Chapter 13

Simulation study

The purpose of this sumulation study is to prove that a bivariate normal distribution can be fitted accurately to a two-way contingency table by employing the ML estimation procedure presented in Part III of this thesis. A total of 1000 samples were simulated from a bivariate normal distribution such that

$$(x, y) \sim \text{BVN}\left(11, 48, 3^2, 8^2, -0.7\right) \ .$$

Each of the data sets consisted of 1000 observations and the descriptive statistics for the sample statistics are listed in Table 13.1. From Table 13.1 it can be concluded that the sample statistics of the simulated data sets correspond very well to the theoretical values.

Table 13.1: Descriptive statistics for the sample statistics.

Stat	Mean	Std.dev	P_{05}	Median	P_{95}
$\overline{x}$	11.008	0.0957	10.849	11.008	11.157
s_x	2.9972	0.0655	2.887	2.998	3.110
$\overline{y}$	47.978	0.2620	47.550	47.970	48.403
s_y	7.9952	0.1765	7.703	7.994	8.291
r	-0.6999	0.0163	-0.7273	-0.7000	-0.6734

The next step will be to cross tabulate each of the bivariate data sets into a two-way contingency table and to fit a bivariate normal distribution to each of the 1000 bivariate grouped data sets. This simulation study was done with of the SAS program *BVNSIM.SAS* listed in the Appendix C4.

13.1 Theoretical distribution

The simulated data sets were all categorised in a two-way contingency table, with the following upper class boundaries

$$\mathbf{x} = \begin{pmatrix} 8 \\ 10 \\ 12 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 45 \\ 50 \\ 55 \end{pmatrix} .$$

The first and last class intervals, for both variables were treated as open ended class intervals and the frequency distribution for the theoretical distribution is given in Table 13.2.

Table 13.2: Theoretical frequency distribution for $\mathrm{BVN}\left(11, 48, 3^2, 8^2, -0.7\right)$ distribution.

X	**Y**				
	$(-\infty, 45)$	$[45, 50)$	$[50, 55)$	$[55, \infty)$	**Total**
$(-\infty, 8)$	4.722	18.436	41.402	94.095	158.655
$[8, 10)$	27.011	55.569	69.373	58.832	210.786
$[10, 12)$	79.095	86.607	65.581	29.834	261.117
$[12, \infty)$	243.002	84.264	34.150	8.026	369.441
Total	353.830	244.876	210.507	190.787	1000

The cumulative relative frequencies for the theoretical distribution, expressed in terms of matrix notation, is

$$\mathbf{\Pi} = \begin{pmatrix} 0.00472 & 0.02316 & 0.06456 & 0.15866 \\ 0.03173 & 0.10574 & 0.21651 & 0.36944 \\ 0.11083 & 0.27144 & 0.44780 & 0.63056 \\ 0.35383 & 0.59871 & 0.80921 & 1.00000 \end{pmatrix} . \tag{13.1}$$

The ML estimators of the 5 parameters of the bivariate normal distribution are all asymptotically normally distributed with standard errors functions of (13.1). The standard errors and percentiles of the ML estimators are listed in Table 13.3.

Table 13.3: Theoretical values for the ML estimators of the bivariate normal distribution.

ML estimate	Standard error	Margin of error	Percentiles		
			P_{05}	Median	P_{95}
$\widehat{\mu}_x$	$\sigma_{\widehat{\mu}_x} = 0.1054$	$1.645\sigma_{\widehat{\mu}_x} = 0.1733$	10.827	11	11.173
$\widehat{\sigma}_x$	$\sigma_{\widehat{\sigma}_x} = 0.1123$	$1.645\sigma_{\widehat{\sigma}_x} = 0.18466$	2.8153	3	3.1733
$\widehat{\mu}_y$	$\sigma_{\widehat{\mu}_y} = 0.2788$	$1.645\sigma_{\widehat{\mu}_y} = 0.45854$	47.541	48	48.459
$\widehat{\sigma}_y$	$\sigma_{\widehat{\sigma}_y} = 0.3065$	$1.645\sigma_{\widehat{\sigma}_y} = 0.50415$	7.4958	8	8.1733
$\widehat{\rho}$	$\sigma_{\widehat{\rho}} = 0.021085$	$1.645\sigma_{\widehat{\rho}} = 0.03468$	-0.7347	-0.7	-0.6653

The descriptive statistics for the ML estimates of the 1000 fitted bivariate normal distributions are summarised in Table 13.4.

Table 13.4: Simulation results of 1000 fitted bivariate normal distributions.

MLE	Theoretical Value	Mean	Std.dev	P_{05}	Median	P_{95}
$\widehat{\mu}_x$	11	11.010	0.1042	10.842	11.008	11.178
$\widehat{\sigma}_{\widehat{\mu}_x}$	0.1054	0.1055	0.0045	0.0980	0.1055	0.1130
$\widehat{\sigma}_x$	3	3.0007	0.1166	2.8063	3.0006	3.1978
$\widehat{\sigma}_{\widehat{\sigma}_x}$	0.1123	0.1125	0.0066	0.1017	0.1124	0.1238
$\widehat{\mu}_y$	48	47.973	0.2829	47.503	47.971	48.426
$\widehat{\sigma}_{\widehat{\mu}_y}$	0.2788	0.2788	0.0121	0.2590	0.2785	0.2996
$\widehat{\sigma}_y$	8	7.9938	0.3203	7.4700	7.9914	8.5373
$\widehat{\sigma}_{\widehat{\sigma}_y}$	0.3065	0.3066	0.0187	0.2763	0.3062	0.3387
$\widehat{\rho}$	-0.7	-0.7006	0.0243	-0.7421	-0.7002	-0.6604
$\widehat{\sigma}_{\widehat{\rho}}$	0.021085	0.0211	0.0013	0.0189	0.0211	0.0231

It is evident from Table 13.4, that the mean for all the ML estimates are remarkably close to the theoretical values. It is also interesting to note that the standard deviation of the 5 ML estimates $\widehat{\mu}_x$, $\widehat{\sigma}_x$, $\widehat{\mu}_y$, $\widehat{\sigma}_y$ and $\widehat{\rho}$ are very close to the mean of its standard errors. E.g. the standard deviation of the $\widehat{\mu}_x$-values is 0.1042 and the mean of the $\widehat{\sigma}_{\widehat{\mu}_x}$-values is 0.1055. The percentiles of the ML estimates in the simulation study (see Table 13.4) correspond extremely well to that of the theoretical distribution given in Table 13.3.

A comparison between the descriptive statistics of the sample statistics of the ungrouped bivariate data sets in Table 13.1 with that of the descriptive statistics of the ML estimates of the grouped data sets tabulated in Table 13.4 shows are very close to each other. This motivates that not too much accuracy is being lost with a grouped data set, when analysed correctly.

The Wald and Pearson goodness of fit statistics were calculated for each of the 1000 estimated bivariate normal distributions. The percentiles of these two statistics are tabulated in Table 13.5 and agrees with a χ^2-distribution with 10 degrees of freedom.

Table 13.5: Percentiles of the Pearson and Wald statistic.

	Percentiles						
	P_5	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	P_{95}
Pearson	3.8374	4.8481	7.1377	9.8363	13.3152	16.7631	18.9273
Wald	4.0572	5.2029	7.6182	10.6859	14.6539	19.3063	23.5933
	Percentiles of a χ^2-distribution with 10 degrees of freedom.						
	$\chi^2_{0.05}$	$\chi^2_{0.10}$	$\chi^2_{0.25}$	$\chi^2_{0.50}$	$\chi^2_{0.75}$	$\chi^2_{0.90}$	$\chi^2_{0.95}$
$\chi^2\,(10)$	3.9403	4.8652	6.7372	9.3418	12.5489	15.9872	18.3070

It can therefore be concluded that the empirical and theoretical distributions of the Pearson and Wald statistics correspond to each other.

Part IV

Chapter 14

Résumé

The main objective of this research is to provide a theoretical foundation for analysing grouped data, taking the underlying continuous nature of the variable(s) into account. Statistical techniques have been developed and applied extensively for continuous data, but the analysis for grouped data has been somewhat neglected. This creates numerous problems especially in the social and economic disciplines, where variables are grouped for various reasons. Due to a lack for the appropriate statistical techniques to evaluate grouped data, researchers are often tempted to ignore the underlying continuous nature of the data and employ e.g. the class midpoint values as an alternative. This leads to an oversimplification of the problem and valuable information in the data is being ignored.

The first part of the thesis demonstrates how to fit a continuous distribution to a grouped data set. By implementing the 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) the ML estimates of the parameters are obtained. The standard errors of the ML estimates are derived from the multivariate delta theorem. It is interesting to note that not much accuracy has been lost by grouping the data, justifying that statistical inference may be done effectively from a grouped data set. The main concern of this part of the thesis was to foster the basic principles. The examples and distributions discussed are merely used to illustrate and explain the philosophy from basic principles. The fit of various other continuous distributions, not mentioned in the thesis, such as the gamma distribution and the lognormal distribution can also be done using the same approach.

The second part of the thesis concentrates on the analysis of generalised linear models where the response variable is presented in grouped format. A cross classification of the independent variables leads to various so-called cells, each containing a frequency distribution of the response variable. Due to the nature of the response variable the usual analysis of variance and covariance models etc. can no longer be applied in the usual sense. A completely new approach, where a specified underlying continuous distribution for the grouped variable is fitted to each cell in the multifactor design is introduced. Certain measures such as the average, median or even any other percentile of the fitted distributions are modelled to explain the influence of the independent variables on the response variable. This evaluation may be done by means of a saturated model where no additional constraints are employed in the ML estimation procedure or by means of any other model where certain structures with regard to the independent variables are incorporated. The main objective is ultimately to provide a satisfactory model that describes the data as effectively as possible, revealing the various trends in the data. Employing the multivariate delta theorem, the standard errors for the ML estimates are calculated, enabling testing of relevant hypotheses. The goodness of fit of the model is evaluated with the Pearson and Wald statistics.

Two applications of multi-factor models are presented. In the first application normal distributions are fitted to the cells in a single factor design. The behavior of the mean of the fitted normal distributions revealed the effect of the single independent variable. Various models are employed to explain the versatility of the technique. Apart from the single factor model a two factor model was employed for data from short term insurance. The positive skewness of the grouped response variable suggested that a log-logistic distribution is to be fitted to the data. The median of the log-logistic distributions was modelled in a two factor model to explain the effect of the independent variable on the response variable. It is also illustrated how to incorporate a grouped independent variable as a covariate or regressor in the model. In the past where researchers might have been restricted to tabulations and graphical representations it is now shown that the possibilities of modelling a grouped response variable in a generalised model are in principle unlimited. The application of a three factor model or any higher order model follows similarly. A typical example pursue from the population census data where the grouped variable income can be explained utilising independent variables such as gender, province, population group, age, education level, occupation, etc.

A final intriguing contribution, given in the third part, is the fit of a bivariate normal distribution to a two-way contingency table. In the case where the underlying distribution of two grouped response variables are jointly normally distributed it is often required to investigate the association between two variables. Traditionally, classical measures such as kappa and McNemar were employed, but are limited in the sense that the complete bivariate structure between the two variables are not revealed. Since all five parameters are estimated, statistical inferences are possible with regard to the marginal as well as the partial distributions. The estimation of the parameter ρ, the correlation coefficient, explains the relationship between the two variables. The calculation of ρ is done by implementing *Sheppard's theorem on median dichotomy (1898)*, which is based on the volumes of the four quadrants of the bivariate normal distribution. It is shown that the calculation of the correlation coefficient, using the standard regression techniques, could lead to incorrect results due to the fact that the required conditions are not met. The method proposed is motivated by a simulation study.

Although various aspects of modelling grouped data are addressed in this thesis, this forms the basic building blocks for the beginning of a completely new and promising field of research with unlimited possibilities and exciting applications to be analysed.