A NOTE ON J. ROY'S "STEP-DOWN PROCEDURE IN MULTIVARIATE ANALYSIS"

by

V. P. Bhapkar

University of North Carolina

This research was supported partly by the Office of Naval Research under Contract No. Nonr-855(06) for research in probability and statistics at Chapel Hill and partly by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command under Contract No. AF 49(638)-213. Reproduction in whole or in part for any purpose of the United States Government is permitted.

Institute of Statistics
Mimeograph Series No. 212
December, 1958

A NOTE ON J. ROY'S "STEP-DOWN PROCEDURE IN MULTIVARIATE ANALYSIS"

By

V. P. Bhapkar

University of North Carolina

1. Introduction and Summary.

Test procedures in multivariate analysis are usually based on the λ-criterion or a criterion in terms of the largest and/or the smallest characteristic roots of certain matrices, each criterion being a special case of the general union-intersection principle. An alternative procedure, called the step-down procedure, has been used by Roy and Bargmann [2] in devising a test of multiple independence between variates distributed according to the multivariate normal law. This procedure again can be derived as a special case of the union-intersection principle. This procedure has been recently applied to multivariate analysis of variance by Roy [1] in deriving new tests of significance and simultaneous confidence-bounds on a number of "deviation-parameters." In this note the same procedure is applied to test multiple independence of normal variates under a general linear model.

2. Notation and Preliminaries.

In the notation of [1], we have a matrix $Y_{n \times p}$ of random variables, such that the rows are distributed independently, each row

*This research was supported partly by the Office of Naval Research under Contract No. Nonr-855(06) for research in probability and statistics at Chapel Hill and partly by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command under Contract No. AF 49(638)-213. Reproduction in whole or in part for any purpose of the United States Government is permitted.

having a p-variate normal distribution with the same variance-covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ which is symmetric and positive-definite.

The expected values are given by

$$\mathcal{E}\, Y_{n \times p} = A_{n \times m} \Theta_{m \times p} \tag{1}$$

where A is a matrix of known constants of rank r, $r \leq (n-p)$, and Θ is a matrix of unknown parameters. We want to test the hypothesis that the p-variables are independent, i.e.,

$$\mathcal{H}_0 : \ \sigma_{ij} = 0 \quad (i \neq j,\ i,j=1,2,\ldots,p). \tag{2}$$

The customary likelihood-ratio test for $\mathcal{H}_0$ is based on p x p matrices of random variables

$$S_e = Y'EY \text{ and } S_h = Y'HY \tag{3}$$

called respectively the sum of products matrix due to error and the sum of products matrix due to hypothesis. Here E and H are n x n symmetric idempotent matrices with non-stochastic elements, E of rank n-r and H of rank p. The test is

$$\text{accept } \mathcal{H}_0, \text{ if } L \equiv \frac{|S_e|}{|S_e + S_h|} > c \tag{4}$$

otherwise reject $\mathcal{H}_0$

where c is a pre-assigned constant depending on the level of significance.

3. The Step-Down Procedure to Test $\mathcal{H}_0$.

We shall denote the i-th columns of the matrices Y and Θ by $\underline{y}_i$ and $\underline{\theta}_i$ respectively and write $Y_i = [\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_i]$ and $\Theta_i = [\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_i]$.

We shall also denote the top left-hand i x i submatrix of Σ by Σ_i.

If Y_i is fixed, the n elements of $\underline{y}_{i+1}$ are distributed indepen dently and normally with the same variance σ^2_{i+1} and expectations given by

(5) $$\mathcal{E}(\underline{y}_{i+1} \mid Y_i) = A\,\underline{\eta}_{i+1} + Y_i\underline{\beta}_i,$$

where $\underline{\beta}_i$ is a column i-vector

(6) $$\underline{\beta}_i = \Sigma_i^{-1}\begin{pmatrix}\sigma_{1,i+1}\\ \sigma_{2,i+1}\\ \vdots\\ \sigma_{i,i+1}\end{pmatrix}$$

and $\underline{\eta}_{i+1}$ is a column m-vector given by

(7) $$\underline{\eta}_{i+1} = \underline{\theta}_{i+1} - \Theta_i\underline{\beta}_i$$

and

(8) $$\sigma^2_{i+1} = \frac{|\Sigma_{i+1}|}{|\Sigma_i|} \qquad i=1,2,\dots,p-1 \quad .$$

We note that $\mathcal{H}_0$ is true if and only if the hypothesis $\mathcal{H}_i$ that $\underline{\beta}_i = 0$ holds for all i = 1,2,..,(p-1). Now the elements of the vectors $\underline{\beta}_i$, $\underline{\eta}_{i+1}$ in (5) may be regarded as unknown parameters and hence, when Y_i is fixed, the hypothesis $\mathcal{H}_i$ that $\underline{\beta}_i = 0$ is a linear hypothesis in univariate analysis with the linear model given by (5).

(9) $$\begin{cases}\text{Now Rank } (Y_i) = i & \text{a.e.}\\ \text{and Rank } (A,\, Y_i) = r+i & \text{a.e.}\end{cases}$$

Also

(10) $\mathcal{H}_i$ implies $i \underset{m}{(0,} \underset{i}{I)} \begin{pmatrix} \underline{\eta}_{i+1} \\ \underline{\beta}_i \end{pmatrix} = \underset{i x 1}{0}$

Furthermore,

(11) $\text{Rank} \begin{pmatrix} A & Y_i \\ 0 & I \end{pmatrix} = r + i = \text{Rank } (A, Y_i).$

Hence $\underline{\beta}_i$ is estimable and the hypothesis $\mathcal{H}_i$ testable. Let $\hat{\underline{\beta}}_i$ be the estimator of $\underline{\beta}_i$, the elements of which are linear functions of elements of $\underline{y}_{i+1}$ and also are minimum variance unbiased estimators of the corresponding elements in $\underline{\beta}_i$. Denote the variance-covariance matrix of $\hat{\underline{\beta}}_i$ by $C_i \sigma^2_{i+1}$ where C_i is an i x i positive-definite matrix. Let $s_i^2/n-r-i$ denote the usual error mean square giving an unbiased estimator of σ^2_{i+1}. Then it is well known that

(12) $$F_i \equiv \frac{(\hat{\underline{\beta}}_i - \underline{\beta}_i)' \, C_i^{-1} \, (\hat{\underline{\beta}}_i - \underline{\beta}_i)/i}{s_i^2/n-r-i}, \quad i=1,2,\ldots,(p-1),$$

is distributed as a variance ratio with i and n-r-i degrees of freedom.

Thus the conditional distribution of F_i, given Y_i, does not involve Y_i and hence $F_1, F_2, \ldots, F_{i-1}$. Therefore, the statistics $F_1, F_2, \ldots, F_{p-1}$ are independently distributed as variance ratios with degrees of freedom i and n-r-i respectively $(i=1,2,\ldots,p-1)$.

For a preassigned constant α_i, $0 < \alpha_i < 1$, let f_i denote the upper 100 α_i percent point of the variance ratio distribution with i and n-r-i degrees of freedom. Then the probability that simultaneously

(13) $F_i \leq f_i, \quad i=1,2,\ldots,p-1$,

is equal to $\prod_{i=1}^{p-1} (1-\alpha_i)$.

Since $\mathcal{H}_0 \Longleftrightarrow \mathcal{H}_i : \underline{\beta}_i = 0 \quad i=1,2,\ldots,p-1$, we utilise (12) and set up the following test procedure for $\mathcal{H}_0$:

(14) accept H_0, if $$u_i \equiv \frac{\hat{\underline{\beta}}_i' C_i^{-1} \hat{\underline{\beta}}_i / i}{s_i^2 / n-r-i} \leq f_i$$

for all $i=1,2,\ldots,p-1$

otherwise reject $\mathcal{H}_0$.

To carry out the test one should first compute u_1. If $u_1 > f_1$, $\mathcal{H}_0$ is rejected. If $u_1 \leq f_1$, u_2 is computed. If $u_2 > f_2$, $\mathcal{H}_0$ is rejected. If $u_2 \leq f_2$, u_3 is computed and so on. The level of significance for this test is obviously $1-\prod_{i=1}^{p-1} (1-\alpha_i)$. One possibility is $\alpha_1=\alpha_2=\ldots=\alpha_{p-1}$. We would prefer choosing α's so that $f_1 = f_2 = \ldots = f_{p-1}$ for reasons discussed in [2].

4. Confidence bounds associated with the test.

Now $F_i \leq f_i \longrightarrow (\hat{\underline{\beta}}_i-\underline{\beta}_i)'(\hat{\underline{\beta}}_i-\underline{\beta}_i) \leq \lambda_{max} (C_i) \ell_i^2 s_i^2$ where $\ell_i^2 = \frac{i f_i}{n-r-i}$.

(15) $$\therefore \underline{a}_i' \hat{\underline{\beta}}_i - \ell_i s_i \lambda_{max}^{1/2} (C_i) \leq \underline{a}_i' \underline{\beta}_i \leq \underline{a}_i' \hat{\beta}_i + \ell_i s_i \lambda_{max}^{1/2} (C_i)$$

for all non-null $\underline{a}_i$ (ixl) such that $\underline{a}_i'\underline{a}_i=1$. This, therefore, implies

(16) $$(\hat{\underline{\beta}}_i' \hat{\underline{\beta}}_i)^{1/2} - \ell_i s_i \lambda_{max}^{1/2} (C_i) \leq (\underline{\beta}_i' \underline{\beta}_i)^{1/2} \leq (\hat{\underline{\beta}}_i' \hat{\underline{\beta}}_i)^{1/2} + \ell_i s_i \lambda_{max}^{1/2} (C_i).$$

We may obtain partial statements by choosing some elements of $\underline{a}_i$ in (15) to be zero. Thus we have the simultaneous confidence bounds given by (16) for all possible subsets of $\underline{\beta}_i$ for all $i=1,2,\ldots,p-1$ with the confidence coeffecient $\geq 1 - \prod_{i=1}^{p-1} (1-\alpha_i)$.

5. Remarks.

(i) It can be easily seen that when Y represents a random sample of size n from N $(\underline{\mu}, \Sigma)$, (5) takes the form

$$(17) \qquad \mathcal{E}(y_{i+1,k} \mid Y_i) = \mu_{i+1} + \sum_{j=1}^{i} \beta_{ij} (y_{jk} - \mu_j),$$

where $\underline{y}_i' = [y_{i1}, y_{i2}, \ldots, y_{in}]$ $\underline{\beta}_i' = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{ii}]$, $i=1,2,\ldots,p-1$ and $k=1,2,\ldots,n$.

If we write $s_{ij} = \sum_{k=1}^{n} (y_{ik}-\bar{y}_i)(y_{jk}-\bar{y}_j)$, then it is well to know that

$$\hat{\underline{\beta}}_i = S_{ii}^{-1} \begin{pmatrix} s_{i+1,1} \\ \vdots \\ s_{i+1,i} \end{pmatrix} = \underline{b}_i$$

$$C_i = S_{ii}^{-1} \quad \text{and}$$

$$s_i^2 = s_{i+1,i+1} - (s_{i+1,1}, \ldots, s_{i+1,i})\, S_{ii}^{-1} \begin{pmatrix} s_{i+1,1} \\ \vdots \\ s_{i+1,i} \end{pmatrix}$$

so that

$$u_i = \frac{\underline{b}_i' S_{ii} \underline{b}_i / i}{s_i^2 / n-1-i} = \frac{r^2_{i+1.1,2,\ldots,i}}{1-r^2_{i+1.1,2,\ldots,i}} \; \frac{n-1-i}{i}$$

where $r_{i+1.1,2,\ldots,i}$ denotes the multiple correlation coefficient of $(i+1)$ with $(1,2,\ldots,i)$; thus giving as a special case the test procedure already obtained in [2]. This is, of course, as it should be.

(ii) In this set up, it is of interest to investigate whether (a) the test of the usual multivariate linear hypothesis of the type

$$(18) \qquad \mathcal{H}_0' : \underset{t\times p}{\Phi} = \underset{t\times m}{B} \; \underset{m\times p}{\Theta} = 0 \qquad (\text{Rank } B = t),$$

where Φ is estimable, and (b) the above test of independence, are quasi-independent. As shown in [1], the step-down test procedure for (18) gives, when Y_i is fixed,

$$(19) \quad F_i' \equiv \frac{(\hat{\underline{\Phi}}_{i+1} - \underline{\Phi}_{i+1})' D_{i+1}^{-1} (\hat{\underline{\Phi}}_{i+1} - \underline{\Phi}_{i+1})/t}{s_i^2/n-r-i} \qquad (i=0,1,2,\ldots,p-1)$$

where $\underline{\Phi}_{i+1} = B\,\underline{\eta}_{i+1}$ and the variance-covariance matrix of $\hat{\underline{\Phi}}_{i+1}$ is $D_{i+1}\,\sigma_{i+1}^2$.

F_i given by (12) and F_i' given by (19), for fixed Y_i, are quasi-independent if the numerators, which are marginally distributed as $\frac{\chi_i^2\,\sigma_{i+1}^2}{i}$ and $\frac{\chi_t^2\,\sigma_{i+1}^2}{t}$ respectively, are independent.

It can be easily verified that χ_i^2 and χ_t^2 are not independent and hence the tests for $\mathcal{H}_0$ and $\mathcal{H}_0'$ are not quasi-independent. It may be noted that when Y_i is fixed, the test of $\underline{\beta}_i = 0$ is the nature of testing significance of covariance, as seen from (5), while the test of $\underline{\Phi}_{i+1} = 0$ is in the nature of covariance-analysis. These two are not quasi-independent.

6. Acknowledgment.

I am indebted to Professor S. N. Roy for suggesting this problem and for suggesting improvements.

References

[1] Roy, J., "Step-down procedure in multivariate analysis," North Carolina Institute of Statistics Mimeograph Series No. 187 (1957).

[2] Roy, S.N., and Bargmann, "Tests of multiple independence and the associated confidence-bounds," *Ann. Math. Stat.*, Vol. 29 (1958) pp. 491-503.