

HYPOTHESIS TESTING IN ANOVA UNDER MULTINOMIAL SAMPLING

By P. G. MOSCHOPOULOS
The University of Texas at Dallas
and
M. L. DAVIDSON
Xerox Corporation, Rochester

SUMMARY. This paper is concerned with hypothesis testing in multiway classification models with fixed effects when the cell sizes follow a multinomial distribution with unknown probability parameters. In particular it is shown that under the multinomial distribution, the sum of squares associated with the hypothesis has a limiting quadratic form distribution that is not always a chi-square when the null hypothesis is true. As a consequence, the conditional (on fixed cell sizes) F -ratios may fail to maintain the correct type I error. Alternative F -ratios are proposed and evaluated under the null hypothesis.

1. INTRODUCTION

There are situations in the analysis of variance in which the cell sizes are random variables instead of being fixed and part of the design of the underlying experiment. This research is motivated from applications in which a simple random sample of fixed size n is drawn from a population and then it is divided into its various subpopulations (cells) according to sample characteristics. Assuming a fixed number m of cells, the simple random sampling scheme results in cell sizes n_1, \dots, n_m that are random variables and follow the multinomial distribution $\text{Mult}(m, n, \pi_1, \dots, \pi_m)$ where π_i is the probability of obtaining a unit from the i -th cell. Practical examples in this regard are available in Gallassi, Frierson and Sharer (1981) and Groat and Neal (1967).

In this paper we study the effect of the randomness of the cell sizes on hypothesis testing under the following assumptions :

- (1) The multinomial probabilities are unknown and positive.
- (2) The random cell sizes are non-correlated with the y -observations on the characteristic of interest.
- (3) Conditional on the cell sizes, the observations \mathbf{y} ($n \times 1$) satisfy a multiway classification model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $X(n \times k)$ is the usual deficient-

AMS (1980) subject classification : Primary : 62J10 ; Secondary : 62G10.

Key words and phrases : Random cell sizes, multinomial distribution, hypothesis testing.

rank matrix of zeros and ones, and $\beta(k \times 1)$ is the vector of the effects (main effects and possibly interaction effects), while $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$.

When the cell sizes are fixed, a large class of hypotheses about the effects is obtained from the expectation of sums of squares of the form $R(\cdot|\cdot)$, see Searle, (1971). These sums of squares are quadratic forms in the sample cell means with matrix of the quadratic forms depending on the cell sizes. As a result, many of the standard ANOVA hypotheses depend on the cell sizes, which is inappropriate when the latter are random variables as in the present setting. In this paper we show that when these hypotheses are reformulated to depend on true cell proportions rather than the sample proportions, the associated sums of squares, and hence the standard F -ratios, have different null distributions.

2. A GENERAL CLASS OF HYPOTHESES

Let $\mu = (\mu_1, \dots, \mu_m)'$ be the vector of the population cell means and consider the reparametrization $\mu = D\beta$ where $D(m \times k)$ is the matrix corresponding to X with one observation per cell. The matrix D , which is relevant to the type of model and is free of the cell sizes may be called the *model matrix*. Its relation to X is $X = GD$ where G is a "membership" matrix with elements $g_{ji} = 1$ if the j -th observation is in the i -th cell and 0 otherwise $i = 1, \dots, m, j = 1, \dots, n_i$. The class of hypotheses considered in this paper is motivated from the following orthogonalizations implied by Searle's $R(\cdot|\cdot)$ notation, see Moschopoulos and Davidson (1982).

Partition $X = [X_1|X_2|X_3]$ and suppose that β_1 and β_2 are the two sub-vectors of β corresponding to X_1 and X_2 . Then, the reduction in the sum of squares for "fitting β_2 over and above β_1 ", in Searle's terminology, is $R(\beta_2|\beta_1)$ where

$$R(\beta_2|\beta_1) = y'X_h(X_h'X_h)^-X_h'y \quad \dots (2.1)$$

where

$$X_h = [I_n - X_1(X_1'X_1)^-X_1'] [X_1|X_2] \quad \dots (2.2)$$

and A^- denotes a generalized inverse of A satisfying $AA^-A = A$, see Rao and Mitra (1971).

Now, on using the fact that the partition of X induces a partition of $D = [D_1|D_2|D_3]$ via the relations $X_i = GD_i, i = 1, 2, 3$, and the relation $G'G = N = \text{diag}(n_1, \dots, n_m)$, we get $X_h = GD_h$ where

$$D_h = [I_m - D_1(D_1'ND_1)^-D_1'N][D_1|D_2]. \quad \dots (2.3)$$

Further, if we let $\bar{y}_i = \sum_j y_{ij}/n_i$ and $\bar{y} = (\bar{y}_1, \dots, \bar{y}_m)'$, and use the relation $G'y = N\bar{y}$, we obtain

$$R(\beta_2|\beta_1) = \bar{y}'ND_h(D_h'ND_h)^-D_h'N\bar{y}. \quad \dots (2.4)$$

When the cell sizes are fixed, the hypothesis tested by the above sum of squares is obtained from its conditional expectation and it has the form

$$H_c : D'_h N \mu = 0. \quad \dots (2.5)$$

It follows that, if some cells are empty, or as in this paper, if the cell sizes are random variables, then the conditional hypothesis H_c may become undefined.

The above deficiency of the $R(\cdot|\cdot)$ notation suggests that, instead of orthogonalizing X_1 and X_2 with Euclidean inner product as in (2.2), we should orthogonalize D_1 and D_2 with inner product matrix relating to the "importance" of the cells, for example the identity matrix treating all the cells equally, or any other weight matrix. Authors who use general weights include Sheffé (1959) and Seber (1977). The new BMDP4V program permits arbitrary cell weights.

The hypothesis (2.5) is now reformulated. Let $\Pi = \text{diag}(\pi_1, \dots, \pi_m)$, and define the inner product on the column space $L(D)$ of D as $x'_1 \Pi x_2$ for $x_1, x_2 \in L(D)$. Define the projectors R_1 on $L(D_1)$ and R_{12} on $L(D_{12})$ where $D_{12} = (D_1 | D_2)$, as

$$R_1 = D_1(D'_1 \Pi D_1)^{-1} D'_1 \Pi, \quad R_{12} = D_{12}(D'_{12} \Pi D_{12})^{-1} D'_{12} \Pi \quad \dots (2.6)$$

which are orthogonal with respect to the above inner product, see Rao and Mitra (1971, p. 112). Then, the unconditional (n_i random) hypothesis is

$$H : \Delta'_h \Pi \mu = 0 \quad \dots (2.7)$$

where

$$\Delta_h = (I_m - R_1) D_{12}. \quad \dots (2.8)$$

Example : Two-way classification without interaction. Here $m = rc$, $\mu_{ij} = \mu + \alpha_i + \gamma_j$, $i = 1, \dots, r$, $j = 1, \dots, c$; $\beta = (\mu, \alpha', \gamma)'$; the matrix D consists of the parts $1_r \times 1_c$, $I_r \times 1_c$ and $1_r \times I_c$ where x denotes the cross-product multiplication of matrices. We choose D_1, D_2 and D_3 as follows.

(1) $D_1 = 0$, $D_2 = 1_r \times 1_c$, $D_3 = [I_r \times 1_c | 1_r \times I_c]$. With this choice, if we order the elements of π and μ by rows, e.g. $\mu = (\mu_{11}, \dots, \mu_{1c}, \mu_{21}, \dots, \mu_{rc})'$, H reduces to the hypothesis that the grand mean is zero, i.e.

$$H_1 : \sum_i \sum_j \pi_{ij} \mu_{ij} = 0.$$

(2) $D_1 = 1_r \times 1_c$, $D_2 = I_r \times 1_c$, $D_3 = 1_r \times I_c$. In this case H reduces to the hypothesis of equality of row means H_2 where

$$H_2 : \sum_j \frac{\pi_{ij}}{\pi_i} \mu_{ij} \quad \text{equal for all } i, \text{ where } \pi_i = \sum_j \pi_{ij}.$$

Note that in this case H_c yields the hypothesis H'_2 (see Yates, 1934), where

$$H'_2 : \sum_j \frac{n_{ij}}{n_{i.}} \mu_{ij} \quad \text{equal for all } i, \text{ where } n_{i.} = \sum_j n_{ij}$$

or, in terms of the α_i 's and γ_j 's the hypothesis, (See Searle, 1971, p. 275)

$$H'_2 : \alpha_i + \sum_j \frac{n_{ij}}{n_{i.}} \gamma_j \quad \text{equal for all } i.$$

$$(3) \quad D_1 = [1_r \times 1_c | I_r \times 1_c], \quad D_2 = [1_r \times I_c], \quad D_3 = 0.$$

Then, both H and H_c yield the hypothesis better known in terms of the γ_j 's namely,

$$H_3 : \gamma_1 = \gamma_2 = \dots = \gamma_c.$$

The form (2.7) covers many hypotheses commonly tested in multiway classification models. More examples, geometrical aspects, and formulation with an arbitrary inner product matrix are found in Moschopoulos (1981) and Moschopoulos and Davidson (1982).

Now, while the distribution of (2.4) under H_c is that of σ^2 times a chi-square random variable $\chi^2(r_h)$ with $r_h = \text{Rank}(D_h)$ degrees of freedom, the distribution under the proper hypothesis H is complicated and it depends on nuisance parameters, i.e. the π_i 's. However, the asymptotic ($n \rightarrow \infty$, $n_i/n \rightarrow \pi_i$) distribution can be obtained and it is given in Theorem 1 of Section 4. In the next section we give some necessary lemmas.

3. ASYMPTOTIC EXPANSIONS

Let $p_i = n_i/n$ and $\mathbf{P} = \text{diag}(p_1, \dots, p_m)$. The following notation is needed in the rest of the paper: Let

$$\begin{aligned} Q_p &= I - D_1(D_1'PD_1)^{-1}D_1'P, & Q_\pi &= I - D_1(D_1'\Pi D_1)^{-1}D_1'\Pi, \\ M_p &= D_{12}(D_h'PD_h)^{-1}D_{12}', & M_\pi &= D_{12}(\Delta_h'\Pi\Delta_h)^{-1}D_{12}', \\ V_p &= D_h(D_h'PD_h)^{-1}D_h', & V_\pi &= \Delta_h'(\Delta_h'\Pi\Delta_h)^{-1}\Delta_h'. \end{aligned}$$

With the above notation, the sum of squares corresponding to the hypothesis H , from (2.4) is obtained as SS_H where

$$SS_H = n\bar{y}'(PQ_p)M_p(Q_p'P)\bar{y}. \quad \dots \quad (3.1)$$

The limiting null distribution of SS_H as $n \rightarrow \infty$, $n_i/n \rightarrow \pi_i$, will be obtained with the help of the following four lemmas. Omitted proofs and details are available from the first author on request.

Lemma 1 : The matrix of the derivatives of PQ_p with respect to a fixed p_i is

$$\frac{\partial(PQ_p)}{\partial p_i} = Q_p' \left(\frac{\partial P}{\partial p_i} \right) Q_p.$$

Proof : Omitted.

Lemma 2 : Consider SS_H with \bar{y} replaced by μ , and let f_1 to be the following function of the m variables p_1, \dots, p_m ,

$$f_1(P) = n\mu'(PQ_p)M_p(Q_p'P)\mu.$$

Then, when H is true, the first non-zero term in the Taylor series expansion of $f_1(P)$ at $P = \Pi$ is

$$T_2 = n\mu'Q_p'(P-\Pi)V_\pi(P-\Pi)Q_p\mu$$

and $f_1(P) = T_2 + O_p(n^{-1})$.

Proof : When H is true, $\Delta_h'\Pi\mu = D_{12}'Q_p'\Pi\mu = 0$ and hence $f_1(\Pi) = 0$. The first order partial derivatives of $f_1(P)$ with respect to p_i are also zero at $p_i = \pi_i, i = 1, \dots, m$, when the null hypothesis is true, since

$$\frac{1}{n} \frac{\partial f}{\partial p_i} = 2\mu' \left[\frac{\partial(PQ_p)}{\partial p_i} \right] D_{12}(D_h'PD_h)^{-1}D_h'P\mu + \mu'PD_h \left[\frac{\partial}{\partial p_i} (D_h'PD_h)^{-1} \right] D_h'P\mu. \dots (3.2)$$

It also follows from (3.2) that the second order partial derivatives with respect to p_i and p_j are given as

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 f}{\partial p_i \partial p_j} &= 2\mu' \frac{\partial^2(PQ_p)}{\partial p_i \partial p_j} M_p Q_p' P \mu + 2\mu' \frac{\partial(PQ_p)}{\partial p_i} \frac{\partial M_p}{\partial p_j} Q_p' P \mu \\ &+ 2\mu' \frac{\partial(PQ_p)}{\partial p_i} M_p \frac{\partial(Q_p'P)}{\partial p_j} \mu + \mu' \frac{\partial(PQ_p)}{\partial p_j} \frac{\partial M_p}{\partial p_i} Q_p' P \mu \\ &+ \mu' PQ_p \frac{\partial^2 M_p}{\partial p_i \partial p_j} Q_p' P \mu + \mu' PQ_p \frac{\partial M_p}{\partial p_i} \frac{\partial(Q_p'P)}{\partial p_j} \mu. \end{aligned}$$

Hence, all but the third term are zero at $p_i = \pi_i, i = 1, \dots, m$, when evaluated under the null hypothesis. The third term, with the help of Lemma 1 becomes

$$2\mu' \frac{\partial(PQ_p)}{\partial p_i} M_p \frac{\partial(Q_p'P)}{\partial p_j} \mu = 2\mu' Q_p' \frac{\partial P}{\partial p_i} Q_p M_p Q_p' \frac{\partial P}{\partial p_j} Q_p \mu.$$

In evaluating this term at $\mathbf{P} = \mathbf{\Pi}$ under H , we replace \mathbf{Q}_p by \mathbf{Q}_π and \mathbf{M}_p by \mathbf{M}_π . Hence, if we let $t_k = p_k - \pi_k$, then the second term in the expansion of $f_1(\mathbf{P})$ is

$$n \sum_i \sum_j t_i t_j \left(\frac{\partial^2 f_1}{\partial p_i \partial p_j} \right)_{\text{at } p_i = \pi_i} = T_2.$$

Since the next term in the expansion is a sum of terms involving the products $(p_i - \pi_i)(p_j - \pi_j)(p_k - \pi_k)$, these terms require $n^{3/2}$ for convergence in distribution while the function has n . Hence,

$$f_1(\mathbf{P}) = T_2 + O_p(n^{-1/2}). \quad \square$$

Lemma 3: Let $f_2(\mathbf{P}) = n\boldsymbol{\mu}'\mathbf{P}\mathbf{Q}_p\mathbf{M}_p\mathbf{Q}_p'\mathbf{P}(\bar{\mathbf{y}} - \boldsymbol{\mu})$. Then under H we have

$$f_2(\mathbf{P}) = n\boldsymbol{\mu}'\mathbf{Q}_\pi(\mathbf{P} - \mathbf{\Pi})\mathbf{V}_\pi\mathbf{\Pi}(\bar{\mathbf{y}} - \boldsymbol{\mu}) + O_p(n^{-1/2}).$$

Proof: Omitted.

The last lemma concerns the limiting distribution of the vector of sample means $\bar{\mathbf{y}}$. Since $n_i = 0$ has positive probability, we give the following.

Definition: Let $U_i = n_i^{1/2}(\bar{y}_i - \mu_i)/\sigma$, $i = 1, \dots, m$ and

$$Z_{in} = \begin{cases} (n/n_i)^{1/2}U_i & \text{if } n_i > 0 \\ cZ_i & \text{if } n_i = 0 \end{cases}$$

where c is an arbitrary constant and $Z_i \sim N(0, 1)$.

Lemma 4: If $\min_i \pi_i > 0$, we have (\xrightarrow{d} for convergence in distribution):

$$(a) \quad Z_{in} \xrightarrow[n \rightarrow \infty]{d} > (1/\pi_i)^{1/2}Z_i$$

$$(b) \quad \mathbf{Z}_n = (Z_{1n}, \dots, Z_{mn})' \xrightarrow[n \rightarrow \infty]{d} \mathbf{\Pi}^{-1/2}\mathbf{Z}, \text{ where } \mathbf{Z} \sim N(0, \mathbf{I}).$$

Proof: Omitted.

4. THE LIMITING NULL DISTRIBUTION OF SS_H

Using the previous lemmas we can now derive the limiting distribution of SS_H when H is true, and discuss the effect of the multinomial distribution on the F -ratios used when the cell sizes are fixed.

Theorem 1: If the hypothesis $H: \Delta_i' \mathbf{\Pi} \boldsymbol{\mu} = \mathbf{0}$ is assumed true, the cell sizes are uncorrelated with the y -observations, and $\min_i \pi_i > 0$, then

$$SS_H \xrightarrow[n \rightarrow \infty]{d} (\sigma \mathbf{\Pi} \mathbf{Y} + \boldsymbol{\Lambda} \mathbf{X})' \mathbf{V}_\pi (\sigma \mathbf{\Pi} \mathbf{Y} + \boldsymbol{\Lambda} \mathbf{X})$$

where $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Pi} - \pi\pi')$, $\mathbf{Y} = \mathbf{\Pi}^{-1}\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Pi}^{-1})$,

\mathbf{X} and \mathbf{Y} are independent, $\pi = (\pi_1, \dots, \pi_m)'$, and $\mathbf{\Lambda}$ is diagonal with elements those of $\mathbf{Q}_\pi\mu$.

Proof: Using the notation of the previous section it is easy to see that SS_H in (3.1) is decomposed as follows :

$$SS_H = n(\bar{\mathbf{y}} - \mu)' \mathbf{P} \mathbf{V}_p \mathbf{P} (\bar{\mathbf{y}} - \mu) + 2n\mu' \mathbf{P} \mathbf{V}_p \mathbf{P} (\bar{\mathbf{y}} - \mu) + n\mu' \mathbf{P} \mathbf{V}_p \mathbf{P} \mu. \quad \dots (4.1)$$

Since $\mathbf{P} \mathbf{V}_p \mathbf{P} \xrightarrow{Pr} \mathbf{\Pi} \mathbf{V}_\pi \mathbf{\Pi}$ (convergence in probability), the first term of the right hand side of (4.1) is

$$n(\bar{\mathbf{y}} - \mu)' \mathbf{P} \mathbf{V}_p \mathbf{P} (\bar{\mathbf{y}} - \mu) = \sigma^2 \mathbf{Z}'_n \mathbf{\Pi} \mathbf{V}_\pi \mathbf{\Pi} \mathbf{Z}_n + o_p(1). \quad \dots (4.2)$$

The second term in the r.h.s. of (4.1) is $2f_2(P)$ and hence by Lemma 3

$$\begin{aligned} 2n\mu' \mathbf{P} \mathbf{V}_p \mathbf{P} (\bar{\mathbf{y}} - \mu) &= 2n\mu' \mathbf{Q}'_\pi (\mathbf{P} - \mathbf{\Pi}) \mathbf{V}_\pi \mathbf{\Pi} (\bar{\mathbf{y}} - \mu) + O_p(n^{-1}) \\ &= 2n(\mathbf{p} - \pi)' \mathbf{\Delta} \mathbf{V}_\pi \mathbf{\Pi} (\bar{\mathbf{y}} - \mu) + O_p(n^{-1}). \quad \dots (4.3) \end{aligned}$$

Finally, the third term is $f_1(P)$ and by Lemma 2 is expressed as

$$n\mu' \mathbf{P} \mathbf{V}_p \mathbf{P} \mu = n(\mathbf{p} - \pi)' \mathbf{\Delta} \mathbf{V}_\pi \mathbf{\Lambda} (\mathbf{p} - \pi) + O_p(n^{-1}). \quad \dots (4.4)$$

On substituting (4.2), (4.3) and (4.4) into (4.1), we have :

$$SS_H = \sigma^2 \mathbf{Z}'_n \mathbf{\Pi} \mathbf{V}_\pi \mathbf{\Pi} \mathbf{Z}_n + 2\sigma \mathbf{X}'_n \mathbf{\Delta} \mathbf{V}_\pi \mathbf{\Pi} \mathbf{Z}_n + \mathbf{X}'_n \mathbf{\Delta} \mathbf{V}_\pi \mathbf{\Lambda} \mathbf{X}_n + o_p(1) + O_p(n^{-1})$$

where $\mathbf{X}_n = n^{1/2}(\mathbf{p} - \pi)$. Since \mathbf{Z}_n and \mathbf{X}_n are uncorrelated and converge in distribution to \mathbf{X} and \mathbf{Y} respectively, the theorem is proved. \square

Corollary 1 : If H is true and $(\mathbf{I} - \mathbf{R}_{12})\mu = \mathbf{0}$, then

$$SS_H \xrightarrow[n \rightarrow \infty]{d} \sigma^2 \chi^2(r_h), \quad r_h = \text{Rank}(\mathbf{\Delta}_h).$$

Proof: When H is true, $\mathbf{R}_{12}\mu = \mathbf{R}_1\mu$ and hence by the given condition $(\mathbf{I} - \mathbf{R}_{12})\mu = \mathbf{0}$ we have $\mathbf{Q}_\pi\mu = \mathbf{0}$, hence $\mathbf{\Lambda} = \mathbf{0}$. \square

It follows from the above corollary, that the condition $(\mathbf{I} - \mathbf{R}_{12})\mu = \mathbf{0}$ is sufficient to guarantee that the effect of the multinomial on testing the hypothesis H is only that of truncation, i.e. for "sufficiently large n ", one may use the usual conditional F -ratio to test H . It can be seen that this is true for the hypothesis H_3 in Section 2. On the other hand, if the condition $(\mathbf{I} - \mathbf{R}_{12})\mu = \mathbf{0}$ is not satisfied, then the conditional F -ratio has different distribution under H . As it is seen from Theorem 1, the limiting distribution of SS_H depends on the π_i 's. In this case, the conditional F -ratio that is used to test H_c is not appropriate to test H . For example, the conditional F -ratio, see Searle (1971, p. 275), that is used for H'_2 or H''_2 , is not appropriate for H_2 . Numerical evaluations in Moschopoulos (1981) show that the conditional test fails to maintain the stated level of significance.

The failure of the conditional test for testing the hypothesis H when the condition $(\mathbf{I}-\mathbf{R}_{12})\boldsymbol{\mu} = \mathbf{0}$ is not satisfied is highlighted by Theorem 2, which gives the mean and variance of SS_H under H .

Theorem 2: *If H is assumed true, then*

$$(a) \quad E(SS_H) = \sigma^2 r_h + tr(\mathbf{V}_\pi \boldsymbol{\Sigma}) + O(n^{-1})$$

$$(b) \quad var(SS_H) = 2r_h \sigma^4 + 4\sigma^2 tr(\mathbf{V}_\pi \boldsymbol{\Sigma}) + 2tr[(\mathbf{V}_\pi \boldsymbol{\Sigma})^2] + O(n^{-1})$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}')\boldsymbol{\Lambda}$.

Proof: The proof is obtained by first conditioning on the n_i 's and then using Lemma 2. It should be noted that the above expressions agree to those obtained from the limiting quadratic form in Theorem 1.

It follows from Theorem 2 that if $(\mathbf{I}-\mathbf{R}_{12})\boldsymbol{\mu} = \mathbf{0}$, then $\boldsymbol{\Lambda} = \mathbf{0}$ and the moments reduce to the well known conditional moments of SS_H . Otherwise, the moments are inflated from variation due to the multinomial. To account for this extra variation in SS_H we propose a test as follows:

Let ε_h be the asymptotic null expectation of SS_H and let $\hat{\varepsilon}_h$ be the estimator of SS_H obtained by estimating σ^2 by its unbiased estimator $\hat{\sigma}^2$, π_i by p_i and μ_i by \bar{y}_i . Then we define our test as

$$F = \frac{SS_H}{\hat{\varepsilon}_h} \quad \dots \quad (4.5)$$

This test has the property that when H is true, the asymptotic expectations of numerator and denominator are equal, see Brown and Forsythe (1974). Moreover, it reduces to the conditional test if $(\mathbf{I}-\mathbf{R}_{12})\boldsymbol{\mu} = \mathbf{0}$.

Next, following Box (1954), we approximate both, numerator and denominator by chi-squares of the form $a\chi^2(b)$ where a and b are determined from the first two moments. This, leads to an F -like approximation for F under the null hypothesis, i.e.

$$F \sim F(df_1, df_2)$$

where

$$df_1 = \frac{2\hat{\varepsilon}_h^2}{S_h^2}, \quad df_2 = \frac{\hat{\varepsilon}_h^2}{r_h^2 \hat{\sigma}^4} (n - \text{Rank}(\mathbf{D}))$$

and S_h^2 is an estimator of the $O(1)$ term in $var(SS_H)$ under H , given by (b) of Theorem 2.

The conditional test has only $\hat{\sigma}^2$ in the denominator and therefore, it does not take into account the extra variation due to the multinomial. This variation may be too serious to ignore. In the case of H_2 , the sum of squares is $SS_H = R(\alpha|\mu)$ in Searle's notation. In this case, it follows from (a) of Theorem 2 that under H_2 we have :

$$E(SS_H) = \sigma^2 r_h + \sum_i \sum_j w_{ij} (\mu_{ij} - \mu_{i.})^2 + O(n^{-1})$$

where $w_{ij} = \pi_{ij}(1 - \pi_{i.})/\pi_{i.}$ and $\mu_{i.} = \sum_j \frac{\pi_{ij}}{\pi_{i.}} \mu_{ij}$.

Hence, the mean of SS_H is now inflated by a weighted average of squared deviations of the cell means from the corresponding row means. This, explains the failure of the conditional test for testing H_2 as mentioned earlier.

The proposed test has been evaluated numerically in a number of models, and for a number of hypotheses implied by H for various choices of D_1 , and D_2 , see Moschopoulos (1981). It follows from these evaluations that the test F with its null approximation as above, provides a practical solution, since it maintains a level of significance sufficiently close to the nominal level.

REFERENCES

- BOX, G. E. (1954): Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Stat.*, 25, 290-302.
- BROWN, M. D., FORSYTHE, A. B. (1974): The small sample behavior of some statistics which test equality of several means. *Technometrics*, 16(1), 129-132.
- GALASSI, J. P., FRIERSON, H. T., JR., SHARER R. (1981): Behavior of high, moderate, and test anxious students during an actual test situation. *Journal of Consulting and Clinical Psychology*, 49, No. 1, 51-62.
- GROAT, H. T., NEAL, A. G. (1967): Social psychological correlates of urban fertility. *Amer. Sociol. Rev.*, 945-959.
- MOSCHOPOULOS, P. G. (1981): Analysis of variance with random sample sizes. Ph.D. thesis. Dept. of Statistics, The University of Rochester.
- MOSCHOPOULOS, P. G., DAVIDSON, M. L. (1982): Defining ANOVA hypotheses by orthogonalization. *Technical report* No. 110, University of Texas at Dallas.
- RAO, C. R., and MITRA, S. K. (1971): *Generalized Inverse of Matrices and Its Application*, New York: Wiley.
- SCHEFFE, H. (1959): *The Analysis of Variance*, New York: Wiley.
- SEARLE, S. R. (1971): *Linear Models*, New York: Wiley.
- SEBER, G. A. F. (1977): *Linear Regression Analysis*, New York, Wiley.
- YATES, F. (1934): The analysis of multiple classifications with unequal numbers in the different classes. *JASA*: 51-56.

Paper received : April, 1984.

Revised : May, 1985.