AN EXTENSION OF BOX'S RESULTS ON THE USE OF THE F DISTRIBUTION IN MULTIVARIATE ANALYSIS

By Seymour Geisser and Samuel W. Greenhouse

National Institute of Mental Health

1. Introduction and summary. The mixed model in a 2-way analysis of variance is characterized by a fixed classification, e.g., treatments, and a random classification, e.g., plots or individuals. If we consider k different treatments each applied to everyone of n individuals, and assume the usual analysis of variance assumptions of uncorrelated errors, equal variances and normality, an appropriate analysis for the set of nk observations x_{ij}, $i = 1, 2, \cdots n$, $j = 1, 2 \cdots k$, is

Source	*D.F.*	*F*
Treatments	$k - 1$	$\dfrac{\text{mean square for treatments}}{\text{mean square for } T \times I}$
Individuals	$n - 1$	
Treat. × Ind.	$(k - 1)(n - 1)$	

where the F ratio under the null hypothesis has the F distribution with $(k - 1)$ and $(k - 1)(n - 1)$ degrees of freedom. As is well known, if we extend the situation so that the errors have equal correlations instead of being uncorrelated, the F ratio has the same distribution. Under the null hypothesis, the numerator estimates the same quantity as the denominator, namely, $(1 - \rho)\sigma^2$, where ρ is the constant correlation coefficient among the treatments. This case can also be considered as a sampling of n vectors (individuals) from a k-variate normal population with variance-covariance matrix

$$V = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & & & \vdots \\ \vdots & & & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}.$$

If we consider this type of formulation and suppose the k treatment errors to have a multivariate normal distribution with unknown variance-covariance matrix (the same for each individual), then the usual test described above is valid for $k = 2$. For $k > 2$, and $n \geqq k$, Hotelling's T^2 is the appropriate test for the homogeneity of the treatment means. However, the working statistician is sometimes confronted with the case where $k > n$, or he does not have the adequate means for computing large order inverse matrices and would therefore like to use the original test ratio which in general does not have the requisite F distribution. Box [1] and [2] has given an approximate distribution of the test ratio to be $F[(k - 1)\epsilon, (k - 1)(n - 1)\epsilon]$ where ϵ is a function of the popula-

Received May 27, 1957; revised January 29, 1958.

tion variances and covariances and may further be approximated by the sample variances and covariances. We show in Section 3 that $\epsilon \geqq (k-1)^{-1}$, and therefore a conservative test would be $F(1, n-1)$.

Box referred only to one group of n individuals. We shall extend his results to a frequently occurring case, namely, the analysis of g groups where the αth group has n_α individuals, $\alpha = 1, 2, \cdots g$, and $\sum_{\alpha=1}^{g} n_\alpha = N$. We will show that the treatment mean square and the treatment $\times$ group interaction can be tested in the same approximate fashion by using the Box procedure.

2. Extension to g groups. Consider a mixed model, k treatments, each applied to N individuals where the N individuals are subdivided into g groups so that we have chosen a random sample of n_α individuals from the αth group. The observations are $x_{ij\alpha}$, $i = 1, \cdots n_\alpha$, $j = 1, \cdots, k$, $\alpha = 1, \cdots, g$ and

$$\sum_{\alpha=1}^{g} n_\alpha = N.$$

Therefore we get the following array for the αth group

$$\begin{matrix} x_{11\alpha} & \cdots & x_{1k\alpha} \\ \vdots & & \vdots \\ x_{n_\alpha 1\alpha} & \cdots & x_{n_\alpha k\alpha} \end{matrix}$$

We may consider the joint distribution of the $x_{ij\alpha}$ to be represented by the vector variable

$$x' = (x_{111} \cdots x_{1k1} \cdots x_{n_1 11} \cdots x_{n_1 k1} \cdots x_{11g} \cdots x_{1kg} \cdots x_{n_g 1g} \cdots x_{n_g kg})$$

where $Ex' = \mu'$ and x' has a kN multivariate normal distribution with variance-covariance matrix

$$\Lambda = \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & V_g \end{pmatrix}$$

and

$$V_\alpha = \begin{pmatrix} V & 0 & \cdots & 0 \\ 0 & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & V \end{pmatrix},$$

where V is a matrix of order k, V_α is of order kn_α and Λ is of order kN.

Let $Ex_{ij\alpha} = \mu_{j\alpha}$ and

$$N^{-1} \sum_{\alpha=1}^{g} n_\alpha \mu_{j\alpha} = \mu_{j.} \text{ is the mean of the } j\text{th treatment,}$$

$$k^{-1} \sum_{j=1}^{k} \mu_{j\alpha} = \mu_{.\alpha} \text{ is the mean of the } \alpha\text{th group, and}$$

$$k^{-1} \sum_{j=1}^{k} \mu_{j.} = N^{-1} \sum_{\alpha=1}^{g} n_\alpha \mu_{.\alpha} = \mu_{..} \text{ the grand mean.}$$

TABLE 1

Source	D.F.	S.S.	F
Treatments	$k - 1$	Q_1	$F_1 = (N - g)Q_1/Q_5$
Groups	$g - 1$	Q_2	$F_2 = (N - g)Q_2/(g - 1)Q_3$
Ind. Within Groups	$N - g$	Q_3	
Treat. × Groups	$(k - 1)(g - 1)$	Q_4	$F_3 = (N - g)Q_4/(g - 1)Q_5$
Treat. × Ind. Within Groups	$(k - 1)(N - g)$	Q_5	
Total	$Nk - 1$		

We will now partition the total sum of squares into 5 constituent sums of squares, as one would usually do with a mixed model that satisfied all the usual analysis of variance assumptions.

Let S be defined as the matrix of the quadratic form representing the correction factor which is the square of the grand total of all the observations divided by kN. S is a $kN \times kN$ matrix whose elements are all $(kN)^{-1}$. Further let a matrix M of sub-matrices $M_{\alpha\beta}$ be denoted as

$$\{M_{\alpha\beta}\} = \begin{pmatrix} M_{11} & \cdots & M_{1g} \\ \vdots & & \vdots \\ M_{g1} & \cdots & M_{gg} \end{pmatrix}.$$

If $M_{\alpha\beta} = 0$, for $\alpha \neq \beta$, let the resulting matrix be denoted by

$$\{M_\alpha\} = \begin{pmatrix} M_1 & 0 & \cdots & 0 \\ 0 & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & M_g \end{pmatrix}.$$

Now let

$$Q_1 = x'Ax = N \sum_{j=1}^{k} (\bar{x}_{.j.} - \bar{x}_{...})^2$$

and

$$A = \{N^{-1}A_{\alpha\beta}\} - S,$$

where $A_{\alpha\beta}$ is the matrix of $n_\alpha \times n_\beta$ matrices each of which is the $k \times k$ identity matrix.

Let

$$Q_2 = x'Bx = k \sum_{\alpha=1}^{g} n_\alpha (\bar{x}_{..\alpha} - \bar{x}_{...})^2,$$

where $B = \{n_\alpha^{-1}B_\alpha\} - S$ and B_α is the matrix of $n_\alpha \times n_\alpha$ matrices each of which is of order $k \times k$, and is of the form

$$E = k^{-1}\mathbf{1}_k\mathbf{1}_k',$$

where $\mathbf{1}_k' = (1, \cdots, 1)$.

Let

$$Q_3 = x'Cx = k \sum_{\alpha=1}^{g} \sum_{i=1}^{n_\alpha} (\bar{x}_{i\cdot\alpha} - \bar{x}_{\cdot\cdot\alpha})^2,$$

where $C = \{n_\alpha^{-1} C_\alpha\}$ and $C_\alpha = n_\alpha E_\alpha - B_\alpha$, where

$$E_\alpha = \begin{pmatrix} E & 0 & \cdots & 0 \\ 0 & & & \vdots \\ \vdots & & & 0 \\ 0 & \cdots & 0 & E \end{pmatrix}.$$

Let

$$Q_4 = x'Dx = \sum_{\alpha=1}^{g} n_\alpha \sum_{j=1}^{k} (\bar{x}_{\cdot j\alpha} - \bar{x}_{\cdot\cdot\alpha} - \bar{x}_{\cdot j\cdot} + \bar{x}_{\cdot\cdot\cdot})^2,$$

where

$$D = \{n_\alpha^{-1}(A_\alpha - B_\alpha)\} + \{N^{-1}(B_{\alpha\beta} - A_{\alpha\beta})\},$$

and here $M_{\alpha\beta}$ is a matrix of $n_\alpha \times n_\beta$ matrices each of order k where here $A_{\alpha\beta}$ refers to the matrix of identity matrices and $B_{\alpha\beta}$ refers to matrices of E's.

Let

$$Q_5 = x'Fx = \sum_{\alpha=1}^{g} \sum_{j=1}^{k} \sum_{i=1}^{n_\alpha} (x_{ij\alpha} - \bar{x}_{\cdot j\alpha} - \bar{x}_{i\cdot\alpha} + \bar{x}_{\cdot\cdot\alpha})^2,$$

where $F = \{n_\alpha^{-1} F_\alpha\}$ and $F_\alpha = n_\alpha(I_\alpha - E_\alpha) + B_\alpha - A_\alpha$, where

$$I_\alpha - E_\alpha = \begin{pmatrix} I - E & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & I - E \end{pmatrix}.$$

Now it is easy to show even for arbitrary V, the basic matrix of Λ, that

$$A\Lambda F = D\Lambda F = B\Lambda C = B\Lambda F = C\Lambda D = 0.$$

Hence by a result due to Carpenter [3], Q_1 and Q_4 are independent of Q_5, Q_2 is independent of Q_3 and Q_5, and Q_3 is independent of Q_4. Further as Box has shown if $Q = (x - \mu)'M(x - \mu)$ where x' has variance-covariance matrix Λ, the sth cumulant of Q, $K_s(Q) = 2^{s-1}(s - 1)! \operatorname{Tr} (\Lambda M)^s$ where Tr stands for the trace of a matrix. Hence by straightforward algebra we get

$$K_1(Q_1) = \operatorname{Tr} V - \operatorname{Tr} EV + N \sum_{j=1}^{k} (\mu_{j\cdot} - \mu_{\cdot\cdot})^2,$$

$$K_1(Q_2) = (g - 1) \operatorname{Tr} EV + \sum_{\alpha=1}^{g} n_\alpha (\mu_{\cdot\alpha} - \mu_{\cdot\cdot})^2,$$

$$K_1(Q_3) = (N - g) \operatorname{Tr} EV,$$

$$K_1(Q_4) = (g - 1)(\operatorname{Tr} V - \operatorname{Tr} EV) + \sum_{\alpha=1}^{g} n_\alpha \sum_{j=1}^{k} (\mu_{j\alpha} - \mu_{j\cdot} - \mu_{\cdot\alpha} + \mu_{\cdot\cdot})^2,$$

$K_1(Q_5) = (N - g)(\text{Tr } V - \text{Tr } EV),$

and

$$K_2(Q_1) = 2 \text{ Tr } (\Lambda A)^2 = 2 \text{ Tr } (V - EV)^2 \text{ if } \mu_{j.} = \mu_{..},$$
$$K_2(Q_2) = 2 \text{ Tr } (\Lambda B)^2 = 2(g - 1) \text{ Tr } (EV)^2 \text{ if } \mu_{.\alpha} = \mu_{..},$$
$$K_2(Q_3) = 2 \text{ Tr } (\Lambda C)^2 = 2(N - g) \text{ Tr } (EV)^2,$$
$$K_2(Q_4) = 2 \text{ Tr } (\Lambda D)^2 = 2(g - 1) \text{ Tr } (V - EV)^2 \text{ if } \mu_{j\alpha} - \mu_{j.} - \mu_{.\alpha} + \mu_{..} = 0,$$
$$K_2(Q_5) = 2 \text{ Tr } (\Lambda F)^2 = 2(N - g) \text{ Tr } (V - EV)^2.$$

From the first cumulants it is clear that under the null hypothesis of no treatment differences, the Expected Mean Square (E.M.S.) for $(k - 1)^{-1}Q_1$ is $(k - 1)^{-1}$ $(\text{Tr } V - \text{Tr } EV)$; under the null hypothesis of no group $\times$ treatment interaction, the E.M.S. of $(g - 1)^{-1}(k - 1)^{-1}Q_4$ is $(k - 1)^{-1}(\text{Tr } V - \text{Tr } EV)$, while the E.M.S. of $(N - g)^{-1}(k - 1)^{-1}Q_5$ is just $(k - 1)^{-1}(\text{Tr } V - \text{Tr } EV)$. Hence under the hypothesis that the treatment means are equal, the numerator and denominator of F_1 estimate the same quantity; and under the hypothesis of no interaction, the numerator and denominator of F_3 estimate the same quantity. Similarly under the hypothesis of no group differences, the numerator and denominator of F_2 estimate the same quantity.

Now using the results of Box ([1], Theorem 6.1) on the approximate distribution of linear sums of chi-square variates, it is clear that F_1 is approximately distributed like $F[(k - 1)\epsilon, (k - 1)(N - g)\epsilon]$ and F_3 is approximately like $F[(g - 1)(k - 1)\epsilon, (k - 1)(N - g)\epsilon]$ while it is obvious that F_2 is exactly distributed like $F(g - 1, N - g)$, where (Box [2])

$$\epsilon = k^2(\bar{v}_{tt} - \bar{v}_{..})^2/(k - 1)\left(\sum_{t=1}^{k}\sum_{s=1}^{k} v_{ts}^2 - 2k\sum_{t=1}^{k} \bar{v}_{t.}^2 + k^2\bar{v}_{..}^2\right)$$

and v_{ts} are the elements of V, $\bar{v}_{tt}$ is the mean of the diagonal terms, $\bar{v}_{t.}$ is the mean of the tth row (or tth column) and $\bar{v}_{..}$ is the grand mean. This result is easily extended to the fixed interactions in an r-way classification where one of the ways is individuals divided into g-groups and the other $r - 1$ classifications are fixed.

3. A lower bound on ϵ. Clearly, the formulation of the degrees of freedom with which we enter the F-table requires the computation of the elements of the variance-covariance matrix. We now present a lower limit on ϵ independent of these elements. This limit, although obvious and simple, may be too conservative.

From Theorem 6.1 Box [1], it is easy to show that

$$\epsilon = (k - 1)^{-1}[\text{Tr } (V - EV)]^2 / \text{Tr } (V - EV)^2,$$
$$\epsilon = (k - 1)^{-1}\left(\sum_{j=1}^{k} \lambda_j\right)^2 \Big/ \sum_{j=1}^{k} \lambda_j^2,$$

where $\lambda_j(j = 1 \cdots k)$ are the latent roots of $(V - EV)$ and are non-negative. But $(\sum \lambda_j)^2 \geqq \sum \lambda_j^2$. Therefore $\epsilon \geqq (k - 1)^{-1}$. Hence, F_1 is conservatively distributed like $F(1, N - g)$ and F_3 is conservatively distributed like $F(g - 1, N - g)$. We also note that if $V = \sigma^2 I$ (the usual analysis of variance assumption) all the roots of $V - EV$ are equal except for one which is equal to zero so that $\epsilon = 1$ in this case.

4. A joint test of groups and treatment × group interaction. In psychological problems it is sometimes necessary to test whether several groups form one cluster. This is equivalent to testing jointly groups and group × treatment interaction. The proposed test here is

$$F_0 = (N - g)Q' / (g - 1)Q,$$

where

$$Q' = Q_2 + Q_4 \quad \text{and} \quad Q = Q_3 + Q_5 .$$

It is clear from Section 2 that the numerator and denominator are independent and

$$K_1(Q') = (g - 1) \operatorname{Tr} V + \sum_{\alpha=1}^{g} n_\alpha (\mu_{\cdot\alpha} - \mu_{\cdot\cdot})^2 + \sum_{\alpha=1}^{g} n_\alpha \cdot \sum_{j=1}^{k} (\mu_{j\alpha} - \mu_{j\cdot} - \mu_{\cdot\alpha} + \mu_{\cdot\cdot})^2, \; K_1(Q) = (N - g) \operatorname{Tr} V;$$

and if $\mu_{\cdot\alpha} = \mu_{\cdot\cdot}$, $\mu_{j\alpha} - \mu_{j\cdot} - \mu_{\cdot\alpha} + \mu_{\cdot\cdot} = 0$, then

$$K_2(Q') = 2(g - 1) \operatorname{Tr} V^2,$$
$$K_2(Q) = 2(N - g) \operatorname{Tr} V^2,$$

and again by using (Theorem 6.1 [1]), F_0 is approximately distributed like $F[(g - 1)k\epsilon', (N - g)k\epsilon']$, where

$$\epsilon' = k\bar{v}_{tt}^2 \Big/ \sum_{t}^{k} \sum_{s}^{k} v_{ts}^2 .$$

Further it is easy to show that $\epsilon' \geqq k^{-1}$ independent of the population variances and covariances and a conservative test would be $F(g - 1, N - g)$. The rationale for this test is that the numerator and denominator of F_0 estimate the same quantity under the null hypothesis of no group effects and no treatment × group effects.

It is of interest to point out and make more explicit the relationship between the foregoing discussion and the general hypothesis in multivariate analysis of the equality of vector means among g populations where all the variables are measured in the same metric. This latter is

$$H_0 (\mu_1 = \mu_2 = \cdots = \mu_g),$$

where $\mu_\alpha' = (\mu_{1\alpha}, \mu_{2\alpha}, \cdots, \mu_{k\alpha})$ is the vector mean of the αth group (i.e., multivariate normal population). But the joint test on groups and group × treatment interaction just presented is in effect also a test for the equality of the g

vector means, since the joint null hypothesis of no interaction and equal group means is equivalent to

$$\mu_{j\alpha} - \mu_{j.} - \mu_{.\alpha} + \mu_{..} = 0 \qquad \text{for all } j,\ \alpha,$$

$$\mu_{.\alpha} = \mu_{..} \qquad \text{for all } \alpha,$$

which is easily seen to imply $\mu_{j\alpha} = \mu_{j.}$ for all α, which is equivalent to $\mu_1 = \mu_2 = \cdots = \mu_g$. Therefore, if the variance-covariance matrices in the g groups can be assumed equal, an approximate test on the hypothesis of equal vector means in multivariate analysis is the F_0 test with ϵ approximated from the sample variances and covariances. It is clear that the conservative F-test which is independent of ϵ can also be used in this case. Furthermore we shall show that if the variance-covariance matrices are not assumed equal, the conservative F-test can be used with the restriction that $n_\alpha = n$.

5. Remarks on unequal variance-covariance matrices. One of the basic assumptions was that each of the N individuals had the same variance-covariance matrix. However if $n_\alpha = n$ for $\alpha = 1, \cdots g$, then we need only assume that individuals in the same group have the same variance-covariance matrix while these variance-covariance matrices may vary from group to group. In this case we get unbiased numerators and denominators of the test ratios as before and the same approximate distributions can be derived, but now the numerator and denominator degrees of freedom have different adjustment factors, each depending upon the different covariance matrices. However it can be easily shown that the lower bounds on these ϵ's are such that F_0, F_1, F_2, and F_3 all have the same conservative F-test, namely, $F(1, n - 1)$.

6. Acknowledgment. We are indebted to the referee for detecting an error and for several suggestions which have improved the exposition.

REFERENCES

[1] G. E. P. Box, "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.*, Vol. 25 (1954), pp. 290–302.

[2] G. E. P. Box, "Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and correlation between errors in the two-way classification," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 484–498.

[3] O. Carpenter, "Note on the extension of Craig's theorem to non-central variates," *Ann. Math. Stat.*, Vol. 21 (1950), pp. 455–457.