Advanced Statistical Methods in Biometric Research

C. RADHAKRISHNA RAO
Professor of Statistics
Indian Statistical Institute
Calcutta

Reprinted with corrections

ANGRAU Central Library
Hyderabad
23037
31.5.69

ANGRAU
519.5
N52RAD
Acc No 23037

AFNER PRESS
A DIVISION OF MACMILLAN PUBLISHING CO., INC.
New York
COLLIER-MACMILLAN PUBLISHERS
London
In memory of
my father
the late C. D. Naik
The statistician is no longer an alchemist expected to produce gold from any worthless material offered him. He is more like a chemist capable of assaying exactly how much of value it contains, and capable also of extracting this amount, and no more. In these circumstances, it would be foolish to commend a statistician because his results are precise or to reprove because they are not. If he is competent in his craft, the value of the result follows solely from the value of the material given him. It contains so much information and no more. His job is only to produce what it contains.

R. A. FISHER
Preface

The ever-increasing need for more searching and finer analyses of statistical data in various domains of human activity has constantly been giving rise to new concepts and improved methods. The vast amount of research carried out during the past few decades in the field of theoretical and applied statistics has been responsible for the discarding or recasting of some of the older methods in statistics and for the creation of a wealth of new statistical tools for the research worker and the routine analyst. The popularity and range of application of any statistical method, however, has always remained in great measure dependent on the logical elucidation it received and the simplicity of procedure it was capable of, mainly because more often a method that is powerful is also complex in theory and procedure.

The object in writing this book is to present a number of statistical techniques, keeping in view the requirements of both the student who questions the basis of a particular method employed and the practical worker who seeks a recipe for the reduction of his data. I have therefore endeavored first to provide a theoretical groundwork for the different methods to satisfy the former and second to illustrate computational procedures by working out a number of problems in full to meet the demands of the latter.

Throughout this book, efforts have been made to integrate a large collection of computational schemes into consistent patterns. Thus the problems of regression and analysis of variance and covariance reduce to fitting of constants by the method of least squares and evaluation of the least sum of squares. The problems of multivariate analysis resolve themselves into an analysis of the dispersion matrix and reduction of determinants. The reduction of a matrix by the method of pivotal condensation emerges as the most useful technique which at the same time does not present any computational difficulty. The check column, properly carried, ensures numerical accuracy. The different computational schemes have been illustrated in detail with the help of original data. These data have also been reproduced, either in full or, in the case of extensive data, in the form of necessary statistics such as totals, sums of squares, products, etc., giving reference to the sources from which they are taken.
The material presented for illustrative purposes has been restricted mainly to anthropometric studies, but without prejudicing the statistical methods used in their applicability to problems in various other branches of knowledge such as general biology, psychology, economics, and other social sciences. The problem of neurotic groups considered in section 9d.1 provides an example from psychology. For the psychological findings appearing in that section I am indebted to Mr. Patrick Slater.

In the development of the text, the first chapter is devoted entirely to mathematical procedures in modern algebra. In the second chapter it is shown that most of the distribution problems connected with univariate and multivariate normal populations can be solved by a fundamental theorem on least squares, whose proof needs only a knowledge of linear transformations.

The third chapter first deals with applications of the least square technique in the estimation of parameters. It next traces as the starting point of exact sampling theory the fundamental discovery by W. S. Gosset, now known as "studentization," that the probability of the error in the observed average expressed as a multiple of the sample standard deviation admits a precise evaluation independently of the unknown standard deviation of the population. Then it proceeds with tests of linear hypotheses. The tests discussed in this chapter, exact for small samples, are due to R. A. Fisher, who, having developed the theory of these tests, also put forward the elegant computational scheme of the analysis of variance table.

The fourth chapter contains some observations on the general theories of estimation and certain applications of the method of maximum likelihood. The scoring system of R. A. Fisher discussed and elaborated in this chapter introduces great simplicity and mechanization in the use of the maximum likelihood principle, thus providing a complete answer to critics who hold that the method of maximum likelihood leads to intractable equations. Certain alternative methods of deriving asymptotically best estimates advocated by some authors neither have general applicability nor would admit mechanized computation.

Problems of specification and associated tests of homogeneity form the subject matter of the fifth chapter. The choice of a mathematical model from which the observations could be deemed to have arisen is of fundamental importance because subsequent statistical computations have to be built on the framework of the chosen model. The choice is necessarily empirical, "but this empiricism could be cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts." Karl Pearson was the first to visualize this approach. The $\chi^2$ goodness
of fit introduced by him has been found extremely useful in many other directions too.

The next chapter gives tests of homogeneity of variances and covariances, and these form the preliminary investigation in multivariate analysis which is detailed in the seventh chapter. The eighth and ninth chapters relate to the utilization of multiple measurements in problems of biological classification. Biologists are usually confronted with the problem of assigning an individual to one of several groups to which he might belong. An objective method which minimizes the errors of classification is provided in the eighth chapter, utilizing the modern theories of inference as developed by J. Neyman and A. Wald. Chapter 9 is devoted to a study of the interrelationships between a number of populations or groups of individuals. The methods of this chapter are based on the researches carried out in the Indian Statistical Institute under the inspiring guidance of P. C. Mahalanobis, who introduced the concept of group distance.

C. R. Rao

Calcutta, India
June 1952
Contents

CHAPTER 1· ALGEBRA OF VECTORS AND MATRICES
1a VECTOR SPACES ................................................. 1
1a.1 Vectors. 1a.2 Linear Independence and Orthogonality. 1a.3
Vector Spaces and the Sweep-Out Method. 1a.4 The Orthogonal
Vector Space and the Deficiency Matrix. 1a.5 Linear Equations.

1b THEORY OF MATRICES AND DETERMINANTS ............... 6
1b.1 Matrices. 1b.2 Partitioned Matrices. 1b.3 Determinants.

1c QUADRATIC FORMS ............................................. 18
1c.1 Definitions. 1c.2 Linear Transformations. 1c.3 Classifica-
tion of Quadratic Forms. 1c.4 The Latent Roots of a Matrix and the
Characteristic Vectors. 1c.5 Pairs of Quadratic Forms. 1c.6 Red-
duction of an Asymmetric Matrix.

1d NUMERICAL APPENDIX ......................................... 29
1d.1 The Evaluation of Determinants, Reciprocals, and Solutions of
Equations.

CHAPTER 2· THEORY OF DISTRIBUTIONS
2a SOME ANALYTICAL METHODS IN DISTRIBUTION PROBLEMS 32
2a.1 Binomial Distribution. 2a.2 Multinomial Distribution. 2a.3
The Poisson Distribution. 2a.4 Normal Distribution. 2a.5 Gamma
Distribution. 2a.6 Beta Distribution. 2a.7 Cauchy Distribution.
2a.8 Pearson’s $F_0$ Distribution. 2a.9 Summary of Results.

2b DISTRIBUTIONS RELATING TO THE UNIVARIATE NORMAL DISTRIBUTION 46
2b.1 Mean and Variance in Normal Samples. 2b.2 Student’s Distribu-
tion. 2b.3 Fisher’s $z$ Distribution. 2b.4 Cochran’s Theorem.
2b.5 Distribution of Non-Central $\chi^2$.

2c MULTIVARIATE NORMAL POPULATIONS .......................... 51
2c.1 The Multivariate Normal Distribution. 2c.2 The Distribution
of a Set of Linear Functions of Normal Variates. 2c.3 The Distribu-
tion of Quadratic Forms.

2d LEAST SQUARES FUNDAMENTAL IN DISTRIBUTION THEORY 58
2d.1 Two Theorems on Least Squares. 2d.2 Multivariate Distri-
butions.
CONTENTS

CHAPTER 3· THE THEORY OF LINEAR ESTIMATION AND TESTS OF HYPOTHESES

3a LIN EAR ESTIMATION
3a.1 Observational Equations. 3a.2 Best Unbiased Estimates. 3a.3 The Necessary and Sufficient Condition for the Existence of an Unbiased Estimate. 3a.4 Normal Equations. 3a.5 Linear Functions with Zero Expectations. 3a.6 Standard Errors of Estimates and Intrinsic Properties of Normal Equations. 3a.7 Principle of Substitution. 3a.8 Observational Equations with Linear Restrictions on Parameters. 3a.9 Observational Equations with Correlated Variables.

3b TESTS OF LINEAR HYPOTHESES
3b.1 Nature of Linear Hypotheses. 3b.2 Test for \( H_0 \) when \( R_0 \) Is Not True.

3c THE COMBINATION OF WEIGHTED OBSERVATIONS
3c.1 Transformation to Unweighted Observations. 3c.2 An Example of Weighted Observations.

3d TESTS OF HYPOTHESES WITH A SINGLE DEGREE OF FREEDOM
3d.1 Student’s t Test. 3d.2 Asymmetry of Right and Left Femora.

3e ANALYSIS OF VARIANCE
3e.1 One-Way Classification. 3e.2 Two-Way Classification with a Single Observation in a Cell. 3e.3 Two-Way Classification with Multiple but Equal Numbers in Cells. 3e.4 Two-Way Classification with Unequal Numbers in Cells.

3f THE THEORY OF STATISTICAL REGRESSION
3f.1 The Concept of Regression. 3f.2 Prediction of Cranial Capacity. 3f.3 Test for the Equality of Regression Equations. 3f.4 The Test for an Assigned Regression Function.

3g THE GENERAL PROBLEM OF LEAST SQUARES WITH TWO SETS OF PARAMETERS
3g.1 Concomitant Variables. 3g.2 Adjustment for Concomitant Variation. 3g.3 An Illustrative Example. 3g.4 A Problem of Inheritance in Man.

CHAPTER 4· THE GENERAL THEORY OF ESTIMATION AND THE METHOD OF MAXIMUM LIKELIHOOD

4a BEST UNBIASED ESTIMATES
4a.1 Estimation by Minimizing the Variance. 4a.2 The Information Limit to Variance: A Single Parameter. 4a.3 Distributions Admitting Estimates with the Information Limit to Variance. 4a.4 Sufficient Statistics and Unbiased Estimates. 4a.5 Distributions Admitting Sufficient Statistics. 4a.6 An Optimum Property of Sufficient Statistics. 4a.7 More Stringent Inequalities for the Variance of an Estimate.
CONTENTS

4a.8 The Case of Several Parameters. 4a.9 Properties of Distributions Admitting Sufficient Statistics: Several Parameters.

4b Estimation by the Method of Maximum Likelihood

4b.1 The Principle of Maximum Likelihood. 4b.2 Consistency and Bias. 4b.3 The Concept of Efficiency. 4b.4 Some Optimum Properties of Maximum Likelihood Estimates.

4c Some Examples of Maximum Likelihood Estimates

4c.1 Improved Estimates of Means from Incomplete Data on Several Variables. 4c.2 The Method of Scoring for the Estimation of Parameters. 4c.3 Combination of Data.

Appendix: Some Limiting Theorems

Chapter 5: Large Sample Tests of Hypotheses with Applications to Problems of Estimation

5a The General Theory of Tests in Large Samples

5a.1 The Nature of Statistical Hypotheses. 5a.2 The Problem of Distribution.

5b Applications of the General Theory

5b.1 The \( \chi^2 \) Test of Departure from a Simple Hypothesis. 5b.2 The \( \chi^2 \) Test of Goodness of Fit. 5b.3 Tests of Homogeneity of Parallel Samples.

5c Contingency Tables

5c.1 The Probability of an Observed Configuration and Tests in Large Samples. 5c.2 Tests of Independence in a Contingency Table. 5c.3 Tests of Independence in Small Samples.

5d Tests in Poisson Populations

5e Transformation of Statistics

5e.1 A General Lemma. 5e.2 The Square Root Transformation of the Poisson Variate. 5e.3 The Sin \(^{-1}\) Transformation of the Binomial Proportion. 5e.4 Other Useful Transformations.

5f Large Sample Standard Errors of Moments

5f.1 Variances and Covariances of Raw Moment Statistics. 5f.2 Large Sample Tests of Difference between Means and an Illustration of the \( F \) Test. 5f.3 Tests of Normality.

Chapter 6: Tests of Homogeneity of Variances and Correlations

6a Homogeneity of Variances

6a.1 Test for a Specified Variance. 6a.2 Test for a Specified Inequality of Two Estimated Variances. 6a.3 The Likelihood Criterion and Its
CONTENTS

6b Homogeneity of Correlations
   6b.1 Exact Test for Zero Correlation. 6b.2 Fisher's Transformation. 6b.3 Test for a Given ρ. 6b.4 Test for the Inequality of Two Correlation Coefficients. 6b.5 Test for the Homogeneity of a Set of Correlation Coefficients. 6b.6 Correction for Bias in the Test for Homogeneity and the Best Estimate of ρ.

CHAPTER 7: TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS

7a Review of Work on Multivariate Analysis

7b Tests with Discriminant Functions
   7b.1 Two Fundamental Distributions. 7b.2 Problems of a Single Sample. 7b.3 Mahalanobis' D² and Problems of Two Samples. 7b.4 Test for an Assigned Discriminant Function. 7b.5 Tests for Discriminant Function Coefficients. 7b.6 The Additional Information Supplied by Some Characters.

7c Generalization of D² and the Large Sample Theory for Several Groups

7d Tests with Wilks' λ Criterion
   7d.1 Analysis of Dispersion and the Theoretical Aspects of the λ Criterion. 7d.2 The Distribution of λ and Its Practical Use. 7d.3 Test of Differences in Mean Values for Several Populations. 7d.4 Internal Analysis of a Set of Variates. 7d.5 Barnard's Problem of Secular Variations in Skull Characters.

CHAPTER 8: STATISTICAL INFERENCE APPLIED TO CLASSIFICATORY PROBLEMS

8a Tests of Null Hypotheses
   8a.1 Problems in Biological Research. 8a.2 Null Hypotheses. 8a.3 Power Function of Neyman and Pearson. 8a.4 Locally Most Powerful Unbiased Tests. 8a.5 Test for a Finite Number of Alternatives. 8a.6 Tests When the Alternatives Are Continuous.

8b Problems of Discrimination
   8b.1 The General Problem. 8b.2 The Discriminant Function of R. A. Fisher. 8b.3 Some Difficulties in the Use of the Best Discriminating Solution. 8b.4 Uncertainty of the A Priori Information That One of the Alternatives Is Correct. 8b.5 The Doubtful Region. 8b.6 Resolution of a Mixed Series Into Two Gaussian Components. 8b.7 Sexing of Osteometric Material. 8b.8 The Problem of Three and More Groups. 8b.9 Application to Multivariate Normal Populations. 8b.10 Allocation of a Number of Individuals to Two or More Groups.
CONTENTS

8c Discriminant Function for Selecting Genetically Desirable Types 329
   8c.1 Prediction Formula for the Genotypic Value. 8c.2 The Genetic Advance.

8d Problems of Optimum Selection 336
   8d.1 A Single Predictor for Dichotomy. 8d.2 The Problem of Differential Predictors.

Appendix A 339
   A4 A Lemma on Power Functions. A5 Two Lemmas Useful in Classification Problems.

Appendix B 345

CHAPTER 9 - THE CONCEPT OF DISTANCE AND THE PROBLEM OF GROUP CONSTELLATIONS

9a Distance between Two Populations 351
   9a.1 The Need for a Distance Function. 9a.2 Mathematical Concepts (Discriminatory Topology). 9a.3 Mahalanobis' Generalized Distance. 9a.4 Karl Pearson's Coefficient of Racial Likeness.

9b An Illustrative Example 357
   9b.1 Calculation of $D^2$. 9b.2 The Determination of Group Constellations.

9c The Use of Canonical Variates in Deriving Group Constellations 364
   9c.1 Graphical Methods of Representing the Groups. 9c.2 The Problem of Maximal Average $D^2$. 9c.3 An Illustrative Example.

9d A Test for Reduction in the Number of Dimensions 370
   9d.1 The Analysis of Neurotic Cases.

APPENDIX: MISCELLANEOUS PROBLEMS 379

INDEX 383
CHAPTER 1

Algebra of Vectors and Matrices

1a Vector Spaces

A set of ordered elements \((x_1, \ldots, x_n)\) is called a vector which may be simply denoted by \(x\). The elements may be \(n\) observations obtained in the order \(x_1, x_2, \ldots, x_n\) from a population, in which case \(x\) is called the vector of observations.

Suppose that these observations are standardized by multiplying each of them by a constant \(c\), otherwise known as a scalar. The resulting vector is \((cx_1, \ldots, cx_n)\) which may be represented by \(cx\). This is the rule for multiplying a vector \(x\) by a scalar \(c\).

The weighted sum of observations with the vector \(w\) of weights \((w_1, \ldots, w_n)\) is

\[
\sum w_i x_i = w_1 x_1 + \cdots + w_n x_n
\]

This may be represented by \(w \cdot x\), denoting the product of two vectors \(w\) and \(x\). The simple average of the observations is the result of multiplying the observation vector \(x\) by the weight vector \((1/n, \ldots, 1/n)\).

The sum of squares of observations is the result of multiplying \(x\) by itself, i.e., \(x \cdot x = x_1^2 + \cdots + x_n^2\).

If \(x = (x_1, \ldots, x_p)\) is the vector of \(p\) measurements on an individual and \(y = (y_1, \ldots, y_p)\) on another, then the sum of the measurements is the vector \((x_1 + y_1, \ldots, x_p + y_p)\) which may be represented by \(x + y\). From the definitions given above we find the vector of average measurements

\[
\frac{1}{2} (x + y) = \left( \frac{x_1 + y_1}{2}, \ldots, \frac{x_p + y_p}{2} \right)
\]

In general,

\[
ax + by + cz = (ax_1 + by_1 + cz_1, \ldots, ax_p + by_p + cz_p)
\]

where \(a, b, c\) are scalars and \(x, y, z\) are vectors. The vector \(0 = (0, \ldots, 0)\) is called the null vector. It is easy to verify that any vector
added to the null one remains unchanged and any vector multiplied by the null vector reduces to zero.

1a.2 Linear Independence and Orthogonality

A set of vectors \((x, y, \ldots)\) are said to be independent if none of them can be expressed as a linear combination of the rest. For instance, the vectors \((1, 0, 1, -1, 2)\), \((3, 2, -1, 1, 2)\), and \((9, 4, 1, -1, 10)\) are not independent since the last vector is the sum of 3 times the first and 2 times the second. Two non-null vectors are said to be orthogonal if their product is zero. A set of orthogonal vectors with real elements are necessarily independent. If not, then

Multiplying by \(x\)

\[ x \cdot x = ay + bz + \cdots \]

\[ x \cdot x = x_1^2 + x_2^2 + \cdots = 0 \]

which means that \(x_1 = x_2 = \cdots = 0\) or \(x\) is a null vector.

1a.3 Vector Spaces and the Sweep-Out Method

The totality of vectors obtained by linear combinations of a set of vectors is called a vector space. Such a totality can be generated by a set of independent vectors called a basis of the vector space. If the number of vectors in a basis is \(m\), then the vector space is said to be of rank or dimensions \(m\). In practical problems it is sometimes necessary to obtain the rank of a vector space formed by the vectors

\[
\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{pmatrix}
\]

The arrangement of the elements \(a_{ij}\) in rows and columns as above is known as a matrix. A convenient way of finding the rank is by a method known as "sweep out," consisting in the following operations:

(i) Any vector having a non-zero value for the first element is taken, and all its elements are divided by the first so that the resulting vector is of the form

\((1, c_2, \cdots, c_n)\)

If the elements of the first column are all zero, then it is omitted to start with.
(ii) From every other vector is subtracted a vector obtained by multiplying \((1, \cdots, c_n)\) by the first element of the former vector so that the resulting vectors, except the one chosen in (i), have zero as their first element. The first column is said to be swept out by the vector called the pivotal row chosen above.

(iii) Omission of the pivotal row and the first column results in a reduced matrix on which operations (i) and (ii) are repeated until a single non-zero row or all null rows are left over. A single non-zero row left over may be regarded as the last pivotal row, in which case the rank of the matrix or of the vector space is equal to the number of pivotal rows. An example is given below with four rows.

<table>
<thead>
<tr>
<th>Column No.</th>
<th>Row No. (1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>(2)</td>
<td>2</td>
<td>-1</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>(3)</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>(4)</td>
<td>0</td>
<td>-2</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

Operations on rows

<table>
<thead>
<tr>
<th>(5)</th>
<th>1</th>
<th>-0.5</th>
<th>2.5</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>2</td>
<td>-4</td>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>-2</td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>-8</td>
<td>-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>8</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>1</td>
<td>1\frac{5}{8}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(13)</td>
<td>0</td>
<td>(11)</td>
<td>-1 \times 8</td>
<td></td>
</tr>
</tbody>
</table>

The rank of the vector space is, therefore, 3, the independent vectors being the pivotal rows

\[
\begin{array}{cccc}
1 & -0.5 & 2.5 & 2 \\
0 & 1 & 2 & 3 \\
0 & 0 & 1 & 1\frac{5}{8} \\
\end{array}
\]

1a.4 The Orthogonal Vector Space and the Deficiency Matrix

The set of vectors orthogonal to all the vectors in a matrix generates an orthogonal vector space since any linear combination in it satisfies
the orthogonality condition. The basis of the orthogonal vector space can be found in a simple way by an extension of the sweep-out process described above. Consider the basis of the vector space in the above example and sweep out the third and second columns using the third and second rows, yielding
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]
which forms another basis for the same vector space. In general, the sweep-out method gives a reduced matrix (after a suitable rearrangement of columns if necessary to bring the unit elements to the diagonal) of the form
\[
\begin{pmatrix}
1 & 0 & \cdots & 0 & b_{11} & \cdots & b_{1k} & 0 \\
0 & 1 & \cdots & 0 & b_{21} & \cdots & b_{2k} & \xi \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & b_{n1} & \cdots & b_{nk}
\end{pmatrix}
\]
All these vectors are independent. Consider the set of independent vectors
\[
D:
\begin{pmatrix}
b_{11} & b_{21} & \cdots & b_{11} & -1 & 0 & \cdots & 0 \\
b_{12} & b_{22} & \cdots & b_{12} & 0 & -1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\
b_{1k} & b_{2k} & \cdots & b_{1k} & 0 & 0 & \cdots & -1
\end{pmatrix}
\]
which is obtained by a rearrangement of the elements in \( S \). The vectors in \( D \) are mutually orthogonal to those in \( S \). Since \((r + k)\) is the total number of independent vectors possible, \( D \) generates all vectors orthogonal to \( S \) and for this reason is termed the deficiency matrix. In the above example, the deficiency matrix consists of the single row
\[
D:
\begin{pmatrix}
-\frac{1}{2} \\
-\frac{1}{2} \\
\frac{1}{2}
\end{pmatrix}
\]
\textbf{Example 1.} In a matrix the number of independent row vectors is equal to the number of independent column vectors.
\textbf{Example 2.} If each vector consists of \( n \) elements, then there cannot be more than \( n \) independent vectors in a set.
\textbf{Example 3.} Any vector of \( n \) elements can be expressed as a linear combination of any given set of vectors with rank \( n \).
Example 4. The rank of the matrix

\[
\begin{bmatrix}
1 & 1 & \cdots & 1 \\
\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n}
\end{bmatrix}
\]

containing \( n \) rows and \( n \) columns, is \( (n - 1) \).

1a.5 Linear Equations

The \( m \) equations in \( n \) unknowns \( x_1, \ldots, x_n \)

\[
a_{11}x_1 + \cdots + a_{1n}x_n = 0 \\
\vdots \quad \vdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n = 0
\]

are called homogeneous linear equations. Since any solution considered as a vector is orthogonal to every vector of the matrix

\[
\begin{bmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{m1} & \cdots & a_{mn}
\end{bmatrix}
\]

the totality of the solution vectors forms a space orthogonal to \( M \). The basis of this orthogonal space is the deficiency matrix \( D \) to be determined as in the previous section. If the rank of \( M \) is \( r \), then the solution space has a basis consisting of \( (n - r) \) independent vectors.

Replacing the zeros in the above equations by \( b_1, \ldots, b_m \), we obtain a set of non-homogeneous equations which can be regarded as homogeneous in \( (n + 1) \) unknowns.

\[
a_{11}x_1 + \cdots + a_{1n}x_n - b_1x_{n+1} = 0 \\
\vdots \quad \vdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n - b_mx_{n+1} = 0
\]

Only those solutions for which \( x_{n+1} \neq 0 \) will yield solutions to the non-homogeneous equations. If \( c = (c_1, \ldots, c_m) \) is any vector orthogonal to column vectors in \( M \), then, multiplying the above equations by
6  ALGEBRA OF VECTORS AND MATRICES

c_1, \ldots, c_m \text{ and adding, we obtain}

\[ x_{n+1}(b_1c_1 + \cdots + b_mc_m) = 0 \]

If \( \Sigma b_ic_i \neq 0 \) for at least one \( c \), then \( x_{n+1} = 0 \). On the other hand, if \( (b_1, \ldots, b_m) \) is dependent on the column vectors in \( M \), then \( \Sigma b_ic_i = 0 \) for all \( c \) orthogonal to \( M \), in which case the number of independent column vectors in \( N \), obtained by adding the column vector with elements \( b_1, b_2, \ldots, b_m \) to \( M \), remains the same as before. This means that the homogeneous equations in \( (n+1) \) unknowns have \( (n+1-r) \) independent solutions. Of these, \( (n-r) \) independent solutions are the vectors in \( D \), the deficiency matrix of \( M \), with zeros added to form the \( (n+1) \)th elements. The one more independent solution must necessarily have a non-zero value of \( x_{n+1} \), for otherwise it leads to a contradiction. Hence the necessary and sufficient condition for the non-homogeneous equations to have a solution is that \( (b_1, \ldots, b_m) \) is dependent on the column vectors or that the ranks of \( N \) and \( M \) are equal.

Example 1. Find the value of \( \delta \) for which the equations in three unknowns

\[
\begin{align*}
2x_1 - x_2 + 5x_3 &= 4 \\
4x_1 + 6x_3 &= 1 \\
-2x_2 + 4x_3 &= 7 + \delta
\end{align*}
\]

admit a solution.

1b  Theory of Matrices and Determinants

1b.1 Matrices

A matrix is, in general, an arrangement of \( pq \) elements in \( p \) rows and \( q \) columns. If \( A \) and \( B \) are two matrices of the form \((p, q)\)

\[
A = \begin{pmatrix}
a_{11} & \cdots & a_{1q} \\
\vdots & \ddots & \vdots \\
a_{p1} & \cdots & a_{pq}
\end{pmatrix}
\quad \text{and} \quad
B = \begin{pmatrix}
b_{11} & \cdots & b_{1q} \\
\vdots & \ddots & \vdots \\
b_{p1} & \cdots & b_{pq}
\end{pmatrix}
\]

then matrix addition is defined by

\[
A + B = \begin{pmatrix}
a_{11} + b_{11} & \cdots & a_{1q} + b_{1q} \\
\vdots & \ddots & \vdots \\
a_{p1} + b_{p1} & \cdots & a_{pq} + b_{pq}
\end{pmatrix}
\]

When there is no ambiguity about the number of rows and columns,
it may be convenient to denote the matrices $A$ and $B$ by $(a_{ij})$ and $(b_{ij})$, in which case $A + B = (a_{ij} + b_{ij})$. This process is known as matrix addition. Just as in the case of vectors, a matrix can be multiplied by a scalar $c$, the law of multiplication being

$$cA = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = (ca_{ij})$$

Consider two matrices $A$ and $B$ such that the number of columns in the first is equal to the number of rows in the second. For instance,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix}$$

are two such matrices. The product $AB$ of $A$ and $B$ is defined to be a matrix whose $(i, j)^{th}$ element is the product of the $i^{th}$ row vector of $A$ and the $j^{th}$ column vector of $B$.

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \end{pmatrix}$$

In general, two matrices $A$ and $B$ can be multiplied in the above manner only when the number of columns in $A$ is equal to the number of rows in $B$. The resulting matrix has the same number of rows as in $A$ and the same number of columns as in $B$.

The product $AB$ is not, in general, equal to the product $BA$. When the product $AB$ is considered, $A$ is said to be post-multiplied by $B$ or $B$ pre-multiplied by $A$.

The product law is associative so that the product of three matrices $A, B, C$ can be done in any of the following ways:

$$ABC = A(BC) = (AB)C$$

For multiplication to be compatible the matrices $A, B, C$ should be of the form $(p, q), (q, r), (r, s)$, in which case the triple product is of the form $(p, s)$. Observe the rule:

$$(p, q)(q, r)(r, s) = (p, s)$$

The matrix $A'$ obtained by interchanging the rows and columns of $A$ is called the transpose of $A$. From the definition it follows that

$$(AB)' = B'A' \quad (ABC)' = C'B'A' \quad \text{etc.}$$
If $\mathbf{x}$ is a row vector, then $\mathbf{x}'$ will be a column vector, in which case the vector multiplication $\mathbf{x} \cdot \mathbf{y}$ of two vectors $\mathbf{x}$ and $\mathbf{y}$ can also be written $\mathbf{xy}'$. Both representations will be used throughout.

A matrix that contains all zero elements is called a null matrix. It is easy to verify that the addition of a null matrix leaves any matrix unaltered whereas multiplication by a null matrix reduces any other matrix to a null matrix.

A matrix with equal number of rows and columns having unity for all its diagonal elements and zero elsewhere is called a unit matrix represented by $I$. It is easy to verify that $AI = A$ and $IA = A$ provided that multiplication is permissible. The distributive law holds for matrix multiplication.

$$A(B + C) = AB + AC$$
$$A(I - A) = 0$$

where $I$ is the unit matrix, then rank $A + \text{rank } (I - A) = n$.

The condition $A(I - A) = 0$ implies that vectors in $A$, being orthogonal to the vectors in $(I - A)$, are independent of the vectors in $(I - A)$.

If $r$ and $s$ are the numbers of independent vectors or the ranks of $A$ and $(I - A)$, then the number of independent vectors in $A$ and $(I - A)$ put together is $(r + s)$. If every row in $(I - A)$ is replaced by the sum of the corresponding rows in $(I - A)$ and $A$, then $n$ rows of the form

$$1 \ 0 \ \cdots \ 0$$
$$0 \ 1 \ \cdots \ 0$$
$$\vdots \ \vdots \ \cdots$$
$$0 \ 0 \ \cdots \ 1$$
PARTITIONED MATRICES

are obtained. These being independent, it follows that the number of independent vectors in \( A \) and \( (I - A) \) is not less than \( n \). Since there are only \( n \) elements in a vector, this is the maximum possible number of independent vectors. Hence \( (r + s) = n \).

**Example 2.** The rank of the product \( AB \) is not greater than the rank of \( A \) or the rank of \( B \).

The product \( AB \) can be obtained from \( B \) by suitable linear combinations of rows. This process does not increase the number of independent rows in \( B \). Therefore rank \( AB \geq \) rank \( B \). Also \( (AB)' = B'A' \), so that the above argument leads to the result rank \( (AB)' \geq \) rank \( A' \), which means that rank \( AB \geq \) rank \( A \).

**Example 3.** Let \( \alpha_1, \ldots, \alpha_s \) be a set of \( r \) vectors generating a vector space. If \( \beta_1, \ldots, \beta_s \) are \( s \) independent vectors, then the maximum number of independent vectors belonging to the \( \beta \) space and lying entirely in the \( \alpha \) space is equal to \( t \), where \( (s - t) \) is the rank of the matrix \( (\delta_i \cdot \beta_j) \), \( (i = 1, \ldots, j = 1, \ldots, s) \), and \( \delta_1, \delta_2, \ldots \) are the independent vectors generating the vector space orthogonal to the \( \alpha \) space.

In other words, there are \( t \) independent linear functions of \( \beta \) which can be expressed as linear functions of \( \alpha \) only.

Consider any linear function

\[
\gamma = l_1\beta_1 + \cdots + l_s\beta_s,
\]

and express the condition that it is orthogonal to \( \delta_1, \delta_2, \ldots \).

\[
l_1\delta_1\beta_1 + \cdots + l_s\delta_s\beta_s = 0 \quad i = 1, 2, \ldots
\]

The number of independent solutions is evidently \( s \) minus the rank of \( (\delta_i \cdot \beta_j) \). Each solution supplies a linear combination of \( \beta \) lying in the \( \alpha \) space. If \( \gamma_1, \ldots, \gamma_t \) are these vectors, then there are \( (s - t) \) more vectors belonging to the \( \beta \) space, \( \gamma_{t+1}, \ldots, \gamma_s \), such that no linear combination of \( \gamma_{t+1}, \ldots, \gamma_s \) belongs to the \( \alpha \) space, for otherwise a contradiction is obtained.

**Example 4.** If the row vectors of a square matrix are all mutually orthogonal and the square of each row vector is unity, so are its column vectors.

1b.2 Partitioned Matrices

Sometimes it is convenient to represent a matrix obtained by the juxtaposition of two or more matrices in a partitioned form. Thus a partitioned matrix \( A \) is represented by

\[
A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}
\]
10 ALGEBRA OF VECTORS AND MATRICES

where the rows in \( P \) equal in number those in \( Q \), the columns in \( P \)
equal those in \( R \), and so on. By definition,

\[
A' = \begin{pmatrix} P' & R' \\ Q' & S' \end{pmatrix}
\]

If

\[
B = \begin{pmatrix} E' & F' \\ G' & H' \end{pmatrix}
\]

then

\[
AB = \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} PE + QG & PF + QH \\ RE + SG & RF + SH \end{pmatrix}
\]

provided that the products \( PE \), etc., are permissible.

1b.3 Determinants

A determinant is a real valued function of the elements of a square matrix. If \( x_1, x_2, \ldots, x_p \) denote the row vectors, then the function may be represented by \( D(x_1, \ldots, x_p) \). We shall choose the function to satisfy the following conditions:

(a) \( D(x_1, \ldots, cx_m, \ldots, x_p) = cd(x_1, \ldots, x_m, \ldots, x_p) \), where \( c \) is a scalar.
(b) \( D(x_1, \ldots, x_m + x_k, \ldots, x_p) = D(x_1, \ldots, x_m, \ldots, x_p) \), for \( m \neq k \).
(c) \( D(e_1, \ldots, e_p) = 1 \), where \( e_1, \ldots, e_p \) constitute the vectors of a unit matrix and are called elementary vectors.

Let \( D \) exist when the following properties hold:

(1) If \( x_i = 0 \), then \( x_i = 0x_i \); hence, by (a), \( D = 0 \), putting \( c = 0 \).
(2) \( D(x_1, \ldots, x_m + cx_k, \ldots, x_p) = D(x_1, \ldots, x_m, \ldots, x_p) \).

This is true when \( c = 0 \). If \( c \) is not zero, then

\[
D(x_1, \ldots, x_m + cx_k, \ldots, x_p)
= \frac{1}{c} D(x_1, \ldots, x_m + cx_k, \ldots, -cx_k, \ldots, x_p) \quad \text{by (a)}
\]

(3) From (2) it follows that, if the rows are dependent, then \( D = 0 \).
As a particular case, if two rows are identical the determinant vanishes.
DETERMINANTS

(4) If two rows are interchanged, then $D$ changes sign.

$$D = D(x_1, \ldots, x_n + x_p, \ldots, x_p)$$

$$= D(x_1, \ldots, x_n + x_p, \ldots, x_m - x_i, \ldots, x_p)$$

$$= D(x_1, \ldots, x_n + x_p, \ldots, -x_m, \ldots, x_p)$$

$$= D(x_1, \ldots, x_i, \ldots, -x_m, \ldots, x_p)$$

$$= -D(x_1, \ldots, x_i, \ldots, x_m, \ldots, x_p)$$

In general, an even permutation of rows does not alter the determinant whereas an odd permutation changes sign.

(5)

$$D(x_1, \ldots, x_n + y, \ldots, x_p)$$

$$= D(x_1, \ldots, x_n, \ldots, x_p) + D(x_1, \ldots, y, \ldots, x_p)$$

If $x_n$ depends on the rest of $x$, the result is established by subtracting a suitable linear combination of other vectors from the $m$th vector in each of the above determinants. If the other vectors are themselves dependent, then each term above is zero. The only alternative is that all $x$ are independent, in which case $y$ is necessarily a linear function of $x$ (because there cannot be more than $p$ independent vectors).

$$y = \Sigma c_i x_i$$

$$D(x_1, \ldots, x_n + \Sigma c_i x_i, \ldots, x_p) = (1 + c_m)D(x_1, \ldots, x_n, \ldots, x_p)$$

$$= D(x_1, \ldots, x_n, \ldots, x_p) + D(x_1, \ldots, c_m x_m, \ldots, x_p)$$

$$= D(x_1, \ldots, x_m, \ldots, x_p) + D(x_1, \ldots, \Sigma c_i x_i, \ldots, x_p)$$

(6) $D(x_1 + y_1, \ldots, x_p + y_p) = \Sigma D(z_1, \ldots, z_p)$, where $z_i$ can be $x_i$ or $y_i$ and the summation is over $2^p$ possible sets, $z_1, \ldots, z_p$. This follows by repeated application of (5).

(7) Any vector $x_n = a_{n1} e_1 + a_{n2} e_2 + \cdots + a_{np} e_p$, where $e_1, \ldots, e_p$ are elementary vectors. Therefore

$$D(x_1, \ldots, x_n, \ldots, x_p) = D(\Sigma a_{11} e_1, x_2, \ldots, x_p)$$

$$= \Sigma a_{11} D(e_1, x_2, \ldots, x_p)$$

$$= \Sigma a_{11} a_{22} D(e_1, e_2, \ldots, x_p)$$

$$\ldots \ldots \ldots \ldots \Sigma a_{11} a_{22} \cdots a_{pp} D(e_1, e_2, \ldots, e_p)$$
In the final summation,
\[ D(e_i, e_j, \cdots, e_k) = 0 \quad \text{whenever two suffixes are equal} \]
\[ = +1 \quad \text{when } i, j, \cdots \text{ is an even permutation of } 1, 2, \cdots, p \]
\[ = -1 \quad \text{when } i, j, \cdots \text{ is an odd permutation} \]
Hence
\[ D = \sum \pm a_{ij} a_{kl} \cdots a_{ap} \]

The function so derived satisfies the conditions (a), (b), and (c) given above and is called the value of the determinant of the square matrix \((a_{ij})\) and is also denoted by \( |a_{ij}| \) or simply \( |A| \).

\[ (8) \]
\[ D = \Sigma a_{mi} D(x_1, \cdots, e_i, \cdots, x_p) \]
\[ = \Sigma a_{mi} A_{mi} \]

where \( A_{mi} \) is the determinant obtained by replacing the \( m \)th vector by \( e_i \) and is called the cofactor of \( a_{mi} \). The minor of \( a_{mi} \) is defined as the determinant obtained by omitting the \( m \)th row and \( i \)th column. The cofactor is obtained from the minor by the relation

\[ (9) \]
\[ \sum a_{mi} A_{ni} = 0 \quad \text{if } m \neq n \]

because this is the value of a determinant with the \( m \)th and \( s \)th rows identical.

From the definition it follows that, when the rank of the matrix \( A \) of the type \((n, n)\) is less than \( n \), then \( |A| = 0 \). To prove the converse, it is necessary to recall the sweep-out method described in 1a.3. When a column is swept out the only operation that changes the value of the determinant is the division by a non-zero element, also called the pivotal element, to obtain the pivotal row. The pivotal row may be moved to the first position if necessary by an interchange of rows in which case the determinant changes sign. Expanding by the first column, it is seen the determinant of this altered matrix is same as the determinant of the reduced matrix of one order less obtained by the omission of the pivotal row and the swept-out column. This means that \( |A| \) is, apart from a sign, equal to the determinant of the reduced matrix at any stage multiplied by the product of the pivotal elements.
Determinants

used up to that stage. If the rank of \( A \) is less than \( n \), a zero row will be encountered at some stage leading to a null value of the reduced matrix. If \( A \) has full rank, then the sweep-out process can be carried out to the last row giving a non-zero value to \( |A| \). Hence,

(i) If \( |A| = 0 \), the rank of \( A \) must be less than \( n \).

(ii) If \( |A| \neq 0 \), the rank of \( A \) is \( n \), or in other words the rows and columns of \( A \) are all independent.

**Example 1.** If \( |A| \), the determinant of \( A \), is not zero, in which case \( A \) is called a non-singular matrix, then there exists a matrix represented by \( A^{-1} \) such that

\[
AA^{-1} = A^{-1}A = I
\]

Defining the matrix

\[
A^{-1} = (a^{ij})
\]

where \( a^{ij} = A_{j} \!/ |A| \), \( A_{j} \) being the cofactor of \( a_{ij} \), it is easy to verify that the above result is true. The matrix \( A^{-1} \) is called the reciprocal of \( A \) and is defined only when \( |A| \neq 0 \).

**Example 2.** If \( X \) is an unknown matrix involved in the equation \( XA = Y \), then \( X = YA^{-1} \), provided that \( A^{-1} \) exists.

**Example 3.** If \( B \) is a matrix of the form \((m, n)\) and \( A \) of the form \((n, n)\) with rank \( n \), then

\[
\text{Rank } B = \text{Rank } BA
\]

Let \( s \) be the rank of \( B \) and \( r \) that of \( BA \). Since \( B \) is the product of \( BA \) and \( A^{-1} \), \( s \) is not greater than \( r \). This in conjunction with the earlier result (example 2, 1b.1) yields \( s = r \).

**Example 4.** If the rank of a matrix \( A \) is \( r \), then all subdeterminants of the order \((r + 1)\) or greater vanish. This is true because there are not more than \( r \) independent rows or columns. Conversely, if all determinants of the order \((r + 1)\) or greater vanish and at least one determinant of the order \( r \) does not vanish, then the rank of \( A \) is \( r \).

**Example 5.** The rank of a matrix \( A \) is unaltered by pre- or post-multiplication by an elementary matrix \( E_r(\lambda) \) where \( E_r(\lambda) \) is defined by \( (e_{ij}) \).

\[
e_{ij} = 1 \quad \text{for all } i \\
e_{ij} = \lambda \quad \text{for } i = r \text{ and } j = s \\
e_{ij} = 0 \quad \text{for other values of } i \text{ and } j
\]

The proposition is true because \( |E_r(\lambda)| \neq 0 \). Pre-multiplication by \( E_r(\lambda) \) means replacing the \( r \)th row of \( A \) by its \( r \)th row + \( \lambda \) times the \( s \)th row. Post-multiplication means replacing the \( s \)th column of \( A \) by the \( s \)th column + \( \lambda \) times the \( r \)th column.
Example 6. If $A$ is a square symmetric matrix, then there exists a non-singular matrix $B$ such that the matrix $BAB'$ is in the diagonal form.

If there is a non-zero diagonal element in $A$, then it may be used as a pivot and the row and column in which it occurs can be swept out, leaving a reduced matrix. This method consists in only row and column additions or, in other words, pre- and post-multiplications by elementary matrices. Since the matrix is symmetrical, the symmetrical elements in a row and column can be swept out by pre- and post-multiplying by elementary matrices which are only transposes. Thus, sweeping out a row and column is equivalent to pre- and post-multiplying by products of elementary matrices which are transposes. The reduced matrix is also symmetrical. The above process can be carried on whenever a non-zero diagonal element can be found. If a diagonal element is zero,
then, by the addition (or subtraction) of a row and the corresponding column a non-zero element can be brought to the diagonal position and the above process continued. This is also a symmetrical operation by the use of elementary matrices, so that it follows that the matrix $A$ can be reduced to the diagonal form by pre- and post-multiplying by $B$ and $B'$ where $B$ is a product of elementary matrices.

It can be easily seen that any non-symmetrical square matrix $A$ can be reduced to the diagonal form by pre- and post-multiplications by matrices which need not be transposes.

Example 7. If the product $AB$ of two square matrices is zero, then either $A = 0$, or $B = 0$, or both $A$ and $B$ are singular matrices.

Example 8. If $A$ and $B$ are square matrices of order $n$ and ranks $r$ and $s$, then

$$\text{Rank } AB \geq r + s - n$$

From example 6 there exist two non-singular matrices $C$ and $D$ such that

$$CAD = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

where $I$ is the unit matrix ($r, r$).

$$\text{Rank of } AB = \text{Rank of } CAB \quad \text{since } C \text{ is non-singular}$$

$$= \text{Rank of } CADD^{-1}B$$

$$= \text{Rank of } \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} B_1$$

where $B_1 = D^{-1}B$ and hence has rank $s$. The last product is a matrix obtained by choosing the first $r$ rows of $B_1$ and the rest consisting of
zero rows. Therefore the rank of $AB$ is equal to the rank of the first $r$ rows of $B_1$. If this is equal to $t$, then the number of dependent rows is $(r - t)$. By considering all rows of $B$, we get $(n - s)$ dependent rows which must not be less than the dependent vectors in a subset.

Hence

$$r - t \leq n - s$$

or

$$t \geq r + s - n$$

**Example 9.** \( |AB| = |A| |B| \) where $A$ and $B$ are two square matrices.

There exist matrices $E$ and $F$ which are products of elementary matrices such that

$$EAF = D$$

where $D$ is diagonal with elements $d_1, \cdots, d_p$ so that $|A| = d_1 \cdots d_p$.

$$AB = AFF^{-1}B$$

$$|AB| = |EAFF^{-1}B|$$

Since the determinant is not altered on multiplication by elementary matrices,

$$|AB| = |DF^{-1}B|$$

$$= d_1 d_2 \cdots d_p F^{-1} B | = d_1 \cdots d_p |B| = |A| |B|$$

**Example 10.** If $A$ is a matrix of the type $(m, n)$, then

$$|AA'| \geq 0 \quad \text{if } m \leq n$$

$$= 0 \quad \text{if } m > n$$

If $m \leq n$, there exists a matrix $B$ of the type $(n - m, n)$ containing row vectors which are orthogonal to those in $A$ (i.e., $AB' = 0$) and satisfying the condition $BB' = I$. Consider the product

$$\begin{pmatrix} B' \\ A' \end{pmatrix} \begin{pmatrix} B' & A' \end{pmatrix} = \begin{pmatrix} B'B & B'A' \\ AB' & AA' \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & AA' \end{pmatrix}$$

Taking determinants,

$$\begin{vmatrix} B' \\ A \end{vmatrix} = |AA'| \geq 0$$

* Consider the equations $xA' = 0$. This has at least $(n - m)$ independent solution vectors which may be replaced by an equivalent set of standardized orthogonal vectors. These vectors form the matrix $B$. 
If \( m > n \), then

\[
(A \ 0) \begin{pmatrix} A' \\ 0 \end{pmatrix} = AA' + 0 = AA'
\]

where 0 stands for a null matrix. Taking determinants,

\[
|AA'| = |A| |0|^2 = 0
\]

**Example 11.** If \( A \) is a matrix of the type \((m, n)\), \( m \leq n \), then:

(i) \( |AA'| \) = the sum of squares of all possible \( m \) columned determinants in \( A \).

(ii) Rank of \( AA' \) = Rank of \( A \).

Let \( A = (a_{ij}) \), in which case

\[
AA' = \sum (a_{ij}a_{jr})
\]

\[
|AA'| = \sum \begin{vmatrix} a_{1r}a_{1s} & \cdots & a_{1r}a_{1t} \\ \cdots & \cdots & \cdots \\ a_{nr}a_{sr} & \cdots & a_{nr}a_{sr} \end{vmatrix}
\]

summed over all \( n^m \) sets of \((r, s, \cdots, t)\); \((r, s, \cdots, t) = 1, 2, \cdots, n\). In this summation it is easy to see that the determinants in which any two of the symbols \((r, s, \cdots, t)\) are equal vanish so that the summation is over sets in which \( r \neq s \cdots \neq t \). Corresponding to any set \((r, s, \cdots, t)\) there are \( m^m \) permutations, the determinants arising out of which can be grouped into a single determinant. Thus

\[
|AA'| = \sum \begin{vmatrix} a_{1r}a_{1s} + \cdots + a_{1r}a_{1t} & \cdots & a_{1r}a_{mr} + \cdots + a_{1r}a_{mt} \\ \cdots & \cdots & \cdots \\ a_{nr}a_{sr} + \cdots + a_{nr}a_{sr} & \cdots & a_{nr}a_{mr} + \cdots + a_{nr}a_{mr} \end{vmatrix}
\]

summed over the \( m \) \( ^m \) combinations \((r, s, \cdots, t)\) from 1 to \( n \),

\[
= \sum \begin{vmatrix} a_{1r} & a_{1s} & \cdots & a_{1t} \\ a_{2r} & a_{2s} & \cdots & a_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ a_{nr} & a_{mr} & \cdots & a_{mt} \end{vmatrix}^2
\]

which proves result (i).

To prove (ii), let the rank of \( A \) be \( r \) so that there are \( r \) independent vectors which may be marked. In the product \( AA' \), the marked rows and the corresponding columns in \( A' \) give rise to a determinant of order \( r \) which, by the above proposition, is equal to the sum of squares of all possible \( r \) columned determinants chosen out of the \( r \) marked rows.
Since the $r$ rows in $A$ are independent, there is at least one determinant containing $r$ independent columns in it, so that the $r$th-order determinant obtained from the marked rows and columns in $A$ and $A'$ is not zero. Therefore, the rank of $AA' < r$. But the rank of the product $AA'$ cannot exceed the rank of $A$, which is $r$. Therefore, the rank of $AA' = \text{the rank of } A$.

Example 12. Defining for any $z_1, \cdots, z_n$

$$S_i = (x_1 - z^i)^2 + (x_2 - z^i)^2 + \cdots + (x_n - z^i)^2 = d_1^i + d_2^i + \cdots + d_n^i$$

$$\bar{z} = \frac{z_1 + z_2 + \cdots + z_n}{n}$$

di = (z_i - \bar{z}) \quad i = 1, 2, \cdots, n$$

show that the determinant

$$\begin{vmatrix}
S_0 & S_1 & \cdots & S_i \\
S_1 & S_2 & \cdots & S_{i+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_i & S_{i+1} & \cdots & S_{i+i}
\end{vmatrix}$$

is not less than zero for all $i$.

This follows from the fact that the above determinant is $|AA'| \geq 0$ (example 10 above), where

$$A = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
d_1^i & d_2^i & \cdots & d_n^i \\
\vdots & \vdots & \ddots & \vdots \\
d_1^i & d_2^i & \cdots & d_n^i
\end{bmatrix}$$

This proves the consistency relations to be satisfied by the moments calculated from any sample of observations.

Example 13. The $j$th moment of a number of variables is defined by $S_j/j!$, where $S_j$ is as in example 12. Two constants $\beta_1$ and $\beta_2$ defined by K. Pearson are

$$\beta_1 = \frac{\mu_j^3}{\mu_j^2} \quad \beta_2 = \frac{\mu_j^4}{\mu_j^2}$$

To show that $\beta_2 \geq 1 + \beta_1$, consider the determinant of example 12 for $i = 2$.

$$\begin{vmatrix}
n & 0 & n \mu_2 \\
0 & n \mu_3 & n \mu_3 \\
n \mu_2 & n \mu_3 & n \mu_4
\end{vmatrix} \geq 0$$
Expanding,
\[ n^2(\mu_2 \mu_4 - \mu_3^2 - \mu_2^2) \geq 0 \]
Hence the result.

1c Quadratic Forms

1c.1 Definitions

The general quadratic form in \( n \) variables \( x_1, \ldots, x_n \) is

\[
\begin{align*}
& a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\
& + a_{21}x_2x_1 + a_{22}x_2^2 + \cdots + a_{2n}x_2x_n \\
& \vdots & \vdots & \vdots \\
& + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_n^2
\end{align*}
\]

where \( a_{ij} = a_{ji} \). Adopting the matrix notation, the above quadratic form can be written

\[ \mathbf{x}^T \mathbf{A} \mathbf{x} \]

where \( \mathbf{x} \) is the vector \((x_1, \ldots, x_n)\) and \( \mathbf{A} \) is the symmetric matrix \((a_{ij})\).

The matrix \( \mathbf{A} \) is called the matrix of the quadratic form \( \mathbf{x}^T \mathbf{A} \mathbf{x} \), and \( \lambda(\mathbf{A}) \), its discriminant.

The rank of the quadratic form \( \mathbf{x}^T \mathbf{A} \mathbf{x} \) is the same as the rank of the matrix \( \mathbf{A} \).

1c.2 Linear Transformations

Let the variables in \( \mathbf{x} \) be transformed to those in \( \mathbf{y} = (y_1, \ldots, y_n) \) by means of the transformation

\[ \mathbf{x} = \mathbf{y} \mathbf{C} \]

(i) Under this transformation the quadratic form \( \mathbf{x}^T \mathbf{A} \mathbf{x} \) changes to \( \mathbf{y}^T \mathbf{C} \mathbf{A} \mathbf{C}^T \mathbf{y} \) so that the matrix of the new form is \( \mathbf{C} \mathbf{A} \mathbf{C}^T \). The discriminant of the transformed quadratic form is \( \lambda(\mathbf{C} \mathbf{A} \mathbf{C}^T) = \lambda(\mathbf{A}) \).

(ii) It has been shown in example 6 of 1b.3 that there exists a matrix \( \mathbf{B}, |\mathbf{B}| \neq 0 \) such that the matrix \( \mathbf{B} \mathbf{A} \mathbf{B}^T \) is in the diagonal form. If this matrix \( \mathbf{B} \) is chosen as the matrix of transformation from \( \mathbf{x} \) to \( \mathbf{y} \), then the quadratic form can be reduced to the form \( c_1y_1^2 + \cdots + c_ry_r^2 \) containing the square terms only. The value of \( |\mathbf{B}| = 1 \) since \( \mathbf{B} \) is a product of elementary matrices which have a unit determinant.

By making a further transformation \( \sqrt{|c_i|} y_i = z_i \) the quadratic form becomes \( \pm z_1^2 \pm z_2^2 \pm \cdots \pm z_r^2 \).

(iii) If the rank of the matrix \( \mathbf{A} \) is \( r \), then the reduced quadratic form contains only \( r \) square terms. This follows, for the ranks of \( \mathbf{A} \) and \( \mathbf{B} \mathbf{A} \mathbf{B}^T \), where \( |\mathbf{B}| \neq 0 \) (example 3, 1b.3), are the same.
(iv) A linear transformation \( x = yC \) is said to be orthogonal if \( CC' = I \). The transformation is non-singular since \( |C|^2 = 1 \). The quadratic form \( x_1^2 + \cdots + x_n^2 = xIx' \) changes over to
\[
yCIC'y' = yCC'y' = y_1^2 + y_2^2 + \cdots + y_n^2
\]
This is referred to as the invariance property of the distance function under an orthogonal transformation. Also let \( x_1 \) and \( x_2 \) be two vectors transforming to \( y_1 \) and \( y_2 \). Then
\[
x_1x_2' = y_1CC'y_2' = y_1y_2'
\]
so that the angles are also invariant. Also, if \( x \) transforms to \( y \) and \( a \) to \( b \), then
\[
(x - a)^2 = (y - b)^2
\]
and
\[
(x_1 - a_1)(x_2 - a_2)' = (y_1 - b_1)(y_2 - b_2)'
\]

1c.3 Classification of Quadratic Forms

The real quadratic form \( xAx' \) is said to be definite if it is positive (or negative) for every set of real values \( x_1, \ldots, x_n \) other than the set \( x_1 = \cdots = x_n = 0 \). A quadratic form which is never negative but which assumes zero value for some non-null values of \( x_1, x_2, \ldots, x_n \) is called semi-positive definite. Similarly, semi-negative forms can be defined.

(i) The definiteness of a quadratic form is invariant under non-singular transformations.

Since the transformation \( x = yB \) is non-singular, there exists the inverse transformation \( y = xB^{-1} \), which establishes a one-to-one correspondence. If the quadratic form is positive (or negative) for a given vector \( x \), then the transformed form is positive (or negative) for the corresponding \( y \), and vice versa. Also \( y = (0, \ldots, 0) \) when and only when \( x = (0, \ldots, 0) \).

(ii) Every real positive definite quadratic form can be transformed by a real transformation matrix of unit modulus to the form
\[
c_{11}y_1^2 + \cdots + c_{nn}y_n^2
\]
where each \( c_{ii} > 0 \).

It is shown in (ii) of 1c.2 that every quadratic form can be reduced to a form consisting of the square terms only. No coefficient is negative, for otherwise it implies that the quadratic form is negative for some values of \( y \neq 0 \) and hence of \( x \). Also, no coefficient can be zero, for, if \( c_{ii} = 0 \), then the quadratic form vanishes for \( y_i \neq 0 \) and others equal to zero, which is contrary to the assumption of the definiteness of the quadratic form.

(iii) The necessary and sufficient condition that a real quadratic form \( xAx' \) is positive definite is that
Let the positive definite quadratic form under the transformation
\[ x = yB \]
be reduced to
\[ C_1y_1^2 + \cdots + C_ny_n^2, \]
where \( C_1, \ldots, C_n \) are positive.

In such a case
\[
BAB' = \begin{pmatrix}
C_1 & 0 & \cdots & 0 \\
0 & C_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & C_n
\end{pmatrix}
\]

Therefore \( \det B \) is a positive definite quadratic form. Consider the set of values of \( y_1, \ldots, y_n \) in which \( y_n = 0 \). Then, from the above argument, it can be shown that
\[
\begin{vmatrix}
a_{11} & \cdots & a_{1(n-1)} \\
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{vmatrix} > 0
\]
and so on, which establishes the necessity of the condition. To prove sufficiency, let
\[
\Delta_1 = \begin{vmatrix}
a_{11} & \cdots & a_{1(n-1)} \\
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{vmatrix}
\]

Since \( a_{11} > 0 \), the first column and row in \( A \) can be swept out. The resulting matrix is
\[
\begin{pmatrix}
a_{11} & 0 & \cdots & 0 \\
0 & b_{22} & \cdots & b_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
0 & b_{n2} & \cdots & b_{nn}
\end{pmatrix}
\]
where \( a_{11}b_{22} = \Delta_2 \), for the value of any subdeterminant including the first row and column is unaltered. Also \( \Delta_2 > 0 \). Hence it follows that \( b_{22} = \Delta_2/a_{11} > 0 \). With \( b_{22} \) as a pivot the second row and column can be swept out. The resulting matrix is
\[
\begin{pmatrix}
a_{11} & 0 & \cdots & 0 \\
0 & b_{22} & \cdots & b_{2n} \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & c_{n2} & \cdots & c_{nn}
\end{pmatrix}
\]
where $a_1b_2c_3 = \Delta_2$. Therefore $c_3 = \Delta_2/\Delta_2 > 0$, and so on. Finally the matrix $A$ can be reduced to the diagonal form

$$
\begin{bmatrix}
\Delta_1 & 0 & \cdots & 0 \\
0 & \Delta_2/\Delta_1 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \Delta_n/\Delta_{n-1}
\end{bmatrix}
$$

where each diagonal element is positive. This shows that the quadratic form $xAx'$ can be transformed to the positive definite form

$$
\Delta_1 y_1^2 + \frac{\Delta_2}{\Delta_1} y_2^2 + \ldots + \frac{\Delta_n}{\Delta_{n-1}} y_n^2
$$

which establishes the sufficiency of the condition. It should be noted that sweeping out is equivalent to pre- and post-multiplication by a product of elementary matrices which are transposes. This product of the elementary matrices provides the transformation matrix.

(iv) The necessary and sufficient condition that a real quadratic form $xAx'$ is negative definite is that

$$
\Delta_1 < 0, \quad \Delta_2 > 0, \quad \Delta_3 < 0, \quad \ldots
$$

This is true, since $-xAx'$ is positive definite.

1c.4 The Latent Roots of a Matrix and the Characteristic Vectors

Let $xAx'$ be a quadratic form in $n$ variables. Let us find a vector $x$ which maximizes $xAx'$ subject to the condition $xx' = 1$. This is obtained by differentiating \* 

$$
xAx' - \lambda(xx' - 1)
$$

\* The following rules of differentiation with respect to vectors (i.e., simultaneously with respect to all the variables) will be useful. Since $xx' = x_1^2 + x_2^2 + \ldots + x_n^2$, 

$$
\left( \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n} \right) (x_1^2 + \ldots + x_n^2) = 2(x_1, \ldots, x_n)
$$

Therefore 

$$
\frac{\partial}{\partial x} xx' = 2x
$$

Similarly 

$$
\frac{\partial}{\partial x} xAx' = 2xA
$$
where \( \lambda \) is a Lagrangian multiplier. The equations are

\[
xA - \lambda x I = 0 \\
x x' = 1
\]

In order that a non-null \( x \) may exist, \( \lambda \) must be chosen to satisfy the determinantal equation

\[
| A - \lambda I | = 0
\]

This is called the characteristic equation of the matrix \( A \). Any value of \( \lambda \) which satisfies this equation is called a latent root, and the \( x \) corresponding to a given \( \lambda \) is called a characteristic vector.

(i) The degree of the characteristic equation for roots other than \( \lambda = 0 \) is equal to the rank of \( A \).

This can be verified by expanding the determinantal equation. The coefficients of \( \lambda^{n-r-1}, \ldots, \lambda, \lambda^0 \), being the sums of determinants containing more than \( r \) columns and rows, will be zero if the rank of \( A \) is \( r \).

(ii) If \( A \) is real, all the roots are real.

Let \( x + iy \) be the characteristic vector corresponding to a complex root \( \lambda \). Then

\[
(x + iy) A - \lambda (x + iy) = 0
\]

Multiplying by \( (x - iy) \)

\[
\lambda (x^2 + y^2) = (x + iy) A (x - iy)' \\
= x Ax' + y A y' + i(y A x' - x A y')
\]

Hence \( \lambda \) is real.

(iii) The value of the quadratic form for a given characteristic vector \( x \) is equal to the value of the latent root \( \lambda \) associated with it.

Since

\[
xA - \lambda x I = 0
\]

Then

\[
x Ax' = \lambda x x' = \lambda
\]

The maximum and minimum values of \( x A x' \) subject to the condition \( x x' = 1 \) are then the largest and the least latent roots.

(iv) If the quadratic form is positive definite, then all the latent roots are positive. This is true, since the latent roots are the values of the quadratic form for some values of the variables.

(v) The characteristic vectors corresponding to two different latent roots are orthogonal.
Let \( \lambda_1, \lambda_2 \) be two roots and \( x, y \) the corresponding vectors. Then

\[
\begin{align*}
\mathbf{x}A - \lambda_1 \mathbf{x} &= 0 \\
\mathbf{y}A - \lambda_2 \mathbf{y} &= 0
\end{align*}
\]

Multiplying the first by \( \mathbf{y}' \), the second by \( \mathbf{x}' \), and subtracting,

\[
(\lambda_1 - \lambda_2)\mathbf{x}\mathbf{y}' = \mathbf{x}A\mathbf{y}' - \mathbf{y}A\mathbf{x}' = 0
\]

From this it follows that \( \mathbf{x}\mathbf{y}' = 0 \), since \( \lambda_1 \neq \lambda_2 \).

(vi) There exists an orthogonal transformation which transforms a quadratic form \( \mathbf{x}A\mathbf{x}' \) into \( \mathbf{y}\Lambda\mathbf{y}' \), where \( \Lambda \) is a diagonal matrix containing all the latent roots of \( A \).

Assume that \( X_i \) is a matrix of the form \((i, n)\), the rows of which are the characteristic vectors corresponding to the latent roots \( \lambda_1, \cdots, \lambda_n \), not all of which need be different. Also, let the rows of \( X_i \) be orthogonal.

We will show that under these conditions there exists a vector \( \mathbf{x} \) which is orthogonal to the rows in \( X_i \) and is a characteristic vector corresponding to the latent root \( \lambda_{i+1} \).

Since the row vectors are orthogonal and normalized

\[
X_iX_i' = I
\]

Since the \( j \)th vector satisfies the relation

\[
\mathbf{x}_jA - \lambda_j \mathbf{x}_j = 0
\]

it follows that

\[
\mathbf{X}_iA - \Lambda \mathbf{X}_i = 0
\]

where \( \Lambda \) is a diagonal matrix containing \( \lambda_1, \cdots, \lambda_n \) in the diagonal.

Let us find a vector \( \mathbf{x} \) \((\mathbf{xx}' = 1)\), orthogonal to the rows in \( X_i \) \( (\text{i.e., } X_i\mathbf{x} = \mathbf{0}) \) and which maximizes \( \mathbf{x}A\mathbf{x}' \). Introducing the vector \( \mu = (\mu_1, \cdots, \mu_n) \) of Lagrangian multipliers, the quantity to be differentiated is

\[
\mathbf{x}A\mathbf{x}' - 2\mu_1 \mathbf{X}_1' \mathbf{x} - \mu \mathbf{X}_i \mathbf{x}'
\]

Differentiating, we obtain

\[
2\mathbf{A}\mathbf{x} - 2\mu_1 \mathbf{X}_1' - 2\mu \mathbf{X}_i = 0 \tag{1c.4.1}
\]

Eliminating \( \mathbf{x} \) and \( \mu \), the equation giving \( \lambda \) is

\[
\begin{bmatrix}
\mathbf{A} - \Lambda & \mathbf{X}_i' \\
\mathbf{X}_i & 0
\end{bmatrix}
\begin{bmatrix}
\mu_1 \\
\lambda
\end{bmatrix} = 0 \tag{1c.4.2}
\]
Pre-multiplying this by
\[
\begin{pmatrix}
I & 0 \\
X_i & \lambda I - \Lambda_i
\end{pmatrix}
\]
we get
\[
\begin{pmatrix}
A - \lambda I & X_i' \\
0 & I
\end{pmatrix}
= | A - \lambda I | = 0
\]
Equation (1c.4.2) can be written
\[
\frac{| A - \lambda I |}{(\lambda - \lambda_1) \cdots (\lambda - \lambda_n)} = 0
\]
Hence all \( \lambda \) satisfying (1c.4.2) are the latent roots of \( | A - \lambda I | = 0 \) other than those already considered. Let us consider the latent root \( \lambda_{i+1} \) and solve for \( x \) and \( \mu \) from the set of equations (1c.4.1). Representing the solution for \( x \) by \( x_{i+1} \), we have
\[
x_{i+1}A = -\mu X_i - \lambda x_{i+1} = 0
\]
\[
X_i x'_{i+1} = 0
\]
Or, multiplying by \( X_i' \),
\[
x_{i+1}A X_i' = -\mu X_i X_i' - \lambda x_{i+1} X_i' = 0
\]
i.e.,
\[
x_{i+1}X_i'\Lambda_i - \mu I - \lambda x_{i+1} X_i' = 0
\]
i.e.,
\[
0\lambda_i - \mu I - 0 = 0
\]
Therefore
\[
\mu = 0
\]
which shows that \( x_{i+1} \) satisfies the equations
\[
x_{i+1} (A - \lambda_{i+1} I) = 0
\]
\[
X_i x'_{i+1} = 0
\]
Therefore \( x_{i+1} \) is a characteristic vector corresponding to \( \lambda_{i+1} \) and is orthogonal to \( X_i \).
Starting from the first characteristic vector, all the \( n \) can be constructed so that there exists an orthogonal transformation \( X_n \) such that
\[
X_n A = \Lambda_n X_n
\]
and hence
\[
X_n A X_n' = \Lambda_n
\]
Corollary. If \( \lambda_i \) is a root of multiplicity \( r \), then there are \( r \) and only \( r \) orthogonal vectors satisfying the equation
\[
x(A - \lambda_i I) = 0
\]
so that the rank of \( A - \lambda_i I \) is \( (n - r) \).

Corresponding to a latent root \( \lambda_i \) of multiplicity \( r_i \), there are, by the above result, \( r_i \) orthogonal characteristic vectors. Since \( (r_1 + r_2 + \cdots) = n \) and there can be only \( n \) orthogonal vectors, it follows that there are only \( r_i \) characteristic vectors corresponding to \( \lambda_i \). Hence the rank of \( (A - \lambda_i I) \) is \( (n - r) \).

To obtain the characteristic vectors corresponding to the root \( \lambda_i \) of multiplicity \( r_i \), the best method is to find the space orthogonal to \( (A - \lambda_i I) \) and choose any set of orthogonal vectors in this space.

1.c.5 Pairs of Quadratic Forms

Let \( A \) be a symmetric and \( B \) a p.d. matrix and \( x \) a vector such that
\[
x(A - \lambda B) = 0
\]
where \( \lambda \) satisfies the determinantal equation \( |A - \lambda B| = 0 \). If \( |B| \neq 0 \), then \( B^{-1} \) exists, in which case \( x \) satisfies the equation
\[
x(AB^{-1} - \lambda I) = 0
\]
(1.c.5.1)

and \( \lambda \) is the latent root of the matrix \( AB^{-1} \). The determination of the vectors satisfying (1.c.5.1) thus reduces to the case considered in the previous section.

(i) If \( \lambda_1 \) and \( \lambda_2 \) are two different roots, then
\[
\begin{align*}
x_1 AB^{-1} - \lambda_1 x_1 &= 0 \\
x_2 A - \lambda_1 x_2 B &= 0 \\
x_2 A - \lambda_2 x_2 B &= 0
\end{align*}
\]
Multiplying the first equation by \( x_2' \) and the second by \( x_1' \) and subtracting,
\[
(\lambda_1 - \lambda_2)x_1 Bx_2' = 0 \quad \text{or} \quad x_1 Bx_2' = 0
\]
(ii) If \( \lambda \) is a root of multiplicity \( r \), then the rank of \( (AB^{-1} - \lambda I) \) can be shown to be \( (n - r) \) as in the previous section. The number of independent vectors satisfying the equation \( x(AB^{-1} - \lambda I) = 0 \) is \( r \). The vectors in this set may be chosen such that any two vectors \( x \) and \( y \) satisfy the relation \( x By' = 0 \). Thus, corresponding to the \( n \) latent roots we obtain \( n \) vectors which may be represented by a matrix \( X_n \).
satisfying the condition that
\[ X_nBX_n' = C \]
a diagonal matrix
Let the leading diagonal of \( C \) contain the elements \( c_1, \ldots, c_n \).
(iii) If \( A_n \) denotes the diagonal matrix containing \( \lambda_1, \ldots, \lambda_n \) in the diagonal, then
\[ X_nABX_n' - A_nX_n = 0 \]
or
\[ X_nA - A_nX_nB = 0 \]
Multiplying by \( X_n' \)
\[ X_nAX_n' = A_nX_nB = A_nC \]
This shows that the transformation \( x = yX_n \) transforms the quadratic forms \( xAX' \) and \( xBY' \) into
\[ c_1\lambda_1y_1^2 + \cdots + c_n\lambda_ny_n^2 \]
and
\[ c_1y_1^2 + \cdots + c_ny_n^2 \]
(iv) Since \( c_i \) are positive the transformation \( \sqrt{c_i}y_i = Y_i \) carries the above quadratic forms to
\[ \lambda_1Y_1^2 + \cdots + \lambda_nY_n^2 \]
and
\[ Y_1^2 + \cdots + Y_n^2 \]
(v) If the quadratic form \( xAx' \) is never negative, then no \( \lambda \) is negative.
\textbf{Example 1.} If \( A \) and \( B \) are symmetric matrices such that \( B \) is positive definite and \( (A - B) \) is positive or semi-positive definite, then \( |A| \geq |B| \).
Consider the equation
\[ |A - B - \lambda B| = 0 \]
where no root is negative since \( (A - B) \) is positive or semi-positive definite and \( B \) is positive definite. Since
\[ |A - B(1 + \lambda)| = 0 \]
it follows that \( |A|/|B| = (1 + \lambda_1) \cdots (1 + \lambda_n) \), where \( \lambda_1, \lambda_2, \ldots, \lambda_n \) are the roots. None of the factors in the product \( (1 + \lambda_1) \cdots (1 + \lambda_n) \) is less than one so that
\[ |A|/|B| \geq 1 \]
which proves the result. This result also shows that the matrix \( A \) is also positive definite.

**Example 2.** The rank of the quadratic form \( \sum (x_i - \bar{x})^2 \), where \( \bar{x} \) is the average of \( x_1, x_2, \ldots, x_n \), is \( (n - 1) \). (Use example 4, la.4.)

**Example 3.** Consider the matrix \((x_{ij})\) of measurements, \( i = 1, 2, \ldots, p; j = 1, 2, \ldots, n \). If

\[
\sum_{i=1}^n x_{ij} = nz_i
\]

then the matrix \((S_{iu})\), \( (i, u = 1, 2, \ldots, p) \), is

(i) positive definite or semi-definite if \( n \geq p \);

(ii) positive semi-definite if \( n < p \); and

(iii) positive definite if the rank of \((x_{ij} - \bar{x})\) is equal to \( p \) and semi-definite if it is less than \( p \).

(Use examples 10 and 11, Ib.3.)

**Example 4.** The solution of

\[
\begin{cases}
\sum a_{li}d_l = 0, \\
\lambda = \sum x_{ij}d_l = dA^{-1}d'
\end{cases}
\]

where \( A = (a_{ij}) \) and \( d = (d_1, \ldots, d_p) \).

**Example 5.** If \( A \) and \( C \) are two matrices such that with respect to a symmetric definite matrix \( \Lambda \)

\[
\Lambda AC' = 0
\]

then the rows in \( A \) are independent of the rows in \( C \).

If any vector \( x \) in \( A \) is dependent on the vectors in \( C \), then \( xAx' = 0 \), which is impossible since \( \Lambda \) is a definite matrix.

**Example 6.** Consider the quadratic form in \( y_1, \ldots, y_t \)

\[
\sum_{i=1}^n (y_1 + d_1y_2 + d_2^2y_3 + \cdots + d_t'y_t)^2
\]

where \( d_1, \ldots, d_t \) are as defined in example 12 of ib.3, and deduce the result about the moments. When the variable \( x \) is continuous, the summation is replaced by integration. We thus deduce the consistency relations to be satisfied by the moments of any distribution.

**1c.6 Reduction of an Asymmetric Matrix**

In population studies it is often necessary to determine the powers of an asymmetric matrix called the generation matrix. If \( A \) represents
the generation matrix of the type \((k, k)\) and \(f_1, f_n\) represent the initial and the \(n\)th generation frequencies in some well-defined classes, then

\[ f_n = f_1 A^n \]

The calculations can be simplified by the following steps. Let \(\lambda_1, \ldots, \lambda_p\) be the distinct roots of the determinantal equation

\[ |A' - \lambda I| = 0 \]

with the multiplicity of the root \(\lambda_i\) being equal to \(m_i, \sum m_i = k\). Let it be possible to find \(m_i\) independent solutions \(x_{i1}, \ldots, x_{im_i}\) of the equation

\[ x_i (A' - \lambda_i I) = 0 \]  \hspace{1cm} (1c.6.1)

If \(P\) stands for the matrix containing the vectors \(x_{ij}\), then

\[ PA' = \lambda P \]

where \(A\) is the matrix containing the latent roots allowing repetitions and arranged in the same order as the corresponding vectors. If \(x_1, \ldots, x_p\) are any characteristic vectors corresponding to the distinct roots \(\lambda_1, \ldots, \lambda_p\), then they are all independent. If there were a relation

\[ \rho_1 x_1 + \cdots + \rho_p x_p = 0 \]  \hspace{1cm} (1c.6.2)

where the terms with zero coefficients are omitted, then from the relations \(x_i A' = \lambda_i x_i\), it follows that

\[ \lambda_1 \rho_1 x_1 + \cdots + \lambda_p \rho_p x_p = 0 \]  \hspace{1cm} (1c.6.3)

Using (1c.6.3), one variable in (1c.6.2) is eliminated and a new relation obtained. This gives rise to a relation similar to (1c.6.3). After repeated eliminations the relation (1c.6.2) reduces to a wrong result that one of the vectors is zero. Therefore no such relation as (1c.6.2) is true. From this it follows that \(P\) is a non-singular matrix in which case

\[ A = (P^{-1} A P)' = Q^{-1} \Lambda Q \]

where \(P' = Q^{-1}\). Hence

\[ A^2 = Q^{-1} \Lambda Q^{-1} \Lambda Q = Q^{-1} \Lambda^2 Q \]

and generally

\[ A^n = Q^{-1} \Lambda^n Q \]

so that an easy rule is provided once the \(P\) matrix is evaluated. Such a simple representation breaks down when it is not possible to find \(m_i\) independent solutions of (1c.6.1) where \(\lambda_i\) is a root of multiplicity \(m_i\).

It is seen that the linear functions of frequencies \(x_{ij} f^\prime\) transform in a
simple way from generation to generation. In fact
\[ x_i (t) = \lambda^* x_i (t) \] (1.6.4)
so that, knowing the initial values of these linear functions, the vector \( f_n \) can be solved from the equations (1.6.4).

1d Numerical Appendix

1d.1 The Evaluation of Determinants, Reciprocals, and Solutions of Equations

The theoretical expression for the value of the determinant given in 1b.3 is not convenient for practical computations. The method of pivotal condensation will be useful in (i) determining the value of a determinant, (ii) solving linear equations, and (iii) obtaining the reciprocal of a matrix. These three techniques are simultaneously illustrated in the example below. Only the relevant computations need be retained in any problem. In this illustration a non-symmetrical matrix is chosen. In most of the statistical computations symmetric matrices are met with. This results in a certain amount of reduction in computations, and the layout of a simplified procedure in such cases is discussed in the text.

References

Notes on the Computations in Table 14.1a

(i) Row 10 is obtained from row 01 by reducing the first element to unity. This is called a pivotal row and the first element, 0.1228 underlined in the table, is the first pivotal element. By multiplying row 10 by 0.1281 and subtracting from row 02, row 11 is obtained in which the first element is zero. Similarly, by eliminating the first element in rows 03 and 04, rows 12 and 13 are obtained. Starting with the reduced matrix in rows 11, 12, and 13, the whole operation is repeated.

(ii) Rows 10, 20, 30, and 40 are pivotal rows, and the product of the four elements

\[0.12280 \times 0.18719 \times 0.21091 \times 0.001625 = 0.00018625\]

is the value of the determinant of the matrix of equations. If only the value of the determinant is needed, the calculations beyond column 4 should be omitted.

(iii) Calculations up to column 5 give the solutions of non-homogeneous equations, and columns 6, 7, 8, and 9 are intended for the solutions of non-homogeneous equations whose right-hand elements are columns of a unit matrix. The four sets of solutions will form the elements of the reciprocal matrix.

(iv) The pivotal row 40 gives the solutions for \(x_5\) and, by substituting this value of \(x_5\) in row 30, row 30' giving the solutions for \(x_3\) is obtained. By substituting the values of \(x_3\) and \(x_4\) in row 20, row 20' is obtained, and so on. Thus the last four rows in the reverse order (in columns 6, 7, 8, and 9) give the reciprocal matrix.

(v) To start, the matrix has only four significant figures. But it is better to keep more places (usually one more) in subsequent calculations to keep a check on the rounding-off errors.

(vi) A sum check provided in the last column ensures accuracy of all the calculations. To start, the sum of the elements in each of the initial rows (here four) is written in the last column. For subsequent operations this is treated as an extra column. Any derived row has the property that the last element is the sum of the other elements. This may be checked whenever a new row is obtained, either by reducing the first coefficient to unity or by sweeping out a first column.
## Table 1d.1.a. Computations for the Evaluation of a Determinant, Reciprocal Matrix, etc.

<table>
<thead>
<tr>
<th>Row No.</th>
<th>Matrix of Linear Equations</th>
<th>Right-Hand Elements</th>
<th>Unit Matrix for the Reciprocal</th>
<th>Sum Check</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>01</td>
<td>0.1228</td>
<td>-0.0508</td>
<td>0.1434</td>
<td>0.2016</td>
</tr>
<tr>
<td>02</td>
<td>0.1261</td>
<td>0.1342</td>
<td>0.0196</td>
<td>0.4703</td>
</tr>
<tr>
<td>03</td>
<td>0.0434</td>
<td>0.2172</td>
<td>0.1034</td>
<td>0.1056</td>
</tr>
<tr>
<td>04</td>
<td>0.2023</td>
<td>0.1523</td>
<td>0.2045</td>
<td>0.5862</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>-0.4136</td>
<td>1.16775</td>
<td>1.64169</td>
</tr>
<tr>
<td>11</td>
<td>0.18719</td>
<td>-0.12999</td>
<td>0.26000</td>
<td>0.20908</td>
</tr>
<tr>
<td>12</td>
<td>0.23515</td>
<td>0.06272</td>
<td>0.03436</td>
<td>1.04327</td>
</tr>
<tr>
<td>13</td>
<td>0.23399</td>
<td>-0.03173</td>
<td>0.52469</td>
<td>0.56526</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>-0.69443</td>
<td>1.3898</td>
<td>1.11160</td>
</tr>
<tr>
<td>21</td>
<td>0.21601</td>
<td>-0.20225</td>
<td>1.30466</td>
<td>0.55780</td>
</tr>
<tr>
<td>22</td>
<td>0.31215</td>
<td>0.1963</td>
<td>-0.30292</td>
<td>0.33228</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>-1.3525</td>
<td>1.30891</td>
<td>4.43094</td>
</tr>
<tr>
<td>31</td>
<td>0.37510</td>
<td>1.10108</td>
<td>-0.91776</td>
<td>-0.49218</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>·</td>
<td>·</td>
<td>·</td>
</tr>
<tr>
<td>30'</td>
<td>1</td>
<td>·</td>
<td>·</td>
<td>·</td>
</tr>
<tr>
<td>20'</td>
<td>1</td>
<td>·</td>
<td>·</td>
<td>·</td>
</tr>
<tr>
<td>20'</td>
<td>1</td>
<td>·</td>
<td>·</td>
<td>·</td>
</tr>
</tbody>
</table>
CHAPTER 2

Theory of Distributions

2a Some Analytical Methods in Distribution Problems

2a.1 Binomial Distribution

Binomial distribution is the simplest problem in the theory of distribution. The observations consist of \( n \) independent stochastic variables, each of which can assume two values, 1 and 0, with probabilities \( p \) and \( q \), \( (p + q = 1) \). If we attach the value 1 to success and 0 to failure in a trial, then the sum of observations in \( n \) independent trials gives the total number of successes. What is the probability distribution of the number of successes in \( n \) trials?

If we denote by \( r \) the number of successes, then

\[
r = x_1 + x_2 + \cdots + x_n
\]

where \( x_i = 1 \) for success and 0 for failure. Since the events are independent, the probability for any series of \( x \) with \( r \), 1 and \( (n - r) \), 0 is \( p^r q^{n-r} \). To obtain the probability of \( r \) successes we have to sum up the probabilities corresponding to all possible series of \( x \) containing 1, \( r \) times and 0, \( (n - r) \) times. This number is \( \binom{n}{r} \), the number of combinations of \( r \) out of \( n \) things. Since each such series has the probability \( p^r q^{n-r} \), the total probability of \( r \) is \( \binom{n}{r} p^r q^{n-r} \). This is the \( r \)th term in the binomial expansion \( (p + q)^n \).

If \( n_1 \) trials give \( r_1 \) successes and an independent set of \( n_2 \) trials gives \( r_2 \) successes, the probability of success being the same for both sets, the probability of \( r = (r_1 + r_2) \) should be the same as that of \( r \) successes in \( n = (n_1 + n_2) \) trials. This can be formally derived in the following manner. The probability of \( r_1 \) and \( r_2 \) is

\[
P(r_1, r_2) = \binom{n_1}{r_1} p^{r_1} q^{n_1-r_1} \binom{n_2}{r_2} p^{r_2} q^{n_2-r_2}
\]

\[
= \binom{n_1}{r_1} \binom{n_2}{r_2} p^{r_1+r_2} q^{n_1+n_2-r_1-r_2}
\]

\[
= \binom{n_1}{r_1} \binom{n_2}{r_2} p^{r_1+r_2} q^{n_1+n_2-r_1-r_2}
\]

\[
= \binom{n_1}{r_1} \binom{n_2}{r_2} p^{r_1+r_2} q^{n_1+n_2-r_1-r_2}
\]
Therefore

\[
P(r = r_1 + r_2) = \sum_{r_1 + r_2 = r} P(r_1, r_2) = p^r q^{n-r} \cdot \sum_{r_1 + r_2 = r} \binom{n_1}{r_1} \binom{n_2}{r_2} = p^r q^{n-r} \binom{n}{r}
\]

Thus the sum of two independent binomial variates with the same \(p\) is also a binomial variate and not otherwise.

Corresponding to a probability distribution there is a distribution function which gives the probability of a variate's assuming a value less than or equal to an assigned value. If \(P(z)\) denotes the probability distribution, the corresponding distribution function will be denoted by \(F(z)\).

In the binomial case

\[
F(r) = \sum_{s=0}^{r} P(s) = \sum_{s=0}^{r} \binom{n}{s} p^s q^{n-s} = \sum_{s=0}^{r} \frac{n!}{s!(n-s)!} p^s q^{n-s} = \frac{n!}{r!(n-r)!} \int_0^r \frac{(t) p^t q^{n-t} \cdot (1 - t)^{n-r-t}}{s!(r-s)!} dt
\]

Putting \(t = 1 - z/q\), the above expression becomes

\[
F(r) = \frac{n!}{r!(n-r)!} \int_0^z x^{n-r-1} (1 - x)^{r-1} dx
\]

This is an incomplete beta function which is extensively tabulated in Tables of Incomplete Beta Function (edited by K. Pearson). If \(r\) is small, then each term in the above summation can be separately calculated and added to obtain \(F(r)\).

Example 1. For a binomial distribution \((p + q)^n\),

\[
E\left(\frac{Y}{n}\right) = p, \quad V\left(\frac{Y}{n}\right) = \frac{pq}{n^2}, \quad \text{cov}\left(\frac{Y}{n}, \frac{n-r}{n}\right) = -\frac{pq}{n}
\]
THEORY OF DISTRIBUTIONS

where $E$ stands for expectation, $V$ for variance, and $\text{cov}$ for covariance.

Example 2. If $\mu_i$ is the $i$th corrected moment of the observed number of successes, then

$$\mu_{i+2} = pq \left| \frac{d}{dp} \mu_i + n(t + 1) \mu_i \right|$$

(Romanovsky, 1925)

Hence

$$\beta_i = \frac{\mu_i^2}{\mu_2} = \frac{(q - p)^2}{npq}$$

$$\beta_2 = \frac{\mu_4}{\mu_2} = 3 + \frac{1 - 6pq}{npq}$$

2a.2 Multinomial Distribution

If there are $k$ mutually exclusive events with probabilities $\pi_1, \pi_2, \cdots, \pi_k$, ($\sum \pi_i = 1$), then the probability of occurrence of $n_1$ events of the first kind, $n_2$ of the second, $\cdots$, etc., in a total of $n$ independent trials is

$$P(n_1, n_2, \cdots, n_k) = \frac{n!}{n_1! \cdots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k}$$

The product $\pi_1^{n_1} \cdots \pi_k^{n_k}$ refers to the probability of events occurring in some order; $n_1/n_1! \cdots n_k!$ represents the number of arrangements of $n_1$ things of one kind, $n_2$ of another kind, and so on. Therefore the total probability of the desired number of events of the various kinds is the product of these two expressions, the argument being similar to that used for the binomial. The above probability is a term of the multinomial expansion.

Since

$$\sum \frac{n!}{n_1! \cdots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k} = (\pi_1 + \cdots + \pi_k)^n$$

Thus

$$\pi_j \sum \frac{n!}{n_1! \cdots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k} = \sum n_i P(n_1, \cdots, n_k)$$

$$= \pi_j \left( \frac{\pi_j}{\sum \pi_j} (\pi_1 + \cdots + \pi_k)^n \right)^{k-1}$$

$$= \pi_j \left( \frac{\pi_j}{\sum \pi_j} (\pi_1 + \cdots + \pi_k)^{k-1} \right)^{n-1}$$

$$= n(n-1) \pi_j (\pi_1 + \cdots + \pi_k)^{n-2}$$
THE POISSON DISTRIBUTION

Similarly

\[ \frac{\partial}{\partial x_i} \sum n_i P(n_1, \ldots, n_k) = \sum \pi_i \frac{\partial}{\partial x_i} P(n_1, \ldots, n_k) \]

\[ = n \pi_i \pi_i + \sum (n-1) \pi_i \pi_i(x_1 + \cdots + x_k) - 2 \]

These results give

\[ E(n_i) = n \pi_i \]

\[ \text{Cov}(n_j, n_j) = E(n_j^2) - E(n_j)E(n_j) = n \pi_i^2 \]

Therefore

\[ V(n_i) = E(n_i^2) - [E(n_i)]^2 = n \pi_i(1 - \pi_i) \]

\[ \text{cov}(n_i, n_j) = E(n_i n_j) - E(n_i)E(n_j) = -n \pi_i \pi_j \]

which are similar to the results obtained for the binomial. From these results the variances and covariances of linear functions of frequencies in \( k \) classes can be obtained in a simple way.

\[ V(l_1 n_1 + \cdots + l_k n_k) = \sum l_i^2 V(n_i) + 2 \sum l_i \text{cov}(n_i, n_j) \]

\[ = \sum l_i^2 n \pi_i (1 - \pi_i) + 2 \sum l_i (n \pi_i \pi_j) \]

Similarly

\[ \text{cov} [(l_1 n_1 + \cdots + l_k n_k) + m_1 n_1 + \cdots + m_k n_k] = n [\sum l_i m_i \pi_i - (\sum l_i)(\sum m_i \pi_i)] \]

Higher moments of the multinomial can be derived by extending the differentiation processes.

2a.3 The Poisson Distribution

This is a discrete distribution where the stochastic variable assumes values 0, 1, 2, \ldots, with the probability for \( r \) equal to

\[ \frac{e^{-\mu} \mu^r}{r!} \]
where $\mu$ is a parameter. Since,

$$e^\mu = \sum_{r=0}^{\infty} \frac{\mu^r}{r!}$$

and

$$\sum_{r=0}^{\infty} \frac{\mu^r}{r!} = \mu \frac{d}{dm} e^\mu = \mu e^\mu$$

$$\sum_{r=0}^{\infty} \frac{\mu^r}{r!} = \mu \frac{d}{dm} (\mu e^\mu) = \mu^2 e^\mu + \mu e^\mu$$

These give

$$E(r) = \mu \quad V(r) = \mu$$

so that the mean and variance of this distribution are equal. The higher moments can be derived from the relation

$$\mu_{r+1} = \mu \mu_{r-1} + \mu \frac{d\mu}{dm}$$

obtained in a manner similar to the corresponding relation in the binomial distribution. This gives

$$\mu_3 = \mu \quad \mu_4 = \mu + 3\mu^2 \quad \text{etc.}$$

Hence

$$\beta_1 = 1 - \frac{1}{\mu} \quad \beta_2 = 3 + \frac{1}{\mu}$$

If $r_1, r_2, \ldots, r_k$ are independent variates from the same Poisson distribution,

$$P(r_1, r_2, \ldots, r_k) = e^{-\mu} \frac{\mu^{r_1+r_2+\cdots+r_k}}{r_1!r_2! \cdots r_k!}$$

$$P(r = r_1 + r_2 + \cdots + r_k) = e^{-\mu} \frac{\mu^{r_1+r_2+\cdots+r_k}}{r_1!r_2! \cdots r_k!}$$

$$= e^{-\mu} \frac{\mu^{r}}{r!} = e^{-\mu} \frac{(k\mu)^r}{r!}$$

which shows that the sum of $k$ Poisson variates, each with parameter $\mu$, is itself a Poisson variate with the parameter $k\mu$.

The conditional distribution

$$P(r_1, r_2, \ldots, r_k \mid r) = \frac{P(r_1, r_2, \ldots, r_k)}{P(r)}$$

$$= \frac{r!}{r_1! \cdots r_k!} \left( \frac{1}{k} \right)^{r_1} \cdots \left( \frac{1}{k} \right)^{r_k}$$
THE POISSON DISTRIBUTION

is a multinomial with index $r$ and probability in each class equal to $1/k$.

If $r_1, \ldots, r_k$ are Poisson variates with parameters $\mu_1, \ldots, \mu_k$, then

$$P(r_1, \ldots, r_k) = \frac{e^{-r} \sum_{i=1}^{k} \frac{r_i^{\mu_i}}{\mu_i!}}{r_1! \cdots r_k!}$$

$$P(r) = r_1 + \cdots + r_k = \frac{e^{-r} \sum_{i=1}^{k} \frac{r_i^{\mu_i}}{\mu_i!}}{r_1! \cdots r_k!}$$

where $\mu = \mu_1 + \cdots + \mu_k$, which shows that the sum of $k$ independent Poisson variates is in general a Poisson variate. This is true in the case of a binomial variate only when the binomial proportions are the same.

The conditional distribution

$$P(r_1, \ldots, r_k | r) = \frac{P(r_1, \ldots, r_k)}{P(r)}$$

$$= \frac{r!}{r_1! \cdots r_k!} \left( \frac{\mu_1}{\mu} \right)^{r_1} \cdots \left( \frac{\mu_k}{\mu} \right)^{r_k}$$

is multinomial with probabilities $\mu_1/\mu, \ldots, \mu_k/\mu$ in the $k$ classes.

In general, any multinomial distribution

$$\frac{n!}{n_1! \cdots n_k!} r_1^{n_1} \cdots r_k^{n_k}$$

can be written as the ratio of

$$P(n_1, \ldots, n_k) = \frac{e^{-nr_1} (nr_1)^{n_1}}{n_1!} \cdots \frac{e^{-nr_k} (nr_k)^{n_k}}{n_k!}$$

to

$$P(n = n_1 + \cdots + n_k) = \frac{e^{-rn} (rn)^n}{n!}$$

which is the relative probability of $k$ Poisson variates with parameters $\mu_i = n \mu_i$, ($i = 1, 2, \ldots, k$), subject to the condition that the sum of the variables is equal to $n$. 
The distribution function for the Poisson distribution is obtained below.

\[ F(r) = e^{-\lambda} \left( 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^r}{r!} \right) \]

\[ = \sum_{k=0}^{r} \frac{e^{-\lambda} \lambda^k}{k!} \int_{0}^{\infty} e^{-x} x^k dx \]

\[ = \frac{e^{-\lambda}}{r!} \int_{0}^{\infty} e^{-x} (\mu + x)^r dx \]

\[ = \frac{1}{r!} \int_{0}^{\infty} e^{-y} y^r dy \]

which is the incomplete gamma integral, tables for which have been edited by K. Pearson.

The Poisson probability can be deduced from the binomial when \( n \), the number of trials, is large and \( p \) is small. For instance, the probability for no success is

\[ P(0) = q^n = (1 - p)^n = \left( 1 - \frac{np}{n} \right)^n \sim e^{-\mu} = e^{-\lambda} \]

\[ P(r + 1) \]

\[ = \frac{\binom{n}{r + 1} p^{r + 1} q^{n - r - 1}}{\binom{n}{r} p^r q^{n - r}} \]

\[ = \frac{n - r p}{r + 1} \frac{\mu}{r + 1} \sim \frac{\mu}{r + 1} \]

where \( \mu = np \). The successive terms of the binomial then tend to

\[ e^{-\lambda}, e^{-\lambda} \frac{\mu}{1!}, e^{-\lambda} \frac{\mu^2}{2!}, \ldots \]

yielding the Poisson series. Thus, when the probability is small and the number of trials is indefinitely large, the Poisson distribution may be used.

2a.4 Normal Distribution

This is a continuous distribution with the probability differential of the stochastic variable \( x \) equal to

\[ N(\mu, \sigma) dx = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \]
The $r$th moment of this distribution is
\[
\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} (x - \mu)^r e^{-\frac{(x - \mu)^2}{2\sigma^2}} dx
\]
\[
= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} y^r e^{-\frac{y^2}{2\sigma^2}} dy \quad \text{putting } y = x - \mu
\]
\[
= 0 \quad \text{if } r \text{ is odd}
\]
\[
= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} z^{(r-1)/2} e^{-\frac{z}{2\sigma^2}} dz \quad \text{putting } z = y^2 \text{ if } r \text{ is even}
\]
\[
= \frac{(2\sigma^2)^{(r+1)/2}}{\sigma \sqrt{2\pi}} I_{(r+1)/2} \left( \frac{r+1}{2} \right) = \sigma^r (r-1)(r-3) \cdots (1)
\]
whence we have the following results:
\[
\mu_2 = \sigma^2 \quad \mu_1 = 0 \quad \mu_4 = 3\sigma^4
\]
\[
\beta_1 = 0 \quad \beta_2 = 3
\]

If $x$ and $y$ are two independent normal variates with mean values $m_1$ and $m_2$ and standard deviations $\sigma_1$ and $\sigma_2$, then their joint distribution is
\[
\text{const. } e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}}
\]
where
\[
Q(x, y) = \frac{(x-m_1)^2}{\sigma_1^2} + \frac{(y-m_2)^2}{\sigma_2^2}
\]
\[
= \frac{[(x-m_1)+(y-m_2)]^2}{\sigma_1^2+\sigma_2^2} + \frac{(x-m_1)-(y-m_2)]^2}{\sigma_1^2+\lambda^2\sigma_2^2}
\]
where $\sigma_1^2 = \lambda\sigma_2^2$. Make the transformation
\[
u = x + y \quad \nu = x - \lambda y
\]
so that the joint distribution of $u$ and $v$ becomes
\[
\text{const. } e^{-\frac{Q(u,v)}{2}} du dv
\]
where
\[
Q(u, v) = \frac{(u-m_1-m_2)^2}{\sigma_1^2+\sigma_2^2} + \frac{(v-m_1+\lambda m_2)^2}{\sigma_1^2+\lambda^2\sigma_2^2}
\]
The distributions of $u$ and $v$ are independent as their joint distribution
turns out to be a product of two functions. The distribution of \( u \) is
\[
\text{const.} \ e^{-\frac{1}{2}(u-m_1-m_2)^2/(\sigma_1^2+\sigma_2^2)} du
\]
which shows that \( u = (x + y) \), the sum of two normal variates, is itself a normal variate with mean equal to the sum of the means and variance equal to the sum of the variances. In general, the sum of \( k \) normal variates is distributed as a normal variate with mean equal to the sum of the individual means and variance equal to the sum of the individual variances.

**Example 1.** If \( x \) is \( N(\mu, \sigma) \), what is the distribution of \( x^2 \)?

The distribution of \( x \) is
\[
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
\]
Let \( y = x^2 \) so that \( dy = 2x \, dx \). The range of \( y \) is from 0 to \( \infty \), and that of \( x \) is from \(-\infty \) to \( \infty \). Corresponding to a given \( y \) there are two values of \( x \) (\( \pm \sqrt{y} \)). The probability density at \( +x \) transforms to
\[
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-2\sqrt{\mu^2+\sigma^2})}{2\sigma^2}} \frac{dy}{2\sqrt{y}}
\]
and that at \(-x \) to
\[
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y+2\sqrt{\mu^2+\sigma^2})}{2\sigma^2}} \frac{dy}{2\sqrt{y}}
\]
so that the total probability differential of \( y \) is the sum of the above two expressions, i.e.,
\[
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y\sqrt{\mu^2+\sigma^2}}{2\sigma^2}} \frac{(e^{\frac{y\sqrt{\mu^2+\sigma^2}}{2\sigma^2}} + e^{-\frac{y\sqrt{\mu^2+\sigma^2}}{2\sigma^2}})}{2\sqrt{y}} dy = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y\sqrt{\mu^2+\sigma^2}}{2\sigma^2}} \cosh \frac{\sqrt{y}\mu}{\sigma} \frac{dy}{\sqrt{y}}
\]
If \( \mu = 0 \), the distribution of \( y \) is
\[
\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y\sigma^2}{2\sigma^2}} \frac{dy}{\sqrt{y}}
\]
which is \( G(1/2\sigma^2, \frac{1}{2}) \) defined in 2a.5.

**2a.5 Gamma Distribution**

The gamma distribution is defined by
\[
G(\alpha, \beta) \, dz = \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\alpha z} z^{\beta-1} dz
\]
BETA DISTRIBUTION

where the range of $x$ is $(0, \infty)$, $\alpha > 0$, and $\gamma \geq 1$. The $r$th raw moment of this distribution is

$$
\int_0^\infty x^r e^{-\alpha x} \beta^\gamma dx = \frac{\Gamma(p + r)}{\Gamma(p)} \frac{1}{\alpha^r}
$$

so that

$$
E(x) = \frac{\gamma}{\alpha}, \quad V(x) = \frac{\gamma}{\alpha^2}
$$

If $\alpha = 1$, the mean and the variance of the gamma variate are equal, as in the case of a Poisson variate.

Let $x$ and $y$ be two independent gamma variates with parameters $(\alpha, \beta)$ and $(\alpha, \theta)$; then their joint distribution is

$$
\text{const. } e^{-\alpha x + \theta y} x^{\beta-1} y^{\theta-1} dx dy
$$

Put $x = r \cos \theta$ and $y = r \sin \theta$, so that $dx dy = 2r \cos \theta \sin \theta dr d\theta$.

The distribution of $r$ and $\theta$ is

$$
\text{const. } e^{-\alpha r + \theta y} (\cos \theta)^{2\beta-1} (\sin \theta)^{2\theta-1} dr d\theta
$$

and that of $r$ alone is $e^{-\alpha r + \theta y} dr$, which is again a gamma variate with parameters $(\alpha, \beta + \theta)$. In general the sum of $k$ gamma variates with parameters $(\alpha, \beta_1), (\alpha, \beta_2), \ldots, (\alpha, \beta_k)$ is distributed as a gamma variate with parameters $(\alpha, \beta_1 + \beta_2 + \ldots + \beta_k)$. The distribution of $z = x/(x+y)$ is that of $\cos^2 \theta$, i.e.,

$$
\frac{1}{\beta(p, q)} e^{p-1(1-z)^{p-1}} dz \quad \beta(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}
$$

which is $B(p, q)$ defined in 2a.6. A special case of the gamma distribution is the $\chi^2$ distribution,

$$
\chi^2(k) = \text{const. } e^{-x^2/2} (\chi^2)^{k/2-1/2} dx^2 = G \left( \frac{1}{2}, k \right)
$$

which is specified by one parameter $k$ known as the degrees of freedom of $\chi^2$.

2a.6 Beta Distribution

A stochastic variable in the range $(0, 1)$ having the probability distribution

$$
B(a, b) dx = \frac{1}{\beta(a, b)} x^{a-1}(1-x)^{b-1} dx
$$

is said to follow the beta distribution with parameters $a$ and $b$. The $r$th
moment about the origin is

$$\frac{1}{\beta(a, b)} \int x^{a+r-1}(1 - x)^{b-1} dx = \frac{\beta(a + r, b)}{\beta(a, b)}$$

The mean and variance of $x$ are

$$\frac{a}{a + b} \quad \text{and} \quad \frac{ab}{(a + b)^2(a + b + 1)}$$

If $x$ and $y$ are two independent beta variates with parameters $(a, b)$ and $(c, d)$, then their joint distribution is

$$\text{const. } x^{a-1}(1 - x)^{b-1} y^{c-1}(1 - y)^{d-1} dx \, dy$$

Put $u = x$ and $z = xy$, so that $\partial(x, y)/\partial(u, z) = 1/u$. The joint distribution of $u$ and $z$ is

$$\text{const. } u^{a-c-d}(1 - u)^{b-1} z^{c-1}(u - z)^{d-1} du \, dz$$

$$= \text{const. } (1 - u)^{b-1} z^{c-1}(u - z)^{d-1} du \, dz \quad \text{if } a = c + d$$

Integrating over $u$, the distribution of $z$ is obtained as

$$\frac{1}{\beta(c, b + d)} z^{c-1}(1 - z)^{b+d-1} dz$$

This shows that the product of two beta variables, with parameters $(a, b)$ and $(c, d)$ such that $a = (c + d)$, is distributed as a beta variable with parameters $(c, b + d)$. In general, the product of beta variables, with parameters $(a_1, b_1), (a_2, b_2), \cdots (a_k, b_k)$ such that $a_i = (a_{i+1} + b_{i+1})$, is distributed as a beta variable with parameters $(a_k, b_1 + \cdots + b_k)$.

### 2a.7 Cauchy Distribution

The Cauchy distribution is defined by

$$\frac{1}{\pi} \frac{dx}{1 + (x - \mu)^2}$$

where $x$ ranges from $-\infty$ to $\infty$. This is a symmetrical distribution with the modal value at $x = \mu$.

$$E(x) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x \, dx}{1 + (x - \mu)^2}$$

This integral does not exist but has the principal value $\mu$ since

$$\lim_{\lambda \to \infty} \pi \int_{-\lambda}^{+\lambda} \frac{x \, dx}{1 + (x - \mu)^2} \quad \text{exists}$$
The second moment

\[ E(x^2) = \int_{-\infty}^{\infty} x^2 \, dx \]

so that the Cauchy variable has infinite variance. This is an example of a continuous distribution for which the mean and variance do not exist.

Consider two independent Cauchy variates \( x \) and \( y \) with parameters \( \mu_1 \) and \( \mu_2 \). Their joint distribution is

\[ \frac{1}{\pi^2 \left[ 1 + (x - \mu_1)^2 \right] \left[ 1 + (y - \mu_2)^2 \right]} \]

Putting \( u = x - \mu_1 \) and \( v = y - \mu_2 \), we find the distribution of \((u + v)/2\) and hence derive that of \((x + y)/2\) by the substitution \( u + v = (x + y - \mu_1 - \mu_2) \). Transforming from \( u, v \) to \( u, z \) connected by the relations \( u + v = 2z \), the joint distribution of \( u \) and \( z \) is given by

\[ \text{const.} \frac{du \, dz}{\left[ 1 + u^2 \right] \left[ 1 + (2z - u)^2 \right]} \]

Integrating term by term with respect to \( u \) from \(-\infty\) to \( \infty \), we obtain

\[ \text{const.} \frac{dz}{(4z^2 + 4)z} \left[ 2z \log (1 + u^2) - 2z \log (1 + (2z - u)^2) \right] \]

The distribution of \( z = (x + y)/2 \) obtained by changing \( z \) to \( z - (\mu_1 + \mu_2)/2 \) in the above expression is

\[ \frac{1}{\pi} \frac{dz}{1 + \left( z - \frac{\mu_1 + \mu_2}{2} \right)^2} \]

which shows that the mean of two Cauchy variables with parameters \( \mu_1 \) and \( \mu_2 \) is distributed as a Cauchy variable with the parameter equal to the mean of the two parameters. In general, the average of \( k \) Cauchy variables with parameters \( \mu_1, \ldots, \mu_k \) is distributed as a Cauchy variable
with the parameter \((\mu_1 + \cdots + \mu_k)/k\). If all \(\mu\) are the same, we obtain the interesting result that the distributions of a single observation and that of the mean of any number of observations are the same.

2a.8 Pearson’s \(F\) Distribution

If a stochastic variable \(X\) has the probability density \(f(x)\), then the variable

\[
\int_{-\infty}^{x} f(z) \, dz
\]

which represents the probability of an observation’s being less than or equal to \(x\), has probability density unity. Since

\[
dy = \frac{d}{dx} \int_{-\infty}^{x} f(z) \, dz = f(x) \, dx
\]

\(f(z) \, dz\) transforms to \(dy\) with the range of \(y\) equal to \((0, 1)\).

Let \(z = -2 \log_{e} y\), in which case \(dz = -(2/y) \, dy\). Since the probability density of \(y\) is \(dy\), that of \(z\) is

\[
y = \frac{1}{2} e^{-z/2} \, dz
\]

which is \(G(y; \frac{1}{2}, 1)\).

If \(y_1, y_2, \ldots, y_k\) are \(k\) probabilities derived from \(k\) independent observations \(x_1, \ldots, x_k\) from \(k\) distributions, all of which may be different from one another, then their joint distribution is

\[
dy_1 \cdots dy_k
\]

If \(Z_i = -2 \log_{e} y_i\), then \(Z_i\) is \(G(y; \frac{1}{2}, 1)\) for all \(i\). If \(P_k\) is defined by \(Z_1 + \cdots + Z_k\), then \(P_k\), being the sum of \(k\) gamma variates with parameters \((\frac{1}{2}, 1)\), is distributed as \(G(y; \frac{1}{2}, k)\) or \(\chi^2(2k)\). This distribution is useful in combining several independent tests.

2a.9 Summary of Results

Some of the important results of this section have been brought together for later use.

(i) If \(n_1, \ldots, n_k\) \((\Sigma n_k = n)\), are the frequencies in \(k\) mutually exclusive classes with probabilities \(\pi_1, \ldots, \pi_k\), then, as shown in 2a.2,

\[
E(n_i) = n \pi_i
\]

\[
V(n_1 + \cdots + n_k) = n(\Sigma \pi_i^2 - (\Sigma \pi_i)^2)
\]

\[
cov [(n_1 + \cdots + n_k), (m_1 + \cdots + m_k)] = n(\Sigma \pi_i \pi_j - (\Sigma \pi_i)(\Sigma \pi_j))
\]

\(2a.9.1\)
(i) Defining as in 2a.4,

\[ N(\mu, \sigma, x) \, dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \]

\[ \int_{-\infty}^{\infty} N(\mu_1, \sigma_1, x_1) \cdots N(\mu_n, \sigma_n, x_n) \, dx_1 \cdots dx_n = N(\mu, \sigma, x) \, dx \]  

(2a.9.2)

where \( \mu = (\mu_1 + \cdots + \mu_n)/n \), \( \sigma^2 = (\sigma_1^2 + \cdots + \sigma_n^2)/n^2 \). The symbol \( \int \) is used to indicate that the integration is over constant values of \( x \).

The distribution of \( y = x^2 \), where \( x \) is \( N(\mu, \sigma) \), is

\[ \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} \cosh \left( \frac{\sqrt{\mu^2 + \sigma^2}}{\sqrt{y}} \right) \, dy \]

\[ = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}} y^{1/2} \left( \frac{1}{2\sigma^2} y + \frac{1}{4!} \left( \frac{\mu}{\sigma^2} \right)^2 y^2 + \cdots \right) dy \]

\[ = e^{-\frac{\mu^2}{2\sigma^2}} \left( \frac{\mu^2}{2\sigma^2} y \right)^{1/2} \frac{1}{\Gamma(1/2)} \left( \frac{1}{2\sigma^2} y + \frac{1}{2} \right) \, dy \]  

(2a.9.3)

(iii) Defining

\[ G(a, p, x) \, dx = \frac{x^p}{\Gamma(p)} e^{-x^2} \, dx \]

the results obtained in 2a.5 are

\[ \int_{-\infty}^{\infty} G(a, p_1, x_1)G(a, p_2, x_2) \cdots dx_1 \, dx_2 \cdots dx_n = G(a, p_1 + \cdots + p_n, x) \, dx \]  

(2a.9.4)

Also

\[ \int_{-\infty}^{\infty} G(a, p, x)G(a, q, y) \, dx \, dy = B(p, q, x) \, dx \]  

(2a.9.5)

and

\[ \int_{-\infty}^{\infty} G(a, p, x)G(a, q, y) \, dx \, dy = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} f^{p-1} \, df \]  

(2a.9.6)

(iv) Defining as in 2a.6,

\[ B(a, b, x) \, dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \, dx \]

\[ \int_{0}^{1} B(a, b, x)B(c, d, y) \, dx \, dy = B(c, b + d, z) \, dz \]  

(2a.9.7)

if \( a = c + d \).
\[ \int_{x_1, \ldots, x_n} B(a_1, b_1, x_1) B(a_2, b_2, x_2) \cdots dx_1 \cdots dx_n = B(a_n, b_1 + \cdots + b_n, z) \, dz \quad (2a.9.8) \]

provided that \( a_1 = a_2 + b_2, a_2 = a_3 + b_3, \ldots \).

2b Distributions Relating to the Univariate Normal Distribution

2b.1 Mean and Variance in Normal Samples

Let \( x_1, \ldots, x_n \) be \( n \) independent observations from a normal population \( N(\mu, \sigma) \). The probability density is

\[
\phi = \Sigma (x_1 - \mu)^2 = n(x - \mu)^2 + S^2
\]

\[
\bar{x} = \frac{(x_1 + \cdots + x_n)}{n}
\]

\[
S^2 = \Sigma (x_i - \bar{x})^2
\]

Consider an orthogonal transformation from \( z \) to \( z' \):

\[
z_1 = \frac{(x_1 + \cdots + x_n)}{\sqrt{n}} = \sqrt{n} \bar{x}
\]

\[
z_i = a_i x_1 + \cdots + a_i x_n \quad i = 2, \ldots, n
\]

Since the vectors of \( z_1 \) and \( z_i \) \( (i \neq 1) \), are orthogonal, \( \sum a_{ij} = 0 \), for \( i = 2, \ldots, n \). Under this transformation

\[
x_1, \ldots, x_n \sim z_1, \ldots, z_n
\]

\[
\mu, \ldots, \mu \sim \sqrt{n} \mu, 0, \ldots, 0
\]

and, as shown in (iv) of 1c.2 (invariance of distance),

\[
\Sigma (x_i - \mu)^2 = (z_1 - \sqrt{n} \mu)^2 + z_2^2 + \cdots + z_n^2
\]

\[
= n(x - \mu)^2 + S^2
\]

Therefore \( S^2 = z_2^2 + \cdots + z_n^2 \). The distribution of \( z_1 \cdots z_n \) is

\[
\text{const.} \, e^{-(z_1^2/n + z_2^2 + \cdots + z_n^2)/2} \, dz_1 \cdots dz_n
\]

which shows that the distributions of \( z_1 \) and \( z_2, \ldots, z_n \) are independent and hence those of \( \bar{x} = z_1/\sqrt{n} \) and \( S^2 = z_2^2 + \cdots + z_n^2 \). The distribu-
Since $z_i^2$ is $G(1/2 \sigma^2, \frac{1}{2})$ for $i \geq 2$, it follows from (2a.9.3) and (2a.9.4) that $z_2^2 + \cdots + z_n^2 = S^2$ is distributed as $G(1/2 \sigma^2, (n - 1)/2)$, i.e.,
\[
\int const. \cdot e^{-S^2/(2\sigma^2)} dS^2 \quad \text{const.} = \frac{1}{\sqrt{2\pi \sigma}}
\]
\[
\text{Since } z_i^2 \text{ is } G(1/2 \sigma^2, \frac{1}{2}) \text{ for } i \geq 2, \text{ it follows from (2a.9.3) and (2a.9.4) that } z_2^2 + \cdots + z_n^2 = S^2 \text{ is distributed as } G(1/2 \sigma^2, (n - 1)/2), \text{ i.e.,}
\]
\[
\int const. \cdot e^{-S^2/(2\sigma^2)} dS^2 \quad \text{const.} = \frac{1}{\sqrt{2\pi \sigma}}
\]
\[
\text{or that } \chi^2 = S^2/\sigma^2 \text{ is } \chi^2(n - 1).
\]

2b.2 Student's Distribution

The joint distribution of $\sqrt{n}(\bar{x} - \mu)$ and $S^2$ is
\[
\cdots N(0, \sigma^2) \cdot d\sqrt{n}(\bar{x} - \mu) \cdot G \left( \frac{1}{2\sigma^2}, \frac{n - 1}{2} \right) dS^2
\]
Student's $t$ statistic is
\[
t = \frac{\sqrt{n}(\bar{x} - \mu)}{s}
\]
where $s^2 = S^2/(n - 1)$ or, squaring and rearranging,
\[
t^2 = \frac{n(\bar{x} - \mu)^2}{s^2} = f
\]
$n(\bar{x} - \mu)^2$ is distributed as $G(1/2 \sigma^2, \frac{1}{2})$, so, from (2a.9.6), the distribution of $f$ is
\[
\frac{\Gamma \left( \frac{n}{2} \right)}{\Gamma \left( \frac{1}{2} \right) \Gamma \left( \frac{n - 1}{2} \right)} f^{-\frac{1}{2}} (1 + f)^{-n/2} df
\]
and hence that of $t$, which is symmetrically distributed, is
\[
\frac{1}{\sqrt{n - 1}} \frac{\Gamma \left( \frac{n}{2} \right)}{\Gamma \left( \frac{1}{2} \right) \Gamma \left( \frac{n - 1}{2} \right)} \left( 1 + \frac{t^2}{n - 1} \right)^{-n/2} dt
\]
This is called Student's *t* distribution based on \((n - 1)\) degrees of freedom. When \(n = 2\), this reduces to Cauchy distribution.

**Non-Null *t* Distribution.** To find the distribution of \(t^2 = \frac{n \hat{x}^2}{s^2}\) when \(\mu \neq 0\), we note that the joint density of \(S^2\) and \(y = ns^2\) is (see 2a.9.3)

\[
e^{-n y / 2s^2} G\left(\frac{1}{2s^2}, \frac{n - 1}{2}, S^2\right) \sum_{j=0} G\left(\frac{1}{2s^2}, r + \frac{1}{2}, y\right)
\]

(2b.2.2)

Making use of (2a.9.6) for each term of the infinite series we find the distribution of \(y / S^2\) to be

\[
e^{-n y / 2s^2} \frac{1}{r!} \frac{1}{2s^2} \frac{r}{(n + r)} \frac{r}{(r + \frac{1}{2})} \frac{n - 1}{2} (1 + f)^{(n/2) + r} df
\]

(2b.2.3)

where \(F_1\) is the hypergeometric function defined above. The distribution of \(t^2 = (n - 1)f\) can be obtained from this. Sometimes it is useful to use the distribution of \(R = S^2 / (y + S^2)\) which can be obtained by applying (2a.9.5) to each term in (2b.2.2).

\[
e^{-n y / 2s^2} \frac{1}{r!} \frac{1}{2s^2} \frac{r}{(n + r)} \frac{r}{(r + \frac{1}{2})} \frac{n - 1}{2} (1 + f)^{(n/2) + r} df\]

(2b.2.4)

**2b.3 Fisher's *z* Distribution**

If \(s_1^2\) and \(s_2^2\) are two independent estimates of variance based on \(n_1\) and \(n_2\) degrees of freedom, then \(S_1^2 = n_1 s_1^2\) and \(S_2^2 = n_2 s_2^2\) have the joint distribution

\[
G\left(\frac{1}{2s_1^2}, \frac{n_1}{2}, S_1^2\right) G\left(\frac{1}{2s_2^2}, \frac{n_2}{2}, S_2^2\right) dS_1^2 dS_2^2
\]

so that their ratio \(f = S_1^2 / S_2^2\) has the distribution

\[
\frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} f^{-1} \left(1 + f\right)^{(n_1 + n_2)/2} df
\]
The distribution of $F = s_1^2/s_2^2 = n_1/n_2$ is

\[ \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{e^{n_1/2 - 1}}{1 + \frac{n_1}{n_2} F^{n_1/2}} \]

This is called the variance ratio distribution. Fisher defines

\[ z = \frac{1}{2} \log \frac{s_1^2}{s_2^2} = \frac{1}{2} \log F \]

so that the $z$ distribution is

\[ \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} e^{-z^2} \frac{e^{n_1/2}}{1 + \frac{n_1}{n_2} e^{2z}} \]

In practice it is convenient to use the $F$ distribution instead of $z$.

2b.4 Cochran's Theorem

Let $x_1, \ldots, x_n$ be $n$ normal variates with zero mean and unit variance. If $x_1^2 + \cdots + x_n^2 = q_1 + \cdots + q_k$, where $q_i$ is a quadratic form of rank $n_i$, then the necessary and sufficient condition that $q_1, q_2, \cdots$ are independently distributed as $\chi^2$ with $n_1, n_2, \cdots$ degrees of freedom is that $n = 2: n_i$.

Since $q_i$ is a quadratic form of rank $n_i$, it can be expressed as (see ii in 1c.2)

\[ \pm l_{ij}^2 \pm l_{ji}^2 + \cdots + l_{ii}^2 \]

where $l_{ij}$ is a linear function of $x_1, \cdots, x_n$. Also by hypothesis

\[ \Sigma x_i^2 = \Sigma q_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \pm l_{ij}^2 \]

If $n = \Sigma n_i$, then there are $n$ linear functions $l_{ij}$, which supply a transformation from $z$ to $l$, viz.,

\[ 1 = xA \]

with $A$ as the matrix of transformation. If $\Delta$ denotes a diagonal matrix with $\pm 1$ in the diagonal then

\[ xx' = \Sigma \pm l_{ij}^2 = 1 \Delta' = xAA'x' = xx' \]

or

\[ A\Delta A' = I \]

with $n$.
Since $|\Delta| \neq 0$, $|A| \neq 0$. This shows that the transformation is non-singular, in which case the positive definite form $xx'$ remains positive even after transformation. Therefore the coefficients in (26.4.1) should all be $+1$. The joint distribution of $l$ is derived from that of $x$

$$\text{const. } e^{-\frac{1}{2}l_1^2 + \cdots + l_n^2} \, dx_1 \cdots dx_n \sim \text{const. } e^{-\frac{1}{2}l_1^2} \, dl_1,$$

which shows that $l_j$ are independently distributed and so are $q_1, q_2, \cdots, q_k$, each of which depends on exclusive subsets of $l_j$. Since $q_i$ is the sum of squares of $n_i$ independent normal variates $l_{1i}, \cdots, l_{ni}$, it is distributed as $\chi^2$ with $n_i$ degrees of freedom. This establishes the sufficiency of the condition.

If $q_i$ is distributed as $\chi^2$ with $n_i$ degrees of freedom, then $\Sigma q_i$, being the sum of $k$ independent $\chi^2$, is also a $\chi^2$ with $\Sigma n_i$ degrees of freedom. But $\Sigma q_i = \Sigma l_i^2$, being the sum of squares of $n$ variates $N(0, 1)$, is a $\chi^2$ with $n$ degrees of freedom. Therefore $n = \Sigma n_i$.

### 26.5 Distribution of Non-Central $\chi^2$

Consider $k$ independent normal variates $x_1, \cdots, x_k$ with means $\mu_1, \mu_2, \cdots, \mu_k$ and standard deviations $\sigma_1, \sigma_2, \cdots, \sigma_k$. The distribution of $\left(\frac{x_1}{\sigma_1}\right)^2 + \left(\frac{x_2}{\sigma_2}\right)^2 + \cdots$ is $\chi^2$ with $k$ degrees of freedom only when $\mu_1 = \cdots = \mu_k = 0$. To find the general distribution make the following transformations.

$$y_i = \frac{x_i}{\sigma_i}, \quad i = 1, 2, \cdots, k,$$

$$z = yA$$

where $A$ is an orthogonal transformation such that the first transformed variable is

$$\left(\frac{x_1}{\sigma_1} + \cdots + \frac{y_k}{\sigma_k}\right) \quad \text{and} \quad z = \Sigma \frac{y_i^2}{\sigma_i^2}.$$

Then $E(z_1) = \mu_1, E(z_2) = \cdots = E(z_k) = 0$, and $z_1, \cdots, z_k$ are distributed as independent normal variates with unit variances.

Now

$$\chi^2 = \Sigma \left(\frac{z_i^2}{\sigma_i^2}\right) = \Sigma y_i^2 = \Sigma z_i^2 = x_1^2 + x_2^2$$
where \( x_1^2 = z_1^2 \), and \( x_2^2 = z_2^2 + \cdots + z_k^2 \). The joint distribution of \( z_1 \) and \( x_2^2 \) is simply
\[
\text{const. } e^{-\frac{1}{2} \left( x_1^2 - n + n_1 + x_2^2 \right)} \, dx_1 \, dx_2
\]
The distribution of \( x_2^2 \) and \( z_1^2 \) is
\[
e^{-\frac{r^2}{2}} \left( \frac{1}{2} \right)^{1/2} \left( \frac{1}{2} \right)^{r-1} \left( \frac{k}{r} \right)^{1/2} G \left( \frac{1}{2}, r + \frac{1}{2} \right) \, dx_2^2
\]
Hence, by using (2a.9.4), each term of the above series can be reduced, yielding the distribution of \( x_2^2 = z_1^2 + x_2^2 \),
\[
e^{-\frac{r^2}{2}} \left( \frac{1}{2} \right)^{1/2} \left( \frac{1}{2} \right)^{r-1} \left( \frac{k}{r} \right)^{1/2} G \left( \frac{1}{2}, r + \frac{1}{2} \right) \, dx_2^2
\]

2c Multivariate Normal Populations

2c.1 The Multivariate Normal Distribution

(i) A set of \( p \) variates \( \mathbf{x} = (x_1, \cdots, x_p) \) is said to follow the \( p \)-variate normal distribution if the joint probability distribution of the variables is
\[
C e^{-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)} \, d\mathbf{x}
\]
and, when \( \mathbf{x} \) is measured from the origin \( \mathbf{0} = (0, \cdots, 0) \), the distribution reduces to
\[
C e^{-\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}} \, d\mathbf{x}
\]
where \( C \) is a constant and the quadratic form \( \mathbf{x}^T \mathbf{A} \mathbf{x} \) is positive definite. The constant is determined from the relation
\[
C \int e^{-\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}} \, d\mathbf{x} = 1
\]
Since \( \mathbf{x}^T \mathbf{A} \mathbf{x} \) is positive definite, there exists a non-singular transformation \( \mathbf{x} = \mathbf{yB} \) such that
\[
\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{y}^T \mathbf{B} \mathbf{A} \mathbf{B}^T \mathbf{y} = \mathbf{y}^T \mathbf{y}
\]
\[
\therefore \mathbf{B} \mathbf{B}^T = \mathbf{I} \quad \text{or} \quad \mathbf{B} = \mathbf{A}^{-1/2}
\]
The Jacobian of the transformation is
\[
\left| \frac{\partial (x_1, \cdots, x_p)}{\partial (y_1, \cdots, y_p)} \right| = \left| \mathbf{B} \right| = \left| \mathbf{A}^{-1/2} \right|
\]
The above integral transforms to

\[ \frac{C}{|A|^\frac{1}{2}} \int e^{-\frac{1}{2} \sum \lambda_i x_i^2} \, dx_1 \cdots dx_y = \frac{C}{|A|^\frac{1}{2}} \int e^{-\frac{1}{2} \sum \lambda_i x_i^2} \, dx_1 \cdots dx_y \]

(ii) Since

\[ \int x_i e^{-\frac{1}{2} \sum \lambda_i x_i^2} \, dx = 0 \]

being an odd function of \( x_i \) and an even function of the rest, it follows that \( \mu_i \) is the mean value of the original variable \( x_i \). If \( R(x_i, x_j) \) is the covariance between the \( i \)th and \( j \)th variables, the expected values of the elements of the matrix \( xx' \) will be the elements of variance covariance or the dispersion matrix to be represented by \( R \).

\[ R = C \int x x' e^{-\frac{1}{2} \sum \lambda_i x_i^2} \, dx \]

If \( x = yB \) such that \( BAB' = I \), the above integral becomes

\[ C |B| \int B'y' y Be^{-\frac{1}{2} \sum \lambda_i y_i^2} \, dy = \frac{1}{(2\pi)^{p/2}} \int B'y' y Be^{-\frac{1}{2} \sum \lambda_i y_i^2} \, dy = B' \left[ \frac{1}{(2\pi)^{p/2}} \int y' y e^{-\frac{1}{2} \sum \lambda_i y_i^2} \, dy \right] B = B' \{I\} B = B'B \]

Since \( BAB' = I, \, A = B^{-1}B^{-1} = (BB)^{-1} \) or \( BB = A^{-1} \). This shows that the dispersion matrix of the variables is the reciprocal of the matrix of the quadratic form in the exponential. Alternatively, given the dispersion matrix \( R \) the probability density can be written as

\[ \frac{1}{|R|^\frac{1}{2}(2\pi)^{p/2}} e^{-\frac{1}{2} x'B^{-1}x} \]

The above proof is due to Nair (1949).
2c.2 The Distribution of a Set of Linear Functions of Normal Variates

Starting from the joint distribution

\[ C e^{-\frac{1}{2}xx' \sigma^2} d\mathbf{x} \]

it is required to find the distribution of \( u = (u_1, \cdots, u_k), \ (k \leq p) \), defined by

\[ u = \mathbf{x}B \]

where \( B \) is a matrix \((p, k)\) of rank \( k \). Make the transformation

\[ u = \mathbf{x}B \quad v = \mathbf{x}A \]

where \( v = (v_1, \cdots, v_{p-k}) \) and \( A \) is a matrix \((p, p-k)\) of rank \((p-k)\) and is such that \( A'\Lambda^{-1}B = 0 \). The total transformation may be represented by \((u, v) = \mathbf{x}(B, A)\). Now the rank of

\[ \begin{pmatrix} B' \\ A' \end{pmatrix} \Lambda^{-1} \begin{pmatrix} B & A \end{pmatrix} = \begin{pmatrix} B'\Lambda^{-1}B & 0 \\ 0 & A'\Lambda^{-1}A \end{pmatrix} \]

is \( p \), the same as that of \((B, A)\). This means that the rank of \( B'\Lambda^{-1}B \) is \( k \) and that \( A'\Lambda^{-1}A \) is \((p-k)\) in which case \((B'\Lambda^{-1}B)^{-1}\) and \((A'\Lambda^{-1}A)^{-1}\) exist. Let the quadratic form \( \mathbf{x}\Lambda\mathbf{x}' \) change over to

\[ uD_1u' + uD_2v' + vD_3v' \]

Substituting \( u = \mathbf{x}B \) and \( v = \mathbf{x}A \), we obtain

\[ BD_1B' + BD_2A' + AD_2A' = \Lambda \]

Pre- and post-multiplying by \( B'\Lambda^{-1} \) and \( \Lambda^{-1}B \),

\[ B'\Lambda^{-1}BD_1B'\Lambda^{-1}B = B'\Lambda^{-1}B \]

or

\[ D_1 = (B'\Lambda^{-1}B)^{-1} \]

Similarly,

\[ D_2 = (A'\Lambda^{-1}A)^{-1} \]

Pre- and post-multiplying by \( B'\Lambda^{-1} \) and \( \Lambda^{-1}A \), we find

\[ B'\Lambda^{-1}BD_2A'\Lambda^{-1}A = 0 \]

or

\[ D_2 = 0 \]

The joint distribution of \( u \) and \( v \) is

\[ \text{const.} \ e^{-\frac{1}{2}(uD_1u' + uD_2v' + vD_3v')} \ du \ dv \]
Integrating out for $v$, the distribution for $u$ is
\[ \text{const. } e^{-\frac{1}{2}(u'(B'A^{-1}B)-1u')} \, du \]
which shows that a set of linear functions of normal variates follows a multivariate normal distribution. The dispersion matrix of $u = XB$ is
\[ \left( (B'A^{-1}B)^{-1} \right)^{-1} = B'A^{-1}B \]
which can be obtained directly without finding the distribution of $u$.

In particular, if $X_1, \ldots, X_p$ follow a $p$-variate normal distribution, the subset $X_1, \ldots, X_k$ follow a $k$-variate normal distribution. It is also of interest to determine the conditional distribution of $X_{k+1}, \ldots, X_p$, given $X_1, X_2, \ldots, X_k$. Consider the partitioned vector
\[ (x_1, x_2) = (x_1, \ldots, x_k \mid x_{k+1}, \ldots, x_p) \]
with its dispersion matrix
\[ E \left( \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) (x_1, x_2) = \begin{pmatrix} A & B \\ B' & C \end{pmatrix} \]
Consider the linear functions
\[ x_2 - x_1 T \]
where the matrix $T$ is chosen such that
\[ E[x_2'(x_2 - x_1 T)] = 0 \]
i.e.,
\[ B - AT = 0 \quad \text{or} \quad T = A^{-1}B \]
\[ E(x_2 - x_1 T)'(x_2 - x_1 T) = E(x_2'x_2) + T'E(x_1'x_1)T - 2E(x_2'x_1)T \]
\[ = C + B'A^{-1}A - 2B'A^{-1}B \]
\[ = C - B'A^{-1}B = D \quad \text{say} \]
The joint distribution of $x_1$ and $y_2 = (x_2 - x_1 T)$ is
\[ \text{const. } e^{-\frac{1}{2}(x_1'A^{-1}x_1 - \frac{1}{2}y_2'D^{-1}y_2)} \, dx_1 \, dy_2 \]
and that of $x_1$ and $x_2$ is
\[ \text{const. } e^{-\frac{1}{2}(x_1'A^{-1}x_1 - \frac{1}{2}(x_2-x_1T)'D^{-1}(x_2-x_1T))} \, dx_1 \, dx_2 \]
which shows that the distribution of $x_2$, given $x_1$, is the second part of the above expression. The matrices $T$ and $D$ are already determined in terms of known quantities.
2c.3 The Distribution of Quadratic Forms

(i) Given the joint distribution

\[ C_0^{-1}\mathbf{x}\mathbf{x}' \, d\mathbf{x} \]

to find the distribution of the quadratic form \( \mathbf{x}\Lambda\mathbf{x}' \) make the transformation \( \mathbf{y} = \mathbf{B} \) such that

\[ \mathbf{x}\Lambda\mathbf{x}' = \mathbf{y}\mathbf{y}' \]

in which case the joint distribution of \( \mathbf{y} \) is

\[ \text{const.} \, e^{-\frac{1}{2}(y_1^2 + \cdots + y_p^2)} \, dy_1, \ldots, dy_p \]

The distribution of \( z_\mathbf{z}' = \sum y_i^2 \), being the sum of squares of \( p \) variates \( N(0, 1) \), is distributed as \( x^2 \) with \( p \) degrees of freedom.

(ii) Suppose that the \( x_i \) are subject to \( k \) restrictions specified by \( \mathbf{xB} = 0 \), where \( \mathbf{B} \) is a matrix \((p, k)\). As in 2c.2 the transformation

\[ u = \mathbf{xB} \quad v = \mathbf{xA} \]

such that \( A'A^{-1}B = 0 \) gives

\[ \mathbf{x}\Lambda\mathbf{x}' = u(B'A^{-1}B)^{-1}u' + v(A'A^{-1}A)^{-1}v' \]

The value of the quadratic form \( \mathbf{x}\Lambda\mathbf{x}' \) subject to the restriction \( u = \mathbf{xB} = 0 \) is \( v(A'A^{-1}A)^{-1}v' \). The distribution \( v \) for any \( u \) is independent of \( u \) and is given by

\[ \text{const.} \, e^{-\frac{1}{2}vDv'} \, dv \quad D_3 = (A'A^{-1}A)^{-1} \]

By the result proved in (i), \( vDv' \) is distributed as \( x^2 \) with \( (p - k) \) degrees of freedom. So the quadratic form \( \mathbf{x}\Lambda\mathbf{x}' \) subject to the condition \( \mathbf{x}\mathbf{B} = 0 \) is distributed as \( x^2 \) with \( (p - k) \) degrees of freedom.

(iii) If \( \mathbf{x}\Lambda\mathbf{x}' = q_1 + \cdots + q_k \) where \( q_i \) is a quadratic form of rank \( p_i \), then the necessary and sufficient condition that \( q \) are independently distributed as \( x^2 \) is that \( p = \Sigma p_i \). This is essentially Cochran's theorem with a general positive definite quadratic form \( \mathbf{x}\Lambda\mathbf{x}' \) instead of \( \mathbf{x}\mathbf{x}' \). The proof remains the same.

(iv) The necessary and sufficient condition that a quadratic form \( \mathbf{xDx}' \) of rank \( k \) is distributed as \( x^2 \) when \( \mathbf{x} \) follows the law, \( \text{const.} \exp \left( -\frac{1}{2}(\mathbf{x}\Lambda\mathbf{x}) \right) d\mathbf{x} \), is that the quadratic form \( \mathbf{x}(\Lambda - D)\mathbf{x}' \) is of rank \( (p - k) \). The sufficiency condition follows from result (iii), since \( \mathbf{x}\Lambda\mathbf{x}' = \mathbf{xDx}' + \mathbf{x}(\Lambda - D)\mathbf{x}' \). To prove the necessity we observe that there exists a transformation \( \mathbf{y} = \mathbf{xB} \) which transforms (see lc.5).

\[ \begin{align*}
\mathbf{x}\Lambda\mathbf{x}' &\to y_1^2 + \cdots + y_p^2 \\
\mathbf{xDx}' &\to \lambda_1 y_1^2 + \cdots + \lambda_k y_k^2
\end{align*} \]

\[ 2.3037 \]

\[ 2.1569 \]
The quadratic form $\mathbf{x} \mathbf{D} \mathbf{x}'$ is thus a linear compound of the squares of independent normal variates $N(0, 1)$. It may be verified that its distribution is $\chi^2$ when and only when each $\lambda$ is equal to unity. Otherwise the distribution is different from $\chi^2$. Take, for instance, $\lambda_1 \gamma_1^2 + \lambda_2 \gamma_2^2$. If this is distributed as $\chi^2$ with 2 degrees of freedom, then

$$E(\lambda_1 \gamma_1^2 + \lambda_2 \gamma_2^2) = \lambda_1 + \lambda_2 = 2$$
$$E(\lambda_1 \gamma_1^2 + \lambda_2 \gamma_2^2)^2 = 3(\lambda_1 + \lambda_2)^2 - 4\lambda_1 \lambda_2$$

$$= 3 \times 4 - 4\lambda_1 \lambda_2 = 8$$

$\lambda_1 \lambda_2 = 1$

which means $\lambda_1 = \lambda_2 = 1$. Similarly, it may be shown by considering $k$ moments that $\lambda_1, \ldots, \lambda_k$ are the roots of an equation $(\lambda - 1)^k = 0$. Therefore $\mathbf{x} \mathbf{D} \mathbf{x}'$ transforms to $\gamma_1^2 + \cdots + \gamma_k^2$, in which case $\mathbf{x}(\Lambda - \mathbf{D}) \mathbf{x}'$ transforms to $y_1^2 + \cdots + y_k^2$ or the rank of $\mathbf{x}(\Lambda - \mathbf{D})$ is $(p - k)$.

(v) Let $\mathbf{x} \mathbf{D}_1 \mathbf{x}'$ and $\mathbf{x} \mathbf{D}_2 \mathbf{x}'$ be two quadratic forms distributed as $\chi^2$ with $k_1$ and $k_2$ degrees of freedom. The necessary and sufficient condition that they are independently distributed is

$$\mathbf{D}_1 \mathbf{A}^{-1} \mathbf{D}_2 = 0.$$

In (iv) it is seen that the transformation $y = \mathbf{x} \mathbf{B}$ transforms

$$\mathbf{x} \mathbf{D} \mathbf{x}' \rightarrow y_1^2 + \cdots + y_k^2$$
$$\mathbf{x}(\Lambda - \mathbf{D}) \mathbf{x}' \rightarrow y_{k+1}^2 + \cdots + y_p^2.$$

If $\mathbf{z}_1 = (y_1, \ldots, y_k)$ and $\mathbf{z}_2 = (y_{k+1}, \ldots, y_p)$, then

$$\mathbf{z}_1 \mid \mathbf{z}_2 = \mathbf{x}(\mathbf{B}_1 \mid \mathbf{B}_2) = \mathbf{x} \mathbf{B}$$

$$\mathbf{x} \mathbf{D}_1 \mathbf{x}' = \mathbf{z}_1 \mathbf{z}_1' = \mathbf{x} \mathbf{B}_1 \mathbf{B}_1 \mathbf{x}'$$
$$\mathbf{x}(\Lambda - \mathbf{D}) \mathbf{x}' = \mathbf{z}_2 \mathbf{z}_2' = \mathbf{x} \mathbf{B}_2 \mathbf{B}_2 \mathbf{x}'$$

Since $\mathbf{z}_1$ and $\mathbf{z}_2$ are independently distributed,

$$\mathbf{B}_1 \mathbf{A}^{-1} \mathbf{B}_2' = 0$$

which gives

$$\mathbf{B}_1' \mathbf{B}_1 \mathbf{A}^{-1} \mathbf{B}_2' \mathbf{B}_2 = 0 \quad \text{or} \quad \mathbf{D} \mathbf{A}^{-1} (\Lambda - \mathbf{D}) = 0$$

$$\mathbf{D} = \mathbf{D} \mathbf{A}^{-1} \mathbf{D}$$

This is another form of the necessary and sufficient* condition for $\mathbf{x} \mathbf{D} \mathbf{x}'$ to be distributed as $\chi^2$.

* It is not difficult to prove sufficiency because $\mathbf{D} = \mathbf{D} \mathbf{A}^{-1} \mathbf{D}$ means that $\mathbf{D} \mathbf{A}^{-1} (\Lambda - \mathbf{D}) = 0$. From Example 5 in 1.5 it follows that $\mathbf{D}$ and $(\Lambda - \mathbf{D})$ have independent row vectors. The rank of the vectors in $\mathbf{D}$ and $(\Lambda - \mathbf{D})$ put together is obviously $p$. Therefore $\text{rank } \mathbf{D} + \text{rank } (\Lambda - \mathbf{D}) = p$. Then (iii) holds good.
THE DISTRIBUTION OF QUADRATIC FORMS

Since \( xD_1x' \) is a \( x^2 \), \( D_1 = D_1\Lambda^{-1}D_1 \). For the same reason \( D_2\Lambda^{-1}D_2 = D_2 \). If \( D_1\Lambda^{-1}D_2 = 0 \), then \( \text{rank } D_1 + \text{rank } D_2 + \text{rank } (\Lambda - D_1 - D_2) = p \). This shows, by (iii), that \( xD_1x' \) and \( xD_2x' \) are independently distributed. On the other hand, if \( xD_1x' \) and \( xD_2x' \) are independently distributed, then \( x(\Lambda - D_1 - D_2)x' \) is also an independent \( x^2 \). Hence there exists a transformation which transforms the quadratic forms \( xD_1x', xD_2x', x(\Lambda - D_1 - D_2)x' \) into sums of squares of independent variables. Proceeding as above it is seen that \( D_1\Lambda^{-1}D_2 = 0 \).

(vi) The necessary and sufficient condition that a linear function \( lx' \) is distributed independently of a quadratic form \( xDx' \) which is a \( x^2 \) is that

\[
1\Lambda^{-1}D = 0
\]

This follows from (iv), since the quadratic forms \( x'llx' \) and \( xDx' \) are independently distributed:

\[
1\Lambda^{-1}D = 0 \quad \text{or} \quad 1\Lambda^{-1}D = 0
\]

(vii) The necessary and sufficient conditions in (v) and (vi) are true under more general conditions. For instance, it is not necessary to assume that the quadratic forms are distributed as \( x^2 \). This has been assumed to obtain simpler proofs of the results. This assumption is not stringent because by using (iv) we can always test whether some given quadratic forms are distributed as \( x^2 \) or not. If they are, then questions of independence arise; otherwise the results are not important.

(viii) The distribution of the quadratic form \( xAx' \) when \( x \) is distributed as

\[
\text{const. } e^{-\frac{1}{2}(x - \Lambda \Lambda^{-1}A^{-1}x') \Lambda^{-1}A^{-1}x'} dx
\]

can be obtained by making the transformation,

\[
x = yA
\]

such that

\[
xAx' = yy' \quad \Lambda \Lambda^{-1} = I \quad \Lambda = \Lambda^{-1}A^{-1}
\]

\[
E(x) = E(y)A \quad \text{or} \quad E(y) = \Lambda^{-1}A^{-1}
\]

The distribution of \( 2y^2 \) is the non-central \( x^2 \) of 2b.5 with the value of

\[
x^2 = (\mu A^{-1})^2 = \mu A^{-1}A^{-1}\mu' = \mu A\mu',
\]

and \( k = \) the number of variables.
2d Least Squares Fundamental in Distribution Theory

2d.1 Two Theorems on Least Squares

Suppose \( y_1, y_2, \ldots, y_n \) are \( n \) independent normal variates with the same variance \( \sigma^2 \) and

\[
E(y_i) = \alpha_{i1} \tau_1 + \cdots + \alpha_{ik} \tau_k \quad i = 1, 2, \ldots, n
\]

where \( \alpha_{ij} \) are elements of a specified matrix \( A \) and \( \tau_1, \tau_2, \ldots, \tau_k \) are unknown parameters.

(i) If \( R_0^2 \) is the minimum value of

\[
\Sigma(y_i - \alpha_{i1} \tau_1 - \cdots - \alpha_{ik} \tau_k)^2
\]

when minimized with respect to \( \tau_1, \ldots, \tau_k \), what is the distribution of \( R_0^2 \)?

Let there be \( r \) independent vectors in the set

\[
a_i = (a_{i1}, a_{i2}, \ldots, a_{in}) \quad i = 1, \ldots, k
\]

Then there exist \( (n - r) \) vectors \( \beta_1, \ldots, \beta_{n-r} \), all orthogonal to \( a_1, \ldots, a_k \), so that \( a_i \cdot \beta_j = 0 \). The \( \beta \) vectors themselves can be chosen to satisfy the conditions \( \beta_i \cdot \beta_j = 1 \) and \( \beta_i \cdot \beta_j = 0 \). They are the vectors of the deficiency matrix considered in 1a.4. The vector \( y \) can be expressed as a linear function of \( a \) and \( \beta \) (example 3 in 1a.4).

\[
y = c_1a_1 + \cdots + c_k a_k + d_1 \beta_1 + \cdots + d_{n-r} \beta_{n-r}
\]

\[
E(y) = \tau_1 a_1 + \cdots + \tau_k a_k
\]

Multiplying by \( \beta_i \) we find

\[
\beta_i \cdot y = d_i \quad \text{and} \quad E(d_i) = \beta_i \cdot E(y) = 0
\]

Also

\[
V(d_i) = \beta_i \cdot \sigma^2 = \sigma^2 \quad \text{cov}(d_i, d_j) = \beta_i \cdot \beta_j \cdot \sigma^2 = 0
\]

Hence \( d_1, \ldots, d_{n-r} \) are all distributed independently in \( N(0, \sigma^2) \), in which case

\[
\frac{(d_1^2 + \cdots + d_{n-r}^2)}{\sigma^2}
\]

is distributed as \( \chi^2 \) with \( (n - r) \) degrees of freedom.

If \( c = (c_1, \ldots, c_k) \), then

\[
[y - E(y)]^2 = (c - \tau)A'A(c - \tau)' + d_1^2 + \cdots + d_{n-r}^2
\]

which attains the minimum value

\[
R_0^2 = d_1^2 + \cdots + d_{n-r}^2
\]
TWO THEOREMS ON LEAST SQUARES

when \((c - \tau) = 0\). Hence \(R_0^2/\sigma^2\) is a \(\chi^2\) with \((n - r)\) degrees of freedom.\(^*\)

(ii) What is the distribution of

\[ R_1^2 = \text{minimum of } \sum (y_i - a_1 r_1 - \cdots - a_r r_r)^2 \]

when minimized with respect to \(r_1, \cdots, r_r\) subject to \(s\) independent conditions

\[
\begin{align*}
1 \cdot \tau &= f_{11} r_1 + \cdots + f_{1r} r_r = g_1 \\
2 \cdot \tau &= f_{21} r_1 + \cdots + f_{2r} r_r = g_2 \\
\vdots \\
R \cdot \tau &= f_{R1} r_1 + \cdots + f_{Rr} r_r = g_r
\end{align*}
\]

(211.1.1)

Starting with the representation

\[ y = c_1 a_1 + \cdots + c_s a_s + d_1 \beta_1 + \cdots + d_{s-r} \beta_{s-r} \]

and multiplying by \(a_1, \cdots, a_s\), we obtain

\[
\begin{align*}
a_1 \cdot y &= c_1 a_1 \cdot a_1 + \cdots + c_s a_s \cdot a_s \\
& \vdots \\
a_s \cdot y &= c_1 a_1 \cdot a_s + \cdots + c_s a_s \cdot a_s
\end{align*}
\]

which may be written in the matrix notation

\[ y A = c A' A \]

If there exists a vector \(l\) such that \(l A' A = p\), then

\[
\begin{align*}
E(c \cdot p) &= E(c A' A l) = E(y A l) \\
&= E(y) A' l = \tau \cdot p \\
V(c \cdot p) &= V(c A' A l) = V(y A l) = l A' A \sigma^2 = p \cdot 1 \sigma^2
\end{align*}
\]

If \(l_1\) and \(l_2\) are two vectors such that \(l_1 A' A = p_1\) and \(l_2 A' A = p_2\), then

\[ \text{cov} (c \cdot p_1, c \cdot p_2) = l_1 \cdot \sigma^2 = l_2 \cdot \sigma^2 \]

Let \(l_1, \cdots, l_s\) be such that

\[ l_s A' A = p_1, \cdots, l_s A' A = p_s \]

and

\[ z_i = c \cdot p_i - p_i \cdot \tau, \cdots, z_i = c \cdot p_i - p_i \cdot \tau \]

* From this it follows that the expected value of the residual sum of squares is \((n - r)\sigma^2\). To prove this it is not necessary to assume that the variables are normally distributed. The result follows from the fact \(V(d_i) = E(d_i^2) = \sigma^2\).
The dispersion matrix of $z$ is

$$Dz^2 = \begin{pmatrix} p_1 \cdot l_1 & \cdots & p_1 \cdot l_n \\ p_2 \cdot l_1 & \cdots & p_2 \cdot l_n \\ \vdots & \ddots & \vdots \\ p_t \cdot l_1 & \cdots & p_t \cdot l_n \end{pmatrix},$$

in which case

$$zDz^{-1}z'$$

is distributed as $\chi^2$ with $t$ degrees of freedom. Also

$$\text{cov}(d_i, c \cdot p_j) = (p_i \cdot l_j A' A) \sigma^2 = 0$$

so that the $d$ and $z$ are uncorrelated. Hence

$$\frac{zDz^{-1}z'}{\sigma^2} + \frac{d_i^2 + \cdots + d_{i-r}^2}{\sigma^2}$$

is distributed as the sum of two independent $\chi^2$s with $t$ and $(n - r)$ degrees of freedom which is the same as $\chi^2$ with $(t + n - r)$ degrees of freedom.

To minimize $(c - \tau)A'(c - \tau)'$ subject to the conditions (2d.1.1), we observe that the conditions (2d.1.1) could be replaced by an equivalent set of $s$ independent linear combinations

$$p_1 \cdot \tau = p_{11} \tau_1 + \cdots + p_{1k} \tau_k = G_1$$

$$\vdots$$

$$p_t \cdot \tau = p_{t1} \tau_1 + \cdots + p_{tk} \tau_k = G_t$$

$$p_{t+1} \cdot \tau = p_{(t+1)1} \tau_1 + \cdots + p_{(t+1)k} \tau_k = G_{t+1}$$

$$\vdots$$

$$p_r \cdot \tau = p_{r1} \tau_1 + \cdots + p_{rk} \tau_k = G_r$$

such that the vectors $p_1, \cdots, p_r$ lie in the space of vectors in the matrix $A' A$ and no linear combination of $p_{r+1}, \cdots, p_s$ lies in the space of $A' A$ (example 3 in lb.1). Let $(s - t)$ be the rank of the matrix

$$(\delta, f_i)$$

where $\delta_1, \delta_2, \cdots$ are vectors orthogonal to those in $A' A$ which are the same as those orthogonal to row vectors in $A$. The vectors $f_1, \cdots, f_s$ belong to the restraining conditions (2d.1.1).

The number of vectors $p_1, \cdots, p_r$ is obtained from the rule, $t = s$ minus the rank of $(\delta, f_i)$. 
It may be observed that (2d.1.1) is being replaced by (2d.1.2) for proving a result and not necessarily for convenience in determining the residual sum of squares which may be obtained in any way.

Using Lagrangian multipliers, we consider the function

\[(c - \tau)A'(c - \tau)' + 2\{\lambda_1(p_1 \cdot \tau - G_1) + \cdots + \lambda_n(p_n \cdot \tau - G_2)\}\]

The minimizing equations are

\[(c - \tau)A'A - \lambda_1p_1 - \cdots - \lambda_np_n = 0\]

or

\[(c - \tau)A'A - \lambda_1p_1 - \cdots - \lambda_n(p_{n+1} + \cdots + \lambda_np_n) = 0\]

This shows that there exists a linear combination of \(p_{n+1}, \ldots, p_1\) which can be expressed in terms of the vectors in \(A'A\), unless \(\lambda_{n+1} = \lambda_{n+2} = \cdots = \lambda_n = 0\). Multiplying the minimizing equation by \((c - \tau)\), we find the optimum value of \((c - \tau)A'(c - \tau)\)' to be \((c - \tau)(\lambda_1p_1 + \cdots + \lambda_np_n)\). Also multiplying the minimizing equation by \(\lambda_1, \ldots, \lambda_n\), defined earlier, we obtain

\[(c - \tau)A'\lambda_1' = \lambda_1p_1 \cdot \lambda_1' + \cdots + \lambda_np_n \cdot \lambda_n'\]

which yields the solution

\[(\lambda_1, \ldots, \lambda_n) = zD^{-1}\]

so that the minimum value of \((c - \tau)A'(c - \tau)'\) is \(zD^{-1}z'\). The minimum value of \([y - E(y)]^2\) subject to the conditions (2d.1.1) is then

\[R_1^2 = zD^{-1}z' + d_1^2 + \cdots + d_{n-t}^2\]

and it is already shown that \(R_1^2/s^2\) is a \(\chi^2\) with \((n - r + t)\) degrees of freedom, where \(t\) is the number of vectors in the \(f\) space depending on the column vectors in the matrix \(A\) and (2d.1.1) holds.

It also follows that the difference \(R_1^2 - R_0^2\) between the conditional and the unconditional minima is distributed as \(s^2\chi^2\) with \(t\) degrees of freedom.

Suppose that \(R_2^2\) is the minimum sum of squares when some more restrictive conditions are given. Using all the conditions, let \(u\) be the extra number of independent vectors which can be expressed as linear combinations of \(a\); then by the above argument

\[R_2^2 = \sum_{i=1}^{u} \sum_{j=1}^{t} \lambda_i^j a_j + d_1^2 + \cdots + d_{n-t}\]
where $1/\theta^2(X_i)$ is inverse to the variance-covariance matrix $(\lambda_{ij})\theta^2$ for $x_1, \cdots, x_{(p+1)}$. This shows that $R_2^2/\theta^2$ is distributed as $\chi^2$ with $(n - r + t + u)$ degrees of freedom. The difference

$$R_2^2 - R_1^2 = \sum_{i,j=1}^{p+1} \lambda_{ij}z_i z_j - \sum_{i,j=1}^{p} d_{ij}z_i z_j$$

is distributed as $\sigma^2\chi^2$ with $u$ degrees of freedom. (See 2c.2 and 2c.3 where the $\chi^2$ is split into two independent components, $\mathbf{x}' \mathbf{x} = u\mathbf{d}\mathbf{u}' + \mathbf{v}' \mathbf{v}'$. One of the parts may be identified with $\sum_{i,j=1}^{p} d_{ij}z_i z_j$.

Consider the special case where $n$ sets of $(p + 1)$ variates $X_1, X_2, \cdots, X_{p+1}$ are such that the conditioned expectation and variance of $X_{p+1}$ are as follows.

$$E(x_{p+1,.}) = \alpha + \beta x_{1.} + \cdots + \gamma x_{.p}$$

$$V(x_{p+1}) = \sigma^2$$

The variate $x_{p+1}$ for given $x_1, \cdots, x_p$ is considered to be normally distributed in the following examples. The values of $x_1, \cdots, x_p$ are taken as fixed quantities.

Example 1. The minimizing equations for $\beta$ coefficients are (writing $b$ for $\beta$)

$$b_i S_{1i} + b_2 S_{2i} + \cdots + b_p S_{pi} = S_{(p+1)i}$$

$i = 1, \cdots, p$

where $S_{ij}$ is the corrected sum of products for the $i$th and $j$th variables.

Example 2. If $(S^{(i)})^p$ is the matrix reciprocal to $(S_{ij})^p (i, j = 1, \cdots, p)$, then the dispersion matrix of $b_1, \cdots, b_p$ is $\sigma^2(S^{(i)})^p$. [Hint: $V(S_{(p+1)i}) = S_{ii}\sigma^2$ and $\text{cov}(S_{(p+1)i}S_{(p+1)j}) = S_{ij}\sigma^2$]

Example 3. The unconditional residual sum of squares is

$$S_{(p+1)(p+1)} = b_i S_{(p+1)i} - \cdots - b_p S_{(p+1)p}$$

Example 4. If $\alpha = 0$, the minimum sum of squares is

$$\frac{|S_{ij} + n\bar{x}_i \bar{z}_j|_{p+1}}{|S_{ij} + n\bar{z}_j|_p}$$

where $\bar{z}_i$ is the average for the $i$th variable.

Example 5. The statistic

$$R_{p+1} = \frac{|S_{ij} + n\bar{x}_i \bar{z}_j|_p}{|S_{ij}|_p} \frac{|S_{ij}|_{p+1}}{|S_{ij} + n\bar{z}_j|_{p+1}}$$
has the distribution
\[ B \left( \frac{n - p - 1}{2}, \frac{1}{2} \right) dR_{p+1} \]
when \( \alpha = 0 \).

Example 6. The joint distribution of \( b_1, b_2, \ldots, b_p \) is
\[ \frac{|S_{ij}|}{(2\pi)^{p/2}} e^{-\frac{1}{2} \sum \sum (\hat{a}_i - \beta_i \hat{b}_i - s_0) \hat{b}_i} \cdots \hat{b}_p \]
(Hint: \( b_i \) are linear functions of \( x_{p+1} \), and their dispersion matrix is as given in example 2.)

Example 7. The distribution of
\[ B^2 = \Sigma b_i \hat{b}_i S_{ij} \]
is that of non-central \( \chi^2 \) (viii in 2c.3)
\[ \text{const.} \cdot e^{-\beta/2} \sum \gamma(\alpha + p/2) dB \]
where
\[ \beta = \frac{\Sigma \Sigma \hat{b}_i \hat{b}_j S_{ij}}{\sigma^2} \]

Example 8. Defining multiple correlation
\[ R^2 = \frac{B^2}{B^2 + W^2} \]
where \( W^2 \) is the residual sum of squares which has the distribution \( G(1/2, (n - p - 1)/2) \), show that the distribution of \( R \) is
\[ \frac{e^{-\beta/2} \beta^{p/2}}{\left( \frac{p}{2} \right)^{n - p - 1/2}} R^{p-1} (1 - R^2)^{(n-p-3)/2} F\left( \frac{n - 1}{2}, \frac{p}{2}, \frac{\beta^2 R^2}{2} \right) dR \]
(Hint: Write the joint distribution of \( B^2 \) and \( W^2 \) and apply (2a.9.5) to each term.)

Example 9. The joint distribution of \( \hat{x}_{p+1}, S_{(p+1)1}, \ldots, S_{(p+1)p}, S_{(p+1)(p+1)} \) can be derived as follows.
The distribution of $x_{p+1}, b_1, \ldots, b_p$, and $W^2$ is the product of the distributions of $x_{p+1}, b_1, \ldots, b_p$, and $W^2$.

\[ \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-x_{p+1}^2/2} dx_{p+1} \]

\[ \frac{|S_{ij}|^{1/2}}{(2\pi)^{p/2}} e^{-\psi_1/2} db_1 \cdots db_p \]

\[ \frac{1}{(2\pi)^{(n-p-1)/2} \Gamma((n-p-1)/2)} e^{-\psi_2/2} (W^2)^{(n-p-3)/2} dW^2 \]

where

\[ \psi_1 = n(x_{p+1} - \alpha - \beta_1 x_1 - \cdots)^2 \]

\[ \psi_2 = \Sigma \Sigma S_{ij} (b_i - \beta_i)(b_j - \beta_j) \]

\[ \psi_2 = W^2 = S_{(p+1)(p+1)} - b_1 S_{(p+1)1} - \cdots - b_p S_{(p+1)p} \]

all adding up to

\[ \psi = \Sigma (x_{(p+1)r} - \alpha - \beta_t x_t - \cdots - \beta_p x_p)^2 \]

The connecting relation between $b$ and $S_{(p+1)i}$ is

\[ S_{(p+1)i} = b_t S_{it} + \cdots + b_p S_{pi} \]

Therefore

\[ D(S_{(p+1)i}, S_{(p+1)2}, \ldots, S_{(p+1)p}) = |S_{ij}|_p \]

Also

\[ \frac{\partial W^2}{\partial S_{(p+1)(p+1)}} = 1 \]

Hence

\[ dx_{p+1} db_1 \cdots db_p dW^2 \sim \frac{1}{|S_{ij}|_p} dx_{p+1} dS_{(p+1)1} \cdots dS_{(p+1)p} dS_{(p+1)(p+1)} \]

The joint distribution can be written

\[ \text{const.} \ e^{-\psi_1 + \psi_2 + \psi_3/2} |S_{ij}|_p \left| \frac{S_{ij}}{|S_{ij}|_p} \right|^{(n-p-3)/2} \times \frac{1}{|S_{ij}|_p} \frac{1}{dS_{(p+1)1} \cdots dS_{(p+1)p} dS_{(p+1)(p+1)}} \]

\[ \text{const.} = \frac{\sqrt{n}}{\sqrt{2\pi}} \left( \frac{\sqrt{2\pi}}{2\pi} \right)^{p/2} \frac{1}{(2\pi)^{(n-p-1)/2} \Gamma((n-p-1)/2)} \left( \frac{n-p-1}{2} \right)^{p/2} \]
where $\varphi_1, \varphi_2, \varphi_3$ can now be expressed in terms of the variables occurring in the differentials.

**Example 10.** Since the above distribution could be obtained by direct integration, it follows on omitting the exponentials

\[
\left( \frac{1}{\sqrt{2\pi}} \right)^n \int_{S_{1(p+1)}} \cdots dS_{n(p+1)} = \text{const.}
\]

where the value of the constant is the same as in example 9 and does not involve any $S_{ij}$.

### 2d.2 Multivariate Distributions

In examples 1 to 10 of 2d.1, what has been considered is only the relative distribution of $x_{p+1}$, given $x_1, \cdots, x_p$, so that the distributions obtained are all relative distributions for fixed values of $x_1 \cdots x_p$. If these distributions are multiplied by the joint distribution $x_1, \cdots, x_p$ and integrated for these variables, then unconditional distributions are obtained. We shall assume that $x_1, x_2, \cdots, x_p$ follow a $p$-variate normal distribution.

**Multiple Correlation Distribution.** For instance, in example 8, the quantity occurring in the distribution is a random variable if $x_1, \cdots, x_p$ are not fixed. Consider the variable $z = \beta_1 x_1 + \cdots + \beta_p x_p$, which is normally distributed, being a linear function of $x_1, \cdots, x_p$.

\[
V(z) = \sum_{i=1}^p \beta_i^2 a_{ij} = \Sigma^2 \quad \text{(say)}
\]

where $a_{ij}$ is the covariance between $x_i$ and $x_j$. Since $\sigma_z^2 = \sum_{r=1}^p (z_r - \bar{z})^2$, where $z_r = \beta_1 x_{1r} + \cdots + \beta_p x_{pr}$ corresponding to the $r$th set, it follows that the distribution of $\beta$ is (see 2b.1)

\[
\text{const.} \cdot e^{-z^2/2\Sigma^2} \beta^{n-2} d\beta
\]

The joint distribution of $R$ and $\beta$ is

\[
P(R, \beta) = P(\beta)P(R|\beta)
\]

\[
= \text{const.} \cdot e^{-z^2/2\Sigma^2} \beta^{n-2} d\beta \cdot e^{-n/2} R^{n-1} (1 - R^2)^{(n-p-3)/2}
\]

\[
\times \frac{n-1}{2} \binom{p}{2} \binom{p}{2} R^2 dR
\]
Expanding $F_1$ and integrating term by term for $\beta$, we obtain the unconditional distribution of $R$.

$$\text{const. } R^{p-1}(1 - R^2)^{(n-p-3)/2} dR \sum_{i=0}^{\infty} \frac{1}{s!} \frac{\Gamma\left(\frac{n-1}{2} + s\right)}{\Gamma\left(\frac{p}{2} + s\right)} \times R^{2s} \int_0^\infty e^{-\left(\sigma^2 + \Sigma^2\right)R^2} \left(\frac{R^2}{2}\right)^{(n-3)/2+s} \, d\sigma^2$$

$$= \text{const. } R^{p-1}(1 - R^2)^{(n-p-3)/2} \sum_{i=0}^{\infty} \frac{1}{s!} \frac{\Gamma\left(\frac{n-1}{2} + s\right)}{\Gamma\left(\frac{p}{2} + s\right)} \times \left(\frac{\Sigma^2}{\sigma^2 + \Sigma^2} R^2\right)^s dR$$

$$= \frac{\gamma^2}{\left(1 - \gamma^2\right)^{(n-1)/2}} \frac{P^p n^p}{P^p n^p - 1} R^{(p+1)n-1} R\left(n - 1, \frac{1}{2} \right),$$

where $\gamma^2 = \Sigma^2 / (\sigma^2 + \Sigma^2)$, the ratio of variance due to regression to total, is the measure of multiple correlation in the population.

**Wishart’s Distribution.** The problem is to find the joint distribution of the corrected sum of squares and products arising out of $n$ sets of observations from a $k$-variate normal population. If

$$x_{11} \quad \cdots \quad x_{1n}$$

$$\vdots \quad \cdots \quad \vdots$$

$$x_{1n} \quad \cdots \quad x_{nn}$$

represent the observations, their probability density is

$$\text{const. } e^{-1/\phi}$$

where $\phi = \Sigma \Sigma^i j \left(\phi_i - \mu_i\right)\left(\phi_j - \mu_j\right) = \phi_1 + \phi_2$

$$\phi_1 = \Sigma \Sigma^i j \left(\phi_i - \mu_i\right)\left(\phi_j - \mu_j\right)$$

$$\phi_2 = \Sigma \Sigma^i j \phi_i$$
Multivariate Distributions

$S_{ij}$ being the corrected sum of products. The joint distribution of $\xi_1$ and $S_{ij}$ is the product of

\[ \text{const. } e^{-\frac{1}{2}(\xi_1 + \psi_{ij})} \] (2d.2.2)

and

\[ \int_{\text{over all } \xi_1 \text{ and } S_{ij}} d\xi_1 \cdots d\xi_k \] (2d.2.3)

The value of (2d.2.3) is equal to

\[ \text{const.} \int_{s_{11},s_{11}} d\xi_1 \cdots d\xi_k \int_{s_{11},s_{11}} d\xi_1 \cdots d\xi_k \cdots \]

which on repeated applications of the result in example 10 of 2d.1 reduces to

\[ \text{const.} S_{11}^{(n-3)/2} dS_{11} \]

\[ \times S_{11}^{-\frac{1}{2}} \left[ \begin{array}{cc} S_{11} & S_{12} \\ S_{12} & S_{22} \end{array} \right]^{(n-1-2)/2} dS_{21} dS_{22} \]

\[ \times |S_{ij}|^{-\frac{1}{2}} \left[ \begin{array}{c} S_{ij} \\ |S_{ij}| \end{array} \right]^{(n-2-3)/2} dS_{31} dS_{32} dS_{33} \]

... 

\[ \times |S_{ij}|^{-(n-1-3)/2} \frac{|S_{ij}|^{(n-k-1)-3/2}}{|S_{ij}|^{k-1}} dS_{1} \cdots dS_{k} \]

\[ = \text{const.} \left[ |S_{ij}|^{n-(k-1)-3/2} dS_{1} \cdots dS_{k} \right] \]

since all the other terms cancel out. This in conjunction with (2d.2.2) gives the distribution of $\xi_1, \xi_2, \cdots$ and that of $S_{ij}$

\[ \text{const. } e^{-\frac{1}{2}(\xi_1 \xi_2 + \psi_{ij})} \left[ \begin{array}{c} S_{ij} \\ \psi_{ij} \end{array} \right]^{(n-k-2)/2} dS_{11} \cdots dS_{kk} \]

This is known as Wishart's distribution with $(n-1)$ degrees of freedom.

Example 1. For a bivariate normal distribution the joint distribution of corrected sums of squares $S_1^2, S_2^2$ and the correlation coefficient $r$ is

\[ \text{const. } e^{-\frac{1}{2}(\xi_1 \xi_2 + \psi_{ij})} \left[ \begin{array}{c} S_1^2 \\ \psi_{ij} \end{array} \right]^{(n-k-2)/2} dS_1 \cdots dS_2 \quad \text{dr} \]
Example 2. If \( \rho = 0 \), the statistic
\[
\tau = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}
\]
is distributed as Student's \( t \) with \( (n - 2) \) degrees of freedom.

Example 3. If \( \rho \neq 0 \), on making the transformation
\[
\tau = \frac{S_1S_2}{\sigma_1\sigma_2}, \quad z = \log \frac{\sigma_1S_1}{\sigma_2S_2}, \quad r = \tau
\]
and integrating for \( \tau \) and \( z \), the distribution of \( \tau \) becomes
\[
\text{const.} \cdot \frac{(1 - r^2)^{(n-4)/2}}{d(r^2)^{n-2}} \left( \frac{\cos^{-1}(\frac{-\rho r}{\sqrt{1 - r^2}})}{\sqrt{1 - r^2}} \right) dr
\]

Example 4. The constant in Wishart's distribution in the special case when \((\lambda_0)\) is a unit matrix is
\[
\left( \frac{n}{2} \right)^{(n-1)/2} \frac{\pi^{(n-1)/2}}{\prod_{k=1}^{n} \Gamma \left( \frac{n - k}{2} \right)}
\]
(Hint: Retain the constant given in example 9 of 2d.1 in evaluating the successive integrals leading to Wishart's distribution.)

Example 5. For any \((\lambda_0)\), by making a suitable transformation the constant is found to be \( \lambda^0 \left( \frac{(n-1)/2}{2} \right) \) times the value in example 4.

Example 6. If \( S_{ij}' \) and \( S_{ij}'' \) are the corrected sums of products in two independent samples of sizes \( n_1 \) and \( n_2 \), then
\[
S_{ij} = S_{ij}' + S_{ij}''
\]
follows Wishart's distribution with \((n_1 + n_2 - 2)\) degrees of freedom.

Example 7. Show that the distribution of the correlation coefficient of a fixed set of \( n \) quantities with a random set of \( n \) independent observations from a normal population is const. \((1 - r^2)^{(n-4)/2} \) dr.

If \((f_1, f_2, \cdots, f_n)\) is the fixed vector, define \( p_i = (f_i - \bar{f})/\sqrt{\Sigma (f_i - \bar{f})^2} \) \( (i = 1, \cdots, n) \). Make the orthogonal transformation
\[
y_1 = \frac{x_1 + \cdots + x_n}{\sqrt{n}}
\]
\[
y_2 = p_1x_1 + \cdots + p_nx_n
\]
and the rest being suitably chosen. It is seen that the required statistic is $y_2/\sqrt{y_2^2 + \cdots + y_n^2}$ where $y_2, \ldots, y_n$ are all independently distributed. The problem is further reduced to determining the distribution of $y_2/\sqrt{y_2^2 + \chi^2}$ where $y_2$ is a normal variate and $\chi^2$ is independently distributed with $(n-2)$ degrees of freedom.

**Example 8.** Find the distribution of

$$\frac{\sum f_i x_i}{\sqrt{\sum (x_i - \bar{x})^2}}$$

where $f_1, f_2, \ldots$ and $x_1, x_2, \ldots$ are as defined in example 7.

**Example 9.** If $x$ and $y$ are independently distributed stochastic variables and at least one has a normal distribution, then the distribution of the correlation coefficient is the same as in example 3 with $\rho = 0$. (Hint: The distribution in example 7 is independent of the fixed vector.)

**Partial Correlation Coefficient.** Let the variates $x_1, x_2, \ldots, x_{p+1}$ be such that, given $x_1, \ldots, x_{p-1},$

$$E(x_p) = a_1 + \beta_1 x_1 + \cdots + \beta_{(p-1)} x_{p-1}$$

$$E(x_{p+1}) = a_2 + \beta_1 (p+1) x_1 + \cdots + \beta_{(p-1)(p+1)} x_{p-1}$$

The correlation between $x_p$ and $x_{p+1}$ for a given set of values of $x_1, \ldots, x_{p-1}$ is called the partial correlation between $x_p$ and $x_{p+1}$. The partial correlation coefficient is estimated by correlating the residual pairs

$$x_p - a_1 - b_1 x_1 - \cdots$$

and

$$x_{p+1} - a_2 - b_1 (p+1) x_1 - \cdots$$

where $a$ and $b$ stand for the values of $\alpha$ and $\beta$ which minimize the residual sum of squares.

The vector $x_p$ containing $n$ values can be represented by

$$x_p = a_1 + b_1 x_1 + \cdots + b_{(p-1)} x_{p-1} + e_p y_1 + \cdots + e_{(n-p)} y_{n-p}$$

where $i = (1, 1, \ldots, 1)$ and $y_1, \ldots, y_{n-p}$ are mutually orthogonal vectors orthogonal to $i$ and $x_p$, and $e_p, e_{p+1} \cdots$ are suitably determined. $x_{p+1}$ has a similar representation from which the following quantities can be constructed.

$$S_{pp'} = e_p^2 + \cdots + e_{(n-p)}^2$$

$$S'_{(p+1)(p+1)} = e_{(p+1)}^2 + \cdots + e_{(p+1)(n-p)}^2$$

$$S_{p(p+1)} = e_{(p+1)} e_{p+1} + \cdots + e_{(n-p)} e_{p+1}$$
which supply the residual sum of squares and products. The estimated partial correlation coefficient is

\[ r' = \frac{S'_{p(p+1)}}{\sqrt{S_{pp} S'_{(p+1)(p+1)}}} \]

It is easy to verify that

\[ E(\epsilon_{p}) = 0 \quad V(\epsilon_{p}) = \sigma_{1}^{2} \]
\[ E(\epsilon_{(p+1)1}) = 0 \quad V(\epsilon_{(p+1)1}) = \sigma_{2}^{2} \]
\[ \text{cov}(\epsilon_{p}, \epsilon_{(p+1)}) = \rho \sigma_{1} \sigma_{2} \]

where \( \rho \) is the partial correlation coefficient.

If \( x_{p} \) and \( x_{p+1} \) for a given \( x_{1}, \ldots, x_{p-1} \) are normally distributed, then \( \epsilon_{p} \) and \( \epsilon_{p+1} \) can be regarded as observations from a bivariate population with correlation \( \rho \), and the distribution of \( r' \) is the same as that derived above with \( (n-1) \) replaced by \( (n-p) = (n-1) - (p-1) \), corresponding to \( (p-1) \) eliminated variables.

**Distribution of \( T, D^{2}, \text{etc.} \)** Consider the following \( n \) sets of observations:

\[ x_{11} \ldots x_{p1} \]
\[ x_{12} \ldots x_{p2} \]
\[ \ldots \ldots \ldots \ldots \ldots \]
\[ x_{1n} \ldots x_{pn} \]

having the joint distribution

\[ \text{const.} \ e^{-\frac{1}{2} \sum_{s=1}^{n} x_{s}^{2}} \prod_{s=1}^{n} d\epsilon_{s} \]

Defining

\[ R_{s} = \left| \frac{S_{ij} + n \epsilon_{s} \epsilon_{i}}{S_{ij}} \right|_{i=1} \]

we find from example 5 in 2d.1 that the distribution of \( R_{s} \) for \( s \neq 1 \) is

\[ B \left( \frac{n - s}{2}, \frac{1}{2} \right) dR_{s} \]

Considering the variables in the order \( x_{p}, \ldots, x_{2} \), the joint distribution of \( R_{p}, R_{p-1}, \ldots, R_{2} \) is

\[ B \left( \frac{n - 2}{2}, \frac{1}{2} \right) dR_{2} B \left( \frac{n - 3}{2}, \frac{1}{2} \right) dR_{3} \cdots B \left( \frac{n - p}{2}, \frac{1}{2} \right) dR_{p} \]

and that of \( R_{1} \) is as shown in (2b.2.4).

\[ e^{-\frac{1}{2} \sum_{r=0}^{n} \frac{1}{r!} B \left( \frac{n-1}{2}, r + \frac{1}{2} \right)} dR_{1} \]
The joint distribution of \( R_1, R_2, \ldots, R_t \) is

\[
e^{-\frac{1}{2} \sum_{r=0}^{m} \frac{1}{r!} \left( n e^2 \right)^r} B \left( \frac{n - l}{2}, r + \frac{1}{2} \right) \int \cdots \int \frac{S_{ij}}{S_{ij} + nz_i z_j} \left| \frac{S_{ij}}{S_{ij} + nz_i z_j} \right|^l dR_1 dR_2 \cdots dR_t
\]

Using the result (2a.9.8) for each term, we find the distribution of

\[
S_t = R_t R_2 \cdots R_t = \left\{ \frac{S_{ij}}{|S_{ij}|} \right\}^l
\]

is

\[
e^{-\frac{1}{2} \sum_{r=0}^{m} \frac{1}{r!} \left( n e^2 \right)^r} B \left( \frac{n - t}{2}, r + \frac{1}{2} \right) \int \cdots \int \frac{S_{ij}}{|S_{ij}|} \left| \frac{S_{ij}}{|S_{ij}|} \right|^l dR_1 dR_2 \cdots dR_t
\]

and is independent of \( S_t \) since \( R_1, \ldots, R_t \) are independent.

In proving Wishart's distribution it was shown that the joint distribution of the means \( \bar{x}_1, \ldots, \bar{x}_p \) and the corrected sum of products, \( S_{ij} \), is

\[
const. \ e^{-\frac{1}{2} \sum \left( S_{ij}^2 + S_{ij} \right) + \sum \left| S_{ij} \right|^l} \int \cdots \int \left| \frac{S_{ij}}{|S_{ij}|} \right|^l \left| \frac{S_{ij}}{|S_{ij}|} \right|^l dR_1 dR_2 \cdots dR_t
\]

The joint distribution of \( S_t \) and \( R \) could be directly derived from the above expression. Hence we obtain the following lemma.

**Lemma.** If the variables \( z_1, \ldots, z_k \) and \( c_{ij} \) \((i, j = 1, \ldots, k)\), have the probability density

\[
const. \ e^{-\frac{1}{2} \sum \left( z_i^2 + c_{ii} \right) + \sum \left| c_{ij} \right|^l} \int \cdots \int \left| \frac{c_{ij}}{|c_{ij}|} \right|^l \left| \frac{c_{ij}}{|c_{ij}|} \right|^l \left| \frac{c_{ij}}{|c_{ij}|} \right|^l dR_1 dR_2 \cdots dR_t
\]

then the statistic

\[
S_t = \left| \frac{c_{ij}}{c_{ij} + z_i z_j} \right|^l
\]

has the probability density

\[
e^{-\frac{1}{2} \sum \frac{1}{r!} \left( n e^2 \right)^r} B \left( \frac{n - t}{2}, r + \frac{1}{2} \right) \int \cdots \int \left| \frac{c_{ij}}{c_{ij} + z_i z_j} \right|^l \left| \frac{c_{ij}}{c_{ij} + z_i z_j} \right|^l dS_t
\]
and

\[
R = \left. \left| \begin{array}{c} c_{ij} + z_i z_j \\ c_{ij} \\ \end{array} \right| \right| \left. \left| \begin{array}{c} c_{ij} + z_i z_j \\ c_{ij} \\ \end{array} \right| \right| 
\]

has the probability density independent of \( S_i \)

\[
B \left( \frac{q - k + 1}{2}, \frac{k - \lambda}{2} \right) dR 
\]

(2d.2.7)

**Example 1.** Show that the statistic

\[
S_i = \left| S_{ij} \right| 
\]

is invariant under linear transformations of the variables \( x_1, \ldots, x_t \).

(Hint: If \( L \) is the transformation matrix \( \left| S_{ij} \right| = \left| L \right| \left| S_{ij} \right| \left| L^t \right| \)).

**Applications of the Lemma.**

(i) Consider a sample of size \( n \) from a \( p \)-variate normal distribution

\[
\text{const.} \, e^{-\frac{1}{2}(x - \mu \lambda)(x - \mu \lambda)^t} \, dx 
\]

To find the distribution of \( S_p \) we first make a linear transformation \( \mathbf{x} = \mathbf{y}L \) such that

\[
(x - \mu)^t \Lambda(x - \mu)^t \sim (y_1 - \mu)^2 + y_2^2 + \cdots + y_p^2 
\]

where \( \mu = \mu \Lambda \mu^t \). Since \( S_p \) is invariant under linear transformations, it has the distribution (2d.2.4) with \( \mu = \mu \Lambda \mu^t \). Suppose that the mean value of every linear function

\[
a_1 x_1 + a_2 x_2 + \cdots + a_p x_p 
\]

uncorrelated with \( x_1, x_2, \ldots, x_t \) is zero; then we can make a linear transformation of the type

\[
y_1 = a_1 x_1 + \cdots + a_t x_t \\
\vdots \\
y_t = a_t x_t \\
y_{t+1} = a_{t+1} x_1 + \cdots + a_{t+t} x_p \\
\vdots \\
y_p = a_{t+p} x_1 + \cdots + a_{2t} x_p 
\]

such that \( y_1, \ldots, y_p \) are uncorrelated. Since \( S_i \) and \( R \) are invariant under the above transformation, it follows that \( R \) has the distribution (2d.2.5) under the assumption that any linear function uncorrelated with \( x_1, \ldots, x_t \) has zero mean value.

(ii) Suppose that two samples of sizes \( n_1 \) and \( n_2 \) are available from two \( p \)-variate normal populations having the same dispersion matrix. Let
\[ d_i = \bar{x}_{i1} - \bar{x}_{i2} = \text{Difference in mean values for the } i\text{th variable in the sample.} \]

\[ s_i = \mu_{i1} - \mu_{i2} = \text{Difference in mean values for the } i\text{th variable in the population.} \]

The dispersion matrix of \( d \) is \( \frac{1}{c} = \frac{1}{[(1/n_1) + (1/n_2)]} \) times that of \( x \), so that the probability density of \( d \) is

\[
\text{const. } e^{-\frac{n}{2}d^T \Sigma^{-1} d}
\]

The \( p(p + 1)/2 \) quantities

\[ S_{ij} = S_{ij}' + S_{ij}'' \]

where \( S_{ij}' \) and \( S_{ij}'' \) are the corrected sums of products in the first and second samples, are distributed in Wishart's distribution with \( (n_1 - 1) \) + \( (n_2 - 1) \) degrees of freedom. Hence the statistic

\[ S_p = \left| S_{ij} \right|_p = \left| S_{ij} + cd_{ij} \right|_p \]

is distributed as in (2d.2.6) with \( q = (n_1 + n_2 - 2) \), \( t = p, f^2 = cd\Sigma' \).

If all linear functions uncorrelated with \( x_1, \ldots, x_t \) have no difference in mean values between the two populations, then, by making a transformation similar to that used in (i) above, it can be shown that

\[ R = \frac{\left| S_{ij} + cd_{ij} \right|_t}{\left| S_{ij} \right|_p} \]

is distributed as in (2d.2.7) with \( q = (n_1 + n_2 - 2) \), \( k = p, t = t \).

The distribution of Mahalanobis' \( D^2 \), connected with \( S_p \) by the relation

\[ S_p = \frac{1}{1 + \frac{c}{(n_1 + n_2 - 2) D^2}} \]

where \( (\Sigma') \) is the matrix reciprocal to \( (\Sigma) = (S_{ij}/(n_1 + n_2 - 2)) \) or that of Hotelling's \( T^2 \) defined by

\[ S_p = \frac{1}{1 + T^2} \]

can be deduced from that of \( S \). When \( s_1 = s_2 = \cdots = 0 \) the distribution of \( S_p \) is

\[ B \left( \frac{n_1 + n_2 - p - 1}{2} \right) dS_p \]
or that of $(1 - S_p)/S_p = T$ is

\[
\text{const.} \frac{T_p^{(p/2)-1}}{(1 + T_p)^{(n_1 + n_2 - 1)/2}} dT_p
\]

which means that

\[
\frac{n_1 + n_2 - p - 1}{p} T_p \quad \text{or} \quad \frac{n_1 + n_2 - p - 1}{p} \frac{c}{n_1 + n_2 - 2} D_p^2
\]

is distributed as a variance ratio with $p$ and $(n_1 + n_2 - p - 1)$ degrees of freedom. Similarly, when the conditions under which the distribution of $R$ is derived hold, the statistic

\[
U = \frac{n_1 + n_2 - p - 1}{p - t} \left( \frac{1}{R} - 1 \right) = \frac{n_1 + n_2 - p - 1}{p - t} \frac{1 + T_p}{1 + T_t - 1}
\]

is distributed as the variance ratio with $(p - t)$ and $(n_1 + n_2 - p - 1)$ degrees of freedom.

References


CHAPTER 3

The Theory of Linear Estimation
and Tests of Hypotheses

3a Linear Estimation

3a.1 Observational Equations

Let \( y_1, \ldots, y_n \) be \( n \) independent stochastic variables with a common unknown variance \( \sigma^2 \) and having as expectations linear functions of \( k \) unknown parameters \( T_1, \ldots, T_k \), the expectation of \( y_i \) being

\[
E(y_i) = a_{i1}T_1 + \cdots + a_{ik}T_k \quad i = 1, \ldots, n \quad (3a.1.1)
\]

where the compounding coefficients \( a_{ij} \) are known. Let the rank of the matrix \( (a_{ij}) \) or the number of independent linear functions on the right side of the equations (3a.1.1) be \( r \). The quantities \( n \) giving the number of observations, \( k \) the number of unknown parameters and \( r \) the rank of the matrix \( (a_{ij}) \) can be quite general and need not satisfy any equality or inequality relationships. Equations such as (3a.1.1) are called the observational equations. Nothing need be assumed at this stage about the actual distribution of the stochastic variables.

3a.2 Best Unbiased Estimates

A linear function \( p_1T_1 + \cdots + p_kT_k \) where the \( p \) coefficients are known is called a linear parametric function. This parametric function is said to be estimable if there exists a linear function \( b_1y_1 + \cdots + b_ny_n \) of the observations such that

\[
E(b_1y_1 + \cdots + b_ny_n) = p_1T_1 + \cdots + p_kT_k
\]

A function \( b_1y_1 + \cdots + b_ny_n \) satisfying the above condition is called an unbiased estimate of \( p_1T_1 + \cdots + p_kT_k \). If no such linear function exists, the parametric function is said to be non-estimable.

An unbiased estimate with the minimum possible variance is said to be the best unbiased estimate. The mathematical discussion of arriving at an estimate with the minimum possible variance out of a large class
of unbiased estimates is known as the theory of linear estimation, which was originally considered by Gauss and later explicitly formulated by Markoff (1912).

3a.3 The Necessary and Sufficient Condition for the Existence of an Unbiased Estimate

If \( b_1y_1 + \cdots + b_ny_n \) is an unbiased estimate of \( p_1r_1 + \cdots + p_kr_k \), then

\[
E(b_1y_1 + \cdots + b_ny_n) = p_1r_1 + \cdots + p_kr_k \tag{3a.3.1}
\]

Using the expected value of \( y \) from the equations (3a.1.1) and equating the coefficients of \( r \) on both sides, the following equations are obtained:

\[
p_i = b_1a_{i1} + \cdots + b_na_{in} \quad i = 1, \cdots, k \tag{3a.3.2}
\]

This means that the set of equations (3a.3.2) treating \( b \) as unknown is soluble. Also, if the equations (3a.3.2) are soluble, then (3a.3.1) holds.

The necessary and sufficient condition for the estimability of the parametric function \( p_1r_1 + \cdots + p_kr_k \) is that the set of equations (3a.3.2) is soluble, the condition for which is given in 1a.5. The vector \( p = (p_1, \cdots, p_k) \) must depend on the row vectors in \( (aij) \).

3a.4 Normal Equations

If the solution of (3a.3.2) is unique, then there is only one unbiased estimate and that is the best possible. In general there will be a multiplicity of solutions, and the one for which the variance is the least has to be chosen. The variance of \( b_1y_1 + \cdots + b_ny_n \) is \( (b_1^2 + \cdots + b_n^2)\sigma^2 \).

The problem, then, reduces to minimizing \( (b_1^2 + \cdots + b_n^2) \) subject to the conditions (3a.3.2). To obtain the restricted minimum we need to consider the expression

\[
\sum_{j=1}^{n} b_j^2 + 2 \sum_{j=1}^{k} \lambda_j(p_j - b_1a_{j1} - \cdots - b_na_{jn}) \tag{3a.4.1}
\]

where \( \lambda_1, \cdots, \lambda_k \) are Lagrangian multipliers and to differentiate with respect to \( b_1, \cdots, b_n \). The minimizing equations are

\[
b_j = \lambda_1a_{j1} + \cdots + \lambda_ka_{jn} \quad j = 1, \cdots, n \tag{3a.4.1}
\]

On eliminating \( b \) in (3a.3.2) with the use of (3a.4.1), the equations giving \( \lambda \) are obtained

\[
p_i = (a_{i1} \lambda_1 + \cdots + (a_{ik} \lambda_k) \quad i = 1, \cdots, k \tag{3a.4.2}
\]

where \( a_i \) is the vector \( (a_{i1}, a_{i2}, \cdots, a_{in}) \) and \( a_i \cdot a_j \) is the vector product \( a_{i1}a_{j1} + \cdots + a_{in}a_{jn} \). It is enough to get a single set of \( (\lambda_1, \cdots, \lambda_k) \) satisfying (3a.4.2) for substitution in (3a.4.1) to obtain the \( b \) coefficients,
since these are unique \( \lambda_1, \ldots, \lambda_k \) satisfy (3a.4.2). The best estimate is
\[
b_1 y_1 + \cdots + b_n y_n = (a_1 \cdot y) \lambda_1 + \cdots + (a_n \cdot y) \lambda_k
\] (3a.4.3)
where \( y = (y_1, \ldots, y_n) \) and \( a_\ell \cdot y = a_1 y_1 + \cdots + a_n y_n \). The condition of unbiasedness gives
\[
\lambda \bar{E}(a_1 \cdot y) + \lambda_2 \bar{E}(a_2 \cdot y) + \cdots = p_1 \tau_1 + p_2 \tau_2 + \cdots
\]
This shows that if the observational equations (3a.1.1) are replaced by
\[
E(a_1 \cdot y) = E(Q_i) = (a_1 \cdot a_i) \tau_1 + \cdots + (a_k \cdot a_i) \tau_k
\] (3a.4.4)
\[
t = 1, \ldots, k
\]
then any linear function of \( Q_1, \ldots, Q_k \), unbiased for a parametric function \( p_1 \tau_1 + \cdots + p_k \tau_k \), is unique as a function of \( y_1, \ldots, y_n \) and is also the best estimate. The equations (3a.4.4) so constructed are called the normal equations.

3a.5 Linear Functions with Zero Expectations

If \( c_1 y_1 + \cdots + c_n y_n \) is a linear function whose expectation is zero, then
\[
c_1 a_{1j} + \cdots + c_n a_{nj} = 0 \quad j = 1, \ldots, k
\] (3a.5.1)
Since the rank of \( (a_{ij}) \) is \( r \), there are \( (n - r) \) independent sets \( (c_1, \ldots, c_n) \) which satisfy the equations (3a.5.1) (see 1aA), which shows that there are \( (n - r) \) independent linear functions of the variables whose expectations are identically zero.

If \( b_1 y_1 + \cdots + b_n y_n \) is the best estimate of a parametric function, then
\[
b_i = \sum_{j=1}^{k} \lambda_j a_{ij} \quad \text{or, in matrix form, zero}
\]
\[
\sum_{i=1}^{k} b_i c_i = \sum_{j=1}^{k} \lambda_j \sum_{i=1}^{k} c_i a_{ij} = 0
\] (3a.5.2)
Equation (3a.5.2) shows that the best estimates of parametric functions are uncorrelated with linear functions whose expectations are zero. The number of independent parametric functions that can be estimated is \( r \), the rank of the matrix \( (a_{ij}) \), and their best estimates are uncorrelated with the \( (n - r) \) linear functions whose expectations are zero.

*If \( \chi_1, \ldots, \chi_n \) is another solution leading to \( b_1', \ldots, b_n' \), then, defining \( \lambda_1 = \chi_1 - \chi_i = d_i \) and \( b_i = b_i' - c_i \), we have
\[
0 = (a_1 \cdot a_i)d_1 + \cdots + (a_n \cdot a_i)d_n \quad i = 1, \ldots, k
\]
\[
c_i = d_i a_1 + \cdots + d_n a_n
\]
from which it follows that \( \Sigma c_i^2 = 0 \). Hence \( c_1 = c_2 = \cdots = c_n = 0 \).
78  LINEAR ESTIMATION AND TESTS OF HYPOTHESES

3a.6 Standard Errors of Estimates and Intrinsic Properties of Normal Equations

Since \( Q_i = (a_i \cdot y) \), it follows that

\[
V(Q_i) = (a_i \cdot a_i) \sigma^2 \quad \text{and} \quad \text{cov}(Q_i, Q_j) = (a_i \cdot a_j) \sigma^2.
\]

Hence the variance of \( Q_i \) is \( \sigma^2 \) times the coefficient of \( \tau_i \) in the \( i \)th normal equation, and the covariance of \( Q_i, Q_j \) is \( \sigma^2 \) times the coefficient of \( \tau_i \) in the \( j \)th normal equation or of \( \tau_j \) in the \( i \)th normal equation. If the best estimate of \( p_1 \tau_1 + \cdots + p_k \tau_k \) is given by \( l_1 Q_1 + \cdots + l_k Q_k \), then

\[
l_i = (a_i \cdot a_i) l_1 + \cdots + (a_i \cdot a_k) l_k \quad i = 1, 2, \ldots, k
\]

Using (3a.6.1),

\[
V(l_1 Q_1 + \cdots + l_k Q_k) = \sigma^2 \Sigma l_i^2 (a_i \cdot a_i)
\]

by virtue of (3a.6.2). Similarly the covariance of \( \Sigma l_i Q_i, \Sigma m_i Q_i \), the best estimates of \( \Sigma p_i \tau_i \) and \( \Sigma q_i \tau_i \), is given by

\[
\sigma^2 \Sigma l_i q_i = \sigma^2 \Sigma m_i q_i
\]

The formulae (3a.6.3) and (3a.6.4) supply an easy method of evaluating the variances and covariances of the best estimates. Only the compounding coefficients of \( Q_1, \ldots, Q_k \) need be determined for the application of these formulae.

If in the first equation the coefficient of \( \tau_1 \) is reduced to unity,

\[
E(Q_1) = \tau_1 + (a_1 \cdot a_1) \tau_2 + \cdots + (a_1 \cdot a_k) \tau_k
\]

and \( \tau_1 \) is eliminated from the rest of the equations by the method of sweep out (see 1a.3), then

\[
E(Q'_i) = E \left[ Q_i - \frac{(a_i \cdot a_i)}{(a_1 \cdot a_i)} Q_1 \right] = \sum_{j=2}^{k} \left( a_j \cdot a_i - \frac{(a_j \cdot a_i)(a_j \cdot a_1)}{(a_1 \cdot a_1)} \right) \tau_j
\]

\( i = 2, \ldots, k \)

These become normal equations for the estimation of parametric func-
tions involving \( r_2, \cdots, r_k \). Also

\[
V(Q'_i) = V \left( Q_i - \frac{\alpha_i \cdot a_i}{a_1 \cdot a_1} Q_i \right)
\]

\[
= \sigma^2 \left( \frac{\alpha_i \cdot a_i}{a_1 \cdot a_1} \right)
= \sigma^2 \times \text{the coefficient of } r_i \text{ in the equation for } Q'_i
\]

Similarly

\[
\text{cov}(Q'_i; Q'_j) = \sigma^2 \times \text{the coefficient of } r_i \text{ in the equation for } Q'_j
\]

\[
= \sigma^2 \times \text{the coefficient of } r_j \text{ in the equation for } Q'_j
\]

These properties may be termed "the intrinsic properties" of normal equations. From this it follows that, if the best estimates of \( r_2 Q_2 + \cdots + r_k Q_k \) and \( a_2 Q'_2 + \cdots + a_k Q'_k \) are \( f_2 Q'_2 + \cdots + f_k Q'_k \) and \( g_2 Q'_2 + \cdots + g_k Q'_k \), then

\[
V(f_2 Q'_2 + \cdots + f_k Q'_k) = (f_2^2 + \cdots + f_k^2) \sigma^2 \quad (3a.6.5)
\]

\[
\text{cov}(f_2 Q'_2, f_2 Q'_2) = (f_2 g_2 + \cdots + f_k g_k) \sigma^2 \quad (3a.6.6)
\]

These properties hold good when the parameters are successively eliminated by the method of sweep out.

3a.7 Principle of Substitution

Some amount of simplification can be effected in the computational methods of the foregoing analysis by adopting the principle of substitution or fitting of constants. Let \( t_1, \cdots, t_k \) be a solution of the equations

\[
Q_i = (a_1 \cdot a_1)t_1 + \cdots + (a_k \cdot a_k)t_k \quad i = 1, 2, \cdots, k \quad (3a.7.1)
\]

If an estimable parametric function \( p_1 r_1 + \cdots + p_k r_k \) is estimated by \( c_1 Q_1 + \cdots + c_k Q_k \), then

\[
\Sigma p_i r_i = \Sigma c_i \{(a_1 \cdot a_1)r_1 + \cdots + (a_k \cdot a_k)r_k \} \quad (3a.7.2)
\]

Substituting \( t \) for \( r \) in (3a.7.2),

\[
\Sigma p_i t_i = \Sigma c_i \{(a_1 \cdot a_1)t_1 + \cdots + (a_k \cdot a_k)t_k \}
= c_1 Q_1 + \cdots + c_k Q_k
\]
by virtue of the equations (3a.7.1). This shows that the best estimate of an estimable parametric function is obtained by substituting for the parameters \( \tau_1, \cdots, \tau_k \) any solution of the equations (3a.7.1). The solution for \( \tau_1 \) will be the best estimate of \( \tau_1 \) only when \( \tau_1 \) is estimable. The equations (3a.7.1) can be obtained formally by minimizing the sum of squares

\[
\frac{1}{2} \sum (y_i - a_0 t_1 - \cdots - a_k t_k)^2
\]

with respect to \( t_1, \cdots, t_k \).

If \( \Sigma \tau_i \) is non-estimable and \( \Sigma \sigma_i \) is homogeneous in \( y \), then \( E(\Sigma \sigma_i) \neq \Sigma \sigma_i \), for otherwise it contradicts the assumption of non-estimability. Also, the result of substituting a solution in an estimable parametric function leads to a homogeneous function of the observations. Hence the necessary and sufficient conditions for \( \Sigma \sigma_i \) to be the best estimate of \( \Sigma \sigma_i \) is that (a) \( \Sigma \sigma_i \) is homogeneous in \( y_1, \cdots, y_n \) and (b) \( E(\Sigma \sigma_i) = \Sigma \sigma_i \). Also, if the result of substituting two different solutions in \( \Sigma \tau_i \) leads to two different values, then \( \Sigma \tau_i \) is non-estimable. This supplies a sufficient condition for non-estimability.

The set of normal equations are consistent in the sense that there always exist solutions \( t_1, \cdots, t_k \) satisfying them. To prove this it is sufficient to show that, if there exist quantities \( d_1, \cdots, d_k \) such that

\[
\Sigma d_i (a_i - a_j) = 0 \quad \text{for all } j,
\]

then \( \Sigma d_i Q_i = 0 \). Now

\[
V(\Sigma d_i Q_i) = \sigma^2 \Sigma d_i V(a_i - a_j) = 0
\]

which shows that the variance of a homogeneous function of stochastic variables is identically zero. This is not possible unless the compounding coefficients identically vanish, in which case \( \Sigma d_i Q_i = 0 \).

Since a single solution is sufficient for the purpose of substitution, we may add a set of consistent, convenient, or conventionally chosen equations to the normal equations and solve them. In many practical situations the normal equations have unique solutions, in which case all parametric functions are estimable.

This aspect of normal equations is not properly brought out in literature. Unnecessary restrictions * have been imposed on the rank of the observational equations to make all the unknown quantities estimable, in which case the normal equations have a unique solution. The above treatment covers the most general case.

* This generalization was first noted by R. C. Bose, who developed a special method for estimating an assigned parametric function. The author (Rao, 1945) has shown that even when all the restrictions are withdrawn the least square technique of deriving normal equations and substituting the solution in a parametric function works. The principle of least squares in estimation and derivation of statistical tests is thus valid under very general conditions.
3a.8 Observational Equations with Linear Restrictions on Parameters

Sometimes it may be known that the parameters $r_1, \ldots, r_k$ in the observational equations (3a.1.1) satisfy some linear restrictions:

$$g_i = r_1 t_1 + \cdots + r_k t_k \quad i = 1, \ldots, m \quad (3a.8.1)$$

In this situation two courses are open. It may be possible to eliminate some of the $r$ parameters in the observational equations with the help of equations (3a.8.1) and obtain a different set of observational equations with fewer $r$ parameters having no restrictions. The theory developed above will then be applicable. Another method is to derive the normal equations by minimizing

$$\frac{1}{2} \sum (y_i - a_1 r_1 - \cdots - a_k r_k)^2$$

subject to the restrictions (3a.8.1). Introducing Lagrangian parameters $\lambda_1, \ldots, \lambda_m$, the normal equations $(m + k)$ in number are

$$Q_i = (a_1^t a_1) t_1 + \cdots + (a_k^t a_k) t_k + l_1 r_1 + \cdots + l_m r_m \quad i = 1, \ldots, k \quad (3a.8.2)$$

and

$$g_i = r_{1j} t_1 + \cdots + r_{kj} t_k \quad j = 1, \ldots, m$$

The best estimate of any estimable parametric function $p_1 r_1 + \cdots + p_k r_k$ is simply $p_1 t_1 + \cdots + p_k t_k$, where $t_1, \ldots, t_k$ are chosen to satisfy equations (3a.8.2).

If the best estimate of $p_1 r_1 + \cdots + p_k r_k$ is obtained as $\tilde{c}_1 Q_1 + \cdots + \tilde{c}_k Q_k + d_1 p_1 t_1 + \cdots + d_m p_m t_m$, then its variance is simply $(\tilde{c}_1^2 + \cdots + \tilde{c}_k^2 + p_1^2 + \cdots + p_m^2) \sigma^2$ as before. Similar expressions hold good for the covariances. The intrinsic properties considered in 3a.6 are also true. Equations (3a.8.2) are always soluble. Also, the best estimates of parametric functions are uncorrelated with the linear functions having zero expectations and with the estimates of the parametric functions on the right-hand side of equations (3a.8.1).

3a.9 Observational Equations with Correlated Variables

In the setup of (3a.1.1) it was assumed that $y_1, \ldots, y_n$ are independent stochastic variables having a common variance $\sigma^2$. This condition can be relaxed by assuming that the dispersion matrix is of the form $\sigma^2 \Lambda$ where the elements of $\Lambda$ are all known and $\sigma^2$ is an unknown multiplier. The observational equations may be written

$$E(y) = \tau \Lambda'$$
where \( A = (a_{ij}) \). The condition of unbiasedness of an estimate \( \beta' \) of \( \theta \) is the same as (3a.3.2)

\[ p = bA \]  

(3a.9.1)

The variance of \( \beta' \) is proportional to \( bA'b' \). Minimizing this expression, the best estimate is found to be

\[ mA'A^{-1} \beta' \]  

(3a.9.2)

where

\[ p = mA'A^{-1} \theta \]

This shows that, if \( \theta \) is estimable, the estimate is given by \( \theta' \) where \( t \) satisfies the equation

\[ \gamma A^{-1} \theta = tA'A^{-1} \theta \]  

(3a.9.3)

which is similar to (3a.4.4). With the new definition \( Q = \gamma A^{-1} \theta \) the results of 3a.5, 3a.6, 3a.7, and 3a.8 hold good in the correlated case also. The equation (3a.9.3) can be obtained by minimizing the expression

\[ \frac{1}{2} \Sigma z \Lambda z = \sum_{i=1}^{n} \left( \frac{y_i - a_1r_1 - \ldots - a_kr_k}{\theta_i - a_1t_1 - \ldots - a_kt_k} \right) \]

where \( (\Lambda^{ij}) = \Lambda^{-1} \). If the \( r \) parameters are subject to some restrictions, then the above expression is minimized subject to these restrictions.

### 3b Tests of Linear Hypotheses

#### 3b.1 Nature of Linear Hypotheses

The data on which tests of significance are based consist of \( n \) independent observations \( y_1, \ldots, y_n \) with a common variance \( \sigma^2 \) and having expectations

\[ E(y_i) = a_i1, \ldots, a_i \]  

(3b.1.1)

where \( a_{ij} \) are known and \( r_1 \) are unknown parameters except that they may be known to satisfy a set of \( s \) restrictions, \( R_0 \).

\[ r_1 t_1 + \ldots + r_k t_k = \gamma_1 \]

\[ R_0: \quad \ldots \quad \ldots \quad \ldots \]  

(3b.1.2)

\[ r_1 t_1 + \ldots + r_k t_k = \gamma_s \]

These linear restrictions can be assumed to be independent, for, if not, they can be replaced by an independent set.

* If \( E(y_i) = a_0 + a_1 r_1 + \ldots + a_k r_k \), then \( (y_i - a_0) \) can be considered to be the stochastic variable.
TEST FOR $H_0$

A linear hypothesis $H_0$ specifies the values of one or more linear functions of parameters.

$$h_1r_1 + \cdots + h_1r_k = \theta_1$$

$$H_0: \quad h_m r_1 + \cdots + h_m r_k = \theta_m$$

(3b.1.3)

As in $R_0$, the linear functions in $H_0$ can be assumed to be independent. Also, if some linear combination of the functions in $H_0$ can be expressed in terms of the conditions in $R_0$, then they can be immediately verified. To start with, $H_0$ may be replaced by a set of equations, no combination of which belongs to $R_0$ (example 3 in 1b.1). Let $H_0$ in (3b.1.3) be such a set.

3b.2 Test for $H_0$

If the hypothesis $H_0$ is to be tested, it is necessary that all vectors $h_1, \cdots, h_m$ in $H_0$ must belong to the vector space generated by the vectors

\[ a_1 = (a_{11}, \cdots, a_{1k}) \]
\[ a_n = (a_{n1}, \cdots, a_{nk}) \]
\[ r_1 = (r_{11}, \cdots, r_{1k}) \]
\[ \vdots \]
\[ r_s = (r_{s1}, \cdots, r_{sk}) \]

This is the condition for estimability of the parametric functions in $H_0$.

Let

(i) $\sigma^2 \chi^2_{R_0 + H_0}$ be the minimum value of $\Sigma(y_i - a_1r_1 - \cdots - a_k r_k)^2$ when $r_i$ are subject to the conditions $R_0$ and $H_0$,

(ii) $\sigma^2 \chi^2_{R_0}$ be the minimum value of $\Sigma(y_i - a_1r_1 - \cdots - a_k r_k)^2$ when $r_i$ are subject to the conditions $R_0$ only, and

(iii) $y_1, y_2, \cdots$ are all normally distributed.

It is shown in 2d.1 that $\chi^2_{R_0}$ is distributed as $\chi^2$ with $(n - r + t)$ degrees of freedom where $r$ is the rank of the space of $a_1, \cdots, a_n$ and $t$ is the number of independent vectors in the space of $r_1, \cdots, r_s$ which lie entirely in the space of $a_1, \cdots, a_n$. When both $R_0$ and $H_0$ are considered, there are $(t + m)$ independent vectors in the space of $r_1, \cdots, r_s, h_1, \cdots, h_m$ which can be expressed in terms of $a_1, \cdots, a_n$. Therefore
\[ X^2_{R_0 + R_1} \] is distributed as \( X^2 \) with \( (n - r + t + m) \) degrees of freedom. Hence, as shown in 2d.1,
\[ X^2_{R_0 + R_1} = X^2_{R_0 + R_1} - X^2_{R_0} \]
is also a \( X^2 \) with \( m \) degrees of freedom, the distribution being valid only when the hypothesis \( H_0 \) is true. If \( \sigma^2 \) is known, then the \( X^2 \) distribution can be used to test \( H_0 \). On the other hand, the ratio
\[ F = \frac{X^2_{R_0 + R_1}}{m} \]
is independent of \( \sigma^2 \) and is distributed as a variance ratio with \( m \) and \( (n - r + t) \) degrees of freedom. Hence the hypothesis \( H_0 \) can be tested, using the \( F \) distribution when \( \sigma^2 \) is unknown.

**Sb.3 Test for \( H_0 \) When \( R_0 \) Is Not True**

In problems of the nature posed in 3b.1 it is often desirable not to take on trust the given restrictions but to test for them if possible. In this case all parametric functions in \( R_0 \) must be estimable. If the restrictions are true and are estimable, then \( X_{R_0}^2 \) is distributed as \( X^2 \) with \( (n - r + s) \) degrees of freedom. On the other hand, the unconditional minimum value \( X_{E}^2 \) of \( (1/\sigma^2) \Sigma (y_i - \alpha_1 r_1 - \cdots - \alpha_k r_k)^2 \) is distributed as \( X^2 \) with \( (n - r) \) degrees of freedom. Hence a test for \( R_0 \) is provided by the variance ratio
\[ \frac{X_{R_0}^2 - X_{E}^2}{X_{E}^2} \]

based on \( s \) and \( (n - r) \) degrees of freedom. If this is significant, then \( R_0 \) cannot be used in testing for \( H_0 \). In such a case all parametric functions in \( H_0 \) must be directly estimable from the observational equations (3b.1.1). The statistic \( X_{R_0}^2 \), the minimum value of \( (1/\sigma^2) \Sigma (y_i - \alpha_1 r_1 - \cdots - \alpha_k r_k)^2 \) subject to \( H_0 \), is distributed as \( X^2 \) with \( (n - r + m) \) degrees of freedom. The test for \( H_0 \) when \( R_0 \) is not true is provided by the variance ratio
\[ \frac{X_{R_0}^2 - X_{E}^2}{X_{E}^2} \]

based on \( m \) and \( (n - r) \) degrees of freedom.

**Example 1.** Expressing the restrictions (3a.8.1) in the matrix form \( g = TR \), show that the condition for the estimability of a parametric function \( PT' \) is that there exist vectors \( b \) and \( c \) such that
\[ p = bA + cR \]
Example 2. The best estimate of $pr'$ is given by $\lambda Q' + c\mathbf{g}'$ where $Q = \mathbf{yA}'$ and $\lambda$, $c$ are such that

$$\lambda \mathbf{A}'\mathbf{A} + c\mathbf{R} = \mathbf{p}$$

$$\lambda \mathbf{R}' = 0$$

Hence deduce the principle of least squares given in 3a.8.

Example 3. Show that the minimum value of

$$\Sigma(y - a_1r_1 - a_2r_2 - \cdots)^2$$

in $\sum y^2 = t_1Q_1 - t_2Q_2 - \cdots$, where $t_1$, $t_2$, ... and $Q_1$, $Q_2$, ... are as in (3a.7.1).

Example 4. The minimum value of the expression in example 3 when the $r$ parameters are subject to the relations (3a.8.1) is

$$\Sigma y^2 - t_1Q_1 - t_2Q_2 - \cdots - t_1g_1 - t_2g_2 - \cdots$$

where $t_i$, $Q_i$, $t_i$ are as in (3a.8.2).

Example 5. Let $x_1$, $x_2$, ..., $x_m$ be the estimates of parametric functions in (3b.1.3) with the dispersion matrix $\sigma^2\mathbf{D}$. If $\mathbf{c} = (c_1, c_2, \cdots)$ is a vector of arbitrary constants, then a linear compound of the $m$ deviations from the hypothetical values in (3b.1.3) is

$$c(x - \mathbf{g})' = c_1(x_1 - \mathbf{g}_1) + c_2(x_2 - \mathbf{g}_2) + \cdots + c_m(x_m - \mathbf{g}_m)$$

with its variance $\sigma^2\mathbf{cDc}'$. Show that the maximum value of the ratio $\{c(x - \mathbf{g})'\sigma^2\mathbf{cDc}'/\sigma^2(x - \mathbf{g})'\sigma^2(x - \mathbf{g})'$. Example 6. Show that $x^2$ defined in 3b.2 is $(x - \mathbf{g})D^{-1}(x - \mathbf{g})'\sigma^2$, the expression derived in example 5. (Hint: Follow the method of 2d.1.)

3c The Combination of Weighted Observations

3c.1 Transformation to Unweighted Observations

The general theory of least squares as discussed above can be extended to observations having unequal variance but with known ratios. The variances $\sigma_1^2$, ..., $\sigma_n^2$ of $y_1$, ..., $y_n$ can be expressed as $w_1\sigma_1^2$, ..., $w_n\sigma_n^2$ where $w_i$ are known quantities and $\sigma^2$ is an unknown parameter. The problem is reduced to the unweighted case by replacing $y_i$ by $z_i$ where

$$z_i = \frac{y_i}{w_i}$$

in which case

$$E(z_i) = \frac{a_{i1}}{w_i} + \cdots + \frac{a_{ik}}{w_i}$$

(3c.1.1)
The general theory is applicable when the observational equations are considered as in (3c.1.1).

3c.2 An Example of Weighted Observations

Consider the observational equations

\[ E(y) = \tau x, \quad V(y) = x^2 \sigma^2 \]

where the \( x_i \) are known but \( \sigma^2 \) is unknown. The transformed equations are

\[ E(x_i) = E\left(\frac{y_i}{x_i}\right) = \tau, \quad V(x_i) = \sigma^2 \left(\frac{1}{x_i^2} \right) \]

The normal equation is

\[ E(Q) = n \tau \]

where \( Q = \Sigma y \). The best estimate of \( \tau \) is

\[ \frac{Q}{n} = \bar{z} \quad V\left(\frac{Q}{n}\right) = \sigma^2 \]

For testing whether \( \tau = \xi \), an assigned quantity, the following analysis of sum of squares is needed.

\[ \bar{z} = \bar{x} \quad \frac{Q}{n} = \sigma^2 \]

In the following data are considered the variables \( y_i \), the dry weight of paddy, and \( x_i \), the green weight obtained from 25 samples. It is desired to test whether the conversion factor from green to dry is \( \frac{3}{4} \). The mean weight of dry paddy increases with the increase in green weight and so also the variance. The constancy of coefficient of variation is a plausible hypothesis, and therefore the method developed above is applicable.

\[ n = 25 \quad \sum \left(\frac{y_i}{x_i}\right) = 17.300 \quad \bar{z} = 0.692 \quad \Sigma x_i^2 = 11.9716 \]

\[ \Sigma (x_i - \bar{x})^2 = 0.2952 \]

\[ n(\bar{z} - \xi)^2 = (0.692 - 0.75)^2 \times 25 = 0.0841 \]

<table>
<thead>
<tr>
<th>Test for a Given Value of ( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.F.</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Residual</td>
</tr>
</tbody>
</table>
The ratio 6.837 with 1 and 24 degrees of freedom is significant at the 5% level, so we reject the hypothesis that the conversion factor is \( \frac{3}{4} \).

### 3d Tests of Hypotheses with a Single Degree of Freedom

#### 3d.1 Student’s t Test

Situations arise in which there is a single series of observations and it is desired to test whether the mean in the population is an assigned quantity \( \mu_0 \). If \( n \) is the size of the sample and \( \bar{z} \) the mean, the variance ratio appropriate to test the above hypothesis is

\[
F = \frac{n(\bar{z} - \mu_0)^2}{\sum z_i^2 - n\bar{z}^2}/(n - 1)
\]

with 1 and \((n - 1)\) degrees of freedom. When the sum of squares in the numerator has a single degree of freedom as above, tables * have been constructed for the statistic

\[
t = \frac{\sqrt{n}(\bar{z} - \mu_0)}{\sqrt{\sum z_i^2 - n\bar{z}^2}/(n - 1)}
\]

This statistic is also useful in testing whether the true mean value is above or below the hypothetical quantity \( \mu_0 \). This test was first proposed by Student,† who demonstrated that an exact test of significance is possible when the standard deviation is unknown. This epoch-making discovery is the starting point of the exact sampling theory as developed by R. A. Fisher.

#### 3d.2 Asymmetry of Right and Left Femora

The mean difference (right femur - left femur) in length between the right and left femora of 36 skeletons of a certain series is found to be 2.0234; the corrected sum of squares of these 36 differences is 418.0875. The estimated variance on 35 degrees of freedom is

\[
\frac{418.0875}{35} = 11.9625
\]

* See Table III in “Statistical Tables for Biological, Agricultural, and Medical Research,” by R. A. Fisher and F. Yates.

† This is the pen name under which W. S. Gosset wrote. The reader is referred to Student’s (1908) original paper where this test was first derived.
If the right and left femora are of equal length on the average, then the observed mean is a chance deviation from the true value, zero. The value of
\[ t = \frac{2.0234}{\sqrt{11.9625}} = 3.5193 \]
is significant at the 5% level so that on the basis of this test the lengths of the right and left femora cannot be considered equal on the average. Actually the probability of a tabulated \( t \) value corresponds to the probability that the absolute value of \( t \) (irrespective of sign) exceeds the tabulated value. If it is known a priori that the alternative hypothesis is that the right femur is longer than the left or if the purpose of the test is to discriminate only the asymmetry due to the bigger length of the right femur, then the sign of \( t \) is important. The hypothesis of equality is rejected in favor of the suggested alternative only when \( t \) exceeds the upper 5% value of \( t \) which corresponds to the 10% tabulated value of \( t \). If the alternative hypothesis is that the left femur has a greater length, then \((-t)\) should exceed the 10% tabulated value for significance. In the above example \( t \) certainly exceeds the upper 5% value of \( t \), showing that the data are in agreement with the suggested alternative that the right femur is longer than the left. If the suggested alternative is the other way, the null hypothesis could not be rejected on the basis of the test utilizing the lower 5% value of \( t \). In any problem the decision to use a two-sided or a one-sided test should be taken in advance before the analysis is undertaken considering the situation arising out of the problem at hand.

These tests are useful in situations where the mean values of two series are to be compared, but the observations are such that there is a one-to-one correspondence between a member of one series and a member of the other. In the above example two measurements belong to the same skeleton. The 36 pairs of measurements give rise to 36 differences which can now be treated as a single series in which the mean is expected to be zero.

On the other hand, the two series may not have any correspondence; for instance, no two measurements are made on the same skeleton, in which case the method of analysis of variance applied to groups (see 3e.1) has to be used. In the second test the variation due to skeletons has also to be taken into account, and therefore the precision of the comparison decreases and such a small difference as the above may go undetected, even in a large sample. The variance of femur length is about 400, in which case the variance for difference in means of two independent series of 36 is
\[ 400(\frac{1}{36} + \frac{1}{36}) = 22.22 \]
whereas the corresponding variance for the correlated series is 11.9625 + 36 = 0.33, which admits a more precise assessment of asymmetry. This aspect should be kept in view while conducting any investigation. When two measurements are to be compared, it may be designed to obtain two correlated series of measurements. The higher the correlation, the greater is the advantage. The association should be positive; otherwise the test based on correlated pairs becomes less efficient.

3e Analysis of Variance

3e.1 One-Way Classification

Let there be \( k \) samples of sizes \( n_1, \ldots, n_k \) from \( k \) populations with unknown means \( \mu_1, \ldots, \mu_k \) and with a common unknown variance \( \sigma^2 \). The hypothesis which may be desired to be tested is

\[
\mu_1 = \mu_2 = \cdots = \mu_k
\]

The observational equations \( n_1 + \cdots + n_k \) in number are given below

<table>
<thead>
<tr>
<th>First Sample</th>
<th>4th Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Variable</td>
<td>Expectation</td>
</tr>
<tr>
<td>( \chi_{11} )</td>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>( \chi_{21} )</td>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>( \chi_{31} )</td>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>( \chi_{41} )</td>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>Total ( T_1 )</td>
<td>( \cdots )</td>
</tr>
</tbody>
</table>

The minimum value of \( \sum (\chi_{ij} - \mu_i)^2 \) subject to the condition of the hypothesis is

\[
\sum (\chi_{ij} - \mu_i)^2 = \frac{(\sum \chi_{ij})^2}{n}
\]

\( n = n_1 + \cdots + n_k \)  \hspace{1cm} (3e.1.1)

which is the total corrected sum of squares of all the observations. The minimum value of \( \sum (\chi_{ij} - \mu_i)^2 \) without any restrictions is

\[
\sum_i \sum_j (\chi_{ij} - \bar{\chi})^2 = \frac{1}{n_k} \left( \sum_j \chi_{ij} \right)^2
\]

\hspace{3cm} (3e.1.2)

The sum of squares due to deviation from the hypothesis

\[
(3e.1.1) - (3e.1.2) = \frac{T_1^2}{n_1} + \cdots + \frac{T_k^2}{n_k} - \frac{T^2}{n} \hspace{1cm} (3e.1.3)
\]

\[
T = T_1 + \cdots + T_k
\]
is obtained from the totals for the samples only. This has \((k - 1)\) degrees of freedom. It is easier to calculate the expressions (3e.1.1) and (3e.1.3) in practice and derive the expression (3e.1.2) by subtraction. The scheme of computation is set out below.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation from hypothesis or between samples (k - 1)</td>
<td>(\frac{T_1^2}{n_1} - \frac{T_2^2}{n_2})</td>
</tr>
<tr>
<td>Residual</td>
<td>(\frac{T_3^2}{n_3})</td>
</tr>
<tr>
<td>Total</td>
<td>(n - 1)</td>
</tr>
</tbody>
</table>

The quantities marked by * are obtained by subtraction. The \(F\) statistic is constructed using the mean squares derived from the above table.

The following data relate to the head breadths of 142 skulls belonging to three series. Can the mean head breadth be considered the same in the three series?

<table>
<thead>
<tr>
<th>Series</th>
<th>Sample Size</th>
<th>Head Breadth Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83</td>
<td>11,277</td>
<td>135.87</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td>7,049</td>
<td>138.22</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>1,102</td>
<td>137.75</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>19,428</td>
<td>136.817</td>
</tr>
</tbody>
</table>

The sum of squares between series is

\[
11,277 \times 135.87 + 7,049 \times 138.22 + 1,102 \times 137.75 - 19,428 \times 136.817 = 238.59
\]

The total sum of squares is found to be 4616.64. The analysis of variance is set out in Table 3e.1p.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between series</td>
<td>2</td>
<td>238.59</td>
<td>119.29</td>
</tr>
<tr>
<td>Residual</td>
<td>139</td>
<td>4378.05</td>
<td>31.50</td>
</tr>
<tr>
<td>Total</td>
<td>141</td>
<td>4616.64</td>
<td></td>
</tr>
</tbody>
</table>

The variance ratio 3.79 with 2 and 139 degrees of freedom is significant at the 5% level so that the mean values cannot be considered equal in all the series.
When between series has 1 degree of freedom the square root of $F$ can be referred to the $t$ distribution with degrees of freedom of the residual. If the two series can be distinguished as the first and the second, then $t$ can be given the same sign as the difference between the averages of the first and second series. Then as in 3d.1 it is possible to test whether the mean of the first series significantly exceeds the other, and vice versa.

3e.2 Two-Way Classification with a Single Observation in a Cell

Let there be $pq$ observations, each of which can be specified in terms of the categories of two classes. The observations may be set out in the following tabular form.

<table>
<thead>
<tr>
<th>Class A</th>
<th>Class B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B_1$</td>
<td>$B_2$</td>
</tr>
<tr>
<td>$A_1$</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$x_{21}$</td>
<td>$x_{22}$</td>
</tr>
<tr>
<td>$A_q$</td>
<td>$x_{q1}$</td>
<td>$x_{q2}$</td>
</tr>
<tr>
<td>Total</td>
<td>$x_{.1}$</td>
<td>$x_{.2}$</td>
</tr>
</tbody>
</table>

The observational equations are known to be

$$E(x_{ij}) = \alpha_i + \beta_j \quad V(x_{ij}) = \sigma^2 \quad (3e.2.1)$$

The $\alpha$ and $\beta$ parameters may be called the effects of the categories in the A and B classes. When only a single observation is present in each cell, it is not possible to test whether the additive setup assumed in (3e.2.1) is correct. If, however, this can be taken to be true, two hypotheses which may be tested from these data are

$$\alpha_1 = \alpha_2 = \cdots = \alpha_q$$
$$\beta_1 = \beta_2 = \cdots = \beta_p \quad (3e.2.2)$$

It is easy to see that the rank of the matrix of equations (3e.2.1) is $(p + q - 1)$ so that not all parametric functions can be estimated. As a matter of fact, the individual parameters $\alpha_i$ and $\beta_j$ are not estimable. But the differences of $\alpha_i$ or $\beta_j$ which are relevant to the hypotheses (3e.2.2) to be tested are estimable. The residual sum of squares is the minimum value of $\sum (x_{ij} - \alpha_i - \beta_j)^2$ with $pq - (p + q - 1)$ degrees
of freedom. This is obtained as
\[ \Sigma (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}.)^2 \]  
(3e.2.3)
The minimum value of the sum of squares \[ \Sigma (x_{ij} - a_i - \beta_j)^2 \] subject to the restriction \( a_1 = \cdots = a_p \) is
\[ \Sigma (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}..) + \frac{1}{p} \Sigma x_i.^2 - \frac{1}{p} x..^2 \]  
(3e.2.4)
with \((pq - p)\) degrees of freedom. The sum of squares between the categories of class \( A \), obtained by subtraction, is
\[ \frac{1}{p} \Sigma x_i.^2 - \frac{1}{pq} x..^2 \]  
(3e.2.5)
with \((q - 1)\) degrees of freedom. The sum of squares due to the categories of class \( B \) is similarly
\[ \frac{1}{q} \Sigma x_j.^2 - \frac{1}{pq} x..^2 \]  
(3e.2.6)
with \((p - 1)\) degrees of freedom. The expressions (3e.2.5) and (3e.2.6) are easy to compute, and the expression (3e.2.3) can be deduced from the equality
\[ \Sigma x_{ij}^2 = x..^2 \]  
- the sum of the expressions (3e.2.3), (3e.2.5), and (3e.2.6)
The scheme of computation is presented below.

| Table 3e.2a. Analysis of Variance, Two-Way Classification |
|-----------------|-----------------|
| D.F.            | S.S.            |
| Between A classes | \( q - 1 \) | \( \frac{1}{p} \Sigma x_i.^2 - \frac{1}{pq} x..^2 \) |
| Between B classes | \( p - 1 \)  | \( \frac{1}{q} \Sigma x_j.^2 - \frac{1}{pq} x..^2 \) |
| Residual        | \((p - 1)(q - 1)\) | \* |
| Total           | \( pq - 1 \)  | \( \Sigma x_{ij}^2 - \frac{1}{pq} x..^2 \) |

* Obtained by subtraction.

The analysis is similar for more complex classifications. The numerical methods of the analysis are given in 3g.3.

3e.3 Two-Way Classification with Multiple but Equal Numbers in Cells
In the last section it was shown that the differences in class effects can be tested when the effects due to the classes are known to be addi-
EQUAL NUMBERS IN CELLS

tive. This can, however, be tested when each cell contains more than
one observation. If there are \( n \) observations in the \((i,j)\)th cell, they
may be represented by

\[ x_{ij1}, x_{ij2}, \ldots, x_{ijn} \]

with a total \( x_{ij} \) and mean \( \bar{x}_{ij} \). The observational equations are

\[ E(x_{ijk}) = \alpha_{ij} \quad V(x_{ijk}) = \sigma^2 \]

The hypothesis to be tested is

\[ \alpha_{ij} = \alpha_i + \beta_j \]

If this is not true, there is said to be interaction, in which case the test
of significance for the class effects cannot be properly interpreted. Dif-
ferences in \( A \) classes might be tested for each \( B \) class. The magnitudes
of these will depend on the nature of the \( B \) class considered. The
residual sum of squares with \( pq(n - 1) \) degrees of freedom is

\[
\min \sum \sum (x_{ijk} - \alpha_{ij})^2 = \sum \sum \sum (x_{ijk} - \bar{x}_{ij})^2
\]

treating \( \alpha_{ij} \) as free parameters. The sums of squares due to the inter-
action and the \( A \) or \( B \) classes are derivable as in 3e.2 by considering
the totals \( x_{ij} \) in each cell as single observations and dividing the final
expressions of sums of squares by \( n \) to reduce to the scale of the original
observations. By considering the obvious identity relations, the scheme
of computation is given in Table 3e.3a.

### TABLE 3e.3a. Analysis of Variance, Two-Way Classification

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between ( A ) classes</td>
<td>( q - 1 ) [ \frac{1}{n} \sum_{p} \sum_{i} z_{i}^2 - \frac{1}{npq} \bar{x}^2 ]</td>
</tr>
<tr>
<td>Between ( B ) classes</td>
<td>( p - 1 ) [ \frac{1}{n} \sum_{q} \sum_{j} z_{j}^2 - \frac{1}{npq} \bar{x}^2 ]</td>
</tr>
<tr>
<td>Interaction</td>
<td>((p - 1)(q - 1)) *</td>
</tr>
<tr>
<td>Between ( pq ) cells</td>
<td>( pq - 1 ) [ \frac{1}{n} \sum \sum z_{ij}^2 - \frac{1}{npq} \bar{x}^2 ]</td>
</tr>
<tr>
<td>Residual</td>
<td>( pq(n - 1) ) *</td>
</tr>
<tr>
<td>Total</td>
<td>( npq - 1 ) [ \sum \sum \sum z_{ijk}^2 - \frac{1}{npq} \bar{x}^2 ]</td>
</tr>
</tbody>
</table>
The sum of squares indicated by * are to be filled in by subtraction. If the interaction is significant when tested against the residual, it may be necessary to test for differences in A classes for every B class, or vice versa. The sum of squares with \((q - 1)\) degrees of freedom due to A classes for the \(j\)th class of B is

\[
\frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 - \frac{1}{nq} \bar{x}_j^2
\]

The mean square corresponding to this is tested against the residual mean square.

If the interaction is not significant, the first two entries of Table 3e.3a can be tested against the residual mean square or the interaction mean square, whichever is greater. This is to guard against any bias due to small effects of interaction which could not be detected by the test but which is indicated when the interaction mean square exceeds the error. When both the mean squares are of the same magnitude, a common estimate can be obtained by adding the degrees of freedom and the sum of squares corresponding to interaction and error.

### 3e.4 Two-Way Classification with Unequal Numbers in Cells

The following notation is used.

- \(x_{ij}\) = the total of all observations in the \((i, j)\)th cell.
- \(\bar{x}_{ij}\) = mean in the \((i, j)\)th cell.
- \(x_j = \sum x_{ij}\) (total for the \(j\)th column).
- \(x_i = \sum x_{ij}\) (total for the \(i\)th row).
- \(\bar{x} = \sum x_{ij}\) (total of all observations).
- \(\bar{x} = \text{mean of all observations.}\)
- \(n_{ij}, n_j, n_i, \text{ and } n\) are the numbers of observations for the \((i, j)\)th cell, \(j\)th column, \(i\)th row, and all the cells, respectively.

The data in Table 3e.4a present the mean values and totals of nasal height of skulls excavated from three different strata by three observers. It is desired to test for the stratum and observer differences.

In problems of this nature it is convenient to set up the figures as in Table 3e.4a for the computation of the various sums of squares. The analysis is carried out in three stages. The total sum of squares with 309 degrees of freedom is found to be 5398.4206. For further calculations the entries in the above table are sufficient.

#### A. The Computation of Between-Cell Sum of Squares

The between-cell sum of squares with \(p q - 1 = 8\) degrees of freedom is \(\Sigma \Sigma x_{ij}^2 - x.. \bar{x}..\) = 931.5204.
UNEQUAL NUMBERS IN CELLS

<table>
<thead>
<tr>
<th>Observer</th>
<th>Strata</th>
<th>XI</th>
<th>(x_i)</th>
<th>(x_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O_1)</td>
<td>(S_1)</td>
<td>1,071.00</td>
<td>1,572.48</td>
<td>913.50</td>
</tr>
<tr>
<td>(O_1)</td>
<td>(S_2)</td>
<td>51.00</td>
<td>60.14</td>
<td>89.85</td>
</tr>
<tr>
<td>(O_1)</td>
<td>(S_3)</td>
<td>32 (21)</td>
<td>18 (32)</td>
<td>50 (18)</td>
</tr>
</tbody>
</table>

| \(O_2\)  | \(S_1\) | 1,965.86 | 2,315.40 | 1,721.52 | 6,003.78 |
| \(O_2\)  | \(S_2\) | 46.83 | 45.40 | 47.82 | 139.05 |
| \(O_2\)  | \(S_3\) | 51 (42) | 51 (51) | 50 (36) | 152 (129) |

| \(O_3\)  | \(S_1\) | 1,219.90 | 2,091.60 | 1,840.20 | 5,151.70 |
| \(O_3\)  | \(S_2\) | 48.76 | 46.48 | 46.23 | 141.47 |
| \(O_3\)  | \(S_3\) | 45 (25) | 45 (45) | 40 (40) | 130 (90) |

| \(x.\) | 4,256.86 | 5,979.48 | 4,484.22 | 14,720.56 |
| \(x.\) | 48.373 | 46.715 | 47.704 | 147.892 |

B. The Computation of the Interaction Sum of Squares. When there are only two classes for A or B, the computation of the interaction sum of squares is very simple.

<table>
<thead>
<tr>
<th>B1</th>
<th>B2</th>
<th>Bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>(z_{11})</td>
<td>(z_{12})</td>
<td>(z_{1p})</td>
</tr>
<tr>
<td>(z_{21})</td>
<td>(z_{22})</td>
<td>(z_{2p})</td>
</tr>
</tbody>
</table>

\[
\text{Difference: Weights} = n_{11} \cdot z_{11} + n_{12} \cdot z_{12} + \cdots + n_{1p} \cdot z_{1p} = \Sigma w_i \cdot z_i
\]

\[
\text{Difference \times Weight} = w_{11} \cdot z_{11} + w_{12} \cdot z_{12} + \cdots + w_{1p} \cdot z_{1p} = \Sigma w_i \cdot z_i
\]

Note: For obtaining \(w_i^2\), we need to multiply \(w_i\) by \(y_i\).
The interaction sum of squares with \((2 - 1)(p - 1)\) degrees of freedom is

\[
\sum w_{ij}^2 = \frac{(\sum w_{ij})^2}{\sum w_i}
\]

In the general case, for any number of A and B classes the absence of interaction means that

\[
a_{ij} + a_{i'j'} - a_{ij'} - a_{i'j} = 0
\]

for all \(i \neq i', j \neq j'\), where \(a_{ij}\) is the expected value in the \((i, j)\)th cell.

The best estimate of this tetrad difference is

\[
\hat{a}_{ij} + \hat{a}_{i'j'} - \hat{a}_{ij'} - \hat{a}_{i'j}
\]

To obtain the sum of squares due to interaction we can directly consider these functions and derive the suitable sum of squares.

There are \((p - 1)(q - 1)\) such independent functions whose variances and covariances can be easily written down. Consider, for instance, a \(3 \times 3\) table with the mean values

\[
\begin{array}{ccc}
\bar{x}_{11} & \bar{x}_{12} & \bar{x}_{13} \\
\bar{x}_{21} & \bar{x}_{22} & \bar{x}_{23} \\
\bar{x}_{31} & \bar{x}_{32} & \bar{x}_{33}
\end{array}
\]

The following four functions indicated by the tetrad differences may be taken

\[
\begin{align*}
y_1 &= x_{11} + x_{22} - x_{21} - x_{12} = 0.43 \\
y_2 &= -x_{22} + x_{12} + x_{23} - x_{13} = 0.81 \\
y_3 &= -x_{22} + x_{21} + x_{32} - x_{31} = -0.85 \\
y_4 &= +x_{22} - x_{23} - x_{32} + x_{31} = -2.67
\end{align*}
\]

If \(\sigma^2(a_{ij})\) is the covariance matrix of \(y_1, y_2, y_3, y_4\), then

\[
\begin{align*}
a_{11} &= \frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} = 0.122286 \\
a_{12} &= -\frac{1}{n_{22}} - \frac{1}{n_{12}} = -0.050858 \\
a_{13} &= -\frac{1}{n_{22}} - \frac{1}{n_{21}} = -0.043417
\end{align*}
\]
UNEQUAL NUMBERS IN CELLS

\[ a_{14} = \frac{1}{n_{22}} = 0.019008 \]
\[ a_{22} = \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{23}} + \frac{1}{n_{13}} = 0.134192 \]
\[ a_{23} = \frac{1}{n_{22}} = 0.019008 \]
\[ a_{24} = -\frac{1}{n_{22}} - \frac{1}{n_{23}} = -0.047386 \]
\[ a_{33} = \frac{1}{n_{22}} + \frac{1}{n_{21}} + \frac{1}{n_{32}} + \frac{1}{n_{31}} = 0.105639 \]
\[ a_{34} = -\frac{1}{n_{22}} - \frac{1}{n_{32}} = -0.041830 \]
\[ a_{44} = \frac{1}{n_{22}} + \frac{1}{n_{23}} + \frac{1}{n_{32}} + \frac{1}{n_{33}} = 0.094008 \]

Table 3e.4y contains the required computations. The matrix \( a_{ij} \) is written with an extended column of \( y \) and reduced by the method of pivotal condensation (1d.1).

The elements below the diagonal are omitted because the matrix is symmetrical at each stage. It is sometimes necessary to retain a large number of decimal places to obtain sufficient accuracy in the final value. More examples illustrating this computational scheme are given in Chapter 7.

The last pivotal value with the sign changed, 125.5049, is the sum of squares for interaction. An alternative way of calculating the interaction sum of squares is by fitting constants by the method of least squares. We need to find the minimum value of

\[ \sum \sum n_{ij} (\bar{x}_{ij} - \alpha_i - \beta_j)^2 \]

If \( \alpha_i \) and \( \beta_j \) are the optimum values, then the minimum value is

\[ \sum \frac{z_{ij}^2}{n_{ij}} - \sum \alpha_i \bar{x}_i - \sum \beta_j \bar{x}_j \]
### Table 3e.4v. Method of Pivotal Condensation for Interaction

#### Sum of Squares

<table>
<thead>
<tr>
<th>Matrix $a_{ij}$</th>
<th>Value of $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.12286$</td>
<td>0.43</td>
</tr>
<tr>
<td>$-0.030868$</td>
<td>$-0.019468$</td>
</tr>
<tr>
<td>$-0.019288$</td>
<td>$-0.047288$</td>
</tr>
<tr>
<td>$0.041830$</td>
<td>0.85</td>
</tr>
<tr>
<td>$0.094608$</td>
<td>2.67</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$1$</td>
<td>$-0.415894$</td>
</tr>
<tr>
<td>$-0.335045$</td>
<td>$0.100345$</td>
</tr>
<tr>
<td>3.546347</td>
<td></td>
</tr>
<tr>
<td>$0.113040$</td>
<td>0.988834</td>
</tr>
<tr>
<td>$0.001531$</td>
<td>$-0.032023$</td>
</tr>
<tr>
<td>$-0.034808$</td>
<td>$-0.097351$</td>
</tr>
<tr>
<td>$-0.094646$</td>
<td>2.738949</td>
</tr>
<tr>
<td>1.512029</td>
<td></td>
</tr>
<tr>
<td>$1$</td>
<td>0.013721</td>
</tr>
<tr>
<td>$-0.347054$</td>
<td>8.747647</td>
</tr>
<tr>
<td>$0.002030$</td>
<td>0.710909</td>
</tr>
<tr>
<td>$-0.034330$</td>
<td>2.205770</td>
</tr>
<tr>
<td>$-0.077849$</td>
<td>10.162000</td>
</tr>
<tr>
<td>$-0.380386$</td>
<td>7.881102</td>
</tr>
<tr>
<td>$-0.064783$</td>
<td>2.666328</td>
</tr>
<tr>
<td>15.764667</td>
<td></td>
</tr>
<tr>
<td>41.15783</td>
<td></td>
</tr>
<tr>
<td>$-125.504044$</td>
<td></td>
</tr>
</tbody>
</table>

The optimum values are obtained from the equations:

$$a_1 a_2 a_3 b_1 b_2 b_3 \text{ Total}$$

<table>
<thead>
<tr>
<th>Marginal</th>
<th>71</th>
<th>129</th>
<th>110</th>
<th>21</th>
<th>32</th>
<th>18</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>21</td>
<td>32</td>
<td>18</td>
<td>3556.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>42</td>
<td>51</td>
<td>36</td>
<td>6003.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>25</td>
<td>45</td>
<td>40</td>
<td>5159.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>42</td>
<td>25</td>
<td>88</td>
<td>4250.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>51</td>
<td>45</td>
<td>128</td>
<td>5679.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>40</td>
<td>94</td>
<td>4484.22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The method of writing down these equations is very simple. Start with any marginal total, say 3556.98, based on 71 observations distributed
UNEQUAL NUMBERS IN CELLS

in the B classes as 21, 32, and 18. This gives the first equation. There are six marginal totals corresponding to A and B classes, giving rise to six equations.

The method of solution is also simple. First reduce the first three equations by making the coefficients of $a_1$, $a_2$, $a_3$ equal to unity. The value of $b_2$ can be assumed to be zero so that the column corresponding to $b_2$ (the last constant) may be omitted.

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.295775</td>
<td>0.450704</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>0.325581</td>
<td>0.385349</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0.227273</td>
<td>0.409091</td>
</tr>
</tbody>
</table>

Subtracting from the fourth row of the original equations 21 (first row above) + 42 (second row above) + 25 (third row above), and similarly from the fifth row subtracting 32 (first row) + 51 (second row) + 45 (third row), we obtain

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>62.432498</td>
<td>36.296717</td>
</tr>
<tr>
<td>-</td>
<td>-36.296717</td>
<td>64.844662</td>
</tr>
</tbody>
</table>

which give simultaneous equations in $b_1$, $b_2$ with solutions:

$b_1 = 0.539250$  $b_2 = -1.170335$  $b_3 = 0$

Substituting them in the first three reduced equations with unit coefficients for $a_1$, $a_2$, $a_3$, we find the values of $a_1$, $a_2$, $a_3$ to be:

$a_1 = 50.490372$  $a_2 = 46.821539$  $a_3 = 47.258943$

The minimum sum of squares is then computed:

Uncorrected between-cell sum of squares

$-a_2(3556.98) - a_2(6003.78) - \cdots$

$-b_1(1256.86) - b_1(5979.48) - \cdots$

$= 125.4490$

The method of solving the above equations is quite general. First, omit one $b$ if the number of $b$ classes is greater than or equal to that of $a$, and reduce the coefficients of $a$ to unity in the first set and eliminate $a$ from the latter set. The resulting equations contain $b$ only, which may be directly solved. Also, the last equation may be omitted or it may be retained as a check, since the column sums should be zero. The $a$ coefficients can be obtained by substitution for $b$ in the first set.
of reduced equations. The value 125.4490 for interaction sum of squares agrees with that obtained earlier only to three significant figures. This is the order of accuracy expected in computations of this nature unless, of course, the calculations are carried to a large number of decimal places. This is unnecessary because the data, to start with, may not be so accurate, in which case there is a limit to the number of significant figures in any computed value. Even in the above case the number of decimal places retained could be reduced, but it is a good arithmetical discipline to assume that the original figures are absolutely correct and carry out the computations to as many decimal places as can be conveniently retained. In this particular case the average of the two values is 125.4769, which is used in Table 3e.44. In general, the first method needs the reduction of a matrix of \((p - 1)(q - 1) + 1\)th order; the second leads to a solution of simultaneous equations of order \((p + q - 1)\). The latter method may be relatively simpler when \(p\) and \(q\) exceed 3.

C. The Computation of Main Effects. If the interaction is significant, then the problem reduces to testing the differences in each row or in each column of the two-way table. On the other hand, when the interaction is not significant, the main effects may be tested by considering all the table entries.

As shown above, the interaction sum of squares is the minimum value of \(\sum n_{ij}(x_{ij} - \alpha_i - \beta_j)^2\) when minimized with respect to \(\alpha\) and \(\beta\). If now this quantity is minimized with the further restriction that \(\alpha_1 = \ldots = \alpha_p\), it is easily seen that the minimum value is

\[
\left\{ \sum \left( \frac{x_{i.}^2}{n_{ij}} \right) - \frac{x^2 \ldots}{n} \right\} - \left\{ \sum \left( \frac{x_{..j}^2}{n_{ij}} \right) - \frac{x^2 \ldots}{n} \right\}
\]

This can be recognized as the total sum of squares between the \(pq\) cells minus the sum of squares between the \(B\) classes, ignoring the classification due to \(A\). If the interaction sum of squares is subtracted from this, the valid sum of squares for testing the differences in \(A\) classes is obtained. The scheme of computation is shown in Table 3e.44.

| TABLE 3e.44. Sum of Squares for A Classes |
|-----------------|-----------------|-----------------|
| D.F.            | S.S.            |                  |
| Between B classes ignoring A | \(p - 1\)       | \(\sum x_{i.}^2 - x^2 \ldots\) |
| Interaction     | \((p - 1)(q - 1)\) | (As obtained in stage B) |
| Between A classes| \(q - 1\)       | *                |
| Total between cells | \(pq - 1\)     | \(\sum x_{ij}^2 - x^2 \ldots\) |

* Obtained by subtraction.
Similarly, the valid sum of squares between B classes is obtained. The mean squares obtained from each of these sums of squares can be tested against the residual or the interaction mean square, whichever is greater. This completes the analysis.

The final analysis of variance is given in Table 3e.4e.

**Table 3e.4e. Complete Analysis of Variance for the Two-Way Data**

<table>
<thead>
<tr>
<th></th>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>S.S.</th>
<th>D.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata, ignoring</td>
<td>2</td>
<td>147.6319</td>
<td>73.8159</td>
<td>317.0758</td>
<td>2</td>
</tr>
<tr>
<td>observers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>4</td>
<td>125.4769</td>
<td>31.3692</td>
<td>125.4769</td>
<td>4</td>
</tr>
<tr>
<td>Observers</td>
<td>2</td>
<td>658.4116</td>
<td>329.2058</td>
<td>85.9459</td>
<td>2</td>
</tr>
<tr>
<td>Between cells</td>
<td>8</td>
<td>931.5204</td>
<td>116.4405</td>
<td>931.5204</td>
<td>8</td>
</tr>
<tr>
<td>Within cells</td>
<td>301</td>
<td>4466.0002 *</td>
<td>14.8402</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>309</td>
<td>5398.4206</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Obtained by subtraction.

The variance ratio for interaction is

\[
\frac{31.3692}{14.8402} = 2.11
\]

which with 4 and 301 degrees of freedom is not significant. The interaction mean square is greater than the within-cell mean square and is therefore used in testing for observer and stratum differences. The variance ratio for observers is

\[
\frac{329.2058}{31.3692} = 10.49
\]

which with 2 and 4 degrees of freedom is significant at the 5% level. The variance ratio for strata is

\[
\frac{85.9459}{31.3692} = 2.74
\]

which with 2 and 4 degrees of freedom is not significant.
Thus the discrepancy in the mean values can be traced to observer differences. Probably the observers used different techniques of measuring nasal height.

The calculations become very simple when the unequal cell numbers are proportionate so that

\[ n_{ij} = \frac{n_{ij}}{n_{..}} \]

In this case the main effects of A and B can be calculated in the usual manner from the marginal totals. Thus the sum of squares due to A is

\[ \sum \frac{x_1^2 - \bar{x}_1^2}{n_{..}} \]

The interaction sum of squares is obtained by subtracting the sum of squares due to A and B from the total for the pq cells.

Suppose that we ignore the observers and look for stratum differences. The analysis will then be in the nature of

<table>
<thead>
<tr>
<th>D.F.</th>
<th>M.S.</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between strata</td>
<td>2</td>
<td>73.8159</td>
</tr>
<tr>
<td>Within strata</td>
<td>307</td>
<td>17.1035</td>
</tr>
</tbody>
</table>

which gives a significant ratio of 4.31, leading to the conclusion that three different groups of people inhabited the three different strata. A closer analysis reveals that observer differences are also important so that some caution is necessary in combining the results of the three different investigators.

3f The Theory of Statistical Regression

3f.1 The Concept of Regression

The theory of regression is concerned with the derivation of the relationship between a set of variables \( x_1, \ldots, x_p \) and the mean value of another variable \( y \) observable with them. The variables \( x_1, \ldots, x_p \) are called concomitant variables, and \( y \) the dependent variable. The equation

\[ m = R(x_1, \ldots, x_p) \]

giving \( m \) the mean value of \( y \) for given \( x \), as a function of \( x \), is called the regression equation of \( y \) on \( x_1, \ldots, x_p \). It is customary to write the above equation as

\[ Y = R(x_1, \ldots, x_p) \]
Various uses of the regression are discussed below with suitable examples.

3f.2 Prediction of Cranial Capacity

One of the uses of the regression equation is for the prediction of the dependent variate for a given set of concomitant variates. For instance, a skull may be broken so that the actual cranial capacity cannot be determined. In such a case the capacity may be predictable if at least some external measurements are available. This requires the construction of the regression equation between the cranial capacity and the observed set of external measurements on complete skulls.

The Regression Equation. Three important measurements from which the cranial capacity ($C$) may be predicted are the glabella-occipital length ($L$), the maximum parietal breadth ($B$), and the basio-bregmatic height ($H'$). Since the magnitude to be estimated is a volume, it is appropriate to set up a regression formula of the type

$$C = \alpha' L^\beta_1 B^\beta_2 H'^\beta_3$$

where $\alpha'$, $\beta_1$, $\beta_2$, and $\beta_3$ are the constants to be estimated. By transforming the variables to

$$y = \log_{10} C \quad x_1 = \log_{10} L \quad x_2 = \log_{10} B \quad x_3 = \log_{10} H'$$

the formula can be written

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where $\alpha = \log_{10} \alpha'$. From this equation the constants are estimated by the method of least squares.

Estimation of the Constants. Using the measurements on the 86 male skulls from the Farringdon Street series (Hooke, 1926), we find the mean values

$$\bar{y} = 3.1685 \quad \bar{x}_1 = 2.2752 \quad \bar{x}_2 = 2.1523 \quad \bar{x}_3 = 2.1128$$

The corrected sums of the products matrix ($S_{ij}$) for $x_1$, $x_2$, $x_3$ is

\[
\begin{array}{ccc}
0.01875 & 0.00848 & 0.00684 \\
0.00848 & 0.02904 & 0.00878 \\
0.00684 & 0.00878 & 0.02886 \\
\end{array}
\]

The corrected sums of products of $y$ with $x_1$, $x_2$, and $x_3$ are, respectively,

$$Q_1 = 0.03030 \quad Q_2 = 0.04410 \quad Q_3 = 0.03629$$
The reciprocal of the matrix \((S_{ij})\) denoted by \((C_{ij})^*\) is obtained by the method of Id.1.

\[
\begin{bmatrix}
64.21 & -15.57 & -10.49 \\
-15.57 & 41.71 & -9.00 \\
-10.49 & -9.00 & 39.88
\end{bmatrix}
\]

The estimates of the parameters are

\[
\begin{align*}
b_1 &= 64.21Q_1 - 15.57Q_2 - 10.49Q_3 = 0.878 \\
b_2 &= -15.57Q_1 + 41.71Q_2 - 9.00Q_3 = 1.041 \\
b_3 &= -10.49Q_1 - 9.00Q_2 + 39.88Q_3 = 0.733 \\
a &= y - b_1x_1 - b_2x_2 - b_3x_3 = -2.618
\end{align*}
\]

The formula for the prediction of cranial capacity \(C\) is

\[
C = 0.00241L_1^{0.878}P_1^{0.041}H^{0.733}
\]

Since the estimate of \(\beta_i\) is \(C_1Q_1 + C_2Q_2 + C_3Q_3\), it follows from (3a.6.5) and (3a.6.6) that

\[
V(b_i) = C_i\sigma^2 \quad \text{and} \quad \text{cov}(b_i, b_j) = C_{ij}\sigma^2
\]

where \(C_{ij}\) are elements of the above matrix.

Tests of Hypotheses. Having estimated these constants, it is relevant to examine how far the concomitant variables are helpful in prediction. If these variables are of no use, then the prediction formula does not depend on them so that \(\beta_1 = \beta_2 = \beta_3 = 0\). This hypothesis may be tested from the above data.

The residual sum of squares with \((n - 4)\) degrees of freedom is the minimum value of

\[
\sum(y_i - \alpha - \beta_1x_{1i} - \beta_2x_{2i} - \beta_3x_{3i})^2
\]

which is

\[
(\Sigma y^2 - n\bar{y}^2) - b_1Q_1 - b_2Q_2 - b_3Q_3 \\
= 0.12692 - 0.878(0.03030) - 1.041(0.04410) - 0.733(0.03629) \\
= 0.12692 - 0.09911 = 0.02781
\]

*To be consistent with the matrix notation the reciprocal \((S_{ij})\) should be written \((S'_{ij})\). In statistical literature this is already known as the \(C\) matrix with the elements \(C_{ij}\).

†The capacity of the Farringdon Street series skulls was determined by tight packing with mustard seed and weighing in the manner described by Macdonell (1904). The formula is applicable only for predicting capacity determined in this way.
PREDICTION OF CRANIAL CAPACITY

If the hypothesis \( \beta_1 = \beta_2 = \beta_3 = 0 \) is true, then the minimum value of \( \sum (y_i - \alpha)^2 \) is \( \sum y_i^2 - \frac{n\bar{y}^2}{n} = 0.12692 \), which is the total sum of squares with \( (n - 1) \) degrees of freedom. The reduction in the above sum of squares is due to regression. The analysis of the sum of squares is shown in Table 3f.2a.

<table>
<thead>
<tr>
<th></th>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>0.09911</td>
<td>0.03303</td>
<td>97.41</td>
</tr>
<tr>
<td>Residual</td>
<td>82</td>
<td>0.02781</td>
<td>0.0003391</td>
<td></td>
</tr>
<tr>
<td>( \sum y_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_i )</td>
<td>85</td>
<td>0.12692</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The variance ratio 97.41 with 3 and 82 degrees of freedom is significant at the 1% level, which shows that the variables considered above are useful in prediction.

It may now be examined whether the three linear dimensions appear to the same degree in the prediction formula. From the estimates it is seen that the index \( \beta_2 \) for maximum parietal breadth is higher than the others. This means that a given ratio of increase in breadth counts more for capacity than the corresponding increase in length or height.

The hypothesis relevant to examine this point is

\[ \beta_1 = \beta_2 = \beta_3 = \beta \] (say)

The minimum value of \( \sum (y_i - \alpha - \beta(x_{1i} + x_{2i} + x_{3i}))^2 \) has to be found out. The normal equation giving the estimate of \( \beta \) is

\[ (S_{11} + S_{22} + S_{33} + 2(S_{12} + S_{23} + S_{31}))b = Q = (Q_1 + Q_2 + Q_3) \]

\[ 0.12485b = 0.11069 \]

\[ b = 0.8866 \]

The minimum value with \( (n - 2) \) degrees of freedom is

\[ (\sum y_i^2 - \frac{n\bar{y}^2}{n}) - bQ = 0.12692 - 0.09814 = 0.02878 \]

<table>
<thead>
<tr>
<th></th>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation from equality</td>
<td>2</td>
<td>0.00097</td>
<td>0.000485</td>
<td>1.430</td>
</tr>
<tr>
<td>Residual</td>
<td>82</td>
<td>0.02781</td>
<td>0.0003391</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>0.02878</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The ratio is not significant, so there is no evidence from the data to conclude that $\beta_1$, $\beta_2$, $\beta_3$ are different. The differences, if any, are likely to be small, and a large collection of measurements may be necessary before anything definite can be said about this. Evolutionists believe that the breadth is increasing relatively more than any other magnitude on the skull. If this is true it is of interest to examine how far the cranial capacity is influenced by the breadth.

So far as the problem of prediction is concerned, the formula

$$ C = 0.002342(LBH)^{0.8865} $$

obtained by assuming $\beta_1 = \beta_2 = \beta_3$, may be as useful as the formula derived without assuming that these are equal. The variance of the estimate $b$ of $\beta$ is $\sigma^2/\Sigma S_{ij}$ where $\sigma^2$ is the estimate based on 84 degrees of freedom with the corresponding sum of squares given in Table 3f.2a.

A simple formula of the type $C = \alpha'LBH'$ is sometimes used for predicting the cranial capacity. A test of the adequacy of such a formula is equivalent to testing the hypothesis

$$ \beta_1 = \beta_2 = \beta_3 = 1 $$

The minimum value of $\Sigma(y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i})^2$, assuming this to be true, is

$$ (\Sigma y_i^2 - n\bar{y}^2) + S_{11} + S_{22} + S_{33} + 2(S_{12} + S_{23} + S_{31}) $$

$$ -2(Q_1 + Q_2 + Q_3) = 0.03039 $$

which has $(n - 1)$ degrees of freedom. The residual has $(n - 4)$ degrees of freedom so that the difference with 3 degrees of freedom is due to deviation from the hypothesis.

**Table 3f.2y. Test of the Hypothesis $\beta_1 = \beta_2 = \beta_3 = 1$**

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation from $\beta_1 = \beta_2 = \beta_3 = 1$</td>
<td>3</td>
<td>0.00258</td>
</tr>
<tr>
<td>Residual</td>
<td>82</td>
<td>0.002781</td>
</tr>
<tr>
<td>Total</td>
<td>85</td>
<td>0.03039</td>
</tr>
</tbody>
</table>

The ratio 2.544 with 3 and 82 degrees of freedom is just below the 5% significance level. It would be of interest to examine this point with more adequate material.

In Table 3f.2y the sum of squares due to deviation from the hypothesis could be directly calculated from the formula, providing a compound
measure of the deviations of the estimates \(b_1, b_2, b_3\) from the expected values 1, 1, 1.

\[
\begin{align*}
\sum \sum S_{ij} (b_i - 1)(b_j - 1) &= (b_1 - 1)[S_{11}(b_1 - 1) + S_{12}(b_2 - 1) + S_{13}(b_3 - 1)] + \cdots \\
&= (b_1 - 1)(Q_1 - S_{12} - S_{13} - S_{11}) + \cdots \\
&= b_1 Q_1 + b_2 Q_2 + b_3 Q_3 - 2(Q_1 + Q_2 + Q_3) + \sum \sum S_{ij} \\
&= 0.09911 - 2(0.11069) + 0.12485 \\
&= 0.00258
\end{align*}
\]

which is the same as that given in Table 3f.2y.

Having found evidence that the \(\beta\) coefficients differ individually from unity, it is of some interest to examine whether the indices add up to 3 while distributing unequally among the three dimensions used. This requires the test of the hypothesis \(\beta_1 + \beta_2 + \beta_3 = 3\). The best estimate of the deviation is \(b_1 + b_2 + b_3 - 3 = 2.652 - 3 = -0.348\).

\[
V(b_1 + b_2 + b_3 - 3) = (\sum \sum \gamma_{ij})s^2 = 75.68s^2
\]

The ratio with 1 and 82 degrees of freedom is

\[
\frac{(0.348)^2 \times 1}{75.68 \times 0.0003391} = 4.72
\]

which is significant at the 5% level. This shows that the number of dimensions of the prediction formula is not 3.

It is often desirable to test whether the inclusion of an extra variable increases the accuracy of prediction. For instance, in the above example we can test whether \(H'\) is necessary when \(L\) and \(B\) have already been considered. This is equivalent to testing whether \(\beta_3 = 0\). The estimate \(b_3 = 0.733\) has the variance \(C_{33}s^2\). The ratio with 1 and 82 degrees of freedom is

\[
\frac{b_3^2}{C_{33}s^2} = (0.733)^2 \times \frac{1}{39.88 \times 0.0003391} = 0.01347 = 39.72
\]

where for \(s^2\) the estimate based on 82 degrees of freedom is used. This is significant at the 1% level, showing that \(H'\) is also relevant.
If $b_3$ were not significant, the sum of squares due to $b_3$

$$\frac{b_3^2}{C_{33}} = 0.01347$$

could be added to the residual sum of squares 0.02781 to obtain a sum 0.04128 based on 83 degrees of freedom, giving the estimate of $\sigma^2 = 0.0004973$.

If $b_3$ is declared to be zero, the best estimates of $b_i$ and $b_2$ have to be revised, starting with the equation

$$y = \alpha + b_1x_1 + \beta_2x_2$$

It is, however, not necessary to start afresh. The $C$ matrix

$$\begin{bmatrix}
64.21 & -15.57 & -10.49 \\
-15.57 & 41.71 & -9.00 \\
-10.49 & -9.00 & 39.88
\end{bmatrix}$$

is reduced by the method of pivotal condensation, starting from the last row and using $C_{33}$ as the pivot.

$$\begin{bmatrix}
64.21 - \frac{(10.49)^2}{39.88} & -15.57 - \frac{(9.00)(10.49)}{39.88} \\
-15.57 - \frac{(9.00)(10.49)}{39.88} & 41.71 - \frac{(9.00)^2}{39.88}
\end{bmatrix}$$

which gives the reduced $C$ matrix for the evaluation of $b_1$ and $b_2$.

$$b_1 = 61.45Q_1 - 17.94Q_2 = 1.071$$

$$b_2 = -17.94Q_1 + 39.68Q_2 = 1.206$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = -1.864$$

The residual sum of squares is

$$\Sigma(y - \bar{y})^2 - b_1Q_1 - b_2Q_2 = 0.0004973$$

which agrees with the value obtained by adding the sum of squares due to $b_3$ to the residual, thus providing a check on the calculation of $b_1$ and $b_2$. The variance-covariance matrix of $b_1, b_2$ is $\sigma^2$ times the new $C$ matrix.

If more variables are omitted, the method of pivotal condensation has to be carried further. The reduced matrix at each stage gives the
Prediction of Cranial Capacity

C matrix appropriate to the retained variables. It would avoid some confusion if the C matrix could be written in the order in which the variables are eliminated before attempting the method of pivotal condensation. Thus if \( x_1, x_2, x_3, x_4 \) are the original variables and if \( x_2 \) and \( x_4 \) are to be eliminated, we may write the C matrix as

\[
\begin{pmatrix}
  C_{22} & C_{24} & C_{21} & C_{23} \\
  C_{42} & C_{44} & C_{41} & C_{43} \\
  C_{12} & C_{14} & C_{11} & C_{13} \\
  C_{32} & C_{34} & C_{31} & C_{33}
\end{pmatrix}
\]

which is obtained by bringing the second and fourth rows and columns to the first two positions. Now this matrix can be reduced by the method of forward pivotal condensation.

If the order in which the variables are to be included in the regression equation is assigned, the successive regression equations can be obtained by following the computational method of Table 7b.6B in Chapter 7.

The Use of the Formula for Predicting the Capacity of a Single Skull.

A skull with \( L = 198.5, B = 147, H' = 131 \), i.e., \( x_1 = 2.298, x_2 = 2.167, x_3 = 2.117 \), will have the estimated log capacity

\[
Y = \hat{g} + b_1 (x_1 - \bar{x}_1) + b_2 (x_2 - \bar{x}_2) + b_3 (x_3 - \bar{x}_3)
\]

\[
= 3.2069
\]

\[
C = \text{antilog} 3.2069 = 1610
\]

\[
V(Y) = \sigma^2 \left( \frac{1}{n} + \frac{\sum (x_i - \bar{x}_i)(x_j - \bar{x}_j) \text{cov} b_i b_j}{n} \right)
\]

\[
= \sigma^2 + \sigma^2 \sum (x_i - \bar{x}_i)(x_j - \bar{x}_j) C_{ij}
\]

\[
= \sigma^2(0.04187) = 0.0003391(0.04187) = 0.0001420
\]

The estimated value of \( \sigma^2 \) is obtained from the residual line in Table 5f.2a.

\[
V(C) = C^2 V(y) \left( \log e 10 \right)^2 \text{ approximately }^\ast
\]

\[
= 195.2 (2.3026)^2 = 1034.94
\]

\(^\ast\)A general method of obtaining the variances of transformed variables is given in Sec.I. The factor involving logarithm to base e because the original logarithms are to the base 10.
It is seen that in the above formula for variance of the estimated value \( \hat{Y} \) the precision of the estimate depends on the closeness of \( \bar{x}_1, \bar{x}_2, \bar{x}_3 \) to the averages \( \bar{x}_1, \bar{x}_2, \bar{x}_3 \) realized in the sample used to construct the prediction formula. In fact, the variance is least, equal to \( \sigma^2/n \), for the estimate when the measurements on the specimen coincide with the average values. The accuracy of prediction diminishes as \( \bar{x}_1, \bar{x}_2, \bar{x}_3 \) depart more and more from the average values, and the prediction may become completely unreliable if \( \bar{x}_1, \bar{x}_2, \bar{x}_3 \) fall outside the range of values observed in the sample.

For prediction with a single variable the regression equation is

\[
Y = a + b(x - \bar{x})
\]

and

\[
V(Y) = \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right) \sigma^2
\]

where \( S_{xx} \) is the corrected sum of squares for \( x \) in the observed sample. As before, accuracy is higher for prediction near the mean. The formula also depends on \( S_{xx} \), the scatter of \( x \) in the sample; the larger the scatter, the higher is the precision of the estimate for any \( x \). Therefore, in choosing the sample for the construction of the prediction formula, the values of \( x \) at the extremities of its range should be observed if a best prediction formula is to be constructed. This is, no doubt, a theoretically sound policy which can be carried out with advantage when it is known for certain that the regression equation is of the linear form. In fact, data collected in such a manner are not suitable for judging whether the regression is linear or not, and there is no reason to believe that linearity of regression is universal. The biometric experience is that the regressions are very nearly linear, deviations from linearity being detectable only in large samples. If this is so, the regression line fitted to the data is only an approximation to the true regression function, and the data should allow a closest possible fit of the straight line to the ideal curve. The best plan for this is to choose \( x \) from all over its range or, preferably, to choose \( x \) at random so that different values of \( x \) may occur with their own probability and exert their influence in the determination of the straight-line fit.

**The Use of the Formula in Estimating the Mean Capacity.** The formula can also be used to estimate the mean cranial capacity of a series of skulls. For this purpose two methods are available. We may estimate the cranial capacity of individual skulls and calculate the mean of these estimates, or we may apply the formula directly to the mean values of \( L, B, H' \) for the series. It is of interest to know whether these two methods give the same results. For this purpose estimates
were made of the mean cranial capacity of an additional 29 male skulls of the Farringdon Street series for which measurements of \( L, B, H' \) but not of \( C \) were available.

For these 29 skulls the mean of \( L \) is 191.1, of \( B \) is 143.1, and of \( H' \) is 129.0.

Applying the formula \( C = 0.00241 L^{0.878} B^{0.041} H'^{0.733} \) to these mean values, we estimate the mean of \( C \) to be 1498.3. If we estimate \( C \) for the 29 skulls individually and take the mean of the 29 estimates, we get an estimate of the mean value of \( C \) equal to 1498.2.

The same estimates were calculated for the 22 male skulls of the Moorfields series (Hooke, 1926) for which all four measurements were available. For these 22 skulls the mean of \( L \) is 189.5, of \( B \) is 142.5, and of \( H' \) is 128.8, giving an estimate of the mean of \( C \) equal to 1479.0. If we estimate \( C \) for the 22 skulls individually and calculate the mean, we get an estimate of the mean of \( C \) equal to 1480.0. Thus it appears that the two methods give very nearly the same estimates.

Are Only Small Skulls Preserved? A point of some interest is that, whereas the mean value of \( C \) for the Farringdon Street series as calculated from 86 measured values is 1481.3, the mean value of \( C \) as estimated by our formula from the 29 skulls for which measurements of \( L, B, H' \) but not of \( C \) are available is 1498.3.

Again, for the Moorfields series the mean of \( L \) is 189.2 based on 44 measurements, of \( B \) is 143.0 based on 46 measurements, and of \( H' \) is 129.8 based on 34 measurements. Applying our formula to these mean values (as we may do with some confidence as shown above), we obtain an estimate of the mean of \( C \) equal to 1490.7. The mean of \( C \) as calculated from 22 measured values is only 1473.8.

The above results suggest that those skulls which are damaged to such an extent that the cranial capacity cannot be measured are on the whole larger than those that remain intact.

This raises a serious issue. Are not the published mean values of cranial capacities gross underestimates? Can a suitable method be suggested to correct these values? One way would be to use the samples providing observations on \( C, L, B, \) and \( H' \) for merely constructing the prediction formula. As observed earlier, the prediction formula, provided the nature of the regression function used is appropriate, could be obtained from samples providing observations on all the measurements although the samples are not drawn at random from the population. For instance, if only small skulls are preserved, the measurements obtained are not strictly random from the population of skulls. Such material is being used just to establish a relationship. Having obtained this formula, the mean values of \( L, B, H' \) obtained from all
LINEAR ESTIMATION AND TESTS OF HYPOTHESES

the available measurements may be substituted to obtain an estimate of the mean capacity. This value will be higher than the average of the available measurements of the cranial capacity but less than the predicted value on the basis of mean $L$, $B$, $H'$ from skulls providing these measurements only.

The extent of underestimation depends on the proportion of the disintegration of the large skulls. This may vary from series to series, and hence for a proper comparison of the mean capacities the correction indicated above may have to be applied.

3f.3 Test for the Equality of Regression Equations

It is very often necessary to test whether regression functions constructed from two series are the same. Thus if the formulae for the prediction of the cranial capacity from a different series is

$$y = a' + b_1'x_1 + b_2'x_2 + b_3'x_3$$

two types of hypotheses may be tested whether, in the expectation,

(i) \[ a = a' \] \[ b_1 = b_1' \] \[ b_2 = b_2' \] \[ b_3 = b_3' \]

(ii) \[ b_1 = b_1' \] \[ b_2 = b_2' \] \[ b_3 = b_3' \] irrespective of whether \( a \) equals \( a' \) or not

If the former is true, then the whole regression function is the same in both series; if the latter is true, the regression functions are the same apart from a change in the constant. These two hypotheses are relevant because many problems arise where a prediction formula constructed from one series may have to be used for a specimen from an entirely different series. An extreme and rather ambitious case of such a use is the prediction of stature of prehistoric men from the length of fossil femur by using a formula connecting the stature of modern man with the length of his long bones (K. Pearson, 1898). Some sort of justification for such a procedure will be available if the first hypothesis is proved to be correct in analogous situations. We first deal with the test procedures when the prediction formulae are available for both the series.

Let the derived quantities for the second series be:

| Sample size: \( n' \) |
| Mean values: \( \bar{x}_1', \bar{x}_2', \bar{x}_3' \) and \( \bar{y}' \) |
| Corrected sums of products: |
| \( \Sigma(x_{i1}' - \bar{x}_1')(y_{i1}' - \bar{y}') = S_{i1}' \) |
| \( \Sigma(x_{i2}' - \bar{x}_1')(y_{i2}' - \bar{y}') = Q_{i2}' \) |
These are sufficient to determine the regression function.

\[ y = a' + b'_1x_1 + b'_2x_2 + b'_3x_3 \]

The residual sum of squares

\[ R_0^2 = \sum(y_1 - \bar{y})^2 - b'_1Q_1' - b'_2Q_2' - b'_3Q_3' \quad \text{(for the second sample)} \]

\[ + \sum(y_2 - \bar{y})^2 - b_1Q_1 - b_2Q_2 - b_3Q_3 \quad \text{(for the first sample)} \]

has \((n' - 4) + (n - 4) = (n + n' - 8)\) degrees of freedom.

We now throw the two samples together and consider them as a single sample of size \((n + n')\) and determine the regression line and residual sum of squares by using the above formula. The necessary quantities can be computed from those already available:

Sample size: \(n + n'\)

Mean values:

\[
\frac{(n \bar{x}_1 + n' \bar{x}_1')}{(n + n')} = \bar{x}'
\]

Corrected sums of products:

\[
S_{ij}' = S_{ij} + \frac{nn'}{n + n'} (\bar{x}_1 - \bar{x}) (\bar{x}_1' - \bar{x}')</n>

\[
Q_i' = Q_i + \frac{nn'}{n + n'} (\bar{x}_1 - \bar{x}) (\bar{y} - \bar{y})'
\]

If \(b_1', b_2', \ldots\) are the regression coefficients, then the residual sum of squares \(R_1^2\) with \((n + n' - 4)\) degrees of freedom is

\[
\sum(y_1 - \bar{y})^2 + \sum(y_2 - \bar{y})^2 + \frac{nn'}{n + n'} (\bar{y} - \bar{y})^2
\]

We can set up the analysis of variance table.

**Table 3f.3a. Analysis of Variance for Testing Equality of Regression Coefficients**

<table>
<thead>
<tr>
<th>Residual Due to</th>
<th>D.F.</th>
<th>S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation from hypothesis</td>
<td>4</td>
<td>*</td>
</tr>
<tr>
<td>Separate regression</td>
<td>(n + n' - 8)</td>
<td>(R_0^2)</td>
</tr>
<tr>
<td>Common regression</td>
<td>(n + n' - 4)</td>
<td>(R_1^2)</td>
</tr>
</tbody>
</table>

* Obtained by subtraction.
The significance of the ratio of mean square due to deviation from hypothesis to residual due to separate regressions is tested. If the object is to test for the equality of the $b$ coefficients only, we calculate the quantities

$$Q_i''' = Q_i + Q_i'$$

and obtain the constants $b_1'''$, $b_2'''$, $b_3'''$ from the equations

$$Q_i''' = b_1'''S_{1i}''' + b_2'''S_{2i}''' + b_3'''S_{3i}''' \quad i = 1, 2, 3$$

and find the residual sum of squares $R_2^2$ with $(n + n' - 5)$ degrees of freedom.

$$R_2^2 = \sum(y_i - \bar{y})^2 + \sum(y_i' - \bar{y}')^2 - b_1'''Q_1''' - b_2'''Q_2''' - b_3'''Q_3'''$$

The test depends on the variance ratio

$$\frac{R_2^2 - R_0^2}{3} \div \frac{R_0^2}{n + n' - 8}$$

with 3 and $(n + n' - 8)$ degrees of freedom.

In biological data it is often found that the mutual correlations and variabilities of measurements are approximately the same for all allied series, in which case the coefficients $b_1$, $b_2$, $b_3$ in the regression formula will not differ much. On the other hand the mean values differ to some extent from series to series, in which case the equality of the constant term means that the expected value of

$$y - \beta_1x_1 - \beta_2x_2 - \beta_3x_3$$

is the same for both the series. This leads us to consider a different problem whether $\alpha = \alpha'$ when $\beta_1 = \beta_1'$, $\beta_2 = \beta_2'$, $\cdots$. A test for this can be immediately obtained from the sums of squares calculated above. The suitable statistic is the variance ratio

$$\frac{R_1^2 - R_2^2}{1} \div \frac{R_2^2}{n + n' - 5}$$

with 1 and $(n + n' - 5)$ degrees of freedom. If the above hypothesis is true, then the difference in the mean value of $y$ could be completely explained by differences in the other variables $x_1$, $x_2$, $x_3$. This problem is considered more fully in Chapter 7 (7b.6). It appears that when a sufficient number of measurements is considered the extra difference contributed by any other measurement independent of the set already considered is negligibly small. In such situations the equality of the dispersion matrix in both series is sufficient to ensure the equality of the regression functions as a whole. A good deal of caution is necessary.
when the prediction formula based on one or two variables is so used. Such a statistical adventure undertaken by Karl Pearson in predicting the stature of prehistoric men is, however, justifiable if we agree with the last statement of his article (K. Pearson, 1898).

“No scientific investigation is final; it merely represents the most probable conclusion which can be drawn from the data at the disposal of the writer. A wider range of facts, or more refined analysis, experiment, and observation will lead to new formulae and new theories. This is the essence of scientific progress.”

3f.4 The Test for an Assigned Regression Function

In 3f.2 it was assumed that the regression function for log capacity is linear in the logarithms of length, breadth, and height. If, at least, for some given sets of values of the independent variables multiple observations on the dependent variable have been observed, the validity of such an assumption can be tested.

In Table 3f.4a are given the mean values of nasal index of people living in various parts of India together with the mean annual temperature and relative humidity of the places.

<table>
<thead>
<tr>
<th>Region</th>
<th>Sample Size (n)</th>
<th>Nasal Index</th>
<th>Temperature</th>
<th>Relative Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample</td>
<td>Mean (j)</td>
<td>Total (nj)</td>
<td>Mean Annual (f)</td>
</tr>
<tr>
<td></td>
<td>Size (n)</td>
<td>Mean (j)</td>
<td>Total (nj)</td>
<td>Mean Annual (f)</td>
</tr>
<tr>
<td>Assam</td>
<td>36</td>
<td>83.0</td>
<td>2,988.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Orissa</td>
<td>40</td>
<td>80.4</td>
<td>3,216.0</td>
<td>80.3</td>
</tr>
<tr>
<td>Bihar</td>
<td>30</td>
<td>80.1</td>
<td>2,403.0</td>
<td>74.8</td>
</tr>
<tr>
<td>Malabar</td>
<td>45</td>
<td>77.0</td>
<td>3,465.0</td>
<td>80.2</td>
</tr>
<tr>
<td>Bombay</td>
<td>36</td>
<td>76.2</td>
<td>1,981.2</td>
<td>77.6</td>
</tr>
<tr>
<td>Madras</td>
<td>35</td>
<td>75.9</td>
<td>2,056.5</td>
<td>81.8</td>
</tr>
<tr>
<td>Punjab</td>
<td>28</td>
<td>71.4</td>
<td>1,992.2</td>
<td>76.4</td>
</tr>
<tr>
<td>United Province</td>
<td>32</td>
<td>80.8</td>
<td>2,585.6</td>
<td>77.2</td>
</tr>
<tr>
<td>Andhra</td>
<td>41</td>
<td>76.8</td>
<td>3,148.8</td>
<td>80.3</td>
</tr>
<tr>
<td>Ceylon</td>
<td>31</td>
<td>80.3</td>
<td>2,400.0</td>
<td>80.2</td>
</tr>
<tr>
<td>Total Mean</td>
<td>344</td>
<td>78.294</td>
<td>26,933.3</td>
<td>78.335</td>
</tr>
</tbody>
</table>
The corrected total sum of squares for nasal index has been found to be 11,140.209 with (344 - 1) degrees of freedom.

If no assumption is made about the regression of nasal index on temperature and relative humidity, the analysis of variance between and within groups is obtained as in Table 3f.4.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>9</td>
<td>$\sum(y_0 - \bar{y})^2</td>
</tr>
<tr>
<td>Within groups</td>
<td>334</td>
<td>$\sum(y - y_0)^2</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>$\sum y = 11,140.209</td>
</tr>
</tbody>
</table>

* Obtained by subtraction.

The mean square for between groups is very large, indicating real differences in nasal index. Can these differences be explained by a regression of nasal index \(y\) on temperature \(t\) and relative humidity \(h\) of the form

\[
y = \alpha + \beta_1 t + \beta_2 h
\]

The normal equations leading to the estimates \(b_1, b_2\) of \(\beta_1, \beta_2\) are

\[
Q_1 = b_1 S_{11} + b_2 S_{12} = \sum (n_i (y_i - \bar{y}) (t_i - \bar{t})) = -1,072.40
\]

\[
Q_2 = b_1 S_{12} + b_2 S_{22} = \sum (n_i (y_i - \bar{y}) (h_i - \bar{h})) = 4,586.57
\]

\[
S_{12} = \sum (n_i (t_i - \bar{t}) (h_i - \bar{h})) = 2,334.91
\]

\[
S_{11} = \sum (n_i (t_i - \bar{t}) (t_i - \bar{t})) = 2,721.06
\]

\[
S_{22} = \sum (n_i (h_i - \bar{h}) (h_i - \bar{h})) = 22,042.32
\]

Solving the above equations, \(b_1, b_2\) are obtained as

\[
b_1 = -0.237113 \quad b_2 = 0.182963
\]
With these values the regression analysis can be set up as in Table 3f.4y.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to regression</td>
<td>2</td>
<td>$b_1Q_1 + b_2Q_2 = 1,096.45$</td>
</tr>
<tr>
<td>Residual about regression</td>
<td>341</td>
<td>*</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>11,140.209</td>
</tr>
</tbody>
</table>

* Obtained by subtraction.

If the hypothesis concerning the regression is true, then the mean squares obtained from "Within groups" of Table 3f.4y and "Residual about regression" of Table 3f.4y will be of the same magnitude. A significant difference would disprove the hypothesis.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation from specified regression</td>
<td>7</td>
<td>2,076.450 *</td>
<td>296.636</td>
</tr>
<tr>
<td>Within groups</td>
<td>334</td>
<td>7,970.309</td>
<td>23.863</td>
</tr>
<tr>
<td>Residual about regression</td>
<td>341</td>
<td>10,046.759</td>
<td></td>
</tr>
</tbody>
</table>

* Obtained by subtraction.

The ratio 12.4 with 7 and 334 degrees of freedom is significant at the 1% level, so that the regression of nasal index on temperature and relative humidity cannot be considered linear. It is also seen from Table 3f.4y that the variance ratio 2.34 with 2 and 341 degrees of freedom is significant, but this does not mean that nasal index depends entirely on weather conditions of the place in which the individuals live. The observed differences may be more complex than can be explained by weather differences, or the nature of dependence on weather may itself be very complicated. Unless such a relationship is discovered and found to fit well, some caution is necessary before concluding that the shape of the nose is determined by temperature, humidity, etc.

In some cases, as in the distribution of heights of father and daughter, it may be desired to test whether the regression of one variable on the other is linear. For the purposes of such a test the range of the independent variable has to be divided into a suitable number of class intervals and the variance of the dependent variable analyzed into
between and within classes. To get an estimate of within variation it is necessary that at least some of the classes contain more than one observation. The regression analysis can be done without the use of the class intervals, or if the data are already grouped the midpoint of the class interval is taken as the value of the independent variable for each observation of the dependent variable in that class. The final test can be carried out as in Table 3f.44.

3g The General Problem of Least Squares with Two Sets of Parameters

3g.1 Concomitant Variables

Suppose that the growth rates of groups of animals receiving different diets are to be compared. The observed differences in growth rates can be attributed to diet only if all the animals treated are similar in some observable aspects such as age, initial weight, parentage, etc., which influence the growth rate. In fact, if the groups of animals receiving different diet differ in these aspects, it is desirable to compare the growth rates after eliminating these differences.

However, it may be noted that no bias is introduced in the experiment if the animals which might differ in these aspects are assigned at random to the groups to be treated differently. This procedure enhances the residual variation calculable from the differences in the growth rates of animals receiving the same treatment and thus decreases the efficiency of the experiment.

If the magnitudes of these additional variables are known, it is possible to eliminate the differences caused by them independently of the treatments both from within and between groups and to test for the pure effects of the treatments with greater efficiency. The computational technique relating to this process is known as the adjustment for concomitant variation.

For significant reduction in the residual variation it must be known that the effects under study are influenced by the concomitant variables. This is important in experimental studies where due consideration is to be given to the cost and time involved in recording the concomitant variables. This can be tested as shown in the example considered in 3g.3.

On the other hand, the concomitant variables chosen must not have been influenced by the treatments under consideration. Sometimes, in assessing the differences in yields of plants treated differently, concomitant variables such as the number of branches or the quantity of straw are chosen. These will be valid only when variations in them produce corresponding variation in the yield of plants treated alike.
3g.2 Adjustment for Concomitant Variation

Assuming the regression of \( y \) on the concomitant variables \( x_1, \ldots, x_k \) to be linear, the observational equations containing the parameters \( r_1, \ldots, r_m \) under consideration and the regression coefficients \( \beta_1, \ldots, \beta_k \) can be written

\[
E(y_i) = \alpha_0 r_1 + \cdots + \alpha_m r_m + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \quad (3g.2.1)
\]

The normal equations giving the best estimates of parametric functions are

\[
Q_{ij} = (\alpha_1 \cdot \alpha_j) t_i + \cdots + (\alpha_m \cdot \alpha_j) t_m \quad j = 1, \ldots, m \quad (3g.2.2)
\]

and

\[
P_{is} = (\alpha_1 \cdot x_s) t_1 + \cdots + (\alpha_m \cdot x_s) t_m \quad s = 1, \ldots, k \quad (3g.2.3)
\]

where \( Q_{ij} = (\alpha_j, y) \) and \( P_{is} = (x_s, y) \). To solve them let us construct the equations

\[
Q_{ij}^{(0)} = (\alpha_1 \cdot \alpha_j) t_i^{(0)} + \cdots + (\alpha_m \cdot \alpha_j) t_m^{(0)} \quad j = 1, \ldots, m
\]

and

\[
Q_{is}^{(s)} = (\alpha_1 \cdot x_s) t_1^{(s)} + \cdots + (\alpha_m \cdot x_s) t_m^{(s)}
\]

where \( Q_{ij}^{(s)} \) is the same function as \( Q_{ij}^{(0)} \) with the variable \( y \) replaced by the \( s \)th concomitant variable \( x_s \). Multiplying the equations in (3g.2.2) by \( t_1^{(1)}, t_2^{(1)}, \ldots, t_m^{(1)} \) and subtracting their total from the first equation in (3g.2.3), we obtain

\[
E_{ij} = P_{ij} - t_1^{(1)} Q_{ij}^{(0)} - \cdots - t_m^{(1)} Q_{jm}^{(0)}
\]

\[
= b_1 (P_{i1} - t_1^{(1)} Q_{1j}^{(1)} - \cdots - t_m^{(1)} Q_{mj}^{(1)}) + b_2 (P_{i2} - t_1^{(1)} Q_{1j}^{(2)} - \cdots - t_m^{(1)} Q_{mj}^{(2)}) + \cdots + b_k E_{jk}
\]

Similarly, the equations

\[
E_{ij} = b_1 E_{ij1} + b_2 E_{ij2} + \cdots + b_k E_{ijk} \quad j = 2, \ldots, k \quad (3g.2.4)
\]

*The regression can be of the type \( \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k \) where \( \beta_1, \ldots, \beta_k \) are functions of the concomitant variables, in which case \( \beta_1, \ldots, \beta_k \) will be treated as separate variables. Thus, if the regression is polynomial in one variable \( \beta_1 x + \beta_2 x^2 + \cdots \), the functions \( x, x^2, \ldots \) are considered as separate variables.
are obtained. In the above equations

\[ E_{ij} = P_{ij} - t_i^{(0)}Q_i^{(0)} - \cdots - t_m^{(0)}Q_m^{(0)} = P_{ij} - t_i^{(0)}Q_i^{(0)} - \cdots - t_m^{(0)}Q_m^{(0)} \]

and

\[ P_{ij} = x_i \cdot x_j \]

These are the residual sums of products, the residuals being obtained from the observational equations for any two variables with the same matrix of equations but different sets of parameters. Having obtained the values of \( b_1, \ldots, b_k \) satisfying equations (3g.2.4), the solution for \( t_i \) is given by

\[ t_i = t_i^{(0)} - b_1t_i^{(1)} - \cdots - b_kt_i^{(k)} \]

This completes the estimation of the parametric functions. The residual sum of squares for the observational equations (3g.2.1) is

\[ \sum y^2 - \sum t_i Q_i^{(0)} = 2b_0 P_{00} = (\sum y^2 - \sum t_i Q_i^{(0)}) - 2b_0 E_{00} = E_{00} + \sum b_i E_{i0} \]

which is a function of the residual sum of squares and products.

There are two types of hypotheses to be tested in the above problem. Do the concomitant variables increase the efficiency of comparisons? The hypothesis to be tested is \( \beta_1 = \cdots = \beta_k = 0 \). If this is true the residual sum of squares is \( E_{00} \), which differs from the pure residual \( E_{00} - b_1E_{01} - \cdots \) by \( b_1E_{01} + b_2E_{02} + \cdots \) which has \( k \) degrees of freedom. The mean square of this is compared with the mean square for the pure residual.

Is there any additional advantage in considering \( x_k \) in conjunction with \( x_1, \ldots, x_{k-1} \)? The hypothesis to be tested is \( \beta_k = 0 \). Omitting \( \beta_k \) in the equations (3g.2.4), let the solutions be \( b_1', \ldots, b_{k-1}' \), so that the residual sum of squares is

\[ E_{00} - b_1'E_{01} - \cdots - b_{k-1}'E_{0k-1} \]

The pure residual is \( E_{00} - b_1E_{01} - \cdots - b_kE_{0k} \). Their difference with 1 degree of freedom supplies the valid sum of squares for testing the above hypothesis. Similarly, we can test whether two or more variables are useful in conjunction with others.

If the hypothesis to be tested is specified by some linear restrictions on \( t_1, t_2, \cdots \), then the residual sum of squares has to be obtained subject to these restrictions. If \( E_{ij} \) represents the residual sum of
products under these restrictions, then the residual sum of squares is

$$E_{oo'} = b_1''E_{01'} - \cdots - b_k''E_{0k'}$$

where $b_1'', \cdots, b_k''$ are the solutions of

$$E_{01'} = b_1''E_{11'} + \cdots + b_k''E_{1k'}$$

$$\cdots$$

$$E_{0k'} = b_1''E_{k1'} + \cdots + b_k''E_{kk'}$$

The sum of squares for testing the above hypothesis is

$$E_{oo'} - b_1''E_{01'} - \cdots - b_k''E_{0k'} = \text{pure residual sum of squares}$$

The degrees of freedom will be equal to the degrees of freedom of the hypothesis. The mean square corresponding to this can be tested against the pure residual. This completes the formal theory. The method is further explained in the illustration considered in 3g.3.

### 3g.3 An Illustrative Example

The following data relate to the initial weights and the growth rates of 30 pigs classified according to pen, sex, and type of food given.

<table>
<thead>
<tr>
<th>Pen</th>
<th>Treat-</th>
<th>Sex</th>
<th>Initial Weight (w)</th>
<th>Growth Rate in Pounds per Week (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>G</td>
<td>48</td>
<td>9.94</td>
</tr>
<tr>
<td>I</td>
<td>B</td>
<td>G</td>
<td>48</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
<td>48</td>
<td>9.75</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>H</td>
<td>48</td>
<td>9.11</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>H</td>
<td>39</td>
<td>9.41</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>H</td>
<td>38</td>
<td>9.62</td>
</tr>
<tr>
<td>II</td>
<td>B</td>
<td>G</td>
<td>32</td>
<td>9.24</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
<td>28</td>
<td>8.96</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>G</td>
<td>32</td>
<td>9.48</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>H</td>
<td>37</td>
<td>8.80</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>H</td>
<td>35</td>
<td>8.51</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>H</td>
<td>38</td>
<td>9.05</td>
</tr>
<tr>
<td>III</td>
<td>C</td>
<td>G</td>
<td>33</td>
<td>7.63</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>G</td>
<td>35</td>
<td>9.32</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>G</td>
<td>41</td>
<td>9.34</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>H</td>
<td>46</td>
<td>8.48</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>H</td>
<td>41</td>
<td>9.32</td>
</tr>
</tbody>
</table>
The problem is to study the effect of food after eliminating the initial weight.

The first step in the analysis is to analyze the sum of squares of both the dependent and independent variables and also the sum of products. The analysis of the sum of products is done in the same manner as the analysis of the sum of squares by adopting the rule that the square of a variable in the latter is replaced by the product of the variables involved in the former.

The total sums of squares and products are

\[ \sum g^2 - n \bar{g}^2 = 16.6068 \]
\[ \sum wg - n \bar{w} \bar{g} = 78.979 \quad 29 \text{ D.F.} \]
\[ \sum w^2 - n \bar{w}^2 = 1108.70 \]

If \( w_i \) and \( g_i \) denote the totals of 6 observations for the \( i \)th pen, then the sums of squares and products for pens are

\[ \frac{1}{6} \sum g_i^2 - n \bar{g}^2 = 4.8518 \]
\[ \frac{1}{6} \sum g_i w_i - n \bar{g} \bar{w} = 39.905 \quad 4 \text{ D.F.} \]
\[ \frac{1}{6} \sum w_i^2 - n \bar{w}^2 = 605.87 \]

Similarly, the sums of squares and products for food and sex can be obtained.

If \( w_{ij} \) and \( g_{ij} \) denote the total of 5 observations for the \( i \)th type of food and \( j \)th sex, then the sums of squares and products for the joint

<table>
<thead>
<tr>
<th>Pen</th>
<th>Treatment</th>
<th>Sex</th>
<th>Initial Weight (w)</th>
<th>Growth Rate in Pounds per Week (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>B</td>
<td>G</td>
<td>46</td>
<td>10.56</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>H</td>
<td>46</td>
<td>9.68</td>
</tr>
<tr>
<td>V</td>
<td>B</td>
<td>H</td>
<td>40</td>
<td>9.20</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>H</td>
<td>42</td>
<td>8.76</td>
</tr>
</tbody>
</table>

The table provides data for analysis contined.
AN ILLUSTRATIVE EXAMPLE

effects of food and sex are

\[
\begin{align*}
\frac{1}{2} \sum_{ij} g_{ij}^2 & - g \Sigma g = 3.2422 \\
\frac{1}{2} \sum_{ij} w_{ij} - w \Sigma w & = -0.885 \\
\frac{1}{2} \sum_{ij} w_{ij} & - w \Sigma w = 59.90 
\end{align*}
\]

If from these the corresponding expressions due to food (2 D.F.) and sex (1 D.F.) are subtracted, the expressions for food \( \times \) sex interaction are obtained. Table 3g.33 gives the whole analysis. The error line is obtained by subtraction from the total.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>( g^2 )</th>
<th>( w^2 )</th>
<th>( w^3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pen</td>
<td>4</td>
<td>4.9607</td>
<td>40.324</td>
</tr>
<tr>
<td>Food</td>
<td>2</td>
<td>2.3242</td>
<td>-0.171</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>0.4538</td>
<td>-4.813</td>
</tr>
<tr>
<td>Food ( \times ) Sex</td>
<td>2</td>
<td>0.4642</td>
<td>4.099</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>8.4039</td>
<td>39.540</td>
</tr>
</tbody>
</table>

\[= E_{00} = E_{01} = E_{11} \]

Total 29 16.6068 78.979 1108.70

There is only one regression constant to be estimated.

\[ E_{11} b = E_{01} \]
\[ 442.93b = 39.540 \]
\[ b = 0.089269 \]

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>( bE_{01} = 3.5297 )</td>
<td>3.5297</td>
</tr>
<tr>
<td>Residual</td>
<td>19</td>
<td>( E_{01} - bE_{01} = 4.8742 )</td>
<td>0.2655</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>( E_{01} = 8.4039 )</td>
<td></td>
</tr>
</tbody>
</table>

The ratio is significant at the 1% level so that the comparisons can be made more efficient by eliminating the concomitant variations.

If the hypothesis specifies that there are no differences in food, then the residual sums of squares and products are obtained by adding the rows corresponding to food and error:

\[ E_{00}' = 10.7281 \quad E_{01}' = 39.369 \quad E_{11}' = 448.33 \]
The new regression coefficient is

\[ b''E_{11} = E_{01} \]

\[ b'' = 0.087813 \]

The residual sum of squares when the hypothesis is true is

\[ E_{00}'' - b''E_{01}'' = 7.2710 \text{ with 21 D.F.} \]

<table>
<thead>
<tr>
<th>Food</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>2.3968</td>
<td>1.1984</td>
</tr>
<tr>
<td>Residual</td>
<td>19</td>
<td>E_{00} - bE_{01} = 4.8742</td>
<td>0.2565</td>
</tr>
<tr>
<td>Food + Error</td>
<td>21</td>
<td>E_{00}'' - b''E_{01}'' = 7.2710</td>
<td></td>
</tr>
</tbody>
</table>

* Obtained by subtraction.

The ratio is significant at the 5% level. To test for food without adjustment for concomitant variation, we have to construct the ratio with 2 and 20 degrees of freedom.

\[ F = \left( \frac{2.3242}{2} \right) \left( \frac{20}{8.4039} \right) = 2.76 \]

which is not significant at the 5% level. The quantities used above are taken from the analysis of variance table (Table 3g.3f). The differences caused by food could be detected when the concomitant variation is eliminated. Similarly, any other effect such as sex or interaction can be tested.

3g.4 A Problem of Inheritance in Man

The methods of analysis of variance and regression are of great value in studying the problems of inheritance. Some aspects of Boas’ data * analyzed by Fisher and Gray (1937), are given in this section for illustrating the methods.

The data consist of measurements on Sicilian children and some of their parents. The first step in problems of this nature is to obtain the measurements on children corrected for age. The measurement \( m \) on any character can be represented by

\[ m = f + bA + E \]

* The data are published in *Materials for the study in inheritance* (New York, Columbia University Press, 1938).
where \( f_i \) is a constant for the \( i \)th fraternity, \( A \) the age of the child, \( b \) the regression coefficient, and \( E \) the deviation from expectation. From the above formulation the covariance between \( A \) and \( m \) can be analyzed into between and within fraternities. If the latter is denoted by \( E_{W} \) and the sum of squares for \( A \) within fraternities by \( E_{W} \), then the best estimate of \( b \) is \( E_{W}/E_{W} \). The measurement corrected for age will be 

\[
y = m - AE_{W}/E_{W}.
\]

In the above data the measurements of all the children were used in calculating the regression coefficient. For studying the problems of inheritance only 752 children whose parents had been also measured were considered. The sum of the corrected statures for these children is 1505, and their sum of squares 3,304,643 mm². The total corrected sum of squares for 751 degrees of freedom is

\[
3,304,643 - (1505)^2 / 752 = 3,304,643 - 3012 = 3,301,631 \text{ mm}^2
\]

The children belonged to 337 different fraternities and 235 different combinations of parental heights. The sum of squares between fraternities with 336 degrees of freedom can be split up into between fraternities with the same parental statures (102 D.F.) and between combinations of parental heights (234 D.F.). The sum of squares between fraternities is obtained by the usual formula for between groups. To derive the sum of squares between 235 different combinations of parental heights, all the children belonging to any combination are considered as forming a group and the between group sum of squares is calculated. The sum of squares between fraternities of the same parental heights is obtained by subtracting the latter sum of squares from the former.

The analysis of variance is given in Table 3g.4a.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within fraternities</td>
<td>415</td>
<td>1,272,150</td>
</tr>
<tr>
<td>Between fraternities with the same parental height</td>
<td>102</td>
<td>406,257</td>
</tr>
<tr>
<td>Between combinations of parental height</td>
<td>234</td>
<td>1,623,224</td>
</tr>
<tr>
<td>Total</td>
<td>751</td>
<td>3,301,631</td>
</tr>
</tbody>
</table>

The mean square for between fraternities with the same parental height is greater, though not significantly, than that for within fra-
ternities, indicating that parents of the same height were not identical genetically. Whether the former mean square is significantly different from the latter or not, it supplies the valid estimate of error for testing any hypothesis concerning the regression of child's stature on those of parents.

In order to test not only for linear regression on the two parents independently but also for theoretically possible deviations due to bias in dominance, the formula chosen was of the form

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

where $x_1$ and $x_2$ stand for the heights of the father and mother and $x_3$ for the product of $x_1 x_2$.

The method of solving for $a$, $b_1$, $b_2$, $b_3$ is the same as that considered in detail in 3f.2. The matrix $(C_{ij})$ giving the variances and covariances of the estimates $b_1$, $b_2$, $b_3$, written in millionths, is

\[
\begin{pmatrix}
0.384227 & -0.041917 & -0.0074896 \\
-0.041917 & 0.438083 & -0.0052708 \\
-0.0074896 & -0.0052708 & 0.0117165
\end{pmatrix}
\]

The values of the coefficients are

$$b_1 = 0.02618850 \quad b_2 = 0.3420271 \quad b_3 = 0.0008596738$$

The sum of squares due to regression (3 D.F.) is 526,452. To test the adequacy of the regression formula chosen, the mean square for deviation from regression is to be compared with the mean square for between fraternities of like parents.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>526,452</td>
<td></td>
</tr>
<tr>
<td>Deviation from regression</td>
<td>234</td>
<td>1,623,224</td>
<td>4748</td>
</tr>
<tr>
<td>Between combinations of parental heights</td>
<td>102</td>
<td>406,257</td>
<td>3983</td>
</tr>
<tr>
<td>Between fraternities of like parents</td>
<td>102</td>
<td>406,257</td>
<td>3983</td>
</tr>
</tbody>
</table>

The ratio for deviation from regression is not significant so that there is no evidence against the inadequacy of the regression formula. The mean square for deviation from regression supplies the valid estimate of error for testing any hypothesis concerning the regression coefficients.
The sum of squares for $b_3$ alone is
\[ \frac{b_3^2}{c_{bb}} = 6308 \]

The sum of squares for regression (2 D.F.) when $b_3$ is not considered is obtained by subtracting this quantity from the total sum of squares due to regression (3 D.F.). The test for the regression coefficients is given in Table 3g.4.

<table>
<thead>
<tr>
<th>D.F.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to $b_3$</td>
<td>1</td>
<td>6,308</td>
<td>6,308</td>
</tr>
<tr>
<td>Linear regression</td>
<td>2</td>
<td>520,144</td>
<td>260,072</td>
</tr>
<tr>
<td>Regression</td>
<td>3</td>
<td>526,452</td>
<td></td>
</tr>
<tr>
<td>Deviation from regression</td>
<td>231</td>
<td>1,096,772</td>
<td>4,748</td>
</tr>
</tbody>
</table>

The ratio for linear regression is significant. Since $b_3$ is positive, there is an indication of negative bias in dominance, i.e., "a situation in which the heterozygotes more nearly, or more frequently, resemble the smaller rather than the larger of the corresponding homozygotes." The ratio for $b_3$, though greater than unity, is not large enough to establish the existence of negative bias in dominance.

The actual regression formula is derived as
\[
y = 1.084225 + 0.2690299(x_1 - \bar{x}_1) + 0.3483146(x_2 - \bar{x}_2) + 0.0008596738(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \ast
\]

The sex difference in selection is
\[
b_2 - b_1 - (\bar{x}_2 - \bar{x}_1)b_3 = 0.0792847
\]
with standard error 0.06559 so that the difference in regression in favor of the mother, though large, is not significant.

References


\* The coefficients $b_3$ or $b_4$ are changed because of using the deviations from the mean in the product term.


CHAPTER 4

The General Theory of Estimation
and the Method of Maximum Likelihood

4a Best Unbiased Estimates

In Chapter 3 problems were considered in which the class of estimates of parametric functions was restricted to linear functions of observations only. Nothing, however, was assumed about the actual distribution functions of these variables, except that their expectations are linear functions of unknown parameters and have a common unknown variance. It is of interest to examine the methods of obtaining the estimates with the minimum possible variance by considering the totality of unbiased estimates. This method of estimation is not necessarily the best. The general problem is that of deriving a function \( t = f(Y_1, \ldots, Y_n) \) of the observations \( Y_1, \ldots, Y_n \) such that with respect to any other function \( t' \) the probabilities satisfy the relationship

\[
P(\theta - \lambda_1 < t < \theta + \lambda_2) < P(\theta - \lambda_1 < t' < \theta + \lambda_2) \quad (4a.1.1)
\]

for all possible \( \lambda_1 \) and \( \lambda_2 \) in an interval \((0, \lambda)\). The choice of the interval may be fixed by other considerations, depending on the frequency and the magnitude of departure of the estimate from the true value allowable in a problem. If the condition \((4a.1.1)\) is satisfied for all \( \lambda \), then a necessary condition is

\[
E(t - \theta)^2 > E(t' - \theta)^2
\]

where \( E \) stands for expectation. If, further, it is assumed that the estimate should be unbiased, then it follows that

\[
V(t) > V(t')
\]

where \( V \) stands for variance.
As no simple solution satisfying the postulate (4a.1.1) exists, the inevitable arbitrariness of the postulates of unbiasedness and minimum variance needs no emphasis. The only justification for choosing the estimate with the minimum possible variance from the class of unbiased estimates is that a necessary condition for (4a.1.1) to hold for all \( \lambda \), with the further requirement \( E(t) = \theta \), is ensured. The condition of unbiasedness is particularly defective in that many biased estimates with smaller variances lose their claim as estimating functions. There are, however, numerous examples where a slightly biased estimate is preferred to an unbiased estimate with a greater variance. Until a unified solution of the problem of estimation is set forth, an estimating function has to be subjected to a critical examination as to its bias, variance, and frequency for a given amount of departure from the true value before utilizing it. The theory of confidence intervals as developed by J. Neyman is a great advance toward such a unified theory.

There is one important aspect which favors unbiased estimates. In biometric investigations it is often necessary to combine the evidence supplied by various sources. Often the evidence is in the nature of an estimate, probably with a standard error attached to it. For instance, two geneticists may be determining the proportion of albino mice produced under certain types of mating. One gives a proportion \( P_1 \pm \varepsilon_1 \), and the other \( P_2 \pm \varepsilon_2 \). If the estimates are unbiased, then a combined unbiased estimate may be reached, and with the accumulation of more and more evidence the true value can be approached through a series of unbiased estimates. On the other hand, if biased estimates are published without any indication as to the nature of the bias involved, nothing definite can be said about the combined estimate. The bias may exceed the standard error at some stage, and ultimately the combined estimate may not be near the true value.

4a.2 The Information Limit to Variance: A Single Parameter

Let \( \phi(x_1, \ldots, x_n, \theta) \) be the probability density of the observations \( x_1, \ldots, x_n \), and \( t(x_1, \ldots, x_n) \) be an unbiased estimate of \( \psi(\theta) \), a function of \( \theta \), the parameter occurring in the probability density. Then

\[
\int \phi \, dv = \psi(\theta)
\]  

(4a.2.1)

where \( dv \) stands for the product of differential elements \( dx_1, \ldots, dx_n \), and a single integral sign stands for the multiple integral. It may be noted that when the variables are discrete the integral sign can be replaced by the summation symbol. If the limits of integration do not
THE INFORMATION LIMIT TO VARIANCE

involve \( \theta \), then differentiation under the integral sign yields

\[
\int \left( \frac{d\phi}{d\theta} \right)^2 d\theta = \frac{d^2}{d\theta^2}
\]

if the above integral exists, which shows that the covariance between \( t \) and \( (1/\phi)(d\phi/d\theta) \) is \( (d\phi/d\theta) \). Since the square of the covariance is not greater than the product of the variances of the two variables, the following relationships (using \( V \) and \( C \) for variance and covariance) are true.

\[
V(t) \left( \frac{1}{\phi} \frac{d\phi}{d\theta} \right)^2 \leq \left( C \left( \frac{1}{\phi} \frac{d\phi}{d\theta} \right)^2 \right)
\]

\[
V(t) \leq \frac{(d\phi/d\theta)^2}{I}
\]

where

\[
I = V \left( \frac{1}{\phi} \frac{d\phi}{d\theta} \right)^2 = E \left[ - \frac{d^2 \log \phi}{d\theta^2} \right]
\]

The quantity \( I \) is the information on \( \theta \) supplied by a sample of \( n \) observations as defined by R. A. Fisher. This gives the following theorem.

**Theorem.** The variance of any unbiased estimate of \( \phi(\theta) \), a function of the unknown parameter \( \theta \), is not less than \( [\phi'(\theta)]^2 / I \), which is defined independently of any method of estimation. The conditions to be satisfied are that the range of the stochastic variable is independent of \( \theta \) and the probability density admits differentiation under the integral sign.

If \( \phi(\theta) = \theta \), then this limit is \( 1/I \). We shall call \( [\phi'(\theta)]^2 / I \) the information limit to variance for the estimation of \( \phi(\theta) \). This is not, however, the minimum attainable in any particular distribution. An unbiased estimate of \( \phi(\theta) \) with the minimum attainable variance will be called a best unbiased estimate. It is incorrect to conclude that an estimate is inefficient if its variance is not equal to \( [\phi'(\theta)]^2 / I \) unless it is ascertained that this minimum is attainable. In fact, it is shown later that there exists a more exact expression for the minimum possible variance.

As a corollary to the above theorem, it follows that the minimum value of

\[
E(T - \theta)^2
\]

for the set of statistics \( T \) with expectation \( \phi(\theta) \) is

\[
[\theta - \phi(\theta)]^2 + \frac{[\phi'(\theta)]^2}{I(\theta)}
\]
Thus if \( b(\theta) \) is the bias in the estimate \( T \) of \( \theta \), then

\[
\phi(\theta) = \theta + b(\theta)
\]

in which case

\[
E(T - \theta)^2 \geq |b(\theta)|^2 + \frac{[1 + b'(\theta)]^2}{I(\theta)}
\]

4.3.3 Distributions Admitting Estimates with the Information
Limit to Variance

The relationship (4a.2.2) may be written as

\[
\int [t - \phi(\theta)]^2 \phi \; d\nu \times \int \left( \frac{1}{\phi} \frac{d\phi}{d\theta} \right)^2 \phi \; d\nu < \left( \int [t - \phi(\theta)] \frac{1}{\phi} \frac{d\phi}{d\theta} \; d\nu \right)^2
\]

It is known from Schwarz's inequality that the equality is attained only when

\[
t - \phi(\theta) = \lambda \frac{1}{\phi} \frac{d\phi}{d\theta}
\]

where \( \lambda \) is a constant depending only on \( \theta \). The solution of this differential equation is

\[
\log \phi = \int \left[ \frac{t}{\lambda} - \frac{\phi(\theta)}{\lambda} \right] d\theta = A + \theta_1 + \theta_2
\]

where \( \theta_1 \) and \( \theta_2 \) are functions of \( \theta \), and \( A \) is independent of \( \theta \). Hence

\[
\phi = \phi_1 \exp (\theta_1 + \theta_2) \tag{4a.3.1}
\]

where \( \phi_1 \) is only a function of the observations. Also if

\[
\phi = \phi_1 \exp (\theta_1 + \theta_2)
\]

then

\[
\int \phi \; d\nu = \int \phi_1 \exp (\theta_1 + \theta_2) \; d\nu = 1
\]

and

\[
\int \phi_1 \exp (\theta_1 + \theta_2) \; d\nu = \exp (-\theta_2)
\]

On differentiating with respect to \( \theta_1 \) twice,

\[
\int \phi_1 \exp (\theta_1) \; d\nu = -\frac{d^2\nu}{d\theta_1^2} \exp (-\theta_2)
\]

\[
\int \phi_1 \exp (\theta_1) \; d\nu = \left[ -\frac{d^2\nu}{d\theta_1^2} + \left( \frac{d\nu}{d\theta_1} \right)^2 \right] \exp (-\theta_2)
\]
Hence

\[ E(t) = -\frac{d\theta_2}{d\theta_1} \quad \text{and} \quad V(t) = -\frac{d^2\theta_2}{d\theta_1^2} \quad (4a.3.2) \]

The information limit for the estimation of \(-\frac{d\theta_2}{d\theta_1}\) is

\[
\left( \frac{d}{d\theta} \frac{d\theta_2}{d\theta_1} \right)^2 + \left( -\frac{d^2\theta_2}{d\theta_1^2} \right)^2 = -\frac{d^2\theta_2}{d\theta_1^2}
\]

which is the same as that derived in (4a.3.2) so that the information limit is attainable.

Hence the necessary and sufficient condition that a distribution admits the estimation of a suitably chosen function of the parameter with variance equal to the information limit is that

\[ \phi = \phi_1 \exp(\theta_1 + \theta_2) \]

where \(\phi_1\) and \(t\) are functions of the observations only and \(\theta_1\) and \(\theta_2\) are functions of \(\theta\) only. The parametric function to be estimated is

\[ -\frac{d\theta_2}{d\theta_1} = -\frac{d\theta_2}{d\theta} \frac{d\theta}{d\theta_1} \]

and the variance of the estimate is

\[ -\frac{d^2\theta_2}{d\theta_1^2} = \left( \frac{d\theta_2}{d\theta} \right)^{-1} \frac{d}{d\theta} \left( -\frac{d\theta_2}{d\theta_1} \right) \]

For any estimate \(t\) which has the minimum variance (4a.2.2),

\[ t - \psi(\theta) = \lambda \frac{d\phi}{d\theta} \]

so that \(\lambda^2 = [\psi'(\theta)]^2/I\), which is a unique function of \(\theta\). This shows that \(t\) is unique as the best unbiased estimate of \(\psi(\theta)\).

Example 1. Let \(x_1, \ldots, x_n\) be \(n\) independent observations from the normal population

\[ \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} dx \]
If \( n\bar{x} = x_1 + \cdots + x_n \), then
\[
\phi = \frac{1}{(\sigma \sqrt{2\pi})^n} e^{-\frac{1}{2}n(\bar{x} - \mu)^2 + \frac{3}{2}(\mu - \bar{x})^2/2\sigma^2}
\]
where
\[
\phi_1 = \frac{1}{(\sigma \sqrt{2\pi})^n} e^{-\frac{1}{2}x_1/2\sigma^2}
\]
which is independent of \( \mu \). Thus \( \phi \) can be expressed in the form \((4a.3.1)\), which means that \( \bar{x} \) is the estimate of
\[
d(\frac{-n\bar{x}^2/2\sigma^2}{-n\bar{x}/(2\sigma)}) = \mu
\]
and has the minimum variance \((4a.2.2)\).

**Example 2.** Consider \( n \) independent observations from the population
\[
\phi = \frac{\alpha^n}{\Gamma(p)} \frac{e^{-\alpha x_p x_p^{-1}}}{x_p^{(x_p)^n}} = \phi_1 e^{-\sum x_1/\alpha}
\]
where \( \phi_1 \) is independent of \( \alpha \) and \( y = (\sum x_i)/pn \), so that the minimum variance \((4a.2.2)\) is attained for the estimate \( y \) of
\[
-d(\frac{\alpha}{-\alpha \log \alpha}) = \frac{1}{\alpha}.
\]
Also
\[
Y(y) = \left[ \frac{d}{da}(-\alpha \log \alpha) \right]^{-1} \frac{d}{da} \left( \frac{1}{n} \right) = \frac{1}{\alpha^2/n}
\]

**Example 3.** The chance of obtaining \( r \) successes in \( n \) independent trials with probability \( x \) for success in a trial is
SUFFICIENT STATISTICS AND UNBIAS: D ESTIMATES

\[ \phi = \binom{n}{r} \pi^r (1 - \pi)^{n-r} \]

\[ d \log \phi = \frac{r - n \pi}{\pi (1 - \pi)} \]

\[ I = \frac{V(r - n \pi)}{\pi^2 (1 - \pi)^2} = \frac{n}{\pi (1 - \pi)} \]

\[ E \left( \frac{r}{n} \right) = \pi \quad \text{and} \quad V \left( \frac{r}{n} \right) = \frac{\pi (1 - \pi)}{n} = \frac{1}{I} \]

This shows that \( r/n \) as the estimate of \( \pi \) has the minimum variance (4a.2.2).

Example 4. The chance of \( n \) independent observations from a Poisson distribution is

\[ \phi = e^{-\mu} \frac{\mu^{x_1} \cdots \mu^{x_n}}{x_1! \cdots x_n!} \]

\[ 1 \frac{d \phi}{\phi \, d \mu} = -n + \frac{x_1}{\mu} + \cdots + \frac{x_n}{\mu} \]

\[ I = V \left( \frac{1}{\mu} \frac{d \phi}{d \mu} \right) = \frac{1}{\mu^2} V(x_1 + \cdots + x_n) = \frac{n \mu}{\mu^2} = \frac{n}{\mu} \]

The estimate \( \hat{x} \) of \( \mu \) has the variance \( \mu/n = 1/I \) so that the minimum variance (4a.2.2) is attained.

4a.4 Sufficient Statistics and Unbiased Estimates

A statistic \( T \) is said to be sufficient for the parameter \( \theta \) if, with respect to any other statistic \( T' \), the joint probability density \( P(T, T') \) of \( T \) and \( T' \) is of the form

\[ P(T, T') = P_1(T, \theta)P_2(T' \mid T) \]

where \( P_1(T, \theta) \) is the probability density for \( T \), and \( P_2(T' \mid T) \), the relative probability density of \( T' \) given \( T \), is independent of \( \theta \). From this definition it can be shown that the necessary and sufficient condition for \( \phi \) to admit a sufficient statistic is

\[ \phi = \Phi(T, \theta) \phi_1(x_1, \cdots, x_n) \quad (4a.4.1) \]

where \( \phi_1 \) as a relative probability density given \( T \) is independent of \( \theta \), and \( T \) is a function of \( x \) only. The statistic \( T \) is said to be sufficient for \( \theta \).
In fact, any function of $T$ having one-to-one correspondence with $T$ will be sufficient for $\theta$.

**Theorem.** If an unbiased estimate and a sufficient statistic exist for $\psi(\theta)$, the best unbiased estimate of $\psi(\theta)$ is an explicit function of the sufficient statistic.

The statement does not imply that there exists a function of the sufficient statistic which attains the minimum variance (4a.2.2). But the best unbiased estimates have to be sought from the functions of the sufficient statistic only.

If $t$ is unbiased for $\psi(\theta)$, then

$$
\psi(\theta) = \int t \phi \, dv = \int \phi(T, \theta) \phi_T \, dv
$$

Integration over the surfaces of the constant $T$ gives

$$
\psi(\theta) = \int f(T) \phi(T, \theta) \, dT
$$

(4a.4.2)

which shows that there exists a function $f(T)$, of the sufficient statistic $T$, which is unbiased for $\psi(\theta)$. Also

$$
\int [t - \psi(\theta)]^2 \phi \, dv = \int [t - f(T)]^2 \phi \, dv + \int [f(T) - \psi(\theta)]^2 \phi(T, \theta) \, dT
$$

since the product term vanishes in virtue of (4a.4.2). Hence

$$
\int [t - \psi(\theta)]^2 \phi \, dv \geq \int [f(T) - \psi(\theta)]^2 \phi(T, \theta) \, dT
$$

i.e.,

$$
V(t) \geq V[f(T)]
$$

which proves the theorem. Thus, when a sufficient statistic exists, we need only search for the best estimates among functions of the sufficient statistic. If there exists a unique function of $T$ unbiased for $\psi(\theta)$, then this is necessarily the best. But, if more than one function of $T$ is unbiased for $\psi(\theta)$, then the one with the least variance has to be chosen. This leads to the corollary.

**Corollary.** If a function $F(T)$ of the sufficient statistic $T$ is unbiased for $\psi(\theta)$ and is also unique, then this is the best unbiased estimate.

* It can be verified that $\bar{x}$ is a sufficient statistic for $\mu$, the mean of a normal distribution. But $\bar{x}$ is not sufficient for $\mu$ for the condition (4a.4.1) is no longer true. The statement, generally made, that any function of a sufficient statistic is also sufficient is not true.
Example 1. The minimum variance unbiased estimate is unique. [Hint: If $T_1$ and $T_2$ are two such estimates, then $V([T_1 + T_2]/2) > V(T_1)$. This shows that the correlation between $T_1$ and $T_2$ is unity.]

Example 2. If $T_1$ and $T_2$ are two unbiased estimates of a parameter with variances $\sigma_1^2$, $\sigma_2^2$ and correlation $\rho$, what is the best unbiased linear combination of $T_1$ and $T_2$ and what is the variance of such a compound?

We have to minimize

$$l_1^2 \sigma_1^2 + 2\rho l_1 l_2 \sigma_1 \sigma_2 + l_2^2 \sigma_2^2$$

subject to the condition $l_1 + l_2 = 1$. This gives, in one step,

$$l_1(\sigma_1^2 - \rho \sigma_1 \sigma_2) = l_2(\sigma_2^2 - \rho \sigma_1 \sigma_2)$$

giving the ratio between $l_1$ and $l_2$.

Example 3. Suppose that $T_1$ in the above example is an unbiased minimum variance estimate and $T_2$ any other unbiased estimate with variance $\sigma_2^2/c$ where $V(T_1) = \sigma^2$. Then the correlation between $T_1$ and $T_2$ is $\sqrt{c}$.

The best linear compound of $T_1$, $T_2$ has the coefficients $l_1$ and $l_2$ satisfying the condition

$$l_1(\sigma_1^2 - \rho \sigma_1 \sigma_2) = l_2(\sigma_2^2 - \rho \sigma_1 \sigma_2)$$

giving the condition that the variance of the best compound is not less than $\sigma^2$ or simply, since $T_1$ cannot be improved upon, $l_2 = 0$, which means $c = \rho \sqrt{c}$.

Example 4. If $T_1$, $T_2$ are two unbiased statistics having the same variance, then their correlation is at least $2\rho - 1$, where $\rho$ is the ratio of the variance of the best estimate to the common variance of $T_1$ and $T_2$.

[Hint: Consider the statistic $T = (T_1 + T_2)/2$ and express the condition $V(T) \geq$ the least variance.]

4a.5 Distributions Admitting Sufficient Statistics

The necessary and sufficient condition that a distribution admits a sufficient statistic is that the probability density can be written in the form

$$\phi = \Phi(T, \theta) \phi_1(x_1, \cdots, x_n)$$

(4a.5.1)

where $\Phi(T, \theta)$ is the density of the statistic $T$ and $\phi_1(x_1, \cdots, x_n)$, the density of the sample given $T$, is independent of $\theta$. It is obviously not enough to state the necessary and sufficient condition as the factorizability of $\phi$ into $\Phi(T, \theta)$ and $\phi_1(x_1, \cdots, x_n)$, where $T$ is a function of $x_1, \cdots, x_n$, and $\phi_1$ is independent of $\theta$ unless the range of $x$ is independent
of $\theta$. For instance consider

$$\phi = \left(\frac{\theta}{e^\theta - 1}\right)^n e^{\theta x_1 + \cdots + x_n} \quad (4a.5.2)$$

where the range of each $x$ is from 0 to $\theta$. This does not admit a sufficient statistic.

Let us now assume that $x_1, \ldots, x_n$ are independent observations from the same distribution so that

$$\phi = p(x_1, \theta) \cdots p(x_n, \theta)$$

If this is factorizable into $\Phi(T, \theta)$ and $\phi_1(x_1, \cdots, x_n)$ then, assuming that the functions are partially differentiable with respect to $\theta$, we obtain

$$\sum \frac{\partial \log p(x_i, \theta)}{\partial \theta} = \frac{\partial \log \Phi(T, \theta)}{\partial \theta} = G(T, \theta) \quad (4a.5.3)$$

Since this holds for all $\theta$, any value of $\theta$ can be substituted in (4a.5.3) to obtain the relation

$$u = \sum u(x_i) = g(T)$$

connecting $T$ and the statistic $u = \sum u(x_i)$. If $g(T)$ and $u(x)$ are differentiable functions, it follows that

$$\frac{\partial u}{\partial x_i} = \frac{du(x_i)}{dx_i} = \frac{dp(T)}{dT} \frac{dT}{dx_i}$$

Also from (4a.5.3)

$$\frac{\partial G(T, \theta)}{\partial \theta} \frac{\partial T}{\partial x_i} = \frac{\partial^2 \log p(x_i, \theta)}{\partial \theta \partial x_i}$$

Therefore, for all $i$

$$\frac{\partial^2 \log p(x_i, \theta)}{\partial \theta \partial x_i} + \frac{du(x_i)}{dx_i} = \frac{\partial G(T, \theta)}{\partial T} + \frac{dp(T)}{dT}$$

$$= \lambda_i(\theta) \quad \text{a function of } \theta \text{ only}$$

Integrating with respect to $T$,

$$G(T, \theta) = \lambda_1(\theta)g(T) + \lambda_2(\theta)$$

and then with respect to $\theta$,

$$\log \phi(x_1, \cdots, x_n) = \theta_1 g(T) + \theta_2 h(x_1, \cdots, x_n)$$

$$\phi(x_1, \cdots, x_n) = \phi_0 e^{\theta_1 g(T) + \theta_2 h(x_1, \cdots, x_n)} \quad (4a.5.4)$$
AN OPTIMUM PROPERTY OF SUFFICIENT STATISTICS

where \( g(T) \), \( \phi_1 \), and \( \phi_2 \) are functions of \( x_1, \ldots, x_n \) only, and \( \theta_1, \theta_2 \) are functions of \( \theta \) only. This is obviously not a necessary and sufficient condition because the only condition used is the factorizability of \( \phi \). In fact the illustration in (4a.5.2) has the same form as (4a.5.4). But when a sufficient statistic exists, the distribution must necessarily be of the form (4a.5.4).

4a.6 An Optimum Property of Sufficient Statistics

In 4a.5 it was shown that the distribution admitting a sufficient statistic is of the form

\[
\phi = \exp \left[ t_1 \theta_1 + \theta_2 + t_2 \right]
\]

Let \( F(t_1) \) be any function of \( t_1 \) with the expectation \( \psi(\theta) \). Consider an alternative function unbiased for \( \psi(\theta) \) but differing from \( F(t_1) \) by \( f(t_1) \).

Then

\[
\int f(t_1) \exp \left[ t_1 \theta_1 + \theta_2 + t_2 \right] \, d\theta = 0
\]

Continuous differentiation, when permissible, yields

\[
\int t_1^k f(t_1) \exp \left[ t_1 \theta_1 + \theta_2 + t_2 \right] \, d\theta = 0 \quad \text{for all } k
\]

or

\[
\int e^{t_1 f(t_1)} \exp \left[ t_1 \theta_1 + \theta_2 + t_2 \right] \, d\theta = 0
\]

From Fourier's inversion theorem it is known that, if \( f(t_1) \exp \left( t_1 \theta_1 + \theta_2 + t_2 \right) \) is continuous, then it must be zero almost everywhere. Since the second expression cannot be zero, it follows that \( f(t_1) \) is unique as an estimate of its expected value, and therefore the best possible. Hence the following theorem is obtained.

**Theorem:** Any function of the sufficient statistic is the best estimate of its expected value under the regularity conditions assumed above.

This is a general demonstration of the properties discussed in examples 1 and 2 below. The result is true under less stringent conditions than those assumed above.

**Example 1.** Consider \( n \) independent observations from the normal population

\[
\phi = \phi_1 \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] dx
\]

\[
\phi_1 = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]
where $\phi_1$ is independent of $\mu$ and $n\bar{x} = x_1 + \cdots + x_n$. This shows that $\bar{x}$ is sufficient for $\mu$. In fact, we have seen that $\bar{x}$ has the minimum variance (4a.2.2) as an estimate of $\mu$. Suppose that instead of $\mu$ we are seeking the estimate of $\mu^2$. Evidently $\bar{x}$ is sufficient for $\mu^2$ or for any function of $\mu$. The parameter $\sigma^2$ being considered known, an unbiased estimate of $\mu^2$ is

$$X = \bar{x}^2 - \frac{\sigma^2}{n}$$

$$V(X) = E(\bar{x}^4) - \left(\mu^2 + \frac{\sigma^2}{n}\right)^2$$

$$= \mu^4 + \frac{6\mu^2\sigma^2}{n} + \frac{3\sigma^4}{n^2} - \mu^4 - \frac{2\sigma^2\mu^2 - \sigma^4}{n^2}$$

$$= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$$

The minimum variance (4a.2.2) is

$$\frac{(d\mu^2/d\mu)^2}{I} = \frac{4\mu^2\sigma^2}{n}$$

since $I = n/\sigma^2$, which is smaller than $V(X)$. But $X$ is the best unbiased estimate since, as shown below, it has the minimum attainable variance. To prove this it will be shown that there exists no function of $\bar{x}$, the sufficient statistic, such that its expectation is $\mu^2$ and its variance less than that of $V(X)$. Let the alternative function differ from $X$ by $f(\bar{x})$. Since both are unbiased, it follows that

$$\exp\left[-\frac{n\mu^2}{2\sigma^2}\right] \int f(\bar{x}) \exp\left[-\frac{n(\bar{x}^2 - 2\mu\bar{x})}{2\sigma^2}\right] d\bar{x} = 0$$

Omitting the term before the integral and differentiating twice with respect to $\mu$, the following relation is obtained.

$$\int \frac{n^2 f(\bar{x}) \bar{x}^2}{\sigma^4} \exp\left[-\frac{n(\bar{x}^2 - 2\mu\bar{x})}{2\sigma^2}\right] d\bar{x} = 0$$

This shows that cov $[f(\bar{x}), \bar{x}^2] = 0$.

$$V[X + f(\bar{x})] = V(X) + V[f(\bar{x})] + 2 \text{cov} \left[f(\bar{x}), \bar{x}^2 - \frac{\sigma^2}{n}\right]$$

$$= V(X) + V[f(\bar{x})]$$

$$\geq V(X)$$
which shows that $X$ is better than any other estimate, or, in other words, it has minimum variance as an estimate of $\mu$.

**Example 2.** Consider $n$ independent observations from the population

\[
\frac{\alpha^p}{\Gamma(p)} e^{-\alpha x x^{p-1}} dx
\]

\[
\phi = \frac{\alpha^p}{[\Gamma(p)]^n} e^{-\alpha x x^{p-1}}
\]

which shows that $\bar{x}$ is sufficient for $\alpha$. Since $E[(np - 1)/\bar{x}] = \alpha$, the statistic $(np - 1)/\bar{x}$ is unbiased for $\alpha$.

\[
V\left(\frac{np - 1}{nx}\right) = \alpha^2 \left| \frac{np - 1}{np - 2} - 1 \right| = \frac{\alpha^2}{np - 2}
\]

The minimum variance (4a.2.2) is

\[
\frac{1}{I} = \frac{\alpha^2}{pn}
\]

which is smaller than $V[(np - 1)/\bar{x}]$. Let an alternative estimate differ from $(np - 1)/\bar{x}$ by $f(\bar{x})$. Then

\[
\int f(\bar{x}) \exp\left[-\alpha \bar{x}\right] d\bar{x} = 0 \quad \text{for all } \alpha > 0
\]

which gives the result that

\[
\int \phi f(\bar{x}) \exp\left[-\alpha \bar{x}\right] d\bar{x} = 0 \quad \text{for all } \alpha
\]

or

\[
\int \phi^{\gamma} f(\bar{x}) \exp\left[-\alpha \bar{x}\right] d\bar{x} = 0, \quad \text{where } i = \sqrt{-1}
\]

From the Fourier inversion theorem we obtain

\[
f(\bar{x}) \exp\left[-\alpha \bar{x}\right] = 0
\]

or $f(\bar{x}) = 0$ almost everywhere, so that the function $(np - 1)/\bar{x}$ is unique as the estimate of $\alpha$.

**Example 3.** Consider $n$ independent observations $x_1, \ldots, x_n$ from a rectangular distribution in the range 0 to $\beta$. The biggest observation $x_n$ has the distribution

\[
x_n \sim x_n = x_{n-1} \quad \text{for } n > 0
\]

\[
x_n \sim x_n = x_{n-1} \quad \text{for } n > 0
\]
so that

\[ E(x_\beta) = \frac{n}{n + 1} \beta \]

or

\[ E\left( \frac{n + 1}{n} x_\beta \right) = \beta \quad V\left( \frac{n + 1}{n} x_\beta \right) = \frac{\beta^2}{n(n + 2)} \]

Also, \( x_\beta \) is sufficient for \( \beta \); hence the unbiased minimum variance estimate must be a function of \( x_\beta \) only. The statistic \( (1 + 1/n)x_\beta \) is unbiased for \( \beta \). It is also unique for, if another statistic differs from this by \( \phi(x_\beta) \), then

\[ \int_0^\beta \phi(x_\beta)x_\beta^{n-1} \, dx_\beta = 0 \]

for all \( \beta \). This means that

\[ \phi(x)x^n = 0 \]

or \( \phi(x) = 0 \) almost everywhere. The statistic \( (1 + 1/n)x_\beta \) is the best unbiased estimate of \( \beta \).

**Example 4.** Consider a rectangular distribution in the range \( \alpha \) to \( \beta \). The biggest and smallest observations \( x_\beta \) and \( x_1 \) form a sufficient set of statistics for \( \alpha \) and \( \beta \). The joint distribution of \( x_\beta \) and \( x_1 \) is

\[ n(n-1)(\beta - \alpha)^{n-1}(x_\beta - x_1)^{n-2} \, dx_\beta \, dx_1 \]

The statistics

\[ T_1 = \frac{n}{n-1} \left( x_\beta - x_1 \right) \]

and

\[ T_2 = \frac{n}{n-1} \left( x_\beta - x_1 \right) \]

are unbiased for \( \alpha \) and \( \beta \), respectively. It can also be shown, as in example 3, that these are the best.

**Example 5.** The unbiased minimum variance estimates of the central point \( (\alpha + \beta)/2 \) and the range \( (\beta - \alpha) \) are

\[ \frac{x_\beta + x_1}{2} \quad \text{and} \quad \frac{n + 1}{n-1}(x_\beta - x_1) \]

**Example 6.** Show that in example 3 any function \( t(x_\beta) \) of \( x_\beta \) unbiased for \( \beta \) has the terminal value

\[ t(\beta) = \left( 1 + \frac{1}{n} \right) \beta \]
4a.7 More Stringent Inequalities for the Variance of an Estimate

It was shown in 4a.2 that the minimum variance for an unbiased estimate of $\psi(\theta)$ is not less than

$$\frac{|\psi'(\theta)|^2}{I}$$

Considering the equality

$$\int t \, dt = \psi(\theta)$$

and differentiating $k$ times

$$\int \frac{d^k \psi}{d\theta^k} \, dt = \frac{d^k \psi}{d\theta^k}$$

which leads to the relationship

$$V(t) \leq \frac{(d^k \psi/d\theta^k)^2}{J_{kk}}$$

where

$$J_{kk} = V\left(\frac{1}{\theta} \frac{d\phi}{d\theta}\right) \quad k = 1, 2, \ldots$$

Thus a chain of relationships can be derived of which the result obtained in 4a.2 is a special case. More generally, if

$$J_{kl} = \text{cov}\left(\frac{1}{\theta} \frac{d^2 \phi}{d\theta^2}, \frac{1}{\theta} \frac{d^2 \phi}{d\theta^2}\right)$$

then the square of the multiple correlation of $t$ on $\frac{1}{\theta} \frac{d\phi}{d\theta}, \frac{1}{\theta} \frac{d^2 \phi}{d\theta^2}, \ldots$ is

$$\left[\Sigma J_{kk} \frac{d^2 \psi(\theta)}{d\theta^2} \frac{d^2 \psi(\theta)}{d\theta^2}\right]^{-1} = V(t)$$

where $(J^{kk})$ is the matrix inverse to $(J_{kk})$. Since this is not greater than unity, we obtain the relationship

$$V(t) \leq \Sigma J_{kk} \frac{d^2 \psi(\theta)}{d\theta^2} \frac{d^2 \psi(\theta)}{d\theta^2}$$

This result is due to Bhattacharya (1947). Since the multiple correlation obtained by considering all the variables of a group is greater than that for a subset, it follows that the lower bound to the variance of an estimate can be improved by the addition of more variables of the type $(1/\phi)(d^2 \phi/d\theta^2)$. 
4a.8 The Case of Several Parameters

Minimal Set of Sufficient Statistics. A set of statistics $T_1, \ldots, T_m$ is said to be a minimal set of sufficient statistics if $m$ is the smallest number for which

$$\phi(x \mid \theta) = P_1(T_1, \ldots, T_m \mid \theta)P_2(x \mid T)$$

(4a.8.1)

where $P_1$ is the probability density of $T_1, \ldots, T_m$, and $P_2$ the probability of the observations given the statistics $T_1, \ldots, T_m$, is independent of $\theta$.

There is no restriction on $m$, which may be greater than, equal to, or less than $k$, the number of parameters involved in $\phi$.

Information Matrix:

Let

$$\phi_{ij} = -\frac{\partial^2 \log \phi}{\partial \theta_i \partial \theta_j} \quad \text{and} \quad E(\phi_{ij}) = I_{ij}$$

The matrix $I = (I_{ij})$, $(i, j = 1, \ldots, k)$, is called the information matrix. If $(S_{ij})$ denotes the information matrix obtained from the distribution $P_1(T \mid \theta)$, then it follows from the definition given in (4a.8.1) that $(I_{ij}) = (S_{ij})$. Also, if $\phi(x, \theta)$ is the probability density corresponding to $n$ independent sets of observations from the same population, then $I_{ij} = nJ_{ij}$, where $J_{ij}$ is the element of the information matrix corresponding to a single set of observations. This is the additive property of information.

The following theorem may now be proved.

Theorem. Let $t_1, \ldots, t_r$ be $r \leq k$ functionally independent statistics such that

(i) $E(t_i) = \psi_i(\theta_1, \ldots, \theta_k)$
(ii) $E(t_i - \psi_i)(t_j - \psi_j) = V_{ij}$

then:

(A) There exist functions $M_1, \ldots, M_r$ of the minimal set of sufficient statistics such that

(a) $E(M_i) = \psi_i(\theta_1, \ldots, \theta_k)$;
(b) if $U = (U_{ij})$ where $U_{ij} = E(M_i - \psi_i)(M_j - \psi_j)$ and $V = (V_{ij})$, then the matrix $(V - U)$ is positive definite or semi-definite.

(B) If the ranges of integration do not involve the parameters, and $I^{-1}$, the inverse of $I$, exists, and $\Delta = (\partial \phi / \partial \theta)$, $(i = 1, \ldots, r, j = 1, \ldots, k)$, then the matrix $V - \Delta I^{-1} \Delta'$, where $\Delta'$ is the transpose of $\Delta$, is positive definite or semi-definite.
THE CASE OF SEVERAL PARAMETERS

Proof. Since

\[
\int t|\phi(x, \theta)\;dv = \int P_1(T | \theta)\;dT \int t|P_2(x | T)\;dv'
\]

\[
= \int M_i(T)P_1(T | \theta)\;dT
\]  

(4a.8.2)

it follows that \(M_i\) is a function of \(T\) only such that

\[
E(M_i) = E(\psi_i) = \psi_i, \quad i = 1, 2, \cdots, r
\]

This proves (A)(a) of the theorem.

Consider \(\Sigma \psi_i \phi_i\) where \(\psi_i\) are arbitrary constants.

\[
E(\Sigma \psi_i) = \Sigma \psi_i \phi_i = E(\Sigma \psi_i M_i)
\]

\[
\int [\Sigma (t_i - \psi_i)]^2 \phi(x | \theta)\;dv
\]

\[
= \int \Sigma (t_i - M_i) + \Sigma (M_i - \psi_i)^2 \phi(x | \theta)\;dv
\]

\[
+ 2 \int [\Sigma (M_i - \psi_i)]P_1(T | \theta)\;dT \int [\Sigma (t_i - M_i)]P_2(x | T)\;dv'
\]

By virtue of the result (4a.8.2), the last term vanishes identically, leaving only two positive quantities. If we retain only the latter, the following relationship is obtained.

\[
V(\Sigma \psi_i) \leq V(\Sigma \psi_i M_i)
\]

or

\[
\Sigma \psi_i \phi_i \leq \Sigma \Sigma \psi_i \psi_i \phi_i \quad \text{or} \quad \Sigma \psi_i \phi_i \leq 0
\]

This means that \((V - U)\) is positive or semi-definite. This proves the result (A)(b) of the theorem.

Since

\[
\int t|\phi(x, \theta)\;dv = \psi_i
\]

\[
\int t|\frac{d\phi}{d\theta_i}\;dv = \frac{d\psi_i}{d\theta_i}
\]
so that by considering the dispersion matrix of \( t_i, \ldots, t_r, \frac{1}{\phi} \frac{d\phi}{d\theta_k} \), we find that the partitioned matrix

\[
\begin{pmatrix}
V & \Delta \\
\Delta' & I
\end{pmatrix}
\]

is positive or semi-definite.*

Consider the determinant

\[
\begin{vmatrix}
\delta_r & -\Delta I^{-1} \\
0 & I^{-1}
\end{vmatrix}
\]

where \( \delta_r \) is the unit square matrix of order \( r \), which is always positive. The product

\[
\begin{vmatrix}
\delta_r & -\Delta I^{-1} \\
0 & I^{-1}
\end{vmatrix} \begin{vmatrix}
V & \Delta \\
\Delta' & I
\end{vmatrix} = \begin{vmatrix}
V - \Delta I^{-1} \Delta' & 0 \\
I^{-1} \Delta' & \delta_r
\end{vmatrix} \geq 0
\]

or \( |V - \Delta I^{-1} \Delta'| \geq 0 \). This result holds true even for a subset of the statistics \( t_1, \ldots, t_r \), which means that the matrix \( V - \Delta I^{-1} \Delta' \) is positive definite or semi-definite. This proves result (B) of the theorem.

A series of corollaries can be obtained from results (A) and (B) of the above theorem.

**Corollary 1.** By considering only the diagonal elements in \( V - \Delta I^{-1} \Delta' \)

\[
V_{ii} < \sum \sum I_{mn} \frac{\partial \psi_i}{\partial \theta_m} \frac{\partial \psi_i}{\partial \theta_n}
\]

(4a.8.3)

where \( I_{mn} \) are the elements of the matrix reciprocal to the information matrix \( (I_{nn}) \). This shows that the variance of the estimate of \( \psi_i \) is not less than a quantity which is defined independently of any method of estimation. This is the generalization to many parameters of the expression derived in (4a.2.2).

If \( \psi_i = \theta_i \) (\( i = 1, \ldots, k \)), the relationship (4a.8.3) reduces to

\[
V_{ii} < I_{ii}
\]

These are not necessarily the minima attainable. Observe that \( I_{ii} \) is greater than \( 1/I_{ii} \), which is the limit obtained in (4a.2.2) for the estimate of \( \theta_i \). When the values of \( \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k \) are known, then

* The matrix \( I \) is the information matrix, not to be misunderstood for the unit matrix introduced in Chapter 1. To distinguish this the unit matrix is here represented by \( I \) which is also an accepted symbol for a diagonal matrix.
the limit (4a.2.2) is applicable. If not, the estimate of $\theta_i$ has to be independent of the above quantities and for this reason the limit is increased.

**Corollary 2.** Since the matrix $(V - U)$ is positive definite or semi-definite, it follows that $V_{ii} < U_{ii}$, ($i = 1, \cdots, n$), which shows that estimates with the minimum attainable variances are explicit functions of sufficient statistics.

**Corollary 3.** Since the matrix $(V - \Delta^{-1}\Delta')$ is positive definite or semi-definite, it follows that (see example 1 in Ic.5)

$$|V| < |\Delta^{-1}\Delta'|$$

The quantity $|V|$ is called the generalized variance of the estimates. The above result shows that this is not less than a quantity which is defined independently of any method of estimation.

**Corollary 4.** Since $(V - U)$ is positive definite or semi-definite, it follows that $|V| < |U|$ (example 1 in Ic.5). This shows that the estimates with the minimum possible generalized variance are functions of the sufficient statistics.

**Corollary 5.** If

$$V_{ii} = \sum\sum I^{\kappa} \frac{\partial \phi_i}{\partial \theta_n} \frac{\partial \phi_i}{\partial \theta_n},$$

in which case the estimate of $\phi_i$ has the minimum variance, then

$$V_{ij} = \sum\sum I^{\kappa} \frac{\partial \phi_i}{\partial \theta_n} \frac{\partial \phi_j}{\partial \theta_n},$$

so that the covariance of this best estimate with any estimate of any other parametric function has a fixed value defined independently of any method of estimation. This follows from the fact that the determinant

$$|V - \sum\sum I^{\kappa} \frac{\partial \phi_i}{\partial \theta_n} \frac{\partial \phi_j}{\partial \theta_n}|,$$

which is a subdeterminant of $|V - \Delta^{-1}\Delta'|$, is not less than zero.

**Example.** Consider $n$ independent observations from the normal population

$$\frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] dx$$
It is easy to verify that the information matrix for $\mu$ and $\sigma^2$ is

$$
(I_{\mu,\sigma^2}) = \begin{bmatrix}
\frac{n}{\sigma^2} & 0 \\
0 & \frac{n}{2\sigma^4}
\end{bmatrix}
$$

with its reciprocal

$$
(I^{*\sigma^2}) = \begin{bmatrix}
\frac{\sigma^2}{n} & 0 \\
0 & \frac{2\sigma^4}{n}
\end{bmatrix}
$$

Since $\bar{x}$ as an estimate of $\mu$ has the minimum possible variance, it follows that any estimate of $\sigma^2$ has zero correlation with $\bar{x}$, since $I^{*\sigma^2} = 0$. This result can be extended to the case of multivariate normal populations where it can be shown that the means are uncorrelated with all possible estimates of the variances and covariances.

As for the estimate of $\sigma^2$, let us consider

$$
\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}
$$

$E(\hat{\sigma}^2) = \sigma^2$

$V(\hat{\sigma}^2) = \frac{2\sigma^4}{n-1}$

which shows that the minimum variance $2\sigma^4/n$ is not attained. But this is the minimum attainable, as shown below. If any estimate of $\sigma^2$ differs from $\hat{\sigma}^2$ by $f(s, \bar{x})$, then

$$
\int f(s, \bar{x}) \exp \left[ \frac{-n(\bar{x} - \mu)^2 + (n-1)s^2}{2\sigma^2} \right] dv = 0
$$

Twice differentiation with respect to $\mu$ leads to the result

$$
E[(\bar{x} - \mu)^2f(s, \bar{x})] = 0
$$

Differentiation with respect to $\sigma^2$ gives

$$
E[(n(\bar{x} - \mu)^2f(s, \bar{x}) + (n-1)s^2f(s, \bar{x})] = 0
$$

or

$$
\text{cov} \{\sigma^2, f(s, \bar{x})\} = 0
$$

Consider

$$
V[\sigma^2 + f(s, \bar{x})] = V(\sigma^2) + V[f(s, \bar{x})]
$$
which means that $V(x^2)$ is the least possible. Thus $\bar{x}$ and $s^2$ are the best unbiased estimates of $\mu$ and $\sigma^2$.

4a.9 Properties of Distributions Admitting Sufficient Statistics: Several Parameters

It has been shown by Koopman (1936) that under some general conditions the distribution function $\phi(x_1, \cdots, x_r; \theta_1, \cdots, \theta_q)$ admitting a set of statistics $T_1, \cdots, T_s$ sufficient for $\theta_1, \cdots, \theta_q$ can be expressed in the form

$$\phi = \exp \left[ \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_q x_q + \theta + X \right]$$

where $X$ depends on $T$ only and $\theta$ on $\theta$ only. Using the relation

$$\int \phi \, dv = 1$$

we find on suitable differentiations

$$E(X_i) = -\frac{\partial^2 \phi}{\partial \theta_i}$$

$$V(X_i) = -\frac{\partial^2 \theta}{\partial \theta_i^2}$$

$$\text{cov}(X_iX_j) = -\frac{\partial^2 \phi}{\partial \theta_i \partial \theta_j}$$

The element $I_{ij}$ of the information matrix for the functions $\theta_i = \theta_0$, $(i = 1, \cdots, q)$, is

$$E\left( -\frac{\partial^3 \log \phi}{\partial \theta_i \partial \theta_j \partial \theta_k} \right) = -\frac{\partial^2 \theta}{\partial \theta_i \partial \theta_j}$$

If $(I_{ij})$ is the matrix inverse to $(I_{ij})$, then the minimum possible variance for an estimate of the parameter $-\partial \theta/\partial \theta_i$ is

$$\Sigma \sum I^{mn} \frac{\partial^2 \phi}{\partial \theta_m \partial \theta_n \partial \theta_i} = \Sigma \sum I^{mn} I_{mn} = I_{ii} = V(X_i)$$

so that the minimum variance (4a.8.3) is attainable for the estimates of $-\partial \theta/\partial \theta_i$, $(i = 1, \cdots, q)$. This shows that, when sufficient statistics equal in number to the unknown parameters exist, it is possible to find functions of parameters which admit estimates with the minimum variance (4a.8.3).
Example. For the example considered in 4a.8, \( \phi \) can be expressed

\[
\phi = \exp \left[ -\frac{n(x - \mu)^2}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2} - n \log \sigma + X \right]
\]

or

\[
\phi = \exp \left[ \frac{n\mu}{\sigma^2} - \frac{(n-1)s^2 + ns^2}{2\sigma^2} - n \log \sigma + X \right]
\]

\[
\theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = -\frac{1}{\sigma^2}, \quad \theta = \left( n \log \sigma + \frac{n s^2}{2\sigma^2} \right)
\]

\[
\frac{d\theta}{d\mu} = -\frac{n\mu}{\sigma^2} = \frac{\partial \theta}{\partial \theta_1} \frac{\partial \theta_1}{d\mu} + \frac{\partial \theta}{\partial \theta_2} \frac{\partial \theta_2}{d\mu} = \frac{d\theta}{\partial \theta_1} \frac{d\theta}{\partial \theta_2}
\]

or

\[
\frac{d\theta}{d\theta_1} = -\frac{d\theta}{d\theta_2} = \frac{n\mu}{\sigma^2}
\]

Similarly,

\[
\frac{d\theta}{d\theta_2} = \frac{n(\mu^2 + \sigma^2)}{\sigma^2}
\]

The parametric functions \( \mu \) and \( \mu^2 + \sigma^2 \) admit estimation with the minimum possible variance. Their estimates are \( \bar{x} \) and \( \{n(i - \mu)^2 + (n-1)s^2 + ns^2\}/n \). It is seen in the previous example that \( \bar{x} \) and \( s^2 \) are the best for \( \mu \) and \( \sigma^2 \). In general, it can be proved that any function of the sufficient set of statistics has the minimum attainable variance as an estimate of its expected value.

This fundamental concept of sufficient statistics is due to R. A. Fisher, who recommended, as a first step in any methodological problem, the replacement of a sample by an exhaustive set of sufficient statistics. It is already known that efficient estimates derived by the method of maximum likelihood are functions of sufficient statistics. The author has shown (Rao, 1945) that minimum variance estimates must necessarily be functions of sufficient statistics. The 1945 paper contains the bulk of the matter (on limits to variance) treated in 4a.

4b Estimation by the Method of Maximum Likelihood

4b.1 The Principle of Maximum Likelihood

If \( \phi \) is the probability density of the observations, then the likelihood \( L \) of the parameters occurring in \( \phi \) is defined to be any function

* In statistical literature the term "likelihood of the observations" is often wrongly used to mean the probability density of the observations. The probability density for a given set of observations may be considered as a function of the parameters which is otherwise termed as the likelihood of the parameters.
proportional to $\phi$, the constant of proportionality being independent of the parameters. The principle of maximum likelihood consists in accepting as the best estimate of the parameters those values of the parameters which maximize the likelihood for a given set of observations. The estimates thus obtained from a primitive postulate satisfy some optimum properties which are considered below.

If $T_1, \ldots, T_m$ constitute a minimal set of sufficient statistics, then $\phi$ is of the form

$$
\phi = P_1(T_1, \ldots, T_m | \theta_1, \ldots, \theta_m)P_2(x_1, x_2, \ldots | T_1, \ldots, T_m)
$$

so that maximizing $\phi$ is equivalent to maximizing $P_1(T | \theta)$. The estimates of the parameters are necessarily functions of these sufficient statistics. This shows that the maximum likelihood estimates satisfy the necessary condition for possessing the minimum attainable variance. Under some conditions, when the number of sufficient statistics is equal to the number of parameters to be estimated, it was shown in 4a.9 that this is a sufficient condition for minimum variance estimates of suitably chosen parametric functions to exist. The existence of sufficient statistics equal in number to the parameters to be estimated is rare, and it is of importance to study what properties the maximum likelihood estimates obey in general.

4b.2 Consistency and Bias

A statistic $t_n$, a function of the $n$ observations in the sample, is said to be a consistent estimate of a parameter $\theta$ if, for any two positive numbers $\delta$ and $\epsilon$, a number $n_0$ exists such that when $n$ exceeds $n_0$ the probability that

$$
|t_n - \theta| > \delta
$$

is less than $\epsilon$. This implies that with the increase in $n$, the sample size, the chance that the difference between the statistic $t_n$ and the parameter $\theta$ will exceed any given amount decreases. If such statistics are used, the accuracy of the estimate increases with the increase in the observations, and ultimately the true value of the parameter is approached.

The property of consistency can be simply expressed as

$$
P(|t_n - \theta| < \delta) > 1 - \epsilon
$$

or $t_n \rightarrow \theta$ stochastically. This must be differentiated from the mathematical limit where the property $|t_n - \theta| < \delta$ holds unconditionally when $n > n_0$. In a stochastic limit the statistic $t_n$ can differ from $\theta$ by more than $\delta$ when $n > n_0$, but it does so with a frequency tending to zero as $n$ becomes large. If $n_0$, so determined, is independent of $\theta$ in an interval $(a, b)$ the consistency is said to be uniform in $(a, b)$. 
It may be noted that consistency does not imply unbiasedness of the statistic for any given sample size; it may not be so even in the limit. Since only moderately large sample sizes are met with in practice, it is important to calculate the bias in any estimate and correct for it, if possible. In some cases the bias may be ignored if it is known to be small enough not to invalidate any inference drawn by using the estimate.

Example 1. Consider the sample of $n$ observations from a normal population. The m.l. (maximum likelihood) estimate of $\sigma^2$ is $\frac{\sum(x_i - \bar{x})^2}{n}$ which has the expected value $(n - 1)\sigma^2/n$, so that there is some underestimation. In such a case the bias can be corrected by using the estimate $\frac{\sum(x_i - \bar{x})^2}{n - 1}$ whose expectation is exactly $\sigma^2$. Such a correction is unimportant when the sample is large.

Example 2. Consider $n$ observations from the Cauchy distribution. It was shown in 2a.7 that the mean of $n$ observations has the same probability density as that of a single observation. The probability of
\[
| \bar{x} - \mu | > \delta
\]
where $\delta$ is an assigned quantity remains the same for any $n$ so that $\bar{x}$ is not consistent for $\mu$.

Example 3. A statistic $T_n$ such that
\[
E(T_n) = \theta_n \to \theta \quad \text{and} \quad V(T_n) \to 0 \quad \text{as} \quad n \to \infty
\]
is consistent.

To prove this we shall first prove a lemma due to Tschebyscheff.

Lemma. If $x$ is a stochastic variable such that $E(x) = \mu$ and $V(x) = \sigma^2$, then
\[
P(\{ | x - \mu | > \delta \}) < \frac{\sigma^2}{\delta^2}
\]
where $\delta^2$ is any assigned quantity.

Proof. If $f(x)$ denotes the probability density of $x$, then
\[
\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) \, dx
\]
\[
\geq \int_{-\infty}^{-\delta} (x - \mu)^2 f(x) \, dx + \int_{\delta}^{+\infty} (x - \mu)^2 f(x) \, dx
\]
\[
\geq \delta^2 \left[ \int_{-\infty}^{-\delta} f(x) \, dx + \int_{\delta}^{+\infty} f(x) \, dx \right]
\]
Hence the lemma. Using this lemma in the above problem we have

$$P(\| T_n - \theta_n \| < \delta > 1 - V(T_n) \| \delta^2$$

If $| T_n - \theta_n | < \delta$, then $| T_n - \theta | < \delta + | \theta - \theta_n |$ and

$$P(\| T_n - \theta \| < \delta + | \theta - \theta_n | > P(\| T_n - \theta_n \| < \delta > 1 - V(T_n) \| \delta^2$$

Since $\theta_n \to \theta$ and $V(T_n) \to 0$, there exists an $n_0$ such that for all $n > n_0$

$$| \theta - \theta_n | < \delta_1 \quad \text{and} \quad V(T_n) < \delta^2$$

If $| T_n - \theta | < \delta + | \theta - \theta_n |$, then $| T_n - \theta | < \delta + \delta_1$; therefore for $n > n_0$

$$P(\| T_n - \theta \| < \delta + \delta_1 > P(\| T_n - \theta \| < \delta + | \theta - \theta_n | > 1 - \varepsilon$$

Since $\delta$ and $\delta_1$ are arbitrary, the result is established. It may be inferred that the stochastic convergence is uniform if the mathematical convergence of $E(T_n)$ and $V(T_n)$ is uniform.

Example 4. Consider the bivariate normal distribution

$$\text{const. exp} - \frac{1}{2(1 - \rho^2)} \left[ \frac{(x - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x - \mu_1)(y - \mu_2)}{\sigma_1 \sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2} \right] dz \, dy$$

The bivariate moment $\mu_n$ is defined by

$$E(z - \mu_1)^n(y - \mu_2)^n = \sigma_1^n \sigma_2^n E(\xi^n)$$

where

$$E(\xi^n) = \text{const.} \int \exp \left[ - \frac{1}{2(1 - \rho^2)} (t^2 - 2 t \xi + \xi^2) \right] t^n d\xi$$

$$= \text{const.} \int \xi^n e^{-t^2/2} dt \int t^n \exp \left[ - \frac{1}{2(1 - \rho^2)} (t - \rho t')^2 \right] d\eta$$

$$= \alpha \nu_{n+1} + \alpha \nu_{n+1} \nu_{n+2} + \cdots + \alpha \nu_{n+1} \nu_{n+2} + \cdots$$

(by symmetry)

where

$$\alpha = (1 - \rho^2)^{1/2}$$
\[ v_t = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^e^{-x^2} \, dx = 0 \quad \text{if } t \text{ is odd} \]

\[ = \frac{[(2-\ell)/2]}{(\ell/2)!} \quad \text{if } t \text{ is even} \]

It is easy to see that the expression for \( E(\xi^t) \) vanishes whenever \((\ell + s)\) is odd. We thus obtain

\[
\begin{align*}
\mu_{20} &= \sigma_1^2 & \mu_{02} &= \sigma_2^2 & \mu_{11} &= \rho \sigma_1 \sigma_2 \\
\mu_{40} &= 3\sigma_1^4 & \mu_{04} &= 3\sigma_2^4 & \mu_{13} &= 3\rho \sigma_1 \sigma_2^3 \\
\mu_{31} &= 3\rho \sigma_1^3 \sigma_2 & \mu_{22} &= (1 + 2\rho^2)\sigma_1^2 \sigma_2^2 & \text{and so on}
\end{align*}
\]

The ratio \( y/x \) is called an index. To determine the moments of \( y/x \) it can be formally expanded.

\[
y = \frac{\mu_2}{\mu_1} \left( 1 + \frac{\eta}{\mu_2} \right) \left( 1 + \frac{\xi}{\mu_1} \right)^{-1} = \frac{\mu_2}{\mu_1} \left( 1 - \frac{\xi}{\mu_1} + \frac{\xi^2}{\mu_1^2} - \cdots + \frac{\eta}{\mu_2} \left( 1 - \frac{\xi}{\mu_1} + \frac{\xi^2}{\mu_1^2} - \cdots \right) \right)
\]

Taking expectations of both sides,

\[
E\left( \frac{y}{x} \right) = \frac{\mu_2}{\mu_1} \left[ 1 + \frac{\mu_{20} + \mu_{40} + \cdots + \mu_{11} - \mu_{31}}{\mu_{1} \mu_{1} \mu_{2}} \right] = \frac{\mu_2}{\mu_1} \left[ 1 + \sum_{t=0}^{\infty} \frac{(2t)!}{t!} \left( \frac{\eta}{\mu_2} \right)^{2t} \left( \frac{\xi}{\mu_1} \right)^t \right]
\]

where \( v_1 \) and \( v_2 \) are the coefficients of variation of \( x \) and \( y \), respectively. The series on the right-hand side converges only when the coefficient of variation \( v_1 \) is small.

If \( n \) pairs of observations on \( x \) and \( y \) are available, we can construct two statistics

\[
T_1 = \frac{1}{n} \sum y_i \quad T_2 = \frac{\bar{y}}{\bar{x}}
\]

\[
E(T_1) = \frac{\mu_2}{\mu_1} \left[ 1 + (v_1 - \rho v_2) \sum_{t=1}^{\infty} \frac{(2t)!}{t!} \left( \frac{v_1}{v_2} \right)^{2t-1} \right]
\]

\[
E(T_2) = \frac{\mu_2}{\mu_1} \left[ 1 + \frac{1}{\sqrt{n}} (v_1 - \rho v_2) \sum_{t=1}^{\infty} \frac{(2t)!}{t!} \left( \frac{v_1}{\sqrt{n}} \right)^{2t-1} \right]
\]
since the coefficient of variation of \( x \) is \( n_1/\sqrt{n} \) and of \( y \) is \( n_2/\sqrt{n} \). It is seen that \( E(T_2) \rightarrow \mu_2/\mu_1 \) while \( E(T_1) \) remains the same for all \( n \) and since both have variances of \( O\left(\frac{1}{n}\right) \), \( T_2 \) converges stochastically to \( \mu_2/\mu_1 \) and \( T_1 \) to some other value.

\( T_1 \) is a biased estimate of \( \mu_2/\mu_1 \) and does not admit a simple correction for bias. Since the bias does not tend to zero as \( n \rightarrow \infty \), it should be considered inconsistent as an estimate of the parametric function \( \mu_2/\mu_1 \). On the other hand, \( T_2 \) is a consistent estimate of the ratio.

In biometric work, extensive use is made of the indices, and in many cases the mean index is calculated by taking the ratio of the mean values of two characters. The estimate so obtained is not strictly comparable with that obtained by taking the average of all indices. The expected value of the index may be different from the ratio of the expected values of the two characters for which \( T_2 \) is a good estimate. The index should be treated as a separate character for evaluation of its constants, mean, standard deviation, etc. A comparison of the indices involves the comparison of a function of both the mean values and second-order moments.

4b.3 The Concept of Efficiency

Of all statistics which converge stochastically to a parameter \( \theta \), the practically important ones are those that converge rapidly. Such statistics give large deviations from the true value less frequently, at least in large samples, thus satisfying the requirements of a good estimate considered in 4b.2. We need, then, a criterion for judging which of two statistics converges more rapidly. If only statistics which are asymptotically normally distributed are considered, then the rapidity of convergence is measured by invariance or the reciprocal of variance. This is because in a normal distribution the probability of a departure exceeding \( \lambda \) times the standard deviation is a decreasing function of \( \lambda \) only, so that the probability of a departure's exceeding a given value decreases with decrease in the variance. Therefore that statistic with the smallest asymptotic variance is preferred and is called the efficient estimate. The efficiency of any other estimate can be measured by the ratio of variance of the efficient estimate to that of the other.

Efficiency, though linked with minimum variance, is essentially a large sample concept. Statistics with minimum variances considered in 4a are, no doubt, the most efficient ones in large samples, in which case variance acquires a special significance. But in small samples there is not sufficient justification for using invariance as the criterion for selecting a good estimate.

It may be argued that the comparison is confined only to the class of statistics which are asymptotically normally distributed. This is no
serious objection in view of the fact that a large class of statistics obeys this property.

The properties of minimum variance estimates considered in examples of 4a.4 hold good for efficient estimates also.

Example 1. Asymptotic distribution of quantiles. Let \( f(x) \) be the probability differential of \( x \), and define

\[
F(x) = \int_{-\infty}^{x} f(z) \, dz
\]

If \( x \) is the quantile of order \( p \), \( 0 < p < 1 \), in a series of \( n \) observations, then \( \mu = [np] \) observations are less than \( x \), and \( (n - \mu - 1) \) are greater than \( x + dx \); the remaining value falls between \( x \) and \( x + dx \). Hence the probability differential of \( x \) is

\[
\frac{n!}{\mu!(n - \mu - 1)!} \left[ F(x) \right]^{\mu-1} \left[ 1 - F(x) \right]^{n-\mu-1} f(x) \, dx
\]

Let \( y = F(x) \) so that \( dy = f(x) \, dx \). The distribution of \( y \) is

\[
\frac{n!}{\mu!(n - \mu - 1)!} y^{\mu-1} (1 - y)^{n-\mu-1} \, dy
\]

Let \( z = \{ \sqrt{n(y - p)} \} / \sqrt{pq} \); then the distribution of \( z \) becomes

\[
\text{const.} \left( 1 + \frac{z \sqrt{q}}{\sqrt{np}} \right)^{[np]} \left( 1 - \frac{z \sqrt{p}}{\sqrt{np}} \right)^{\mu-[np]} \, dz = \text{const.} \ e^z \, dz
\]

where

\[
\psi = [np] \log \left( 1 + z \frac{\sqrt{q}}{\sqrt{np}} \right) + (n - [np] - 1) \log \left( 1 - z \frac{\sqrt{p}}{\sqrt{pq}} \right) = -\frac{z^2}{2}
\]

which shows that \( y \) is asymptotically normal with \( p \) as mean and \( pq/n \) as variance. Since \( y = F(x) \), assuming that the inverse function exists, we can write

\[
x = F^{-1}(y) = \phi(y)
\]

or, expanding at \( y = p \),

\[
x = \phi(p) + (y - p)\phi'(p) + \frac{(y - p)^2}{2!} \phi''(p) + \ldots
\]

Neglecting terms of the order \((y - p)^2\), we find that \( x \) is distributed normally in large samples with mean \( \phi(p) \) and variance \( (\phi'(p))^2 \lambda(y) \).
MAXIMUM LIKELIHOOD ESTIMATES

The value of \( \phi(p) = t \) is determined from the relation

\[
p = \int_{-\infty}^{t} f(x) \, dx
\]

Also, since

\[
\phi'(p) = \frac{1}{F'(t)} = \frac{1}{f(t)}
\]

\[
V(x) = [f(t)]^{-2} V(y) = [f(t)]^{-2} \frac{P_{Y}}{n}
\]

It must be noted that convergence to normality is rapid only for quantiles of order \( p \) near about \( \frac{1}{2} \). For lower or higher quantiles normality is realized only in very large samples.

Example 2. Efficiency of the median as an estimate of the mean of a normal population. For a normal population the average of the observations is the most efficient estimate of the mean, and its variance is \( \sigma^{2}/n \). The asymptotic variance of the median is

\[
\frac{[f(x)]^{-2}}{4n}
\]

where \( f(x) \), the ordinate at the mean of the normal distribution, is equal to \( 1/\sqrt{2\pi} \). The variance of the median is

\[
\frac{\pi \sigma^{2}}{2n}
\]

so that its efficiency is \( 2/\pi \) or 63.7%.

4b.4 Some Optimum Properties of Maximum Likelihood Estimates

Let \( f(x, \theta) \, dx \) be the probability distribution from which a sample of size \( n \), \( (x_1, \ldots, x_n) \), is observed. We shall denote \( f(x, \theta) \) by \( f_i \) and the product by \( L \). The following assumptions are made.

(i) The derivatives \( \frac{\partial \log L}{\partial \theta} \) and \( \frac{\partial^2 \log L}{\partial \theta^2} \) exist and are continuous for every \( \theta \) in a range \( R \), including the true value, and for almost all \( x \). For every \( \theta \) in \( R \),

\[
\left| \frac{\partial L}{\partial \theta} \right| < F_1(x) \quad \left| \frac{\partial^2 L}{\partial \theta^2} \right| < F_2(x)
\]

where \( F_1 \) and \( F_2 \) are integrable functions over \((-\infty, +\infty)\).
(ii) The derivative \( \frac{\partial^3 \log L}{\partial \theta^3} \) exists and is such that

\[
\left| \frac{\partial^3 \log L}{\partial \theta^3} \right| < M(z) \quad E[|M(z)|] < K \quad \text{(a positive quantity)}
\]

(iii) For every \( \theta \) in \( R \),

\[
\int_{-\infty}^{+\infty} - \frac{\partial^2 \log L}{\partial \theta^2} L \, dx = I(\theta)
\]

is finite and non-zero.

(iv) The range of integration is independent of \( \theta \).

Under these conditions the following theorems will be proved.

Theorem 1. With probability approaching unity as \( n \to \infty \), the likelihood equation \( \frac{\partial \log L}{\partial \theta} = 0 \) has a solution which converges in probability to the true value \( \theta_0 \). (Dugue, 1937.)

Theorem 2. Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tending to unity as the sample size tends to infinity. (Huzurbazar, 1948.)

Theorem 3. A consistent solution of the likelihood equation is asymptotically normally distributed about the true value \( \theta \). (Cramér, 1946.)

Some of the limiting theorems used in this connection are given without proof in an appendix at the end of this chapter. Under the conditions assumed we have, following Cramér (1946),

\[
\frac{\partial \log L}{\partial \theta} = \left. \left( \frac{\partial \log L}{\partial \theta} \right) \right|_{\theta = \theta_0} + (\theta - \theta_0) \left( \frac{\partial^2 \log L}{\partial \theta^2} \right)_{\theta_0} + \frac{(\theta - \theta_0)^2}{2} \left( \frac{\partial^3 \log L}{\partial \theta^3} \right)_{\theta_0}
\]

where \( \theta' \) lies in \( (\theta, \theta_0) \). Dividing both sides by \( n \), we have

\[
\frac{1}{n} \frac{\partial \log L}{\partial \theta} = \frac{B_0}{n} + B_1 (\theta - \theta_0) + \frac{B_2}{2} (\theta - \theta_0)^2
\]

where

\[
B_0 = \frac{1}{n} \sum_1^n \left( \frac{\partial \log f_i}{\partial \theta} \right)_{\theta_0} \quad E(B_0) = 0
\]

\[
B_1 = \frac{1}{n} \sum_1^n \left( \frac{\partial^2 \log f_i}{\partial \theta^2} \right)_{\theta_0} \quad E(B_1) = - \frac{I(\theta)}{n} \quad \text{where } I(\theta) \text{ is the information}
\]

\[
B_2 = \frac{1}{n} \sum_1^n \left( \frac{\partial^3 \log f_i}{\partial \theta^3} \right)_{\theta_0}
\]

By Kintchine’s theorem the quantities $B_0$, $B_1$, and $B_2$ stochastically converge to their mean values. Given two quantities $\delta$ and $\epsilon$, it is possible to find $n > n_0(\delta, \epsilon)$ such that

$$P_1 = P\{|B_0| \geq \delta^2| < \frac{\epsilon}{3}$$

$$P_2 = P\{|B_1 \geq -\frac{1}{2} I(\theta)| < \frac{\epsilon}{3}$$

$$P_3 = P\{|B_2| \geq 2K| < \frac{\epsilon}{3}$$

The probability that the sample point is such that the inequalities

$$|B_0| < \delta^2, \quad |B_1| < -\frac{1}{2} I(\theta), \quad |B_2| < 2K$$

are simultaneously satisfied is evidently greater than $1 - P_1 - P_2 - P_3 (= 1 - \epsilon)$. Let $\delta$ denote the set of such points.

For $\theta = \theta_0 \pm \delta$

$$\frac{\partial \log L}{\partial \theta} = B_0 \pm \delta B_1 + \frac{1}{2} B_2 \delta^2$$

For every point in $S$, $B_0 + \frac{1}{2} B_2 \delta^2 < (K + 1)\delta^2$, and $B_1 \delta < -\frac{1}{2} I(\theta)\delta^2/2n$. If $\delta < [I(\theta)/2n(K + 1)]$, the sign of the whole expression for $\theta = \theta_0 \pm \delta$ will be determined by $B_0$ so that $(\partial \log L/\partial \theta) > 0$ for $\theta = \theta_0 - \delta$ and $(\partial \log L/\partial \theta) < 0$ for $\theta = \theta_0 + \delta$. Since $\partial \log L/\partial \theta$ is a continuous function of $\theta$ in $R$, for almost all $\theta$, it follows that the likelihood equation has a root in $\theta_0 \pm \delta$ with probability tending to unity. This establishes theorem 1.

Let $\hat{\theta}$ be a consistent solution of the likelihood equation so that

$$P\{|\hat{\theta} - \theta_0| < \gamma| \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

By the mean value theorem

$$\frac{1}{n} \frac{\partial^2 \log L}{\partial \theta^2} - \frac{1}{n} \frac{\partial^2 \log L}{\partial \theta_0^2} = \frac{1}{n} \left( \frac{\partial^2 \log L}{\partial \theta_3^2} \right)(\hat{\theta} - \theta_0)$$

Hence the modulus of the left-hand side is less than $K_1$ with probability tending to unity, which means that $(1/n)(\partial^2 \log L/\partial \theta^2)$ converges in probability to $(1/n)(\partial^2 \log L/\partial \theta_0^2)$ which tends to $-i(\theta_0)$, where $i(\theta_0)$ is the information per single observation. Therefore, for any arbitrarily
small $\epsilon'$ we have

$$P\left( \frac{1}{n} \sum \frac{\partial^2 \log L}{\partial \theta^2} < -i(\theta_0) + \epsilon' \right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

The quantity $-i(\theta_0)$ is fixed and negative and for small $\epsilon'$, $-i(\theta_0) + \epsilon'$ is also negative. Therefore

$$P\left( \frac{1}{n} \sum \frac{\partial^2 \log L}{\partial \theta^2} < 0 \right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

i.e.,

$$P\left( \frac{\partial^2 \log L}{\partial \theta^2} < 0 \right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

This shows that the probability that the likelihood is a maximum at $\hat{\theta}$ approaches certainty as $n$ tends to $\infty$, thus establishing theorem 2.

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two consistent solutions of the likelihood equation, it follows from Rolle's theorem that $\frac{\partial^2 \log L}{\partial \theta^2} = 0$ has at least one solution $\hat{\theta}_3$ lying between $\hat{\theta}_1$ and $\hat{\theta}_2$. Since $\hat{\theta}_1$ and $\hat{\theta}_2$ converge to $\theta_0$ in probability, $\hat{\theta}_3$ also does. Therefore

$$P\left( \frac{\partial^2 \log L}{\partial \theta^2} < 0 \right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

which is a contradiction because $\frac{\partial^2 \log L}{\partial \theta^2} = 0$. We thus obtain the following corollary.

**Corollary.** A consistent solution of the likelihood equation is unique.

Wald (1949) has recently proved that the solution of the likelihood equation which makes the likelihood function an absolute maximum is necessarily consistent. This is more powerful than theorem 2 which says that the likelihood has a relative maximum at the unique consistent solution of the likelihood equation. This proof is not given here.

Let the consistent solution of the likelihood equation be denoted by $\hat{\theta}$. We have

$$\sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{I(\theta_0)}} \sum \frac{\partial \log f_i}{\partial \theta} \delta_n$$

The denominator of the right-hand fraction converges to 1 in probability. By the Lindeberg-Levy theorem the sum $\sum \delta_n(\partial \log f_i/\partial \theta)$ is asymptotically normal with zero mean and variance $I(\theta)$. Therefore the
IMPROVED ESTIMATES FROM INCOMPLETE DATA

numerator is asymptotically normal with zero mean and unit variance. Finally, it follows from the convergence theorem 1 in the appendix that the ratio \( \sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) \) is asymptotically normal with zero mean and unit variance. This means that \( \theta \) is distributed about \( \theta_0 \) with the variance \( 1/I(\theta_0) \). Since this is the minimum possible variance, we have the following result, when \( E(\theta) \) and \( \sqrt{V(\theta)} \) exist.

**Corollary.** The consistent solution of the maximum likelihood equation is fully efficient.

4c Some Examples of Maximum Likelihood Estimates

4c.1 Improved Estimates of Means from Incomplete Data on Several Variables

Suppose that each individual of a population is characterized by two measurements but in a sample only one measurement is recorded in some cases. Thus, out of a total of \( N = (n_1 + n_2 + n) \) individuals observed, \( n_1 \) of them may provide the first measurement alone, \( n_2 \) the second alone, and \( n \) both the measurements. The characteristics of, say, the first measurement, such as the mean, scatter, etc., can be estimated from the available set of \( n_1 \) observations on the first measurement alone. If the two measurements are correlated, it is possible that the observations on the second measurement may throw some more information on the characteristics of the first. If this is so, the estimates of the characteristics of any particular measurement obtained by taking all the data into account will be more accurate than those obtained from the available set of observations on that particular measurement alone.

Assuming normal distributions for the measurements, the probability density of the observations can be written

\[
\text{const.} \exp \left\{ \frac{1}{2} \left[ \frac{1}{\sigma_1^2} \left( x - \mu_1 \right)^2 + \frac{1}{\sigma_2^2} \left( y - \mu_2 \right)^2 - 2 \rho \left( x - \mu_1 \right) \left( y - \mu_2 \right) \right] \right\}
\]

where \( x \) and \( y \) are the two measurements, \( \mu_1 \) and \( \mu_2 \) their mean values, \( \sigma_1 \) and \( \sigma_2 \) the standard deviations and the correlation coefficient between \( x \) and \( y \), \( \Sigma_1 \) the summation over the \( n_1 \) observations, \( \Sigma_2 \) the summation over the \( n_2 \) observations, and \( \Sigma \) the summation over the common set of \( n \) observations.
Equating the derivatives of the logarithm of the likelihood with respect to \( \theta_1 \) and \( \theta_2 \) to zero,

\[
S_1 = \frac{\Sigma_1(x - \mu_1)}{\sigma_1^2} + \frac{1}{1 - \rho^2}\left(\frac{\Sigma(x - \mu_1) - \rho \Sigma(y - \mu_2)}{\sigma_1 \sigma_2}\right) = 0
\]

\[
S_2 = \frac{\Sigma_2(y - \mu_2)}{\sigma_2^2} + \frac{1}{1 - \rho^2}\left(\frac{\Sigma(y - \mu_2) - \rho \Sigma(x - \mu_1)}{\sigma_1 \sigma_2}\right) = 0
\]

or

\[
\mu_1 \left[\frac{n_1 + n}{\sigma_1 + \sigma_1(1 - \rho^2)}\right] - \frac{\rho}{1 - \rho^2} \mu_2 = \frac{\Sigma_1 x}{\sigma_1} + \frac{1}{1 - \rho^2}\left(\frac{\Sigma x - \rho \Sigma y}{\sigma_1 \sigma_2}\right)
\]

\[
\mu_2 \left[\frac{n_2 + n}{\sigma_2 + \sigma_2(1 - \rho^2)}\right] - \frac{\rho}{1 - \rho^2} \mu_1 = \frac{\Sigma_2 y}{\sigma_2} + \frac{1}{1 - \rho^2}\left(\frac{\Sigma y - \rho \Sigma x}{\sigma_1 \sigma_2}\right)
\]

These are simultaneous equations giving the estimates of \( \mu_1 \) and \( \mu_2 \) when \( \sigma_1, \sigma_2, \) and \( \rho \) are known. The equations giving the estimates of \( \sigma_1, \sigma_2, \) and \( \rho \) are complicated, but the following estimates may be safely used when the sample sizes are moderately large.

\[
\sigma_1^2 = \frac{\Sigma x^2 + \Sigma y^2 - (\Sigma_1 x + \Sigma_2 y)^2}{n_1 + n} \div (n_1 + n - 1)
\]

\[
\sigma_2^2 = \frac{\Sigma x^2 + \Sigma y^2 - (\Sigma_1 x + \Sigma_2 y)^2}{n_2 + n} \div (n_2 + n - 1)
\]

\[
\rho = \frac{\Sigma xy - (\Sigma x)(\Sigma y)}{n} + \sqrt{\frac{\Sigma x^2 - (\Sigma x)^2}{n} \cdot \frac{\Sigma y^2 - (\Sigma y)^2}{n}}
\]

To obtain the measures of accuracy of the estimates of \( \mu_1 \) and \( \mu_2 \) the information matrix is derived.

\[
(I_{ii}) = \begin{bmatrix} E(S_1^2) & E(S_1 S_2) \\ E(S_2 S_1) & E(S_2^2) \end{bmatrix}
\]

\[
= \begin{bmatrix} \frac{n_1}{\sigma_1^2} + \frac{n}{\sigma_1^2(1 - \rho^2)} & -\frac{n \rho}{\sigma_1 \sigma_2(1 - \rho^2)} \\ -\frac{n \rho}{\sigma_1 \sigma_2(1 - \rho^2)} & \frac{n_2}{\sigma_2^2} + \frac{n}{\sigma_2^2(1 - \rho^2)} \end{bmatrix}
\]

\[
(I_{ii}) = \begin{bmatrix} \Delta & \Delta \\ \Delta & \Delta \end{bmatrix}
\]

\[
\Delta = \frac{n_1}{\sigma_1^2} + \frac{n}{\sigma_1^2(1 - \rho^2)} + \frac{n_2}{\sigma_2^2} + \frac{n}{\sigma_2^2(1 - \rho^2)}
\]

\[
\Delta = \frac{n_1(1 - \rho^2) + n_2}{\sigma_1^2} + \frac{n_1(1 - \rho^2) + n_2}{\sigma_2^2}
\]
where $\Delta = (n + n_1)(n + n_2) - \rho^2 n_1 n_2$. If $m_1$ and $m_2$ are the estimates of $\mu_1$ and $\mu_2$, then

$$V(m_1) = \frac{n_2(1 - \rho^2) + n}{(n + n_1)(n + n_2) - \rho^2 n_1 n_2} \sigma_1^2$$

$$V(m_2) = \frac{n_1(1 - \rho^2) + n}{(n + n_1)(n + n_2) - \rho^2 n_1 n_2} \sigma_2^2$$

The estimate of $\mu_1$ obtained from the mean of $(n + n_1)$ observations on the first measurement has the variance $\sigma_1^2/(n + n_1)$ so that the efficiency of this estimate is

$$\frac{n + n_1}{(n + n_1)(n + n_2) - \rho^2 n_1 n_2} \left\{ \frac{1 - \rho^2}{n + n_2} \right\} + \left\{ \frac{1 - \frac{n_1 n_2}{(n + n_1)(n + n_2)} \rho^2}{n + n_2} \right\}$$

Of 188 skeletons of Anglo-Saxons (Münter, 1936) 103 provided maximum lengths of both the right and left femora, 48 of the right femur alone, and 37 of the left. The 151 observations on the right femur gave the estimates

Mean = 463.3  $s_1 = 22.4$ *  

and the 140 observations on the left femur gave the estimates

Mean = 465.7  $s_2 = 24.4$  

The 103 pairs of observations gave the estimated correlation $r = 0.9835$. The equations for the combined estimates $m_1$ and $m_2$ of $\mu_1$ and $\mu_2$ are

$$m_1 \left( \frac{n_1}{s_1^2} + \frac{n}{s_1^2(1 - r^2)} \right) - m_2 \frac{r}{1 - r^2} \frac{1}{s_2} = \frac{\Sigma x}{s_1} + \frac{1}{s_2} \left( \frac{\Sigma x - \Sigma y}{r} \right)$$

$$m_1 \frac{r}{1 - r^2} \frac{n}{s_1} + m_2 \left( \frac{n_2}{s_2} + \frac{n}{s_2(1 - r^2)} \right) = \frac{\Sigma x}{s_2} + \frac{1}{s_2} \left( \frac{\Sigma y - \Sigma x}{1 - r^2} \right)$$

In this example

$$\frac{n_1}{s_1} = \frac{48}{22.4} = 2.1428$$

$$\frac{n_2}{s_2} = \frac{37}{24.4} = 1.5164$$

$$\frac{n}{s_1(1 - r^2)} = \frac{22.4}{103} \frac{1}{1 - 0.9835^2} = 140.4989$$

$$\frac{n}{s_2(1 - r^2)} = \frac{24.4}{103} \frac{1}{1 - 0.9835^2} = 128.0826$$

*The estimates of $s_1$, $s_2$, and $r$ are represented by $s_1$, $s_2$, and $r$, respectively.
ESTIMATION AND MAXIMUM LIKELIHOOD

\[ \Sigma x = 21851.4 \]  
\[ \Sigma y = 16999.3 \]  
\[ \Sigma z = 48096.9 \]  
\[ \Sigma w = 48198.7 \]

(The sum of observations on 48 maximum lengths of right femora)

(The sum of observations on 37 maximum lengths of left femora)

(The sum of 103 observations on the right femur from the common set)

(The sum of 103 observations on the left femur from the common set)

With these values the equations can be written

\[ 142.6418m_1 - 126.5544m_2 = 7221.5778 \]
\[ -138.1807m_1 + 130.4990m_2 = -3470.9599 \]

The solutions are

\[ m_1 = 462.43 \quad m_2 = 463.06 \]

The variance of \( m_1 \) is

\[ \frac{n_1(1 - r^2) + n}{(n + n_1)(n + n_2) - r^2n_1n_2} s_1^2 = \frac{104.21093}{19422.1245} s_1^2 = 0.0053655s_1^2 \]

The standard error is

\[ \sqrt{0.0053655s_1} = 0.07324 \times 22.4 = 1.64 \]

The standard error of the average of 151 observations is

\[ \frac{22.4}{\sqrt{151}} = 1.79 \]

which is greater than the standard error of the maximum likelihood estimate \( m_1 \), which is as efficient as the mean of about 1/0.0053655 = 186 observations. Similarly, the standard error of \( m_2 \) is 1.82 and is as efficient as a mean based on about 185 observations. There is, however, some loss of efficiency due to errors in the estimates of \( s_1, s_2, \) and \( r \), but this is very small.

This technique can be employed in many situations. For instance, the maximum length including spine cannot be obtained for all femora since the spine is usually broken. If the maximum lengths including and excluding spine are available for some femora, and only excluding spine for others, the best estimates of the means of both the measurements can be obtained by following the above procedure. If there are no observations for the maximum length including spine alone, the quantity \( n_1 \) is equated to zero in the above equations. It may happen
as in 3f.2 that the skeletons providing measurements on both the right and left femora may be undersized. Using this portion of the material two prediction formulae can be derived, one for predicting the length of the left femur given that of the right, and another for predicting the length of the right femur given that of the left. With the help of the first prediction formula the average length of the left femur can be estimated for those skeletons providing measurement on the right femur only. Let this be \( \bar{l}_1 \) based on \( n_1 \) right femur measurements. The direct averages of the length of the left femur for the \( n \) skeletons providing both the measurements and \( n_2 \) skeletons only the left femur length are denoted by \( l \) and \( l_2 \). If all three types of material are random samples from the original skeletal population, then the three estimates \( \bar{l}_1, l, l_2 \) should agree, in which case the estimate obtained by the method of maximum likelihood is the best. If they do not agree, then the estimate of the mean left femur length of the skeletal population may be obtained as

\[
\frac{n_1 \bar{l}_1 + n \bar{l} + n_2 l_2}{n_1 + n + n_2}
\]

Similarly the mean length of the right femur can be estimated.

4c.2 The Method of Scoring for the Estimation of Parameters

The maximum likelihood equations are usually complicated so that the solutions cannot be obtained directly. A general method in such cases is to assume a trial solution and derive linear equations for small additive corrections. The process can be repeated till the corrections become negligible. A great mechanization is introduced by adopting the method known as the scoring system for obtaining the linear relations connecting the additive corrections.

The quantity \( \frac{d \log L}{d \theta} \), where \( L \) is the likelihood of the parameter \( \theta \), is defined as the efficient score for \( \theta \). The maximum likelihood estimate of \( \theta \) is that value of \( \theta \) for which the efficient score vanishes. If \( \theta_0 \) is the trial value of the estimate, then expanding \( \frac{d \log L}{d \theta} \) and retaining only the first power of \( \theta - \theta_0 \),

\[
\frac{d \log L}{d \theta} = \frac{d \log L}{d \theta_0} + \theta - \theta_0 \left( \frac{d^2 \log L}{d \theta^2} \right)
\]

where \( I(\theta_0) \), the information at the value \( \theta = \theta_0 \) is the expected value of \( \frac{d^2 \log L}{d \theta^2} \). In large samples the difference between \( -I(\theta_0) \) and \( \frac{d^2 \log L}{d \theta_0^2} \) will be of \( O(1/n) \), where \( n \) is the number of observations,
so that the above approximation holds to the first order of small quantities. The correction \( \delta \) is obtained from the equation

\[
\delta \theta = \frac{d \log L}{d \theta} \quad \delta \theta = \frac{d \log L}{d \theta} - I(\theta)
\]

The first approximation is \( (\theta_0 + \delta \theta) \), and the above process can be repeated with this as the new trial value.

**Example 1.** Consider a sample of size \( n \) from the Cauchy distribution.

\[
\frac{1}{1 + x^2}
\]

The likelihood equation is

\[
\frac{d \log L}{d \theta} = \frac{2(x - \theta)}{1 + (x - \theta)^2} = 0
\]

The efficient score for any \( \theta \) is

\[
S(\theta) = \frac{2(x - \theta)}{1 + (x - \theta)^2}
\]

Information for a single observation is \( \frac{1}{2} \) so that the asymptotic variance is \( \frac{2}{n} \) and the additive correction to a trial value \( \theta_0 \) is

\[
\frac{2S(\theta_0)}{n}
\]

This process can be continued until a stable value is attained. Fortunately in the above example \( I(\theta) \) happened to be independent of \( \theta \) so that \( I(\theta) \) need not be calculated at each trial value.

**Example 2.** Score and information for grouped data.

Let \( \pi_1, \pi_2, \ldots, \pi_k \) (\( \sum \pi_i = 1 \)) be the probabilities in \( k \) mutually exclusive classes, and suppose that \( \pi_i = \phi_i(\theta) \) so that all the proportions \( \pi_i \) are defined as functions of a single parameter \( \theta \). If \( f_1, f_2, \ldots, f_k \) are the observed frequencies, then

\[
\log L = f_1 \log \pi_1 + \cdots + f_k \log \pi_k
\]

The score at \( \theta \) is

\[
\frac{\partial \log L}{\partial \theta} = \frac{f_1 \pi_1}{\pi_1} + \cdots + \frac{f_k \pi_k}{\pi_k}
\]
Information is the variance of $\partial \log L / \partial \theta$ which is a linear function of the frequencies. Hence by (2a.9.1)

$$I(\theta) = f \sum_{i=1}^{k} \frac{1}{s_i} \left( \frac{\partial s_i}{\partial \theta} \right)^2$$

$$f = f_1 + \cdots + f_k$$

The quantities

$$\frac{1}{s_i} \frac{\partial s_i}{\partial \theta} \quad \text{and} \quad \frac{1}{s_i} \left( \frac{\partial s_i}{\partial \theta} \right)^2$$

which may be called the score and information supplied by the ith class, are to be derived in any particular problem before proceeding with the problem of estimation.

If two factors are linked with a recombination fraction $p$, then the intercrosses $AB/ab \times AB/ab$ (coupling) and $Ab/ab \times Ab/ab$ (repulsion) give rise to the following expected proportions and information as given by Mather (1938).

<table>
<thead>
<tr>
<th>Class</th>
<th>Coupling</th>
<th>Repulsion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability</td>
<td>Score</td>
</tr>
<tr>
<td>$AB$</td>
<td>$3 - 2p + p^2$</td>
<td>$1 \frac{d\sigma}{d\theta}$</td>
</tr>
<tr>
<td></td>
<td>$3 - 2p + p^2$</td>
<td>$-2(1 - p)$</td>
</tr>
<tr>
<td>$Ab$</td>
<td>$2p - p^2$</td>
<td>$2(1 - p)$</td>
</tr>
<tr>
<td>$ab$</td>
<td>$2p - p^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$1 - 2p + p^2$</td>
<td>$2(1 - p)$</td>
</tr>
</tbody>
</table>

Information: $\frac{2(3 - 4p + 2p^2)}{p(2 - p)(3 - 2p + p^2)}$ $\frac{2(1 + 2p^2)}{(2 + p^2)(1 - p^2)}$

The amount of information can be used to judge the relative efficiency of one type of cross with respect to the other for the estimation of the recombination fraction. For instance, if $p = \frac{1}{4}$, the amounts of information for coupling and repulsion are 3.7909 and 1.1636, respectively. This means that, using intercrosses with repulsion, the number of off-
springs needed will be three times as great as that for coupling to estimate the recombination fraction with the same precision.

Consider coupling data with values for $AB$, $Ab$, $aB$, and $ab$ as shown below. With the trial value $p = 0.21$, the score and information are obtained.

<table>
<thead>
<tr>
<th></th>
<th>$4x$</th>
<th>$\frac{4dx}{dp}$</th>
<th>$\frac{1dx}{dp}$</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AB$</td>
<td>2.6241</td>
<td>-1.58</td>
<td>-0.60211</td>
<td>125</td>
</tr>
<tr>
<td>$Ab$</td>
<td>0.3759</td>
<td>1.58</td>
<td>4.20325</td>
<td>18</td>
</tr>
<tr>
<td>$aB$</td>
<td>0.3759</td>
<td>1.58</td>
<td>4.20325</td>
<td>30</td>
</tr>
<tr>
<td>$ab$</td>
<td>0.6241</td>
<td>-1.58</td>
<td>-2.53164</td>
<td>34</td>
</tr>
</tbody>
</table>

Absolute sum 11.54025 197

Information per observation $= \frac{1.58}{4} (11.54025) = 4.55840$

Efficient score $= -0.60211(125) + 4.20325(18 + 20)$

$= -2.53164(34)$

$= -1.61601$

Correction term $= \frac{-1.61601}{197(4.55840)} = -0.0018$

Second approximation $= 0.21 - 0.0018 = 0.2082$

The correction is small so that the process may not be repeated. The variance of the estimate is given by

$$\frac{1}{197(4.55840)} = 0.00111357$$

A better estimate of the variance is the reciprocal of the information at the value $p = 0.2082$

The scores for trial values from 1 to 50% are given in Table XIV of Fisher and Yates (1948). They can be directly used by retaining two decimal places at each stage of approximation, and finally when two places are stabilized a complete calculation with more places may be carried out.

**Example 3. Scoring for several parameters.**

The method of scoring developed in example 2 can be extended to the case of the simultaneous estimation of several parameters. If $\theta_1$,
SCORING FOR ESTIMATION OF PARAMETERS

\( \theta_1, \cdots, \theta_q \) are the parameters, the \( i \)th efficient score is defined by

\[
S_i = \frac{\partial \log L}{\partial \theta_i} \quad i = 1, 2, \cdots
\]

where \( L \) is the likelihood of the parameters, and the information matrix is defined by \( (I_{ij}) \) where

\[
I_{ij} = E(S_i S_j)
\]

If the values of the efficient scores and information at the trial values \( \theta_1^0, \cdots, \theta_q^0 \) are indicated with index zero, then small additive corrections to the trial values are given by the simultaneous equations

\[
\begin{align*}
I_{11}^0 \delta \theta_1 + I_{12}^0 \delta \theta_2 + \cdots + I_{1q}^0 \delta \theta_q &= S_1^0 \\
I_{21}^0 \delta \theta_1 + I_{22}^0 \delta \theta_2 + \cdots + I_{2q}^0 \delta \theta_q &= S_2^0 \\
&\vdots \\
I_{q1}^0 \delta \theta_1 + I_{q2}^0 \delta \theta_2 + \cdots + I_{qq}^0 \delta \theta_q &= S_q^0
\end{align*}
\]

This operation is repeated with corrected values each time until stable values of \( \theta_1, \cdots, \theta_q \) are obtained. The variance of the final estimate \( \tilde{\theta}_i \) of \( \theta_i \) is given by \( I_{ii} \), the co-factor of \( I_{ii} \) in the determinant \( |I_{ij}| \).

In the case of grouped distributions with \( x_i \) and \( f_i \) as probability and frequency in the \( i \)th class,

\[
S_i = \sum f_i \frac{\partial x_i}{\partial \theta_i}
\]

and

\[
I_{ii} = \sum \frac{1}{x_i} \frac{\partial x_i}{\partial \theta_i} \frac{\partial x_i}{\partial \theta_i}
\]

so that the calculations become simple as illustrated below.

**Blood Groups, ABO System.** Every human being can be classified into one of the four blood groups \( O, A, B, AB \). The inheritance of these blood groups is controlled by three allelomorphic genes—\( O, A, B \)—of which \( O \) is recessive to \( A \) and \( B \). If \( r, p, \) and \( q \) are gene frequencies of \( O, A, \) and \( B \), then the expected probabilities of the six genotypes (four phenotypes) in random mating will be

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genotype</th>
<th>Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O )</td>
<td>( OO )</td>
<td>( r^2 )</td>
</tr>
<tr>
<td>( A )</td>
<td>( AA )</td>
<td>( p^2 )</td>
</tr>
<tr>
<td></td>
<td>( AO )</td>
<td>( 2pr )</td>
</tr>
<tr>
<td>( B )</td>
<td>( BB )</td>
<td>( q^2 )</td>
</tr>
<tr>
<td></td>
<td>( BO )</td>
<td>( 2qr )</td>
</tr>
<tr>
<td>( AB )</td>
<td>( AB )</td>
<td>( 2pq )</td>
</tr>
</tbody>
</table>
ESTIMATION AND MAXIMUM LIKELIHOOD

If \( O, A, B \) and \( AB \) are the observed frequencies adding to \( N \), the problem is to estimate the gene frequencies \( p, q, \) and \( r \). A rough estimate is supplied by

\[
\begin{align*}
\hat{r} &= \sqrt{\frac{O}{N}} \\
\hat{p} &= 1 - \sqrt{\frac{O + B}{N}} \\
\hat{q} &= 1 - \sqrt{\frac{O + A}{N}}
\end{align*}
\]

These may not necessarily add to unity, whereas the true values should.

Let \( D \) denote the deviation

\[
-D = p' + q' + r' - 1
\]

Better estimates due to Bernstein are obtained as follows.

\[
\begin{align*}
\hat{r} &= (1 + \frac{1}{2}D)(r' + \frac{1}{2}D) \\
\hat{p} &= (1 + \frac{1}{2}D)p' \\
\hat{q} &= (1 + \frac{1}{2}D)q'
\end{align*}
\]

where \( p', q', r', \) and \( D \) are as defined above.

There is still some deviation, \((1 - p - q - r) = \frac{1}{4}D^2\). If this is small, then Bernstein's method supplies fairly good estimates. We shall now show how these estimates can be improved by the method of maximum likelihood, using the frequencies \( O = 176, A = 182, B = 60, \) and \( AB = 17. \) Approximate solutions obtained by Bernstein's method are

\[
\begin{align*}
p &= 0.26449 \\
q &= 0.09317 \\
r &= 0.64234
\end{align*}
\]

The probabilities and derivatives, with respect to the independent parameters \( p, q \) in the general case, are

<table>
<thead>
<tr>
<th>Derivatives</th>
<th>Probability</th>
<th>( \frac{\partial r}{\partial p} )</th>
<th>( \frac{\partial r}{\partial q} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O )</td>
<td>( r^2 )</td>
<td>-2r</td>
<td>-2r</td>
</tr>
<tr>
<td>( A )</td>
<td>( p(p + 2r) )</td>
<td>2r</td>
<td>-2p</td>
</tr>
<tr>
<td>( B )</td>
<td>( q(q + 2r) )</td>
<td>-2q</td>
<td>2r</td>
</tr>
<tr>
<td>( AB )</td>
<td>( 2pq )</td>
<td>( 2q )</td>
<td>( 2p )</td>
</tr>
</tbody>
</table>
The probabilities and coefficients for the calculation of efficient scores at the approximate values obtained above are set out below.

<table>
<thead>
<tr>
<th>Coefficients for Scores</th>
<th>Probability</th>
<th>$\frac{1}{\theta}$</th>
<th>$\frac{1}{\theta}$</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>0.41260</td>
<td>-3.11362</td>
<td>-3.11362</td>
<td>176</td>
</tr>
<tr>
<td>A</td>
<td>0.40974</td>
<td>3.13543</td>
<td>-1.27104</td>
<td>182</td>
</tr>
<tr>
<td>B</td>
<td>0.12838</td>
<td>-1.45217</td>
<td>10.00685</td>
<td>60</td>
</tr>
<tr>
<td>AB</td>
<td>0.04928</td>
<td>3.75086</td>
<td>10.73007</td>
<td>17</td>
</tr>
</tbody>
</table>

The scores are

\[ \phi_p = (-3.11362)176 + (3.13543)182 + (-1.45217)60 + (3.75086)17 = -0.20444 \]

\[ \phi_q = (-3.11362)176 + (-1.27104)182 + (10.00685)60 + (10.73007)17 = -0.09321 \]

The information matrix \( \star \) for a single observation is

\[ I_{pp} = 9.00315 \quad I_{pq} = 2.47676 \quad I_{qq} = 23.21612 \]

Small corrections \( \delta_p \) and \( \delta_q \) to \( p \) and \( q \) are given by

\[ N(9.00315 \delta_p + 2.47676 \delta_q) = -0.20444 \]

\[ N(2.47676 \delta_p + 23.21612 \delta_q) = -0.09321 \]

The inverse of the information matrix per single observation is

\[ I^{pp} = 0.114430 \quad I^{pq} = -0.012208 \quad I^{qq} = 0.044376 \]

The solutions are

\[ \delta_p = \frac{I^{pq} \phi_p + I^{qq} \phi_q}{N} = -0.00005116 \]

\[ \delta_q = \frac{I^{qp} \phi_p + I^{qq} \phi_q}{N} = -0.00000377 \]

The corrections are hardly necessary in this particular case. If the corrections are not small, the whole process has to be repeated with the
second approximations. It is important to note that after some stage the information matrix need not be recalculated for each approximation. Only the new scores have to be calculated at each stage and used in conjunction with the same inverse matrix of information (kept constant from some stage) to obtain closer approximations. When convergence is achieved, the information matrix and scores may be calculated for the last approximate values and the final approximation obtained.

The maximum likelihood estimates and the variances are

\[
\begin{align*}
    p &= 0.26444 \\
    q &= 0.09317 \\
    r &= 0.64239 \\
    \frac{\text{IPP}}{N} &= \frac{V(p)}{N} = 0.00026305 \\
    \frac{\text{IPP}}{N} &= \frac{V(q)}{N} = 0.00010202 \\
    \frac{\text{IPP}}{N} &= \frac{V(r)}{N} = 0.00030893
\end{align*}
\]

4c.3 Combination of Data

The advantage of the scoring system can be best seen in the mechanism it introduces when various sets of data giving information on some parameters have to be combined for estimation. If \( L \) is the joint likelihood based on all the data and \( L_i \) for the \( i \)th part, then

\[
L = L_1 L_2 \cdots
\]

\[
\frac{\partial \log L}{\partial \theta} = \frac{\partial \log L_1}{\partial \theta} + \frac{\partial \log L_2}{\partial \theta} + \cdots
\]

which shows that the efficient scores are additive. Also, if \( I_{\alpha} \) is an element of the information matrix for the whole body of data and \( I_{\alpha j} \) for the \( j \)th part, then

\[
I_{\alpha} = I_{\alpha 1} + I_{\alpha 2} + \cdots
\]

Thus, to obtain the best estimates it is necessary to replace each part of the data by the scores and information matrix at a trial value and obtain the total scores and information matrix by simple addition. The correction to trial values can be obtained by solving simultaneous equations as shown in the previous section.

Appendix: Some Limiting Theorems

A general convergence theorem (Cramér, 1946). Let \( \xi_1, \xi_2, \cdots \) be a sequence of random variables with distribution functions \( F_1, F_2, \cdots \).
Suppose that $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$. Let $\eta_1, \eta_2, \cdots$ be another sequence of random variables, and suppose that $\eta_n$ converges in probability to a constant $c$. If

$$X_n = \xi_n + \eta_n \quad Y_n = \xi_n \eta_n \quad Z_n = \frac{\xi_n}{\eta_n}$$

then the distribution functions

- of $X_n \rightarrow F(x - c)$
- of $Y_n \rightarrow F\left(\frac{x}{c}\right)$ if $c > 0$
- of $Z_n \rightarrow F(cx)$ if $c > 0$

The theorem covers the case of $c < 0$ also, in which case the variables $-\eta_1, -\eta_2, \cdots$ would be considered.

The theorem is proved for $Z_n$, the proof being similar for the rest. The set $S$ of points satisfying $\xi_n/\eta_n \leq x$ consists of two non-overlapping sets:

- $S_1$: $\frac{\xi_n}{\eta_n} \leq x \quad |\eta_n - c| \leq \epsilon$
- $S_2$: $\frac{\xi_n}{\eta_n} \leq x \quad |\eta_n - c| > \epsilon$

Thus

$$P_n(S) = P_n(S_1) + P_n(S_2)$$

For every $\epsilon$

$$P_n(S_2) < P_n(|\eta_n - c| > \epsilon) \rightarrow 0$$

by hypothesis, in which case $P_n(S)$ lies within the limits

$$P_n[\xi_n \leq (c \pm \epsilon)x \quad |\eta_n - c| \leq \epsilon]$$

This differs from the corresponding quantity

$$P_n[\xi_n \leq (c \pm \epsilon)x] = P_n[(c \pm \epsilon)x]$$

by less than $P_n(|\eta_n - c| > \epsilon)$. As $n \rightarrow \infty$, $P_n(|\eta_n - c| > \epsilon) \rightarrow 0$. $P_n(S)$ is thus enclosed between two limits which can be made to lie as close as possible to $F(cx)$ by choosing $\epsilon$ small. The theorem is thus proved. It may be noted that no condition of independence of the variables involved is assumed.

Slutsky’s theorem (1925). If $\xi_n, \eta_n, \cdots, \rho_n$ are random variables converging in probability to constants $x, y_1, \cdots, r$, respectively, any rational function $R(\xi_n, \eta_n, \cdots, \rho_n)$ converges in probability to the constant
ESTIMATION AND MAXIMUM LIKELIHOOD

$R(x, y, \cdots, r)$, provided that the latter is finite. It follows that any power $R^k(\xi_n, \eta_n, \cdots, \rho_n)$ with $k > 0$ converges in probability to $R^k(x, y, \cdots, r)$.

Khintchine's theorem. Let $\xi_1, \xi_2, \cdots$ be independent random variables, all having the same distribution function $F(x)$, and suppose that $F(x)$ has a finite mean $m$. Then the variable $\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$ converges in probability to $m$.

Levy's theorem (1925). A necessary and sufficient condition for the convergence of the sequence $\{F_n(x)\}$ of distribution functions to a distribution function $F(x)$ is that, for every $t$, the sequence $\{\phi_n(t)\}$ of characteristic functions converges to the limit $\phi(t)$ which is continuous for the special value $t$ and is the characteristic function for $F(x)$.

Central limit theorems. (1) Lindberg and Levy (1922, 1925). If $\xi_1, \xi_2, \cdots$ are independent random variables, all having the same probability distribution, and if $m$ and $\sigma$ denote the mean and standard deviation of every $\xi_n$, then the sum $\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$ is asymptotically normally distributed with mean $nm$ and standard deviation $\sigma \sqrt{n}$.

(2) Liapounoff (1901). Let $\xi_1, \xi_2, \cdots$ be independent random variables with means and standard deviations $m_v$ and $\sigma_v$, $(v = 1, 2, \cdots)$. Suppose that the third absolute moment of $\xi_v$ about its mean $m$ is finite for every $v$. If $\rho/\sigma \to 0$ as $n \to \infty$, where

$$\rho^3 = E(|\xi_v - m_v|^3)$$

then the sum $\sum_{v=1}^{n} \xi_v$ is asymptotically normal with mean $m = m_1 + m_2 + \cdots$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \cdots$.

References


BHATTACHARYA, A. (1946). On some analogues of the amount of information and their uses in statistical estimation. (In three parts.) Sankhya, 8, 1, 201, 315.


REFERENCES


CHAPTER 5

Large Sample Tests of Hypotheses
with Applications to Problems
of Estimation

5a The General Theory of Tests in Large Samples

5a.1 The Nature of Statistical Hypotheses

If the probability differential of a set of stochastic variables contains \( k \) unknown parameters, the statistical hypotheses concerning them may be simple or composite. The hypothesis leading to a complete specification of the values of the \( k \) parameters is a simple hypothesis, and the one leading to a collection of admissible sets a composite hypothesis. In this chapter are discussed tests of these two types of hypotheses on the basis of a large number of observations from any probability distribution satisfying some mild restrictions and also their use in problems of estimation.

5a.2 The Problem of Distribution

There are two problems of distribution that are useful in deriving tests of significance for simple and composite hypotheses. Let

\[ x_1, \ldots, x_p; y_1, \ldots, y_p; \ldots \]

be independent sets of observations from probability laws with densities represented by \( f_1(x \mid \theta), f_2(y \mid \theta), \ldots \) such that each function contains at least one of the unknown parameters \( \theta_1, \ldots, \theta_k \). The likelihood of the parameters which is the same as the probability density of the observed sets of data is

\[ L = f_1(x \mid \theta)f_2(y \mid \theta) \ldots \]

As defined in Chapter 4 the \( i \)th efficient score is represented by

\[ \phi_i = \frac{\partial \log L}{\partial \theta_i} \]

176
The mean values of these scores are zero. Their covariance matrix is represented by \((a_{ii})\), and its reciprocal by \((a_{ii})\). Let there exist a positive quantity \(\eta\) such that

\[
E\left(\frac{1}{f_i} \right)^2 = i = 1, \ldots, k \tag{5a.2.1}
\]

are finite. Under these conditions, if the non-vanishing terms in the sequence \(\theta \log f_i/\partial \theta_i, (i = 1, 2, \ldots)\), for any \(j\) form a sufficiently large set, it follows from an extension of the central limit theorem to many variables that the limiting distribution of \(\phi_1, \ldots, \phi_k\) at the true values \(\theta_1, \ldots, \theta_k\) tends to the multivariate normal form with zero mean and covariance matrix \((a_{ij})\).

\[
\text{const. } e^{-\frac{1}{2}Q} \, d\phi_1 \cdots d\phi_k
\]

where

\[
Q = \sum a_{ij} \phi_i \phi_j
\]

Hence \(Q\) is distributed, in large samples, as \(\chi^2\) with \(k\) degrees of freedom when the true values of the parameters are \(\theta_1, \ldots, \theta_k\).

If the probability densities \(f_1, f_2, \ldots\) are the same, it is enough for the limiting properties to hold that

\[
E(f_i) = a_{ij} \text{ is finite for every } j \text{ which is less restrictive than the condition (5a.2.1).}
\]

Suppose that the \(\theta\) parameters are subject to \(s\) restrictions defined by \(s\) independent relations.

\[
\psi_i(\theta_1, \ldots, \theta_k) = 0 \quad i = 1, 2, \ldots, s \tag{5a.2.2}
\]

The maximum likelihood estimates are given by

\[
\phi_i + \sum_j \lambda_j \frac{\partial \psi_j}{\partial \phi_i} = 0 \quad i = 1, 2, \ldots, k \tag{5a.2.3}
\]

where \(\lambda_j\) are Lagrangian multipliers. Let \(\phi_1, \ldots, \phi_k\) be the maximum likelihood estimates. Since the set of equations (5a.2.3) involve \((k-s)\) linear restrictions on \(\phi_i(\theta)\), it is expected that the statistic

\[
\chi^2 = \sum a_{ij}(\theta) \phi_i(\theta) \phi_j(\theta)
\]

is distributed as \(\chi^2\) with \(s\) degrees of freedom which is \((k-s)\) less than the degrees of freedom for true values.

This can be demonstrated if we assume that the restrictions (5a.2.2) specify \(s\) of the parameters \(\theta_{k-s+1}, \ldots, \theta_k\) (say) as functions of \((k-s)\)
free parameters $\theta_1, \cdots, \theta_{k-1}$, so that the likelihood is an explicit function of these parameters only, and further that the joint distribution of $\theta_1, \cdots, \theta_{k-1}$, tends to the multivariate normal form in large samples with variances and covariances of $O(n^{-1})$. It is known that the latter assumption is true provided that the probability laws satisfy the condition (5a.2.1), and further that the maximum likelihood estimates are uniformly consistent (Wald, 1949). This does not seem to be a necessary condition, and the approach to normality is probably true under less stringent conditions.

Let us take the case of two parameters and one restrictive condition which may be taken as $\theta_2 = w(\theta_1)$. The differential coefficient $\partial \theta_2 / \partial \theta_1$ is denoted by $\lambda(\theta_1)$. The maximum likelihood estimates satisfy the equations

$$\phi_1(\theta) + \lambda(\theta_1) \phi_2(\theta) = 0 \quad \theta_2 = w(\theta_1) \quad (5a.2.4)$$

If the given relation is true, then the statistic

$$x_0^2 = \sum \partial_0^j(\theta) \phi_0(\theta) \phi_0(\theta) \quad (5a.2.5)$$

depends only on $\theta_1$ and is distributed as $\chi^2$ with 2 degrees of freedom at the true value of $\theta_1$. The expression (5a.2.5) treated as a function of $\theta_1$ may be expanded in the neighborhood of $\theta_1$. The first term is

$$x_1^2 = \sum \partial_0^j(\theta) \phi_0(\theta) \phi_0(\theta)$$

The second term is

$$-2(\theta_1 - \theta_1)[\phi_1(\theta) \{a_{11}(\alpha_{11} + \lambda \alpha_{12}) + a_{12}(\alpha_{12} + \lambda \alpha_{22})\}$$

$$+ \phi_2(\theta) \{a_{22}(\alpha_{12} + \lambda \alpha_{22}) + a_{11}(\alpha_{11} + \lambda \alpha_{12})\}]$$

$$= -2(\theta_1 - \theta_1)\{\phi_1(\theta) + \lambda \phi_2(\theta)\} = 0 \quad (5a.2.6)$$

by virtue of (5a.2.4). In the expression (5a.2.6) terms $O(n^{-1})$ only have been retained, $\partial \phi_1 / \partial \theta_1$ being replaced by $\alpha_{11}$, and terms of the type

$$(\theta_1 - \theta_1) \frac{\partial \alpha_{12}}{\partial \theta_1} \phi_1 \phi_2$$

being omitted as they are of $O(n^{-1})$.

The third term can be easily shown to be

$$x_3^2 = (\theta_1 - \theta_1)[a_{22}(\theta) + 2\lambda \alpha_{12}(\theta) + \lambda^2 \alpha_{22}(\theta)]$$

Neglecting terms of higher order of smallness, we obtain

$$x_0^2 = x_1^2 + x_2^2$$

Since

$$\frac{1}{V(\theta_1 - \theta_1)} = a_{11}(\theta) + 2\lambda \alpha_{12}(\theta) + \lambda^2 \alpha_{22}(\theta)$$

$$x_0^2 = x_1^2 + x_2^2$$
it follows that $\chi^2_2$ is distributed in large samples as $\chi^2$ with 1 degree of freedom.

It can be demonstrated by expanding $\phi(\theta)$ in powers of $(\theta_i - \hat{\theta}_i)$ that $(\theta_i - \hat{\theta}_i)$ and $\phi(\theta)$ tend to be uncorrelated in large samples, so that $\chi^1$ and $\chi^2$ are independently distributed in the limiting case.

Since $\chi^2_0$ is distributed as $\chi^2$ with 2 degrees of freedom and $\chi^2_2$ with 1 degree of freedom, it follows that the residual part $\chi^2_1$ is distributed as $\chi^2$ with 1 degree of freedom.

For $s$ relations and $(k \geq s)$ parameters, $\chi^2$ can be expressed as a function of $(k - s)$ parameters and split into two portions, one of which is $\chi^2_0$ with $(k - s)$ degrees of freedom measuring the discrepancy of the $(k - s)$ estimated parameters from their true values, and another $\chi^2_1$ with $s$ degrees of freedom measuring the departures from the assigned relationships.

5b Applications of the General Theory
5b.1 The $\chi^2$ Test of Departure from a Simple Hypothesis

The problem is to test whether $n_1, \ldots, n_k, (\Sigma n_i = n)$, the frequencies in $k$ classes, are in accordance with some hypothetical proportions $\pi_1, \ldots, \pi_k, (\Sigma \pi_i = 1)$. The probability of the sample on the given hypothesis is

$$L = \frac{n!}{n_1! \cdots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k}$$

There are only $(k - 1)$ independent parameters which may be taken as $\pi_1, \ldots, \pi_{k-1}$. The efficient scores are

$$\phi_i = \frac{\partial \log L}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n}{\pi}$$

Their variances and covariances are

$$\alpha_{ii} = n \left( \frac{1}{\pi_i} + \frac{1}{\pi} \right)$$

using the formula (2a.9.1)

$$\alpha_{ij} = \frac{n}{\pi_k}$$

using the formula (2a.9.1)

$$| \alpha_{ij} | = \frac{n^{i-1}}{\pi_1 \cdots \pi_k}$$

$$\alpha_{ii} = \pi_i(1 - \pi_i)$$

$$\alpha_{ij} = -\frac{\pi_i \sigma_j}{n}$$
The $\chi^2$ statistic is

$$\sum \pi_i (1 - \pi_i) \left( \frac{n_i - n \pi_i}{n \pi_i} \right)^2 - \sum \pi_j \pi_k \left( \frac{n_j - n \pi_j}{n \pi_j} \right) \left( \frac{n_k - n \pi_k}{n \pi_k} \right)$$

where $d_i = n_i - n \pi_i$, $i = 1, \ldots, k$. The above expression reduces to

$$\frac{d_1^2}{n \pi_1} + \cdots + \frac{d_k^2}{n \pi_k} = \sum \frac{(O - E)^2}{E}$$

As shown in 5a.2 the large sample distribution of the above statistic is that of $\chi^2$ with $(k - 1)$ degrees of freedom because the test is based on $(k - 1)$ independent parameters. We shall, however, present an alternative way of deriving its distribution.

In 2a.3 it was shown that the multinomial distribution is equivalent to a product of Poisson distributions subject to the condition that the sum of the variates is $n$. If each of the individual cell frequencies is large, then the Poisson probabilities could be replaced by the normal approximation. If $X_i = (n_i - n \pi_i)/\sqrt{n \pi_i}$, then the approximate distribution of $X_1, \ldots, X_k$ is subject to the condition

$$\sqrt{n \pi_1} X_1 + \sqrt{n \pi_2} X_2 + \cdots + \sqrt{n \pi_k} X_k = (n_1 - n \pi_1) + \cdots + (n_k - n \pi_k) = \Sigma n_i - n = 0$$

Therefore by using the result in 2c.3 the distribution of $\chi^2$ is that of the sum of squares of $k$ variates $N(0, 1)$ subject to one homogeneous linear restraint, i.e., $\chi^2$ with $(k - 1)$ degrees of freedom.

If the deviations of the observed from the expected frequencies are subject to $t$ linear homogeneous restrictions on the total, then $\Sigma x^2$ is a $\chi^2$ with $(k - t)$ degrees of freedom.

**Example 1.** Bateson gives the following data concerning the segregation of two genes for purple-red flower color and long-round pollen shape in sweet pea.

The results are from an intercross so that the expected frequencies on the hypothesis of independent segregation are in the ratio 9:3:3:1. Are the data in agreement with the expected frequencies?
THE $\chi^2$ TEST OF DEPARTURE

<table>
<thead>
<tr>
<th>Purple-Long</th>
<th>Red-Long</th>
<th>Purple-Round</th>
<th>Red-Round</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>296</td>
<td>27</td>
<td>19</td>
<td>85</td>
</tr>
<tr>
<td>Expected</td>
<td>364.7817</td>
<td>9.1054</td>
<td>4.5090</td>
<td>270.7260</td>
</tr>
</tbody>
</table>

$\chi^2 = \sum \frac{(O - E)^2}{E}$

This is significant on 3 degrees of freedom, showing that there is a departure from the expected. For the test to hold, it is necessary that each expected cell frequency should be at least greater than 5. If any such frequency is small, two suitable cells can be combined to form one cell with a higher frequency.

*Example 2.* The number of deaths due to cholera is 350 out of a total of 976 due to all causes in a certain week. Is cholera on the increase if it accounted for $\frac{1}{4}$ of the deaths in the last week?

The expected number of deaths due to cholera on the basis of $\frac{1}{4}$ proportion is $976/3 = 325.3$. The value of $\chi^2$ with 1 degree of freedom is

$$\frac{(350)^2}{325.3} + \frac{(626)^2}{650.6} = 976$$

or

$$\chi^2 = \frac{[350] - 626]^2}{976} = 2.80$$

the probability of exceeding which is just less than 10%. This probability is not appropriate in answering the problem whether cholera is on the increase. $\chi^2$ gives the probability for deviations both in excess or defect of the expected, whereas only the probability of deviations in excess of the expected is relevant to this problem. To determine this we observe that

$$\sqrt{\chi^2} = 1.6733$$

can be used as a normal deviate with zero mean and unit standard deviation. The probability of a normal deviate's exceeding this value is less than 5%, which shows that cholera is on the increase.

*Example 3.* Test whether the frequencies 8, 3, 1 could have arisen from a trinomial with equal probabilities.

The expected values 4, 4, 4 are all small so that the $\chi^2$ approximation cannot be used. If we ignore this condition, the $\chi^2$ with 2 degrees of freedom is

$$\chi^2 = \frac{8^2}{4} + \frac{3^2}{4} + \frac{1^2}{4} - 12 = 6.50$$

which has a probability between 2 and 5%. In problems where the expected frequencies are small in some cells, they can be combined with other cells so as to have frequencies at least greater than 5 in each
class. Such a procedure is not possible in the present example because all the expectations are small. This necessitates the evaluation of the probability of the observed distribution 8, 3, 1 and those yielding a larger $X^2$ on the hypothesis of equal probabilities of the three classes of events.

The probability for any observed distribution $x, y, z$ is

$$\frac{12! (1)^{12}}{x!y!z!3}$$

There are 91 partitions of 12 for which the probabilities are calculated below.

<table>
<thead>
<tr>
<th>Partitional Type</th>
<th>Number</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 0 0</td>
<td>3</td>
<td>0.001882</td>
</tr>
<tr>
<td>11 1 0</td>
<td>6</td>
<td>0.002258</td>
</tr>
<tr>
<td>10 2 0</td>
<td>6</td>
<td>0.003124</td>
</tr>
<tr>
<td>10 1 1</td>
<td>3</td>
<td>0.004484</td>
</tr>
<tr>
<td>9 3 0</td>
<td>6</td>
<td>0.004740</td>
</tr>
<tr>
<td>9 2 1</td>
<td>6</td>
<td>0.004740</td>
</tr>
<tr>
<td>8 4 0</td>
<td>6</td>
<td>0.009314</td>
</tr>
<tr>
<td>8 3 1</td>
<td>6</td>
<td>0.009736</td>
</tr>
<tr>
<td>8 2 2</td>
<td>3</td>
<td>0.015589</td>
</tr>
<tr>
<td>7 5 0</td>
<td>6</td>
<td>0.024160</td>
</tr>
<tr>
<td>7 4 1</td>
<td>6</td>
<td>0.024160</td>
</tr>
<tr>
<td>7 3 2</td>
<td>6</td>
<td>0.01490</td>
</tr>
<tr>
<td>6 6 0</td>
<td>3</td>
<td>0.01783</td>
</tr>
<tr>
<td>6 5 1</td>
<td>6</td>
<td>0.02043</td>
</tr>
<tr>
<td>6 4 2</td>
<td>6</td>
<td>0.02608</td>
</tr>
<tr>
<td>6 3 3</td>
<td>3</td>
<td>0.03130</td>
</tr>
<tr>
<td>5 5 2</td>
<td>3</td>
<td>0.03130</td>
</tr>
<tr>
<td>5 4 3</td>
<td>6</td>
<td>0.05216</td>
</tr>
<tr>
<td>4 4 4</td>
<td>1</td>
<td>0.06520</td>
</tr>
</tbody>
</table>

The sum of probabilities for $X^2$'s exceeding the value 6.5 corresponding to (8, 3, 1) is 0.0484, which exceeds the probability obtained by the $X^2$ approximation. The approximation overestimated significance. Even according to the exact test, the probability being less than 5%, the hypothesis may be rejected.

Example 4. Out of 8 fossils discovered, 2 and 6 were identified as belonging to male and female. Is this compatible with the sex ratio 1:1?

The expected values are small, as in example 3, so that an exact treatment is needed. The problem is the same as that of finding the probability of 6 or more heads or tails in 8 tosses with an unbiased coin. The total chance for 6, 7, and 8 heads is

$$(^8 C_6 + ^8 C_7 + ^8 C_8)2^{-8} = \frac{37}{2^8}$$
The total for heads as well as tails is twice the above probability equal to \(37/2^7 = 0.289\), which is quite high, so that there is no definite evidence against the 1:1 sex proportion. In a general case the term-by-term evaluation may be a difficult job. It is shown in 2a.1 that the sum of the probabilities for 0, 1, \(\cdots\), \(r\) successes is given by the incomplete \(\beta\)-integral,

\[
\frac{n!}{r!(n-r-1)!} \int_0^1 x^{n-r-1}(1-x)^r \, dx
\]

where \(q = 1 - p\) is the chance of a failure. This function is tabulated in the incomplete beta tables edited by K. Pearson. In the above problem

\[
p = q = \frac{1}{2} \quad n = 8 \quad r = 2
\]

In the notation of incomplete beta tables *

\[
\begin{align*}
\text{index } p &= (n - r - 1) + 1 = 6 \\
\text{index } q &= r + 1 = 3 \\
x &= \text{probability } q = 0.5
\end{align*}
\]

The tabular entry for \(x = 0.5\), index \(p = 6\), and index \(q = 3\) is 0.1445312, which is the probability for 0, 1, and 2 heads. By symmetry the probability for tails is also the same, so that the required total is 0.2890624, which agrees with the value obtained above.

5b.2 The \(\chi^2\) Test of Goodness of Fit

The general problem in judging goodness of fit is to test whether the cell proportions can be expressed as functions of a fewer number of parameters. Thus, if \(O, A, B,\) and \(AB\) represent the four blood group classes, it may be desired to test whether the cell frequencies can be expressed in terms of gene frequencies of \(O, A,\) and \(B,\) or two independent parameters \(p\) and \(q\) (example 3 in 4c.2). Here the values of \(p\) and \(q\) are not known, but what is needed is a test of the consistency relations among the probability expressions for the four classes.

If the observed frequency distribution has arisen from a normal distribution, the probability \(\pi_i\) in the \(r\)th class bounded by \(e_i\) and

* There are three quantities to be entered into this table. Index \(p\) is equal to 1 plus the power of \(x\) in the above integral, and index \(q\) is 1 plus the power of \((1 - x)\). The entry \(x\) of the table is the upper limit of the integral. The probabilities \(p\) and \(q\) are not to be confused with indices \(p, q\) of the table.
so that the cell probabilities could be expressed in terms of two parameters \( \mu \) and \( \sigma \).

There are in general \((k - 1)\) independent proportions \( \pi_1, \ldots, \pi_{k-1} \) specifying the distribution in \( k \) classes. If these proportions can be expressed as functions of \( s \) independent parameters, then all the \((k - 1)\) proportions can be expressed as functions of \( s \) suitably chosen proportions. Thus there are \((k - 1 - s)\) restrictions on \( \pi_1, \ldots, \pi_{k-1} \). If we construct

\[
\chi^2 = \sum \sum d_i^2 \phi_i = \sum \frac{(O - E)^2}{E}
\]

over all classes and substitute for \( \pi_1, \ldots, \pi_{k-1} \) their best estimates subject to the \((k - 1 - s)\) restrictions, then the \( \chi^2 \) has \((k - 1 - s)\) degrees of freedom. This can be used to test whether the specification is correct or not.

To estimate \( \pi_1, \ldots, \pi_{k-1} \) subject to \((k - 1 - s)\) restrictions, it is enough to estimate \( s \) parameters in terms of which \( \pi_1, \ldots, \pi_{k-1} \) are completely defined. The degrees of freedom for \( \chi^2 \) is \((k - 1)\) minus \( s \), the number of parameters estimated. It is necessary for the formula of the degrees of freedom to hold that the parameters should be estimated by the most efficient method (e.g., maximum likelihood).

In example 3 of 4e.2 the estimates of the blood-group gene frequencies are found by the method of maximum likelihood. To test whether the proportions in the four blood-group classes can be expressed as functions of gene frequencies, the \( \chi^2 \) is calculated as below. Using the estimates \( p = 0.2644, q = 0.0932, \) and \( r = 0.0424 \), the expected values are derived.

\[
\begin{array}{ccc}
\text{O} & n(p^2) = 179.51 & 179.51 \\
A & n(p^2 + 2pr) = 182.18 & 182.18 \\
B & n(q^2 + 2qr) = 55.87 & 55.87 \\
AB & n(2pq) = 21.44 & 21.44 \\
\hline
\text{Total} & n & 435.00 & 435.00 \\
\hline
\text{Expected} & 435 & 435.00 & 435.00 \\
(Observed)^2/Expected & 172.56 & 185.90 & 13.48 \\
\hline
\chi^2 = 1.37 & (1 \text{ D.F.})
\end{array}
\]

The \( \chi^2 \) with \((3 - 2) = 1\) degree of freedom is not significant, thus indicating that the cell expectations could be expressed in terms of the gene frequencies.
6b.3 Tests of Homogeneity of Parallel Samples

Let the frequencies in \( k \) classes for two samples be

<table>
<thead>
<tr>
<th>Classes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>( n_1 ) ( n_2 ) ( \cdots ) ( n_k ) ( n )</td>
</tr>
<tr>
<td>Second</td>
<td>( n'_1 ) ( n'_2 ) ( \cdots ) ( n'_k ) ( n' )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_1 + n'_1 ) ( n_2 + n'_2 ) ( \cdots ) ( n_k + n'_k ) ( n + n' )</td>
</tr>
</tbody>
</table>

These classes may refer to a discrete classification or to intervals of a continuous variable. Nothing being known about the actual distributions, how can it be tested that the two samples have arisen from the same population?

If \( \pi_1, \cdots, \pi_k \) and \( \pi'_1, \cdots, \pi'_k \) are the cell proportions in populations from which the samples are drawn, then the hypothesis to be tested is \( \pi_i = \pi'_i \) \((i = 1, \cdots, k - 1)\). If \( \pi_i \) and \( \pi'_i \) are known, then the \( \chi^2 \) test of departure from the expected is

\[
\text{\( \chi^2 \)} = \sum \frac{(n_i - n\pi_i)^2}{n\pi_i} + \sum \frac{(n'_i - n'\pi'_i)^2}{n'\pi'_i}
\]

which has \((k - 1) + (k - 1)\) degrees of freedom. If \( \pi_i = \pi'_i \), there are \((k - 1)\) restrictions, and the best estimate of the common value is

\[
\hat{\pi}_i = \frac{n_i + n'_i}{n + n'}
\]

If this value is substituted in the above expression, \( \chi^2 \) reduces to

\[
\text{\( \chi^2 \)} = \frac{1}{n'n'} \sum \frac{(n(n' - n\pi'))^2}{n_i + n'_i}
\]

which has now \((k - 1)\) degrees of freedom. This tests the departure from the equality of proportions. If

\[
p_1 = \frac{n_1}{n_1 + n'_1}, \quad p_2 = \frac{n_2}{n_2 + n'_2}, \cdots
\]

and

\[
p = \frac{n}{n + n'}
\]

then the above \( \chi^2 \) could also be written

\[
\chi^2 = \sum_{i=1}^{k} \frac{n(n' - n\pi')}{p(1-p)}
\]
which is convenient to calculate if the problem needs the evaluation of the $p_i$ also.

**Example 1.** The distributions in four blood-group classes $O$, $A$, $B$, and $AB$ of 140 Christians who are army cadets and 295 other Christians are given below.

<table>
<thead>
<tr>
<th></th>
<th>$O$</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Army cadets</td>
<td>56</td>
<td>60</td>
<td>18</td>
<td>6</td>
<td>140</td>
</tr>
<tr>
<td>Others</td>
<td>120</td>
<td>122</td>
<td>42</td>
<td>11</td>
<td>295</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>176</td>
<td>182</td>
<td>60</td>
<td>17</td>
<td>435</td>
</tr>
</tbody>
</table>

$p_0 = 56(0.3182) + 60(0.3297) + 18(0.3000) + 6(0.3529) = 45.1186$

$np = 45.0520, p(1-p) = 0.24994$

$x^2 = 0.0666/0.24994 = 0.2664$

The probability that $x^2$ exceeds 0.2664 with 3 degrees of freedom is greater than 95%, showing thereby that the two samples may be considered to have arisen from the same population.

**Example 2.** The test given above is not necessarily the best and is recommended only when nothing is known about the frequency distributions. In the above example of blood groups it is known that the frequencies can be expressed in terms of two independent gene frequencies (example 3, in 4c.2). If $p$, $q$ and $p'$, $q'$ are the parameters appropriate for the two samples, then the test of agreement between the two samples reduces to that of testing the hypothesis $p = p'$ and $q = q'$. The $x^2$ test for hypothetical values of $p$, $q$ and $p'$, $q'$ has 4 degrees of freedom, whereas if they are estimated subject to the conditions $p = p'$, $q = q'$ the resulting $x^2$ has only 2 degrees of freedom. There are only two essential comparisons needed, and the $x^2$ with 2 degrees of freedom is sufficient for this purpose. But in the example worked out above, the $x^2$ test of discrepancy has 3 degrees of freedom, one more than that of the hypothesis specifying two relationships. The $x^2$ test with 2 degrees of freedom is to be preferred, and this is possible because the nature of distributions is known.

In general, if the distributions are specified by $r$ parameters, then the $x^2$ test for equality of parameters given in 5a.2 has $r$ degrees of freedom.
If the discrepancies in the $k$ classes arise owing to the parameters' being different, then the test based on a direct comparison of the estimated parameters appears to be reasonable.

For carrying out the proposed test it is necessary to estimate the gene frequencies from the totals. This is worked out in example 3 of 4c.2. The estimates of $p$, $q$ are

$$p = 0.26444 \quad q = 0.09317$$

and the inverse to the information matrix (example 3 in 4c.2) per single observation is

$$I_{pp} = 0.114430 \quad I_{pq} = -0.012208 \quad I_{qq} = 0.044376$$

The scores $\phi_1$ and $\phi_2$ for the two samples and $\chi^2$ are calculated below.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$n$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>140</td>
<td>10.37497</td>
<td>-7.27341</td>
<td>0.11704</td>
</tr>
<tr>
<td>2</td>
<td>295</td>
<td>-10.37497</td>
<td>7.27341</td>
<td>0.05554</td>
</tr>
<tr>
<td>Total</td>
<td>435</td>
<td>0</td>
<td>0</td>
<td>0.17258</td>
</tr>
</tbody>
</table>

In the last column the values of $\chi^2$ differ only in the multiplier $1/n$ so that the total $\chi^2$ can be simply obtained from the formula

$$\chi^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi_{i1}^2 + \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} \phi_{i2}^2$$

$$= \left( \frac{1}{n_1} \right) (0.114430(10.37497)^2 + 2(-0.012208)(10.37497)(-7.27341) + 0.044376(7.27341)^2)$$

$$= 0.17258$$

The $\chi^2$ with 2 degrees of freedom is small so that the data do not provide any evidence for differences in gene frequencies. In fact, the probability of exceeding the observed value with 2 degrees of freedom is just over 90%, which is smaller than the corresponding probability of example 1 above. In general, the test given in example 2 is more sensitive than the overall test of example 1. The common estimates of $p$ and $q$ are as found above.

Example 3. The tests proposed above can be extended to the general case of testing whether a number of samples come from the same popu-
lation. When nothing is known about the distributions except that they are identical the \( \chi^2 \) is based on \((k - 1)(s - 1)\) degrees of freedom where \( k \) is the number of cells and \( s \) is the number of samples. If \( n_{ij} \) denotes the frequency in the \( i \)th cell for the \( j \)th sample and \( n_i = \Sigma n_{ij} \), \( n_{..} = \Sigma n_{ij} \), then the expected value of \( n_{ij} \) is \( n_i \cdot n_{..}/n \), when the hypothesis is true. The \( \chi^2 \) on \((k - 1)(s - 1)\) degrees of freedom is

\[
\sum \frac{(n_{ij} - n_i \cdot n_{..}/n)^2}{n_i \cdot n_{..}/n}
\]

The test should, however, be modified if the nature of the distribution is known. The following example illustrates the method.

**Table 5b.3p. Distribution of Animals Bred for Linkage between Two Factors A and B**

<table>
<thead>
<tr>
<th>Sex of Heterozygotes</th>
<th>Phase</th>
<th>Sex of Animals Bred</th>
<th>Phenotype</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( AB )</td>
<td>( Ab )</td>
</tr>
<tr>
<td>( \sigma^+ \sigma^+ )</td>
<td>Coupling</td>
<td>( \sigma^+ )</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \sigma^+ )</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Repulsion</td>
<td>( \sigma^+ )</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \sigma^+ )</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>( \sigma^- \sigma^- )</td>
<td>Coupling</td>
<td>( \sigma^- )</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \sigma^- )</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Repulsion</td>
<td>( \sigma^- )</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \sigma^- )</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>131</td>
<td>110</td>
</tr>
</tbody>
</table>

First it is necessary to test whether there is sex difference within a mating type. The results are from backcrosses so that the probabilities in the four classes are \((1 - p)/2, p/2, p/2, (1 - p)/2\) for coupling and \(p/2, (1 - p)/2, (1 - p)/2, p/2\) for repulsion. The score and information for the recombination fraction \( p \) are

\[
S = -\frac{(AB) + (Ab) + (aB) + (ab)}{1 - p} \quad \text{and} \quad I = \frac{n}{p(1 - p)}
\]
TESTS OF HOMOGENEITY OF PARALLEL SAMPLES

for coupling and

\[ S = \frac{(AB) + (ab)}{p} - \frac{(Ab) + (aB)}{1 - p} \quad \text{and} \quad I = \frac{n}{p(1 - p)} \]

for repulsion. The \( \chi^2 \) is \( S^2 / I \).

To test for homogeneity of the first two samples

\[
\begin{array}{cccccc}
\varphi & 12 & 13 & 11 & 8 & 44 \\
\sigma & 13 & 15 & 16 & 16 & 60 \\
\end{array}
\]

we obtain the estimate of \( p \) from the totals.

\[ p = \frac{28 + 27}{104} = \frac{55}{104} \]

\[ S_1 = \frac{20}{1 - p} + \frac{24}{p} = \frac{(49 \times 24 - 55 \times 20)}{104} \]

\[ I_1 = \frac{44}{p(1 - p)} = \frac{55 \times 49}{104^2 \times 44} \]

\[ S_1^2 = \frac{(49 \times 24 - 55 \times 20)^2}{104 \times 55 \times 49} = \frac{5776 \times 1}{1} \]

\[ I_1^2 = \frac{44 \times 55 \times 49}{1 - p} = \frac{5776 \times 1}{2095 \times 44} \]

\[ I_2^2 = \frac{5776 \times 104}{2095 \times 60 \times 44} = 0.0814 \]

Similarly, we obtain the following four values of \( \chi^2 \) to test for sex homogeneity within mating types.

<table>
<thead>
<tr>
<th>Mating Type</th>
<th>( \chi^2 )</th>
<th>D.F.</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varphi \varphi (C) )</td>
<td>0.0844</td>
<td>1</td>
<td>( &gt;0.75 )</td>
</tr>
<tr>
<td>( \varphi \varphi (R) )</td>
<td>0.0575</td>
<td>1</td>
<td>( &gt;0.45 )</td>
</tr>
<tr>
<td>( \sigma \sigma (C) )</td>
<td>0.0251</td>
<td>1</td>
<td>( &gt;0.87 )</td>
</tr>
<tr>
<td>( \sigma \sigma (R) )</td>
<td>0.0089</td>
<td>1</td>
<td>( &gt;0.84 )</td>
</tr>
<tr>
<td>Total</td>
<td>0.7059</td>
<td>4</td>
<td>( &gt;0.95 )</td>
</tr>
</tbody>
</table>

The total \( \chi^2 \), 0.7059 with 4 degrees of freedom, and the individual \( \chi^2 \)’s with 1 degree of freedom have very high probabilities, thus showing re-
markable agreement within mating types. The probability of observing a $\chi^2$ less than 0.7059 with 4 degrees of freedom is less than 5%, which shows that such good agreement can be expected very rarely. This might lead the experimenter to suspect his material. In such a case nothing can be said for or against the offered hypothesis. For instance, an unconscious bias as to the nature of things to be expected on the part of the experimenter may result in wrong recording.

Accepting, however, the fact that there is no sex difference within mating types, we may proceed to test whether any difference is caused by the mating types. The total frequencies are

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma\sigma$ (C)</td>
<td>25 28 27 24 104</td>
<td>0.0015</td>
</tr>
<tr>
<td>$\sigma\sigma$ (R)</td>
<td>26 23 23 35 107</td>
<td>0.7982</td>
</tr>
<tr>
<td>$\sigma\sigma$ (C)</td>
<td>48 35 40 37 160</td>
<td>2.1757</td>
</tr>
<tr>
<td>$\sigma\sigma$ (R)</td>
<td>32 24 24 21 111</td>
<td>0.7359</td>
</tr>
<tr>
<td>Total</td>
<td>482</td>
<td>3.7093</td>
</tr>
</tbody>
</table>

The common value of $p$ is obtained by equating the total score to zero.

$$S_1 + S_2 + S_3 + S_4 = \frac{25 + 24}{1 - p} + \frac{28 + 27}{p}$$

$$= \frac{23 + 23}{1 - p} + \frac{26 + 35}{p}$$

$$= \frac{48 + 37}{1 - p} + \frac{35 + 40}{p}$$

$$= \frac{24 + 24}{1 - p} + \frac{32 + 31}{p} = 0$$

or

$$\frac{228}{1 - p} + \frac{254}{p} = \frac{254}{182}$$

The values of $\chi^2$ in the last column above are calculated for the value of $p = 254/482$. For instance, the first value is

$$\frac{(49 \times 254 - 55 \times 228)^2}{254 \times 228 \times 104} = 0.0015$$

The total $\chi^2$ is 3.7093, which is not significant on 3 degrees of freedom, thus indicating close agreement of the four samples. The best estimate
of the recombination fraction is

$$\frac{2}{4} \text{ or } 52.7\%$$

thus indicating the possibility of the recombinants' exceeding 50%.

5c Contingency Tables

5c.1 The Probability of an Observed Configuration and Tests in Large Samples

If the individuals of a population can be described as belonging to one of $r$ categories, $A_1, \ldots, A_r$ with respect to an attribute $A$, and to one of $s$ categories, $B_1, \ldots, B_s$ with respect to an attribute $B$, and so on, then we have a frequency distribution of individuals in $r \times s \times \cdots$ classes, a typical class being represented by $A_i B_j \cdots$. If there are $k$ attributes on the total, the arrangement described above is called a $k$-fold contingency table.

In this section various problems connected with two attributes are discussed, the treatment being similar in the general case. Let the observed frequency in the class $A_i B_j$ be denoted by $n_{ij}$ and the probability by $\pi_{ij}$. Also let

$$n_{11} + n_{12} + \cdots + n_{1s} = n_1, \quad \pi_{11} + \pi_{12} + \cdots + \pi_{1s} = \pi_1,$$

$$n_{1j} + n_{2j} + \cdots + n_{rj} = n_j, \quad \pi_{1j} + \pi_{2j} + \cdots + \pi_{rj} = \pi_j,$$

$$n_1 + n_2 + \cdots = n_+, \quad n_+ + n_2 + \cdots = n_+,$$

$$\pi_1 + \pi_2 + \cdots = \pi_+, \quad \pi_+ + \pi_2 + \cdots = 1.$$

The probability of the observed frequencies is

$$n_! \prod_{i=1}^{r} \left( \frac{\pi_{ij}}{n_{ij}} \right)^{n_{ij}} \cdot n_+ ! \prod_{j=1}^{s} \left( \frac{\pi_{i.}}{n_{i.}} \right)^{n_{i.}} \cdot \cdots \cdot \prod_{i=1}^{r} \prod_{j=1}^{s} \left( \frac{\pi_{ij}}{n_{ij}} \right)^{n_{ij}} \cdot \cdot \cdot$$

If $\pi_{ij} = \pi_1 \pi_2$, then the above expression becomes

$$n_! \prod_{i=1}^{r} \left( \frac{\pi_{i.}}{n_{i.}} \right)^{n_{i.}} \cdot n_+ ! \prod_{j=1}^{s} \left( \frac{\pi_{j.}}{n_{j.}} \right)^{n_{j.}} \cdot \prod_{i=1}^{r} \prod_{j=1}^{s} \left( \frac{\pi_{ij}}{n_{ij}} \right)^{n_{ij}} \cdot \cdot \cdot$$

The first two expressions give the probability of the marginal totals,
and the third gives the probability of the class frequencies for fixed values of the marginal totals.

\[ P(n_{ij} \mid n_i, n_j) = \frac{n_i! \prod n_{ij}!}{n_j!} \prod \frac{1}{n_{ij}!} \quad (5c.1.1) \]

It is interesting to observe that the above expression is independent of the hypothetical values of the proportions, provided, of course, that the attributes are independent, i.e., that the probability \( r_{ij} \) is the product of the probabilities for the \( i \)th category of the first attribute and the \( j \)th category of the second attribute.

In some situations, especially in designed experiments, one set of marginals is determined in advance. Thus, for instance, we might choose a number of individuals and inoculate them against an infection. Another chosen number of individuals could be kept as controls. Both groups supply the number of individuals infected and not infected from which a \( 2 \times 2 \) contingency table can be set up. In general, if the row totals are fixed in advance, then, assuming the same set of probabilities \( p_1, \ldots, p_s \) for different categories in each row, the joint probability of the observations is

\[ P(n_{ij} \mid n_i) = \prod_{i=1}^{r} \frac{n_i!}{n_{i1}! \ldots n_{is}!} p_1^{n_{i1}} \cdots p_s^{n_{is}} \]

The probability of the marginal totals \( n_{ij} \) in this case is

\[ P(n_{j} \mid n_i) = \frac{n_{j}!}{n_{j1}! \ldots n_{js}!} p_1^{n_{j1}} \cdots p_s^{n_{js}} \]

which is obtained by summing the previous expression over \( n_{ij} = \sum_{j} n_{ij} \) for \( j = 1, \ldots, s \). Hence

\[ P(n_{ij} \mid n_i, n_j) = \frac{P(n_{ij} \mid n_i)}{P(n_j \mid n_i)} = \frac{\prod n_{ij}! \prod n_{ij}! \frac{1}{n_{ij}!}}{n_{i1}! \cdots n_{is}!} \]

which is the same as the expression \( (5c.1.1) \) obtained in the general case.

5c.2 Tests of Independence in a Contingency Table

If the probabilities \( r_{ij} \) of the cells in a contingency table are assigned, then to test the hypothesis that the data are in agreement with these hypothetical values the statistic

\[ \chi^2 = \sum \sum \frac{(n_{ij} - n \cdot r_{ij})^2}{n \cdot r_{ij}} = \sum \frac{(O - E)^2}{E} \quad \text{(over all classes)} \]
can be used as $X^2$ with $(rs - 1)$ degrees of freedom, the only restriction being $\sum n_{ij} = n$. If the attributes are independent, then the cell probabilities satisfy the relations

$$\forall i \quad \pi_{ij} = \pi_i \pi_j \quad \text{for all } i \text{ and } j$$

How can this hypothesis be tested on the basis of the observed data? Two situations may arise.

(i) The hypothetical probabilities $\pi_i$ and $\pi_j$ specifying the marginal distributions may be known, in which case we are required to examine whether the cell probabilities could be constructed by the above law.

(ii) The hypothetical proportions of the marginal frequencies not being known, we are required to test whether the attributes are independent.

In the first problem we have the total

$$\chi^2 = \sum \left( \frac{n_{ij} - n_i \pi_j}{n_i \pi_j} \right)^2 \quad \text{D.F.} = rs - 1$$

which measures the overall discrepancy of the observed from the expected. From this we can single out two components

$$\chi_1^2 = \sum_i \left( \frac{n_i - n_i \pi_i}{n_i \pi_i} \right)^2 \quad \text{D.F.} = r - 1$$

$$\chi_2^2 = \sum_j \left( \frac{n_j - n_j \pi_j}{n_j \pi_j} \right)^2 \quad \text{D.F.} = s - 1$$

which measure the discrepancy of the observed marginal frequencies from the expected. With these statistics we can test whether the observed marginal totals are as expected. On subtracting $\chi_1^2$ and $\chi_2^2$ from the total, we obtain

$$\chi_3^2 = \chi^2 - \chi_1^2 - \chi_2^2$$

$$= \sum \left[ \frac{n_{ij} - n_i \pi_j - \pi_j (n_i - n_i \pi_j) - \pi_i (n_i - n_i \pi_i)}{n_i \pi_j} \right]^2$$

It may be noted that $\chi_3^2$ is equal to $\chi^2$, $\chi_1^2$ and $\chi_2^2$ being zero, when the frequencies are subject to the restrictions

$$n_i - n_i \pi_i = 0 \quad \text{for } i = 1, \ldots, r$$
out of which \((r - 1)\) are independent, and

\[ n_{ij} - n \cdot x_{ij} = 0 \quad j = 1, \ldots, s \]

out of which \((s - 1)\) are independent. The degrees of freedom for \(x^2\) when the frequencies are subjected to \((r - 1) + (s - 1) + 1\) (for the restriction \(\sum n_{ij} = n\)) restrictions are \((rs - r - s + 1) = (r - 1)(s - 1)\). Therefore \(x^2_3\) is distributed as \(x^2\) with \((r - 1)(s - 1)\) degrees of freedom. This component is used to test the departure from independence.

As an example, consider the fourfold contingency table

<table>
<thead>
<tr>
<th></th>
<th>(B_1)</th>
<th>(B_2)</th>
<th>(a + c)</th>
<th>(b + d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A_1)</td>
<td>(a)</td>
<td>(b)</td>
<td>(a + b)</td>
<td>(c + d)</td>
</tr>
<tr>
<td>(A_2)</td>
<td>(c)</td>
<td>(d)</td>
<td>(a + c)</td>
<td>(b + d)</td>
</tr>
</tbody>
</table>

with the marginal hypothetical proportions \((p_1, q_1)\) for \(A\) and \((p_2, q_2)\) for \(B\). The total \(x^2\) with 3 degrees of freedom is

\[
x^2 = \frac{(a - np_1p_2)^2}{np_1p_2} + \frac{(b - np_1q_2)^2}{np_1q_2} + \frac{(c - np_2q_1)^2}{np_2q_1} + \frac{(d - np_2q_2)^2}{np_2q_2}\]

The components are

\[
x^2_1 = \frac{(a + b - np_1)^2}{bp_1q_1} \quad x^2_2 = \frac{(b + d - np_2)^2}{bp_2q_2} \quad x^2_3 = \frac{1}{np_1p_2}\]

with 1 degree of freedom each.

**Example 1.** Bateson found the following distribution of sweet pea plants obtained from an intercross so that the marginal frequencies are expected to be in the ratio 3:1. If the two characters, flower color and
pollen shape, are independently inherited, then the cell frequencies are expected to be in the ratio 9:3:3:1.

<table>
<thead>
<tr>
<th>Pollen Shape</th>
<th>Flower Color</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Purple</td>
</tr>
<tr>
<td>Long</td>
<td>296</td>
</tr>
<tr>
<td>Round</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>315</td>
</tr>
</tbody>
</table>

It was seen (example 1, 5b.1) that the total $\chi^2$ of discrepancy with 3 degrees of freedom is 222.1221, which is very high. The first component is

$$X_1^2 = \frac{(a + b - np1)^2}{np1q1} = \frac{(323 - \frac{3}{11} \times 427)^2}{427 \times \frac{3}{11} \times \frac{3}{11}} = 0.0945$$

which is quite small for 1 degree of freedom, showing that the single factor segregation for pollen shape is as expected. The second component is

$$X_2^2 = \frac{(a + c - np2)^2}{np2q2} = \frac{(315 - \frac{3}{11} \times 427)^2}{427 \times \frac{3}{11} \times \frac{3}{11}} = 0.3443$$

which again is quite small. The third component is

$$X_3^2 = \frac{(a - b - c + d)^2}{np1q1q2} = \frac{(296 - 27 \times 3 - 19 \times 3 + 85 \times 9)^2}{427 \times 3 \times 3 \times 1 \times 1} = 221.6833$$

which is very high for 1 degree of freedom. The total $\chi^2$ is

$$0.0945 + 0.3443 + 221.6833 = 222.1221$$

thus agreeing with the total calculated earlier. It is seen that the whole discrepancy is concentrated in one component with a single degree of freedom. This shows that the departure of the observed from the expected cell frequencies is due to the dependence of the characters inherited but not to single factor segregations. The success of all sta-
statistical tests lies in isolating such components which are most efficient for judging the points at issue.

Suppose the hypothetical values of the marginal probabilities are not known. Then we estimate their values on the hypothesis

\[ \pi_{ij} = \pi_i \pi_j \]

The best estimates are

\[ \hat{\pi}_i = \frac{n_i}{n} \quad \text{and} \quad \hat{\pi}_j = \frac{n_j}{n} \]

These values may be substituted in the total \( \chi^2 \).

\[
\chi^2 = \sum \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}
\]

Since \((r - 1) + (s - 1)\) parameters have been estimated, the above \( \chi^2 \) has \((rs - 1) - (r - 1) - (s - 1) = (r - 1)(s - 1)\) degrees of freedom. At the estimated values the components \( \chi_1^2 \) and \( \chi_2^2 \) have zero values so that \( \chi_3^2 = \chi^2 \). Thus \( \chi_3^2 \) measures the departure from independence.

This test is useful in two situations:

(i) Suppose that in example 1 above it is found that the marginal frequencies deviate significantly from the expected. This indicates that the assigned marginal probabilities may not be correct. In fact, if the single factor segregations are disturbed owing to unequal viability of the two types of plants, then the marginal frequencies will not be in the ratio 3:1. In such a case the third component \( \chi_3^2 \) loses its importance, or, in other words, the significance of \( \chi_3^2 \) may be due to the use of wrong proportions. The best course is then to substitute the estimated proportions and use the test obtained above.

(ii) The second situation is when nothing is specified about the marginal proportions. In this case only a test for independence is possible.

It may be observed that the test of independence in a contingency table is the same as the test of homogeneity in parallel samples described in 5b.3.

If the hypothetical proportions are not known in the above example, then the \( \chi^2 \) for testing independence is

\[
\frac{(a - (a + b)(a + c)/n)^2}{(a + b)(a + c)/n} + \frac{(b - (a + b)(b + d)/n)^2}{(a + b)(b + d)/n} + \ldots + \ldots
\]
which reduces to

\[ \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{427(296 \times 85 - 27 \times 19)^2}{315 \times 112 \times 323 \times 104} = 209.3095 \]

This gives a \( \chi^2 \) higher than that obtained by using the hypothetical values of the marginal proportions. Such discrepancies will not in general lead to contradictory conclusions. The earlier test makes use of the information supplied by a total of 427 plants in testing for independence, whereas the latter makes use of the information supplied by the set of configurations having the same marginal totals. Some marginal totals, as in the present case, are more informative than the average, whereas others are less.

**Example 2.** The following data give the number of skulls excavated in three different seasons and the sex distribution as sexed by investigator A working in the first two seasons and by B working in the third season.

<table>
<thead>
<tr>
<th>Seasons</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>162</td>
</tr>
<tr>
<td>( \varphi )</td>
<td>110</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>272</td>
</tr>
</tbody>
</table>

Let us assume that in each season the excess of males is due to a random deviation from the expected equal numbers for the two sexes. The expected values are

\[ \sigma^* 136 \quad 152.5 \quad 205 \]
\[ \varphi^* 136 \quad 152.5 \quad 205 \]

giving a total of \( \chi^2 = 9.9412 + 9.9180 + 0.2440 = 20.1032 \), a high value for 3 degrees of freedom. Individually the deviations in the first two seasons \( (\chi^2 = 9.9412 \text{ and } 9.9180) \) are significant. If the sex ratio is 1:1 on the total, the \( \chi^2 \) resulting from the marginal is

\[ \frac{(552)^2}{493.5} + \frac{(435)^2}{493.5} - 987 = 13.8092 \]

which leaves \( \chi^2 = 20.1032 - 13.8092 = 6.2940 \) with 2 degrees of free-
dom for testing the differences in the sex ratio in the three seasons. This is no doubt significant, showing differences in sex ratio, but the test is not strictly valid owing to the fact that the marginal totals are not compatible with the sex ratio 1:1; the $\chi^2 = 13.8692$ being significant for 1 degree of freedom. Having observed that there is an overall discrepancy from the sex ratio 1:1 in all three seasons put together, we might ask whether the sex ratio is the same for the three seasons although it may not be 1:1. A straight test of independence for fixed marginals or homogeneity of parallel samples (5b.3) can now be calculated. This gives $\chi^2$ equal to 6.3222 with 2 degrees of freedom, showing significant differences in sex ratios in the three seasons. The agreement of this $\chi^2$ with the earlier value of 0.2310 obtained by using the hypothetical ratios is, perhaps, accidental.

One must be careful in drawing conclusions from data of this nature. It must be observed that the skulls are sexed by a subjective method of anatomical appreciation. Different investigators have different methods of sexing, leading to different ratios. The observed discrepancy in the sex ratio for different seasons may be due to the investigator in the third season being different. The observed proportions are

<table>
<thead>
<tr>
<th>Season</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>0.5056</td>
<td>0.5002</td>
<td>0.5122</td>
<td>0.5203</td>
</tr>
</tbody>
</table>

The same investigator working in the first two seasons shows a smaller proportion in the second though not significantly different from the first, the $\chi^2$ with 1 degree of freedom being 0.0175. But it is just as well that he thought that his method gave an excess of males in the first season and tried to alter his method consciously or unconsciously in the second season. The discrepancy between the investigators is then tested by the $\chi^2$ test of independence from the table:

<table>
<thead>
<tr>
<th>First and Second Seasons</th>
<th>Third Season</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
<td>342</td>
</tr>
<tr>
<td>$v$</td>
<td>235</td>
</tr>
</tbody>
</table>

The $\chi^2$ with 1 degree of freedom is 6.3055, which is significant. Thus, out of a total $\chi^2$ of 6.3222 with 2 degrees of freedom, measuring the differences in sex ratios, 1 degree alone accounts for 6.3055, which shows that the whole discrepancy between seasons is due only to the discrepancy between investigators. This might indicate a difference in the method of sexing or that the investigators' proportions referred to dif-
Different strata from which the skulls are excavated and there might be stratum differences.

It should also be observed that the deviation of the overall sex ratio from 1:1 may be due to a wrong technique of sexing.

Example 3. Consider the following data, collected from a number of schools, regarding speech defects ($S_1$, $S_2$, $S_3$) and physical defects ($P_1$, $P_2$, $P_3$) of school children.

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>45</td>
<td>26</td>
<td>12</td>
</tr>
<tr>
<td>$P_2$</td>
<td>32</td>
<td>50</td>
<td>21</td>
</tr>
<tr>
<td>$P_3$</td>
<td>4</td>
<td>10</td>
<td>31</td>
</tr>
</tbody>
</table>

The expected values on the hypothesis of independence are:

<table>
<thead>
<tr>
<th></th>
<th>30.982</th>
<th>38.447</th>
<th>11.571</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32.814</td>
<td>40.820</td>
<td>12.286</td>
</tr>
<tr>
<td></td>
<td>19.124</td>
<td>23.733</td>
<td>7.143</td>
</tr>
</tbody>
</table>

The $\chi^2$ with $2 \times 2 = 4$ degrees of freedom is 34.8828, which is significant. It is seen that, although the frequency in one cell is as small as 4, the expected is large enough for the $\chi^2$ approximation to hold. But the frequency also should be large enough for the approximation to be good. In such cases it is reasonable to combine two cells by adding their frequencies and treat them as one cell for purposes of tests of significance.

In the above example 4 and 10 may be added to yield an observed frequency of 14 with the corresponding expected 11.571 + 12.286 = 23.857. The new $\chi^2$ is 33.573. Although the summation is now taken over one cell less, theoretically 1 degree of freedom is not lost. So to use the new $\chi^2$ with $(4 - 1) = 3$ degrees of freedom is to overestimate significance, whereas to consider it as with 4 degrees of freedom is to underestimate significance. Although definite conclusions can be drawn either when the new $\chi^2$ is not significant for 3 degrees of freedom or when it is significant for 4 degrees of freedom as in the present case, it is not possible to say anything when the new $\chi^2$ is significant for 3 degrees and not for 4 degrees of freedom. In such a situation, for the purpose of the $\chi^2$ test a new set of expected values may be obtained by considering the cells $S_1P_2$ and $S_3P_3$ as constituting a single cell. If $p_1$, $q_1$, $r_1$ and $p_2$, $q_2$, $r_2$ are the marginal proportions for physical defects and speech defects, then the probability of the observed frequencies on the hypothesis of independence is

$$\text{const.} \cdot (p_1p_2)^4(p_1q_2)^6(p_1r_2)^2 \cdots [(p_2 + q_2)q_2]^{14}(r_1^2)^{12}$$
The maximum likelihood estimates are

\[ \begin{align*}
217p_1 &= 83 \\
217q_1 &= 103 \\
217r_1 &= 31 \\
217r_2 &= 50 \\
70p_2 &= 77q_2
\end{align*} \]

which give

\[ p_2 = 0.387308 \quad q_2 = 0.382278 \quad r_2 = 0.230114 \]

The estimates of \( p_1, q_1, r_1 \) are the same as before. The new expectations are

\[ \begin{align*}
32.147 &= 31.720 \\
20.832 &= 20.757 \\
7.143 &= 7.183 \\
(12.000 + 11.851) &= 7.143
\end{align*} \]

and the \( \chi^2 \) with \( (7 - 4) \) degrees of freedom is 31.2472. This test is valid in the sense that when significance is noted the hypothesis of independence is rejected.

Fisher recommended a test based on likelihood, which is more appropriate when the cell frequencies are small. This is defined by

\[ L = \sum \frac{O - E}{E} \]

which in tests of independence in a contingency table can be written

\[ L = 2 \sum \frac{O \log \frac{O}{E} - E \log \frac{E}{N}}{N} \]

In this case \( L \) is approximately distributed as \( \chi^2 \) with \( (r - 1)(s - 1) \) degrees of freedom. The value of \( L \) in the above problem is 30.4418, which is significant on 4 degrees of freedom. Even the use of \( L \) requires a large sample, in which case \( \chi^2 \) and \( L \) tend to equivalence and there is no theoretical justification of one in preference to the other. In small samples the statistic \( L \) may be more appropriate, but it cannot be used unless its distribution is known. Therefore some such technique as that followed above may be used. It must be emphasized that the object of the test is first to establish departure from independence in a general way. For this it is enough to use a valid test which is simple to compute. Afterwards more refined tests may be used to examine some portions of the contingency table. For instance, we may inquire whether the two physical defects \( P_2 \) and \( P_3 \) and the speech defects \( S_1 \) and \( S_2 \) are associated. This needs a refined technique discussed in the next section.
mated do not enter into the large sample distribution of the $\chi^2$ statistic. But in small samples it might happen that the $\chi^2$ approximation breaks down and/or the unknown parameters appear in the exact distribution of $\chi^2$. In the latter case no exact test of significance is possible, owing to the presence of the unknown parameters in the probability distribution. Such unknown parameters are called nuisance parameters.

One way of getting rid of the nuisance parameters is to compare the particular observed sample, not with the whole population of samples with which a comparison might be made if the exact values of the nuisance parameters were known, but with a subpopulation selected with reference to the sample in such a way that the distribution of a statistic in this subpopulation does not involve any unknown parameter (Bartlett, 1937; Hotelling, 1940). For instance, it is shown that on the hypothesis of independence of two attributes the probability of cell frequencies, given the marginal totals, is

$$P(n_{ij} | n_i, n_j) = \frac{\prod n_{ii} \prod n_{jj} 1}{n!}$$

which does not contain the hypothetical values $\pi_i$ and $\pi_j$. The distribution of $\chi^2$ may then be found, using the conditional distribution of the cell frequencies. This admits the possibility of determining the exact probabilities in tests of independence.

Consider the fourfold table with frequencies $a, b, c, d$. The probability of the observed configuration, given the marginal totals, is

$$\frac{(a+b)(a+c)(b+d)(c+d)!}{a!b!c!d!n!}$$ (5c.3.1)

If $a, b, c, d$ are considered as four independent Poisson variates having the joint probability

$$e^{-m_1} \frac{m_1^a}{a!} e^{-m_2} \frac{m_2^b}{b!} e^{-m_3} \frac{m_3^c}{c!} e^{-m_4} \frac{m_4^d}{d!}$$ (5c.3.2)

where

$$m_1 = \frac{(a+b)(a+c)}{n}$$
$$m_2 = \frac{(b+d)(a+b)}{n}$$
$$m_3 = \frac{(a+c)(c+d)}{n}$$
$$m_4 = \frac{(b+d)(c+d)}{n}$$
then the probability of the totals \((a + b), (a + c), (b + d), (c + d)\) is
\[
e^{-\left(m_1 + m_2 + m_3 + m_4\right)} \frac{(a + b)^{m_1}(a + c)^{m_2}(b + d)^{m_3}(c + d)^{m_4}}{n^m} \times \sum_{a!b!c!d!} \frac{n!}{(a + b)! (a + c)! (b + d)! (c + d)!}
\]
(5c.3.3)
where the summation is taken over constant values of \((a + b), (a + c), (b + d), (c + d)\). Since
\[
\sum_{a!b!c!d!} \frac{n!}{(a + b)! (a + c)! (b + d)! (c + d)!}
\]
the probability of \(a, b, c, d\) for given values of \((a + b), (a + c), (b + d), (c + d)\) is the ratio of the expression (5c.3.2) to (5c.3.3),
\[
\frac{(a + b)! (a + c)! (b + d)! (c + d)!}{n!a!b!c!d!}
\]
which is the same as the expression (5c.3.1). Thus the probability of a given configuration in a contingency table for given marginal totals may be considered as a relative probability of four Poisson variates subject to three independent restrictions. If the values of \(m_1, m_2, m_3, m_4\) or the expected cell frequencies are large, the Poisson distributions tend to normality, in which case the \(\chi^2\) statistic is distributed as \(\chi^2\) with 1 degree of freedom. On the other hand, if the expectations are small the continuous distribution of \(\chi^2\) cannot be used. In such a case a direct approach is to calculate the sum of the probabilities of the observed and the less probable configurations and to reject the hypothesis if this sum is small (either below 5 or 1%). The following illustrations explain the method.

Example 1. Do the following data on sociability (S) and non-sociability (NS) of soldiers recruited in cities (C) and villages (V) suggest that city soldiers are more sociable than village soldiers?

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>NS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

The smaller frequencies in one diagonal suggest that city soldiers are more sociable. But it must be ascertained whether such a configuration as the observed and those indicating a higher degree of sociability
of the city recruits can occur by chance if in fact there was no difference in the sociabilities of the city and the village recruits. Since for fixed marginals the probability of a given configuration \(a, b, c, d\) in the four cells is

\[
\frac{17! 20! 19! 18!}{37!} \frac{1}{a b c d!}
\]

we find that the probabilities for configurations with 4, 3, 2, 1, and 0 in the north-east corner cell (these being less favorable to the hypothesis of independence and more to the alternative suggested) are, respectively, 0.005218, 0.005966, 0.003907, 0.001097, and 0.000175, adding up to 0.0059. The chances are very small, thus indicating that city soldiers are more sociable than village soldiers.

If the cell frequencies are not small, this result could be established by calculating \(X^2\) for testing independence and determining the probability of a normal deviate with zero mean and unit variance exceeding \(X\) (example 2 in 5b.1). In the above case

\[
X^2 = \frac{37(13 \times 14 - 6 \times 4)^2}{17 \times 20 \times 19 \times 18} = 7.9435
\]

so that the normal probability is 0.0025, which is smaller than the actual value 0.0059, the discrepancy being due to smallness of the sample.

Yates suggested that by calculating \(X^2\) from a table obtained by increasing the smaller frequency \(*\) by \(\frac{1}{2}\) without altering the marginal totals a closer approximation to the actual probability is realized. In the present example, the new \(X^2\), said to be corrected for continuity, is 6.1922. The value of \(X = 2.4884\), so that the normal probability is 0.0064, which is closer to the actual value than in the case of the uncorrected \(X^2\).

A slightly different method suggested by V. M. Dandekar involves the calculation of \(X_0^2\), \(X_{-1}^2\), and \(X_1^2\) for the observed configuration and those obtained by increasing and decreasing the smallest frequency by unity. From these a corrected \(X^2\) can be obtained from the formula

\[
X_0^2 - X_{-1}^2 = X_1^2 - X_0^2
\]

*The reference is to the smaller frequency in the diagonal (6,4) under consideration indicating nonsociability of the village recruits. In the general test of independence \(\frac{1}{2}\) is to be added to a frequency so as to obtain a smaller \(X^2\).
In the present example \( x_0^2 = 7.9435, x_1^2 = 12.0995, x_{-1}^2 = 4.6587, \) and
\[
\chi^2 = \frac{7.9435 - 4.6587}{12.0995 - 4.6587} (12.0995 - 7.9435) = 6.1086
\]
\[
\sqrt{\chi^2} = 2.4715
\]
The normal probability is 0.0068, which is also close to the actual value.

The likelihood test
\[
L = 220 \log_e \frac{O}{E}
\]
in this case gives the value 8.2811. The value of \( \chi \) is \( \sqrt{8.2811} = 2.8778 \) so that the probability is much smaller than the actual value. Thus the likelihood test does not improve the situation.

Example 2. In example 1, the object of investigation was to study whether city soldiers are more sociable than village soldiers. This necessitated consideration of the deviations from the expected in one way only. But, in general, if the object is to discover association between two attributes without specifying the nature of the association, it is necessary to consider all possible deviations from the expected. Thus, if a plant can be classified with respect to one of two flower colors and one of two pollen shapes, we pose the question whether the pollen shape and flower color are independent. In such a case we have to determine the total probability of the observed configuration and those less probable than this. Let us consider the same data as in example 1. The configurations less probable than the observed and indicating association in one way only have the values 3, 2, 1, and 0 in the north-east cell of the table. The probabilities of these configurations and the observed have been calculated in that example, and they add up to 0.0059. The configurations less probable than the observed but indicating association the other way are those which have 4, 3, 2, 1, and 0 in the north-west corner and their probabilities are, respectively, 0.022088, 0.031864, 0.058733, 0.081132, and 0.00023, adding up to 0.0023. The total probability is then 0.0059 + 0.0023 = 0.0082, which is small, thus indicating departure from independence.

If the sample is large, we can find the probability by directly entering the uncorrected \( \chi^2 \) in a \( \chi^2 \) table with 1 degree of freedom. In the present example \( \chi^2 = 7.9435 \) with the associated probability about 0.005, which is smaller than the actual value 0.0082.

Using Yates’ correction for continuity, the \( \chi^2 = 6.1922 \), with the corresponding probability 0.0128, which is higher than the actual value.

To extend Dandekar’s correction to this case we first note that the
\( x^2 \) values below and above the observed \( x^2 = 7.9435 \) are 6.0598 and 9.7448, corresponding to the partitions 5 12 4 13 and and 14 6 15 5. The corrected \( x^2 \) is

\[
\begin{align*}
\hat{x}^2 &= 7.9435 - 6.0598 \times 2 \\
&= 7.9435 - (9.7448 - 7.9435) \\
&= 7.9435 - 1.8013 \\
&= 6.1422
\end{align*}
\]

which gives a probability 0.0072, almost exactly equal to the actual value. In general, Dandekar's correction is slightly better than that of Yates, although the correction is simpler in Yates' method. In testing for linkage on the basis of data classified according to two factors, it is enough to test for association one way if it is known that the recombination fraction is less than \( \frac{1}{2} \). It is now known that the recombination fraction can exceed \( \frac{1}{2} \), as demonstrated by Fisher. So it is better first to disprove the hypothesis of independence without inquiring as to the nature of association. Further it must be noted that departure from independence may occur owing to various causes in experimental data, and it is better to use a test which gives a direct appraisal of the data as to its compatibility with the hypothesis of independence.

5d Tests in Poisson Populations

It was shown in 2a.3 that the probability of \( k \) independent Poisson variates can be written as the product of

\[
P(X_1 + \cdots + X_k) = e^{-\mu} \prod_{i=1}^{k} \frac{(\mu_i)^{x_i}}{x_i!}
\]

and

\[
P(X_1, \ldots, X_k | \Sigma X) = \frac{(\Sigma X)!}{X_1! \cdots X_k!} \left( \frac{\mu_1}{\Sigma \mu} X_1 \right)^{x_1} \cdots \left( \frac{\mu_k}{\Sigma \mu} X_k \right)^{x_k}
\]

the latter being the multinomial probability. Testing whether the observations have come from the same Poisson population is equivalent to testing whether in a series of \( (X_1 + \cdots + X_k) \) trials the frequencies \( X_1, \ldots, X_k \) could arise from a multinomial distribution with equal probabilities in the \( k \) classes, since \( \mu_i/\Sigma \mu = 1/k \), if \( \mu_i \) are all equal. For instance, to test whether the observations 8, 3, 1 could have arisen from the same Poisson population, the test explained in example 3 of 5b.1 could be carried out. If the individual values are not small, then the
\( x^2 \) test of departure from expected values could be used. The expected value is \( X = \frac{\sum X_i}{k} \), corresponding to each \( X_i \), so that

\[
\chi^2 = \frac{\sum (X_i - \bar{X})^2}{\bar{X}} = \frac{\sum X_i^2}{X} - k\bar{X}
\]

This has \((k - 1)\) degrees of freedom. We can also test any hypothesis specifying the proportions \( \lambda_i = \mu_i/\Sigma \mu \). This is equivalent to testing whether the frequencies could have arisen from a multinomial population with proportions \( \lambda_1, \ldots, \lambda_k \) so that the \( \chi^2 \) with \((k - 1)\) degrees of freedom is

\[
\chi^2 = \frac{\sum (X_i - \lambda_i \Sigma X)^2}{\lambda_i \Sigma X}
\]

**Example 1.** Four samples of sizes 120, 100, 100, 125 from Poisson populations gave the following mean values: 251/120, 323/100, 180/100, 426/125. Do the populations have the same mean values?

It is seen in 2a.3 that the sum of \( n \) independent Poisson variates is also a Poisson variate with mean value equal to \( n \) times the mean of the original population. Considering the sums observed and assuming equal mean \( \mu \) for all the four populations, we have the following.

<table>
<thead>
<tr>
<th>Sum</th>
<th>251</th>
<th>323</th>
<th>180</th>
<th>426</th>
<th>1180</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>120</td>
<td>100</td>
<td>100</td>
<td>125</td>
<td>445</td>
</tr>
</tbody>
</table>

The test for equality of mean values is equivalent to testing whether the sums are in the ratio 120:100:100:125. The expected distribution of 1180 is

\[
318.20 \quad 265.17 \quad 265.17 \quad 331.46
\]

so that \( \chi^2 = 81.12 \), which is significant for 3 degrees of freedom. The mean values cannot be considered to be the same for all populations.

**Example 2.** Obtain the distribution of entries in a two-way table for fixed marginals if the observation in the \((i,j)\)th cell is regarded as a Poisson variate with mean value \( r_i \delta_j \). Assume that \( \Sigma r_i = r \) and \( \Sigma \delta_j = \delta \).

The joint distribution of \( n_{ij} \), \( i = 1, \ldots, r; j = 1, \ldots, s \), is

\[
e^{-\Sigma r_i \delta_j \Pi} \frac{\left( r_i \delta_j \right)^{n_{ij}}}{n_{ij}!} \times e^{-\Sigma r_i \Pi} \frac{\left( r_i \right)^{n_i}}{n_i!} \times e^{-\Sigma \delta_j \Pi} \frac{\left( \delta_j \right)^{n_j}}{n_j!} \times 1 \times \frac{\Pi n_i ! \Pi n_j !}{n !}
\]

\[
e^{-\Sigma r_i \delta_j \Pi} \frac{\left( r_i \delta_j \right)^{n_{ij}}}{n_{ij}!} \times \frac{n_i ! \Pi n_i !}{n !}
\]
The last expression is the desired probability, the first three representing
the joint probability of the observed marginals. The relative probability
is the same as in a contingency table with two independent attributes
and fixed marginals. In any problem a test of independence can be
carried out to test the hypothesis that the mean of the Poisson popu­
lation for the (i, j)th cell can be written as the product of two parameters
specific for the ith row and the jth column. If this is true, then the
marginal totals can be used to test whether \( r_i \) are identical or \( \beta_j \) are
identical.

In an analysis of randomized blocks, when the plot yields can be
considered to be Poisson variates it seems reasonable to set up the
product hypothesis \( r_i \beta_j \), \( r_i \) representing the treatment effect, and \( \beta_j \) the
block effect. The adequacy of the product hypothesis can be tested
before testing for treatment differences.

### 6e Transformation of Statistics

#### 6.1 A General Lemma

Let the joint distribution of the statistics \( T_1, \ldots, T_k \) tend to the
\( k \)-variate normal form with mean values \( \theta_1, \ldots, \theta_k \) and dispersion
matrix \( \gamma^{-1}(\sigma_{ij}) \), where \( \sigma_{ij} \) are finite and \( n \) is the sample size. This
means that the variables \( \sqrt{n}(T_i - \theta_i) \), \( \ldots \), \( \sqrt{n}(T_k - \theta_k) \) are in the
limit distributed as a \( k \)-variate normal distribution with zero mean
values and dispersion matrix \( (\sigma_{ij}) \).

**Lemma:** If \( f(T_1, \ldots, T_k) \) is a continuous function with continuous
first partial derivatives then the variable

\[
u = \sqrt{n}[f(T_1, \ldots, T_k) - f(\theta_1, \ldots, \theta_k)]
\]
is distributed normally in the limit with zero mean and variance

\[
\Sigma\Sigma_{\sigma_{ij}} \frac{df}{d\theta_i} \frac{df}{d\theta_j}
\]

Since \( f(T_1, \ldots, T_k) \) has continuous partial derivatives in the neighbor­
hood of \( \theta_1, \ldots, \theta_k \), expanding by the mean value theorem we get

\[
f(T_1, \ldots, T_k) = f(\theta_1, \ldots, \theta_k) + \Sigma(T_i - \theta_i) \left( \frac{df}{d\theta_i} + \eta_i \right)
\]

where \( \eta_i \to 0 \) as \( T_i \to \theta_i \). Now

\[
u = \sqrt{n}\Sigma(T_i - \theta_i) \frac{df}{d\theta_i} = \sqrt{n}\Sigma(T_i - \theta_i)\eta_i
\]
so that \( u \) and \( v \) have the same limiting distribution if \( v_n(T_i - \theta_0) \to 0 \) stochastically. To prove this it is enough to show that \( v_n(T_i - \theta_0) \to 0 \) stochastically for all \( i \).

\[
P(\left| v_n(T_i - \theta_0) \right| < \epsilon) \\
> P(\left| v_n(T_i - \theta_0) \right| < \epsilon') \\
> P(\left| v_n(T_i - \theta_0) \right| < \epsilon'') \\
> P(\left| v_n(T_i - \theta_0) \right| < \epsilon'') - P(\left| v_i \right| > \epsilon') > 1 \text{ for any } \epsilon' \text{ by choosing } A \text{ and } n \text{ sufficiently large. Also since } T_i \to \theta_i \text{ stochastically and } \eta_i \to 0 \text{ as } T_i \to \theta_i \text{ mathematically, it follows that } P(\left| v_i \right| > \epsilon'') > \epsilon' \text{ for large } n, \text{ however small } \epsilon' \text{ may be. If now } \epsilon' \text{ is chosen such that } \epsilon'' = \epsilon', \text{ then}
\]

\[
P(\left| v_n(T_i - \theta_0) \right| < \epsilon'') - P(\left| v_i \right| > \epsilon') > 1 - \epsilon'' > 1 - \epsilon\]

since \( \epsilon' \) and \( \epsilon'' \) are arbitrary. This proves the required result. The statistic \( \Sigma(T_i - \theta_i) (\partial f / \partial \theta_i) \), being a linear function of statistics which tend to be normally distributed, is itself normally distributed in the limit with

\[
E \left[ \Sigma(T_i - \theta_i) (\partial f / \partial \theta_i) \right] = \Sigma E'(T_i - \theta) (\partial f / \partial \theta_i) \rightarrow 0 \quad \text{under certain conditions.}
\]

and

\[
V \left[ \Sigma(T_i - \theta_i) (\partial f / \partial \theta_i) \right] = \frac{1}{n} \sum_{ij} \frac{\partial f}{\partial \theta_j} \frac{\partial f}{\partial \theta_i} \sigma_{ij}
\]

Therefore \( f(T_1, \ldots, T_k) \) has the asymptotic mean and variance

\[
f(\theta_1, \ldots, \theta_k) \quad \text{and} \quad \frac{1}{n} \sum_{ij} \frac{\partial f}{\partial \theta_j} \frac{\partial f}{\partial \theta_i} \sigma_{ij} \quad (5e.1.1)
\]

In particular let \( T \) be a statistic such that \( \sqrt{n} (T - \theta) / \sqrt{\psi(\theta)} \) has a limiting standard normal distribution, which we may describe as \( T \) having a limiting normal distribution with asymptotic mean, \( \theta \) and asymptotic variance \( (\partial f / \partial \theta_i) \sigma_{ij} / \partial \theta_j \sigma_{ij} \), which is also called standard error \( (\partial f / \partial \theta_i) \sigma_{ij} / \partial \theta_j \sigma_{ij} \) of \( T \). From the above analysis, it follows that if \( F(T) \) is a differentiable function of \( T \) then

\[
\sqrt{n} \left[ F(T) - F(\theta) \right] \sim \sqrt{(F'(\theta))^2 \psi(\theta)}
\]

where \( F' \) denotes the derivative, has a limiting standard normal distribution. Or, in other words \( F(T) \) has a limiting normal distribution with mean \( F(\theta) \) and a.v. or standard error

\[
[F'(\theta)]^2 \psi(\theta) / n
\]
SQUARE ROOT TRANSFORMATION OF POISSON VARIATE

In some cases \( w(O)/n \) the a.v. of \( T \), may be independent of \( \theta \). Otherwise it may be necessary to transform the statistic \( T \) such that the new statistic has an asymptotic variance independent of \( \theta \). Let \( F(T) \) be the transformation needed; then

\[
\text{a.v.}\{ F(T) \} = \frac{1}{n} \{ F'(\theta) \}^2 \psi(\theta)/n
\]

On equating this to a constant, the following differential equation is obtained.

\[
\frac{dF}{d\theta} = \frac{c}{\sqrt{\psi(\theta)}}
\]

or

\[
F = \int \frac{c \, d\theta}{\sqrt{\psi(\theta)}}
\]

This result is applied in deriving the following transformations.

5e.2 The Square Root Transformation of the Poisson Variate

If \( x \) is a Poisson variate, then

\[
E(x) = \mu \quad V(x) = \mu
\]

The functional form of the transformation is supplied by

\[
F(\mu) = \int c \frac{d\mu}{\sqrt{\mu}}
\]

by choosing \( c \) suitably. The transformed variable \( \sqrt{x} \) has the asymptotic mean and variance

\[
\sqrt{\mu} \quad \text{and} \quad \frac{1}{4}
\]

when \( \mu \) is large. It was found by Anscombe (1948) that the transformation \( \sqrt{x + b} \) where \( b \) is a suitably determined constant has some theoretical advantages. Let \( (x - \mu) = t \) and \( (\mu + b) = \mu' \); then by using Taylor's expansion we find

\[
\sqrt{x + b} = \sqrt{\mu'} \left[ 1 + a_1 \frac{t}{\mu'} - a_2 \left( \frac{t}{\mu'} \right)^2 + \cdots \right]
\]

where

\[
a_s = (-1)^{s+1}(1)(-3) \cdots \frac{(-2s + 3)}{s!2^s}
\]

Observing that the Poisson moments are

\[
E(t) = 0 \quad E(t^2) = \mu \quad E(t^3) = \mu \quad E(t^4) = 3\mu^2 + \mu
\]
we find, by taking expectations of both sides of the expansion

\[ E(\sqrt{x} + b) = \sqrt{\mu} + b - \frac{1}{8\sqrt{\mu}} + \frac{24b - 7}{128\mu\sqrt{\mu}} + \cdots \]

\[ V(\sqrt{x} + b) = \frac{1}{4} \left\{ 1 + \frac{3 - 8b}{8\mu} + \frac{32b^2 - 52b + 17}{32\mu^2} + \cdots \right\} \]

which, on choosing the value \( b = \frac{3}{8} \), reduce to

\[ E(\sqrt{x} + \frac{3}{8}) = \sqrt{\mu} + \frac{3}{8} - \frac{1}{8\sqrt{\mu}} + \frac{1}{64\mu\sqrt{\mu}} + \cdots \]

\[ V(\sqrt{x} + \frac{3}{8}) = \frac{1}{4} \left( 1 + \frac{1}{16\mu^2} + \cdots \right) \]

The variance of \( \sqrt{x} + \frac{3}{8} \) is more stable than that of \( \sqrt{x} \) because the second term in the expansion of the variance of \( \sqrt{x} + \frac{3}{8} \) is \( \mathcal{O}(1/\mu^2) \).

5e.3 The Sin^{-1} Transformation of the Binomial Proportion

The binomial proportion \( r/n \) has the mean value \( \mu \) and variance \( \mu(1 - \mu)/n \). The transformation is obtained by solving the equation

\[ F(\mu) = \int \frac{e^{\sqrt{n}}}{{\sqrt{n}}(1 - \mu)} d\mu \]

\[ = \sin^{-1} \sqrt{\mu} \quad \text{choosing } c \text{ suitably} \]

\[ a.v. \left( \sin^{-1} \left( \frac{r}{n} \right) \right) = \frac{1}{4n} \]

It is shown by Anscombe (1948) that a slightly better transformation is

\[ \sin^{-1} \left( \frac{r + \frac{3}{8}}{n + \frac{3}{2}} \right) \]

which has the asymptotic variance \( 1/(4n + 2) \). If \( n \) is large, the simpler transformation \( \sin^{-1} \sqrt{r/n} \) can be used; for moderately large sample sizes the refined transformation \( \sin^{-1} \sqrt{(r + \frac{3}{8})/(n + \frac{3}{2})} \) may be used.
Example. R. A. Fisher (1949) found the following recombination fractions between undulated and agouti loci in house mice. The data relate to backcrosses so that the estimate of the fraction is the ratio of recombinants to the total offsprings.

<table>
<thead>
<tr>
<th>Classes of Heterozygous Parents</th>
<th>( A'^{A}A^{a} )</th>
<th>( A'^{A'}A' )</th>
<th>( A'^{A'}A )</th>
<th>( A'^{A'}a )</th>
<th>( A'^{A}a )</th>
<th>( A^{A}A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>182</td>
</tr>
<tr>
<td>( \phi )</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>194</td>
</tr>
</tbody>
</table>

*The number in the denominator gives the number of animals contributing to the ratio.

The data for any heterozygote supply a test with 1 degree of freedom for sex difference in the recombination fraction. The (independence) \( \chi^2 \) for sex difference is 0.0905. Similarly, we obtain the individual \( \chi^2 \) for each of the ten types of heterozygotes and obtain the sum 4.827 which has nearly 90% probability for 10 degrees of freedom, indicating no sex difference. If we ignore sex differences, then on pooling the data over sex we obtain a contingency table for heterozygotes and the nature of combinations (old or new). This supplies a \( \chi^2 = 16.315 \) with 9 degrees of freedom showing significant differences in the recombination fractions for the various heterozygotes.

This example is not suitable for further analysis on sex differences. Suppose that it is found that sex differences exist in all the ten types of mating. Then the further problem arises as to whether the sex difference is the same in all the cases. That is, we need to test whether there is interaction between sex and the nature of the heterozygote. This can be done by using the angular transformation and then applying analysis of variance. Corresponding to each observed proportion \( p \), an angle \( \phi \) is determined such that \( p = \sin^2 \phi \). If \( \phi \) is given in degrees, as in Table 12 of the Fisher and Yates tables, then \( \phi \) has the variance...
$100/n\sigma^2$ or approximately 820.7/n. The 20 angles and the necessary computational steps are as given below.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$w$</th>
<th>$dw$</th>
<th>$iFw$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$9$</td>
<td>$14.4$</td>
<td>$11.8$</td>
<td>$13.0$</td>
</tr>
<tr>
<td>$15.3$</td>
<td>$15.4$</td>
<td>$11.2$</td>
<td>$14.9$</td>
</tr>
<tr>
<td>$d$</td>
<td>$w$</td>
<td>$dw$</td>
<td>$iFw$</td>
</tr>
<tr>
<td>$-0.9$</td>
<td>$-3.6$</td>
<td>$1.8$</td>
<td>$0.2$</td>
</tr>
<tr>
<td>$77.12$</td>
<td>$60.93$</td>
<td>$55.19$</td>
<td>$91.20$</td>
</tr>
<tr>
<td>$-69.41$</td>
<td>$-219.35$</td>
<td>$171.34$</td>
<td>$18.24$</td>
</tr>
<tr>
<td>$62.47$</td>
<td>$789.65$</td>
<td>$308.41$</td>
<td>$3.65$</td>
</tr>
<tr>
<td>$9$</td>
<td>$12.7$</td>
<td>$7.9$</td>
<td>$11. 9$</td>
</tr>
<tr>
<td>$10.4$</td>
<td>$9.6$</td>
<td>$14.2$</td>
<td>$7.9$</td>
</tr>
<tr>
<td>$d$</td>
<td>$w$</td>
<td>$dw$</td>
<td>$iFw$</td>
</tr>
<tr>
<td>$2.3$</td>
<td>$-1.7$</td>
<td>$-2.3$</td>
<td>$4.3$</td>
</tr>
<tr>
<td>$98.14$</td>
<td>$85.42$</td>
<td>$112.16$</td>
<td>$83.98$</td>
</tr>
<tr>
<td>$225.72$</td>
<td>$-145.21$</td>
<td>$-257.97$</td>
<td>$361.11$</td>
</tr>
<tr>
<td>$519.16$</td>
<td>$246.86$</td>
<td>$593.33$</td>
<td>$1522.79$</td>
</tr>
</tbody>
</table>

The $\chi^2$ with 10 degrees of freedom for testing sex differences in all the ten types is

$$4434.10 \div 820.7 = 5.40$$

This is slightly above the value 4.827 obtained earlier by a direct $\chi^2$ analysis. From the 10 degrees of freedom $\chi^2$ we subtract the $\chi^2$ with 1 degree due to overall sex difference. The residual

$$5.40 - 0.05 = 5.35$$

is the $\chi^2$ with 9 degrees of freedom for testing whether sex difference depends on the type of the heterozygote. This interaction $\chi^2$ is not significant, nor is that due to sex difference. In such a case, the differences in the various types of heterozygotes can be studied by summing over sex. To complete the analysis, however, we determine the $\chi^2$ due to differences in heterozygotes, eliminating sex difference.

The total $\chi^2$ with $(20 - 1) = 19$ degrees of freedom is

$$\left\{ \frac{\Sigma n \sigma^2 - (\Sigma n \sigma)^2}{\Sigma n} \right\} \div 820.7$$
the summation extending over the 20 angles. The following computations are made.

\[
\begin{array}{cccc}
\Sigma n & \Sigma n_0 & \Sigma n_0^2 & \chi^2 \\
\varphi & 1731 & 21,470.9 & 274,666.21 & 8,346.43 & 10.17 \\
\sigma & 1843 & 22,699.4 & 290,576.08 & 10,997.81 & 13.40 \\
Overall & 3574 & 44,170.3 & 556,242.29 & 19,351.02 & 23.58 \\
\end{array}
\]

The total of \( \chi^2 \) for \( \varphi \) and \( \sigma \)

\[10.17 + 13.40 = 23.57\]

has 18 degrees of freedom. Subtracting the interaction component of 5.35 on 9 degrees of freedom, the residual \( \chi^2 \) with 9 degrees of freedom for testing differences in heterozygotes eliminating sex is

\[23.57 - 5.35 = 18.22\]

which is significant. This can also be calculated in a slightly different way. The sex \( \chi^2 \), ignoring differences in heterozygotes, is obtained as follows.

\[
820.7 \chi^2 = \frac{21,470.9^2 + 22,699.4^2 - 44,170.3^2}{1731 + 1843 - 3574}
\]

\[\chi^2 = 0.01\]

the values being obtained from \( \Sigma n_0 \) and \( \Sigma n \) over all heterozygotes. The valid \( \chi^2 \) for heterozygotes is obtained by subtracting from the total the above value and the interaction sum of squares. This leads to the value, 23.58 - 0.01 - 5.35 = 18.22, the same as before.

| Table 5e.3b. Analysis of \( \chi^2 \) |
| Degrees of Freedom |
|-----|-----|
| Sex | 1 | 0.05 |
| Heterozygotes | 9 | 18.22 |
| Interaction | 9 | 5.35 |
| Total | 19 | 23.58 |

The total \( \chi^2 \) with 19 degrees of freedom is significant, showing overall differences. The various components of \( \chi^2 \) do not add up to the total because the proportions are based on different numbers. If the interaction component is high, then a precise test will be to calculate the
mean $\chi^2$ for heterozygotes and interaction

\[ \frac{18.22}{9} = 2.024 \quad \frac{5.35}{9} = 0.594 \]

and determine the variance ratio

\[ \frac{2.024}{0.594} = 3.407 \]

with 9 and 9 degrees of freedom. This is significant on the 5% level.

5e.4 Other Useful Transformations

The estimated variance

\[ s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \]

where $x_1, \ldots, x_n$ are $n$ random observations from a normal population, has the expected value $\sigma^2$ and variance $2\sigma^4/(n - 1)$. The transformation needed to make the variance independent of $\sigma$ is

\[ F(\sigma) = \int c \sqrt{n - 1} \frac{d\sigma^2}{\sqrt{2\pi}} = \log \sigma^2 \quad \text{choosing } c \text{ suitably} \]

\[ \text{a.v. (log } s^2) = \frac{2}{n - 1} \]

The estimated correlation from $n$ pairs of observations from a bivariate normal distribution has the asymptotic variance

\[ \frac{(1 - \rho^2)^2}{n - 1} \]

The necessary transformation is

\[ F(\rho) = \int c \sqrt{n - 1} \frac{d\rho}{1 - \rho^2} \]

\[ = \tanh^{-1} \rho \quad \text{choosing } c \text{ suitably} \]

\[ \text{a.v. (tanh}^{-1} \rho) = \frac{1}{n - 1} \]

The uses of these transformed statistics with slight refinements will be discussed in Chapter 6.
5f Large Sample Standard Errors of Moments

5f.1 Variances and Covariances of Raw Moment Statistics

Any population may be defined with respect to a number of frequency classes with probabilities \( \pi_1, \ldots, \pi_k \) and the corresponding values of a variate \( x_1, \ldots, x_k \). A continuous distribution can be considered to contain an infinite number of class intervals, each of length \( dx \), the differential element. In such a case the raw moments in the population are

\[ \mu_r = \pi_1 x_1^r + \cdots + \pi_k x_k^r \quad r = 1, 2, \ldots \]

If \( n_1, n_2, \ldots, n_k \), \( (\Sigma n_i = n) \), are the observed frequencies in the \( k \) classes, then the \( r \)th sample moment about the origin is

\[ O_r = \frac{n_1 x_1^r + \cdots + n_k x_k^r}{n} \]

which is a linear function of the frequencies. If \( O_s \) is the \( s \)th sample moment, then using the results (2a.9.1) we find

\[ E(O_r) = \Sigma \pi_i x_i^r = \mu_r \]

\[ V(O_r) = \frac{1}{n} \left[ \Sigma \pi_i x_i^{2r} - (\Sigma \pi_i x_i^r)^2 \right] = \frac{\mu_{2r} - \mu_r^2}{n} \]

\[ \text{cov} (O_r, O_s) = \frac{1}{n} \left[ \Sigma \pi_i x_i^r x_i^s - (\Sigma \pi_i x_i^r)(\Sigma \pi_i x_i^s) \right] = \frac{\mu_{r+s} - \mu_r \mu_s}{n} \]

If the origin is the population mean, then in the above expressions the moments \( \mu_r \) about an arbitrary origin will be replaced by \( \mu_r \), the population \( r \)th moment about the mean.

The \( r \)th moment about the sample mean is

\[ m_r = O_r - \binom{r}{1} O_{r-1} \mu_1 + \binom{r}{2} O_{r-2} \mu_2^2 - \cdots + (-1)^{r} \mu_1^r \]

Since this is invariant for origin, we can consider \( O_r \) to be the sample
raw moments about the population mean as the origin. We now use formula (5e.1.1) for determining the asymptotic variance (a.v.) of \( m_r \).

\[
\frac{\partial m_r}{\partial \mu_r} = 1, \quad \frac{\partial m_r}{\partial \mu_r - \mu} = (-1)^r \binom{r}{r} \mu_1^r
\]

\[
\frac{\partial m_r}{\partial \mu_1} = -r \mu_{r-1} + \binom{r}{2} \mu_{r-2} (2 \mu_1) - \cdots
\]

Using the approximation

\[
m_r - \mu_r \sim (O_r - \mu) - r \mu_{r-1} (O_1 - \mu_1)
\]

since all the other derivatives vanish at the expected values, we find

\[
a.v. (m_r) = E(m_r - \mu_r)^2 \sim V(O_1) - 2r \mu_{r-1} \text{cov} (O_r - \mu) (O_1 - \mu_1) + r^2 \mu_{r-1}^2 V(O_1)
\]

\[
= \frac{1}{n} [\mu_{r+1} - \mu_r^2 - 2r \mu_{r+1} \mu_{r-1} + r^2 \mu_{r-3}^2 - \mu_1]
\]

Similarly

a. \( \text{cov} (m_r, m_s) = \frac{1}{n} [\mu_{r+s} - \mu_r \mu_s + r \mu_{r+s-1} \mu_{r-1} - r \mu_{r-1} \mu_{s+1} - 4 \mu_{r+1} \mu_{s-1}] \)

As special cases we obtain the large sample variances:

a. v. (\( m_2 \)) = \( \frac{1}{n} (\mu_4 - \mu_2^2) \)

a. v. (\( m_3 \)) = \( \frac{1}{n} (\mu_6 - \mu_3^2 - 6 \mu_2 \mu_4 + 9 \mu_4^2) \)

a. v. (\( m_4 \)) = \( \frac{1}{n} (\mu_8 - \mu_4^2 - 8 \mu_3 \mu_5 + 16 \mu_2 \mu_4^2) \)

\( V(O_1, \text{about the origin}) = \frac{\mu_4}{n} \)

the last formula being exact.

Example. The asymptotic variance of the coefficient of variation

\( 100 \sqrt{\frac{m_1}{O_1}}, \) where \( O_1 \) is the average about the origin, is

\[
\eta^2 \left( \frac{\mu_4 - \mu_2^2}{4 \mu_2^2} + \frac{\mu_2}{\mu_1^2} - \frac{\mu_2}{\mu_2 \mu_1^2} \right)
\]

where \( \eta \) is the population value \( 100 \sqrt{\frac{\mu_4}{\mu_1}}. \)
5.2 Large Sample Tests of Difference between Means and
an Illustration of the \( P \) Test

Let \( x_1, x_2 \) be the mean values and \( s_1^2, s_2^2 \) the estimated variances in
two samples of sizes \( n_1 \) and \( n_2 \), respectively. The standard errors of
\( x_1 \) and \( x_2 \) are \( \sqrt{s_1^2/n_1} \) and \( \sqrt{s_2^2/n_2} \), and hence that of \( x_1 - x_2 \) is
\( \sqrt{s_1^2/n_1 + s_2^2/n_2} \). If \( n_1 \) and \( n_2 \) are large, the statistic
\[
  w = \frac{x_1 - x_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}
\]
can be used as a normal deviate with zero mean and unit standard
deviation to test the hypothesis that the samples are drawn from popu­
lations having the same mean, nothing being specified about the varia­
tions.

It is not necessary that the original populations be normal, provided
that the samples are large enough. How large the sample should be
depends on the nature of the populations. Populations with highly
skew or multimodal distributions require very large samples.

In a feeding experiment with pasteurized and unpasteurized milk,
Elderton (1933) found the following values of \( w \) and the associated
normal probabilities. The samples were large so that the difference in
mean statures divided by the standard error of difference could be used
as a normal deviate.

<table>
<thead>
<tr>
<th>Age Group of Boys</th>
<th>Observed ( w = w_o )</th>
<th>Probability ( P(w \geq w_o) )</th>
<th>( \log_10 P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6%</td>
<td>2.69</td>
<td>0.0035726</td>
<td>3.5529644</td>
</tr>
<tr>
<td>7%</td>
<td>-0.71</td>
<td>0.7611479</td>
<td>1.8814690</td>
</tr>
<tr>
<td>8%</td>
<td>1.24</td>
<td>0.1074877</td>
<td>0.0335888</td>
</tr>
<tr>
<td>9%</td>
<td>1.04</td>
<td>0.823841</td>
<td>0.8706849</td>
</tr>
<tr>
<td>10%</td>
<td>1.06</td>
<td>0.1445723</td>
<td>1.1600851</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5.1428922</td>
<td><strong>= -5.8371478</strong></td>
<td></td>
</tr>
</tbody>
</table>

It is found that the probabilities are less than 5% only in two cases.
Should we then say that pasteurized milk is beneficial only for the age
groups 6\% and 9\% and not for the other age groups? The positive
deviations (of \( w \)) in 4 out of 5 cases indicate a taller stature for boys fed
on pasteurized milk, and the nonsignificance, in some cases, may be due
to the inadequacy of the numbers or to the presence of high variance.
in the stature of boys chosen for the experiment. On the other hand, it may be argued that 2 significant cases out of 5 could arise by chance, even when pasteurized milk does not affect the stature of boys. In such cases the evidences supplied by the various groups have to be combined to answer the problem of differences caused by feed. This can be done by calculating the statistic

\[ P_A = -2 \sum \log P_i \]

\[ = -2 \log 10 \sum \log 10 P_i \]

\[ = -2(2.3026)(-5.85712) = 26.9732 \]

which is distributed as \( \chi^2 \) with 5 \( \times \) 2 degrees of freedom, as shown in 2a.8. In the above problem \( \chi^2 = 26.9732 \), which as \( \chi^2 \) with 10 degrees of freedom is significant on the 1% level. This shows that pasteurized milk causes some difference in stature in general. It is difficult to say from these data alone that in any particular age group pasteurized milk has no effect.

6f.3 Tests of Normality

Biometric measurements relating to homogeneous groups usually have unimodal distributions in which the asymmetry is small and the kurtosis is approximately the same as that of normal distribution. Asymmetry is measured by \( \beta_1 \), where

\[ \beta_1 = \frac{m_3^2}{m_2^3} \]

and \( m_2 \) and \( m_3 \) being the second and third sample moments. Deviations from normal kurtosis are measured by \( \beta_2 \), where

\[ \beta_2 = \frac{m_4}{m_2^2} - 3 \]

\( m_4 \) being the fourth moment. Deviations from these normal values, when significant, are important features of distributions of biometric measurements.

If the observed frequency distribution is given, we can fit a normal curve to it and test for the goodness of fit. This is useful in testing for all departures of the observed frequency distribution from the expected on the normal basis, but it is quite insensitive in testing for some specific aspects of the distribution, such as symmetry and kurtosis. If deviations other than those in symmetry and normal kurtosis are
not important, then in any problem the observed \(\hat{b}_1\) and \(\hat{b}_2\) can be tested for the expected values 0 and 3. If \(n\) is the sample size, the asymmetry can be tested by using the statistic

\[ w_1 = \pm \sqrt{\frac{\hat{b}_1(n + 1)(n + 3)}{6(n - 2)}} \]  

(5f.3.1)

as a normal deviate with zero mean and unit variance when \(n\) is large.

The sign in the above statistic is the same as that of \(m_3\), the third moment. In the same way, the probability of departure from normality in kurtosis can be tested by using the normal deviate

\[ w_2 = \left( \frac{\hat{b}_2 - 3}{n + 1} \right) \sqrt{\frac{(n + 1)^2(n + 3)(n + 5)}{24n(n - 2)(n - 3)}} \]  

(5f.3.2)

The following table gives for 23 samples the values of \(n\), \(\sqrt{\hat{b}_1}\), and \(P(\sqrt{\hat{b}_1})\), the probability of \(\hat{b}_1\)'s being smaller than the observed on the normal hypothesis.

<table>
<thead>
<tr>
<th>Sample</th>
<th>(n)</th>
<th>(\sqrt{\hat{b}_1})</th>
<th>(P(\sqrt{\hat{b}_1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86</td>
<td>0.1552</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
<td>0.1025</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>107</td>
<td>-0.5442</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>139</td>
<td>-0.1010</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>168</td>
<td>-0.3829</td>
<td>0.11</td>
</tr>
<tr>
<td>6</td>
<td>150</td>
<td>0.0100</td>
<td>0.48</td>
</tr>
<tr>
<td>7</td>
<td>124</td>
<td>-0.2000</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>187</td>
<td>-0.0100</td>
<td>0.71</td>
</tr>
<tr>
<td>9</td>
<td>113</td>
<td>0.0000</td>
<td>0.50</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>0.4118</td>
<td>0.92</td>
</tr>
<tr>
<td>11</td>
<td>94</td>
<td>-0.0770</td>
<td>0.35</td>
</tr>
<tr>
<td>12</td>
<td>173</td>
<td>0.0173</td>
<td>0.54</td>
</tr>
</tbody>
</table>

There are about 5 values (marked with an asterisk in the table above) significant on the 5% level against the expected 2% (\(= 1.15\)). To test for overall significance the \(P_o\) test explained in 5f.2 may be carried out. The value of \(P_o\) is 69.4880, which as \(\chi^2\) with 2 \(\times 23 = 46\) degrees of freedom is significant, thus indicating skewness of the nasal height distributions. On the whole the distribution of nasal height is negatively skew.

The values of \(\hat{b}_2\) can be treated in an exactly similar way, using the formula (5f.3.2).
References


CHAPTER 6

Tests of Homogeneity of Variances and Correlations

6a  Homogeneity of Variances

6a.1  Test for a Specified Variance

It is sometimes necessary to test whether an estimated variance is in agreement with a specified hypothetical variance. If $\sigma^2$ is the estimate based on $n$ degrees of freedom,* of the hypothetical variance $\sigma'^2$, then the statistic

$$\chi^2 = \frac{ns^2}{\sigma'^2}$$

can be used as $\chi^2$ with $n$ degrees of freedom to test the above hypothesis.

The variance for head breadth calculated from measurements on 29 crania from Jebel Moya (Sudan) is 48.5632. Could this sample have arisen from a homogeneous cranial population with a head breadth variance of 18.2313? The degrees of freedom in this case are 28, one less than the number of observations, and

$$\chi^2 = \frac{ns^2}{\sigma'^2} = 77.2485$$

The object of inquiry is whether the Jebel Moya population belongs to a homogeneous group. Heterogeneity would increase the internal variance so that high values of $\chi^2$ would indicate significance. The observed value exceeds the 5% value of $\chi^2$ with 28 degrees of freedom so that the hypothesis is rejected. The hypothetical value considered in this case

* The estimate $s^2$ is obtained by dividing the corrected sum of squares by the degrees of freedom equal to one less than the sample size if only a single sample is available.
is the variance for head breadth derived from a large series of crania from Egypt and Sudan. The higher variance in the present case suggests that the cranial population of Jebel Moya is heterogeneous.

We could also test by considering the lower tail of the $\chi^2$ distribution whether an observed variance is significantly less than the hypothetical variance. Thus, for a hypothetical variance of 90, the $\chi^2$ in the above problem would be 15.6481, the probability of exceeding which is more than 95%. The probability of obtaining a $\chi^2$ less than the observed is therefore less than 5%, so that the observed variance is significantly less than the assigned one.

Suppose that it is not known that the specified value 18.2313 of the variance is for a homogeneous population, and it is desired to know whether the observed variance could have reasonably arisen from a population with a hypothetical variance of 18.2313. In a problem of this nature we are interested in both small and high values of $\chi^2$ that might arise, both disproving the null hypothesis. The test procedure in this case is slightly complicated, and we may consider two situations, depending on the sample size.

(i) When the sample size is large (greater than 30): When the sample size is large, $\chi^2$ tends to be normally distributed with degrees of freedom $n$ as mean and variance $2n$ so that the test reduces to

$$A_1\quad \left| \frac{\chi^2 - n}{\sqrt{2n}} \right| \geq \lambda \quad \text{as normal deviate, considering both tails.}$$

where $\lambda$ is the 5% or 1% value of the normal deviate, considering both tails. The approach to normality of $\chi^2$ is slow so that the above approximation may not hold good in moderately large samples. A better approximation is to use $\sqrt{2\chi^2}$ as a normal variate with mean $\sqrt{2n - 1}$ and unit variance, in which case the test is

$$A_2\quad | \sqrt{2\chi^2} - \sqrt{2n - 1} | \geq \lambda$$

A third and a fairly accurate approximation is to use $(\chi^2/n)^{0.5}$ as a normal variate with mean $(1 - 2/9n)$ and variance $2/9n$ so that the test is

$$A_3\quad \left| \sqrt{\frac{2n}{9}} \left( \frac{\chi^2}{n} \right)^{0.5} + \frac{2}{9n} - 1 \right| \geq \lambda$$
### Table 6a.1a. The Admissible Range of the $\chi^2$ Distribution

<table>
<thead>
<tr>
<th>D.F.</th>
<th>$x^2_{0.05}$</th>
<th>$x^2_{0.01}$</th>
<th>$x^2_{0.05}$</th>
<th>$x^2_{0.01}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.043163</td>
<td>7.8155</td>
<td>0.091341</td>
<td>11.3458</td>
</tr>
<tr>
<td>2</td>
<td>0.084260</td>
<td>9.2822</td>
<td>0.1746</td>
<td>13.2866</td>
</tr>
<tr>
<td>3</td>
<td>0.2061</td>
<td>11.10930</td>
<td>0.1011</td>
<td>15.1251</td>
</tr>
<tr>
<td>4</td>
<td>0.6972</td>
<td>12.8008</td>
<td>0.2609</td>
<td>16.9004</td>
</tr>
<tr>
<td>5</td>
<td>0.9889</td>
<td>14.3709</td>
<td>0.4466</td>
<td>18.6480</td>
</tr>
<tr>
<td>6</td>
<td>1.4250</td>
<td>15.8684</td>
<td>0.7854</td>
<td>20.2980</td>
</tr>
<tr>
<td>7</td>
<td>1.9026</td>
<td>17.9222</td>
<td>1.1214</td>
<td>21.0306</td>
</tr>
<tr>
<td>8</td>
<td>2.4136</td>
<td>18.8616</td>
<td>1.4984</td>
<td>23.3040</td>
</tr>
<tr>
<td>9</td>
<td>2.9299</td>
<td>20.3008</td>
<td>1.9017</td>
<td>25.1352</td>
</tr>
<tr>
<td>10</td>
<td>3.5160</td>
<td>21.7290</td>
<td>2.3450</td>
<td>26.6200</td>
</tr>
<tr>
<td>11</td>
<td>4.0907</td>
<td>23.1341</td>
<td>2.8061</td>
<td>28.1830</td>
</tr>
<tr>
<td>12</td>
<td>4.7004</td>
<td>24.5244</td>
<td>3.2916</td>
<td>29.6900</td>
</tr>
<tr>
<td>13</td>
<td>5.3170</td>
<td>25.9012</td>
<td>3.7960</td>
<td>31.1662</td>
</tr>
<tr>
<td>14</td>
<td>5.9472</td>
<td>27.2580</td>
<td>4.3162</td>
<td>32.6140</td>
</tr>
<tr>
<td>15</td>
<td>6.5910</td>
<td>28.6140</td>
<td>4.8325</td>
<td>34.0905</td>
</tr>
<tr>
<td>16</td>
<td>7.2448</td>
<td>29.9552</td>
<td>5.4048</td>
<td>35.5376</td>
</tr>
<tr>
<td>17</td>
<td>7.9001</td>
<td>31.2851</td>
<td>5.9670</td>
<td>36.9760</td>
</tr>
<tr>
<td>18</td>
<td>8.5842</td>
<td>32.6070</td>
<td>6.5448</td>
<td>38.3886</td>
</tr>
<tr>
<td>19</td>
<td>9.2952</td>
<td>33.9188</td>
<td>7.1307</td>
<td>39.8012</td>
</tr>
<tr>
<td>20</td>
<td>9.9080</td>
<td>35.2200</td>
<td>7.7300</td>
<td>41.3149</td>
</tr>
<tr>
<td>21</td>
<td>10.5554</td>
<td>36.5274</td>
<td>8.3340</td>
<td>42.8880</td>
</tr>
<tr>
<td>22</td>
<td>11.1001</td>
<td>37.8281</td>
<td>8.9518</td>
<td>43.0670</td>
</tr>
<tr>
<td>23</td>
<td>11.6538</td>
<td>39.1046</td>
<td>9.5772</td>
<td>45.3533</td>
</tr>
<tr>
<td>24</td>
<td>12.2072</td>
<td>40.3872</td>
<td>10.2072</td>
<td>46.7084</td>
</tr>
<tr>
<td>25</td>
<td>12.7886</td>
<td>41.6575</td>
<td>10.8475</td>
<td>48.0600</td>
</tr>
<tr>
<td>26</td>
<td>13.1500</td>
<td>42.9286</td>
<td>11.4916</td>
<td>49.4104</td>
</tr>
<tr>
<td>27</td>
<td>14.0769</td>
<td>44.1900</td>
<td>12.1446</td>
<td>50.7600</td>
</tr>
<tr>
<td>28</td>
<td>15.7136</td>
<td>45.4552</td>
<td>12.8044</td>
<td>52.0688</td>
</tr>
<tr>
<td>29</td>
<td>16.4575</td>
<td>46.7103</td>
<td>13.4676</td>
<td>53.4334</td>
</tr>
<tr>
<td>30</td>
<td>17.2050</td>
<td>47.9610</td>
<td>14.1360</td>
<td>54.7680</td>
</tr>
</tbody>
</table>

Consider an example where $x^2 = 43.773$ and $n = 30$. The normal deviates corresponding to $A_1$, $A_2$, and $A_3$ are

\[
A_1 = \frac{43.773 - 30}{\sqrt{60}} = 1.7781 \quad \text{Probability} = 0.075
\]

\[
A_2 = \sqrt{87.540 - \sqrt{59}} = 1.6755 \quad \text{Probability} = 0.095
\]

\[
A_3 = \sqrt{\left(\frac{43.773}{30}\right)^2 + \frac{2}{9 \times 30} - 1} = 1.6452 \quad \text{Probability} = 0.100
\]
The most accurate of the three approximations is the last one; in fact it gives almost an exact value to the probability of $\chi^2$'s exceeding 43.773, which is 0.05 as seen from $\chi^2$ tables. The probability 0.100 corresponding to $A_3$ is very high so that the value of $\chi^2 = 43.773$ does not give sufficient indication as to whether the observed variance differs from the assigned value.

(ii) When the sample size is small (less than 30): If the observed $\chi^2$ is above the lower 5% value or below the upper 5% value, no further analysis is needed; the hypothesis cannot be rejected. The doubt arises only when $\chi^2$ is beyond these limits, in which case the limits to non-significant values of $\chi^2$ have to be determined.

Table 6a.1a gives the admissible range of $\chi^2$ as determined by the locally most powerful unbiased test (see 8a.4) of Neyman and Pearson (1939). The value of $\chi^2 = 77.2485$ in the case of the Jebel Moya population lies outside the 5% admissible range (15.7136, 45.4552) for 28 degrees of freedom so that the null hypothesis is rejected.

6a.2 Test for a Specified Inequality of Two Estimated Variances

The statistical analysis of data often leads to a number of estimated variances of which it is desirable to test the homogeneity. This problem is considerably simple when there are only two estimated variances and it is desired to test whether a particular estimate significantly exceeds the other. Thus one might have the estimates of the head length variances $s_1^2 = 42.302$ and $s_2^2 = 34.658$ based on $n_1 = 24$ and $n_2 = 30$ degrees of freedom for males and females. The important question to be asked, in this connection, is whether the male head length is, as commonly believed, more variable than the female head length. The statistic to be constructed for this purpose is

$$F = \frac{s_2^2}{s_1^2} = 1.2205$$

which can be entered in the variance ratio table with $n_1 = 24$ and $n_2 = 30$ degrees of freedom. The 5% value of $F$ is 1.89, which is greater than the observed $F$, so that there is no evidence against the equality of variances in the two sexes.

It may be noted that this does not prove that the male and female head lengths are equally variable. The evidence supplied by the above data may not be sufficient to detect the difference, if any. The ratio such as the observed could be expected not infrequently in samples of the above sizes when, in fact, the variances are equal. To detect the difference as indicated by the above ratio a very large number of degrees of freedom, i.e., a large sample, would be necessary.
SPECIFIED INEQUALITY OF TWO ESTIMATED VARIANCES

If sample sizes are large, the test for equality of standard deviations can be carried out in a simple way. If any estimated variance $\hat{\sigma}^2$ based on $n$ degrees of freedom is transformed by the relation (see 5e.4)

$$y = \log \sigma$$

then $y$ tends to be normally distributed with

$$\text{Mean} = \log \sigma \quad \text{and} \quad \text{Variance} = \frac{1}{2n}$$

so that the variance is independent of $\sigma$. Kemsley (1950) found the following standard deviations for heights of males and females.

<table>
<thead>
<tr>
<th>Age</th>
<th>Standard Deviation</th>
<th>Sample Size</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>$\varphi$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>14</td>
<td>3.66</td>
<td>2.68</td>
<td>305</td>
</tr>
<tr>
<td>14.5</td>
<td>3.71</td>
<td>2.40</td>
<td>892</td>
</tr>
<tr>
<td>15.5</td>
<td>3.57</td>
<td>2.50</td>
<td>1404</td>
</tr>
<tr>
<td>16.5</td>
<td>3.07</td>
<td>2.35</td>
<td>1644</td>
</tr>
<tr>
<td>17.5</td>
<td>2.95</td>
<td>2.50</td>
<td>1280</td>
</tr>
<tr>
<td>18.5</td>
<td>2.86</td>
<td>2.23</td>
<td>786</td>
</tr>
<tr>
<td>19.5</td>
<td>2.87</td>
<td>2.34</td>
<td>591</td>
</tr>
<tr>
<td>20.5</td>
<td>2.70</td>
<td>2.31</td>
<td>1487</td>
</tr>
<tr>
<td>21.5</td>
<td>2.65</td>
<td>2.34</td>
<td>459</td>
</tr>
<tr>
<td>22.5</td>
<td>2.62</td>
<td>2.37</td>
<td>477</td>
</tr>
</tbody>
</table>

The standard error of the difference is $\sqrt{0.0003307} = 0.05751$, and the ratio of the difference to standard error is $w = 17.30$, which as a normal deviate has a very small probability, thus showing that the variabilities are different for $\sigma^2$ and $\varphi$, females being less variable than males. This is in accordance with an observed fact in biological material that the standard deviation depends on the mean size of organisms: the higher the mean, the greater is the scatter. The female dimensions are smaller than the male, and this is reflected in the standard deviation. We can make a closer examination by considering the data for various ages since the overall standard deviation may be affected by the age composition of the samples for males and females.

*The ages in the original table extend up to 74.5, and only the first 10 have been chosen to illustrate the test.
It is found that at each age the statistic $w$ exceeds the upper 5% value of the normal deviate, showing that female variability is less at each age. If samples were not so numerous as those above, it would be difficult to detect the difference in variabilities at each age. In such a situation the normal probabilities for exceeding $w$ could be calculated at each age and the $F$ test (5.2) carried out to combine the evidences supplied by all age groups.

6a.3 The Likelihood Criterion and Its Use

With two estimated variances a situation different from the one considered in 6a.2 may arise. The estimated variances of nasal indices may be available for two series of male skulls. The question to be asked is whether the variances in the cranial populations from which the two series are samples can be considered equal. In the absence of any knowledge as to the possible inequality relationship between the two population variances, the following statistic

$$L = \frac{(n_1 + n_2)^{1/(n_1 + n_2)}}{n_1 s_1^2 + n_2 s_2^2}$$

which is the ratio of the weighted geometric mean to the arithmetic mean of the estimates, may be constructed. This lies between 0 and 1, the value 0 being reached when the ratio of one estimate to the other is large, and 1 when the two are equal. A small value of $L$ would thus indicate a difference in the population variances.

Instead of $L$ it is convenient, from the point of view of computations, to consider the statistic (due to Bartlett, 1934)

$$M = -n \log_e L = n \log_e \left( \frac{n_1 s_1^2 + n_2 s_2^2}{n} \right) - n_1 \log_e s_1^2 - n_2 \log_e s_2^2$$

where $n = n_1 + n_2$. $M$ varies from 0 to $\infty$, with small values of $L$ corresponding to high values of $M$.

This can be extended to the case where $k$ estimated variances have to be tested for homogeneity. If $s_1^2, \cdots, s_k^2$ are the estimated variances with $n_1, \cdots, n_k$ degrees of freedom, then the statistic to be constructed is

$$M = n \log_e \left( \frac{n_1 s_1^2 + \cdots + n_k s_k^2}{n} \right) - n_1 \log_e s_1^2 - \cdots - n_k \log_e s_k^2$$

where $n = n_1 + \cdots + n_k$. The distribution of $M$ is approximately chi-squared with $k-1$ degrees of freedom.
The probability of \( M \)'s exceeding the observed value \( M_o \) can be expanded in an asymptotic series, the first six terms of which are given here.

\[
\frac{1}{6} \sum_{i=0}^{6} \beta_i P_{k-1+2i}(M_o)
\]

where \( P_{k-1+2i}(M_o) \) stands for the probability \(^*\) of the \( \chi^2 \) with \((k - 1 + 2i)\) degrees of freedom exceeding \( M_o \). The values of \( \beta \) are

\[
\begin{align*}
\beta_0 &= 1 \\
\beta_1 &= \frac{1}{3} c_1 \\
\beta_2 &= \frac{1}{3} \beta_1^2 \\
\beta_3 &= -\frac{1}{3} c_3 + \frac{1}{3} \beta_1^3 \\
\beta_4 &= \beta_1 \beta_3 - \frac{1}{3} \beta_1 \beta_2^2 \\
\beta_5 &= \frac{\gamma_2 \gamma_3 - \gamma_2 \gamma_4 \beta_2 - \frac{1}{2} \beta_1 \gamma_3}{3} \\
\beta_6 &= \frac{\gamma_2 \beta_5 \beta_2 - \frac{1}{2} \beta_1 \gamma_2 \beta_3 - \frac{1}{2} \beta_1 \beta_2}{3}
\end{align*}
\]

6a.4 Practical Applications

The exact formula derived above for evaluating the probability of exceeding the observed value \( M_o \) need not always be used in practice. The approximations which can be profitably used in many situations are given below with suitable illustrations.

First Approximation. If the observed value \( M_o \) is less than the 5% (or 1%) value of \( \chi^2 \) with \((k - 1)\) degrees of freedom, then the hypothesis of equality of variances cannot be rejected on the 5% (or 1%) level. This is due to the fact that the exact 5% and 1% limits of \( M \) are beyond the corresponding values of the \( \chi^2 \) limits.

Table 6a.4a gives a set of 10 estimates of variance, calculated from 10 samples of weight records of schoolboys of similar age but from different forms. It is desired to test whether there are any real "form differences" in the weight dispersion of the boys.

\(^*\) These values can be obtained from Tables of the Incomplete \( t \)-Function, edited by K. Pearson.
TABLE 6a. The Estimated Variances and Evaluation of the $M$-Statistic
(Hartley and Pearson, 1946)

<table>
<thead>
<tr>
<th>Form No.</th>
<th>No. of Boys</th>
<th>Variance</th>
<th>D.F. = $n_k$</th>
<th>$\log s_k^2$</th>
<th>$n_k \log s_k^2$</th>
<th>$n_k s_k^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>51</td>
<td>9</td>
<td>3.03</td>
<td>35.4</td>
<td>459</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>78</td>
<td>14</td>
<td>4.36</td>
<td>61.0</td>
<td>1092</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>91</td>
<td>20</td>
<td>4.51</td>
<td>90.2</td>
<td>1820</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>52</td>
<td>22</td>
<td>3.35</td>
<td>88.9</td>
<td>1144</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>101</td>
<td>14</td>
<td>4.62</td>
<td>64.7</td>
<td>1414</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>36</td>
<td>10</td>
<td>3.58</td>
<td>35.8</td>
<td>360</td>
</tr>
<tr>
<td>7</td>
<td>31</td>
<td>41</td>
<td>30</td>
<td>2.71</td>
<td>111.3</td>
<td>1230</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>76</td>
<td>14</td>
<td>4.33</td>
<td>60.6</td>
<td>1064</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>64</td>
<td>2</td>
<td>4.16</td>
<td>8.3</td>
<td>128</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>52</td>
<td>5</td>
<td>4.53</td>
<td>22.6</td>
<td>465</td>
</tr>
<tr>
<td>Total</td>
<td>140</td>
<td></td>
<td>140 = $n$</td>
<td></td>
<td>576.8</td>
<td>9176</td>
</tr>
</tbody>
</table>

\[ M = n \log \left( \frac{\sum n_k s_k^2}{n} \right) - \sum n_k \log s_k^2 \]
\[ = 140 \log \left( \frac{9176}{140} \right) - 576.8 \]
\[ = 140 \times 4.183 - 576.8 = 8.8 \]

The 5% value of $x^2$ with $(10 - 1) = 9$ degrees of freedom is 15.51 so that the observed value 8.8, being less than this, cannot be considered significant. No further calculations are needed in this case. The best estimate of the common variance is \( \frac{\sum n_k s_k^2}{n} = \frac{9176}{140} = 65.54 \).

Second Approximation. Bartlett (1934) suggested the use of

\[ M' = \frac{M}{1 + c_1(k - 1)} \]

as $x^2$ with $(k - 1)$ degrees of freedom, where $c_1 = \sum (1/n_k) - (1/n)$. This approximation tends to increase the probability for an observed $M'$ so that if $M'$ is significant on any desired level then $M$ is certainly so.

The following data give a number of variances estimated on different degrees of freedom. These have been calculated from the yields of rice observed in successive blocks (columns) when a rectangular lattice is superimposed on a big field. It is desired to test whether the block variance is independent of its size, i.e., the number of cells it contains.
PROBLEMS REQUIRING AN EXACT TREATMENT

TABLE 6a.4. Variances for Different Block Sizes

<table>
<thead>
<tr>
<th>No. of Cells in a Block</th>
<th>Mean Variance for Blocks of the Same Size</th>
<th>D.F.</th>
<th>$\log_e n$</th>
<th>$n t \log_e n$</th>
<th>$n t^2$</th>
<th>$1/n_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>48.76</td>
<td>26</td>
<td>3.8870</td>
<td>101.9660</td>
<td>1.26776</td>
<td>0.038461</td>
</tr>
<tr>
<td>23</td>
<td>101.97</td>
<td>44</td>
<td>4.6250</td>
<td>203.5000</td>
<td>4.48668</td>
<td>0.022727</td>
</tr>
<tr>
<td>35</td>
<td>122.67</td>
<td>170</td>
<td>4.8096</td>
<td>817.6320</td>
<td>20.85330</td>
<td>0.005582</td>
</tr>
<tr>
<td>45</td>
<td>94.39</td>
<td>264</td>
<td>4.5474</td>
<td>1200.5136</td>
<td>24.91896</td>
<td>0.003788</td>
</tr>
<tr>
<td>51</td>
<td>78.40</td>
<td>200</td>
<td>4.3618</td>
<td>872.3600</td>
<td>15.08000</td>
<td>0.006000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>704</td>
<td></td>
<td>3195.0676</td>
<td>67.20730</td>
<td>0.075858</td>
</tr>
</tbody>
</table>

\[ M = 704 \log_e \left( \frac{67.20730}{704} \right) - 3195.0676 \]

\[ = 704(4.5589) - 3195.0676 = 14.3276 \]

This value is greater than the 1% value of $\chi^2$ with 4 degrees of freedom.

In this case the alternative statistic is

\[ M' = \frac{M}{1 + c_1/(k - 1)} \]

where $c_1 = 0.075858 - 0.001420 = 0.074438$.

\[ M' = \frac{14.3276}{1 + 0.074438/12} = 14.2393 \]

This also exceeds the 1% value of $\chi^2$ with 4 degrees of freedom so that the variances cannot be considered equal.

This shows that the variance is a function of the block size. Contrary to expectation, this is not an increasing function of the block size. The decrease in variance after a certain size of the block may be due to some periodicity in the fertility gradient.

6a.5 Problems Requiring an Exact Treatment

It is seen in 6a.4 that in any practical situation a decision can be reached if $M \leq 5\%$ value of $\chi^2$ with $(k - 1)$ degrees of freedom (not significant) and $M' = M/(1 + c_1/(k - 1)) \geq 5\%$ value of $\chi^2$ with $(k - 1)$ degrees of freedom (significant). There may arise cases where $M' < 5\%$ value of $\chi^2 < M$. The formula given in 6a.3 then has to be
used. In practice the situation presented above will occur only when the \( M \) statistic is just near its significance limit. The decision as to the acceptance or rejection of the hypothesis depends on the further use of these estimated variances in statistical analysis. This being so, the evaluation of the probability to a high degree of accuracy rarely will be necessary in practical problems.

It may be noted that all the tests based on the \( F \) and \( L \) statistics given in 6a.2 and 6a.3 can be extended to test for an assigned ratio \( \rho_1 : \rho_2 : \cdots : \rho_k \) of the hypothetical variances. The only modification needed is to replace \( s_i^2 \) by \( s_i^2 / \rho_i \).

6b Homogeneity of Correlations

6b.1 Exact Test for Zero Correlation

It has been shown (example 2, in 2d.2) that, when the correlation in a bivariate population is zero, the statistic

\[
t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}
\]

where \( r \) is the correlation coefficient calculated on a sample of size \( n \), is distributed as

\[
\frac{1}{\sqrt{n-2} \left( \frac{n-2}{2} - \frac{1}{2} \left( 1 + \frac{r^2}{n-2} \right)^{(n-1)/2} \right)^{0.05}} \sim t
\]

This is the same as the \( t \) distribution with \((n - 2)\) degrees of freedom. To test whether the population correlation coefficient is zero, the statistic \( t \) defined above is calculated and tested for significance by the use of the \( t \) table.

The correlation between frontal breadth and head breadth calculated from 18 crania of a series is 0.6521. Are the two dimensions, frontal breadth and head breadth, uncorrelated? The value of \( t \) is

\[
t = \frac{0.6521}{\sqrt{1 - (0.6521)^2}} \sqrt{16} = 4(0.8602)
\]

\[
= 3.4408
\]

with \( 18 - 2 = 16 \) degrees of freedom. This exceeds 2.120, the 5% value of \( t \), so that the observed correlation can be interpreted as establishing an association between frontal breadth and head breadth.
6b.2 Fisher’s tanh⁻¹ Transformation

It was shown in 5c.4 that the tanh⁻¹ transformation of the correlation coefficient \( r \) gets rid of the unknown parameter \( p \) in the expression for variance. Accordingly we consider the transformed values \( \xi \) and \( z \) instead of \( p \) and \( r \).

\[
\xi = F(p) = \frac{1}{2} \log \frac{1 + p}{1 - p} = \tanh^{-1} p
\]

\[
z = F(r) = \frac{1}{2} \log \frac{1 + r}{1 - r} = \tanh^{-1} r
\]

Putting \( z - \xi = x \), the distribution of \( x \) may be derived from the distribution of \( r \). The first four moments of \( z \) were found by R. A. Fisher and later revised by A. K. Gayen.

\[
\mu_1' = \frac{\rho}{2(n - 1)} \left\{ 1 + \frac{5 + \rho^2}{4(n - 1)} + \cdots \right\}
\]

\[
\mu_2 = \frac{1}{n - 1} \left\{ 1 + \frac{4 - \rho^2}{2(n - 1)} + \frac{22 - 6\rho^2 - 3\rho^4}{6(n - 1)^2} + \cdots \right\}
\]

\[
\mu_3 = \frac{\rho^3}{(n - 1)^3} + \cdots
\]

\[
\mu_4 = \frac{1}{(n - 1)^2} \left\{ 3 + \frac{14 - 3\rho^2}{n - 1} + \frac{181 - 48\rho^2 - 21\rho^4}{4(n - 1)^2} + \cdots \right\}
\]

\[
\beta_1 = \frac{\rho^5}{(n - 1)^5} + \cdots
\]

\[
\beta_2 = 3 + \frac{2}{n - 1} + \frac{4 + 2\rho^2 - 3\rho^4}{(n - 1)^2} + \cdots
\]

Since \( \beta_1 \) and \( (\beta_2 - 3) \) are small, even for moderate \( n \), it follows that \( (x - \xi) \) can be considered to be approximately a normal variate with

Mean = \( \frac{\rho}{2(n - 1)} \)

Variance = \( \frac{1}{n - 1} + \frac{4 - \rho^2}{2(n - 1)^2} = \frac{1}{n - 3} \)
6b.3 Test for a Given $p$

In a sample of 28 the correlation coefficient is found to be 0.6521. Can such a value have arisen from a population in which the coefficient has the value 0.7211?

$$z = \frac{1}{2} \log \frac{1 + r}{1 - r} = 0.7790$$

$$\text{Mean } z = \frac{1}{2} \log \frac{1 + p}{1 - p} + \frac{p}{2(n - 1)}$$

$$= 0.9100 + \frac{0.7211}{54} = 0.9233$$

The normal deviate is

$$\sqrt{n - 3} (z - \text{mean } z) = \sqrt{28 - 3} (0.7790 - 0.9233)$$

$$= 5(-0.1443) = -0.7215$$

The chance of exceeding the value 0.7215 in either direction is about 35% so that the hypothesis cannot be rejected.

The correction term $p/2(n - 1)$ for the mean $z$ is unimportant if $n$ is large. The probability will be more precisely obtained in any case by its inclusion.

6b.4 Test for the Equality of Two Correlation Coefficients

Two samples consisting of $n_1$ and $n_2$ observations give the correlation coefficients $r_1$ and $r_2$. Are these values compatible with the hypothesis that the samples arose from two populations having the same correlation coefficient? Let

$$z_1 = \frac{1}{2} \log \frac{1 + r_1}{1 - r_1}$$

and

$$z_2 = \frac{1}{2} \log \frac{1 + r_2}{1 - r_2}$$

The statistic $z_1 - z_2$ is distributed about the mean

$$\frac{\rho}{2(n_1 - 1)} = \frac{\rho}{2(n_2 - 1)}$$

(6b.4.1)

* These values can be directly obtained from the Fisher-Yates tables (transformation of $r$ to $z$).
where $\rho$ is the common correlation coefficient, with variance

$$\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

If the samples are not small or if $n_1$ and $n_2$ are not very different, the statistic

$$\frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}$$

can be used as a normal deviate. The more exact method given in 6b.6 may be necessary when the value of (6b.4.1) is not small.

**6b.5 Test for the Homogeneity of a Set of Correlation Coefficients**

Let $r_1, \cdots, r_k$ be $k$ correlation coefficients based on samples of sizes $n_1, \cdots, n_k$. By means of the tanh$^{-1}$ transformation, the quantities $z_1, \cdots, z_k$ corresponding to $r_1, \cdots, r_k$ can be obtained. If the bias in mean $z$ can be neglected, the test for homogeneity of the correlation coefficients is equivalent to the test of equality of the mean values of $z$.

The scheme of computation is as follows.

**Table 6b.5a. Test of Homogeneity of a Set of Correlation Coefficients**

<table>
<thead>
<tr>
<th>Sample No.</th>
<th>Sample Size</th>
<th>Correlation Coefficient</th>
<th>$r = z$</th>
<th>tanh$^{-1}$ of Variance $n - 3$</th>
<th>$(n - 3)z$</th>
<th>$(n - 3)z^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_1$</td>
<td>$r_1$</td>
<td>$z_1$</td>
<td>$n_1 - 3$</td>
<td>$(n_1 - 3)z$</td>
<td>$(n_1 - 3)z^2$</td>
</tr>
<tr>
<td>k</td>
<td>$n_k$</td>
<td>$r_k$</td>
<td>$z_k$</td>
<td>$(n_k - 3)$</td>
<td>$(n_k - 3)z$</td>
<td>$(n_k - 3)z^2$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>$N$</td>
<td>$T_1$</td>
<td>$T_2$</td>
<td></td>
</tr>
</tbody>
</table>

The best estimate of tanh$^{-1} \rho$ when the various coefficients are homogeneous is $T_1/N$. The statistic for testing homogeneity is

$$\chi^2 = T_2 - \frac{T_1^2}{N}$$

which can be used as $\chi^2$ with $(k - 1)$ degrees of freedom.
As an example, let the correlations obtained from 6 samples of sizes 10, 14, 16, 20, 25, 28 be 0.318, 0.106, 0.253, 0.340, 0.116, 0.112. Can these be considered homogeneous?

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.318</td>
</tr>
<tr>
<td>14</td>
<td>0.106</td>
</tr>
<tr>
<td>16</td>
<td>0.253</td>
</tr>
<tr>
<td>20</td>
<td>0.340</td>
</tr>
<tr>
<td>25</td>
<td>0.116</td>
</tr>
<tr>
<td>28</td>
<td>0.112</td>
</tr>
</tbody>
</table>

To determine if the correlations are homogeneous, we calculate:

\[ T_1 = \frac{18.2310}{95} \approx 0.191905 \]

\[ T_2 = T_1 \frac{18.2310}{95} - 3.4986 = 1.3876 \]

The value 1.3876 as \( x^2 \) with 5 degrees of freedom is not significant, so the correlations may be considered homogeneous.

### 6b.6 Correction for Bias in the Test for Homogeneity and the Best Estimate of \( \rho \)

When the sample sizes are not large and not nearly equal, there is a certain amount of bias (extraneous to the hypothesis tested) introduced in the \( x^2 \) statistic used in 6b.5. This is due to neglecting the term \( \rho/2(n-1) \) in the mean value of \( z \). Even if the bias introduced in the \( x^2 \) statistic is small, the bias introduced in the best estimate of \( \rho \) when \( x^2 \) is not significant will not be small when compared to the standard error of the estimate. This can be corrected by following a slightly different procedure.

Since \( z \) can be considered as a normal deviate with mean \( \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} + \rho/2(n-1) \) and the variance \( 1/(n-3) \), the score for \( \rho \) obtained from \( k \) samples is

\[ S = \Sigma(n_i - 3) \left[ \frac{1}{1 - \rho^2} + \frac{1}{2(n_i - 1)} \right] \left[ z_i - \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} - \frac{\rho}{2(n_i - 1)} \right] \]

(6b.6.1)
and the information

\[ I = E(S^2) = \Sigma(n_i - 3) \left( \frac{1}{1 - \rho^2} + \frac{1}{2(n_i - 1)} \right)^2 \]  

(6b.6.2)

If the value of \( \rho \) obtained in the last section is taken as a first approximation, then the additive correction \( \delta \rho \) to this value is given by

\[ \delta \rho = \frac{S_0}{I_0} \]

where \( S_0 \) and \( I_0 \) are the values of (6b.6.1) and (6b.6.2) calculated at the approximate value chosen. This process may be repeated till the correction becomes negligible. Having obtained the best estimate \( \hat{\rho} \) of \( \rho \), the \( \chi^2 \) statistic with \( (k - 1) \) degrees of freedom for testing homogeneity is

\[ \chi^2 = \Sigma(n_i - 3) \left( z_i - \frac{1}{2} \log_e \frac{1 + \hat{\rho}}{1 - \hat{\rho}} - \frac{\hat{\rho}}{2(n_i - 1)} \right)^2 \]

References


CHAPTER 7

Tests of Significance
in Multivariate Analysis

7a Review of Work on Multivariate Analysis

Attempts have been made in recent years to generalize the univariate analysis of variance technique to the case of multiple variates. The extension of the theory has been slow, and only a few methods have been made available for practical use. The starting point of these researches, given by Wishart in 1928, is the simultaneous sampling distribution of the variances and covariances in samples from a multivariate normal population. A few years later Hotelling (1931) found the distribution of a quantity $T$ which is a natural extension of Student’s distribution to a sample from a multivariate normal population.

Wilks (1932), following the likelihood ratio method (Neyman and Pearson, 1928, 1931; Pearson and Neyman, 1930), obtained suitable generalizations in the analysis of variance applicable to several variables. The statistic $A$ proposed by him has been found useful in a variety of problems. Bartlett (1934) applied it for testing the significance of treatments with respect to two variables in a varietal trial and indicated its general use in multivariate tests of significance. Wilks (1935) and Hotelling (1935) found it useful in testing the independence of several groups of variates. Wilks's statistic supplied some of the basic tests in multivariate analysis, but the problem of tabulation has not been tackled except in some limited cases (Wald and Brookner, 1941). A very useful approximation has been suggested by Bartlett (1938), who further demonstrated its use in another paper (1947).

A new line of research was initiated by Fisher (1936) with his introduction of the discriminant function analysis. It has been shown that a set of multiple measurements may be used to provide a discriminant function linear in the observations having the property that, better than any other linear function, it will discriminate between any two chosen classes such as taxonomic species, the two sexes, and so on.
The introduction of the discriminant function led to a new method of deriving test criteria suitable for multiple variates. The problem is reduced to the case of a single variate by using a linear compound of the several variables, where the compounding coefficients are chosen to maximize the value of a statistic suitable for a single variate. The application of this method to test the differences in mean values for several groups gave rise to the theory of canonical roots of determinantal equations (Roy, 1939; Fisher, 1939; Hsu, 1939). The distribution of the individual roots and the exact nature of tests require further study. Wilks's statistic, which is a symmetric function of the canonical roots, may be considered as providing an overall test of the hypothesis concerned.

In this chapter a unified approach to the problem of tests of significance in multivariate analysis is developed. The concept of analysis of dispersion, which is a natural extension of the univariate analysis of variance, has been found useful in discussing multivariate problems.

In presenting the various tests of significance it has been found convenient to consider the problems arising out of a single sample and two samples in the first stage. They depend on simple tests of significance requiring the use of variance ratio tables alone and are of very great importance in practice. The use of Wilks's statistic in multivariate analysis involving more than two samples is considered in the second stage. Two powerful approximations have been found for the exact distribution of the $\Lambda$ statistic. A number of examples have been worked out to explain the computational procedure.

**7b Tests with Discriminant Functions**

**7b.1 Two Fundamental Distributions**

The method of discriminant functions in deriving test criteria has been found extremely useful in multivariate analysis. The problem is reduced to that of a single variable by choosing a linear compound of the original variables and constructing a statistic suitable for the univariate case. The maximized value of this statistic obtained by a suitable choice of the compounding coefficients is taken as the appropriate test criterion. The distribution of the statistics thus derived in problems involving a single sample and two samples depends on the two fundamental distributions considered below.

Let $(w_{ij})$, $(i, j = 1, 2, \ldots, p)$, be the matrix giving the estimates, on $n$ degrees of freedom, of the elements in the dispersion matrix $(a_{ij})$ of $p$ normally correlated variables. The definition of $w_{ij}$ implies that it has been calculated from a certain sum of products by dividing by the
appropriate degrees of freedom. Let \( d_1, d_2, \ldots, d_p \) be \( p \) normal variates with the same dispersion matrix \((a_{ij})\) but distributed independently of \( w_{ij} \). Considering only the first \( r \) variables, \( d_1, \ldots, d_r \), the statistic \( T_r \) is defined by

\[
n T_r = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} d_i d_j
\]

(7b.1.1)

where \((w_{ij})\) is the matrix reciprocal to \((w_{ij})\), \((i, j = 1, 2, \ldots, r)\). It was shown in 2d.2 that, when \( E(d_1) = \cdots = E(d_r) = 0 \), the statistic

\[
\begin{align*}
\left| \frac{w_{ij}}{w_{ij} + d_i d_j} \right| &= 1 \\
\frac{1}{w_{ij} + d_i d_j} &= 1 + T_r
\end{align*}
\]

is distributed in the beta form

\[
B\left(\frac{n - r + 1}{2}, \frac{r}{2}\right)
\]

in which case \( T_r \) has the distribution

\[
\frac{T^2}{(1 + T_r)^{(n+1)/2}} dT
\]

This shows that

\[
n - r + 1 \quad T_r
\]

can be referred to a variance ratio table with \( r \) and \( (n - r + 1) \) degrees of freedom.

It was further shown that, if the conditional means of \( d_{r+1}, \ldots, d_p \) given \( d_1, \ldots, d_r \) have the constant terms zero (i.e., are linear homogeneous in \( d_1, \ldots, d_r \), \( E(d_i) \) being not necessarily zero when \( i = 1, \ldots, r \), the statistic

\[
\begin{align*}
\left| \frac{w_{ij}}{w_{ij} + d_i d_j} \right| &= 1 + T_r \\
\left| \frac{w_{ij} + d_i d_j}{w_{ij}} \right| &= \frac{1 + T_r}{1 + T_p} = (U_{p-r} + 1)^{-1}
\end{align*}
\]

is distributed as \( B((n - p + 1)/2, (p - r)/2) \). This shows that

\[
\frac{n + 1 - p}{p - r} U_{p-r}
\]

(7b.1.2)

can be used as a variance ratio with \( (p - r) \) and \( (n + 1 - p) \) degrees of freedom. The statistic \( T_p \) is calculated from the formula (7b.1.1) by using all the \( p \) variables.

All the tests of significance considered in this section depend on the use of the statistics defined in (7b.1.1) and (7b.1.2).
7b.2 Problems of a Single Sample

Student's test connected with pairs of observations admits generalization in two directions.

The first is to test whether the means of $p$ correlated variables are the same on the basis of a sample of size $N$ from a $p$-variate population. When the test shows differences in mean values, there arises the question of deciding whether an assigned contrast involving the $p$ variates differs from the best contrast as determined from the data.

If $x_{1i}, x_{2i}, \ldots, x_{pi}$ are the observations on the $i$th individual, then they may be replaced by a linear compound $z_i = l_1 x_{1i} + \ldots + l_p x_{pi}$ where $l_i$ satisfy the condition $l_1 + \ldots + l_p = 0$. The problem of determining the best contrast reduces to that of determining the compounding coefficients $l_1, \ldots, l_p$ such that the ratio of mean $z$ to standard deviation of $z$ is a maximum. An alternative method which has some practical advantage is as follows.

By arbitrary choice of constants we construct $(p - 1)$ independent linear combinations of the variables $x_1, \ldots, x_p$,

$$y_j = m_{1j} x_1 + \ldots + m_{pj} x_p$$

such that $\sum m_{ij} = 0$ for $j = 1, 2, \ldots, (p - 1)$. Choosing a linear compound of $x$ with coefficients adding to zero is the same as choosing a linear compound of $y$ without any restriction on the compounding coefficients. If the linear compound is

$$\lambda_1 y_1 + \lambda_2 y_2 + \ldots + \lambda_{p-1} y_{p-1}$$

then the quantity to be maximized is

$$v = \frac{(\lambda_1 \bar{y}_1 + \ldots + \lambda_{p-1} \bar{y}_{p-1})^2}{\lambda_1^2 + \ldots + \lambda_{p-1}^2}$$

where

$$w_{ij} = \frac{1}{N-1} \sum_{i=1}^{N} (y_{ir} - \bar{y}_i)(y_{jr} - \bar{y}_j)$$

Observe that only the ratios of $\lambda$ are uniquely determinable, the equations giving $\lambda$ may be written

$$\lambda_1 w_{11} + \ldots + \lambda_{p-1} w_{p-1} = \bar{y}_i \quad \text{for} \quad i = 1, 2, \ldots, (p - 1)$$

with the solution

$$\lambda_1 = w^{11} \bar{y}_1 + \ldots + w^{p-1} \bar{y}_{p-1} \quad \text{for} \quad i = 1, 2, \ldots, (p - 1)$$

where the matrix $(w^{ij})$ is reciprocal to $(w_{ij})$. This supplies the best
linear compound of \( y \), which on transformation to \( x \) gives the best contrast determinable from the data.

The maximum value of \( v \) is given by

\[
\Sigma_j g_j = \Sigma \Sigma u^{ij} \hat{g}_j
\]

If \( T_{p-1} = N(\Sigma \Sigma u^{ij} \hat{g}_j)/(N - 1) \), then, on the hypothesis that all \( x \) have the same mean value, the conditions required for the use of the statistic (7b.1.1) are satisfied so that

\[
T_{p-1}(N - p + 1) = \frac{1}{(p - 1)}
\]

can be used as a variance ratio with \((p - 1)\) and \((N - p + 1)\) degrees of freedom to test the above hypothesis.

The statistic \( T_{p-1} \) is invariant for all sets of coefficients chosen to construct \( y \) from \( x \) so that in any practical problem either conveniently or conventionally chosen linear contrasts of \( x \) may be used to define \( y \).

To test whether the best contrast as determined from the data is in agreement with an assigned contrast \( \eta_1 x_1 + \cdots + \eta_p x_p \) or \( \eta_1 y_1 + \cdots + \eta_{p-1} y_{p-1} \) in terms of \( y \), we proceed as follows.

The appropriate statistic for testing the significance of the assigned contrast is

\[
T_1 = \frac{N(\eta_1 g_1 + \cdots + \eta_{p-1} g_{p-1})^2}{(N - 1)(\Sigma \Sigma \eta_1 \eta_2)}
\]

where \( T_1(N - 1) \) is a variance ratio with 1 and \((N - 1)\) degrees of freedom. The appropriate statistic for all the \((p - 1)\) contrasts is \( T_{p-1} \), considered before. The hypothesis specifies that all contrasts uncorrelated to the assigned one have zero mean so that the conditions for the use of the statistic (7b.1.2) are satisfied. Hence

\[
\frac{(N - p + 1)}{(p - 2)} \left( \frac{1 + T_{p-1}}{1 + T_1} - 1 \right)
\]

can be used as a variance ratio with \((p - 2)\) and \((N - p + 1)\) degrees of freedom to test the above hypothesis.

The above test can be generalized to answer the problem whether a set of \( k \) assigned contrasts contain the best contrast. In this case the statistic

\[
\frac{(N - p + 1)}{(p - k - 1)} \left( \frac{1 + T_{p-1}}{1 + T_k} - 1 \right)
\]

can be used as a variance ratio with \((p - k - 1)\) and \((N - p + 1)\) degrees of freedom.
PROBLEMS OF A SINGLE SAMPLE

Example 1. The data of Table 7b.2a consist of weights of cork borings taken from the north (N), east (E), south (S), and west (W) directions of the trunk for 28 trees in a block of plantations. The problem is to test whether the bark deposit varies in thickness and hence in weight in the four directions. It was suggested that the bark deposit is likely to be uniform in N and S directions and also uniform but less so in E and W directions, so that \( (N - E - W + S) \) can be taken as the best contrast. This can, however, be tested from the given data as shown below.

<table>
<thead>
<tr>
<th>N</th>
<th>E</th>
<th>S</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>66</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>60</td>
<td>53</td>
<td>66</td>
<td>63</td>
</tr>
<tr>
<td>56</td>
<td>57</td>
<td>64</td>
<td>58</td>
</tr>
<tr>
<td>41</td>
<td>29</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>34</td>
<td>26</td>
</tr>
<tr>
<td>29</td>
<td>39</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>42</td>
<td>43</td>
<td>31</td>
<td>25</td>
</tr>
<tr>
<td>37</td>
<td>40</td>
<td>31</td>
<td>25</td>
</tr>
<tr>
<td>33</td>
<td>29</td>
<td>27</td>
<td>36</td>
</tr>
<tr>
<td>32</td>
<td>30</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>63</td>
<td>45</td>
<td>74</td>
<td>63</td>
</tr>
<tr>
<td>54</td>
<td>46</td>
<td>60</td>
<td>52</td>
</tr>
<tr>
<td>47</td>
<td>51</td>
<td>52</td>
<td>43</td>
</tr>
</tbody>
</table>

It has been found in similar studies that there exists a significant correlation between contrasts such as \( (N - E) \) and \( (S - W) \) so that the method of fitting constants for the four directions and the individual trees by the method of least squares is not appropriate. The three contrasts arising out of the four weights may then be treated as three correlated variables, in which case the theory developed above is applicable.

It is interesting to observe that the individual weights in Table 7b.2a are exceedingly asymmetrically distributed. This does not, however, invalidate the test so long as the contrasts are normally distributed. In fact, the distribution of the individual weights depends on the nature of plants and the variation between plants. If the above condition is satisfied, it is not necessary that the individual weights should follow
any distribution law of the known type. It may be sometimes necessary to make a transformation (such as log, square, or cube root) of the variables under consideration to ensure that the contrasts of the transformed variables are symmetrically distributed if the contrasts of the original variables are not so.

As observed earlier, the contrasts may be conveniently or conventionally chosen. In the above example we may choose the simple set of contrasts

\[ y_1 = N - E - W + S \quad y_2 = S - W \quad y_3 = N - S \]

The mean values and estimates of variances and covariances based on 27 degrees of freedom for \( y_1, y_2, y_3 \) are

\[
\begin{align*}
\bar{y}_1 &= 8.8571 \\
\bar{y}_2 &= 4.5000 \\
\bar{y}_3 &= 0.8571 \\
(\mu_{ij}) &= \begin{bmatrix} 128.7200 & 61.4076 & -21.0211 \\ 61.4076 & 56.9259 & -28.2963 \\ -21.0211 & -28.2963 & 63.5344 \end{bmatrix}
\end{align*}
\]

The coefficients of the best linear function \( \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3 \) are given by the equations

\[
\begin{align*}
128.7200\lambda_1 + 61.4076\lambda_2 - 21.0211\lambda_3 &= 8.8571 \\
61.4076\lambda_1 + 56.9259\lambda_2 - 28.2963\lambda_3 &= 4.5000 \\
-21.0211\lambda_1 - 28.2963\lambda_2 + 63.5344\lambda_3 &= 0.8571
\end{align*}
\]

Solving, \( \lambda_1 = 0.05620, \lambda_2 = 0.04415, \lambda_3 = 0.05174 \), so that the best contrast is

\[
\lambda_1(N - E - W + S) + \lambda_2(S - W) + \lambda_3(N - S)
\]

\[
= 0.10794N - 0.05620E - 0.10035W + 0.04861S
\]

or, by multiplying the coefficients by 10 (arbitrarily),

\[
1.0794N - 0.5620E - 1.0035W + 0.4861S
\]

The statistic for testing the hypothesis of equality of means is

\[
T_{p-1} = \frac{N}{N - 1} \left( \lambda_1 \bar{y}_1 + \lambda_2 \bar{y}_2 + \lambda_3 \bar{y}_3 \right) = \frac{28}{27} \left( \frac{0.740790}{6.4019} \right) = 0.768226
\]

\[
T_{p-1} = \frac{(N - p + 1)}{(p - 1)} \frac{0.768226(28 - 4 + 1)}{3} = 6.4019
\]

The quantity 6.4019 as a variance ratio with 3 and 25 degrees of freedom
is significant at the 1% level so that the bark deposit cannot be considered uniform in the four directions.

The assigned contrast is represented by \( y_1 \). To test for its significance the statistic is

\[
T_1 = \frac{N}{N - 1} \frac{\hat{\theta}^2}{\sigma_{W1}^2} = \frac{28(8.5571)^2}{27(128.7200)} = 0.632020
\]

The quantity \( (N - 1)T_1 = 17.0645 \) as the variance ratio with 1 and 27 degrees of freedom is significant.

To test whether the assigned contrast agrees with that estimated from the data, the statistic \( U \) defined in (7b.1.2) has to be calculated.

\[
U_{z,1} = \frac{N - p + 1}{p - 2} \left( \frac{1 + T_{p-1}}{1 + T_1} - 1 \right)
\]

\[
= \frac{25}{2} \left( \frac{1.768226}{1.632020} - 1 \right) = 1.0431
\]

This value as the variance ratio with 2 and 25 degrees of freedom is small so that the evidence supplied by the data is not sufficient to reject the assigned contrast as not the best, although the ratios of the coefficients in the estimated contrast depart considerably from those assigned.

Another problem connected with a single sample is to test for the significance of the departures of the observed mean values from those assigned. Let \( \bar{z}_1, \cdots, \bar{z}_p \) be the mean values based on a sample of size \( N \), and \( \xi_1, \cdots, \xi_p \) the assigned values. If \((w_{ij})\) is the covariance matrix of \( z_1, \cdots, z_p \) estimated on \( n \) degrees of freedom, then

\[
T_p = \frac{N}{n} \sum w_{ij} (\bar{z}_i - \xi_i)(\bar{z}_j - \xi_j)
\]

The variance ratio with \( p \) and \( (n + 1 - p) \) degrees of freedom to test the above hypothesis is

\[
T_p \left( \frac{n + 1 - p}{p} \right) = \frac{N(n + 1 - p)}{np} \sum w_{ij} (\bar{z}_i - \xi_i)(\bar{z}_j - \xi_j)
\]

In many problems both the mean values and the covariance matrix are estimated from the same sample, in which case \( n = N - 1 \).

**Example 2.** Consider the covariance matrix

\[
\begin{array}{ccc}
128.7200 & 61.4076 & -21.0211 \\
61.4076 & 56.9259 & -28.2963 \\
-21.0211 & -28.2963 & 63.5344
\end{array}
\]
estimated on 27 degrees of freedom and mean values 8.8571, 4.5000, 0.8571 based on 28 observations as in example 1 above. Suppose that it is required to test whether the calculated averages agree with the assigned values 5, 1, and -2. The deviations are

\[
8.8571 - 5 = 3.8571 \quad 4.5000 - 1 = 3.5000 \quad 0.8571 + 2 = 2.8571
\]

to evaluate the quadratic form

\[
\sum \omega_{ij} d_i d_j
\]

we follow the form adopted in Id.1 and sweep out the matrix.

Dispersion Matrix
\[
\begin{pmatrix}
128.7200 & 61.4076 & -21.0211 \\
56.9259 & -28.2963 & 63.5344 \\
0 & 0 & 0
\end{pmatrix}
\]

Deviations
\[
\begin{pmatrix}
3.8571 \\
3.5000 \\
2.8571 \\
0
\end{pmatrix}
\]

This gives the last pivotal quantity 0.6529 (with a negative sign) which is the value of the quadratic form. The variance ratio with 3 and (27 + 1 - 3) degrees of freedom is

\[
\frac{28}{27} \frac{27 + 1 - 3}{3} \frac{0.6529}{81} = 5.6423
\]

which is significant on the 1% level, indicating departure from the expected.

The second generalization of Student's \(t\) is concerned with testing, on the basis of a sample of size \(N\) from a \(2p\)-variate population containing the variables \(y_1, \ldots, y_p, y_{p+1}, \ldots, y_{2p}\) whether the mean values of \(y_i\) and \(y_{i+p}\) are the same for all \(i\). The \(2p\) variates can be reduced to \(p\) variates

\[
z_1 = y_{p+1} - y_1 \quad z_2 = y_{p+2} - y_2 \quad \cdots \quad z_p = y_{2p} - y_p
\]

in which case the problem reduces to the one considered above. The variance ratio with \(p\) and \((N - p)\) degrees of freedom is

\[
\frac{N}{N - p} \frac{N - 1}{p} \sum \omega_{ij} d_i d_j
\]

where \((\omega_{ij})\) is the dispersion matrix of \(z_1, \cdots, z_p\) based on \((N - 1)\) degrees of freedom.

The test is useful in various situations. Suppose that we want to test for asymmetry of organisms. The sets \(y_1, \cdots, y_p\) and \(y_{p+1}, \cdots, y_{2p}\) will then correspond to the same measurements on the right and
left sides of an organism. Another interesting study is whether the first born in a family differs from the second born. To illustrate the method, a random sample of 25 families has been chosen from Dr. G. P. Frets's data giving the head lengths and breadths of all sons and daughters in a large number of families in Germany. For effective

**Table 7b.23. The Measurements on the First and Second Adult Sons in a Sample of 25 Families (Data by G. P. Frets)**

<table>
<thead>
<tr>
<th>Head Length</th>
<th>Head Breadth</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Son</td>
<td>Second Son</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>179</td>
</tr>
<tr>
<td>195</td>
<td>201</td>
</tr>
<tr>
<td>181</td>
<td>185</td>
</tr>
<tr>
<td>183</td>
<td>188</td>
</tr>
<tr>
<td>176</td>
<td>171</td>
</tr>
<tr>
<td>208</td>
<td>192</td>
</tr>
<tr>
<td>189</td>
<td>190</td>
</tr>
<tr>
<td>197</td>
<td>186</td>
</tr>
<tr>
<td>188</td>
<td>197</td>
</tr>
<tr>
<td>192</td>
<td>187</td>
</tr>
<tr>
<td>179</td>
<td>186</td>
</tr>
<tr>
<td>183</td>
<td>174</td>
</tr>
<tr>
<td>174</td>
<td>185</td>
</tr>
<tr>
<td>190</td>
<td>195</td>
</tr>
<tr>
<td>188</td>
<td>187</td>
</tr>
<tr>
<td>163</td>
<td>161</td>
</tr>
<tr>
<td>195</td>
<td>183</td>
</tr>
<tr>
<td>186</td>
<td>173</td>
</tr>
<tr>
<td>181</td>
<td>182</td>
</tr>
<tr>
<td>175</td>
<td>165</td>
</tr>
<tr>
<td>192</td>
<td>185</td>
</tr>
<tr>
<td>174</td>
<td>178</td>
</tr>
<tr>
<td>176</td>
<td>176</td>
</tr>
<tr>
<td>187</td>
<td>200</td>
</tr>
<tr>
<td>190</td>
<td>187</td>
</tr>
</tbody>
</table>

Mean difference 1.88 1.48
The dispersion matrix of the differences estimated on 24 degrees of freedom is

\[
\begin{pmatrix}
  0.803 & 11.52 \\
  11.52 & 24.01
\end{pmatrix}
\]

with its inverse

\[
\begin{pmatrix}
  0.015999 & -0.007677 \\
  -0.007677 & 0.045332
\end{pmatrix}
\]

\[
\sum \sum w_{ij}d_{ij}
= 0.015999(1.88)^2 - 2(0.007677)(1.88)(1.48) + 0.045332(1.48)^2
= 0.113121
\]

\[
\frac{N}{N-1} \frac{N-p}{p} (0.113121) = \frac{25}{24} \frac{23}{2} (0.1131) = 1.3548
\]

This is not significant as a variance ratio with 2 and 23 degrees of freedom. There is no difference in the dimensions of the first son and the second as judged by the above sample. The method described above for such studies is quite general and can be applied to any number of characters.

7b.3 Mahalanobis' $D^2$ and Problems of Two Samples

Let $N_1$ and $N_2$ be the samples drawn from two populations, each characterized by $p$ variates. The sample means for the $i$th character are represented by $\bar{x}_{1i}$ and $\bar{x}_{2i}$ for the first and second samples, respectively. The estimated value of the covariance is given by

\[
(N_1 + N_2 - 2)\omega_{ij} = \sum_{i=1}^{N_1} (x_{1it} - \bar{x}_{1i})(x_{jit} - \bar{x}_{jt})
+ \sum_{i=1}^{N_2} (x_{2it} - \bar{x}_{2i})(x_{j2t} - \bar{x}_{j2})
\]

the right-hand expression being the sum of the corrected sums of products for two samples. Mahalanobis' (1936) distance between the two populations as estimated from the sample on the basis of the $p$ characters is

\[
D_p^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_{ij}(\bar{x}_{1i} - \bar{x}_{2i})(\bar{x}_{1j} - \bar{x}_{2j})
\]

where $(\omega_{ij})$ is the reciprocal of $(\omega_{ij})$, $(i, j = 1, 2, \cdots, p)$. The exact distribution of $D^2$ on the hypothesis specifying real differences in mean

*The subscript $p$ in the symbol $D_p^2$ denotes the number of characters used. The suffix may be omitted unless various $D^2$ values based on different sets or numbers of characters are to be kept distinct in any problem.
values is derived in 2.1.2. To test the hypothesis specifying no difference in mean values of the \( p \) characters for the two populations, the statistic
\[
\frac{N_1N_2(N_1 + N_2 - p - 1)}{p(N_1 + N_2)(N_1 + N_2 - 2)} D^2
\]
can be used as a variance ratio with \( p \) and \( (N_1 + N_2 - 1 - p) \) degrees of freedom.

As observed earlier, the above test can be derived in an interesting way suggested by R. A. Fisher. If the \( p \) measurements are replaced by a linear compound
\[
y = l_1x_1 + \cdots + l_p x_p
\]
then the ratio of between to within variance of \( y \) from the two samples is
\[
\frac{N_1N_2 (l_1d_1 + \cdots + l_pd_p)^2}{N_1 + N_2 \Sigma \Sigma l_il_jw_{ij}}
\]
Maximizing this, we find that the coefficients of the best linear function separating the two groups are obtained as solutions of the equations
\[
\begin{align*}
l_1w_{11} + l_2w_{12} + \cdots & = d_1\mu \\
l_1w_{21} + l_2w_{22} + \cdots & = d_2\mu \\
& \vdots \\
l_1w_{p1} + l_2w_{p2} + \cdots & = d_p\mu
\end{align*}
\]
where \( \mu \) is a constant. Observing that only ratios of \( \mu \) can be uniquely determined, we can replace \( \mu \) by unity and solve the above equations. Multiplying the above equations by \( l_1, l_2, \cdots \) and adding, we find
\[
\Sigma \Sigma l_il_jw_{ij} = l_1d_1 + \cdots + l_pd_p = \Sigma \Sigma w^idy_i = D^2
\]
The optimum ratio is then
\[
\frac{N_1N_2}{N_1 + N_2} \Sigma \Sigma w^idy_i = \frac{N_1N_2}{N_1 + N_2} D^2
\]
The significance of this can be tested as shown above.

Example. The following tables (7b.3a and 7b.3b), reproduced from Fisher (1938), give the mean values based on 50 observations each and the covariance based on \((50 + 50 - 2)\) degrees of freedom for four characters in two species of plants \textit{Iris versicolor} and \textit{Iris setosa}.

The solution of the equations (see Table 7b.6d)
\[
l_1w_{1i} + \cdots + l_4w_{4i} = d_i \quad i = 1, 2, 3, 4
\]
is obtained as

\[ l_1 = -3.0692 \quad I_2 = -18.0006 \quad l_3 = 21.7641 \quad l_4 = 30.7510 \]

so that the discriminant function is

\[-3.0692x_1 - 18.0006x_2 + 21.7641x_3 + 30.7510x_4\]

The value of \( D^2 \) is

\[ I_1 d_1 + \cdots + I_p d_p = 103.2119 \]

To test for the differences in mean values the statistic is

\[ \frac{N_1 N_2 (N_1 + N_2 - 1 - 4) D^2}{(N_1 + N_2)(N_1 + N_2 - 2)} = \frac{103.2119}{50 \times 98 \times 4} = 81.8^2 (26.3205) = 625.3256 \]

which as a variance ratio with 4 and 95 degrees of freedom is significant.

<table>
<thead>
<tr>
<th>Character</th>
<th>Iris versicolor</th>
<th>Iris setosa</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal length ((x_1))</td>
<td>5.036</td>
<td>5.006</td>
<td>0.030</td>
</tr>
<tr>
<td>Sepal width ((x_2))</td>
<td>2.770</td>
<td>3.428</td>
<td>-0.658</td>
</tr>
<tr>
<td>Petal length ((x_3))</td>
<td>4.260</td>
<td>1.462</td>
<td>2.798</td>
</tr>
<tr>
<td>Petal width ((x_4))</td>
<td>1.326</td>
<td>0.246</td>
<td>1.080</td>
</tr>
</tbody>
</table>

The method of evaluating the \( D^2 \) given above is useful because the best discriminating function is also found out during the process. This is useful in problems of classification as treated in the next chapter.

**Tab. 7b.3a.** Observed Mean Values Based on 50 Observations Each for the Two Species

**Table 7b.3a.** The Pooled Covariance Matrix \((\Sigma_{ij})\) Based on 98 Degrees of Freedom

The method of evaluating the \( D^2 \) given above is useful because the best discriminating function is also found out during the process. This is useful in problems of classification as treated in the next chapter.

**Tab. 7b.4.** Test for an Assigned Discriminant Function

In the last section the discriminant function for *Iris versicolor* and *Iris setosa* based on four measurements was found to be

\[-3.0524x_1 - 18.0231x_2 + 21.7656x_3 + 30.8446x_4\]

with the value of \( D^2 = 103.2119 \). Since the mean measurements for...
versicolor exceed those for setosa except in sepal width \((x_2)\), a discriminant function of the type
\[
x_1 - x_2 + x_3 + x_4
\]
may be suggested. In such a case it might be of interest to know whether the discriminant function derived above is an improvement over the assigned simpler function. If the assigned function is represented by \(y\), then
\[
D_y^2 = \frac{\langle \hat{y}_1 - \hat{y}_2 \rangle^2}{V(y)}
\]
where \(\hat{y}_1\) and \(\hat{y}_2\) are the mean values of \(y\) for the two species.

\[
\hat{g} = \hat{x}_1 - \hat{x}_2 + \hat{x}_3 + \hat{x}_4
\]
\[
\hat{g}_1 - \hat{g}_2 = d_1 - d_2 + d_3 + d_4
\]
\[
= 0.930 - 0.638 + 2.798 + 1.080 = 5.466
\]
\[
V(y) = V(x_1) + V(x_2) + V(x_3) + V(x_4) - 2 \text{ Cov}(x_1x_2) + 2 \text{ Cov}(x_1x_3) + 2 \text{ Cov}(x_1x_4)
\]
\[
+ 2 \text{ Cov}(x_2x_3) - 2 \text{ Cov}(x_2x_4)
\]
\[
= w_{11} + w_{33} + w_{44} - 2w_{12} - 2w_{13} - 2w_{14} + 2w_{34}
\]
\[
+ 2w_{44} = 0.482295
\]
using the values of \(w_{ij}\) given in Table 7b.3J3.

\[
D_y^2 = \frac{29.8771}{0.482295} = 61.9479
\]

To test whether the assigned discriminant function is in agreement with that derived from the data, the significance of the statistic
\[
U = \frac{1 + N_1N_2D_y^2/(N_1 + N_2)(N_1 + N_2 - 2)}{1 + N_1N_2D_y^2/(N_1 + N_2)(N_1 + N_2 - 2)/1 + W_{26.3295} - 1 = 0.6265}
\]
\[
= 15.8030
\]
has to be tested. The value of the statistic
\[
U(N_1 + N_2 - 1 - 4) = 0.6265 \times 95
\]
\[
= 19.8392
\]
as a variance ratio with 3 and 96 degrees of freedom is significant at the 1% level. This shows that the assigned function is not the best discriminant of the two species.

In general, if the assigned discriminant function is

\[ y = a_1 x_1 + \cdots + a_p x_p \]

then

\[ D_2^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{V(y)} \]

where \( V(y) = \Sigma \sigma_{ij} w_{ij} \). If \( w_{ij} \) are estimated on \( n \) degrees of freedom and the mean values are based on \( N_1 \) and \( N_2 \) observations for the two groups, then

\[ U = \frac{1 + N_1 N_2 D_2^2 / (N_1 + N_2)n}{1 + N_1 N_2 D_2^2 / (N_1 + N_2)n} - 1 \]

and

\[ U(n - p + 1) \]

\[ \frac{p - 1} \]

can be used as a variance ratio with \((p - 1)\) and \((n - p + 1)\) degrees of freedom. This test is due to Fisher.

7b.5 Tests for Discriminant Function Coefficients

If four samples of sizes \( N_1, N_2, N_3, \) and \( N_4 \) from populations \( A, B, C, \) and \( E \) are available, we can test whether the discriminant functions between \( A, B \) and \( C, E \) are significantly different by an extension of the test criterion discussed above. It is a necessary condition of the test that the variances and covariances are identical in the four populations \( A, B, C, \) and \( E \). No reasonably simple test can be constructed to establish the equivalence of the discriminant functions when this condition is not satisfied.

Let \( \Sigma \) be the dispersion matrix based on \( (N_1 + N_2 + N_3 + N_4 - 4) \) degrees of freedom. If \( d_1, \cdots, d_p \) are the differences in mean values for \( A \) and \( B \), and \( d_1', \cdots, d_p' \) are those for \( C \) and \( E \), the test for equality of discriminant functions and the associated distances is identical with the testing of the hypotheses

\[ E(d_i) = E(d_i') \quad i = 1, 2, \cdots, p \]

\[ E(d_i) = E(-d_i') \quad i = 1, 2, \cdots, p \]

The variance ratios with \( p \) and \( n = (N_1 + N_2 + N_3 + N_4 - 3 - p) \)
degrees of freedom for the two cases are

\[ \frac{n}{p n + p - 1} \sum \omega_i (d_i - d'_i)(d_j - d'_j) \]

and

\[ \frac{n}{p n + p - 1} \sum \omega_i (d_i + d'_i)(d_j + d'_j) \]

where

\[ \frac{1}{f(N)} = \frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4} \]

The equality of discriminant functions is indicated if at least one of the statistics is not significant. Similar tests can be constructed for judging the differences in discriminant functions in parallel samples from two populations or between \( A, B \) and \( A, C \).

If the equality of discriminant function coefficients are to be tested without considering the associated distance function, a suitable statistic is

\[ g(N) \sum \omega_i (d_i - \lambda d'_i)(d_j - \lambda d'_j) \]  \hspace{1cm} (7b.5.1)

where

\[ \frac{1}{g(N)} = \frac{1}{N_1} + \frac{1}{N_2} + \lambda^2 \left( \frac{1}{N_3} + \frac{1}{N_4} \right) \]

and \( \lambda \) is chosen to minimize (7b.5.1). This minimum value may be used as \( \chi^2 \) with \( (p - 1) \) degrees of freedom when \( n \) is large.

Standard errors of discriminant function coefficients have been evaluated in an attempt to judge the significance of any single coefficient. There is some difficulty in this approach because discriminant function coefficients are not unique in the sense that they are the estimates of definite population parameters. What is unique is the ratio of any two coefficients, and an exact test is possible to test for an assigned ratio. For instance, if the ratio for the \( i \)th and \( j \)th characters is \( p \), then we have to test whether the distance based on the \( (p - 1) \) characters

\[ x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p, x_i + px_j \]

is the same as that based on all the \( p \) characters

\[ x_1, \ldots, x_p \]

The statistic is

\[ U = \frac{1 + \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D_p^2}{1 + \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D_p^{p-1}} - 1 \]
and to test for its significance
\[
\frac{N_1 + N_2 - p - 1}{U}
\]
can be used as a variance ratio with 1 and \((N_1 + N_2 - p - 1)\) degrees of freedom.

7b.6 The Additional Information Supplied by Some Characters

Table 7b.6a gives the mean values of femur and humerus lengths of 20 Indian and 27 Anglo-Indian skeletons.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean Length of</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Femur</td>
<td>Humerus</td>
</tr>
<tr>
<td>Anglo-Indians</td>
<td>460.4</td>
<td>335.1</td>
</tr>
<tr>
<td>Indians</td>
<td>444.3</td>
<td>323.2</td>
</tr>
<tr>
<td>Difference</td>
<td>16.1</td>
<td>11.9</td>
</tr>
</tbody>
</table>

The pooled estimates (on 45 degrees of freedom) of standard deviations are 23.7 and 18.2 and of correlation 0.8675. The \(D^2\) based on the femur alone is 0.4614, which yields a significant variance ratio 5.301 with 1 and 45 degrees of freedom. But the \(D^2\) based on the two characters, femur and humerus lengths, is 0.4777, leading to the variance ratio 2.685 which is not significant on 2 and 44 degrees of freedom. Here appears to be a dangerous situation where the inclusion of an extra character is not beneficial in discriminating between two populations. This leads us to the problem of studying the nature and number of characters which may be of use in discriminating between the groups. The first step in such a study is to develop a test to judge the significance of the additional distance contributed by the inclusion of some extra characters. The addition of such characters which do not increase the distance between the groups in the population will weaken the test. Even a small increase in distance will be helpful if the sample size is large. For instance, in the above example, with 10 more observations and an equal division of the sample size between the two groups, the observed \(D^2\) would have been significant.

Two problems then arise: first, to test whether the inclusion of some extra characters increases the distance in the population, and second to estimate the additional distance and determine for what sample size this addition is useful. There is yet another practical issue which is relevant in problems of the next chapter where a number of measure-
ments are obtained for assigning an individual to one of two groups. The error committed in such a classification depends on the distance between the two groups, and an extra character added may increase the distance only by a trifle, in which case it may not be worth while to measure an extra character.

To solve the first problem let \( p \) be the number of basic characters to which are added \( q \) more characters. Let samples of sizes \( N_1 \) and \( N_2 \) be available for the two groups containing measurements on all the \( p + q \) characters. If \( D^2_{p+q} \) is of the same order as \( D^2_p \), then the ratio

\[
R = \frac{1 + \frac{N_1N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D^2_{p+q}}{1 + \frac{N_1N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D^2_p}
\]

is about unity. A high value of this ratio would indicate that \( D^2_{p+q} \) is significantly greater than \( D^2_p \) so that the \( q \) characters supply some additional information. The actual test is to use

\[
U_{q,p} = \frac{N_1 + N_2 - p - q - 1}{q} \cdot \frac{1}{U_{q,p}}
\]

where \( U_{q,p} = (R - 1) \) as a variance ratio with \( q \) and \((N_1 + N_2 - p - q - 1)\) degrees of freedom.

In the example of *Iris versicolor* and *Iris setosa* we might ask the question whether sepal and petal lengths alone are sufficient for discrimination. In other words, does the inclusion of widths increase the distance? For this we need the value of \( D^2 \) based on the lengths only. It is useful to obtain the corresponding discriminant function also. The successive evaluation of \( D^2 \)'s and discriminant functions can be carried out as illustrated in Table 7b.6. This is essentially a method of pivotal condensation developed by Aitken (1933) but slightly modified to effect economy in entries.

The \( D^2 \) corresponding to lengths only is 76.7082, and \( D^2_4 = 103.2119 \), so that

\[
U_{q,p} = \frac{1 + 50 \times 50(103.2119)/100 \times 98}{1 + 50 \times 50(76.7082)/100 \times 98} - 1
\]

\[
= \frac{1 + 26.3205}{1 + 19.5084} - 1 = 0.3287
\]
### Table 7b.06

| Row | No. | $z_1$ | $z_2$ | $z_3$ | $z_4$ | Difference in Means ($d$) | Sum Including the Indented | Check Excluding the Indented |
|-----|-----|------|------|------|------|----------------|----------------|----------------|---|
| 01  | 2.0153 | 0.0996 | 0.0222 | 0.0311 | | 2.798 | 1.5502 | 2.1009 |
| 02  | 0.1256 | 0.0472 | 0.0368 | | | 0.658 | 0.3723 | |
| 03  | 0.1211 | 0.0222 | | | | 1.680 | 1.2300 | 3.1500 |
| 04  | 0.0251 | | | | | 0.000 | |

**1.** Rows 10, 20, 30, and 40 are the pivots at each stage.

**2.** After sweeping out the first column, fill in the first column by the elements of the first pivot row. These are indented as shown above. Retain these elements in sweeping out the second column at the second stage. In the reduced matrix fill up the second column by elements in the second pivot row. Retain them at subsequent stages.

**3.** The sums in the last but one column are used in obtaining the elements of the check column at each stage of reduction.

**4.** At the stage of obtaining $L(x)$ and $D(x)$ an additional check is available, because $L_0(d) = D_0$. Thus from row 23, 11, 101250 (0.039) + 31, 101250(0.278) = 76.705656, and so on.

**5.** If only successive values of $D(x)$ are needed, as in the test for additional information, then the entries below the diagonal are unnecessary. One check column will do.
The variance ratio with 2 and \((N_1 + N_2 - p - q - 1) = 95\) degrees of freedom is

\[
\frac{95}{95} (0.3287) = 15.6132
\]

which is significant, showing that widths are useful in addition to the lengths. The additional distance in such a case is determined by

\[
D^2_{p+\bar{q}} - D_p^2 = 103.2119 - 76.7082 = 26.5037
\]

A question may be asked as to why the difference \(D^2_{p+\bar{q}} - D_p^2\) could not be tested directly. The distribution of this difference involves the population value of the distance based on the first \(p\) characters, and unless this is known no exact test of significance can be made. On the other hand the statistic \(U_{q,p}\), which also gives a comparison of the two \(D^2\) values, is distributed in a simpler manner, and there is not the problem of any nuisance parameter (56.3). If the samples are large and the population value \(\Delta_{p+}\) of \(D_p^2\) is not large, the distribution of the difference \(D^2_{p+\bar{q}} - D_p^2\) is independent of \(\Delta_{p+}\) to a large extent. In such a case

\[
(W_{q,p}) = \frac{N_1 + N_2 - p - q - 1}{N_1 + N_2 - 1}
\]

\[(7b.6.1)\]

can be used approximately as a variance ratio with \(q\) and \((N_1 + N_2 - p - q - 1)\) degrees of freedom, where

\[
W_{q,p} = \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} (D^2_{p+\bar{q}} - D_p^2)
\]

In the above example,

\[
W_{q,p} = 26.3295 - 19.5684 = 6.7611
\]

\[
\frac{6.7611}{6.7611} = 0.83616
\]

which is a very high variance ratio compared to that obtained on the basis of \(U_{q,p}\). The approximation here is very crude (it always overestimates significance), especially because \(D_{q,p}\) happens to be very high.

Cochran and Bliss (1948) considered a situation where initial intelligent quotients (I.Q.) can be used as concomitant variables in studying the differences introduced by two types of training. For this it is suggested that a sample may be divided at random into two groups, each of which is required to take a different training. This means that with respect to initial I.Q. values the two groups can be regarded as having come from the same population so that an exact test based on \(W_{q,p}\) is possible. The exact distribution of \(W_{q,p}\) even in this case, is a bit
256 TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS

complicated, but a good approximation to this is the variance ratio considered above (76.6.1).

It is interesting to observe that $U_{q,p}$ in any case is distributed independently of the first $p$ variables. The only condition needed for the test is that, given $p$ measurements, the expected value of any other measurement is a linear combination of the first $p$. If this is so, it is not necessary that the first $p$ variables be observed at random. If the problem is to test whether $q$ additional characters discriminate between the two populations independently of a basic set, it might be profitable to select samples from the two populations such that they agree on the average, as far as possible, in the basic set of $p$ characters. In the problem of Cochran and Bliss the initial I.Q. values may be used to effect such a division. The only test available in such a case is that based on $U_{q,p}$. If the sample is divided at random into two groups, the test based on $W_{q,p}$ is theoretically more accurate.

The following conclusions will be useful in studying the problems of this nature.

1. Whatever may be the number of characters chosen to discriminate between two populations, it is profitable to divide the samples equally between the two populations.

2. Unless the samples are large, it is not profitable to consider nearly related measurements in tests for discrimination.

3. To judge the significance of increase in distance due to the inclusion of $q$ extra characters to a basic set of $p$, it is advantageous (from the point of view of the test) to choose the individuals from the two populations such that they may agree, as far as possible, in the average measurements of the basic set.

In 3f.3 some comments were made as to the applicability of a prediction formula found from one series for an individual of an unrelated series. Thus a question might be asked whether a prediction formula for cranial capacity deduced from the measurements on an Anglo-Saxon skull can be used to predict the capacity of an Indian skull, providing measurements on its length, breadth, and height. This is a very important problem because, in the absence of suitable data, it may be necessary to use the formula derived from a different series. As observed earlier, one condition may be that the internal relationships of the measurements should be the same for the two series, which means that the variances and covariances of the measurements should be the same. This requirement is very often satisfied with biometric material; the differences, when they exist, are quite small. This is not enough. It must be known that the two series are such that the whole distance between them should be capable of being explained by the differences in the characters used for prediction. In other words, if
x₁, x₂, ⋯, xₚ are the characters used for the prediction of a character y, then these measurements must be such that the additional distance due to y on eliminating x₁, ⋯, xₚ is theoretically zero. Only in such cases can a single formula be applicable for two series. This can be verified by the methods developed here. In fact the methods are applicable in a more general case requiring the prediction of more than one character with the help of a basic set.

7c Generalization of $D²$ and the Large Sample Theory for Several Groups

Let there be k multivariate populations A₁, A₂, ⋯, Aₖ from which samples of sizes $N₁, N₂, \cdots, Nₖ$ are available for $(p + q)$ characters. The common covariance matrix assumed to be known or estimated on a large number of degrees of freedom is represented by $(a_{ij})_p$ for the first $p$ characters and by $(a_{ij})_{p+q}$ for all the $(p + q)$ characters. The inverse of $(a_{ij})_p$ is represented by $(a_{ij})^{-p}$, and that of $(a_{ij})_{p+q}$ by $(a_{ij})_{p+q}^{-p}$. Let $\bar{x}_i$, $\bar{x}_j$, ⋯, be the mean values of the $i$th character in the first, second, etc., populations.

It is shown in 2c.3 that the statistic,

$$V_{pk} = \sum_{i,j=1}^{p} a_{ij} N_i (\bar{x}_{ir} - \bar{x}_i)(\bar{x}_{jr} - \bar{x}_j)$$

where $\bar{x}_i = (\Sigma N_i \bar{x}_{ir})/(\Sigma N_i)$, can be used as $x²$ with $p(k - 1)$ degrees of freedom to test the hypothesis that the mean values are the same in all the $k$ populations for these $p$ characters. The statistic $V$ is a suitable generalization of Mahalanobis' $D²$ in its classical form.

When this test indicates differences in mean values it is in some problems necessary to test whether the observations on $q$ additional characters supply independent information for discrimination. The statistic for testing the differences in means for all the $p + q$ characters is

$$V_{(p+q)k} = \sum_{i,j=1}^{p+q} a_{ij} N_i (\bar{x}_{ir} - \bar{x}_i)(\bar{x}_{jr} - \bar{x}_j)$$

which can be used as $x²$ with $(p + q)(k - 1)$ degrees of freedom. The $q(k - 1)$ additional degrees of freedom bring in the contribution

$$V_{(p+q)k} - V_{pk}$$

and the significance of this difference can be appropriately used to judge the significance of the information supplied by the additional characters. This difference can be used as $x²$ with $q(k - 1)$ degrees of freedom, as shown below.
The hypothesis that the new characters do not lead to further discrimination of the populations specifies that any linear function of the \((p + q)\) characters uncorrelated with each of the \(p\) characters has the same mean value for all the \(k\) populations. There are \(q\) such linear functions and, if they are treated as \(q\) variables, a \(\chi^2\) with \(q(k - 1)\) degrees of freedom can be constructed to test the above hypothesis. The above method of taking the difference is only an alternative way of calculating this \(\chi^2\); for, \(V_{(p+q)k}\) calculated from all the \((p + q)\) characters, being invariant under linear transformations of the variables, is equal to \(V_{pk} + \chi^2\) calculated from the \(p\) original characters and the \(q\) linear functions chosen to be uncorrelated with each of the \(p\) characters.

In the above derivation it has been assumed that the variances and covariances are known, and the distributions are asymptotically true when they are estimated on a large number of degrees of freedom. When more than two populations are involved the pooled estimates of the covariances usually have a sufficiently large number of degrees of freedom to validate the use of the asymptotic distributions. More exact tests for cases involving small numbers of degrees of freedom are given in the next section.

7d Tests with Wilks's A Criterion

7d.1 Analysis of Dispersion and the Theoretical Aspects of the A Criterion

In the univariate analysis of variance, tests of significance reduce to the comparison of two independently distributed mean squares. One of the mean squares is an unbiased estimate of the variance to which a single observation in any particular class is subject and is called the error variance. The other is an unbiased estimate only when the null hypothesis which is being tested is correct and may be called the mean square due to deviation from the hypothesis. The test depends only on the individual degrees of freedom of two mean squares.

When each sample supplies \(p\) mutually correlated variables, there are \(p\) total sums of squares and \(p(p - 1)/2\) total sums of products which can be analyzed into various categories. This process, which involves the technique of analyzing the variances and covariances of multiple correlated variables, may be termed the analysis of dispersion. The term dispersion was originally used by P. C. Mahalanobis to indicate the scatter of a set of observations as measured by the variances and covariances. Following this terminology, the total dispersion may be said to be analyzed into dispersion due to various categories.

If we represent the total sums of products by the matrix \(S = (S_{ij})\), then the analysis of dispersion consists in analyzing each element such
as $S_{ij}$, according to the usual procedure, into various categories with the corresponding distribution of degrees of freedom. The dispersion due to any category supplies the sum of products (S.P.) matrix which on division by the degrees of freedom gives the mean product (M.P.) matrix. The S.P. matrix leading to unbiased estimates of the variances and covariances to which a single set of variables is subject is called the S.P. matrix due to error. This error matrix may be denoted by $W$ with $w$ as its degrees of freedom. In the analysis of dispersion the S.P. matrix due to any other category leads to unbiased estimates of variances and covariances only when the null hypothesis regarding that category is true. This may be called the S.P. matrix due to deviation from the hypothesis. If such a matrix is represented by $Q$ with $q$ as its degrees of freedom, then the problem of testing the null hypothesis consists in comparing the matrices $(1/w)W$ and $(1/q)Q$. The simultaneous comparison of the estimates of the variances and covariances appears to be a natural extension of the comparison of variances in the case of a single variate.

The appropriate test criteria for comparison may be obtained by extending the method of discriminant function analysis. A linear compound of the variables is taken, and the compounding coefficients are chosen such that the ratio of mean squares due to deviation from hypothesis and due to error for this variable is a maximum. The ratio $f^2$ which comes out as a root of the determinantal equation

$$\left| Q - \frac{q}{w} f^2 W \right| = 0$$

may be used as the appropriate test criterion. If $|W| \neq 0$, the number of non-zero roots of this equation is equal to the number of variables under consideration or $q$, the number of degrees of freedom of $Q$, whichever is smaller. An adequate comparison of $(1/q)Q$ and $(1/w)W$ must involve the tests of significance of all the roots. If $f_1, f_2, \cdots$ represent the various roots, it is easy to verify that

$$\frac{(1 + \frac{q}{w} f_1^2)(1 + \frac{q}{w} f_2^2) \cdots}{|W + Q|} = \frac{|W + Q|}{|W|}$$

The ratio $|W|/|W + Q|$ denoted by $\Lambda$ decreases as the magnitude of the roots increase, and a significantly small value of $\Lambda$ may be taken as providing the significance of one or more of the roots. This is the underlying theory of the $\Lambda$ criterion arrived at by Wilks (1932) by using the likelihood ratio method and later extended by Bartlett (1934) for general use in multivariate analysis.

However, this does not provide a satisfactory test, for when only one or a smaller number of roots than the total indicate real differences, their
significance may be obscured by the use of the overall test. Its use can be recommended only in situations where small deviations from the hypothesis can be ignored.

7d.2 The Distribution of $\Lambda$ and Its Practical Use

The following notations will be used throughout this and the subsequent sections.

<table>
<thead>
<tr>
<th>Table 7d.2a. Analysis of Dispersion for $p$ Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>(1) Deviation from hypothesis</td>
</tr>
<tr>
<td>(2) Error</td>
</tr>
<tr>
<td>(3) Total</td>
</tr>
</tbody>
</table>

$\Lambda = \frac{|W|}{|W + Q|}$

If the number of variables involved is $p$, then, assuming that the elements of $W$ are distributed independently of those of $Q$, it is easy to derive the $th$ moment of $\Lambda$.

$$E(\Lambda^t) = \prod_{i=0}^{n-p} \frac{\Gamma\left(\frac{1}{2}(n-i)\right)\Gamma\left(\frac{1}{2}(n-q-i)+t\right)}{\Gamma\left(\frac{1}{2}(n-q-i)\right)\Gamma\left(\frac{1}{2}(n-i)+t\right)}$$

The tests based on the exact distributions given by Wilks (1932) and Nair (1939) for some particular cases obtained by a comparison of moments are reproduced below.

<table>
<thead>
<tr>
<th>Nature of the Test</th>
<th>Variance Ratio</th>
<th>Degrees of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q = 1$, for any $p$</td>
<td>$\frac{1 - \Lambda}{\Lambda}$</td>
<td>$\frac{n - p}{p}$</td>
</tr>
<tr>
<td>$q = 2$, for any $p$</td>
<td>$\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}}$</td>
<td>$\frac{n - p - 1}{p}$</td>
</tr>
<tr>
<td>$p = 1$, for any $q$</td>
<td>$\frac{1 - \Lambda}{\Lambda}$</td>
<td>$\frac{n - q}{q}$</td>
</tr>
<tr>
<td>$p = 2$, for any $q$</td>
<td>$\frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}}$</td>
<td>$\frac{n - q - 1}{q}$</td>
</tr>
</tbody>
</table>
For other values of $p$ and $q$, the exact values of the probabilities can be obtained by the use of $\chi^2$ tables alone. Defining

$$V = -m \log_e \Lambda = -\left( n - \frac{p + q + 1}{2} \right) \log_e \Lambda$$

the distribution function of $V$ can be obtained in the asymptotic form

$$P_{pq} + \frac{\gamma_2}{m^2} (P_{pq+4} - P_{pq}) + \frac{1}{m^4}$$

$$\times \left\{ \gamma_4 (P_{pq+8} - P_{pq}) - \gamma_2^2 (P_{pq+4} - P_{pq}) \right\} + \cdots$$

where $P_{pq+r}$ is the distribution function of $\chi^2$ with $(pq + r)$ degrees of freedom. If $m$ is large, the first approximation consists in using $V$ as $\chi^2$ with $pq$ degrees of freedom. For obtaining the second and third approximations the expressions for $\gamma_2$ and $\gamma_4$ are

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

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

In many practical problems the first approximation $V$ suggested by Bartlett can be used. A better approximation is as follows.

Defining the statistic

$$y = \Lambda^{1/s} \quad s = \sqrt{\frac{p^2q^2 - 4}{p^2 + q^2 - 5}}$$

the distribution function of $y$ can be written

$$B \left( \frac{ms}{2} + \lambda, r \right) + \frac{c \Gamma \left( \frac{ms}{2} + \lambda + r \right)}{\Gamma \left( \frac{ms}{2} + \lambda + r + 4 \right)} \left[ B \left( \frac{ms}{2} + \lambda, r + 4 \right) - B \left( \frac{ms}{2} + \lambda, r \right) \right] + \cdots$$
where
\[
\begin{align*}
\epsilon &= \frac{\Gamma(r+4)}{\Gamma(r)} \left\{ \frac{\gamma_4^s}{16r(r+1)(r+2)(r+3) - (r-1)(5r-7)} - \frac{1}{10 \times 10 \times 36} \right\} \\
\gamma &= \frac{pq}{2} \\
\lambda &= -\frac{pq - 2}{4}
\end{align*}
\]
and \(B(t, u)\) is the distribution function of the beta variable. The first term above offers a powerful approximation, the second term being \(O(1/m^4)\). In this case the statistic
\[
1 - A^\frac{1}{m} \frac{ms + 2\lambda}{2r}
\]
can be used as a variance ratio with \(2r\) and \((ms + 2\lambda)\) degrees of freedom. The quantity \((ms + 2\lambda)\) need not be integral.

**7d.3 Test of Differences in Mean Values for Several Populations**

Let \(A_1, A_2, \ldots, A_k\) be \(k\) populations from which samples of sizes \(N_1, \ldots, N_k\) for \(p\) correlated variables are available. The dispersion has to be analyzed into between and within populations. The S.P. matrix due to within populations (or the error) has \((N_1 + \cdots + N_k - k)\) degrees of freedom, and that due to between populations has \((k-1)\) degrees of freedom. If these are represented by \(W\) and \(Q\), then the statistic to be used for testing the differences in mean values is
\[
V = -m \log A
\]
where
\[
\begin{align*}
A &= \frac{\left| W \right|}{\left| W + Q \right|} \\
m &= n - \frac{p + q + 1}{2} \\
n &= N_1 + \cdots + N_k - 1 \\
q &= k - 1
\end{align*}
\]
The exact probability of \(V\)'s exceeding the observed value can be calculated as explained in 7d.2.

*Example 1.* Table 7d.3a gives the analysis of dispersion for the three characters, head length \((x_1)\), height \((x_2)\), and weight \((x_3)\), measured on
DIFFERENCES IN MEAN VALUES

140 schoolboys, of almost the same age, belonging to six different schools in an Indian city.

**Table 7d.3a. Analysis of Dispersion**

<table>
<thead>
<tr>
<th>Dispersion Due to</th>
<th>D.F.</th>
<th>$x_1^2$</th>
<th>$x_2^2$</th>
<th>$x_3^2$</th>
<th>$x_{12}$</th>
<th>$x_{13}$</th>
<th>$x_{23}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between schools</td>
<td>5</td>
<td>752.0</td>
<td>151.3</td>
<td>1,612.7</td>
<td>1003.7</td>
<td>2671.2</td>
<td>4123.6</td>
</tr>
<tr>
<td>Within schools</td>
<td>134</td>
<td>12,809.3</td>
<td>1499.6</td>
<td>21,009.6</td>
<td>214.2</td>
<td>521.3</td>
<td>401.2</td>
</tr>
<tr>
<td>Total</td>
<td>139</td>
<td>13,561.3</td>
<td>1650.9</td>
<td>22,622.3</td>
<td>1217.9</td>
<td>3192.5</td>
<td>4524.8</td>
</tr>
</tbody>
</table>

\[ A = \begin{bmatrix} W \cdot S \cdot S \cdot W \end{bmatrix} \]

\[ V = \log_e A = \log_e (139^{15} \cdot 0.0375) \]

\[ m = 139 - \frac{1}{2}(5 + 3 + 1) = 134.5 \]

\[ V = -m \log_e A = (134.5)(0.193724) = 26.0559 \]

Using \( V \) as \( x^2 \) with \( pq = 15 \) degrees of freedom, the first approximation comes out as

\[ P_{15} = 0.0375 \]

The second term is

\[ \frac{P_2}{m^2} (P_{10} - P_{15}) \]
\begin{align*}
\gamma_2 &= \frac{29 \times 15}{48} \quad \text{and} \quad \gamma_3 = \frac{29 \times 15}{48(131.5)^2} = 0.00050096 \\
\gamma_2 \left( P_{1(9)} - P_{1(8)} \right) &= 0.00050096(0.1285 - 0.0375) = 0.00001574
\end{align*}

This correction to the first approximation affects only the fourth decimal place so that correction is hardly necessary. The observed value of $V$ is significant at the $0.05$ level, showing thereby that boys of various schools differ in physique. This appears to be generally true since boys belonging to different social strata attend different schools.

To use the variance ratio approximation we find $2r = 15$, $s = 2.07$, $m+2 = 352.61$. The variance ratio 1.77 with 15 and 352.61 degrees of freedom is significant at the $0.05$ level.

7d.4 Internal Analysis of a Set of Variates

Let $X_1, \cdots, X_p, X_{p+1}, \cdots, X_{p+s}$ be $(s + p)$ correlated variables for which samples of sizes $N_1, \cdots, N_k$ are available from $k$ populations. If the differences in mean values of these $(s + p)$ variables are to be tested for significance, then the method given in 7d.3 can be used. An important problem that arises in biometry is to test whether the variables, say $X_{s+1}, \cdots, X_{s+p}$, bring out further differences in populations when the differences due to $X_1, \cdots, X_s$ are removed.

It is apparent in problems of this nature that some of the variables in the set $X_1, \cdots, X_s$ might be in the nature of concomitant variables which have been observed in association with the dependent variables or which might have been chosen to have some specified values. An illustration of such an analysis is found in a problem where three dependent variables $g$, $h$, and $i$, corresponding to linear, parabolic, and cubic terms of growth curves of pig weights, are considered together with a concomitant variable giving the initial weight of pigs. It was desired to test whether the variables $h$ and $i$ bring out further differences in food treatments when the differences due to $g$ and $w$ are eliminated.

The problem is identical with that posed above, with $g, w$ forming the first set and $h, i$ the second set of variables.

There is a third set of problems in which it is desired to test whether the differences in $s$ groups characterized by $(s + p)$ measurements can be explained by variations in $s$ assigned linear functions of these measurements. If $y_1, \cdots, y_{s+p}$ are the $(s + p)$ variables and

\begin{align*}
L_1 &= m_1y_1 + \cdots + m_{1+p}y_{1+p} \\
L_2 &= m_2y_1 + \cdots + m_{2+p}y_{2+p}
\end{align*}


are the assigned linear functions, then we can replace the \((s + p)\) variables \(y_1, \ldots, y_{s+p}\) by \(x_1, \ldots, x_{s+p}\), defined by
\[
\begin{align*}
x_1 &= L_1, \quad \ldots, \quad x_s = L_s \\
x_{s+1} &= m_{s+1,1}y_1 + \cdots + m_{s+1,s+p}y_{s+p} \\
&\vdots \\
x_{s+p} &= m_{s+p,1}y_1 + \cdots + m_{s+p,s+p}y_{s+p} 
\end{align*}
\]
where the coefficients in \(x_{s+1}, \ldots, x_{s+p}\) are chosen arbitrarily subject to the condition that the determinant \(\det \{ m_{ij}, \ 1 \leq i, j \leq (s + p) \}\) is not zero. This latter condition ensures that the transformation from the \(y\) to the \(x\) leads to one-to-one correspondence. Once again, the problem is reduced to that of considering the differences in \(x_{s+1}, \ldots, x_{s+p}\) when those due to \(x_1, \ldots, x_s\) are removed. The proposed test is independent of the compounding coefficients used to define the set \(x_{s+1}, \ldots, x_{s+p}\) so that, in any practical problem, they may be conveniently or conventionally chosen.

In all these cases, the problem is one of analyzing the dispersion of the variables \(x_{s+1}, \ldots, x_{s+p}\) when the dispersion due to \(x_1, \ldots, x_s\) is removed. This can be done by following the covariance technique suitable for \(p\) dependent variables and \(s\) independent variables as indicated in 3g.2.

Let
\[
(S_{ij}) = (Q_{ij}) + (W_{ij}), \quad i, j = 1, 2, \ldots, (s + p)
\]
be the analysis of dispersion for all the \((s + p)\) variates due to deviation from hypothesis and error with the corresponding distribution of degrees of freedom as
\[
n' = q + (n' - q)
\]
The S.P. matrix due to error for the variables \(x_1, \ldots, x_s\) to be eliminated is
\[
\begin{bmatrix}
W_{11} & \cdots & W_{1s} \\
\vdots & \ddots & \vdots \\
W_{s1} & \cdots & W_{ss}
\end{bmatrix}
\]
and its inverse is represented by
\[
\begin{bmatrix}
W^{11} & \cdots & W^{1s} \\
\vdots & \ddots & \vdots \\
W^{s1} & \cdots & W^{ss}
\end{bmatrix}
\]
The S.P. matrix due to error for $x_{s+1}, \ldots, x_{s+p}$ when corrected for $x_1, \ldots, x_s$, is given by $W(p+s, 1, \ldots, s+p; 1, \ldots, s)$ or simply $W(p \mid s)$ where

$$W(p \mid s) = \begin{bmatrix}
W_{s+1,s+1} & \cdots & W_{s+1,s+p} \\
\vdots & \ddots & \vdots \\
W_{s+p,s+1} & \cdots & W_{s+p,s+p}
\end{bmatrix}$$

This form, which involves the evaluation of a triple product of matrices, appears to be convenient for computation as illustrated in the next section. Another way of obtaining this matrix $W(p \mid s)$ is to start with the complete matrix $(W_{ij})$, $(i, j = 1, 2, \ldots, s, s+1, \ldots, s+p)$ and reduce it $s$ times by the method of pivotal condensation starting from the element $W_{11}$. Replacing $W$ by $S$, we have the formula for computing the S.P. matrix due to "deviation from hypothesis + error" for $x_{s+1}, \ldots, x_{s+p}$ when corrected for $x_1, \ldots, x_s$. If this is represented by $S(p \mid s)$, then the required criterion is

$$W(p \mid s) \quad S(p \mid s)$$

The degrees of freedom for $W(p \mid s)$ are $(n' - q - s)$, and that for $S(p \mid s)$ are $(n' - s)$, so that in standard notation the parameters associated with $A$ are

$$n = n' - s \quad p = p' \quad q = q$$

The test can be carried out as discussed in 7d.2.

7d.5 Barnard's Problem of Secular Variations in Skull Characters

The problem of measuring secular variations in skull characters considered by Barnard (1935) is of immense importance to the anthropologists. It is, however, of interest to examine the methods employed by her in the light of the latest developments in multivariate analysis. The two problems involved in her study are:
The selection of a smaller number, out of seven skull characters, which give significant information, as far as is possible, as to changes taking place with time in four series of Egyptian skulls; and

(ii) The determination of an expression, linear in measurements, which characterizes most effectively an individual skull with respect to the progressive secular changes.

To answer problem (i), Barnard first chose basialveolar length and nasal height as two basic characters which, independently of each other, show significant variation in the four series. To choose further characters she considered the problem of testing the significance of the linear regression of the mean values of an added character with time (corresponding to the four series) when that part of the regression due to the two basic characters is removed. This meant the choice of characters with special reference to the average linear rate of change of the individual means with time. If the choice of characters is to be with reference to the complete nature of changes taking place with time, then what is needed is an internal analysis of the characters to decide whether the configuration of the four series as determined by several characters is the same as that indicated by a smaller number. Barnard's method should, of course, be preferred if the regressions were known to be linear. This can, however, be tested from the data.

Taking the four measurements

\[
\begin{align*}
X_1 &= \text{basialveolar length} \\
X_2 &= \text{nasal height} \\
X_3 &= \text{maximum breadth} \\
X_4 &= \text{basibregmatic height}
\end{align*}
\]

the relevant data are summarized in Tables 7d.5a and 7d.5b, which give the means for the four series and the analysis of dispersion.

<table>
<thead>
<tr>
<th>Character</th>
<th>Series</th>
<th>( N_1 = 91 )</th>
<th>( N_2 = 162 )</th>
<th>( N_3 = 70 )</th>
<th>( N_4 = 75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>133.582418</td>
<td>134.265432</td>
<td>134.371429</td>
<td>135.306667</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>98.307692</td>
<td>95.462963</td>
<td>95.857143</td>
<td>95.040000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>50.835165</td>
<td>51.148148</td>
<td>50.100000</td>
<td>52.093333</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>133.000000</td>
<td>134.882716</td>
<td>133.648978</td>
<td>131.499667</td>
</tr>
</tbody>
</table>
Table 7d.5. Analysis of Dispersion (S. P. Matrix)

<table>
<thead>
<tr>
<th>Dispersion due to</th>
<th>Between</th>
<th>Within</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 D.F.</td>
<td>394 D.F.</td>
<td>397 D.F.</td>
</tr>
<tr>
<td>( X_1 )</td>
<td>123.180628</td>
<td>9661.997470</td>
<td>9785.178088</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>486.343863</td>
<td>5073.115027</td>
<td>5559.458890</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>100.411505</td>
<td>3938.320351</td>
<td>4038.731856</td>
</tr>
<tr>
<td>( X_4 )</td>
<td>640.733891</td>
<td>8741.508820</td>
<td>9382.242720</td>
</tr>
<tr>
<td>( X_1 X_2 )</td>
<td>-231.375635</td>
<td>445.573301</td>
<td>214.197666</td>
</tr>
<tr>
<td>( X_1 X_3 )</td>
<td>87.305348</td>
<td>1130.623000</td>
<td>1217.928348</td>
</tr>
<tr>
<td>( X_1 X_4 )</td>
<td>-128.763484</td>
<td>2148.584210</td>
<td>2019.820924</td>
</tr>
<tr>
<td>( X_2 X_3 )</td>
<td>-107.505618</td>
<td>1239.221990</td>
<td>1131.716372</td>
</tr>
<tr>
<td>( X_2 X_4 )</td>
<td>125.313318</td>
<td>2255.812722</td>
<td>2381.126040</td>
</tr>
<tr>
<td>( X_3 X_4 )</td>
<td>-137.580764</td>
<td>1271.054662</td>
<td>1133.473898</td>
</tr>
</tbody>
</table>

Example 1. Do the characters \( X_3 \) and \( X_4 \) show significant variation in the four series independently of the variation due to the characters \( X_1 \) and \( X_2 \)?

The method developed in 7d.4 is directly useful in this problem. The S.P. matrix within for the basic characters \( X_1 \) and \( X_2 \) is

\[
\begin{pmatrix}
9661.997470 & 445.573301 \\
445.573301 & 9073.115027
\end{pmatrix}
\]

Its inverse is

\[
\begin{pmatrix}
1.037332 & -0.050942 \\
-0.050942 & 1.104659
\end{pmatrix}
\]

The within S.P. matrix for \( X_3, X_4 \) due to \( X_1, X_2 \) is given by the triple product

\[
\begin{pmatrix}
W_{13} & W_{14} \\
W_{14} & W_{24}
\end{pmatrix}
\begin{pmatrix}
W_{11} & W_{12} \\
W_{21} & W_{22}
\end{pmatrix}
\begin{pmatrix}
W_{13} & W_{14} \\
W_{14} & W_{24}
\end{pmatrix}
\]

\[
= 10^{-4}
\begin{pmatrix}
1130.623900 & 1239.221990 \\
2148.584210 & 2255.812722
\end{pmatrix}
\begin{pmatrix}
W_{11} & W_{12} \\
W_{21} & W_{22}
\end{pmatrix}
\begin{pmatrix}
W_{13} & W_{14} \\
W_{14} & W_{24}
\end{pmatrix}
\]

\[
= 10^{-4}
\begin{pmatrix}
1109.70904 & 1311.321492 \\
2113.879535 & 2382.450625
\end{pmatrix}
\begin{pmatrix}
W_{13} & W_{14} \\
W_{14} & W_{24}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
287.967620 & 534.238796 \\
534.238796 & 991.621041
\end{pmatrix}
\]
SECULAR VARIATIONS IN SKULL CHARACTERS

The within S.P. matrix for $X_3$ and $X_4$ after correcting for $X_1$ and $X_2$ is

$$
\begin{pmatrix}
W_{33} & W_{34} \\
W_{43} & W_{44}
\end{pmatrix} - \begin{pmatrix}
W_{13} & W_{14} \\
W_{14} & W_{24}
\end{pmatrix} \begin{pmatrix}
W_{11} & W_{12} \\
W_{21} & W_{22}
\end{pmatrix} \begin{pmatrix}
W_{13} & W_{14} \\
W_{23} & W_{24}
\end{pmatrix}
$$

$$
= \begin{pmatrix} 3938.320351 & 1271.054662 \\
1271.054662 & 8741.508829 \end{pmatrix} - \begin{pmatrix} 287.967620 & 534.238796 \\
534.238796 & 991.621041 \end{pmatrix}
$$

$$
= \begin{pmatrix} 3650.353731 & 736.815866 \\
736.815866 & 7749.887788 \end{pmatrix} = W(2 \mid 2)
$$

This has $394 - 2 = 392$ degrees of freedom. Similarly, $S(2 \mid 2)$ with $397 - 2 = 395$ degrees of freedom is

$$
\begin{pmatrix} 3809.335190 & 611.798381 \\
611.798381 & 8393.755848 \end{pmatrix}
$$

$$
\Lambda = \begin{vmatrix} W(2 \mid 2) \\ S(2 \mid 2) \end{vmatrix} = \begin{pmatrix} 0.27746934 \\ 0.31600332 \end{pmatrix} = 0.878058
$$

$$
V = -m \log \Lambda \quad m = n - p + q + 1 = 395 - 2 + 3 + 1 = 392
$$

$$
V = -392 \log (0.878058) = 51.39
$$

This value of $V$ with $pq$ equal to 6 degrees of freedom is significant so that $X_3$ and $X_4$ may be considered as discriminating the series independently of $X_1$ and $X_2$.

The above method could be simplified by starting with the full matrices $W$ and $S$ and reducing them by the method of pivotal condensation. The four pivotal elements for $W$ are

$$
10^4(0.966200, 0.903257, 0.365033, 0.760117)
$$

and for $S$

$$
10^4(0.978358, 0.955477, 0.380933, 0.829550)
$$

The value of $\begin{vmatrix} W(2 \mid 2) \end{vmatrix}$ is the product of the last two pivotal elements

$$
10^8(0.365033)(0.760117) = 0.277469 \times 10^8
$$

Similarly, $\begin{vmatrix} S(2 \mid 2) \end{vmatrix} = 0.316003 \times 10^8$. Thus we obtain the same value of $\Lambda$ as above.

Example 2. Taking the relative times between the series in the proportion 2:1:2, can the variation of the characters be accounted for by the linear regression of individual characters with time?
In order to obtain the regression with time, the values of $t$, the time variable, may be taken as $-5$, $-1$, 1, and 5 for the individuals of the first, second, third, and fourth series, respectively. The calculation of individual regressions involves the quantities,

$$
\Sigma (t - \bar{t})^2 = 4307.66832
$$

$$
\Sigma x_1(t - \bar{t}) = 718.76286 \quad \Sigma x_2(t - \bar{t}) = -410.10194
$$

$$
\Sigma x_3(t - \bar{t}) = -1407.26075 \quad \Sigma x_4(t - \bar{t}) = -733.42758
$$

The matrix $R$ with 1 degree of freedom giving the squares and products due to regression is given in Table 7d.5y.

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>119.030358</td>
<td>-234.810812</td>
<td>68.428235</td>
<td>-122.377258</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-234.810812</td>
<td>459.734449</td>
<td>-133.975163</td>
<td>-149.601596</td>
</tr>
<tr>
<td>$x_3$</td>
<td>68.428235</td>
<td>-133.975163</td>
<td>39.042852</td>
<td>-69.824358</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-122.377258</td>
<td>-149.601596</td>
<td>-69.824358</td>
<td>124.874099</td>
</tr>
</tbody>
</table>

In the above table

$$
R_{11} = \frac{[\Sigma x_1(t - \bar{t})]^2}{\Sigma (t - \bar{t})^2}
$$

$$
R_{12} = \frac{[\Sigma x_1(t - \bar{t})][\Sigma x_2(t - \bar{t})]}{\Sigma (t - \bar{t})^2}
$$

and so on

With these results we analyze the dispersion of which a typical product $(x_1 x_2)$ is chosen below for illustration. To test the hypothesis that the regressions are linear we compare $W$ and $Q + W$.

<table>
<thead>
<tr>
<th></th>
<th>D.F.</th>
<th>S.P. Matrix $(x_1 x_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>-234.810812 $(R_{11})$</td>
</tr>
<tr>
<td>Deviation from regression</td>
<td>2</td>
<td>3.430177 $(Q_{11})$</td>
</tr>
<tr>
<td>Total (between series)</td>
<td>3</td>
<td>-231.376635 $(R_{11} + Q_{11})$</td>
</tr>
<tr>
<td>Within series</td>
<td>394</td>
<td>445.573201 $(W_{11})$</td>
</tr>
<tr>
<td>Total</td>
<td>397</td>
<td>214.197666 $(S_{11})$</td>
</tr>
<tr>
<td>Deviation from regression</td>
<td>396</td>
<td>449.008478 $(Q_{11} + W_{11})$</td>
</tr>
<tr>
<td>Within series</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Obtained by subtraction. The complete matrix $(Q_{ij} + W_{ij})$ obtained by the above method is given in Table 7d.5e. This is the total S.P. matrix minus the matrix $R$ due to regression.
REFERENCES

TABLE 7d.5a. Matrix \((Q + W)\) with 396 Degrees of Freedom

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>966.247740</td>
<td>449.008478</td>
<td>1149.501013</td>
<td>2142.197474</td>
</tr>
<tr>
<td>2</td>
<td>449.008478</td>
<td>9099.726441</td>
<td>1265.691535</td>
<td>2231.524444</td>
</tr>
<tr>
<td>3</td>
<td>1149.501013</td>
<td>1265.691535</td>
<td>4049.689004</td>
<td>1203.298256</td>
</tr>
<tr>
<td>4</td>
<td>2142.197474</td>
<td>2231.524444</td>
<td>1203.298256</td>
<td>9257.368621</td>
</tr>
</tbody>
</table>

\[
\Lambda = \frac{|W|}{|Q + W|} = \frac{0.24209054 \times 10^{12}}{0.26873816 \times 10^{12}} = 0.90307436
\]

\[
V = - \left( 396 - \frac{2 + 4 + 1}{2} \right) \log_e (0.90307436)
\]

= 40.02

The \(\chi^2\) approximation has \(p \times q = 2 \times 4\) degrees of freedom, since \(Q\) has 2 degrees of freedom and there are four variables. The result is significant so that the regressions cannot be considered linear.

This test can be extended to examine whether a parabolic regression with time can explain the differences in mean values. The matrix \(Q\) giving the deviation from regression has then 1 degree of freedom and \(R\) due to regression 2.

To determine the coefficients of a linear compound which characterizes most effectively the secular changes in progress, Barnard maximized the ratio of the square of unweighted regression of the compound with time. It is doubtful whether such a linear compound can be used to specify an individual skull most effectively with respect to progressive changes, since linear regression with time does not adequately explain all the differences in the four series.

References


272 TESTS OF SIGNIFICANCE IN MULTIVARIATE ANALYSIS


CHAPTER 8

Statistical Inference Applied to Classificatory Problems

8a Tests of Null Hypotheses

8a.1 Problems in Biological Research

There are two types of problems confronted in biological research. The first is that of specifying an individual as a member of one of many groups to which he can possibly belong, as when a taxonomist has to assign an organism to its proper species or subspecies or an anthropologist is faced with the problem of sexing a skull or a jawbone. The second is the problem of classification of the groups themselves into some significant system based on the configuration of the various characteristics. The need of this is felt in the study of systematics and the evolution of species. A number of species or subspecies may have to be arranged in a hierarchical order showing the closeness of some and the distinctness of others. Such a representation superimposed on a geographical classification may, it is suggested, be of use in tracing the evolution of various species or subspecies.

The solution of these problems requires the development of a suitable theory of statistical inference and the formulation of some practical rules of procedure which the biologist can profitably use.

To start, it is useful to distinguish problems of discrimination from those of testing of hypotheses. Recently there has been a tendency to treat both these problems on an equal footing, and this has no doubt caused a good deal of confusion. In testing of hypotheses we have a clearly stated null hypothesis and a comparatively undefined set of alternatives. The emphasis is more on the null hypothesis, which may be rejected or provisionally accepted. When a null hypothesis is rejected no decision is made about the actual alternative hypothesis. But in problems of discrimination we have a class of alternative hypotheses out of which one has to be chosen. Although it is a question of rejecting the null hypothesis at a given risk in the former problem, it is a question
of balancing between wrong and correct decisions in the latter problem. Although the a priori probabilities have no place, even conceptually, in problems of testing a null hypothesis, they are essential for a satisfactory solution of the problems of discrimination. In all scientific investigations both problems are important.

8a.2 Null Hypotheses

Consider the following problem, which frequently crops up in biological research.

A specimen is observed, and it is desired to know, on the basis of some morphological measurements, whether it belongs to a previously classified group whose characteristics are either known or estimated from a sample of individuals from that group. In such a problem there are only two possibilities: the new find belongs either to a known group or to an unknown group. The alternatives to the specified one are obviously undefined. The new group might be one whose existence has yet to be established.

Thus, when a fossil is discovered the paleontologist inquires whether it is a specimen from a known collection. Such an inquiry is often made with the hope of obtaining a negative answer, in which case the fossil could be taken as a new specimen.

The investigator may not be successful in distinguishing a new specimen from a previous collection because the answer depends on the evidence supplied by the observed specimen. The only safeguard offered by a statistical test in such a case is that it checks the investigator from rushing to a hasty conclusion unless the evidence is strong enough. If specimens like the observed or differing to a greater extent than the observed from the characteristics of a known group form a reasonable proportion in the group, then the evidence for rejecting the null hypothesis cannot be considered conclusive. Only when this proportion is small can we take the risk of asserting that the observed specimen belongs to a new group. How small this proportion or level of significance should be depends on the risk involved in asserting that the null hypothesis is wrong when in fact it is true. The choice of level of significance is arbitrary in this sense,* but once it is fixed the rule of procedure is determined exactly. Thus it may be possible to refute any statement made about the observed specimen. Such an inference is possible only when some risk is allowed.

* It is not arbitrary in the sense that we are assuming one value when in fact it should be something else. It is one which is chosen by the investigator. Thus if the consequences of rejecting a true hypothesis involve a great loss it is reasonable to keep the level of significance as low as possible.
On the other hand, it is almost impossible to assert that the new find belongs to a specified group. To make such a statement we must ascertain whether the chance of the observed specimen's arising from any other group is small. This is clearly not possible when the alternative groups are undefined ones.

There is clearly no scope for the introduction of a priori probabilities in this case. However perfect our past knowledge may be about the species that have been already studied and their relative numbers, nothing can be said about the new species to be discovered. When these new species are considered as alternatives to a null hypothesis tested, there is no method of attaching a priori probabilities to the alternatives.

Sometimes the a priori probabilities are introduced not as objective quantities measured by observed frequencies but as measuring merely psychological tendencies. If this is so we need further rules of procedure for choosing the a priori probabilities themselves. One can recall the efforts made by Jeffreys (1948) in this connection. To remove some apparent contradictions in Bayes's postulate of equal ignorance, Jeffreys advocates the use of certain invariant functions of the parameters occurring in a probability distribution as a priori weights. Even here no argument is put forward for using particular invariant functions of the parameters. In fact, different choices lead to different results so that no objective theory could be built up on the lines of inverse probability.

To take another example, a geneticist inquires on the basis of observed data whether two factors are segregating independently. If he can disprove this with some confidence, then he acquires some basis on which to plan future experiments, to estimate the intensity of their linkage, and to study the relationship of the two factors under consideration with others. If data are not sufficiently numerous, loose linkages go undetected and it is only by repeated experimentation and accumulation of evidence supplied by other factors linked with the former that some definite conclusion can be arrived at.

The alternative to the hypothesis of independence in the above problem is linkage with all possible values of the recombination fraction (lying between 0 and 1). To the experimenter it is definite knowledge if he can disprove the hypothesis of independence. Only then will he proceed to inquire what the value of the recombination fraction is and try to obtain an estimate. To ask for a priori probabilities of the alternative recombination fractions before attempting to answer the problem posed is to believe that from previous experience the frequencies with which various recombination fractions occur can be deduced. But
there may not be sufficient reason to believe that the frequencies so derived correspond to the total frequencies obtainable from all possible factors known and unknown.

In the problem of the paleontologist the alternatives are completely undefined whereas in the problem of the geneticist the alternatives are known, viz., that the recombination fraction lies between 0 and 1. But in both types of problems there is no scope for the introduction of a priori probabilities.

The null hypothesis is one which is chosen by the experimenter appropriate to his inquiry. When sufficient evidence gathers against this during the experimental work, he rejects it. He is not trying to balance between the evidences supplied by the data on the various alternatives.

Whether a particular null hypothesis is rejected or not, there is a class of null hypotheses which are not contradicted by the data at a given level of significance. Any hypothesis outside this class is rejected. The class of null hypotheses acceptable to the data supplies us with what may be called a fiducial set. When the hypotheses refer to the values of a parameter, the fiducial set will be in the nature of an interval called the fiducial interval (Fisher, 1947). The fiducial set of hypotheses may be asserted to contain the true hypothesis because the chance of its being left out is small (equal to the percentage level of significance chosen). Thus, although it is not possible to accept any single hypothesis, it is possible to restrict the scope of inquiry to only a subset of all possible alternatives. Any further discrimination among the alternatives in the fiducial set has necessarily to be based on insufficient evidence. No statement of confidence can be made about a single hypothesis chosen by any rule of procedure as the most appropriate for the data, and consequently such a procedure does not possess a scientific basis of inference.

If the problem needs the choice of a single hypothesis, then what should be the nature of the answer? We might try to formulate a rule of procedure which selects a hypothesis which is as near as possible to the true hypothesis and which in large samples differs very little from the true one with probability approaching certainty. The procedure of choosing that hypothesis which maximizes the likelihood, advocated by Fisher, conforms to the above requirement to a large extent. Thus the two methodological problems, testing of hypothesis and estimation, admit neat solutions independent of the probabilities a priori.

As an example for the determination of the fiducial interval, consider a sample \( x_1, \ldots, x_n \) from a normal population. If \( \mu \) is the true mean value, then
\[
I = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]
POWER FUNCTION OF NEYMAN AND PEARSON

is distributed as $t$ with $(n - 1)$ degrees of freedom. All values of $\mu$ not acceptable to the data at the 5% level of significance satisfy the inequality

$$\left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| \geq t_{0.05}$$

where $t_{0.05}$ is the 5% significant value of $t$. This gives two values

$$\bar{x} - \frac{st_{0.05}}{\sqrt{n}} \quad \text{and} \quad \bar{x} + \frac{st_{0.05}}{\sqrt{n}}$$

beyond which all values of $\mu$ are incompatible with the observed data. In such a case we could assert subject to a small risk that the true value of $\mu$ lies in the above interval.

Similarly, if the fiducial interval for $\sigma^2$ is needed, then two equations are considered:

$$\frac{(n - 1)s^2}{\sigma^2} = x_1^2 \quad \text{and} \quad \frac{(n - 1)s^2}{\sigma^2} = x_2^2$$

giving

$$\sigma_1^2 = \frac{(n - 1)s^2}{x_1^2} \quad \text{and} \quad \sigma_2^2 = \frac{(n - 1)s^2}{x_2^2}$$

where $(x_1^2, x_2^2)$ is the critical interval given in Table 6a.1 for $\chi^2$ with $(n - 1)$ degrees of freedom.

8a.3 Power Function of Neyman and Pearson

Various attempts have been made to build up a consistent theory from which all tests of significance can be deduced as solutions to precisely stated mathematical problems. It is difficult to argue whether such a theory exists or not, but formal theories leading to a clear understanding of the problems are nonetheless important. One such theory, contributed by Neyman and Pearson (1933), is an important development because it unfolded the various complex problems in testing of hypotheses and led to the construction of general theories in problems of discrimination, sequential tests, etc.

Any rule of procedure by which we can reject or accept a given hypothesis $H_0$ consists in a division of all possible samples into two groups, one opposed to $H_0$ and the other not unfavorable to it. Whenever a sample of the first category occurs, we reject the hypothesis $H_0$. As observed earlier, the frequency of the samples in this category, when $H_0$ is true, ought to be small so that the chance of rejecting the hypothesis when it is true is small. Let this chance be fixed as $\alpha$ (a small assigned quantity). Corresponding to any procedure such as the
above, there is a frequency with which the samples of the first category appear under a different hypothesis $H$. This frequency, denoted by $\beta(H)$, is called the power of the test procedure associated with an alternative hypothesis $H$.

This function $\beta(H)$ is fundamental in the theory of Neyman and Pearson. It gives us the frequency with which various alternative hypotheses could be detected, or, to be more exact, the frequency with which the hypothesis $H_0$ is rejected when a different hypothesis $H$ is true. If the sample is to be given a fair chance of rejecting $H_0$, when it is not true, the division of the samples must be such that the frequency of those in the first category is as high as possible under any different hypothesis. To start with, let us determine the maximum possible frequency of detection associated with a given alternative hypothesis.

Let $f_H(x)$ denote the probability density of the observations $x$ under any hypothesis $H$. The sample observations $x_1, \ldots, x_n$ may be represented by a point in a space of $n$ dimensions, in which case the rule of procedure suggested above results in a division of the space into two regions, $w$ for rejecting the hypothesis and the rest for not favoring any alternative. Then, for a given $H$, what is $w$ such that

$$\int_{w} f_H(x) \, dv = a \quad \text{(an assigned quantity)} \tag{8a.3.1}$$

and

$$\int_{w} f_H(x) \, dv \quad \text{is a maximum?}$$

Applying the result of lemma A1 in Appendix A, we find the best region $w_0$ is defined by

$$f_H \geq \lambda f_{H_0} \quad \text{inside } w_0$$

and

$$f_H \leq \lambda f_{H_0} \quad \text{outside } w_0$$

in which case the maximum $\beta(H)$ is

$$\beta(H) = \int_{w_0} f_H(x) \, dv$$

where $\lambda$ is determined to satisfy 8a.3.1.

We are in a happy situation if the same region $w_0$ makes $\beta(H)$ a maximum for all $H$. In this situation $w_0$ is independent of $H$, and the knowledge of any particular alternative $H$ which may be true does not help us in improving the test. The $w_0$ satisfying this property is said to be a uniformly most powerful critical region, and when this exists the test procedure is above criticism since nothing has been assumed about
the alternatives. The existence of such tests can be easily verified because in this case the boundary of the critical region, \( f_H / f_{H_0} = \text{constant} \), can be expressed without the use of any unknown quantities entering in \( f_H \).

Example 1. Consider \( n \) independent observations from a normal distribution \( N(\mu, \sigma^2) \). Let the null hypothesis be \( \mu = \mu_0 \).

\[
f_H = c \exp \left( -\frac{\sum (x_i - \mu)^2}{2\sigma^2} \right)
\]

\[
f_{H_0} = c \exp \left( -\frac{\sum (x_i - \mu_0)^2}{2\sigma^2} \right)
\]

\[
\log \frac{f_H}{f_{H_0}} = \frac{\mu - \mu_0}{\sigma^2} (x_1 + \cdots + x_n) + n \frac{\mu_0^2 - \mu^2}{2\sigma^2}
\]

The relationship

\[
\frac{f_H}{f_{H_0}} \geq \lambda
\]

reduces to

\[
x(\mu - \mu_0) \geq k
\]

or

\[
x \geq k_1 \quad \text{if } \mu > \mu_0
\]

\[
x \leq k_2 \quad \text{if } \mu < \mu_0
\]

A uniformly most powerful test exists only when it is known that the alternative value \( \mu \) is greater or smaller than the assigned value \( \mu_0 \). The test simply depends on the distribution of the mean, \( \bar{x} \), on the null hypothesis. The distribution of \( \bar{x} \) is

\[
c \exp \left( -\frac{n(\bar{x} - \mu_0)^2}{2\sigma^2} \right) d\bar{x}
\]

which involves another parameter \( \sigma^2 \) so that the test can be carried out only when the hypothetical value of the standard deviation is known. When it is not known, a suitable device is necessary to make the test independent of \( \sigma \).

Example 2. The best region for testing the hypothesis \( H_0 \) against a single alternative \( H \) is bounded by the surface of a constant value of a function of the minimal set of sufficient statistics.

When a set of sufficient statistics \( T_1, T_2, \cdots, T_k \) exists,

\[
f_H = P(T \mid H)P(x \mid T)
\]

and

\[
f_{H_0} = P(T \mid H_0)P(x \mid T)
\]
and hence the ratio $f_H/f_{H_0}$ is equivalent to $P(T \mid H)/P(T \mid H_0)$ which is a function of $T_1, \ldots, T_k$ only.

Example 3. The probability of $r$ successes in $n$ trials of a binomial population with proportion $p$ is

$$\binom{n}{r} p^r q^{n-r}$$

If the null hypothesis is $p = p_0$, then $\log (f_H/f_{H_0}) \geq \log \lambda$ reduces to

$$r \left( \log \frac{p}{p_0} - \log \frac{q}{q_0} \right) \geq k$$

or

$$r \geq k_1 \quad \text{if} \quad p \geq p_0$$

and

$$r \leq k_2 \quad \text{if} \quad p \leq p_0$$

so that the uniformly most powerful test exists only when it is known that the alternative is greater or smaller than the assigned value.

In this example it is presumed that the best test is offered by the ratio $f_H/f_{H_0}$ even when the stochastic variable is discontinuous. The difficulty arises owing to the fact that there may not exist a $\lambda$ such that the probability that $f_H \geq f_{H_0}$ is exactly equal to the assigned value, the percentage level of significance. The setup appropriate for discrete probability densities is to determine a class of events $E$ such that

$$P_H(E) \leq \alpha$$

and

$$P_H(E) \quad \text{is a maximum}$$

The class of events so determined constitutes the critical set, and the happening of any event in this set disproves the null hypothesis. It is easy to see that in the case of continuous distributions the equality relation in (8a.3.2) is attained. Under the new setup the events in the critical set are those for which the ratio $(f_H/f_{H_0}) \geq \lambda$, where $\lambda$ is the minimum value such that the total probability of the events on the null hypothesis does not exceed $\alpha$. The proof is similar to that in the case of continuous distributions.

8a.4 Locally Most Powerful Unbiased Tests

Uniformly most powerful tests exist very rarely so that in most cases there will not be a single region which is the best for all alternative hypotheses. As a first step in making the test independent of the alternative hypotheses, Neyman and Pearson introduced the concept of the
locally most powerful unbiased test, applicable to cases where the hypothesis is specified by the value of a parameter occurring in the probability distribution. Assuming differentiation under the integral sign, the solution depends on the existence of a region $w$ such that

$$\int_w f(x \mid \theta_0) \, dx = \alpha \quad (8a.4.1)$$

$$\int_w f'(x \mid \theta_0) \, dx = 0 \quad (8a.4.2)$$

and

$$\int_w f''(x \mid \theta_0) \, dx \quad \text{is a maximum} \quad (8a.4.3)$$

$\theta_0$ is the value of the parameter under the null hypothesis.

It follows from the lemma in A1 (Appendix A) that a region $w_0$

inside which

$$f'(\theta_0) \geq k_1 f(\theta_0) + k_2 f'(\theta_0)$$

outside which

$$f'(\theta_0) \leq k_1 f(\theta_0) + k_2 f'(\theta_0)$$

where $k_1$ and $k_2$ are determined to satisfy the conditions (8a.4.1) and (8a.4.2), maximizes the integral in (8a.4.3).

This ensures maximum power only for alternatives in the immediate neighborhood of the null hypothesis. This is not a good solution unless the power is quite high for alternatives more distant from the null hypothesis also. In fact, if a locally most powerful test has a very low power beyond a certain range near the null hypothesis, no investigator would be tempted to use it. There is no provision in the method of derivation of a locally powerful test to safeguard against this. This method therefore cannot be considered as general but can be regarded only as a means by which test criteria can be derived for possible comparison with any other offered test procedure.

**Example 1.** The locally most powerful test for the null hypothesis $\sigma = \sigma_0$ is defined by

$$s^2 \geq s^2_1 \quad \text{and} \quad s^2 \leq s^2_2$$

such that

$$\left( \frac{n_s}{\sigma_0^2} \right)^{n_s/2} e^{-n_s/2s^2_1} = \left( \frac{n_s}{\sigma^2} \right)^{n_s/2} e^{-n_s/2s^2_2}$$

and

$$\int_0^{s_1^2} P(s^2) \, ds^2 + \int_{s_1^2}^{s_2^2} P(s^2) \, ds^2 = \alpha$$
where

\[ P(\hat{s}^2) = \text{const.} \cdot e^{-n\hat{s}^2/2} (\hat{s}^2)^{(n-2)/2} \, d\hat{s}^2 \]

and \( \hat{s}^2 \) is the estimate of \( s^2 \) based on \( n \) degrees of freedom.

Example 2. Among locally unbiased tests the power for any alternative value \( \sigma \) is greatest in the above case.

Example 3. What is the locally unbiased most powerful test for a given ratio of two hypothetical variances estimated from two independent samples from normal populations.

[Hint: Start with the variance ratio \( F \) distribution on \( n_1, n_2 \) degrees of freedom, assuming a hypothetical ratio \( \rho \). The best test leads to the condition that a quadratic expression in \( F \) is greater than zero, so that the test is \( F \geq F_1 \) and \( F \leq F_2 \). To determine the relation between \( F_1 \) and \( F_2 \), express the condition of unbiasedness. This leads to a condition of the form \( F_1/(1 + cF_1) = F_2/(1 + cF_2) \).]

Example 4. Prove that among locally unbiased tests the test derived in example 3 is uniformly most powerful.

The test derived in examples 1 and 2 is used in Chapter 6 in testing whether a calculated variance is in agreement with an assigned value. The test derived in examples 3 and 4 is useful in testing whether two estimated variances are equal in their expectation. In Chapter 6, a different test based on the \( L \) statistic was used. But these two are equivalent.

8a.5 Test for a Finite Number of Alternatives

Consider a null hypothesis \( H_0 \) and a set of alternatives \( H_1, H_2, \ldots \). Let the power of the best possible test for \( H_0 \) when \( H_i \) is the only alternative be denoted by \( \gamma_i(\alpha) \), where \( \alpha \) denotes the level of significance. Any region \( w \) suggested as the critical region for testing \( H_0 \) will have

\[ \int_w d\theta = \beta(\alpha) \]

as the power for the alternative \( H_i \). In no case can \( \beta \) exceed \( \gamma \), but there may exist a single region such that \( \beta_i = \gamma_i \) for all \( i \), in which case a uniformly most powerful test exists.

If this is not so, various alternatives have been suggested. One is to choose a region which maximizes the minimum \( \beta \) (Neyman and Pearson, 1933). Such a procedure may give undue preference to the hypotheses nearer \( * \) to the null hypothesis. It may be felt that a method

* A hypothesis \( H_i \) can be said to be nearer than \( H_j \) to \( H_0 \) if \( \gamma_i < \gamma_j \). With this concept a suitable distance function between two hypotheses can be defined as shown in Chapter 9.
which effectively controls the errors of not accepting a nearer hypothesis when it is true will be good enough for distant hypotheses.

On the other hand, we may take the view that in the course of experimentation it is necessary to detect a distant hypothesis as early as we can. If, in fact, a distant hypothesis were true and the critical region had been so chosen as to give this hypothesis the maximum possible power, then it could be discovered with the minimum possible number of observations. If a nearer hypothesis were true, a larger experiment would be necessary to detect it. In such a case the experimenter might consider himself unlucky on the choice of his subject or might regard the consequences of accepting \( H_0 \) when, in fact, an alternative close to it is true as less serious than when the alternative is distant.

A compromise solution may be suggested if the experimenter can assign a priori probabilities for the various alternatives. This means that he has a knowledge of a series of similar experiments and the frequencies of various types of alternatives. When such a knowledge is imperfect or if the experimenter is not sure that the particular experiment he is conducting belongs to the same group of experiments that have been conducted before, no unique solution is possible. In the absence of any information about the a priori probabilities, as a compromise between the two views of maximizing the minimum power or giving more weight to distant hypotheses, the following solution is suggested.

The critical region is chosen such that the common ratio

\[
\frac{\beta_1(a)}{\gamma_1(a)} = \frac{\beta_2(a)}{\gamma_2(a)} = \cdots
\]

is a maximum where \( \beta \) and \( \gamma \) are as defined above. This method supplies a system of weights to be attached to the powers due to various alternatives, the weights being the individual maximum powers. This region has the following two properties.

(i) The distant hypotheses have necessarily more power than the nearer hypotheses.

(ii) The individual maximum powers are now reduced in the same proportion with the provision that this proportion is as small as possible.

If \( f_0, f_1, f_2, \cdots \) denote the probability densities for the hypotheses \( H_0, H_1, H_2, \cdots \), then the region satisfying the above requirements is deducible from the lemma proved in Appendix A4. The inside of this region \( w \) is defined by

\[
f_0 \leq \lambda_1 f_1 + \lambda_2 f_2 + \cdots
\]
where $A_1, A_2, \ldots$ are determined from the relations

\[
\int_{-\infty}^{\infty} f_A \, dv = \alpha \quad \text{(8a.5.1)}
\]

and

\[
\frac{1}{\gamma_1} \int_{-\infty}^{\infty} f_A \, dv = \frac{1}{\gamma_2} \int_{-\infty}^{\infty} f_A \, dv = \cdots \quad \text{(8a.5.2)}
\]

The solution deduced above is not useful in practice because of the difficulty in evaluating the constants. It may be convenient to consider the region complementary to

\[
f_0 \geq \mu f_1 \quad \text{(8a.5.3)}
\]

as the critical region, the quantities $\mu_1, \mu_2, \ldots$ being determined to satisfy the relations (8a.5.1) and (8a.5.2).

**8a.6 Tests When the Alternatives Are Continuous**

The foregoing theory could be extended to the case where the alternatives can be specified by parameters with continuous variation. The following definitions will be useful.

A region $W$ which gives equal power to all hypotheses equidistant, i.e., having the same power of detection from the null hypothesis, is called the distance power region. A test based on a distance power region $W_0$ is said to be uniformly the best distance power test if:

(i) The size of the region $W_0$ with respect to the null hypothesis is $\alpha$ (an assigned value).

(ii) $W_0$ is a distance power region.

(iii) For any specified alternative hypothesis the power associated with the region $W_0$ is not less than the power for any other region satisfying requirements (i) and (ii).

Let $\Delta$ denote the distance of a hypothesis $H$ from $H_0$, the null hypothesis. Then a distance power region satisfies the condition

\[
\int_{-\infty}^{\infty} f_H \, dv = \phi(\Delta) \quad \text{a function of } \Delta \text{ only}
\]

If the parameters entering in the alternative hypothesis are denoted symbolically by $\theta$ and in the null hypothesis by $\theta_0$, then

\[
\int_{-\infty}^{\infty} f(\theta) \, dv = \phi(\Delta) \quad \text{and} \quad \int_{-\infty}^{\infty} f(\theta_0) = \alpha \quad \text{(8a.6.1)}
\]
Let us define the inside of the region by

\[ f(\theta_0) \leq \int_{\Delta = \text{const.}} \lambda(\theta) f(\theta) \, ds \cdots \quad (8a.6.2) \]

where the integral is taken over the surface \( \Delta = \text{constant} \). Let there exist a positive function \( \lambda(\theta) \) such that the conditions (8a.6.1) are satisfied. The region \( \nu_0 \), if it exists, is the best distance power region for alternatives on the surface \( \Delta = \text{constant} \). This follows from the lemma in Appendix A4 extended to an infinite set of alternatives. If the relationship (8a.6.2) is independent of the alternative used, then we obtain a uniformly best distance power test. It is seen that the region (8a.6.2) is the same as the region which has the best average power for alternatives on the surface \( \Delta = \text{constant} \) and for an assigned a priori probability density \( \lambda(\theta) \) of the parameters. Although in the theory of average power tests there is no justification for choosing a particular type of the density function \( \lambda(\theta) \) on which the test generally depends, the function \( \lambda(\theta) \) is suitably determined in constructing distance power tests. The determination of such a function, even if its existence is known, may be a difficult task. Once it is determined by trial or otherwise, the optimum property of the test is immediately established.

It is of interest to examine the critical region obtained by extending the results in (8a.5.3) to the case of alternative hypotheses specified by parameters with continuous variation. The outside of such a critical region is defined by

\[ f(\theta_0) < \lambda(\theta, \Delta) f(\theta) \]

for all \( \theta \) on the surface \( \Delta(\theta) = \Delta \), where \( \Delta \) is the specified distance of the alternative from the null hypothesis. If, owing to considerations of symmetry, the function \( \lambda(\theta, \Delta) \) could be replaced by a function of \( \Delta \) only, then the critical region is the outside of the envelope of the surfaces

\[ f(\theta_0) = \text{const.} \]

\[ f(\theta) \]

for variations in \( \theta \) on the surface \( \Delta(\theta) = \Delta \). This is the likelihood ratio test developed by Neyman and Pearson (1928).

Example 1. Let \( z_1, \ldots, z_n \) be \( n \) independent normal variates having zero mean on the null hypothesis. For any alternative hypothesis specifying the mean of \( z_i \) as \( \mu_i \), the best test is

\[ \mu_1 z_1 + \cdots + \mu_n z_n \geq k \]

and the associated power is a function of \( \mu^2 = \mu_1^2 + \cdots + \mu_n^2 \). So we can define the distance of the alternative hypothesis from the null by \( \mu^2 \).
The best distance power test, if it exists, is given by

\[ e^{-\frac{\sum x_i^2}{2\sigma^2}} \leq \int e^{-\frac{\sum (x_i - \mu_i)^2}{2\sigma^2}} f(\mu_1, \ldots, \mu_n) \, d\mu_1 \ldots d\mu_n \]

\( f(\mu_1, \ldots, \mu_n) \) can be chosen to be a constant, in which case because of symmetry between \( x \) and \( \mu \) the test reduces to

\[ \sum x_i^2 \geq k \]

which is a distance power test, the distribution of \( \sum x_i^2 \) being that of a noncentral \( \chi^2 \) involving only the parameter \( \nu^2 \).

**Example 2.** Following the method of example 1, construct a distance power test to examine whether \( p \) correlated variables have assigned mean values. This leads to Hotelling’s \( T \) with a known dispersion matrix, in which case \( T \) has a \( \chi^2 \) distribution.

When the variances and covariances are not known in examples 1 and 2 above, a slightly different technique has to be followed. The best region, besides satisfying the above property, must be similar to the sample space with respect to the unknown variances and covariances. That is, the integral of the probability density over such a region is equal to a constant \( c \), whatever may be the value of the variances and covariances. We shall not enter into the mathematics of the construction of similar regions but simply note that all the univariate and multivariate tests considered in the earlier chapters are all best distance power tests.

### 8b Problems of Discrimination

#### 8b.1 The General Problem

We now come to a group of problems where a priori probabilities are needed for a satisfactory solution and the null hypothesis does not play a prominent part but is sometimes posed to arrive at a decision subject to a small risk.

Thus when a question is asked whether a skull or a jawbone belonged to a male or a female, there are evidently two alternative hypotheses and one has to be chosen. Here a procedure is needed by which the individual specimen can be assigned to one or the other of the groups. In any such rule of procedure, errors are inevitable unless the ranges of measurements for the two groups are completely different.

We first answer the following problem. Suppose an individual is drawn from a mixed population consisting of two distinct groups of individuals in the ratio \( \tau_1 : \tau_2 \), \( \tau_1 + \tau_2 = 1 \). If \( \alpha_1 \) is the chance of
wrongly classifying an individual of the first group by following any rule of procedure, and \( \alpha_2 \) the corresponding chance for the second group, then the probability of wrongly classifying an individual chosen at random is \((\pi_1\alpha_1 + \pi_2\alpha_2)\). Evidently that procedure is the best which minimizes this error.

If the individual admits \( p \) measurements, then we need a division of the \( p \) dimensional space into two regions, \( R_1 \) and \( R_2 \), such that when the point representing the \( p \) measurements falls in \( R_1 \) the individual is assigned to the first group, and otherwise to the second. If \( f_1(x|\theta_1) \) and \( f_2(x|\theta_2) \) represent the probability densities, then the chance of committing an error is

\[
\pi_1\int_{R_1} f_1(x) \, dx + \pi_2\int_{R_1} f_2(x) \, dx.
\]

We need such a division for which the above value is a minimum. Following a lemma given in Appendix A2, we find that the best regions are

\[
\begin{align*}
R_1 & \cap \quad \pi f_1 \geq \pi f_2 \\
R_2 & \cap \quad \pi f_2 \geq \pi f_1
\end{align*}
\]

where the symbol \( \cap \) stands for "defined by." This supplies a mutually exclusive division of the space into two regions, \( R_1 \) and \( R_2 \). The case where the equality occurs can be decided by considering the corresponding relationship when one measurement (chosen at random from the available \( p \)) is omitted.

If, in any problem, there is scope for the introduction of a risk function specifying the loss incurred in a wrong classification, then the best solution which minimizes the expected risk can be determined as follows. Let \( r_1 \) be the loss resulting in assigning an individual of the first group to the second, and \( r_2 \) of the second to the first. Then the expected loss is

\[
\pi_1\alpha_1 r_1 + \pi_2\alpha_2 r_2.
\]

The best solution is

\[
\begin{align*}
R_1 & \cap \quad \pi r f_1 \geq \pi r f_2 \\
R_2 & \cap \quad \pi r f_2 \geq \pi r f_1
\end{align*}
\]

8b.2 The Discriminant Function of R. A. Fisher

In the cases considered in the above section it is seen that the boundary separating the two regions in the space is defined by a constant value of the likelihood ratio. If the probability densities are multivariate normal with the same dispersion matrix \((a_{ij})\) and mean values, \( \mu_1, \cdots, \mu_r \),
and 

where the matrix \( (a_{ij}) \) is reciprocal to \( (a_{ij}) \). Simplifying the above expression, the surface of a constant likelihood ratio can be expressed

\[
\sum_{i=1}^{p} (a_{i1}d_1 + \cdots + a_{ip}d_p)x_i = \text{const.} \tag{8b.2.1}
\]

where \( d_j = \mu_{j1} - \mu_{j2} \), \( (j = 1, 2, \ldots, p) \). The regions in the \( p \) dimensional space are thus separated by a hyperplane whose equation is (8b.2.1) for a suitably determined constant. An individual for whom the value of the left-hand function exceeds the constant value chosen is assigned to the first group and when it is smaller he is assigned to the second group.

The linear function of the measurements deduced above is called the discriminant function, first introduced by R. A. Fisher, who suggested the following computational procedure.

If there is only one character, then the problem of classification is very simple; all individuals with the value of that character exceeding a suitably determined value could be assigned to one group, and the rest to the other. The multiple character case is reduced to that of a single variate by replacing the several measurements by a suitably chosen linear compound. If \( x_1, x_2, \ldots, x_p \) are the measurements, then an arbitrary linear compound is \( l_1x_1 + \cdots + l_px_p \). The coefficients \( l_1, \ldots, l_p \) may be chosen such that the linear compound affords the maximum discrimination between the two groups.

The function \( l_1x_1 + \cdots + l_px_p \) has the variance

\[
\Sigma a_{ij}ld_j \tag{8b.2.2}
\]

and the square of the difference in mean values of this compound for the two groups is

\[
(l_1d_1 + \cdots + l_pd_p)^2 \tag{8b.2.3}
\]

The coefficients may be chosen such that the difference in mean values is a maximum, subject to the condition that the variance (8b.2.2) is a constant (say, unity). This is also equivalent to maximizing the ratio of (8b.2.3) to (8b.2.2) without any condition on the compounding coefficients. Introducing a Lagrangian multiplier \( \lambda \) and differentiating the expression

\[
\Sigma l_ijd_j - \lambda \Sigma a_{ij}ld_j
\]
we obtain the equations

\[ l_1 a_{11} + l_2 a_{12} + \cdots + l_p a_{1p} = k a_1 \]
\[ l_1 a_{21} + l_2 a_{22} + \cdots + l_p a_{2p} = k a_2 \]
\[ \vdots \]
\[ l_1 a_{p1} + l_2 a_{p2} + \cdots + l_p a_{pp} = k a_p \]

where \( k = (l_1 d_1 + \cdots + l_p d_p)/\lambda \). Observing that the above equations can supply only ratios of \( l_1, \ldots, l_p \), we may substitute \( k = 1 \) and solve the above equations. The final values of \( l_1, \ldots, l_p \) may be adjusted by multiplying each of them by a constant \( \theta \) where

\[ \theta^2 \sum l_i a_{ii} = 1 \]

This is unnecessary because the constant separating the values of the discriminant function for classification into the two groups can be adjusted suitably. The linear equations obtained above have the solutions

\[ l_i = a_1 d_1 + \cdots + a_p d_p \quad i = 1, 2, \ldots, p \]

thus giving the same linear function derived as the ratio of the two likelihoods. Thus Fisher's linear discriminant function is the best for classification when the distributions are multivariate normal and the dispersion matrices are the same. If the dispersion matrices are different, then the likelihood ratio surface is defined by the quadratic expression

\[ \sum (a^2 (x_i - \mu_i) (x_j - \mu_j) - \beta^2 (x_i - \mu_i) (x_j - \mu_j)) = \text{const.} \]

where \((a^2)\) and \((\beta^2)\) are the inverses of the dispersion matrices corresponding to the two populations.

8b.3 Some Difficulties in the Use of the Best Discriminating Solution

The elegant solution obtained in 8b.1 has many limitations so far as the practical applications are concerned.

(i) The parameters occurring in the probability distributions are not usually known. The only solution is to obtain their best possible estimates and substitute them for the unknown values in setting up the discriminant function. This introduces some additional errors in classification, depending on the paucity of the available material for the estimation of the parameters.
(ii) The a priori probabilities explicitly occurring in the best solution may not be known, and in some cases they may not be estimable from the available data.

(iii) In any problem the finite number of alternatives into which an observed individual has to be classified is assigned by the investigator. In some cases, such as sexing, there are only two alternatives possible. But in general it may be necessary to test whether the a priori information that an individual belongs to one of the given groups is correct or not. Only when the a priori information is ascertained to be correct can we proceed to decide to which of the given groups he is likely to belong.

(iv) Even by following the best procedure of classification it may not be possible to assert with confidence that any individual has been correctly classified. Can any provision be made to identify at least those cases which are less likely to be misclassified?

(v) Suppose that it is known that an individual has been taken at random from only one group and it is not known whether it is the first or the second group. Should he be treated in the same way as an individual drawn from a mixed population?

(vi) What is the nature of a risk function in biometric investigations?

(vii) Suppose that some quick decisions are needed. Is there any simple method of arriving at a discriminating function which is a good approximation to the ideal one?

These problems are discussed in the following sections with suitable illustrations.

8b.4 Uncertainty of the A Priori Information That One of the Alternatives Is Correct

For discriminatory analysis it must be known that an individual belongs to one or the other of two groups. Such knowledge has to be inferred from external evidence; the association of artifacts with a burial sometimes provides it for skeletal remains. In questions of sexing bones, the chances of identification are limited to two possibilities, male and female. However, where the external evidence is slight or equivocal, the assignment of an individual to one of two groups may be subject to another kind of error, viz., the wrong assumption that he belongs to one of the two groups when, in fact, he comes from a third unknown group. In the absence of any definite knowledge about the nature of the third group, we may have to examine by means of the internal evidence supplied by the measurements on the individual whether he could be considered as belonging to either of the two groups; that is, we examine whether there is any evidence to suggest that the individual could
not have come from one or other of the two groups. Consider the following problem.

In August, 1939, a relatively complete male human skeleton was recovered from the ditch of an Iron Age camp on Highdown Hill, Goring by Sea, in the course of excavations conducted under the auspices of the Worthing Archaeological Society. Fragments associated with the bones suggest that the burial could not have taken place later than the very beginning of the Iron Age in Sussex, about 500 B.C. The camp went out of use not later than 250 B.C., and the remains themselves can be assigned to a 500 B.C. “invasion” horizon. It is doubtful, however, whether their owner was a Bronze Age “defender” or an Iron Age “invader.” The principal question to be considered in the present context is whether the Highdown skull is more likely to have belonged to a Bronze Age or to an Iron Age population.

An attempt is made here to answer this problem by utilizing the published data concerning the Bronze Age and the Iron Age represented by Romano-British crania from Maiden Castle. The characteristics of these groups have been computed from scanty material, so that the conclusion regarding the Highdown skull cannot be treated as final. This example has been chosen merely to illustrate the method.

In solving the problem whether the Highdown skull belongs to the Bronze Age or to the Iron Age, we can test separately the two null hypotheses: (1) it belongs to the Bronze Age, and (2) it belongs to the other group. If neither of the two hypotheses can be rejected on the 5% level, there is sufficient justification to proceed with the problem of assigning the skull to one of the two groups.

It must be noted that in such a procedure we are not testing the combined null hypothesis that the specimen belongs to one or the other of the groups at the 5% level. Of the 5% of the rejected cases under one hypothesis, some are accepted under the second hypothesis so that when the 5% level is used for the two hypotheses separately we will be judging the combined null hypothesis at a lower level.

An adjustment could be made in the test procedure to correct this by defining the critical region (of rejection)

\[
\Sigma \Sigma a^i(x_i - \mu_1)(x_j - \mu_1) \geq c
\]

\[
\Sigma \Sigma a^i(x_i - \mu_2)(x_j - \mu_2) \geq c
\]

where \((a^i)\) is the reciprocal of the dispersion matrix \((a_{ij})\); \(\mu_1, \mu_2\) are mean values for the two groups; \(x_1, x_2, \ldots\) are the measurements on the individual; and \(c\) is chosen such that the total density of the region
is a (the assigned value) for each of the probability distributions. The existence of \( c \) satisfying the above condition follows from symmetry. It is difficult to find its actual value. In practice the procedure outlined earlier of testing the two hypotheses separately may be followed.

Table 8b.4a gives the mean values of male English Bronze Age (Morant, 1926) and Maiden Castle (Goodwin and Morant, 1940) cranial measurements.

<table>
<thead>
<tr>
<th>Character</th>
<th>Highborn</th>
<th>Bronze</th>
<th>Maiden</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>198.2</td>
<td>184.5 (45) *</td>
<td>188.6 (23)</td>
</tr>
<tr>
<td>B</td>
<td>148.1</td>
<td>149.9 (89)</td>
<td>140.8 (24)</td>
</tr>
<tr>
<td>H'</td>
<td>142.0</td>
<td>134.9 (25)</td>
<td>137.1 (22)</td>
</tr>
<tr>
<td>G'H</td>
<td>72.4</td>
<td>69.1 (30)</td>
<td>72.4 (14)</td>
</tr>
<tr>
<td>GB</td>
<td>95.8</td>
<td>98.0 (11)</td>
<td>95.2 (18)</td>
</tr>
<tr>
<td>NH,L</td>
<td>48.2</td>
<td>49.1 (13)</td>
<td>51.9 (16)</td>
</tr>
</tbody>
</table>

* The figures in parentheses indicate numbers on which the averages are based.

L = head length from the glabella in the median sagittal plane.
B = head breadth on the parietal bones perpendicular to L.
H' = head height from the basion to the bregma.
G'H = upper facial height from nasion to the alveolare.
GB = bimaxillary breadth between the zygomatics.
NH,L = nasal height from the nasion to the left nariale.

In the present problem, since only a few of the Bronze Age standard deviations have been published and none are available for the Maiden Castle, the variance-covariance matrix obtained from a long series of Farringdon Street (English male) crania measured by Hooke (1926) is used in the analysis. Such a procedure is not strictly correct, but as observed elsewhere the dispersion matrix remains sensibly constant, provided that the series are not completely unrelated. Since we are using it only to construct the coefficients of the discriminant function, slight variations in the elements of the dispersion matrix do not matter.

To study this problem in some detail and to judge the importance of the various characters used for discrimination it is advantageous first to obtain a set of uncorrelated functions of the original variables. This can be easily obtained by using the dispersion matrix with a unit matrix appended to it and reducing the dispersion matrix by the method of sweep out as shown in Table 8b.4b. The theoretical discussion associated with this procedure is treated in Appendix B.
<table>
<thead>
<tr>
<th>Row No</th>
<th>L</th>
<th>B</th>
<th>H'</th>
<th>Q'H</th>
<th>GB</th>
<th>NHL</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>41.73</td>
<td>10.78</td>
<td>7.78</td>
<td>3.85</td>
<td>9.63</td>
<td>6.79</td>
<td>1</td>
</tr>
<tr>
<td>02</td>
<td>34.81</td>
<td>12.08</td>
<td>6.64</td>
<td>6.12</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>03</td>
<td>23.09</td>
<td>3.85</td>
<td>1.37</td>
<td>4.51</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>04</td>
<td>18.90</td>
<td>7.65</td>
<td>2.46</td>
<td>8.00</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>05</td>
<td>19.34</td>
<td>4.35</td>
<td>1.07</td>
<td>3.28</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>06</td>
<td>37.94</td>
<td>4.35</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.3645</td>
<td>0.6096</td>
<td>0.0277</td>
<td>0.1827</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>32.035</td>
<td>5.853</td>
<td>11.962</td>
<td>4.152</td>
<td>2.974</td>
<td>1</td>
<td>0.2662</td>
</tr>
<tr>
<td>12</td>
<td>44.169</td>
<td>3.1763</td>
<td>4.056</td>
<td>2.974</td>
<td>1</td>
<td>1</td>
<td>0.2662</td>
</tr>
<tr>
<td>13</td>
<td>19.441</td>
<td>2.379</td>
<td>8.803</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>30.737</td>
<td>2.432</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>3.622</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.17784</td>
<td>0.34099</td>
<td>0.19956</td>
<td>0.15633</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>33.140</td>
<td>1.1996</td>
<td>1.1582</td>
<td>7.499</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>15.6574</td>
<td>1.1402</td>
<td>7.3207</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>30.1793</td>
<td>2.1710</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>8.0590</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>0.00818</td>
<td>0.06006</td>
<td>0.00196</td>
<td>0.00276</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>15.5412</td>
<td>1.0003</td>
<td>7.516</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>30.1212</td>
<td>2.3407</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>7.7662</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>0.10294</td>
<td>0.40391</td>
<td>0.00253</td>
<td>0.00323</td>
<td>0.00438</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>34.5555</td>
<td>1.3883</td>
<td>4.4507</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>4.4507</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.00357</td>
<td>0.00297</td>
<td>0.00158</td>
<td>0.00294</td>
<td>0.00281</td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>4.3426</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. The elements below the diagonal are omitted because the matrices at all stages are symmetrical.
2. The rows 10, 20, 30, 40, 50 represent the pivotal rows at each stage of reduction. Following each pivotal row is given the reduced matrix. The first row in the reduced matrix equals a linear function of the variables whose variance is the pivotal element (underlined) in that row. Thus Vx, Vx, Vx, Vx, and Vx have been previously defined, and 0.00253 is their variance.
3. The computations can be correctly represented by omitting the matrix for functions of original variables and accommodating the figure below the diagonal in the left-hand matrix as shown in Table 7b.42.
Obtaining the expressions for \( Y_1, Y_2, \ldots, Y_6 \) from Table 8b.4 and dividing them by the corresponding standard deviations, we obtain the following uncorrelated transformed variables with unit variances.

\[
Y_1 = 0.1548L - 0.0456B + 0.1767G'
\]
\[
Y_2 = -0.0201L - 0.0369B + 0.2079G'
\]
\[
Y_3 = -0.0291L - 0.0369B + 0.2079G'
\]
\[
Y_4 = 0.0010L - 0.0854B - 0.0131G'
\]
\[
Y_5 = -0.0346L - 0.0176B + 0.0094G'
\]
\[
Y_6 = -0.0503L + 0.0209B - 0.0409G'
\]

<table>
<thead>
<tr>
<th>Table 8b.4. Values of Transformed Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highbourn</td>
</tr>
<tr>
<td>Skull</td>
</tr>
<tr>
<td>( y_1 )</td>
</tr>
<tr>
<td>( y_2 )</td>
</tr>
<tr>
<td>( y_3 )</td>
</tr>
<tr>
<td>( y_4 )</td>
</tr>
<tr>
<td>( y_5 )</td>
</tr>
<tr>
<td>( y_6 )</td>
</tr>
</tbody>
</table>

To test whether the Highbourn skull belongs to the Bronze Age we calculate the sum of squares of differences:

\[
\chi^2 = (30.681 - 28.561)^2 + (17.131 - 18.074)^2 + \cdots + (-7.606 + 5.479)^2
\]

\[
= 12.694
\]

which can be used as \( \chi^2 \) with 6 degrees of freedom if the Bronze Age means and dispersion matrix have been obtained from a large sample. On the other hand, if the Bronze Age means are based on a sample of size \( N \) and the dispersion matrix is estimated on \( f \) degrees of freedom, then

\[
F = \chi^2 \frac{(f + 1 - p)N}{f p(N + 1)}
\]
can be used as a variance ratio with \( p \) and \((f + 1 - p)\) degrees of freedom, \( p \) being the number of characters. For the purposes of the above example we shall treat the mean values and dispersion matrix as known so that the \( \chi^2 \) test can be used. The value 12.694 is just significant on the 5\% level.

Similarly, the \( \chi^2 \) for the Highbrown skull and the Maiden Castle series is

\[
(1.486)^2 + (0.852)^2 + \cdots + (-2.319)^2 = 9.091
\]

which has a probability of more than 15\%. By this criterion the Highbrown skull could be assigned to the Maiden Castle series. Since \( \chi^2 \) is only just significant in the other case, we might construct the discriminant function and decide the issue.

Since all \( y \) are uncorrelated, we can very easily construct the discriminant function. For instance, the one based on the first three values of \( y \) is

\[-0.634y_1 + 1.795y_2 - 0.674y_3\]

where the coefficients are the differences in mean values of \( y \) for the Bronze Age and the Maiden Castle series. The discriminant function based on \( y_1, y_2, y_3, y_4 \) is obtained by adding \(-1.589y_4\) to the above expression, and so on. The discriminant function with all the characters is

\[-0.634y_1 + 1.795y_2 - 0.674y_3 - 1.589y_4 + 0.491y_5 - 0.192y_6\]

which has the mean values

\[-4.300 \quad \text{and} \quad 2.581\]

for the Maiden Castle and the Bronze Age series, with the middle value \(-0.859\). Suppose we follow the procedure of assigning all individuals with values of the discriminant function above \(-0.859\) to the Bronze Age, and all others to the Maiden Castle series. Then the error in classification corresponds to the area above \(-0.859\) for a normal distribution with mean \(-4.300\) and variance

\[
D^2 = (-0.634)^2 + (1.795)^2 + \cdots + (-0.192)^2 = 6.881
\]

which is the sum of squares of the discriminant function coefficients. The normal deviate with zero mean and unit standard deviation is

\[
\frac{4.300 - 0.859}{\sqrt{6.881}} = \frac{3.441}{\sqrt{6.881}} = \frac{\sqrt{6.881}}{2} = \frac{D}{2} = 1.312
\]
with a probability of about 0.095, which is the error of wrong classification associated with any group. The value of the discriminant function for the Highdown skull is \(-2.661\) which assigns him to the Maiden Castle series.

In Table 8b.4, the probability of error in using the first \(i = 1, 2, \ldots, 6\) characters is given in column (7). It is seen that the probability of error decreases with the increase in the number of characters, although such a decrease is inappreciable in some cases. For instance, the addition of the last character \(NII.L\) leading to the calculation of \(y_5\) does not add much, the decrease in error being very small, less than 2 in 1000 or so.

The evaluation of the discriminant function coefficients and the tests of significance used above are easily carried out with the use of transformed variates. An alternative way is to adopt the computational scheme of Table 7b.6 in Chapter 7 where successive values of \(D^2\) and the discriminant functions were obtained. The variance of the discriminant function (treated as a linear combination of the measurements) is \(D^2\). This is useful in computing the frequencies of wrong classification.

8b.5 The Doubtful Region

In using the discriminant function in the example of 8b.4, the critical value separating the individuals of the two groups was obtained as the middle value of the mean values of the function for the two groups. The frequency of wrong classification for any group is 9.5%. Suppose that an individual is drawn at random from two such populations (considered above) mixed in the proportions \(\pi_1: \pi_2\). The probability of wrong assignment of an individual in such a case is

\[
\pi_1(0.095) + \pi_2(0.095) = 0.095
\]

which is the same as that for a single group. This is not, however, the minimum possible error when \(\pi_1\) and \(\pi_2\) are known, the minimum being associated with the section

\[
\lambda = \frac{L_1 + L_2}{2} + \log_\pi \pi_2 - \log_\pi \pi_1 \quad (8b.5.1)
\]

where \(L_1\) and \(L_2\) are the mean values of the discriminant function \(L\) for the two groups, the higher value being associated with the first group. If \(L > \lambda\), then the individual is assigned to the first group; otherwise to the second. Suppose that \(\pi_1: \pi_2 :: 1:2\), then

\[
\lambda = -0.859 + 0.693 = -0.166
\]
in which case the errors of wrong classifications for the two groups are

\[ a_1 = 0.15 \quad \text{and} \quad a_2 = 0.05 \]

giving a total error

\[ \pi_1 a_1 + \pi_2 a_2 = \frac{(0.15 + 0.10)}{3} = 0.08 \quad \text{nearly} \]

Although the error is nearly 8%, the error of misclassification for an individual of the first group is as high as 15%, so that an individual assigned to the second group cannot be asserted to belong to the second group since his chances of belonging to the first group are as high as 15%.

Consider another situation where a doctor wants to discriminate between two types of neurotic, psychopaths and obsessinals, on the basis of some tests. If the test scores of properly diagnosed neurotics are available, then, assuming that the ratio of the two types of patient admitted into the hospital in the past represents the ratio in the general population, the doctor can set up the criterion (86.5.1). By following this procedure he can minimize the number of cases of wrong diagnosis.

But in problems like this the groups overlap to a large extent so that even by following the best procedure the percentage of wrong classifications remains quite high. By increasing the number of characters this percentage could be made smaller and smaller but not always below an irreducible minimum because of the correlations between the characters.

Furthermore, a stage may be reached at which the cost involved in further examination will not be commensurate to the reduction in the number of wrong classifications. But, subject to a given cost, the indicators * can be chosen so as to minimize the number of wrong classifications. Thus one has to balance between the errors committed and the time or money available.

By following this procedure it may be difficult to assert that an individual belongs to one group or the other unless the groups are well separated, in which case the proportion of wrong classifications will be low. On the other hand, one may take the view that, whatever may be the basis of judgment, in some cases it should be possible to give a decisive answer (subject to a small risk) whereas in others no decision or only provisional decisions can be made. The latter group comprises the doubtful cases which need further examination.

* For instance, there are two types of jaundice which are difficult to distinguish. One calls for a surgical treatment; the other for medical treatment. A discriminant function based on two biochemical tests is used in practice to ensure a greater certainty of diagnosis for far less laboratory work.
Cases also arise where the question asked is whether a selected individual can be asserted to belong to one particular group out of a given number of possibilities. Consider the problem of the Highdown skull. The grave findings associated with the skull excavated from the "invasion horizon" do not give any conclusive evidence as to whether the skull belonged to a Bronze Age "defender" or to an Iron Age "invader." It may or may not be possible to give a definite answer in such a problem. The case has to be judged on its individual merits, with consideration given to the probability of the individual's having come from one group or the other.

Sitting on the fence is a scientific attitude if it means looking for further evidence and better methods of judgment to be able to give a definite answer.

Consider a doctor who has a routine method of diagnosing a disease or discriminating among a number of diseases. Although by following this procedure he commits the least possible errors, he would like to be more confident about his diagnosis in some selected cases. If the routine method does not give him sufficient assurance, he may supplement it by further tests.

For any specially chosen case like this or for an individual find such as the Highdown skull, the rule of procedure suggested should necessarily be independent of the a priori probabilities used in the general problem of discrimination. First, such a priori probabilities may not be available; in the case of the Highdown skull it is not possible to know the proportions of the Bronze and the Iron Age cranial population. Second, even if such knowledge is available from previous experience, this is not strictly applicable in a case not chosen at random from a mixed population. For instance, the proportions applicable to the Highdown skull may depend on the numbers of Bronze and Iron Age warriors who went down fighting and not on the general proportion.

Thus a problem involving only one individual must be distinguished from a problem in which a number of individuals have to be classified into a given number of groups by means of suitable criteria. The latter supplies a provisional answer to the former, but for definite answers suitable criteria have to be developed.

Let us suppose that for the best solution of assigning individuals to the first group if

\[ \pi_1 f_1(x \mid \theta_1) \geq \pi_2 f_2(x \mid \theta_2) \]

and to the second group if

\[ \pi_2 f_2(x \mid \theta_2) \geq \pi_1 f_1(x \mid \theta_1) \]
the expected proportions of wrongly classified individuals of the first and second groups are $a_1$ and $a_2$, respectively. If $a_1$ and $a_2$ are small, then we can assert for any given individual that he is rightly classified. Otherwise we may follow the procedure of assigning an individual to the first group if

$$f_1(x | \theta_1) \geq A f_2(x | \theta_2)$$

and remain in doubt if

$$A f_2(x | \theta_2) > f_1(x | \theta_1) > B f_2(x | \theta_2)$$

The quantities $A$ and $B$ are chosen such that the probabilities of wrong decisions are at assigned levels. The diagram below (Figure 1) shows the nature of the decisions that could be made after ascertaining the value of the ratio or its logarithm.

In the region $R_2$ the individual can be asserted (at a given risk) to belong to the second group; in $D_2$ he can be provisionally assigned to the second group; and similarly for $R_1$ and $D_1$. In doubtful cases it may be possible to measure more characters and thus bring in further evidence to decide the issue.

In the example of the Highbrough skull with $\tau_1 = \frac{1}{2}$ and $\tau_2 = \frac{3}{2}$, it has already been shown that the point of section is $-0.166$ so that if the discriminant value exceeds $-0.166$ the individual is assigned to the Bronze Age. The point corresponding to the 5% level of errors for the Maiden Castle series is

$$-4.300 + 1.645D = 0.016 \quad D = 2.624$$

so that unless the discriminant value exceeds 0.016 an individual cannot be asserted to belong to the Bronze Age, although provisionally he will be put in the Bronze Age as soon as the value exceeds $-0.166$. Similarly, the 5% value for the Bronze Age is

$$2.581 - 1.645D = -1.735$$
and unless the value of the discriminant function is below \(-1.735\) the individual cannot be asserted to belong to the Maiden Castle series. The value for the Highdown skull is \(-2.661\) so that he can be confidently assigned to the Maiden Castle series.

8b.6 Resolution of a Mixed Series into Two Gaussian Components

In the previous sections were considered problems of determining the group to which an individual belongs when the distributions in the alternative groups are known. There may arise cases where a collection of individuals is observed but no information is available as to the distributions in the groups from which they have arisen or the proportion of mixture. The general problem is then to determine the characteristics of the various groups and also the proportion of mixture from the available set of measurements. This information may be finally used to specify the group of each individual, if necessary.

Considering only two groups in which a certain character is distributed normally, the statistical problem reduces to that of estimating from the observed frequency distribution the two mean values \(\mu_1, \mu_2\), standard deviations \(\sigma_1, \sigma_2\), and the proportion of mixture \(\pi\). The estimation of these five parameters by the method of moments was discussed by Pearson (1894). The estimates depend on a suitably chosen root of a nonic (ninth-degree equation) constructed from the first five moments of the observed frequency distribution.

In many problems it is reasonable to suppose that \(\sigma_1 = \sigma_2\), in which case there are only four parameters to be estimated. If the method of moments is followed, the first four moments are sufficient, for it has been shown that the estimates depend on the negative root of a simple cubic equation constructed from the first four moments. In practice, where large numbers are involved, the estimates obtained by the method of moments, though not efficient, may serve the purpose at hand. Where higher efficiency is aimed at, the estimates have to be found by the method of maximum likelihood. The numerical computations involved in this method are very complicated.

Whatever the method of estimation employed, the numerical computations become much simpler when the standard deviations are assumed to be equal. The simplifying assumption may introduce bias in the estimates when, in fact, the standard deviations differ. Such estimates are, however, more accurate than those obtained without this assumption when the bias in any estimate is smaller than its standard error. If the mean values and the proportion of mixture are to be estimated with a higher precision, small differences in the standard deviations can be ignored.
Estimation by the Method of Moments. The rule of estimating the parameters by the method of moments consists in equating the moments as calculated from the observations to functions of parameters representing the moments in the population. Since the expectations of calculated moments are not the same as the moments in the population, this method might introduce a little bias in the estimating equations. This bias can be avoided by equating the calculated moments to their expected values. Instead of this, we can choose the system of $k$-statistics of Fisher (defined in *Statistical Methods for Research Workers*) and equate them to their expectations which are the cumulants of the distribution.

If $s_2$, $s_3$, and $s_4$ are the second, third, and fourth moments about the mean, and $s_1$ the first moment about the origin, as calculated by the usual method, the first four $k$-statistics derivable from them are given by

$$k_1 = s_1$$

$$k_2 = \frac{n}{(n-1)} s_2$$

$$k_3 = \frac{n^2}{(n-1)(n-2)} s_3$$

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)} \left\{ (n+1)s_4 - 3(n-1)s_2^2 \right\}$$

If the moments are calculated from grouped data with class interval $h$, the quantities $\frac{1}{2} h^2$ and $\frac{1}{3} h^4$ have to be subtracted from the expressions for $k_2$ and $k_4$, respectively.

If $p$, $m_1$, $m_2$, and $s$ denote the estimates of $\pi$, $\mu_1$, $\mu_2$, and $\sigma$, the common standard deviation, then the estimating equations by this method are

$$1 = p + q$$

$$k_1 = pm_1 + qm_2$$

$$k_2 = s^2 + pd_1^2 + qd_2^2$$

$$k_3 = pd_1^3 + qd_2^3$$

$$k_4 = pd_1^4 + qd_2^4 - 3(pd_1^2 + qd_2^2)^2$$

where $q = 1 - p$, $d_1 = m_1 - k_1$, and $d_2 = m_2 - k_1$. From the definition $x = d_1d_2$, the value of $x$ is obtained as the negative root of the cubic

$$x^3 + \frac{1}{2}k_4x + \frac{1}{2}k_2^2 = 0$$
If $x$ is the required root, then $d_1$ is given by the negative root of the quadratic

$$d_1^2 + \frac{k_3}{x} d_1 + x = 0$$

and $d_2$ by $-(k_3/x) - d_1$. The estimates $m_1$, $m_2$, $p$, and $s$ are given by

$$m_1 = k_1 + d_1 \quad m_2 = k_1 + d_2 \quad p = \frac{d_2}{(d_2 - d_1)} \quad s^2 = k_3 + x$$

The fundamental cubic equation

$$x^3 + \frac{1}{2}k_4 x + \frac{1}{2}k_3^2 = 0$$

introduced above has a single negative root greater than $(-k_3)$. The best method of determining the root is to start with a trial value and obtain the correction by Newton's method of approximation. Since the equation is in a reduced form with the coefficient of $x^2$ absent, it is easy to guess the root correct to the nearest integer. If $x_1$ stands for the trial value, then the additive correction $\Delta x_1$ is given by

$$\left[ 3x_1^2 + \frac{k_4}{2} \right] \Delta x_1 = -x_1^3 - \frac{1}{2}k_4 x_1 - \frac{1}{2}k_3^2$$

The process is repeated until the expression on the right-hand side becomes very small. The data of Table 8b.6a give the frequency distribution of heights of 454 plants of two different types grown on the same

**Table 8b.6a.** The Frequency Distribution of Height in Centimeters for 454 Plants

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5-8.5</td>
<td>3</td>
</tr>
<tr>
<td>9.5</td>
<td>9</td>
</tr>
<tr>
<td>10.5</td>
<td>21</td>
</tr>
<tr>
<td>11.5</td>
<td>40</td>
</tr>
<tr>
<td>12.5</td>
<td>59</td>
</tr>
<tr>
<td>13.5</td>
<td>76</td>
</tr>
<tr>
<td>14.5</td>
<td>79</td>
</tr>
<tr>
<td>15.5</td>
<td>89</td>
</tr>
<tr>
<td>16.5</td>
<td>46</td>
</tr>
<tr>
<td>17.5</td>
<td>30</td>
</tr>
<tr>
<td>18.5</td>
<td>13</td>
</tr>
<tr>
<td>19.5</td>
<td>7</td>
</tr>
<tr>
<td>20.5</td>
<td>2</td>
</tr>
</tbody>
</table>

454
plot. The plants are indistinguishable except at the flowering stage. The problem is to estimate the mean height of the two types of plant, their common standard deviation, and the proportion of mixture.

The values of cumulants after adjustment for grouping are

\[ k_1 = -0.244493 \quad \text{about 14 as the origin} \]
\[ k_2 = 4.973963 \]
\[ k_3 = 0.728751 \]
\[ k_4 = -5.314741 \]

\[ \frac{1}{2} k_4 = -2.657370 \quad \frac{1}{2} k_3^2 = 0.265539 \]

The fundamental cubic is

\[ x^3 - 2.657370x - 0.265539 = 0 \]

Taking \(-1.65\) as a trial root, we find the correction \(\delta x\) is given by

\[ [3(1.65)^2 - 2.657370] \delta x = -(-1.65)^3 + 2.657370(-1.65) - 0.265539 \]
\[ 5.510130 \delta x = -0.158074 \]

Similarly, the second correction is 0.000707 so that the second approximation is \(-1.678688 + 0.000707 = -1.677981\). The quadratic giving \(d_1\) is

\[ d_1^3 - 0.434302d_1 - 1.677981 = 0 \]

which yields the negative root

\[ d_1 = -1.096203 \]
\[ d_2 = 1.096203 + 0.434302 = 1.530595 \]

The estimates of \(m_1\) and \(m_2\) about 14 as the origin are

\[ m_1 = -1.096203 - 0.244493 = -1.340786 \]
\[ m_2 = 1.530595 - 0.244493 = 1.286102 \]

\[ p = \frac{d_2}{(d_2 - d_1)} = 0.582665 \]
\[ q = (1 - p) = 0.417335 \]
\[ s^2 = 4.973963 - 1.677981 = 3.297982 \]
\[ s = 1.816055 \]
This completes the estimation of the four parameters by the method of moments.

The expressions for standard errors of these estimates are very complicated, but it appears that the estimate of the proportion of mixture will have the highest percentage of error whereas the estimates of means and standard deviation will be fairly reliable in large samples.

A good deal of caution is needed in resolving a mixed series.

(i) It must be ascertained that the population is a mixture of two homogeneous groups only.

(ii) Departure of the individual distributions from the Gaussian type introduces serious errors in the estimates.

(iii) Samples must be large enough for a successful resolution into two components.

8b.7 Sexing of Osteometric Material

In anthropometric work the problem of sexing often arises. The sex of an excavated skull or jawbone can be ascertained with a high chance of success if the associated pelvis is also found. Sometimes the sex of the skeleton may be determined by external evidence such as beads and other ornaments found buried with the body. Anthropologists of the "trained eye" school claim that a skull can be properly sexed by anatomical appreciation. In such cases, the conclusions may be marred by a subjective bias unless the features of the specimen examined are so striking as to leave the question of its identity in little doubt.

In any case two types of situations have to be faced. Some skeletal remains can be definitely sexed, but how can the other bones belonging to the same series be sexed? All those bones which have been sexed supply the basic material with which suitable discriminant functions can be constructed. The sex proportion can also be ascertained from the basic material. The problem, then, reduces to the simplest one of discrimination between two groups when the individual distributions and the a priori probabilities are known.

It may be necessary to obtain discriminant functions based on different sets and subsets of characters to suit the various bones which have to be sexed. For instance, a skull may be in a broken condition so that only the length and breadth of the cranium can be measured. A decision is then made with the help of a discriminant function based on length and breadth of the cranium only. Some skulls may admit facial and nasal measurements as well, in which case the discriminant function based on all these measurements has to be used. The method of transformed characters and the construction of discriminant functions adopted in 8b.5 are very useful in this connection, provided that
SEXING OF OSTEOMETRIC MATERIAL

the order in which the characters are added is suitably determined.
For instance, consider the five measurements of the cranium: length
(L), breadth (B), frontal breadth (B'), height (H'), and circumference
(S). The order here is obviously L, B, B', H', and S, for there may be
some skulls providing measurements on L, B, and B' alone and not on
the rest, whereas skulls admitting the measurements of H' and S must
necessarily supply the measurements L, B, and B'.

We have another situation when adequate material is not available
for the construction of the discriminant function. Then an approximate
function can be tried, the simplest of which is of the type

\[ \pm \frac{x_1 - \mu_1}{\sigma_1} \pm \frac{x_2 - \mu_2}{\sigma_2} \pm \ldots \]

where \( \pm \) or \(-\) is chosen according as the male mean for \( x_1 \) is greater or
smaller than the female mean. The values of the standard deviations
\( \sigma_1, \sigma_2, \ldots \) of \( x_1, x_2, \ldots \) need not be known exactly. Values from any
related series can be used because what is important is the relative order
of the various standard deviations. This formula does not make use of
the actual mean values but only of their inequality relationship. We
shall term the above expression as the "general size factor."

In the problem of sexing, the size factor can be conveniently employed
because the inequalities relating to various measurements of the male
and female are known. Most of the linear measurements have higher
values for males, whereas angles and some indices have higher values
for females. The series of values of the size factor calculated for each
specimen to be sexed can first be arranged in decreasing order of mag­
nitude and then divided in a given sex ratio assigning the higher values
to males and the smaller ones to females.

Some difficulty arises when the sex ratio is unknown. The series of
size factor values may then be treated as in Sb.6 for resolution into two
Gaussian components. This supplies the sex ratio and other constants
which may be useful in setting up the best procedure for discrimination
based on a single characteristic, viz., the size factor, or only the sex ratio
may be used to divide the series into males and females, as suggested
above.

As observed earlier, the resolution into Gaussian components is not
always a happy proposition because the estimation involves the calcu­
lated higher moments which are subject to large standard errors. An
alternative procedure is as follows.

First we observe that the standard deviation of the size factor \( \sigma^2 \) for
a homogeneous group (either \( \sigma^2 \sigma^2 \) or \( \sigma^2 \sigma^2 \)) can be obtained from any
related series. If $k_2$, $k_3$ are the second and third cumulants of the mixed series of the size factor, then the estimating equations are

\[ p_1 d_1 + p_2 d_2 = 0 \]
\[ d_1 d_2 = -(k_2 - \sigma^2) \]
\[ d_1 + d_2 = \frac{k_3}{(k_2 - \sigma^2)} \]
\[ p = \frac{d_2}{(d_2 - d_1)} \]

where $d_1 = m_1 - k_1$ and $d_2 = m_2 - k_1$, as defined before. The use of the fourth moment is avoided by this procedure, but it remains to be seen what error is committed by choosing a wrong $\sigma$. Consider the previous example with

\[
\begin{align*}
  k_1 &= -0.244493 & k_2 &= 4.975963 & k_3 &= 0.728751 \\
  \sigma^2 &= 4, & 3.297982, & 3
\end{align*}
\]

Assuming three different values

\[
\begin{align*}
  q^2 &= 4, 3.297982, 3
\end{align*}
\]

the equations are

\[
\begin{align*}
  d_1 d_2 &= -0.975963, -1.677981, -1.975963 \\
  d_1 + d_2 &= 0.746699, 0.434302, 0.368808
\end{align*}
\]

with solutions

\[
\begin{align*}
  d_1 &= -0.682753, -1.096293, -1.233329 \\
  d_2 &= 1.429452, 1.530595, 1.602137 \\
  p &= 0.676758, 0.582665, 0.565035
\end{align*}
\]

of which the middle one corresponds to the solution already obtained. It appears that $p$ is stable for small variations in $\sigma$.

Another method is to avoid the calculation of even the third moment but to use the median instead. The computations will be slightly heavier.

If the mean values are available, then a better approximation to the discriminant function is

\[ k_1 x_1 + k_2 x_2 + \cdots + k_p x_p \]

where $k_i = d_i / \sigma_i^2$, $d_i$ being the difference in mean values of the $i$th character. This is exactly the discriminant function when the correlations are neglected.
Example 1. Construct the discriminant function, assuming differences \( d_1, d_2, \ldots, d_p \) in the mean values and a correlation matrix with all correlations equal to a constant value \( r \). The function can be expressed in terms of two linear functions of the variables

\[
P = \frac{x_1}{s_1} + \frac{x_2}{s_2} + \cdots + \frac{x_p}{s_p}
\]

and

\[
Q = k_1x_1 + \cdots + k_px_p
\]

where

\[
k_i = \frac{d_i}{s_i^2}
\]

The second factor is called the shape factor by Penrose (1947) who used a slightly different form.

Example 2. Consider a correlation matrix of the type

\[
\begin{pmatrix}
A & B \\
B & C
\end{pmatrix}
\]

where \( A \) and \( C \) are two equicorrelation matrices and \( B \) is a matrix with all its elements constant. Here the discriminant function depends on two sets of size and shape factors and on another factor which may be called the bipolar factor.

8b.8 The Problem of Three and More Groups

In 8b.5 it is seen that, if measurements on a certain number of characters are available for two groups, it is possible to construct a discriminant function which affords the maximum discrimination between them. This function is useful in assigning with a certain degree of confidence an individual or individuals to one or the other of the two groups to which they are known to belong. In taxonomic problems there arise cases where an individual specimen is known to belong to one of three or more groups and has to be assigned to its proper group. Thus a plant may have to be specified as \textit{Iris versicolor}, \textit{Iris setosa}, or \textit{Iris virginica}. This problem is approached by the extension of the discriminant function analysis developed with special reference to two groups.

The General Theory for Three Groups. Let the probability densities in the three groups be represented by \( f_1(x, \theta_1), f_2(x, \theta_2), f_3(x, \theta_3) \), where \( x \) stands for the available set of measurements and \( \theta \) for the parameters. First we shall consider the general problem of classifying a collection of individuals drawn from a mixed population containing individuals of the three groups in the proportions, \( \pi_1, \pi_2, \pi_3 \), \( (\pi_1 + \pi_2 + \pi_3 = 1) \).
Any individual characterized by \( p \) measurements can be represented by a point in a \( p \)-dimensional space. The problem of classifying an observed collection of individuals is the same as the division of the space into three mutually exclusive regions, \( R_1, R_2, R_3 \), with the rule of procedure of assigning an individual \( I \), represented by a point in \( R_i \), to the \( i \)th group. If the probability that an individual of the \( i \)th group will fall in \( R_i \) is \( \beta_i \), then the expected value of the proportion of wrong classifications is

\[
\alpha = 1 - (\beta_1 + \beta_2 + \beta_3)
\]

The errors will be a minimum for that choice of regions \( R_1, R_2, R_3 \) for which \( \beta_1 + \beta_2 + \beta_3 \) is a maximum. Such regions, if they exist, may be termed the "best possible" regions. The following theorem establishes the existence and nature of the best possible regions.

**Theorem 1.** The regions defined by

\[
R_1 \cap \tau_1 f_1 \geq \tau_2 f_2, \quad \tau_1 f_1 \geq \tau_3 f_3 \\
R_2 \cap \tau_2 f_2 \geq \tau_1 f_1, \quad \tau_2 f_2 \geq \tau_3 f_3 \\
R_3 \cap \tau_3 f_3 \geq \tau_1 f_1, \quad \tau_3 f_3 \geq \tau_2 f_2
\]

constitute the best possible system of mutually exclusive regions.

The result follows from the lemma proved in Appendix A2. It is interesting to observe that this solution is the same as that obtained from Bayes's theorem on posterior probabilities.

If every individual is equally likely to be drawn from any group, the best regions are

\[
R_1 \cap f_1 \geq f_2, \quad f_1 \geq f_3 \\
R_2 \cap f_2 \geq f_1, \quad f_2 \geq f_3 \\
R_3 \cap f_3 \geq f_1, \quad f_3 \geq f_2
\]

These regions may be used for classifying an observed collection of individuals when nothing is known about \( \beta_1, \beta_2, \beta_3 \), the proportions of mixture. This is the maximum likelihood method in the problem of classification. We choose that hypothesis for which the likelihood is a maximum.

By adopting this procedure the probability of an individual of the first group being rightly assigned is \( \int_{R_1} f_1 \; dv = \beta_1 \), and the probabilities of the individuals of the second and third groups being wrongly assigned to the first group are

\[
\alpha_{12} = \int_{R_2} f_2 \; dv \quad \text{and} \quad \alpha_{13} = \int_{R_3} f_3 \; dv
\]
Since, in $R_1$, $f_1 \geq f_2$, $f_1 \geq f_3$, it follows that

$$\beta_1 \geq \text{the greater of } a_{12} \quad \text{and} \quad a_{13}$$

If $a_{12}$ and $a_{13}$ are small, we can assert with some confidence that an individual falling in $R_1$ is correctly classified. If they are not small, it is pertinent to inquire whether there exists a region $C_1$ such that

$$\int_{C_1} f_1 \, dv \quad \text{is a maximum}$$

subject to the conditions that

$$\int_{C_1} f_2 \, dv \quad \text{and} \quad \int_{C_1} f_3 \, dv$$

are both not greater than a quantity $a_1$, chosen to be small, say 0.01 or 0.05. If an observed specimen falls in such a region, then the hypothesis that it belongs to the second or the third groups may be rejected, in which case it is assigned to the first group. The existence and nature of such regions are established by theorem 2.

**Theorem 2.** Region $C_1$ satisfying the condition

$$\int_{C_1} f_1 \, dv \quad \text{is a maximum}$$

subject to the restrictions

$$a_1 \geq \int_{C_1} f_2 \, dv \quad \text{and} \quad \int_{C_1} f_3 \, dv$$

is defined by

$$f_1 \geq a f_2 + b f_3$$

where $a$ and $b$ are suitably chosen.

The proof of this theorem follows from the lemma of Neyman and Pearson given in Appendix A1. To apply this lemma consider two quantities $a_{12}$, $a_{13}$, both less than the assigned quantity $a_1$, and choose a region such that

$$\int_{\omega} f_1 \, dv \quad \text{is a maximum}$$

subject to the conditions

$$\int_{\omega} f_2 \, dv = a_{12} \quad \int_{\omega} f_3 \, dv = a_{13}$$
310 STATISTICAL INFERENCE IN CLASSIFICATORY PROBLEMS

The inside of such a region is defined by

\[ f_1 \geq a'f_2 + b'f_3 \]

where \( a' \) and \( b' \) are properly chosen. Let the maximized value of

\[ \int f_1 \, dv, \]

which is evidently a function of \( \alpha_{12}, \alpha_{13} \), be represented by

\( \beta(\alpha_{12}, \alpha_{13}) \). This is not, in general, an increasing function of \( \alpha_{12} \) and

\( \alpha_{13} \); therefore the maximum value is not necessarily attained when

\( \alpha_{12} = \alpha_{13} = \alpha_1 \). If, now, the function \( \beta(\alpha_{12}, \alpha_{13}) \) is maximized with

respect to \( \alpha_{12}, \alpha_{13} \), subject to the conditions \( \alpha_{12}, \alpha_{13} \leq \alpha_1 \), we obtain two values, \( \alpha_{12}^0, \alpha_{13}^0 \), corresponding to the optimum solution. Denoting the values of \( \alpha', b' \), corresponding to \( \alpha_{12}^0, \alpha_{13}^0 \), by \( a, b \), the required

region may be written

\[ f_1 \geq af_2 + bf_3, \]

which proves theorem 2.

It is easy to see that at least one of the values \( \alpha_{12}^0, \alpha_{13}^0 \) coincides

with the boundary value \( \alpha_1 \). Consider the best region corresponding to

\( \alpha_{12}, \alpha_{13} \) with both less than \( \alpha_1 \). If \( \alpha_{12} \geq \alpha_{13} \), it is always possible to

add a region in which \( \alpha_{12} \) is increased to \( \alpha_1 \), and \( \alpha_{13} \) to a value \( \leq \alpha_1 \). If this is not possible, a region in which \( \alpha_{12} \leq \alpha_{13} \) can

be added such that at least one value, \( \alpha_{12} \) or \( \alpha_{13} \), reaches the value \( \alpha_1 \).

The value of \( \beta(\alpha_{12}, \alpha_{13}) \) is increased in any case.

Having obtained regions \( R_1 \) and \( C_1 \) as given in theorems 1 and 2, we

may specify an individual falling in the \( C_1 \) region as belonging to the

first group, and an individual falling in \( D_1 = R_1 - C_1 \) as likely to be­

long to the first group. Regions \( R_2, C_2 \) and \( R_3, C_3 \) can be similarly

constructed.

If the proportions \( \tau_1, \tau_2, \tau_3 \) considered in theorem 1 are known, then

region \( C_1 \) is determined by

\[ f_1 \geq a(\tau_2 f_2 + \tau_3 f_3) \]

where \( a \) is chosen such that \( \int_{c_1} (\tau_2 f_2 + \tau_3 f_3) = a \tau_1 \) and so on. The

position is shown in Figure 2.

A certain amount of simplification results if the best region \( C_1 \) is

replaced by

\[ C_1' \cap f_1 \geq Af_2, \quad f_1 \geq Bf_3, \]

where \( A \) and \( B \) are chosen such that

\[ \int_{c_1'} f_1 \, dv \]

is a maximum.
subject to the conditions

\[ \int_{C_1} f_1 \, dv \leq \alpha_1 \quad \int_{C_2} f_2 \, dv \leq \alpha_1 \]

This region is not the best possible, but it is likely to be a good approximation.

In some situations it may be necessary to find regions \( R_1, R_2, R_3 \) such that the errors of classification are the same for each group or are to be in given ratios \( \pi_1 : \pi_2 : \pi_3 \). The existence and nature of such regions are established by the following theorem.

**Theorem 3.** The system of regions

\[
\begin{align*}
R_1 & \cap \quad a_1 \geq b_2 \quad a_1 \geq c_3 \\
R_2 & \cap \quad b_2 \geq c_3 \quad b_2 \geq a_1 \\
R_3 & \cap \quad c_3 \geq a_1 \quad c_3 \geq b_2
\end{align*}
\]

where \( a, b, c \) are suitably chosen, are the best possible if the errors of classification for the three groups are to be in an assigned ratio.

Let \( R_1', R_2', R_3' \) be any other set of regions for which the errors of classification are in the assigned ratio. The region common to \( R'_1 \) and
It is first easy to see that for such regions

\[ \alpha_1 = \alpha_2 = \alpha_3 \]

for, if an inequality relationship is true, say \( \alpha_2 > \alpha_3 \), it is possible to reallocate the regions \( R_2 \) and \( R_3 \) such that \( \alpha_2 \) is decreased and \( \alpha_3 \) is increased, thus reducing the maximum \( \alpha \). No improvement is possible by this method when \( \alpha_1 = \alpha_2 = \alpha_3 \), in which case we can choose the
regions with the help of theorem 3 to minimize this common value, i.e.,
when \( p_1 = p_2 = p_3 \).

The minimax requirement is to some extent unrealistic. Consider a
situation where two of the three groups are close together and the other
is quite distant. If the individuals of the distant group are considered,
the chance of misclassification should be small whereas the chance of
error for any one of the closer groups should be high. No compromise
is served by equalizing these errors. As observed earlier the maximum
likelihood solution can be used when nothing is known about the a
priori probabilities. When these regions are used, the requirement
stated above is automatically satisfied. Also, there does not exist any
other set of regions which is uniformly better than this set in the sense
that the errors are smaller for each group.

Theorem 2 led us to the construction of 4 mutually exclusive regions
with the help of which an observed specimen can be assigned either to
a particular group or to none. In some problems it may be necessary
to construct a system of 7 regions, 3 for assigning an observed specimen
to particular groups, 3 others for specifying it as belonging to one of
two of the groups, and the remaining one for making no decision. To
construct these regions we set up three regions \( W_1, W_2, W_3 \) for not accepting
respectively the first, the second, and the third groups, as the possible
ones from which the observed specimen has arisen. The boundary
surfaces of these regions determine by mutual intersection the required
system of 7 regions. The region outside \( W_1, W_2, W_3 \) is the doubtful re-
gion; the intersection of \( W_i \) and \( W_j \) is the region for specifying an indi-
vidual belonging to the \( k \)th group, \( (k \neq i \neq j) \); and the region outside
\( W_i, W_j \) but inside \( W_k \) is for either the \( i \)th or the \( j \)th group. Some methods
of constructing regions \( W_1, W_2, W_3 \) are discussed below.

Regions \( W_i \) when \( \tau_1, \tau_2, \tau_3 \) are known: If \( \tau_1, \tau_2, \tau_3 \) considered in
theorem 1 are known, then region \( W_1 \) is such that

\[
\int_{W_1} \rightarrow f_1 \, dv = \alpha_1 \quad \text{(a small assigned quantity)}
\]

and

\[
\int_{W_1} (\tau_2 f_2 + \tau_3 f_3) \, dv \quad \text{is a maximum}
\]

The boundary surface of such a region is

\[
\tau_1 f_1 \leq a(\tau_2 f_2 + \tau_3 f_3)
\]

where the constant \( a \) is suitably determined. Similarly, \( W_2 \) and \( W_3 \) can
be constructed.
Regions \( w_i \) independent of any a priori information: Consider the region \( w'_i \) such that
\[
\int_{w'_i} f_1 \, dv = \alpha_i
\]
and
\[
\int_{w'_i} f_2 \, dv = \int_{w'_i} f_3 \, dv = \beta \quad \text{is a maximum}
\]
Having determined \( \beta \), it is possible to construct the region \( w_i \) such that
\[
\int_{w_i} f_1 \, dv = \alpha_i
\]
\[
\int_{w_i} f_1 \, dv = \beta \quad i \neq 1
\]
\[
\int_{w_i} f_j \, dv \quad \text{is a maximum} \quad j \neq i
\]
where \( i = 2 \) if
\[
\int_{w_2} f_2 \, dv \leq \int_{w_2} f_3 \, dv
\]
and \( i = 3 \) otherwise. The boundary surface of such a region is of the form
\[
f_1 \leq af_2 + bf_3
\]
where \( a \) and \( b \) are suitably determined.

There is an alternative method of constructing \( w_1, w_2, w_3 \) which may be useful in some practical situations. Let \( \beta_2 \) and \( \beta_3 \) be the maximum values of
\[
\int_{w_2} f_2 \, dv \quad \text{and} \quad \int_{w_3} f_3 \, dv
\]
subject to the conditions
\[
\int_{w_2} f_1 \, dv = \alpha_1 \quad \int_{w_3} f_1 \, dv = \alpha_1
\]
where \( w_2 \) and \( w_3 \) are the regions corresponding to the maxima. Region \( w_1 \) may be determined such that
\[
\int_{w_1} f_1 \, dv = \alpha_1
\]
\[
\frac{1}{\beta_2} \int_{v_1} f_2 \, dv = \frac{1}{\beta_3} \int_{v_1} f_3 \, dv \quad \text{is a maximum}
\]

The boundary surface of such a region is again of the form

\[f_1 \leq a_2 + b_3\]

where \(a\) and \(b\) are suitably determined.

Denoting the populations or possible alternatives by \(H_1, H_2, H_3\), the rule of procedure is as indicated in Figure 3.

**Figure 3.** Division of the space for seven possible decisions.
8b.9 Application to Multivariate Normal Populations

For multivariate normal populations the probability density for the rth group is

\[ f_r = \text{const. exp} - \frac{1}{2} \left\{ \sum X_j (x_i - \mu_{ir})(x_j - \mu_{jr}) \right\} \]

The surfaces of constant likelihood ratios are defined by

\[ \sum_j \left( \sum \lambda_r^s (\mu_{ir} - \mu_{jr}) \right) x_j = \text{const.} \]

These surfaces can also be defined in terms of what may be called linear discriminant scores defined in terms of the constants for the rth group only.

\[ L_r = \sum_j \left( \sum \lambda_r^j (\mu_{ir} - \mu_{jr}) \right) x_j - \frac{1}{2} \sum \lambda_r^j \mu_{ir} \mu_{jr} \]

A constant likelihood ratio corresponds to a constant difference in the discriminant scores. If the a priori probabilities are \( \pi_1, \pi_2, \pi_3 \) for the three groups, then the rule of procedure is to assign an observed individual to that group for which

\[ L_r + \log \pi_r \]

is a maximum.

Example 1. The scores in three tests, A, B, and C, of 256 army recruits classified by their neurotic condition have the mean values as given in Table 8b.9a.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety state</td>
<td>114</td>
<td>2.0228</td>
<td>1.1667</td>
</tr>
<tr>
<td>Hysteria</td>
<td>32</td>
<td>3.0308</td>
<td>1.2424</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>32</td>
<td>3.8125</td>
<td>1.8438</td>
</tr>
<tr>
<td>Obsession</td>
<td>17</td>
<td>4.7059</td>
<td>1.5882</td>
</tr>
<tr>
<td>Personality change</td>
<td>5</td>
<td>1.4000</td>
<td>0.2000</td>
</tr>
<tr>
<td>Normal</td>
<td>55</td>
<td>0.6000</td>
<td>0.1455</td>
</tr>
</tbody>
</table>

The dispersion matrix within the groups and its reciprocal are given below.

\[
\begin{align*}
\text{Within Dispersion Matrix (} \lambda^j \text{)} & \quad \text{Reciprocal (} \lambda^j \text{)} \\
A & B & C & A & B & C \\
A & 2.300851 & 0.251578 & 0.474169 & 0.543234 & -0.200195 & -0.420813 \\
B & 0.251578 & 0.607466 & 0.033774 & -0.200195 & 1.725807 & 0.005767 \\
C & 0.474169 & 0.033774 & 0.595094 & -0.420813 & 0.005767 & 2.012357
\end{align*}
\]
APPLICATION TO MULTIVARIATE NORMAL POPULATIONS

For any group the linear discriminant score is

\[ L = l_1 A + l_2 B + l_3 C - \frac{1}{2}(l_1 m_1 + l_2 m_2 + l_3 m_3) \]

where

\[ l_1 = \lambda^{11} m_1 + \lambda^{12} m_2 + \lambda^{13} m_3 \]
\[ l_2 = \lambda^{21} m_1 + \lambda^{22} m_2 + \lambda^{23} m_3 \]
\[ l_3 = \lambda^{31} m_1 + \lambda^{32} m_2 + \lambda^{33} m_3 \]

\( m_1, m_2, m_3 \) are the mean values of \( A, B, C \), and the elements \( \lambda^{ij} \) belong to the reciprocal of the dispersion matrix. For the anxiety state group

\[ m_1 = 2.9298 \quad m_2 = 1.1667 \quad m_3 = 0.7281 \]
\[ l_1 = 0.5432(2.9298) - 0.2002(1.1667) - 0.4208(0.7281) = 1.0515 \]
\[ l_2 = -0.2002(2.9298) + 1.7258(1.1667) + 0.0558(0.7281) = 1.4676 \]
\[ l_3 = -0.4208(2.9298) + 0.0558(1.1667) + 2.0124(0.7281) = 0.2975 \]
\[ \frac{1}{2}(l_1 m_1 + l_2 m_2 + l_3 m_3) = \frac{1}{2}(1.0515(2.9298) + \cdots) = 2.5047 \]

Hence the discriminant score for the anxiety state is

\[ L = 1.0515A + 1.4676B + 0.2975C - 2.5047 \]

For purposes of classification the expression to be calculated is \( L + \log \pi \), where \( \pi \) denotes the relative frequency of cases of anxiety state. The discriminant scores involving the relative frequencies are given in Table 8b.9B.

The present data are not a representative sample of officers serving in the Army and the Navy. The sample of neurotic officers has not been exposed to any known bias, and the proportions between the numbers in the various groups may be fairly representative; but the number of normal officers is grossly under-represented. It seems impossible to obtain a reliable general estimate of the risk that a man will be referred to a hospital for the treatment of a neurosis while he is serving in the Army or the Navy as an officer; but the indications are that, even under conditions of very severe stress, it is not more than 2 to 3 per cent. For proportional representation over 100 times as many normal cases should have been reported.

In Table 8b.9B the formulae have been given in terms of general relative frequencies and also for the particular values realized in the sample although they are subject to systematic as well as to chance errors. The formulae are unsuitable for practical use unless reliable estimates of the relative frequencies are available.
### TABLE 8b.9. The Linear Discriminant Scores for Various Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Coefficients of Measurements</th>
<th>Constant Term</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A)</td>
<td>(B)</td>
</tr>
<tr>
<td>Normal Personality change</td>
<td>0.2050</td>
<td>0.1481</td>
</tr>
<tr>
<td>Anxiety state</td>
<td>0.7204</td>
<td>0.0649</td>
</tr>
<tr>
<td>Hysteria</td>
<td>1.0615</td>
<td>1.4676</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>1.3599</td>
<td>2.4641</td>
</tr>
<tr>
<td>Obsession</td>
<td>1.7680</td>
<td>1.8611</td>
</tr>
</tbody>
</table>

\* \(r_1 = 0.21454, r_2 = 0.01053, r_3 = 0.44531, r_4 = 0.12691, r_5 = 0.12500, r_6 = 0.06641\).

Given the measurements \(A, B, C\) of an individual, we calculate the linear discriminant scores \(L_1 \cdots L_6\), corrected for a priori probabilities, and assign him to the group for which his score is highest. If the a priori probabilities are not known, the maximum likelihood method leads to the rule of assigning an individual to that group for which \(L\) is highest.

### Example 2.
Table 8b.9y gives the statistical constants for three groups of individuals measured by D. N. Majumdar. These groups are considered again in Chapter 9 where the constants for all available characters are given.

### TABLE 8b.9y. Statistical Constants for Three Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Stature Mean Value</th>
<th>Height</th>
<th>Depth</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahmin</td>
<td>164.51</td>
<td>86.43</td>
<td>25.49</td>
<td>51.24</td>
</tr>
<tr>
<td>Artisan</td>
<td>160.53</td>
<td>81.47</td>
<td>23.84</td>
<td>48.62</td>
</tr>
<tr>
<td>Korwa</td>
<td>158.17</td>
<td>81.16</td>
<td>21.44</td>
<td>46.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Sitting</th>
<th>Nasal</th>
<th>Naosal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahmin</td>
<td>(St)</td>
<td>(NSt)</td>
<td>(MSt)</td>
</tr>
<tr>
<td>Artisan</td>
<td>164.51</td>
<td>86.43</td>
<td>25.49</td>
</tr>
<tr>
<td>Korwa</td>
<td>158.17</td>
<td>81.16</td>
<td>21.44</td>
</tr>
</tbody>
</table>

Dispersion Matrix:

\[
\begin{bmatrix}
32.96 & 7.43 & 1.78 & 3.97 \\
10.24 & 1.17 & 2.43 \\
3.06 & 1.78 \\
12.22 &
\end{bmatrix}
\]
APPLICATION TO MULTIVARIATE NORMAL POPULATIONS

By inverting the dispersion matrix we can obtain the linear discriminant scores, as in the illustration of neurotic groups, and use them for classification. In order to determine the probabilities of wrong classification for each group it is necessary to go through a slightly complicated procedure. If there are only three groups, the four (in general, $p$) measurements can be replaced by two independent linear functions, given which the relative distributions in all the groups become identical. The problem is thus reduced to a two-variable case. The two independent functions can be obtained in a number of ways. One simple method is to calculate the discriminant functions for any two pairs of groups. The computational method is to write down the dispersion matrix with two appended columns:

$$
\begin{array}{cccc}
32.95 & 7.43 & 1.78 & 3.97 \\
10.24 & 1.17 & 2.43 & 4.96 \\
3.96 & 1.78 & 1.63 & 2.40 \\
12.25 & 2.62 & 1.90 & \\
\end{array}
$$

The two sets of solutions give the two discriminant functions for the pairs Brahmin, Artisan and Artisan, Korwa,

$$
X = -0.0039S_1 + 0.4301S_H + 0.3293N_D + 0.0819N_H
$$

$$
Y = 0.0476S_1 - 0.1036S_H + 0.7679N_D + 0.0486N_H
$$

and the discriminant function for Brahmin, Korwa is $X + Y$, which have the following mean values.

$$
\begin{array}{ccc}
& X & Y \\
Brahmin & 49.1224 & 70.0630 \\
Artisan & 46.2467 & 66.1173 \\
Korwa & 45.1766 & 63.0317 \\
\end{array}
$$

The discriminant function for Artisan – Korwa gives the rule for distinguishing Artisan from Korwa when

$$
Y \geq \frac{19.8706 + 17.8551}{2} = 18.8628
$$

Similarly, Artisan is distinguished from Brahmin when

$$
X \leq \frac{49.1224 + 46.2467}{2} = 47.6845
$$
Therefore, the maximum likelihood method of classification for Artisan is \( Y \geq 18.8628, X \leq 47.6845 \). Similarly, by considering the discriminant functions between Brahmin, Artisan and Brahmin, Korwa, the rule for Brahmin is

\[
X \geq 47.6845 \quad X + Y \geq 66.5473
\]

For Korwa the rule is

\[
Y \leq 18.8628 \quad X + Y \leq 66.5473
\]

obtained by considering the two discriminant functions separating Korwa from Artisan and Korwa from Brahmin. For instance consider an individual with

\[
St = 162.00 \quad SH = 84.00 \quad ND = 24.00 \quad NH = 49.00
\]

The value of \( X = 47.4129, Y = 19.8198, \) and \( X + Y = 67.2327 \). Since \( 19.8198 > 18.8628 \) and \( 47.4129 < 47.6845 \) the individual is assigned to the Artisan group.

Figure 4 gives the two-dimensional chart for \( X \) and \( Y \) with respect to which the individuals can be classified. The point \((X, Y)\) for the

\[
\begin{align*}
A_i &\quad (46.25, 19.87) \\
&\quad (47.41, 19.82)

d\quad (49.12, 20.94)
\end{align*}
\]

**Figure 4.** The regions separating the three groups.
observed individual is represented by $I$ and the mean values by $B_M$, $A_M$, and $K_M$.

To determine the errors of classification we need find the variances and covariances of $X$ and $Y$ which can be simply obtained from the mean values.

\[
V(X) = B_X - A_X = 49.1224 - 46.2467 = 2.8757
\]
\[
V(Y) = A_Y - K_Y = 19.8706 - 17.8551 = 2.0155
\]
\[
\text{cov}(X, Y) = A_X - K_X = 46.2467 - 45.1767
\]
\[
= B_Y - A_Y = 20.9406 - 19.8706 = 1.0700
\]
\[
V(X + Y) = V(X) + V(Y) + 2 \text{cov}(X, Y)
\]
\[
= 2.8757 + 2.0155 + 2(1.0700) = 7.0312
\]
\[
\text{cov}(X, X + Y) = V(X) + \text{cov}(X, Y)
\]
\[
= 2.8757 + 1.0700 = 3.9457
\]
\[
\text{cov}(Y, X + Y) = V(Y) + \text{cov}(X, Y)
\]
\[
= 2.0155 + 1.0700 = 3.0855
\]

The correlation matrix of $X$, $Y$ and $X + Y$ is

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th>$Y$</th>
<th>$X + Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>0.4459</td>
<td>0.8810</td>
<td></td>
</tr>
<tr>
<td>$Y$</td>
<td></td>
<td>0.8200</td>
<td></td>
</tr>
<tr>
<td>$X + Y$</td>
<td>1.69</td>
<td>1.42</td>
<td>2.65</td>
</tr>
</tbody>
</table>

The proportion of right classification for Brahmin is

\[ P(X \geq 47.6845, X + Y \geq 66.5473) \]

which give the deviates

\[
\frac{47.6845 - 49.1224}{1.69} = -0.85 \quad \text{and} \quad \frac{66.5473 - 70.0630}{2.65} = -1.33
\]

The probability for $h > 0.85$, $k > 1.33$ and $r = 0.88$ is given in Part II of Pearson’s tables for statisticians and biometricians. This value is approximately 0.085. The required probability for wrong classification is

\[ P(h > 0.85) + P(k > 1.33) - P(h > 0.85, k > 1.33) \]
\[ = 0.195 + 0.092 - 0.085 = 0.202 \]
where the first two probabilities are obtained from univariate normal tables.

Similarly, for Korwa the deviates are

\[ h = 0.71 \quad \text{and} \quad k = 1.33, \quad r = 0.82 \]

and the probability for correct classification is

\[ A \leq 0.71 \quad \text{and} \quad k \leq 1.33 \]

The tabular value gives the probability 0.085 for \( h > 0.71 \) and \( k > 1.33 \).

The probability for wrong classification is

\[
P(h > 0.71) + P(k > 1.33) - P(h > 0.71, k > 1.33) = 0.239 + 0.092 - 0.085 = 0.246
\]

Similarly, for Artisan the deviates are

\[ h = 0.85 \quad \text{and} \quad k = -0.71, \quad r = 0.45 \]

and the probability for correct classification is

\[ P(h < 0.85, k > -0.71) \]

Since one of the deviates is negative, the tabular entry for \( h > 0.85 \) and \( k > 0.71 \) has to be obtained, taking \( r \) to be negative. For \( r = -0.44 \) the value is 0.013. The probability for wrong classification is

\[
P(h > 0.85) + P(k > 0.71) - P(h > 0.85, k > 0.71) = 0.195 + 0.239 - 0.013 = 0.421
\]

8b.10 Allocation of a Number of Individuals to Two or More Groups

Suppose that \( n_1 \) and \( n_2 \) posts have to be filled in the Navy and the Air Force, a candidate being chosen on the basis of his performance in a test. Assuming that the distribution of test scores for those who are fit for the Navy and the Air Force are available from past experience, how can this knowledge be used for the most efficient selection?

In the actual population the relative proportions of candidates suitable for the Navy may be different from \( n_1:n_2 \). The procedure for selection must be such that, whatever may be the actual proportion in the population, the division of a sample of \((n_1 + n_2)\) individuals in the assigned ratio \( n_1:n_2 \) should involve the least possible errors. A similar problem is the allocation of a given number of skulls into two sexes in a given ratio which is determined from some a priori considerations. This may be only an estimated proportion and hence may not represent the true sex ratio. Whatever criterion is chosen, some male skulls will be
classified as female and vice versa. A procedure which gives the least value to the expected number of wrong classifications in either group may be regarded as the best one.

There may arise another situation. Two samples of sizes \( n_1 \) and \( n_2 \) drawn independently from the first and second groups may get mixed. The difference between the first and second problems is that in the latter every sample is known to consist of \( n_1 \) individuals from the first group and \( n_2 \) from the second whereas in the former no such information is available, the sample being drawn at random from a mixed population.

Solution to the First Problems. Let \( x_1, \ldots, x_n \) represent the measurements on \( n \) individuals. As observed earlier, \( x_i \) will stand for all the available set of measurements on the \( i \)-th individual. The probability of the set is

\[
\prod_{i=1}^{n} [\pi_1 f_1(x_i) + \pi_2 f_2(x_i)]
\]

Consider the following set of functions \( \delta_i(x_1, \ldots, x_n), \delta'_i(x_1, \ldots, x_n) \quad i = 1, \ldots, n \) which can be represented simply as \( \delta_i, \delta'_i \), satisfying the conditions

\[
\delta_i = 0 \text{ or } 1 \quad \delta_i + \delta'_i = 1
\]

and

\[
\sum_i \delta_i = n_1 \quad \sum_i \delta'_i = n_2
\]

where \( n_1 \) and \( n_2 \) are the specified numbers. If the individual with measurements \( x_i \) is assigned to the first group when \( \delta_i = 1 \) and to the second when \( \delta'_i = 1 \), then the above set of functions constitutes a decision rule. The problem is then to construct the above functions such that the expected risk associated with this decision rule is a minimum.

To calculate the expected risk we need know as a datum of the problem the loss of assigning an individual of one group to another. If \( r_{12} \) represents the loss in assigning an individual of the first group to the second and \( r_{21} \) the loss in the other case, then the quantity to be minimized is the expected value of

\[
\sum_{i=1}^{n} (\delta_i a_{1i} + \delta'_i a_{2i})
\]

where

\[
a_{1i} = \frac{r_{12} \pi_2 f_2(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)} \quad a_{2i} = \frac{r_{12} \pi_1 f_1(x_i)}{\pi_1 f_1(x_i) + \pi_2 f_2(x_i)}
\]

The expected loss will be a minimum if \( \delta_i \) and \( \delta'_i \) are chosen such that the expression (8b.10.1) has the least value. The problem is the same.
as that treated in lemma 2 of Appendix A5 leading to the solution
\[ \delta_i = 1 \quad \text{if} \quad a_{ii} + \mu_1 \leq a_{ij} + \mu_2 \]
\[ \delta_i' = 1 \quad \text{if} \quad a_{ii} + \mu_1 \geq a_{ij} + \mu_2 \]
where \( \mu_1 \) and \( \mu_2 \) are suitably chosen to satisfy the condition \( \Sigma_4 = n_1 \).

Now \( (a_{ii} - a_{ij}) \leq (\mu_2 - \mu_1) \) implies that
\[ \frac{\tau_{21} f_2(x_i) - \tau_{12} f_1(x_i)}{\tau_{11} f_1(x_i) + \tau_{21} f_2(x_i)} \leq \mu \]
or
\[ \frac{f_1(x_i)}{f_2(x_i)} \geq \lambda \]
so that the decision rule reduces to the evaluation of the likelihood ratios \( \lambda = f_1(x_i)/f_2(x_i) \) and assigning all the individuals with highest \( n_1 \) values of the ratios to the first group and the rest to the second. Fortunately the decision rule is independent of the a priori probabilities and also the loss function.

Corresponding to every decision rule we can set up a density function of the observations by considering all individuals assigned to the first group as having been drawn at random from the first group and similarly for the second. By using lemma 1 in Appendix A5 we find that the best decision rule found above maximizes the corresponding probability density. As shown below this forms the basis on which the solution of the second problem depends.

**Solution to the Second Problem.** In this problem the mixture is known to consist of \( n_1 \) individuals drawn from the first group and \( n_2 \) from the second. The observations \( x_1, \ldots, x_n \) could have arisen in \( \binom{n}{n_1} \) ways, any subset of \( n_1 \) observations belonging to the first group. The probability density of the observations is equal to the sum of the densities associated with \( \binom{n}{n_1} \) ways of splitting the sample. If \( x_a, x_b, \ldots \) and \( x_p, x_q, \ldots \) represent a division into two groups of sizes \( n_1 \) and \( n_2 \), then the probability density of the observations can be written as
\[ P(x_1, \ldots, x_n) = \sum f_1(x_a)f_2(x_b) \cdots f_2(x_p)f_2(x_q) \cdots \] (8b.10.2)
where the summation is over \( \binom{n}{n_1} \) such terms. Corresponding to any one of \( \binom{n}{n_1} \) possible decision rules the loss relative to the given set of
observations is $1/P(x_1, \ldots, x_n)$ times

$$\Sigma'l(a, b, \ldots; p, q, \ldots)f_1(x_a)f_1(x_b)\ldots f_2(x_p)f_2(x_q)\ldots \quad (8b.10.3)$$

where $l(a, b, \ldots; p, q, \ldots)$ is the loss incurred in adopting a given decision rule when in fact $x_a, x_b, \ldots$ come from the first group and $x_p, x_q, \ldots$ from the second. The loss will generally be a function of the number of wrong classifications only. That decision rule for which the expression $(8b.10.3)$ is the least is the best in the sense that it minimizes the expected loss. The solution depends on the evaluation of the expression $(8b.10.3)$ for each of the $\binom{n}{n_1}$ decision rules, and this makes the application a little difficult in practice even when $n$ is small.

As an alternative we may try to minimize the maximum loss incurred by following a decision rule. If we suppose that the loss function is proportional to the number of wrong classifications, then the maximum loss occurs when all the individuals in the smaller group are wrongly classified. With this assumption it can be shown that the division of the sample corresponding to the maximum probability density supplies the best possible solution to the problem. This solution may be referred to as the maximum likelihood solution; we consider the $\binom{n}{n_1}$ ways of splitting the sample as associated with $\binom{n}{n_1}$ different hypotheses concerning the individuals in the sample and choose that hypothesis which has the maximum likelihood.

To prove the property referred above, consider any other decision rule leading to a division

$$x_{a1}, x_{a2}, \ldots, x_{b1}, x_{b2}, \ldots$$

and $x_{a1}, x_{a2}, \ldots, x_{b1}, x_{b2}, \ldots$ of the sample into two groups of sizes $n_1$ and $n_2$ and compare with the division

$$x_{a1}, x_{a2}, \ldots, x_{b1}, x_{b2}, \ldots$$

associated with the maximum density. The measurements classified in the same way by the two decision rules are represented by $x_{a1}, x_{a2}, \ldots$ for the first group and by $x_{b1}, x_{b2}, \ldots$ for the second group. By definition

$$\frac{f_1(x_{a1})}{f_2(x_{b1})} \geq \frac{f_1(x_{b2})}{f_2(x_{a2})} \quad (8b.10.4)$$
and the same is true for the product of a number of ratios involving $x_a$ and the product of the same number of ratios involving $x_b$.

Let $n_2 \leq n_1$ without loss of generality. In this case the maximum loss occurs, by following the first decision rule when

$$x_{a1}, x_{a2}, \ldots, x_{a_{n2}}, x_{a_{n1}} \ldots$$

arise from the first group in which case a subset $n_2$ out of

$$x_{b1}, x_{b2}, \ldots, x_{b_{n1}}, x_{b_{n2}} \ldots$$

arise from the second group. Let this subset be

$$x_{b1}, x_{b2}, \ldots, x_{b_{n1}}, x_{b_{n2}} \ldots$$

By replacing the subscript $c$ by $b$ we obtain the corresponding situation for the proposed maximum likelihood decision rule. The difference between the above two probability densities associated with maximum errors for the two rules is, apart from a common multiplier, equal to

$$[f_1(x_{a1})f_1(x_{a2}) \cdots f_2(x_{b1})f_2(x_{b2}) \cdots - f_1(x_{a1})f_1(x_{a2}) \cdots f_2(x_{b1})f_2(x_{b2}) \cdots]$$

which is not less than zero according to (8b.10.4). By considering all subsets of $n_2$ observations out of (8b.10.6) we exhaust all possible ways in which the observations leading to maximum error according to the first rule can arise. To each such case there is a corresponding division leading to the maximum error for the proposed rule. But this division leads to a smaller probability density. The total chance of maximum error relative to the given set of observations is thus a minimum for the maximum likelihood decision rule.

The Problem of Three Groups. As in the case of two groups we consider two situations, firstly when the sample consists of $n$ individuals observed at random from a mixed population and secondly when the sample is a mixture of $n_1$ individuals drawn from the first group, $n_2$ from the second, and $n_3$ from the third. The problem in either case is to select $n_1$ individuals for the first group, $n_2$ for the second, and $n_3$ for the third where $n_1 + n_2 + n_3 = n$.

Let $f_1(x), f_2(x), f_3(x)$ represent the probability densities of $x$ for the three groups and $\pi_1, \pi_2, \pi_3$ the proportions of mixture in the general population. The loss in assigning a person to the $i$th group when, in fact, he belongs to the $j$th group is denoted by $r_{ji}$. The a posteriori risks in assigning an individual with measurements $x_i$ to the first, sec-
Consider a set of functions
\[
\delta_i = 0 \text{ or } 1 \\
\delta_i' = 0 \text{ or } 1 \\
\delta_i'' = 0 \text{ or } 1
\]
such that
\[
\sum \delta_i = n_1 \\
\sum \delta_i' = n_2 \\
\sum \delta_i'' = n_3
\]
They define a decision rule if the individual with measurements $x_i$ is assigned to the first group when $\delta_i = 1$, to the second when $\delta_i' = 1$, and to the third when $\delta_i'' = 1$. The a posteriori risk for such a selection procedure is
\[
\sum \binom{\delta_i}{a_{1i}} + \binom{\delta_i'}{a_{2i}} + \binom{\delta_i''}{a_{3i}}
\]
The best decision rule is one that minimizes the above expression. This is exactly the problem solved in lemma 2 of Appendix A5. The best solution is
\[
\delta_i = 1 \quad \text{when } a_{1i} + \lambda_1 \leq a_{2i} + \lambda_2, \quad a_{1i} + \lambda_1 \leq a_{3i} + \lambda_3 \\
\delta_i' = 1 \quad \text{when } a_{2i} + \lambda_2 \leq a_{1i} + \lambda_1, \quad a_{2i} + \lambda_2 \leq a_{3i} + \lambda_3 \\
\delta_i'' = 1 \quad \text{when } a_{3i} + \lambda_3 \leq a_{1i} + \lambda_1, \quad a_{3i} + \lambda_3 \leq a_{2i} + \lambda_2
\]
where $\lambda_1, \lambda_2, \lambda_3$ are determined such that $\sum \lambda_i = n_1, \sum \lambda_i' = n_2,$ and $\sum \lambda_i'' = n_3$. As it stands, the problem of determination of $\lambda_1, \lambda_2, \lambda_3$ appears to be complicated. There is a geometrical device which is helpful in the solution of the problem. In higher dimensional cases involving four or more groups the geometrical method cannot be ap-
CLASSIFICATORY PROBLEMS

For three groups we replace $a_{i1}, a_{i2}, a_{i3}$ by two coordinates

$$X_i = a_{i1} - a_{i2} \quad Y_i = a_{i1} - a_{i3}$$

and represent the $n$ points $(X_i, Y_i)$ on a two-dimensional chart with rectangular axes. The problem is to determine a point $(X_0, Y_0)$ on this chart such that the regions formed by the lines $X = X_0$, $Y = Y_0$, and $Y - Y_0 = X - X_0$ contain the requisite number of points. This can be done by moving three thin rods $OX', OY', OZ'$ fixed at the point $O$ as shown in Figure 5, with $OX'$ and $OY'$ parallel to the $X$ and $Y$ axes, and arrive at the required division by trial and error. It will help in this process if the numbers falling in the three regions are recorded for a few positions of the frame with the frame marked on the chart.

To solve the second problem when the sample consists of a mixture of $n_1$ individuals drawn from the first group, $n_2$ from the second, and $n_3$ from the third we can set up the total risk relative to the given set of observations for each of $n_1!n_2!n_3!$ possible decision rules and choose that rule for which the risk is a minimum. This is very difficult in practice so that a simplified procedure is needed. As before we may choose that decision rule which leads to a division of the sample with the maximum probability density. This rule possesses an important property that

![Figure 5. The arrangement of three thin rods leading to the required division.](image)
the probability of the maximum number of wrong classifications for any one group is as small as possible.

The method of arriving at the required division is first to obtain the quantities

\[ a_{1i} = \log f_1(x_i) \quad a_{2i} = \log f_2(x_i) \quad a_{3i} = \log f_3(x_i) \]

\[ i = 1, \ldots, n \]

and plot the points

\[ X_i = a_{2i} - a_{1i} \quad Y_i = a_{3i} - a_{1i} \]

and proceed geometrically as in Figure 5. For this division the probability density will be a maximum.

The problems treated in 8b fit in with the general decision function theory developed by Wald (1949).

8c Discriminant Function for Selecting Genetically Desirable Types

8c.1 Prediction Formula for the Genotypic Value

When quantitative characters are involved, it is considerably difficult to select genetically desirable types in breeding work because heritable differences are to some extent masked by non-heritable or environmental variations. The problem then arises as to what is the best indicator of the genotypic value of any individual line. Suppose the desired quality in the plant is yield. The observed yield is no doubt a good measure, but, if the factors influencing the yield affect to some extent other observable characters of the plant, then these latter characters can also be used in assessing the strength of factors responsible for yield. This can also be looked on as a problem of prediction: How can the genotypic value with respect to some characteristic be predicted when measurements on a number of observable characters are available?

This problem can be extended further to cases where the quality of a line is determined not by a single character but by a given linear compound of the genotypic values corresponding to a number of characteristics. The coefficients of the linear compound are fixed by the relative worth of these characters in assessing the quality of the line as a whole.

Thus, in poultry, the annual number of eggs laid \((x_1)\), the size of the egg \((x_2)\), and the age at maturity \((x_3)\) are some of the important factors to be considered. To what extent these can be combined in a breed depends on the genetic relationships among these characters, and in any breeding program best use should be made of the available material.
If \( \psi_1, \psi_2, \psi_3 \) represent the genotypic values of the three characters mentioned above, the breeder’s interest is in the value of a linear compound \( a_1\psi_1 + a_2\psi_2 + a_3\psi_3 \) which corresponds to the commercial value of the bird for properly chosen values of the compounding coefficients. For instance, Panse (1916) considers three sets of weights of which one is

\[ a_1 = 8 \quad a_2 = 5 \quad a_3 = -2 \]

These weights depend on the relative importance to be attached to each of these characters and are assigned by the animal breeder. The cash return from each bird depends on the number of eggs laid and also on the size of the egg. The age at maturity is also important when the cost of feeding the bird in the period from its hatching to the date of laying the first egg is considered. For this reason the age at maturity is given a negative weight. The highest weight is given to the annual number of eggs laid.

To estimate the above linear compound it may be useful, as observed earlier, to consider other characters in which the animal breeder is not directly interested. In the material on poultry analyzed by Panse (1946) the body weight \( (X_4) \) is also available and the best predictor may be written as

\[ b_0 + b_1\bar{x}_1 + \cdots + b_4\bar{x}_4 \]

where \( \bar{x}_1, \bar{x}_2, \bar{x}_3, \) and \( \bar{x}_4 \) are the mean values for a sire. In fact, any number of extraneous characters can be considered in building up the prediction formula, and it is not necessary that the formula be linear in the measurements although linearity introduces a great simplification in actual computation. To determine these coefficients, Smith (1936) maximized the regression between the two linear compounds \( a_1\psi_1 + a_2\psi_2 + a_3\psi_3 \) and \( b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + b_4\bar{x}_4 \), which is equivalent to minimizing the sum of squares with due weights

\[ \sum_{r=1}^{n}(a_1\psi_{1r} + a_2\psi_{2r} + a_3\psi_{3r} - b_0 - b_1\bar{x}_{1r} - \cdots - b_4\bar{x}_{4r})^2 \]

where \( n_r \) is the sample size on which the mean values \( \bar{x}_{1r}, \cdots, \bar{x}_{4r} \) are based for the \( r \)th sire and \( \psi_{1r}, \psi_{2r}, \psi_{3r} \) are the genotypic values for the \( r \)th sire.

The minimizing equations for \( b_1, b_2, b_3, \) and \( b_4 \) are

\[ \begin{align*}
&b_1B_{11} + b_2B_{12} + b_3B_{13} + b_4B_{14} = a_1G_{11} + a_2G_{12} + a_3G_{13} \\
&b_1B_{21} + b_2B_{22} + b_3B_{23} + b_4B_{24} = a_1G_{21} + a_2G_{22} + a_3G_{23} \\
&b_1B_{31} + b_2B_{32} + b_3B_{33} + b_4B_{34} = a_1G_{31} + a_2G_{32} + a_3G_{33} \\
&b_1B_{41} + b_2B_{42} + b_3B_{43} + b_4B_{44} = a_1G_{41} + a_2G_{42} + a_3G_{43}
\end{align*} \]
where \( B_{ij} \) is the sum of products between sires and

\[
G_{ij} = \sum_{r} (y_{ir} - \bar{y}_{i}) (y_{jr} - \bar{y}_{j})
\]

\[
\bar{y}_{i} = \frac{\sum_{r} y_{ir}}{n_{r}} \quad \text{and} \quad \bar{y}_{j} = \frac{\sum_{r} y_{jr}}{n_{r}}
\]

The actual values of \( G_{ij} \) are not known, but their expected values are obtained from the equations

\[
E(B_{ij}) = E(G_{ij}) + (k - 1)\sigma_{ij}
\]

where \( \sigma_{ij} \) is the expected covariance between the \( i \)th and \( j \)th characters within a sire. If \( S_{ij} \) is the sum of products within a sire, then an estimate of \( G_{ij} \) is

\[
wG_{ij} \sim wB_{ij} - (k - 1)S_{ij}
\]

where \( w \) is the degrees of freedom within sires, and \( (k - 1) \) the degrees of freedom between sires.

The analysis of sum of squares and products, the estimated values of \( G_{ij} \), and other related values are given in Table 8c.1a.

**Table 8c.1a. Analysis of Dispersion**

<table>
<thead>
<tr>
<th>Due to</th>
<th>Between ( (B_{ij}) )</th>
<th>Within ( (S_{ij}) )</th>
<th>( wB_{ij} - (k - 1)S_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>( 6,476.4 )</td>
<td>67,797.3</td>
<td>352,594.2</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( 982.10 )</td>
<td>2,440.14</td>
<td>163,240.14</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( 37,422 )</td>
<td>317,982</td>
<td>2,070,074</td>
</tr>
<tr>
<td>( \delta )</td>
<td>( 335.16 )</td>
<td>3,600.21</td>
<td>16,124.22</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( 678.16 )</td>
<td>307.53</td>
<td>131,160.54</td>
</tr>
<tr>
<td>( \xi )</td>
<td>( 6,043.8 )</td>
<td>35,898.6</td>
<td>712,223.4</td>
</tr>
<tr>
<td>( \omega )</td>
<td>( 750.54 )</td>
<td>172.86</td>
<td>153,278.58</td>
</tr>
<tr>
<td>( \phi )</td>
<td>( 825.86 )</td>
<td>404.01</td>
<td>160,341.72</td>
</tr>
<tr>
<td>( \pi )</td>
<td>( 708.12 )</td>
<td>1,370.82</td>
<td>161,523.60</td>
</tr>
</tbody>
</table>

\( \alpha \beta \gamma \delta \eta \xi \omega \) : 67,289.60 616,014.75 4,901,003.10 *

\( \alpha \beta \gamma \delta \eta \xi \omega \) : 8,684.06 14,335.32 1,544,801.58 *

\( \alpha \beta \gamma \delta \eta \xi \omega \) : 67,289.60 616,014.75 4,901,003.10 *

\( \alpha \beta \gamma \delta \eta \xi \omega \) : 5,829.18 7,089.76 1,065,408.54 *

* These quantities can be obtained in two ways: (i) directly from the \( G_{ij} \) values above, and also (ii) from \( B_{ij} \) and \( S_{ij} \) values of the linear compounds, thus providing a check. These form the right-hand expressions of the equations for \( b \). Since only the ratios of \( b \) are important, these quantities are divided by 1000 and corrected to three decimal places to obtain the same order of figures as those in the equations.
The equations are
\[
\begin{align*}
6476.4b_1 + 678.16b_2 - 6043.8b_3 + 750.54b_4 &= 4901.003 \\
678.16b_1 + 982.10b_2 + 825.86b_3 + 248.22b_4 &= 1544.802 \\
-6043.8b_1 + 825.86b_2 + 37422b_3 + 708.12b_4 &= -11036.227 \\
750.54b_1 + 248.22b_2 + 708.12b_3 + 335.16b_4 &= 1065.409
\end{align*}
\]
yielding the solutions
\[
\begin{align*}
b_1 &= -0.16274 \\
b_2 &= 1.12795 \\
b_3 &= -0.41387 \\
b_4 &= 3.58233
\end{align*}
\]
Strangely, the character which is scored quite high has a negative coefficient in the discriminant function, which leads us to suspect the reliability of the material. If, however, body weight, which has a suspiciously low variance, is omitted from consideration, the new weights are
\[
\begin{align*}
b_1 &= 0.33484 \\
b_2 &= 1.57346 \\
b_3 &= -0.27556
\end{align*}
\]
leading to the discriminant function
\[
0.33484x_1 + 1.57346x_2 - 0.27556x_3
\]
against a straight selection function
\[
8x_1 + 5x_2 - 2x_3
\]
It is seen that although the number of eggs is scored high in the latter function its selective value is not so high as that of the egg weight as shown by the calculated discriminant function.

First, it is necessary to test whether the genotypic value \(a_1\psi_1 + a_2\psi_2 + a_3\psi_3\) is significantly different between the sires. An answer to this is provided by an analysis of the linear compound \(a_1x_1 + a_2x_2 + a_3x_3\) between and within the sires, as shown in Table 8c.2a. The discriminant function obtained above is a better estimate of the genotypic value \(a_1\psi_1 + a_2\psi_2 + a_3\psi_3\), and it is unnecessary to test whether the estimated function sufficiently discriminates between the sires. What is more important is the construction of an index to measure the advantages in selection by means of the discriminant function and also a test of its significance.

8c.2 The Genetic Advance

Let \(\psi\) be a characteristic as measured by a character \(x\) with mean \(\mu\) and variance \(\sigma^2\). Let the regression of \(\psi\) on \(x\) be \(\beta\); then the expected value of \(\psi\) for a given \(x\) is
\[
\psi + \beta(x - \mu)
\]
THE GENETIC ADVANCE

where \( \bar{\psi} = E(\psi) \) for all \( x \). If \( x \) is normally distributed, then the expected value of \( \psi \) for all \( x \) exceeding the upper \( q \)th part is

\[
\frac{1}{q} \int_{z_q}^{\infty} \left( \bar{\psi} + \beta(x - \mu) \right) e^{-\frac{(x - \mu)^2}{2\sigma^2}} dx = \bar{\psi} + \frac{\beta\sigma}{q} \quad (1)
\]

where \( z \) is the ordinate to the normal curve at \( z_q \), the abscissa corresponding to the upper \( q \)th part of the normal curve.

Suppose that a large number of plant lines are available and a \( q \)th part of them is chosen for further propagation by the above method. The genetic advance then is

\[
\frac{\beta\sigma}{q}
\]

and, since \( z/q \) is common for all selection procedures, the intensity of genetic advance depends on \( \beta \sigma \), which is equivalent to \( \text{cov}(\psi_2)/\sigma \).

For selection of poultry the values of the discriminant function at the mean values are calculated and the sires corresponding to the highest values are chosen. It is seen that, if the means are based on a large number of observations corresponding to each sire, then the prediction of the genotypic value is more accurate and consequently the genetic advance is higher, the maximum genetic advance being available when the mean values are known exactly. So in the problem of evaluation of the genetic advance it is required to know how much experimentation is contemplated to assess the value of each sire. Suppose that the mean values for each sire are based on a sample of size \( n \). The genetic advance associated with any linear compound

\[
c_1\bar{\psi}_1 + c_2\bar{\psi}_2 + c_3\bar{\psi}_3
\]

is

\[
\text{cov}(c_1\bar{\psi}_1 + c_2\bar{\psi}_2 + c_3\bar{\psi}_3)
\]

\[
\sqrt{V(c_1\bar{\psi}_1 + c_2\bar{\psi}_2 + c_3\bar{\psi}_3)}
\]

The numerator has the expected value

\[
c_1(c_1\bar{\psi}_{11} + c_2\bar{\psi}_{12} + c_3\bar{\psi}_{13})
\]

\[
+ c_2(c_1\bar{\psi}_{12} + c_2\bar{\psi}_{22} + c_3\bar{\psi}_{23})
\]

\[
+ c_3(c_1\bar{\psi}_{13} + c_2\bar{\psi}_{23} + c_3\bar{\psi}_{33})
\]
and the square of the denominator $V(c_1 x_1 + \cdots + c_5 x_5)$ is

$$\frac{\sum \sum c_i c_j g_{ij}}{n}$$

where $g_{ij}$ is the expected covariance between the $i$th and $j$th characters within the sires and $g_{ij}$ is the covariance between the $i$th and $j$th genotypic values between the sires. The estimated values of $g_{ij}$ and $g_{ij}$ are substituted in the above expressions. These estimates are obtained by a direct analysis of variance and covariance of the linear compound $c_1 x_1 + c_2 x_2 + c_3 x_3$ and the straight selection function $a_1 x_1 + a_2 x_2 + a_3 x_3$.

The mean square between sires ($S_1$) for any linear compound $c_1 x_1 + c_2 x_2 + c_3 x_3$ has the expectation

$$\sum \sum c_i c_j g_{ij} + \sum \sum c_i c_j r_{ij}$$

where

$$\lambda = \frac{(S_1 - \sum n_i^2 / \sum n_i)}{(k - 1)}$$

In the above expression, $n_1, n_2, \cdots$ are the observations on the first, second, $\cdots$ sires. The mean square within sires ($S_2$) has the expectation $\sum \sum c_i c_j r_{ij}$, so that

$$\sum \sum c_i c_j r_{ij} \sim \frac{(S_1 - S_2)}{\lambda}$$

Hence the variance of the mean of the linear compound based on $n$ observations is

$$V(c_1 x_1 + c_2 x_2 + c_3 x_3) \sim \frac{S_1 - S_2}{\lambda} + \frac{\sum \sum c_i c_j r_{ij}}{n}$$

Similarly,

$$\text{cov} (a_1 x_1 + a_2 x_2 + a_3 x_3, c_1 x_1 + c_2 x_2 + c_3 x_3) = \frac{D_1 - D_2}{\lambda}$$

where $D_1$ and $D_2$ are the mean sums of products between and within sires for the two linear compounds, $a_1 x_1 + a_2 x_2 + a_3 x_3$ and $c_1 x_1 + c_2 x_2 + c_3 x_3$.

We shall illustrate the method for the linear compound determined above. The analysis of variance and covariance for the calculated function $b_1 x_1 + b_2 x_2 + b_3 x_3 = Y_1$ and the straight selection function $a_1 x_1 + a_2 x_2 + a_3 x_3 = Y_2$ is shown in Table 8c.2a.
### The Genetic Advance

335

#### Table 8c.2a. Analysis of Variance and Covariance of $Y_1$ and $Y_2$

<table>
<thead>
<tr>
<th></th>
<th>$Y_1^2$</th>
<th>$Y_2^2$</th>
<th>$Y_1^2Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.F.</td>
<td>S.S.</td>
<td>M.S.</td>
<td>S.S.</td>
</tr>
<tr>
<td>Between sires</td>
<td>14</td>
<td>7,112.875</td>
<td>509.0025</td>
</tr>
<tr>
<td>Within sires</td>
<td>201</td>
<td>44,449.820</td>
<td>221.1433</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>296.9192</td>
<td>24.521.8</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>20.791</td>
<td>1,776.948</td>
</tr>
<tr>
<td>Within sires</td>
<td></td>
<td>15.357</td>
<td>2,363.881</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30.148</td>
<td>4,140.829</td>
</tr>
</tbody>
</table>

(i) The values are obtained by the following formulae.

\[
7,112.875 = b(6001.003) + b_1(1044.801) + b_2(-11,036.227) \text{ (the values used in the equations for b from Table 8c.1a)}
\]

\[
44,449.820 = \sum b(\delta)
\]

\[
819.867.3 = a(6,739.80) + a(8,864.66) + a(-110,065.10)
\]

\[
60,004.85 = b(\quad) + b_1(\quad) + b_2(\quad)
\]

\[
6,842.060.1 = a(614,044.70) + a(14,339.32) + a(-201,132.73)
\]

\[
482,649.77 = b(\quad) + b_1(\quad) + b_2(\quad)
\]

(ii) The ratio of mean squares for $Y_2$ is $58,561.9/34,040.1 = 1.73$ which is on the 5% significant level for 14 and 201 degrees of freedom. Only when this is significant can we proceed to estimate the genotypic value.

The actual values of $n_1, n_2, \cdots, n_{15}$ available for $k = 15$ sires are not known in the above experiment. Let us suppose that the values are such that

\[
\lambda = \frac{(\Sigma n_r - \Sigma n_r^2/\Sigma n_r)}{(k - 1)} = 13.3
\]

and let us find the genetic advance for the value of $n = 14.4$, which is the average size in the above example. As observed earlier, we can calculate the genetic advance for any $n_i$ depending on the intensity of experimentation.

The index of genetic advance for straight selection associated with $a_1x_1 + a_2x_2 + a_3x_3$ is

\[
\lambda = \frac{\text{Difference}}{\text{Within}} \text{ for } Y_2^2 = \frac{1776.9}{\sqrt{4140.83}} = 27.613
\]
The numerator and denominator are both obtained from the column under $Y_1^2$ in Table 8c.2a. The expression for $b_1X_1 + b_2X_2 + b_3X_3$ is

$$\frac{\text{Difference}}{\lambda} \text{ for } Y_1Y_2 = \frac{183.166}{\sqrt{\frac{\text{Difference} + \text{Within}}{\lambda} \text{ for } Y_1^2}} = \frac{30.467}{\sqrt{36.148}}$$

so that the genetic advance is $\frac{30.467}{\sqrt{36.148}} - 1 \times 100 = 10.33\%$ higher for the discriminant function. It is difficult to test how far this observed increase is significant.

8d Problems of Optimum Selection

8d.1 A Single Predictor for Dichotomy

Birnbaum and Chapman (1950) considered a problem of selecting candidates on the basis of $p$ admission scores $y_1, \ldots, y_p$. The object is to select those whose performance is expected to be better in the final test. The offered solution does not refer to a case where the scores of a number of individuals $N$ have been observed but to a hypothetical set of individuals applying for the admission test. The former problem is often met with because the question asked is who out of a number of individuals whose admission scores are available should be admitted. Let the scores of $N$ individuals be represented by

$$y_{11}, y_{21}, \ldots, y_{p1}$$

$$\ldots$$

$$y_{1N}, y_{2N}, \ldots, y_{pN}$$

To answer this problem we need to know the expected performance in the final test of an individual with the admission scores $y_{1i}, \ldots, y_{pi}$. Let this expected performance be

$$x_i = \phi(y_{1i}, \ldots, y_{pi})$$

which actually stands for the regression equation of the final performance on the initial scores. The regression function, which may be of any complicated type, supplies us with the expected performances $x_1, \ldots, x_N$ of the candidates, and these latter scores form the basis for selection. The regression function can be estimated on the basis of the previous information.
For instance, if a given number of \( k \) seats are available, then the best plan is to admit \( k \) candidates corresponding to the \( k \) largest values of \( x \) because this maximizes the expected performance under the condition that \( k \) have to be chosen.

A second alternative may be to admit as many as possible with the restriction that the expected average performance is not less than an assigned number \( z_0 \). The best plan is then to order the \( x \) scores in a decreasing order and find the cumulative averages from the top and admit all those for whom the cumulative average is greater than or equal to \( z_0 \). Obviously under such a selection procedure the maximum number is admitted subject to the condition that the expected average performance of the selected candidates is not less than \( z_0 \).

If the restriction is that the average performance of the chosen candidates should exceed a given value \( z_0 \) with a probability greater than \( \beta \), then we start with the highest score of \( x \) and go on adding the others in the decreasing order of \( x \) till the required probability remains greater than \( \beta \). The calculations are not simple, however.

If we consider a hypothetical set of candidates, a situation that may arise when the statistician is asked to give a uniform rule for independent recruitment at various places without specifying the numbers to be selected from each place, then what is needed is the determination of a critical value \( z_0 \) leading to the selection of all individuals with the expected \( x \) score (calculated on the basis of the admission \( y \) score) greater than or equal to \( z_0 \). For this the distribution of the expected score \( x \) as a function of \( y \) has to be studied. Let this be \( f(x) \). If the criterion is that the maximum number of candidates has to be admitted subject to the condition that the expected average performance is not less than \( z_0 \), then \( z_0 \) is determined from the formula

\[
\int_{z_0}^{\infty} xf(x) \, dx = z_0 \int_{z_0}^{\infty} f(x) \, dx
\]

in which case the expected proportion admitted is

\[
\int_{z_0}^{\infty} f(x) \, dx
\]

8d.2 The Problem of Differential Predictors

In the problem treated in 8b.10 it was assumed that the individuals belong to separate groups characterized by distinct distributions. What was needed there was the classification of a collection of individuals into the distinct groups to which they belong. But situations arise where an individual cannot be said to belong to a distinct group, as when we
have to judge the relative usefulness of a person in two jobs, A and B. To give another instance, it may be necessary to determine on the basis of a student’s score in an admission test whether he should be allowed to take a course in mathematics or physics. What we need in such cases is a set of predictors measuring the success of a candidate in various careers on the basis of the initial scores and decide on a profitable course of action. The general problem may be stated as follows.

Out of $N$ applicants, $n_1$ have to be selected for the first job, $n_2$ for the second, \ldots, $n_k$ for the $k$th, given their scores in some suitably designed tests. If the number of applications exceed the number of jobs to be filled, then the rest of the applications $n_{k+1}$ have to be rejected. This problem arising out of a study by Brogden (1946) admits a neat solution by the use of lemmas in A5.

Let $x_1, \ldots, x_p$ be the scores for $p$ items of a test used to predict the success of a candidate in $k$ careers. For a proper treatment of the problem it is necessary that success should be measurable quantitatively, in which case the success in two different jobs could be compared. For instance it may be possible to predict (if past records are available) on the basis of initial scores or to find out by direct practical tests, if possible, how much worth of goods a person can produce in various types of jobs. If it is the admission of a student into one of various alternative courses, the success may be measured by the number of marks (properly standardized) the student is expected to secure at the end of the course on the basis of the initial score. These quantities measuring a person’s success in $k$ given situations are represented by $s_1(x)$, $s_2(x)$, \ldots, $s_k(x)$ which are necessarily functions of the initial scores $x_1, \ldots, x_p$. Let there be $N$ applicants whose success scores are given below, with a zero column representing the success score when the applicant is not selected for any job.

\[
\begin{array}{c c c c c}
\text{\textit{s}}_{11} & \text{\textit{s}}_{12} & \ldots & \text{\textit{s}}_{1k} & 0 \\
\text{\textit{s}}_{21} & \ldots & \ldots & \ldots & \ldots \\
\text{\textit{s}}_{N1} & \text{\textit{s}}_{N2} & \ldots & \text{\textit{s}}_{Nk} & 0 \\
\end{array}
\]

Let us choose $n_1$ values from the first column, $n_2$ from the second, \ldots and $n_{k+1}$ from the last, such that the sum of all these values is a maximum. The method of determining these values is given in A5. The candidates corresponding to the $n_i$ values chosen from the $i$th column are selected for the $i$th job. The candidates corresponding to $n_{k+1}$ zero values from the last column are not chosen for any job.
A GENERALIZATION OF THE NEYMAN-PEARSON LEMMA

Appendix A

A1 A Lemma of Neyman and Pearson

Let $F_0, F_1, \ldots, F_m$ be a set of integrable functions defined in the whole space of $(x_1, \ldots, x_n)$ and $w$ any region such that

$$\int_w F_i \, dv = C_i \quad i = 1, 2, \ldots, m$$

where $dv$ stands for the volume element $dx_1 \cdots dx_n$ and $C_i$ are assign constants. Let $w_0$ be a region within which

$$F_0 \geq k_1 F_1 + \cdots + k_m F_m$$

and outside which

$$F_0 \leq k_1 F_1 + \cdots + k_m F_m$$

where $k_i$ are determined such that $w_0$ satisfies (A1).

The lemma states that for any region $w$ satisfying (A1) the following relationship holds.

$$\int_w F_0 \, dv \leq \int_{w_0} F_0 \, dv$$

Let the common part of the regions $w$ and $w_0$ be denoted by $w_{w_0}$. The region $w - w_{w_0}$ is the part of $w$ not common to $w_0$. From condition (A1.1) it follows that

$$\int_{w - w_{w_0}} F_i \, dv = \int_{w_0 - w_{w_0}} F_i \, dv$$

Consider the difference

$$\Delta = \int_{w_0} F_0 \, dv - \int_{w} F_0 \, dv = \int_{w_0 - w_{w_0}} F_0 \, dv - \int_{w - w_{w_0}} F_0 \, dv$$

$$\geq \int_{w_0 - w_{w_0}} (2k_i F_i) \, dv - \int_{w - w_{w_0}} (2k_i F_i) \, dv$$

$$= 0 \quad \text{due to (A1.4)}$$

Hence the lemma is proved.

A2 A Generalization of the Neyman-Pearson Lemma

(i) Let $R_{1}', R_{2}', \ldots$ be a set of mutually exclusive regions covering the whole space such that with respect to the integrable functions $g$.
The values \( g_i \) are constant.

(ii) Consider the system of regions \( \cap = \) defined by

\[
R_k \cap F_k \leq F_s \quad s = 1, 2, \ldots \quad k = 1, 2, \ldots
\]  

(A2.2)

where

\[
F_k = \phi_k + \lambda_k g_1 + \lambda_{k2} g_2 + \cdots
\]

\( \phi_k \) being some assigned functions. We now prove that the value of the integral

\[
\int_{R_i} \phi_1 \, dv + \int_{R_i} \phi_2 \, dv + \cdots
\]

subject to the conditions in (A2.1) is a minimum for the set of regions defined in (A2.2).

The intersection of the regions \( R_i \) and \( R_i' \) is represented by \( R_{ij} \). It follows from definition that

\[
\int_{R_{ij}} F_{ij} \, dv \leq \int_{R_{ij}} F_{1} \, dv + \int_{R_{ij}} F_{2} \, dv + \cdots
\]

Writing down the above relationship for all \( R_i \) and adding, we obtain

\[
\int_{R_i} F_{1} \, dv + \int_{R_i} F_{2} \, dv + \cdots \leq \int_{R_{i'}} F_{1} \, dv + \int_{R_{i'}} F_{2} \, dv + \cdots
\]

or

\[
\int_{R_i} \phi_1 \, dv + \int_{R_i} \phi_2 \, dv + \cdots \leq \sum \lambda_{ij} g_{ij}
\]

because of the conditions in (A2.1). The above result is established.

Suppose that the sum

\[
\int_{R_i} \phi_1 \, dv + \int_{R_i} \phi_2 \, dv + \cdots
\]

has to be maximized. Then the regions are

\[
R_1 \cap F_1 \geq F_2, F_1 \geq F_3, \ldots
\]

\[
R_2 \cap F_2 \geq F_1, F_2 \geq F_3, \ldots
\]

\[ \cdots \]

\[ \cdots \]
A SLIGHT VARIATION OF LEMMA A1

Suppose that no conditions such as (A2.1) are specified. Then the regions are
\[ R_1 \cap \phi_1 \geq \phi_2, \phi_2 \geq \phi_3, \cdots \]
\[ R_2 \cap \phi_2 \geq \phi_3, \phi_3 \geq \phi_4, \cdots \]
\[ \cdots \]

A3 A Slight Variation of Lemma A1

Let \( F_0, F_1, \cdots, F_m \) be a set of integrable functions such that, with respect to a positive function \( p(x) \sim 1 \),
\[ \int_S F_i p(x) \, dv = C_i \quad i = 1, \cdots, m \quad (A3.1) \]
over the whole space \( S \).

Consider the special form of the function \( p(x) \)
\[ p(x) = 0 \quad \text{when} \quad F_0 \leq k_1 F_1 + \cdots + k_m F_m \quad (A3.2) \]
\[ = 1 \quad \text{when} \quad F_0 \geq k_1 F_1 + \cdots + k_m F_m \]
where \( k_1, \cdots, k_m \) are determined to satisfy the condition (A3.1). Out of all \( p(x) \) the integral
\[ \int_S F_0 p(x) \, dv \]
is a maximum for the special form chosen in (A3.2).

Let \( D \) be a region inside which \( F_0 \leq k_1 F_1 + \cdots + k_m F_m \) and outside which the reverse relation holds. For any general \( p(x) \)
\[ \int_S F_0 p(x) \, dv = \int_D F_0 p(x) \, dv + \int_{S-D} F_0 p(x) \, dv \]
\[ \leq \int_D F_0 p(x) \, dv + \int_{S-D} (\Sigma k_i F_i) p(x) \, dv \]
\[ = \int_D F_0 p(x) \, dv + \int_{S-D} (\Sigma k_i F_i) p(x) \, dv - \int_D (\Sigma k_i F_i) p(x) \, dv \]
\[ = \int_D F_0 p(x) \, dv + \int_D (\Sigma k_i F_i) [1 - p(x)] \, dv \]
\[ \leq \int_D F_0 p(x) \, dv + \int_D F_0 [1 - p(x)] \, dv = \int_D F_0 \, dv \]
A4 A Lemma on Power Functions

Let \( f_1, f_2, \ldots \) be a finite number of probability densities alternative to \( f_0 \) which is specified by the null hypothesis. Let \( w \) be any region satisfying the conditions

\[
\int_w f_0 \, dv = \alpha \quad \text{(A4.1)}
\]

and

\[
\frac{1}{\alpha_1} \int_w f_1 \, dv = \frac{1}{\alpha_2} \int_w f_2 \, dv = \cdots \quad \text{(A4.2)}
\]

where \( \alpha_1, \alpha_2, \ldots \) are positive assigned quantities.

Out of all regions satisfying the conditions A4.1 and A4.2, the region \( w_0 \)

Inside which \( f_0 \leq \lambda_1 f_1 + \lambda_2 f_2 + \cdots \)

Outside which \( f_0 \geq \lambda_1 f_1 + \lambda_2 f_2 + \cdots \)

where \( \lambda_1, \lambda_2, \ldots \) are determined such that the above conditions are satisfied, gives the highest common value to the quantities in (A4.2).

Proof. Let \( \beta \) and \( \beta_0 \) be the common values (A4.2) associated with the regions \( w \) and \( w_0 \), and denote by \( w w_0 \) the region common to \( w \) and \( w_0 \). Then we have

\[
(\lambda_1 \alpha_1 + \lambda_2 \alpha_2 + \cdots) \beta_0 = \int_{w w_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
\geq \int_{w - w w_0} f_0 \, dv + \int_{w w_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
= \int_{w - w w_0} f_0 \, dv + \int_{w w_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
\geq \int_{w - w w_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
+ \int_{w w_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
= \int_{w} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv
\]

\[
= (\lambda_1 \alpha_1 + \lambda_2 \alpha_2 + \cdots) \beta
\]

If \( (\lambda_1 \alpha_1 + \lambda_2 \alpha_2 + \cdots) \) is positive, then \( \beta_0 \geq \beta \). To prove that \( \lambda_1 \alpha_1 \)
TWO LEMMAS USEFUL IN CLASSIFICATORY PROBLEMS

\( \int_{v_0} (\lambda_1 f_1 + \lambda_2 f_2 + \cdots) \, dv \geq \int_{a_1} f_0 \, dv \)

i.e.,

\( (\lambda_1 a_1 + \lambda_2 a_2 + \cdots) \beta_0 \geq \alpha \)

Since \( \beta_0 \) and \( \alpha \) are positive, it follows that \( (\lambda_1 a_1 + \lambda_2 a_2 + \cdots) \) is necessarily positive. The lemma is proved.

This lemma gives us a method of determining a region with respect to which the powers of the various alternative hypotheses are in assigned ratios, and, subject to this condition, every alternative hypothesis has the maximum power.

A5 Two Lemmas Useful in Classificatory Problems

Consider an array of elements

\[
\begin{array}{cccc}
  a_{11} & a_{21} & \cdots & a_{p1} \\
  a_{12} & a_{22} & \cdots & a_{p2} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{1n} & a_{2n} & \cdots & a_{pn}
\end{array}
\]

consisting of \( p \) columns and \( n \) rows. Let \( P \) denote the product and \( S \) the sum of \( n \) elements chosen one from each row such that the total number of elements coming from the first column is equal to a specified non-zero value \( n_1 \), from the second \( n_2 \), and so on from the \( p \)th column \( n_p \). Obviously \( n \geq p \) and \( n = n_1 + \cdots + n_p \).

Lemma 1. If the elements \( a_{ij} \) are not negative and if there exist quantities \( \lambda_1, \ldots, \lambda_p \) such that each element \( a_{ik} \) of the \( n \) elements chosen from the \( i \)th column satisfies the relationships

\[
\lambda_i a_{ik} \geq \lambda_j a_{jk} \quad j = 1, \ldots, p \tag{A5.1}
\]

and if similar relationships are satisfied for all \( i = 1, \ldots, p \) with the same set \( \lambda_1, \ldots, \lambda_p \) then for this choice of \( n_1, \ldots, n_p \) elements the product \( P \) defined above is a maximum.

To prove this consider any other choice of \( n_i \) elements

\[
a_{im1}, a_{iim2}, \ldots
\]

from the \( i \)th column. If we remember that the subscript refers to the row number, the elements from the \( m_1 \)th, \( m_2 \)th, \ldots rows occurring in the proposed selection may be represented by

\[
a_{i(m_1)}, a_{i(m_2)}, \ldots
\]
from which by definition it follows that the product
\[ a_{i_1m_1}, a_{i_2m_2}, \ldots \]
is not less than
\[ \frac{\lambda_1}{\lambda_{i_1}}, \frac{\lambda_2}{\lambda_{i_2}}, \frac{\lambda_3}{\lambda_{i_3}}, \ldots \]

Since each \( \lambda \) is positive by definition, division by \( \lambda \) does not change the inequality sign. Considering all the groups of elements in the second selection we find
\[ \prod_{i} a_{i_1m_1} a_{i_2m_2} \cdots \geq \prod_{i} \frac{\lambda_1}{\lambda_{i_1}}, \frac{\lambda_2}{\lambda_{i_2}}, \frac{\lambda_3}{\lambda_{i_3}}, \ldots \]
\[ \geq \prod_{i} a_{i_1m_1} a_{i_2m_2} \cdots \]
since
\[ \prod_{i} \frac{\lambda_1}{\lambda_{i_1}}, \frac{\lambda_2}{\lambda_{i_2}}, \frac{\lambda_3}{\lambda_{i_3}}, \ldots = 1 \]
the terms in the numerator canceling with those in the denominator in the final product.

**Corollary 1.1.** If the object is to maximize the product \( P \) without the restriction on the number of elements coming from each column, then obviously the best procedure is to choose the biggest element from each row.

**Corollary 1.2.** The product \( P \) will be a minimum if the inequality relationship (A.5.1) is reversed.

**Corollary 1.3.** If \( p = 2 \), the method described in lemma 1 reduces to evaluating the ratios
\[ \frac{a_{11}}{a_{21}}, \frac{a_{12}}{a_{22}}, \ldots, \frac{a_{1n}}{a_{2n}} \]
and arranging them in descending order of magnitude and choosing the elements in the numerator from the first \( n_1 \) ratios and the elements in the denominator from the second \( n_2 \) ratios.

**Lemma 2.** If there exist quantities \( \mu_1, \ldots, \mu_p \) such that each element \( a_{ik} \) of the \( n_i \) elements chosen from the \( r \)th column satisfies the relationships
\[ a_{ik} + \mu_i \geq a_{jk} + \mu_j \quad j = 1, \ldots, p \]
and if similar relationships hold for all \( i = 1, \ldots, p \) with the same set \( \mu_1, \ldots, \mu_p \), then for this choice of \( n_1, \ldots, n_p \) elements the sum \( S \) defined above is a maximum.
TRANSFORMATION FOR MULTIVARIATE COMPUTATIONS

The result follows from lemma 1 by considering \( \exp(a_{ik}) \) and maximizing the product. The existence of \( \lambda_1, \ldots, \lambda_p \) leads to the existence of positive quantities \( \lambda_1, \ldots, \lambda_p \) used in lemma 1. The sum \( S \) is minimized when the reverse relationships hold good.

Appendix B

B1 On a Transformation Useful in Multivariate Computations

In multivariate analysis one is often confronted with the task of inverting a covariance matrix which is laborious when the number of variates exceeds four or five. This and the further use of the elements of the inverse matrix in the computation of statistical constants and test criteria can be considerably simplified by working with a set of transformed variates derivable from the original variates. The method of construction of these transformed variates and the mechanization it introduces on the computational side are given in this appendix with special reference to the statistical methods used in Chapters 8 and 9.

Let \( X_1, \ldots, X_p \) be the original variables, and \( \lambda_{ij} \) the covariance between the \( i \)th and \( j \)th variates. The transformed variables \( Y_1, Y_2, \ldots \) are defined by

\[
Y_1 = X_1 \\
Y_2 = X_2 - a_{21} Y_1 \\
Y_3 = X_3 - a_{32} Y_2 - a_{31} Y_1 \\
\vdots \\
Y_p = X_p - a_{p,p-1} Y_{p-1} - \cdots - a_{p1} Y_1
\]

The constants \( a_{ij} \) are chosen such that \( Y_i \) are independent. The actual evaluation of these coefficients is carried out in successive stages so that, if the coefficients in \( Y_1, \ldots, Y_i \) are known, any coefficient in \( Y_{i+1} \) can be calculated in a simple manner.

To find \( a_{21} \), the covariance of \( Y_1, Y_2 \) denoted by \( \text{cov}(Y_1 Y_2) \) has to be equated to zero.

\[
\text{cov}(Y_1 Y_2) = \text{cov}(X_1 X_2) - a_{21} V(Y_1) = 0
\]

\[
= \lambda_{21} - a_{21} \lambda_{11} = 0
\]

\[
a_{21} = \frac{\lambda_{21}}{\lambda_{11}}
\]

\[
V(Y_2) = \lambda_{22} - \lambda_{21} a_{21}
\]
where $V$ denotes variance. For $Y_3$, $a_{21}$ and $a_{22}$ are to be calculated in order. With the constants $b_{ij}$ as defined below introduced merely to facilitate computation, the steps may be given as follows:

$$b_{31} = \lambda_{31} \quad a_{31} = \frac{b_{31}}{V(Y_3)}$$
$$b_{32} = \lambda_{32} - a_{21}b_{31} \quad a_{32} = \frac{b_{32}}{V(Y_3)}$$
$$V(Y_3) = \lambda_{33} - b_{31}a_{31} - b_{32}a_{32}$$

To find $Y_4$, the steps are:

$$b_{41} = \lambda_{41} \quad a_{41} = \frac{b_{41}}{V(Y_4)}$$
$$b_{42} = \lambda_{42} - a_{21}b_{41} \quad a_{42} = \frac{b_{42}}{V(Y_2)}$$
$$b_{43} = \lambda_{43} - a_{21}b_{42} - a_{31}b_{41} \quad a_{43} = \frac{b_{43}}{V(Y_3)}$$
$$V(Y_4) = \lambda_{44} - b_{41}a_{41} - b_{42}a_{42} - b_{43}a_{43}$$

With $Y_1, \ldots, Y_{i-1}$ known, the steps for the evaluation of $Y_i$ are:

$$b_{il} = \lambda_{il} \quad a_{il} = \frac{b_{il}}{V(Y_i)}$$
$$b_{ij} = \lambda_{ij} - a_{21}b_{ij} \quad a_{ij} = \frac{b_{ij}}{V(Y_j)} \quad j < i - 1$$
$$V(Y_i) = \lambda_{ii} - \sum_{j=1}^{i-1} a_{ij}b_{ij}$$

The method needs checking at each stage since the constants derived at any stage depend on those previously calculated. Errors may accumulate due to rounding off in earlier calculations, but the accuracy can
be maintained by retaining a sufficient number of decimal places at each stage.

It is unnecessary to express \( Y \) as a function of \( z \) only. This would mean another set of successive operations starting with \( Y'_{1} = z_{1} \) and substituting for \( Y_{1} \) in \( Y'_{2} \), for \( Y_{1}, Y_{2} \) in \( Y'_{3} \), and so on. In any problem \( Y_{1}, \ldots, Y_{n} \) will be successively calculated, and for this the transformation derived above can be directly used. If \( Y_{i} \) has to be directly calculated from the original measurements, then the computational method given in B2 is much simpler.

**B2 An Alternative Computational Scheme**

An alternative method which directly yields the functions of \( x \) is suggested by the following theoretical considerations. Let the dispersion matrix of \( x_{1}, x_{2}, \ldots, x_{p} \) be

\[
\begin{pmatrix}
\lambda_{11} & \cdots & \lambda_{1p} \\
\vdots & \ddots & \vdots \\
\lambda_{p1} & \cdots & \lambda_{pp}
\end{pmatrix}
\]

Consider the extended matrix

\[
\begin{pmatrix}
\lambda_{11} & \cdots & \lambda_{1p} & x_{1} \\
\vdots & \ddots & \vdots & \vdots \\
\lambda_{p1} & \cdots & \lambda_{pp} & x_{p}
\end{pmatrix}
\]

Taking \( \lambda_{11} \) as the first pivotal element, replace the first row by

\[
\begin{pmatrix}
1 & \lambda_{12} & \cdots & \lambda_{1p} & x_{1} \\
\lambda_{11} & \lambda_{11} & \vdots & \vdots & \vdots
\end{pmatrix}
\]

Sweeping out the first column and using the first pivotal row, we obtain the reduced matrix

\[
\begin{pmatrix}
\lambda_{2'} & \cdots & \lambda_{2'p'} & x_{2}' \\
\vdots & \ddots & \vdots & \vdots \\
\lambda_{p'2} & \cdots & \lambda_{p'p'} & x_{p}'
\end{pmatrix}
\]

where

\[
\lambda_{ij}' = \frac{\lambda_{ij} - \lambda_{11} \lambda_{ij}}{\lambda_{11}} \quad x_{i}' = z_{i} - \frac{\lambda_{11} x_{1}}{\lambda_{11}}
\]
348 STATISTICAL INFERENCE IN CLASSIFICATORY PROBLEMS

Now

\[ V(x') = V(x) - \frac{2\lambda_{11}}{\lambda_{11}} \text{cov} (x_1x_1) + \left( \frac{\lambda_{11}}{\lambda_{11}} \right)^2 V(x_1) \]

\[ = \lambda_{11} - \lambda_{11} = \lambda_{11} \]

Similarly,

\[ \text{cov} (x'x') = \lambda_{11} \]

This shows that the reduced matrix at any stage is the dispersion matrix of the new variables on the right-hand side, provided that the first matrix is the dispersion matrix of the original variables. This property has been discussed in 3a.6 in connection with the solution of normal equations and their intrinsic properties. Also

\[ \text{cov} (x_1x') = \text{cov} (x_1x) - \frac{\lambda_{11}}{\lambda_{11}} V(x_1) \]

\[ = \lambda_{11} - \lambda_{11} = 0 \]

so that the new variables are all uncorrelated with the variable of the pivotal row. We now consider the second pivotal row

\[ \begin{pmatrix} 1 & \lambda_{21}' & \cdots & \lambda_{p1}' \\ \lambda_{22}' & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots \\ \lambda_{p2}' & \cdots & \cdots & \cdots \end{pmatrix} \]

and find the further reduced matrix

\[ \begin{pmatrix} \lambda_{31}'' & \cdots & \lambda_{3p}'' & x_3'' \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_{p3}'' & \cdots & \lambda_{pp}'' & x_p'' \end{pmatrix} \]

We thus obtain the variables

\[ x_1', x_2', x_3'', \cdots \]

with variances

\[ \lambda_{11}, \lambda_{22}', \lambda_{33}'', \cdots \]

They are all mutually uncorrelated as shown above, and further \( x_2' \) depends on \( x_1 \) and \( x_2 \), and \( x_3'' \) on \( x_1, x_2, \) and \( x_3 \), only, and so on. Thus the transformation is of the type considered in B1. We thus obtain a relatively simple scheme for obtaining uncorrelated linear functions of the original variables, provided that the number and variables to be included are fixed in advance. In the earlier method the transformed variables are calculated one after the other so that we are free to choose
the variable to be added at any stage and in any order we like. There are a few problems where the decision to add a new character depends on tests to be made with the help of the transformed variates up to that stage. In situations like this only the earlier method is open to us. It is enough to compute the transformation \((B1.1)\) in such a case since successive values of \(Y_1, Y_2, \ldots\) will be obtained. There is no need to express \(Y\) as functions of \(x\) only. In problems where a transformation of a chosen set of correlated variables is required, the alternative method of \(B2\) is better.

Having obtained \(Y_1, Y_2, \ldots, Y_5\) (say) directly as functions of the original variables, if we want to extend the transformation to a sixth variable \(z\) we write

\[
Y_6 = x_6 - a_{61}Y_1 - a_{62}Y_2 - a_{63}Y_3 - a_{64}Y_4 - a_{65}Y_5
\]

as in the earlier method. The coefficients are determined from the equations

\[
\text{cov} (x_6, Y_i) = a_{6i}V(Y_i)
\]

Since \(Y_i\) is a known function of \(x\), it is easy to calculate \(\text{cov} (x_6, Y_i)\) and \(V(Y_i)\) is already available.

References


CHAPTER 9

The Concept of Distance
and the Problem of Group Constellations

9a Distance between Two Populations

9a.1 The Need for a Distance Function

One important object of obtaining biometric measurements is to study the possibilities of classifying different groups of individuals in the form of a significant pattern. Here we are concerned not with the individual variations within a group which played a prominent part in the investigations of Chapter 8, but with the group characteristics or the statistical constants related to the distributions of measurements. The configuration of several groups or, to be more precise, of the group characteristics may admit a description in terms of a few group constellations and their interrelationships. The groups within a constellation must necessarily be closer, in some sense, to one another than those belonging to different constellations. Such a description, based only on measurements, quantitative or qualitative in character, may be of use in the study of evolution of the various groups.

A word of caution is necessary. Although it is possible to refute any statement concerning the relationships of some groups, it cannot be asserted that any closeness as indicated by a study of measurements alone is due to some common stock from which the groups have evolved. Historical and ethnological evidence and also geographical contiguity of localities inhabited by various groups have to be considered in interpreting the observed differences.

The first step in the problem of group constellations is the construction of an index by which we can measure the resemblance between two groups. With such an index it is possible to speak of a generalized distance between two groups and to compare the distances between any two pairs of groups. We may, then, be able to say that groups $G_1$ and $G_2$ resemble each other more than $G_2$ and $G_3$ or $G_3$ and $G_4$, and so on.
If groups $G_1$ and $G_2$ are close together and $G_3$ is distant from both, we can talk of $G_1$ and $G_2$ as forming a cluster. It may be that all the distances between $G_1$, $G_2$, and $G_3$ are small but the distances of these from the others are large. Then $G_1$, $G_2$, and $G_3$ can be considered to be a closely associated cluster of groups. By sorting out such clusters it may be possible to arrange the various groups in some simplified pattern.

9a.2 Mathematical Concepts (Discriminatory Topology)

In 8b statistical criteria were developed for specifying an individual as a member of one of two groups to which he can possibly belong. These criteria depend on the evaluation of the likelihood ratio and on the assignment of all individuals providing a ratio higher than a predetermined value $\lambda$ to one group, and the rest to the other. Errors are inevitable in such a procedure, and for any given $\lambda$ the chances of incorrect classification of individuals of the first and second groups are calculable. For a suitable choice of $\lambda$ these errors can be made equal, thus supplying the least possible proportion $\alpha$ of individuals who are liable to be misclassified. In other words the two groups can be said to overlap to the extent of 100\% per cent. The overlap is a maximum when the two groups are identical, in which case the method of classification reduces to a toss of the coin. The overlap decreases with an increase in the divergence between the two groups. If the two groups are distinct in the sense that the ranges of measurements are non-overlapping in the two cases, then the individuals have a distinct identity so that no error is committed. The percentage of overlap is thus zero. The extent of separation or divergence between two groups can thus be judged by $\alpha$, the least proportion of overlapping individuals who are liable to be misclassified. If a measure of divergence is necessary to express the amount of separation, one might choose a decreasing function of $\phi$ so that the zero value of $\phi$ may correspond to the maximum distance. One such function is $(1 - \phi)$.

This satisfies the two fundamental postulates* of distance in topological spaces:

(i) The distance between two groups is not less than zero.

(ii) The sum of distances of a group from two other groups is not less than the distance between the two other groups (triangle law of distance).

The first postulate follows since $\alpha \leq 1$, in which case the distance $1 - \alpha \geq 0$. To prove the second, we consider three groups, $G_1$, $G_2$, $G_3$. Let $R_1(1, 2), R_2(1, 2)$ be the best divisions of the space corresponding to $G_1$ and $G_2$. Similar definitions hold for $R_1(1, 3), R_3(1, 3), R_2(2, 3)$.

*The postulate that the distance of a group with itself is zero does not hold and is not relevant for our purpose.
Defining
\[ \int_{R_i(2,3)} f_i \, dv = 1 - \alpha_{ij} \]
the proposition required to be proved may be stated as follows.

\[ (1 - \alpha_{12}) \gg (1 - \alpha_{13}) + (1 - \alpha_{23}) \]

From definition it follows that
\[ \int_{R_i(2,3)} f_i \, dv \lesssim \int_{R_i(2,3)} f_i \, dv \]
so that
\[ \int_{R_i(2,3)} f_i \, dv + \int_{R_i(2,3)} f_j \, dv \lesssim 1 \]
Hence
\[ \int_{R_i(2,3)} f_i \, dv \gg 1 > \int_{R_i(2,3)} f_i \, dv + \int_{R_i(2,3)} f_j \, dv \]
and
\[ \int_{R_i(2,3)} f_2 \, dv \gg 1 > \int_{R_i(2,3)} f_2 \, dv + \int_{R_i(2,3)} f_3 \, dv \]
Adding, \( 2(1 - \alpha_{12}) \gg 2(1 - \alpha_{12}) + 2(1 - \alpha_{23}) \), which proves the desired result. The distance function defined above must satisfy some further empirical requirements if it is to be of any value in biological classifications.

(i) The distance must not decrease when additional characters are considered.
(ii) The increase in distance by the addition of some characters to a suitably chosen set must be relatively small so that the group constellations arrived at on the basis of the chosen set are not distorted when additional characters are considered.

The first requirement is reasonable since adding some characters to a basic set must necessarily reduce the errors of classification. In fact, this requirement is satisfied when the distance function is as chosen above. Let \( P_1(x_1, \cdots, x_p) \) and \( P_2(x_1, \cdots, x_p) \) denote the probability densities of two groups with \( R_1(p) \) and \( R_2(p) \) as the best divisions of the \( p \)-space. When an additional character is considered, the probability densities can be written \( P_1(x_1, \cdots, x_{p+1}) \) and \( P_2(x_1, \cdots, x_{p+1}) \) with \( R_1(p+1) \) and \( R_2(p+1) \) as the best division of the \( (p+1) \)-space. If \( \Omega \) denotes the region obtained by considering \( R_1(p) \) and the complete
range for $x_{p+1}$, then by definition it follows that

$$\int_{R_1(p+1)} P_2(x_1, \ldots, x_{p+1}) \, dv' \leq \int_0^p P_2(x_1, \ldots, x_{p+1}) \, dv'$$

$$= \int_{R_1(p)} P_2(x_1, \ldots, x_p) \, dv$$

so that if $\alpha$ and $\alpha'$ represent the proportions of overlapping individuals in the two cases it follows that $\alpha' \leq \alpha$, which proves the result.

The second requirement has been introduced merely as a practical necessity. Owing to considerations of cost in obtaining the information and the numerical reduction of data, there must be some limit to the number of characters used in order to arrive at stable judgments. This should be empirically verified in any situation.

9a.3 Mahalanobis' Generalized Distance

Consider two multivariate normal populations with a common dispersion matrix $(\lambda_{ij})$ and having mean values $\mu_{11}, \ldots, \mu_{1p}$ and $\mu_{21}, \ldots, \mu_{2p}$ so that the probability densities $f_1(x)$ and $f_2(x)$ are

$$f_1 = \text{const. exp} \left\{ -\frac{1}{2} \sum_1^p (x_i - \mu_{1i})(x_j - \mu_{1j}) \right\}$$

$$f_2 = \text{const. exp} \left\{ -\frac{1}{2} \sum_1^p (x_i - \mu_{2i})(x_j - \mu_{2j}) \right\}$$

where $(\lambda_{ii})$ is reciprocal to $(\lambda_{ij})$. The surfaces of constant likelihood ratios are defined by

$$L(x) = \sum_i (\sum_k \lambda_{ki} x_k) x_i = \text{const.}$$

where $d_i = \mu_{1i} - \mu_{2i}$, $(i = 1, \ldots, p)$. The common value of the least proportion of wrong classifications is

$$\int_{L(x) = \alpha} f_2 \, dv = \int_{L(x) = \alpha} f_1 \, dv$$

This can be determined by considering only the distribution of $L(x)$, which, being a linear function of $x_i$, is normal. The mean $L(x)$ in the first group is

$$L(\mu_1) = \sum (\sum \lambda_{ki} d_k) \mu_{1i}$$
and in the second group

\[ L(\mu_2) = \sum_i (\Sigma x_i d_i) \mu_i \]

\[ V[L(x)] = D^2 = \Sigma \Sigma x_i d_i = L(\mu_1) - L(\mu_2) \]

\[ a = \int_{L(x) \leq s} e^{-\frac{1}{2}(L(x) - L(\mu_1))^2/2 + L(\mu_1)} \frac{dL(x)}{D\sqrt{2\pi}} \]

By symmetry it follows that

\[ a = \frac{L(\mu_1) + L(\mu_2)}{2} \]

On making the transformation \( y = [L(x) - L(\mu_1)]/D \), the above integral reduces to

\[ a = \frac{1}{\sqrt{2\pi}} \int_{D}^{D} e^{-\frac{1}{2}y} dy \]

and hence

\[ 1 - a = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{D/2} e^{-\frac{1}{2}y} dy \]

which shows that \( 1 - a \) is an increasing function of \( D \). The measure \( 1 - a \) may then be conveniently replaced by the functionally dependent quantity \( D^* \) which was first introduced by P. C. Mahalanobis. The above analysis supplies a logical derivation of this tool suggested on intuitive considerations. It also shows that Mahalanobis' generalized distance function is applicable only to groups in which the measurements are normally distributed.

9.4 Karl Pearson's Coefficient of Racial Likeness

In 1921, in a paper by Miss M. L. Tidesley, Karl Pearson proposed a measure of racial likeness (C.R.L.) which has been used since then by anthropologists of the biometric school for purposes of classifying skeletal remains (Morant, 1923). If \( n_1 \) and \( n_2 \) denote the number of observations on which the means \( m_{1i} \) and \( m_{2i} \) of the ith character for the first and second groups are based, \( s_i \) the standard deviation of the ith

* \( D \) is the ordinary Euclidean distance in a space defined by a set of oblique axes.

Therefore it also satisfies the triangular law of distance.
character, and \( p \) is the number of characters used, then C.R.L. is defined by

\[
C.R.L. = \frac{1}{p} \sum_{i=1}^{p} \frac{n_{1i}n_{2i}}{n_{1i} + n_{2i}} \left( \frac{m_{1i} - m_{2i}}{s_i} \right)^2
\]

This is meant to be an estimate of a measure of distance between two populations. Since this estimate depends to a large extent on the sample sizes a reduction factor is employed for comparing the C.R.L. values arising out of two pairs. The reduced coefficient of racial likeness (R.C.R.L.) is

\[
R.C.R.L. = \frac{n_1 + n_2}{n_1n_2} \frac{1}{p} \sum_{i=1}^{p} \frac{n_{1i}n_{2i}}{n_{1i} + n_{2i}} \left( \frac{m_{1i} - m_{2i}}{s_i} \right)^2 - 1
\]

where

\[
\frac{n_1}{p} = \frac{\sum n_{1i}}{p}, \quad \frac{n_2}{p} = \frac{\sum n_{2i}}{p},
\]

It may be seen that

\[
E(R.C.R.L.) = \frac{1}{n_1n_2} \frac{n_1 + n_2}{p} \sum_{i=1}^{p} \frac{n_{1i}n_{2i}}{n_{1i} + n_{2i}} \left( \frac{\mu_{1i} - \mu_{2i}}{\sigma_i} \right)^2
\]

where \( \mu_{1i} \) and \( \mu_{2i} \) are the population mean values and it is assumed that \( s_i \) is not subject to variations. This is not independent of the sample sizes unless \( n_{1i} = n_1 \) and \( n_{2i} = n_2 \) for all \( i \).

With skeletal material there is a good deal of variation in the number of observations available for various characters, which makes the above coefficient ill-suited to such applications. In the expected value, which is supposed to measure the differences in the group characteristics, the various characters are not weighted according to any criteria; the weights are determined by the number of observations available. This, at the same time, makes the coefficient uncomparable unless by some happy coincidence the relative numbers of observations are such that in two R.C.R.L. values to be compared the weights happen to be the same. Also, the R.C.R.L. values calculated from the same two groups in two different situations with different sample observations for the various characters may be widely different. The effect of sample size in the estimate of \( D^2 \) is not very serious and can be easily corrected if necessary, as shown later, by subtracting some value which depends solely on the sample sizes and whose value is negligible (tending to zero) in large samples. Comparability is retained because the weights attached to the various characters are not functions of sample sizes.
CALCULATION OF $D^2$  

C.R.L. differs from $D^2$ in another important aspect. In C.R.L. all the characters are treated as independent. All biological experience points to the fact that some amount of correlation exists between any two characters, and the effect of this is to make C.R.L. increase rapidly with an increase in the number of characters. It is difficult to say from the changes in C.R.L. whether some newly added characters supply additional information for purposes of discrimination. It was seen in the examples of Chapter 8 that, if differences in some characters have already been considered, the absolute differences in some other characters do not throw any light on discrimination between the groups. There is no provision in the construction of C.R.L. to distinguish which characters are more useful for discrimination. For instance, when a character highly correlated with a set of characters is used in addition to those in the set, the C.R.L. may be considerably altered although the change may be inappreciable when the high correlation is taken into account. The change in $D^2$ may not be appreciable if such superfluous characters are used. If any character alters $D^2$ considerably, it may be taken to be of additional value in discrimination. In view of the fact that the $D^2$ statistic satisfies some logical requirements, it may be used in preference to the C.R.L.

9b An Illustrative Example

9b.1 Calculation of $D^2$

The mean values of 9 characters for 12 castes and tribes of the United Provinces (Mahalanobis, Majumdar, and Rao, 1949) are given in Table 9b.1a, and the pooled intragroup correlations and standard deviations in Table 9b.1b. The first step in the analysis is the evaluation of all possible $D^2$ values. The formula

$$D^2 = \sum \sum \sum \sum d_i d_j$$

for the computation of $D^2$ is not useful since it requires the inversion of a ninth-order determinant and then the evaluation of $9(9 + 1)/2$ terms whose sum is $D^2$. In an earlier example, the classification of the Highdown skull, it was found convenient to work with a set of uncorrelated characters constructed from the original measurements. The $D^2$ with such transformed variables reduces to the evaluation of a simple sum of squares. Two simple methods of obtaining this transformation are given in Appendix B of Chapter 8, and one method is actually illustrated in the problem of the Highdown skull. The other method is followed in the present example. If the original characters expressed in standard deviation units are represented by lower-case letters, then
## Table 9b.1a. Mean Values by Groups and Characters

<table>
<thead>
<tr>
<th>Group</th>
<th>Head Length</th>
<th>Head Breadth</th>
<th>Browridge Breadth</th>
<th>Nasal Height</th>
<th>Nasal Breadth</th>
<th>Nasal Depth</th>
<th>Stature</th>
<th>Siting Height</th>
<th>Frontal Breadth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>n</td>
<td>Mean</td>
<td>n</td>
<td>Mean</td>
<td>n</td>
<td>Mean</td>
<td>n</td>
</tr>
<tr>
<td>Brahmin (Beri, B)</td>
<td>86</td>
<td>191.92</td>
<td>80</td>
<td>139.58</td>
<td>88</td>
<td>133.58</td>
<td>80</td>
<td>81.34</td>
<td>85</td>
</tr>
<tr>
<td>Brahmin (Other, B)</td>
<td>92</td>
<td>191.35</td>
<td>92</td>
<td>139.50</td>
<td>92</td>
<td>132.28</td>
<td>91</td>
<td>50.40</td>
<td>90</td>
</tr>
<tr>
<td>Chatter (Ch)</td>
<td>139</td>
<td>192.58</td>
<td>139</td>
<td>131.72</td>
<td>139</td>
<td>131.70</td>
<td>137</td>
<td>52.72</td>
<td>137</td>
</tr>
<tr>
<td>Muslin (M)</td>
<td>167</td>
<td>200.78</td>
<td>168</td>
<td>137.40</td>
<td>168</td>
<td>131.22</td>
<td>168</td>
<td>81.38</td>
<td>168</td>
</tr>
<tr>
<td>Buxu (B)</td>
<td>148</td>
<td>196.10</td>
<td>150</td>
<td>136.18</td>
<td>150</td>
<td>132.55</td>
<td>150</td>
<td>52.06</td>
<td>150</td>
</tr>
<tr>
<td>Habra (H)</td>
<td>124</td>
<td>186.94</td>
<td>124</td>
<td>137.40</td>
<td>124</td>
<td>131.18</td>
<td>124</td>
<td>50.35</td>
<td>124</td>
</tr>
<tr>
<td>Bhil (Bh)</td>
<td>187</td>
<td>181.87</td>
<td>180</td>
<td>137.42</td>
<td>180</td>
<td>131.18</td>
<td>187</td>
<td>65.40</td>
<td>187</td>
</tr>
<tr>
<td>Dom (D)</td>
<td>113</td>
<td>186.40</td>
<td>113</td>
<td>137.12</td>
<td>113</td>
<td>132.64</td>
<td>113</td>
<td>50.34</td>
<td>113</td>
</tr>
<tr>
<td>Akir (A)</td>
<td>68</td>
<td>187.45</td>
<td>67</td>
<td>138.12</td>
<td>68</td>
<td>131.70</td>
<td>68</td>
<td>48.98</td>
<td>68</td>
</tr>
<tr>
<td>Kurnel (A)</td>
<td>94</td>
<td>188.86</td>
<td>94</td>
<td>137.80</td>
<td>94</td>
<td>131.82</td>
<td>94</td>
<td>40.22</td>
<td>94</td>
</tr>
<tr>
<td>Other Arians (Aa)</td>
<td>173</td>
<td>187.09</td>
<td>173</td>
<td>138.84</td>
<td>173</td>
<td>131.20</td>
<td>173</td>
<td>48.72</td>
<td>173</td>
</tr>
<tr>
<td>Kedar (Aa)</td>
<td>57</td>
<td>188.83</td>
<td>57</td>
<td>136.29</td>
<td>57</td>
<td>130.70</td>
<td>57</td>
<td>48.42</td>
<td>57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>Mean</th>
<th>n</th>
<th>Mean</th>
<th>n</th>
<th>Mean</th>
<th>n</th>
<th>Mean</th>
<th>n</th>
<th>Mean</th>
</tr>
</thead>
</table>
Table 9b.15. Pooled Estimates of Correlations and Standard Deviations

<table>
<thead>
<tr>
<th></th>
<th>HL</th>
<th>HB</th>
<th>B.B</th>
<th>NH</th>
<th>NB</th>
<th>ND</th>
<th>St</th>
<th>SH</th>
<th>FB</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL</td>
<td>0.1982</td>
<td>0.2792</td>
<td>0.1758</td>
<td>0.1930</td>
<td>0.1537</td>
<td>0.2969</td>
<td>0.2551</td>
<td>0.2270</td>
<td></td>
</tr>
<tr>
<td>HB</td>
<td>0.5407</td>
<td>0.1735</td>
<td>0.1413</td>
<td>0.1308</td>
<td>0.1927</td>
<td>0.2069</td>
<td>0.4461</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.B</td>
<td>0.1852</td>
<td>0.2720</td>
<td>0.1575</td>
<td>0.2891</td>
<td>0.2929</td>
<td>0.4030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH</td>
<td>0.0438</td>
<td>0.2910</td>
<td>0.1974</td>
<td>0.2170</td>
<td>0.1139</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>0.1139</td>
<td>0.1412</td>
<td>0.1182</td>
<td>0.1831</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>0.1774</td>
<td>0.2094</td>
<td>0.1243</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St</td>
<td>0.9440</td>
<td>0.2173</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard deviation 6.60 4.50 4.58 3.50 2.57 1.75 5.74 3.20 3.92

the transformed variables \( y_1, y_2, \ldots, y_9 \), which are all uncorrelated and which have standard deviation unity, are as given below.

\[
y_1 = Y_1 = hl
\]

\[
0.980162y_2 = Y_2 = hb - 0.198200Y_1
\]

\[
0.827092y_3 = Y_3 = b_b - 0.505209Y_2 - 0.279200Y_1
\]

\[
0.970863y_4 = Y_4 = nh - 0.097610Y_3 - 0.144326Y_2 - 0.175800Y_1
\]

\[
0.953943y_5 = Y_5 = nb + 0.023230Y_4 - 0.246667Y_3 - 0.107261Y_2 - 0.193000Y_1
\]

\[
0.944814y_6 = Y_6 = nd - 0.096434Y_5 - 0.258066Y_4 - 0.094404Y_3 - 0.104439Y_2 - 0.153700Y_1
\]

\[
0.924222y_7 = Y_7 = st - 0.084080Y_6 - 0.045612Y_5 - 0.122960Y_4 - 0.211917Y_3 - 0.144918Y_2 - 0.269000Y_1
\]

\[
0.788187y_8 = Y_8 = sh - 0.507938Y_7 - 0.115274Y_6 - 0.018904Y_5 - 0.141858Y_4 - 0.217927Y_3 - 0.160069Y_2 - 0.265100Y_1
\]

\[
0.826753y_9 = Y_9 = fb - 0.102016Y_8 - 0.054612Y_7 - 0.027845Y_6 - 0.044100Y_5 + 0.006243Y_4 - 0.333334Y_3 - 0.417533Y_2 - 0.227000Y_1
\]
The normalized mean values of the characters are given in Table 9b.17, and the mean values of the transformed characters \( y_1, \ldots, y_9 \) in Table 9b.18. First the values of \( Y_1, Y_2, \ldots \) are obtained by substituting the values from Table 9b.17, and then by division by the corresponding standard deviations \( y_1, \ldots, y_9 \) are derived. The \( D^2 \) corresponding to any two groups is the sum of squares of the differences in the values of \( y_1, \ldots, y_9 \). The values of \( D^2 \) are given in Table 9b.18 where corresponding to each group the other groups are arranged in increasing order of \( D^2 \).
### Table 9b.1e. Values of $D^2$ (Based on 9 Characters) Arranged in Increasing Order of Magnitude

<table>
<thead>
<tr>
<th>Brahmin (Bastl, B1)</th>
<th>Brahmin (Other, B2)</th>
<th>Bhantu (C1)</th>
<th>Habru (C2)</th>
<th>Dom (D)</th>
<th>Bil (Bh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2 0.27</td>
<td>B1 0.27</td>
<td>C2 1.32</td>
<td>A1 1.26</td>
<td>Bh 1.15</td>
<td>D 1.15</td>
</tr>
<tr>
<td>A1 1.17</td>
<td>A1 0.78</td>
<td>A1 2.68</td>
<td>C1 1.32</td>
<td>C2 2.11</td>
<td>A2 1.75</td>
</tr>
<tr>
<td>A2 1.48</td>
<td>A2 1.03</td>
<td>A2 2.98</td>
<td>A1 1.53</td>
<td>A2 2.51</td>
<td>A3 2.23</td>
</tr>
<tr>
<td>A3 2.13</td>
<td>A3 1.47</td>
<td>A3 3.35</td>
<td>B2 1.63</td>
<td>A3 2.41</td>
<td>A4 2.24</td>
</tr>
<tr>
<td>B2 2.23</td>
<td>C2 1.63</td>
<td>B1 3.48</td>
<td>A1 1.67</td>
<td>M 2.47</td>
<td>C2 2.43</td>
</tr>
<tr>
<td>M 2.86</td>
<td>M 2.62</td>
<td>B3 3.61</td>
<td>D 2.11</td>
<td>A4 2.66</td>
<td>A5 2.53</td>
</tr>
<tr>
<td>D 2.86</td>
<td>A4 2.72</td>
<td>A4 4.20</td>
<td>B1 2.23</td>
<td>B2 2.81</td>
<td>M 2.16</td>
</tr>
<tr>
<td>Ch 3.06</td>
<td>D 2.81</td>
<td>M 4.46</td>
<td>E 2.43</td>
<td>B1 2.86</td>
<td>B2 3.82</td>
</tr>
<tr>
<td>A4 3.30</td>
<td>Ch 2.87</td>
<td>D 4.52</td>
<td>A4 2.87</td>
<td>A5 2.91</td>
<td>B2 4.45</td>
</tr>
<tr>
<td>C1 3.48</td>
<td>C1 3.61</td>
<td>Bh 5.08</td>
<td>M 3.74</td>
<td>Ch 3.84</td>
<td>Ch 5.02</td>
</tr>
<tr>
<td>Bh 4.45</td>
<td>Bh 3.82</td>
<td>Ch 5.25</td>
<td>Ch 4.68</td>
<td>C1 4.32</td>
<td>C1 5.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chattri (Ch)</th>
<th>Muslim (M)</th>
<th>Ahir (A1)</th>
<th>Karmi (A2)</th>
<th>Other Artisan (A3)</th>
<th>Kahar (A4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 0.40</td>
<td>Ch 0.49</td>
<td>A2 0.30</td>
<td>A3 0.12</td>
<td>A4 0.12</td>
<td>A5 0.43</td>
</tr>
<tr>
<td>A2 2.12</td>
<td>A4 0.90</td>
<td>A3 0.49</td>
<td>A1 0.30</td>
<td>A4 0.43</td>
<td>A5 0.90</td>
</tr>
<tr>
<td>A4 2.24</td>
<td>A2 1.34</td>
<td>B2 0.78</td>
<td>A4 0.58</td>
<td>A1 0.49</td>
<td>M 0.90</td>
</tr>
<tr>
<td>A3 2.72</td>
<td>A2 1.45</td>
<td>B1 1.17</td>
<td>B2 1.03</td>
<td>M 1.45</td>
<td>A1 1.32</td>
</tr>
<tr>
<td>B2 2.87</td>
<td>A2 2.45</td>
<td>C2 1.26</td>
<td>M 1.34</td>
<td>B2 1.47</td>
<td>Bh 2.24</td>
</tr>
<tr>
<td>B1 5.06</td>
<td>D 2.47</td>
<td>A1 1.52</td>
<td>B2 1.82</td>
<td>C2 1.67</td>
<td>Ch 2.24</td>
</tr>
<tr>
<td>A1 3.88</td>
<td>B2 2.62</td>
<td>M 2.45</td>
<td>C2 1.53</td>
<td>Bh 1.75</td>
<td>D 2.66</td>
</tr>
<tr>
<td>D 3.84</td>
<td>B1 2.86</td>
<td>Bh 2.53</td>
<td>Ch 2.12</td>
<td>B1 2.13</td>
<td>B2 2.72</td>
</tr>
<tr>
<td>C2 4.00</td>
<td>Bh 3.16</td>
<td>C2 2.08</td>
<td>Bh 2.53</td>
<td>D 2.31</td>
<td>C2 2.87</td>
</tr>
<tr>
<td>Bh 5.02</td>
<td>C2 3.74</td>
<td>D 2.91</td>
<td>D 2.41</td>
<td>Ch 2.72</td>
<td>B2 3.30</td>
</tr>
<tr>
<td>C1 6.25</td>
<td>C1 4.46</td>
<td>Ch 3.38</td>
<td>C1 2.98</td>
<td>C1 3.35</td>
<td>C1 4.20</td>
</tr>
</tbody>
</table>

### 9b.2 The Determination of Group Constellations

A scrutiny of Table 9b.1e reveals that there is a pattern exhibited by the 12 groups. The Brahmins B1 and B2 cluster together in almost each column and may be treated as a single unit, and also the four Artisans A1, A2, A3, and A4, which appear to be linearly arranged in the order A2, A3, A4, with A4 being most distant from A1. Bhatu (C1) and Habru (C2) go together, with C2 being nearer to the Brahmin and
Artisan clusters and C1 far removed from them. Bhil (Bh) and Dom (D) are closer, and so also are Muslim (M) and Chattri (Bh).

![Diagram of clusters and their interrelationships](image)

The average $D^2$ within and between clusters found above is given in Table 9b.2a. The configuration of the clusters and their mutual relationships is approximately indicated in Figure 1 where the square root of average $D^2$ (of Table 9b.2a) represents the distance.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Groups</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Ch</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1, A2, A3, A4</td>
<td>0.57</td>
<td>1.76</td>
<td>2.07</td>
<td>2.07</td>
<td>2.38</td>
</tr>
<tr>
<td>B</td>
<td>Bh, B2</td>
<td>1.76</td>
<td>0.27</td>
<td>2.74</td>
<td>2.85</td>
<td>3.48</td>
</tr>
<tr>
<td>C</td>
<td>C1, C2</td>
<td>2.57</td>
<td>2.74</td>
<td>1.32</td>
<td>4.53</td>
<td>3.58</td>
</tr>
<tr>
<td>Ch</td>
<td>Ch, M</td>
<td>2.07</td>
<td>2.85</td>
<td>4.53</td>
<td>0.49</td>
<td>3.62</td>
</tr>
<tr>
<td>D</td>
<td>D, Bh</td>
<td>2.38</td>
<td>3.48</td>
<td>3.58</td>
<td>3.62</td>
<td>1.15</td>
</tr>
</tbody>
</table>

No formal rules can be laid down for finding the clusters because a cluster is not a well-defined term. The only criterion appears to be that any two groups belonging to the same cluster should at least on
the average show a smaller $D^2$ than those belonging to two different clusters. A simple device suggested by K. D. Tocher is to start with two closely associated groups and find a third group which has the smallest average $D^2$ from the first two. Similarly, the fourth is chosen to have the smallest average $D^2$ from the first three, and so on. If at any stage the average $D^2$ of a group from those already listed appears to be high, then this group does not fit in with the former groups and is therefore taken to be outside the former cluster. The groups of the first cluster are then omitted, and the rest are treated similarly.

It is also useful to calculate the change in average $D^2$ within a cluster due to the inclusion of an additional group. If the change is appreciable, then the newly added group has to be considered as outside the cluster. The calculations arising from the $D^2$ table (9b.1c) are given in Table 9b.2.

### Table 9b.2. Computational Scheme for Finding Clusters

<table>
<thead>
<tr>
<th>Group Added to a Cluster</th>
<th>$D^2$</th>
<th>No. of Terms Terms</th>
<th>Increase in $D^2$ per Cluster</th>
<th>Average $D^2$ ($\sum D^2/n$)</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0, A_1$</td>
<td>0.12</td>
<td>1</td>
<td>...</td>
<td>0.12</td>
<td>$A_0, A_1$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0.91</td>
<td>3</td>
<td>0.30</td>
<td>0.20</td>
<td>$A_2$</td>
</tr>
<tr>
<td>$A_3$</td>
<td>3.44</td>
<td>6</td>
<td>0.84</td>
<td>0.57</td>
<td>$A_3$</td>
</tr>
<tr>
<td>$B_1$</td>
<td>9.44</td>
<td>10</td>
<td>1.50</td>
<td>0.94</td>
<td>$B_1$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>0.27</td>
<td>1</td>
<td>...</td>
<td>0.27</td>
<td>$B_2$</td>
</tr>
<tr>
<td>$C_1, C_2$</td>
<td>4.13</td>
<td>3</td>
<td>1.93</td>
<td>1.38</td>
<td>$C_1, C_2$</td>
</tr>
<tr>
<td>$M, Ch$</td>
<td>6.71</td>
<td>3</td>
<td>3.15</td>
<td>2.24</td>
<td>$M, Ch$</td>
</tr>
<tr>
<td>$B_0, D$</td>
<td>1.15</td>
<td>1</td>
<td>...</td>
<td>1.15</td>
<td>$B_0, D$</td>
</tr>
<tr>
<td>$C_1, C_2$</td>
<td>5.69</td>
<td>3</td>
<td>2.27</td>
<td>1.90</td>
<td>$C_1, C_2$</td>
</tr>
<tr>
<td>$C_0, C_1$</td>
<td>1.32</td>
<td>1</td>
<td>...</td>
<td>1.32</td>
<td>$C_0, C_1$</td>
</tr>
</tbody>
</table>

There is a sharp increase in the average increase in $D^2$ when $M$ is added to $A_0, A_1, A_2, A_3$. This indicates that $A_1, A_2, A_3, A_4$ form a cluster; therefore these are omitted and the rest are considered. Thus the other clusters are determined.
For comparison of $D^2$ values it must be ascertained that they estimate the corresponding population distances. This is one reason why C.R.L. was found to be not useful. The bias in $D^2$ when the dispersion is estimated on a large number of degrees of freedom is given by

$$\sum \sum \lambda_i n_{ij1} + n_{ij2}$$

where $n_{ij}$ is the number of samples supplying observations on both the $i$th and the $j$th character in the first group, and similarly $n_{ij2}$ for the second group. The value $n_{ii} = n_{ij}$, the number of observations on the $i$th character. This quantity, which depends only on the sample sizes and variances and covariances, can be calculated and subtracted from $D^2$. If $n_{ii} = n_1$ for all $i$ and $n_{jj} = n_2$ for all $j$, then the bias is simply

$$\frac{n_1 + n_2}{p}$$

where $p$ is the number of characters. If $n_1$ and $n_2$ are individually large, the correction is trivial and need not be carried out. Also, if the sample numbers are such that the quantities

$$\frac{n_1 + n_2}{n_1 n_2}$$

are of the same order for all pairs of groups under consideration, then no correction is needed because the $D^2$ values become comparable. In the illustration chosen above, the value of $p(n_1 + n_2)/n_1 n_2$ is very nearly the same and also small for all pairs of groups, and hence no correction was carried out.

9c The Use of Canonical Variates in Deriving Group Constellations

9c.1 Graphical Methods of Representing the Groups

The method of finding the group constellations, as described in 9b, becomes much simpler if the groups are characterized by two or three measurements. In the case of two characters $x_1$ and $x_2$, we can represent the mean values, expressed in standard deviation units, of the $k$ populations under consideration on a two-dimensional chart with axes inclined at an angle $\cos^{-1} r$, where $r$ is the correlation between $x_1$ and $x_2$. In such a chart, the distance between two points is equivalent, apart from a constant multiplier, to Mahalanobis' $D$ between the two populations represented by the two points. This is valuable as it facilitates
THE PROBLEM OF MAXIMAL AVERAGE $D^2$

the study of group constellations and also serves as a pictorial representation of the configuration of various groups.

For three characters we can construct a three-dimensional model representing the characters along three mutually inclined axes. In order that the distance between two points might be equal to the $D$, apart from a constant multiplier, between the two populations represented by them, the angle between the axes corresponding to $x_1$ and $x_2$ should be chosen as $\cos^{-1} r_{12,3}$ and the scale of $x_1$, $x_2$, $x_3$ be properly adjusted.

If the object of representation is only to measure $D$ by the actual distance in space, we can transform the characters to independent variables, in which case they can be represented along three mutually orthogonal axes. This method of representation fails when we are dealing with more than three characters. In such cases it might be useful to examine whether the configuration of the mean values with respect to $p > 3$ characters can be preserved, as far as possible, by representing the groups with respect to two or three suitably chosen functions of the $p$ characters. A convenient measure for examining the adequacy of such a simpler representation is given by the ratio of the sum of squares of all possible $k(k-1)/2$ distances arising out of the $k$ populations in the simpler representation to the sum of squares of the $p$-dimensional representation. The former sum of squares is not greater than the latter, and the two representations are identical when the ratio is unity. When this ratio is close to unity, the simpler model might be considered as a fair representation of the groups in the total character space.

9c.3 The Problem of Maximal Average $D^2$

The general problem to be solved is what are the best $t (< p)$ linear combinations of the $p$ variates which make the sum of all possible $D^2$ values arising out of a number of populations as calculated with these $t$ variates a maximum?

Let

$$
m_{11} \cdots m_{p1} \\
m_{12} \cdots m_{p2} \\
\vdots \hspace{1cm} \vdots \\
m_{1k} \cdots m_{pk}
$$

represent the mean values of $p$ characters for the first, second, \ldots, $k$th populations, and $A$ the common dispersion matrix. Suppose that $t = 1$ and the required linear function is $l_1 x_1 + l_2 x_2 + \cdots + l_p x_p$ with its variance $I A I'$ where $I$ is the vector $(l_1, l_2, \cdots, l_p)$. The $D^2$ with respect
to \(d_1 \chi_1 + \cdots + d_p \chi_p\) between the \(i\)th and \(j\)th populations is

\[
[l_1(m_{1i} - m_{1j}) + \cdots + l_p(m_{pi} - m_{pj})]^2
\]

The sum of all possible \(D^2\) values is proportional to

\[
\Sigma \Sigma l_i d_{ij}
\]

where

\[
b_{ij} = \Sigma(m_{ij} - \bar{m}_i)(m_{ij} - \bar{m}_j)
\]

and \(\bar{m}_i = (\Sigma m_{ij})/k\). This is nothing but between populations variance with respect to the chosen linear compound.

To find the best linear function, \(\Sigma \Sigma l_i d_{ij}\) is maximized, subject to the condition \(\Sigma d_{ij} \lambda_{ij} = 1\). Introducing the Lagrangian multiplier \(\lambda\), the vector \(l\) is obtained as the solution of

\[
(B - \lambda A) = 0
\]

where \(B = (b_{ij})\), \(A = (\lambda_{ij})\), and \(\lambda\) is the greatest root of the equation (see 1c.4).

Suppose that we want the best two functions \(l'x\) and \(m'x\) where \(l\) and \(m\) can be chosen to satisfy the conditions \(d_{ll}' = 1 = mAm'\) and \(d_{lm}' = 0\). Introducing Lagrangian multipliers \(c_1, c_2, c_3\), the expression to be differentiated is

\[
IB' + mBm' - c_1 lA' - 2c_1Am' - c_2 mAm'
\]

The equations are

\[
IB - c_1 lA - c_2 mA = 0
\]

\[
mB - c_2 mA - c_3 lA = 0
\]

The value of \(c_3\) can be shown to be zero by multiplying the first equation by \(m'\) and the second by \(l'\) and adding. This shows that \(c_1\) and \(c_2\) are roots of the equation

\[
|B - \lambda A| = 0
\]

Since the maximized value of \(IB' + mBm'\) is \(c_1 + c_2\), it follows that \(c_1\) and \(c_2\) correspond to the first two largest roots. Also the vectors corresponding to any two roots satisfy the condition \(1Am' = 0\) (see 1c.3), so that \(l'x\) and \(m'x\) are uncorrelated. The best two linear functions are thus the first two canonical vectors associated with the matrices \(B\) and \(A\) as discussed in 1c.3.

Similarly, the first \(t\) canonical variates give the best \(t\) linear functions. The sum of all possible \(D^2\) values with respect to all the \(p\) characters is proportional to sum of the roots \(\lambda_1 + \cdots + \lambda_p\), and the corresponding
AN ILLUSTRATIVE EXAMPLE

sum for the $t$ largest canonical variates is $\lambda_1 + \lambda_2 + \cdots + \lambda_t$, so that
the adequacy of the fit is judged by the smallness of the sum of the residual roots $\lambda_{t+1} + \cdots + \lambda_p$ or its ratio to the total.

9c.3 An Illustrative Example

In the example of 9b, the original variables have been already transformed into an uncorrelated set, so that, if the transformed values (for which $\Lambda = I$, the unit matrix) are used, the problem reduces to the determination of the latent roots and vectors of the between product sum matrix. From the table of mean values of the transformed characters the between group sum of squares and products are calculated (Table 9c.3a). It may be observed that the mean values of Table 9c.3a are so adjusted that their sum for any character over all the groups is zero, in which case no correction is necessary for the raw sum of squares and products.

One method of determining the canonical vectors is by iteration. We start with a trial vector, multiply each row of the matrix with this vector, and thus obtain a derived vector which is a better approximation than the trial one. The convergence will be quicker if we start not with the original matrix but with a suitable power of it (Hotelling, 1936). The canonical vectors associated with a symmetric matrix, $\Lambda$, and $\Lambda^{2P}$, a suitable power of $\Lambda$, are the same; the canonical roots of $\Lambda^{2P}$ are those of $\Lambda$ raised to the power $2P$. If $I$ is the vector associated with the root $\lambda$ of $\Lambda$, then

$$1(A - \lambda I) = 0$$

Multiplying by $(A + \lambda I)$, the equation reduces to

$$1(A^2 - \lambda^2 I) = 0$$

<table>
<thead>
<tr>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
<th>$y_5$</th>
<th>$y_6$</th>
<th>$y_7$</th>
<th>$y_8$</th>
<th>$y_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.296</td>
<td>0.067</td>
<td>-0.567</td>
<td>0.709</td>
<td>-0.991</td>
<td>0.074</td>
<td>-0.561</td>
<td>0.014</td>
<td>0.682</td>
</tr>
<tr>
<td>0.067</td>
<td>0.499</td>
<td>0.113</td>
<td>0.311</td>
<td>-0.158</td>
<td>0.403</td>
<td>0.239</td>
<td>0.708</td>
<td>-0.104</td>
</tr>
<tr>
<td>-0.567</td>
<td>0.113</td>
<td>0.417</td>
<td>0.070</td>
<td>0.270</td>
<td>0.214</td>
<td>0.254</td>
<td>0.157</td>
<td>-0.489</td>
</tr>
<tr>
<td>0.709</td>
<td>0.311</td>
<td>0.070</td>
<td>1.472</td>
<td>-0.503</td>
<td>0.437</td>
<td>0.192</td>
<td>-0.424</td>
<td>-0.412</td>
</tr>
<tr>
<td>-0.091</td>
<td>-0.158</td>
<td>0.270</td>
<td>-0.503</td>
<td>1.295</td>
<td>0.187</td>
<td>0.571</td>
<td>-0.604</td>
<td>0.806</td>
</tr>
<tr>
<td>0.074</td>
<td>0.403</td>
<td>0.314</td>
<td>0.437</td>
<td>0.187</td>
<td>0.731</td>
<td>0.447</td>
<td>0.498</td>
<td>-0.039</td>
</tr>
<tr>
<td>-0.361</td>
<td>0.239</td>
<td>0.254</td>
<td>0.192</td>
<td>0.571</td>
<td>0.447</td>
<td>1.022</td>
<td>0.488</td>
<td>-0.052</td>
</tr>
<tr>
<td>0.014</td>
<td>0.708</td>
<td>0.197</td>
<td>-0.424</td>
<td>-0.604</td>
<td>0.498</td>
<td>0.488</td>
<td>3.073</td>
<td>-1.140</td>
</tr>
<tr>
<td>0.882</td>
<td>-0.104</td>
<td>-0.489</td>
<td>-0.412</td>
<td>0.806</td>
<td>-0.020</td>
<td>-0.059</td>
<td>-1.140</td>
<td>2.722</td>
</tr>
</tbody>
</table>
Multiplying by \((A^2 + \lambda^2 I)\),
\[
I(A^4 - \lambda^4 I) = 0
\]
and so on. Hence the result stated above is true.

From matrix \(A\) of Table 9c.3a first \(A^2\), the square of \(A\), is obtained and then \(A^4\) by squaring \(A^2\). Finally, by squaring \(A^4\), \(A^8\) is obtained, as in Table 9c.3a.

Choosing the trial vector \((1, \ldots, 1)\), the first approximation is

\[
-2581, 12940, 7558, 4949, -18114, 9445, 7663, 50071, -41284
\]

which are simply the column totals. If we divide by the highest quantity in the set, the vector reduces to

\[
-0.0515, 0.2524, 0.1509, 0.0988, -0.3618, 0.1886, 0.1530, 1, -0.8245
\]

Multiplying each row of \(A^8\) by this vector, we derive a second approximation, and so on; the operations are repeated until stable values are obtained. After five operations the following vector is obtained.

\[
-0.1849, 0.2400, 0.1865, 0.0507, -0.3008, 0.1816, 0.1855, 1, -0.8755
\]

The highest value used in the last stage of division is 206,926, and this gives the eighth power of the first canonical root.

\[
\lambda_1^8 = 206,926 \quad \text{or} \quad \lambda_1 = 4.620
\]

This vector is standardized by dividing each element by the square root of the sum of squares of all the elements. The standardized vector is

\[
-0.129, 0.167, 0.130, 0.035, -0.210, 0.127, 0.129, 0.698, -0.611
\]
AN ILLUSTRATIVE EXAMPLE

From the \((i, j)\)th element of the matrix \(A^8\) is subtracted the product \(\lambda_1^8 \times i\)th element \(\times j\)th element of the first vector to obtain the reduced matrix given in Table 9c.3y. The above process of choosing a trial vector and finding better approximations is repeated on this reduced matrix.

TABLE 9c.3y. The Reduced Matrix Eliminating the First Root and the Vector

| 19521 | 1306 | 3907 | 6883 | -10266 | -22 | -5202 | 1304 | 503 |
| 1306 | 151 | -315 | 543 | -778 | 61 | -354 | 190 | 151 |
| 3907 | -315 | 976 | -1656 | 2538 | 24 | 1302 | -346 | -138 |
| 6883 | 543 | -1656 | 3247 | 4554 | -10 | -2292 | 331 | -11 |
| -10266 | -778 | 2538 | -4554 | 6783 | 109 | 3494 | -678 | 115 |
| -22 | 61 | 24 | -10 | 106 | 98 | 121 | 88 | 135 |
| -5202 | -354 | 1302 | -2292 | 3494 | 121 | 1844 | -298 | 17 |
| 1304 | 190 | -346 | 331 | -678 | 88 | -298 | 413 | 381 |
| 503 | 151 | -138 | -11 | 115 | 135 | 17 | 381 | 416 |

The second vector is found to be

\(0.722, 0.062, -0.103, 0.342, -0.508, -0.004, -0.259, 0.058, 0.013\)

and the second root is

\[(27,347)^{1/2} = 3.629\]

The first two canonical vectors supply the best two linear functions.

The sum of squares for \(z_1 = 4.619\) and for \(z_2 = 3.630\). They correspond to the first two canonical roots, thus providing a check on the calculations. These two roots alone account for a total variation of 4.619 \times 3.630 = 8.249 out of a total of 13.529, which is the sum of diagonal elements in matrix \(A\). The percentage of variation absorbed is 61, so that a two-dimensional representation gives a fairly accurate picture of the configuration of the groups in the nine-dimensional space. The two-dimensional representation with the canonical variates as coordinate

<table>
<thead>
<tr>
<th>(y_1)</th>
<th>(y_2)</th>
<th>(y_3)</th>
<th>(y_4)</th>
<th>(y_5)</th>
<th>(y_6)</th>
<th>(y_7)</th>
<th>(y_8)</th>
<th>(y_9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(z_1)</td>
<td>-0.129</td>
<td>0.167</td>
<td>0.130</td>
<td>0.035</td>
<td>-0.210</td>
<td>0.127</td>
<td>0.129</td>
<td>0.698</td>
</tr>
<tr>
<td>(z_2)</td>
<td>0.722</td>
<td>0.062</td>
<td>-0.103</td>
<td>0.342</td>
<td>-0.508</td>
<td>-0.004</td>
<td>-0.259</td>
<td>0.058</td>
</tr>
</tbody>
</table>

From the mean values of \(y_1, y_2, \ldots\) given in Table 9b.14 the mean values of \(z_1, z_2\) are calculated as shown in Table 9c.34.

<p>| Table 9c.34. Mean Values of the Canonical Variates |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|</p>
<table>
<thead>
<tr>
<th>(R_1)</th>
<th>(R_2)</th>
<th>(C_b)</th>
<th>(C_h)</th>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(B_h)</th>
<th>(D)</th>
<th>(A_1)</th>
<th>(A_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1)</td>
<td>0.437</td>
<td>0.380</td>
<td>-0.509</td>
<td>-0.568</td>
<td>-1.088</td>
<td>0.020</td>
<td>-0.410</td>
<td>-0.357</td>
<td>0.372</td>
</tr>
<tr>
<td>(x_2)</td>
<td>0.335</td>
<td>0.423</td>
<td>0.918</td>
<td>0.465</td>
<td>0.001</td>
<td>-0.133</td>
<td>-1.217</td>
<td>-0.783</td>
<td>0.028</td>
</tr>
</tbody>
</table>

The sum of squares for \(z_1 = 4.619\) and for \(z_2 = 3.630\). They correspond to the first two canonical roots, thus providing a check on the calculations. These two roots alone account for a total variation of 4.619 \times 3.630 = 8.249 out of a total of 13.529, which is the sum of diagonal elements in matrix \(A\). The percentage of variation absorbed is 61, so that a two-dimensional representation gives a fairly accurate picture of the configuration of the groups in the nine-dimensional space. The two-dimensional representation with the canonical variates as coordinate
axes is given in Figure 2. A third root may absorb some more variation, in which case a three-dimensional representation will supply an almost true picture of the configuration of the groups.

Figure 2. U.P. Anthropometric survey group constellations in the (λ1-λ2) chart.

9d A Test for Reduction in the Number of Dimensions

9d.1 The Analysis of Neurotic Cases

In the previous section no sampling problem was considered, the object being to obtain the best method of representation of p-dimensional data in a smaller number of dimensions. No question was asked whether the variation between the groups is entirely confined to a smaller space or whether it can be explained by variations with respect to a smaller number of hypothetical characters. Such a question would require the estimation of these hypothetical characters, if they exist, and a test of significance for the residual variation. The algebra is the same as in the previous section, but sample sizes come into the calcula-
The analysis of neurotic cases when tests of significance are considered. The method is illustrated by an example discussed elsewhere by the author and Patrick Slater who supplied the material and who also provided a suitable psychological interpretation of the statistical analysis.

The mean values of scores on 5 types of neurotics and normal cases and the within dispersion matrix have already been given in Table 8b.9a. For the following analysis we need the between and within product sum matrices.

<table>
<thead>
<tr>
<th>Table 9d.la. Analysis of Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.P. Matrix between Groups</td>
</tr>
<tr>
<td>5 D.F.</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>367.7248</td>
</tr>
<tr>
<td>161.7773</td>
</tr>
<tr>
<td>76.2389</td>
</tr>
</tbody>
</table>

A glance at the mean Table 8b.9a shows that the most conspicuous single difference is between the normal cases and the neurotics, but there are also marked differences between the groups of neurotics. The five cases of post-traumatic personality change approximate the normal cases. Closely similar to one another but distinctly different from the normal and the remaining neurotic groups are the hysterias and anxiety states. Still further removed from the normal are the obsessional and psychopathic cases; but between these two groups some differences appear to exist. Whereas the obsesionals exhibit more pointers per person than the psychopaths, this difference is wholly due to an excess of symptoms of inadequacy and shyness; there is less evidence of instability among them than among the psychopaths. This suggests that, in addition to variations in the degree to which the neurotic groups differ from normal, a further source of variation may be found and may prove useful for differential diagnosis. We may ask the following problem.

Is there sufficient evidence to demonstrate variation between the groups in more than one dimension, or can the observed differences between them be treated as differences simply in degree?

The determinantal equation for \( \lambda \), giving the canonical variances, is

\[
\text{Between matrix} - \lambda \text{ Within matrix} = 0
\]

\[
\begin{align*}
367.7248 - \lambda & 575.2127 161.7773 - \lambda 62.8946 76.2389 - \lambda 118.5423 \\
161.7773 - \lambda & 62.8946 76.4732 - \lambda 151.8666 33.0330 - \lambda 8.9436 = 0 \\
76.2389 - \lambda & 118.5423 33.0330 - \lambda 17.7109 - \lambda 148.7735
\end{align*}
\]
Various methods have been suggested for obtaining the solutions of this equation and the canonical vectors associated with the roots. To apply the method described in the previous section it is necessary to multiply the determinantal equation by the reciprocal of the dispersion matrix given after Table 8.9a. This is equivalent to multiplying the between sum of products matrix by the reciprocal of the dispersion matrix and subtracting 250A times the unit matrix (250 being the degrees of freedom of the within matrix). The resulting equation is

\[
\begin{bmatrix}
135.2909 - \mu & 58.6725 & 27.3495 \\
209.8339 & 101.4341 - \mu & 42.7341 \\
7.7001 & 2.6617 & 5.4009 - \mu
\end{bmatrix} = 0
\]

where \( \mu = 250A \). This is of the form already considered. In the above example it is convenient to expand the determinant because the resulting equation is a simple cubic.

\[
\mu^3 - 242.1261 \mu^2 + 2365.84813 \mu - 5455.761654 = 0
\]

The three roots of this equation are

\[
\mu_1 = 232.0312 \quad \mu_2 = 6.4481 \quad \mu_3 = 3.6468
\]

The total variation is 242.1261, out of which the variations absorbed by the 3 roots are 95.8%, 2.7%, and 1.5%; so the second and third are relatively unimportant.

Since the dispersion matrix is estimated on a large number of degrees of freedom, the total variation 242.1261 is approximately distributed as \( \chi^2 \) with \( p(k - 1) \) (the number of characters \( \times \) 1 minus the number of groups), equal to 15 degrees of freedom. This is an alternative test to the A criterion (see 7d.1) for judging the overall group differences.

If the whole variation is concentrated in the first canonical (hypothetical) variate, then the residual variation

\[242.1261 - 232.0312 = 10.0949\]

should be attributed to chance. The first canonical variance is distributed as \( \chi^2 \) with \( (p + k - 2) \) = 7 degrees of freedom, so that the residual is a \( \chi^2 \) with \( (15 - 7) \) = 8 degrees of freedom. The value of \( \chi^2 = 10.0949 \) has more than 20% probability on 8 degrees of freedom, thus providing no evidence of variation in the other dimensions.

If this \( \chi^2 \) is significant, we proceed to test whether the first two canonical variates are sufficient to explain the total variance. In this case the significance of the sum of the other variances has to be tested.
The distribution of the degrees of freedom among the various roots is
\[ p(k - 1) = (p + k - 2) + (p + k - 4) + \cdots \]
each term being 2 less than the previous one. The degrees of freedom for the residual in any case are the total minus the degrees of freedom for the roots accounting for the total variation on the specified hypothesis. This is equivalent to the sum of degrees of freedom of the smallest canonical variances added to form the residual.

The \( x^2 \) approximation can be slightly improved (Bartlett, 1948) by using \( [N - \frac{1}{2}(p + k)] \log_e(1 + \lambda) \) instead of \( \mu = 250X \) itself, where \( N \) is the total sample size for all groups put together. In the present example:

<table>
<thead>
<tr>
<th>Root</th>
<th>Term</th>
<th>( x^2 )</th>
<th>D.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>( [255 - \frac{1}{2}(p + k)] \log_e(1 + \lambda) )</td>
<td>165.14</td>
<td>7</td>
</tr>
<tr>
<td>Second</td>
<td>( [255 - \frac{1}{2}(p + k)] \log_e(1 + \lambda) )</td>
<td>6.35</td>
<td>5</td>
</tr>
<tr>
<td>Third</td>
<td>( [255 - \frac{1}{2}(p + k)] \log_e(1 + \lambda) )</td>
<td>3.02</td>
<td>3</td>
</tr>
</tbody>
</table>

The \( x^2 \) for the residual after the first root is eliminated is 6.35 + 3.62 = 9.97, with 5 + 3 = 8 degrees of freedom. In the previous approximation the residual \( x^2 \) is 10.0949, which is very nearly the same. Although no exact test is known, the two tests described above are equivalent in large samples.

In the present example only the first root is significant, but in problems of this nature it is of practical importance to consider at least some of the smaller roots in determining the configuration of the various groups. It might happen that the variation in the dimension corresponding to a smaller root is concentrated among a few of the groups, in which case very large samples would be necessary to establish significance. Any noticeable difference between two groups in the mean values of the canonical variate corresponding to such a root cannot be strictly interpreted as real when the overall test does not establish the significance of this root. But this additional analysis may throw some light on the prospects of future investigations. In the present case the canonical variates corresponding to the first two roots have been calculated. The configuration of the various groups is indicated in Figure 3.

The coefficients \( k_1, k_2, k_3 \) of the best linear fit or the first canonical variate are obtained from the equations
\[
\begin{align*}
(135.2909 - 232.0312)k_1 + 58.6725k_2 + 27.3195k_3 &= 0 \\
209.8339k_1 + (101.4343 - 232.0312)k_2 + 42.7341k_3 &= 0 \\
7.7001k_1 + 2.6617k_2 + (5.4009 - 232.0312)k_3 &= 0
\end{align*}
\]
The matrix of these equations is obtained by substituting for $\mu$ the maximum root 232.0312 in the matrix of the determinantal equation for $\mu$. Putting $k_3 = 1$ arbitrarily, we find that the proportional values of $k_1$ and $k_2$ are $k_1 = 18.84886$, $k_2 = 30.61225$.

The variance of $k_1A + k_2B + k_3C$ is $a_1k_1^2 + a_2k_2^2 + a_3k_3^2 + 2a_{12}k_1k_2 + 2a_{13}k_1k_3 + 2a_{23}k_2k_3$, where $a_{ij}$ are the elements of the dispersion matrix within the groups. Using the values of $k_1$, $k_2$, $k_3$ obtained as above, we find the variance as 1697.6281 or the standard deviation as $\sqrt{1697.6281} = 41.2023$. Dividing $k_1$, $k_2$, $k_3$ by this value we find the standardized best linear function $0.4575A + 0.7430B + 0.0243C$.

Similarly, the standardized linear function for the second dimension is $0.4071A - 1.0473B + 0.3292C$. The mean values of these variates for all the groups are given in Table 9d.18.

<table>
<thead>
<tr>
<th>Group</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.3879</td>
<td>0.1637</td>
</tr>
<tr>
<td>Personality change</td>
<td>2.2249</td>
<td>0.2105</td>
</tr>
<tr>
<td>Anxiety state</td>
<td>2.3227</td>
<td>0.1130</td>
</tr>
<tr>
<td>Hysteria</td>
<td>3.1339</td>
<td>-0.1115</td>
</tr>
<tr>
<td>Obsession</td>
<td>3.3601</td>
<td>0.6204</td>
</tr>
</tbody>
</table>

These values can be used for a pictorial representation of the groups as shown in Figure 3. In the first dimension the normal group occupies a position at one end of the scale; the neurotic groups are spread out towards the other extreme, the small group of cases of post-traumatic personality change being the only one which is not clearly distinct.
from normal. At the other extreme are the psychopathic personalities and the obsessionals; in terms of this dimension they lie very close together. The anxiety states and the hysterias are also found to approximate one another closely, but they lie a considerable distance from the other groups.

The preponderance of the part of the total variation between the groups which occurs in this dimension is the most striking finding in the analysis. That it is the first of the dimensions found, that each of the original scores makes a positive contribution to the variation in it, and, above all, that in it the normal group appears at one pole and the various neurotic groups diverge all towards the other—these facts suggest that it can be identified with the general factor among neurotic characteristics described by Eliot Slater, Eysenck, and other writers. Whether this general dimension of neuroticism indicates the existence of any unitary psychological trait or whether it is simply a reflection of the fact that most neurotic characteristics are non-specific and are found with varying degrees of frequency among all neurotic states is a controversial question upon which it is unnecessary to enter here. Mayer Gross, Moore, and Patrick Slater prefer the latter alternative.

Although the variation in the second dimension is very much smaller, the arrangement of the groups it discloses invites some psychological consideration. The equation defining variation in this dimension contrasts the scores for inadequacy with the scores for instability by giving them opposite signs. At one extreme is the obsessional group, which in terms of average scores is the most highly inadequate but not the most unstable; at the other extreme is the psychopathic group—the most unstable but not the most inadequate. The psychological picture presented by this arrangement of the groups is a familiar one: obsessional cases are notoriously fixed in their habits; psychopaths are notoriously irresponsible and unreliable. What is surprising is not that some contrast of this kind was found but that the observations exhibit so little variation in this respect. But, as the score is based on a summation of three pointers only, it is more likely that variation in this dimension is insignificant than that it has been insufficiently accurately measured.

Example 1. If $B$ and $W$ denote the between and within sum of product matrices and $T$ the total, then the roots of

$$|W - \psi T| = 0$$

and

$$|B - \lambda W| = 0$$
are connected by the relation

$$\psi = \frac{1}{1 + \lambda}$$

and the canonical vectors

$$I(W - \psi T) = 0 \quad \text{and} \quad I(B - \lambda W) = 0$$

are the same.

**Example 2.** The equation for $\psi$ can be deduced as follows. Evaluate the determinant

$$u(\psi) = |W - \psi T|$$

for $\psi = 0, 1, 2, 3, \ldots$ and construct the equation for $\psi$ by the method of finite differences.

[Hint: If $u_0, \Delta u_0, \Delta^2 u_0, \ldots$ denote the forward differences, then

$$u(\psi) = u_0 + \psi \Delta u_0 + \frac{\psi(\psi - 1)}{2!} \Delta^2 u_0 + \frac{\psi(\psi - 1)(\psi - 2)}{3!} \Delta^3 u_0 + \cdots.$$]

**Example 3.** If $\psi_0$ is an approximate solution of

$$u(\psi) = u_0 + \psi \Delta u_0 + \frac{\psi(\psi - 1)}{2!} \Delta^2 u_0 + \cdots$$

then the correction $\delta$ to $\psi_0$ is given by

$$-u(\psi_0) = \delta \left[ \Delta u_0 + (\psi_0 + \psi_0 - 1) \frac{\Delta^2 u_0}{2} \right.$$

$$+ \frac{d}{d\psi_0} \psi(\psi - 1)(\psi - 2) \frac{\Delta^3 u_0}{3!} + \cdots \right]$$

**Example 4.** All the solutions for $\psi$ lie between 0 and 1. [Hint: $\psi = 1/(1 + \lambda)$]

Examples 1 to 4 supply an alternative method of evaluating the canonical variances without reducing the determinantal equation $|B - \lambda W| = 0$ to the form $|C - \mu I| = 0$. It is convenient to find first the values of $\psi$, which all lie between 0 and 1. It is easy to guess an approximate root and then obtain the correction terms by iteration.

If in any problem all the roots are needed, they may be obtained by Graff’s root-squaring method, which is illustrated below. If the equation is

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_n = 0$$

then the correction $\delta$ to $\psi_0$ is given by

$$-u(\psi_0) = \delta \left[ \Delta u_0 + (\psi_0 + \psi_0 - 1) \frac{\Delta^2 u_0}{2} \right.$$

$$+ \frac{d}{d\psi_0} \psi(\psi - 1)(\psi - 2) \frac{\Delta^3 u_0}{3!} + \cdots \right]$$
then the equation whose roots are squares of the above roots is computed in the following manner

\[
\begin{array}{cccc}
x^p & x^{p-1} & x^{p-2} & \cdots \\
a_0^t & 2a_0a_2 & 2a_0a_4 & \cdots \\
-a_1^t & -2a_1a_3 -2a_1a_5 & \cdots \\
a_2^t & 2a_2a_4 & 2a_2a_6 & \cdots \\
& \cdots & \cdots & \cdots \\
\end{array}
\]

Total \( b_0 \) \( b_1 \) \( b_2 \) \( b_3 \) \( b_4 \) \cdots

This gives the first approximation to the square of the roots

\[
\psi_1^2 \sim \frac{b_1}{b_0}, \quad \psi_2^2 \sim \frac{b_2}{b_1}, \quad \psi_3^2 \sim \frac{b_3}{b_2}, \quad \cdots
\]

Starting from this equation, the first operation is repeated. If \( c_0, c_1, \cdots \) are the coefficients, then

\[
\psi_1^4 \sim \frac{c_1}{c_0}, \quad \psi_2^4 \sim \frac{c_2}{c_1}, \quad \psi_3^4 \sim \frac{c_3}{c_2}, \quad \cdots
\]

By using logarithms, the values \( \psi_1, \psi_2, \cdots \) at any stage are obtained. The process can be terminated as soon as stable values of \( \psi_1, \psi_2, \cdots \) are obtained. Each time, the powers of roots obtained are 2, 4, 8, 16, 32, \cdots, and so on, and the convergence will depend on the separation of roots in the equation. This method is suitable only when all the roots are different. In problems of the nature discussed in this chapter, cases of exactly equal roots rarely occur.

Example 5. If \( b \) is an arbitrary vector, show that

\[
\lim_{n \to \infty} \frac{bA^n}{\lambda_1^n} = x_1
\]

where \( \lambda_1 \) is the dominant root and \( x_1 \) is the first characteristic vector arising out of the determinantal equation

\[
| A - \lambda I | = 0
\]

This justifies the iterative method adopted on p. 368.

Example 6. Taking an arbitrary vector \( k \), form the matrix \( P \), whose rows are the vector,

\[
k, kA, \cdots, kA^n
\]

where \( A \) is a symmetric matrix of order \( p \). Solve the equations

\[
bP = 0
\]
and show that the expansion of $|A - \lambda I| = 0$ is

$$b_p \lambda^p + b_{p-1} \lambda^{p-1} + \cdots + b_0 = 0$$

where $b = (b_0, \ldots, b_p)$.

[Hint: First show that $A$ satisfies the equation

$$b_p A^p + \cdots + b_0 I = 0$$

References


Appendix.

Miscellaneous Problems

The following problems are intended to serve as exercises and discussion problems.

1. Distinguish between a mathematical limit and a stochastic limit. Show that if \( x_n \xrightarrow{\text{stoch}} x \) stochastically then the limiting distribution of \( x_n \) is the same as the distribution of \( x \).

2. If \( x \) is a stochastic variable assuming only non-negative values and \( E(x) = t \), then
\[
P(x < t^2) > 1 - \frac{1}{\lambda^2}
\]

3. Find the general form of the probability law which has the median as the maximum likelihood estimator of a parameter.

4. On the basis of a sample of size \( n \) how do you test whether the observations have arisen from a rectangular population with an unknown range?

This is equivalent to testing whether the middle \((n-2)\) observations have a rectangular distribution in the range determined by the smallest and the biggest observation.

5. From the joint distribution of \( x \) and \( S^2 \) as defined in 2b.1 find the distribution of \( \sqrt[n]{s} \) when \( \mu \neq 0 \) and show how this can be used to test for an assigned value of the coefficient of variation in the population.

Derive a locally most powerful unbiased test for the null hypothesis that the coefficient of variation has an assigned value.

6. Show without actually working out the distribution that, for samples from a multivariate normal distribution, the distribution of the computed multiple correlation coefficient contains only the corresponding population parameter and a similar result is true for the partial correlation coefficient.

7. Prove that
\[
a^T G a' = (a^T G^T a')^\frac{1}{2} (a^T G T^{-1} G a')^\frac{1}{2}
\]
where \( a \) is an arbitrary vector and \( T \) and \( G \) are positive definite matrices. Make use of the fact that the ratio of the left-hand expression to that of the right is the correlation between the vectors \( x = a D \) and \( y = a G D^{-1} \) where \( DD' = T \).

Hence, following the notions of 8c.2, show that the genetic advance
as determined by the discriminant function is greater than that for the straight selection function.

8. Calculate the information limits to the variances of unbiased estimates of $\mu_1$ and $\mu_2$ in the problem of 4e.1 and show that they are actually attained for the maximum likelihood estimates.

9. A coin is tossed a number of times till $z$ heads appear. Find the average and the variance of the number of tosses. Determine a recurrence relation for the evaluation of higher moments.

10. Using the data considered in 7d.5, obtain the function of the measurements which characterizes most effectively the secular changes in progress. Assume that regressions with time are linear.

Hint: The probability density for any time $t$ can be written

$$\text{const.} \exp - \frac{1}{2} \sum \lambda_i (x_i - a_i - b_i t)(x_i - a_i - b_i t)$$

The likelihood ratio for two different time points gives the best discriminating function. In this case it is linear in the measurements. Estimate the constants.

11. Give any method by which the number of fishes in a pond can be estimated.

In an investigation to estimate the number of births in a locality during a particular month the following data are obtained: 350 families reported no birth and 100 reported a single birth of which 25 took place in the government hospital. Knowing the exact number of births in the hospital, how do you estimate the total number of births in the locality? Find the standard error of this estimate.

12. Comment on the following anecdote: On examining the chart giving the standard heights and associated weights of persons a husband remarked that his wife weighs more than she should. The wife retorted saying that she is not so tall as she should be.

13. A total number of 130 heads was observed when 100 rupee coins and 100 half-rupee coins were thrown together. Knowing that the rupee coin is unbiased, what can you say about the half-rupee coin?

14. Show that under conditions of random mating the blood group frequencies in the four classes $O, A, B, AB$ can be expressed in terms of gene frequencies only.

15. The following data relate to the distribution of age at death in a particular year classified according to civil conditions.

<table>
<thead>
<tr>
<th>Civil Condition</th>
<th>Age at Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelor</td>
<td>0-15 15-30 30-45 45-60 60-75 Over 75</td>
</tr>
<tr>
<td>Married</td>
<td>10 70 100 120 50 45 10</td>
</tr>
<tr>
<td>Widowed</td>
<td>12 45 30 20</td>
</tr>
</tbody>
</table>
What conclusions can you draw about the effect of civil condition on the age at death? What more is needed for a proper comparison of these distributions?

16. How do you collect material to study whether the correlation coefficient between the measurements of head length and head breadth on the skull is greater than the correlation coefficient between the corresponding measurements on the living? What is the appropriate statistic to test the above null hypothesis if all four measurements are available for each of n individuals?

17. Find on the basis of a sample of size n the best estimate of the correlation coefficient in a bivariate normal population when both the variables have the same mean and variance. This coefficient is known as the intra-class correlation.

Suppose that p measurements are available on each of two brothers from a number of families. Determine the best linear compound of the measurements which has the highest intra-class correlation coefficient between the brothers.

Obtain the large and small sample distributions of the maximum correlation and indicate their use in tests of significance.

18. The following estimates of the correlation coefficients between intelligence test scores were found in an investigation to study the relative influences of environmental and heredity factors.

<table>
<thead>
<tr>
<th></th>
<th>Two Brothers</th>
<th>Two Brothers</th>
<th>Twins</th>
<th>Twins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reared</td>
<td>Living</td>
<td>Reared</td>
<td>Living</td>
</tr>
<tr>
<td></td>
<td>Apart</td>
<td>Together</td>
<td>Apart</td>
<td>Together</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.235</td>
<td>0.342</td>
<td>0.451</td>
<td>0.513</td>
</tr>
<tr>
<td>Sample size</td>
<td>50</td>
<td>40</td>
<td>45</td>
<td>55</td>
</tr>
</tbody>
</table>

Comment on these figures, using tests of significance wherever necessary.

19. It is said that man's stature is variable during the day and that the range is half an inch, the maximum stature being observable when the man gets up in the morning and the minimum when he retires to bed. How will you proceed to study this variability? Also comment on the precision with which the stature has to be measured for such studies.

20. To determine the sex ratio three procedures were adopted: (1) A number of boys were asked to state the number of brothers (including himself) and sisters they have. (2) A number of girls were asked to state the number of sisters (including herself) and brothers they have. (3) Finally a number of parents were asked to give the number of sons
and daughters they have. The totals of brothers, sisters, and families for the three procedures are given below.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brothers</td>
<td>203</td>
<td>140</td>
<td>120</td>
</tr>
<tr>
<td>Sisters</td>
<td>130</td>
<td>220</td>
<td>115</td>
</tr>
<tr>
<td>Total families</td>
<td>80</td>
<td>95</td>
<td>60</td>
</tr>
</tbody>
</table>

Test whether the family size is the same in the three investigations. Find the estimates of the sex ratio for the three procedures, and test whether they are significantly different.
Index

(Names occurring in references are indicated by r, and those in footnotes by n.)

Aitken, A. C., 29r, 127r, 253, 271r
Analysis of covariance, concomitant variation, 119
Analysis of dispersion, 258
Analysis of variance, concomitant variation, 119
equality of regression equations, 112
least square method, 83
one-way classification, 89
significance of regression coefficients, 105
test for an assigned regression function, 115
two-way classification, 91
unequal numbers in cells, 94
Anscombe, F. J., 209, 210, 220r
Anthropological illustrations, asymmetry and kurtosis of nasal height distributions, 219
asymmetry of right and left femora, 87
cranial capacity, 103
differences in boys of different schools, 202
differences of the first and second born, 245
differences in sexing of skulls, 197
differences in variabilities of males and females, 225
estimation from incomplete data, 161
femur and humerus differences in groups, 262
group constellations of castes in United Provinces, 357
group constellations using canonical variates, 370
group differences in head breadth, 90
group differences in nasal height, 94
inheritance in man, 134

Anthropological illustrations, regression of nasal index on weather factors, 115
secular variations in skull characters, 206
sexing of osteometric material, 304
stature of prehistoric man, comments, 115
use of indices, 155
Asymmetric matrix, canonical reduction, 27
Asymptotic distribution, of quantiles, 156
of the median, 157
of maximum likelihood estimates, 157
Backcross data, 188
Barnard, M. M., 266, 267, 271, 271r
Bartlett, M. S., 201, 220r, 226, 228, 235r, 236, 250, 201, 271r, 373, 378r
Bateson, 180, 194
Bayes' theorem, 308
Bernstein, F., 170, 174r
Beta variable, distribution, 41
distribution of product, 42
moments of, 42
Bhattacharya, A., 143, 174r
Bias in an estimate, 151
Binomial proportion, estimation, 134
sin⁻¹ transformation, 210
Binomial variate, distribution, 32
distribution function, 33
moments of, 34
Biological illustrations, differential thickness of bark in different directions, 241
feeding experiment on pigs, 121
milk feeding experiment, 217
see also Anthropological illustrations,
Genetic illustrations, Psychological illustrations
INDEX

Birnbaum, Z. W., 336, 340
Bivariate normal distribution, estimation of parameters, 161
moments of, 153
Bliss, C. I., 255, 256, 271r
Blood groups, O, A, B, comparison of gene frequencies, 187
consistency test, 184
estimation of gene frequencies, 169
Box, 124
Bose, R. C., 74r, 80n
Brogden, H. E., 338, 340r
Brookner, R. J., 236, 272r
Canonical reduction, of asymmetric matrix, 27
do symmetric matrix, 24
Canonical variates, illustrations, 367, 370
use in group constellations, 364
Cauchy population, distribution, 42
distribution of the mean in samples, 43
efficient estimate of the parameter, 166
inconsistent estimate of the parameter, 152
Central limit theorem, 174
Chapman, D. G., 336
Characteristic vectors, 21
Chi-square, $x^2$, approximations to distribution of, 222
distribution of least squares, 58
distribution of sum of, 41
general large sample theory of $x^2$ test, 179
goodness of fit test, 183
limits to $x^2$ in test for specified variance, 223
multivariate tests of, 257, 294, 372
non-central distribution of, 50, 57
ratio of, 45
test of homogeneity, 185
tests in contingency tables, 192
Classificatory problems, allocation of individuals to groups in a fixed ratio, 322
differential predictors, 337
discriminant function, 287
doubtful region, 206
in genetic selection, 329
in job selection, 336
Classificatory problems, with more than two groups, 307
resolution of a mixed series, 300
single predictor, 336
with two groups, 286
uncertainty of the a priori information, 290
Cochran's theorem, 49, 55
Cochran, W. G., 74r, 253, 256, 271r
Coefficient of correlation, see Correlation coefficient
Coefficient of racial likeness, C.R.L., 355
reduced form, R.C.R.L., 356
Combination of data from various sources, 172
Combination of tests, $P_A$, 41, 217
Concomitant variation, adjustment for, 119
Conditional tests, 201, 205
Confidence interval, 276
Consistency in estimation, 151
Consistency relation of moments, in population, 27
in samples, 17
Contingency tables, with small samples, 200
tests of independence, 192
Convergence theorems, 172, 173
Correction for grouping, 301
Correlation coefficient, estimate from parallel samples, 234
multiple, distribution of, 63, 65
partial, distribution of, 69
simple, distribution of, 68
tanh$^{-1}$ transformation, 231
tests of significance, 230, 232
Covariation data, 167
Covariance analysis, concomitant variation, 119
Cramer, H., 158, 172, 174r
Cranial capacity, estimation of mean, 110
prediction formula for, 103
preservation of skulls with small, 111
$D^2$(Mahalanobis), applications, 246, 354, 357
distribution, 79
generalization to many groups, 257
INDEX

Dandekar's correction for continuity, 203
see also Yates' correction, 203

Deficiency matrix, 3

Degrees of freedom, of \( \chi^2 \), 177, 179
of least squares, 84

Dependence of vectors, 2

Determinantal equation, 22

Determinants, 10

reduction of, 31

Discriminant function, construction of, 247
difficulties in its use, 289
in job selection problems, 336, 337
as ratio of likelihoods, 287
successive evaluation of these functions, 254
test for an assigned, 248
test for equality of functions, 250
test for its coefficients, 251

Discriminant score, linear, 310

Discrimination of neurotic conditions, 316

Discriminatory problems, see Classificatory problems

Discriminatory topology, 252

Dispersion matrix, 52

Dissection of a compound frequency curve, 300

Distance between two populations, 352

Mahalanobis' \( D^2 \), 70, 246, 354, 357

Distribution, of \( \chi^2 \) goodness of fit, 177, 179
of correlation coefficient, 68
of Fisher's \( z \), 48
of Hotelling's \( T \), 70
of \( L \) and \( M \) statistics for homogeneity of variances, 226, 227
of \( \Lambda \) statistic (Wilks), 261
of least squares, 58
of linear functions of normal variates, 53
of Mahalanobis' \( D^2 \), 70
of maximum likelihood estimates, asymptotic, 127
of mean and variance in normal samples, 46
of multiple correlation coefficient, 63, 65
of non-central \( \chi^2 \), 50

Distribution, of non-null \( t^2 \), 48
of \( \phi \) of Pearson, 44
of partial correlation coefficient, 69
of product of beta variables, 42
of quadratic forms, 45, 55
of quantiles, asymptotic, 156
of regression coefficients, 63
of square of a normal variable, 40
of Students' \( t \), 47
of \( U \)-statistic, comparison of two \( D^2 \) values, 74
of \( U \)-statistic, generalization of \( D^2 \), 257
of variances and covariances (Wishart), 65
of \( W \)-statistic, the difference of two \( D^2 \) values, 255

Distribution of sum of variates, Cauchy,
43

gamma, 41

normal, 39

Poisson, 36, 37

Distribution function, binomial, 33

Poisson, 38

Doubtful region in classification, 296

Dugue, 158, 174

Efficiency of an estimate, 155

Elderton, W. E., 220

Elementary matrix, 13

Equations, linear, homogeneous, 5	nonhomogeneous, 6

numerical solutions, 31

Estimation, linear, with correlated variables, 81

intrinsic properties of normal equations, 79

normal equations, 78

observational equations, 75

of regression coefficients, 103

with restrictions on parameters, 81

standard errors of estimates, 78

with two sets of parameters, 118

with weighted observations, 85

method of maximum likelihood, 150

combination of data, 172

consistency and bias, 151

efficiency, 155

estimation from incomplete data, 161

method of scoring, 165
Estimation, method of maximum likelihood, optimum properties of estimates, 157
the primitive postulate, 151
relation to sufficient statistics, 151
minimum variance, information limit to variance, 130
ea lower bound to variance, 143
the problem of several parameters, 144
relation to sufficient statistics, 135, 139
Eysenck, 375

F, f distributions, 45, 48, 49
Factors of neurosis, 270
Fairfield Smith, H., 330, 349r
Fiducial interval, 276
Fisher, R. A., 74r, 87, 87n, 124, 127r, 128r, 131, 130, 139, 147, 175r, 200, 211, 220r, 221, 222, 231, 235, 247, 271r, 272r, 276, 287, 288, 301
Fisher's z distribution, 48
Frets, G. P., 245, 272r

Gamma variate, distribution of sum of variates, 41
estimation of parameters of distribution of, 134, 141
moments of, 41
Gauss, K. S., 76
Generalized distance, see D^2
Generalized variance, 174
Genetic advance, 332
Genetic illustrations, comparison of gene frequencies, 186
comparison of recombination fractions, 188, 211
consistency of blood group frequencies, 184
detection of linkage, 181, 185
discrimination between Iris versicolor and Iris setosa, 248
estimation of blood group frequencies, 169
estimation of linkage, coupling and repulsion, 167
inheritance in man, 124
selection of poultry by discriminant function, 330

Genetic selection, discriminant function in, 329
Gouldwin, C. N., 292, 349r
Gosset, W. S. (Student), 87n, 128r
Gray, H., 124, 127r
Gross, Mayer, 375
Grouping correction, 301
Hartley, H. O., 228, 235r
Homogeneity tests, contingency table, 185
of correlations, 233
of variances, 226
Hoek, B. G. E., 103, 111, 128r, 292, 349r
Hotelling, H., 74r, 79, 111, 128r, 236, 272r, 280, 307, 378r
Hotelling's T, 70
Hsu, P. L., 74r, 237, 272r
Huzurbazar, V. S., 158, 175r
Hypothesis, composite, 176
null, 174
simple, 176

Illustrations, see Anthropological, Biological, Genetic, and Psychological
Incomplete data, estimation from, 161
Index, consistent estimate of, 155
moments of, 154
Information, 131
Information limit, 131
Intercross data, 189
Intrinsic properties of normal equations, 79
Invariance, under orthogonal transformation, 19
of some statistics, 72
Jeffreys, H., 275, 349r
k-statistic, 301
Kemeny, W. F. F., 225, 235r
Kintchine's theorem, 174
Kolodziejczyk, St., 128r
Koopman, B. O., 149, 175r
Kurtosis, test for deviation from normal, 219
INDEX

L test, 226, 227
Large sample, standard errors of, index, 154
median, 157
moments, 215
quantiles, 156
transformed statistics, 207
Large sample tests, \( \chi^2 \) goodness of fit, 179
of deviation from symmetry and normal kurtosis, 219
difference between means, 217
homogeneity of parallel samples, 185
homogeneity of Poisson samples, 205
independence in contingency tables, 192
with transformed statistics, 207
Latent roots of a matrix, 21
Least squares, in estimation, 80, 81, 82
evaluation of, 85
two fundamental distributions in, 58
Levi, F. W., 29r
Levy, P., 175r
Levy's theorem, 174
Liapounoff's theorem, 174
Likelihood, 150
Likelihood ratio test, 226, 259, 286
Limiting distribution, of binomial variate, 38
of maximum likelihood estimates, 157
of quantiles, 156
Limiting theorems, central limit, Lindberg-Levy's, 174
convergence of sum, product and ratio, 177
Khintchine's, 174
Levy's, 174
Liapounoff's, 174
Slutsky's, 173
Linear equations, homogeneous, 5
nonhomogeneous, 5
numerical methods of solution, 31
Linear estimation, see Estimation, linear
Linear hypothesis, nature of, 82
test of, 83
Linear transformation, 18
Lindberg, J. W., 175r
Lindberg-Levy theorem, 160, 174
M test, homogeneity of variances, 227

Macdonell, W. R., 104n, 129r
Mahalanobis, P. C., 246, 258, 272r, 355, 357, 379r
Mahalanobis' \( D^2 \), see \( D^2 \)
Majumdar, D. N., 318, 327, 378r
Markoff, A. A., 76, 129r
Martin, E. S., 349r
Matther, K., 167, 175r
Matrix, characteristic vectors of, 21
the determinant of, 10
elementary, 13
latent roots of, 21
of linear transformations, 18
numerical computations with, 31
partitioned, 9
of a quadratic form, 18
rank of, 2
reciprocal, 13
reduction of an asymmetric form, 27
sweeping out a, 2
unit matrix, 8
Maximum likelihood, see Estimation, method of maximum likelihood
Merril, A. S., 175r
Minimax, some comments on, 313
Minimum variance, see Estimation, minimum variance
Mises, R. V., 312, 349r
Mixed series, resolution of, 300
Moments, of beta distribution, 42
of binomial distribution, 34
of bivariate normal, 153
of gamma distribution, 41
of multinomial distribution, 35
of normal distribution, 39
of Poisson distribution, 36
of moment statistics, 215
Moore, 375
Meant, G. M., 292, 349r, 355, 378r
Multinomial distribution, moments of, 35
Multiple correlation, distribution of, 63, 65
test for, 104
Multivariate analysis, of dispersion, 259
generalization of \( D^2 \), 257
internal analysis of variates, 264
of neurotic patients, 370
problem of secular variations in skull characters, 206
INDEX

Multivariate analysis, problems of a single sample, 230
problems of two samples, 246
review of work on, 236
successive evaluation of discriminant functions, 254
test for additional information due to some characters, 252
test for an assigned contrast of p correlated variables, 243
test for an assigned discriminant function, 248
test for an assigned ratio of discriminant function coefficients, 251
test for asymmetry, 245
test for differences in mean values, 262
test for differences in two samples, 248
test for equality of discriminant functions in parallel samples, 251
test for equality of means of p correlated variables, 240
use of the W-statistic, difference of two $D^2$ values, 255
Wilks's criterion, 258

Multivariate normal population, distribution of linear functions from, 53
distribution of quadratic forms in, 55
variances and covariances in, 52

Multivariate distributions, Hotelling's $T^2$, 70
Mahalanobis' $D^2$, 70
multiple correlation, 63, 65
partial correlation coefficient, 69
$U$-statistic, comparison of two $D^2$ values, 74
variances and covariances, 66
Wilks's $A$ criterion, 260

Münter, A. H., 128r, 163, 175r

Nair, U. S., 52, 74r, 260, 272r

Nandi, H. K., 74r

Neurosis, the discrimination problem of, 316
factors of, 370

Neyman, J., 130, 175r, 224, 235r, 236, 272r, 277, 278, 280, 282, 285, 349r, 350r
Neyman-Pearson lemma, 309, 339

Neyman-Pearson theory of hypothesis testing, 278
Noncentral $x^2$ distribution, 50
Non-null $t$ distribution, 48
Normal equations, 76
intrinsic properties, 79
Normal population, distribution of mean and variance, 46
distribution of the square, 40
distribution of sum of variates, 39
estimation of parameters, 133, 139, 140, 144, 150

moments of, 39

see also Multivariate normal population

Multivariate normal population

Null hypothesis, 274
Null vector, 1

Observational equations, 75
see also Estimation, linear

Orthogonal transformation, 19
Orthogonal vector space, 3

$P_X$ distribution, 44
application of, 217

Pairs of quadratic forms, 25

Pareto, V. G., 330, 350r
Partial correlation coefficient, 69, 365
Partitioned matrix, 9

Pearson, E. S., 224, 228, 235r, 236, 272r, 277, 278, 280, 282, 285, 349r, 350r
Pearson, K., 33, 38, 74r, 112, 115, 128r, 227r, 300, 321, 350r, 355
Pearson's (Karl) $P_X$, 44
coefficient of racial likeness, 355
Perrone, L. S., 307, 350r

Pivotal condensation, 29

Poisson variate, conditional distributions, 36, 37
distribution function, 38
distribution of the sum, 36, 37
estimation of the mean, 133
large sample tests, 205
moments of, 36

Power function, 277
a lemma on, 342

Prediction formula, 103
for cranial capacity, 103

Psychological illustrations, discrimination of neurotic conditions, 316
INDEX

Psychological illustrations, factors of neurosis, 370
sociability of village and city recruits, 202
speech and physical defects, 100
Quadratic forms, classification of, 19
distribution of, 56
reduction of two forms, 25
Quantiles, asymptotic distribution, 156
Range, estimation of, 142
Rank of a matrix, 2
Rao, C. R., 74r, 128r, 150, 175r, 220r,
272r, 316, 350r, 357, 378r
Reciprocal matrix, 13
numerical evaluation, 31
Recombination fraction, 167
Regression coefficients, distribution of,
63
estimation of, 103
test for, 104
Residual sum of squares, see Least squares
Scoring method, 165
combination of data, 172
in estimating linkage, 167
gene frequencies, 169
Similar region, 286
Sin^{-1} transformation, 210
Slater, Elliot, 375
Slater, Patrick, 316, 350r, 371, 375, 378r
Slutsky, E., 175r
Sutsky's theorem, 173
Statistics used in tests of significance, see
\chi^2, D^2, F, \phi, \kappa, M, \phi_1, \tau, T, U, V,
W, \xi (Wilk's), z
Student (W. S. Gosset), 87, 87n, 128r
Student's distribution, 47
Student's generalization of t, 239
Student's t test, 87
Sufficient statistics, 135
distributions admitting, 137, 149
an optimum property of, 139
minimal set of, 144
Sweep-out method, 2
Student's, distribution, 47
non-null distribution, 48
Student's, test, 87
T, Hotelling's, applications, 239
distribution, 70
Tanh^{-1} transformation, 231
Tests, distance power, 281
likelihood ratio, 226, 259, 285
locally most powerful unbiased, 280
uniformly most powerful, 278
see also Analysis of variance, Chi square, Homogeneity tests, Large sample tests, Multivariate analysis
Tildesley, M. L., 355, 378r
Tocher, K. D., 303
Transformation of statistics, expression
for variance under, 207
log of standard deviation, 214
sin^{-1} root of binomial proportion, 219
square root of Poisson variate, 208
\Phi^{-1} of correlation coefficient, 214
Tchebyshoff's lemma, 152
Turnbull, H. W., 20r
U-statistic, application, 245, 250, 251,
253
distribution, 74
Unbiased estimation, general, 129
an important aspect of, 130
linear, 73
Unbiased minimum variance, see Estimation, minimum variance
Uniformly most powerful test, 278
V-statistic, generalization of D^2, 257
Variance, distribution of, in normal samples, 46
information limit to, in estimation,
130, 143, 144
log transformation, 214
test for a given inequality of variances,
221
test for homogeneity of variances, 221
test for a specified value of, 221
unbiased minimum variance, estimate of, 148
see also Analysis of variance
Vectors, linear independence of, 2
orthogonality of, 2
Vector space, basis of, 2
INDEX

Vector space, orthogonal, 3
rank of, 2

Von Mises, R., 312, 349r

W-statistic, difference of two $D^2$ values, 255
Wald, A., 160, 175r, 178, 229r, 236, 272r, 329, 350r
Wilks, S. S., 236, 259, 260, 272r

Wilks' $\Lambda$ criterion, 218
Wishart, J., 74r, 121, 128r, 236, 272r
Wishart's distribution, 66

Yates, F., 87n, 158, 175r, 203, 229r
Yates' correction for continuity, 203
see also Dandekar's correction, 203

z distribution, 48