SOME TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS

by

JUDITH MAWDSLEY MAHAFFEY

B. S., Kansas State University, 1963

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

Approved by:

R. M. Feyerherm
Major Professor

TABLE OF CONTENTS

HISTORY AND INTRODUCTION

Tests of significance in multivariate analysis refers to various methods of testing hypotheses in which samples are drawn from one or several multivariate populations. The most important of such methods is the analogue to analysis of variance for a single variate. This technique has been referred to as analysis of dispersion. It concerns partitioning sums of squares and cross products of a set of samples from several multivariate populations. The null hypothesis that particular mean vectors of these populations are equal, given that the populations have equal covariance matrices, is tested.

Criteria to test for equality of covariance matrices have been developed. In addition, there have been other criteria for testing hypotheses about such matrices formulated. A criterion to test if a subset of the variables brings out further differences in populations when differences resulting from the rest of the variables are removed has been proposed, and the associated method has been called internal analysis. Finally, two useful tests have been derived to test hypotheses about a single population. The equality of means; of variances and covariances; or of means, variances, and covariances of the variates can be tested. Also, a criterion has been devised to test the mutual independence of subsets of the variates.

In these tests the problem of finding percentage points is difficult because the exact distribution is unknown or too complicated. Approximations have been derived from the moments of these criteria by two methods of approach. First, the moments have been used to fit a Pearson-type curve. This usually gives a good approximation but also involves tedious and complicated computations. Second, the moment expression has been used to obtain

a chi-square approximation to a function of the logarithm of the statistic.

Beginnings in the theory of multivariate analysis date back to 1917 when R. A. Fisher found the sampling distribution of the elements of the covariance matrix for a bivariate normal population. The extension to the case of multiple variates occurred in 1928 when Wishart found the distribution bearing his name. In 1931, Hotelling found the distribution of a statistic T^2, a natural multivariate extension of the univariate case of Student's t distribution. Next, a multivariate analogue to the variance ratio test was developed. Wilks (1932), using the likelihood ratio method that had recently been proposed by Neyman and Pearson, obtained distribution generalizations of many univariate statistics including that of the analysis of variance test criterion. This work formed a basis for many developments in multivariate hypothesis testing.

Generally speaking the difficulties found in hypothesis testing were not in the derivation of a criterion, but in finding its exact distribution when the null hypothesis was true and in evaluating percentage points for practical use. Thus, for the next twenty-five years there were numerous derivations of the distributions of the criteria and of good and workable approximations to them.

This report will be devoted to a discussion of the topics mentioned above.

WILK'S 1932 CONTRIBUTION

Wilks (1932) defined the generalized variance of a sample of k items from a p-variate normal population to be the p^{th} order determinant $|s_{ij}|$ where

$$s_{ij} = \frac{1}{k} \sum_{\alpha=1}^{k} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j),$$

$$\bar{x}_i = \frac{1}{k} \sum_{\alpha=1}^{k} x_{i\alpha} \qquad i, j = 1, 2, \ldots, p,$$

and $x_{i\alpha}$ is the value of the i^{th} variate for the α^{th} item in the sample. Wilks then found the general moment and the distribution of $|s_{ij}|$. He proceeded to derive the general moment of the distribution of the ratio of two generalized sample variances. Both of these distributions were in integral form which he succeeded in evaluating only for $p = 1$ and $p = 2$.

In 1931 Pearson and Neyman had proposed a maximum likelihood criterion λ_H for testing the hypothesis that n' samples are drawn from a subclass ω of a class Ω of admissible populations, where Ω is the class of all sets of n' univariate normal populations and ω is the subclass in each set for which the n' populations have the same means and variances. The maximum of the likelihood function of all n' populations of class Ω is

$$M_\Omega = c(s_o^2)^{-\frac{N}{2}} \prod_{\beta=1}^{n'} (s_\beta^2)^{\frac{n_\beta - 3}{2}}$$

where c is a constant depending on the n_β's, the number of individuals in the β^{th} sample, s_β^2 is the variance of the β^{th} sample, s_o^2 is the pooled variance of all n' samples, and $N = \sum_{\beta=1}^{n} n_\beta$. The maximum in ω is

$$M_\omega = c \prod_{\beta=1}^{n'} (s_\beta^2)^{-\frac{n_\beta}{2}} \prod_{\beta=1}^{n'} (s_\beta^2)^{\frac{n_\beta - 3}{2}},$$

and the likelihood ratio is

$$\lambda_H = \frac{M_\omega}{M_\Omega} = \prod_{\beta=1}^{n'} \left(\frac{s_\beta^2}{s_o^2} \right)^{\frac{n_\beta}{2}} .$$

Wilks found the generalization of this to be

$$\lambda_{H(p)} = \prod_{\beta=1}^{n'} \left(\frac{|s_{ij\beta}|}{|s_{ijo}|} \right)^{\frac{n_\beta}{2}}$$

where $|s_{ij\beta}|$ is the generalized variance of the β^{th} sample and $|s_{ijo}|$ is the generalized variance of the sample formed by combining all n' samples. Note that the β^{th} sample consists of n_β p-variate observations.

To test the hypothesis H(p)' that ω' is the class of all sets of p-variate normal populations in which the covariance matrices are equal, irrespective of the means, Wilks used the criterion

$$\lambda_{H(p)'} = \prod_{\beta=1}^{n'} \left(\frac{|s_{ij\beta}|}{|s_{ij}'|} \right)^{\frac{n_\beta}{2}}$$

where

$$s_{ij}' = \frac{1}{N} \sum_{\beta=1}^{n'} n_\beta s_{ij\beta} .$$

Wilks also derived the general moments of these criteria, and evaluated the distribution functions for some special cases. These are as follows:

	Variance Ratio	Degrees of freedom
$n'=2$, for any p	$\frac{1-\lambda}{\lambda} \cdot \frac{N-p-1}{p}$	p and $(N-p-1)$
$n'\geq 3$, for any p	$\frac{1-\sqrt{\lambda}}{\sqrt{\lambda}} \cdot \frac{N-p-2}{p}$	$2p$ and $2(N-p-2)$
$p=1$, for any n	$\frac{1-\lambda}{\lambda} \cdot \frac{N-n'}{n'-1}$	$(n'-1)$ and $(N-n')$
$p=2$, for any n	$\frac{1-\sqrt{\lambda}}{\sqrt{\lambda}} \cdot \frac{N-n'-1}{n'-1}$	$2(n'-1)$ and $2(N-n'-1)$

The quantity λ denotes either of the criteria.

ANALYSIS OF DISPERSION

In univariate analysis of variance, tests of significance reduce to comparing two independent mean squares. One of these is an estimate of the variance to which a single observation is subject and can usually be referred to as error variance. The other is such an estimate only when the null hypothesis is true, and may be called the mean square due to deviation from hypothesis. In the multivariate case, if each observation consists of p variates, then there are p total sums of squares and $1/2\ (p)(p-1)$ sums of cross products. Analysis of dispersion consists of analyzing each of these into independent categories. The dispersion due to any category gives a generalized variance. As in the univariate case an error term is available, which consists of estimates of the variances and covariances of a set of variables. The corresponding matrix due to any other category leads to such estimates only when the null hypothesis is true. Thus, as in the univariate case, it is possible to compare independent "mean squares" in analysis of dispersion.

Many publications have appeared in statistical literature on the analysis of dispersion for particular experimental layouts and designs. Presently, only the general linear hypothesis will be considered, since the theory and methodology can readily be modified for a particular case.

Let n be the number of samples Y_i ($i=1, 2, \ldots, n$), where Y_i is the (px1) vector of the observations on the p correlated variables, such that the Y_i' are normally and independently distributed with mean vector μ_i' and covariance matrix Σ. Thus, if Y is the (nx1) vector of the Y_i', Y is a (nxp) observation matrix. The joint distribution of the Y_i would be

$$L = \frac{1}{(2\pi)^{\frac{1}{2}np} |\Sigma|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \operatorname{Sum}_{i=1}^{n} (Y_i' - \mu_i') \Sigma^{-1} (Y_i - \mu_i)\right\}.$$

Let the model be denoted $Y = A\phi + \varepsilon$. Then A is the (nxm) design matrix whose elements describe the design under which the data are obtained and which has rank $r \leq m < n$. If A is singular, it must be constrained so that it has an inverse, or so that ϕ is linearly estimable. The effects of the factors involved in the experiment are denoted as ϕ, an (mxp) matrix of unknown parameters. The rows of the matrix ε are assumed to be normally and independently distributed with mean Z and covariance matrix Σ, where Z is a null matrix. Also, $p \leq (n-r)$, so that the sample error matrix is positive definite almost everywhere.

The null hypothesis to be tested is that $C\phi = Z$ where C is an (sxm) matrix of rank $s \leq r < m < n$, which operates on ϕ to give differences in corresponding parameters which are being tested for equality. It is usual to make the first row of C a vector of units, but this is unnecessary. The

alternative hypothesis to be considered is that at least one of the equalities in the system $C\phi = Z$ does not hold.

By the operations admissible in matrix algebra, the quadratic form of the likelihood function may be manipulated as follows:

$$(\text{tr}) \sum_{i=1}^{n} (Y_i' - \mu_i') \sum^{-1} (Y_i - \mu_i) =$$

$$\text{tr} \sum^{-1} \sum_{i=1}^{n} (Y_i - \mu_i)(Y_i' - \mu_i') =$$

$$\text{tr} \sum^{-1} (Y' - \phi'A')(Y - A\phi) =$$

$$\text{tr} \sum^{-1} [Y'Y - Y'A(A'A)^{-1} A'Y + (\hat{\phi}' - \phi')A'A(\hat{\phi} - \phi)] =$$

$$\text{tr} \sum^{-1} [E + (\hat{\phi}' - \phi') A'A(\hat{\phi} - \phi)]$$

where $E = Y'Y - Y'A(A'A)^{-1} A'Y$ and tr denotes the trace of the matrix. In the first equation the matrix has dimension (1x1) and thus taking the trace does not alter the expression. In general, L will be maximized when

$\phi = \hat{\phi} = (A'A)^{-1} A'Y$. Thus, in Ω

$$L = \frac{1}{(2\Pi)^{\frac{np}{2}} |\sum|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \text{tr} \sum^{-1} E\right\}.$$

It can be shown that $\sum$ is maximized by $\hat{\sum} = \frac{1}{n} E$. Thus,

$$L(\Omega) = \frac{n^{\frac{n}{2}}}{(2\Pi)^{\frac{np}{2}} |E|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} np\right\} .$$

Under the null hypothesis the quadratic form can be partitioned as follows:

$$\text{tr } \Sigma^{-1} (Y' - \phi'A')(Y - A\phi) =$$

$$\text{tr } \Sigma^{-1} \Big\{(Y'Y - Y'A(A'A)^{-1} A'Y) + (Y'A(A'A)^{-1} C'[C(A'A)^{-1} C']^{-1} C \cdot$$

$$(A'A)^{-1} A'Y) + (\hat{\hat{\phi}}' - \phi')A'A(\hat{\hat{\phi}} - \phi)\Big\} =$$

$$\text{tr } \Sigma^{-1} [E + H + (\hat{\hat{\phi}}' - \phi')A'A(\hat{\hat{\phi}} - \phi)]$$

where

$$\hat{\hat{\phi}} = (A'A)^{-1} A'Y - (A'A)^{-1} C^{-1} [C(A'A)^{-1} C']^{-1} C(A'A)^{-1} A'Y$$

$$= (A'A)^{-1} [I - C^{-1} [C(A'A)^{-1} C']^{-1} C(A'A)^{-1}] A'Y$$

and $H = Y'A(A'A)^{-1} C'[C(A'A)^{-1} C']^{-1} C(A'A)^{-1} A'Y$.

In ω, letting $\phi = \hat{\hat{\phi}}$, and $\hat{\hat{\Sigma}} = \frac{1}{n}(H+E)$, the maximum likelihood estimates, it is found that

$$L(\omega) = \frac{n^{\frac{n}{2}}}{(2\Pi)^{\frac{np}{2}} |H+E|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} np\right\} .$$

Therefore,

$$\lambda = \frac{L(\omega)}{L(\Omega)} = \left(\frac{|E|}{|H+E|} \right)^{\frac{n}{2}}$$

or

$$-2 \ln \lambda = -n \ln \left(\frac{|E|}{|H+E|} \right) .$$

It should be noted that E is the matrix due to error and H is the matrix due to the hypothesis. Also, H is symmetric and at least positive semidefinite. Thus, the criterion can take on values only on the unit interval. A necessary and sufficient condition for a criterion to be unity is that the null hypothesis be identically true. When the null hypothesis is "accepted", the average value of the criteria will be less than one under repeated application, but it will be larger than when the hypothesis is rejected.

A basic theorem has been proven by Wilks (1938). It states:

> If a population with a variate x is distributed according to the probability function $f(x,\theta_1,\ldots,\theta_h)$, such that optimum estimates $\hat{\theta}_i$ of the θ_i exist which are distributed in large samples according to
>
> $$\frac{|c_{ij}|^{\frac{1}{2}}}{(2\pi)^{\frac{h}{2}}} \exp\left\{ -\frac{1}{2} \sum_{i,j=1}^{h} c_{ij} Z_i Z_j \right\} (1+\phi)\, dZ_1 \ldots dZ_h ,$$
>
> where
>
> $$Z_i = (\hat{\theta}_i - \theta_i) \sqrt{n} ,$$
>
> $$c_{ij} = -E\left(\frac{\partial^2 \ln F}{\partial\theta_i \partial\theta_j} \right) ,$$

$||c_{ij}||$ is positive definite, E denotes mathematical expectation, and ϕ is of order $1/\sqrt{n}$, then when the hypothesis that $\theta_i = \theta_{oi}$, i=m+1, m+2,...,h is true, the distribution of -2 ln L, where L is a likelihood ratio criterion is, except for terms of order $1/\sqrt{n}$, distributed like a chi-square with (h-m) degrees of freedom.

The degrees of freedom is the difference in the number of parameters used to maximize the likelihood function in Ω and ω. This was proven by showing that the characteristic function of -2 ln L is identical to that of the chi-square distribution. Thus

$$-n \ln\left(\frac{|g|}{|H+E|}\right) \sim c[p(k-1)]$$

where c(m) denotes the chi-square density function with m degrees of freedom.

$$V = -\left(n - \frac{p+q+1}{2}\right) \ln\left(\frac{|g|}{|H+E|}\right) \sim c[p(k-1)]$$

where q is one less than the number of populations k.

Rao (1948) obtained the asymptotic distribution of V in the form

$$P_{pq} + \frac{\gamma_2}{m^2}(P_{pq+4} - P_{pq}) + \frac{1}{m^4}\left\{\gamma_4(P_{pq+8} - P_{pq}) - \gamma_2^2(P_{pq+4} - P_{pq})\right\} + \ldots$$

Where P_{pq+s} is the density function of chi-square with pq+s degrees of freedom, and

$$m = n - \frac{1}{2}(p+q+1),$$

$$\gamma_2 = \frac{pq}{48}(p^2+q^2-5),$$

and

$$\gamma_4 = \frac{\gamma_2^2}{2} + \frac{pq}{1920}\left[3p^4 + 3q^4 + 10p^2q^2 - 50(p^2+q^2) + 159\right].$$

Other values of the γ_i may be calculated if a more exact probability is desired. The first term consists of the approximation suggested by Bartlett, and the second term is of order $1/m^2$.

Rao (1951) obtained a closer approximation by considering that if Y is the s^{th} root of λ, where

$$s = \left(\frac{p^2+q^2-4}{p^2+q^2-5}\right)^{\frac{1}{2}}$$

then the distribution function of Y can be written as

$$\beta\left(\frac{ms}{2}+\theta,r\right) + c\,\frac{\Gamma\left(\frac{ms}{2}+\theta+r\right)}{\Gamma\left(\frac{ms}{2}+\theta+r+4\right)} \left\{ \beta\left(\frac{ms}{2}+\theta,\ r+4\right) - \beta\left(\frac{ms}{2}+\theta,\ r\right) \right\} + \ldots$$

where

$$r = \frac{pq}{2},$$

$$\theta = \frac{pq-2}{4},$$

$$c = \frac{\Gamma(r+4)}{\Gamma(r)} \left\{ \frac{\lambda_4 s^4}{16r(r+1)(r+2)(r+3)} - \frac{(r-1)(5r-7)}{5760} \right\},$$

and $\beta(t,u)$ is the distribution function of the beta. The first term offers a powerful approximation, the second being of order $1/m^4$. Also, Rao has shown that

$$\frac{1-\lambda^{\frac{1}{s}}}{\lambda^{\frac{1}{s}}} \cdot \frac{ms+2\theta}{2r}$$

can be used as a variance ratio with $2r$ and $(ms+2\theta)$ degrees of freedom. The quantity $(ms+2\theta)$ need not be integral. The exact distribution functions given by Wilks (1932) are also applicable in analysis of dispersion. They can be used by letting $q=n' - 1$ and $n=N-1$.

Other work in this area has been done by Kabe (1962), Katti (1961), and Roy (1951). Kabe shows that the distribution of several test criteria which can be expressed in terms of gamma functions is identical with the distribution of a linear function of gamma variates. Katti shows that random orthogonal transformations can be used to show that the likelihood ratio is distributed as the product of independent beta variables. Roy derived the series expansion of the distribution of a number of likelihood criteria, and demonstrated the accuracy of using the first term as an approximation.

INTERNAL ANALYSIS

Let $x_1, \ldots, x_s, x_{s+1}, \ldots, x_{s+p}$ be $s+p$ dependent variables for which samples of size $n_1, \ldots, n_k$ are taken from k normal populations. It is possible to test if the variables $x_{s+1}, \ldots, x_{s+p}$ bring out further differences in populations when differences due to $x_1, \ldots, x_s$ are removed. This was done by Rao (1948) and is accomplished by following technique developed by Wishart. Let $(T_{ij}) = (H_{ij}) + (E_{ij})$, $i,j=1, 2, \ldots, s+p$ be the analysis of dispersion for the $s+p$ variates with degrees of freedom $n^* = q + (n^* - q)$. The sum of products matrix due to error for the variables $x_1, \ldots, x_s$ is to be eliminated. Partition the matrix (E_{ij}) as follows:

$$\begin{bmatrix} E_{11}^* & E_{12}^* \\ E_{21}^* & E_{22}^* \end{bmatrix}$$

where E_{11}^* is (sxs). Then the sum of products matrix due to error for $x_1, \ldots, x_s$ is given by

$$E(p/s) = E_{22}^* - E_{21}^*(E_{11}^*)^{-1} E_{12}^* .$$

Similarly, T(s/p) can be found and the criterion is the ratio of the former to the latter. The associated degrees of freedom are (n^*-q-s) and (n^*-s).

The approximations in internal analysis are the same as those in analysis of dispersion except for adjustments in the number of variates, since internal analysis is essentially a conditional analysis of dispersion. Again this procedure can be extended to any partitioning of the variances and covariances.

TESTS OF EQUALITY IN A SINGLE POPULATION

Using the method of maximum likelihood, Wilks (1946) developed test criteria L_{MVC}, L_{VC}, and L_M for testing the following hypotheses about the parameters of a single p-variate normal population:

H_{MVC}: The means are equal and the variances and covariances are equal;

H_{VC}: The variances and covariances are equal, irrespective of the means;

H_M: The means are equal, given that the variances and covariances are equal.

The sample criteria turn out to be the following:

$$L_{MVC} = L_{VC} L_M^{p-1} ,$$

$$L_{VC} = \frac{|s_{ij}|}{(s^2)^p (1-r)^{p-1} (1+(p-1)r)} ,$$

and

$$L_M = \frac{s^2(1-r)}{(s^2)(1-r) + \frac{1}{p-1}\sum_{i=1}^{p}(\bar{x}_i - \bar{x})^2};$$

respectively,

where

$$s^2 = \frac{1}{p}\sum_{i=1}^{p} s_{ii},$$

$$r = \frac{1}{s^2 p(p-1)} \sum_{i \neq j=1}^{p} s_{ij},$$

$$\bar{x} = \frac{1}{p}\sum_{i=1}^{p} \bar{x}_i,$$

and s_{ij} and $\bar{x}_i$ are as previously defined. Thus, there are k p-variate observations. The range of each criterion is the unit interval.

To find these criteria Wilks used the maximum likelihood method, obtaining a likelihood ratio λ_i by appropriate operations. The test criteria were the $2/k^{th}$ root of λ_{MVC} and λ_{VC}, and the $2/k(p-1)^{st}$ root of λ_M, denoted by L_{MVC}, L_{VC}, and L_M. To use the three criteria their distribution under the null hypothesis had to be found. Wilks did this for $p=2$ and $p=3$ for L_{MVC} and L_{VC}, and generally for L_M. The distribution functions were found from the general moments by formulating functions similar to moment generating functions and evaluating particular derivatives. Since the criteria can have values between zero and one, the moments determined the distribution uniquely. For L_M, when H_M is true, the distribution is

$$dF(L_M) = \frac{\Gamma\left[\frac{1}{2}k(p-1)\right]}{\Gamma\left[\frac{1}{2}(k-1)(p-1)\right]\ \Gamma\left[\frac{1}{2}(p-1)\right]} L_M^{\frac{1}{2}(k-1)(p-1)-1} (1-L_M)^{\frac{1}{2}(p-1)-1} dL_M .$$

For p=2, the sampling distribution functions of L_{MVC} and L_{VC} can be found to be

$$dF(L_{MVC}) = \tfrac{1}{2}(k-2)(L_{MVC})^{\frac{1}{2}(k-4)} \quad dL_{MVC}$$

$$dF(L_{VC}) = \frac{\Gamma\left[\frac{1}{2}(k-1)\right]}{\Pi^{\frac{1}{2}}\ \Gamma\ \frac{1}{2}(k-2)} \quad L_{VC}^{\frac{1}{2}(k-4)} \quad \left[1-L_{VC}\right]^{-\frac{1}{2}} \quad dL_{VC};$$

and for p=3,

$$dF(L_{MVC}) = \frac{\Gamma(k)}{2\Gamma(k-3)} \quad \left[L_{MVC}^{\frac{1}{2}}\right]^{k-4} \quad \left[1 - L_{MVC}^{\frac{1}{2}}\right]^2 \quad dL_{MVC}^{\frac{1}{2}} ,$$

$$dF(L_{VC}) = \frac{\Gamma(k-1)}{\Gamma(k-3)} \quad \left[L_{VC}^{\frac{1}{2}}\right]^{k-4} \quad \left[1 - L_{VC}^{\frac{1}{2}}\right] \quad dL_{VC}^{\frac{1}{2}} .$$

These distribution functions are valid only if the respective hypotheses are true. It should be noted that to test H_M, an F-ratio can also be used. This is

$$F[p-1,\ (p-1)(k-1)] = \frac{(k-1)(p-1)(1-L_M)}{(p-1)L_M} = \frac{\frac{s_1}{p-1}}{\frac{s_2}{(p-1)(k-1)}}$$

where

$$s_1 = k \sum_{i=1}^{p} (\bar{x}_i - \bar{\bar{x}})^2,$$

$$s_2 = \sum_{\alpha=1}^{k} \sum_{i=1}^{p} (x_{i\alpha} - \bar{x}'_\alpha - \bar{x}_i + \bar{\bar{x}})^2,$$

and

$$\bar{x}'_\alpha = \frac{1}{p} \sum_{\alpha=1}^{k} x_{i\alpha} .$$

This is equivalent to an analysis of variance to test "row" effects in a (pxk) layout where rows are associated with the p variates and columns with the k items in the sample.

Other criteria to test special hypotheses about a single p-variate normal sample have been given by Bhapkar (1959), Fertig and Leary (1936), and Mauchly (1940). Bhapkar gives a criterion to test that all covariances are zero. The criterion of Fertig and Leary may be used to test that the means, variances, and covariances have specified values. Mauchly's test criterion is used to test that the variances are equal and the covariances are all zero.

Varma (1951) developed methods to find the exact distributions of L_{MVC} and L_{VC} in the form of infinite series, which are numerically workable in calculating significance to any level of accuracy. Beginning with the moments and using various complex transformations, he showed that

$$dF(L_{VC}) = \frac{\Gamma(k-1) \left[L_{VC}^{\frac{1}{2}}\right]^{k-5} \left[1 - L_{VC}^{\frac{1}{2}}\right]}{2\Gamma[(k-3)]} \quad dL_{VC}^{\frac{1}{2}}$$

$$dF(L_{MVC}) = \frac{\Gamma(k) \left[L_{MVC}^{\frac{1}{2}}\right]^{k-5} \left[1 - L_{MVC}^{\frac{1}{2}}\right]^2}{4\Gamma(k-3)} \quad dL_{MVC}^{\frac{1}{2}} .$$

By using Sterlings' asymptotic expansion for $\Gamma(x+h)$ and a factorial series suggested by Nair, these can be evaluated. The one and five percent probability levels for the criteria are given for p=4,5,6,7 and various values of k.

The preceeding theorem by Wilks (1938) is also applicable to these criteria. Thus,

$$-k \ln L_{MVC} \sim C[\tfrac{1}{2} p(p+3)-3]$$
$$-k \ln L_{VC} \sim C[\tfrac{1}{2} p(p+1)-2]$$
$$-k \ln L_{M} \sim C[p-1] .$$

Tukey and Wilks (1946) derived approximate distributions for L_{MVC} and L_{VC} in the form of Pearson's Type I distribution. Approximate percentage points can be obtained by using tables of the incomplete beta function. They justified this by showing that the distribution of a product of beta variables can be approximated by a single beta variable,

$$\frac{1}{\beta(u,v)} (L^{r})^{\frac{1}{r}u-1} (1 - L^{r})^{\frac{1}{r}v-1} ,$$

where L denotes any one of the criteria, v and r depend on p, and u depends also on k. Some special cases are as follows:

p	criterion	v	u
2	L_{MVC}	$\frac{1}{2}(k-2)$	1
2	L_{VC}	$\frac{1}{2}(k-2)$	$\frac{1}{2}$
3	$L_{MVC}^{\frac{1}{2}}$	k-3	3
3	$L_{VC}^{\frac{1}{2}}$	k-3	2
p	L_{M}	$\frac{1}{2}(k-1)(p-1)$	$\frac{1}{2}(p-1)$.

Pearson's tables of the Incomplete Beta Functions are entered with v and u. If Thompson's tables are employed, 2v and 2u are used. Tukey and Wilks have tabled other values of v and u for particular values of p and r.

INDEPENDENCE OF SUBSETS OF VARIABLES

Wilks (1935) developed a criterion to test the independence of groups of variates. If k individuals are measured on p variables with a multivariate normal distribution, and if these variables are grouped into t groups, the m^{th} group having n_m of the variates, the hypothesis to be tested is that the t sets are mutually independent. The criterion is

$$\lambda_I = \frac{|s_{ij}|}{\prod_{m=1}^{t} |s_{ijm}|} = \frac{|r_{ij}|}{\prod_{m=1}^{t} |r_{ijm}|} ,$$

where r_{ij} is the sample correlation coefficient of all individuals between the i^{th} and j^{th} variables, $|s_{ijm}|$ is the determinant of the covariance matrix of the m^{th} set of variables, and $|r_{ijm}|$ is the determinant of the correlations within the m^{th} group. Again, the values of this criterion lie on the unit interval. Wilks found the distribution of this criterion for special groupings into two and three subsets. In the case of division into two groups, one of one variable, the other of p-1, λ_I reduces to $1-R^2$ where R is the multiple correlation coefficient.

Wald and Brookner (1941) have shown that the distribution of the criterion can be obtained in a beta series, and the distribution of the logarithmic statistic in a gamma series. Wilks has proven that $k(1-\lambda_I)$ is distributed as chi-square with γ degrees of freedom, where

$$\gamma = \frac{p(p+1)}{2} - \sum_{m=1}^{t} \frac{n_m(n_m+1)}{2}$$

if k is large compared with p. Further, the probability of obtaining a value of λ_I less than an observed value, say λ, is

$$\Pr\{\lambda_I \leq \lambda\} = 1 - I\left(\frac{k(1-\lambda)}{\sqrt{2\lambda}}, \frac{\gamma-2}{2}\right)$$

where

$$I(u,v) = \frac{1}{\Gamma(v+1)} \int_0^{u(v+1)^{\frac{1}{2}}} e^{-x} x^v \, dx$$

is the incomplete gamma function.

TESTS ON COVARIANCE MATRICES

In addition to Wilks' 1932 criterion to test the equality of covariance matrices, Plackett (1947) has developed a test which is essentially the same as the analysis of variance procedures developed by Fisher. To test the equality of a set of parameters is equivalent to testing whether $(n-1)$ orthogonal linear functions of the parameters each vanish. If $p=1$ and a typical observation from the m^{th} distribution is t_m $(m = 1, 2, \ldots, n)$, then n mutually orthogonal linear functions can be formed, one being

$$u = t_1 + t_2 + \ldots + t_n .$$

If the $(n-1)$ covariances of u and all of the other linear functions are zero, then the variances of the n distributions must be equal. Thus, the problem becomes one of testing the independence of two groups of variates, one of p equations and one of the $p(n-1)$ other orthogonal linear functions.

If the β^{th} population has sample covariance matrix $s_{ij\beta}$ $(\beta=1,2,\ldots,n)$, the hypothesis to be tested is that $s_{ij\beta} = s_{ij\varepsilon}$, $\beta \neq \varepsilon = 1,2, \ldots, n$. The

observations may be written in a matrix of dimensions (kxnp). Thus the i^{th} observations on the β^{th} population will appear in column $(i-1)n+\beta$. Let the observation in the α^{th} sample in this column be denoted as $x_{i\alpha}^{\beta}$. The order of the elements in a column is assumed to be random; if this is a faulty assumption, it is necessary to rearrange segments of the rows of the observation matrix since the criterion depends on a random arrangement. Plackett proves the following theorem:

$$W(n,p) = \frac{n^{2p}}{S(n,p)\ S^{-1}(n,p)}$$

is distributed like Wilks' statistic for testing the hypothesis that two groups of variates of sizes p and p(n-1), known to have been drawn from a np-variate normal population, are mutually independent. If the groups are in fact mutually independent, then $s_{ij\beta} = s_{ij\epsilon}$.

The symbol S(n,p) denotes the sum of all n^{2p} signed minors formed by rows $\beta_1, \beta_2, \ldots, \beta_p$ and columns $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$ of a matrix $G = H'H$ where H is the (kxnp) matrix in which an element is

$$h_{i\alpha}^{\beta} = x_{i\alpha}^{\beta} - \frac{1}{k} \sum_{\alpha=1}^{k} x_{i\alpha}^{\beta} \quad .$$

In application, this criterion is restricted to equal sample size k when $k > np$.

To test for homoscedasticity in the univariate case, Box (1949) extends the criteria that Bartlett (1937) proposed as a test statistic,

$$M = N \ln S - \sum_{\beta=1}^{k} V_\beta \ln S_\beta,$$

where

$$s = \frac{1}{N} \sum_{\beta=1}^{k} V_\beta s_\beta \ , \ N = \sum_{\beta=1}^{k} V_\beta \ ,$$

and s_β is the unbiased estimate of the variance in the β^{th} group based on sums of squares having v_β degrees of freedom. Suppose $s_{ij\beta}$ is the unbiased estimate of the variance or covariance between the i^{th} and j^{th} variables in the β^{th} sample based on sums of squares having v_β degrees of freedom. If there are k such samples and

$$s_{ij} = \frac{1}{N} \sum_{\beta=1}^{k} v_\beta s_{ij\beta}, \quad N = \sum_{\beta=1}^{k} v_\beta$$

then Bartlett's criterion can be generalized as

$$M = N \ln |s_{ij}| - \sum_{\beta=1}^{k} (v_\beta \ln |s_{ij\beta}|)$$

$$= - N \ln L_1$$

where

$$L_1 = \prod_{\beta=1}^{k} \left(\frac{|s_{ij\beta}|}{|s_{ij}|} \right)^{\frac{v_\beta}{N}} .$$

Box (1949) has provided several approximations to this criterion. By finding the general moment of a function of this statistic; the characteristic function of $c'M$, where c' is a constant less than or equal to one; and using an inversion theorem, the probability density function of $c'M$ was found as an infinite series which can be approximated by the chi-square distribution. Box found that if $p > 1$

$$\frac{M}{c'} \sim C[\tfrac{1}{2}(n-1)p(p+1)],$$

where

$$\frac{1}{c'} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{\beta=1}^{k} \frac{1}{\nu_\beta} - \frac{1}{N} \right) .$$

If the degrees of freedom are equal this becomes

$$\frac{1}{c'} = 1 - \frac{(2p^2 + 3p - 1)(k+1)}{6(p+1)k\nu_\beta} .$$

This approximation is inadequate when p is large and ν_β is small.

Box has also shown that

$$\frac{f_2 M}{f_1 (b - M)} \sim F(f_1, f_2),$$

where

$$b = \frac{f_1}{1 - A_1 - f_1/f_2} \qquad \text{if } A_2 > A_1^2$$

$$= \frac{f_2}{1 - A_1 + 2/f_2} \qquad \text{if } A_2 < A_1^2 ,$$

$$A_1 = \frac{2p^2 + 3p - 1}{6(k-1)(p+1)} \left(\sum_{\beta=1}^{k} \frac{1}{\nu_\beta} - \frac{1}{N} \right) ,$$

$$A_2 = \frac{(p-1)(p+2)}{6(k-1)} \left(\sum_{\beta=1}^{k} \frac{1}{\nu_\beta^2} - \frac{1}{N^2} \right) ,$$

$$f_1 = \frac{1}{2} p(p+1)(k-1),$$

and

$$f_2 = \frac{f_1 + 2}{|A_2 - A_1^2|} .$$

This approximation is easily computed with unequal sample sizes.

George (1945) finds the exact distribution for the criterion developed by Wilks to test for homogeneity of variance for special cases of equal and unequal sample sizes. Nair (1938) has given a formal solution to the distribution of the $2/n^{th}$ root of L_1 and for the case of samples of equal sizes the exact expression has been derived in certain instances. Bishop (1939) has developed an empirical method of fitting a Type I curve. He also obtained an approximation for the general case in terms of the chi-square distribution. Gnanadesikan (1960) gives methods to test equality of more than two dispersion matrices against certain alternatives. A method analagous to least squares has been used to derive the criteria.

DISCUSSION

The methods of multivariate analyses are useful for many reasons. First, in many cases with little additional cost to an experimenter, more information can be obtained on each individual item. Then, if additional measurements provide no additional information, as can be tested, the loss would be slight as compared to the conceivable possible gain for purposes of prediction or control. Second, in many disciplines experiments which are multivariate in nature are being performed, but univariate methods of analysis are being applied by assuming repeated measurements to be independent. Not only are basic assumptions being violated, but the total information in the sample is not being utilized, and faulty conclusions can result. Third, with the availability of high speed computers, the cost and difficulty of calculation are minimized. Further, time is no longer a major factor in the choice of a method of analysis.

Applications of the above methods are unlimited. Kendall (1961) gives examples in many areas, illustrating many of the different methods. In fact, every experimental discipline has use for multivariate analysis.

ACKNOWLEDGEMENT

The writer wishes to express her appreciation to Dr. Arlin M. Feyerherm of the Department of Statistics, Kansas State University, for his helpful suggestions and advice during the preparation of this report.

REFERENCES

Bhapkar, V. P. (1959). A note on multiple independence under multivariate normal linear models. Ann. Math. Statist. 30 1248-1251.

Bishop, D. J. (1939). On a comprehensive test of the homogeneity of variances and covariances in multivariate problems. Biometrika 31 31-55.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. Biometrika 36 317-346.

Fartig, J. W. and Leary, M. V. (1936). On a method of testing the hypothesis that an observed sample of n variables of size N has been drawn from a specified population of the same number of variables. Ann. Math. Statist. 7 113-121.

George, A. (1945). On the accuracy of the different approximations to the L_1 distribution. Sankhya 7 20-26.

Gnanadesikan, R. (1959). Equality of more than two variances and of more than two dispersion matrices against certain alternatives. Ann. Math. Statist. 30 177-184.

Kabe, D. G. (1962). On the exact distribution of a class of multivariate test criteria. Ann. Math. Statist. 33 1197-1200.

Katti, S. K. (1961). Distribution of the likelihood ratio for testing multivariate linear hypotheses. Ann. Math. Statist. 32 333-335.

Kendall, M. G. (1963). A Course in Multivariate Analysis. Griffin and Company Limited, London.

Mauchly, J. W. (1940). Significance test for sphericity of a normal multivariate distribution. Ann. Math. Statist. 11 204-209.

Nair, U. S. (1938). The application of moment functions in the study of distribution laws in statistics. Biometrika 30 274-294.

Plackett, R. L. (1947). An exact test for the equality of variances. Biometrika 34 311-319.

Rao, C. R. (1948). Tests of significance in multivariate analysis. Biometrika 35 58-79.

_____ (1951). An asymptotic expansion of the distribution of Wilks' criterion. Bull. Inst. Int. Statist., Part II 33 177-180.

Roy, J. (1951). The distribution of certain likelihood criterion useful in multivariate analysis. Bull. Inst. Int. Statis., Part II 33 219-230.

Smith, H., Gnanedesikan, R., and Hughes, J. B. (1962). Multivariate analysis of variance. Biometrics 18 22-40.

Tukey, John W., and Wilks, S. S. (1946). Approximation of the distribution of the product of beta variables by a single beta variable. Ann. Math. Statis. 17 318-324.

Varma, Bhaskara K. (1951). On the exact distribution of Wilks' L_{MVC} and L_{VC} criteria. Bull. Inst. Int. Statis. Part II 33 181-194.

Wald, A. and Brookner, R. J. (1941). On the distribution of Wilks' statistic for testing independence of several groups of variables. Ann. Math. Statis. 12 137-152.

Wilks, S. S. (1932). Certain generelizations in analysis of variance. Biometrika 24 471-494.

_____ (1935). On the independence of k sets of normally distributed statistical variables. Econometrika 3 309-326.

_____ (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statis. 9 60-62.

_____ (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. Ann. Math. Statis. 17 257-281.

SOME TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS

by

JUDITH MAWDSLEY MAHAFFEY

B. S., Kansas State University, 1963

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1965

The derivation and use of criteria to test hypotheses about the parameters in a multivariate population were given impetus by the work of S. S. Wilks. One of the most significant results was analysis of dispersion.

In analysis of dispersion the criterion is used to test the hypothesis that the mean vectors of n populations are equal. It is assumed that the covariance matrices are equal; criteria to test this assumption are given. A general procedure of analysis of dispersion is outlined under the general linear model and the assumption of normality. Internal analysis, which is a conditional analysis of dispersion, is also discussed briefly.

Various tests of equality of the parameters in a single population are presented. Also, the criterion to test for independence of subsets of the variates in a single population is given.

Various approximations to the distributions of all the above criteria are given. Those approximations that are superior because of ease of computation or because of goodness of fit are enumerated more fully.