UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Statistical analysis of grouped data

by

Gretel Crafford

**Submitted in partial fulfillment of
the requirements for the degree**

Philosphiae Doctor

**in the
Faculty of Natural & Agricultural Sciences
University of Pretoria
Pretoria**

February 2007

Declaration

I declare that the thesis that I hereby submit for the degree Philosphiae Doctor at the University of Pretoria has not previously been submitted by me for degree purposes at any other university.

Signature ______________________ Date ______________________

Acknowledgements

I am grateful to Professor NAS Crowther for encouraging me to embark on this study. His guidance played a major role in the accomplishment of this challenging research.

I would also like to thank Professor CF Smit for all the valuable discussions and for suggesting helpful improvements.

I am grateful for a grant awarded to me by the University of Pretoria, which provided financial assistance required to complete this research.

To my colleagues, my sincere appreciation for your help and understanding.

Finally, I would like to thank my family for their love, support and encouragement.

Summary

The maximum likelihood (ML) estimation procedure of *Matthews and Crowther* (1995: *A maximum likelihood estimation procedure when modelling in terms of constraints.* South African Statistical Journal, 29, 29-51) is utilized to fit a continuous distribution to a grouped data set. This grouped data set may be a single frequency distribution or various frequency distributions that arise from a cross classification of several factors in a multifactor design. It will also be shown how to fit a bivariate normal distribution to a two-way contingency table where the two underlying continuous variables are jointly normally distributed. This thesis is organized in three different parts, each playing a vital role in the explanation of analysing grouped data with the ML estimation of *Matthews and Crowther*.

In Part I the ML estimation procedure of *Matthews and Crowther* is formulated. This procedure plays an integral role and is implemented in all three parts of the thesis. In Part I the exponential distribution is fitted to a grouped data set to explain the technique. Two different formulations of the constraints are employed in the ML estimation procedure and provide identical results. The justification of the method is further motivated by a simulation study. Similar to the exponential distribution, the estimation of the normal distribution is also explained in detail. Part I is summarized in Chapter 5 where a general method is outlined to fit continuous distributions to a grouped data set. Distributions such as the Weibull, the log-logistic and the Pareto distributions can be fitted very effectively by formulating the vector of constraints in terms of a linear model.

In Part II it is explained how to model a grouped response variable in a multifactor design. This multifactor design arise from a cross classification of the various factors or independent variables to be analysed. The cross classification of the factors results in a total of T cells, each containing a frequency distribution. Distribution fitting is done simultaneously to each of the T cells of the multifactor design. Distribution fitting is also done under the additional constraints that the parameters

of the underlying continuous distributions satisfy a certain structure or design. The effect of the factors on the grouped response variable may be evaluated from this fitted design. Applications of a single-factor and a two-factor model are considered to demonstrate the versatility of the technique.

A two-way contingency table where the two variables have an underlying bivariate normal distribution is considered in Part III. The estimation of the bivariate normal distribution reveals the complete underlying continuous structure between the two variables. The ML estimate of the correlation coefficient ρ is used to great effect to describe the relationship between the two variables. Apart from an application a simulation study is also provided to support the method proposed.

Contents