

Statistical analysis of grouped data

by

Gretel Crafford

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

Philosophiae 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 Philosophiae 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

1	Introduction	1
I	Fitting distributions to grouped data	3
2	The ML estimation procedure	4
2.1	Formulation	4
2.2	Estimation	6
2.3	Goodness of fit	8
3	The exponential distribution	9
3.1	Direct set of constraints	10
3.2	Constraints in terms of a linear model	15
3.3	Simulation study	19
4	The normal distribution	21
4.1	Direct set of constraints	22
4.2	Constraints in terms of a linear model	28

4.3	Simulation study	32
5	The Weibull, log-logistic and Pareto distributions	35
5.1	The Weibull distribution	35
5.2	The log-logistic distribution	38
5.3	The Pareto distribution	41
5.4	Generalization	43
II	Linear models for grouped data	48
6	Multifactor design	49
6.1	Formulation	50
6.2	Estimation	53
7	Normal distributions	56
7.1	Estimation of distributions	56
7.2	Equality of variances	62
7.3	Multifactor model	64
7.4	Application: Single-factor model	66
7.4.1	Model 1: Unequal variances	67
7.4.2	Model 2: Equal variances	72
7.4.3	Model 3: Ordinal factor	74
7.4.4	Model 4: Regression model	77

8	Log-logistic distributions	81
8.1	Estimation of distributions	82
8.2	Multifactor model	85
8.3	Application: Two-factor model	88
8.3.1	Model 1: Saturated model	89
8.3.2	Model 2: No interaction model	98
8.3.3	Model 3: Regression model with no interaction	103
8.3.4	Model 4: Regression model with interaction	109
III	Bivariate normal distribution	115
9	Bivariate grouped data	116
9.1	Formulation	117
9.2	Estimation	119
10	The bivariate normal distribution	120
10.1	Joint distribution	120
10.2	Marginal distributions	121
10.3	Standard bivariate normal distribution	121
10.4	Conditional distributions	123
10.5	Bivariate normal probabilities	124
10.5.1	Calculation of bivariate normal probabilities	124

10.5.2 Calculation of ρ 128

11 Estimating the bivariate normal distribution 132

11.1 Bivariate normal probabilities 132

11.2 Parameters 135

11.2.1 Marginal distribution of \mathbf{x} 135

11.2.2 Marginal distribution of \mathbf{y} 136

11.2.3 Joint distribution of \mathbf{x} and \mathbf{y} 137

11.3 Vector of constraints 138

11.3.1 Marginal distribution of \mathbf{x} 139

11.3.2 Marginal distribution of \mathbf{y} 139

11.3.3 Joint distribution of \mathbf{x} and \mathbf{y} 140

11.4 Matrix of Partial Derivatives 141

11.4.1 Marginal distribution of \mathbf{x} 141

11.4.2 Marginal distribution of \mathbf{y} 142

11.4.3 Joint distribution of \mathbf{x} and \mathbf{y} 143

11.5 Iterative procedure 149

11.6 ML estimates 150

11.6.1 ML estimates of the natural parameters 151

11.6.2 ML estimates of the original parameters 152

11.7 Goodness of fit 153

12 Application	154
12.1 ML estimation procedure	155
12.1.1 Unrestricted estimates	157
12.1.2 ML estimates	162
13 Simulation study	168
13.1 Theoretical distribution	169
IV	172
14 Résumé	173
V Appendix	178
A SAS programs: Part I	179
A.1 EXP1.SAS	179
A.2 EXP2.SAS	180
A.3 EXPSIM.SAS	182
A.4 NORM1.SAS	184
A.5 NORM2.SAS	185
A.6 NORMSIM.SAS	187
A.7 FIT.SAS	190



B SAS programs: Part II	195
B.1 FACTOR1.SAS	195
B.2 FACTOR2.SAS	200
C SAS Programs: Part III	206
C.1 Phi0.SAS	206
C.2 Phi.SAS	207
C.3 BVN.SAS	209
C.4 BVNSIM.SAS	218