Cairo University
Faculty of Economics and Political Science
Department of Statistics

On Missing Values in Classification Analysis

By

Mahmoud Al.Said Mahmoud

Under the Supervision of

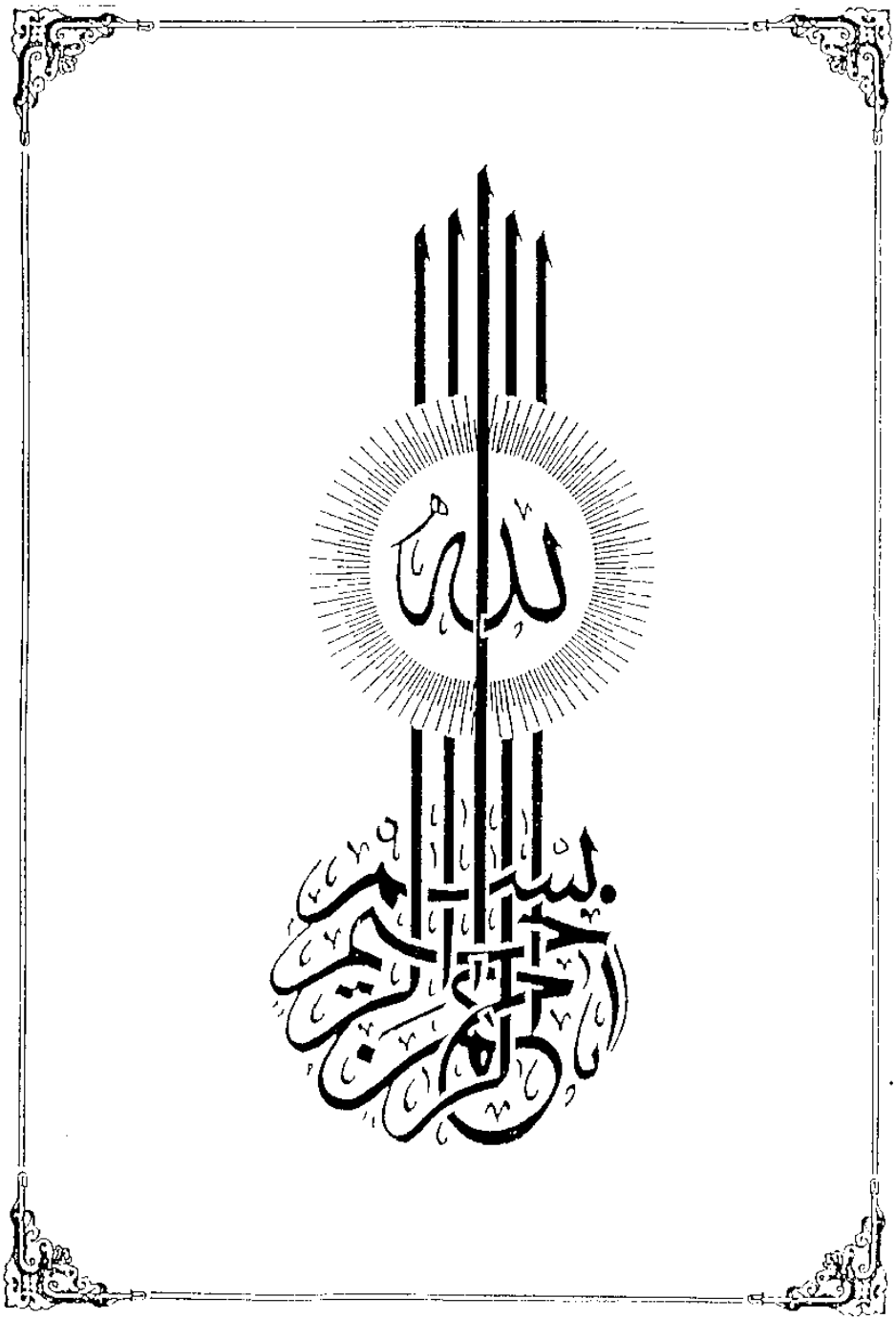
Prof. Adel M. Zaher

Department of Statistics
Faculty of Economics
and Political Science
Cairo University

A.Prof. Ahmed H. Haroun

Department of Applied Statistics
Institute of Statistical
Studies and Research
Cairo University

A Thesis submitted to the Faculty of Economics and Political Science (Department of Statistics) in partial fulfillment of the requirement for
The M.Sc. Degree in Statistics
1997



Acknowledgment

I am indebted to Prof. Adel Zaher for the considerable help, encouragement and valuable advice he has given me through his kind supervision of this dissertation. I would also like to express my gratitude to Dr. Ahmed Haroun for his effort and advice through the duration of this study. Their help and encouragement were particularly important to me and provided a constant source of inspiration.

Finally, I would like to give special thanks to my wife and my family for their support and prayers at all times.

Mahmoud Al Said.

Abstract

Classification Analysis is concerned with the problem of classifying a subject to one of several distinct groups on the basis of a set of measurements. For example, in business, a bank loan officer wishes to classify loan applicants to low risk credit customers or high risk credit customers on the basis of a set of variables. The most well-known and widely used rules of classification are presented through this thesis.

The presence of missing values in a data set used for building a classification rule is a serious problem that may face the investigator when applying the classification analysis (CA) to a practical situation. Many procedures have been developed to handle the missing values when applying (CA).The default method of handling missing values in (CA) used by many statistical packages (for example, SAS, Minitab and SPSS) is to omit all units containing missing values. Thus, considerable information may be lost due to the reduction of the sample size. The most commonly used procedures for this purpose are displayed in a separate chapter.

Many studies dealt with this problem in case of two multivariate populations with equal covariance matrices while a few studies treated it in case of two multivariate populations with unequal covariance matrices. The present study deals with the problem of classification analysis with missing values in case of two or more multivariate normal populations with equal and unequal covariance matrices through a simulation study. Three rules of classification and five methods of handling missing values are considered. The objective of this study is to compare the different

methods of handling missing values with respect to their ability in obtaining a "good" classification rule. In this thesis, two patterns of missing values are considered and the mechanism that lead to the presence of missing values is assumed to be missing at random (MAR).

Seven factors are taken into consideration. These factors are the sample size, the number of populations, the number of variables, the distance between the populations, the covariance matrices (equal or unequal), the pattern of missing values and the percentage of missing values. The impact of each factor on the methods of handling missing values is studied. A Minitab macro was designed to run the necessary calculations. This macro is displayed in a separate appendix.

Table of Contents

Chapter (1) :

Introduction.