b^{(\alpha)}_i - v^{(\alpha)}_i$ by $\ell^{(\alpha)}_i$. In general, suppose that $n^{(\alpha)}_{ir}$ is the number of c-plets $(\underset{\sim}{X}_{1t_1}, \ldots, \underset{\sim}{X}_{ct_c})$ such that $X^{(\alpha)}_{it_i}$ has rank r among $\{X^{(\alpha)}_{kt_k}, k = 1,2,\ldots,c\}$, $r = 1,2,\ldots,c$.

Then we note that $v^{(\alpha)}_i = n^{(\alpha)}_{i1}$ and $b^{(\alpha)}_i = n^{(\alpha)}_{ic}$. Further, let

$$w^{(\alpha)}_i = \sum_{r=1}^{c} (r-1)\, n^{(\alpha)}_{ir} \ .$$

Finally, if $R^{(\alpha)}_{ij}$ denotes the rank of $X^{(\alpha)}_{ij}$ among $\{X^{(\alpha)}_{kt}, t=1,\ldots,n_k$ and $k=1,\ldots,c\}$, let

$$\bar{R}^{(\alpha)}_i = \sum_{j=1}^{n_i} R^{(\alpha)}_{ij}/n_i \ .$$

Then the test-statistics now being proposed are:

$$\text{(2.1)} \qquad V = N(2c-1) \sum_{i=1}^{c} p_i(\underset{\sim}{u}_i-\bar{\underset{\sim}{u}})' \, \underset{\sim}{\Sigma}_V^{-1}(\underset{\sim}{u}_i - \bar{\underset{\sim}{u}}) \ ,$$

where $N = \Sigma_i n_i$, $p_i = n_i/N$, $u^{(\alpha)}_i = v^{(\alpha)}_i/n_1n_2\ldots n_c$, $\underset{\sim}{u}'_i = (u^{(1)}_i, \ldots, u^{(p)}_i)$, $\bar{\underset{\sim}{u}} = \Sigma_i p_i \underset{\sim}{u}_i$ and $\underset{\sim}{\Sigma}_V$ is a matrix given by (4.1).

$$\text{(2.2)} \qquad B = N(2c-1) \sum_{i=1}^{c} p_i(\underset{\sim}{u}_i-\bar{\underset{\sim}{u}})' \, \underset{\sim}{\Sigma}_B^{-1} (\underset{\sim}{u}_i-\bar{\underset{\sim}{u}}) \ ,$$

where $u^{(\alpha)}_i = b^{(\alpha)}_i/n_1n_2\ldots n_c$ and $\underset{\sim}{\Sigma}_B$ is given by (4.2).

$$\text{(2.3)} \qquad L = \frac{N(2c-1)(c-1)^2\binom{2c-2}{c-1}}{2c^2\{\binom{2c-2}{c-1} - 1\}} \sum_{i=1}^{c} p_i(\underset{\sim}{u}_i-\bar{\underset{\sim}{u}}) \, \underset{\sim}{\Sigma}_L^{-1} (\underset{\sim}{u}_i-\bar{\underset{\sim}{u}}),$$

where $u^{(\alpha)}_i = \ell^{(\alpha)}_i/n_1n_2\ldots n_c$ and $\underset{\sim}{\Sigma}_L$ is given by (4.3).

$$\text{(2.4)} \qquad W = \frac{12N}{c^2} \sum_{i=1}^{c} p_i \, (\underset{\sim}{u}_i-\bar{\underset{\sim}{u}})' \, \underset{\sim}{\Sigma}_W^{-1} (\underset{\sim}{u}_i-\bar{\underset{\sim}{u}}) \ ,$$

where $u_i^{(\alpha)} = w_i^{(\alpha)}/n_1 n_2 \ldots n_c$ and $\underset{\sim}{E}_W$ is given by (4.4).

$$(2.5) \qquad H = \frac{12}{N} \sum_{i=1}^{c} p_i \left(\bar{\underset{\sim}{R}}_i - \frac{N+1}{2} \underset{\sim}{j}\right)' \underset{\sim}{E}_W^{-1} \left(\bar{\underset{\sim}{R}}_i - \frac{N+1}{2} \underset{\sim}{j}\right),$$

where $\underset{\sim}{j}' = (1)_{1 \times c}$.

The V-test consists in rejecting H_o at a significance level α if V exceeds some predetermined number V_α; the same thing holds for the other statistics. In the next section it is shown that, when H_o is true, each of these statistics is asymptotically distributed as a χ^2 variable with p(c-1) degrees of freedom. Thus large sample approximations for V_α, B_α, L_α, W_α and H_α are provided by the upper α-point of the χ^2 distribution with p(c-1) degrees of freedom. These tests are thus seen to be the multivariate analogues of the corresponding univariate tests.

If $\bar{\underset{\sim}{X}}_i = \Sigma_j \underset{\sim}{X}_{ij}/n_i$, $\bar{\underset{\sim}{X}} = \Sigma_{i,j} \underset{\sim}{X}_{ij}/N$, and the populations are p-variate normal with covariance matrix $\underset{\sim}{\Sigma}$, then it is well-known that the statistic

$$(2.6) \qquad \sum_i n_i \, (\bar{\underset{\sim}{X}}_i - \bar{\underset{\sim}{X}})' \, \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{X}}_i - \bar{\underset{\sim}{X}})$$

is distributed as χ^2 with p(c-1) degrees of freedom under H_o and, hence,

$$(2.7) \qquad \sum_i n_i \, (\bar{\underset{\sim}{X}}_i - \bar{\underset{\sim}{X}})' \, \underset{\sim}{S}^{-1} \, (\bar{\underset{\sim}{X}}_i - \bar{\underset{\sim}{X}})$$

is asymptotically distributed as χ^2 with p(c-1) degrees of freedom under H_o, where S is the covariance matrix obtained from the samples. The statistics being offered now are thus seen to be rank-order analogues of (2.7).

3. <u>The asymptotic distribution of a test-criterion under H_o</u>. Let

$$(3.1) \qquad U_i^{(\alpha)} = \frac{1}{n_1 n_2 \ldots n_c} \sum_{t_1=1}^{n_1} \cdots \sum_{t_c=1}^{n_c} \Phi_i^{(\alpha)}(\underset{\sim}{X}_{1t_1}, \underset{\sim}{X}_{2t_2}, \ldots, \underset{\sim}{X}_{ct_c})$$

$$i = 1,2,\ldots,c$$
$$\alpha = 1,2,\ldots,p$$

and suppose that

$$(3.2) \qquad \Phi_i^{(\alpha)}(\underset{\sim}{X}_1,\underset{\sim}{X}_2,\ldots,\underset{\sim}{X}_c) = \Phi_i(X_1^{(\alpha)},X_2^{(\alpha)},\ldots,X_c^{(\alpha)}),$$

and further

$$(3.3) \qquad \Phi_i(x_1,x_2,\ldots,x_c)=\Phi(x_i;\ \{x_j,\ j \neq i\}),$$

which means that Φ is a function of x_i and the set of remaining x_j's and, moreover, is symmetric in x_j's, $j = 1,2,\ldots,c$ except i. We shall also assume that Φ is bounded. Then, the following special cases (3.4), (3.5), (3.6) and (3.7) will lead to the u's in the statistics (2.1), (2.2), (2.3) and (2.4), respectively:

$$(3.4) \qquad \Phi_i(x_1,x_2,\ldots,x_c) = \begin{cases} 1 \text{ if } x_i < x_j,\ j=1,\ldots,c \text{ except } i \\ 0 \text{ otherwise} \end{cases}$$

$$(3.5) \qquad \Phi_i(x_1,\ldots,x_c) = \begin{cases} 1 \text{ if } x_i > x_j,\ j=1,\ldots,c \text{ except } i \\ 0 \text{ otherwise} \end{cases}$$

$$(3.6) \qquad \Phi_i(x_1,\ldots,x_c) = \begin{cases} 1 \text{ if } x_i > x_j,\ j=1,\ldots,c \text{ except } i \\ -1 \text{ if } x_i < x_j,\ j=1,\ldots,c \text{ except } i \\ 0 \text{ otherwise} \end{cases}$$

$$(3.7) \qquad \Phi_i(x_1,\ldots,x_c) = \begin{cases} r-1, \text{ if the rank of } x_i \text{ among } \{x_1,x_2,\ldots,x_c\} \\ \text{is } r, \text{ so that} \end{cases}$$

$$\Phi_i(x_1,\ldots,x_c) = \sum_{j=1}^{c} h(x_i-x_j) \quad ,$$

where the function h is defined by

$$h(y) = 1 \quad \text{if } y > 0$$
$$\quad = 0 \quad \text{otherwise} .$$

From (3.1) it is seen that $U_i^{(\alpha)}$ is a generalized U-statistic corresponding to $\Phi_i^{(\alpha)}$. For studying the asymptotic distributions we shall write $U_{iN}^{(\alpha)}$ for $U_i^{(\alpha)}$ based on the samples of total size N; also the random variables will be denoted by capital letters and the variables held fixed will be denoted by the corresponding small letters. Let

$$(3.8) \qquad \underset{\sim}{U}_N' = (\underset{\sim}{U}_{1N}', \underset{\sim}{U}_{2N}', \ldots, \underset{\sim}{U}_{cN}'), \; \underset{\sim}{U}_{iN}' = (U_{iN}^{(1)}, \ldots, U_{iN}^{(p)}), \; \underset{\sim}{\eta}' = (\underset{\sim}{\eta}_1', \underset{\sim}{\eta}_2', \ldots, \underset{\sim}{\eta}_c'),$$

where

$$\underset{\sim}{\eta}_i' = (\eta_i^{(1)}, \ldots, \eta_i^{(p)}) \; , \quad \eta_i^{(\alpha)} = \mathcal{E}[\Phi_i^{(\alpha)}(\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_c)]$$

and $\underset{\sim}{X}_i$'s are independent random variables with c.d.f. F_i, $i = 1,2,\ldots,c$, respectively. From the c-sample version (e.g. see [1]) of Hoeffding's theorem [9] concerning U-statistics it then follows that in the limit as $n_i \longrightarrow \infty$ in such a way that $n_i = np_i$, the p's being fixed positive numbers such that $\Sigma_i p_i = 1$, $N^{1/2}[\underset{\sim}{U}_N - \underset{\sim}{\eta}]$ is normally distributed with zero mean vector and covariance matrix $\underset{\sim}{\Sigma}$, where

$$(3.9) \qquad \underset{\sim}{\Sigma} = \begin{bmatrix} \underset{\sim}{\Sigma}_{11} & \underset{\sim}{\Sigma}_{12} & \cdots & \underset{\sim}{\Sigma}_{1c} \\ \underset{\sim}{\Sigma}_{21} & \underset{\sim}{\Sigma}_{22} & \cdots & \underset{\sim}{\Sigma}_{2c} \\ \cdot & \cdot & \cdot & \cdot \\ \underset{\sim}{\Sigma}_{c1} & \underset{\sim}{\Sigma}_{c2} & \cdots & \underset{\sim}{\Sigma}_{cc} \end{bmatrix}$$

$$\underset{\sim}{\Sigma}_{ij} = \left[\sigma_{ij}^{(\alpha,\beta)} \right] \; , \quad \alpha,\beta = 1,2,\ldots,p \; ,$$

$$(3.10) \qquad \sigma_{ij}^{(\alpha,\beta)} = \sum_{k=1}^{c} \frac{1}{p_k} \zeta_{(k)}^{(\alpha,\beta)}(i,j) \quad ,$$

and

$$(3.11) \qquad \zeta_{(k)}^{(\alpha,\beta)}(i,j) = \mathcal{E}[\Phi_i^{(\alpha)}(\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_c) \; \Phi_j^{(\beta)}(\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_c)] - \eta_i^{(\alpha)} \eta_j^{(\beta)},$$

where $\underset{\sim}{X}_i, \underset{\sim}{Y}_i$ are independent random variables with c.d.f. F_i $(i=1,2,\ldots,c)$, except that $\underset{\sim}{Y}_k = \underset{\sim}{X}_k$.

Now, when H_o holds, $F_1 = F_2 = \ldots = F_c = F$, say. We shall denote by $F^{(\alpha,\beta)}$ the marginal c.d.f. of the $\underline{\alpha}^{th}$ and $\underline{\beta}^{th}$ components of $\underset{\sim}{X}$. Here, and hereafter in this section, $\underset{\sim}{X}$'s, $\underset{\sim}{Y}$'s are independent random variables each with c.d.f. F.

Theorem 3.1. Suppose the functions $\Phi_i^{(\alpha)}$, $i=1,2,\ldots,c$ and $\alpha=1,2,\ldots,p$ satisfy the following conditions:

(i) $\Phi_i^{(\alpha)}(\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots,\underset{\sim}{x}_c) = \Phi_i(x_1^{(\alpha)}, x_2^{(\alpha)},\ldots,x_c^{(\alpha)})$

(ii) $\Phi_i(x_1,x_2,\ldots,x_c) = \Phi(x_i; \{x_j, j\neq i\})$

(iii) There exists a constant A such that $|\Phi| < A$

(iv) $\Sigma_i \Phi_i(x_1,x_2,\ldots,x_c) = cd$, where d is some constant; it is the only linear constraint on Φ_i's.

(v) The statistic $\Phi(X_1; \{X_j, j=2,3,\ldots,c\})$ is distribution-free in the class of continuous univariate distributions, i.e., the probability distribution of $\Phi(X_1; \{X_j, j=2,\ldots,c\})$, where X's are independent random variables each with univariate continuous c.d.f. G, is independent of G.

(vi) The common distribution of the independent random variables $\underset{\sim}{X}$'s is nonsingular.

Then as $n_i \longrightarrow \infty$, in such a way that $n_i = np_i$, $i=1,\ldots,c$, the p's being fixed positive numbers such that $\Sigma_i p_i = 1$, and under H_o, the asymptotic distribution of the random variable.

$$(3.12) \qquad T_N = \frac{N(c-1)^2}{\mu c^2} \sum_{i=1}^{c} p_i \, (\underset{\sim}{U}_{iN} - \overline{\underset{\sim}{U}}_N)' \underset{\sim}{\Gamma}^{-1} (\underset{\sim}{U}_{iN} - \overline{\underset{\sim}{U}}_N),$$

where $\underset{\sim}{U}_{iN}$ is given by (3.1) and (3.8), $\overline{\underset{\sim}{U}}_N = \Sigma_i p_i \underset{\sim}{U}_{iN}$,

$$\mu = \mathcal{E}[\Phi_1^{(\alpha)}(X_1, X_2, \ldots, X_c)\Phi_1^{(\alpha)}(X_1, Y_2, \ldots, Y_c)] - d^2,$$

and $\rho = (\rho_{\alpha\beta})$, $\alpha,\beta, = 1,2,\ldots,p$ with

$$\rho_{\alpha\beta} = \text{Corr. Coeff.}[\Phi_1^{(\alpha)}(X_1, X_2, \ldots, X_c), \Phi_1^{(\beta)}(X_1, Y_2, \ldots, Y_c)] \tag{3.13}$$

is χ^2 with degrees of freedom $p(c-1)$.

Proof. The theorem will follow by the application of the multisample extension of Hoeffding's theorem and noting that under H_o, $F_1 = F_2 = \ldots = F_c = F$, say.

We first see that, in view of conditions (i), (ii) and (v), $\mathcal{E}[\Phi_i^{(\alpha)}(X_1, \ldots, X_c)]$ is a constant, d, independent of any univariate continuous c.d.f. G. Now

$$\begin{aligned} \psi_{i,k}^{(\alpha)}(x, y) &= \mathcal{E}\Phi_i^{(\alpha)}(X_1, X_2, \ldots, X_c | X_i = x,\ X_k = y) - d \\ &= \mathcal{E}\Phi(x^{(\alpha)}; \{y^{(\alpha)}, X_j^{(\alpha)},\ j = 1, \ldots, c \text{ except } i \text{ and } k\}) - d \\ &= \psi^{(\alpha)}(x^{(\alpha)}, y^{(\alpha)}) \text{ say, for all } i \neq k \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} \psi_i^{(\alpha)}(x) &= \mathcal{E}\Phi_i^{(\alpha)}(X_1, \ldots, X_c | X_i = x) - d \\ &= \mathcal{E}\psi^{(\alpha)}(x^{(\alpha)}, Y^{(\alpha)}) = \psi^{(\alpha)}(x^{(\alpha)}), \text{ say for all } i. \end{aligned} \tag{3.15}$$

Then from (3.11), under H_o,

$$\begin{aligned} \zeta_{(i)}^{(\alpha,\alpha)}(i,i) &= \mathcal{E}[\psi_i^{(\alpha)}(X)]^2 = \mathcal{E}[\psi^{(\alpha)}(X^{(\alpha)})]^2 = \mu, \text{ say}, \\ \zeta_{(i)}^{(\alpha,\beta)}(i,i) &= \mathcal{E}[\psi^{(\alpha)}(X^{(\alpha)})\psi^{(\beta)}(X^{(\beta)})] = \int \psi^{(\alpha)}(x)\psi^{(\beta)}(y)\,dF^{(\alpha,\beta)}(x,y) \\ &= \mu\,\rho_{\alpha\beta}, \end{aligned} \tag{3.16}$$

where $\rho_{\alpha,\beta}$ is the correlation coefficient of $\psi^{(\alpha)}(X)$ and $\psi^{(\beta)}(Y)$, where (X,Y) has c.d.f. $F^{(\alpha,\beta)}$ and, in general, will depend on F. We also note that $\rho_{\alpha\beta}$ can be expressed in the form (3.13). Also, if for $i \neq k$

$$\psi_{i,k}^{(\alpha)}(\underset{\sim}{y}) = \mathcal{E}[\psi_{i,k}^{(\alpha)}(\underset{\sim}{X},\underset{\sim}{y})] = \mathcal{E}[\psi^{(\alpha)}(X^{(\alpha)},y^{(\alpha)})] = w^{(\alpha)}(y)^{(\alpha)}, \text{ say}$$

then

$$\zeta_{(k)}^{(\alpha,\alpha)}(i,i) = \mathcal{E}[\psi_{i,k}^{(\alpha)}(\underset{\sim}{Y})]^2 = \mathcal{E}[w^{(\alpha)}(Y^{(\alpha)})]^2 = w, \text{ say}$$

(3.17) and

$$\zeta_{(k)}^{(\alpha,\beta)}(i,i) = \mathcal{E}[\psi_{i,k}^{(\alpha)}(\underset{\sim}{Y})\psi_{i,k}^{(\beta)}(\underset{\sim}{Y})] = \mathcal{E}[w^{(\alpha)}(Y^{(\alpha)})w^{(\beta)}(Y^{(\beta)})].$$

But from (i) and (iv) we have

$$\sum_{i=1}^{c} \Phi_i^{(\alpha)}(x_1,\ldots,x_c) = cd,$$

and hence

$$\mathcal{E}[\sum_{i=1}^{c} \Phi_i^{(\alpha)}(\underset{\sim}{x}_1,\underset{\sim}{x}_2,\underset{\sim}{X}_3,\ldots,\underset{\sim}{X}_c)] = cd,$$

which gives

$$(3.18)\qquad \psi^{(\alpha)}(x_1^{(\alpha)},x_2^{(\alpha)}) + \psi^{(\alpha)}(x_2^{(\alpha)},x_1^{(\alpha)}) + \sum_{i=3}^{c}\mathcal{E}\Phi(X_i^{(\alpha)};\{x_1^{(\alpha)},x_2^{(\alpha)},X_j^{(\alpha)},j=3,\ldots,c \text{ except } i\}) = (c-2)d.$$

If we let

$$\eta^{(\alpha)}(x_1^{(\alpha)},x_2^{(\alpha)}) = \mathcal{E}[\Phi(X_3^{(\alpha)};\{x_1^{(\alpha)},x_2^{(\alpha)},X_4^{(\alpha)},\ldots,X_c^{(\alpha)}\}] - d,$$

then $\eta^{(\alpha)}(x_1^{(\alpha)},x_2^{(\alpha)}) = \eta^{(\alpha)}(x_2^{(\alpha)}, x_1^{(\alpha)})$

and $$\mathcal{E}[\eta^{(\alpha)}(X_1^{(\alpha)},x_2^{(\alpha)})] = w^{(\alpha)}(x_2^{(\alpha)}) \quad ,$$

(3.18) then gives

(3.19) $$\psi^{(\alpha)}(x_1^{(\alpha)},x_2^{(\alpha)}) + \psi^{(\alpha)}(x_2^{(\alpha)},x_1^{(\alpha)}) + (c-2)\eta^{(\alpha)}(x_1^{(\alpha)},x_2^{(\alpha)}) = 0.$$

Replacing $x_1^{(\alpha)}$ by $X_1^{(\alpha)}$ and taking expectation we have

$$w^{(\alpha)}(x_2^{(\alpha)}) + \psi^{(\alpha)}(x_2^{(\alpha)}) + (c-2)w^{(\alpha)}(x_2^{(\alpha)}) = 0 ,$$

i.e.,

(3.20) $$\psi^{(\alpha)}(x^{(\alpha)}) = -(c-1)w^{(\alpha)}(x^{(\alpha)}) .$$

From (3.17), (3.20) and (3.16) it then follows that

(3.21) $$\mu = (c-1)^2 w \text{ and } \zeta_{(k)}^{(\alpha,\beta)}(i,i) = w\, \rho_{\alpha\beta} .$$

Similarly, for $i \neq j$,

(3.22) $$\zeta_{(j)}^{(\alpha,\alpha)}(i,j) = \mathcal{E}[\psi_{ij}^{(\alpha)}(\underset{\sim}{X})\psi_j^{(\alpha)}(\underset{\sim}{X})] = \mathcal{E}[w^{(\alpha)}(X^{(\alpha)})\psi^{(\alpha)}(X^{(\alpha)})]$$

$$= -(c-1)w \quad , \text{ in view of (3.20),}$$

and

$$\zeta_{(j)}^{(\alpha,\beta)}(i,j) = \mathcal{E}[\psi_{ij}^{(\alpha)}(\underset{\sim}{X})\psi_j^{(\beta)}(\underset{\sim}{X})] = \mathcal{E}[w^{(\alpha)}(X^{(\alpha)})\psi^{(\beta)}(X^{(\beta)})]$$

$$= -(c-1)w\rho_{\alpha\beta} .$$

Finally, for $i \neq j \neq k$

$$\zeta_{(k)}^{(\alpha,\alpha)}(i,j) = \mathcal{E}[\psi_{ik}^{(\alpha)}(\underset{\sim}{X})\psi_{jk}^{(\alpha)}(\underset{\sim}{X})] = \mathcal{E}[w^{(\alpha)}(X^{(\alpha)})]^2 = w,$$

(3.23) and

$$\zeta_{(k)}^{(\alpha,\beta)}(i,j) = \mathcal{E}[\psi_{ik}^{(\alpha)}(\underset{\sim}{X})\psi_{jk}^{(\beta)}(\underset{\sim}{X})] = \mathcal{E}[w^{(\alpha)}(X^{(\alpha)})w^{(\beta)}(X^{(\beta)})]$$

$$= w\rho_{\alpha\beta} .$$

Thus from (3.10), (3.16), (3.17), (3.21), (3.22) and (3.23) it follows that, under H_o,

$$\sigma_{ii} = \frac{\mu}{p_i} + \sum_{k\neq i} (\frac{1}{p_i})w = w[q + \frac{c(c-2)}{p_i}]$$

(3.24) and

$$\sigma_{ij} = -(c-1)w[\frac{1}{p_i} + \frac{1}{p_j}] + \sum_{k\neq i,j} (\frac{1}{p_k})w = w[q - \frac{c}{p_i} - \frac{c}{p_j}], \quad i \neq j$$

where $q = \Sigma_i(1/p_i)$, and $\sigma_{ij}^{(\alpha,\beta)} = \sigma_{ij}\rho_{\alpha\beta}$.

Let $\underset{\sim}{\Sigma}^* = (\sigma_{ij})$, $i,j = 1,2,\ldots,c$, where σ's are given by (3.24). Then

(3.25) $$w^{-1}\underset{\sim}{\Sigma}^* = c^2\underset{\sim}{P}^{-1} - c\underset{\sim}{q}\underset{\sim}{j}' - c\underset{\sim}{j}\underset{\sim}{q}' + q\underset{\sim}{J} ,$$

where $\underset{\sim}{P}$ = diagonal $(p_i, i=1,2,\ldots,c)$, $\underset{\sim}{J} = (1)_{c\times c}$,

$$\underset{\sim}{j} = (1)_{c\times 1} \text{ and } \underset{\sim}{q}' = (1/p_1, \ldots, 1/p_c).$$

Then, from (3.9), under H_o, $\underset{\sim}{\Sigma}_{ij} = \sigma_{ij}\underset{\sim}{\rho}$ and

(3.26) $$\underset{\sim}{\Sigma} = \begin{bmatrix} \sigma_{11}\underset{\sim}{\rho} \cdots \sigma_{1c}\underset{\sim}{\rho} \\ \cdots \\ \sigma_{c1}\underset{\sim}{\rho} \cdots \sigma_{cc}\underset{\sim}{\rho} \end{bmatrix} = \underset{\sim}{\rho} \otimes \underset{\sim}{\Sigma}^* ,$$

the Kronecker (or the direct) product of matrices $\underset{\sim}{\rho}$ and $\underset{\sim}{\Sigma}^*$. Thus by the extension of Hoeffding's theorem it follows that, under the conditions of Theorem 3.1, $\sqrt{N}(\underset{\sim}{U}_N - d\underset{\sim}{j})$ has a limiting normal distribution with zero mean vector and covariance matrix given by (3.26); here again $\underset{\sim}{j}$ stands for a column vector of the appropriate order with unit elements.

Now from (i), (ii) and (iv) it follows that $\Sigma_i \underset{\sim}{U}_{iN} = c d\underset{\sim}{j}$ and, moreover, it is the only linear restriction on $\underset{\sim}{U}_{iN}$'s. Hence the distribution of $\underset{\sim}{U}_N$ is singular of rank p(c-1) provided the distribution F of $\underset{\sim}{X}$ is nonsingular in the sense that the unit probability mass is not contained in any lower dimensional space. The limiting normal distribution of $\sqrt{N}(\underset{\sim}{U}_N - d\underset{\sim}{j})$ is then also singular of rank p(c-1) which is, thus, also the rank of $\underset{\sim}{\Sigma}$ given by (3.26). In fact, it can be verified that $\underset{\sim}{\Sigma}^* \underset{\sim}{j} = \underset{\sim}{0}$ and that $\underset{\sim}{\Sigma}^*$ is of rank (c-1). It is then seen that, as expected, $\underset{\sim}{\rho}$ is nonsingular; it can be singular if and only if there is some linear constraint on the random variables $w(X^{(\alpha)})$, $\alpha=1,2,\ldots,p$ defined in (3.17), which is impossible since the distribution of $\underset{\sim}{X}$ is assumed to be nonsingular.

If we consider $\underset{\sim}{U}'_{0N} = (\underset{\sim}{U}'_{1N},\ldots,\underset{\sim}{U}'_{c-1,N})$, then $\sqrt{N}(\underset{\sim}{U}_{0N} - d\underset{\sim}{j})$ is asymptotically normal with zero mean vector and covariance matrix

$$\underset{\sim}{\Sigma}_0 = \underset{\sim}{\rho} \otimes \underset{\sim}{\Sigma}_0^* ,$$

where $\underset{\sim}{\Sigma}_0^* = (\sigma_{ij}),\ i,j = 1,2,\ldots,c-1$. Also

$$\underset{\sim}{\Sigma}_0^{-1} = \rho^{-1} \otimes \underset{\sim}{\Sigma}_0^{-1*}. \quad \text{Let } \underset{\sim}{\Sigma}_0^{-1*} = (\sigma^{ij}).$$

Then

$$N(\underset{\sim}{U}_{0N} - d\underset{\sim}{j})'\underset{\sim}{\Sigma}_0^{-1}(\underset{\sim}{U}_{0N} - d\underset{\sim}{j}) = N \sum_{i,j=1}^{c-1}\sum \sigma^{ij}(\underset{\sim}{U}_{iN} - d\underset{\sim}{j})'\underset{\sim}{\rho}^{-1}(\underset{\sim}{U}_{jN} - d\underset{\sim}{j})$$

$$= N \sum_{k,j=1}^{c-1}\sum \sigma^{ij} \operatorname{tr} \underset{\sim}{\rho}^{-1}(\underset{\sim}{U}_{jN} - d\underset{\sim}{j})(\underset{\sim}{U}_{iN} - d\underset{\sim}{j})'$$

$$= N \operatorname{tr} \underset{\sim}{\rho}^{-1} \sum_{k,j=1}^{c-1}\sum \sigma^{ij}(\underset{\sim}{U}_{jN} - d\underset{\sim}{j})(\underset{\sim}{U}_{iN} - d\underset{\sim}{j})'$$

which, after simplification as in [1], can be shown to reduce to

$$\frac{N}{wc^2} \operatorname{tr} \underset{\sim}{\rho}^{-1}\left[\sum_{i=1}^{c} p_i(\underset{\sim}{U}_{iN} - \overline{\underset{\sim}{U}}_N)(\underset{\sim}{U}_{iN} - \overline{\underset{\sim}{U}}_N)'\right] ,$$

which is equal to the expression (3.12). The theorem then follows.

Remark. As far as the above theorem is concerned, the condition (iii) can be replaced by the weaker condition that $\mathcal{E}[\Phi(X_1,X_2,\ldots,X_c)]^2$ be finite. The functions Φ defined by (3.4) to (3.7) satisfy the stronger condition (iii).

We note that in T_n, given by (3.12), even though μ is independent of F, $\rho_{\alpha\beta}$ does depend, in general, on F and thus T_n is not a distribution-free statistic. But we can construct unbiased and consistent estimators $e_{\alpha\beta}$ of $\rho_{\alpha\beta}$, i.e., an unbiased and consistent estimator $\underset{\sim}{E} = (e_{\alpha\beta})$ of $\underset{\sim}{\rho}$ as follows:

Let

$$\mu e_{\alpha\beta(i)} = \frac{1}{n_i(n_i-1)\ldots(n_i-2c+2)} \sum_p \Phi(X_{ij_1}^{(\alpha)};\{X_{ij_k}^{(\alpha)},\ k = 2,3,\ldots,c\})$$
$$\times \Phi(X_{ij_1}^{(\beta)};\{X_{ij_k}^{(\beta)},\ k = c+1,\ldots,2c-1\}) - d^2 ,$$

where p denotes the summation over all permutations of (2c-1) integers $(j_1,j_2,\ldots,j_{2c-1})$ that can be chosen out of $(1,2,\ldots,n_i)$. It can be seen that $e_{\alpha\beta(i)}$ is a U-statistic and, hence (see e.g., [7] p. 142) a minimum variance unbiased estimator, that can be formed from the i^{th} sample, of $\rho_{\alpha\beta}$. Moreover it is well-known that $e_{\alpha\beta(i)} \xrightarrow{(p)} \rho_{\alpha\beta}$ as $n_i \longrightarrow \infty$. So we can have an unbiased and consistent estimator of $\rho_{\alpha\beta}$ given by $\Sigma_i\ e_{\alpha\beta(i)}/c$. We shall prefer another consistent and unbiased estimator

$$e_{\alpha\beta} = \frac{\sum_{i=1}^{c} n_i(n_i-1)\ldots(n_i-2c+2)e_{\alpha\beta(i)}}{\sum_{i=1}^{c} n_i(n_i-1)\ \ldots\ (n_i-2c+2)} . \tag{3.27}$$

If we let

$$T_N^* = \frac{N}{\mu}\frac{(c-1)^2}{c^2}\sum_{i=1}^{c} p_i(\underset{\sim}{U}_{iN} - \overline{\underset{\sim}{U}}_N)'\underset{\sim}{E}^{-1}(\underset{\sim}{U}_{iN}-\overline{\underset{\sim}{U}}_N) , \tag{3.28}$$

it can be easily shown that $T_N - T_N^* \xrightarrow{(p)} 0$, under the conditions of Theorem (3.1) and hence T_N^* also has a limiting χ^2 distribution with p(c-1) degrees of freedom. Thus we have

Theorem 3.2. Under the conditions of Theorem (3.1), the statistic T_N^* given by (3.28) has a limiting χ^2 distribution, under H_o, with degrees of freedom p(c-1).

4. Special Cases. (i) For Φ given by (3.4), $d = c^{-1}$ and

$$\psi^{(\alpha)}(x) = P[x < X_j^{(\alpha)},\ j = 2,3,\ldots,c] - c^{-1}$$

$$= [1 - F^{(\alpha)}(x)]^{c-1} - c^{-1} \quad ,$$

$$\mu = \int_{-\infty}^{\infty} [1 - F^{(\alpha)}(x)]^{2c-2}\, dF^{(\alpha)}(x) - c^{-2}$$

$$= (c-1)^2/c^2(2c-1) \ ,$$

$$\mu\rho_{\alpha\beta} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [1 - F^{(\alpha)}(x)]^{c-1}[1-F^{(\beta)}(y)]^{c-1} dF^{(\alpha,\beta)}(x,y) - c^{-2},$$

so that, if $N^* = \sum_{i=1}^{c} n_i(n_i-1)\ldots(n_i-2c+2)$, we have

(4.1) $$c^{-2} + \mu^e{}_{\alpha\beta} = \frac{1}{N^*}\ [\text{The number of } (2c-1)\text{-tuples } \underset{\sim}{X}_{ij_1},\ldots,\underset{\sim}{X}_{ij_{2c-1}} \text{ such that } X_{ij_1}^{(\alpha)} < X_{ij_k}^{(\alpha)},\ k = 2,\ldots,c \text{ and } X_{ij_1}^{(\beta)} < X_{ij_k}^{(\beta)},\ k = c+1,\ldots,2c-1],$$

and we get the V-statistic (2.1) by suppressing N in the subscript of U.

(ii) For Φ given by (3.5), $d = c^{-1}$,

$$\psi^{(\alpha)}(x) = [F^{(\alpha)}(x)]^{c-1} - c^{-1},$$

$$\mu = (c-1)^2/c^2(2c-1) \ ,$$

$$\mu\rho_{\alpha\beta} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F^{(\alpha)}(x)]^{c-1}[F^{(\beta)}(y)]^{c-1}\, dF^{(\alpha,\beta)}(x,y) - c^{-2},$$

and we have

(4.2) $c^{-2} + \mu e_{\alpha\beta} = \frac{1}{N^*}$ [The number of (2c-1)-tuples $\underset{\sim}{X}_{ij_1},\ldots,\underset{\sim}{X}_{ij_{2c-1}}$ such that $X^{(\alpha)}_{ij_1} > X^{(\alpha)}_{ij_k}$, $k = 2,\ldots,c$ and $X^{(\beta)}_{ij_1} > X^{(\beta)}_{ij_k}$ $k = c+1,\ldots,2c-1$,

and supressing N we get the B-statistic (2.2).

(iii) Similarly, for Φ given by (3.6), $d = 0$

$$\psi^{(\alpha)}(x) = P[x > X^{(\alpha)}_j,\ j = 2,\ldots,c] - P[x < X^{(\alpha)}_j,\ j = 2,\ldots,c]$$

$$= [F^{(\alpha)}(x)]^{c-1} - [1 - F^{(\alpha)}(x)]^{c-1},$$

$$\mu = \frac{1}{2c-1}\,\frac{2(c-1)!(c-1)!}{(2c-1)!} + \frac{1}{2c-1} = \frac{2}{(2c-1)\binom{2c-2}{c-1}}\left[\binom{2c-2}{c-1} - 1\right],$$

$$\mu\rho_{\alpha\beta} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{[F^{(\alpha)}(x)]^{c-1} - [1-F^{(\alpha)}(x)]^{c-1}\}\{[F^{(\beta)}(y)]^{c-1} - [1-F^{(\beta)}(y)]^{c-1}\}\, dF^{(\alpha,\beta)}(x,y),$$

and we have

(4.3) $$\mu e_{\alpha\beta} = \frac{1}{N^*}[N^*_1(\alpha,\beta) + N^*_2(\alpha,\beta) - N^*_3(\alpha,\beta) - N^*_4(\alpha,\beta)],$$

where $N^*_1(\alpha,\beta)$ and $N^*_2(\alpha,\beta)$ are the numbers of (2c-1)-tuples mentioned in (4.1) and (4.2), respectively; $N^*_3(\alpha,\beta)$ is the number of (2c-1)-tuples with $X^{(\alpha)}_{ij_1} > X^{(\alpha)}_{ij_k}$, $k = 2,3,\ldots,c$ and $X^{(\beta)}_{ij_1} < X^{(\beta)}_{ij_k}$, $k = c+1,\ldots,2c-1$, and $N^*_4(\alpha,\beta)$ is the number like $N^*_3(\alpha,\beta)$ obtained from inequalities with signs reversed, and we get the L-statistic (2.3).

(iv) Finally for Φ given by (3.7), $d = (c-1)/2$,

$$\psi^{(\alpha)}(x) = \sum_{j=2}^{c} P[X_j^{(\alpha)} < x] - (c-1)/2 = (c-1)F^{(\alpha)}(x) - (c-1)/2,$$

$$\mu = (c-1)^2 \int_{-\infty}^{\infty} [F^{(\alpha)}(x)]^2 \, dF^{(\alpha)}(x) - (c-1)^2/4 = (c-1)^2/12,$$

$$\mu\rho_{\alpha\beta} = (c-1)^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{(\alpha)}(x)F^{(\beta)}(y) \, dF^{(\alpha,\beta)}(x,y) - (c-1)^2/4,$$

that is

$$\rho_{\alpha\beta} = 3 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [2F^{(\alpha)}(x)-1][2F^{(\beta)}(y)-1] \, dF^{(\alpha,\beta)}(x,y),$$

which is the "grade correlation coefficient" ([7], p. 259) between $X^{(\alpha)}$ and $X^{(\beta)}$, and we have

(4.4) $$3 + e_{\alpha\beta} = 12 \, \Sigma_i \nu_i^{(\alpha,\beta)} / \Sigma_i n_i(n_i-1)(n_i-2) \ ,$$

where $\nu_i^{(\alpha,\beta)}$ is the number of triplets $\underset{\sim}{X}_{ij_1}$, $\underset{\sim}{X}_{ij_2}$, $\underset{\sim}{X}_{ij_3}$ such that $X_{ij_1}^{(\alpha)} > X_{ij_2}^{(\alpha)}$ and $X_{ij_1}^{(\beta)} > X_{ij_3}^{(\beta)}$, and we get the W-statistic (2.4). In this case

$$U_i^{(\alpha)} = \frac{1}{n_1 n_2 \cdots n_c} \sum_{j=1}^{c} \frac{n_1 n_2 \cdots n_c}{n_i n_j} \sum_{t_i=1}^{n_i} \sum_{t_j=1}^{n_j} h(x_{it_i}^{(\alpha)} - x_{jt_j}^{(\alpha)})$$
$$= \sum_{j=1}^{c} \nu_{ij}^{(\alpha)}/n_i n_j \ ,$$

where $\nu_{ij}^{(\alpha)}$ is the number of pairs $(\underset{\sim}{x}_{it_i}, \underset{\sim}{x}_{jt_j})$ such that $x_{it_i}^{(\alpha)} > x_{jt_j}^{(\alpha)}$.

If $n_1 = n_2 = \cdots = n_c = n$, say, then

(4.5) $$U_i^{(\alpha)} = \frac{1}{n^2} [n\bar{R}_i^{(\alpha)} - \frac{n(n-1)}{2}] = \frac{c}{N}[\bar{R}_i^{(\alpha)} - \frac{n+1}{2}].$$

It has been observed [2] that, in the univariate case, even with unequal n_i's, the H-statistic [11] is the same as the one obtained from the W-statistic by making the transformation (4.5) and, hence, in the multivariate case it is conjectured that the appropriate H-statistic will be given from the W-statistic by using (4.5), i.e., by (2.5). A rigorous proof for this conjecture can be given by considering the asymptotic normal distribution of $N^{-1/2}[\underset{\sim}{\bar{R}}_N - \underset{\sim}{j}(N+1)/2]$, where $\underset{\sim}{\bar{R}}'_N = (\underset{\sim}{\bar{R}}'_{1N}, \ldots, \underset{\sim}{\bar{R}}'_{cN})$, under H_o; the asymptotic normal distribution under H_o can be obtained either by an appeal to the multivariate extension of the Wald-Wolfowitz theorem (see e.g., [7] p. 239), or by noting that

$$n_i\bar{R}_i^{(\alpha)} = [n_i(n_i+1)/2] + \Sigma_j \nu_{ij}^{(\alpha)}$$
$$= [n_i(n_i+1)/2] + n_i\Sigma_j n_j U_{ij}^{(\alpha)} ,$$

where

$$U_{ij}^{(\alpha)} = \frac{1}{n_i n_j} \sum_{t_i=1}^{n_i} \sum_{t_j=1}^{n_j} h(x_{it_i}^{(\alpha)} - x_{jt_j}^{(\alpha)}) ,$$

is a two-sample U-statistic corresponding to the function $\Phi(x,y) = h(x-y)$, obtained for the sample pair (i,j), and then making an appeal to the joint limiting normal distribution of $\sqrt{N}[\underset{\sim}{U}_{ij_N} - \underset{\sim}{j}(1/2)]$'s.

5. Remarks. It is seen that the U's occurring in these statistics are in the nature of "between sample" comparisons while the e's are in the nature of "within sample" comparisons. The comparisons themselves are with respect to a particular function Φ defined appropriately in each case.

In the univariate case it has been observed [1] that the V-statistic is more efficient than the H-statistic (or the L and B statistics), in the Pitman sense, for populations bounded below (e.g., exponential distribution $f(y,\alpha) = e^{-(y-\alpha)}$, $y \geq \alpha$). It is expected that the B-statistic is similarly more efficient for populations bounded above (e.g., reversed exponential distribution $f(y,\alpha) = e^{(y-\alpha)}$, $y \leq \alpha$). Both of these are fairly efficient (and the L-statistic is even much more so) for distributions bounded on both sides (e.g., uniform distribution $f(x,\alpha,\beta) = 1/(\beta-\alpha)$, $\alpha \leq x \leq \beta$). The W-statistic is seen [2] to be as efficient as the H-statistic and these two appear to be more efficient for unbounded distributions.

It is conjectured that the same will be true for the corresponding multivariate analogues. These are also expected to be consistent against the relevant class of alternatives and, especially against the class of translation alternatives. Work is in progress on these problems and will be presented in a subsequent communication. Since the distributions are assumed to be continuous, the probability that any two observations are equal is zero. But, in practice, ties do occur. This problem will also be considered in the next communication.

REFERENCES

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

[2] Bhapkar, V. P. "A nonparametric test for the several sample location problem," North Carolina Institute of Statistics, Mimeo Series No. 411 (1964).

[3] Bhapkar, V. P. "Some nonparametric median procedures," Ann. Math. Stat. vol. 32 (1961), pp. 846-863.

[4] Chatterjee, S. K. and P. K. Sen. "Nonparametric tests for the bivariate two-sample location problem," Calcutta Stat. Asso. Bull. vol. 13 (1964) pp. 18-58.

[5] Deshpande, Jayant V."A nonparametric test based on U-statistics for several samples," (Abstract) Ann. Math. Stat., vol. 34 (1963), p. 1624.

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

[7] Fraser, D. A. S. Nonparametric Methods in Statistics, John Wiley and Sons, New York, 1957.

[8] Hodges, J. L., Jr. "A bivariate sign test ," Ann. Math. Stat. vol. 26 (1955), pp. 523-527.

[9] Hoeffding, Wassily. "A class of statistics with asymptotic normal distributions," Ann. Math. Stat., vol. 19 (1948), pp. 293-325.

[10] Kiefer, J. "K-sample analogues of the Kolmogorov-Smirnov and Cramer-v. Mises tests," Ann. Math. Stat. vol. 30 (1959), pp. 420-447.

[11] Kruskal, William H. "A nonparametric test for the several sample problem," Ann. Math. Stat., vol. 23 (1952), pp. 525-540.

[12] Kruskal, William H. and W. Allen Wallis, "Use of ranks in one-criterion variance analysis," J. Amer. Stat. Asso. vol. 47 (1952), pp. 583-621.

[13] Mood, Alexander McFarlane. Introduction to the Theory of Statistics, McGraw-Hill Book Co., New York, 1950.

[14] Vincze, I. "On two-sample tests based on order statistics," Proc. Fourth Berkeley Symp. on Prob. and Math. Stat., University of California, Berkeley, pp. 695- .