MULTIVARIATE THEORY FOR GENERAL STEPWISE METHODS[1]

By A. P. Dempster

Harvard University

0. Summary. This paper presents null hypothesis distribution theory for certain methods of significance testing applicable to multivariate data. This theory is derived in Sections 2 and 3 using simple geometrical reasoning. Section 3 derives theory related to stepwise methods in a form sufficiently general to include both the standard methods of Section 4 and certain new methods based on principal variables of Section 5. The generality in Section 3 is made possible by the independence results proved in Section 2.

1. Introduction. In multivariate analysis the data to be analyzed usually consists of the joint observation of a set of p variables on each of a set of n individuals. For example, a variable might be a length in *cm.* of a specified bone which is found in all humans, and a set of data might provide values of these lengths for a specified set of n human subjects. Such data would become multivariate when it provided values for several such variables on the same set of n subjects. The concept of *variable* will be used a great deal in this paper in the interpretation of various proposed procedures. Since the word variable is used in different senses in different parts of mathematics, a brief discussion of the present use is appropriate.

First, it is important to distinguish between the terms *variable* and *random variable*. As commonly used in mathematical statistics, the latter term refers to a measurable function over a probability measure space. The analogous interpretation of the word variable, as used here without the modifying adjective, is as a function over a space with no associated measure structure. In order not to complicate the mathematical structure any sooner than necessary, I prefer to carry out the discussion in terms of variables rather than random variables wherever possible. This is in line with a general principle of model-building in specific applied situations, namely that first one identifies the variables of interest and then one tries to find the laws (probability or otherwise) governing these variables.

Thus, in its mathematical connotation, a variable V is a function from an arbitrary set I to the real line. The set I corresponds to the set of individuals on which the variable may conceivably be observed, and the values taken on by V correspond to the observations on the variable for the members of I. The notation used will assign different symbols to a variable V and an observed value X of the variable. I believe that users of statistical methods do think naturally in

Received December 19, 1960; revised November 3, 1962.

[1] This research has been supported by the United States Navy through the Office of Naval Research, under Contract Nonr 1866(37). Reproduction in whole or in part is permitted for any purpose of the United States Government.

terms of a variable V without explicitly thinking of sets of observations on V, and it is therefore appropriate that mathematical statistics should allow separate notation for V. In this paper symbols like V and U will denote variables and *not* real-valued observations on these variables, but bold-face $\mathbf{V}$ and $\mathbf{U}$ may represent vectors of observations on V and U.

Multivariate considerations require a set of p variables $V_1, V_2, \cdots, V_p$ which are functions over the same space I. These variables are usually regarded as directly observable on a subset of I. Real functions of $V_1, V_2, \cdots, V_p$ are again real functions over I, and so many new variables can be constructed from the directly observable $V_1, V_2, \cdots, V_p$. In multivariate statistics consideration is often restricted to linear functions, i.e., the p-dimensional vector space $\mathfrak{V}$ of variables

$$U = \boldsymbol{\alpha}(V_1, V_2, \cdots, V_p)' \tag{1.1}$$

where $\boldsymbol{\alpha}$ is any $1 \times p$ vector of real coefficients. The variables $V_1, V_2, \cdots, V_p$ form a basis of $\mathfrak{V}$.

Many of the techniques of multivariate data analysis have as a basic aim to find from data particular variables in $\mathfrak{V}$, e.g., a sample best linear discriminator, pairs of variables resulting from a sample canonical correlation analysis, and variables resulting from a sample principal component analysis. Without the terminology and notation introduced here it is necessary to describe such variables either in terms of coefficient vectors $\boldsymbol{\alpha}$ as in (1.1) or in terms of sets of sample observations, and the indirectness of both of these types of description seems to me awkward and not in accord with applied thinking.

The variables resulting from a sample principal component analysis will be called *principal variables*, as will be made precise in Section 5. Principal variables play a special role in this paper. The distribution theory to be presented, although more general in scope, was motivated by a consideration of *stepwise significance testing procedures based on principal variables*. By this is meant a sequence of test criteria which are naturally thought of in terms of the sample observations on the principal variables. The main idea is to show that certain null hypothesis distributions are unaffected by the fact that the principal variables themselves are computed from the data.

All of the test criteria considered in this paper are based on a pair of matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ which are $p \times p$ sample dispersion matrices corresponding to variables $V_1, V_2, \cdots, V_p$. When the usual normality assumptions are made for the sample observations, and the null hypotheses are specified, then $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent Wishart matrices with common $\boldsymbol{\Sigma}$ and degrees of freedom n_1 and n_2. Such models arise in two different contexts. In multivariate analysis of variance $\mathbf{S}_1$ and $\mathbf{S}_2$ are "between" and "within" dispersion matrices, and the alternative hypothesis is that $\mathbf{S}_1$ has a non-central Wishart distribution. The other context is that of testing for equality of covariance matrices, where $\mathbf{S}_1$ and $\mathbf{S}_2$ refer to different covariances matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ under the alternative hypothesis.

The choice of a test criterion in these situations is usually made on heuristic grounds, since considerations of power are very difficult mathematically and, in any case, do not lead to unique optimum tests. The stepwise methods of this paper constitute a class of testing methods, where several particular members of the class appear to have heuristic appeal. A general member of the class involves a sequence of independent tests on variables $U_1, U_2, \cdots, U_r$ where

$$(U_1, U_2, \cdots, U_r)' = \mathbf{C}(V_1, V_2, \cdots V_p)' \tag{1.2}$$

and the $r \times p$ matrix $\mathbf{C}$ satisfies $\mathbf{CSC}' = \mathbf{I}$ where $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

One particular member of the class is defined by requiring $\mathbf{C}$ to be triangular with zeros above the main diagonal. This will be called the standard stepwise method. It has been discussed by J. Roy [10] and is discussed further in Section 4 of this paper. Another particular member chooses $U_1, U_2, \cdots, U_r$ to be variables resulting from a principal component analysis of $\mathbf{S}$. This method allows the resulting principal variables to be tested one at a time according to a sequence of independent tests. Section 5 gives more details about the approach to testing, which represents a new method of data analysis.

The distribution theory of this paper is of limited originality. The results of Lemmas 2.2 and 2.4 and Theorem 2.1 are essentially given in Section 8.4 of James [6], except that James's methods do not obviously apply to the case $p > n$ which is permitted in the present theory. Similar independence results are implicit in certain derivations of density functions as given by S. N. Roy [12], and are made explicit in some derivations of densities given by Khatri [7]. An attempt has been made to give new and simple derivations for the theory presented here. These derivations are mostly geometrical in nature. It is interesting to note that some early derivations in the multivariate field made heavy use of geometrical terms, e.g., [4] and [11], and the habit of relying almost entirely on matrices was a subsequent development [1], [12]. The present geometrical approach has the fundamental difference with the early geometrical approach that, whereas the latter was concerned with considering small volume elements and deriving densities, probability density functions have no essential role in this paper.

2. Some independence relationships. A random $p \times n$ matrix $\mathbf{X}$ will be called *spherically distributed about the origin*, or simply *spherically distributed*, if, for any $n \times n$ orthogonal matrix $\mathbf{B}$, $\mathbf{XB}$ has the same distribution as $\mathbf{X}$. The briefer terminology will cause no confusion in this paper. The following lemma is well known [6] and its simple proof is omitted.

LEMMA 2.1. *Suppose the columns of a $p \times n$ matrix $\mathbf{X}$ represent a random sample of size n from the normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{0}$ is the $p \times 1$ vector of zeros and $\mathbf{\Sigma}$ is an arbitrary $p \times p$ covariance matrix. Then $\mathbf{X}$ is spherically distributed.*

The results of this paper are based on the next lemma, c.f. James [6], which is perhaps less well known, and whose proof, assuming Lemma 2.1, is presented.

LEMMA 2.2. *Suppose the hypotheses of Lemma* 2.1 *hold. Then the conditional distribution of* $\mathbf{X}$, *given* $\mathbf{S} = \mathbf{X}\mathbf{X}'$, *is spherical.*

PROOF. From Lemma 2.1, the conditional distribution of $\mathbf{X}\mathbf{B}$, given $(\mathbf{X}\mathbf{B})(\mathbf{X}\mathbf{B})'$, is the same as the conditional distribution of $\mathbf{X}$, given $\mathbf{X}\mathbf{X}'$, where $\mathbf{B}$ is any $n \times n$ orthogonal matrix. But $(\mathbf{X}\mathbf{B})(\mathbf{X}\mathbf{B})' = \mathbf{X}\mathbf{X}'$ and therefore the conditional distribution of $\mathbf{X}\mathbf{B}$, given $\mathbf{X}\mathbf{X}'$, is the same as the conditional distribution of $\mathbf{X}$, given $\mathbf{X}\mathbf{X}'$, as required.

Before proceeding, some geometrical terminology is introduced. Suppose E is the n-dimensional Euclidean vector space of $1 \times n$ vectors with the standard definition of the inner product of $\mathbf{A}$ and $\mathbf{B}$ as $\mathbf{A}\mathbf{B}'$. The rows of a $p \times n$ sample matrix $\mathbf{X}$ can be regarded as p vectors $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ in E. To say that $\mathbf{X}$ is spherically distributed is to say that $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ have a joint distribution which is invariant under any orthogonal linear transformation of E leaving the origin fixed. Such $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ will be called *jointly spherically distributed.* The matrix $\mathbf{S} = \mathbf{X}\mathbf{X}'$ determines the *configuration* of (i.e., the set of lengths and angles among) $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$, but not their absolute orientation. Thus, in geometrical language, the foregoing lemmas state that, if $\mathbf{X}$ is a random sample of n from $N(\mathbf{0}, \mathbf{\Sigma})$, then the rows $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ of $\mathbf{X}$ are jointly spherically distributed in E both unconditionally, and conditionally, given their configuration $\mathbf{S}$.

LEMMA 2.3. *Suppose* $\mathbf{X}$ *is any spherically distributed* $p \times n$ *random matrix such that* $\mathbf{S} = \mathbf{X}\mathbf{X}'$ *is constant. Then the distribution of* $\mathbf{X}$ *is uniquely determined.*

PROOF. The rows of $\mathbf{X}$ define vectors $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ in E. $\mathbf{V}_1$ has fixed length l_1 determined by $\mathbf{S}$ and, since it has a spherical distribution, its distribution is uniquely determined as the uniform distribution over the surface of the sphere of radius l_1 in E. For $i = 2, 3, \cdots, p$, $\mathbf{V}_i$ has a component $\mathbf{V}_{i(12\cdots(i-1))}$ in the subspace of E spanned by $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{i-1}$, and a component $\mathbf{V}_{i\cdot 12\cdots(i-1)}$ in the subspace $E_{\cdot 12\cdots(i-1)}$ of E orthogonal to $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{i-1}$. If $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{i-1}$ are fixed along with $\mathbf{S}$, then (i) $\mathbf{V}_{i(12\cdots(i-1))}$ is fixed and (ii) $\mathbf{V}_{i\cdot 12\cdots(i-1)}$ has a fixed length l_i in a fixed subspace $E_{\cdot 12\cdots(i-1)}$ of E. Moreover, the overall spherical symmetry in the model implies that the conditional distribution of $\mathbf{V}_{i\cdot 12\cdots(i-1)}$, given $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{i-1}$, must obey spherical symmetry in $E_{\cdot 12\cdots(i-1)}$, i.e., this distribution is uniquely determined as the uniform distribution over the surface of the sphere of radius l_i in $E_{\cdot 12\cdots(i-1)}$. Consequently the conditional distribution of $\mathbf{V}_i = \mathbf{V}_{i(12\cdots(i-1))} + \mathbf{V}_{i\cdot 12\cdots(i-1)}$, given $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_{i-1}$, is uniquely determined, for $i = 2, 3, \cdots, p$. This completes the proof.

The distribution determined by Lemma 2.3 may be called *the spherical distribution of* $\mathbf{X}$ *with fixed configuration matrix* $\mathbf{S}$. The following Lemma 2.4 is a slight extension of Lemma 2.3 which follows immediately from Lemma 2.3.

LEMMA 2.4. *Suppose* $\mathbf{X}$ *is any spherically distributed* $p \times n$ *random matrix such that* $\mathbf{S} = \mathbf{X}\mathbf{X}'$ *is constant. Suppose* $\mathbf{C}$ *is any constant* $r \times p$ *matrix. Then* $\mathbf{C}\mathbf{X}$ *has the spherical distribution with fixed configuration matrix* $\mathbf{C}\mathbf{S}\mathbf{C}'$.

Lemma 2.4 will be applied in the special case where $r = \text{rank } \mathbf{S}$ and $\mathbf{C}$ is chosen such that the rows $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ of $\mathbf{C}\mathbf{X}$ are orthonormal, i.e., such that

$\mathbf{CSC}' = \mathbf{I}$. The unique distribution of $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ will now be called the spherical distribution in E of the orthonormal set $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$. (This distribution is also described by James [6] as the invariant distribution over a Stiefel manifold.)

The following Theorems 2.1 and 2.2 are basic for the rest of the paper. Using the notation of Anderson [1], a random matrix $\mathbf{S}$ is said to have the *Wishart distribution* $W(\mathbf{\Sigma}, n)$ provided it has the same distribution as $\mathbf{XX}'$ where the columns of the $p \times n$ matrix $\mathbf{X}$ are distributed like a random sample of n from $N(\mathbf{0}, \mathbf{\Sigma})$.

THEOREM 2.1 (c.f. James [6]). *Suppose the columns of a $p \times n$ matrix $\mathbf{X}$ represent a random sample of n from the normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ has rank $f \leqq p$. Suppose $r = \min(f, n)$. Suppose that, for each $\mathbf{S} = \mathbf{XX}'$ of rank r ($\mathbf{S}$ has rank r with probability one), an $r \times p$ matrix $\mathbf{C} = \mathbf{C}(\mathbf{S})$ is defined such that the rows $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ of $\mathbf{CX}$ are orthonormal. Then the set $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ is an orthonormal set distributed spherically in E, and is independent of $\mathbf{S}$ which has the $W(\mathbf{\Sigma}, n)$ distribution.*

PROOF. Theorem 2.1 is an immediate consequence of Lemmas 2.2 and 2.4.

THEOREM 2.2. *Suppose the $p \times p$ random matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are independently distributed like $W(\mathbf{\Sigma}, n_1)$ and $W(\mathbf{\Sigma}, n_2)$, respectively, where $\mathbf{\Sigma}$ has rank $f \leqq p$. Suppose $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ and $n = n_1 + n_2$. Suppose $r = \min(f, n)$ and $r \times p$ matrix $\mathbf{C} = \mathbf{C}(\mathbf{S})$ is defined such that $\mathbf{CSC}' = \mathbf{I}$. Then $\mathbf{CS}_1\mathbf{C}'$ and $\mathbf{CS}_2\mathbf{C}'$ are distributed independently of $\mathbf{S}$ according to a distribution which is free of both $\mathbf{C}(\mathbf{S})$ and $\mathbf{\Sigma}$. In particular the characteristic roots and vectors of $\mathbf{CS}_1\mathbf{C}'$ have this property.*

PROOF. A particular realization of the independent $W(\mathbf{\Sigma}, n_1)$ and $W(\mathbf{\Sigma}, n_2)$ Wishart matrices may be found from two independent samples $\mathbf{X}_1$ and $\mathbf{X}_2$ of sizes n_1 and n_2 from $N(\mathbf{0}, \mathbf{\Sigma})$, i.e., one may set $\mathbf{S}_1 = \mathbf{X}_1\mathbf{X}_1'$ and $\mathbf{S}_2 = \mathbf{X}_2\mathbf{X}_2'$. The proof of Theorem 2.2 rests on the fact that, if the theorem is true for this particular realization of $\mathbf{S}_1$ and $\mathbf{S}_2$, then it must be true in general, because the conclusions of the theorem concern only functions of $\mathbf{S}_1$ and $\mathbf{S}_2$. Given such $\mathbf{X}_1$ and $\mathbf{X}_2$, they may be pooled to provide a single sample $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ of n from $N(\mathbf{0}, \mathbf{\Sigma})$. If the rows of $\mathbf{X}$ define vectors $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ in the Euclidean vector space E, then the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the components of $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_p$ in two mutually perpendicular subspaces $E^{(1)}$ and $E^{(2)}$ of E, when $E^{(1)}$ and $E^{(2)}$ have dimensions n_1 and n_2. From Theorem 2.1, the rows $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ of $\mathbf{CX}$ are orthonormal with the uniform distribution over the unit sphere in E, and are independent of $\mathbf{S}$. But $\mathbf{CS}_i\mathbf{C}'$ is simply the inner product matrix of the orthogonal projections of $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ into $E^{(i)}$, for $i = 1, 2$. Hence, being determined by $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$, the matrices $\mathbf{CS}_1\mathbf{C}'$ and $\mathbf{CS}_2\mathbf{C}'$ share with $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ the property of being distributed independently of $\mathbf{S}$ according to a distribution which is the same for any choice of $\mathbf{C}(\mathbf{S})$.

3. Basic statistics for stepwise procedures. Suppose $\mathbf{S}_1$ and $\mathbf{S}_2$ are $p \times p$ nonnegative definite dispersion matrices arising from data in one of the contexts indicated in Section 1. Following the familiar Gram-Schmidt procedure, each of

the three matrices $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ may be diagonalized to yield

$$\mathbf{D}_1 = \mathbf{T}_1\mathbf{S}_1\mathbf{T}_1', \quad \mathbf{D}_2 = \mathbf{T}_2\mathbf{S}_2\mathbf{T}_2' \quad \text{and} \quad \mathbf{D} = \mathbf{TST}', \tag{3.1}$$

where $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$ are diagonal $p \times p$ matrices, and $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{T}$ are triangular $p \times p$ matrices with all diagonal entries unity and all entries above the diagonal zero. Suppose the ith diagonal elements of $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$ are denoted $d_{ii}^{(1)}$, $d_{ii}^{(2)}$ and d_{ii}, respectively, for $i = 1, 2, \cdots, p$. Having $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$, one may compute

$$P_i = d_{ii}^{(2)}/d_{ii}, \tag{3.2}$$

$$Q_i = d_{ii}^{(1)}/(d_{ii}^{(1)} + d_{ii}^{(2)}), \quad \text{and} \tag{3.3}$$

$$R_i = (d_{ii}^{(1)} + d_{ii}^{(2)})/d_{ii}, \tag{3.4}$$

for $i = 1, 2, \cdots, p$. If any denominators are zero, the corresponding P_i, Q_i or R_i should be regarded as undefined. Note that $R_1 = 1$.

Consider now a generalization of the above. Suppose $\mathbf{C} = \mathbf{C}(\mathbf{S})$ is any $r \times p$ matrix determined by $\mathbf{S}$. Rather than diagonalize $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}$, one may diagonalize $\mathbf{CS}_1\mathbf{C}'$, $\mathbf{CS}_2\mathbf{C}'$ and $\mathbf{CSC}'$ to produce diagonal matrices $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}$ which generalize the $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$ defined in (3.1). Similarly, from the generalized $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$, one may define generalized P_i, Q_i and R_i using (3.2), (3.3) and (3.4), for $i = 1, 2, \cdots, r$. For the remainder of this section, the symbols $d_{ii}^{(1)}$, $d_{ii}^{(2)}$, d_{ii}, P_i, Q_i and R_i should be understood in their generalized sense.

THEOREM 3.1. *Suppose* $\mathbf{S}_1$ *and* $\mathbf{S}_2$ *are independently distributed like* $W(\mathbf{\Sigma}, n_1)$ *and* $W(\mathbf{\Sigma}, n_2)$, *respectively, where* $\mathbf{\Sigma}$ *has rank* $f \leqq p$. *Suppose* $n = n_1 + n_2$, $r = \min(f, n)$, $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ *and* $\mathbf{C} = \mathbf{C}(\mathbf{S})$ *is any* $r \times p$ *matrix such that* $\mathbf{CSC}'$ *has rank* r. *Suppose* P_i, Q_i *and* R_i *are defined as above from* $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}$ *and* $\mathbf{C}$. *Then* (i) *the statistics* P_i, Q_i *and* R_i *for* $1 \leqq i \leqq r$ *are distributed independently of* $\mathbf{S}$, (ii) *the statistics* P_i *for* $1 \leqq i \leqq r$ *are mutually independently distributed like beta random variables with parameters* $((n_2 - i + 1)/2, n_1/2)$, *and* (iii) *the statistics* Q_i *for* $1 \leqq i \leqq r$ *and* R_i *for* $2 \leqq i \leqq r$ *are all mutually independenty distributed like beta random variables with parameters* $((n_1 - i + 1)/2, (n_2 - i + 1)/2)$ *for* Q_i *and* $((n_1 + n_2 - 2i + 2)/2, (i - 1)/2)$ *for* R_i. (*If any of these parameters is zero or negative, then the corresponding statistics are either constant or undefined.*)

PROOF. Theorem 3.1 is simply an extension of Theorem 2.2, except that Theorem 3.1 does not suppose that $\mathbf{CSC}' = \mathbf{I}$. In general, however, there exists a triangular matrix $\mathbf{T}^*$ with unity along the diagonal and a diagonal matrix $\mathbf{D}^*$ such that $\mathbf{C}^*\mathbf{SC}^{*\prime} = \mathbf{I}$ where $\mathbf{C}^* = \mathbf{D}^*\mathbf{T}^*\mathbf{C}$. Replacing $\mathbf{C}$ by $\mathbf{T}^*\mathbf{C}$ does not change $d_{ii}^{(1)}$, $d_{ii}^{(2)}$ or d_{ii} at all, and further replacing $\mathbf{T}^*\mathbf{C}$ by $\mathbf{D}^*\mathbf{T}^*\mathbf{C}$ multiplies each of $d_{ii}^{(1)}$, $d_{ii}^{(2)}$ and d_{ii} by the same constant which leaves P_i, Q_i and R_i unchanged. The following proof of Theorem 3.1 is an extension of the proof of Theorem 2.2, using the same notation and definitions except that $\mathbf{C}$ is replaced by $\mathbf{C}^*$ defined above.

Part (i) of Theorem 3.1 is now seen to be a direct consequence of Theorem 2.2, i.e., P_i, Q_i and R_i are functions of $\mathbf{C}^*\mathbf{S}_1\mathbf{C}^{*\prime}$ and $\mathbf{C}^*\mathbf{S}_2\mathbf{C}^{*\prime}$ only and so are

independent of $\mathbf{S}$. Some additional geometrical terminology will be used in proving parts (ii) and (iii).

Suppose $E^{(1)}_{\cdot 12\cdots(i-1)}$, $E^{(2)}_{\cdot 12\cdots(i-1)}$ and $E_{\cdot 12\cdots(i-1)}$ denote subspaces of $E^{(1)}$, $E^{(2)}$ and E, respectively, which are orthogonal to $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_{i-1}$. These definitions hold for $i = 2, 3, \cdots, p$ and, by convention, $E^{(1)}_{\cdot 12\cdots(i-1)}$, $E^{(2)}_{\cdot 12\cdots(i-1)}$ and $E_{\cdot 12\cdots(i-1)}$ for $i = 1$ may be taken to be $E^{(1)}$, $E^{(2)}$ and E. This notation has been introduced in order to give concise geometrical interpretations to the elements of $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}$, i.e., $d^{(1)}_{ii}$, $d^{(2)}_{ii}$ and d_{ii} are the squared lengths of the components of $\mathbf{W}_i$ in $E^{(1)}_{\cdot 12\cdots(i-1)}$, $E^{(2)}_{\cdot 12\cdots(i-1)}$ and $E_{\cdot 12\cdots(i-1)}$, respectively, for $i = 1, 2, \cdots, r$. Similarly P_i, Q_i and R_i may be interpreted geometrically in an obvious manner as squared cosines of angles. The distributions of these angles are uniquely determined by the spherical symmetry of the distribution of the $\mathbf{W}_i$. First it should be noted that, since $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r$ are spherically distributed in E, the subspaces $E^{(1)}_{\cdot 12\cdots(i-1)}$, $E^{(2)}_{\cdot 12\cdots(i-1)}$ and $E_{\cdot 12\cdots(i-1)}$ have dimensions $n_1 - i + 1$, $n_2 - i + 1$ and $n_1 + n_2 - i + 1$ with probability one. These dimensions determine the parameters of the beta distributions in Theorem 3.1. The conditional distribution of $\mathbf{W}_i$, given $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_{i-1}$, is spherical in the $(n_1 + n_2 - i + 1)$-dimensional subspace $E_{\cdot 12\cdots(i-1)}$. Since P_i is the squared cosine of the angle between $\mathbf{W}_i$ and the $(n_2 - i + 1)$-dimensional subspace $E^{(2)}_{\cdot 12\cdots(i-1)}$, it follows that the conditional distribution of P_i, given $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_{i-1}$, is the beta distribution with parameters $((n_2 - i + 1)/2, n_1/2)$. Since this conditional distribution does not depend on $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_{i-1}$, which in turn determine $P_i, P_2, \cdots, P_{i-1}$, it follows that P_i has the stated beta distribution and is independent of $P_1, P_2, \cdots, P_{i-1}$, for $i = 1, 2, \cdots, r$. Part (iii) of Theorem 3.1 follows in a similar fashion, but the details are omitted.

4. Standard stepwise procedures. In Section 3, the quantities $d^{(1)}_{ii}$, $d^{(2)}_{ii}$, d_{ii}, P_i, Q_i and R_i were first defined in a special case and then were generalized. In the special case, $d^{(1)}_{ii}$, $d^{(2)}_{ii}$, and d_{ii} are of the nature of residual sums of squares of the variable V_i after fitting a linear combination of $V_1, V_2, \cdots, V_{i-1}$. Thus, testing methods based on P_i, Q_i and R_i in the special case may be regarded as tests based on V_i after correcting for $V_1, V_2, \cdots, V_{i-1}$. For example, one might consider a sequence of independent tests based on $P_1, P_2, \cdots, P_p$ in the special case. These methods may be called *standard stepwise procedures*. With these methods it is assumed that $V_1, V_2, \cdots, V_p$ have been pre-ordered according to some *a priori* sense of importance.

For the case of multivariate analysis of variance the standard stepwise procedure based on $P_1, P_2, \cdots, P_p$ is essentially a consequence of a procedure proposed by C. R. Rao [8], p. 73 or [9], p. 264. If $\mathbf{S}_1$ and $\mathbf{S}_2$ denote between groups and within groups dispersion matrices, then, in the notation of this paper, Rao proposes the test statistic

$$\Lambda_1 = \prod_{i=1}^{s} P_i \tag{4.1}$$

for the first s variables, and the test statistic

$$\Lambda_2 = \prod_{i=s+1}^{s+t} P_i \tag{4.2}$$

for the following variables $V_{s+1}, V_{s+2}, \cdots, V_{s+t}$ after correcting for $V_1, V_2, \cdots, V_s$. By taking $t = 1$ and repeating the test (4.2) for $s = 1, 2, \cdots, p-1$ one gets from Rao's method the sequence of independent tests based on $P_1, P_2 \cdots, P_p$.

The standard stepwise procedure for multivariate analysis of variance is given more explicitly by J. Roy [10]. In place of $\mathbf{S}_1$ and $\mathbf{S}_2$ with n_1 and n_2 degrees of freedom, Roy's notation uses S_h and S_e with t and $n - r$ degrees of freedom for between and within dispersion matrices. Roy proposes the test statistics u_i in his Equation (22), where it could be shown that

$$u_i = (1 - P_i)/P_i \cdot (n_2 - i + 1)/n_1 \qquad \text{for } 1 \leqq i \leqq p, \tag{4.3}$$

i.e., the u_i are the F-type variables corresponding to the beta-type variables P_i. A rigorous derivation of (4.3) is omitted since it would require giving a precise definition to Roy's $\hat{\phi}_i$ which in turn would require introducing a great deal of notation. In principle, however, such a derivation is easy, for one need only interpret Roy's u_i geometrically in E, i.e., one need only show that $u_i \times n_1/(n_2 - i + 1)$ is $\tan^2\theta$ for the angle θ such that P_i is $\cos^2\theta$.

Roy also provides standard stepwise procedures for testing the equality of covariance matrices. These tests are equivalent to tests based on Q_i and R_i, i.e., it could be shown that the test statistics given by Roy's formula (50) are

$$\begin{aligned} \frac{\hat{\delta}_i' C_i^{-1} \hat{\delta}_i}{s_{i+1}^2} &= \frac{1 - R_{i+1}}{R_{i+1}} \cdot \frac{i}{n_1 + n_2 - 2i} \qquad \text{for } 1 \leqq i \leqq p-1, \\ \frac{n_2 - i + 1}{n_1 - i + 1} \left(\frac{s_i^{(1)}}{s_i^{(2)}}\right)^2 &= \frac{Q_i}{1 - Q_i} \cdot \frac{n_2 - i + 1}{n_1 - i + 1} \qquad \text{for } 1 \leqq i \leqq p. \end{aligned} \tag{4.4}$$

(In the left hand side of each of the Equations (4.4) I have presumed to correct Roy's formulas (50), (i) by setting $\delta_i = 0$ which is surely true under H_0, and (ii) by squaring $s_i^{(1)}/s_i^{(2)}$ in (50) to agree with (45).) The P_i, Q_i and R_i variables are preferred in this paper because they are simpler and more natural, both to compute and to interpret geometrically.

Note that Theorem 3.1 provides the null hypothesis distribution theory for the standard stepwise procedures based on P_1, Q_1 and R_1.

5. Stepwise procedures on principal variables. Section 1 mentioned the aim of basing tests on principal variables. The term *principal variable* will now be defined and the tests based on them will be described. It will be assumed, to start with, that $\mathbf{S}_1$ and $\mathbf{S}_2$ are between and within dispersion matrices arising in an analysis of variance situation. The *principal component analysis is to be carried out on* $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

The principal component analysis of $\mathbf{S}$ relative to a reference inner product

matrix $\mathbf{K}$ is carried out as follows. First, one chooses the $p \times p$ positive definite matrix $\mathbf{K}$ within certain limitations to be discussed below. Then one finds the *principal variables*

$$U_i = \boldsymbol{\alpha}_i(V_1, V_2, \cdots, V_p)' \tag{5.1}$$

where U_1 is chosen to maximize the ratio of norms $\lambda_1 = \boldsymbol{\alpha}_1\mathbf{S}\boldsymbol{\alpha}_1'/\boldsymbol{\alpha}_1\mathbf{K}\boldsymbol{\alpha}_1'$, and U_i is chosen to maximize the ratio of norms $\lambda_i = \boldsymbol{\alpha}_i\mathbf{S}\boldsymbol{\alpha}_i'/\boldsymbol{\alpha}_i\mathbf{K}\boldsymbol{\alpha}_i'$ subject to the condition that $\boldsymbol{\alpha}_i\mathbf{K}\boldsymbol{\alpha}_j' = 0$ for $i > j$. The associated λ_i may be called *principal values*. Using this definition it is seen that the principal variables have certain extremal properties regarding their dispersion relative to reference matrix $\mathbf{K}$. It is these extremal properties which give the method heuristic practical appeal.

An equivalent description of this analysis is to define the λ_i as the positive roots, in decreasing order, of

$$\det(\mathbf{S} - \lambda\mathbf{K}) = 0 \tag{5.2}$$

and the corresponding $\boldsymbol{\alpha}_i$ to be roots of the linear equations

$$(\mathbf{S} - \lambda_i\mathbf{K})\boldsymbol{\alpha}_i' = \mathbf{0}. \tag{5.3}$$

In this form the principal component analysis defined here is seen to be a more general form, allowing various $\mathbf{K}$, of the analysis proposed by Hotelling [5]. The terms principal variable and principal value are introduced to avoid any ambiguity in the term principal component.

The subsequent testing methods allow $\mathbf{K}$ to be a function of $\mathbf{S}$, although not of $\mathbf{S}_1$ and $\mathbf{S}_2$, subject only to the requirement of positive-definiteness. This range of choice is so wide to allow that any set U_i are principal variables for some $\mathbf{K}$ provided only that they are uncorrelated relative to $\mathbf{S}$, i.e., provided that $\boldsymbol{\alpha}_i\mathbf{S}\boldsymbol{\alpha}_j' = 0$ for $i \neq j$. In practice, however, $\mathbf{K}$ is most often chosen to be the diagonal matrix with the same diagonal elements as $\mathbf{S}$. In this case Equation (5.2) is the same as

$$\det(\mathbf{R} - \lambda\mathbf{I}) = 0, \tag{5.4}$$

where $\mathbf{R}$ is the correlation matrix corresponding to $\mathbf{S}$. An alternative special choice is to set $\mathbf{K} = \mathbf{I}$. Either of these special choices may be used when the problem is expressed in terms of a new basis $V_1^*, V_2^*, \cdots, V_p^*$ of $\mathcal{V}$, whose relation to the original basis $V_1, V_2, \cdots V_p$ does not depend on $\mathbf{S}$. This provides wide, but not all-inclusive, limits on the choice of $\mathbf{K}$.

In any principal component analysis, the $\boldsymbol{\alpha}_i$ are determined only up to a scale factor which may be chosen so that

$$\boldsymbol{\alpha}_i\mathbf{K}\boldsymbol{\alpha}_i' = \lambda_i^{-1}. \tag{5.5}$$

One may then define $\mathbf{C} = \mathbf{C}(\mathbf{S})$ from

$$\mathbf{C} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_r)' \tag{5.6}$$

to satisfy $\mathbf{CSC}' = \mathbf{I}$. Thus Theorems 2.1 and 2.2 may be applied with this $\mathbf{C}$, and

also the general stepwise theory of Section 3 may be applied. The statistics $P_1, P_2, \cdots, P_r$ calculated using this choice of $\mathbf{C}$ come from $\mathbf{CS}_1\mathbf{C}'$, $\mathbf{CS}_2\mathbf{C}'$ and $\mathbf{I}$, i.e., from the dispersion matrices of the principal variables $U_1, U_2, \cdots, U_r$, so that tests based on these $P_1, P_2, \cdots, P_r$ are naturally regarded as *stepwise procedures based on principal variables*. These stepwise statistics are, in general, different from the $P_1, P_2, \cdots, P_p$ found in the standard stepwise procedure, but, according to Theorem 3.1, the same null hypothesis distribution theory holds.

The author feels that tests of this type can be of practical interest, for, if the first few principal variables are imagined to be important in some sense, it is natural to want to check whether they are important relative to an analysis of variance hypothesis. Also, these tests have certain nice mathematical properties which help to give them appeal in certain situations.

The first such property is that, under the null hypothesis, the test statistics $P_1, P_2, \cdots, P_r$ are independent of the set of principal values $\lambda_1, \lambda_2, \cdots, \lambda_r$. This is immediate because, from Theorem 3.1, the P_i are independent of $\mathbf{S}$ and the λ_i are determined by $\mathbf{S}$. Thus, in weighting the individual tests based on the P_i to produce a single test with a given significance level, one may use $\lambda_1, \lambda_2, \cdots, \lambda_r$ to determine the weights. For example, if it turned out that $\lambda_1/\sum_1^p \lambda_i$ were near unity, one might then decide to base a test on the first principal variable only.

The second important property of these tests is that they remain applicable for $p > n_2$, in fact for arbitrarily large p. The usual criteria, based on the roots of

$$\det(\mathbf{S}_1 - \nu\mathbf{S}_2) = 0 \tag{5.7}$$

are undefined in this case. Moreover, the standard stepwise procedure breaks down after incorporating the first n_2 variables, and provides no way of using the information in the remaining $p - n_2$ variables. On the other hand, the principal variables $U_1, U_2, \cdots, U_{n_2}$ can still be tested stepwise and do weight in all p of the original variables. The tests based on $P_1, P_2, \cdots, P_{n_2}$ offer competitors for the testing methods described in [2] and [3].

Several disadvantages of the method based on principal variables, as opposed to the standard method, are as follows. The general distribution theory in Theorem 3.1 is less easily used when the null hypothesis fails. For example, in the standard procedure $P_1, P_2, \cdots, P_s$ retain their null hypothesis distributions provided the null hypothesis holds for the first s variables only, whereas no such simple statement can be made in the case of the principal variable method. Likewise Roy [10] is able to make explicit confidence statements about the parameters of the alternative hypotheses. Such confidence statements are possible in principle with the principal variable method, but in practice they would require carrying out the principal component analysis for each possible set of alternative parameters. This appears to be computationally infeasible. Thus the distribution theory for stepwise tests based on principal variables appears to be limited in practice to testing the full null hypothesis.

Finally, it should be remarked that stepwise tests based on principal variables are also possible for testing the null hypothesis of equality of covariance matrices, i.e., one would again do a principal component analysis on $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$ and define $\mathbf{C}$ as in (5.6). As test statistics one would use the Q_i and R_i associated with this $\mathbf{C}$. These tests share the same advantages and disadvantages as those based on the P_i.

REFERENCES

[1] ANDERSON, T. W. (1958). *Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[2] DEMPSTER, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29** 995–1010.

[3] DEMPSTER, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16** 41–50.

[4] FISHER, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc. London Ser. A* **121** 654–673.

[5] HOTELLING, HAROLD (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417–441 and 498–520.

[6] JAMES, A. T. (1954). Normal multivariate analysis and the orthogonal group. *Ann. Math. Statist.* **25** 40–75.

[7] KHATRI, C. G. (1959). On the mutual independence of certain statistics. *Ann. Math. Statist.* **30** 1258–1262.

[8] RAO, C. RADHAKRISHNA (1948). Tests of significance in multivariate analysis. *Biometrika* **35** 58–79.

[9] RAO, C. RADHAKRISHNA (1952). *Advanced Statistical Methods in Biometric Research.* Wiley, New York.

[10] ROY, J. (1958). Step-down procedure in multivariate analysis. *Ann. Math. Statist.* **29** 1177–1187.

[11] ROY, S. N. (1942). The sampling distribution of p-statistics and certain allied statistics on the non-null hypothesis. *Sankhyā* **6** 15–34.

[12] ROY, S. N. (1957) *Some Aspects of Multivariate Analysis.* Wiley, New York.