ON ANALYSIS OF VARIANCE FOR THE K-SAMPLE PROBLEM

by

Dana Quade

University of North Carolina

Institute of Statistics Mimeo Series No. 453

December 1965

This research was supported by the National Institutes of Health Grant No. GM-10397

DEPARTMENT OF BIOSTATISTICS

UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

ON ANALYSIS OF VARIANCE FOR THE K-SAMPLE PROBLEM[1]

by Dana Quade

University of North Carolina

1. INTRODUCTION. Suppose we have $k \geq 2$ random samples, possibly multivariate, where X_{ij} is the j-th observation in the i-th sample for $1 \leq i \leq k$, $1 \leq j \leq n_i$, and there are $N = \Sigma n_i$ observations in all. Let the distribution function of X_{ij} be G_i. The null hypothesis to be tested is

$$H_o: \quad G_1 \equiv G_2 \equiv \ldots \equiv G_k.$$

We shall be particularly concerned with the large sample situation where $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i < k$. Define

$$G = \sum_{i=1}^{k} p_i G_i ;$$

then our null hypothesis can be reexpressed as

$$H_o: G_i \equiv G, \quad i=1,2,\ldots,k.$$

The general approach we have in mind is an extended version of one-way analysis of variance. Let $f_N(x_1, x_2, \ldots, x_N)$ be a function of N arguments which is symmetric in the last (N-1) of them. Next, corresponding to each observation X_{ij} define a score

$$y_{ij} = f_N(X_{ij}, [\text{(N-1) X's other than } X_{ij}]).$$

---

[1] Supported by the National Institutes of Health Grant No. GM-10397

Then we perform an ordinary analysis of variance based on the scores: that is, we calculate

$$F = \frac{(N-k)\,\Sigma n_i(\bar{y}_i - \bar{\bar{y}})^2}{(k-1)\,\Sigma\Sigma(y_{ij} - \bar{\bar{y}})^2},$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{\bar{y}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$, and test H_o by referring F to the F-distribution with (k-1, N-k) degrees of freedom.

In the remainder of this paper we present conditions under which such a test will be asymptotically valid, and we show that several of the tests proposed in the literature are essentially of this type.

2. THE NULL-HYPOTHESIS DISTRIBUTION. In accordance with Chernoff and Teicher [2], we shall say that n random variables are interchangeable if their joint distribution function is symmetric. Then clearly the scores as defined above are interchangeable random variables if the null hypothesis is true. Now, by a completely straightforward extension of Theorem 1 of Chernoff and Teicher, using the multivariate extension of the Wald-Wolfowitz-Noether limit theorem, we can obtain the following

Theorem 1. For every sufficiently large N let $\{Z_{ij}^{(N)}\}$ be a set of N interchangeable random variables, where $1 \le j \le n_i^{(N)}$, $\sum_{i=1}^{k} n_i^{(N)} = N$, and as $N \to \infty$ $n_i^{(N)}/N \to p_i > 0$ for $1 \le i \le k$. Suppose also that for all N

(i) $$\sum_{i=1}^{k}\sum_{j=1}^{n_i^{(N)}} Z_{ij}^{(N)} = 0$$

(ii) $$E[Z_{ij}^{(N)}]^2 = 1$$

and that as $N \longrightarrow \infty$

(iii) $$\frac{1}{\sqrt{N}} \max_{i,j} |Z_{ij}^{(N)}| \longrightarrow 0 \text{ in probability,}$$

(iv) $$\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i^{(N)}} [Z_{ij}^{(N)}]^2 \longrightarrow 1 \text{ in probability.}$$

Then the k random variables

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{n_i^{(N)}} Z_{ij}^{(N)}, \quad i=1,2,\ldots,k$$

have asymptotically a k-variate normal distribution with zero mean vector and variance matrix

$$V = \begin{bmatrix} p_1-p_1^2 & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2-p_2^2 & & -p_2p_k \\ \vdots & & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k-p_k^2 \end{bmatrix}.$$

Application of the preceding theorem to our problem gives

Theorem 2. If the hypothesis is true, and if $N \overset{as}{\longrightarrow} \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i \leq k$ we have

Assumption A: $$\frac{\max_{i,j} |y_{ij} - \bar{\bar{y}}|}{\sqrt{\Sigma\Sigma\ (y_{ij} - \bar{\bar{y}})^2}} \longrightarrow 0 \text{ in probability,}$$

then the analysis of variance statistic F based on the scores has asymptotically the F-distribution with (k-1, N-k) degrees of freedom.

Proof. Define

$$Z_{ij} = \frac{y_{ij} - \bar{\bar{y}}}{\sqrt{\frac{1}{N}\Sigma\Sigma(y_{ij} - \bar{\bar{y}})^2}}$$

for $1 \le i \le k$, $1 \le j \le n_i$. Then the analysis of variance statistic based on the Z's is identically the same as the statistic based on the y's. But it can be verified immediately that the Z's are interchangeable and satisfy all the conditions (i) - (iv) of Theorem 1. Hence the k random variables

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{n} Z_{ij}, \quad i=1,2,\ldots,k,$$

have asymptotically a k-variate normal distribution with zero mean vector and variance matrix V. It then follows that the random variable

$$\sum_{i=1}^{k} [\sum_{j=1}^{n_i} Z_{ij}]^2/n_i$$

has asymptotically a χ^2-distribution with (k-1) degrees of freedom. Hence

$$\frac{1}{N-k} \sum_{i=1}^{k} [\sum_{j=1}^{n_i} Z_{ij}]^2/n_i \longrightarrow 0$$

in probability, and thus

$$\frac{1}{N-k} \{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Z_{ij}^2 - \sum_{i=1}^{k} [\sum_{j=1}^{n_i} Z_{ij}]^2/n_i\} \longrightarrow 1$$

in probability. Then by Cramér's convergence theorem ([3], p. 254) we have that (k-1)F is asymptotically a $\chi^2(k-1)$ also. But this is equivalent to saying that F itself is asymptotically an F(k-1, N-k).

The reader may note that we have passed up several opportunities for obtaining a statistic with asymptotic $\chi^2(k-1)$ distribution. He may then ask, why put the test in terms of F rather than the simpler χ^2? This may be partly a matter of personal preference. However, statement in terms of F may be intuitively appealing to experimenters accustomed to applying analysis of variance to the k-sample problem, and it provides a unified approach, since an ordinary analysis of variance based on untransformed univariate data is clearly a special case of analysis of variance based on scores. Alternatively, the test could be regarded as a conditional one, given the scores; this will be a particularly useful device if N is small. Then we have a permutation test based on the analysis of variance statistic, and the results of Pitman [5] indicate that the F-distribution affords an excellent approximation to the required conditional distribution. However, our approach throughout is to consider the test as unconditional.

3. SCORES RELATED TO U-STATISTICS. Although the general definition used for scores so far leads to a satisfactory result under the null hypothesis, it is not evident how this may be extended to the case where the alternative holds true. In order to obtain such an extension, we shall suppose in this section that the scores have the more specialized form

$$y_{ij} = \Sigma_{ij}\, \phi(X_{ij}, [m\ X's]),$$

where the function $\phi(x_0, x_1, \ldots, x_m)$ is symmetric in its last m arguments, $1 \leq m \leq \min(n_i)$, and the summation extends over all possible combinations of m

observations other than X_{ij}.

We shall also make the following

Assumption B1: $$\eta = \int \cdots \int \phi^2(x_0, x_1, \ldots x_m) dG(x_0)\, dG(x_1) \ldots dG(x_m) < \infty .$$

This is equivalent to assuming that

$$E[\phi^2(X_0, X_1, \ldots X_m)] < \infty$$

whenever $X_0, X_1, \ldots, X_m$ are independent and their distributions are all taken from $G_1, G_2, \ldots, G_k$. Let $\{m_r\}$ be any ordered set of k nonnegative integers $(m_1, m_2, \ldots, m_k)$ such that $\Sigma m_r = m$, and define

$$\eta_i^{\{m_r\}} = E[\phi^2(X_0, X_1, \ldots, X_m)]$$

where $X_0, X_1, \ldots X_m$ are independent, X_0 comes from G_i, and m_r of the other m X's come from G_r for $1 \leq r \leq k$. Then we may write

$$\eta = \sum_{i=1}^{k} p_i \; \Sigma^* \; R^{\{m_r\}} \; \eta_i^{\{m_r\}}$$

where

$$R^{\{m_r\}} = \frac{m! p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}}{m_1! m_2! \cdots m_k!}$$

and Σ^* indicates summation over all $\binom{m+k-1}{m}$ possible choices of $\{m_r\}$. Hence

$$\eta > \left[\min_i p_i\right]^{m+1} \max_{i, \{m_r\}} \eta_i^{\{m_r\}} ,$$

or

$$\eta_i^{\{m_r\}} < \eta[\min_i p_i]^{-m-1}$$

for all i, $\{m_r\}$.

We note that Assumption B1 is trivially satisfied if ϕ is bounded. Since we shall need this stronger condition for the results of Section 4, we state it here as

Assumption B2: $$|\phi(x_0, x_1, \ldots, x_m)| \leq c.$$

(The constant c may depend on G.) Assumption B2 is satisfied in all the examples to follow.

A few further definitions will be useful. Let

$$\theta_i^{\{m_r\}} = E[\phi(X_0, X_1, \ldots, X_m)]$$

where $X_0, X_1, \ldots, X_m$ are independent, X_0 comes from G_i, and m_r of the other m X's come from G_r for $1 \leq r \leq k$. Then

$$\theta_i^{\{m_r\}} = \int \theta^{\{m_r\}}(x)\, dG_i(x)$$

where

$$\theta^{\{m_r\}}(x) = \int \cdots \int \phi(x, x_{11}, x_{12}, \ldots, x_{1m_1}, x_{21}, \ldots, x_{km_k}) \prod_{r=1}^{k} \prod_{u=1}^{m_r} dG_r(x_{ru}).$$

Define also

$$\theta(x) = \int\ldots\int\phi(x,x_1,x_2,\ldots,x_m)\ dG(x_1)\ dG(x_2)\ldots dG(x_m) = \Sigma^*\ R^{\{m_r\}}\theta^{\{m_r\}}(x),$$

$$\theta_i = \int\theta(x)\ dG_i(x),$$

$$\theta = \int\theta(x)\ dG(x) = \sum_{i=1}^{k} p_i\theta_i,$$

and finally

$$\zeta_i = \int\theta^2(x)\ dG_i(x),$$

$$\zeta = \int\theta^2(x)\ dG(x) = \sum_{i=1}^{k} p_i\ \zeta_i\ .$$

Theorem 3. If Assumption B1 holds, then as $N\longrightarrow\infty$ and $n_i/N\longrightarrow p_i > 0$ for $1 \leq i \leq k$ the k random variables

$$N^{\frac{1}{2}}\{\binom{N-1}{m}^{-1}\ \bar{y}_i - \theta_i\}$$

have asymptotically a joint normal distribution with zero mean vector and finite variance matrix.

Proof. We may write

$$y_{ij} = \Sigma^*\ \Sigma_{ij}^{\{m_r\}}\ \phi(X_{ij},[m\ X's])$$

where the inner summation extends over all possible combinations of m observations other than X_{ij} such that m_r of them come from the r-th sample for $1 \leq r \leq k$. Then

$$\sum_{j=1}^{n_i} y_{ij} = \binom{N-1}{m} \Sigma^* \frac{n_i - m_i}{m_i+1} \; R_N^{\{m_r\}} \; U_i^{\{m_r\}}$$

where

$$R_N^{\{m_r\}} = \binom{N-1}{m}^{-1} \prod_{r=1}^{k} \binom{n_r}{m_r} ,$$

$$U_i^{\{m_r\}} = \left[\prod_{r=1}^{k} \binom{n_r}{m_r+\delta_{ir}} \right]^{-1} \sum_{j=1}^{n_i} \Sigma_{ij}^{\{m_r\}} \phi(X_{ij}, [m \; X's]),$$

and δ_{ir} is Kronecker's delta. For $0 \leq u \leq m$ define

$$\phi_u(x_0, x_1, \ldots, x_m)$$

$$= \sum_{r=0}^{u} \phi(x_r, x_1+\delta_{1r}(x_0-x_1), x_2+\delta_{2r}(x_0-x_2), \ldots, x_u+\delta_{ur}(x_0-x_u), x_{u+1}, \ldots, x_m),$$

so that ϕ_u is symmetric in its first (u+1) arguments and also in its last (m-u) arguments. Then

$$\sum_{j=1}^{n_i} \Sigma_{ij}^{\{m_r\}} \phi(X_{ij}, [m \; X's]) = \Sigma_i^{\{m_r\}} \phi_{m_i}([(m_i+1) \; X_i's], [(m-m_i) \; X's])$$

where the summation extends over all possible combinations of (m+1) observations such that the first (m_i+1) of them come from the i-th sample and that m_r of the others come from the r-th sample for $1 \leq r \leq k$, $r \neq i$. Thus

$$U_i^{\{m_r\}} = \left[\prod_{r=1}^{k} \binom{n_r}{m_r+\delta_{ir}} \right]^{-1} \Sigma_i^{\{m_r\}} \phi_{m_i}([(m_i+1) \; X_i's], [(m-m_i) \; X's])$$

has the form of a k-sample generalized U-statistic, and hence, using Lemma 3.1 of Bhapkar [1], it follows that the $k\binom{m+k-1}{m}$ random variables

$$N^{\frac{1}{2}} \left(U_i^{\{m_r\}} - E[U_i^{\{m_r\}}]\right)$$

have asymptotically a joint normal distribution with zero mean vector and finite variance matrix. Then the same result must hold true also for the k random variables

$$N^{\frac{1}{2}}(Q_i - E[Q_i]),$$

where

$$Q_i = \frac{1}{(m_i+1)} \Sigma^* R^{\{m_r\}} U_i^{\{m_r\}}$$

for $1 \leq i \leq k$. Now

$$E[U_i^{\{m_r\}}] = (m_i+1)\, \theta_i^{\{m_r\}} ,$$

and hence

$$E[Q_i] = \Sigma^* R^{\{m_r\}} \theta_i^{\{m_r\}} = \theta_i .$$

Also,

$$\binom{N-1}{m}^{-1} \bar{y}_i - Q_i = \Sigma^* \left[\frac{n_i-m_i}{n_i} R_N^{\{m_r\}} - R^{\{m_r\}}\right] \frac{U_i^{\{m_r\}}}{m_i+1} ,$$

so

$$\text{var}[\binom{N-1}{m}^{-1} \bar{y}_i - Q_i] \leq \binom{m+k-1}{m}^2 \max_{\{m_r\}} [\frac{n_i - m_i}{n_i} R_N^{\{m_r\}} - R^{\{m_r\}}]^2 \max_{\{m_r\}} \text{var}[\frac{U_i^{\{m_r\}}}{m_i+1}]$$

$$= O(1) \times o(1) \times O(\tfrac{1}{N}) ,$$

and thus the random variables $N^{\frac{1}{2}}\{\binom{N-1}{m}^{-1} \bar{y}_i - Q_i\}$ converge to zero in mean square. Then the asymptotic joint distribution of the random variables $N^{\frac{1}{2}}\{\binom{N-1}{m}^{-1} \bar{y}_i - \theta_i\}$ must be the same as that of the random variables $N^{\frac{1}{2}}\{Q_i - E[Q_i]\}$, and the theorem is proven.

In the next two theorems we extend the argument used by Sen [7] in establishing the structural convergence of U-statistics. For this we shall need to consider the random variables $\theta(X_{ij})$ for $1 \leq i \leq k$, $1 \leq j \leq n_i$. These variables are mutually independent, the expected value of $\theta(X_{ij})$ is

$$E[\theta(X_{ij})] = \int \theta(x)\, dG_i(x) = \theta_i$$

as defined previously, and the variance of $\theta(X_{ij})$ is

$$\text{var}[\theta(X_{ij})] = \zeta_i - \theta_i^2 .$$

We have also

Theorem 4. If Assumption B1 holds, then as $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i \leq k$ we have

$$E[\binom{N-1}{m}^{-1} y_{ij} - \theta(X_{ij})]^2 = O(\tfrac{1}{N})$$

uniformly for all i, j.

Proof. Starting from y_{ij} as expressed in the proof of Theorem 3, we find that

$$\binom{N-1}{m}^{-1} y_{ij} = \Sigma^* \frac{n_i - m_i}{n_i} R_N^{\{m_r\}} V_{ij}^{\{m_r\}}$$

where $R_N^{\{m_r\}}$ is as defined previously, and

$$V_{ij}^{\{m_r\}} = \left[\prod_{r=1}^{k} \binom{n_r - \delta_{ir}}{m_r} \right]^{-1} \Sigma_{ij}^{\{m_r\}} \phi(X_{ij}, [m\ X's]).$$

Suppose for the moment that X_{ij} is fixed. Then, conditioned on X_{ij}, $V_{ij}^{\{m_r\}}$ is a k-sample generalized U-statistic, and we have

$$E[V_{ij}^{\{m_r\}} | X_{ij}] = \theta^{\{m_r\}}(X_{ij}).$$

Now consider estimating the quantity $\theta^{\{m_r\}}(X_{ij})$ not by $V_{ij}^{\{m_r\}}$ but by the mean of independent statistics of the form $\phi(X_{ij}, [m\ X's])$. The number of such independent statistics which can be obtained from the data is the largest integer not greater than $\min_r \left(\frac{n_r - \delta_{ir}}{m_r}\right)$; by hypothesis, this number will be at least equal to $Np_o/2m$, where $p_o = \min_i p_i$, for all sufficiently large N. The variance of any one of the independent statistics will be

$$\text{var}[\phi(X_{ij}, [m\ X's]) | X_{ij}] \leq E[\phi^2(X_{ij}, [m\ X's]) | X_{ij}] = \eta^{\{m_r\}}(X_{ij}),$$

say, where

$$\eta^{\{m_r\}}(x) = \int \cdots \int \phi^2(x, x_{11}, x_{12}, \ldots, x_{1m_1}, x_{21}, \ldots, x_{km_k}) \prod_{r=1}^{k} \prod_{u=1}^{m_r} dG_r(x_{ru}).$$

Hence the variance of their mean will be no greater than $2m\, \eta^{\{m_r\}}(X_{ij})/Np_o$. But since a U-statistic has minimum variance among all unbiased estimates of its expected value it follows that

$$\mathrm{var}[V_{ij}^{\{m_r\}} \mid X_{ij}] = E[(V_{ij}^{\{m_r\}} - \theta^{\{m_r\}}(X_{ij}))^2 \mid X_{ij}] \le \frac{2m\,\eta^{\{m_r\}}(X_{ij})}{N\,p_o} \quad .$$

Then, integrating over the distribution of X_{ij}, we find unconditionally that

$$E[V_{ij}^{\{m_r\}} - \theta^{\{m_r\}}(X_{ij})]^2 \le \frac{2m\eta}{Np_o^{m+2}} = O(\tfrac{1}{N})$$

uniformly for all $i, j, \{m_r\}$. Note that

$$\theta(x) = \Sigma^* R^{\{m_r\}}\, \theta^{\{m_r\}}(x).$$

Hence

$$\binom{N-1}{m}^{-1} y_{ij} - \theta(X_{ij}) = \Sigma^*[\frac{n_i - m_i}{n_i} R_N^{\{m_r\}}\, V_{ij}^{\{m_r\}} - R^{\{m_r\}}\, \theta^{\{m_r\}}(X_{ij})]$$

$$= \Sigma^* \frac{n_i - m_i}{n_i} R_N^{\{m_r\}}\, [V_{ij}^{\{m_r\}} - \theta^{\{m_r\}}(X_{ij})]$$

$$+ \Sigma^*[\frac{n_i - m_i}{n_i} R_N^{\{m_r\}} - R^{\{m_r\}}]\, \theta^{\{m_r\}}(X_{ij})$$

$$= A + B,$$

say, and

$$E[\binom{N-1}{m}^{-1} y_{ij} - \theta(x_{ij})]^2 \le 4 \max\,(E[A^2],\ E[B^2]).$$

But

$$E[A^2] \le \binom{m+k-1}{m}^2 \max_{i,\{m_r\}} \left[\frac{n_i - m_i}{n_i} R_N^{\{m_r\}}\right]^2 \max_{i,j,\{m_r\}} E[V_{ij}^{\{m_r\}} - \theta^{\{m_r\}}(X_{ij})]^2$$

$$= O(1) \times O(1) \times O\left(\frac{1}{N}\right)$$

uniformly for all i, j, and also

$$E[B^2] \le \binom{m+k-1}{m}^2 \max_{i,\{m_r\}} \left[\frac{n_i - m_i}{n_i} R_N^{\{m_r\}} - R^{\{m_r\}}\right]^2 \max_{i,j,\{m_r\}} E[\theta^{\{m_r\}}(X_{ij})]^2$$

$$= O(1) \times O\left(\frac{1}{N}\right) \times O(1)$$

uniformly for all i, j. Hence the theorem follows immediately.

At this point it is convenient to define

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

for $1 \le i \le k$, and also

$$s^2 = \sum_{i=1}^{k} \frac{(n_i - 1)s_i^2}{N-k} = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad .$$

Then we have

Theorem 5. If Assumption B1 holds, then as $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \le i \le k$ we have

$$s_i^2 \binom{N-1}{m}^{-2} \longrightarrow \zeta_i - \theta_i^2$$

in probability, and

$$s^2 \binom{N-1}{m}^{-2} \longrightarrow \zeta - \sum_{i=1}^{k} p_i \theta_i^2$$

in probability.

Proof. Let

$$a_{ij} = \theta(X_{ij}) - \theta_i, \quad b_{ij} = \binom{N-1}{m}^{-1} y_{ij} - \theta_i.$$

Then $n_i^{-1} \sum_{j=1}^{n_i} a_{ij}^2 \longrightarrow (\zeta_i - \theta_i^2)$ in probability by Khinchin's form of the law of large numbers, and $n_i^{-1} \sum_{j=1}^{n_i} (a_{ij} - b_{ij})^2 \longrightarrow 0$ in probability by Theorem 4. Then by Lemma 1 of Sen [7] $n_i^{-1} \sum_{j=1}^{n_i} b_{ij}^2 \longrightarrow (\zeta_i - \theta_i^2)$ in probability, and certainly

$$\frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left[\binom{N-1}{m}^{-1} y_{ij} - \theta_i\right]^2 \longrightarrow (\zeta_i - \theta_i^2)$$

in probability also. But

$$s_i^2 \binom{N-1}{m}^{-2} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left[\binom{N-1}{m}^{-1} y_{ij} - \theta_i\right]^2 - \frac{n_i}{n_i - 1} \left[\binom{N-1}{m} \bar{y}_i - \theta_i\right]^2$$

and the second of these two sums converges to zero in probability by Theorem 3. Thus the first part of the present theorem is proven; and the second part follows immediately from it.

Theorem 6. If the hypothesis is true, and if furthermore Assumption B1 and

Assumption C: $\zeta - \theta^2 > 0$

both hold, then as $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i \leq k$ the analysis of variance statistic F based on the scores has asymptotically the F-distribution with (k-1, N-k) degrees of freedom.

Proof. We present a direct proof instead of reducing the present theorem to Theorem 2. Under the hypothesis, the scores are interchangeable. We have that the expected value of any score is $E[y_{ij}] = \binom{N-1}{m} \theta$ for all i, j; let us write σ_N^2 for the variance of any score, and ρ_N for the correlation between any two scores. Now let $\underline{Z} = (Z_1, Z_2, \ldots, Z_k)'$, where the Z's are the random variables in the statement of Theorem 3. Then $E[\underline{Z}] = 0$ and it is easily verified that

$$\text{var}[\underline{Z}] = \Sigma_N = N\binom{N-1}{m}^{-2} \sigma_N^2 [(1-\rho_N)D_N^{-1} + \rho_N J]$$

where $D_N = \text{diag}\,(n_1, n_2, \ldots, n_k)$ and J is the matrix in which every element is unity. Let $A_N = \binom{N-1}{m}^2 (ND_N - \underline{n}\,\underline{n}')/N^2\sigma_N^2\,(1-\rho_N)$, where $\underline{n} = (n_1, n_2, \ldots, n_k)'$. Then $A_N\Sigma_N = I - D_N J/N$ is an idempotent matrix of rank (k-1). Hence, since by Theorem 3 the joint distribution of the Z_i's is asymptotically k-variate normal, it follows that the quadratic form

$$\underline{Z}'A_N\underline{Z} = \frac{\Sigma n_i(\bar{y}_i - \bar{\bar{y}})^2}{\sigma_N^2(1-\rho_N)}$$

is asymptotically distributed as $\chi^2(k-1)$. A little algebra will show that $E[s^2] = \sigma_N^2(1-\rho_N)$. Hence by Theorem 5 and Cramér's convergence theorem we have that

$$(k-1)F = \frac{(\underline{Z}'A_N\underline{Z})\ \sigma_N^2(1-\rho_N)}{s^2}$$

is asymptotically distributed as $\chi^2(k-1)$. And this is equivalent to saying that F is asymptotically F(k-1, N-k).

We note that making Assumption C is equivalent to saying that the random variable $\theta(x)$ should not equal a constant with probability one when X is a random observation from G.

Theorem 7. If Assumptions B1 and C hold, then as $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i \leq k$ the test of H_o based on the analysis of variance statistic F is consistent against any alternative for which

$$\Lambda = \sum_{i=1}^{k} p_i(\theta_i - \theta)^2 > 0.$$

Proof. From Theorem 3 it follows that under the stated conditions the random variable

$$\frac{\Sigma\ n_i(\bar{y}_i - \bar{\bar{y}})^2}{N\binom{N-1}{m}^2} = \frac{(k-1)F\ s^2}{N\binom{N-1}{m}^2} \longrightarrow \Lambda \qquad \text{———}.$$

in probability. Hence, using Theorem 5,

$$\frac{(k-1)F}{N} \longrightarrow \frac{\Lambda}{\zeta - \Sigma p_i\theta_i^2}$$

in probability, and thus, for sufficiently large N, F is certain to exceed the critical value F* corresponding to any α.

We remark that if for each N the alternative H_N is true, where under H_N

$$\theta_i = \theta + \frac{\delta_i}{\sqrt{N}} + 0(\frac{1}{\sqrt{N}}),$$

then under reasonable regularity conditions the asymptotic distribution of F will be noncentral F with (k-1, N-k) degrees of freedom and noncentrality parameter

$$\Delta = \frac{\Sigma p_i(\delta_i - \delta)^2}{\zeta - \theta^2}$$

where $\delta = \Sigma p_i \delta_i$. However, we shall not attempt to present a general theorem.

4. SCORES RELATED TO MODIFIED U-STATISTICS. Suppose the function ϕ of the preceding section depends on an unknown parameter τ (which may be vector-valued), that is

$$\phi(x_0, x_1, \ldots, x_m) = \phi(\tau;\ x_0, x_1, \ldots, x_m).$$

Then we may consider using an analysis of variance test based on modified scores

$$z_{ij} = \Sigma_{ij}\ \phi(t_N;\ X_{ij},\ [(m-1)\ X's])$$

where τ has been estimated by a statistic t_N which is a symmetric function of all the N observations. We shall require the following

Assumption D: If $X_0, X_1, \ldots, X_N$ are independent observations from G, then

$$E[\phi(\tau;\ X_0, X_1, \ldots, X_m) - \phi(t_N;\ X_0, X_1, \ldots, X_m)]^2 = 0(\frac{1}{N}).$$

Then we may state

Theorem 8. If the hypothesis is true, and if Assumptions B2, C, and D all hold, then as $N \longrightarrow \infty$ and $n_i/N \longrightarrow p_i > 0$ for $1 \leq i \leq k$ the analysis of variance statistic F based on the modified scores has asymptotically the F-distribution with (k-1, N-k) degrees of freedom.

Proof. Define scores

$$y_{ij} = \Sigma_{ij}\, \phi(\tau;\, X_{ij},\, [(m-1)\ X's])$$

as in section 3. Then by Theorem 5 we have that

$$\frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \binom{N-1}{m}^{-2} \longrightarrow \zeta - \theta^2$$

in probability as $N \longrightarrow \infty$. And

$$y_{ij} - z_{ij} = \Sigma_{ij}\{\phi(\tau; X_{ij}, [(m-1)X's]) - \phi(t_N; X_{ij}, [(m-1)X's])\},$$

so by Assumption D

$$E[y_{ij} - z_{ij}]^2 = \binom{N-1}{m}^2 0(\tfrac{1}{N}),$$

and hence

$$\frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}) - (z_{ij} - \bar{z})]^2 \binom{N-1}{m}^{-2} \leq \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - z_{ij})^2 \binom{N-1}{m}^{-2} \longrightarrow 0$$

in probability. But then by Sen's lemma we have

$$\frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z})^2 \binom{N-1}{m}^{-2} \longrightarrow \zeta - \theta^2$$

in probability, where $\zeta - \theta^2 > 0$ by Assumption C. Now, using Assumption B2 we find that

$$\max_{i,j} \left| z_{ij} - \bar{z} \right| < \binom{N-1}{m} c$$

and thus we see immediately that the modified scores satisfy Assumption A. Finally, since under the hypothesis these scores are interchangeable random variables, the present theorem follows from Theorem 2.

5. EXAMPLES

Example 1. (Kruskal-Wallis Test). Let X_{ij} be univariate, and define the corresponding score as

$$y_{ij} = R_{ij} - 1$$

where R_{ij} is the rank of X_{ij} among all the N X's, average ranks being used in case of ties. Then

$$y_{ij} = \Sigma_{ij}\, \phi(X_{ij}, X) = \Sigma_{ij}\, \psi(X_{ij} - X)$$

where

$$\psi(z) = \begin{cases} 0 & z < 0 \\ \frac{1}{2} & z = 0 \\ 1 & z > 0 \end{cases} .$$

The analysis of variance statistic turns out to be

$$F = \frac{(N-k)H}{(k-1)(N-1-H)} ,$$

where

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{\left[\sum_{j=1}^{n_i} R_{ij}\right]^2}{n_i} - 3(N+1)$$

is the familiar Kruskal-Wallis statistic, usually treated as $\chi^2(k-1)$. (It may be noted that F is one of the alternative expressions considered by Wallace [8], who showed that the approximation to the conditional distribution afforded by treating it as F(k-1, N-k) is better than the usual one is, at least for certain very small sample sizes, although not so good as various other more complicated approximations.) In the standard method it is necessary to make special adjustments for ties; however, in our approach equivalent adjustments are made automatically as a part of the computation. Finally, we have

$$\theta_i = P\{X_i > X\} + \tfrac{1}{2}P\{X_i = X\}, \quad \theta = \tfrac{1}{2}$$

where X_i and X are observations randomly chosen from G_i and G respectively.

Example 2. (Rank Analysis of Covariance). Let $X_{ij} = (X_{ij}^{(0)}, X_{ij}^{(1)}, \ldots, X_{ij}^{(r)})$ where $X^{(0)}$ is the response and $X^{(1)}, \ldots, X^{(p)}$ are concomitant variables. Let $R_{ij}^{(u)}$ be the rank of $X_{ij}^{(u)}$ among all the N $X_{ij}^{(u)}$'s for $0 \leq u \leq p$. Then the rank analysis of covariance procedure proposed by the author [6] consists of performing an ordinary analysis of variance on scores

$$y_{ij} = [R_{ij}^{(0)} - \frac{N+1}{2}] - \sum_{u=1}^{p} c_u [R_{ij}^{(u)} - \frac{N+1}{2}]$$

where the c's are any suitable set of constants. These scores can be obtained as $y_{ij} = \Sigma_{ij}\, \phi(X_{ij}, X)$ where

$$\phi(X_{ij},X) = [\psi(X_{ij}^{(0)} - X^{(0)}) - \tfrac{1}{2}] - \sum_{u=1}^{p} c_u[\psi(X_{ij}^{(u)} - X^{(u)}) - \tfrac{1}{2}].$$

No matter what c's have been selected, we have

$$\theta_i = \tfrac{1}{2}P\{X_i^{(0)} > X^{(0)}\} - \tfrac{1}{2}P\{X_i^{(0)} < X^{(0)}\}\ ,\ \theta = 0$$

where $X_i^{(0)}$ and $X^{(0)}$ are observations randomly chosen from $G_i(x^{(0)}, \infty,\ldots,\infty)$ and $G(x^{(0)},\infty,\ldots,\infty)$ respectively. If $c_1=\ldots=c_p=0$ then the test is equivalent to the modified Kruskal-Wallis test of Example 1.

In [6] it is shown that, from the standpoint of asymptotic relative efficiency, the optimal choice for $\underline{c}'= (c_1,\ldots,c_p)$ is $\underline{c}' = \underline{\gamma}'$ where $\underline{\gamma}'$ minimizes $[\underline{\gamma}'\Lambda\,\underline{\gamma} - 2\underline{\gamma}'\underline{\eta}]$, $\underline{\eta}$ is the vector of covariances between $R_{ij}^{(0)}$ and $(R_{ij}^{(1)},\ldots,R_{ij}^{(p)})$, and Λ is the variance matrix of $(R_{ij}^{(1)},\ldots,R_{ij}^{(p)})$. Suppose we choose $\underline{c}'$ to minimize $[\underline{c}'L\underline{c} - 2\underline{c}'\underline{r}]$ where $\underline{r}$ and L are the sample estimates of η and Λ. Then we are using modified scores as in Section 4. Assumption B2 is satisfied since $|\phi| < 1 + \Sigma\,|\gamma_u|$, Assumption C is satisfied unless the multiple correlation between $R_{ij}^{(0)}$ and $(R_{ij}^{(1)},\ldots,R_{ij}^{(p)})$ is unity, and Assumption D since the c's are continuous functions of U-statistics. Thus the test with estimated optimal c's is also asymptotically valid.

Example 3. (Mood's Squared-Rank Test). Let X_{ij} be univariate and define the corresponding score as

$$y_{ij} = (R_{ij} - 1)(N - R_{ij}),$$

where R_{ij} is again the rank of X_{ij} among all the N x's, provided that X_{ij} is not tied with any other X; let the score attached to each member of a group of tied observations be the mean of the various scores they would receive if the ties were

broken arbitrarily. Then

$$y_{ij} = \Sigma_{ij}\, \phi(X_{ij}, [2\ X's])$$

where

$$\phi(x_0, x_1, x_2) = \begin{cases} 1 & \text{if } x_1 < x_0 < x_2 \text{ or } x_2 < x_0 < x_1 \\ \frac{1}{2} & \text{if } x_0 = x_1 \neq x_2 \text{ or } x_0 = x_2 \neq x_1 \\ 1/3 & \text{if } x_0 = x_1 = x_2 \\ 0 & \text{otherwise.} \end{cases}$$

If G is continuous then θ_i is the probability that an observation randomly chosen from G_i will lie between two observations randomly chosen from G; if the hypothesis is true then $\theta_i = \theta = 1/3$.

Now, the squared-rank test proposed by Mood [4] for the case where k = 2 and there are no ties is based on treating the statistic

$$W = \sum_{j=1}^{n_1} (R_{1j} - \tfrac{N+1}{2})^2$$

as normally distributed with mean $E[W] = n_1(N+1)(N+2)/12$ and variance $V[W] = n_1 n_2 (N+1)(N+2)(N-2)/180$; equivalently,

$$W^* = \frac{180\{W - E[W]\}^2}{n_1 n_2 (N+1)(N+2)(N-2)}$$

may be treated as χ^2 with 1 degree of freedom. After some algebra it can be shown that for this special case the analysis of variance statistic based on the scores is

$$F(1,N-2) = \frac{(N-2)W^*}{(N-1-W^*)} \quad .$$

Note, however, that as in Example 1 the extension to $k > 2$ is made immediate through the general approach, and again that no special adjustment for ties is required.

Example 4. (2-by-k Contingency Tables) a) Let the observations X_{ij} be elements of any space Ω and let Ω_0 be any prespecified subspace of Ω. Then the data may be summarized as a contingency table with k rows, corresponding to the k samples, and 2 columns, corresponding to observations in or not in Ω_0; let m_i be the number of observations from the i-th sample which fall in Ω_0, and let $M = \Sigma m_i$. The standard test is based on treating the statistic

$$X^2 = \frac{1}{M(N-M)} \sum_{i=1}^{k} \frac{(m_i N - n_i M)^2}{n_i}$$

as $\chi^2(k-1)$. Let $\phi(X_{ij},X) = 1$ or 0 according as $X_{ij} \in \Omega_0$ or $\notin \Omega_0$; then the score $y_{ij} = \Sigma_{ij}\ \phi(X_{ij},X) = (N-1)$ or 0 according as $X_{ij} \in \Omega_0$ or $\notin \Omega_0$, and the analysis of variance statistic turns out to be

$$F = \frac{(N-k)X^2}{(k-1)(N-X^2)} \quad .$$

The parameter θ_i is then just the probability that a random observation from the i-th population will fall in Ω_0.

b) Now suppose that the subspace Ω_0 is defined in terms of some unknown parameter. As a definite instance, consider the extension of the median test. Here Ω is the real line and Ω_0 is the portion to the right of μ, the population

median. We use modified scores

$$z_{ij} = \Sigma_{ij}\ \phi(\tilde{\mu};\ X_{ij},\ X)$$

where $\phi(\tilde{\mu};\ X_{ij},\ X)$ is 1 or 0 according as X_{ij} does or does not exceed $\tilde{\mu}$, the sample median. That this test is asymptotically valid may be shown by the method of Section 4, although in this case it is easier to apply Theorem 2 directly. Other instances may be adduced.

REFERENCES

[1] BHAPKAR, V. P. (1961). A nonparametric test for the problem of several samples. Ann. Math. Stat. 32 1108-17.

[2] CHERNOFF, H., and TEICHER, H. (1958). A central limit theorem for sums of interchangeable random variables. Ann. Math. Stat. 29 118-30.

[3] CRAMÉR, H. (1951). Mathematical Methods of Statistics Princeton University Press.

[4] MOOD, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Ann. Math. Stat. 25 514-22.

[5] PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any populations, III. The analysis of variance test. Biometrika 29 322-335.

[6] QUADE, D. (1965). Rank Analysis of Covariance. Submitted to J. Amer. Statist. Assoc.

[7] SEN, P. K. (1960). On Some Convergence Properties of U-statistics. Calcutta Statist. Assoc. Bull. 10 1-18.

[8] WALLACE, D. L. (1959). Simplified beta-approximations to the Kruskal-Wallis H test. J. Amer. Statist. Assoc. 54 225-30.