THE APPLICATION OF DIRECTIONAL METHODS IN P DIMENSIONS

by

Jorge Holguin

B.Sc., Javeriana University, Bogota, Colombia, 1974

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics

SIMON FRASER UNIVERSITY

April 1980

APPROVAL

Name: Jorge Holguin

Degree: Master of Science

Title of Thesis: The application of directional methods in p dimensions.

Chairman: S.K. Thomason

M.A. Stephens
Senior Supervisor

C. Villegas

K.L. Weldon

R. Koopman
External Examiner
Associate Professor
Department of Psychology
Simon Fraser University

Date Approved: April 15, 1980

Title of Thesis/Dissertation:

The application of directional methods in p-dimensions.

Author: ______________________

(signature)

JORGE HOLGUIN

(name)

APRIL 10 - 1980

(date)

ABSTRACT

In recent years there have been many advances in the analysis of directional data expressed as two and three dimensional unit vectors. The extension of these methods of analysis to p dimensions is presented in this thesis. Although directions in higher dimensions do not have a physical interpretation, data which can be recorded as p-dimensional unit vectors arise in many areas; an example is when the data are in terms of p continuous proportions.

The von Mises and Fisher distributions have been widely used in the analysis of directional data in two and three dimensions respectively; these are described in Chapter 1. The extension of these distributions to higher dimensions is given in Chapter 2. Tests are given for hypotheses of interest in the analysis of groups of p-dimensional unit vectors. In directional uses, a common technique is a form of analysis of variance introduced by Watson. It is shown that this technique can be used with p-dimensional unit vectors. Further, it can be developed also for a two way layout with a natural extension to a multi-way layout. The analysis of variance can also be expressed in terms of angles between the vectors and their resultants, and this is a useful representation with many types of data.

It is also often of interest to examine the clustering of unit vectors; a simple method of clustering is given in Chapter 3. The results obtained in the examples are compared to those found by means of standard algorithms.

Chapter 4 is a chapter of worked data sets ; several examples are worked through to demonstrate both the p-dimensional ANOVA techniques and the clustering method.

As a result of working with these techniques, a number of problems have been identified. These are briefly discussed in Chapter 5.

ACKNOWLEDGEMENTS

My thanks are extended to the Statistics group of the Mathematics Department; in particular Dr. C. Villegas and Dr. K.L. Weldon who helped me during the course of my studies, and especially Dr. M.A. Stephens who has supervised this thesis with great patience and always gave me encouragement during my graduate work.

Thanks to John Spinelli for the friendship and interest of a fellow graduate student.

Ms. Sylvia Holmes typed this thesis with great speed and accuracy, and I thank her for her tolerance for all the extra bits I kept giving her.

My final and important thanks are due to the Mathematics Department, and particularly the two chairmen Dr. Norman Reilly and Dr. Manohar Singh for supporting me in the Department, and to Ms. Judy Easton and Ms. Betty Dwyer for their kindness which has made my life here as a student most pleasant. I hope my contribution to the Department has merited the support I have been given.

TABLE OF CONTENTS

CHAPTER 1

The von Mises and Fisher distributions.

Useful distributions which have been used for the analysis of directional data are the von Mises distribution in two dimensions, and the Fisher distribution in three dimensions. In this chapter we describe the von Mises and Fisher distributions. In Chapter 2, they are generalized to p dimensions.

1.1. Directional Data.

Suppose the direction taken by a bird when released from a point O is denoted by a unit vector OP starting at the centre, O , of a circle of radius one and finishing at a point P on the circumference of the circle. The vector OP is an example of a piece of directional data, in two dimensions. In three dimensions, the point P would be on a sphere; an example of directional data in three dimensions is the direction of magnetization of a rock sample.

In higher dimensions, a p-dimensional directional sample value is denoted by a unit vector OP starting at the centre O of a hypersphere of radius one and finishing at a point P on the surface of the hypersphere. Although a p-dimensional unit vector does not have a physical interpretation in terms of direction, a set of vectors whose components are continuous proportions might be usefully analyzed using techniques for directional data. Examples of such vectors of continuous proportions are the vector giving the proportions of time spent in different activities by a student and the proportions of a company's production allocated to various outputs.

1.2. The von Mises distribution.

Let OP be a typical unit vector in two dimensions as described above, and let OA be a fixed unit vector (it can be thought of as pointing to the North Pole); suppose θ is the angle between OP and OA. Then the von Mises distribution for θ is given by:

$$f(\theta) = \frac{1}{2\Pi\ I_{(0)}(k)} \exp(k \cos \theta);\ -\Pi \leq \theta \leq \Pi\ , \qquad (1.1)$$

where $I_{(0)}(k)$ is the imaginary Bessel function of order zero and argument k .

The density (1.1) is symmetrical, with a mode at $\theta = 0$; thus the vector OA is called the modal vector. The constant k is a precision or concentration parameter; when k is zero, the density is uniform over the circle, i.e., the points P_i are uniformly distributed on the circumference. A sample of vectors OP_i is then often said to be randomly distributed over the circle. When k is large, the vectors are clustered around the modal vector OA .

The density (1.1) can be extended to place its mode along an arbitrary vector OA at $\theta = \alpha$;

$$f(\theta) = \frac{1}{2\Pi\ I_0(k)} \exp\{k \cos (\theta-\alpha)\},\quad -\Pi \leq \theta \leq \Pi\ , \qquad (1.2)$$

If the direction cosines of the vectors OA and OP are (a_1, a_2) and (x_1, x_2) respectively, the density is given by

$$f(x_1, x_2) = \frac{1}{2\Pi\ I_0(k)} \exp\{k(x_1 a_1 + x_2 a_2)\}. \qquad (1.3)$$

The von Mises distribution has found many uses to describe directional data clustered around a mode; for example, the flights of migratory birds mentioned in Section 1.1, or the progress of animals or insects toward a certain point.

1.3. The Fisher Distribution.

The Fisher distribution is the analogue of the von Mises distribution when the vector OP is in three dimensions, i.e., O is the centre of a sphere of radius one, and P is a point on the surface of the sphere. The vector OP will be denoted by spherical polar coordinates (θ, ϕ). Suppose OA, the modal vector, is the origin for θ . The Fisher distribution for (θ, ϕ) is given by:

$$f(\theta, \phi) = \frac{k \sin\theta}{4\Pi \sinh(k)} \exp(k \cos\theta), \quad (0 \le \theta \le \Pi\ ;\ 0 \le \phi \le 2\Pi) \qquad (1.4)$$

where k is, as before, the concentration parameter. Note that the density is symmetrical around $\theta = 0$. If we wish the modal vector OA to lie along an arbitrary vector, the description of the density is much more complicated and it will not be written in full. If the direction cosines of the vectors OA and OP are (a_1, a_2, a_3) and (x_1, x_2, x_3) respectively, the density per unit area is given by:

$$f(x_1,x_2,x_3) = \frac{k}{4\Pi \sinh(k)} \exp(k(x_1a_1 + x_2a_2 + x_3a_3)) \qquad (1.5)$$

i.e., like the von Mises distribution, it is proportional to $\exp(k \cos\theta)$. Fisher suggested that this is a useful distribution in the earth sciences when doing studies of palaeo-magnetic data and sedimentary geology.

1.4. Statistics associated with the Fisher distribution.

In the next two sections we give some basic estimation results and tests derived by Fisher and others for the three dimensional case; analogous results for the two dimensional case can be inferred immediately. Suppose a sample of N unit vectors OP_i is given on the surface of a sphere; let (x_{i1},x_{i2},x_{i3}) be the direction cosines of the i-th ($i = 1,\ldots,N$) observation. The resultant (denoted by $\underline{R}$) of the set of N vectors is defined as the vector with components:

$$\underline{R} = (X_1,X_2,X_3) = (\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \sum_{i=1}^{N} x_{i3}) \qquad (1.6)$$

The length R of the resultant is given by

$$R = (X_1^2 + X_2^2 + X_3^2)^{1/2} = (\underline{R}\,\underline{R}^t)^{1/2} \qquad (1.7)$$

where $\underline{R}^t$ denotes the transpose of $\underline{R}$.

The maximum likelihood estimate of the direction of the modal vector is the direction of the resultant. (Watson, 1956). Let $\hat{OA} = (\hat{a}_1, \hat{a}_2, \hat{a}_3)$ denote the maximum likelihood estimate of the modal vector $OA = (a_1, a_2, a_3)$; we then have

$$\hat{OA} = (\hat{a}_1, \hat{a}_2, \hat{a}_3) = (X_1, X_2, X_3) = \underline{R} \tag{1.8}$$

The maximum likelihood estimate $\hat{k}$ of k is a function of R; namely, it is the solution of the equation (Watson, 1956):

$$\coth(k) - \frac{1}{k} = \frac{R}{N} . \tag{1.9}$$

The left hand side of (1.9) is a monotonic increasing function of k and its value changes from 0 to 1 as k runs from 0 to ∞. When R/N is near unity, (1.9) has the approximate solution

$$\hat{k} = \frac{N}{N - R} \tag{1.10}$$

The accuracy of (1.10) is sufficient for practical purposes (Watson, 1956), for $k > 3$.

When the modal vector is known, the maximum likelihood estimate $\hat{k}$ of k satisfies the equation

$$\coth(k) - \frac{1}{k} = \frac{\sum_{i=1}^{N} \cos \theta_i}{N} = \frac{X}{N} \tag{1.11}$$

where θ_i is the angle between the i-th observed vector OP_i and the modal vector, and X is the projection of $\underline{R}$ on the modal vector. Thus here $\hat{k}$ is given approximately by:

$$\hat{k} = \frac{N}{N - X} \tag{1.12}$$

1.5. Tests of significance for the Fisher distribution.

Practical application of Fisher's distribution is aided by a series of significance tests analogous to those in use for the normal distribution. Some of the tests for the modal vector will be outlined in this section. More detailed descriptions of these tests, and of tests concerning the concentration parameter k, are in Watson (1956) and Stephens (1962, 1967, 1969).

1. Test of a Given Modal Vector.

This test is used when we wish to test that the modal vector is equal to a particular vector $\underline{A}_0$. Let X be the length of the projection of the resultant on the assumed modal vector $\underline{A}_0$. The following identifications are made:

$2k(N - X)$ = dispersion of the sample about $\underline{A}_0$.

$2k(N - R)$ = dispersion of the sample about $\underline{R}$.

For large k, Watson (1956) showed that $2k(N - X)$ and $2k(N - R)$ are distributed approximately as χ^2_{2N} and $\chi^2_{2(N - 1)}$ respectively.

By analogy with the identity in normal samples

$$\sum_{i=1}^{N} (x_i-\mu)^2 = \sum_{i=1}^{N} (x_i-\bar{x})^2 + N(\bar{x}-\mu)^2$$

we write

$$2k(N - X) = 2k(N - R) + 2k(R - X) \qquad (1.13)$$

$$\text{i.e., } \chi^2_{2N} = \chi^2_{2(N-1)} + \chi^2_2 \qquad (1.14)$$

and therefore the quotient

$$\frac{(N-1)(R-X)}{(N-R)} \qquad (1.15)$$

will have the F-distribution with 2 and 2(N - 1) degrees of freedom. The null hypothesis is rejected if the statistic is larger than the percentage point corresponding to the F-distribution with 2 and 2(N - 1) degrees of freedom, at the appropriate significance level. Thus, intuitively, if the direction of $\underline{R}$ is very different from that of A_0 we will obtain a small X which in turn will produce a large test statistic leading to the rejection of the null hypothesis.

2. Comparison of two modal vectors.

Suppose that samples of size N_1 and N_2 are drawn from two populations and that a test is to be made that their modal vectors are identical. Let R_1 and R_2 denote the length of the resultants of the first and second samples respectively, and as before let R be the length of the total resultant. Assuming that both populations have equal values of k, we may write

$$2k(N-R) = 2k(N_1-R_1) + 2k(N_2-R_2) + 2k(R_1+R_2-R) \qquad (1.16)$$

where $2k(N-R)$, $2k(N_1-R_1)$ and $2k(N_2-R_2)$ are distributed approximately as $\chi^2_{2(N-1)}$, $\chi^2_{2(N_1-1)}$ and $\chi^2_{2(N_2-1)}$ respectively. The parallel χ^2 identity is therefore

$$\chi^2_{2(N-1)} = \chi^2_{2(N_1-1)} + \chi^2_{2(N_2-1)} + \chi^2_2 \, . \qquad (1.17)$$

This suggests that

$$Z = \frac{(N-2)(R_1+R_2-R)}{N-R_1-R_2} \approx F_{2,2(N-2)} \, . \qquad (1.18)$$

The statistic in (1.18) has an immediate intuitive interpretation; if the mean vectors are very different, $R_1 + R_2$ will be much greater than R and the left hand side of (1.18) will be large. Hence the hypothesis will be rejected for large Z. There is a natural extension to more than two samples, which will be treated in the next chapter for p dimensions.

CHAPTER 2

The von Mises distribution in p-dimensions.

2.1. The von Mises and Fisher distribution extended to p-dimensions.

The extension of the von Mises and Fisher distributions to higher dimensions was made by Stephens (1962). This extension is sometimes referred to as the von Mises distribution in p-dimensions. The theory of this distribution is as follows, taken largely from Stephens (1962).

Suppose an observation in p-dimensions is recorded by a unit vector OP starting at the centre O and finishing at P, on the surface of a hypersphere of radius one. Let $\underline{x}$ be the vector OP, and suppose $\underline{x}$ has components:

$$\underline{x} = (x_1, x_2, \ldots, x_p).$$

It is convenient to transform the vector $\underline{x}$ into polar coordinates; these are defined by the radius r (here $r \equiv 1$), and by angles θ_i, components of

$$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_{p-1})$$

where

$$\left.\begin{aligned} x_1 &= \cos\theta_1 \\ x_j &= \cos\theta_j \prod_{i=1}^{j-1} \sin\theta_i \;,\; (j = 2,\ldots,p-1); \\ x_p &= \prod_{i=1}^{p-1} \sin\theta_i \end{aligned}\right\} \begin{aligned} &0 \le \theta_j \le \Pi \\ &(j = 1,2,\ldots,p-2) \\ &0 \le \theta_{p-1} \le 2\Pi \end{aligned}$$

The von Mises density for P in p-dimensions or equivalently of OP is given by:

$$f_p(\theta_1,\theta_2,\ldots,\theta_{p-1}) = C_p(k)\exp(k\cos\theta_1)\sin^{p-2}\theta_1\cdot\sin^{p-3}\theta_2\cdots\sin\theta_{p-2} \quad (2.1)$$

where $0 \le \theta_i \le \pi$, $i = 1,\ldots,p-2$; $0 \le \theta_{p-1} \le 2\pi$ and $k > 0$. The constant term is given by:

$$C_p(k) = \frac{k^{p/2-1}}{I_{(p/2-1)}(k)(2\Pi)^{p/2}} \quad \text{when } k \neq 0$$

where $I_{(m)}(k)$ denotes the imaginary Bessel function of order m and argument k . When p is odd, $C_p(k)$ can be written as a function of sinh(k) and cosh(k).

This density function represents a distribution of vectors symmetrical about the modal vector OA , which lies along $\theta_1 = 0$. The general density function for the mode lying along an arbitrary vector is very complicated. The constant k is, again, a concentration parameter; for large k , the vectors are tightly clustered about OA , and for $k = 0$, the distribution is uniform over the surface of the hypersphere.

The extension of the von Mises distribution to p-dimensions permits its use in much more general situations, for example the analysis of any type of data that can be represented as unit vectors on a hypersphere clustered around a constant vector. In this thesis we propose to use the distribution for data which can be represented by such a cluster of unit vectors. For these vectors k will be large, and we now investigate properties of the distribution for this case.

2.2. Properties of the distribution when k is large.

Let $\underline{v}$ be a unit vector with coordinates $(x_1, x_2, \ldots, x_p)$, and with polar coordinates $(\theta_1, \theta_2, \ldots, \theta_{p-1})$ and $r = 1$.

If the modal vector is along the North pole, x_1 is the component of $\underline{v}$ on the modal vector and θ_1 is the angle between $\underline{v}$ and the modal vector. For large k , the vectors are tightly clustered around the modal vector, and hence there is a high probability of small θ_1 . So $\cos\theta_1 \approx 1 - \theta_1^2/2$ and $\sin\theta_1 \approx \theta_1$. The density of θ_1 becomes:

$$f(\theta_1) \approx C \cdot \exp(k)\exp(-k\theta_1^2/2)\,\theta_1^{p-2}, \quad 0 \le \theta_1 \le \Pi . \qquad (2.2)$$

The quantity $k\theta_1^2$ has approximately a chi-squared distribution with p-1 degrees of freedom, i.e. :

$$k\theta_1^{\ 2} = 2k(1 - \cos\theta_1) \approx \chi^2_{p-1} \qquad (2.3)$$

Since $\cos\theta_1 = x_1$, this may be written

$$2k(1 - x_1) \approx \chi^2_{p-1} \ . \qquad (2.4)$$

Because of the symmetry around the central vector, the other coordinates x_j , $j \geq 2$, have identical normal distributions:

$$x_j \approx N(0, \tfrac{1}{k}) \ , \quad j = 2,\ldots,p \ . \qquad (2.5)$$

2.3. <u>Notation for a sample</u>.

There will frequently be cases where there are several samples of the unit vectors, so we first give the notation for a sample.

Suppose a sample of N unit vectors consists of vectors $OP_i = \underline{v}_i$, $i = 1,\ldots,N$; a typical vector $\underline{v}_i$ has components $(x_{i1}, x_{i2},\ldots,x_{ip})$ and its polar coordinates are $(\theta_{i1}, \theta_{i2},\ldots,\theta_{i(p-1)})$.

The resultant (denoted by $\underline{R}$) of the set of N vectors has components

$$(X_1, X_2,\ldots,X_p) = (\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2},\ldots, \sum_{i=1}^{N} x_{ip}) \ .$$

The length of the resultant is given by

$$R = (X_1^2 + X_2^2 + \ldots + X_p^2)^{1/2} = (\underline{R}\,\underline{R}^t)^{1/2} .$$

2.4. Estimates of k and of the modal vector.

The maximum likelihood estimator of the concentration parameter k is given by the equation

$$\frac{I_{p/2}(\hat{k})}{I_{p/2-1}(\hat{k})} = \frac{R}{N} ;$$

for k large this equation becomes

$$1 - \frac{p-1}{2\hat{k}} = \frac{R}{N} .$$

If the modal vector OA is known, R is replaced by X , the component of $\underline{R}$ on OA . The maximum likelihood estimator of OA is the direction of $\underline{R}$ as described in Section 1.4 for three dimensions.

2.5. Distributions of statistics derived from a single sample.

In this section we investigate some distributions of sample statistics, for large k . For a typical vector $OP_i = \underline{v}_i$ as shown below, let θ_i be the angle between OP_i and OA , the modal vector, and let ϕ_i be the angle between OP_i and the resultant $\underline{R}$, which estimates OA.

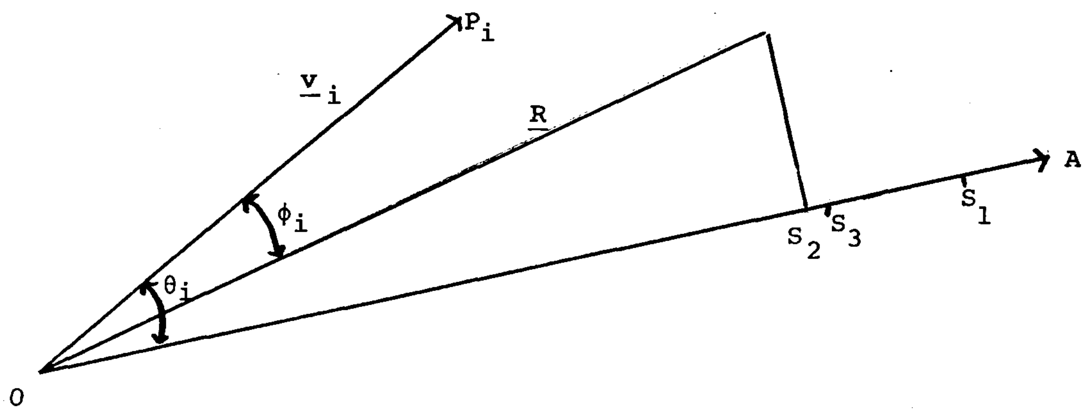

Diagram of population and sample vectors.

Figure 2.1.

Figure 2.1 shows a vector $\underline{v}_i$, the modal vector OA and the resultant $\underline{R}$. Let distance $OS_1 = N$ along OA, and let $X = OS_2$ be the projection of $\underline{R}$ on OA and let OS_3 be the same length as $\underline{R}$. Then clearly $N-X = S_1S_2$ and $N-R = S_1S_3$, and both these quantities are measures of the dispersion of the set of vectors. For large k , we have from (2.4)

$$2k(N-X) = \sum_{i=1}^{N} 2k(1 - x_{i1}) \approx \chi^2_{N(p-1)} . \qquad (2.7)$$

Further, $X_j = \sum_{i=1}^{N} x_{ij} \approx X(0, N/k)$, for $j = 2,\ldots,p$, from (2.5).

Hence $X_j^2 \approx N\,\chi_1^2/k$ $(j = 2,\ldots,p)$ and

$$R^2 - X^2 = \sum_{j=2}^{p} X_j^2 \approx \chi^2_{p-1}\,(\tfrac{N}{k})$$

or

$$\frac{k}{N}\,(R^2-X^2) \approx \chi^2_{p-1} \;.$$

Since $R \approx X \approx N$, this becomes $2k(R-X) \approx \chi^2_{p-1}$. Watson's identity is

$$2k(N-X) = 2k(N-R) + 2k(R-X) \qquad (2.8)$$

corresponding to:

$$\chi^2_{(p-1)N} = \chi^2_{(p-1)(N-1)} + \chi^2_{(p-1)} \;. \qquad (2.9)$$

This leads to the approximation for the statistic Z_1,

$$Z_1 = \frac{(N-1)(R-X)}{(N-R)} \approx F_{(p-1),(p-1)(N-1)} \;. \qquad (2.10)$$

This result can be put in terms of the angles θ_i and ϕ_i ; we have

$$k\,\Sigma_i\,\theta_i^2 = \chi^2_{(p-1)N} \quad \text{and} \quad k\,\Sigma_i\,\phi_i^2 = \chi^2_{(p-1)(N-1)} \,;$$

the first equation comes from (2.7) and the second from (2.9); these are comparable in normal theory (with $\sigma = 1$) to

$$\Sigma_i (x_i - \mu)^2 = \chi^2_N \quad \text{and} \quad \Sigma_i (x_i - \bar{x})^2 = \chi^2_{N-1} ;$$

in each case the second expression replaces the mean by its estimate, and there is a corresponding drop in degrees of freedom of χ^2. Z_1 is used in testing that a given A_0 is the modal vector, analogous to the test in Section 1.5. The left hand side of (2.10) is calculated as the test statistic and compared with the F-distribution with (p-1) and (p-1)(N-1) degrees of freedom. Large values will lead to rejecting the given A_0 as the modal vector.

2.6. Notation for several samples.

When several samples of unit vectors are given, questions might arise whether they have the same modal vectors, same concentration parameters, etc. Let the i-th group ($i = 1,2,\ldots,q$) have modal vector OA_i, and concentration parameter k_i. Let $\underline{v}_{ij}$, $j = 1,\ldots,N_i$, be the set of unit vectors in the i-th group, so that N_i is the number of vectors in the group, and let R_i be the length of the resultant vector $\underline{R}_i$ of the group. Let $N = \Sigma_i N_i$, and let R be the length of the resultant $\underline{R}$ of all the vectors treated as one large group.

2.7. Comparison of several modal vectors.

Suppose q different groups (samples) of unit vectors are given and we wish to test whether all the samples come from populations with the same modal vector, assuming they have the same value of k. The following results come from (2.9);

$$2k(N_1-R_1) \approx \chi^2_{(p-1)(N_1-1)}$$

$$2k(N_2-R_2) \approx \chi^2_{(p-1)(N_2-1)}$$

$$\vdots$$

$$2k(N_q-R_q) \approx \chi^2_{(p-1)(N_q-1)}$$

and

$$2k(N-R) \approx \chi^2_{(p-1)(N-1)}$$

We write the identity

$$2k(N-R) = 2k(N_1-R_1) + 2k(N_2-R_2) + \ldots + 2k(N_q-R_q) + 2k(R_1+R_2+\ldots+R_q-R)$$

and, again by analogy with the analysis of variance we obtain

$$2k(R_1+R_2+\ldots+R_q-R) \approx \chi^2_{(p-1)(q-1)} \; ; \qquad (2.11)$$

hence the quotient

$$Z_2 = \frac{(N-q)(\Sigma_i R_i - R)}{(q-1)(N-\Sigma_i R_i)} . \qquad (2.12)$$

will have approximately the F distribution with $(p-1)(q-1)$ and $(p-1)(N-q)$ degrees of freedom. Therefore to test whether the different groups have the same modal vector, the statistic Z_2 is calculated and compared with this F distribution. Large values of Z_2 will be significant, indicating that the $\underline{R}_i$ vectors point in different directions.

The above analysis is essentially a one-way Analysis of Variance which can be set up in the usual tabular form;

Table 2.1

ANOVA table in terms of resultants.

Sum of squares		d.f.	test
Between groups	$\Sigma_i R_i - R$	$(p-1)(q-1)$	Z_2
Within groups	$N-\Sigma_i R_i$	$(p-1)(N-q)$	
Total	$N-R$	$(p-1)(N-1)$	

Note that throught the table, 2k has been omitted before the terms under "Sum of Squares"; since only ratios will be used for tests, this does not affect the calculations. This is analogous to omitting σ^2 throughout an ANOVA table.

2.8. Examples of ANOVA for resultants.

In the first set (Table A.1.1, found in Appendix 1) we have the hours of time spent in eight different activities by 130 students of Simon Fraser University. The data were requested for one day only and do not represent the overall activity pattern. However, it is used here as an illustration of the general methodology.

An activity pattern is converted to a unit vector as follows. The hours are first converted to proportions p_i for the i-th activity and then $x_i = \sqrt{p_i}$ is the i-th component of the unit vector.

Analysis

1. We first examine whether there are differences in activity patterns due to the sex of the students. The data is split into two groups; Group 1 for women and Group 2 for men. The total resultant R for the 130 students is 117.1987; other results are shown in Table 2.2.

The test statistic Z_2 is less than 1 so we do not reject the hypothesis that there is no difference in activity pattern between the sexes. Whenever Z_2 is greater than 1, the statistic is converted to a standard normal variable. Several transformations have been examined in detail in Appendix 2 of this thesis. The three most accurate are those of Peizer and Pratt (1968), Carter (1947) and Paulson (1942).

2. The activity patterns were next examined according to the age of the students. The results are in Table 2.3.

Table 2.2

Results for test between sexes.

Group	Sex	N_i	$\hat{k}_i$	R_i
1	Female	56	35.664	50.5041
2	Male	74	35.744	66.7540

ANOVA TABLE

Sum of squares			d.f.	
Between groups	$\Sigma_i R_i - R$	= 0.0601	7	Z_2 = 0.5980
Error	$N - \Sigma_i R_i$	= 12.7412	896	
Total	$N - R$	= 12.8013	903	

Table 2.3

Results for test between age groups.

Group	Age	N_i	$\hat{k}_i$	R_i
1	less than 21	47	36.2069	42.4567
2	21 - 25	61	36.8403	55.2047
3	more than 25	22	38.1845	19.9835

ANOVA Table

Sum of Squares			d.f.	Test Statistic
Between groups	$\Sigma_i R_i - R$	= 0.4462	14	Z_2 = 2.2931
Error	$N - \Sigma_i R_i$	= 12.3551	889	
Total	$N - R$	= 12.8013	903	

The test statistic is 2.2931 with 14 and 889 degrees of freedom, and the corresponding Z-score is 2.6241 (Peizer and Pratt), 2.6265 (Carter) or 2.6173 (Paulson). At the .05 level of significance, we reject the null hypothesis that the students in the different age groups have similar activity patterns.

2.9. One way ANOVA for angles.

Watson's one-way ANOVA leads to a test based on the group resultants and total resultant, as described above. We now show the analogous test based directly on the angles between the group resultants and the vectors in the corresponding groups, and between the vectors and the total resultant.

Let N be the total number of unit vectors OP_i $(i = 1,\ldots,N)$ starting at the centre and finishing on the surface of a p-dimensional hypersphere. Let q be the number of groups in which the vectors are split. Let $\underline{v}_{ij}$ be the j^{th} vector in group i, let ϕ_{ij} denote the angle between the total resultant $\underline{R}$ and $\underline{v}_{ij}$, and let α_{ij} denote the angle between $\underline{v}_{ij}$ and the resultant $\underline{R}_i$ of the i-th group. When k is large, ϕ_{ij} and α_{ij} are very small, and so, using the approximations $\alpha^2_{ij} \approx 2(1 - \cos \alpha_{ij})$ and $\phi^2_{ij} \approx 2(1 - \cos \phi_{ij})$, we obtain

$$\Sigma_{ij}\alpha^2_{ij} \approx 2\,\Sigma_{ij}\,(1-\cos \alpha_{ij}) = 2(N-\Sigma_{ij} \cos \alpha_{ij})$$

and

$$\Sigma_{ij}(\phi^2_{ij}-\alpha^2_{ij}) \approx 2\Sigma_{ij}(\cos\alpha_{ij} - \cos\phi_{ij})$$

As $\Sigma_{ij}\cos\phi_{ij}$ is the sum of projections of all N vectors on their resultant and $\Sigma_j\cos\alpha_{ij}$ is the sum of the projections of the vectors in the i-th group on the resultant of the i-th group, it follows that

$$\Sigma_{ij}\cos\phi_{ij} = R$$

and

$$\Sigma_j\cos\alpha_{ij} = R_i \ , \ i = 1,\ldots,q \ .$$

Substituting these results in

$$2k(N-R) \approx \chi^2_{(p-1)(N-1)}$$

$$2k(N-\Sigma_i R_i) \approx \chi^2_{(p-1)(N-q)}$$

$$2k(\Sigma_i R_i - R) \approx \chi^2_{(p-1)(q-1)} \ .$$

We obtain

$$2k(N-\Sigma_{ij}\cos\theta_{ij}) \approx 2k(\Sigma_{ij}\theta^2_{ij}/2) \approx \chi^2_{(p-1)(N-1)}$$

$$2k(N-\Sigma_{ij} \cos \alpha_{ij}) \approx 2k(\Sigma_{ij} \alpha^2_{ij}/2) \approx \chi^2_{(p-1)(N-q)} ,$$

$$2k \Sigma_{ij}(\cos \theta_{ij} - \cos \alpha_{ij}) \approx 2k\Sigma_{ij}(\theta^2_{ij}/2-\alpha^2_{ij}/2) \approx \chi^2_{(p-1)(q-1)} .$$

Then Watson's test statistic

$$Z_2 = \frac{(N-q)(\Sigma_i R_i - R)}{(q-1)(N-\Sigma_i R_i)}$$

becomes, in terms of angles,

$$Z_2^1 = \frac{(N-q)\{\Sigma_{ij}(\phi^2_{ij}-\alpha^2_{ij})\}}{(q-1)(\Sigma_{ij}\alpha^2_{ij})} .$$

Z_2^1 will therefore have an F-distribution with $(p-1)(q-1)$ and $(p-1)(N-q)$ degrees of freedom. The null hypothesis that the q modal vectors are equal is rejected if the test statistic F is greater than the percentage point of the F-distribution with $(p-1)(q-1)$ and $(p-1)(N-q)$ degrees of freedom, at the appropriate significance level. In terms of angles, the ANOVA table is shown in Table 2.4.

Table 2.4

ANOVA table in terms of angles.

Sum of Squares		d.f.	test
Between groups	$\Sigma_{ij}(\phi^2_{ij} - \alpha^2_{ij})$	(p-1)(q-1)	Z^1_2
Within groups	$\Sigma_{ij}\ \alpha^2_{ij}$	(p-1)(N-q)	
Total	$\Sigma_{ij}\ \phi^2_{ij}$	(p-1)(N-1)	

2.10. Examples for one way ANOVA with angles.

The following examples will illustrate the one way ANOVA for angles. The groups compared are the same as in Section 2.8 so that both results may be compared.

1. The one way ANOVA table for angles to test for difference between women and men is shown in Table 2.5.

Table 2.5

ANOVA table for difference between sexes.

Sum of Squares		d.f.	test
Between groups	$\Sigma_{ij}(\phi^2_{ij}-\alpha^2_{ij})$ = .1256	7	Z^1_2 = .6161
Error	$\Sigma_{ij}\alpha^2_{ij}$ = 26.1101	896	
Total	$\Sigma_{ij}\phi^2_{ij}$ = 26.2357	903	

The test statistic is .6161 with 7 and 896 degrees of freedom. The corresponding normal score is -0.6523 (Peizer and Pratt), -0.6492 (Carter), or -0.6579 (Paulson). At the .05 level of significance, we do not reject the null hypothesis that women and men spend the time in a similar way.

2. The one way ANOVA for angles to test for difference between the three age groups is given in Table 2.6.

Table 2.6

ANOVA table of difference between age groups.

Sum of Squares			d.f.	test
Between groups	$\Sigma_{ij}(\phi^2_{ij}-\alpha^2_{ij})$	= .9434	14	Z^1_2 = 2.3686
Error	$\Sigma_{ij}\alpha^2_{ij}$	= 25.2923	889	
Total	$\Sigma_{ij}\phi^2_{ij}$	= 26.2357	903	

The test statistic Z^1_2 is 2.3686 with 14 and 889 degrees of freedom. The corresponding normal score is 2.7310 (Peizer and Pratt), 2.7410 (Carter), or 2.7286 (Paulson). At the .05 level of significance we reject the null hypothesis that the students in the different age groups have the same activity pattern.

2.11. Two way ANOVA.

In the above, the sample was classified into groups by one criterion, and a one-way ANOVA made to examine whether activity patterns differ between groups. This analysis will now be extended to classification by two criteria. Suppose the sample items are classified in two ways, by classification 1 with I groups and classification 2 with J groups. If a sample item (for illustration, a student) falls into group i of classification 1 and group j of classification 2, the associated vector of activity proportions will be placed in cell (i, j) in row i, column j, of a two way table. Extending our previous notation, we write $\underline{v}_{ijk}$ for the k-th vector in cell (i, j). Let N_{ij} be the number of vectors in cell (i, j) and let R_{ij} be the length of the resultant in this cell. Let $R_{i.}$ be the length of the resultant of all vectors in row i, i.e., of all items in group i of the first classification and similarly let $R_{.j}$ be the resultant of all vectors in column j. Suppose the total resultant has length $R_{..}$; a table may be constructed as in Table 2.7.

Table 2.7

Resultants for two-way classification.

Classification 2 (Columns)

Classification 1 (rows)

	1	2	3	4	. . .	J	Total
1	R_{11}	R_{12}	R_{13}	R_{14}	. . .	R_{1J}	$R_{1\cdot}$
2	R_{21}	R_{22}	R_{23}		. . .	R_{2J}	$R_{2\cdot}$
3	R_{31}	. . .					
⋮							
I	R_{I1}	. . .				R_{IJ}	$R_{I\cdot}$
Total	$R_{\cdot 1}$	$R_{\cdot 2}$				$R_{\cdot J}$	$R_{\cdot\cdot}$

The following can be written

$$2k(N - R_{\cdot\cdot}) = 2k[(N_{11} - R_{11}) + (N_{12} - R_{12}) + \ldots + (N_{1J} - R_{1J})] + \cdots$$

$$+ 2k[(N_{I1} - R_{I1}) + (R_{I2} - R_{I2}) + \ldots + (N_{IJ} - R_{IJ})]$$

$$+ 2k[(R_{11} + R_{12} + \ldots + R_{1J} - R_{1\cdot}) + (R_{21} + R_{22} + \ldots + R_{2J} - R_{2\cdot})$$

$$+ \ldots + (R_{I1} + R_{I2} + \ldots + R_{IJ} - R_{I\cdot})] + 2k(R_{1\cdot} + R_{2\cdot\cdot} + \ldots + R_{I\cdot} - R_{\cdot\cdot}).$$

The results obtained in equations (2.8) and (2.9) now give

$$2k(N - R_{\cdot\cdot}) \approx \chi^2_{(p-1)(N-1)}$$

$$2k(\Sigma_j R_{1j} - R_{1\cdot}) \approx \chi^2_{(p-1)(J-1)}$$

$$2k(\Sigma_j R_{2j} - R_{2\cdot}) \approx \chi^2_{(p-1)(J-1)}$$

$$\vdots$$

$$2k(\Sigma_j R_{Ij} - R_{I\cdot}) \approx \chi^2_{(p-1)(J-1)}$$

$$2k(\Sigma_i R_{i\cdot} - R_{\cdot\cdot}) \approx \chi^2_{(p-1)(I-1)}$$

and $$2k(N_{ij} - R_{ij}) \approx \chi^2_{(p-1)(N_{ij}-1)}; \; i = 1,2,\ldots,I; \; j = 1,2,\ldots,J.$$

We obtain

$$2k\{\Sigma_{ij}(N_{ij} - R_{ij})\} \approx \chi^2_{(p-1)(N-IJ)}$$

i.e., $$2k(N - \Sigma_{ij}R_{ij}) \approx \chi^2_{(p-1)(N-IJ)} \; .$$

2.12. Test for no difference between rows.

The above approximations can be used, as before for the one way ANOVA, to give tests for the difference between rows or between columns within rows. Thus the analysis will be similar to what is usually called a nested analysis of variance.

Under the null hypothesis that there is no difference between rows, the quotient

$$Z_3 = \frac{(N - IJ)(\Sigma_i R_{i\cdot} - R_{\cdot\cdot})}{(I - 1)\ (N - \Sigma_{ij} R_{ij})} \tag{2.13}$$

has an F-distribution with $(p-1)(I-1)$ and $(p-1)(N-IJ)$ degrees of freedom. The null hypothesis is rejected if Z_3 is greater than the percentage point corresponding to the F-distribution at the chosen level of significance α .

The quotient

$$Z_{4i} = \frac{(N - IJ)(\Sigma_j R_{ij} - R_{i\cdot})}{(J - 1)\ (N - \Sigma_{ij} R_{ij})} \quad i = 1,2,\ldots,I \tag{2.14}$$

has an F-distribution with $(p-1)(J-1)$ and $(p-1)(N-IJ)$ degrees of freedom. The null hypothesis that there is no difference between the columns within row i should be rejected for large values of Z_{4i} .

With the above analysis, it is not possible in the same table to decide if there is an overall difference between columns. If this is to be examined, the nested layout is set up again, but with the rows nested within columns. Then the difference between rows within columns, and the difference between columns themselves, will be tested with F-distributions analogous to the ones described above.

The two way ANOVA table is shown in Table 2.8.

Table 2.8

Two way ANOVA in terms of resultants.

Sum of squares		d.f.	test
Between rows	$\Sigma_i R_{i\cdot} - R_{\cdot\cdot}$	(p-1)(I-1)	Z_3
Between columns;			
Within row 1	$\Sigma_j R_{1j} - R_{1\cdot}$	(p-1)(J-1)	Z_{41}
Within row 2	$\Sigma_j R_{2j} - R_{2\cdot}$	(p-1)(J-1)	Z_{42}
⋮			
Within row I	$\Sigma_j R_{Ij} - R_{I\cdot}$	(p-1)(J-1)	Z_{4I}
Error	$N - \Sigma_{ij} R_{ij}$	(p-1)(N-IJ)	
Total	$N - R_{\cdot\cdot}$	(p-1)(N-1)	

2.13. Examples for two way ANOVA in terms of resultants.

When classifying the students according to their age (rows) and sex (columns), six cells are obtained. The number of students in each cell and the cell, row and column resultants are given in Table 2.9.

Table 2.9

Cell sizes and resultants for classification age-sex.

Cell sizes

Sex

Age		Females	Males	Total
	< 21	N_{11} = 19	N_{12} = 28	$N_{1\cdot}$ = 47
	21-25	N_{21} = 28	N_{22} = 33	$N_{2\cdot}$ = 61
	> 25	N_{31} = 9	N_{32} = 13	$N_{3\cdot}$ = 22
	Total	$N_{\cdot 1}$ = 56	$N_{\cdot 2}$ = 74	N = 130

Resultants

Sex

Age		Females	Males	Total
	< 21	R_{11} = 17.1612	R_{12} = 25.3094	$R_{1\cdot}$ = 42.4567
	21-25	R_{21} = 25.3636	R_{22} = 29.9985	$R_{2\cdot}$ = 55.2047
	> 25	R_{31} = 8.2024	R_{32} = 11.8276	$R_{3\cdot}$ = 19.9835
	Total	$R_{\cdot 1}$ = 50.5041	$R_{\cdot 2}$ = 66.7540	$R_{\cdot\cdot}$ = 117.1987

1. The two way ANOVA table for sex within age is given in Table 2.10.

Table 2.10

ANOVA table in terms of resultants for sex within age.

Sum of squares			d.f.	test
Between ages	$\Sigma_i R_{i\cdot} - R_{\cdot\cdot}$	= 0.4462	14	$Z_3 = 2.2793$
Between sexes				
Within age group 1	$\Sigma_j R_{1j} - R_{1\cdot}$	= 0.0139	7	$Z_{41} = 0.1420$
Within age group 2	$\Sigma_j R_{2j} - R_{2\cdot}$	= .1574	7	$Z_{42} = 1.1081$
Within age group 3	$\Sigma_j R_{3j} - R_{3\cdot}$	= 0.0465	7	$Z_{43} = 0.4782$
Error	$N - \Sigma_{ij} R_{ij}$	= 12.1373	868	
Total	$N - R_{\cdot\cdot}$	= 12.8013	903	

Using Peizer and Pratt's, Carter's and Paulson's formulae, we obtain the following Z-scores for the test statistics:

$Z_3 \approx 2.6023, 2.6044, 2.5958$; $Z_{41} \approx -2.5673, -2.5373, -2.5042$;

$Z_{42} \approx 0.3704, 0.3549, 0.3705$; $Z_{43} \approx -1.0397, -1.0283, -1.0439$. The only test statistic which is significant (at the $\alpha = .05$ level) is Z_3. Therefore, we reject the null hypothesis that the different age groups have a similar activity pattern, but we do not reject the hypothesis that there is a difference between males and females within any of the age groups.

2. If we now switch rows and columns, the ANOVA table for age within sex can be constructed as in Table 2.11.

Table 2.11

ANOVA table in terms of resultants for age within sex.

Sum of squares		d.f.	test
Between sexes	$\Sigma_j R_{\cdot j} - R_{\cdot\cdot} = 0.0595$	7	$Z_3 = 0.6079$
Between age groups.			
Within females	$\Sigma_i R_{i1} - R_{\cdot 1} = 0.2230$	14	$Z_{41} = 1.1391$
Within males	$\Sigma_i R_{i2} - R_{\cdot 2} = 0.3815$	14	$Z_{42} = 1.9488$
Error	$N - \Sigma_{ij} R_{ij} = 12.1373$	868	
Total	$N - R_{\cdot\cdot} = 12.8013$	903	

The Z-scores (using Peizer and Pratt's, Carter's and Paulson's formulae) are given by: $Z_3 \approx -0.6735, -0.6610, -0.6791$; $Z_{41} \approx 0.4711, 0.4655, 0.4719$; $Z_{42} \approx 2.0756, 2.0720, 2.0745$. Z_3 is not significant; we do not reject the hypothesis that males and females have a similar activity pattern. When considering differences between age groups, only Z_{42} is significant at the $\alpha = .05$ level of significance; hence, there is a difference between age groups within the males.

The two way ANOVA gives a breakdown of the information contained in the one way ANOVA. In Example 2, Section 2.8, we found that there was a difference between age groups; the difference was narrowed down using the two way ANOVA, and we can now see that it is mainly due to a difference within the males.

The two way ANOVA layout can also be extended to 3 or more classifications.

2.14. The two way ANOVA in terms of angles.

It is possible to put the above results again in terms of angles.

Let α_{ijk} be the angle between vector $\underline{v}_{ijk}$ and the resultant $\underline{R}_{ij}$ of group (i, j); let ϕ_{ijk} be the angle between $\underline{v}_{ijk}$ and $\underline{R}_{i\cdot}$, where $\underline{R}_{i\cdot}$ is the resultant of vectors $\underline{R}_{ij}$, $j = 1,2,\ldots,J$; and let γ_{ijk} be the angle between $\underline{v}_{ijk}$ and the total resultant $\underline{R}_{\cdot\cdot}$.

The table analogous to Table 2.8 is Table 2.12.

Table 2.12

Two way ANOVA in terms of angles.

Sum of squares		d.f.	test
Between rows	$\Sigma_{ijk}(\gamma_{ijk}^2 - \phi_{ijk}^2)$	(p-1)(I-1)	z_3^1
Between columns;			
Within row 1	$\Sigma_{jk}(\phi_{1jk}^2 - \alpha_{1jk}^2)$	(p-1)(J-1)	z_{41}^1
Within row 2	$\Sigma_{jk}(\phi_{2jk}^2 - \alpha_{2jk}^2)$	(p-1)(J-1)	z_{42}^1
⋮			
Within row I	$\Sigma_{jk}(\phi_{Ijk}^2 - \alpha_{Ijk}^2)$	(p-1)(J-1)	z_{4I}^1
Error	$\Sigma_{ijk}\alpha_{ijk}^2$	(p-1)(N-IJ)	
Total	$\Sigma_{ijk}\gamma_{ijk}^2$	(p-1)(N-1)	

Tests of significance are made in a similar way as in equations (2.13) and (2.14).

2.15. Examples for two way ANOVA in terms of angles.

The following examples will illustrate the two way ANOVA for angles. The classification into cells is the same as in Section 2.13 so that the results may be compared.

1. The ANOVA table in terms of angles for the tests corresponding to the model sex within age is given in Table 2.13.

Table 2.13

ANOVA table in terms of angles for sex within age.

Sum of squares			d.f.	Test Statistic
Between ages	$\Sigma_{ijk}(\gamma_{ijk}^2 - \phi_{ijk}^2)$	= 0.9434	14	$Z_3^1 = 2.3545$
Between sexes				
Within age 1	$\Sigma_{jk}(\phi_{1jk}^2 - \alpha_{1jk}^2)$	= 0.0290	7	$Z_{41}^1 = 0.1448$
Within age 2	$\Sigma_{jk}(\phi_{2jk}^2 - \alpha_{2jk}^2)$	= 0.3265	7	$Z_{42}^1 = 1.6298$
Within age 3	$\Sigma_{jk}(\phi_{3jk}^2 - \alpha_{3jk}^2)$	= 0.0953	7	$Z_{43}^1 = 0.4757$
Error	$\Sigma_{ijk}\alpha_{ijk}^2$	= 24.8417	868	
Total	$\Sigma_{ijk}\gamma_{ijk}^2$	= 26.2357	903	

Peizer and Pratt's, Carter's and Paulson's approximations give the following Z-scores for the test statistics: $Z_3^1 \approx 2.7151$, 2.7187, 2.7070; $Z_{41}^1 \approx -2.5462$, -2.5162, -2.4851; $Z_{42}^1 \approx 1.1587$, 1.1388, 1.1624; $Z_{43}^1 \approx -1.0474$, -1.0358, -1.0516. Z_3^1 is the only test statistic which is significant (with $\alpha = .05$). Therefore, we reject the null hypothesis that the three age groups have a similar activity pattern, and we do not reject the hypothesis that there is a difference between males and females within any of the age groups.

2. If the rows and the columns are switched, the ANOVA table for age within sex is as in Table 2.14.

Table 2.14

ANOVA table in terms of angles for age within sex.

Sum of squares		d.f.	test
Between sexes	$\Sigma_{ijk}(\gamma^2_{ijk} - \phi^2_{ijk}) = 0.1256$	7	$z^1_3 = 0.6269$
Between age groups			
Within females	$\Sigma_{ik}(\phi^2_{i1k} - \alpha^2_{i1k}) = 0.4681$	14	$z^1_{41} = 1.1683$
Within males	$\Sigma_{ik}(\phi^2_{i2k} - \alpha^2_{i2k}) = 0.8005$	14	$z^1_{42} = 1.9979$
Error	$\Sigma_{ijk}\alpha^2_{ijk} = 24.8417$	868	
Total	$\Sigma_{ijk}\gamma^2_{ijk} = 26.2357$	903	

Peizer and Pratt's, Carter's and Paulson's approximations give the following Z-scores for the test statistics: $z^1_3 \approx -0.6246$, -0.6221, -0.6302; $z^1_{41} \approx 0.5409$, 0.5350, 0.5419; $z^1_{42} \approx 2.1573$, 2.1545, 2.1556. The test statistic z^1_3 is not significant; we do not reject the null hypothesis that males and females have a similar activity pattern. When considering differences between age groups, only z^1_{42} is significant at the $\alpha = .05$ level of significance; we reject the null hypothesis that there is no difference between age groups within the males.

The results given by this method are essentially the same ones obtained with the two way ANOVA for resultants. The use of the two way ANOVA for angles may be preferred in cases where the user feels more comfortable working with angles, or, when it is possible to give an interpretation to angles in p-dimensions.

2.16. Goodness-of-fit.

The analysis described so far assumes that the observations satisfy the p-dimensional von Mises distribution. In order to test whether the data comes from such a population we use the distributional results described in Section (2.2). The important two results are the distribution of the angle θ_1 between a typical vector and the modal vector, and the distribution of the component of a typical vector at right angles to the modal vector. The distribution of θ_1 is given by

$$f(\theta_1) = C \cdot \exp(k \cos \theta_1) \sin^{p-2} \theta_1 \quad 0 \le \theta_1 \le \pi \ . \tag{2.15}$$

Since the modal vector is not precisely known, θ_1 must be estimated by ϕ_1 , the angle between a typical vector and the resultant $\underline{R}$ of the sample. The goodness of fit test is then in two parts: (a) The set of angles ϕ_1 is tested to come from the population (2.15) using the usual Pearson χ^2 test. (b) The (2.15) components at right angles to the resultant vector are tested to be uniform on the hypersphere of dimension p-1, using the Rayleigh test for uniformity, described for example, in Watson (1956) or in a recent survey of such tests by Prentice (1978). For a typical sample, let $\underline{R}$, the resultant of the N vectors, be defined as before: $\underline{R} = \Sigma_i \underline{v}_i$ and R be the length of the resultant. Then $\underline{u} = \underline{R}/R$ is the unit vector along $\underline{R}$. For a typical unit vector $\underline{v}_i$, the component along $\underline{R}$ is

$$\underline{p}_i = (\underline{v}_i \cdot \underline{u})\underline{u} ,$$

and the component at right angles to $\underline{R}$ is $\underline{y}_i = \underline{v}_i - \underline{p}_i$.

The new vectors $\{\underline{y}_i , i = 1,\ldots,N\}$ are such that $\underline{y}_i \cdot \underline{R} = 0$; this gives a useful check on computer calculations. These vectors lie on the (p-1) dimensional subspace S_{p-1} which is orthogonal to $\underline{R}$. The Rayleigh test examines whether the directions of these vectors are uniform over the (p-1)-dimensional hypersphere, so that they must first be transformed into vectors of unit length $\underline{u}_i = \underline{y}_i/y_i$, i = 1,...,N, where y_i is the length of $\underline{y}_i$. The resultant of this set of unit vectors is $\underline{T} = \Sigma_i \underline{u}_i$; let T be the length of $\underline{T}$. On the null hypothesis of uniformity, the test statistic

$$Z = \frac{(p-1)}{N} T^2$$

is asymptotically distributed as χ^2 with p-1 degrees of freedom. Therefore, the null hypothesis that the vectors $\{\underline{u}_i\}$ are uniformly distributed over the surface of the (p-1)-dimensional hypersphere is rejected if Z is larger than $\chi^2_{p-1}(\alpha)$.

The above two tests, the χ^2 test for the distribution for θ_1, and the Rayleigh test for the component at right angles to the resultant, together provide a good omnibus test that the vectors come from the von Mises distribution.

This two-part distributional test is applied to each cell of the two-way analysis of variance table described in Section 2.11. This is analogous to applying the test for normality in the usual two way analysis of variance table.

2.17. Examples.

In Table 2.15 we give the test statistics for the Pearson's χ^2 test on the set of angles ϕ_1 and for Rayleigh's test on the components at right angles to the resultant. Both tests were done for each one of the cells obtained when classifying the students according to their age and sex.

The tests for θ_1 are not significant at $\alpha = .1$; the Rayleigh test statistics are extremely small (significant in the lower tail). The results suggest that the distributional assumptions for θ_1 are satisfactory, but it appears that the components around the estimated modal vector are more regular than expected, or else in groups which cancel each other and produce a very small resultant. This may be due, at least in part, to the fact that the modal vector is estimated and "too good" a fit is obtained. Watson has refered to the robustness of the methods and concludes that for large k they appear to be robust. Intuitively a very small value of Z is less worrying than a very large one; but the subject needs further examination.

Table 2.15

Goodness of fit for classification age-sex.

Cell	Pearson's χ^2	d.f.	Rayleigh's test	d.f.
1,1	3.1005	2	0.5778	7
1,2	2.9083	3	1.1526	7
2,1	6.3231	4	1.3625	7
2,2	1.7856	5	1.8005	7
3,1	1.0925	1	0.6333	7
3,2	1.4501	2	0.8560	7

2.18. Test for constant k .

In the different methods of analysis shown in the previous sections, the concentration parameter k is assumed to be constant over all the cells in the table; this is analogous to the assumption of constant variance in the usual analysis of variance. The null hypothesis $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_q^2$ is usually tested using Bartlett's test; in this section we will describe Bartlett's test and then show an adaptation that can be used to test the null hypothesis $k_1 = k_2 = \ldots k_q$.

In using Bartlett's test we first calculate the joint estimate $S^2 = \Sigma_i v_i S_i^2 / \Sigma_i v_i$, where S_i^2 is the estimate of the variance in the i-th sample and v_i is its degrees of freedom. The test statistic is then:

$$B = \frac{1}{C} (v \,\ell n S^2 - \Sigma_i v_i \,\ell n S_i^2) \tag{2.16}$$

where

$$C = 1 + \frac{\Sigma_i (1/v_i) - 1/v}{3(q-1)} \quad \text{and} \quad v = \Sigma_i v_i \ .$$

For values of v_i of 5 or more the distribution of B is approximately χ^2 with q-1 degrees of freedom. Hence we would reject the hypothesis $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_q^2$ if the value of B were greater than $\chi^2_{(q-1)}(\alpha)$, where α is the chosen significance level. The quantity $1/C$ is less than one since C is always greater than one; hence, if the value of B is not significant when $C = 1$, it is unnecessary to include the term $1/C$.

The test statistic B is a function of v_i and S_i; the sample variances S_i are such that $S_i \approx \chi^2_{v_i} / v_i$, (i.e., a chi-square variable divided by its degrees of freedom). From Section 2.7 we have

$$2k(N_i - R_i) \approx \chi^2_{(p-1)(N_i-1)} \ , \ i = 1,\ldots,q \ .$$

Replacing S_i by $2k(N_i - R_i)/(p-1)(N_i-1)$ and v_i by $(p-1)(N_i-1)$ in equation 2.16, we obtain

$$B = \frac{1}{C}\{v\ \ell n(\frac{N-\Sigma_i R_i}{v}) - \Sigma_i v_i\ \ell n(\frac{N_i-R_i}{v_i})\}\ ,$$

where

$$C = 1 + \frac{\Sigma_i(1/v_i) - 1/v}{3(q-1)}\ ,$$

$$v_i = (N_i-1)(p-1) \qquad \text{and} \qquad v = \Sigma_i v_i = (N-q)(p-1)\ .$$

Therefore, as in the test for the equality of variances, the null hypothesis $k_1 = k_2 = \ldots = k_q$ will be rejected if the test statistic B is greater than $\chi^2_{q-1}(\alpha)$ where α is the chosen level of significance.

Example. In Section 2.13 we had classified the students according to their age and sex, the $\hat{k}$ values for the six cells are given in Table 2.16.

Table 2.16

$\hat{k}$ values for age and sex

		Age		
		< 21	21-25	> 25
Sex	Females	36.1655	37.1720	39.4926
	Males	36.4224	38.4809	38.8098

The test statistic B is equal to 1.7103, which is not significant when compared with χ^2_5; the hypothesis of equality of the k values is not rejected. In this case, the values of $\hat{k}$ are very similar and the test is almost not necessary; however, it gives an illustration of the method.

CHAPTER 3

Clustering.

3.1. Introduction.

Clustering techniques can be helpful in the analysis of p-dimensional unit vectors. In this chapter we present a method of clustering unit vectors which can be performed rather quickly without the use of a computer. The method is based on the dot products between pairs of individuals, which is a natural similarity measure for unit vectors.

The basic problem is as follows. Given a sample of N subjects, for each of which p variables are measured, a classification scheme is to be devised for grouping the subjects into g classes such that the members of any one class are similar to each other.

3.2. Distance Function.

The distance between the unit vectors $\underline{v}_i = OP_i$ and $\underline{v}_j = OP_j$ will be the metric defined by:

$$d(\underline{v}_i, \underline{v}_j) = \theta_{ij},$$ where θ_{ij} is the smaller angle between $\underline{v}_i$ and $\underline{v}_j$.

Since the hypersphere has radius 1, θ_{ij} is also the shortest (Euclidean) distance between the two points P_i and P_j on the surface of the hypersphere, and this angle is a distance.

The i-th and j-th individuals are assigned to the same cluster (i.e., $\underline{v}_i$ and $\underline{v}_j$ are similar) if the distance between the unit vectors $\underline{v}_i$ and $\underline{v}_j$ is "sufficiently small" and to different clusters if the distance between the pair of points is "sufficiently large".

3.3. Similarity.

As a complement to the notion of distance between $\underline{v}_i$ and $\underline{v}_j$, there is the idea of similarity between the two unit vectors. A non-negative real valued function $S(\underline{v}_i, \underline{v}_j) = S_{ij}$ is a <u>similarity measure</u> if:

(a) $0 \leq S(\underline{v}_i, \underline{v}_j) < 1$ for $\underline{v}_i \neq \underline{v}_j$;

(b) $S(\underline{v}_i, \underline{v}_j) = 1$ if and only if $\underline{v}_i = \underline{v}_j$;

(c) $S(\underline{v}_i, \underline{v}_j) = S(\underline{v}_j, \underline{v}_i)$ for all $\underline{v}_i, \underline{v}_j$.

Let $\underline{v}_i \cdot \underline{v}_j$ be the dot product between $\underline{v}_i$ and $\underline{v}_j$; then

$$\underline{v}_i \cdot \underline{v}_j = \sum_{k=1}^{p} x_{ik} x_{jk} .$$

There is a direct one to one correspondence between the distance metric given by θ_{ij} and the dot product $\underline{v}_i \cdot \underline{v}_j$, given by

$$\underline{v}_i \cdot \underline{v}_j = \cos \theta_{ij} .$$

As all the coordinates of the unit vectors are positive we have $0 \le \underline{v}_i \cdot \underline{v}_j \le 1$, $(i,j = 1,\ldots,N)$. In this case $\underline{v}_i \cdot \underline{v}_j$ is a suitable similarity measure. The pairwise similarities $S(\underline{v}_i , \underline{v}_j) = S_{ij}$ can be arranged in the similarity matrix shown in Figure 3.1.

$$\begin{bmatrix} 1 & S_{12} & \cdots & S_{1N} \\ S_{21} & 1 & \cdots & S_{2N} \\ \vdots & & & \\ S_{N1} & S_{N2} & \cdots & 1 \end{bmatrix}$$

Figure 3.1

Similarity Matrix

We say that the unit vectors $\underline{v}_i$ and $\underline{v}_j$ are similar, and so belong to the same cluster, if the similarity measure S between them is greater than y, where y is a value less than 1. In the examples given below y takes values between .90 and .95. We now discuss a procedure to divide a group of subjects into clusters.

3.4. Clustering Procedure.

To start the clustering procedure we select the pair of individuals, say $\underline{v}_1$ and $\underline{v}_2$, having the largest dot product in the similarity matrix. The matrix is then examined, and all those individuals (excluding $\underline{v}_2$) having a dot product with $\underline{v}_1$, greater than a given value y, are selected. From these we pick the one ($\underline{v}_3$, say) having the largest dot product with $\underline{v}_1$; if S_{32} is greater than y, $\underline{v}_3$ becomes a member of this cluster. From the remaining previously selected individuals we pick the individual ($\underline{v}_4$, say) having the largest dot product with $\underline{v}_1$; if all the dot products between $\underline{v}_4$ and the members of the cluster are greater than y, $\underline{v}_4$ becomes a member of the cluster. Those individuals which have a dot product with a member of the cluster not greater than y are eliminated. Once the first cluster is complete (i.e., all its elements are such that all the pairwise dot products are greater than y), a new cluster is started by picking, from the remaining individuals, the pair of vectors having the largest dot product.

3.5. Examples of clustering.

The second data set consists of consumption of selected foods in 45 countries. It has been obtained from the U.N. Statistical Yearbook (1971). Table A.1.2 gives the daily per capita consumption of the selected foods in the 45 countries reduced to unit vectors.

a. For $y = .95$ the following classification is obtained (the two countries underlined are those having the largest dot product, and were used to start clusters):

Cluster 1: Argentina, Australia, Austria, Canada, Costa Rica, Czechoslovakia, Denmark, England, Greece, Israel, Italy, Netherlands, New Zealand, Poland, Spain, Soviet Union, United States.

Cluster 2: Cuba, Cyprus, Honduras, Japan, Lebanon, Philippines, Portugal, Singapore, Turkey, Yugoslavia.

Cluster 3: Bolivia, Brazil, Colombia, Venezuela.

Cluster 4: Algeria, Egypt, India, Mexico, Saudi Arabia.

Cluster 5: China, Kenya.

Cluster 6: South Africa, Yemen.

Cluster 7: Ethiopia, Thailand.

Cluster 8: Congo, Gabon.

Cluster 9: Liberia.

b. With $y = .9$ we obtain the following clusters:

Cluster 1: Argentina, Australia, Austria, Canada, Colombia, Costa Rica, Cuba, Cyprus, Czechoslovakia, Denmark, England, Greece, Honduras, Israel, Italy, Lebanon, Netherlands, New Zealand, Poland, Singapore, South Africa, Spain, Soviet Union, United States, Venezuela, Yugoslavia.

Cluster 2: Algeria, Egypt, Ethiopia, India, Japan, Mexico, Philippines, Saudi Arabia, Thailand, Turkey.

Cluster 3: Congo, Gabon.

Cluster 4: Bolivia, Brazil, Kenya, Portugal.

Cluster 5: China, Liberia.

Cluster 6: Yemen.

3.6. Comparisons.

The results shown in Section 3.5 were compared to those obtained by means of the CLUSTAN package, a standard system of clustering algorithms. The package contains several options for the calculation of distances and for the algorithm or method of clustering. We used the following three different options for the calculation of distances between groups in terms of distances between pairs.

a. Nearest neighbour. The distance between two groups is defined as the distance between their nearest members.

b. Furthest neighbour. The distance between two groups is defined as the distance between their most remote pair of individuals.

c. Group average. The distance between two groups is defined as the average of the distances between all pairs of individuals in the two groups.

For evaluating the distances between pairs in the above options, there are several similarity and distance measures available. The dot product is one of them, nevertheless, the Euclidean distance was the one used in the results shown in this section.

From the different methods of clustering available in the package (hierarchic fusion, monothetic division, iterative relocation, etc.), we chose hierarchic fusion. At the beginning of this method, each individual is considered as a separate cluster. In the first iteration, the two closest individuals (according to the distance options previously selected; e.g., nearest neighbour and

Euclidean distance) are fused into a new cluster. In each one of the subsequent iterations all pairwise distances between clusters are recalculated, and the pair of clusters having the smallest distance is fused. The sets of clusters obtained using the three different methods to calculate the distances are:

a. Hierarchic fusion/nearest neighbour method - 8 clusters.

Cluster 1: Algeria, Argentina, Australia, Austria, Brazil, Canada, Colombia, Costa Rica, Cuba, Cyprus, Czechoslovakia, Denmark, Egypt, England, Greece, Honduras, India, Israel, Italy, Japan, Lebanon, Mexico, Netherlands, New Zealand, Philippines, Poland, Portugal, Saudi Arabia, Singapore, South Africa, Turkey, Soviet Union, United States, Venezuela, Yemen, Yugoslavia, Spain.

Cluster 2: Bolivia.

Cluster 3: China.

Cluster 4: Congo, Gabon.

Cluster 5: Ethiopia.

Cluster 6: Liberia.

Cluster 7: Kenya.

Cluster 8: Thailand.

b. Hierarchic fusion/furthest neighbour method - results for 8 clusters.

Cluster 1: Algeria, Egypt, Ethiopia, India, Philippines, Saudi Arabia, Thailand.

Cluster 2: Costa Rica, Cuba, Czechoslovakia, Greece, Honduras, Israel, Italy, Japan, Poland, Portugal, Singapore, Spain, Turkey, Soviet Union.

Cluster 3: Argentina, Australia, Austria, Canada, Denmark, England, Netherlands, New Zealand, United States.

Cluster 4: Cyprus, Lebanon, Mexico, South Africa, Yemen, Yugoslavia.

Cluster 5: Bolivia, Liberia.

Cluster 6: China, Kenya.

Cluster 7: Brazil, Colombia, Venezuela.

Cluster 8: Congo, Gabon.

c. Hierarchic fusion/group average method - results for 4 clusters.

Cluster 1: Argentina, Australia, Austria, Brazil, Canada, Colombia, Costa Rica, Cuba, Cyprus, Czechoslovakia, Denmark, England, Greece, Honduras, Israel, Italy, Japan, Lebanon, Mexico, Netherlands, New Zealand, Poland, Portugal, Singapore, South Africa, Spain, Turkey, Soviet Union, United States, Venezuela, Yemen, Yugoslavia.

Cluster 2: Algeria, Egypt, Ethiopia, India, Philippines, Saudi Arabia, Thailand.

Cluster 3: Bolivia, China, Kenya, Liberia.

Cluster 4: Congo, Gabon.

Although all the results obtained in Sections 3.5 and 3.6 are different, there are some basic similarities.

a. Argentina, Australia, Austria, Canada, Denmark, England, Netherlands, New Zealand, United States are always members of the same cluster.

Other groups which are always members of the same cluster are

b. Algeria, Egypt, India, Saudi Arabia.

c. Cyprus, Lebanon and Yugoslavia.

d. Costa Rica, Czechoslovakia, Greece, Israel, Italy, Poland, Spain, Soviet Union.

e. Colombia and Venezuela.

f. Congo and Gabon. These two countries are always in a cluster by themselves.

Of all 45 countries, Canada and New Zealand are the closest.

The clustering method introduced in Section 3.4 is essentially a hierarchical technique using the nearest neighbour method: a similarity matrix is computed, and at the beginning of the procedure each individual is considered as a separate cluster; a cluster center is formed by taking the closest pairs and individuals are agglomerated to these centers, in an ordered way that depends on how close they are to the center. The main difference between the method in Section 3.4 and the one used by CLUSTAN is that in the former, the pairwise similarity measures are not recalculated after each step.

If the aim in the cluster analysis is to look for natural groupings in the data, it is not important to pre-determine the number of clusters wanted in the solution. But it sometimes happens that the number of clusters to be obtained is fixed. If this is so, an appropriate selection of y will generally lead to the desired number of clusters; it may be that the correct value of y must be obtained by several trials.

3.7. Comments.

The method of clustering proposed in this section is rather informal, and we present it as an example of the use of the dot product as a natural similarity measure for unit vectors. The choice of the critical value y is arbitrary and it will affect the clusters obtained. In more sophisticated algorithms, such as the ones used in CLUSTAN, it is possible to see how the clustering is affected by different critical values.

CHAPTER 4

Examples of directional techniques.

In this chapter we do a more detailed analysis of the data sets previously introduced, the data on activity patterns of the students and the data on consumption of selected foods in 45 countries. Two new sets of data, products marketed by lumber companies in Canada and a set of ranked preferences expressed as unit vectors, are introduced and analyzed using some of the techniques discussed in Chapters 2 and 3.

4.1. Analysis of activity patterns of students.

In order to do a more detailed analysis of the students' activity pattern, the 130 students were split into groups according to the following classifications:

a. Sex:
   1 - females
   2 - males

b. Age:
   1 - less than 21
   2 - between 21 and 25
   3 - more than 25

c. Living arrangements:
   1 - students living alone
   2 - students living in a marriage like relationship
   3 - other (residence, coop house, etc.)

d.	Major subject:	1 - Economics and Commerce
		2 - Psychology
		3 - Geography
		4 - Criminology
		5 - Mathematics and Computing Science
		6 - Not declared
		7 - Joint major
		8 - Other
e.	Job:	1 - students with full or part time job
		2 - students without a job
f.	Year:	1 - students in first year
		2 - students in second year
		3 - students in third or fourth year.

Table 4.1 gives the cell size, estimate of the concentration parameter k and resultant for the groups in each of the classifications.

The activity patterns of the 130 students were analyzed using the one and two way ANOVA for resultants. Table 4.2 shows a summary of some of the results. The left hand side of the table gives the conclusion obtained from the one way ANOVA to the test for difference in activity pattern between the groups of classification 1.

The model for the two way analysis assumes a classification to be nested within another classification; the right hand side of the table gives the conclusions for the test of difference between

Table 4.1

Statistics for classifications of student data.

Classification	Groups	N_i	$\hat{k}_i$	R_i
Sex	females	56	35.6640	50.5041
	males	74	35.7440	66.7540
Age	< 21	47	36.2069	42.4567
	21 - 25	61	36.8403	55.2047
	> 25	22	38.1845	19.9835
Living	alone	18	37.4440	16.3175
	marriage	28	34.8764	25.1901
	other	84	38.1513	76.2938
Major	Econ/Comm.	53	42.1503	48.5991
	Psychology	11	43.1244	10.1244
	Geography	5	66.8862	4.7384
	Criminology	4	44.7705	3.6873
	Math/Comp. Sci.	6	98.6914	5.6352
	Not declared	24	34.1536	21.5436
	Joint major	4	88.2347	3.8413
	Other	23	35.8558	20.7549
Job	Yes	56	34.1215	50.2558
	No	74	39.4952	67.4422
Year	First	52	35.5432	46.4911
	Second	49	38.4072	44.5347
	Third and Fourth	29	42.6957	26.6227

Table 4.2

Summary of results for activity patterns of students.

(A result in the second column should read "between classification 1 within classification 2")

Between Classification 1	Conclusion	Within Classification 2	Conclusion
Sex	NS		
Age	S	Sex: females	NS
		males	S
		Living: alone	NS
		marriage	NS
		other	NS
		Job: Yes	NS
		No	S
		Year: First	S
		Second	S
		Third and Fourth	NS
Living	S	Sex: females	NS
		males	S
		Age: < 21	NS
		21 - 25	S
		> 25	NS
		Job: Yes	S
		No	S
		Year: First	NS
		Second	S
		Third and Fourth	S

Table 4.2

(Continuation)

Between Classification 1	Conclusion	Within Classification 2	Conclusion
Job	S	Sex: females	S
		males	S
		Age: < 21	NS
		21 - 25	S
		> 25	S
		Living: alone	NS
		marriage	S
		other	NS
		Year: First	S
		Second	S
		Third and Fourth	NS
		Major: Econ/Comm.	NS
		Psychology	NS
		Geography	NS
		Criminology	S
		Math/Comp. Sci.	NS
		Not declared	S
		Joint major	S
		Other	S

Table 4.2

(Continuation)

Between Classification 1	Conclusion	Within Classification 2	Conclusion
Year	S	Sex: females	NS
		males	S
		Age: < 21	S
		21 - 25	NS
		> 25	NS
		Living: marriage	S
		alone	NS
		other	NS
		Job: Yes	S
		No	S
Major	S	Sex: females	S
		males	S
		Job: Yes	S
		No	S

the groups in classification 1 within classification 2. The significance level used throughout the table is $\alpha = .05$, and NS and S stand for "not significant" and "significant" respectively.

It can be seen from Table 4.2 that "sex" is the only classification where there is not a significant difference between the groups. For all the other classifications, the two way ANOVA shows in general where the differences lie. For instance, the difference between the groups in classification "age" can be found to be significant in the males, in those students who do not have a job and in the first and second year students. However, it is possible that a significant difference is found between the groups of classification 1, but the tests obtained in the two-way ANOVA for classification 1 within the groups of classification 2 will not be significant. None of the test statistics for the analysis of "age" within the groups of classification "living" is significant.

4.2. Analysis of consumption of selected foods.

In Chapter 3 we presented clusterings for the data set on selected foods for 45 countries. We now give a further analysis of this set of data using the one way ANOVA technique.

Test for difference between regions.

The countries were split into 7 groups defined by geographical regions. Oceania and North America were pooled into the same group due to the large values of the concentration parameter k, (4,858.2 and 2,212.9 for Oceania and North America respectively). The group obtained after pooling them still had a large value of k

(1,336.3) when compared to the other groups. Bartlett's test for the equality of concentration parameters was used and the test statistic obtained was B = 66.2491 which is significant when compared to χ^2_6. Nevertheless we include the one way ANOVA because Z_2 is highly significant (see Table 4.3). We reject the null hypothesis that proportional consumption of foods is similar for the 7 groups.

Table 4.3

Statistics and ANOVA table for difference between regions.

Group	Region	N_i	$\hat{k}_i$	R_i
1	Oceania/North America	4	1336.2991	3.9910
2	Latin America	9	44.8009	8.3973
3	Eastern Europe	4	178.8693	3.9329
4	Western Europe	8	78.7312	7.6952
5	Africa	8	16.3001	6.5276
6	Near East	6	105.9430	5.8301
7	Far East	6	108.4621	5.8340

ANOVA Table

Sum of squares		df	test
Between groups	$\Sigma_i R_i - R = 42.2082 - 39.5037 = 2.7045$	36	$Z_2 = 6.1352$
Error	$N - \Sigma_i R_i = 45 - 42.7018 = 2.7918$	228	
Total	$N - R = 45 - 39.5037 = 5.4963$	264	

Test for difference between protein groups.

The countries were split into three groups according to the average daily protein consumption. The groups are as follows:

Group 1: Protein consumption below recommended minimum (less than 51 grammes): Bolivia, Colombia, Congo, Honduras, India, Liberia, Thailand.

Group 2: Average recommended protein consumption (51 - 70 grammes). Algeria, Argentina, Brazil, China, Costa Rica, Cuba, Gabon, Kenya, Lebanon, Mexico, Philippines, Saudi Arabia, Singapore, Venezuela, Yemen.

Group 3: Protein consumption above average (more than 70 grammes). Australia, Austria, Canada, Cyprus, Czechoslovakia, Denmark, Egypt, England, Ethiopia, Greece, Israel, Italy, Japan, Netherlands, New Zealand, Poland, Portugal, South Africa, Spain, Turkey, Soviet Union, United Stated, Yugoslavia.

The statistics and ANOVA table are given in Table 4.4. The test statistic Z_2 is significant; we reject the null hypothesis that the proportional food intake is similar for the three groups.

Another grouping was obtained for daily per capita calorie consumption and we shall examine these groups also for significant differences. Note that both the protein and calorie consumption information was obtained from a different U.N. publication and do not form part of the original data.

Table 4.4

Statistics and ANOVA table for differences between protein groups.

Group	Protein consumption	N_i	$\hat{k}_i$	R_i
1	< 51 grammes	7	20.5857	5.9799
2	51 - 70 grammes	15	27.1992	13.3455
3	> 70 grammes	23	40.2225	21.2845

ANOVA Table

	Sum of squares	df	test
Between groups	$\Sigma_i R_i - R$ = 40.6010 - 39.5037 = 1.0973	12	Z_2 = 5.2919
Error	$N - \Sigma_i R_i$ = 45 - 40.6010 = 4.3990	252	
Total	$N - R$ = 45 - 39.5037 = 5.4963	264	

Test for difference between calorie groups.

The countries were next split according to the daily per capita calorie consumption. The three groups obtained are:

Group 1: Below average recommended (less than 2,001):

Algeria, Bolivia, Honduras, India, Philippines.

Group 2: Average recommended (2,001 - 3,000):

Australia, Brazil, China, Colombia, Congo, Costa Rica, Cuba, Cyprus, Egypt, Ethiopia, Gabon, Greece, Israel, Italy, Japan, Kenya, Lebanon, Liberia, Mexico, Portugal, Saudi Arabia, Singapore, South Africa, Spain, Thailand, Turkey, Venezuela, Yemen.

Group 3: Above average (more than 3,000):

Argentina, Austria, Canada, Czechoslovakia, Denmark, England, Netherlands, New Zealand, Poland, Soviet Union, United States, Yugoslavia.

Table 4.5 gives the results for the test. We reject the null hypothesis that the proportional food intake is similar for the three calorie groups.

Table 4.5

Test for difference between calorie groups.

Group	Calorie consumption	N_i	$\hat{k}_i$	R_i
1	< 2,001	5	44.7467	4.6648
2	2,001 - 3,000	28	25.3749	24.6896
3	> 3,000	12	76.5153	11.5295

ANOVA Table

	Sum of squares	df	test
Between groups	$\Sigma_i R_i - R = 40.8839 - 39.5037 = 1.3802$	12	$Z_2 = 7.0419$
Error	$N - \Sigma_i R_i = 45 - 40.8839 = 4.1161$	252	
Total	$N - R = 45 - 39.5037 = 5.4963$	264	

4.3. Analysis of lumber companies.

The third data set was brought to the author by Professor Schwindt of the Department of Economics. The data concerns the products produced by various companies in the lumber industry. There are twenty-nine companies in all marketing wood based products in Canada. The production, expressed in Canadian dollars, of the eight most important products in the forest industry (i.e., newsprint, market pulp, wrapping paper, paperboard, fine paper, sanitary and tissue paper, lumber, plywood) is given for each one of the companies in Table A.1.3, Appendix 1.

Table A.1.3 shows that while some companies are very diversified (i.e., they produce many of the important wood based products), other companies limit their products to only one or two categories.

The more diversified companies tend to be more vertically integrated; they are using the output of one stage of production as an input to the next stage. Therefore, a company that has production in all eight categories would gain considerable advantages in the market.

As none of the companies in the sample is diversified to the extent of marketing all eight products, it was decided to obtain a "measure of diversification" that would allow the 29 companies to be compared to an ideally diversified company. For each company the production in dollars was converted to a percentage of the total, for each of the eight products. For a typical company, the component in the i-th direction, x_i is the square root of the proportion for the i-th product, $i = 1,\ldots,8$. Thus each company is an 8-dimensional unit vector on the surface of the hypersphere of unit radius. The ideal company was obtained from the average of the total Canadian shipments (for the 8 selected products) over a ten year period. The similarity measure between a typical company and the ideal company is then the scalar product between these two companies. If the scalar product is close to one, the company considered is highly diversified, while if the scalar product is close to zero, the company is not very diversified. Table 4.6 shows, in descending order the similarity for the 29 firms. It can be seen that the large companies (with respect to the production in millions of dollars) tend to be more integrated than the smaller companies. This is mainly due to the fact that small companies do not market many of the products included in the analysis, and therefore the unit vectors corresponding to these companies have several zero components.

Table 4.6

Similarity measure between lumber companies and the ideal company.

(The scalar product is the scalar product between the given company and the ideal company).

Firm	Scalar Product	Firm	Scalar Product	Firm	Scalar Product
MACB	.9593	CRES	.7531	WELD	.6781
CRZE	.9465	CFPR	.7465	KIMB	.6039
BCFP	.9035	CIPA	.7454	KPNP	.5737
DOMT	.8309	BCAS	.7253	BOWN	.5727
CONS	.8151	NPTI	.7225	WHON	.5272
GLPA	.8139	WEYE	.7127	DOMA	.5272
ABIT	.8122	FCOS	.7113	WEST	.5272
ONPA	.7835	CCEL	.6950	SCOT	.4612
REED	.7797	RAYC	.6824	ROLL	.2366
PRCO	.7746	EDDY	.6810		

Test for difference between Canadian and Foreign owned companies.

The first analysis is to examine whether there was a difference in the proportions of products marketed by companies operating with foreign capital (namely CIPA, CRZE, WELD, REED, ONPA, BOWM, BCAS, SCOT, CRES, KIMB, WEYE, NPTI) and those operating with Canadian capital. The results appear in Table 4.7. The test statistic is not at all significant, so that it appears likely that we can accept that the proportions of products marketed by companies in the two groups are the same. This implies that there is no difference in diversification between the groups.

Table 4.7

Statistics and ANOVA table for foreign/Canadian classification.

Group	Capital	N_i	$\hat{k}_i$	R_i
1	Foreign	13	12.1200	9.2459
2	Canadian	16	10.2127	10.5166

ANOVA Table

	Sum of squares	df	test
Between groups	$\Sigma_i R_i - R = 19.7625 - 19.4463 = 0.3162$	7	$Z_2 = 0.9245$
Error	$N - \Sigma_i R_i = 29 - 19.7625 = 9.2375$	189	
Total	$N - R = 29 - 19.4463 = 9.5537$	196	

Test for difference between large, medium and small companies.

The companies were also split into three groups, according to their production in millions of dollars.

The statistics and ANOVA table are given in Table 4.8. For the test for no difference between groups, the normal approximation for the test statistic $Z_2 = 2.6450$ with 14 and 182 degrees of freedom is 2.9529 (Peizer and Pratt), 2.9577 (Carter), or 2.9431 (Paulson), therefore Z_2 is highly significant. We reject the null hypothesis that there is no difference in the proportional outputs for the three groups of companies. Bartlett's test for the equality of concentration parameters gave a test statistic B = 15.8195 which is significant when compared to the percentage points of the χ^2 distribution with 2 degrees of freedom, but the Z_2 is sufficiently large that the above conclusion is still valid.

Table 4.8

Statistics and ANOVA table for classification "production".

Group	Production	N_i	$\hat{k}_i$	R_i
1	less than 250	18	9.9247	11.6522
2	250 - 500	6	18.7212	4.8783
3	more than 500	5	37.3068	4.5389

ANOVA Table

Sum of squares		df	test
Between groups	$\Sigma_i R_i - R = 21.0614 - 19.4463 = 1.6151$	14	$Z_2 = 2.6450$
Error	$N - \Sigma_i R_i = 29 - 21.0614 = 7.9386$	182	
Total	$N - R = 29 - 19.4463 = 9.5537$	196	

Clustering.

The companies were clustered using the technique described in Chapter 3. The following clusters were obtained for y = .8.

Cluster 1: WHON, DOMA,WEST.

Cluster 2: FCOS, CCEL, RAYC, CRES, NPTI, WEYE.

Cluster 3: MACB, BCFP, CRZE.

Cluster 4: ABIT, CONS, CIPA, PRCO, REED, ONPA.

Cluster 5: SCOT, KIMB.

Cluster 6: ROLL.

Cluster 7: CFPR, WELD.

Cluster 8: KPNP, BOWM.

Cluster 9: DOMT, BCAS.

Cluster 10: EDDY.

Cluster 11: GLPA.

These clusters correspond to how integrated the companies are. Cluster 3 contains the three most integrated companies (their dot product with the ideal company being greater than 0.9); other very integrated companies are found as members of clusters 4, 9 and 11. Clusters 2, 7 and 10 contain moderately integrated companies and clusters 1, 5, 6 and 8 contain the companies which are not strongly integrated.

4.4. Analysis of data on occuptional prestige.

In this section we do the analysis of a set of sociological data obtained from Professor Charles Jones of McMaster University. The set originates in the ranks and ratings given by 48 subjects to 16 occupations, according to different criteria. The occupational titles included are as follows:

1. Church of Scotland Minister
2. Comprehensive School Teacher
3. Qualified Actuary
4. Chartered Accountant
5. Male Psychiatric Nurse
6. Ambulance Driver
7. Building Site Labourer
8. Machine Tool Operator
9. Country Solicitor
10. Civil Servant
11. Commercial Traveller
12. Policeman
13. Carpenter
14. Lorry Driver
15. Rail Porter
16. Barman.

The 48 subjects were asked to rank or rate the occupations according to 4 different criteria. In the first part, subjects were asked to rank-order the 16 occupations for the criteria "degree of general standing in the community" (social standing criterion) and "prestige or rewards which the job-holders ought to receive" (rewards criterion). In the rating task, subjects were told to consider each occupation and award them a score according to two criteria:

"usefulness to society" (social usefulness criterion) and "estimated income received" (earnings criterion).

Ratings and rank orderings were transformed into 3-dimensional unit vectors by means of a multidimensional scaling method explained in Coxon and Jones (1978). The original data can be found in Coxon and Jones (1979), and the data in terms of unit vectors is given in Table A.1.4.

The 48 subjects are grouped as follows:

Group 1: 8 Church of Scotland Ministers
(initial letter A in Table A.1.4)
3 Episcopalian ministers
(initial letter B)
1 school teacher
(initial letter C)

Group 2: 6 actuaries
(initial letter D)
6 chartered accountants
(initial letter E)

Group 3: 2 male psychiatric nurses
(initial letter K)
8 ambulance drivers
(initial letter L)
2 policemen
(initial letter M)

Group 4: 3 joiners

(initial letter P)

3 plasterers

(initial letter Q)

5 burner fitters

(initial letter R)

1 ship's joiner

(initial letter S).

It was desired to test whether the four groups of subjects gave the same ranks or ratings (according to the 4 criteria) to the 16 occupations. A one way ANOVA for resultants was done for each criterion. The statistics and results are shown in Table 4.9. None of the test statistics is significant; we do not reject the null hypothesis that the 4 groups of subjects give the same ranks and ratings to the 16 occupations.

Table 4.9

Statistics and results for occupational prestige data.

Social Usefulness Criterion

Group	N_i	$\hat{k}_i$	R_i
1	12	5.4068	9.7806
2	12	7.6916	10.4399
3	12	9.6733	10.7594
4	12	4.6393	9.4134

ANOVA Table

	Sum of squares	df	test
Between groups	0.7735	6	$Z_2 = 1.4915$
Error	7.6067	88	
Total	8.3802	94	

Rewards Criterion

Group	N_i	$\hat{k}_i$	R_i
1	12	7.1736	10.3279
2	12	10.8301	10.8919
3	12	3.6442	8.7077
4	12	3.6182	8.6834

ANOVA Table

	Sum of squares	df	test
Between groups	0.6873	6	$Z_2 = 1.0735$
Error	9.3903	88	
Total	10.0776	94	

Social Standing Criterion

Group	N_i	$\hat{k}_i$	R_i
1	12	5.173	9.6805
2	12	19.3659	11.3804
3	12	33.0498	11.6369
4	12	3.6820	8.7409

ANOVA Table

	Sum of squres	df	test
Between groups	0.2377	6	$Z_2 = 0.5313$
Error	6.5674	88	
Total	6.7991	94	

Earnings Criterion

Group	N_i	$\hat{k}_i$	R_i
1	12	10.1633	10.8193
2	12	49.9144	11.7596
3	12	9.3177	10.7121
4	12	13.4571	11.1083

ANOVA Table

	Sum of squres	df	test
Between groups	0.3777	6	$Z_2 = 1.5383$
Error	3.6007	88	
Total	3.9784	94	

CHAPTER 5

Suggestions for further work.

In the previous chapters we made several implicit assumptions for the use of the methodology, namely:

1. All the cells in the two way ANOVA should have at least one individual.
2. The unit vectors in a data set should not have many components equal to zero.
3. The concentration parameter k should be constant over the cells.

The robustness of the techniques when these assumptions fail needs further work, and the following comments may be useful.

5.1. Two way ANOVA with cells having zero individuals.

Suppose we have two classifications and we choose to do a two way ANOVA for classification 2 (columns) within classification 1 (rows). Furthermore, suppose that cell (h, k) contains no individuals (as shown in Figure 5.1). In this case we can use an ANOVA table in which the degrees of freedom have been adjusted to take into account a cell with zero individuals (Table 5.1). Similar adjustments can be made when several cells have zero individuals.

Classification 2

Classification 1	1	2	k	J
1	N_{11}	N_{12}	N_{1k}	N_{1J}
2	N_{21}		⋮	
h	. . . . .	. . . . .	0	
I	N_{I1}			

Figure 5.1

Two way classification with zero individuals in one cell.

Table 5.1

Adjusted ANOVA table for "columns within rows"

Sum of squares		Usual df.	Adjusted df.
Between rows	$\Sigma_i R_i - R_{..}$	$(p-1)(I-1)$	$(p-1)(I-1)$
Between columns			
Within row 1	$\Sigma_j R_{1j} - R_{1.}$	$(p-1)(J-1)$	$(p-1)(J-1)$
Within row 2	$\Sigma_j R_{2j} - R_{2.}$	$(p-1)(J-1)$	$(p-1)(J-1)$
⋮			
Within row h	$\Sigma_j R_{hj} - R_{h.}$	$(p-1)(J-1)$	$(p-1)(J-2)$
⋮			
Within row I	$\Sigma_j R_{Ij} - R_{I.}$	$(p-1)(J-1)$	$(p-1)(J-1)$
Error	$N-\Sigma_{ij} - R_{ij}$	$(p-1)(N-IJ)$	$(p-1)(N-IJ+1)$
Total	$N-R_{..}$	$(p-1)(N-1)$	$(p-1)(N-1)$

5.2. Unit vectors with components equal to zero.

The effect of zero components is not known; however it suggests that the vectors may lie in a hypersphere of lower dimension. A solution to this problem is to pool components together. This is what was done with the data on activity patterns of students; there were originally 13 dimensions, but as many of the components were zero some components were pooled. The components corresponding to "family activities", "personal activities" and "job" were put together into a new component called "non course activities"; similarly, the components corresponding to "socializing with a group", "socializing with a person of the same sex", "socializing with a person of the opposite sex" and "other" were pooled into a new component called "socializing". The one way analysis of variance for the original data was done, and gave exactly the same conclusions as those for 8 components presented in Section 4.1.

In the data on lumber companies introduced in Section 4.3 many individuals present several components equal to zero; some companies have all but one component equal to zero. The original 8 components were pooled into 4 components: "Wrapping paper", "Paper Board", "Fine Paper" and "Sanitary and Tissue Paper" were pooled into "Paper"; "Lumber" and "Plywood" were pooled into "Wood". The analysis using 4 components is given in Tables 5.2 and 5.3. The conclusions are the same as those obtained in Section 4.3, and we give the results in detail so that comparisons can be made. Further work needs to be done on the effect of zero components.

Table 5.2

Statistics and ANOVA table for foreign/Canadian classification.

(4 components)

Group	Capital	N_i	$\hat{k}_i$	R_i
1	Foreign	13	5.5998	9.5177
2	Canadian	16	4.4831	10.6465

ANOVA Table

Sum of squares		df	test
Between groups	$\Sigma_i R_i - R$ = 20.1643 - 19.9696 = 0.1947	3	Z_2 = 0.5949
Error	$N - \Sigma_i R_i$ = 29 - 20.1643 = 8.8357	81	
Total	$N - R$ = 24 - 19.9696 = 9.0304	84	

Table 5.3

Statistics and ANOVA table for classification "production".

(4 components)

Group	Production	N_i	$\hat{k}_i$	R_i
1	less than 250	18	4.5070	12.0093
2	250 - 500	6	7.8957	4.8601
3	more than 500	5	14.8239	4.4941

ANOVA Table

Sum of squares		df.	test
Between groups	$\Sigma_i R_i - R$ = 21.3635 - 19.9696 = 1.3939	6	Z_2 = 2.3729
Error	$N - \Sigma_i R_i$ = 29 - 21.3635 = 7.6365	78	
Total	$N - R$ = 29 - 19.9696 = 9.0304	84	

5.3. Failure of the assumption of constant k.

When k is not constant from cell to cell, the situation is roughly analogous to a normal-theory analysis of variance with the error variance varying from cell to cell. It is known that this hetereogeneity of variance can affect the results, and various procedures have been suggested. For example, the cell means are sometimes weighted in proportion to the residual variance estimates, or a transformation is made of the original data. For vectors, it does not appear that straightforward adjustments can be made. Thus, when the tests are done in the usual way, the apparent conclusions should be interpreted with reserve, according to the degree of hetereogeneity of the k values. This too is a subject which needs further investigation.

Appendix 1

Data sets.

The four sets of data used in this thesis are given in Tables A.1.1 - A.1.4.

Table A.1.1

Time in hours spent in eight activities by 130 students at Simon Fraser University.

The activities considered are:

I. Sleeping.

II. Travelling to school (includes waiting for bus, looking for a parking space, etc.).

III. Attending lectures (seminars and tutorials).

IV. Studying (library, writing papers, reading for a course, etc.).

V. Sports.

VI. Socializing.

VII. Eating meals.

VIII. Non-course activities (personal activities, family/household activities, job).

The students were assigned codes according to the following characteristics:

Column 1: Sex: females = 1

males = 2

Column 2: Age: less than 21 = 1

between 21 and 25 = 2

more than 25 = 3

Column 3: Living arrangements: living alone = 1

marriage-like relationship = 2

other = 3

Column 4: Job: students who have a job = 1

students who do not have a job = 2

Column 5: Year: first year = 1

second year = 2

third or fourth year = 3/4

Column 6: Major: Economics/Commerce = 1

Psychology = 2

Geography = 3

Criminology = 4

Mathematics/Computing Science = 5/6

Not declared = 7

Joint major = 8

Other = 9

Table A.1.1

Activities

Codes	I	II	III	IV	V	VI	VII	VIII
2 1 3 2 2 1	7.25	0.75	3.00	5.50	1.50	2.00	2.50	1.50
2 2 3 2 1 7	4.50	0.50	3.00	4.00	1.00	3.00	6.00	2.00
2 1 3 1 3 9	7.76	0.71	4.71	4.71	1.18	1.41	1.18	2.35
1 3 2 1 2 2	9.25	1.00	2.25	7.50	0.75	0.75	0.75	1.75
2 1 1 1 1 1	7.00	0.75	4.00	4.00	1.00	2.25	4.00	1.00
1 2 2 1 1 2	6.96	0.52	2.09	4.87	0.70	0.52	6.95	1.39
2 1 3 1 1 1	4.50	0.75	6.00	7.25	0.75	1.00	2.00	1.75
2 1 1 2 1 1	9.00	0.75	3.00	7.00	0.75	3.50	0.0	0.0
1 2 3 1 3 1	8.25	0.25	3.00	5.00	1.00	4.00	1.00	1.50
2 2 3 2 1 1	6.22	1.78	4.44	5.33	2.67	1.78	1.77	0.0
2 2 3 1 1 8	8.00	0.25	0.75	4.50	3.00	0.75	0.75	6.00
2 1 3 2 2 4	8.16	2.97	2.97	3.96	1.24	1.98	1.73	0.99
1 1 3 1 1 4	9.00	1.75	2.00	2.00	1.00	0.0	0.0	8.25
1 2 3 2 3 1	8.50	1.00	2.00	2.00	3.00	3.00	1.00	3.50
1 1 3 2 1 4	6.00	0.50	7.00	6.00	1.50	0.0	2.00	1.00
1 1 1 2 1 7	7.00	0.75	4.00	5.00	1.50	3.00	1.00	1.75
2 3 2 2 1 8	7.50	0.75	3.00	4.00	0.50	0.25	2.00	6.00
1 1 3 2 1 1	7.00	1.75	2.00	4.00	0.75	6.00	2.50	0.0
1 2 3 1 2 9	5.50	0.50	3.00	3.00	1.00	1.00	7.00	3.00
2 1 3 2 1 7	8.00	0.50	2.00	3.00	1.50	0.0	2.00	7.00
2 1 3 1 2 1	10.75	0.50	2.75	0.25	0.75	3.25	0.0	5.75
1 1 1 2 1 1	10.75	0.0	3.00	2.00	0.50	3.00	0.0	4.75

Table A.1.1

(Continuation)

Codes	I	II	III	IV	V	VI	VII	VIII
2 3 3 1 1 1	7.00	1.00	4.00	6.00	1.00	0.0	0.50	4.50
2 1 3 2 1 9	8.91	0.25	2.97	4.95	0.49	1.98	1.48	2.97
2 1 3 2 1 1	7.85	1.31	0.65	0.87	1.75	5.23	4.37	1.96
2 1 3 2 3 9	8.00	0.50	2.50	6.00	0.75	1.00	0.0	5.25
2 3 3 1 2 9	7.38	0.46	2.77	4.62	1.38	1.38	5.07	0.92
1 2 3 1 4 2	7.00	0.50	4.00	7.50	1.75	2.00	0.75	0.50
2 1 3 2 1 1	8.00	0.50	1.00	2.00	1.00	2.50	4.00	5.00
2 3 2 1 2 1	6.00	2.50	2.00	5.00	2.00	0.0	3.50	3.00
2 2 3 1 3 9	5.00	1.00	5.00	5.00	0.25	0.0	0.50	7.25
2 2 3 2 2 1	6.00	0.50	2.00	11.00	0.0	1.00	0.50	3.00
2 1 3 2 1 7	6.50	0.50	3.00	4.00	2.00	3.00	5.00	0.0
2 2 3 2 1 1	6.72	1.92	3.84	4.32	1.44	1.44	3.36	0.96
2 3 2 2 2 9	3.96	0.25	0.0	14.85	0.99	0.0	3.96	0.0
2 2 1 2 2 9	10.00	0.0	0.0	4.50	2.00	0.0	2.50	5.00
2 3 2 1 2 9	7.00	0.50	0.0	4.00	1.00	1.00	1.00	9.50
2 1 3 2 1 8	7.00	0.50	3.00	1.00	0.50	1.00	3.00	8.00
2 2 3 1 1 2	9.00	1.25	0.0	4.00	0.75	3.00	3.50	2.50
2 3 2 2 2 1	8.00	0.25	3.50	9.00	1.50	0.25	1.50	0.0
1 1 3 2 1 7	7.00	0.50	4.50	1.00	1.00	6.00	3.00	1.00
2 2 1 1 1 1	10.00	1.00	1.00	2.00	1.00	1.00	3.00	5.00
2 2 3 1 3 7	4.00	2.25	0.0	4.00	0.75	0.0	0.0	13.00
2 2 3 2 3 5	8.00	0.50	2.00	7.00	1.00	1.50	1.00	3.00

Table A.1.1.

(Continuation)

Codes	I	II	III	IV	V	VI	VII	VIII
2 2 1 1 3 5	7.50	1.50	2.75	4.25	1.25	1.50	1.75	3.50
2 2 3 1 1 7	7.00	1.00	4.00	5.00	1.00	0.0	0.50	5.50
2 2 3 2 2 1	7.50	1.50	4.00	5.00	1.50	1.00	1.50	2.00
1 3 1 2 4 7	8.00	0.0	3.00	4.00	1.00	0.25	1.50	6.25
2 2 3 1 3 9	6.50	2.50	6.00	3.00	1.00	1.00	3.00	1.00
2 1 3 2 2 7	8.00	1.00	4.00	2.00	1.00	3.00	4.00	1.00
2 3 1 1 1 7	9.00	0.75	2.00	3.00	1.00	3.00	2.75	2.50
2 1 3 2 1 7	7.27	0.24	4.85	0.97	1.94	3.88	3.88	0.97
1 2 1 2 1 8	4.00	2.00	3.00	5.00	1.00	2.50	1.00	5.50
2 1 3 2 1 2	7.06	1.18	2.82	1.88	0.71	4.70	5.64	0.0
2 1 1 2 2 1	7.05	0.22	1.54	3.52	1.76	1.54	8.36	0.0
2 1 3 2 1 9	8.00	0.50	5.00	6.50	1.50	1.50	0.50	0.50
1 1 1 2 1 1	4.50	0.0	0.0	5.00	2.00	1.50	7.50	3.50
2 1 3 2 1 7	8.00	0.25	3.50	1.50	1.00	2.00	7.50	0.25
1 2 3 1 1 3	4.00	0.25	0.0	4.00	1.00	0.0	3.00	11.75
1 3 3 2 1 2	7.76	1.94	2.91	6.79	2.42	0.0	2.18	0.0
1 1 3 2 1 1	6.24	0.96	2.88	9.60	1.92	0.48	1.44	0.48
1 1 3 2 1 1	4.53	0.23	3.62	8.15	0.45	0.68	3.63	2.72
1 2 3 2 2 1	8.00	1.25	3.00	5.00	2.50	0.50	2.00	1.75
2 1 3 2 1 9	5.00	0.25	2.50	10.00	0.50	0.25	1.75	3.75
2 2 2 2 2 3	9.00	0.25	1.00	2.00	2.00	1.00	4.50	4.25
2 1 3 2 3 1	8.00	1.00	3.00	8.00	1.00	1.00	2.00	0.0

Table A.1.1

(Continuation)

Codes	I	II	III	IV	V	VI	VII	VIII
1 2 3 1 1 6	5.00	0.75	0.0	2.50	1.00	0.50	8.00	6.25
2 1 3 1 1 7	5.00	1.00	5.00	6.00	2.00	2.00	2.75	0.25
1 2 3 2 1 1	6.00	1.00	5.00	7.00	2.00	0.50	2.00	0.50
2 3 3 1 4 6	7.84	2.45	1.96	4.90	2.94	0.0	1.96	1.96
1 1 3 2 2 1	8.00	1.25	4.00	0.75	1.25	3.50	3.00	2.25
1 1 3 2 1 1	10.00	0.50	5.00	4.00	2.00	0.0	0.0	2.50
2 1 3 2 1 1	10.00	0.25	5.50	4.00	2.25	0.0	0.0	2.00
1 3 2 2 1 2	9.00	1.00	4.00	2.00	1.50	3.50	1.00	2.00
2 2 3 2 1 1	8.00	0.50	4.00	6.00	0.25	0.0	0.75	4.50
2 2 3 2 1 1	8.00	1.00	4.00	7.25	1.00	0.0	2.25	0.50
2 2 3 1 3 1	6.00	0.50	4.00	8.00	1.50	1.50	1.75	0.75
2 2 3 2 2 2	7.00	0.50	2.00	6.00	1.00	1.00	1.25	5.25
1 2 1 2 1 1	4.76	0.60	3.17	3.17	1.59	1.59	8.73	0.40
1 3 2 1 3 1	5.18	0.24	1.41	2.82	0.71	1.88	3.29	8.47
2 1 3 1 1 7	5.76	0.38	7.68	1.73	2.30	0.0	6.15	0.0
2 2 3 2 2 1	6.50	0.25	5.00	5.00	1.00	0.25	4.00	2.00
2 2 1 1 3 5	9.00	0.50	4.00	5.00	1.00	1.50	0.50	2.50
1 3 2 2 2 9	8.00	0.50	4.00	8.00	2.50	0.0	1.00	0.0
1 1 3 2 2 2	9.00	0.50	4.00	2.75	2.00	2.00	1.00	2.75
1 2 3 1 1 1	8.50	1.00	3.00	7.00	1.25	0.0	3.25	0.0
2 2 3 2 2 1	8.50	1.00	3.00	6.00	1.00	1.00	3.25	0.25
2 2 2 1 2 9	8.52	0.90	3.59	1.79	0.90	2.69	4.49	1.12

Table A.1.1.

(Continuation)

Codes	I	II	III	IV	V	VI	VII	VIII
2 1 3 2 2 9	5.00	1.00	2.00	9.00	0.0	2.00	3.50	1.50
1 3 2 1 2 7	6.00	0.50	4.00	4.00	1.00	0.50	2.00	6.00
2 3 2 1 2 7	6.37	0.49	2.94	2.94	1.96	2.45	1.96	4.90
2 3 2 1 2 7	7.00	1.00	3.00	2.00	1.00	4.00	3.00	3.00
1 1 2 1 4 9	7.50	0.50	2.00	6.00	1.00	1.00	2.00	4.00
1 2 2 1 2 7	6.50	2.00	3.00	0.0	1.00	1.50	1.00	9.00
1 1 3 1 1 7	7.00	1.00	3.00	5.00	2.00	0.50	4.00	1.50
1 3 2 1 2 7	8.00	2.00	2.00	2.00	1.00	2.00	7.00	0.0
1 3 2 2 2 9	6.37	0.49	4.90	5.88	0.98	1.96	1.96	1.47
2 3 2 1 1 7	7.92	0.74	1.98	1.98	1.48	0.0	0.98	8.91
1 2 3 2 2 1	7.50	0.75	3.00	6.50	1.00	0.75	2.50	2.00
2 2 2 1 3 3	5.00	2.50	1.00	6.50	1.00	0.50	1.00	6.50
1 2 3 2 2 1	6.00	2.00	2.00	4.00	0.0	3.00	5.00	2.00
1 2 3 1 3 2	5.50	2.50	0.0	4.50	0.0	2.00	7.00	2.50
1 2 3 1 3 1	8.00	0.50	3.00	4.00	1.25	4.25	1.25	1.75
2 1 3 1 2 1	5.00	0.50	5.00	6.00	0.50	3.00	2.50	1.50
1 2 3 1 2 9	6.00	0.50	2.50	2.50	1.00	1.25	7.00	3.25
2 2 2 1 2 9	5.00	0.75	2.75	2.75	1.00	1.25	6.75	3.75
2 2 3 1 4 1	7.00	0.75	4.00	7.00	2.00	1.75	1.25	0.25
1 1 3 2 1 7	6.50	0.50	3.00	4.00	1.75	3.50	4.75	0.0
1 2 3 2 3 5	7.50	0.75	2.00	7.00	1.00	1.75	1.00	3.00
1 2 3 1 1 7	7.00	2.00	4.00	4.75	0.75	0.0	0.25	5.25

Table A.1.1.

(Continuation)

Codes	I	II	III	IV	V	VI	VII	VIII
2 2 3 1 2 9	4.89	1.69	4.51	2.26	3.63	3.63	2.64	0.75
1 2 3 1 3 9	6.93	2.97	5.44	2.97	0.99	0.99	2.72	0.99
1 1 1 2 2 1	7.11	0.22	1.56	3.56	1.78	1.56	8.22	0.0
2 1 3 2 2 1	7.18	0.22	1.57	3.59	1.79	1.35	8.30	0.0
1 1 1 2 2 1	6.92	0.43	1.51	3.46	1.73	1.51	8.44	0.0
2 2 1 2 2 1	6.61	0.66	1.76	2.64	1.76	1.98	8.36	0.22
1 1 3 2 3 1	7.60	1.90	2.85	7.60	0.95	0.95	2.14	0.0
2 2 3 2 3 1	8.08	0.95	2.85	7.60	1.43	1.66	1.43	0.0
1 1 3 2 3 1	9.00	1.00	2.00	6.00	1.00	1.00	2.00	2.00
1 2 2 1 2 9	7.71	0.86	3.43	1.71	1.50	2.57	4.08	2.15
1 2 3 2 2 9	4.00	1.00	2.00	7.00	0.0	2.50	2.50	5.00
2 3 2 1 2 7	6.00	0.42	3.00	3.00	1.00	0.75	3.00	6.83
1 2 3 2 2 1	7.50	0.75	3.00	6.00	1.00	0.92	2.75	2.08
2 2 2 1 3 3	6.00	2.00	1.00	5.00	1.00	0.50	1.75	6.75
1 2 1 2 2 1	5.00	2.50	1.00	3.50	0.0	3.50	4.50	4.00
2 2 3 2 2 4	8.00	0.50	6.00	3.00	1.00	3.00	1.50	1.00
1 2 3 2 2 1	7.50	0.75	3.00	5.00	1.00	0.75	2.75	3.25
2 2 2 1 3 3	5.00	2.50	1.00	6.00	1.00	0.50	1.00	7.00
1 2 2 2 2 1	5.70	1.90	1.90	4.28	0.0	2.61	4.75	2.85
1 2 2 1 3 2	5.50	2.50	0.50	4.50	0.0	2.00	6.75	2.25

Table A.1.2

Daily per capita consumption in grammes of selected foods in 45 countries during 1968-1969 , converted to unit vectors.

The selected foods are:

Cereals:	flour and milled rice.
Starchy foods:	potatoes, sweet potatoes, cassava, manioc flour, potato flour, and other root flour. It also includes plantains and bananas when considered staple foods.
Sugar:	refined sugar, crude sugar, syrups, honey and other sugar products.
Seeds:	including also shelled equivalent for nuts, pulses and cocoa beans.
Meats:	poultry and game; expressed in terms of dressed carcass weight, including edible offals.
Milk:	milk and milk products excluding butter.
Fats:	fats and oils.

Table A.1.2

		Cereals	Starches	Sugar	Seeds	Meat	Milk	Fats
1. Algeria	(ALG)	0.8045	0.2775	0.2806	0.1433	0.1940	0.3361	0.1602
2. Argentina	(ARG)	0.4501	0.4282	0.2844	0.0817	0.4889	0.5001	0.1926
3. Australia	(AUS)	0.4053	0.3273	0.3050	0.0877	0.4388	0.6409	0.1561
4. Austria	(AUT)	0.4317	0.3771	0.2603	0.1051	0.3741	0.6337	0.2254
5. Bolivia	(BOL)	0.5293	0.7025	0.2421	0.1082	0.2674	0.2629	0.1234
6. Brazil	(BRA)	0.4309	0.6393	0.2904	0.2656	0.2716	0.4057	0.1102
7. Canada	(CAN)	0.3484	0.3724	0.3026	0.0814	0.4097	0.6627	0.1857
8. China	(CHI)	0.7202	0.5754	0.1157	0.2286	0.2510	0.1098	0.0968
9. Colombia	(COL)	0.4459	0.5752	0.3465	0.1211	0.2724	0.4993	0.1091
10. Congo	(CON)	0.1853	0.9553	0.0787	0.1232	0.1186	0.1006	0.0855
11. Costa Rica	(CSR)	0.5159	0.3819	0.3619	0.1846	0.2749	0.5633	0.1726
12. Cuba	(CUB)	0.5385	0.4571	0.3630	0.1900	0.3292	0.4592	0.1169
13. Cyprus	(CYP)	0.6195	0.3258	0.2711	0.2049	0.3582	0.4866	0.1707
14. Czechoslovakia	(CZE)	0.5132	0.4725	0.2792	0.0778	0.3555	0.5191	0.1825
15. Denmark	(DEN)	0.3533	0.3777	0.2986	0.0679	0.3350	0.6896	0.2255
16. Egypt	(EGP)	0.8196	0.1922	0.2548	0.1892	0.1979	0.3665	0.1296
17. England	(ENG)	0.3677	0.4329	0.2972	0.1097	0.3704	0.6310	0.2037
18. Ethiopia	(ETH)	0.7822	0.3356	0.1054	0.2912	0.2715	0.3006	0.1236
19. Gabon	(GAB)	0.1624	0.9460	0.0765	0.0733	0.2286	0.0988	0.0733
20. Greece	(GRE)	0.5243	0.3656	0.2156	0.1954	0.3036	0.6099	0.2058
21. Honduras	(HON)	0.6273	0.4187	0.2860	0.2162	0.2250	0.4821	0.1395
22. India	(IND)	0.7695	0.2468	0.2622	0.2766	0.0790	0.4257	0.1185
23. Israel	(ISR)	0.5157	0.2913	0.3045	0.1529	0.3845	0.5818	0.2112
24. Italy	(ITA)	0.5535	0.3301	0.2520	0.1493	0.3353	0.5815	0.2192
25. Japan	(JAP)	0.6573	0.4409	0.2814	0.2375	0.2218	0.3950	0.1800
26. Kenya	(KEN)	0.6299	0.5684	0.1863	0.2733	0.2406	0.3276	0.0725
27. Lebanon	(LEB)	0.6493	0.2548	0.2787	0.1854	0.3070	0.5196	0.1922
28. Liberia	(LIB)	0.5424	0.7987	0.0833	0.0833	0.1530	0.1381	0.1062
29. Mexico	(MEX)	0.6725	0.2043	0.3606	0.3031	0.2561	0.4328	0.1761
30. Netherlands	(NET)	0.3556	0.4085	0.2991	0.0970	0.3271	0.6719	0.2216

		Cereals	Starches	Sugar	Seeds	Meat	Milk	Fats
31. New Zealand	(NWZ)	0.3606	0.3555	0.2749	0.0777	0.4295	0.6690	0.1721
32. Philippines	(PHI)	0.7570	0.3823	0.2881	0.1680	0.2672	0.2910	0.1152
33. Poland	(POL)	0.4942	0.4704	0.2500	0.0798	0.4966	0.5933	0.1597
34. Portugal	(POR)	0.5773	0.5195	0.2479	0.2008	0.2772	0.4169	0.2102
35. Saudi Arabia	(SAU)	0.8252	0.0995	0.2439	0.1676	0.8299	0.3941	0.1149
36. Singapore	(SIN)	0.6288	0.3669	0.3535	0.1961	0.3002	0.4385	0.1550
37. South Africa	(SAF)	0.6796	0.2141	0.3293	0.1208	0.3432	0.4811	0.1407
38. Spain	(SPA)	0.4774	0.5135	0.2490	0.1671	0.3237	0.5211	0.4047
39. Thailand	(THA)	0.8309	0.3097	0.2154	0.2725	0.2393	0.1669	0.0879
40. Turkey	(TUR)	0.7063	0.3448	0.2077	0.1919	0.2026	0.4801	0.1747
41. Soviet Union	(USS)	0.5261	0.4944	0.2618	0.1108	0.2618	0.5548	0.1461
42. United States	(USA)	0.3465	0.2887	0.3110	0.1238	0.4488	0.6635	0.2066
43. Venezuela	(VEN)	0.4934	0.5667	0.3187	0.1831	0.2786	0.4400	0.1695
44. Yemen	(YPR)	0.6490	0.1628	0.3662	0.1424	0.2678	0.5328	0.2052
45. Yugoslavia	(YUG)	0.6493	0.3882	0.2364	0.1483	0.2806	0.4877	0.1817

Table A.1.3

Production of the 29 largest forest product firms in Canada for 1977. (The proportional production of the "ideal company" is also given at the bottom of the table).

Table A.1.3

Company		Products (in Million dollars) 1	2	3	4	5	6	7	8	Total
1. MacMillan Bloedell	(MACB)	420	200	8	34	15	0	221	160	1058
2. Abitibi Paper Co.	(ABIT)	342	39	0	32	104	0	19	0	536
3. Domtar Ltd.	(DOMT)	122	133	81	98	188	0	20	0	642
4. Cons. Bathurst	(CONS)	345	80	33	127	0	0	16	0	601
5. Canadian Int. Paper	(CIPA)	344	147	0	114	0	0	0	0	610
6. B.C. Forest Products	(BCFP)	82	188	0	0	0	0	145	36	451
7. Crown Zellerbach	(CRZE)	78	86	17	12	0	0	85	56	334
8. Canadian Forest Prod	(CFPR)	0	61	0	0	0	0	141	33	235
9. Weldwood of Canada	(WELD)	0	39	0	0	0	0	99	115	253
10. Price Co.	(PRCO)	323	6	22	21	0	1.5	34	0	407.5
11. Reed Paper	(REED)	106	48	27	36	3	0	0	0	220
12. Ontario Paper	(ONPA)	252	57	0	0	0	0	0	18	327
13. Great Lakes Paper	(GLPA)	131	155	0	0	0	0	15	0	301
14. Fraser Companies	(FCOS)	0	66	0	10	0	0	12	0	88
15. Kruger Pulp & Paper	(KPNP)	164	0	0	23	0	0	0	0	187
16. Eddy Paper Co.	(EDDY)	0	53	15	7	90	43	25	0	233
17. Bowater	(BOWM)	184	0	0	0	0	10	2	0	196
18. Boise Cascade	(BCAS)	78	30	0	0	110	0	4	0	222
19. Canadian Cellulose	(CCEL)	0	155	0	0	0	0	54	0	209
20. Rolland Paper	(ROLL)	0	0	0	0	65	0	0	0	65
21. Rayonier Canada	(RAYC)	0	192	0	0	0	0	51	0	243
22. Whonnock Ind.	(WHON)	0	0	0	0	0	0	56	0	56
23. Scott Paper	(SCOT)	0	66	0	0	0	78	2	0	146
24. Doman Ind.	(DOMA)	0	0	0	0	0	0	59	0	59
25. West Fraser Timber	(WEST)	0	0	0	0	0	0	91	0	91
26. Crestbrook	(CRES)	0	44	0	0	0	0	30	6	80
27. Northwood Pulp & T.	(NPTI)	0	89	0	0	0	0	133	0	222
28. Kimberley Clark Can.	(KIMB)	0	133	0	0	0	52	18	0	203
29. Weyerhauser	(WEYE)	0	133	0	0	0	0	74	0	207
Ideal Company (Proportion)	(IDEA)	.27	.25	.03	.07	.05	.01	.28	.05	1

1. Newsprint (NEWS)
2. Market Pulp (PULP)
3. Wrapping Paper (WRAP)
4. Paper Board (PBOA)
5. Fine Paper (FINE)
6. Sanitary and Tissue Paper (TISS)
7. Lumber (LUMB)
8. Plywood (PLYW)

Table A.1.4

Occupational prestige.

Table A.1.4a gives the unit vectors for the criteria "Social Usefulness" and "Rewards". Table A.1.4b gives the unit vectors for the "Social Standing Criterion" and the "Earnings Criterion".

Table A.1.4a

Subject	Social Usefulness I	II	III	Rewards I	II	III
A01	-0.039	0.995	0.094	0.787	0.595	-0.164
A02	0.397	0.562	-0.726	0.631	0.734	-0.251
A03	0.623	0.322	-0.713	0.133	0.950	0.281
A04	0.780	0.621	0.081	0.441	0.872	0.212
A05	0.048	0.971	-0.234	0.690	0.588	-0.423
A06	0.454	0.854	-0.252	0.995	0.051	0.087
A07	-0.226	0.916	0.333	0.264	0.961	0.088
A08	-0.052	0.846	-0.531	0.753	0.487	-0.442
B09	0.861	0.505	0.066	0.862	0.487	-0.140
B10	-0.335	0.786	-0.520	0.241	0.920	-0.310
B11	-0.074	0.997	0.034	0.668	0.710	-0.221
C12	0.804	0.579	0.136	0.958	-0.282	0.049
D13	0.582	0.768	-0.269	0.998	0.055	0.011
D14	0.330	0.944	0.018	0.868	0.382	0.317
D15	0.051	0.999	0.004	0.921	0.376	0.039
D16	0.398	0.678	-0.619	0.943	0.330	-0.044
D17	0.913	0.407	-0.038	0.979	0.200	-0.045
D18	0.837	0.228	-0.497	0.828	0.373	0.419
E19	-0.068	0.816	-0.574	0.953	-0.302	-0.012
E20	0.221	0.883	-0.415	0.142	0.925	-0.353
E21	0.284	0.955	0.087	0.826	0.514	0.232
E22	0.736	0.645	-0.208	0.857	0.441	-0.266
E23	0.952	0.303	0.028	0.993	0.103	0.061
E24	0.370	0.574	-0.730	0.994	0.113	0.010

Table A.1.4a

(Continuation)

Subject	I	II	III	I	II	III
K25	0.702	0.672	-0.235	0.915	0.325	0.237
K26	0.732	0.670	0.123	0.836	0.511	-0.200
L27	0.357	0.883	0.305	0.051	0.949	0.311
L28	0.228	0.972	0.065	0.772	0.614	-0.166
L29	0.757	0.644	0.113	0.960	0.143	0.239
L30	-0.145	0.950	0.276	-0.577	-0.758	0.303
L31	0.962	0.271	-0.026	0.990	0.121	-0.073
L32	0.407	0.913	-0.021	0.993	0.102	-0.052
L33	-0.210	0.977	-0.036	-0.027	0.956	0.293
L34	0.397	0.908	0.135	0.861	0.503	0.072
M35	0.033	0.979	0.204	0.854	0.489	-0.175
M36	0.626	0.744	-0.233	0.963	0.253	-0.094
P37	0.617	0.780	-0.109	0.919	0.390	0.056
P38	0.558	0.741	-0.373	0.896	-0.438	0.070
P39	-0.248	0.901	0.355	0.897	-0.438	0.065
Q40	-0.279	0.896	0.345	-0.171	0.917	0.361
Q41	0.711	0.622	-0.329	-0.006	0.901	0.433
Q42	0.778	0.547	0.311	0.955	0.229	0.191
R43	0.718	0.688	-0.102	0.674	0.673	-0.306
R44	0.736	0.466	-0.491	0.408	0.856	-0.318
R45	0.906	-0.300	0.299	0.980	-0.199	0.014
R46	0.884	0.467	-0.021	0.817	0.557	0.151
R47	-0.284	0.897	0.338	-0.140	0.961	0.240
S48	0.469	0.883	0.010	0.746	0.665	0.046

Table A.1.4b

Subject	Social Standing I	II	III	Earnings I	II	III
A01	0.669	0.681	-0.299	0.957	-0.258	-0.129
A02	0.993	0.073	0.092	0.656	-0.501	-0.565
A03	0.907	0.0	-0.421	0.800	-0.598	-0.039
A04	0.104	0.961	0.257	0.864	-0.321	-0.386
A05	0.826	0.564	0.016	0.833	-0.553	-0.018
A06	0.938	-0.342	-0.048	0.844	-0.535	0.036
A07	0.010	0.986	0.167	-0.104	0.055	-0.993
A08	0.934	-0.354	0.044	0.896	-0.421	-0.139
B09	0.945	-0.274	0.180	0.895	-0.446	0.013
B10	0.979	0.173	-0.106	0.903	-0.427	-0.054
B11	0.975	0.181	-0.132	0.842	-0.518	-0.147
C12	0.877	0.479	-0.027	0.926	-0.329	-0.187
D13	0.972	-0.232	0.041	0.767	-0.567	-0.301
D14	0.984	0.104	0.142	0.853	-0.508	-0.123
D15	0.907	0.421	-0.019	0.749	-0.661	0.058
D16	0.968	0.221	0.121	0.924	-0.370	0.100
D17	0.999	0.034	-0.003	0.891	-0.452	0.034
D18	0.941	-0.107	-0.320	0.914	-0.406	-0.025
E19	0.993	-0.049	-0.108	0.913	-0.373	0.167
E20	0.922	0.357	-0.149	0.842	-0.529	0.108
E21	0.664	0.747	0.022	0.873	-0.474	0.114
E22	0.892	0.450	-0.049	0.875	-0.483	0.024
E23	0.939	0.332	-0.087	0.814	-0.572	0.094
E24	0.988	-0.116	0.106	0.831	-0.374	-0.412

Table A.1.4b

(Continuation)

Subject	I	II	III	I	II	III
K25	0.990	0.143	-0.007	0.769	-0.638	0.040
K26	0.866	0.482	0.137	0.459	-0.882	0.109
L27	0.973	0.143	0.182	0.621	-0.759	0.197
L28	0.791	0.595	-0.142	0.712	0.346	-0.611
L29	0.903	0.404	0.144	0.928	-0.357	0.105
L30	0.781	0.617	-0.097	0.936	0.153	-0.318
L31	0.962	0.250	0.110	0.995	0.046	0.084
L32	0.991	0.109	-0.077	0.937	-0.322	0.137
L33	0.997	-0.056	-0.061	0.895	-0.430	0.117
L34	0.893	0.449	-0.009	0.888	-0.459	-0.013
M35	0.857	0.514	0.039	0.809	-0.481	-0.339
M36	0.948	0.263	0.181	0.969	-0.219	-0.114
P37	0.906	0.418	0.064	0.773	-0.594	0.223
P38	0.896	-0.438	0.070	0.875	-0.480	0.067
P39	0.897	-0.438	0.065	0.888	-0.456	0.066
Q40	-0.115	0.928	0.354	0.874	-0.478	0.085
Q41	0.905	-0.122	-0.407	0.819	-0.553	0.153
Q42	0.890	0.409	0.200	0.983	-0.173	0.061
R43	0.601	0.760	-0.247	0.964	-0.255	-0.075
R44	0.584	0.792	-0.177	0.983	0.175	0.045
R45	0.994	0.035	0.106	0.995	-0.081	-0.065
R46	0.804	0.539	-0.251	0.836	-0.540	0.101
R47	-0.449	0.645	0.618	0.825	-0.551	0.127
S48	0.847	0.433	0.309	0.737	0.535	-0.414

Appendix 2

Normal approximations to the F-distribution.

A.2.1. Introduction.

When the one way and two way ANOVA for resultants and angles are calculated the test statistics obtained require the percentage points of the F-distribution with degrees of freedom larger than those usually tabulated.

In the examples shown here it can be seen that n_2, the number of degrees of freedom in the denominator, is very large, while n_1 , the number of degrees of freedom in the numerator, is small. This indicates that the percentage points of the F-distribution can be approximated by using a χ^2-distribution with n_1 degrees of freedom:

$$F(n_1,n_2;\ \alpha) \approx \chi^2(n_1;\ \alpha)/n_1\ .$$

However, in the general case, when p , the number of dimensions, is very large, the degrees of freedom in the numerator will also be large and the χ^2-approximation can not be used. It was therefore decided to find an approximation to the F-significance points for this situation. We report on eight methods that have appeared, over a span of many years, in the literature. All the methods considered take a value F and convert it to a new value z; if F has the $F(n_1,\ n_2)$ distribution, z will have, to a good approximation, a standard normal distribution. Of particular interest is to see if the upper tail points of $F(n_1,\ n_2)$ taken from the tables, convert accurately to the standard normal points. Results were compared only for values n_1 and n_2 greater than or equal to 20. Other comparisons have been given by Peizer and Pratt (1968) and Ling (1978).

A.2.2. Transformations from F to Standard Normal Values.

The eight transformations examined are listed below.

a. Abramowitz and Stegun (1967) list three approximations. Two of these are Z_2 and Z_3 below, but they also give the following approximation recommended for use with large values of n_1 and n_2:

$$Z_1 = \frac{F - \frac{n_2}{n_2-2}}{\frac{n_2}{n_2-2}\sqrt{\frac{2(n_1+n_2-2)}{n_1(n_2-4)}}} .$$

b. The square root approximation (Z_2) is a modification of an approximation suggested by Pinkham (1957) after investigation of the first four moments. It is given by

$$Z_2 = \frac{\sqrt{(2n_2-1)\frac{n_1}{n_2}F} - \sqrt{2n_1-1}}{\sqrt{1 + \frac{n_1}{n_2}F}}$$

c. The cube root approximation given by Paulson (1942) is

$$Z_3 = \frac{F^{1/3}(1 - \frac{2}{9n_2}) - (1 - \frac{2}{9n_1})}{\sqrt{\frac{2}{9n_1} + \frac{2}{9n_2} F^{2/3}}}, \text{ for } n_2 \geq 3.$$

If it is desired to use the lower tail of the F distribution n_1 should also be greater than or equal to 3 (Paulson, 1942).

d. Mudholkar and Yogendra (1976) use the cumulants of (-log X) where X has a Beta distribution with parameters (p, q): these cumulants are approximated by a chi-square variable, as was done by Patnaik (1949) or Pearson (1959). This is in turn subjected to the Wilson-Hilferty cube root transformation to obtain a normal approximation for (-log X). A direct three moment normal approximation for (-log X) can also be constructed (Sankaran, 1959). If X has the Beta distribution with parameters (p, q), the r-th cumulant of (-log X) is

$$K_r = (r-1)! \sum_{j=0}^{q-1} (p+j)^{-r}, \quad r = 1,2,\ldots \quad .$$

If F has F-distribution with (n_1, n_2) degrees of freedom then $X = (1 + n_1F/n_2)^{-1}$ has the Beta distribution with parameters (p, q), where $p = n_2/2$ and $q = n_1/2$. If n_1 is odd the cumulants are approximated by

$$K_r = (r-1)! \sum_{j=0}^{q-3/2} (p+j)^{-r} + (1/2)(p+q-1/2)^{-r}, \ r = 1,2,\ldots \ .$$

The approximation of Pearson type has the form

$$Z_4 = [\{\frac{-(\log X+b)}{av}\}^{1/3} - (1 - \frac{2}{9v})](\frac{2}{9v})^{-1/2} ,$$

where $B = K_3/K_2^{3/2}$, $v = 8/B^2$

$$a = (K_2/2v)^{1/2} , b = K_1 - va .$$

The Sankaran approximation involves the determination of a constant h such that the leading term in the third cumulant of $\{(-\log X)/K_1\}^h$ vanishes. It is given by

$$Z_5 = \frac{\{(-\log X)/K_1\}^h - \mu}{\sigma} .$$

where $h = 1 - K_1K_3/3K_2^2$

$$\mu = 1 - K_3h/6K_1K_2$$

$$\sigma = h^2K_2/K_1^2 .$$

The approximation obtained by following Patnaik's method becomes

$$Z_6 = \{(\frac{-\log X}{av})^{1/3} - (1 - \frac{2}{9v})\}(\frac{2}{9v})^{-1/2} ,$$

where $a = K_2/2K_1$

$$v = 2K_1^2/K_2 .$$

e. Peizer and Pratt (1968) give a normal approximation to the Beta distribution and its relatives, in particular the binomial, Pascal, negative binomial, F, t, Poisson, Gamma and χ^2 distributions. The approximate normal deviate for the F distribution is

$$Z_7' = d'\left\{\frac{1+qg(\frac{S}{np}) + pg(\frac{T}{nq})}{(n+1/6)pq}\right\}^{1/2}$$

where $S = (n_2 - 1)/2, \quad T = (n_1 - 1)/2$;

$p = n_2/(n_1 F + n_2), \quad q = 1 - p$

$n = (n_1 + n_2 - 2)/2$

$d' = S + 1/6 - (n + 1/3)p$

and $g(x) = (1 - x)^{-2}(1 - x^2 + 2x \log x)$.

Transformation Z_7, the one used in the comparisons, is a refinement of Z_7' using $d = d' + .02\{q/(S+.5)-p/(T+.5)+(q-.5)/(n+1)\}$ instead of d'. The function $g(x)$ is tabulated in Peizer and Pratt's paper so that the approximation can be calculated quickly.

f. Carter (1947) gives a normal approximation based on approximations to the third and fourth cumulants. He recommends its use for large values of n_1 and n_2 and beyond the range of the published tables. Carter actually gives F from z, but inverting we calculate z from F as follows. First calculate

$s = 1/(n_1-1) + 1/(n_2-1)$ and $t = 1/(n_1-1) - 1/(n_2-1)$; then

$a = t^2/36 - s^2/24$;

$b = Qt/3 + t^2(1-s)/9 - s/2+s^2/8$;

and $c = 2Qt(1-s)/3 + t^2(1-s)^2/9 + Q^2$, with $Q = \log F$.

Finally the normal variable is

$$Z_8 = (-1)^p\{(-b-\sqrt{b^2-4ac})/2a\}^{1/2};$$

$p = 0$ if F is in the upper tail, and $p = 1$ if F is in the lower tail.

A.2.3. Comparisons.

The eight approximations where compared for a wide range of degrees of freedom and upper and lower tail probabilities (0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25). The degrees of freedom examined consisted of all possible combinations with n_1, n_2 = 20, 30, 40, 60, 120.

Because of the well-known identity $F(n_1, n_2; \alpha) = 1/F(n_2, n_1; 1-\alpha)$ where $F(n_1, n_2; \alpha)$ refers to the upper tail α-level percentage point of F with n_1 (numerator) and n_2 (denominator) degrees of freedom, it is not strictly necessary to examine the lower tail separately; this was however done as a check on calculations. Our reported results below refer to the upper tail points. The technique of comparison followed was to insert a known $F(n_1, n_2; \alpha)$ as F in the formulae, and calculate $Z_1,\ldots,Z_8$; the values were then compared to the standard normal values. The exact percentage points of F were taken from Biometrika Tables for Statisticians, Vol. 1, Table 1. A typical set of results is shown in Tables A.2.1 and A.2.2. From these and similar tables (25 in all) the conclusions are as follows.

Table A.2.1.

Normal approximations for $F(60,60;\alpha)$; α = 0.250, 0.100, 0.050, 0.025.

α =	0.250	0.100	0.050	0.025
True z-score	0.67449	1.28155	1.64485	1.95996
z_1	0.57162	1.31510	1.82306	2.30635
z_2	0.67372	1.27711	1.63546	1.94420
z_3	0.67463	1.28179	1.64482	1.95959
z_4	0.67444	1.28153	1.64472	1.95980
z_5	0.67440	1.28126	1.64416	1.95886
z_6	0.67281	1.28240	1.64776	1.96511
z_7	0.67438	1.28156	1.64482	1.95997
z_8	0.67438	1.28152	1.64474	1.95984

Table A.2.2

Normal approximations for $F(60,60;\alpha)$; α = 0.010, 0.0050, 0.0025, 0.0010.

α =	0.0100	0.0050	0.0025	0.0010
True z-score	2.32635	2.5783	2.80703	3.09023
Z_1	2.92459	3.38381	3.83974	4.44194
Z_2	2.30012	2.54026	2.76104	3.02910
Z_3	2.32529	2.57403	2.80433	3.08625
Z_4	2.32614	2.57552	2.80658	3.08973
Z_5	2.32459	2.57344	2.80393	3.08625
Z_6	2.33447	2.58615	2.81950	3.10564
Z_7	2.32639	2.57583	2.80695	3.09016
Z_8	2.32621	2.57562	2.80671	3.08992

a. Seven of the eight statistics examined give very good approximations; the exception is Z_1. The extent to which Z_1 differs from the others in accuracy is illustrated by, for example, the values for $F(60,60;.10)$ shown in Table A.2.1. It can be seen that in all the comparisons the error in Z_1 is noticeably greater than for the other statistics.

b. The comparisons corresponding to α-values 0.25, 0.10, 0.05 and 0.025 are listed apart from those corresponding to α-values 0.01, 0.005, 0.0025 and 0.001. Thus the more used significance levels may be compared separately from the extreme tail levels. For each combination of (n_1, n_2) and α, the statistics were ranked 1 (the best), 2 and 3 by the values $|z_i - z_e|$ where z_e was the expected normal value. Table A.2.3 gives the number of times each statistic was ranked 1, 2 or 3, for α-values in the upper tail. The table is divided into four parts, corresponding to smaller and larger values of n_1 and n_2. An approximation with rank 1, 2 or 3 is given a score of 3, 2 or 1 respectively. S is the sum of these scores.

Table A.2.3

Ranks for normal approximations.

n_2	α		$n_1 = 20,30,40$ 1	2	3	S		$n_1 = 60,120$ 1	2	3	S
$n_2 = 20,30,40$	.25, .1, .05, .025	z_8	26	8	2	96	z_8	31	5	0	103
		z_7	7	8	10	47	z_7	3	17	16	59
		z_6	0	1	2	4	z_6	0	0	4	4
		z_5	1	0	6	9	z_5	0	0	1	1
		z_4	1	9	11	32	z_4	2	14	14	48
		z_3	1	10	5	28	z_3	0	0	1	1
		z_2	0	0	0	0	z_2	0	0	0	0
	.01, .005, .0025, .001	z_8	19	3	1	85	z_8	16	8	0	64
		z_7	2	2	4	14	z_7	7	16	1	54
		z_6	1	3	3	12	z_6	0	0	10	10
		z_5	0	3	7	13	z_5	0	0	0	0
		z_4	0	1	1	3	z_4	0	0	4	4
		z_3	2	12	18	48	z_3	0	0	9	9
		z_2	0	0	0	0	z_2	1	0	0	3

Table A.2.3.

(Continuation)

n_2	α		n_1 = 20,30,60					n_1 = 60,120			
			1	2	3	S		1	2	3	S
n_2 = 60,120	.25, .1, .05, .025	Z_8	17	5	1	62	Z_8	22	2	0	70
		Z_7	1	7	2	19	Z_7	2	16	6	44
		Z_6	0	0	1	1	Z_6	0	0	0	0
		Z_5	1	0	0	3	Z_5	0	0	11	11
		Z_4	4	3	11	29	Z_4	0	6	7	19
		Z_3	1	9	9	30	Z_3	0	0	0	0
		Z_2	0	0	0	0	Z_2	0	0	0	0
	.01, .005, .0025, .001	Z_8	6	4	3	29	Z_8	14	1	1	45
		Z_7	1	6	3	18	Z_7	1	16	0	35
		Z_6	0	0	0	0	Z_6	0	0	1	1
		Z_5	1	4	1	12	Z_5	0	0	2	2
		Z_4	4	2	3	19	Z_4	1	0	9	12
		Z_3	4	0	6	18	Z_3	0	0	3	3
		Z_2	0	0	0	0	Z_2	0	0	0	0

It can be seen that the best approximations are Z_8 and Z_7; those two statistics appear frequently in Table A.2.3. Statistic Z_3 and Z_4 also appear very often. At the other extreme, Z_2 appeared only once, and statistics Z_5 and Z_6 appeared only rarely.

c. Overall, it is clear that Z_7 and Z_8 are to be preferred, for the relatively high values of n_1 and n_2 which we consider; they are convenient if good computing facilities are available since the formulae are complicated.

d. A very useful result is the accuracy of the older approximation Z_3 , since this is easily computable on a hand - or desk - calculator; although it is not often the best approximation, it often places in the best three all along the tail, and will be accurate enough for most practical purposes.

e. Most of the approximations were devised to give points in the upper tail of F . If points are needed in the lower tail, the question arises whether the approximations should be used as given, or whether the well known identity $F(n_1, n_2, \alpha) = 1/F(n_2, n_1; 1-\alpha)$ should be used. The implication of this identity is that if Z_1^* is the value obtained by any one of the approximations above, corresponding to $F(n_1, n_2; \alpha)$, and if Z_2^* is a value obtained corresponding to $G = 1/F(n_2, n_1; 1-\alpha)$, then Z_1^* should equal $-Z_2^*$. Thus for a small value of F , one could either calculate Z_1^* directly, using the approximation, or calculate Z_2^* and take its negative. Statistics Z_2, Z_3, Z_7 and Z_8 have the property that either method will give the same result, and this adds to the appeal of approximations Z_3, Z_7 and Z_8. We have investigated for the whole range of n_1 and n_2 considered here, these two methods for the other approximations. For approximation Z_4, it appears to be almost always better to use

G and $-Z_2^*$, rather than to calculate Z_1^* directly. For approximation Z_6, the results are somewhat inconclusive. It appears to be better to use G whenever $n_1 \leq n_2$ but to use the lower tail directly, if $n_1 > n_2$.

f. The overall pattern of this examination suggests that when computing facilities are available, Z_8 or Z_7 are to be preferred; but when a hand calculator or desk calculator is used, Z_3 gives very good results indeed.

BIBLIOGRAPHY

[1] Abramowitz, Milton and Stegun, Irene, ed., (1964). Handbook of Mathematical Functions with Formulas, Graphs and Mathematical tables. National Bureau of Standards Applied Mathematics Series No. 55, Government Printing Office, Washington, D.C., p. 947.

[2] Carter, A.H., (1947). Approximation to percentage points of the Z-distribution. Biometrika, 34, p. 352.

[3] Coxon, A.P. and Jones, C.L., (1978). The images of occupational prestige. London: Academic Press.

[4] Coxon, A.P. and Jones, C., (1979). Measurements and meanings. London: Academic Press.

[5] Ling, R.F., (1978). A study of the accuracy of some approximations for t, χ^2 and F tail probabilities. Journal of the American Statistical Association, 73, p. 274.

[6] Mudholkar, Govind and Chaubey, (1976). Some refinements of the Wise approximation for the beta and F distributions. Utilitas Mathematica, 10, p. 199.

[7] Patnaik, P.B., (1949). The Noncentral Chi-square and F Distributions and their Applications. Biometrika, 36, p. 202.

[8] Paulson, Edward, (1942). An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 13, p. 233.

[9] Pearson, E.S., (1959). Note on the Approximation to the Distribution of Noncentral Chi-square. Biometrika, 46, p. 364.

[10] Peizer, David and Pratt, John, (1968). A normal approximation for binomial, F, Beta and other common related tail probabilities I. Journal of the American Statistical Association, 63, p. 1416.

[11] Pinkham, Roger, (1957). The first four moments of a transformed Beta variable. Memorandum Report 61, Statistical Research Group, Princeton University.

[12] Prentice, M.J., (1978). On invariant tests of uniformity for directions and orientation. Annals of Statistics, 6, p. 169.

[13] Sankaran, M., (1959). On the Noncentral Chi-square Distribution. Biometrika, 45, p. 235.

[14] Stephens, M.A., (1962). The statistics of directions; the von Mises and Fisher distribution. Ph.D. Thesis, University of Toronto.

[15] Stephens, M.A., (1969). Multisample tests for the Fisher distribution for directions. Biometrika, 56, p. 169.

[16] Stephens, M.A., (1967). Tests for the dispersion and for the modal vector of a distribution on a sphere. Biometrika, 54, p. 221.

[17] Watson, G.S., (1956). Analysis of dispersion on a sphere. Monthly Notices of the Royal Astronomical Society. Geophysical Supplement 7, p. 153.