THE DISTRIBUTION OF THE EXTREME MAHALANOBIS' DISTANCE FROM THE SAMPLE MEAN

by

Yvonne Germaine Marie Ghislaine Cuttle

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS
in the Department
of
MATHEMATICS

We accept this thesis as conforming to the standard required from candidates for the degree of MASTER OF ARTS.

Members of the Department of
MATHEMATICS

THE UNIVERSITY OF BRITISH COLUMBIA
April, 1956

ABSTRACT

The problem of classification in multivariate analysis is considered. The distribution of the extreme Mahalanobis' distance from the sample mean has been derived for a special case of the bivariate problem, and for this special case the cumulative distribution has been partially tabulated.

The characteristic function of the joint distribution of the Mahalanobis' distances from the sample mean has also been derived.

A brief discussion of the one-dimensional problem and its solution has been included.

ACKNOWLEDGEMENT

We are deeply indebted to Dr. Stanley W. Nash of the Mathematics Department at the University of British Columbia for suggesting the topic and rendering invaluable assistance in the development of this thesis. Our further thanks are due to Dr. S. A. Jennings of the Department of Mathematics for his helpful suggestions.

We are also pleased to acknowledge the support of the National Research Council of Canada which made this research possible.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

INTRODUCTION

In Anthropology, Biology and other sciences the following problem often occurs: Given k groups of objects G_1, G_2, ..., G_k of which samples of sizes n_1, n_2, ..., n_k are taken, p normally distributed characters being measured on these objects, determine on the basis of this data

(1) whether the groups G_1, G_2, ..., G_k all belong to the same population

(2) if all the groups do not belong to the same population, which groups belong together to form clusters and which are from different populations.

The first part of this problem is solved in the most general case and several methods are available to answer the second part of this problem in the special case $p = 1$. For $p > 1$, the second part of this problem has not been solved satisfactorily and certainly not rigorously, although a subjective method of attack has been advocated by K. D. Tocher and is presented in ref. 1 p. 363.

We have attempted in this paper to give a more rigorous approach to the second part of the problem and have succeeded in solving a special case.

CHAPTER ONE

THE ONE-DIMENSIONAL CASE AND ITS SOLUTION

1.1 Notation

Call G_1, G_2, ... G_k the groups or treatments about which we want to test some hypothesis; let n_1, n_2, ... n_k be the sample sizes of the groups, let $\bar{x}_1$, $\bar{x}_2$, ... $\bar{x}_k$ be the sample means of a normally distributed character measured in the objects in the groups. Further, let $\bar{x}$ be the grand mean of the measurements, s^2 be an independent estimate of the variance of the measurements and $s_{\bar{x}}^2 = \frac{s^2}{n}$ the estimate of the variance of the mean $\bar{x}$ where $n = \sum_{i=1}^{k} n_i$.

The problem consists of deciding on the basis of the above information which groups or treatments are significantly different.

Several tests are available to solve this problem. They can be roughly classified as "Multiple Range Tests" and "Multiple F Tests". A brief discussion of these tests for the purpose of illustrating their nature is given below. A more detailed expose and illustrations of the various tests can be found in ref. 2 pp. 18-45.

We have attempted in Chapter 2 and Chapter 3 to generalize, to some extent one of the one-dimensional tests to the multi-dimensional problem. Tukey's multiple F test appeared to be the one which would most easily carry over to the general case, and it is with that test in mind that we approached the problem.

1.2 Multiple Range Tests

In what follows we can assume without loss of generality that the means $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_k$ have been ranked, $\bar{x}_1$ being the smallest mean and $\bar{x}_k$ the largest mean.

A. Student-Newman-Keul test (ref. 3)

The Studentized range

$$q = \frac{\bar{x}_{max} - \bar{x}_{min}}{s_{\bar{x}}} = \frac{\text{range}}{\text{standard deviation}}$$

is considered. The distribution of $q_{(\alpha, n)}$, where α is the level of significance and n the number of degrees of freedom associated with $s_{\bar{x}}$, has been tabulated by J. M. May for various values of α. The tables can also be found in ref. 2 p.22-23.

The test suggested by Newman is as follows:

Step 1: Choose a level significance α, (usually .05 or .1).

Step 2: Compute

$$W_n = q_{(\alpha, n)} \, s_{\bar{x}}$$

$$W_{n-1} = q_{(\alpha, n-1)} \, s_{\bar{x}}$$

$$\vdots$$

$$W_2 = q_{(\alpha, 2)} \cdot s_{\bar{x}}$$

Step 3: Compare $\bar{x}_n - \bar{x}_1$ with W_n. If $\bar{x}_n - \bar{x}_1$ is less than W_n, the process terminates and we assert that the groups belong to the same population at a level of significance α.

If $\bar{x}_n - \bar{x}_1 > W_{n\ 1}$ we state that $\bar{x}_n$ is different from $\bar{x}_1$ or that the corresponding groups G_n and G_1 are significantly different. We then proceed to compare $\bar{x}_{n-1} - \bar{x}_1$ and $\bar{x}_n - \bar{x}_2$ with W_{n-1}. If both $\bar{x}_{n-1} - \bar{x}_1$ and $\bar{x}_n - \bar{x}_2$ are less than W_{n-1} the process terminates.

If, say $\bar{x}_n - \bar{x}_2 > W_{n-1}$, we state that $\bar{x}_n$ is different from $\bar{x}_2$ (or G_n different from G_2) and proceed to compare $\bar{x}_n - \bar{x}_3$ and $\bar{x}_{n-1} - \bar{x}_2$ with W_{n-2}. This process continues until the actual ranges of subsets of i means do not exceed W_i.

Note that it is not necessary to compare a subset of means which is contained in a larger subset, the range of which is less than the calculated W. We could have therefore dispensed with the comparison of $\bar{x}_{n-1} - \bar{x}_2$ and W_{n-2} since the subset ($\bar{x}_{n-1}, \bar{x}_{n-2}, \ldots, x_2$) is contained in ($\bar{x}_{n-1}, \bar{x}_{n-2}, \ldots \bar{x}_2, \bar{x}_1$), the range of which was found to be less than W_{n-1} .

This is one of the easiest tests to perform.

B. Other multiple range tests

Duncan (ref. 5) suggested table values somewhat different from $q_{(\alpha, n)}$. The use of Duncan's procedure tends to decrease the number of type II errors.

Other variations of this test were proposed by Tukey (ref. 6), the use of which would decrease the number of type I errors.

1.3 Multiple F tests

Multiple F tests combine the use of ranges with variance-ratios.

A. Duncan's multiple F test (cf. ref. 5 and 7)

The first stage of Duncan's procedure is to perform a multiple range test as was done above using instead of $q_{(\alpha, p)}$ tabular values somewhat different, R^1_p (ref. 7 or 2). Once the multiple range test has been performed calculate $ss_p = \frac{1}{2} R_p'^2$ p = 2, 3 ... n which gives the sum of squares significant at level α , obtained from the least significant range. Suppose ($\bar{x}_1^1$ $\bar{x}_2^1$ ... $\bar{x}_r^1$) is a group of ranked means for which the multiple range test has failed to show any hetrogenity. The second stage of the test consists of applying the following rule: $\bar{x}_r^1 - \bar{x}_1^1$ is significant if $\bar{x}_r^1 - \bar{x}_1^1 > R_2^1$ and if the sum of squares of all combinations of means out of $\bar{x}_1^1$ $\bar{x}_2^1$ ... $\bar{x}_2^1$ including $\bar{x}_1^1$ $\bar{x}_r^1$ exceed ss_p , p being the number of means in the combination.

The sum of square among the r means $\bar{x}_1^1$ $\bar{x}_2^1$ ... $\bar{x}_r^1$ is $ss_{\bar{x},r} = \dfrac{\sum_{i=1}^{r} \bar{x}_i'^2 - \left(\sum_{i=1}^{r} \bar{x}_i'\right)^2}{r}$

Duncan showed that $\bar{x}_r^1 - \bar{x}_1^1 > R_m^1$ together with $ss_{x,r} > ss_r$ implies that the sum of squares among m means or less out of the r means exceeds the corresponding ss_p , so that in most cases it is not necessary to calculate the sum of square for all possible combinations.

B. <u>Scheffé's test</u> (ref. 8)

In addition to being applicable in testing the difference between two means, Scheffé's test may be used to judge all comparisons of the form $a_1 \bar{x}_1 + a_2 \bar{x}_2 + \ldots + a_n \bar{x}_n$ where the a's are constants with the condition $\sum_{i=1}^{n} a_i = 0$ The standard error of the comparison is $s_c = s_{\bar{x}} \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}$

Define $S = \sqrt{(n-1)\, F_\alpha(n-1, f)}$ where n is the number of means
f is the number of
degrees of freedom

of the error variance.

Scheffe proves that the value of the comparison is significant at the α level if $a_1 \overline{x}_1 + a_2 \overline{x}_2 + \dots + a_n \overline{x}_n > Ss_c$. This test has a larger type II error than Duncan's test, but it has smaller type I error.

C. <u>Tukey's Gap Straggler and variance test (ref. 9)</u>

Rather than considering the range of a group of, say, k means and comparing it to the tabulated values of the Studentized range, Tukey considers the extreme deviate say $\overline{x}^{1}$ from the grand mean m of the group of k means. He shows empirically that

(1) $$\frac{\frac{|\bar{x}' - m|}{s_{\bar{x}}} - \frac{6}{5}\log_{10} k}{3\left(\frac{1}{4} + \frac{1}{n}\right)}$$ n = d.f. associated with $s_{\overline{x}}$ (for $k > 3$)

and

(2) $$\frac{\frac{|\bar{x}' - m|}{s_{\bar{x}}} - \frac{1}{2}}{3\left(\frac{1}{4} + \frac{1}{n}\right)}$$ (for $k = 3$)

are distributed approximately as normal deviates. The exact distribution of an extreme deviate from the sample mean has been given by K. R. Nair (ref. 10).

<u>Tukey's test as given in ref. 9.</u>

<u>Step 1</u>: Choose a level of significance α.

<u>Step 2</u>: Calculate the difference which would have been significant if there were only two means, i.e. $\sqrt{2}\ s_{\bar{x}}\, t_{(n,\alpha)}$ where t is the Student's t with n d. f.

Step 3: Arrange the means in order of magnitude and consider any gap wider than $\sqrt{2}\, s_{\bar{x}}\, t_{(n,\alpha)}$ as a group boundary.

If no group contains more than two means the process terminates.

Step 4: In each group of three or more find the grand mean m, the most straggling mean $\bar{x}'$ and compute the value (1) or (2) as the case may be. Separate any straggling mean for which this is significant at the two sided significance level α for the normal distribution.

Step 5: If step 4 changes any group, repeat the process until no further means are separated. The means separated off from one side of a group form a new group. If any of the new groups so formed contains three or more means apply step 4 and 5 to this new group.

Step 6: Calculate the sum of squares of deviations from the group mean and the corresponding mean square for each group of three or more means resulting from step 5. Using $s^2_{\bar{x}}$ as denominator calculate the variance ratio and apply the F test. If the ratio is found significant we assert that there is an overall difference among the means of that group.

CHAPTER TWO

THE MULTI-DIMENSIONAL CASE

2.1 Definitions and fundamental assumptions

Suppose we have k groups (of objects) G_1, G_2, ... G_k, of which samples of size n_1, n_2, ... n_k are taken; p normally distributed characters are measured on these objects. We denote the sample means of the characters by

$$\begin{array}{llll} \bar{x}_{11}, & \bar{x}_{12}, & \cdots, & \bar{x}_{1k} \\ \bar{x}_{21}, & \bar{x}_{22}, & \cdots & \\ \vdots & & & \\ \bar{x}_{p1}, & \cdots & , & \bar{x}_{pk} \end{array}$$

Throughout this paper we will assume that the covariance matrix (α_{ij}) of these measurements is known or estimated for a large number of degrees of freedom. We denote the inverse of this matrix by (α^{ij}).

We will make extensive use of the following statistic

$$V = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij} \sum_{r=1}^{k} n_r \, (\bar{x}_{ir} - \bar{x}_i)(\bar{x}_{jr} - \bar{x}_j)$$

where $\bar{x}_i = \dfrac{\sum_{r=1}^{k} n_r \bar{x}_{ir}}{n}$ and $n = \sum_{r=1}^{k} n_r$

If we let k = 2 , we get, after some manipulation

$$V = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij} \, (\bar{x}_{i1} - \bar{x}_{i2})(\bar{x}_{j1} - \bar{x}_{j2})$$

or

$$V = \frac{n_1 n_2}{n_1 + n_2} \, D^2$$

D^2 is known as the Mahalanobis distance. D^2 is, to some extent, a measure of the distance between two groups. V is a generalization of Mahalanobis D^2. D^2 was shown to be distributed as χ^2 with p degrees of freedom, and V as χ^2 with $p(k-1)$ degrees of freedom.

2.2 A possible approach to the multidimensional problem

A. Generalization of Tukey's method

The statistic V can be used to test the null hypothesis that the groups belong to the same multinormal population as follows:

If the observed value of V is larger than the tabulated χ^2 with $p(k-1)$ d.f. at the α level of significance, we reject the null hypothesis and assert that the groups do not all belong to the same population.

We are then left with the problem of classifying the groups into clusters of groups belonging to the same population.

As stated in Chapter 1, we will try to generalize Tukey's method and more particularily step 4 of his procedure. Tukey uses in his test the extreme deviate from the grand mean, the exact distribution of which is given by K. R. Nair. We will use the extreme generalized distance from the centroid of all the groups, and our problem will be then to find the distribution of such a distance.

B. Generalization of Nair's approach to the distribution of the extreme deviate from the sample mean

I Nair's distribution

We will give here only a short account of Nair's work (ref. 10).

In order to find the distribution of the extreme deviate among the ordered normal N (0,1) variates $x_1 \ldots x_n$ from their mean $\bar{x}$, Nair writes down the joint distribution of the x's

$$\frac{n!}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \prod_{i=1}^{n} dx_i$$

By a suitable orthogonal transformation this reduces to

$$\frac{n}{(\sqrt{2\pi})^n} \exp\left(-\frac{n\bar{x}^2}{2} - \frac{1}{2}\sum_{i=1}^{n-1} \frac{z_i^2}{i(i+1)}\right) d\bar{x} \prod_{i=1}^{n-1} dz_i$$

and integrating out $\bar{x}$ he gets

$$\frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1} \frac{z_i^2}{i(i+1)}\right) \prod_{i=1}^{n-1} dz_i$$

where it may be shown that $z_{n-1} = n(x_n - \bar{x}) = nu$

u being the extreme deviate from the mean $\bar{x}$, and

$$0 \leqslant z_1 \leqslant \cdots \leqslant z_{n-2} \leqslant z_{n-1} = nu$$

The distribution of u may then be obtained by integrating out $z_1, \ldots, z_{n-2}$ Finally the distribution of u can be written

$$f_n(u) = \frac{n\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \exp\left[\frac{-nu^2}{2(n-1)}\right] G_{n-2}(nu)$$

where

$$G_0(x) = 1 \qquad G_r(x) = \int_0^x \exp\left[\frac{-t^2}{2r(r+1)}\right] G_{r-1}(t)\, dt$$

II Generalization of Nair's approach

Under the null hypothesis that all the groups belong to the same population, all measurements have the same multinormal distribution. The means

$$\begin{array}{llll} \bar{x}_{11}, & \bar{x}_{12}, & \cdots & \bar{x}_{1k} \\ \bar{x}_{21} & & & \\ \vdots & & & \\ \bar{x}_{p1} & \cdots & & \bar{x}_{pk} \end{array}$$

from different groups are independent, but in general the observations from different characters are correlated. Since we know the covariance matrix (α_{ij}) we can replace the observations $(\bar{x}_{1r}, \cdots, \bar{x}_{pr})$ by linear combinations $(\bar{y}_{1r}, \cdots, \bar{y}_{pr})$ which are uncorrelated. The covariance matrix of the y's is a diagonal matrix denoted by (λ_i) Moreover we can assume without loss of generality that the true centroid of the distribution is $\mu_1 = 0$, $\mu_2 = 0$, $\cdots$ $\mu_p = 0$ The joint distribution of the $\bar{y}$'s is then

$$f(\bar{y}_{11}, \ldots, \bar{y}_{p1}; \cdots; \bar{y}_{1k}, \ldots, \bar{y}_{pk}) \prod_{i=1}^{p} \prod_{r=1}^{k} d\bar{y}_{ir} =$$

$$= C \exp\left(-\frac{1}{2} \sum_{r=1}^{k} n_r \sum_{i=1}^{p} \frac{1}{\lambda_i} y_{ir}^2\right) \prod_{i=1}^{p} \prod_{r=1}^{k} d\bar{y}_{ir}$$

where

$$C = \prod_{i=1}^{p} \prod_{r=1}^{k} \sqrt{\frac{n_r}{2\pi\lambda_i}}$$

Now,

$$\sum_{r=1}^{k} n_r \sum_{i=1}^{p} \frac{1}{\lambda_i} \bar{y}_{ir}^2 = \sum_{r=1}^{k} n_r \sum_{i=1}^{p} \frac{1}{\lambda_i} (\bar{y}_{ir} - \bar{y}_i)^2 + n \sum_{i=1}^{p} \frac{1}{\lambda_i} \bar{y}_i^2$$

Note that

$$\frac{1}{n_r} V_r = \sum_{i=1}^{p} \frac{1}{\lambda_i} (\bar{y}_{ir} - \bar{y}_i)^2$$

is the Mahalanobis distance between the r^{th} group and the observed centroid of all the groups. The largest of these V_r 's is thus the extreme distance the distribution of which we want to find.

We can write.

$$f(\bar{y}_{11},\dots,\bar{y}_{p1};\dots;\bar{y}_{1k},\dots,\bar{y}_{pk})\prod_{i=1}^{p}\prod_{r=1}^{k}d\bar{y}_{ir} = C\exp\left(-\tfrac{1}{2}\sum_{r=1}^{k}V_r - \tfrac{1}{2}n\sum_{i=1}^{p}\tfrac{1}{\lambda_i}y_i^2\right)\prod_{i=1}^{p}\prod_{r=1}^{k}d\bar{y}_{ir}$$

Consider now the following orthogonal transformation

$$u_i = \frac{1}{\sqrt{k}}\sum_{r=1}^{k}\bar{y}_{ir} = \sqrt{k}\,\bar{y}_i \qquad i = 1 \text{ to } p$$

$$v_{ij} = \frac{1}{\sqrt{j(j+1)}}\left(j\,\bar{y}_{i,j+1} - \sum_{s=1}^{j}\bar{y}_{is}\right) \qquad j = 1 \text{ to } k-1$$

The inverse transformation is

$$\bar{y}_{ir} = \frac{1}{\sqrt{k}}u_i + \frac{r-1}{\sqrt{(r-1)r}}v_{i,r-1} - \sum_{j=r}^{k-1}\frac{1}{\sqrt{j(j+1)}}v_{ij}$$

and we note that

$$V_r = n_r\sum_{i=1}^{p}\frac{1}{\lambda_i}\left[\frac{r-1}{\sqrt{(r-1)r}}v_{i,r-1} - \sum_{j=r}^{k-1}\frac{1}{\sqrt{j(j+1)}}v_{ij}\right]^2$$

is independent of the u's .

The distribution of the u's and v's is then

$$g(u_1,\dots,u_p;v_{11},\dots,v_{p,k-1})\prod_{i=1}^{p}du_i\prod_{i=1}^{p}\prod_{j=1}^{k}dv_{ij} =$$

$$= C\exp\left(-\tfrac{1}{2}\sum_{r=1}^{k}V_r - \tfrac{1}{2}\sum_{i=1}^{p}\frac{u_i^2}{k\lambda_i}\right)\prod_{i=1}^{p}du_i\prod_{i=1}^{p}\prod_{j=1}^{k-1}dv_{ij}$$

Integrating out the u's from $-\infty$ to $+\infty$ we get

$$h(v_{11},\dots v_{p,k-1})\prod_{i=1}^{p}\prod_{j=1}^{k-1}dv_{ij} =$$

$$= C_1\exp\left(-\tfrac{1}{2}\sum_{r=1}^{k}V_r\right)\prod_{i=1}^{p}\prod_{j=1}^{k-1}dv_{ij}$$

where

$$C_1 = \left(\frac{k}{n}\right)^{\frac{p}{2}} \left(\frac{1}{2\pi}\right)^{\frac{p(k-1)}{2}} \frac{\prod_{r=1}^{k} n_r^{\frac{p}{2}}}{\prod_{i=1}^{p} \lambda_i^{\frac{k-1}{2}}}$$

The problem is now to find the joint distribution of the V_r's.

2.3 Two methods for obtaining the joint distribution of the generalized distances of the groups from their observed centroid

A. Introduction of additional variables

Considering the distribution obtained in 2.2

$$h(v_{11}, \dots, v_{p,k-1}) = C_1 \exp\left(-\frac{1}{2}\sum_{r=1}^{k} V_r\right) \prod_{i=1}^{p}\prod_{j=1}^{k-1} dv_{ij}$$

where the V_r's are functions of the v's , it will be noticed that we have p (k-1)v's but only k V_r's . A change of variables from the v's to the V_r's as they stand is therefore not possible.

A device sometimes used under such circumstances would be to introduce additional V_r's which are functions of the v's and integrate them out later on. We would then be left with the desired joint distribution of the V_r's . This method proves successful in the special case $p = 2$, $k = 3$ which is considered in Chapter 3. In other cases the integrations of the additional variables could not be performed. Numerical integration is obviously not applicable here.

B. The Characteristic function of the distribution

The joint characteristic function of the functions $V_1 \dots V_k$ of the variables $v_{11} \dots v_{p(k-1)}$ is defined to be the expected value of $\exp\left(i \sum_{r=1}^{k} t_r V_r \right)$ that is

$$C_1 \int_{-\infty}^{\infty} \dots \int \exp\left(i \sum_{r=1}^{k} t_r V_r \right) h\left(v_{11}, \dots, v_{p.(k-1)} \right) \prod_{i=1}^{p} \prod_{j=1}^{k-1} d v_{ij}$$

where C_1 and $h\ (v_{11}, \dots, v_{p(k-1)})$ are given in the preceding section.

If this Characteristic function turns out to be a known Fourier transform, the joint distribution of the $V_1 \ \dots \ V_k$ will be the inverse of this transform.

We consider first the special case $k = 3$, $p = 2$ and $n_1 = n_2 = n_3 = n_0$. The Characteristic function is then

$$\varphi_2\ (t_1, t_2, t_3) =$$

$$= C_1 \int\int\int\int\int_{-\infty}^{\infty} \exp\left\{ n_0 \left[V_1 (it_1 - \tfrac{1}{2}) + V_2 (it_2 - \tfrac{1}{2}) + V_3 (it_3 - \tfrac{1}{2}) \right] \right\} dv_{11}\, dv_{12}\, dv_{21}\, dv_{22}$$

where $$C_1 = \left(\frac{1}{2\pi}\right)^2 \frac{n_0^2}{\lambda_1 \lambda_2}$$

Expressing the V's in terms of the v's , $\varphi_2(t_1, t_2, t_3)$ becomes

$$C_1 \int\int_{-\infty}^{\infty} \exp\left\{ n_0 \frac{1}{\lambda_1} \left[\left(\frac{1}{\sqrt{2}} v_{11} + \frac{1}{\sqrt{6}} v_{12} \right)^2 (it_1 - \tfrac{1}{2}) + \left(\frac{1}{\sqrt{2}} v_{11} - \frac{1}{\sqrt{6}} v_{12} \right)^2 (it_2 - \tfrac{1}{2}) + \frac{2}{3} v_{12}^2 (it_3 - \tfrac{1}{2}) \right] \right\} dv_{11}\, dv_{12}$$

$$\times \int\int_{-\infty}^{\infty} \exp\left\{ n_0 \frac{1}{\lambda_2} \left[\left(\frac{1}{\sqrt{2}} v_{21} + \frac{1}{\sqrt{6}} v_{22} \right)^2 (it_1 - \tfrac{1}{2}) + \left(\frac{1}{\sqrt{2}} v_{21} - \frac{1}{\sqrt{6}} v_{22} \right)^2 (it_2 - \tfrac{1}{2}) + \frac{2}{3} v_{22}^2 (it_3 - \tfrac{1}{2}) \right] \right\} dv_{21}\, dv_{22}$$

These two double integrations are quite similar except for minor changes in the constants.

Omitting all factors not involving v_{11} , the integration with respect to v_{11} reduces to

$$I_1 = \int_{-\infty}^{\infty} \exp\left\{ n_0 \frac{1}{\lambda_1}\left[\left(\tfrac{1}{2} v_{11}^2 + \tfrac{1}{\sqrt{3}} v_{11} v_{12}\right)\left(it_1 - \tfrac{1}{2}\right) + \left(\tfrac{1}{2} v_{11}^2 - \tfrac{1}{\sqrt{3}} v_{11} v_{12}\right)\left(it_2 - \tfrac{1}{2}\right)\right]\right\} dv_{11}$$

which after some manipulations yields

$$I_1 = K_1 \exp\left[- \frac{n_0}{6\lambda_1} \frac{(it_1 - it_2)^2 v_{12}^2}{(it_1 + it_2 - 1)}\right]$$

where

$$K_1 = \frac{\sqrt{2\pi\lambda_1}}{\sqrt{n_0 (it_1 + it_2 - 1)}}$$

Integrating with respect to v_{12} we get

$$I_2 = \int_{-\infty}^{\infty} \exp\left\{ \frac{n_0}{6\lambda_1} v_{12}^2 \left[it_1 + it_2 + 4\, it_3 - 3 - \frac{(it_1 - it_2)^2}{(it_1 + it_2 - 1)}\right]\right\} dv_{12}$$

This in turn yields

$$I_2 = K_2 = \frac{-\sqrt{6\pi\lambda_1}}{\sqrt{n_0\left[it_1 + it_2 + 4\, it_3 - 3 - \frac{(it_1 - it_2)^2}{(it_1 + it_2 - 1)}\right]}}$$

The first double integration gives us

$$K_1 K_2 = \frac{\sqrt{12}\ \pi \lambda_1}{n_0 \sqrt{\left[it_1 + it_2 + 4\, it_3 - 3 - \frac{(it_1 - it_2)^2}{(it_1 + it_2 - 1)}\right](it_1 + it_2 - 1)}}$$

The second double integration is performed similarly and given an analogous expression with λ_1 replaced by λ_2 Finally the characteristic function is found to be

$$\varphi_2(t_1, t_2, t_3,) = \frac{3}{(it_1 + it_2 - 1)\left[it_1 + it_2 + 4\, it_3 - 3 - \frac{(it_1 - it_2)^2}{(it_1 + it_2 - 1)}\right]}$$

This expression could be simplified, but for the purpose of generalization it is convenient to leave it in this form. This function generalizes readily to 3 groups of p characters giving $\varphi_p(t_1, t_2, t_3) = \left[\varphi_2(t_1, t_2, t_3)\right]^{\frac{p}{2}}$

In a similar fashion the special cases $k = 4$ and $k = 5$ were worked out and a pattern was observed which enabled us to write the characteristic function $\varphi_p(t_1, t_2, \ldots, t_k)$ as follows

$$\varphi_p(t_1, t_2, \cdots, t_k) = \frac{1}{2^{p(k-1)}} \left\{ \prod_{r=1}^{k-1} \frac{(-1)^{k-1} r(r+1)}{a_r} \right\}^{\frac{p}{2}}$$

where

$$a_r = \sum_{s=1}^{r} it_s + r^2 it_{r+1} - \frac{r(r+1)}{2} - \sum_{h=1}^{r-1} d_h$$

$$d_h = \frac{b_h^2}{a_h}$$

$$b_h = \sum_{s=1}^{h} it_s - h\,it_{h+1} - \sum_{l=1}^{h-1} d_l$$

$$b_0 = 0 \qquad d_0 = 0$$

This Characteristic function applies generally except for the restriction $n_1 = n_2 = \cdots n_k = n_0$. For $k > 3$ $\varphi_p(t_1, t_2, \cdots, t_k)$ becomes very complicated and quite hard to handle. But even for the simplest case $k = 3$ we were unable to recognize $\varphi_p(t_1, t_2, t_3)$ as a familiar Fourier transform.

In Chapter 3, the joint distribution in the special case $k = 3$, $p = 2$ is shown to be

$$f(V_1, V_2, V_3)\, dV_1\, dV_2\, dV_3 = \frac{3}{4\pi} \frac{\exp\left[-\frac{1}{2}(V_1 + V_2 + V_3)\right]}{\sqrt{2V_1V_2 + 2V_2V_3 + 2V_3V_1 - V_1^2 - V_2^2 - V_3^2}}\, dV_1\, dV_2\, dV_3$$

Formally, this is the inverse transform of $\varphi_2(t_1, t_2, t_3)$.

CHAPTER THREE

A SPECIAL CASE AND ITS SOLUTION

3.1 The joint distribution of the generalized distances of the groups from their observed centroid

Specializing the results of . 2.2 B II to the special case p = 2 k = 3 , the joint distribution of

$(\bar{y}_{11}, \bar{y}_{12}, \bar{y}_{13}, \bar{y}_{21}, \bar{y}_{22}, \bar{y}_{23})$ is

$$C \exp\left(-\frac{1}{2}\sum_{r=1}^{3} V_r - \frac{1}{2} n \sum_{i=1}^{2} \frac{1}{\lambda_i} \bar{y}_i^2\right) \prod_{i=1}^{2} \prod_{r=1}^{3} d\bar{y}_{ir}$$

where

$$V_r = n_r \sum_{i=1}^{2} \frac{1}{\lambda_i} (\bar{y}_{ir} - \bar{y}_i)^2$$

and

$$C = \left(\frac{1}{2\pi}\right)^3 \frac{n_1 n_2 n_3}{(\lambda_1 \lambda_2)^{\frac{3}{2}}}$$

Consider the orthogonal transformation

$$u_1 = \frac{1}{\sqrt{3}}(\bar{y}_{11} + \bar{y}_{12} + \bar{y}_{13}) \qquad u_2 = \frac{1}{\sqrt{3}}(\bar{y}_{21} + \bar{y}_{22} + \bar{y}_{23})$$

$$v_{11} = \frac{1}{\sqrt{2}}(-\bar{y}_{11} + \bar{y}_{12}) \qquad v_{21} = \frac{1}{\sqrt{2}}(-\bar{y}_{21} + \bar{y}_{22})$$

$$v_{12} = \frac{1}{\sqrt{6}}(-\bar{y}_{11} - \bar{y}_{12} + 2\bar{y}_{13}) \qquad v_{22} = \frac{1}{\sqrt{6}}(-\bar{y}_{21} - \bar{y}_{22} + 2\bar{y}_{23})$$

of which the inverse transformation is

$$\bar{y}_{11} = \frac{1}{\sqrt{3}} u_1 - \frac{1}{\sqrt{2}} v_{11} - \frac{1}{\sqrt{6}} v_{12} \qquad \bar{y}_{21} = \frac{1}{\sqrt{3}} u_2 - \frac{1}{\sqrt{2}} v_{21} - \frac{1}{\sqrt{6}} v_{22}$$

$$\bar{y}_{12} = \frac{1}{\sqrt{3}} u_1 + \frac{1}{\sqrt{2}} v_{11} - \frac{1}{\sqrt{6}} v_{12} \qquad \bar{y}_{22} = \frac{1}{\sqrt{3}} u_2 + \frac{1}{\sqrt{2}} v_{21} - \frac{1}{\sqrt{6}} v_{22}$$

$$\bar{y}_{13} = \frac{1}{\sqrt{3}} u_1 + \frac{2}{\sqrt{6}} v_{12} \qquad \bar{y}_{23} = \frac{1}{\sqrt{3}} u_2 + \frac{2}{\sqrt{6}} v_{22}$$

The distribution of the u's and the v's is

$$C \exp\left(-\frac{1}{2}\sum_{r=1}^{3} V_r - \frac{1}{2} n \sum_{i=1}^{2} \frac{1}{\lambda_i} \frac{u_i^2}{3}\right) \prod_{i=1}^{2} \prod_{j=1}^{2} dv_{ij} \prod_{i=1}^{2} du_i$$

where the V's are functions of the v's only.

Integrating out the u's we get

$$C \exp\left(-\tfrac{1}{2}\sum_{r=1}^{3} V_r\right) \prod_{i=1}^{2}\prod_{j=1}^{2} dv_{ij} \iint_{-\infty}^{\infty} \exp\left[-\frac{n}{6}\left(\frac{1}{\lambda_1}u_1^2 + \frac{1}{\lambda_2}u_2^2\right)\right] du_1\, du_2 =$$

$$= C_1 \exp\left(-\tfrac{1}{2}\sum_{r=1}^{3} V_r\right) \prod_{i=1}^{2}\prod_{j=1}^{2} dv_{ij} = C_1 \exp\left[-\tfrac{1}{2}(V_1+V_2+V_3)\right] dv_{11}\, dv_{12}\, dv_{21}\, dv_{22}$$

where $C_1 = \left(\frac{1}{2\pi}\right)^2 \frac{3\, n_1 n_2 n_3}{n\,\lambda_1 \lambda_2}$

Call $\frac{n_1\, n_2\, n_3}{n} = N$ and consider the transformation

$$V_1' = \frac{n_2 n_3}{n} V_1 = N\left[\frac{1}{\lambda_1}\left(\frac{1}{\sqrt{2}}v_{11} + \frac{1}{\sqrt{6}}v_{12}\right)^2 + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{2}}v_{21} + \frac{1}{\sqrt{6}}v_{22}\right)^2\right]$$

$$V_2' = \frac{n_1 n_3}{n} V_2 = N\left[\frac{1}{\lambda_1}\left(\frac{1}{\sqrt{2}}v_{11} - \frac{1}{\sqrt{6}}v_{12}\right)^2 + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{2}}v_{21} - \frac{1}{\sqrt{6}}v_{22}\right)^2\right]$$

$$V_3' = \frac{n_1 n_2}{n} V_3 = N\left[\frac{1}{\lambda_1}\left(\frac{2}{\sqrt{6}}v_{12}\right)^2 + \frac{1}{\lambda_2}\left(\frac{2}{\sqrt{6}}v_{22}\right)^2\right]$$

$$V_4' = \quad = N\,\frac{1}{\lambda_1}\left(\frac{1}{\sqrt{2}}v_{11}\right)^2$$

where V_4^1 is a new variable introduced in order to perform the change of variables from the v's to the V's .

V_4^1 satisfies the inequality $0 \leqslant V_4' \leqslant \frac{1}{2}V_1' + \frac{1}{2}V_2' - \frac{1}{4}V_3'$ After lengthy algebraic manipulations the inverse transformation is found to be

$$\frac{1}{2}\,\frac{1}{\lambda_1}\,N\,v_{11}^2 = V_4'$$

$$\frac{1}{2}\,\frac{1}{\lambda_2}\,N\,v_{21}^2 = \frac{1}{2}V_1' + \frac{1}{2}V_2' - \frac{1}{4}V_3' - V_4'$$

$$(1) \qquad \sqrt{\frac{2}{3}\,\frac{1}{\lambda_1}\,N}\;v_{12} = \pm\left[\frac{(V_1'-V_2')\sqrt{V_4'} \pm \sqrt{\left(\frac{1}{2}V_1'+\frac{1}{2}V_2'-\frac{1}{4}V_3'-V_4'\right)\left(2V_1'V_2'+2V_2'V_3'+2V_3'V_1'-V_1'^2-V_2'^2-V_3'^2\right)}}{V_1'+V_2'-\frac{1}{2}V_3'}\right]$$

$$(2) \qquad \sqrt{\frac{2}{3}\,\frac{1}{\lambda_2}\,N}\;v_{22} = \pm\left[\frac{(V_1'-V_2')\sqrt{\frac{1}{2}V_1'+\frac{1}{2}V_2'-\frac{1}{4}V_3'-V_4'} \pm \sqrt{V_4'\left(2V_1'V_2'+2V_2'V_3'+2V_3'V_1'-V_1'^2-V_2'^2-V_3'^2\right)}}{V_1'+V_2'-\frac{1}{2}V_3'}\right]$$

To eliminate extraneous solutions the following restrictions on the signs in (1) and (2) must be introduced: the signs in front of the expressions (1) and (2) are the signs of v_{11} and v_{21} respectively; the signs in front of the root sign in the expressions (1) and (2) must be opposite.

The Jacobian of this transformation is

$$J = \frac{3\sqrt{3}\ \lambda_1^2 \lambda_2^2}{8N^4 \left(v_{12} v_{21}^2 v_{11} - v_{11}^2 v_{22} v_{21} \right)}$$

or in terms of the V's

$$J = \frac{3 \lambda_1 \lambda_2}{N^2 \sqrt{V_4' \left(\frac{1}{2} V_1' + \frac{1}{2} V_2' - \frac{1}{4} V_3' - V_4' \right) \left(2 V_1' V_2' + 2 V_2' V_3' + 2 V_3' V_1' - V_1'^2 - V_2'^2 - V_3'^2 \right)}}$$

The joint distribution of (V_1^1 V_2^1 V_3^1 V_4^1) is then found to be

$$C_2 \ \frac{\exp\left[-\frac{1}{2}(V_1 + V_2 + V_3) \right] \qquad dV_1'\, dV_2'\, dV_3'\, dV_4'}{\sqrt{2 V_1' V_2' + 2 V_2' V_3' + 2 V_3' V_1' - V_1'^2 - V_2'^2 - V_3'^2}\ \sqrt{V_4' \left(\frac{1}{2} V_1' + \frac{1}{2} V_2' - \frac{1}{4} V_3' - V_4' \right)}}$$

where

$$C_2 = \frac{9}{4\pi^2 N}$$

To find the joint distribution of (V_1^1 V_2^1 V_3^1) we integrate out V_4^1 over its range which is o to $\frac{1}{2}V_1' + \frac{1}{2}V_2' - \frac{1}{4}V_3'$:

$$C_2 \ \frac{\exp\left[-\frac{1}{2}(V_1 + V_2 + V_3) \right] \ dV_1'\, dV_2'\, dV_3'}{\sqrt{2 V_1' V_2' + 2 V_2' V_3' + 2 V_3' V_1' - V_1'^2 - V_2'^2 - V_3'^2}} \int_0^{\frac{1}{2}V_1' + \frac{1}{2}V_2' - \frac{1}{4}V_3'} \frac{dV_4'}{\sqrt{V_4' \left(\frac{1}{2} V_1' + \frac{1}{2} V_2' - \frac{1}{4} V_3' - V_4' \right)}}$$

This integration is easily performed and gives

$$\int_0^{\frac{1}{2}V_1' + \frac{1}{2}V_2' - \frac{1}{4}V_3'} \frac{dV_4'}{\sqrt{V_4' \left(\frac{1}{2} V_1' + \frac{1}{2} V_2' - \frac{1}{4} V_3' - V_4' \right)}} = \pi$$

The joint distribution of (V_1 V_2 V_3) is then found to be

$$\frac{q}{4\pi n} \quad \frac{\exp\left[-\frac{1}{2}(V_1+V_2+V_3)\right] \; dV_1\, dV_2\, dV_3}{\sqrt{\frac{2}{n_1 n_2} V_1 V_2 + \frac{2}{n_2 n_3} V_2 V_3 + \frac{2}{n_3 n_1} V_3 V_1 - \frac{1}{n_1^2} V_1^2 - \frac{1}{n_2^2} V_2^2 - \frac{1}{n_3^2} V_3^2}}$$

3.2 The distribution of the extreme deviate from the centroid

Let us restrict the problem further by assuming the number of observations to be the same for all groups, i.e. $n_1 = n_2 = \dots = n_k$, Call n_0 this common value.

The joint distribution of V_1, V_2, V_3 in this case specializes to

$$f(V_1, V_2, V_3) = \frac{3}{4\pi} \quad \frac{\exp\left[-\frac{1}{2}(V_1+V_2+V_3)\right]}{\sqrt{2V_1V_2 + 2V_2V_3 + 2V_3V_1 - V_1^2 - V_2^2 - V_3^2}}$$

The variates V_1, V_2, V_3 are always positive and it is easy to check that V_1, V_2, V_3 do not assume values outside the cone defined by

$$(\ast) \qquad 2V_1V_2 + 2V_2V_3 + 2V_3V_1 - V_1^2 - V_2^2 - V_3^2 \geqslant 0$$

The distribution f (V_1, V_2, V_3) is therefore always real and positive.

We can assume without loss of generality that the variates have been ordered say $0 \leqslant V_1 \leqslant V_2 \leqslant V_3 \leqslant t$. The density of these ordered variates is $3!\, f$ (V_1, V_2, V_3).

We are interested in the distribution of the extreme deviate from the centroid, V_3 ; in other words we want to find the probability G (t) that $V_3 \leqslant t$.

$$G(t) = \left(\frac{3}{4\pi}\right) 3! \int_{V_3=0}^{t} \int_{V_2=\frac{V_3}{4}}^{V_3} \int_{V_1=(\sqrt{V_3}-\sqrt{V_2})^2}^{V_2} \frac{\exp\left[-\frac{1}{2}(V_1+V_2+V_3)\right] dV_1\, dV_2\, dV_3}{\sqrt{2V_1V_2+2V_2V_3+2V_3V_1-V_1^2-V_2^2-V_3^2}}$$

The lower limits for V_1 and V_2 are obtained from the restriction (*) on the variates and the inequalities $0 \leqslant V_1 \leqslant V_2 \leqslant V_3 \leqslant t$. G (t) is well defined by the above expression but the integration is hard to perform and not suitable for numerical integration. In order to remedy this state of affairs consider the orthogonal transformation

$$w = \frac{1}{\sqrt{3}}\left(V_1 + V_2 + V_3\right)$$

$$u = \frac{1}{\sqrt{2}}\left(-V_1 + V_2\right)$$

$$v = \frac{1}{\sqrt{6}}\left(-V_1 - V_2 + 2V_3\right)$$

the inverse of which is

$$V_1 = \frac{1}{\sqrt{3}} w - \frac{1}{\sqrt{2}} u - \frac{1}{\sqrt{6}} v$$

$$V_2 = \frac{1}{\sqrt{3}} w + \frac{1}{\sqrt{2}} u - \frac{1}{\sqrt{6}} v$$

$$V_3 = \frac{1}{\sqrt{3}} w + \frac{2}{\sqrt{6}} v$$

Under this transformation the distribution

$$3!\, f(V_1, V_2, V_3)\, dV_1\, dV_2\, dV_3$$ becomes

$$f'(u, v, w)\, dv\, du\, dw = \frac{9}{2\pi} \frac{\exp\left(-\frac{\sqrt{3}}{2} w\right)}{\sqrt{w^2 - 2u^2 - 2v^2}}\, du\, dv\, dw$$

Follow this by the transformation

$$w = \frac{2}{\sqrt{3}} \zeta$$

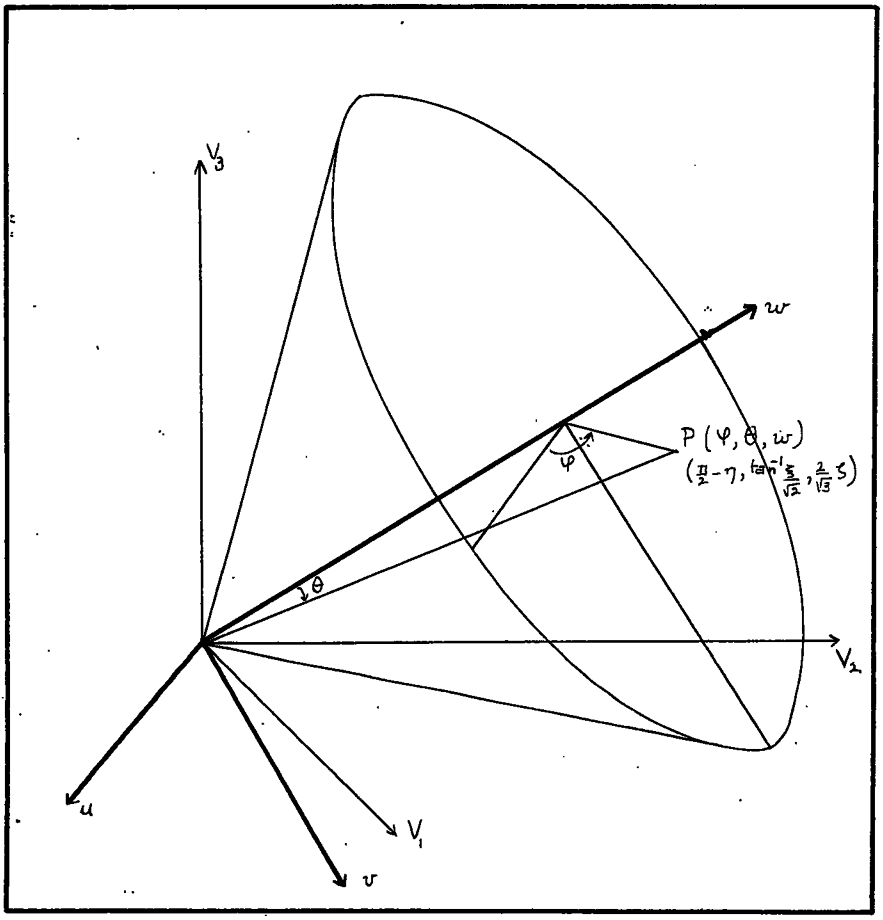
$$u = \sqrt{\frac{2}{3}}\, \zeta \xi \sin \eta$$

$$v = \sqrt{\frac{2}{3}}\, \zeta \xi \cos \eta$$

The Jacobian of this transformation is $\quad |J| = \frac{4}{3\sqrt{3}} \zeta^2 \xi$

and we find the distribution of ζ, ξ, η to be

$$g(\zeta, \xi, \eta)\, d\zeta\, d\xi\, d\eta = \frac{3}{\pi} \frac{\xi}{\sqrt{1-\xi^2}}\, \zeta \exp(-\xi)\, d\zeta\, d\xi\, d\eta$$


V_3
w
$P(\varphi, \theta, w)$
$(\frac{\pi}{2}-\eta, \tan^{-1}\frac{\xi}{\sqrt{2}}, \frac{2}{\sqrt{3}}\zeta)$
φ
θ
V_2
u
V_1
v

(fig. 1)

This change of variable is, roughly speaking, a change to cylindrical coordinates as shown in (fig. 1) where we have set $\varphi = \frac{\pi}{2} - \eta$, $\theta = \tan^{-1}\frac{\xi}{\sqrt{2}}$, $w = \frac{2}{\sqrt{3}}\zeta$

The transformation is defined and single valued if $\zeta \neq 0$ and $\xi < \infty$ $(\theta \neq \frac{\pi}{2})$. It can be easily verified that the angle at the vertex of the cone is $\frac{\pi}{2}$ so that $\theta \leqslant \frac{\pi}{4}$ and ξ is therefore always finite. The ranges of ζ, ξ, η taken independently are $0 \leqslant \zeta < \infty$, $0 \leqslant \xi \leqslant 1$ $0 \leqslant \eta \leqslant 2\pi$ However if we let $V_3 \leqslant t$, the limits on ζ, ξ and η are no longer independent, for

$$V_3 = \frac{1}{\sqrt{3}} w + \sqrt{\frac{3}{2}}\, v = \frac{2}{3}\zeta\left(1 + \xi\cos\eta\right) \leqslant t$$

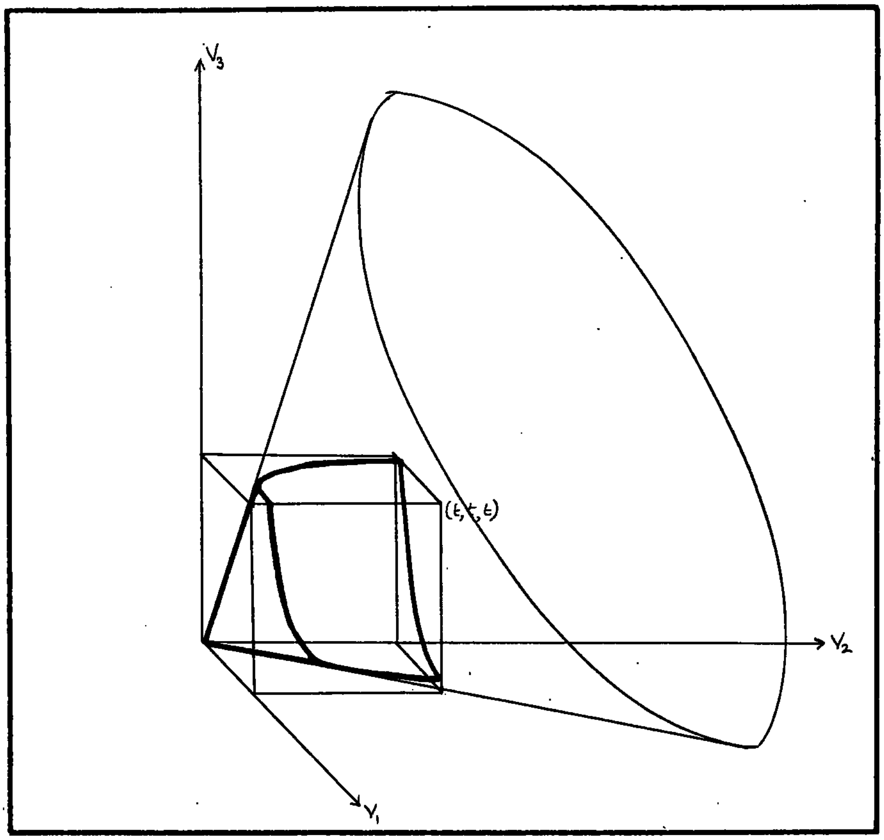
and therefore
$$\zeta \leqslant \frac{3t}{2(1 + \xi\cos\eta)}$$

The inequality $V_1 \leqslant V_2 \leqslant V_3$ give limits for η : $V_1 \leqslant V_2$ implies $u \geqslant 0$ or $0 \leqslant \eta \leqslant \pi$, $V_2 \leqslant V_3$ implies $-\frac{1}{\sqrt{2}}u + \sqrt{\frac{3}{2}}\, v \geqslant 0$ or $\tan\eta \leqslant \sqrt{3}$ and the limits on are thus $0 \leqslant \eta \leqslant \frac{\pi}{3}$

The probability G (t) of getting $0 \leqslant V_1 \leqslant V_2 \leqslant V_3 \leqslant t$ becomes finally

$$G(t) = \frac{3}{\pi}\int_{\xi=0}^{1} \frac{\xi}{\sqrt{1-\xi^2}} \int_{\eta=0}^{\frac{\pi}{3}} \int_{\zeta=0}^{\frac{3t}{2(1+\xi\cos\eta)}} \zeta \exp(-\zeta)\, d\zeta\, d\eta\, d\xi$$

The region of integration for the unordered V's is shown in (fig. 2)



(fig.2)

The integration with respect to ζ gives

$$\int_{\zeta=0}^{\frac{3t}{2(1+\xi\cos\eta)}} \zeta \exp(-\zeta)\,d\zeta = 1 - \exp\left[\frac{-3t}{2(1+\xi\cos\eta)}\right]\left[\frac{3t}{2(1+\xi\cos\eta)}+1\right]$$

The other integrations have to be evaluated numerically.

three decimal places.

The double numerical integration:

$$\int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} \int_0^{\frac{\pi}{3}} \left[\frac{3t}{2(1+\xi\cos\eta)} + 1\right] \exp\left[\frac{-3t}{(1+\xi\cos\eta)}\right] d\eta \, d\xi$$

was performed using seven equally spaced values of η : $\eta_i = \frac{i\pi}{18}$, i = o to 6 and eleven equally spaced values of ξ : $\xi_i = \frac{i}{10}$, i = o to 10. The function

$$f(\eta,\xi) = \left[\frac{3t}{2(1+\xi\cos\eta)} + 1\right] \exp\left[\frac{-3t}{2(1+\xi\cos\eta)}\right]$$

was first evaluated for all seventy seven pairs of values (η_i, ξ_j) using the "Tables of the Exponential Function e^x." prepared by the National Bureau of Standards. The integration with respect to η keeping ξ fixed, $\int_0^{\frac{\pi}{3}} f(\eta,\xi_j) d\eta$ was performed using Cotes' numbers, thus fitting a polynomial equation of degree six to the seven points $[\eta_i, f(\eta_i,\xi_j)]$ i = o to 6. This was done for the eleven values of ξ. The first integration thus yields eleven values of a function $A(\xi)$ Using the method of least squares and a set of orthogonal polynomials, we constructed a polynomial $P(\xi)$ fitting the eleven points $A(\xi_i)$, i = o to 10, thus approximately $A(\xi)$.

The integration $\int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} P(\xi)\, d\xi =$

$$= \int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} \left(a_0 + a_1\xi + a_2\xi^2 + a_3\xi^3 + a_4\xi^4 + a_5\xi^5\right) d\xi$$

can be split into six integrations each of which can be performed by using tables of the Beta-function (e.g.: "Tables of the Incomplete Beta-Function" edited by Karl Pearson).

At least six decimal places were carried throughout the computation but the accuracy is reduced by approximating $A(\xi)$ by $P(\xi)$. Bounds on G (t) can be given as follows

$$\min_i \left[\frac{A(\xi_i)}{P(\xi_i)}\right] \int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} P(\xi)\, d\xi \leqslant$$

$$\leqslant \int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} A(\xi)\, d\xi \leqslant$$

$$\leqslant \max_i \left[\frac{A(\xi_i)}{P(\xi_i)}\right] \int_0^1 \frac{\xi}{\sqrt{1-\xi^2}} P(\xi)\, d\xi$$

The results of this computation are summarized in the following table.

Table of the probability G(t) of getting a value for the extreme deviate at least as large as t

	t	G(t)	
(1)	2.667	.686	
	3.000	.743	*
	3.333	.791	*
	3.667	.832	*
(2)	4.000	.866	
	4.333	.894	*
	** 4.424	.900	
	4.667	.916	*
	5.000	.934	*
(3)	5.333	.948	
	** 5.394	.950	
	5.667	.959	*
	6.000	.967	*
	6.333	.974	*
(4)	6.667	.980	

* These probabilities were obtained by parabolic interpolation through the points (1) (2) (3) (4) .

** The values of t yielding a probability of .90 and .95 were obtained by linear interpolation.

Conclusion

The results obtained in this paper enable us to construct a solution to the following problem:

Samples of equal size n_0 are taken from three groups. Two normally distributed characters are measured on these objects, $\bar{x}_{11}, \ldots \bar{x}_{23}$ denoting the mean values of these measurements. The covariance matrix of these measurements $A = (\alpha_{ij})$ is known, or estimated on a large number of degrees of freedom. On the basis of these measurements decide whether the groups belong to the same population and if they do not, which are different from the others.

The solution we propose is as follows:

Step 1: Choose a level of significance α.

Step 2: Uncorrelate the measurements. To this end find the orthogonal matrix $B = (b_{ij})$ such that $BAB' = \Lambda$ where $\Lambda = (\lambda_i)$ is a diagonal matrix. Then perform a transformation from the x_{ij}'s to a new set of variates y_{ij}. Compute $\bar{y}_{11}, \ldots, \bar{y}_{23}$ the means of the new uncorrelated variates.

Step 3: Compute

$$V_1 = n_0 \sum_{i=1}^{2} \frac{1}{\lambda_i} (\bar{y}_{i1} - \bar{y}_i)^2$$

$$V_2 = n_0 \sum_{i=1}^{2} \frac{1}{\lambda_i} (\bar{y}_{i2} - \bar{y}_i)^2$$

$$V_3 = n_0 \sum_{i=1}^{2} \frac{1}{\lambda_i} (\bar{y}_{i3} - \bar{y}_i)^2$$

$$V = \sum_{r=1}^{3} V_r$$

where

$$\bar{y}_i = \frac{\sum_{r=1}^{3} \bar{y}_{ir}}{3}$$

Rank the V's say $V_1^1 \leqslant V_2^1 \leqslant V_3^1$

Step 4: Compare V to χ_α^2 with 4 d.f. If $V < \chi_{\alpha,4}^2$ the groups are asserted to belong to the same population, and the process terminates.

If $V \geqslant \chi_{\alpha,4}^2$ proceed to compare V_3^1 with t_α (tabular value given in Chapter 3). If $V_3^1 < t_\alpha$, no group is separated from the cluster although there is an overall difference among the groups, and the process terminates.

If $V_3^1 \geqslant t_\alpha$ we assert that at a level of significance α the group corresponding to V_3^1 does not belong to the same population as the other two groups.

Then proceed to compute $V' = \frac{n_0}{2}\sum_{i=1}^{2}\frac{1}{\lambda_i}(\bar{y}_{i1'} - \bar{y}_{i2'})^2$ and compare V' with χ_α^2 with 2 d.f.

If $V' < \chi_{\alpha,2}^2$ we assert that the groups corresponding to V_1^1 and V_2^1 belong to the same population.

If $V' \geqslant \chi_{\alpha,2}^2$ we assert that each of the groups belongs to a different population.

Although a solution is given only for the special case 3 groups - 2 characters, it covers a somewhat wider range of problems. In many instances the configuration of the mean values with respect to p characters can be preserved by representing the groups with respect to two suitably chosen functions of the p characters. Methods have been devised for constructing such functions and for testing the adequacy of the representation, (see for example ref. 1 p. 365). The main restriction to the solution given is thus in the number of groups.

As stated previously in the paper, the joint distribution of the deviates from the centroid in the general case is not readily available by the method used in the special case. We do suspect the form of the general distribution to be quite similar to that of the special case, but we have been unable to justify this guess so far.

We suggest that some more research could be carried in the following directions

(1) Try to increase the number of groups

(2) Try to increase the number of characters

(3) Try to invert the characteristic function of the joint distribution

(4) Guessing the joint distribution try to show that its characteristic function coincides with that given in Chapter 2

(5) Extend these results to the case where the covariance matrix (α_{ij}) is not known, that is, find the Studentized form of the distribution of the extreme deviate from the centroid of the groups.

BIBLIOGRAPHY

1. C. R. Rao, Advanced Statistical Methods in Biometric Research, J. Wiley & Sons (1952)

2. W. T. Federer, Experimental design, MacMillan Co. (1955)

3. D. Newman, The distribution of range in samples from a normal population, Biometrika 31 20-30, (1939)

4. E. S. Pearson & H. O. Hartley, Biometrika tables for statisticians, Cambridge University Press (1954)

5. D. B. Duncan, Multiple range and multiple F. test, Mimeo. Tech. Report no. 6, Va. Polytechnical Inst. (Sept. 1953)

6. J. W. Tukey, The problem of multiple comparisons, Ditto, Princeton University (1953)

7. D. B. Duncan, A significance test for difference between ranked treatments in an analysis of variance, Va. J. Sci 2: 171-189 (1951)

8. H. Scheffé, A method for judging all contrasts in the analysis of variance, Biometrika 40: 87-104, (1953)

9. J. W. Tukey, Comparing individual means in the analysis of variance, Biometrica 5: 99-114, (1949)

10. K. R. Nair, The distribution of the extreme deviate from the sample mean and its studentized form, Biometrika 35: 118-144 (1948)