Statistical Analysis of Management Data

Hubert Gatignon

Statistical Analysis of Management Data

Second Edition

Springer

Hubert Gatignon
Boulevard de Constance
INSEAD
77305 Fontainebleau
France
hubert.gatignon@insead.edu

ISBN 978-1-4419-1269-5 e-ISBN 978-1-4419-1270-1
DOI 10.1007/978-1-4419-1270-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009937584

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my daughters, Aline and Valérie

Preface to Second Edition

This second edition reflects a slight evolution in the methods for analysis of data for research in the field of management and in the related fields in the social sciences. In particular, it places a greater emphasis on measurement models. This new version includes a separate chapter on confirmatory factor analysis, with new sections on second-order factor analytic models and multiple group factor analysis. A new, separate section on analysis of covariance structure discusses multigroup problems that are particularly useful for testing moderating effects. Some fundamental multivariate methods such as canonical correlation analysis and cluster analysis have also been added. Canonical correlation analysis is useful because it helps better understand other methodologies already covered in the first version of this book. Cluster analysis remains a classic method used across fields and in applied research.

The philosophy of the book remains identical to that of its original version, which I have put in practice continuously in teaching this material in my doctoral classes. The objectives articulated in Chapter 1 have guided the writing not only of the first edition of this book but also of this new edition.

In addition to all the individuals I am indebted to and who have been identified in the first edition of this book, I would like to express my thanks to the cohorts of students since then. The continuous feedback has helped select the new material covered in this book with the objective to improve the understanding of the material. Finally, I would like to thank my assistant of 15 years, Georgette Duprat, whose commitment to details never fails.

Preface

I am very indebted to a number of people without whom I would not have envisioned this book. First, Paul Green helped me tremendously in the preparation of the first doctoral seminar I taught at the Wharton School. The orientations and objectives set for that book reflect those he had for the seminar on data analysis, which he used to teach before I did. A second individual, Lee Cooper at UCLA, was determinant in the approach I used for teaching statistics. As my first teacher of multivariate statistics, the exercise of having to program all the methods in APL taught me the benefits of such an approach for the complete understanding of this material. Finally, I owe a debt to all the doctoral students in the various fields of management, both at Wharton and INSEAD, who have, by their questions and feedback, helped me develop this approach. I hope it will benefit future students in learning these statistical tools, which are basic to academic research in the field of management especially. Special thanks go to Bruce Hardie, who helped me put together some of the databases, and to Frédéric Dalsace, who carefully identified sections that needed further explanation and editing. Also, my research assistant at INSEAD, Gueram Sargsyan was instrumental in preparing the examples used in this manual to illustrate the various methods.

Contents

Chapter 1
Introduction

1.1 Overview

This book covers multivariate statistical analyses that are important for researchers in all fields of management, whether finance, production, accounting, marketing, strategy, technology, or human resources management. Although multivariate statistical techniques such as those described in this book play key roles in fundamental disciplines of the social sciences (e.g., economics and econometrics or psychology and psychometrics), the methodologies particularly relevant and typically used in management research are the focus of this study.

This book is especially designed to provide doctoral students with a theoretical knowledge of the basic concepts underlying the most important multivariate techniques and with an overview of actual applications in various fields. The book addresses both the underlying mathematics and *problems of application*. As such, a reasonable level of competence in both statistics and mathematics is needed. This book is not intended as a first introduction to statistics and statistical analysis. Instead, it assumes that the student is familiar with basic univariate statistical techniques. The book presents the techniques in a fundamental way but in a format accessible to students in a doctoral program, to practicing academicians and data analysts. With this in mind, it may be recommended to review some basic statistics and matrix algebra such as those provided in the following books:

Green, Paul E. (1978), *Mathematical Tools for Applied Multivariate Analysis*, New York: Academic Press, [Chapters 2–4].

Maddala, Gangadharrao S. (1977), *Econometrics*, New York: McGraw Hill Inc., [Appendix A].

This book offers a clear, succinct exposition of each technique with emphasis on when each technique is appropriate for use and how to use it. The focus is on the essential aspects that a working researcher will encounter. In short, the focus is on using multivariate analysis appropriately through understanding of the foundations of the methods to gain valid and fruitful insights into management problems. This book presents methodologies for analyzing primary or secondary data typically

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4419-1270-1_1,

used by academics as well as analysts in management research and provides an opportunity for the researcher to have hands-on experience with such methods.

1.2 Objectives

The main objectives of this book are

1. To develop the student's knowledge of the technical details of various techniques for analyzing data.
2. To expose students to applications and "hands-on" use of various computer programs. This experience will enable students to carry out statistical analyses of their own data. Commonly available software is used throughout the book, as much as possible, across methodologies to avoid having to learn multiple systems, each presenting their own specific data manipulation instructions. However, not a single data analysis software performs all the analyses presented in the book. Therefore, three basic statistical packages are used: SAS, LIMDEP, and LISREL.

1.2.1 Develop the Student's Knowledge of the Technical Details of Various Techniques for Analyzing Data

The first objective is to prepare the researcher with the basic technical knowledge required for understanding the methods, as well as their limitations. This requires a thorough understanding of the fundamental properties of the techniques. Basic knowledge means that the book will not deal in-depth with the methodologies. This depth should be acquired through specialized, more advanced books on the specific topics. Nevertheless, this book provides enough details of what is the minimum knowledge expected of a doctoral candidate in management studies. "Basic" should not be interpreted as a lower level of technical expertise. It is used to express the minimum knowledge expected of an academic researcher in management. The objective is to train the reader to understand the technique, to be able to use it, and to have the sufficient knowledge to understand the more advanced technique that can be subsequently found in other books.

1.2.2 Expose Students to Applications and "Hand-On" Use of Various Computer Programs for Carrying Out Statistical Analyses of Data

Although the basic statistical theories corresponding to the various types of analysis are necessary, they are not sufficient to carry out research. The use of any technique requires the knowledge of the statistical software corresponding to these analyses.

It is indispensable that students learn both the theory *and the practice* of using these methods *at the same time*. A very effective, albeit time consuming, way to ensure that the intricacies of a technique are mastered is by programming the software oneself. A quicker way is to ensure that the use of the software coincides with the learning of the theory by associating application examples with the theory and by doing some analysis oneself.

Consequently, in this book, each chapter comprises four parts. The first part of each chapter presents the methods from a theoretical point of view with the various properties of the method. The second part shows an example of an analysis with instructions on how to use a particular software program appropriate for that analysis. The third part is an assignment so that students can actually practice the method of analysis. The data sets for these assignments are described in Appendix C and can be downloaded from the personal web page of Hubert Gatignon at: http://www.insead.edu/. Finally, the fourth part consists of references of articles in which such techniques are used appropriately, and which serve as templates. Selected readings could have been reprinted in this book for each application. However, few articles illustrate all the facets of the techniques. By providing a list of articles, each student can choose the applications that correspond best to his or her interests. By accessing multiple articles in the area of interest, the learning becomes richer. All these articles illustrating the particular multivariate techniques used in empirical analysis are drawn from major research journals in the field of management.

1.3 Types of Scales

Data used in management research are obtained from existing sources (secondary data) such as the data published by Ward for automobile sales in the USA or from vendors who collect data, such as panel data. Data are also collected for the explicit purpose of the study (primary data): survey data, scanner data, or panels.

In addition to this variety of data sources, differences in the type of data that are collected can be critical for their analysis. Some data are continuous measures such as, for example, the age of a person, with an absolute starting point at birth, or the distance between two points. Some commonly used data do not have such an absolute starting point. Temperature is an example of such a measure. Yet in both cases, i.e., temperatures and distances, multiple units of measurement exist throughout the world. These differences are critical because the appropriateness of data analysis methods varies, depending on the type of data at hand. In fact, very often, the data may have to be collected in a certain way in order to be able to test hypotheses using the appropriate methodology. Failure to collect the appropriate type of data would prevent performing the test.

In this chapter, we discuss the different types of scales, found in management research, for measuring variables.

1.3.1 Definition of Different Types of Scales

Scales are quantitative measures of a particular construct, usually not observed directly. Four basic types of scales categorize measurements used in management:

- Ratio
- Interval
- Rank order or ordinal
- Categorical or nominal

1.3.2 The Impact of the Type of Scale on Statistical Analysis

The nature of analysis depends, in particular, on the scale of the variable(s). Table 1.1 summarizes the most frequently used statistics which are permissible according to the scale type. The order of the scales in the table from Nominal to Ratio is hierarchical, in the sense that statistics which are permissible for a scale above are also permissible for the scale in question. For example, a median is a legitimate statistic for an ordinal scale variable and is also legitimate for an interval

Table 1.1 Scales of measurement and their properties

Scale	Mathematical group structure	Permissible statistics	Typical examples
Nominal	Permutation group $y = f(x)$ [$f(x)$ means any one-to-one correspondence]	• Frequency distribution • Mode	• Numbering of brands • Assignment of numbers to type of products or models • Gender of consumers • Organization types
Ordinal	Isotonic group $y = f(x)$ [$f(x)$ means any increasing monotonic function]	• Median • Percentiles • Order (Spearman) correlations • Sign test	• Order of entry • Rank order of preferences
Interval	General linear group $y = a + bx$ $b > 0$	• Mean • Average deviation • Standard deviation • Product–moment correlation • t test • F test	• Likert scale items (agree–disagree) • Semantic scale items (ratings on opposite adjectives)
Ratio	Similarity group $y = cx$ $c > 0$	• Geometric mean • Coefficient of variation	• Sales • Market share • Advertising expenditures

Sources: Adapted from Stevens (1962), p. 25; Stevens (1959), p. 27; and Green and Tull (1970), p. 181.

or ratio scale. The reverse is not true; for example, a mean is not legitimate for an ordinal scale.

1.4 Topics Covered

This book covers the major methods of analysis that have been used in the recent management research literature. A survey of the major journals in the various fields of management was conducted to identify these methods. This survey revealed interesting observations.

It is striking that the majority of the analyses involve the estimation of a single equation or of several equations independently of one another. Analyses involving a system of equations represent a very small percentage of the analyses reported in these articles. This appears, at first sight, surprising given the complexity of management phenomena. Possibly, some of the simultaneous relationships analyzed are reflected in methodologies that consider explicitly measurement errors; these techniques appear to have advanced over the recent years. This is why the methodologies used for measurement modeling receive special attention in the book. Factor analysis is a fundamental method found in a significant number of studies, typically to verify the unidimensionality of the constructs measured. The more advanced aspects such as second-order factor analysis and multiple group factor analysis have gained popularity and have also been discussed. Choice modeling has been an important topic, especially in the field of Marketing and also in the other fields of Management, with studies estimating probit or logit models. Still, a very small percentage of articles use these models for ordered choice data (i.e., where the data reflects only the order in which brands are preferred from best to worse). Analysis of proximity data concerns few studies, but cluster analysis and multidimensional scaling remain the favorite methods for practice analysts.

Therefore, the following topics were selected. They have been classified according to the type of the key variable or variables which is or are the center of the interest in the analysis. Indeed, as discussed in Chapter 2, the nature of the criterion (also called dependent or endogenous) variable(s) determines the type of statistical analysis that may be performed. Consequently, the first issue to be discussed concerns the nature and properties of variables and the process of generating scales with the appropriate statistical procedures, subsequently followed by the various statistical methods of data analysis.

Introduction to Multivariate Statistics and Tests About Means

- Multivariate Analysis of Variance

Multiple Item Measures

- Reliability Alpha
- Principal Component Analysis
- Exploratory Factor Analysis

- Confirmatory Factor Analysis
- Second-Order Factor Analysis
- Multigroup Factor Analysis

Canonical Correlation Analysis
Single Equation Econometrics

- Ordinary Least Squares
- Generalized Least Squares
- Tests of Homogeneity of Coefficients: Pooling Tests

System of Equations Econometrics

- Seemingly Unrelated Regression
- Two-Stage Least Squares
- Three-Stage Least Squares

Categorical Dependent Variables

- Discriminant Analysis
- Quantal Choice Models: Logit

Rank-Ordered Data

- Conjoint Analysis
- Ordered Probit

Analysis of Covariance Structure
Analysis of Similarity Data

- Cluster Analysis
- Multidimensional Scaling

1.5 Pedagogy

There are three key learning experiences necessary to be able to achieve these objectives:

1. Sufficient knowledge of statistical theory to be able to understand the methodologies, when they are applicable, and when they are not appropriate.
2. An ability to perform such analyses with proper statistical software.
3. An understanding of how these methodologies have been applied in management research.

This book differs from others in that no other book on multivariate statistics or data analysis addresses the specific needs of doctoral education. The three aspects mentioned above are weighted differently. This book emphasizes the first aspect of the methodology by providing the mathematical and statistical analyses necessary to fully understand them. This can be contrasted with other books that prefer primarily or exclusively a verbal description of the method.

This book favors the understanding of the rationale for modeling choices, issues, and problems. While the verbal description of a method may be better accessible to a wider audience, it is often more difficult to follow the rationale, which is based on mathematics. For example, it is difficult to understand the problem of multicollinearity without understanding the effect on the determinant of the covariance matrix, which needs to be inverted. The learning that results from verbal presentation tends, therefore, to be more mechanical.

This book also differs in that, instead of choosing a few articles to illustrate the applications of the methods, as would be found in a book of readings (sometimes with short introductions), a list of application articles is provided from which the reader can choose. Articles tend to be relatively easy to access, especially with services available through the Internet. The list of references covers a large cross section of examples and a history of the literature in this domain.

Finally, the examples of analyses are relatively self-explanatory and, although some explanations of the statistical software used are provided with each example, this book does not intend to replace the instruction manuals of those particular software packages. The reader is referred to those for details.

In summary, this book focuses on the first aspect of understanding the statistical methodology while providing enough information to the reader for developing skills in performing the analyses and in understanding how to apply them to management research problems.

More specifically, the learning of this material involves two parts: learning of the statistical theory fundamental to the technique and learning of how to use the technique. Although there may be different ways to combine these two experiences, it is recommended to first learn the theory by reading the sections in which the methodologies are presented and discussed. Then, the statistical computer package (e.g., SAS, LIMDEP, LISREL, and other specialized packages) used to apply the methodology is presented in the context of an example. Students can then apply the technique using the data sets available from the personal page of Hubert Gatignon at http://www.insead.edu/. Finally, application issues can be illustrated by other applications found in prior research and listed at the end of each chapter.

In addition to the books and articles included with each chapter, the following books are highly recommended to develop further one's skills in different methods of data analysis. Each of these books is highly specialized and covers only a subset of the methods presented in this book. However, they are indispensable complements to gain proficiency in the techniques used in research.

Bibliography

Green, Paul E. and Donald S. Tull (1970), *Research for Marketing Decisions*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Greene, William H. (1993), *Econometric Analysis*, New York: MacMillan Publishing Company.

Hanssens, Dominique M., Leonard J. Parsons and Randall L. Shultz (1990), *Market Response Models: Econometric and Time Series Analysis*, Norwell, MA: Kluwer Academic Publishers.

Judge, George G., William E. Griffiths, R. Carter Hill, Helmut Lutkepohl and Tsoung-Chao Lee (1985), *The Theory and Practice of Econometrics*, New York: John Wiley and Sons.

Stevens, Stanley S. (1959), "Measurement, Psychophysics and Utility," in C. W. Churchman and P. Ratoosh, eds., *Measurement: Definitions and Theories*, New York, NY: John Wiley and Sons, Inc.

Stevens, Stanley S. (1962), "Mathematics, Measurement and Psychophysics," in S. S. Stevens, ed., *Handbook of Experimental Psychology*, New York, NY: John Wiley and Sons, Inc.

Chapter 2
Multivariate Normal Distribution

In this chapter, we define univariate and multivariate normal distribution density functions and then we discuss tests of differences of means for multiple variables simultaneously across groups.

2.1 Univariate Normal Distribution

Just to refresh memory, in the case of a single random variable, the probability distribution or density function of that variable x is represented by Equation (2.1):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.1}$$

2.2 Bivariate Normal Distribution

The bivariate distribution represents the joint distribution of two random variables. The two random variables x_1 and x_2 are related to each other in the sense that they are not independent of each other. This dependence is reflected by the correlation ρ between the two variables x_1 and x_2. The density function for the two variables jointly is

$$\Phi(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\} \tag{2.2}$$

This function can be represented graphically as in Fig. 2.1:

The *Isodensity contour* is defined as the set of points for which the values of x_1 and x_2 give the same value for the density function Φ. This contour is given by Equation (2.3) for a fixed value of C, which defines a constant probability:

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} = C \tag{2.3}$$

H. Gatignon, *Statistical Analysis of Management Data*,
DOI 10.1007/978-1-4419-1270-1_2,

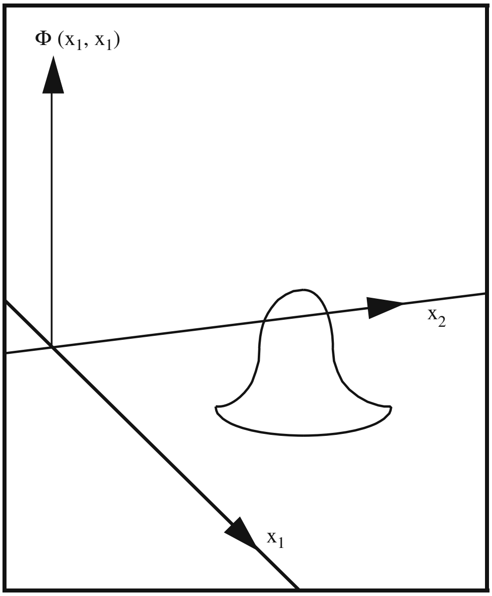

Fig. 2.1 The bivariate normal distribution

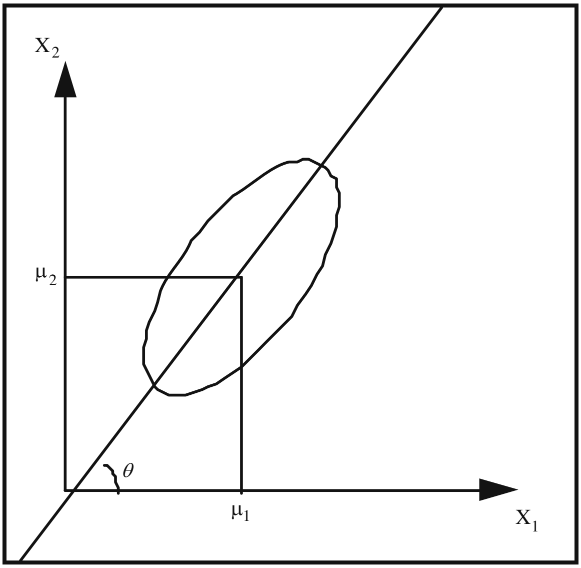

Fig. 2.2 The locus of points of the bivariate normal distribution at a given density level

Equation (2.3) defines an ellipse with centroid (μ_1, μ_2). This ellipse is the locus of points representing the combinations of the values of x_1 and x_2 with the same probability, as defined by the constant C (Fig. 2.2).

For various values of C, we get a family of concentric ellipses (at a different cut, i.e., cross section of the density surface with planes at various elevations) (see Fig. 2.3).

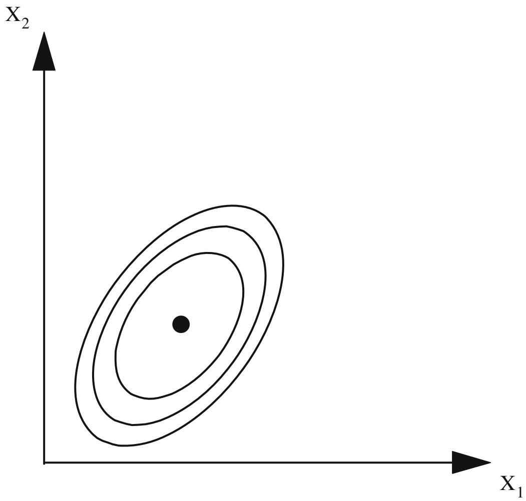

Fig. 2.3 Concentric ellipses at various density levels

The angle θ depends only on the values of σ_1, σ_2, and ρ but is independent of C. The higher the correlation between x_1 and x_2, the steeper the line going through the origin with angle θ, i.e., the bigger the angle.

2.3 Generalization to Multivariate Case

Let us represent the bivariate distribution in matrix algebra notation in order to derive the generalized format for more than two random variables.

The covariance matrix of (x_1, x_2) can be written as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{2.4}$$

The determinant of the matrix Σ is

$$|\Sigma| = \sigma_1^2\sigma_2^2\left(1 - \rho^2\right) \tag{2.5}$$

Equation (2.3) can now be re-written as

$$C = [x_1 - \mu_1, x_2 - \mu_2]\,\Sigma^{-1}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{2.6}$$

where

$$\Sigma^{-1} = 1/\left[\sigma_1^2\sigma_2^2(1 - \rho^2)\right]\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1 - \rho^2}\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \tag{2.7}$$

Note that $\Sigma^{-1} = |\Sigma|^{-1} \times$ matrix of cofactors.
Let

$$\mathbf{X} = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

then $\mathbf{X}'\Sigma^{-1}\mathbf{X} = \chi^2$, which is a quadratic form of the variables $\mathbf{x}$ and is, therefore, a chi-square variate.

Also, because $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$, $|\Sigma|^{1/2} = \sigma_1\sigma_2\sqrt{(1-\rho^2)}$, and consequently,

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = (2\pi)^{-1}|\Sigma|^{-1/2} \tag{2.8}$$

the bivariate distribution function can be now expressed in matrix notation as

$$\Phi(x_1, x_2) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{X}'\Sigma^{-1}\mathbf{X}} \tag{2.9}$$

Now, more generally with p random variables $(x_1, x_2, \ldots, x_p)$, let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

The density function is

$$\Phi(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-\frac{1}{2}} e^{\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]} \tag{2.10}$$

For a fixed value of the density Φ, an ellipsoid is described. Let $\mathbf{X} = \mathbf{x} - \mu$. The inequality $\mathbf{X}'\Sigma^{-1}\mathbf{X} \leq \chi^2$ defines any point within the ellipsoid.

2.4 Tests About Means

2.4.1 Sampling Distribution of Sample Centroids

2.4.1.1 Univariate Distribution

A random variable is normally distributed with mean μ and variance σ^2:

$$x \sim N\left(\mu, \sigma^2\right) \tag{2.11}$$

After n independent draws, the mean is randomly distributed with mean μ and variance σ^2/n:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{2.12}$$

2.4.1.2 Multivariate Distribution

In the multivariate case with p random variables, where $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, $\mathbf{x}$ is normally distributed following the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ:

$$\mathbf{x} \sim N(\mu, \Sigma) \tag{2.13}$$

The mean vector for the sample of size n is denoted by

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

This sample mean vector is normally distributed with a multivariate normal distribution with mean μ and covariance Σ/n:

$$\bar{\mathbf{x}} \sim N\left(\mu, \frac{\Sigma}{n}\right) \tag{2.14}$$

2.4.2 Significance Test: One-Sample Problem

2.4.2.1 Univariate Test

The univariate test is illustrated in the following example. Let us test the hypothesis that the mean is 150 (i.e., $\mu_0 = 150$) with the following information:

$$\sigma^2 = 256;\ n = 64;\ \bar{x} = 154$$

Then, the z score can be computed as

$$z = \frac{154 - 150}{\sqrt{256/64}} = \frac{4}{16/8} = 2$$

At $\alpha = 0.05$ (95% confidence interval), $z = 1.96$, as obtained from a normal distribution table. Therefore, the hypothesis is rejected. The confidence interval is

$$\left[154 - 1.96 \times \frac{12}{6}, 154 + 1.96 \times \frac{12}{6}\right] = [150.08, 157.92]$$

This interval excludes 150. The hypothesis that $\mu_0 = 150$ is rejected. If the variance σ had been unknown, the t statistic would have been used:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{2.15}$$

where s is the observed sample standard deviation.

2.4.2.2 Multivariate Test with Known Σ

Let us take an example with two random variables:

$$\Sigma = \begin{bmatrix} 25 & 10 \\ 10 & 16 \end{bmatrix} \qquad n = 36$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 20.3 \\ 12.6 \end{bmatrix}$$

The hypothesis is now about the mean values stated in terms of the two variables jointly:

$$H: \quad \mu_0 = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$$

At the alpha level of 0.05, the value of the density function can be written as below, which follows a chi-squared distribution at the specified significance level α:

$$n\,(\mu_\circ - \bar{\mathbf{x}})'\,\Sigma^{-1}\,(\mu_0 - \bar{\mathbf{x}}) \sim \chi_p^2\,(\alpha) \tag{2.16}$$

Computing the value of the statistics,

$$|\Sigma| = 25 \times 16 - 10 \times 10 = 300$$

$$\Sigma^{-1} = \frac{1}{300}\begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix}$$

$$\chi^2 = 36 \times \frac{1}{300}\,(20 - 20.3,\ 15 - 12.6)\begin{bmatrix} 16 & -10 \\ -10 & 25 \end{bmatrix}\begin{bmatrix} 20 - 20.3 \\ 15 - 12.6 \end{bmatrix} = 15.72$$

The critical value at an alpha value of 0.05 with two degrees of freedom is provided by tables:

$$\chi_{p=2}^2\,(\alpha = 0.05) = 5.991$$

The observed value is greater than the critical value. Therefore, the hypothesis that $\mu = \begin{bmatrix} 20 \\ 15 \end{bmatrix}$ is rejected.

2.4.2.3 Multivariate Test with Unknown Σ

Just as in the univariate case, Σ is replaced with the sample value $\mathbf{S}/(n-1)$, where **S** is the sum-of-squares-and-cross-products (SSCP) matrix, which provides an unbiased estimate of the covariance matrix. The following statistics are then used to test the hypothesis:

$$\text{Hotelling: } T^2 = n\,(n-1)\,(\bar{\mathbf{x}} - \mu_0)'\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \mu_0) \tag{2.17}$$

where, if

$$\underset{n\times p}{\mathbf{X}^d} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots \\ \vdots & \vdots & \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots \end{bmatrix}$$

$$\mathbf{S} = \mathbf{X}^{d'}\mathbf{X}^d$$

Hotelling showed that

$$\frac{n-p}{(n-1)\,p}T^2 \sim F^p_{n-p} \tag{2.18}$$

Replacing T^2 by its expression given above

$$\frac{n\,(n-p)}{p}\,(\bar{\mathbf{x}} - \mu_0)'\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \mu_0) \sim F^p_{n-p} \tag{2.19}$$

Consequently, the test is performed by computing the expression above and comparing its value with the critical value obtained in an F table with p and $n-p$ degrees of freedom.

2.4.3 Significance Test: Two-Sample Problem

2.4.3.1 Univariate Test

Let us define $\bar{x}_1$ and $\bar{x}_2$ as the means of a variable on two unrelated samples. The test for the significance of the difference between the two means is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{or} \quad t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2\left(\frac{n_1+n_2}{n_1 n_2}\right)} \tag{2.20}$$

where

$$s = \frac{\sqrt{(n_1-1)\frac{\sum_i x^2_{1i}}{n_1-1} + (n_2-1)\frac{\sum_i x^2_{2i}}{n_2-1}}}{(n_1-1)+(n_2-1)} = \sqrt{\frac{\sum_i x^2_{1i} + \sum_i x^2_{2i}}{n_1+n_2-2}} \tag{2.21}$$

s^2 is the pooled within groups variance. It is an estimate of the assumed common variance σ^2 of the two populations.

2.4.3.2 Multivariate Test

Let $\bar{\mathbf{x}}^{(1)}$ be the mean vector in sample $1 = \begin{bmatrix} \bar{x}_1^{(1)} \\ \bar{x}_2^{(1)} \\ \vdots \\ \bar{x}_p^{(1)} \end{bmatrix}$ and similarly for sample 2.

We need to test the significance of the difference between $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$. We will consider first the case where the covariance matrix, which is assumed to be the same in the two samples, is known. Then we will consider the case where an estimate of the covariance matrix needs to be used.

Σ Is Known (The Same in the Two Samples)

In this case, the difference between the two group means is normally distributed with a multivariate normal distribution:

$$\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \sim N\left(\mu_1 - \mu_2, \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \tag{2.22}$$

The computations for testing the significance of the differences are similar to those in Section 2.4.2.2 using the chi-square test.

Σ Is Unknown

If the covariance matrix is not known, it is estimated using the covariance matrices within each group but pooled.

Let $\mathbf{W}$ be the within-groups SSCP (sum of squares cross products) matrix. This matrix is computed from the matrix of deviations from the means on all p variables for each of n_k observations (individuals). For each group k,

$$\underset{n_k \times p}{\mathbf{X}}^{d(k)} = \begin{bmatrix} x_{11}^{(k)} - \bar{x}_1^{(k)} & x_{21}^{(k)} - \bar{x}_2^{(k)} & \cdots \\ x_{12}^{(k)} - \bar{x}_1^{(k)} & x_{22}^{(k)} - \bar{x}_2^{(k)} & \cdots \\ \vdots & \vdots & \\ x_{1n_k}^{(k)} - \bar{x}_1^{(k)} & x_{2n_k}^{(k)} - \bar{x}_2^{(k)} & \cdots \end{bmatrix} \tag{2.23}$$

For each of the two groups (each k), the SSCP matrix can be derived:

$$\mathbf{S}_k = \mathbf{X}^{d(k)'}_{p \times n_k} \mathbf{X}^{d(k)}_{n_k \times p} \tag{2.24}$$

The pooled SSCP matrix for the more general case of K groups is simply:

$$\underset{p \times p}{\mathbf{W}} = \sum_{k=1}^{K} \underset{p \times p}{\mathbf{S}_k} \tag{2.25}$$

In the case of two groups, K is simply equal to 2.

Then, we can apply Hotelling's T, just as in Section 2.4.2.3, where the proper degrees of freedom depending on the number of observations in each group (n_k) are applied.

$$T^2 = \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)' \left[\frac{\mathbf{W}}{n_1 + n_2 - 2} \frac{n_1 + n_2}{n_1 n_2}\right]^{-1} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \tag{2.26}$$

$$= \frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)' \mathbf{W}^{-1} \left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \tag{2.27}$$

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)\, p} T^2 \sim F^p_{n_1 + n_2 - p - 1} \tag{2.28}$$

2.4.4 Significance Test: K-*Sample Problem*

As in the case of two samples, the null hypothesis is that the mean vectors across the K groups are the same and the alternative hypothesis is that they are different.

Let us define Wilk's likelihood-ratio criterion:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{2.29}$$

where $\mathbf{T}$ = total SSCP matrix, $\mathbf{W}$ = within-groups SSCP matrix.

W is defined as in Equation (2.25). The total SSCP matrix is the sum of squared cross products applied to the deviations from the grand means (i.e., the overall mean across the total sample with the observations of all the groups for each variable). Therefore, let the mean centered data for group k be noted as

$$\underset{n_k \times p}{\mathbf{X}}^{d^*(k)} = \begin{bmatrix} x^{(k)}_{11} - \bar{x}_1 & x^{(k)}_{21} - \bar{x}_2 & \cdots \\ x^{(k)}_{12} - \bar{x}_1 & x^{(k)}_{22} - \bar{x}_2 & \cdots \\ \vdots & \vdots & \\ x^{(k)}_{1n_k} - \bar{x}_1 & x^{(k)}_{2n_k} - \bar{x}_2 & \cdots \end{bmatrix} \tag{2.30}$$

where $\bar{x}_j$ is the overall mean of the j's variate.