Aspects of Categorical Data Analysis

by

Yogarani Govender

Durban, 1998.

Submitted in fulfilment of the
requirements for the degree of
Master of Science
in the
Department of Statistics
at the University of Natal

Preface

The research presented in this dissertation was supervised by Dr Glenda Matthews in the department of Statistics, at the University of Natal in Durban. I declare that the contents of this dissertation is my own work. Information that was quoted from other researcher's work have been duly referenced.

Acknowledgements

I am deeply indebted to my supervisor Dr Glenda Matthews for her assistance and directional guidance. To Jackie, thank you for the help with Latex. I want to also take this opportunity to extend my gratitude to other members of the department who have assisted me in my studies thus far.

To my parents and sisters, I am extremely thankful for your support, understanding and encouragement in my quest for academic growth. To my husband, Naren, I am grateful to you for the use of the computer and an even greater thank you for the concern, motivation and support that you have rendered at all times.

Further thanks go to my colleague, Professor Themba Dube, for helping me obtain financial assistance and to the Foundation for Research and Development for sponsoring my studies.

Abstract

The purpose of this study is to investigate and understand data which are grouped into categories. At the onset, the study presents a review of early research contributions and controversies surrounding categorical data analysis. The concept of sparseness in a contingency table refers to a table where many cells have small frequencies. Previous research findings showed that incorrect results were obtained in the analysis of sparse tables. Hence, attention is focussed on the effect of sparseness on modelling and analysis of categorical data in this dissertation.

Cressie and Read (1984) suggested a versatile alternative, the power divergence statistic, to statistics proposed in the past. This study includes a detailed discussion of the power-divergence goodness-of-fit statistic with areas of interest covering a review on the minimum power divergence estimation method and evaluation of model fit. The effects of sparseness are also investigated for the power-divergence statistic. Comparative reviews on the accuracy, efficiency and performance of the power-divergence family of statistics under large and small sample cases are presented. Statistical applications on the power-divergence statistic have been conducted in SAS (Statistical Analysis Software).

Further findings on the effect of small expected frequencies on accuracy of the X^2 test are presented from the studies of Tate and Hyer (1973) and Lawal and Upton (1976).

Other goodness-of-fit statistics which bear relevance to the sparse multinomial case are discussed. They include Zelterman's (1987) D^2 goodness-of-fit statistic, Simonoff's (1982, 1983) goodness-of-fit statistics as well as Koehler and Larntz's tests for log-linear models. On addressing contradictions for the sparse sample case under asymptotic conditions and an increase in sample size, discussions are provided on Simonoff's use of nonparametric techniques to find the variances as well as his adoption of the jackknife and bootstrap technique.

Contents

Chapter 1

Introduction

Data which is arranged into various groups and where counts are taken to determine the number of individuals belonging to each category is collectively known as categorical data.

Research output over the past 20 years highlights that the testing and defining of models for discrete multivariate data has been a popular research area. The idea of grouping data into classes and then counting group frequencies has resulted in many applications. A few examples in the sciences include : characterising survey responses (always, sometimes, never); medical reports on reactions of patients to treatment (mild, moderate, severe, remission) and in industry, the reporting of the failure of equipment under quality control tests (mechanical, electrical, no problems detected).

Initially the statistical analysis of discrete multivariate data was more concerned with model development and it was assumed that the adequacy of the model could be assessed by employing the traditional goodness-of-fit tests, example : Pearson's χ^2 or the loglikelihood ratio statistic G^2, using a chi-squared critical value.

Read and Cressie (1988) contend that this is a poor approximation. Hence they have defined and used the *power-divergence family of statistics* to analyse, compare and describe the behaviour and merits of the traditional goodness-of-fit tests.

In chapter 2, a historical outline of the χ^2 distribution is presented together with an indepth discussion of early controversies surrounding categorical data analysis.

Chapter 3 presents the necessary concepts and notation for modelling and testing discrete multivariate data. The power divergence statistic proposed by Read and Cressie is presented. Further areas of discussion include : the modelling of cross-classified data under independence; loglinear models for two and three dimensions and methods of parameter estimation.

Chapter 4 looks at the influence of sparseness assumptions on significance levels and accuracy. On discussing large sparse multinomial distributions, statistics proposed by Zelterman (1987) and Simonoff (1987) will be introduced and examined. Findings by Koehler (1986) on the use of the loglinear models under multinomial sampling in the sparse contingency table case will also be presented. Further aspects of discussion include jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. Finally, a report weighing the benefits between the Pearson χ^2 and the loglikelihood ratio statistic G^2 for sparse contingency tables sums up the chapter.

Chapter 5 deals with the effects of small expected frequencies. Research findings and recommendations from Tate and Hyer (1973) as well as Chapman (1976) will be presented. Furthermore, comparisons will be made between the X^2 and G^2 statistics under conditions of small expected frequencies.

Chapter 6 deals with the issue of small samples. The relevance of the properties that were previously discussed for large samples will now be examined for the small sample case and efficiency evaluations will be carried out.

Chapter 7 looks at applications in SAS for small samples with small cell frequencies and cases containing sampling zeros. The IML and Catmod procedures were used with sensitivity analyses being carried out on the data sets containing sampling zeros to investigate if there is a substantial difference in the parameter estimates when a zero cell is adjusted by a pre-selected small constant.

Chapter 2

The Chi-squared Distribution : Historical Documentation

The chi-square test statistic plays an important role in categorical data analysis. For this reason, the aim of this chapter is to first present a historical preview on the chi-square distribution and to thereafter show how the early developments in the chi-square distribution progressed to give greater contributions to categorical data analysis. Literature surveys disclose that the early work done on categorical data analysis was surrounded by controversy. Thus, a twentieth century tour of categorical data analysis is included to highlight the developments of categorical data analysis together with a discussion on some aspects of this controversy.

2.1 Early Contributions : The Pearson χ^2

Pascal, Fermat, Huygens, Cardano among other scientists studied games of chance and obtained many useful results. De Moivre later rewrote some of the earlier results in a more approximate form using an equation now associated with James Stirling. Utilising areas under the normal curve he obtained approximations to terms of the binomial distribution and thereafter expressed the binomial terms in an exponential form. He discovered that neighbouring terms could be obtained from areas under the normal curve by equating the Riemann sums to integrals. This was noted as the first work done on the calculation of areas under the normal curve by quadrature.

It was de Moivre and Laplace who showed that

$$\chi = \frac{(m - Np)}{(Npq)^{\frac{1}{2}}}, \tag{2.1}$$

was asymptotically standard normal where m denotes the number of successes in N independent trials with p being a probability of a success in each trial.

Squaring (2.1) gives

$$\begin{aligned}
\chi^2 &= \frac{(m - Np)^2}{(Npq)} \\
&= \frac{(m - Np)^2(p + q)}{Npq}, \text{ since } p + q = 1 \\
&= \frac{(m - Np)^2}{Np} + \frac{(m - Np)^2}{Nq} \\
&= \frac{(m - Np)^2}{Np} + \frac{(Np - m)^2}{Nq} \\
&= \frac{(m - Np)^2}{Np} + \frac{(N - m - N + Np)^2}{Nq} \\
&= \frac{(m - Np)^2}{Np} + \frac{(N - m - N(1 - p))^2}{Nq} \\
&= \frac{(m - Np)^2}{Np} + \frac{(N - m - Nq)^2}{Nq},
\end{aligned} \tag{2.2}$$

since $q = 1 - p$.
Pearson (1900) generalized (2.2) to obtain

$$\chi^2 = \sum_{i=1}^{k} \frac{(a_i - Np_i)^2}{Np_i}, \tag{2.3}$$

where a_i = observed number in the ith cell of a multinomial distribution which has k categories. Lancaster (1969 , p. 2) reports that it was Bienyamé who obtained the distribution of the sum of squares of n independently distributed normal variables in the gamma function form. Bienyamé evaluated the integral

$$P(U^2 < \gamma^2) = \frac{2}{\Gamma(\frac{1}{2}n})\int_0^{\gamma} u^{n-1}e^{-u^2}du.$$

He also presented work on the distribution of a linear form in the class frequencies of a multinomial distribution. In later work Bienyamé gave formulae that indicated that he was proposing to extend the normal approximation to the multinomial distribution. Later he presented a version of the multivariate central limit theorem as well as a χ^2 distribution for a sum of standardized squares.

Karl Pearson was initially responsible for highlighting the use of the χ^2 distribution. Other scientists who made significant contributions were Lexis and Sheppard. Sheppard looked at feasible goodness-of-fit tests for the multinomial distribution. He suggested a method to test the goodness-of-fit by calculating the difference between the observed frequency and the expected frequency for each cell of a frequency table and thereafter to check how often it exceeded the probable error. He concluded that in a good fit this would happen less often than in a bad fit. It was Pearson, who presented the widely used goodness of fit test by using the variance-covariance matrix of the multinomial distribution in the quadratic form approach rather than the awkward form of the variances and covariances for a two-way contingency table as was done in the work of Sheppard.

2.2 Summary of Karl Pearson's Work

Pearson introduced a number of theoretical statistical distributions which are now known as Pearson's system of frequency curves. One of these distributions is the so called Type III distribution which includes the gamma and chi-squared distributions. On noticing that the normal curve was being accepted as fitting a set of data even when it was not an accurate fit, Pearson realized the necessity to identify a goodness of fit test which would determine which distributions described a frequency distribution well and which did not.

In his paper, Pearson (1900), starts by considering

$$\chi^2 = \mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y},$$

where $\mathbf{Y}$ is an $n \times 1$ vector whose elements are random variables and $\mathbf{V}$ is an $n \times n$ positive definite variance-covariance matrix. Then Pearson's lemma states that there is a linear transformation, $\mathbf{Y} \rightarrow \mathbf{Z}$, such that

$$\mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \,. \tag{2.4}$$

Further if the joint distribution of $Y_1, \ldots, Y_n$ is given by

$$f(y_1, \ldots, y_n) = c\exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}\right) \,, \tag{2.5}$$

where c is a constant, then $\mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y} = \chi^2$ is distributed as the sum of squares of n independently distributed standard normal variables. This follows from Pearson's lemma, since the joint distribution of the Z_i's is

$$g(z_1, \ldots, z_n) = c\exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{z}\right) \tag{2.6}$$

which is the probability density function of n independent standard normal variables. Furthermore we have

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

Pearson refers to χ as the generalized probable error.

For the multinomial distribution, Pearson then went on to define the observed and expected frequencies $\acute{m}_i$ and m_i, respectively and differences $e_i = \acute{m}_i - m_i$, $i = 1, \ldots, n+1$, where

$$\sum_{i=1}^{n+1} e_i = 0, \tag{2.7}$$

and

$$\mathrm{Var}(e_i) = Np_i(1-p_i) = \frac{N(1-\frac{m_i}{N})}{(\frac{m_i}{N})} = \sigma_i{}^2\ , \tag{2.8}$$

and

$$\mathrm{Cov}(e_i, e_j) = \frac{-m_i m_j}{N} = -Np_i p_j = \sigma_i \sigma_j r_{ij}\ . \tag{2.9}$$

Because of the identity in (2.7), only the first n of the e_i were considered to be "variables" by Pearson. Under the assumption that he would always work with large numbers, Pearson considered the errors to be approximately distributed as normal variables having an $n \times n$ covariance matrix, $\mathbf{V}$, given by (2.8) and (2.9). He also further assumed that normal variables have a joint normal distribution.

Using the assumption of joint normality, Pearson expressed the joint distribution of the first n of the set e_i by an equation of the form (2.4). Trying to express $\chi^2 = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e}$ in a form so that computations could be carried out with ease, involved intricate work for deriving $\mathbf{V}^{-1}$. Pearson determined the elements of $\mathbf{V}^{-1}$ by an application of the theory of partial correlation.

Pearson's research in 1904 involved the study of contingency tables which considered the underlying distribution to potentially have a joint normal distribution. He defined "contingency" and other related numerical measures. The "first coefficient of contingency ", r was defined as

$$r = \pm\phi(1+\phi^2)^{-1/2}, \text{ with } \phi^2 = \frac{\chi^2}{N}.$$

It was also Pearson who showed that χ^2 was unaffected by a reordering of the marginal classes.

2.3 Findings and Accomplishments of R.A. Fisher

Fisher worked on testing hypotheses in regression analysis. He considered cases where a joint normal distribution was assumed and where the sums of squares resulted in quadratic forms of normal variables. On reading Pearson's *Mathematical contributions to the Theory of Evolution,* he derived the distribution of the correlation coefficient. He also derived the distribution for the non-central and central χ^2 as well as the distribution of the correlation ratio, Pearson's η.

Other contributions made by Fisher in 1922 include the concepts of sufficiency of estimators and measuring the effectiveness of statistical tests. He is further acclaimed for the first fundamental proof of the asymptotic distribution of the χ^2 statistic when parameters are estimated from the data. Another contribution by Fisher, was to partition χ^2 into different components which made the comparison of nested models possible.

His work also provided insight into a measure of second order interaction for contingency tables of higher dimensions. Further contributions include an alternative method of partitioning the overall χ^2 in higher order tables. The theory of two dimensional contingency tables also owes much to Fisher.

2.4 Early Controversies of Categorical Data Analysis

Agresti (1996, p. 257) contends that "the beginnings of categorical data analysis were often shrouded in controversy". He states that although key figures in the development of statistical science made significant contributions, they were frequently engaged in disputes with one another. Among these scientists, one can include Karl Pearson, G Udny Yule, and R A Fisher.

A literature survey suggests that the earliest methods for analysing categorical data were developed in England. Therefore, for the purposes of this study, the historical overview of the evolution of categorical data will commence in London.

2.4.1 Contentions arising from the Pearson - Yule Alliance

By 1900, Karl Pearson (1857–1936) was recognized by the statistical community for his work on the Pearson curves, finding the product-moment estimate of the correlation coefficient and its standard error. Documentation of Pearson's research work show that he wrote articles on an assortment of subjects which include art, religion, philosophy, socialism, women's rights, physics, genetics, and evolution. This versatile ability earned him the title of being called a renaissance man.

Literature on categorical data analysis in the early 1900's concentrated on the discussions and debates regarding suitable choices of summary indices for describing association. He recognized that association could be measured by approximating a measure, like correlation. In 1904 Pearson described the term *contingency* as "a measure of the total deviation of the classification from independent probability" (Agresti, (1996, p. 258)) and he further defined measures to describe its extent.

George Udny Yule (1871 – 1951), an Englishman and associate of Pearson's concluded his investigations in multiple regression and partial correlation coefficients, and ventured into an examination of association in contingency tables. He believed that many categorical variables are fundamentally discrete and he further represented indices in terms of cell counts, without assuming underlying continuum. One such measure attributed to him is the odds ratio θ and a transformation of it to the $[-1,+1]$ scale,

$$Q = \frac{(\theta - 1)}{(\theta + 1)},$$

now described as *Yule's Q* . Yule said the following with regard to Pearson's assumptions of underlying normality for certain measures, "at best the normal coefficient can only be said to give us in cases like these, a hypothetical correlation between supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work" (Agresti, (1996, p. 258)).

Karl Pearson did not accept Yule's criticism graciously and he and D. Heron responded by filling more than 150 pages of Pearson's journal *(Biometrika)*with a bitter response to Yule's criticism. In addition, they responded negatively to Yule's book, *An Introduction to the Theory of Statistics* (Griffin, 1911) despite the positive reception it received from the statistical community. They found fault with the above-mentioned book and declared that

"If Mr Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory.... [Yule's Q] has never been and never will be used in any work done under his [Pearson's] supervision..... We regret having drawn attention to the manner in which Yule has gone astray at every stage in his treatment of association, but criticism of his methods has been thrust on us not only by Mr Yule's recent attack, but also by the unthinking praise which has been bestowed on a text-book which at many points can only lead statistical students hopelessly astray".

It was later seen that Pearson and Yule had legitimate arguments since most nominal variables did not possess an obvious or noticeable underlying continuous distribution. However, many applications did relate to an underlying continuum, and could be used for model building and inference toward that continuum.

2.4.2 Disputes in the Pearson R.A. Fisher Association

Fisher (1890-1962) introduced the idea of *degrees of freedom* and proposed that for tests of independence in $I \times J$ tables, X^2 had $(I-1)(J-1)$ degrees of freedom. On the contrary, Pearson claimed that any application of his statistic, had degrees of freedom equal to the number of cells minus 1, or $IJ - 1$ for a two way table. Fisher pointed out that an additional $(I-1)(J-1)$ constraints on the fitted values arose when estimating hypothesized cell probabilities using estimated row and column probabilities thus modifying the distribution of X^2.

Pearson's criticism of Fisher's findings was even more vicious than his attack on Yule's claims. He stated the following "I hold that such a view [Fisher's] is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of Royal Statistical Society* ....I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us" (Agresti (1996, p. 259)).

Pearson declared that the use of row and column sample proportions to estimate unknown probabilities was inconsequential for large sample distributions. Fisher tried unsuccessfully to get his rebuttal published by the Royal Statistical Society and he withdrew his membership.

Unfortunately, it was only shortly thereafter that, statisticians realized that Fisher was correct but Fisher was embittered over this and his dealings with Pearson. When writing about Pearson, he said : "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age" (Agresti, 1996, p. 260). In an article in 1926, he verified his earlier assertions regarding the degrees of freedom for the chi-squared distributions by using 12000 2×2 tables randomly generated by Karl Pearson's son E. S.

Pearson. He showed that the sample mean of X^2 was 1.00001, which is nearer to his earlier prediction of 1.0 for the formula of $E(X^2)$ than Pearson's $IJ - 1 = 3$.

Fisher's contributions to areas such as design of experiments and analysis of variance; his introduction of concepts of sufficiency, information, and optimal properties of maximum likelihood estimators earned him the recognition that he deserved amongst the statistical fraternity. Realising the restrictions of large sample statistical methods for laboratory work, he was one of the first to aid the work on small samples and hence promoted the research done by W. S. Gosset on the t distribution.

Working on applications in toxicology with binary responses, Chester Bliss used and made famous the probit model and hence contributed to some work on model building for categorical data analysis. Fisher presented an algorithm for obtaining maximum likelihood estimates of parameters for the probit model in the appendix of one of Bliss's articles in 1935. This algorithm is usually described as *Fisher scoring.*

Canonical correlation methods for contingency tables was another area of interest to Fisher. He allotted scores to rows and columns of a contingency table in such a manner that a maximum correlation is obtained.

2.5 Logistic Regression and the Loglinear Model

Many categorical variables consist of only two categories, for example (yes, no) or (dead, alive), giving rise to binary data. A binary response can be defined in terms of a Bernoulli variable with the probability of a success, denoted by π, and the probability of a failure, denoted by $(1 - \pi)$. When observing n independent observations on a binary response with parameter π, the number of successes has a binomial distribution with parameters n and π. Consider the simple case of one explanatory variable X. To show that the value of π changes as the value of X changes, write $\pi(x)$. For example as one increases the level of toxicity, say X, so the probability that an insect dies, $\pi(x)$, will increase. Often the relationship between $\pi(x)$ and x is nonlinear with a fixed change in X having a smaller impact when π is nearer 0 or 1 rather than at the centre of the range. In most cases the nonlinear relationship consists of $\pi(x)$ increasing continuously as x decreases or vice versa resulting in an "S" shaped curve. A function having this shape is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

which is called the logistic regression function. This function may also be written in the form

$$\log\left[\frac{\pi}{(1-\pi)}\right] = \alpha + \beta x.$$

R.A. Fisher and Frank Yates proposed $\log\left[\frac{\pi}{(1-\pi)}\right]$ as a plausible transformation of a binomial parameter for analysing binary data. This type of transformation was described by the term "logit".

Other models which may be considered are of the form $\pi(x) = F(\alpha + \beta x)$, where $F(\cdot)$ is a continuous distribution function. When $F(\cdot)$ is the standard normal cumulative distribution function, $\Phi(\cdot)$, then the model is

$$\pi(x) = \Phi(\alpha + \beta x) \text{ or } \Phi^{-1}(\pi(x)) = \alpha + \beta x\ .$$

This model is called a probit model. The early work of Joseph Berkson, showed that the fit for the logistic regression model and the probit model are similar.

Sir David R. Cox expanded on logistic regression through his 1970 book, *The Analysis of Binary data.* Another variation of the logit model, the Rasch model, which is used in psychometric testing, was introduced by George Rasch in the same time period. Many powerful developments took place in categorical data analysis during the quarter century following the end of World War

II. A few findings include expressions derived by H. Cramér and C. R. Rao for large sample distributions of parameter estimators in models for categorical data analysis. Jerzey Neyman initiated the family of best asymptotically normal (BAN) estimators which possess similar properties as the large sample properties of maximum likelihood estimators. During the early 1950's, William Cochran generalized (Cochran's Q) of McNemars's test which compared proportions in several matched samples and further explained the partitioning of the chi-squared statistics into parts that described various elements of association. Explanations on appropriate sample size for the chi-squared approximations to work well for the X^2 statistic was another contribution by him as well as the test of conditional independence for $2 \times 2 \times K$ tables.

Bartlett investigated the aspect of interaction in contingency tables. These findings were extended to multiway tables in articles by J. N. Darroch, I.J Good, L Goodman, H. O. Lancaster, N. Mantel, R. L. Plackett and S. Roy. Some instrumental work done by Birch in 1963−1965 showed how to obtain maximum likelihood estimates of cell probabilities in three-way tables under various conditions. Earlier theoretical findings of Cramér and Rao on large sample distributions for categorical data models were also discussed and expanded. Birch's articles sparked off substantial research on loglinear models between 1965 and 1975.

For an $r \times c$ contingency table, with classification variables A and B, a saturated loglinear model for the data is given by

$$\log(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

where $\sum\limits_i \lambda_i^A = 0$, $\sum\limits_j \lambda_j^B = 0$ and $\sum\limits_{i,j} \lambda_{ij}^{AB} = 0$. This model is easily extended to higher order contingency tables.

The next decade's research concerning loglinear and logit modelling were centered at three American Universities namely : Chicago, Harvard and North Carolina. Leo Goodman, of the University of Chicago, presented his findings on partitioning of chi-squared models for square tables, stepwise logit and loglinear model-building procedures, as well as specialized models for ordinal data. He also wrote articles for social science journals and hence made loglinear and logit methods popular approaches for applications.

Some of Goodman's students at Chicago made many useful contributions. One such student is Shelby Haberman. He completed his Ph.D dissertation making sound theoretical contributions to loglinear models. Areas covered in his discussions include residual analysis, loglinear models for ordinal variables and theoretical results for models.

Later, Stephen Fienberg and William Cochran, students of Fredrick Mosteller were involved in related research on maximum likelihood methods for loglinear and logit models. The bulk of this research was kindled by problems which arose after analyzing large multivariate data sets in the National Halothane study. This study compared halothane with other anaesthetics in order to ascertain if halothane was more likely than other anaesthetics to cause death due to liver damage. Mosteller (1968) explained how loglinear models could be used for smoothing in multidimensional discrete data sets. Gary Koch looked at studies in weighted least squares. This approach was applied to cases when maximum likelihood methods presented problems.

Later advancements in categorical data analysis include modelling of ordinal data by Leo Goodman in 1979; Cyrus Metha and Nitin Patel's algorithm for implementing exact-small sample methods, generation of graphical models for multiway contingency tables, conditional likelihood methods for multiway contingency tables and method of conditional likelihood for modelling of odds ratios and methodology for longitudinal and multivariate categorical responses.

John Nelder and R. W. M. Wedderburn's research on generalized linear models is construed to be a significant contribution since it combines logistic and probit regression models for binomial data and loglinear models for Poisson data with concepts of regression and ANOVA for normal response data. The fitting of GLM's required Fisher scoring which was developed further by Fisher in 1935 when he used maximum likelihood fitting of probit models. Once again it is seen that Fisher's contributions were indeed very wide.

Apart from the work done on model building, Cressie and Read concentrated on describing and assessing goodness of fit. They investigated the conventional tests for goodness of fit and also suggested the alternative of the *power-divergence family* of statistics.

From 1950 onwards more research was done on small-sample comparisons between the X^2 and G^2 test statistics under the null model, as well as examining the effect of modified test statistics on various methods of parameter estimation and determining the efficiency of the modified test statistics.

Chapter 3

Statistics Measuring Goodness-of-fit : Discrete Multivariate Data

The fit of a model can be measured by comparing the expected frequencies for each outcome with the observed frequencies from the sample. Cressie and Read (1984) suggested the power-divergence statistic as an alternative to statistics proposed in the past. In this chapter a notation for modelling and testing discrete multivariate data will be presented. Section 3.1 discusses the multinomial distribution. Section 3.2 presents the definition of a model for a multinomial vector $\boldsymbol{\pi}$ and examines model fit whilst section 3.3 discusses the power divergence family of statistics introduced by Read and Cressie (1988). In section 3.4, the case of no parameter estimation is examined. Two methods used for estimating unknown parameters, the minimum power divergence estimation approach as well as the method of maximum likelihood are described in section 3.5. Independence and homogeneity in two dimensional tables for the case of parameter estimation are topics examined in section 3.6. Lastly, section 3.7, highlights different selection methods to ensure that the best loglinear model is chosen.

3.1 The Multinomial Distribution

The multinomial distribution plays a central role in the analysis of categorical data and will be discussed in this section.
Suppose an outcome of a trial can be classified into one of k mutually exclusive categories $A_i, i = 1, \dots k$, and that the probability of being classified into category A_i is $P(A_i) = \pi_i$, $i = 1, \dots, k$. If there are n independent trials and X_i is the number of outcomes classified in category A_i, $i = 1, \dots k$, then the random vector $\mathbf{X} = (X_1, X_2, ..., X_k)$ is said to have a *multinomial distribution* with parameters n, $\boldsymbol{\pi} = (\pi_1, ...\pi_k)$, which is denoted by $\mathrm{Mult}_k(n, \boldsymbol{\pi})$. The probability function of $\mathbf{X} = (X_1, ..., X_k)$ is given by

$$P(\mathbf{X} = \mathrm{x}) = n! \prod_{i=1}^{k} \frac{\pi_i^{x_i}}{x_i!}, \tag{3.1}$$

where $0 \leq x_i \leq n$, $0 \leq \pi_i \leq 1$; $i = 1, ..., k$ and $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} \pi_i = 1$.

As an example, consider the question:
"Should major political decisions be based on a nationwide referendum ?" Suppose that there are $k = 3$ answer categories with possible responses; A_1 ="agree", A_2 = "disagree" and A_3 ="no opinion". If n people are interviewed and asked their opinion on the question above, then the classification of the answers can be summarized by the random vector $\mathbf{X} = (X_1, X_2, X_3)$ where X_i = the number of times that A_i is observed, with $\sum_{i=1}^{3} X_i = n$.

In the case of a two-way frequency table with r levels for classification variable A and c levels for classification B, let X_{ij} denote the number of outcomes classified in cell (i, j), and π_{ij} denote the probability that an outcome is classified in cell (i, j). If there are n independent trials then the random vector

$\mathbf{X} = (X_{11}, \ldots, X_{1c}, \ldots, X_{r1} \ldots, X_{rc})$ will have a multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1c}, \ldots, \pi_{r1} \ldots, \pi_{rc})$, i.e. $\mathbf{X} \sim \text{Mult}(n, \boldsymbol{\pi})$.

The moment generating function of $\mathbf{X}$ is given by

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= E[e^{\mathbf{t}'\mathbf{X}}] = E[\exp \sum_{i=1}^{k} t_i X_i] \\ &= \sum_{\mathbf{x} \in \mathbf{A}} \frac{n!}{x_1! x_2! \cdots x_k!} (\pi_1 e^{t_1})^{x_1} (\pi_2 e^{t_2})^{x_2} \cdots (\pi_k e^{t_k})^{x_k} \\ &= (\pi_1 e^{t_1} + \pi_2 e^{t_2} + \cdots + \pi_k e^{t_k})^n \,. \end{aligned}$$

The mean of X_i is

$$\begin{aligned} E(X_i) &= \frac{\partial}{\partial t_i} M_{\mathbf{X}}(\mathbf{t}) \Big]_{\mathbf{t}=\mathbf{0}} \\ &= n(\pi_1 e^{t_1} + \cdots + \pi_k e^{t_k})^{n-1} \pi_i e^{t_i}]_{\mathbf{t}=\mathbf{0}} = n\pi_i \end{aligned}$$

It follows that

$$\begin{aligned} E(X_i^2) &= \frac{\partial^2}{\partial t_i^2} M_{\mathbf{X}}(\mathbf{t}) \Big]_{\mathbf{t}=\mathbf{0}} \\ &= n(n-1)(\pi_1 e^{t_1} + \cdots + \pi_k e^{t_k})^{n-2} (\pi_i e^{t_i})^2 + n(\pi_1 e^{t_1} + \cdots + \pi_k e^{t_k})^{n-1} \pi_i e^{t_i}]_{\mathbf{t}=\mathbf{0}} \\ &= n(n-1)\pi_i^2 + n\pi_i \,. \end{aligned}$$

$$\begin{aligned} E(X_i X_j) &= \frac{\partial^2}{\partial t_i \partial t_j} \mathbf{M}_{\mathbf{X}}(\mathbf{t}) \Big]_{\mathbf{t}=\mathbf{0}} \\ &= n(n-1)\ (\pi_1 e^{t_1} + \cdots + \pi_k e^{t_k})^{n-2} \pi_i e^{t_i} \pi_j e^{t_j}]_{\mathbf{t}=\mathbf{0}} \\ &= n(n-1)\pi_i \pi_j \quad \text{for } i \neq j. \end{aligned}$$

Thus

$$\begin{aligned}\operatorname{var}(X_i) &= E(X_i^2) - (EX_i)^2\\ &= n(n-1)\pi_i^2 + n\pi_i - n^2\pi_i^2\\ &= n\pi_i^2 + n\pi_i = n\pi_i(1-\pi_i)\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(X_i, X_j) &= E(X_iX_j) - E(X_i)E(X_j)\\ &= n(n-1)\pi_i\pi_j - (n\pi_i)(n\pi_j) = -n\pi_i\pi_j\ .\end{aligned}$$

For the random vector $\mathbf{X}$ we have $E(\mathbf{X}) = n\boldsymbol{\pi}$ and

$$\begin{aligned}\operatorname{cov}(\mathbf{X}) &= \begin{pmatrix} n\pi_1(1-\pi_1) & -n\pi_1\pi_2 & \cdots & -n\pi_1\pi_k \\ -n\pi_2\pi_1 & n\pi_2(1-\pi_2) & \cdots & -n\pi_2\pi_k \\ \vdots & \vdots & & \vdots \\ -n\pi_k\pi_1 & -n\pi_k\pi_2 & & n\pi_k(1-\pi_k) \end{pmatrix}\\[2ex] &= n\left[\begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ 0 & 0 & \cdots & \pi_k \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{pmatrix} (\pi_1, \pi_2 \cdots \pi_k)\right]\\[2ex] &= n[\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}'].\end{aligned}$$

Thus if $\mathbf{P} = \mathbf{X}/n$ then

$$\begin{aligned}\operatorname{cov}(\mathbf{P}) &= E\left[(\mathbf{X}/n - \boldsymbol{\pi})(\mathbf{X}/n - \boldsymbol{\pi})'\right]\\ &= \frac{1}{n^2}\operatorname{cov}(\mathbf{X}) = \frac{1}{n}[\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}'].\end{aligned}$$

Another important result is that as $n \to \infty$ the multinomial random vector will have an asymptotic normal distribution. This result is stated and proved in Theorem 3.1.

Let $\mathbf{X}_n = (X_{n1}, X_{n2}, \ldots, X_{nk})' \sim \text{Mult}(n; \boldsymbol{\pi})$ and $\mathbf{U}_n = (\mathbf{X}_n - n\boldsymbol{\pi})/\sqrt{n} = \sqrt{n}(\mathbf{P} - \boldsymbol{\pi})$, where $\mathbf{P} = \frac{\mathbf{X}_n}{n}$. It then follows that

$$E(\mathbf{U}_n) = \mathbf{0} \text{ and } \operatorname{cov}(\mathbf{U}_n) = n\operatorname{cov}(\mathbf{P}) = \mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}'\ .$$

Theorem 3.1:
$\mathbf{U}_n \xrightarrow{d} \mathbf{U}$, where $\mathbf{U}$ has the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}'$.

Proof: The m.g.f of $\mathbf{U}_n$ is given by

$$\begin{aligned} M_{\mathbf{U}_n}(\mathbf{t}) &= E[e^{\mathbf{t}'\mathbf{U}_n}] = E[e^{\mathbf{t}'(\mathbf{X}_n - n\boldsymbol{\pi})/\sqrt{n}}] \\ &= e^{-\mathbf{t}'\boldsymbol{\pi}\sqrt{n}} M_{\mathbf{X}_n}(\mathbf{t}/\sqrt{n}) \\ &= e^{-\mathbf{t}'\boldsymbol{\pi}\sqrt{n}} \left(\sum_{j=1}^{k} \pi_j e^{t_j/\sqrt{n}} \right)^n \end{aligned}$$

Note that $e^{-\mathbf{t}'\boldsymbol{\pi}\sqrt{n}} = \left(e^{-\mathbf{t}'\boldsymbol{\pi}/\sqrt{n}}\right)^n$. Thus

$$M_{\mathbf{U}_n}(\mathbf{t}) = \left(\sum_{j=1}^{k} \pi_j e^{(t_j - \mathbf{t}'\boldsymbol{\pi})/\sqrt{n}} \right)^n$$

Since $e^x = 1 + x + x^2/2 + o(x^2)$, as $x \to 0$, we have

$$\begin{aligned} M_{\mathbf{U}_n}(\mathbf{t}) &= \left[\sum_{j=1}^{k} \pi_j \{1 + (t_j - \mathbf{t}'\boldsymbol{\pi})/\sqrt{n} + \tfrac{1}{2}(t_j - \mathbf{t}'\boldsymbol{\pi})^2/n + o(n^{-1})\} \right]^n \\ &= \left[1 + \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \pi_j (t_j - \mathbf{t}'\boldsymbol{\pi}) + \frac{1}{2n} \sum_{j=1}^{k} \pi_j (t_j - \mathbf{t}'\boldsymbol{\pi})^2 + o(n^{-1}) \right]^n \end{aligned}$$

Now

$$\begin{aligned} \sum_{j=1}^{k} \pi_j (t_j - \mathbf{t}'\boldsymbol{\pi}) &= \sum_{j=1}^{k} \pi_j t_j - \mathbf{t}'\boldsymbol{\pi} \sum_{j=1}^{k} \pi_j \\ &= \mathbf{t}'\boldsymbol{\pi} - \mathbf{t}'\boldsymbol{\pi} = 0 \end{aligned}$$

and

$$\sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi})^2 = \sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi})(t_j - \mathbf{t}'\boldsymbol{\pi})$$

$$= \sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi})t_j - \sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi})\mathbf{t}'\boldsymbol{\pi}$$

$$= \sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi})t_j - 0\ , \text{ since } \sum_{j=1}^{k} \pi_j(t_j - \mathbf{t}'\boldsymbol{\pi}) = 0, \text{ as above}$$

$$= \sum_{j=1}^{k} \pi_j t_j t_j - \left(\sum_{j=1}^{k} \pi_j t_j\right)^2$$

$$= \sum_{j=1}^{k} t_j \pi_j t_j - \sum_{j=1}^{k} \sum_{j'=1}^{k} \pi_j \pi_{j'} t_j t_{j'} = \mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t}\ .$$

Thus $M_{\mathbf{U}_n}(\mathbf{t}) = \left(1 + \frac{1}{n} \cdot \frac{1}{2}\mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t} + o(n^{-1})\right)^n$, so that

$$\begin{aligned} \ln M_{\mathbf{U}_n}(\mathbf{t}) &= n \ln\left\{1 + \tfrac{1}{n} \cdot \tfrac{1}{2}\mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t} + o(n^{-1})\right\} \\ &= n\left[\tfrac{1}{n} \cdot \tfrac{1}{2}\mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t} + o(n^{-1})\right] \\ &= \tfrac{1}{2}\mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t} + o(1) \quad \text{as } n \to \infty \end{aligned}$$

since $\ln(1 + x) = x + o(x)$ as $x \to 0$. Thus

$$\lim_{n\to\infty} M_{\mathbf{U}_n}(\mathbf{t}) = e^{\frac{1}{2}\mathbf{t}'(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{t}},$$

which is the m.g.f. of the $N(\mathbf{0}, \mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')$ distribution.

This result may also be expressed as follows:

If $\mathbf{X} \sim \text{Mult}(n; \boldsymbol{\pi})$, then $\mathbf{X}$ has approximately the $N(n\boldsymbol{\pi}, n(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}'))$ distribution for n large.

3.2 Defining a Model and Model Fit

The simplest model for the multinomial probability vector $\boldsymbol{\pi}$, is the null model described by :

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0, \tag{3.2}$$

with $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, ..., \pi_{0k})$ representing the hypothesized probability vector and expected cell frequencies given by $n\boldsymbol{\pi}_0 = (n\pi_{01}, n\pi_{02}, ..., n\pi_{0k})$. Let $\mathbf{x} = (x_1, x_2, ..., x_k\ )$, where x_i denotes the number of observed replies in cell A_i and $\sum_{i=1}^{k} x_i = n$. Model fit can be measured by comparing the expected frequency of the ith cell, $n\pi_{0i}$ with the observed frequency x_i. The null model is rejected if the difference between the observed and expected frequencies becomes too large.

Popular goodness-of-fit statistics used to test $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ include :

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - n\pi_{0i})^2}{n\pi_{0i}}, \tag{3.3}$$

which was proposed by Pearson and the loglikelihood ratio statistic given by

$$G^2 = 2\sum_{i=1}^{k} x_i \log\left(\frac{x_i}{n\pi_{0i}}\right). \tag{3.4}$$

$X^2 = 0 = G^2$ only if there are no differences between x_i and $n\pi_{0i}$. An increase in the discrepancy between $\mathbf{x}$ and $n\boldsymbol{\pi}_0$, results in an increase of the values of X^2 and G^2.

If $\mathbf{x}$ is replaced by the multinomial vector $\mathbf{X}$ in (3.3) and (3.4), then X^2 and G^2 may be viewed as random variables. Pearson reported that X^2 exhibited properties of a chi-squared distribution with $k - 1$ degrees of freedom, when the sample size n increased, under the null model described previously in (3.2). This result is stated and proved in the following theorem.

Theorem 3.2
Consider the hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ ($\boldsymbol{\pi}_0$ a fixed value).

$$X^2 = \sum_{i=1}^{k} \frac{(X_i - n\pi_{0i})^2}{n\pi_{0i}} \sim \chi^2(k-1).$$

Proof:
Write X^2 as follows:

$$\begin{aligned} X^2 &= \sum_{i=1}^{k} \frac{n^2\left(\frac{X_i}{n} - \pi_{0i}\right)^2}{n\pi_{0i}} \\ &= n(\mathbf{P} - \boldsymbol{\pi}_0)'\mathbf{D}_{\pi_0}^{-1}(\mathbf{P} - \boldsymbol{\pi}_0) \\ &= \sqrt{n}(\mathbf{P} - \boldsymbol{\pi}_0)'\mathbf{D}_{\pi_0}^{-1}(\mathbf{P} - \boldsymbol{\pi}_0)\sqrt{n} \\ &= \mathbf{U}_n'\mathbf{D}_{\pi_0}^{-1}\mathbf{U}_n\ . \end{aligned}$$

Thus $\mathbf{U}_n \xrightarrow{d} \mathbf{U}$ where $\mathbf{U} \sim N(\mathbf{0}, \mathbf{D}_{\pi_0} - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')$. By Theorem 3.1, (replacing $\boldsymbol{\pi}$ by $\boldsymbol{\pi}_0$), it follows that

$$X^2 = \mathbf{U}_n'\mathbf{D}_{\pi_0}^{-1}\mathbf{U}_n \xrightarrow{d} \mathbf{U}'\mathbf{D}_{\pi_0}^{-1}\mathbf{U}\ .$$

Using Corollary 2s.2 of Searle (1971, p.69), we have the result: $\mathbf{X} \sim N(\mathbf{0}, \mathbf{V})$ then, for singular or non-singular $\mathbf{V}$, $\mathbf{X}'\mathbf{A}\mathbf{X} \sim \chi^2(tr(\mathbf{AV}))$ if and only if $\mathbf{VAVAV} = \mathbf{VAV}$. Firstly

$$\begin{aligned} tr(\mathbf{AV}) &= tr[\mathbf{D}_{\pi_0}^{-1}(\mathbf{D}_{\pi_0} - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')] \\ &= tr[\mathbf{I} - \mathbf{D}_{\pi_0}^{-1}\boldsymbol{\pi}_0\boldsymbol{\pi}_0'] \\ &= tr(\mathbf{I}) - tr(\boldsymbol{\pi}_0'\mathbf{D}_{\pi_0}^{-1}\boldsymbol{\pi}_0) \\ &= k - tr(1), \quad \text{since } \boldsymbol{\pi}'\mathbf{D}_{\pi}^{-1}\boldsymbol{\pi} = 1 \\ &= k - 1\ . \end{aligned}$$

To show that **VAVAV** = **VAV**, note that

$$\begin{aligned}\mathbf{AV} &= \mathbf{D}_{\pi}^{-1}(\mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}') = \mathbf{I} - \mathbf{D}_{\pi}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}' \\ \text{and } \mathbf{VAV} &= (\mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}')(\mathbf{I} - \mathbf{D}_{\pi}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}') \\ &= \mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}' - (\mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}')(\mathbf{D}_{\pi}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}') \\ &= \mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}' - \boldsymbol{\pi}\boldsymbol{\pi}' + \boldsymbol{\pi}\boldsymbol{\pi}'\mathbf{D}_{\pi}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}' \\ &= \mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}' \text{ , since } \boldsymbol{\pi}'\mathbf{D}_{\pi}^{-1}\boldsymbol{\pi} = 1 \text{ .}\end{aligned}$$

Now

$$\begin{aligned}\mathbf{VAVAV} &= (\mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}')(\mathbf{I} - \mathbf{D}_{\pi}^{-1}\boldsymbol{\pi}\boldsymbol{\pi}') \\ &= \mathbf{D}_{\pi} - \boldsymbol{\pi}\boldsymbol{\pi}' = \mathbf{VAV} \text{ ,}\end{aligned}$$

which is the required result.
Thus $\chi^2 = \mathbf{U}_n'\mathbf{D}_{\pi_0}^{-1}\mathbf{U}_n \xrightarrow{d} \chi^2(k-1)$.

Null models that do not completely specify the null probability vector $\boldsymbol{\pi}_0$, create problems. For example, (Read and Cressie, (1988, p10)), discuss a model $\log \pi_i = \alpha + \beta i$, which has two unspecified parameters α and β. These parameters are called nuisance parameters and must be estimated from the sample. Once the parameters have been estimated, the estimates for the expected frequencies $n\hat{\pi}_i$ can be found and the goodness-of-fit statistics given in (3.3) and (3.4) can be calculated.

Estimation entails the null model being expressed by

$$H_0 : \boldsymbol{\pi}\epsilon\boldsymbol{\Pi}_0, \tag{3.5}$$

where $\boldsymbol{\Pi}_0$ refers to a set of hypothesized probability vectors for $\boldsymbol{\pi}$. Estimating nuisance parameters can be viewed as choosing an element of the set $\boldsymbol{\Pi}_0$ which proves most consistent with the sample data. The resulting estimated probability vector is denoted by $\hat{\boldsymbol{\pi}}$.

Fisher (1924) illustrated that in the case of one nuisance parameter occurring,

$$X^2 = \sum_{i=1}^{k} \frac{(X_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i}, \tag{3.6}$$

has a chi-squared distribution with $k-2$ degrees of freedom when (3.5) is true for increasing sample size. As highlighted in the earlier discussion on the controversy between Fisher and Pearson, Pearson had formerly prescribed $k-1$ degrees of freedom.

Studies undertaken by Cramer (1946) which involved generalisations to s parameters yielded

$$G^2 = 2\sum_{i=1}^{k} X_i \log\left[\frac{X_i}{n\widehat{\pi}_i}\right], \tag{3.7}$$

and (3.6) are asymptotically chi-squared with $k-s-1$ degrees of freedom, in the event that certain regularity conditions on $\boldsymbol{\pi}$ and k hold and that (3.5) is true.

Recent additions to the family of goodness-of-fit statistics include the Freeman-Tukey Statistic which was defined as

$$F^2 = 4\sum_{i=1}^{k} \left(\sqrt{X_i} - \sqrt{n\widehat{\pi}_i}\right)^2, \tag{3.8}$$

by Fienberg (1979) and Moore (1986). Two further statistics are the modified loglikelihood ratio statistic,

$$GM^2 = 2\sum_{i=1}^{k} n\widehat{\pi}_i \log\left(\frac{n\widehat{\pi}_i}{X_i}\right), \tag{3.9}$$

and the Neyman-modified statistic (Neyman , 1949)

$$NM^2 = \sum_{i=1}^{k} \frac{(X_i - n\widehat{\pi}_i)^2}{X_i}. \tag{3.10}$$

Research findings have indicated that under the conditions described earlier, the abovementioned three statistics possess the same asymptotic chi-squared distribution as X^2 and G^2. Discrepancies between statistics do arise in terms of equivalence under descriptions of finite sample size thus causing controversy on the topic of choice of an appropriate statistic. Read and Cressie (1988) furnished the power-divergence family of goodness-of-fit statistics which will be introduced and discussed in the subsequent section.

3.3 The Power Divergence Statistic

In accordance with the previous methods for measuring the fit of a model by comparing the expected frequency for each category with the observed frequency, Read and Cressie (1988) defined the *power-divergence statistic* as:

$$2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}\right) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} X_i \left[\left(\frac{X_i}{n\widehat{\pi}_i}\right)^{\lambda} - 1\right], \qquad -\infty < \lambda < \infty\ , \tag{3.11}$$

where λ is the family parameter, and is chosen by the user. The power-divergence statistic is zero when the observed and expected frequencies are equal for every outcome, (for any value of λ). The statistic is always positive for all other cases and gets larger as the discrepancy between the observed and expected frequencies increases.

The statistic $2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}\right)$ measures the divergence of $\frac{\mathbf{X}}{n}$ from $\widehat{\boldsymbol{\pi}}$ through a weighted sum of powers of the terms $\dfrac{X_i}{n\widehat{\pi}_i}$ for $i = 1, ..., k$. In other words it measures how far the empirical probability diverges from the probability distribution under the hypothesis, hence the name *power-divergence statistic* .

The following notation

$$2I^{\lambda}\left(\mathbf{X} : \widehat{\mathbf{m}}\right) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} X_i \left[\left(\frac{X_i}{\widehat{m}_i}\right)^{\lambda} - 1\right]; \qquad -\infty < \lambda < \infty \tag{3.12}$$

is adopted by Read & Cressie (1988) when comparing the cell frequency vector

$\mathbf{X}$ with the estimated expected frequency vector, $\widehat{\mathbf{m}} = n\widehat{\boldsymbol{\pi}}$; where $\sum_{i=1}^{k} X_i = \sum_{i=1}^{k} \widehat{m}_i$. Closer examination of the previous equation reveals that the total sample size n is eliminated.

If $\lambda = 1$ is substituted in (3.12), the power-divergence statistic simplifies to the χ^2 statistic. This is verified as follows : For $\lambda = 1$ the power-divergence statistic is

$$\begin{aligned}
2I^1(\mathbf{X}_i : \widehat{\mathbf{m}}) &= \frac{2}{1(2)} \sum_{i=1}^{k} X_i \left(\frac{X_i}{\widehat{m}_i} - 1 \right) \\
&= \sum_{i=1}^{k} X_i \left(\frac{X_i - \widehat{m}_i}{\widehat{m}_i} \right) \\
&= \sum_{i=1}^{k} X_i \left(\frac{X_i - \widehat{m}_i}{m_i} \right) - \sum (X_i - \widehat{m}_i) \quad \text{since } \sum X_i = \sum \widehat{m}_i \cdot \\
&= \sum (X_i - \widehat{m}_i) \left[\frac{X_i}{\widehat{m}_i} - 1 \right] \\
&= \sum (X_i - \widehat{m}_i) \left(\frac{X_i - \widehat{m}_i}{\widehat{m}_i} \right) = \sum_{i=1}^{k} \frac{(X_i - \widehat{m}_i)^2}{\widehat{m}_i} = \chi^2
\end{aligned}$$

Further inspection of (3.12) shows that the equation is undefined for $\lambda = -1$ and $\lambda = 0$. But if $2I^{\lambda}(\mathbf{X} : \widehat{\mathbf{m}})$ is defined as the continuous limits of (3.12) as $\lambda \to -1$ and $\lambda \to 0$, then $2I^{\lambda}(\mathbf{X} : \widehat{\mathbf{m}})$ is continuous in λ.

The loglikehood ratio statistic is a special case of the power-divergence when $\lambda \to 0$ in (3.12). We use the fact that

$$\log(t) = \lim_{h \to 0} (t^h - 1)/h$$

Now

$$\lim_{\lambda\to 0} 2I^{\lambda}(\mathbf{X}:\widehat{\mathbf{m}}) = \lim_{\lambda\to 0}\sum_{i=1}^{k} X_i\left[\left(\tfrac{X_i}{\widehat{m}_i}\right)^{\lambda}-1\right]\tfrac{1}{\lambda}\left(\tfrac{2}{\lambda+1}\right)$$

$$= 2\sum_{i=1}^{k} X_i\log\left(\tfrac{X_i}{\widehat{m}_i}\right) = G^2\,.$$

For the case $\lambda \to -1$,

$$\lim_{\lambda\to -1} 2I^{\lambda}(\mathbf{X}:\widehat{\mathbf{m}}) = \lim_{\lambda\to -1}\frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} X_i\left[\left(\frac{X_i}{\widehat{m}_i}\right)^{\lambda}-1\right].$$

Let $y = \lambda+1$, then if $\lambda\to -1$, $y\to 0$. The limit above then becomes

$$\lim_{y\to 0}\tfrac{2}{(y-1)y}\sum_{i=1}^{k} X_i\left[\left(\tfrac{X_i}{\widehat{m}_i}\right)^{y-1}-1\right] = \lim_{y\to 0}\tfrac{2}{(y-1)(y)}\sum_{i=1}^{k}\left[\left(\frac{X_i}{\widehat{m}_i}\right)^{y}\widehat{m}_i - X_i\right]$$

$$= \lim_{y\to 0}\frac{2}{y(y-1)}\sum_{i=1}^{k}\widehat{m}_i\left[\left(\frac{X_i}{\widehat{m}_i}\right)^{y}-1+1-\frac{X_i}{\widehat{m}_i}\right]$$

$$= \lim_{y\to 0}\frac{2}{y(y-1)}\sum_{i=1}^{k}\widehat{m}_i\left[\left(\frac{X_i}{\widehat{m}_i}\right)^{y}-1\right] + \lim_{y\to 0}\frac{2}{y(y-1)}\sum_{i=1}^{k}(\widehat{m}_i - X_i)$$

$$= \lim_{y\to 0}\frac{2}{(y-1)}\sum_{i=1}^{k}\widehat{m}_i\left[\left(\frac{X_i}{\widehat{m}_i}\right)^{y}-1\right]\frac{1}{y}\ ,\quad \text{since } \sum(X_i-\widehat{m}_i)=0$$

$$= -2\sum_{i=1}^{k}\widehat{m}_i\log\left(\frac{X_i}{\widehat{m}_i}\right)$$

$$= 2\sum_{i=1}^{k}\widehat{m}_i\log\left(\frac{\widehat{m}_i}{X_i}\right) = GM^2\,.$$

Hence, equations (3.6), (3.7) and (3.9) may be expressed as :

$$2nI^{1}\left(\frac{\mathbf{X}}{n}:\widehat{\boldsymbol{\pi}}\right) = X^2\,,$$

$$2nI^{0}\left(\frac{\mathbf{X}}{n}:\widehat{\boldsymbol{\pi}}\right)=G^2\ ,$$

and

$$2nI^{-1}\left(\frac{\mathbf{X}}{n}:\widehat{\boldsymbol{\pi}}\right)=GM^2\ .$$

Similarly, substituting $\lambda = -\frac{1}{2}$ and $\lambda = -2$ in (3.12) gives rise to two other statistics that are linked to the power-divergence statistic through the index λ. They are the F^2 and NM^2 statistics respectively, with

$$2nI^{-\frac{1}{2}}\left(\frac{\mathbf{X}}{n}:\widehat{\boldsymbol{\pi}}\right)=F^2\ ,$$

and

$$2nI^{-2}\left(\frac{\mathbf{X}}{n}:\widehat{\boldsymbol{\pi}}\right)=NM^2\ .$$

The power-divergence statistic is seen to consolidate and unify goodness-of-fit tests which were previously considered in isolation.

3.4 Case of no Parameter Estimation

The simple null hypothesis for $\boldsymbol{\pi}$ is: $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, where $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, ..., \pi_{0k})$ is completely specified and each $\pi_{0i} > 0$.

From Theorem 3.2, under the null model above, the following results have been obtained.

(i) $\sqrt{n}(\frac{\mathbf{X}}{n} - \boldsymbol{\pi}_0)$ converges in distribution to a multivariate normal distribution as $n \to \infty$.

(ii) $X^2 = \sum_{i=1}^{k} \dfrac{(X_i - n\pi_{0i})^2}{n\pi_{0i}}$ converges in distribution as $n \to \infty$ to the quadratic form of the multivariate normal random vector in (i).

(iii) In the case of (i) and (ii), X^2 converges in distribution to a central chi-squared random variable with $k-1$ degrees of freedom.

From the results (i) to (iii) above, and the null model, it follows that $P(X^2 \geq c) \rightarrow P(\chi^2_{k-1} \geq c)$ for any $c \geq 0$ and $n \rightarrow \infty$, and

$$P\left(X^2 \geq \chi^2_{1-\alpha}(k-1)\right) \rightarrow \alpha \quad \text{as } n \rightarrow \infty. \tag{3.13}$$

Thus the null hypothesis will be rejected at a 100α percent significance level, if the value of X^2 exceeds the critical value $\chi^2_{1-\alpha}(k-1)$. The 100αth percentile for the chi-square distribution with ν degrees of freedom, $\chi^2_{1-\alpha}(\nu)$, is defined as follows:

$$P(\chi^2_\nu \leq \chi^2_{1-\alpha}(\nu)) = 1 - \alpha \, .$$

The Pearson X^2 test requires that the sample size, n in (i) is large enough for (3.13) to be true and further that the number of cells be fixed in order for k to be small in comparison with n. As a consequence of k being small in relation to n, it ensures that each expected cell frequency would be large since $n\pi_{0i} \rightarrow \infty$ for each $i = 1, ..., k$.

The following theorem shows that the power divergence statistic has the same asymptotic distribution as the Pearson X^2 statistic, under $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$.

Theorem 3.3 :

$$2nI^\lambda\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) = 2nI^1\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) + o_p(1) = X^2 + o_p(1), \qquad -\infty < \lambda < \infty.$$

Proof:

The power-divergence statistic, (for $\lambda \neq 0$ or $\lambda \neq -1$), under H_0, is

$$\begin{aligned}
2nI^{\lambda}\left(\tfrac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) = &= \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} X_i\left[\left(\frac{X_i}{n\pi_{0i}}\right)^{\lambda} - 1\right] \\
&= \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} n\pi_{0i}\left[\left(\frac{X_i}{n\pi_{0i}}\right)^{\lambda+1} - 1\right] \\
&= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{k} \pi_{0i}\left[\left(1 - \frac{n\pi_{0i}}{n\pi_{0i}} + \frac{X_i}{n\pi_{0i}}\right)^{\lambda+1} - 1\right] \\
&= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{k} \pi_{0i}\left[\left(1 + \frac{X_i - n\pi_{0i}}{n\pi_{0i}}\right)^{\lambda+1} - 1\right] \\
&= \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{k} \pi_{0i}\left[(1+V_i)^{\lambda+1} - 1\right] ,
\end{aligned}$$

where $V_i = \dfrac{X_i - n\pi_{0i}}{n\pi_{0i}}$.

Now consider $f(V_i) = (1+V_i)^{\lambda+1}$ and expand in a Taylor series, i.e.

$$\begin{aligned}
f(V_i) &= 1 + f'(0)V_i + \frac{f''(0)}{2!}V_i^2 + \ldots \\
&= 1 + (\lambda+1)V_i + \lambda(\lambda+1)\frac{V_i^2}{2} + o_p\left(\tfrac{1}{n}\right)
\end{aligned}$$

where $o_p(\frac{1}{n})$ depends on λ and converges in probability to zero faster than $\frac{1}{n}$ as $n \longrightarrow \infty$ because $\sqrt{n\pi_{0i}}V_i$ is asymptotically normally distributed as $n \longrightarrow \infty$, $i = 1, \ldots, k$. This follows from (i) since $\frac{\mathbf{X}}{\sqrt{n}} = \boldsymbol{\pi} + o_p\left(\frac{1}{\sqrt{n}}\right)$

Thus

$$2nI^{\lambda}\left(\tfrac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) = \; = \frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{k}\pi_{0i}\left[(\lambda+1)V_i + \lambda(\lambda+1)\frac{V_i^2}{2} + o_p\left(\tfrac{1}{n}\right)\right]$$

$$= n\left[\sum_{i=1}^{k}\pi_{0i}V_i^2 + o_p\left(\tfrac{1}{n}\right)\right] ,$$

since

$$\sum_{i=1}^{k}\pi_{0i}V_i = \sum_{i=1}^{k}\pi_{0i}\left(\frac{X_i - n\pi_{0i}}{n\pi_{0i}}\right) = \frac{\sum_{i=1}^{k}X_i - n\sum_{i=1}^{k}\pi_{0i}}{n} = \frac{n-n}{n} = 0 \, .$$

Thus

$$2nI^{\lambda}\left(\tfrac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) = \sum_{i=1}^{k}n\pi_{0i}\left(\frac{X_i - n\pi_{0i}}{n\pi_{0i}}\right)^2 + o_p(1)$$

$$= \sum_{i=1}^{k}\frac{(X_i - n\pi_{0i})^2}{n\pi_{0i}} + o_p(1) = X^2 + o_p(1) \, .$$

The power divergence statistic is hence found to have the same asymptotic distribution as Pearson's X^2 statistic with

$$P\left(2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) \geq \chi^2_{1-\alpha}(k-1)\right) \longrightarrow \alpha \quad \text{as } n \longrightarrow \infty, \qquad (3.14)$$

for each $\lambda \,\epsilon\, (-\infty, \infty)$ and each $\alpha \,\epsilon\, (0,1)$.

The null hypothesis will thus be rejected if X^2 or $2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right)$ exceeds the value $\chi^2_{1-\alpha}(k-1)$.

In conclusion it is seen that properties of the power-divergence family of statistics being equivalent to the asymptotic equations derived earlier, requires that the number of cells k be fixed and the expected cell frequencies $n\pi_{0i}$ should be large for all $i = 1, \ldots, k$.

3.5 Estimating Parameters

In the case where a model contains unknown parameters, these parameters will have to be estimated from the data. Calculation of the expected cell frequencies is only done after the unknown parameters have been estimated. Two methods that will be considered include the method of maximum likelihood and the minimum power divergence estimation approach.

3.5.1 Maximum Likelihood Approach

The maximum likelihood approach is a frequently used estimation technique. To illustrate the method consider a two-dimensional contingency table. If a full multinomial sampling scheme is used then the log likelihood function is

$$\log L(\boldsymbol{\pi}) = k_1 + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{ij})$$

where k_1 is not a function of π_{ij}. Note that

$$\begin{aligned}\log L(\boldsymbol{\pi}) &= k_1 + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log\left(\frac{n\pi_{ij}}{n}\right) \\ &= k_1 + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(m_{ij}) - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(n)\ .\end{aligned}$$

If a Poisson sampling scheme is used, then the likelihood function is

$$L(\mathbf{m}) = \prod_{i=1}^{r}\prod_{j=1}^{c}[\exp(-m_{ij})m_{ij}^{x_{ij}}]/x_{ij}!$$

and

$$\log L(\mathbf{m}) = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(m_{ij}) - \sum_{i=1}^{r}\sum_{j=1}^{c} m_{ij} + k_2\ ,$$

where k_2 is not a function of m_{ij}.

For a product-multinomial distribution with fixed row marginals, x_{i+}, the likelihood function is

$$L(\boldsymbol{\pi}) = \frac{\prod_{i=1}^{r} x_{i+}!}{\prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{c} \pi_{ij}^{x_{ij}}$$

and the log likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\pi}) &= k_3 + \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log(\pi_{ij}) \\ &= k_3 + \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log(m_{ij}) - \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log(n) \ , \end{aligned}$$

where k_3 is not a function of m_{ij}.

It is evident that all three sampling procedures have a log likelihood proportional to

$$K(\mathbf{m}) = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log(m_{ij}) \ .$$

Maximum likelihood estimation will require a constrained maximization of $K(\mathbf{m})$ with respect to the frequency m_{ij}, subject to the constraints imposed on the m_{ij} by the hypothesized model. For example the model $\log(\pi_{ij}) = \mu + \alpha_i + \beta_j$, $i = 1, \ldots, r$, $j = 1, \ldots, c$, imposes the constraints

$$m_{ij} = n\pi_{ij} = n \exp[\mu + \alpha_i + \beta_j], \ \ i = 1, \ldots, r \ ; \ j = 1, \ldots, c \ .$$

If the method of maximum likelihood is used, then note that

$$\begin{aligned}
K(\mathbf{m}) - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(x_{ij}) &= \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(m_{ij}) - \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(x_{ij}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log\left(\frac{m_{ij}}{x_{ij}}\right) \\
&= -\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log\left(\frac{x_{ij}}{m_{ij}}\right) = -\tfrac{1}{2}G^2
\end{aligned}$$

Thus maximizing the kernel of the loglikelihood function, $K(\mathbf{m})$, is equivalent to minimizing the loglikelihood ratio statistic, G^2.

3.5.2 Minimum Power-divergence Estimation Approach

In order to use the power-divergence family of statistics in a multidimensional table, arrange the frequencies appropriately into a column vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, where k will now be the number of cells in the table. The power-divergence statistic will be used in the form

$$2I^{\lambda}(\mathbf{x} : \widehat{\mathbf{m}}) = \frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{k} x_i \left[\left(\frac{x_i}{\widehat{m}_i}\right)^{\lambda} - 1\right] ; -\infty < \lambda < \infty \qquad (3.15)$$

with the limit as $\lambda \to -1$ and $\lambda \to 0$ being used for $\lambda = -1$ and $\lambda = 0$, respectively and the subscript i runs from 1 to k, the total number of cell in the contingency table, to ensure that the statistic can be applied to multidimensional tables.

Read and Cressie (1988) further define the general null model for $\mathbf{m}$ as

$$H_0 : \mathbf{m} \in M_0 \qquad (3.16)$$

where $M_0 \subset M$ describes the set of all frequency vectors satisfying the constraints of the hypothesized model. As a consequence, the maximum likelihood method is said to be similar to minimizing (3.15) with regard to $\mathbf{m}\,\epsilon M_0$ using $\lambda = 0$. Equation (3.15) could be minimized for different values of λ; thus prompting the definition of the minimum power divergence estimate of $\mathbf{m}\,\epsilon M_0$ to be expressed as the $\widehat{\mathbf{m}}^{\lambda}$ which satisfies

$$I^{\lambda}(\mathbf{x} : \widehat{\mathbf{m}}^{(\lambda)}) = \inf_{\mathbf{m} \epsilon M_0} I^{\lambda}(\mathbf{x} : \mathbf{m}); \qquad -\infty < \lambda < \infty\,. \tag{3.17}$$

The estimate $\widehat{\mathbf{m}}^{\lambda}$ is unique since $I^{\lambda}(\mathbf{x} : \mathbf{m})$ is strictly convex. Read and Cressie (1988) show that a minimum power divergence estimator is a best asymptotic estimator (BAN). A BAN estimator must have the following three properties:

(a) The estimator converges to the true value of the evaluated parameter as $n \longrightarrow \infty$.

(b) They are asymptotically normally distributed.

(c) They are asymptotically efficient, since no other estimator can have a smaller variance as $n \longrightarrow \infty$.

The null hypothesis H_0 can be reparameterized by assuming that the vector $\boldsymbol{\pi} \in \boldsymbol{\Pi}_0$, is a function of s parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_s) \in \boldsymbol{\Theta}_0)$, where $s < k - 1$. This means that there is a function $\mathbf{f}(\boldsymbol{\theta})$ that maps each element of the subset $\boldsymbol{\Theta}_0 \subset R^s$ into the subset $\boldsymbol{\Pi}_0$. Thus H_0 can be reparameterized in terms of the pair $(\mathbf{f}, \Theta_0)$ as

$$H_0 : \text{There is a } \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \text{ such that } \boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}).$$

In order to ensure that the minimum power divergence statistic exists and converges to $\boldsymbol{\theta}$, as $n \longrightarrow \infty$, it is necessary to specify certain regularity conditions on $\mathbf{f}$ and $\boldsymbol{\Theta}_0$ under H_0.

The six regularity conditions of Birch (1964) are given in Read and Cressie (1988, Appendix A5). Birch's regularity conditions require that $\mathbf{f}$ satisfies certain smoothness requirements and that $\pi_i = f_i(\boldsymbol{\theta})$ is positive for all i. These conditions ensure that the model has under H_0 has s parameters and not fewer and are also necessary for deriving the asymptotic distribution of the power-divergence statistic under H_0.

The results presented earlier for the general hypothesis hold in this case by using the asymptotic normality of the BAN estimator $\widehat{\boldsymbol{\pi}}$. These can be stated as follows :

(i^*) If $\mathbf{X}$ is a random vector with a multinomial distribution $\text{Mult}_k(n, \boldsymbol{\pi})$ and $H_0 : \boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}) \in \boldsymbol{\Pi}_0$ for some $\boldsymbol{\theta}$ then $\sqrt{n}(\frac{\mathbf{X}}{n} - \widehat{\boldsymbol{\pi}})$ converges in distribution to a multivariate normal random vector as $n \longrightarrow \infty$ under the restriction that $\mathbf{f}$ satisfies Birch's regularity conditions and $\widehat{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_0$ is a BAN estimator of $\mathbf{f}(\boldsymbol{\theta})$.

(ii^*) $X^2 = \sum_{i=1}^{k} \dfrac{(X_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i}$ can be written as a quadratic form in $\sqrt{n}(\frac{\mathbf{X}}{n} - \widehat{\boldsymbol{\pi}})$, and X^2 converges in distribution, as $n \longrightarrow \infty$, to the quadratic form of the multivariate normal random vector in (i^*).

(iii^*) In the case of (i^*) and (ii^*) X^2 converges in distribution to a chi-squared distribution wth $k - s - 1$ degrees of freedom.

Read and Cressie (1988) prove that $X^2 = \sum_{i=1}^{k} \dfrac{(X_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i}$ has a χ^2_{k-s-1} distribution under H_0 and that

$$2nI^{\lambda}(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}) = 2nI^{1}(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}) + o_p(1),$$

i.e. the power-divergence statistic has asymptotically the same distribution as X^2. This means that

$P(2nI^{\lambda}(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}) \geq c) \longrightarrow P({\chi^2}_{k-s-1} \geq c)$, as $n \longrightarrow \infty$ for each $\lambda \in (-\infty, \infty)$, $c \geq 0$.

Thus for testing H_0 at the $100\alpha\%$ level of significance,
if $c = {\chi^2}_{1-\alpha}(k - s - 1)$ where $\alpha \in (0, 1)$, then

$$P\left(2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}\right) \geq {\chi^2}_{1-\alpha}(k - s - 1)(\alpha)\right) \longrightarrow \alpha \quad \text{as } n \longrightarrow \infty, \qquad (3.18)$$

Also, if $\widehat{\boldsymbol{\pi}}^{(\lambda)} \epsilon \boldsymbol{\Pi}_0$ is the minimum power divergence estimator of $\boldsymbol{\pi}$, then for each $\lambda \epsilon (-\infty, \infty)$ and each $\alpha \epsilon$ (0,1),

$$P\left(2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \widehat{\boldsymbol{\pi}}^{(\lambda)}\right) \geq \chi^2_{1-\alpha}(k-s-1)\right) \rightarrow \alpha \quad \text{as } n \rightarrow \infty. \qquad (3.19)$$

3.6 Case of Parameter Estimation

Goodness-of-fit tests are generally used to check if a set of data comes from a given distribution or class of distributions. However, another use is to observe and interpret associations or relationships between two or more random variables. The model of independence examines whether categorical variables are independent and this model will require parameter estimation.

The model of marginal homogeneity will also be discussed and it will be shown that the maximum likelihood estimates for the models of independence and marginal homogeneity are the same.

3.6.1 Independence and Homogeneity in Two Dimensional Contingency Tables

Consider a two-dimensional table comprising of r rows, c columns with rc cells obtained from the cross-classification of two variables say A and B. The r rows represent the r categories of the variable A, denoted by A_1, A_2,...,A_r whilst the c columns represent the c categories of variable B, denoted by B_1, B_2,...,B_c. The cell frequency in the ith row and jth column is denoted by x_{ij}.

The concept of marginal totals plays an important role in testing independence. The row total for row i is expressed as $x_{i+} = \sum_{j=1}^{c} x_{ij}$; where $i = 1, ..., r$ and the column total for column j is denoted by $x_{+j} = \sum_{i=1}^{r} x_{ij}$; where $j = 1, ..., c$.

These totals are referred to as the marginal totals for the frequency table. The total for all the cells is denoted by $n = x_{++} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}$. The total may also be expressed as $n = \sum_{i=1}^{r} x_{i+} = \sum_{j=1}^{c} x_{+j}$.

The study of independence involves the calculation of π_{ij}, which is the probability of observing an individual in $A_i \cap B_j$ or otherwise referred to as the joint probability for the cells.

Looking specifically at independence : two variables A and B are said to be statistically independent if

$$P(A_i \cap B_j) = P(A_i)P(A_j), \text{ for } i = 1, \ldots, r;\ j = 1, \ldots c.$$

For the two-dimensional frequency table, this can be expressed as

$$\pi_{ij} = \pi_{i+}\pi_{+j} \text{ for } fori = 1, ..., r;\ j = 1, ..., c,$$

where π_{i+} and π_{+j} are the unknown marginal probabilities.

The model of independence or no association is

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad , i = 1, ..., r;\ j = 1, ..., c, \tag{3.20}$$

Let the random variable X_{ij}, $i = 1, ..., r$; $j = 1, ..., c$ denote the observed frequency for the cell (i, j), and let π_{ij} denote the probability that an outcome is classified in cell (i, j). Then $\mathbf{X} = (X_{11}, X_{12}, ..., X_{1c}, X_{r1}, X_{r2}, ..., X_{rc})$ has a multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, ..., \pi_{1c}, ..., \pi_{r1}, \pi_{r2}, ..., \pi_{rc})$. The expected value of X_{ij} is $E(X_{ij}) = m_{ij} = n\pi_{ij}$ Hence (3.21) can be expressed in terms of the expected frequencies, as

$$H_0 : m_{ij} = n\pi_{i+}\pi_{+j}, \tag{3.21}$$

The unknown marginal probabilities must be estimated in order to test (3.21) and this is done by using the maximum likelihood estimates of π_{i+} and π_{+j}, namely $\dfrac{x_{i+}}{n}$ and $\dfrac{x_{+j}}{n}$, respectively. These estimates are obtained as follows:

If a multinomial sampling procedure is used, then the likelihood function is the multinomial probability function, i.e.

$$L(\boldsymbol{\pi}) = \frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} x_{ij}} \prod_{i=1}^{r}\prod_{j=1}^{c} \pi_{ij}^{x_{ij}}$$

The log likelihood function is

$$\log L(\boldsymbol{\pi}) = c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{ij}) \ ,$$

where c is a constant which is not a function of the π_{ij}. Under H_0, the log likelihood function is

$$\log L(\boldsymbol{\pi}) = c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{i+}\pi_{+j}) \ .$$

The maximum estimates of π_{i+} and π_{+j} are found by maximizing $\log L(\boldsymbol{\pi})$ subject to the constraints $\sum_{i=1}^{r}\pi_{i+} = 1$ and $\sum_{j=1}^{c}\pi_{+j} = c$. In order to do this consider the two Lagrange multipliers λ_1 and λ_2 and maximize the log likelihood function with the two constraints, i.e. maximize

$$\log L(\boldsymbol{\pi}) = c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{i+}\pi_{+j}) - \lambda_1\left(\sum_{i=1}^{r}\pi_{i+} - 1\right) - \lambda_2\left(\sum_{j=1}^{c}\pi_{+j} - 1\right)$$

$$= c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{i+}) + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{+j}) - \lambda_1\left(\sum_{i=1}^{r}\pi_{i+} - 1\right) - \lambda_2\left(\sum_{j=1}^{c}\pi_{+j} - 1\right)$$

with respect to π_{i+} and π_{+j}.

The partial derivative with respect to π_{i+} is

$$\frac{\partial \log L(\boldsymbol{\pi})}{\partial \pi_{i+}} = \sum_{j=1}^{c} x_{ij}\frac{1}{\pi_{i+}} - \lambda_1 \ ,$$

which must be set to zero. This gives

$$\sum_{j=1}^{c} x_{ij} \frac{1}{\pi_{i+}} = \lambda_1 \text{ or } \sum_{j=1}^{c} x_{ij} = \lambda_1 \pi_{i+}$$

Now

$$\sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} = \lambda_1 \sum_{i=1}^{r} \pi_{i+} = \lambda_1 \text{ , since } \sum_{i=1}^{r} \pi_{i+} = 1 \text{ .}$$

Hence $\lambda_1 = n$ and $\pi_{i+} = \frac{1}{n} \sum_{j=1}^{c} x_{ij} = \frac{x_{i+}}{n}$. Thus the maximum likelihood estimate is $\widehat{\pi}_{i+} = \frac{x_{i+}}{n}$. Similarly setting $\frac{\partial}{\partial \pi_{+j}} \log L(\boldsymbol{\pi}) = 0$ and solving for π_{+j} we get $\widehat{\pi}_{+j} = \frac{x_{+j}}{n}$.

Consequently, the expected value of X_{ij} is expressed :

$$\widehat{m}_{ij} = n \left(\frac{x_{i+}}{n}\right) \left(\frac{x_{+j}}{n}\right) = \frac{x_{i+} x_{+j}}{n}. \tag{3.22}$$

The null hypothesis in (3.22) can now be tested by using the power divergence family of statistics, defined from (3.12) as :

$$2I^{\lambda}(\mathbf{x} : \widehat{\mathbf{m}}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \left[\left(\frac{x_{ij}}{\widehat{m}_{ij}} \right)^{\lambda} - 1 \right] , \qquad -\infty < \lambda < \infty \tag{3.23}$$

with limits as $\lambda \to -1$ and $\lambda \to 0$ being used for $\lambda = -1$ and $\lambda = 0$.

The degrees of freedom for the chi-square distribution are:

$$k - s - 1 = (rc - 1) - (r - 1) - (c - 1) = (r - 1)(c - 1).$$

Thus the power-divergence statistic and X^2 have approximately a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

When the row totals are constrained by the sample layout, then the appropriate model in question is the model of homogeneity of proportions. The hypothesis of homogeneity of proportions may be expressed as

$$H_0 : m_{ij} = x_{i+} \pi_{+j}, \tag{3.24}$$

with the maximum likelihood estimate of π_{+j}, the marginal proportion $\frac{x_{+j}}{n}$. The maximum likelihood estimate for π_{+j} is found by maximizing

$$\log L(\boldsymbol{\pi}) = c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(x_{i+}) + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{+j})$$

subject to the constraint $\sum_{j=1}^{c} \pi_{+j} = 1$. This is done by introducing the Lagrange multiplier λ_3 such that $\lambda_3\left(\sum_{j=1}^{c}\pi_{+j} - 1\right) = 0$ and maximizing

$$\log L(\boldsymbol{\pi}) = c + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(x_{i+}) + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log(\pi_{+j}) - \lambda_3\left(\sum_{j=1}^{c}\pi_{+j} - 1\right)$$

with respect to π_{+j}.

$$\frac{\partial \log L(\boldsymbol{\pi})}{\partial \pi_{+j}} = \sum_{i=1}^{r} x_{ij}\frac{1}{\pi_{+j}} - \lambda_3 ,$$

which is set to zero, giving $\sum_{i=1}^{r} x_{ij} = \lambda_3\pi_{+j}$. Now

$$\sum_{j=1}^{c}\sum_{i=1}^{r} x_{ij} = \lambda_3\sum_{j=1}^{c}\pi_{+j} = \lambda_3 ,$$

since $\sum_{j=1}^{c}\pi_{+j} = 1$. Hence

$$\lambda_3 = n \text{ and } \widehat{\pi}_{+j} = \frac{\sum_{i=1}^{r} x_{ij}}{n} = \frac{x_{+j}}{n} .$$

Thus $\widehat{m}_{ij} = x_{i+}\widehat{\pi}_{+j} = \frac{x_{i+}x_{+j}}{n}$, which is the same result obtained for the maximum likelihood estimator for the independence model.

The sampling procedure adopted for model of homogeneity of proportions, is the product-multinomial model, with the row totals, x_{i+} fixed, while for the hypothesis of independence the full multinomial sampling procedure is assumed. The full-multinomial distribution has the total sample size, n fixed. Since the expressions for the maximum likelihood estimators for the test of homogeneity are the same as those for the test of independence, there is no difference between the two tests as long as the method of maximum likelihood is used to estimate the expected cell frequencies.

The minimum power divergence estimator for the independence model is found as follows:
Express H_0 : $\mathbf{m} \in M_0$ for the contingency table, where the π_{ij} must be estimated as,

$$H_0 : \boldsymbol{\pi} \in \boldsymbol{\Pi}_0, \tag{3.25}$$

where $\boldsymbol{\Pi}_0$ is a set of values for $\boldsymbol{\pi}$.

When testing independence for a two dimensional contingency table; $\boldsymbol{\Pi}_0 = \{\pi_{ij} : \pi_{ij} > 0;\ \sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij} = 1;\ \pi_{ij} = \pi_{i+}\pi_{+j}\}$. Estimation of $\boldsymbol{\pi}$ can be achieved by selecting the value of $\widehat{\boldsymbol{\pi}} \in \boldsymbol{\Pi}_0$ which is closest to $\frac{\mathbf{x}}{n}$ with regard to the measure $2nI^{\lambda}\left(\frac{\mathbf{x}}{\mathbf{n}} : \widehat{\boldsymbol{\pi}}\right)$; hence yielding the power divergence estimate defined in (3.27) ie : $\widehat{\boldsymbol{\pi}}^{(\lambda)}$ which fulfills

$$I^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}^{(\lambda)}\right) = \inf_{\boldsymbol{\pi} \in \boldsymbol{\Pi}_0} I^{\lambda}\left(\frac{\mathbf{x}}{n} : \boldsymbol{\pi}\right) \qquad -\infty < \lambda < \infty. \tag{3.26}$$

The chi-squared distribution, for large samples, is appropriate if one degree of freedom is subtracted for each parameter that is estimated.

The versatility of the power-divergence statistic is highlighted when fitting log-linear models to cross-classified data as well as the estimating of unknown model parameters.

3.6.2 Adopting the Loglinear Models Approach

Loglinear models are used mainly when at least two classification variables are used for a frequency table. For a $r \times c$ table, which classifies n subjects on two responses, recall that the joint probabilities π_{ij} for the cells is determined by the row and column marginal totals when the variables are independent. Hence $\pi_{ij} = \pi_{i+}\pi_{+j} \;\; i = 1, ..., r; \; j = 1, ..., c$ and the related expression for expected frequencies is $m_{ij} = n\pi_{i+}\pi_{+j}$ for all i and j. The model of independence (3.21) and the model of homogeneity (3.25) possess a linear structure when logarithms are applied to the expected cell frequencies. The resulting model is

$$\log(m_{ij}) = \log(n) + \log(\pi_{i+}) + \log(\pi_{+j}) \tag{3.27}$$

and

$$\log(m_{ij}) = \log(n) + \log\left(\frac{x_{i+}}{n}\right) + \log(\pi_{+j}). \tag{3.28}$$

Further generalising simplifies l_{ij} to

$$l_{ij} \equiv \log(m_{ij}) = u + u_{1(i)} + u_{2(j)} \;, \tag{3.29}$$

where

$$u = \frac{l_{++}}{rc} \;, \; u + u_{1(i)} = \frac{l_{i+}}{c} \;, \; u + u_{2(j)} = \frac{l_{+j}}{r}$$

with departures from u described by $u_{1(+)} = u_{2(+)} = 0$.

In keeping with the previous models namely (3.22) and (3.25), the model for independence or homogeneity has $(r-1)(c-1)$ degrees of freedom.

Loglinear Models : Two Dimensional Tables

When the number of parameters equals the number of cells in a contingency table, the model is described as being saturated. For the two dimensional case,

$$l_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \tag{3.30}$$

summarizes the most general loglinear model. The interaction term $u_{12(ij)}$ identifies deviation from independence (or homogeneity) and further fulfills the condition that $u_{12(i+)} = u_{12(+j)} = 0$, for every i and j. Now, taking $u_{12(ij)} = l_{ij} - \frac{l_{i+}}{c} - \frac{l_{+j}}{r} + \frac{l_{++}}{rc}$ jointly with the constraints $u_{1(+)} = u_{2(+)} = 0$ and $u_{12(i+)} = u_{12(+j)} = 0$ yields rc parameters and hence has $rc - rc = 0$ degrees of freedom, hence the description of a saturated model.

Loglinear Models For Three Dimensional Tables

Expanding the saturated two dimensional loglinear model to three dimensions gives

$$l_{ijl} = \log(m_{ijl}) = u+u_{1(i)}+u_{2(j)}+u_{3(l)}+u_{12(ij)}+u_{13(il)}+u_{23(jl)}+u_{123(ijl)} \ , \quad (3.31)$$

where m_{ijl} represents the expected cell frequency in the ith row, jth column and lth layer of the table.

In addition to the constraints applied to the two dimensional case, further constraints that apply include : $u_{3(+)} = 0$, $u_{13(i+)} = u_{13(+l)} = u_{23(j+)} = u_{23(+l)} = 0$ and $u_{123(ij+)} = u_{123(i+l)} = u_{123(+jl)} = 0$.

The model for which variables 1, 2 and 3 are completely independent of one another is the ideal model and is described as follows :

$$H_0 : l_{ijl} = u + u_{1(i)} + u_{2(j)} + u_{3(l)}. \quad (3.32)$$

If the above model is not suitable because dependence exists between variables then one of the following three models can be used :

When variables 1 and 2 are dependent on each other but are jointly independent of variable 3, the appropriate model is

$$H_0 : l_{ijl} = u + u_{1(i)} + u_{2(j)} + u_{3(l)} + u_{12(ij)} \ . \quad (3.33)$$

The second model is

$$H_0 : l_{ijl} = u + u_{1(i)} + u_{2(j)} + u_{3(l)} + u_{12(ij)} + u_{13(il)} \ , \tag{3.34}$$

is suitable when variables 1 and 2 are dependent and variables 1 and 3 are dependent whilst variables 2 and 3 are conditionally independent given variable 1.

Lastly, the third option is

$$H_0 : l_{ijl} = u + u_{1(i)} + u_{2(j)} + u_{3(l)} + u_{12(ij)} + u_{13(il)} + u_{23(jl)} \ , \tag{3.35}$$

which has no three factor interaction but there is pairwise association between variables 1, 2 and 3.

Although other methods have been used for analyzing cross-classified data, it was found that the loglinear model is preferred in most cases with the main reason being the abundant availability of computer programs for analysing such data.

3.7 Choosing the Appropriate Loglinear Model

Model fit can be examined through calculation of the goodness-of-fit test statistic from the power-divergence family $2I^{\lambda}(\mathbf{x} : \widehat{\mathbf{m}}\ )$ in (3.16). A prerequisite step is to calculate the expected cell frequencies, $\mathbf{m}$ for the hypothesized loglinear model. A model is thought to be an appropriate choice if the value of $2I^{\lambda}(\mathbf{x} : \widehat{\mathbf{m}})$ is not "too large". When the hypothesized model is true, it is deduced that the test statistic follows a chi-squared distribution. The degrees of freedom is equal to the difference between the total number of cells in the table and the number of parameters minus a further one degree of freedom. Hence, the description "too large" encompasses values that lie in the upper tail of the chi-squared distribution. However, this assumption is not always suitable.

The choice of λ depends on the kind of variations from the null model, for example when the minimum expected cell frequency is no less than one, Read & Cressie suggest that $\lambda = \frac{2}{3}$ is the appropriate choice.

By default, fitting loglinear models requires that all lower order effects be incorporated before a higher order effect is included. In other words, a model can only use the interaction term u_{123} if all the terms u_{12}, u_{13}, u_{23} , u_1, u_2, u_3 and u are already included in the model. This is called a *hierarchical model* and it is frequently used unless a nonhierarchical model is deemed fit.

Selection of a model could be based on one of the following techniques : stepwise selection, selection based on measures of marginal and partial association or selection based on standardized parameter estimates. A further expansion on these selection methods will be presented shortly for a model of order v which contains all the interaction terms.

Stepwise selection employs three steps. Firstly, for a t dimensional table, all uniform models of order v, that is, a model which includes all interaction terms involving v variables , for $1 \leq v \leq t$ must be filled. Secondly, the smallest value of v for which the uniform model of order v presents the best fit to the model must be identified. Thirdly, either a forward, stepwise selection or a backward elimination method must be used to choose the interaction terms involving v variables for the appropriate model.

Selection using measures of marginal and partial association involves examining the terms to ensure that only the necessary few terms are included in the model. A partial association statistic looks at the difference between models at each stage of the deletion of an interaction term from a uniform model of some order v. Thus, in the case of a three dimensional table, the partial association of u_{12} is measured by comparing the difference between the fit of the hierarchical model containing the terms u_{12}, u_{13}, u_{23} with the model in which u_{12} is deleted.

On the other hand, determining the effect of a given interaction term from a marginal table by collapsing over all variables not included in the interaction describes the procedure based on marginal association. Marginal association of u_{12} is evaluated by finding the goodness-of-fit statistic for testing $u_{12} = 0$ for variables 1 and 2.

If significant answers are obtained for both the marginal and partial association tests, this would imply that the term in question should be included in the model. On the other hand, if both tests are insignificant, then the term should be excluded. However, if one test is significant whilst the other is insignificant then further investigations need to be carried out on the model.

Lastly, selection based on standardized parameter estimates, requires that all possible terms in the saturated model be estimated first and then their standardized estimates should be compared. The terms with the largest standardized estimates should be included in the model.

Usage of the three methods mentioned above does not guarantee that the "correct" model is chosen; therefore all prior knowledge about the problem and the data should be used and the fit of the model should be assessed thereafter.

Chapter 4

Evaluating Model Fit using the Power Divergence Statistic

Acceptance or rejection of a model is determined by either calculating the 100% α value or examining how efficient the test statistic is. The initial sections of this chapter address the above criteria for large samples with increasing sample size but fixed cell number. The influence of sparseness assumptions yield different findings and the usual conditions defining acceptance or rejection of a model are no longer appropriate. Hence, discussions on the sparse sample case are presented. The D^2 goodness-of-fit statistic proposed by Daniel Zelterman (1987) is introduced and deliberated in section 4.3. Thereafter Koehler and Larntz's (1980) proposed goodness-of-fit statistic is explained and findings of their study under sparseness assumptions are presented. In section 4.5, a goodness of fit statistic introduced by Simonoff (1982) is highlighted and discussed regarding issues concerning the effect of parameter estimation as well as sparseness assumptions. Thereafter, a discussion on goodness-of-fit tests for loglinear models under sparse conditions undertaken by Koehler (1986) is summarized addressing topics of independence and accuracy assessments. A further section investigates the use of jackknife and bootstrap methods in sparse multinomials. The study undertaken by Simonoff (1986) involved the use of nonparametric techniques to estimate variances. Finally the last section in this chapter looks at a comparative study between Pearson X^2 and the loglikelihood ratio statistic G^2 under conditions of sparseness.

4.1 Influence of Sparseness Assumptions on Significance Levels and Accuracy

The common experience when working with experiments in the social and biological sciences is that one frequently obtains large sparse arrays full of zeros and ones. With the number of multinomial cells, k, increasing without limit, there is a change in the dimension and in the probability space. Therefore, the property that the expected cell frequencies become large with n, as was assumed under asymptotic theory where the cell probabilities were fixed, cannot be assumed.

It is highlighted by Hoeffding (1965, p. 371−372) that equivalence or supremacy of the loglikelihood ratio test G^2 $(\lambda = 0)$ over Pearson's X^2 $(\lambda = 1)$ "are subject to the limitation that k is fixed or does not increase rapidly with n. Otherwise the relation between the two tests may be reversed." It is further found that X^2 and G^2 have different asymptotic normal assumption distributions when the rate at which $k \rightarrow \infty$ is constrained ($\frac{n}{k}$ must remain finite). For a fixed k , the asymptotic distributions for both statistics is chi-squared with degrees of freedom comparable to k. Hence the property of asymptotic normality. With an increase in the degrees of freedom, the chi-squared variable gets closer to a normal variable in distribution.

4.1.1 Case of the Equiprobable Model

Although $\mathbf{X}$ is a multinomial probability vector since k is no longer fixed, the cell probabilities and the sample size need to be expressed as functions of k.

Change in notation yields $\mathbf{X}_k = (X_{1k}, \ldots, X_{kk}) \sim \text{Mult}_k(n_k, \boldsymbol{\pi}_k)$ and the null model is expressed as

$$H_0 : \boldsymbol{\pi}_k = \frac{1}{k}\mathbf{1}, \tag{4.1}$$

where $\mathbf{1}$ = (1,1,...,1) is a $1 \times k$ vector.

Using the sparseness assumptions of Holst (1972) the following result is stated and proved (Read and Cressie, 1988 : p. 58). Suppose $n_k \to \infty$ as $k \to \infty$ so that $\frac{n_k}{k} \to a \quad (0 < a < \infty)$.

Then for any $c \geq 0$

$$P\left[\frac{\left(2n_k I^{\lambda}(\frac{\mathbf{X}_k}{n_k} : \frac{1}{k}\mathbf{1}) - {\mu_k}^{(\lambda)}\right)}{{\sigma_k}^{(\lambda)}} \geq c\right] \to P[N(0,1) \geq c], \qquad \text{as } k \to \infty \tag{4.2}$$

when hypothesis (4.9) holds and $\lambda > -1$.

Under the assumption that $\mathbf{X}_k$ is a multinomial random vector, the statistic

$$S_k = \sum_{i=1}^{k} h_k(X_{ik}, \frac{i}{k}), \tag{4.3}$$

where $h_k(\cdot, \cdot)$ is a real measurable function, is used so that Holst's result reads as follows (Read and Cressie, (1988, p. 174)) :

Theorem 4.1

"Define

$$\mu_k = \sum_{i=1}^{k} E[h_k(Y_{ik}, \frac{i}{k})]$$

and

$$\sigma_k{}^2 = \sum_{i=1}^{k} \text{var}[h_k(Y_{ik}, \frac{i}{k})] - \frac{[\sum_{i=1}^{k} cov[Y_{ik}, h_k(Y_{ik}, \frac{i}{k})]]^2}{n}, \tag{4.4}$$

where the $Y_{ik}^{'}s$ are independent Poisson random variables with means $n_k \pi_{0ik}$ and $i = 1, \ldots, k$.

Assume the following :

(a) n_k and $k \to \infty$, such that $\dfrac{n_k}{k} \to a$ $(0 < a < \infty)$;

(b) $k\pi_{0ik} \leq c < \infty$ for some nonnegative number c; $i = 1, ..., k$ and all k;

(c) $\mid h_k(v, x) \mid \leq \alpha \text{ exp } (\beta\nu)$ for $0 \leq x \leq 1$; $\nu = 0, 1, 2, ..., \alpha$ and β real;

(d) $0 < \lim\limits_{n_k \to \infty} \inf \dfrac{\sigma_k{}^2}{n_k} \leq \lim\limits_{n_k \to \infty} \sup \dfrac{\sigma_k{}^2}{n_k} < \infty$.

Then $\dfrac{(S_k - \mu_k)}{\sigma_k}$ is asymptotically a standard normal random variable, as $k \to \infty$."

The following theorem shows that although S_k is a sum of dependent random variables thus preventing the standard central limit theorems from being applied, it is possible under certain instances to ensure that S_k has the same asymptotic limit as $S_k{}^* = \sum_{i=1}^{k} h_k(Y_{ik}, \frac{i}{k})$, where the $Y_{ik}^{\prime}s$ are independent Poisson random variables and have the same means as the multinomial X_{ik}'s.

Cressie and Read (1984) applied this theorem to the equiprobable hypothesis and a summary of their findings is presented. For the equiprobable hypothesis (4.1) condition (b) of the theorem is satisfied immediately since $\pi_{0ik} = \frac{1}{k}$ for each $i = 1, ..., k$.

Now defining $h_k(X_{ik}, \frac{i}{k})$ as :

$$h_k\left(X_{ik}, \frac{i}{k}\right) = \begin{cases} \dfrac{2}{\lambda(\lambda+1)}\dfrac{n_k}{k}\left[\left(\dfrac{X_{ik}}{\frac{n_k}{k}}\right)^{\lambda+1} - 1\right]; & \lambda \neq 0,\ \lambda > -1 \\ 2X_{ik}\log\left(\dfrac{X_{ik}k}{n_k}\right); & \lambda = 0, \end{cases} \tag{4.5}$$

for any given $\lambda > -1$, condition (c) of the theorem is satisfied and hence condition (d) is also satisfied. Asymptotic normality of the power divergence statistic (4.2) is obtained by substituting (4.5) for $h_k(\cdot,\cdot)$ into (4.4) with the $Y_{ik}'s$ being independent and identically distributed Poisson variables with mean $\frac{n_k}{k}$.

This indicates further that under the sparseness assumptions, the members of the power-divergence family are no longer asymptotically equivalent. When a zero cell frequency is observed then $2n_kI^\lambda\left(\frac{\mathbf{x}_k}{n_k} : \boldsymbol{\pi}_k\right)$ is undefined for $\lambda \leq -1$, since positive powers of $\frac{n_k\pi_{ik}}{x_{ik}}$ are needed where x_{ik} =0. Using a similar approach it is found that $\mu_k{}^{(\lambda)}$ and $[\sigma_k{}^{(\lambda)}]^2$ are not defined for $\lambda \leq -1$ because Y_k has a positive probability of being 0.

4.1.2 Accuracy of Possible Alternatives to the Equiprobable Model

Read and Cressie (1988) observed that the power-divergence family members were asymptotically equally efficient when testing against local alternative models that converge to the null but under the sparseness assumptions this is no longer the trend.

When testing the model

$$H_0 : \boldsymbol{\pi}_k = \frac{1}{k}\mathbf{1} \tag{4.6}$$

against

$$H_{1,k} : \boldsymbol{\pi}_k = \frac{1}{k}\mathbf{1} + \frac{\boldsymbol{\delta}}{n_k{}^{\frac{1}{4}}} \tag{4.7}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_k)$ and $\sum_{i=1}^{k} \delta_i = 0$, the power-divergence statistic has a normal distribution under $H_{1,k}$. Pearson's X^2 test ($\lambda = 1$) was found to be maximally efficient for the power-divergence family for testing (4.6). The loss in efficiency for using any λ other than λ =1 was calculated by Cressie and Read (1988) and it was found that for $\lambda > 3$, the accuracy swiftly declined whereas $-1 < \lambda \leq 3$ seemed to produce the best results. It was further found that for n_k large as compared to k, the variations in the λ values were minimal.

Comparisons between Pearson's X^2 (λ =1) and the loglikelihood ratio $G^2(\lambda = 0)$ tests under the equiprobable null hypothesis were carried out by Koehler and Larntz (1980) using Monte Carlo power comparisons when $\frac{n_k}{k}$ lies in the range $[\frac{1}{4}$ to 5]. Their findings were consistent with Read and Cressie that X^2 is slightly more effective for local alternatives. On the other hand for the case when one or two cells have very small probabilities and the rest are almost equivalent, G^2 was observed to yield better findings. Alternatively, with an increase in the number of near-zero alternative probabilities, X^2 prevails as the better statistic. This finding substantiates the advice by Read and Cressie (1988) to use $\lambda = \frac{2}{3}$.

4.2 Zelterman's D^2 Goodness of Fit Statistic

A goodness-of-fit statistic D^2 proposed by Daniel Zelterman (1987) for applications to large sparse multinomial distributions is discussed in this section. Comparing the Pearson's X^2 and the D^2 statistics he concluded that both were approximately normally distributed when the sample size is not large in relation to the number of categories. Further findings include that the test based on D^2 possesses reasonable power when the X^2 test exhibits properties of biasedness.

The D^2 statistic originates from the loglikelihood ratio statistic for testing a sequence of multinomial hypotheses against a sequence of local alternatives which are Dirichlet mixtures of multinomials.

Zelterman's findings and conclusions are as follows : when sample size n was large, in comparison with k, the number of multinomial categories, X^2 and $D^2 + k$ function closely to chi-squared random variables and vary minimally under the null hypothesis with similar conclusions being yielded quite often.

Under conditions of sparseness arising from n and k both large, both X^2 and D^2 behave like normal random variables with means and variances that differ from those of the chi-squared distribution. Also, in the sparse distribution case, X^2 and D^2 are not alike and X^2 rejects the null hypothesis very seldomly.

4.2.1 General Mathematical Notation of the Statistic

For each $k = 1, 2 \ldots$, let $\mathbf{x}_{(k)} = \{x_{i(k)} : i = 1, \ldots, k\}$, represent a multinomial vector with probability vector $\boldsymbol{\pi}_{(k)} = \{\pi_{i(k)} > 0 : \sum_{i=1}^{k} \pi_{i(k)} = 1\}$ and parameters $n_{(k)} = \sum_{i=1}^{k} x_{i(k)}$ when $k = 1, 2, \ldots$. The null hypothesis $H_0 : \boldsymbol{\pi}_{(k)} = \boldsymbol{\pi}_{0(k)}$ with $\boldsymbol{\pi}_{0(k)} = \left\{ \pi_{0i(k)} > 0 : \sum_{i=1}^{k} \pi_{0i(k)} = 1 \right\}$, describes a completely specified sequence of probability vectors. The alternative hypothesis is $H_a : \boldsymbol{\pi}_{(k)} = \mathbf{q}_{(k)} = \{q_{i(k)}\}$.

The study undertaken considered each $\mathbf{q}_{(k)}$ as an unobservable realization on a set of random variables assuming values near $\boldsymbol{\pi}_{(k)}$, with high probability. It is assumed that for k sufficiently large, the $q_{i(k)}$ can be expressed as

$$q_{i(k)} = \pi_{i(k)}(1 + \tau_{i(k)} {n_{(k)}}^{-\frac{1}{2}}), \tag{4.8}$$

The constants $\tau_{i(k)}$ are such that $max_i \mid \tau_{i(k)} \mid$ is bounded as $k \to \infty$ and

$\sum_{i=1}^{k} \tau_{i(k)}\pi_{i(k)} = 0$ for all k. The statistic obtained from the theorem is asymptotically only a function of the variances of mixing distributions under asymptotic conditions. For unspecified mixing variances a D^2 statistic is proposed. The D^2 statistic is not a member of the power-divergence family of statistics studied by Cressie and Read (1984).

Extending the loglikelihood ratio for tests of the alternative hypotheses with no estimated parameters gave rise to the Dirichlet mixtures as a special case. Zelterman's derivation of the statistic involved testing a sequence of simple hypotheses against a similar sequence of multinomial alternative hypotheses. He found that the variances of the mixing distributions decrease to zero as the number of multinomial categories increases without limit.

Let $n_{(k)}\pi_{i(k)} = m_{i(k)}$, where $\mathbf{m}_{(k)} = \{m_{i(k)} > 0\}$ are real known constants. Consider unconditional tests for known values of $\mathbf{m}$. For the sequence under H_a, let $\pi_{i(k)}$ be expressed as

$$\pi_{i(k)} = \frac{\theta_{i(k)}}{\theta_{1(k)} + \cdots + \theta_{k(k)}}, \quad i = 1, 2, \ldots, k\ , \tag{4.9}$$

where $\theta_{i(k)}$ described unobservable realisations of mutually independent, non-negative valued random variables each possessing distribution functions $H_{i(k)}(\cdot)$.

Under the assumption that the distribution functions $H_{i(k)}$ satisfy

$\int_0^\infty \theta dH_{i(k)}(\theta) = m_{i(k)}$, and $\mu_{i(k)} = \int (\theta - m_{i(k)})^6 \text{ dH }_{i(k)}(\theta)$ being finite for all $i = 1, \ldots, k$, and $k = 1, 2, \ldots$; and defining ${\sigma_{i(k)}}^2 = \int (\theta - m_{i(k)})^2 dH_{i(k)}(\theta)$ and $V_k = \frac{1}{2}\sum_{i=1}^{k} \frac{{\sigma_{i(k)}}^4}{{m_{i(k)}}^2}$, an important theorem which applies to many mixing distributions and more significantly to applications in the Dirichlet multinomial mixture will be stated. The proof however can be found in Zelterman (1986).

Theorem 4.2

Suppose that

(i) $0 < \liminf V_k \leq \limsup V_k < \infty$,

(ii) $\lim\limits_{k} \max\limits_{i} \dfrac{m_{i(k)}}{n_{(k)}} = 0$,

(iii) $\lim\limits_{k} \sum\limits_{i=1} \mu_{i(k)} = 0$,

(iv) and for some $\epsilon > 0$ and all k sufficiently large $\min\limits_{i} m_{i(k)} > \epsilon$.

Then the loglikelihood ratio for testing these hypotheses satisfies

$$\log \Lambda_k = \frac{1}{2} \sum_{i=1}^{k} \frac{[(x_{i(k)} - m_{i(k)})^2 - x_{i(k)}]\sigma_{i(k)}{}^2}{m_{i(k)}{}^2} - \frac{1}{2} V_k + o_{P_0}(1). \tag{4.10}$$

The first term of (4.10) has mean o(1) and variance $V_k + o(1)$.

Zelterman also mentions the following finding in his article : if $\{H_{i(k)}(\cdot)\}$ represents independent gamma distributions with means $m_{i(k)}$ and variances $\sigma_{i(k)}{}^2$, then assuming that for some value of k, there exists a real number $M > 0$, which is not a function of k or i, he has shown in detail that under certain conditions, the conditions of the theorem are satisfied and that the loglikelihood ratio, for the gamma variables, satisfies (4.10).

A case of interest is where $\dfrac{\sigma_{i(k)}{}^2}{m_{i(k)}}$ is independent of i. In this case the distribution of $\boldsymbol{\pi}_{(k)}$ in (4.9) is Dirichlet for every k. Dirichlet mixtures of multinomials take the form of posterior distributions in the Bayesian analysis of multinomials with Dirichlet priors.

Zelterman further reports that the likelihood ratio (4.10) for testing Dirichlet mixtures of multinomials satisfies

$$\log \Lambda_k = \frac{1}{2} c k^{-\frac{1}{2}} \sum_{i=1}^{k} \frac{[(x_{i(k)} - m_{i(k)})^2 - x_{i(k)}]}{m_{i(k)}} - \frac{1}{4} c^2 + o_{P_0}(1), \tag{4.11}$$

when $k^{\frac{1}{2}} \frac{\sigma_{i(k)}{}^2}{m_{i(k)}} = c^2$ for all $i = 1, 2, \ldots, k$ and $k = 1, 2, \ldots$. The term $o_{P_0}(1)$ in (4.11) is with respect to the sequence of multinomial null hypotheses.

Zelterman defines

$$D^2 = \sum_{i=1}^{k} \frac{\left[(x_{i(k)} - m_{i(k)})^2 - x_{i(k)}\right]}{m_{i(k)}}, \tag{4.12}$$

He observed that for k large, $(2k)^{-\frac{1}{2}} D^2$ behaves approximately as a standard normal variable under multinomial sampling.

Also, when $X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$, then $D^2 = X^2 - \sum_{i=1}^{k} \frac{x_i}{m_i}$, and under the multinomial sampling, $E\left(\sum_{i=1}^{k} \frac{x_i}{m_i}\right) = k$ and $\text{var}\left(\sum_{i=1}^{k} \frac{x_i}{m_i}\right) = \sum_{i=1}^{k} m_i{}^{-1} - \frac{k^2}{n}$. Thus X^2 and $D^2 + k$ are differ slightly from one another when all m_i are nearly equal or when few of the m_i are small.

Writing

$$D^2 = \sum_{i=1}^{k} \frac{\left[(x_i - \frac{1}{2}) - m_i\right]^2}{m_i} - k - \frac{1}{4} \sum_{i=1}^{k} m_i{}^{-1},$$

provides another relationship between D^2 and X^2, where $\frac{1}{2}$ must be subtracted from each observation prior to testing.

He concluded that G^2 and D^2 may have better normal approximations than X^2 when the data is sparse due to the reduced skewness and kurtosis.

When $x_i = 1$ and the corresponding mean m_i is small, then the contribution to D^2 is

$$\frac{(1 - m_i)^2 - x_i}{m_i} = \frac{1 - 2m_i + m_i^2 - x_i}{m_i}$$

$$= m_i - 2 + \tfrac{1}{m_i} - \tfrac{x_i}{m_i},$$

which if m_i is small will be $m_i - 2$ and similarly for G^2, the contribution for an $x_i = 1$, is $-2\log m_i$, but the contribution to X^2 is ${m_i}^{-1} - 2 + m_i$, which could be very large.

4.2.2 Testing with Estimated Parameters

Zelterman further addresses the topic of testing for independence of rows and columns. Let $\mathbf{x}_{(k)} = \{x_{ij(k)} : i = 1, \ldots, r_k;\ j = 1, \ldots, c_k\}$ represent a multinomial vector with $r_{(k)} \times c_{(k)}$ cells and parameter $n_{(k)} = \sum\limits_{ij} x_{ij(k)}$ and $\pi_{ij(k)} > 0$, where $\sum\limits_{ij} \pi_{ij(k)} = 1$.

The mean of $x_{ij(k)}$ under the hypothesis of independence is

$$E_0(x_{ij(k)}) = n_{(k)}\pi_{i+(k)}q_{jk}\ , \tag{4.13}$$

with $\pi_{i+(k)} > 0$ and $\pi_{+j(k)} > 0$, being unknown probabilities such that $\sum\limits_{i} \pi_{i+(k)} = \sum\limits_{j} \pi_{+j(k)} = 1$.

Inference on the model of independence is conditional on marginal totals defined by $x_{i+(k)} = \sum\limits_{j} x_{ij(k)}$ and $x_{+j(k)} = \sum\limits_{i} x_{ij(k)}$. Defining $I =$ the number of $x_{i+(k)}$ which are nonzero and $J =$ the number of nonzero $x_{+j(k)}$ for each $k = 1, 2, \ldots$ and applying Theorem 4.2, if $\boldsymbol{\pi}_{(k)} = \pi_{ij(k)}$ behaves like a Dirichlet vector under the simple hypotheses, then a test with maximum asymptotic power rejects H_0 for large values of

$$D^2 = \sum_{ij} \frac{\left[(x_{ij(k)} - m_{ij(k)})^2 - x_{ij(k)}\right]}{m_{ij(k)}}, \tag{4.14}$$

where $m_{ij(k)} = n_{(k)}\pi_{i+(k)}\pi_{+j(k)}$.

If the $m_{ij(k)}$ are unknown, then the following statistic can be considered :

$$\widehat{D}^2 = \sum^{*} \frac{\left[(x_{ij(k)} - \widehat{m}_{ij(k)})^2 - x_{ij(k)}\right]}{\widehat{m}_{ij(k)}}, \tag{4.15}$$

where $\widehat{m}_{ij(k)} = \frac{x_{i+(k)}x_{+j(k)}}{n_{(k)}}$ is the maximum likelihood estimate of $m_{ij(k)}$, as was shown in chapter 3. The * in the summation sign above represents the summation over only those values of i and j for which $\widehat{m}_{ij(k)}$ is positive.

Zelterman (1983, p. 628) showed that the "asymptotic correlation between D^2 and $\widehat{D}^2$ is unity under the null hypothesis" under general conditions of $r_{(k)}$ and $c_{(k)}$ both being proportional to k and all $m_{ij(k)}$ bounded above and away from zero. Hence it can be concluded that $\widehat{D}^2$ will be approximately normally distributed and have maximal asymptotic power when D^2 does.

4.3 Koehler and Larntz's Goodness-of-fit Statistics for Sparse Multinomials

Koehler and Larntz (1980) question the validity of the traditional asymptotic rules with regard to expected cell frequency and goodness of fit tests. They mention the following example Koehler and Larntz (1980, p. 336) "in a multidimensional contingency table analysis the full table is often collapsed over categories and/or variables to avoid problems caused by small expected frequencies. Thus for larger samples sizes, instead of increasing the expected frequencies, the analyst increases the number of variables in the table."

Holst (1972, p. 137) expresses that for the goodness-of-fit problem of examining whether a sample has come from a given population, "it is rather unnatural to keep k fixed when $n \rightarrow \infty$; instead we should have that $k \rightarrow \infty$ when $n \rightarrow \infty$". These ideas were echoed by Bishop, Fienberg and Holland (1975, p. 416) in the following description "Typically, multinomial data arrive in the form of a cross-classification of discrete variables. In many situations there are a large number of variables which can be used to cross-classify each observation, and if all variables are used the data would be spread too thinly over the cells in the resulting multidimensional contingency table. Thus, if the investigator uses a subset of the variables to keep the average number of observations from becoming too small; he is in effect choosing k so that $\frac{n}{k}$ is moderate." Koehler and Larntz look at goodness-of-fit when the number of cells expand with a similar growth in sample size.

4.3.1 A Look at Asymptotic Normality

$(X_1, X_2, ..., X_k)$ is the multinomial random vector described in the previous chapter with probability parameter $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_k)$ such that $n = \sum_{i=1}^{k} X_i$ and $1 = \sum_{i=1}^{k} \pi_i$. Under their investigation of asymptotic normality, Koehler and Larntz (1980) allow the number of cells to expand with a similar increase in sample size, and find that both ${X_k}^2$ (3.3) and ${G_k}^2$ (3.4) possess asymptotic normal distributions under the conditions that allow both n and k to become large without essentially forcing $\min_{1 \leq i \leq k} n\pi_i \to \infty$.

The test for a uniform distribution described on a fixed interval which is divided into a number of subintervals of equal length was considered. If some specific expected frequency λ is needed for each subinterval, then k subintervals are used for a sample size of n, where k is chosen such that $\frac{n}{k}$ is close to λ. An increase in n would in effect lead to an increase in k.

To examine the limiting distribution of goodness-of-fit statistics for multinomials of increasing dimension, a sequence of Poisson random vectors is defined since the asymptotic moments are expressed by use of independent Poisson frequencies. Thus for each multinomial vector $(X_{1k}, X_{2k}, ..., X_{kk})$; $(Y_{1k}, Y_{2k}, ..., Y_{kk})$ is described as a vector of independent Poisson random variables such that $E(Y_{ik}) = E(X_{ik})$.

Morris (1966, 1975) generalized a conditioning argument by Steck (1957) and hence achieved a central limit theorem for sums of functions of multinomial frequencies by adopting a method that demanded that sums of functions of independent Poisson frequencies possess a limiting normal distribution. Thus, the asymptotic normality of the sum under the multinomial distribution can be found by conditioning on the sum of independent Poisson frequencies.

The case of the null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ being true was considered. The necessary and sufficient conditions for asymptotic normality as $k \rightarrow \infty$ were firstly that $\min\limits_{1 \leq i \leq k} \pi_{ik} = o(1)$ as $k \rightarrow \infty$ and secondly that $n_k \pi_{ik}$ is uniformly bounded below by some constant. Koehler and Larntz then showed that when the null hypothesis is true, the asymptotic mean for the Pearson statistic is described as

$$\mu_{P,k} = k, \tag{4.16}$$

whilst

$$\sigma_{P,k}{}^2 = 2k + \sum_{j=1}^{k} \frac{(1 - k^{-1}\pi_{jk})}{n_k \pi_{jk}}. \tag{4.17}$$

It is further stated that exact moments for the Pearson statistic were obtained by Haldane (1937) and is expressed as

$$E({X_k}^2) = \mu_{P,k} - 1 \tag{4.18}$$

and

$$\operatorname{var}({X_k}^2) = \sigma_{P,k}{}^2 - 2\left[1 + \frac{k-1}{n_k}\right]. \tag{4.19}$$

Another finding is that $\sigma_{P,k}{}^2$ and hence $\operatorname{var}({X_k}^2)$ can be greater than the chi-squared variance on $k - 1$ degrees of freedom when the expected frequencies are not all equal.

The asymptotic moments for the likelihood ratio statistic were also described by using independent Poisson random variables. By defining the Poisson information kernel by

$$I(y,m) = \begin{cases} y\log(\frac{y}{m}) - y + m, & \text{if } y > 0 \\ \\ m, & \text{if } y = 0 \end{cases} \tag{4.20}$$

Koehler expresses the first two asymptotic moments as

$$\mu_{LR,k} = 2\sum_{i=1}^{k} E[I(Y_{jk}, n_k\pi_{jk})] \tag{4.21}$$

and

$$\sigma_{LR,k}{}^2 = 4\sum_{j=1}^{k} \text{var}[I(Y_{jk}, n_k\pi_{jk})] - n_k\gamma_k{}^2 \tag{4.22}$$

where $\gamma_k = \frac{2}{n_k}\sum_{j=1}^{k} \text{cov}[I(Y_{jk}, n_k\pi_{jk}), Y_{jk}]$.

Graphs for the expected value and variance of the Poisson Information Kernel as a function of the expected cell size and covariance between the Poisson Information Kernel and the observed value as a function of expected cell size were generated and presented in Koehler and Larntz's paper.

The graph of E[I(Y,m)] for the Poisson random variable Y with mean m, exhibits a swift descent of E[I(Y,m)] as $m \to 0$ which reinforces that $\mu_{LR,k}$ can be much smaller than the chi-squared mean when many expected frequencies are smaller than one half. Another observation is that when most expected frequencies are between 1 and 5, $\mu_{LR,k}$ is considerably bigger than $k-1$ and the mean of the likelihood ratio statistic is close to $k-1$ when almost all expected frequencies are large.

The graphs of $\text{var}[I(Y,m)]$ and $\text{cov}[I(Y,m),Y]$ show that the asymptotic variance can be much smaller than $2(k-1)$ when expected frequencies are smaller than one. On the other hand, it is larger than $2(k-1)$ when most expected frequencies are moderate. These findings are warnings that the chi-squared approximation for the likelihood ratio statistic may result in overblown critical levels when most expected frequencies are moderate and very cautious critical values when most expected frequencies are smaller than one-half.

Pearson and likelihood ratio statistics are found to possess different limiting normal distributions as $k \rightarrow \infty$ because of the varying influence of the very small observed counts on the statistics. Koehler and Larntz (1980, p. 338) state "For a cell with an expected frequency larger than one, an observed count of zero or one makes a larger minimum contribution to ${G_k}^2$ than ${X_k}^2$. Consequently, when most expected cell frequencies are in the range of 1.0 to 5.0 the first two moments for ${G_k}^2$ are larger than those for ${X_k}^2$. The contribution to ${X_k}^2$ for a nonzero count can be quite large when the expected frequency is less than one, however, and the first two moments for ${X_k}^2$ are larger than the corresponding moments for ${G_k}^2$ when a sufficient number of expected frequencies are less than one."

Also, accuracy of the asymptotic chi-squared and normal approximations for cell sizes when expected frequencies do not exceed five were assessed by use of a Monte Carlo study. The aim was to pinpoint when the normal approximation is more precise to warrant the extra calculation, to investigate the effect of deviations from the condition stipulated by the Central Limit Theorem on the accuracy of the asymptotic approximations and to further identify and isolate situations when the exact means and variances leads to better normal approximations. Detail description of the procedure is given in Koehler and Larntz (1980, p. 339). Monte Carlo power calculations indicated that the normal approximation was more accurate for the ${G_k}^2$ statistic standardized with $\mu_{LR,k}$ and $\sigma^2{}_{LR,k}$ than for ${X_k}^2$ standardized with $\mu_{p,k}$ and $\sigma^2{}_{p,k}$. In the case of the unsymmetrical null hypothesis, either test was found to be assertive. Another observation was that the Pearson statistic is dominant over a smaller area as compared to the symmetrical null hypothesis case. The normal approximation for ${X_k}^2$ and ${G_k}^2$ were stated to provide "computationally inexpensive power approximations. Monte Carlo results indicate that that it is not uncommon for the power approximations to be too large by as much as 20 percent for moderate power and moderate cell sizes. The discrepancy is generally smaller for ${G_k}^2$ than for ${X_k}^2$ " (Koehler and Larntz (1980, p. 343)). The Pearson test was found to possess some optimal power properties in the symmetrical case when the number of cells were quite big so the Pearson goodness-of-fit based on the traditional chi-squared approximation is preferred in the symmetrical case.

4.4 Simonoff's Goodness-of-fit Statistic : Sparse Multinomials

The general goodness-of-fit problem tests the null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, where $\boldsymbol{\pi}_0$ is some completely specified probability vector, against all possible alternatives. Simonoff suggested a new approach for testing goodness-of-fit which restricts the null distribution of $\boldsymbol{\pi}_0$ such that it satisfies a smoothness constraint so that knowledge contained in nearby cells can be used to provide concise estimations of probabilities in a particular cell. Taking into consideration that frequent parametric forms like the uniform, normal or gamma occur hence this is not a confining assumption. The improved statistic presented, is based on the MPE (maximum posterior estimator) estimates of Simonoff (1982, 1983). As found in earlier work, under the assumption of smoothness, the estimates are consistent under the sparse asymptotic framework. The proposed test was judged to be more powerful than the standard tests for sparseness since the frequency estimates $\widehat{\pi}_i = \frac{xn_i}{n}$ do not possess the above-mentioned characteristic.

The suggested statistic considered a random vector $\mathbf{x}$ generated from $\boldsymbol{\pi}$ by a multinomial likelihood :

$$\log(\mathbf{x} \mid \boldsymbol{\pi}) = \sum_{i=1}^{k} x_i \log \pi_i, \tag{4.23}$$

where $\sum_{i=1}^{k} \pi_i = 1$ and $\mathbf{x}$ is some random vector.

Simonoff (1982) defined the MPE, $\widehat{\boldsymbol{\pi}}$ as the value of $\boldsymbol{\pi}$ that maximizes

$$L(\boldsymbol{\pi} \mid \mathbf{x}) = \sum_{i=1}^{k} x_i \log \pi_i - \beta \sum_{i=1}^{k-1} \left(\log \frac{\pi_i}{\pi_{i+1}} \right)^2 , \tag{4.24}$$

with $\sum_{i=1}^{k} p_i = 1$ and $\beta \geq 0$.

Simonoff (1983) showed that the MPE is consistent in a sparse asymptotic framework. The consistency of $\widehat{\boldsymbol{\pi}}$, where $\boldsymbol{\pi}_0$ is the null distribution of $\boldsymbol{\pi}$, implies that $\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0$, will find even tiny departures from the null hypothesis if the deviations are from a case which possesses smoothness. The standardized value

$$z_i = \frac{(\widehat{\pi}_i - \pi_{0i})}{\pi_{0i}}, \tag{4.25}$$

is utilized to construct the test statistic

$$M^2 = \sum_{i=1}^{k} \left(\frac{z_i - \mu_0(z_i)}{\sigma_0(z_i)} \right)^2 . \tag{4.26}$$

The subscript 0 denotes that the test statistic is under the null hypothesis. M^2 is more or less equal to the sum of dependent χ_1^2 random variables when the null hypothesis is true. Departure from the null leads to increased values of $\mid z_i - \mu_0(z_i) \mid$, hence rejection occurs for large values of M^2.

The usage of mean and variance occurs in a moderate capacity since the distribution of M^2 is not asymptotically normal. The z_i's exhibited evidence of being dependent and hence M^2 does not have a chi-squared distribution. Simonoff however, verifies that the z_i's are closely distributed as gamma random variables.

The accuracy of the approximation was analysed under the uniform null distribution by examining the true significance level of M^2 when using a gamma approximation, as a function of k. It was found that the gamma random variable with parameters $a = 2.5$ and $b = 2.5 + 0.6a$ fitted the distribution of M^2 well. The value of $a = 0.05$ was estimated from computer simulations. The actual test was repeated for $n = k = 20, 30, 40, ..., 100$. The standard errors of the a levels were observed to be smaller than 0.0095. Use of the gamma distribution leads to a tolerant test; however for $k > 40$ this tolerance is almost insignificant.

The connection between the gamma approximation and the chi-squared approximation allows for the critical value of the gamma $(a,\ b)$ distribution to be rewritten in terms of the critical value of a chi-square random variable. In particular, a gamma $(a\ , b)$ random variable is equal to $\frac{b}{2}$ times the corresponding critical value of a χ^2_{2a} random variable. Hence, the critical values of M^2 can be determined from a table of critical values for chi-square. Koehler and Larntz (1980) had stated earlier that many problems can be recreated in terms of the uniform situation, hence the gamma approximation has wide utility and relevance.

When $n \neq k$, M^2 relies on k for a wide range of n because of the standardization of the z_i. Various simulations with changing sizes of n indicate that as n increases the critical value decreases. However, for nonuniform π_i values in the range $\left[\frac{1}{(3k)}, \frac{3}{(2k)}\right]$ the approximated values are accurate to within ±0.5. Once the values are out of this range, M^2 takes on notably higher values than the gamma approximation would verify.

4.4.1 Effect of Parameter Estimation

As discussed in previous chapters, the issue of unknown parameters needs to be first addressed by estimating the parameters and then decreasing the degrees of freedom by the number of parameters estimated. Effect of parameter estimation was analysed through use of simulations and Simonoff (1985) discovered that for small values of $k(\leq 35)$, there was conformity of the critical values with parameter estimation to those without estimation for k, reduced by the number of parameters being evaluated. This is not consistent with larger values of k and hence it is suggested that significance be evaluated for each situation individually.

The procedure adopted for the case when parameters are estimated involved initial estimation of parameters and this yields a fully specified null distribution and an observed value of M^2. The simulation method was broken down into two stages : "(a) generate a multinomial vector based on the specified null distribution, and (b) from the generated vector, estimate the parameters and calculate M^2" which is described in Simonoff (1985, p. 673). Both steps were repeated for about 500 to 1000 times and then the observed M^2 is weighed against the simulated null distribution of M^2, and the tail probability is estimated which essentially means the null distribution of M^2, when parameters are estimated, was simulated.

In conclusion, Simonoff (1985) had proposed a new goodness-of-fit statistic for sparse multinomials and he showed that if the null and alternative distributions have properties of smoothness, then the proposed test is more effective than the standard tests.

4.5 Loglinear Models for Sparse Data : Koehler's Findings

Koehler (1986) looked at the fit of loglinear models under multinomial sampling in the sparse contingency table case. Amongst the traditional assumptions that minimum expected cell frequencies be increasing without bound with increasing sample size, the assumption of the categories being fixed was inappropriate. As an example, Koehler and Larntz (1986) consider a study of a rare disease. Invariably, this results in information on many issues but a limited sample size. Sample size being small yields unrealistic results due to the large number of variables being included in the study. The trend is to form smaller tables or join related categories to condense that table in terms of sparseness. Seeing that sample size is effected, it generated interest in investigating the effect on choice of categories and the asymptotic properties of goodness-of-fit statistics as the number of categories increased with sample size but without an increase in expected frequencies.

Asymptotic normality of the likelihood ratio goodness-of-fit statistic for loglinear models with closed form estimates was investigated for contingency tables. Asymptotic normality was shown to hold under the following sufficient conditions furnished by Koehler :

(a) the sample size increases with a growth in the number of categories,

(b) increase in the sample size should occur at a quicker pace than the number of parameters estimated under the null hypothesis.

The condition of all expected frequencies becoming large as the sample size increases is not essential. In comparison with other goodness-of-fit statistics; the Pearson statistic is not seriously influenced or distorted by small expected frequencies as in the chi-square approximation case.

Haberman's (1977) work on large sparse contingency tables yielded that when the difference in the degrees of freedom for the two models is substantially less than the total number of observations, the regular chi-squared approximation is suitable for Pearson and likelihood ratio test statistics.

4.5.1 Examining Independence In Sparse Tables

The following notation was adopted for a sequence of multinomial distributions with an increasingnumber of categories.Let t_k represent the number of categories for the kth multinomial in the sequence and $\mathbf{n}_k = (n_{1k}, n_{2k}, ..., n_{t_k k})$ denoted the vector of random frequencies with $\mathbf{p}_k = (p_{1k}, p_{2k}, ..., p_{t_k k})$ describing the corresponding vector of probabilities. The total sample size is $n_k = \sum_{i=1}^{t_k} n_{ik}$. The vector of expected frequencies is expressed $\mathbf{m}_k = (m_{1k}, m_{2k}, ..., m_{t_k k})$ where $m_{ik} = n_k p_{ik}$.

For a two dimensional case, a test of independence was carried out on a series of tables which were increasing in size. It was observed that the kth table in the sequence possessed t_{1k} rows and t_{2k} columns and hence a total of $t_k = t_{1k} \times t_{2k}$ categories. An increase in the number of categories results from a simultaneous increase in the number of rows and columns. The set of indexes denoting the rows and columns of the kth two-dimensional table is represented by $I_k = \{(i,j) : i = 1,2,...,t_{1k}$ and $j = 1,2,...,t_{2k}\}$ whilst $\{n(\mathbf{i}, I_k) : \mathbf{i} \epsilon I_k\}$ is the set of observed frequencies. The total sample size $\{n_k = \sum_{i \epsilon I_k} n(\mathbf{i}, I_k)\}$ is fixed with a matching set of probabilities represented by $\{p(\mathbf{i}, I_k) : \mathbf{i} \epsilon I_k\}$. Expected frequencies were expressed as $m(\mathbf{i}, I_k) = n_k p(\mathbf{i}, I_k)$ for $\mathbf{i} \epsilon I_k$. Row totals and column totals described by $\{n(\mathbf{i}, I_{1k}) : \mathbf{i} \epsilon I_{1k}\}$ and $\{n(\mathbf{i}, I_{2k}) : \mathbf{i} \epsilon I_{2k}\}$, respectively, were indexed by sets $\{I_{1k} = (i,+) : i = 1,2,...,t_{1k}\}$ and $\{I_{2k} = (+,j) : j = 1,2,...,t_{2k}\}$. The corresponding expected frequencies for the row and column margins were denoted by $\{m(\mathbf{i}, I_{1k}) : \mathbf{i} \epsilon I_{1k}\}$ and $\{m(\mathbf{i}, I_{2k}) : \mathbf{i} \epsilon I_{2k}\}$.

The expected frequencies under the null hypothesis are

$$m_0(\mathbf{i}, I_k) = \frac{m(\mathbf{i}, I_{1k}) m(\mathbf{i}, I_{2k})}{n_k} \tag{4.27}$$

The role of $\mathbf{i}$ is as follows. For $\{m_o(\mathbf{i}, I_k), \mathbf{i} = (i,j)\}$ represents an element of I_k but for $\{m(\mathbf{i}, I_{1k}), \mathbf{i} = (i,+)\}$ it is an element of I_{1k} and denotes the corresponding row. The maximum likelihood estimates for the expected frequencies are given by

$$\widehat{m}_0(\mathbf{i}, I_k) = \frac{n(\mathbf{i}, I_{1k})n(\mathbf{i}, I_{2k})}{n_k} \tag{4.28}$$

hence the likelihood ratio goodness-of-fit test statistic for independence is written as

$$G_k^2 = 2\sum_{i\epsilon I_k} n(\mathbf{i}, I_k)\log\left[\frac{n(\mathbf{i}, I_k)}{\widehat{m}_0(\mathbf{i}, I_k)}\right]. \tag{4.29}$$

Koehler (1986) states further that with an increase in the number of categories resulting from a simultaneous increase in the rows and columns, if the conditions of the following theorem are satisfied, then $\frac{(G_k^2-\mu_k)}{\sigma_k}$ has a limiting standard normal distribution where μ and σ_k are described in the following theorem from Koehler (1986, p 485).

Theorem 4.3 "Suppose the hypothesized model for the kth table" has a multiplicative probability structure of the form

$$m_0(\mathbf{i}, I_k) = \frac{\prod_{j=1}^{d_k} m(\mathbf{i}, I_{jk})}{\prod_{j=1}^{d_k-1} m(\mathbf{i}, J_{jk})} \tag{4.30}$$

"and suppose that $t_k \longrightarrow \infty$ as $k \longrightarrow \infty$ in such a way that

(i) there is a fixed $\epsilon > 0$ such that $n_k p(\mathbf{i}, I_k) = m(\mathbf{i}, I_k) > \epsilon$ for all $\mathbf{i}\epsilon I_k$ and all k.

(ii) $\max p(\mathbf{i}, I_k) = o(1)$,

(iii) $d_k t_k^{-1} \sum_{j=1}^{d_k} t(I_{jk}) + t_k{}^{-2} \sum_{j=1}^{d_k} t^4(I_{jk}) = o(1)$,

(iv) $\sigma_k{}^{-1} \sum_{i\epsilon I_k} [n(\mathbf{i}, I_k) - m(\mathbf{i}, I_k)] \times \log\left[\frac{m(\mathbf{i}, I_k)}{m_0(\mathbf{i}, J_k)}\right] = o_p(1)$.

Then $\dfrac{(G_k^2 - \mu_k)}{\sigma_k}$ has a limiting standard normal distribution where the location parameter

$$\mu_k = \mu(I_k) - \sum_{j=1}^{d_k} \mu(I_{jk}) + \sum_{j=1}^{d_k-1} \mu(J_{jk}) + 2\sum_{i \epsilon I_k} m(\mathbf{i}, I_k) \log\left[\frac{m(\mathbf{i}, I_k)}{m_0(\mathbf{i}, I_k)}\right] \quad (4.31)$$

and

$$\sigma_k{}^2 = 4\sum_{i \epsilon I_k} \text{var}\,[Y(\mathbf{i}, I_k)] - 4n_k\gamma_k{}^2.'' \quad (4.32)$$

In the location parameter μ_k,

$$\mu(I_k) = 2\sum_{i \epsilon I_k} EY(\mathbf{i}, I_k)$$

and

$$\mu(I_{jk}) = 2\sum_{i \epsilon I_{jk}} EY(\mathbf{i}, I_{jk})\ .$$

$\mu(I_k)$ represents the sum of the expectations of the information kernels and $\mu(I_{jk})$ is the sum of expectations of the information kernels for a specific margin.

In the theorem, (i) and (ii) are the appropriate smoothness conditions with condition (ii) stipulating the chance of an event in any category decreasing as the number of categories increase, and condition (i) safeguards against any probabilities converging to zero as total sample size increases. Requirement (iii) secures that the number of estimated parameters grows at a gradual rate, compared to the total number of categories. Application of the theorem can be carried out on a sequence of alternatives that converge at a rapid rate to the sequence of null hypotheses as $k \to \infty$ as a result of (iv).

Koehler also shows that G_k^2 possesses the property of asymptotic normality.

When the total sample size n_k and the total number of categories in the table, t_k, get bigger, category size can be increased by expanding the number of variables, or increasing the number of groups for some variables or even increasing both. The value of d_k varies with an increase in the number of variables in the model.

4.5.2 Accuracy Assessments

The normal and chi-squared approximations for X_k^2 and G_k^2 were found and thereafter accuracy assessments were made with the aid of Monte Carlo studies. The algorithm that was adopted to generate tables of frequencies appears in Koehler (1977). Tables of different sizes were tested for the hypothesis of independence. Monte Carlo studies showed that the following modifification of the asymptotic mean

$$\mu_k = 2\sum_{i\epsilon I_k} EY(\mathbf{i}, I_k) - \sum_{j=1}^{d_k}\left[2\sum_{i\epsilon I_{jk}} EY(\mathbf{i}, I_{jk}) - 1\right] + \sum_{j=1}^{d_k-1}\left[2\sum_{i\epsilon J_{jk}} EY(\mathbf{i}, J_{jk}) - 1\right], \tag{4.33}$$

improved the approximation for smaller tables. In estimating the unknown parameters it was found that the expected frequencies became large under the conditions of theorem 4.3 in conjunction with $2\sum EY(\mathbf{i}, I_{jk})$ quickly converging to $t(I_{jk})$ and $2\sum EY(\mathbf{i}, J_{jk})$ rapidly converging to $t(J_{jk})$. Hence, the asymptotic mean in (4.33) is rewritten

$$\mu_k = 2\sum_{i\epsilon I_k} EY(\mathbf{i}, I_k) - \sum_{j=1}^{d_k}[t(I_{jk}) - 1] + \sum_{j=1}^{d_k-1}[t(J_{jk}) - 1]. \tag{4.34}$$

Koehler (1986) writes that " The latter formula is the asymptotic mean when no parameters are estimated minus the number of parameters estimated in fitting the loglinear model." The asymptotic mean in (4.34) and the asymptotic variance were described as the exact asymptotic moments in the Monte Carlo study. Their estimates were achieved from $EY(\mathbf{i}, I_k)$ and var$[(Y(\mathbf{i}, I_k)]$ at the value of the maximum likelihood estimate of the corresponding expected cell frequency.

Monte Carlo studies carried out on tables of dimensions $k \times k$ and $2 \times k \times k$ yielded the following findings by Koehler :

In the $k \times k$ case, a 6 $\times$ 6 table with 36 categories and sample sizes of 18, 72 and 180 were used. Also 10 $\times$ 10 tables with sample sizes of 50, 200 and 500 and lastly a 20 $\times$ 20 table was considered with sample size 200 only.

Thereafter, situations based on varying sets of marginal probabilities were discussed. Findings from the simulations were as follows : the chi-squared approximation was not effective for sparse tables in most cases. Tables possessing a majority of expected frequencies less than 0.5, yielded results which described the ${G_k}^2$ statistic as stochastically smaller than the chi-squared random variable with $(k-1)^2$ degrees of freedom. If most expected frequencies have a value between 1 and 4, then the G_k^2 statistic rejects H_0 too often. The chi-squared approximation was quite accurate for the 6 $\times$ 6 table with sample size of 180 in which all the expected frequencies are 5.

On the contrary, the normal approximation using large sample moments yielded much better findings than the chi-squared approximation but moderate findings for the 6 $\times$ 6 tables. The use of estimated moments resulted in larger rejection levels being obtained as opposed to cases when exact asymptotic moments were used. When most expected frequencies exceeded 1, normal approximations with exact and estimated moments were almost alike. Expected frequencies less than 1, resulted in the normal approximation with estimated moments yielding large rejection levels as a result of the negative bias of the estimator for $EY(\mathbf{i}, I_k)$ obtained by replacing $m(\mathbf{i}, I_k)$ with $\widehat{m}(\mathbf{i}, I_k)$.

Koehler (1986, p. 489) concludes his study of $k \times k$ tables with " The chi-squared approximation for the Pearson statistic is quite accurate for case 1 where all the expected frequencies are equal, but inaccurate for case 3 which possessed either moderately large and very small expected frequencies. The rejection rates for the normal approximation with exact asymptotic moments are similar for the chi-squared approximation."

In the $2\times\ 5\ \times\ 5$ tables, with sample sizes of 25 and 100 and $2\ \times\ 10\ \times 10$ tables samples of sizes 100, 200 and 400 were taken. The chi-squared approximation for ${G_k}^2$ was thought to be erratic as opposed to the normal approximation for ${G_k}^2$ which was quite accurate when the exact asymptotic moments were used. Use of the estimated asymptotic moments proved to provide sound findings when the total sample size was almost double the total number of categories. When all expected frequencies are the same the normal and chi-squared approximations for the Pearson statistic were consistent with the findings from the $k \times k$ case study. The normal approximation was slightly better for total sample size being as large as the number of categories. Neither approximation was favourable for the case of number of categories being twice as large as the sample size.

Koehler (1986, p. 489) summarizes his findings by stating that "It is not possible to make specific recommendations about the accurate use of the normal and chi-squared approximations for the ${G_k}^2$ and ${X_k}^2$ statistics from the limited Monte Carlo study discussed here ". However, he generalizes that the accuracy of the chi-squared approximation for ${G_k}^2$ in sparse tables is unreliable for testing the fit of loglinear models. Concentration of expected cell frequencies between 0.5 and 4 boosts the Type I error. Tables with expected frequencies nearly equal found the chi-squared approximation to be more reliable for the Pearson statistic whereas the chi-squared approximations for ${X_k}^2$ were obtained for sparse tables. A frequent observation was that the limiting normal distribution was more precise for ${G_k}^2$ than for ${X_k}^2$.

4.6 Adopting Jackknife and Bootstrap Methods for Sparse Multinomials

Tests for $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ against all possible alternatives where $\mathbf{X} = (X_1, ..., X_k)$ is $\text{Mult}_k(n, \boldsymbol{\pi})$, initially saw Pearson's statistic (3.3) and the loglikelihood statistic (3.4) being generally recommended. Further work by Cressie and Read (1984) on the family of power-divergence statistics yielded a more appropriate statistic with $\lambda = \frac{2}{3}$.

With an increase in sample size, $n \to \infty$, the asymptotic χ^2 distribution for the statistic no longer holds for the sparse case. Morris (1975) observed that ${X_k}^2$ and ${G_k}^2$ were asymptotically normal with different mean and variance under a simple hypothesis for increasing sample size. Cressie and Read (1984) showed that the power divergence statistic mirrored such properties. The variance of the statistic is required to prove these resuts under sparse conditions.

Simonoff (1986) used nonparametric techniques to estimate such variances. His findings showed that the bootstrap does not produce an unchanging or constant variance estimate but the jackknife and the "categorical jackknife" each gives a consistent estimate. Categorical jackknifing involves deletion of cells instead of the deletion of observations.

4.6.1 Approximating Variance

Conditions for the sparse asymptotic environment include $n, k \to \infty$ with $0 < \gamma_1 < \frac{n}{k} < \gamma_2 < \infty$. The null hypothesis is assumed to be true and further assumptions were that there exists $M \in (1, \infty)$ such that $0 < (Mk)^{-1} < \pi_i < \frac{M}{k} < 1$ for all i where M is positive. Another assumption is that the maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ is consistent. Also, that $\mathbf{p}(\boldsymbol{\theta})$ have bounded second derivatives with respect to $\boldsymbol{\theta}$ is another essential condition.

Then, $\widehat{\pi}_i = \frac{x_i}{n}$ is defined to be the frequency estimate.

The concept of bootstrapping introduced by Effron (1979) entailed the estimation of the variance of a statistic under some unknown distribution F by its variance under the hypothetical distribution $\widehat{F}$. For categorical data, this involves analysing the statistic as if the true distribution of $\mathbf{X}$ was multinomial $(n, \widehat{\boldsymbol{\pi}})$ instead of $(n, \boldsymbol{\pi})$. The bootstrap distribution in a parametric bootstrap is based on the parametric estimates $\mathbf{p}(\widehat{\boldsymbol{\pi}})$ instead of $\widehat{\boldsymbol{\pi}}$.

Previous research on jackknifing summarized the jackknife estimate as :

$$\mathrm{var}_J(T) = \sum_j \frac{(P_j - T_{(\cdot)})^2}{(n(n-1))},$$

where $T_{(\cdot)} = \sum_j \frac{P_j}{n}$ and $P_j = nT - (n-1)T_{(j)}$ where $T_{(j)}$ represents the value of the statistic evaluated after the jth observation is deleted from the data set. The ith cell contains x_i identical pseudovalues. Varying the notation of Simonoff (1986) slightly, the ith cell is described

$$J_i = \sum_j \frac{(x_j - np_j(\widehat{\boldsymbol{\theta}}))^2}{p_j(\widehat{\boldsymbol{\theta}})} - \frac{((x_i - 1) - (n-1)p_i(\widehat{\boldsymbol{\theta}}_{(i)}))^2}{p_i(\widehat{\boldsymbol{\theta}}_{(i)})} - \sum_{j \neq i} \left[\frac{(x_j - (n-1)p_j(\widehat{\boldsymbol{\theta}}_{(i)}))^2}{p_j(\widehat{\boldsymbol{\theta}}_{(i)})} \right] \qquad (4.35)$$

Now (4.46) can be expressed as

$$J_i \doteq 1 - p_i(\widehat{\boldsymbol{\theta}})^{-1} + \frac{2(x_i - np_i(\widehat{\boldsymbol{\theta}}))}{p_i(\widehat{\boldsymbol{\theta}})}, \text{ for } i = 1, ..., k.$$

Therefore $X^2{}_{(\cdot)} \doteq \sum_i \frac{x_i J_i}{n} = 2X^2 - \sum_i \frac{\widehat{\pi}_i}{p_i(\widehat{\boldsymbol{\theta}})} + 1$ and the jackknife estimate of the variance is expressed

$$s_J{}^2 \doteq \sum_i \frac{x_i(J_i - X^2{}_{(\cdot)})^2}{(n(n-1))}.$$

${s_J}^2$ was expressed as

$$\frac{\sum_i x_i \left(1 - p_i(\widehat{\boldsymbol{\theta}})^{-1} + 2\frac{(x_i - np_i(\widehat{\boldsymbol{\theta}}))}{p_i(\widehat{\boldsymbol{\theta}})} - 2X^2 + \sum_j \frac{\widehat{\pi}_j}{p_j(\widehat{\boldsymbol{\theta}})} - 1\right)^2}{(n(n-1))}. \qquad (4.36)$$

He states further that the calculation of $E({s_J}^2)$ shows that ${s_J}^2$ is too large. Hence for the purpose of his study he restated (4.47) as a smaller estimate :

$$\bar{s}_J^2 = \frac{{s_J}^2}{2} + \left[\frac{(n-2)}{2n(n-1))}\right]\left(\sum_i p_i(\widehat{\boldsymbol{\theta}})^{-1} - k^2\right). \qquad (4.37)$$

Comparing the structures of X^2, G^2 and $2nI^{\frac{2}{3}}$ in this case, he states that a jackknife approach in which cells, rather than observations are deleted one at a time is an appropriate choice. This is verified by the property that the cell counts become independent asymptotically under the sparse conditions. Simonoff (1986) jackknifed $U = \frac{X^2}{k}$ instead of X^2 since $\text{var}(X^2) = k^2\text{var}(\frac{X^2}{k})$ is not very different in structure.

4.6.2 Evaluation and Assessment of Estimator's Performance

Measuring the performance of various estimators was based on the bias, standard deviation (SD), and root mean squared error (RMSE) of variance estimators for X^2 as well as $2nI^{\frac{2}{3}}$, suggested by Cressie and Read. Cell values were all divided by k. Simulations were undertaken for 500 multinomial responses with the same set of replicated multinomials being used for each pair (k, n). Use of an unbiased but more variable estimate leads to a test with better size, on average.

Simonoff's (1986) simulation results illustrate that the bootstrap is an extremely poor variance estimate for this case study. The bootstrap variance is basically $\mathrm{Var}(X^2)$ under the alternative model $\mathbf{X} \sim \mathrm{Mult}_k(n, \widehat{\boldsymbol{\pi}})$ instead of $\mathbf{X} \sim \mathrm{Mult}_k(n, \boldsymbol{\pi})$. When the data is sparse ($\frac{n}{k} < 5$) heavily biased variance estimates are obtained. This is consistent with Cressie and Read's asymptotic variance which leads to a heavily biased variance estimate even for $n = 10k$.

On the other hand, the jackknife furnishes a favourable variance estimate for sparse data ($n < 5k$) and is therefore preferred over the categorical jackknife due to minimal fluctuations. The bias properties of the parametric bootstrap and the asymptotic formula are quite similar but Cressie and Read's asymptotic variance is preferred as the parametric bootstrap has changeable bias properties. Accuracy improves with an increase in the table size for given constant $\frac{n}{k}$.

On comparing the results of estimator behaviour for X^2 and $2nI^{\frac{2}{3}}$, Simonoff concluded that the biases and standard deviations are a bit smaller for $2nI^{\frac{2}{3}}$ than for X^2.

The restriction, of the standardized version of a goodness-of-fit statistic, that the number in the denominator should equal the square root of the estimate of the *null* variance, is a disadvantage of the jackknife. The asymptotic formula provided by Cressie and Read (1984), is acceptable but if the null is not true then the probability estimates will be incorrect. The estimate is still considered to be closer to the null variance than the jackknife. Simonoff (1986) states in conclusion that in the case of few cells, the jackknife should be used if $n < 10k$. On the other hand the jackknife for sparse data should be used when $n < 5k$ whilst $\mathrm{Var}(2nI^{\lambda})$ is recommended for nonsparse data.

4.7 Comparisons between Pearson X^2 and G^2 under Sparse Conditions

Discussions on the asymptotic normality of the two statistics, small-sample studies, the effect of parameter estimation and conditional tests form the basis of this comparative study.

4.7.1 Similarities and Differences under Asymptotic Normality

X^2 and G^2 are asymptotically normal for sparse samples but their asymptotic means and variances are unequal unlike the findings for the fixed cell assumptions. Koehler and Larntz (1980) attribute this to the varying effect of very small observed counts on X^2 and G^2. For expected frequency larger than 1, the first two moments of G^2 are larger those for X^2 when many expected cell frequencies exist between 1 and 5 for observed frequencies of 0 or 1. When expected frequencies are less than 1 then the opposite occurs.

4.7.2 Small-Sample Studies

Koehler and Larntz (1980) observed the following findings in their comparative study : for the equiprobable null hypothesis, the chi-squared approximation for X^2 did not produce distorted information in the case of small expected frequencies of almost 0.25 with $k \geq 3$, $n \geq 10$ and $\frac{n^2}{k} \geq 10$. However, G^2 was not well approximated by the chi-squared distribution when $\frac{n}{k} \leq 5$. Moderate intervals were obtained for the chi-square approximation when $\frac{n}{k} < 0.5$ with the "liberal" intervals being yielded for $\frac{n}{k} > 1$. They advise that the normal approximation be used for G^2 and that the X^2 based on the traditional chi-squared be used for the test involving the equiprobable model.

Zelterman examined the hypothesis of cell frequencies coming from a Poisson distribution with equal means and found that Zelterman (1984) reports that the G^2 and Freeman-Tukey T^2 are better approximated by the normal distribution as compared to X^2. That X^2 is better approximated by the chi-square distribution than the normal distribution under the equiprobable model was a contribution of Koehler and Larntz (1980). This was also preferred to using G^2 with the normal approximation. Cressie and Read (1984), however, contend that the normal approximation is inadequate as compared to the chi-squared approximation for both X^2 and G^2 when $10 \leq n \leq 20$ and $2 \leq k \leq 6$.

In the case of unequal probabilities G^2 was the approved choice for the normal approximation under conditions where most expected frequencies are smaller than 5, $n \geq 15$ and $\frac{n^2}{k} \geq 10$. Expected frequencies which are exceptionally small distort the accuracy of the normal approximation for X^2 however the normal approximation for G^2 is unaffected. When a large number of expected frequencies are less than one, the yielded critical values for X^2 being unrestricted and G^2 exhibited the same shortfalls as observed in the equiprobable case.

4.7.3 Comparison for Parameter Estimation

Earlier documentation on Koehler and Larntz's work indicate that they initially assumed that all hypotheses were simple and require no parameter estimation. Thereafter Koehler (1986) looked at the level of precision of the normal approximation to the G^2 by scrutinising the outcomes of parameter estimation on loglinear models. He concluded that firstly, X^2 is closely approximated by the chi-square approximation, unlike G^2, when the expected cell frequencies are less than 1. Inferior chi-squared approximations for X^2 are are obtained when the expected cell frequency range in size from being quite large to very small. In most instances the normal approximation was more accurate for G^2 than for X^2. Lastly, replacing the expected frequencies by the maximum likelihood estimates yields large biases for the moments of G^2 quite often and hence Koehler motivates that estimates with an insignificant or negligible bias be developed.

4.7.4 Comparisons under Conditional Tests

In addressing the shortfalls of the asymptotic approximation, McCullagh (1986) suggested that by conditioning on the sufficient statistic for the nuisance parameters, the distribution dependent on the unknown parameters could be eliminated. He constructed a normal approximation for the conditional tests, based on X^2 and G^2, under the restriction that the number of estimated parameters do not change as k increases. His study of X^2 suggests that the normal approximation is inadequate.

4.7.5 Evaluating and Measuring Efficiency

Efficiency comparisons between X^2 and G^2 can be made by studying their power functions. The topic of efficiency under small-sample studies and other sparse conditions will be examined in the next two subsections.

X^2 was found to be more robust in small samples than in other studies undertaken by West and Kempthorne (1971). Koehler and Larntz (1980) obtained results consistent with West and Kempthorne whilst Goldstein, Wolf and Dillon (1976) compared the X^2, G^2 and Freeman Tukey F^2 statistics and obtained almost similar results for all three statistics. This prompted further power computations for X^2, G^2 and F^2. Wakimoto, Odaka and Kang (1987) reveal that there is a significant discrepancy between the powers of the test statistics, which were similar to the *bump* and *dip* classification obtained by Read and Cressie (1988), in their study of power computations for the above-mentioned statistics. For the power-divergence statistic, the *bump* alternative is a description used when the power function increases with λ for $\delta > 0$. When the power function decreases with λ for $\delta < 0$, a *dip* occurs when $k > 2$. Read and Cressie (1988) recommend X^2 for *bumps* and F^2 for *dips* whilst G^2 is found to lie between X^2 and F^2.

Examining discrete data obtained from parts of continuous distributions, Kallenberg, Oosterhoff and Schriever (1985) show that X^2 and G^2 possess equivalent power for testing the equiprobable hypothesis when the number of cells is small. With an increase in the cell size, their findings highlight that X^2 is better than G^2 for heavy tailed cases and G^2 is better than X^2 for light tailed options.

4.7.6 Sparseness Assumptions

For the equiprobable null hypothesis, Holst (1972) and Ivchenko and Medvedev (1978) report that X^2 is more robust than G^2 under tests for local alternatives. On the other hand, the null hypotheses with unequal cell probabilities yields that G^2 or X^2 is preferred depending on the situation.

Recalling Zelterman's (1986, 1987) D^2 statistic,

$$D^2 = X^2 - \sum_{i=1}^{k} \frac{X_i}{n\pi_i} \tag{4.38}$$

derived from the loglikelihood ratio statistic for testing a sequence of multinomial null hypotheses against a sequence of local alternatives which are Dirichlet mixtures of multinomials, it is also recalled that he claimed that D^2 is not a member of the power-divergence family, however Cressie and Read (1988) claim otherwise.

They contend that the following modified statistic

$$D^2 = \sum_{i=1}^{k} \frac{(X_i - \frac{1}{2} - n\pi_i)^2}{n\pi_i} - k - \sum_{i=1}^{k} \frac{1}{4n\pi_i}, \tag{4.39}$$

consists of a generalized X^2 statistic, (the first term), with the latter two terms being independent of the data. Expanding the power divergence family of statistics to include

$$\sum_{i=1}^{k} h^{\lambda}(X_i + c, m_i + d), \tag{4.40}$$

where

$$h^{\lambda}(u, v) = \frac{2}{\lambda(\lambda + 1)} \left[u \left\{ \left(\frac{u}{v} \right)^{\lambda} - 1 \right\} + \lambda(v - u) \right], \tag{4.41}$$

indicates that D^2 is similar to the member of the power-divergence family with $\lambda = 1$, $c = -\frac{1}{2}$ and $d = 0$. D^2 was shown to have fairly tolerable asymptotic power when the test based on X^2 is biased.

Chapter 5

Consequences of Small Expected Frequencies

The effects of small expected frequencies will be examined in this chapter. Section 5.1 deals with Tate and Hyer's (1973) findings whilst section 5.2 presents and discusses findings of Lawal and Upton. Research findings by Chapman (1976) will be documented in section 5.3.

5.1 Small Expected Frequency and Accuracy of the X^2 Test

The research study undertaken by Tate and Hyer (1973) considered the accuracy of the chi-square distribution as an approximation to the multinomial. Large sample methods have been popular in the analysis of tests like the Pearson X^2 test.

For the Pearson X^2 statistic it is known that, provided that the expected frequencies are not too small, then the X^2 statistic is approximately distributed as a chi-square with $k-1$ degrees of freedom. The acceptable size of the "small" expected frequencies has been a widely investigated topic. Previous findings include : Fisher (1941, p. 82) suggestion that "no expectation be less than five" to avoid invalidating the chi-square approximation. Cramer (1946, p. 420) stated that "expectations should be at least ten," whilst Kendall (1952, p. 292) claims that the estimation "confidently be applied when the theoretical cell frequencies are, say, not less than 20." Cochran (1952) suggests that in goodness-of-fit tests of the normal or Poisson distributions, expectations should be at least one at either one or both tails. Yarnold (1970, p. 865) suggested that in the single multinomial with no estimated parameters "If the number of classes s, is three or more, and if r denotes the number of expectations less than five, then the minimum may be as small as $\frac{5r}{s}$". Statistical literature suggests that expectations should be five or more. Tate and Hyer (1973, pp. 836–837) state the following with regard to the significance of an expected frequency of five : "Although any recommendation is in part arbitrary, the number five may well have originated in the experience that, when expectations are about five or more, the binomial distribution is usually well fitted by the normal curve and that, consequently, expectations of five or more satisfy the assumption of normal binomial distributions in categories."

5.1.1 The X^2 Test : Issues of Precision

To gain some insight into the accuracy of the chi-square approximation, Neyman and Pearson (1931) compared the exact multinomial probabilities with chi-square probabilities of X^2, Shanawany (1936) looked at the agreement between the exact and chi-square probabilities under different distributions whilst Van der Waerden (1957) carried out similar investigations for small skew distributions and uniform distributions. Their conclusions were similar, as each found that the chi-square approximation was adequately sufficient even in the extreme cases.

Additional experiments on assessing accuracy of the X^2 test were undertaken by Slakter (1966) who obtained the distributions of X^2 when the null hypothesis is true that the π_i are each equal to $\frac{1}{k}$ for n of sizes 10, 15 and 50. Expected frequencies range from 5 when ($n = 50$ and $k = 10$) to 0.05 when ($n = 10$ and $k = 200$) using 10 000 random samples. He computed the probability values of X^2 in the hypothetical sampling distributions having chi-square probabilities of 0.01, 0.05 and 0.1 He concluded that the exactness or precision of the approximations were unperturbed by the size of expectations.

Roscoe and Byars (1971) found the distribution of X^2 in 10 000 random samples using various associations between n and k namely, $n =$ 10, 15, 20, 30, 50 and 100 and $k =$ 2, 3, 4, 5, 6, 8 and 10. The X^2 values were calculated when the expected frequency was $\frac{n}{k}$. Comparisons for X^2 and the 0.05 and 0.01 tabled values of chi-square, with attention being paid to the number of rejections in each set of 10 000 samples, yielded the following response : (Tate and Hyer (1973, p. 837)) "acceptable approximations were obtained with expected frequencies as small as one". His findings are consistent with Slakter's conclusions and hence indicate that the chi-square approximation is extremely powerful in tests of goodness-of-fit for the discrete uniform distribution with minimum expected frequencies presenting no complications.

A further finding was that the X^2 statistic is quite close to the chi-square distribution irrespective of size. However, the topic of accuracy of X^2 did not address the single multinomial case.

5.1.2 Differences between the Multinomial and X^2 Tests

In addition to the investigations discussed in the previous sections, further accuracy assessments of the chi-square distribution as an approximation to the multinomial distribution were undertaken by Tate and Hyer (1973). They explain that the multinomial distributions were contrived by k taking on values from 3 to 7 and n changing so that the expectations spanned from one to not fewer than five, with parameters each equal to $\frac{1}{k}$. Further, thirty six distributions were constructed, with $k = 3$ and n ranging from 4 to 12, with parameters 0.10, 0.25, 0.65, 0.10, 0.35, 0.55, 0.25, 0.25, 0.50 and 0.30, 0.50, 0.20 respectively. After calculating the exact cumulative probabilities and chi-square probabilities "the absolute percentage errors in the chi-square probabilities and the number of times they underestimated the exact probabilities in the $[0.005 - 0.009]$, $[0.010 - 0.050]$, $[0.051 - 0.100]$, $[0.101 - 0.150]$ and $[0.151 - 0.205]$" (Tate and Hyer (1973, p. 837)) regions were determined.

Results obtained indicated that there was no reduction in the mean error when there was a growth in the size of the expected values. The largest errors had appeared for expectation approximately equal to five. Further, the error decreased as the exact probability grew but also increased in accordance with an increase in the number of categories.

Tate and Hyer (1973, p. 838) conclude that "Insofar as one may generalize from the distributions studied, the X^2 test is not satisfactory if close approximations to exact probabilities are needed and expectations fewer than 10. Even when expectations exceed 10, the approximations may be poor. Regarding expectations, there appears to be no more justification for the 'five-or-more' rule-of-thumb than a 'one-or-more' rule in using X^2 to test the hypothesis that the parameters of a multinomial distribution have specified values against the alternative that at least one parameter is not as specified."

The main source of discrepancy was the number of outcomes resulting in the same X^2 value having differing cumulative multinomial probabilities. Another approach that was utilized involved calculating the rank-order coefficients of correlation between the X^2 and multinomial probabilities of individual outcomes over the region $[0.01 - 0.10]$ and the extent of disagreement between the chi-square and exact probabilities was observed.

Tate and Hyer (1973) state that the X^2 test of goodness-of-fit is not reliable for small expectations. Differences between nominal and actual levels of significance of the test varies, thus no formal conclusions about power were furnished.

Cochran (1952) regarded the inaccuracy to be unimportant and Roscoe and Byars (1971) agree with this conclusion but comment further that researchers would ideally prefer limits that are slightly less confining. Good, Gover and Mitchell (1970) considered an approximation to be satisfactory if $P(\chi^2)$ is within a factor of two of $P(X^2)$. The studies undertaken involved goodness-of-fit tests with no parameters being estimated.

5.1.3 Radlow and Alf's Assessment Approach

Contentions by Radlow and Alf (1975) that discrepancies exist in Tate and Hyer's (1973) findings due to the use of an inappropriate model led them to suggest an alternate multinomial test. They attribute the discrepancies to the finding that the objective function for arranging experimental outcomes has been altered when comparisons were made with the χ^2 test.

Tate and Hyer's (1973) approach involved the multinomial test ordering terms by their probabilities, unlike the χ^2 test which orders terms by deviations from the null hypothesis. The method is favoured if events of lower probability exhibit more discrepancy from the null hypothesis which is seldomly true.

Another multinomial test which orders experimental outcomes in the same way as the χ^2 test is proposed. Referred to as the "exact χ^2 test" and expressed, as "$P_{m\chi^2}$", the procedure adopted is as follows:

1. The probability of each outcome under the null hypothesis is computed using the multinomial frequency distribution.

2. The χ^2 values of each outcome under the null hypothesis are calculated.

3. Outcomes are ordered using the χ^2 values.

4. Cumulative probabilities are computed commencing with the one associated with the largest value for χ^2.

5. The null hypothesis is rejected at the α level if the cumulative probability corresponding to the outcome is equal to or less than α.

On examining the accuracy of the approximate χ^2 test, Radlow and Alf (1975, p. 813) state that "If Tate and Hyer were using an incorrect ordering of experimental outcomes, large discrepancies should be found between their multinomial test and the χ^2 test even for large expected values. This is the test they reported. Moreover, if the exact χ^2 test described here provides the correct ordering of experimental outcomes, discrepancies from the approximate χ^2 test should be smaller then those obtained by the Tate and Hyer method, and only negligible errors should appear when expected frequencies are large. This is precisely the result obtained."

Their findings also confirm that for small expected cell frequencies the proposed exact χ^2 test should be used and a further corollary is obtained. The corollary states that the one-sample χ^2 test supplies better approximations than the Neyman Pearson (1931) statistic and the Tate and Hyer statistic.

5.2 Log Normal Approximation to the Distribution of the X^2 Statistic

In testing the null hypothesis that $X_1, \ldots, X_k$, the observed frequencies in k classes are distributed according to a multinomial with probabilities $\pi_1, \ldots, \pi_k$, the commonly used X^2 statistic, proves to be unreliable if the expected frequencies become too small. Lawal and Upton (1980) also questioned the permissible size of the expected frequency for undistorted approximation of X^2 by the χ^2_{k-1} distribution. They recommend a log normal approximation to the distribution of X^2, claiming that the approximation is reliable under restrictions that the smallest expectation is bigger than $\frac{r}{d^{\frac{3}{2}}}$, with r describing the number of expectations less than 5, and d representing the number of degrees of freedom.

5.2.1 Earlier Findings on Size of Expectations

Yarnold (1970) suggested that the following description should be the basis of allocating some clarity as to what is meant by "small" expected frequency: "If the number of classes k is three or more, and if r denotes the number of expectations less than five, then the minimum expectation may be as small as $\frac{5r}{k}$" (Lawal and Upton, 1980, p. 447). Lawal and Upton's intention involved developing an approximation to the distribution of X^2 so that it can be applied even to cases outside Yarnold's restrictions.

Pearson's (1932) approximation to the distribution of X^2 described the exact variance of X^2 as

$$\operatorname{var}(X^2) = 2(k-1) + \frac{(R - k^2 - 2k + 2)}{n}, \tag{5.1}$$

where $R = \sum \pi_i{}^{-1}$. He found that when R and k, the number of classes, were small in comparison to n, the total frequency, then

$$\mathrm{var}(X^2) = 2(k-1) \tag{5.2}$$

which is the variance of the approximating ${\chi^2}_{k-1}$ distribution, yielded good approximations.

When small expected frequencies arise, $\frac{R}{n}$ is large and leads to X^2 possessing a greater variance than the approximating χ^2. Nass (1959) concluded that the value of cX^2 should be looked at in comparison with percentage points of the χ^2_d distribution, with c and d being obtained so that there is acquiescence between both means and variances only to be discredited by Yarnold (1970) who found minimal improvement on the usual χ^2 approximation in his study involving small expected frequencies. He assessed the performance of the $C(m)$ distribution (Cochran, 1942). The C(m) approximation works in that portion of the parameter space where the χ^2 approximation fails. It was derived under the assumption that as $n \to \infty$, some of the expectations are finite whilst the others are very large. The limiting distribution of X^2 is called the $C(m)$ distribution and the $C(m)$ approximation for $P(X^2 \geq c)$ is called the probability under this limiting distribution.

The $C(m)$ distribution involved r of the cells possessing an exact Poisson distribution with the remaining cells having an exact χ^2 distribution :

$$C(m) = \sum_{i=1}^{r} \frac{(U_i - m_i)^2}{m_i} + \chi^2. \tag{5.3}$$

In the above distribution, U_i has a Poisson distribution with expectation m_i, and χ^2 is a random variable having a chi-square distribution with $k-r-1$ degrees of freedom. Furthermore, $U_1, ..., U_r$ and χ^2 are independently distributed.

Yarnold looked at changes in performance of the upper tail of the χ^2 approximation to the distribution of X^2 for both small and large expectations. The χ^2 approximation yielded misleading findings when some expectations were large and others were small whilst the $C(m)$ approximation yielded accurate results and the error in the χ^2 approximation is well approximated by the difference between the χ^2 and $C(m)$ approximations for $P(X^2 \geq c)$. His observations showed that changes in the sizes of the large expectations had minimal consequences. Hence he concluded that the $C(m)$ distribution was a good alternative in cases when his rule did not find the usual χ^2 approximation to be suitable.

5.2.2 Approximations Using the Log Normal

Studies undertaken by Lawal and Upton (1980) on the distribution of $\sum_{i=1}^{r} \frac{(U_i - m_i)^2}{m_i}$, indicated that this term has the shape of the upper tail of a log normal distribution, but they were unsuccessful in proving a connection between the tail areas of the log normal and the X^2 distributions through theoretical methods. The two-parameter log normal distribution was used for approximations to the upper tail of a χ^2 distribution and the fit was considered to be acceptable.

Their approach described the following : "Z has a log normal distribution with parameters μ and σ^2 and are related to the mean and variance of X^2 by

$$\mu = \theta - \frac{1}{2}\psi, \quad \sigma^2 = \psi - \theta, \tag{5.4}$$

where

$$\theta = 2\log[E(X^2)] = 2\log(k-1), \tag{5.5}$$

and

$$\psi = \log[E(X^2)^2 + \mathrm{var}(X^2)] = \log\left[k^2 - 1 + \frac{(R - k^2 - 2k + 2)}{n}\right]. \tag{5.6}$$

If U_α is the upper α point of a unit normal random variable, the above implies that the upper α point for X^2 is estimated to be

$$exp(\mu + \sigma U_\alpha), \tag{5.7}$$

while

$$P(X^2 > z) = \Phi\left[\frac{(\log z - \mu)}{\sigma}\right], \tag{5.8}$$

where $\Phi(\cdot)$ is the unit normal distribution function." (Lawal and Upton (1980, p. 449)).

5.2.3 Effect of Infinite Samples

The effectiveness of the log normal approximation was gauged through comparisons with the $C(m)$ distribution. The exact tail probabilities of the $C(m)$ distribution corresponding to the log normal critical values from (5.8) were calculated for the ranges that were considered suitable by Cochran (1952) for this type of approximation to be successful. The χ^2 approximation yielded poor results quite consistently and the accuracy of the χ^2 approximation worsened when the number of cells which are small, increased. This is also echoed in Yarnold's rule which connects the admissible size of the smallest expectation to the proportion of cells which are small. Thus the log normal approximation seems to be uninfluenced by the number of small cells, providing that the number of expectations less than 5 is restricted to be less than $k - 1$ where k represents the number of cells.

5.2.4 Effect of Finite Samples

Yarnold (1970) illustrated that the distribution of X^2 was almost identical to the $C(m)$ distribution, even when n is quite small. For the case of infinite samples, it was seen that the log normal approximation performed quite successfully for the upper tail of the $C(m)$ distribution. Seeing that the distribution of X^2 is a step function, the number of possible values for X^2 would be finite. Thus the most apparent discrepancy of any approximation to the X^2 would be expected for the case when there are few values for X^2 which arise when the number of cells, k, is small, and when most of the cells have equal expectations.

The following situations were considered : firstly, when r cells had small expectations and the remaining cells had large expectations. The second case investigated performance when the expectations varied quite distinctly in size. In both cases the log normal approximation fared well immaterial of the variation in the number and size of expected frequencies.

5.2.5 Recommendations for Use of the Log Normal Approximation

X^2 is discrete, especially when the number of cells are small and they possess equal expectations. Therefore it may sometimes occur that there is no critical value which leads to a tail probability close to one of the standard levels. Therefore, uncertainty exists about the critical values derived from the log normal approximation corresponding to a tail probability in a specified interval. However, results from experiments undertaken by Lawal and Upton (1980, p. 452) suggest that the following method described below yields suitable tail probabilities between 0.03 and 0.07 at the nominal 0.05 level and between 0.003 and 0.02 at the nominal 0.01 level. Bounds of 0.04 and 0.06 and between 0.005 and 0.015 are obtained for cases in which power was not extreme.

Lawal and Upton (1980, p. 452)) recommend the following rules:
"(i) in cases where Yarnold's rule for the χ^2 approximation is satisfied, this should be used;

(ii) in other cases, the smallest expectation should be greater than $\frac{r}{d^{\frac{3}{2}}}$, where r is the number of expectations less than 5 and d is the number of degrees of freedom."

5.3 Comparisons between X^2, G^2, and the Multinomial : Small Expected Frequencies

Comparisons were made between the exact multinomial, and exact χ^2 probabilities for X^2 and G^2 by using the significance levels obtained for the abovementioned statistics in cases with small expected frequencies occuring in almost 30 classes. The eventual findings indicate that the χ^2 probabilities for G^2 were nearer to the multinomial than the corresponding X^2 probabilities.

Chapman (1976) explains that the test of homogeneity can be used to test if the categories in a contingency table contain the same number of observations. Choice of the relevant test depends on the data set. The exact multinomial probabilities, and exact χ^2 probabilities for the statistics X^2 and G^2 were used for the calculation of significance levels for various partitions of n into k classes. Thereafter a comparison was made taking the levels from each of the three criteria.

An alternative approach for testing the same hypothesis is

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - (\sum_{i=1}^{k} \frac{x_i}{k}))^2}{(\sum_{i=1}^{k} \frac{x_i}{k})}. \tag{5.9}$$

The partitions x_i were ordered by their X^2 values. Partitions which are further away from the null hypothesis give rise to larger values of X^2, which is approximately χ^2 with $k-1$ degrees of freedom if the expected value per class, $(\sum_{i=1}^{k} \frac{x_i}{k})$, is fairly large.

An additional test measuring departures from homogeneity, is the loglikelihood ratio criterion described by,

$$G^2 = 2\sum_{i=1}^{k} x_i \log_e \left(\frac{x_i}{(\sum_{i=1}^{k} \frac{x_i}{k})} \right). \tag{5.10}$$

Here partitions were ranked by their G^2 values. Like χ^2 G^2 has an approximate chi-squared distribution with $(k-1)$ degrees of freedom for substantially large expected values.

In addition to Tate and Hyer's (1973) findings discussed in previous sections, Cochran (1936) suspected that the G^2 distribution may be expressed more accurately by a continuous curve than the X^2 distribution since there is greater chance for many different arrangements to give the same X^2 value than the same G^2 value. Fisher (1950) found that the X^2 test was affected to a greater extent than G^2, for many classes with large observations. Conahan (1970) observed that for $k \geq 5$ and expected value being at least 3, the likelihood ratio test was found to be superior to the X^2. Good approximations to the exact multinomial were obtained when the expected value is greater than or equal to 10. Finally, when $k \geq 5$, the multinomial test was preferred for expected values less than 3 whilst the chi-squared approximation, G^2 was suitable for expected frequencies greater than or equal to 3.

Partitions were generated with the use of Lehmer's (1964) algorithm for the calculation of significance levels. Many restrictions were imposed to minimize the number of divisions and hence reduce execution time when the number of observations was very large. Partitions were discarded if they contributed little to the significance levels for other partitions.

5.4 Comparing X^2 and G^2

The differences between the X^2, G^2, χ^2 and exact multinomial probabilities were observed with the use of the modification of Tate and Hyer's method of breaking down data to a simpler form. Probability differences between the exact multinomial and two chi-squared approximations were examined according to the difference caused by varying expectation per class, number of classes and multinomial probability, by the size and range of the differences and variations in the number of underestimates of the multinomial probability. Variance for each partition was assumed to be the same so that the averages were weighted by the square root of the number of partitions used to form them. It was suggested that further analyses need to be undertaken to determine if a trend arises with varying expectation or the number of classes.

Generally, the exact G^2 probabilities yielded values closer to the multinomial probabilities than for X^2, while the differences between the exact and chi-squared probabilities for X^2 were not as large as the corresponding G^2 differences. An increase of expectation from two to five, led to the size of the differences between the exact X^2 and the multinomial increasing. However, there was a decrease in the differences between the exact and chi-squared probabilities for X^2 and a decrease in the size of both types of differences for G^2 especially for larger multinomial probabilities.

The general pattern that was noted is as follows : "the probabilities from the exact log likelihood, the log likelihood used as a χ^2 approximation, and the multinomial initially tend to be smaller than those from the exact X^2, and X^2 as a χ^2 approximation.... This relationship becomes more pronounced as the expectation per class increases to five and appears to be independent of the multinomial probability levels" (Chapman, 1976, p. 860).

Chapter 6

The Effects of Small Sample Size

In this chapter, the appropriateness of properties previously highlighted for large samples will be examined for the small sample case. Topics that will be covered in the discussion include :

(i) Suitability of asymptotic results for small samples.

(ii) Differences between asymptotic significance levels and exact significance levels for the power divergence statistic.

(iii) Further approximation methods relevant to small samples.

(iv) What sample size does "small" define ?

(v) Comparing relative efficiency of the power-divergence family members for small samples with large sample results.

(vi) General overview and suggestions for small sample statistics.

6.1 Asymptotic Moments and Asymptotic Significance Levels

Read and Cressie contend that the accuracy of the small sample asymptotic significance levels can be determined through comparisons between the asymptotic moments of the test statistic with small sample expressions for these moments.

They examined the null model as defined below :

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0, \tag{6.1}$$

where $\boldsymbol{\pi}_0$ represents a completely specified probability vector with k, the number of cells, being fixed. In the Chapter 3 it was shown that the power divergence statistic $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$ is chi-squared with $k-1$ degrees of freedom. The first three moments of a chi-squared statistic with $k-1$ degrees of freedom can be found provided that $\lambda > -1$.

The mean and variance calculated for these expansions were described (Read and Cressie (1988, p. 65)) as :

$$\begin{aligned} E\left[2nI^{\lambda}\left(\tfrac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right)\right] &= [k-1] + n^{-1}[(\lambda-1)(2-3k+t)/3 \\ &\quad +(\lambda-1)(\lambda-2)(1-2k+t)/4] + O(n^{-1}) \end{aligned} \tag{6.2}$$

and Var $\left[2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right)\right]$

$$\begin{aligned} &= [2k-2] + n^{-1}[(2-2k-k^2+t) + (\lambda-1)(8-12k-2k^2+6t) \\ &+(\lambda-1)^2(4-6k-3k^2+5t)/3 + (\lambda-1)(\lambda-2)(2-4k+2t)] + O(n^{-1}) \end{aligned} \tag{6.3}$$

where $t = \sum\limits_{i=1}^{k} {\pi_{0i}}^{-1}$.

Examining each component of (6.2) and (6.3) identifies the first terms on the right hand side as the mean and variance of the chi-squared random variable with $k-1$ degrees of freedom. The second term describes correction terms of order n^{-1}. Read and Cressie (1988) further express the correction terms for the mean and variance as $f_m(\lambda, k, t)$ and $f_v(\lambda, k, t)$ respectively.

The terms for f_m and f_v regulate the rate of convergence of the mean and variance of the power divergence statistic to the mean and variance of the chi-square random variable with $k-1$ degrees of freedom. Thus the values of $\lambda > -1$ for which f_m and f_v are almost zero are of significance.

On examining the equiprobable hypothesis

$$H_0 : \boldsymbol{\pi} = \frac{1}{k}\mathbf{1}, \tag{6.4}$$

$t = \sum_{i=1}^{k} {\pi_{0i}}^{-1} = k^2$, it was reported that for $k \geq 2$,

$$f_m(\lambda, k, k^2) = 0 \text{ when } \lambda = 1 \text{ or } \lambda = 2 - \frac{4(k-2)}{3(k-1)}$$

and

$$f_v(\lambda, k, k^2) = 0 \text{ when } \lambda = \frac{5k - 1 \pm [3(3k^2 - 2k + 7)]^{\frac{1}{2}}}{2(4k-5)}.$$

Solving for λ in $f_m(\lambda, k, k^2) = 0$ and $f_v(\lambda, k, k^2)$ for increasing values of k, resulted in the values of λ being unaffected by the number of cells, k. They further remark that for cell size greater than 50, the outcomes were consistently unchanging whereas, the value of $\lambda = 1$ (Pearson X^2 statistic) reduces the correction term for the mean over all values of k but does this for the variance for only large values of k.

When $t = k^2$ for the equiprobable hypothesis, Pearson's $X^2(\lambda = 1)$ yielded the smallest correction terms for $k \geq 20$. For cases when t strongly influences k^2, selecting $\lambda \in [0.61; 0.67]$ is reported to give the smallest mean and variance correction terms.

Read and Cressie's Moment Corrected Statistic

$P(\chi^2_{k-1} \geq c)$ represents the distribution tail function of the chi-squared distribution or the asymptotic significance level for the critical value c.

Read and Cressie (1988) found that the precision of the small sample approximation to the significance level, can be improved through the use of a moment-corrected distribution tail function formulated from the moment corrected statistic

$$\frac{2nI^{\lambda}(\frac{\mathbf{x}}{n} : \boldsymbol{\pi}_0) - \mu_\lambda}{\sigma_\lambda} \qquad -\infty < \lambda < \infty, \tag{6.5}$$

with $\mu_\lambda = (k-1)(1-\sigma_\lambda) + \dfrac{f_m(\lambda, k, t)}{n}$ and $\sigma_x{}^2 = 1 + \dfrac{f_v(\lambda, k, t)}{2(k-1)n}$.

The mean and variance of this moment corrected statistic is reported to exist for $\lambda > -1$ which is similar to the chi-squared mean and variance, $k-1$ and $2(k-1)$ respectively. Unlike the power-divergence statistic, the corrected statistic exists for $\lambda \leq -1$. The moment corrected distribution tail function is defined by Read and Cressie as $T_C(c) = T_\chi\left(\frac{c-\mu_\lambda}{\sigma_\lambda}\right)$ where $T_\chi(c) = P(\chi^2_{k-1} \geq c)$. Therefore T_c provides a more precise approximation to the small sample significance level of the test based on the power divergence statistic.

6.2 Comparisons between other Approximations to the Exact Significance Level

Rejection of a null hypothesis is based on comparisons between the test statistic and the critical value. Read and Cressie (1988) looked at testing $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ using the test statistic $2nI^{\lambda}(\frac{\mathbf{x}}{n} : \boldsymbol{\pi}_0)$ with $\mathbf{x}$ being the observed value of the multinomial random vector $\mathbf{X}$. Due to the discrete nature of the power divergence statistic a constant c_α must be found for a specified value of λ in order for

$P(2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0) \geq c_{\alpha} \mid H_0) \geq \alpha$ and P(2n $I^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0) > c_{\alpha} \mid H_0) < \alpha$.

Letting $P(2nI^{\lambda}(\frac{\mathbf{X}}{\mathbf{n}} : \boldsymbol{\pi}_0) \geq c \mid H_0) = T_E(c)$ for $c \geq 0$ they express $T_E(c_{\alpha}) \geq \alpha$ and $T_E(c_{\alpha} + \epsilon) < \alpha$; $\epsilon > 0$, as a reformulation of the above probability statements. $T_E(c)$ represents the exact distribution tail function of the test statistic $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$, assuming that H_0 is true.

Further alternatives which could be used as approximations to $T_E(c)$ are reported by Read and Cressie (1988, p. 70) to be :

(i) $T_{\chi}(c)$: the chi-squared distribution tail function.

(ii) $T_C(c)$: the moment corrected chi-squared distribution tail function.

(iii) $T_S(c)$: the second-order-corrected chi-squared distribution tail function

(iv) $T_N(c)$: defined as $P\left(N(0,1) \geq \frac{(c - {\mu_k}^{(\lambda)})}{{\sigma_k}^{(\lambda)}}\right)$, the tail function of the normal distribution.

6.2.1 Assessing Model Accuracy

The attainment of the exact distribution tail function T_E for $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$ was achieved after expressing all possible combinations $\mathbf{x}$ =$(x_1, ..., x_k)$ of n observations arranged into k cells. $T_E(c)$ considered only values of $\mathbf{x}$ for which $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0) \geq c$, where c is a stipulated critical value and thereafter adding their respective multinomial probabilites. The equiprobable hypothesis has been a common area of focus in most small sample studies because firstly, it produces very sensitive tests; secondly, application of the probability integral transformation breaks down numerous goodness-of-fit problems to essentially evaluating the fit of the uniform distribution on [0,1], and lastly, few calculations are required for T_E.

Read and Cressie (1988) selected the commonly used critical value of $c = \chi^2_{1-\alpha}(k-1)$ since it was not dependent on λ. On comparing the sizes of $\mid T_X(\chi^2_{1-\alpha}(k-1)) - T_E(\chi^2_{1-\alpha}(k-1)) \mid$ when the number of cells range from 2 to 6 and values of n spanned the interval of 10 to 50, Read and Cressie (1988, p. 71) observed "that for $10 \leq n \leq 20$ the chi-squared approximation $T_X(\chi^2_{1-\alpha}(k-1)) = \alpha$ is accurate for $T_E(\chi^2_{1-\alpha}(k-1))$ provided that $\lambda \,\epsilon\, [\frac{1}{3}, \frac{3}{2}]$".

These results parallel Larntz's (1978) findings for λ. It was further observed by graphical means that the accuracy of the normal approximation yielded inferior results for many values of n and k. Also noticed was that for an with an increase in n, there is an interval of λ for which the chi-squared critical value can be used to approximate the exact value. However when n is kept constant an increase in the number of cells, k causes greater amounts of error in the significance level to increase for tests using λ outside the interval $[\frac{1}{3}, \frac{3}{2}]$.

Finally based on suggestions by Larntz (1978) and Fienberg (1980, p. 172) on minimum expected cell size for Pearson's X^2, Read and Cressie (1988, p. 72) conclude that "the traditional chi-squared critical value $\chi^2_{1-\alpha}(k-1)$ can be used with accuracy for general k (when testing the equiprobable hypothesis) provided that $\min\limits_{1 \leq i \leq k} n\pi_i \geq 1$, and $\lambda \epsilon\, [\frac{2}{3}, 1]$".

It is thus found that the power divergence statistic with $\lambda = \frac{2}{3}$ and $X^2(\lambda = 1)$ is well approximated by the equiprobable hypothesis whereas the significance levels for $G^2(\lambda = 0)$ and the Freeman-Tukey statistic $F^2(\lambda = -\frac{1}{2})$ did not fare well in terms of approximation by the chi-squared significance level.

6.2.2 Approach for Assessing Accuracy

Commenting that accuracy assessments of the chi-squared approximation, when null models have parameters which must be estimated, is an under-researched area, Read and Cressie highlight approaches that were used in the past. One such estimation method for nuisance parameters required conditioning on sufficient statistics for these parameters in order to eliminate the conditional distribution depending on unknown parameters. Other algorithms can be used as well.

An algorithm was presented by Mehta and Patel in 1986 to compute $r \times c$ tables whose associated probabilities are conditioned on rows and columns. Tables whose probabilities were less than the probability of the observed table were calculated thus improving efficiency for large tables. Another algorithm proposed by Agresti (1979) takes a sample from the set of all possible tables and is useful for approximating the attained significance level for very large tables.

Indepth comparisons between X^2 and G^2 yield that $\lambda = 1$ is preferable to $\lambda = 0$ in terms of the accuracy of the chi-squared significance level.

Rudas (1986) and Hosmane (1987) used simulation with $\lambda = \frac{2}{3}$ and obtained a test statistic which is close to X^2 with regard to small-sample accuracy of the chi-squared significance levels. However G^2 was observed to reject the null hypothesis too frequently in the cases considered. This again supports Cressie and Read's conclusions for those cases where no parameters are estimated and the minimum expected cell frequency is no less than 1.

6.3 Assessing Efficiency : Comparing Power of X^2 and G^2

Assessing how efficient the power-divergence statistic is for small samples required the identification of the previous asymptotic results which would be approximately correct for small samples. Frosini (1976) commented that in the case of the Pearson X^2 statistic, earlier conditions which were established to be appropriate, are very confining in small samples. The approach preferred by Read and Cressie (1988) involved calculating the exact power $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$ for each λ and to thereafter make contrasts without any connection with asymptotic results.

The case of the equiprobable null model is again considered. In this particular case, one of the k probabilities is in a state of disorder whilst the remainder are harmonized so that they still add up to 1 and is expressed by Read and Cressie (1988) as

$$H_1 : \pi_i = \begin{cases} \dfrac{\frac{(1-\delta)}{(k-1)}}{k}, & i = 1, ..., k-1 \\ \dfrac{(1+\delta)}{k}, & i = k \end{cases} \tag{6.6}$$

where $-1 \leq \delta \leq k-1$.

Power computations for each test statistic $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$, requires the calculation of the critical value c_{α} for some chosen α value. This describes a "size-α test". The size of the approximation error is dependent on λ, hence an error in the estimation of the exact size of the test distorts comparisons with λ-dependent approximation errors. In keeping with the critical values being discrete, the significance levels are also discrete. Hence it is unusual that an exact size α test will exist for every λ. This situation can be contended with by use of the appropriate randomized size-α test which can always be obtained thus allowing comparisons of power functions for all λ values.

The randomized size-α test is defined by Read and Cressie (1988, p. 77) as follows :

Let $c_{\lambda}(\alpha)$ be a value that can be obtained by $2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0)$ in order for $P(2nI^{\lambda}(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0) > c_{\lambda}(\alpha) \mid H_0) = \alpha_{1,\lambda}$ and

$$P\left(2nI^{\lambda}\left(\frac{\mathbf{X}}{n} : \boldsymbol{\pi}_0\right) \geq c_{\lambda}(\alpha) \mid H_0\right) = \alpha_{2,\lambda} \tag{6.7}$$

with $\alpha_{1,\lambda} < \alpha \leq \alpha_{2,\lambda}$.

For outcome $\mathbf{x}$, the randomized size-α test rejects H_0 with probability

$$\begin{array}{cl} 1 & \text{if } 2nI^{\lambda}\left(\frac{\mathbf{x}}{\mathbf{n}} : \boldsymbol{\pi}_0\right) > c_{\lambda}(\alpha), \\ \dfrac{\alpha - \alpha_{1,\lambda}}{\alpha_{2,\lambda} - \alpha_{1,\lambda}} & \text{if } 2nI^{\lambda}\left(\frac{\mathbf{x}}{\mathbf{n}} : \boldsymbol{\pi}_0\right) = c_{\lambda}(\alpha), \\ 0 & \text{if } 2nI^{\lambda}\left(\frac{\mathbf{x}}{\mathbf{n}} : \boldsymbol{\pi}_0\right) < c_{\lambda}(\alpha). \end{array}$$

The exact size of the test was further written :

$$1 \cdot \alpha_{1,\lambda} + \frac{\alpha - \alpha_{1,\lambda}}{\alpha_{2,\lambda} - \alpha_{1,\lambda}} \cdot (\alpha_{2,\lambda} - \alpha_{1,\lambda}) = \alpha.$$

Expressing $\beta_{1,\lambda}$ and $\beta_{2,\lambda}$ as :

$$P(2nI^{\lambda}(\frac{\mathbf{x}}{n} : \boldsymbol{\pi}_0) > c_{\lambda}(\alpha) \mid H_1) = \beta_{1,\lambda}$$

$$P(2nI^{\lambda}(\frac{\mathbf{x}}{n} : \boldsymbol{\pi}_0) \geq c_{\lambda}(\alpha) \mid H_1) = \beta_{2,\lambda}$$

for H_1 in (6.7), the power of the randomized test obtained by Read and Cressie (1988) is :

$$\beta_{\lambda} = 1 \cdot \beta_{1,\lambda} + \frac{\alpha - \alpha_{1,\lambda}}{\alpha_{2,\lambda} - \alpha_{1,\lambda}} \cdot (\beta_{2,\lambda} - \beta_{1,\lambda}).$$

Three values of the disturbance factor δ ie : $\delta = 1.5$, $\delta = 0.5$ and $\delta = -0.9$ were considered by Read and Cressie (1988) to depict the bounds of the alternative model (6.7) with a midpoint. Results regarding power of the randomized size-α test yielded the following observations noted by Read and Cressie (1988, p. 77) : when " $k = 2$, the power-divergence statistic has identical critical values and power functions for every value of $\lambda \ \epsilon \ [-5, 5]$. For $k > 2$, the power function increases with λ for $\delta > 0$ (which represents a bump alternative) but decreases with λ for $\delta < 0$ (which represents a dip alternative)". Large $\mid \lambda \mid$ values are associated with small rates of change in the power function. Bump alternatives are detected by using large λ values but for dip type descriptions power improves with large and negative λ. If an individual cell contains a zero observed frequency, the power-divergence statistic with $\lambda \geq -1$ will be undefined and it was suggested that $\lambda > -1$ be used.

6.4 Choice of a Suitable Test Statistic

For critical values which can be estimated with very little difficulty and at the same time providing adequate safeguarding against extreme bump and dip models, it is advised that $\lambda = \frac{2}{3}$ be used as it is found to be the best value for λ is when n is small. Other findings include "For large n, it is an excellent compromise between the loglikelihood ratio statistic $G^2(\lambda = 0)$ which is optimal for nonlocal (fixed) alternatives... and the Pearson $X^2(\lambda = 1)$ statistic which is optimal under sparseness assumptions" (Read and Cressie, (1988, pp. 79–80)). It is further suggested that the critical value be estimated from the chi-square approximation only if $n \geq 10$ and the minimum expected cell frequency more than 1. Expected cell frequencies which are approximately equivalent result in the chi-square approximation being calculated even when expectations are as low as $\frac{1}{4}$. No fixed recommendation can be given for the situation when expected cell frequencies are very small while others are greater than 1.

Using $\mid \lambda \mid$ values greater than 5 should be avoided as the growth in power becomes very small and the approximations used to obtain the critical value for rejecting the null hypothesis becomes more unreliable. Stipulated requirements for values of λ outside the interval $[\frac{1}{3}, \frac{3}{2}]$ are the calculation of the moment corrected approximation based on the moment-corrected distribution tail function whenever $n \geq 10$ and the minimum expected cell frequency is no less than 1. Zero counts led to the power-divergence statistic being undefined for $\lambda \leq -1$, therefore Read and Cressie write that values of $\lambda > -1$, should be used for large sparse arrays. This recommendation will be looked at critically and analysed in the practical applications later.

Chapter 7

Applications using the SAS Software

This chapter looks at the performance of the chi-squared tests for contingency tables under conditions of sparseness. The power-divergence statistic was introduced by Cressie and Read (1988) with the key purpose of investigating and curbing the limitations of the chi-squared tests under conditions of sparseness. The choice of λ was based on different factors. One of the aspects was the expected frequency. The appropriateness of this choice, is examined through applications using a loglinear program of Crowther and Joubert (1988), which utilizes the IML procedure of the SAS software system. Source code for the power-divergence analysis was appended to the IML loglinear program in order to compare the values of the Pearson X^2, likelihood ratio statistic G^2, and the power divergence series. Residuals were also included to provide additional insight into the interpretations. This program appears in the Appendix.

The hypothesis of independence and conditional independence in loglinear models is investigated for frequency tables with small and zero frequencies. The IML loglinear program as well as PROC CATMOD of the SAS system was used for analysis of data sets.

The IML loglinear program presented problems when a matrix, comprising of the columns associated with parameters set equal to zero under independence, became singular, thus PROC CATMOD was also used for the analysis of frequency tables.

Examples presented cover the following areas : small samples, samples with small expectations, tables containing cells with zero cell counts, which are sampling zeros.

7.1 Small Samples and Small Cell Frequencies

The IML procedure of Crowther and Joubert (1988), for estimating the parameters in a loglinear model, was used for the case of a small sample as well as tables with small cell counts, to test the hypothesis of independence and conditional independence by fitting the appropriate models.

7.1.1 Model Fit and Interaction or Independence

In the first example, a small sample was drawn from a slightly larger sample which inspected the relationship of a respondent's reaction to a support pressure group with respect to local area noise. The category counts in the first sample, Fingleton (1984, p. 6) appeared to be quite large and a loglinear analysis of the data yielded expected values which were all above 1 thus avoiding complications of small cell counts or small expected frequencies.

A much smaller sample (Fingleton, 1984, p. 8) was taken from the aforesaid population. The following table contains the data of interest. Responses A_1, A_2, B_1, and B_2 are explained as follows :

"A_1 : Respondent supports the Anti-Stansted airport pressure group".
"A_2 : Respondent does not support pressure group."
"B_1 : Respondent considers local area noisy."
"B_2 : Respondent does not consider local area noisy."

Table 1

	B_1	B_2
A_1	5 (1.69)	2 (5.31)
A_2	2 (5.31)	20 (16.69)

Fingleton observed that the degree of association between pressure group support and noise perception was similar for the above table and the slightly larger table from which the above-mentioned smaller sample was drawn. He arrived at this conclusion by comparing the cross-product ratios for the samples in table 1 and table 1a, below, respectively.

Table 1a contains the data for the larger sample.

Table 1a

	B_1	B_2
A_1	21	25
A_2	3	77

Expected cell frequencies for the independence model appear in brackets in Table 1. Testing the hypothesis of independence, yielded a loglikelihood ratio statistic, $G^2 = 10.27$ and $X^2 = 11.27$. For $\alpha = 0.05$, the null hypothesis is rejected, since $G^2 > \chi^2_{0.95}(1) = 3.84$. There is therefore sufficient evidence to conclude that the variables "pressure support group" and "noise perception" are dependent.

Cressie and Read (1988) recommend that the critical value be calculated from the chi-squared approximation provided that $n \geq 10$ and minimum expected cell frequency is no less than 1. Examination of the expected frequencies in the data indicate that all values were above 1. Hence $\lambda = \frac{2}{3}$ was used in the calculation of the power divergence statistic, for which $2nI^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}\right) = 10.72$. The data was also analysed for different values oflambda. Computations of the power-divergence statistic for lambda ranging between 0 and 2 indicate that the power-divergence statistic is closer to G^2 for values of lambda less than $\frac{2}{3}$ whereas it is closer to the X^2 statistic for values of lambda greater than $\frac{2}{3}$. At $\lambda = \frac{2}{3}$, there seems to be a good compromise between X^2 and G^2 thus justifying our choice of lambda. Our findings are consistent with those of Rudas (1986), who showed that using $\lambda = \frac{2}{3}$ resulted in a test statistic whose accuracy of the chi-square significance level is very similar to that of X^2.

Fienberg (1979) supports the use of the χ^2 test in small samples if the minimum expected cell frequency is approximately 1. If sample size is about 4 to 5 times the number of cells in the table then appropriate p-values are obtained. In the above example, the sample size is 29, the number of cells equals 4 and the expected frequency is almost 1 for only one cell. The p-value obtained is 0.013 which is in keeping with the above statement concerning appropriate values of p on the basis of expected cell frequency and sample size.

The second small sample case study ($n = 52$), Christensen (1990, p. 35) with data displayed in Table 2 below, identifies males between the ages of 11 and 30 who underwent knee operations using orthroscopic surgery and the result of the surgery. Categories for type of injury include descriptions of : twisted knee (T), direct blow (D), or both (B). The results for surgery were described by the following groups : excellent (E), good (G) and fair or poor (F-P).

Table 2

RESULT				
	E	G	F-P	Totals
Twist	21 (21.46)	11 (9.69)	4 (4.84)	36
Direct	3 (4.17)	2 (1.88)	2 (0.94)	7
Both	7 (5.37)	1 (2.42)	1 (1.21)	9
Totals	31	14	7	52

Output of the test produced the following results: $X^2 = 3.228842$, and $G^2 = 3.173231$. Expected values for the independence model appear within brackets in Table 2. From the table it can be seen that only one expected value is less than 1. Provided that the sample size is large, comparisons can be made between X^2 and the χ^2 distribution with 4 degrees of freedom. However, since the sample size is small, using either the X^2 or G^2 statistics will result in distorted conclusions.

Fienberg's conclusion is that if the sample size is 4 to 5 times the cell size, (which is the case in this example), then the χ^2 test yields a p-value with the appropriate order of magnitude. However, this cannot be applied in this example since one of the estimated expected cell frequencies is less than one.

If the hypothesis of independence is tested at the 5% level of significance, i.e. $\alpha = 0.05$, then $\chi^2_{0.95}(4) = 9.488$. Since $X^2 = 3.228842 < 9.4888$, H_0 will not be rejected. There is thus insufficient evidence to indicate a dependence in the relationship between type of injury and the result of the operation.

In order to investigate the value of the power-divergence statistic for different values of λ, $2nI^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}\right)$ was calculated for $-1 < \lambda < 1$. The table below gives the values of $2nI^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}\right)$ for selected values of λ for the independence model.

Table 2a

λ	$2nI^{\lambda}\left(\frac{\mathbf{x}}{n}:\widehat{\boldsymbol{\pi}}\right)$
-0.95	3.337088
-0.50	3.23.553
-0.30	3.201095
-0.10	3.180189
-0.05	3.176422
0.50	3.172612
0.55	3.175660
0.60	3.179274
0.65	3.181295
2/3	3.184981
0.70	3.188210
0.75	3.193535
0.80	3.199435
0.85	3.202601
0.90	3.212970
0.95	3.220612
X^2	3.228842
G^2	3.173231

On comparing the Pearson X^2 and G^2 statistics with the changing values of the power-divergence statistic, the following comments are made : λ values between 0.5 and 0.95 produce $2nI^{\lambda}\left(\frac{\mathbf{x}}{n}:\widehat{\boldsymbol{\pi}}\right)$ values which deviate slightly from each other. Hence an appropriate choice of λ could be any λ value in this interval.

However, on closer examination, it is seen that as λ approaches $\frac{2}{3}$, the power-divergence statistic yields values closer and closer to G^2 but the deviations between these values increase for values greater than $\frac{2}{3}$. On the other hand, the power-divergence values gets closer to X^2 for values greater than $\frac{2}{3}$. These findings are to be expected since the power-divergence statistic approaches X^2 as $\lambda \longrightarrow 1$ and approaches G^2 as $\lambda \longrightarrow 0$. It was further noted that at $\frac{2}{3}$, the X^2 and G^2 values compare quite closely to the power-divergence statistic with the value of G^2 being nearer to the value of the power-divergence statistic.

The next example from Bishop, Fienberg and Holland (1975, p. 148) is an example of a small sample with small cell frequencies. The information presented in table 3 below gives the classification of the reaction of lymphoma patients to chemotherapy classified according to sex and cell type.

Table 3

		Variable 1	
Variable 3	Variable 2	No Response	Response
Nodular	Male	1 (3)	4 (2)
	Female	2 (4.8)	6 (3.2)
Diffuse	Male	12 (7.8)	1 (5.2)
	Female	3 (2.4)	1 (1.6)

The total sample size is 30 and there are cells present with small frequencies. The table contains 8 cells and if we use the traditional chi-square tests on this data set it would lead to distorted p-values since sample size is still less than the number of cells multiplied by either 4 or 5. The three variables of interest are cell type, either nodular or diffuse, sex and response to the treatment. A further observation is that patients with nodular disease seem to respond better than patients with diffuse disease under the treatment offered.

Various models were fitted to test if interaction exists between the variables. The example highlights the versatility of the G^2 statistic when tables are condensed. The ability to subtract G^2 values as a result of partitioning G^2 is also used. The conditional tests of hypothesis could have been also tested using the difference between the X^2 values as both G^2 and X^2 are distributed as χ^2 with the appropriate degrees of freedom. In this case however, only the G^2 statistic is used in the analysis. The advantages of G^2 cannot, however, be generalized to the power-divergence statistic.

The models considered investigate the relationship between :

(i) sex and response to the treatment,

(ii) sex and cell type, and

(iii) cell type and response to treatment.

To test (i) a model with $u_{123} = 0$ is found. Thereafter the model with $u_{12} = u_{123} = 0$ is obtained. The overall difference between $G^2(V_1V_2, V_2V_3, V_1V_3)$ and $G^2(V_3V_1, V_3V_1V_2)$, indicates whether interaction exists between the variables or not.

The output generated by the loglinear programme utilizing Proc IML for the test of no three factor interaction yields the following values : $G^2 = 0.650743$ with a p-value of 0.4191847. The second model yields a G^2 value of 0.809475 with a p-value of 0.667152. The hypothesis that $u_{12} = 0$ is tested by finding the difference $G^2(V_3V_2V_1|V_2V_1) = G^2(V_1V_2) - G^2(V_3V_2V_1) = 0.809475 - 0.650743 = 0.158732 < \chi^2_{0.95}(1) = 3.84$. This indicates that there is insufficient evidence to conclude that the u_{12} term is nonzero. Hence there appears to be no interaction between sex and response.

On testing the second model for interaction between sex and cell type, the usual approach would have been to first get the model for no three factor interaction, that is, obtain a model for $u_{123} = 0$ and thereafter to obtain the model for $u_{23} = u_{123} = 0$ with the difference between the models used to test whether there is an interaction between sex and cell type. This approach is however omitted since it is already established that $u_{12} = 0$. Instead the

G^2 value is found for the model in which $u_{23} = u_{12} = u_{123} = 0$, thereafter subtracting the G^2 value for the model with $u_{12} = u_{123} = 0$. This difference is then used to test whether $u_{23} = 0$.

The model with $u_{23} = u_{12} = u_{123} = 0$ gives a G^2 value of 5.316720 with a p-value of 0.150021. Testing $u_{23} = 0$, requires the difference, $5.316720 - 0.809475 = 4.507245$. Since this value exceeds $\chi^2_{0.95}(1) = 3.84$, this indicates that there is evidence of interaction between sex and cell type. In other words there is a relationship between the patient's sex and whether they suffer with nodular or diffuse disease.

The value for the difference contained in Bishop, Fienberg and Holland appears to be in error, hence the inconsistency in the final answer for the differences between their finding and our answer. Lastly, a test was conducted along similar lines to test if any interaction exists between the type of cell disease and the response which essentially means that we have to assess the magnitude of the interaction between cell type and response. Again we find G^2 values for two models and subtract one from the other to obtain overall interaction.

Taking G^2 for the model in which $u_{13} = u_{12} = u_{123} = 0$ and subtracting the G^2 value for the model in which $u_{12} = u_{123} = 0$ gives $14.829709 - 0.809475 = 14.020315$. Once again this value exceeds $\chi^2_{0.95}(1)$ and it can be concluded that there is an interaction effect between the type of cell disease and the response to the treatment.

We have thus found that u_{13} and u_{23} are not zero, thus cell type is related to both sex and response.

7.2 Cells With Zero Counts

Contingency tables with zero cell frequencies are obtained quite frequently. These cells of zero magnitude cause complications in loglinear applications. Some cells will always have a zero frequency, because of the fact that it is impossible to classify an individual in a particular cell. For example the number of males suffering from menstrual problems. Such zeros are called fixed or structural zeros. They have a true probability of zero and cell size will always be zero immaterial of sample size. The other type of zero frequency is a random zero or sampling zero. Cells containing such zeros have the possibility of containing a positive cell count and their true probability is positive. With an increase in sample size, these cells could increase to positive counts.

Some of the problems presented by random zeros include asymptotic results being invalid and maximum likelihood estimates of the parameter not existing. Sampling or random zeros occur more often than structural zeros. Tables containing structural zeros are called incomplete tables. For the purpose of this study, our discussion will be restricted to sampling or random zeros. Cells with zero counts affect the existence of maximum likelihood estimates in loglinear models. Parameter estimates having values of positive or negative infinity indicate that the likelihood function keeps increasing as the parameter moves toward positive or negative infinity.

Empty cells or cells of zero count also lead to poor approximation of the goodness-of-fit statistics. The problem which arises when taking the logarithm of the zero frequency is addressed by adding a very small constant like 10^{-8} to the zero cell. The size of the constant in terms of "smallness" is not restricted. Some authors suggest that the zero cell frequency should be adjusted by adding a constant of 0.5. The effect of the adjustment to empty cells on parameter estimates and goodness-of-fit statistics was investigated. A data set was analysed with different "small" values for a constant added to the zero frequency cell and comparisons were made. Special attention was paid to the case where cells with zero frequency are adjusted by 0.5.

An example relating to sampling zeros appears in Agresti (1996, p. 186). Table 4, obtained from the 1991 General Social Survey, investigated whether there is any association between job satisfaction (S) and Income (I), grouped by gender (G).

Table 4

		JOB SATISFACTION			
Gender	Income	Very Dissatisfied	A Little Dissatisfied	Moderately Satisfied	Very Satisfied
Female	< 5000	1	3	11	2
	5000-15000	2	3	17	3
	15000-25000	0	1	8	5
	> 25000	0	2	4	2
Male	< 5000	1	1	2	1
	5000-15000	0	3	5	1
	15000-25000	0	0	7	3
	> 25000	0	1	9	6

The table contains small cell frequencies and sampling zeros, hence it is suspected that the usual chi-squared tests may not be appropriate. Agresti's approach checked if the $I * S$ term measuring association can be taken out of the model with (IS, GI, GS) interactions using the GENMOD procedure of SAS. This is achieved by obtaining the G^2 values for both the (GI, GS) and (IS, GI, GS) models. The model for $G^2[(GI, GS)|(IS, GI, GS)]$ yields the eventual result. His reported findings highlighted $G^2(GI, GS) = 19.4$ with 18 degrees of freedom and $G^2(IS, GI, GS) = 7.1$ with 9 degrees of freedom. The test statistic for the hypothesis of conditional independence is hence

$G^2[(GI,GS)|(IS,GI,GS)] = 19.4 - 7.1 = 12.3$ with a total of $18 - 9 = 9$ degrees of freedom. The p-value of 0.2 suggests that there is insufficient evidence to indicate that association exists.

Validation of Agresti's findings were accomplished through analyses using the Catmod and the IML loglinear programme. Comparisons were carried out between results yielded by the CATMOD procedure and the IML routine. It was observed that as soon as zero cells were made extremely small (e.g. 10^{-15}), the IML loglinear procedure encountered a singular matrix when fitting certain models, hence analyses were run using PROC CATMOD to check if similar problems were experienced. General discussions as well as findings for each procedure with respect to the above-mentioned example follow.

7.2.1 The Catmod Procedure

PROC CATMOD can utilize the maximum likelihood analysis or the weighted least squares analysis. This procedure treats any zero in the data set as a *structural zero*, but if there is more than one population, then a zero can be treated as a sampling zero. When the algorithm of PROC CATMOD is required to evaluate the logarithm of a zero in the estimation routine, it automatically adjusts to take a logarithm of some small value in order to continue. If it is known that the zero cell frequencies are sampling zeros then a frequency say, $f = 0$, can be adjusted by inserting an "if statement" :
if $f = 0$, then $f = f + 0.00000001$ in the data section of the program.

7.2.2 Maximum Likelihood Approach

The maximum likelihood approach was first used to measure the (GI, GS) interaction in order to make comparisons with Agresti's findings. On examining the goodness of fit for the data contained in the maximum likelihood analysis of variance table from the output of PROC CATMOD, it is seen that the results are the same as those of Agresti. Zero cell frequencies were adjusted by adding the constant 0.000000001, thus obtaining $G^2 = 19.37$ with 18 degrees of freedom. The standard errors obtained for each of the parameters were small. These results are the same as those reported by Agresti.

A sensitivity analysis was conducted by comparing results of analyses when zero cells were adjusted to the following values to $10^{-8}, 10^{-15}, 10^{-16}, 10^{-17}$ $10^{-18}, 10^{-20}, 10^{-25}$, and 10^{-30}.

The following findings were noted for the above scenario : examining the analysis of maximum-likelihood estimates for each parameter for the above adjustments on the zero cell frequency indicates that the outputs were identical for all the cases. In other words, the values of the parameter estimates, standard errors, chi-squared values and p-values remain the same. The only difference in output arose for the calculation of the maximum-likelihood predicted values for response functions and frequencies. Large absolute values of residuals occurred for functions 9, 21, 25, 26 and 29. This finding was consistent with all of the above trials.

PROC CATMOD thus seems to be stable when an extremely small value is used to replace a zero cell frequency and converges consistently to the same parameter estimates.

7.2.3 Weighted Least Squares Approach

In the weighted least squares approach the weighted residual sums of squares for the model are made as small as possible.

The weighted least squares statement was applied in two ways : firstly, a small value is assigned to the zero cell frequencies by the user by setting the frequency f equal to some value say, 0.00000001 and secondly the zero cell frequencies were left unadjusted. The latter analysis allows SAS to automatically take care of the zero cell frequency. The following observations were made :

When f is stipulated as some value like $f = f + 0.000000001$ the output provides values for $n - 1 = 31$ cells. On the other hand when f is not assigned some adjusted value, the output presented indicates that the 6 cells containing zero cell frequencies were omitted from the analysis with findings being presented for only $n - 1 = 25$ cells. This confirms the fact that PROC CATMOD takes zeros as structural zeros.

To investigate the effect of varying sizes of the adjustment made by the user to the zero cell through assignment some small value to f, analyses were run for the following assignments to zero frequencies : $10^{-8}, 10^{-15}, 10^{-20}$, and 10^{-25}. The parameter estimates were not affected by the size of the constant added to zero cell.

Although in large samples with categorical data, the weighted least squares estimators have similar properties to the maximum likelihood estimators, in the case of small samples with zero cell frequencies, it is suggested that the maximum likelihood approach be used as the CATMOD procedure calculates the observed response function for the weighted least squares analysis.

Finally, on using either the weighted least squares approach or the maximum likelihood approach with PROC CATMOD, the following point needs to be noted : it appears that CATMOD treats all zeros as structural zeros hence special care must be taken if one is working with data that contains sampling zeros so that they are adjusted appropriately to avoid them being incorrectly analysed as structural zeros.

7.2.4 The ADDCELL Option

A further option available in PROC CATMOD is the ADDCELL statement. It is stated as follows : ADDCELL = some number, where the number should be positive. This is said to facilitate automatic adjustment to cells with zero frequency. It has no effect on maximum likelihood analyses but can be used for the weighted least squares approach. Applications were carried out on the model testing interaction between the $G*I$ and $G*S$ effects using the weighted least squares approach. The following cases were compared :

(a) weighted least squares with f assigned 10^{-15},

(b) weighted least squares with ADDCELL statement assigned to 10^{-15} and

(c) weighted least squares with no adjustments being made to zero cells, i.e. a zero appears in the data set and is not adjusted.

In case (a) the output yields a residual of 8.20 with 18 degrees of freedom and a p-value of 0.9756. There is a change in the findings for (b) : the residual = 8.20 but the degrees of freedom are now 12, since the CATMOD analysis has taken a zero as a structural zero. The residual occurs with a p-value of 0.7695. The output for case (c) is identical to (b)'s findings. It appears that SAS handles the ADDCELL option in exactly the same way as when the weighted least squares analysis is run with no adjustment to a zero cell frequency.

Hence, the ADDCELL statement in the weighted least squares analysis seems to treat zero cells as structural zeros. These findings are inconsistent with the description of the ADDCELL option given by the SAS manual, as it needs some positive value to be assigned to it in order for the program to execute. However, the output indicates that it ignores the assigned value and executes as if no adjustments were made to the zero cells. It is therefore advised that the ADDCELL statement be avoided and that some small value be assigned to a zero cell in the SAS program statements, to ensure that the analysis is done on all the cells.

7.3 The IML Procedure

The results obtained by Agresti (1996) for the (IS, GI, GS) interaction model and those obtained by the loglinear program using PROC IML were identical. The results yielded a G^2 value of 7.093 with 9 degrees of freedom. Ideally we would like the value of the adjustment for zero cell frequencies to be as small as possible. However, PROC IML either experiences problems when a singular matrix occurs or very large *StandLH* values are found for the G, $G * I$ and $G * S$ effects. This problem is eliminated by keeping the adjustments to cells to approximately $10^{-6}, 10^{-7}$ or at most 10^{-8}. The analysis of the model for (GI, GS) using the IML procedure with the loglinear approach yields $G^2 = 19.368408$ with a p-value of 0.369475, which are both identical to Agresti's findings. If the adjusted cell size is set to 10^{-15} or smaller, then analysis is not possible as a singular matrix occurs in the estimation procedure.

7.4 Comparisons Between the IML and Catmod Procedures

Outputs for the (IS, GI, GS) model were obtained for varying sizes of adjustments made to zero cell frequencies by using the loglinear approach in IML and thereafter through use of the Catmod procedure. The maximum likelihood approach was used in these analyses.

As reported earlier, Proc Catmod does not seem to be sensitive to the magnitude of the constant that is added to a zero cell. But the recommendation is to use a value such as 10^{-8} as the adjustment. The value of the likelihood ratio statistic, for the above-mentioned model, is 7.09 with a p-value of 0.8972 and 13 degrees of freedom.

The loglinear program utilizing the IML procedure on the other hand, consistently yields a likelihood ratio value of 7.093491 with 9 degrees of freedom and a probability value of 0.627386 with a corresponding Pearson chi-square value of 6.605016 with 9 degrees of freedom and a probability value of 0.678167, as long as the adjustment made to zero cells is made no smaller than 10^{-8}. If the adjustment to zero cells is made smaller than 10^{-8}, then instabilities occur in the IML program. These instabilities can affect the *standLH* values for some adjusted values. It is therefore recommended that adjustments to zero cells be made no smaller than 10^{-8}.

Further analysis was conducted for adjustments which were larger than 10^{-8} up to 0.5. In the case of the adjusted cell being assigned the value of 0.5, a chi-square value of 5.172430 with a probability value of 0.819025 was observed whilst the $G^2 = 4.879436$ with a corresponding p-value of 0.844689. The power-divergence statistic value of 5.046544 for $\lambda = \frac{2}{3}$ was closer to the the Pearson chi-square than G^2, whilst the difference between the Pearson chi-square and G^2 statistic increased.

The fluctuations in the values of the goodness-of-fit statistics, as the adjusted cell gets closer to 0.5, suggests that large adjusted cell sizes are not favoured hence such adjustments should be avoided, rather adjustments which are closer to 0 and that yield more consistent results should be considered.

These findings are justified by our results obtained when the hypothesis, $\lambda^{GI} = \lambda^{GS} = \lambda^{GIS} = 0$ was tested. Using an adjustment of 0.5 gave $G^2 = 24.417746$ with a p-value of 0.27326 and $X^2 = 24.268192$ with a p-value of 0.280246, (df = 21). This indicated that the model fitted was adequate for the data. On the other hand, an adjustment of 10^{-8} yielded the following results for the above analysis : $G^2 = 31.91945$ with a p-value of 0.084679 whilst $X^2 = 30.376529$ with a corresponding p-value of 0.059659. These findings indicate that the model is not really adequate for the data and that the fit is not good, in contrast with the conclusion for the adjustment of 0.5. This distinct difference in conclusions cautions one to choose the size of the adjustment carefully. Also, there is a big difference between the X^2 and G^2 values for each of these adjusted cell sizes. Hence it is highlighted once more that zero cells should be adjusted to 10^{-8} rather than using the larger adjustment of 0.5.

Using the above recommendation that the preferred adjustment for cells having a zero frequency be 10^{-8}, further analysis was carried out to observe the effect of changing lambda on the power-divergence statistic. Once more, values of λ in the interval 0.5 to 0.95 were used. The following table gives the values of $2nI^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}\right)$ for the selected values of λ for the (IS, GI, GS) model.

Table 4a

λ	$2nI^{\lambda}\left(\frac{\mathbf{x}}{n} : \widehat{\boldsymbol{\pi}}\right)$
0.50	6.650788
0.55	6.633281
0.60	6.619133
0.65	6.608119
2/3	6.604546
0.70	6.600043
0.75	6.594735
0.80	6.592044
0.85	6.591837
0.90	6.593996
0.95	6.598420
X^2	6.605016
G^2	7.093490

As expected, when λ approaches 0, the power-divergence statistic yields values closer to G^2. For values of λ closer to 1, the power-divergence statistic is closer to X^2. It is suggested that λ can be chosen to be any value in the above interval but $\lambda = \frac{2}{3}$ results in a power-divergence statistic which seems to provide a good compromise between $G^2(\lambda \rightarrow 0)$ and $X^2(\lambda \rightarrow 1)$.

Conclusion

It is hoped that this dissertation has presented the chosen topics on aspects of categorical data analysis in an organised manner and that the area of sparse contingency tables has been adequately addressed.

Areas of future research could include parameter estimation and hypothesis testing for the power-divergence statistic under the sparseness assumption. One particular avenue of interest is to investigate the use of jackknifing as an estimation method under conditions of sparseness. This study contains a brief discussion on jackknifing in section 4.7 of chapter four.

Another topic that has been addressed in detail in articles not covered here, is the use of the Akaike's Information Criterion as measure of goodness of fit. This criterion can also be investigated and compared with the power-divergence statistic.

Lastly, from the applications undertaken, it is noted that the procedure introduced by Crowther and Joubert (1988) presented problems, when a matrix comprising of the columns associated with parameters set equal to zero under independence, became singular. This problem was avoided by using adjustments of about $10^{-6}, 10^{-7}$, or at best 10^{-8}. Future research could consider addressing the problems experienced by the IML routine regarding the size of adjustments and the occurrence of large *StandLH* values.

Bibliography

[1] Agresti, A., (1996). *An Introduction into Categorical Data Analysis.* Wiley.

[2] Agresti, A., Wackerley, D., and Boyett, J.M. (1979). *Exact Conditional Tests for Cross-classifications : Approximation of Attained Significance Levels.* Psychometrika 44, 75–83.

[3] Beatty, G., (1983). *Salary Survey of Mathematicians and Statisticians in Proceedings of the Section on Survey Research Methods.* American Statistical Association, 743–747.

[4] Birch, M. W., (1964). *A New Proof of the Pearson-Fisher Theorem.* Annals of Mathematical Statistics 35, 817–824.

[5] Bishop, Y. W., Fienberg, S. E., and Holland, P.W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA : MIT Press.

[6] Chapman, J. W., (1976). *A Comparison of the χ^2, -2 log R and Multinomial Probability Criteria for Significance Tests when Expected Frequencies are Small.* Journal of American Statistical Association, 71, 854–862.

[7] Christensen, R. (1990). *Log-Linear Models.* Springer-Verlag, New York, Inc.

[8] Cochran, W. G., (1936). *The χ^2 Distribution for the Binomial and Poisson Series, with Small Expectations.* Annals of Eugenics 7 No. 2, 207–17.

[9] Cochran, W. G., (1942). *The χ^2 Correction for Continuity.* Iowa State College. Journal of Sci. 16, 421−436.

[10] Cochran, W. G., (1952). *The χ^2 Test of Goodness of Fit.* Annals of Mathematical Statistics, 3, 315−345.

[11] Conahan, M. A., (1970). *The Comparative Accuracy of the Likelihood Ratio and X^2 as Approximations to the Exact Multinomial Test.* Unpublished Ph.D. dissertation, Department of Education, Lehigh University.

[12] Cramer, H., (1946). *Mathematical Methods of Statistics.* Princeton, NJ : Princeton University Press.

[13] Cressie N. A. C., and Read T. R. C., (1984). *Multinomial Goodness-of-fit Tests.* Journal of Royal Statistical Society Series B 46, 440−464.

[14] Crowther, N. A. S., Joubert, H. M., (1988). *Statistical Modelling of Categorical Data.* Institute for Statistical Research WS−41. (HSRC), Pretoria.

[15] Effron, B., (1979). *Bootstrap Methods : Another Look at the Jackknife.* The Annals of Statistics, 7, 1−26.

[16] Fienberg, S. E., (1979). *The Use of the Chi-squared Statistics for Categorical Data Problems.* Journal of the Royal Statistical Society Series B 41, 54−64.

[17] Fingleton, B. (1984). *Models of Category Counts.* Cambridge University Press.

[18] Fisher R. A., (1924). *The Conditions Under which χ^2 Measures the Discrepancy between Observations and Hypothesis.* Journal of the Royal Statistical Society 87, 442−450.

[19] Fisher R. A., (1941). *Statistical Methods for Research Workers, 8th ed.*, Edinburgh : Oliver and Boyd, Ltd, 1941.

[20] Fisher R. A., (1950). *The Significance of Deviations from Expectations in a Poisson Series.* Biometrics, 6, 17–24.

[21] Frosini B. V., (1976). *On the Power Function of the χ^2 Test.* Metron 34, 3–36.

[22] Goldstein M., Wolf E., Dillon W.,(1976). *On a Test of Independence for Contingency Tables.* Communications in Statistics - Theory and Methods 5, 159–169.

[23] Good I. J., Gover T., N., and Mitchell G., J., (1970). *Exact Distributions for X^2 and for the Likelihood-ratio Statistic for the Equiprobable Multinomial Distribution.* Journal of the American Statistical Association 65, 267–283.

[24] Gurian J. M., Cornfield J., Mosimann J. E., (1964). *Comparisons of Power for Some Exact Multinomial Significance Tests.* Psychometrika No. 4, 409–19.

[25] Haberman S. J., (1974)., *The Analysis of Frequency Data.* Chicago, University of Chicago Press.

[26] Haberman S.J. (1977)., *Log-linear Models and Frequency Tables with Small Expected Cell Counts.* Annals of Statistics 5, 1148 –1169).

[27] Haldane J. B. S., (1937). *The Exact Value of the Moments of the Distribution of χ^2, used as a Test of Goodness of Fit, When Expectations are Small.* Biometrika 29, 133–143.

[28] Haldane J. B. S., (1939). *The Mean and Variance of χ^2, When used as a Test of Homogeneity, when Expectations are Small.* Biometrika 31, 346–355.

[29] Hoeffding W., (1965). *Asymptotically Optimal Tests for Multinomial Distributions.* Annals of Mathematical Statistics 36, 369–408.

[30] Holst L., (1972). *Asymptotic Normality and Efficiency for Certain Goodness-of-fit Tests.* Biometrika 59, 137–145, 699.

[31] Hosmane B., (1987). *An Empirical Investigation of Chi-square Tests for the Hypothesis of No Three-factor Interaction in $I \times J \times K$ Contingency Tables.* Journal of Statistical Computation and Simulation 28, 167–178.

[32] Ivechenko G. I., and Mevedev Y. I., (1978). *Separable Statistics and Hypothesis Testing. The case of Small Samples.* Theory of Probability and Its Applications 23, 764–775.

[33] Kallenberg W. C. M., Oosterhoff J., and Schriver B. F., (1985). *The Number of Classes in Chi-squared Goodness-of-fit Tests.* Journal of the American Statistical Association 80, 959–968.

[34] Kendall M. G., (1952). *The Advanced Theory of Statistics.* Vol 1, 5th ed., London : Griffin.

[35] Koehler K. J., (1977). *Goodness-of-fit Statistics for Large Sparse Multinomials.* Ph.D. Dissertation, School of Statistics, University of Minnesota, Minneapolis, MN.

[36] Koehler K. J., and Larntz K., (1980). *An empirical Investigation of Goodness-of-fit Statistics for Sparse Multinomials.* Journal of the American Statistical Association 75, 336–344.

[37] Koehler K. J., (1986). *Goodness-of-fit Tests for Log-Linear Models in Sparse Contingency Tables.* Journal of the Americal Statistical Association 81, 483–493.

[38] Lancaster H. O., (1969). *The Chi-Squared Distribution.* New York, John Wiley.

[39] Larntz K., (1978). *Small-sample Comparisons of Exact Levels for Chi-squared Goodness-of-fit Statistics.* Journal of the American Statistical Association 73, 253–263.

[40] Lawal H. B., and Upton G. J. G., (1980). *An Aproximation to the Distribution of χ^2 Goodness-of-fit Statistic for use with Small Expectations.* Biometrika 67, 447–53.

[41] Lehmer D. H., (1964). *The Machine Tools of Combinatorics, in E.F. Beckenbach, ed., Applied Combinatorial Mathematics.* New York : John Wiley & Sons, inc., 25-6.

[42] Mehta C. R. and Patel N. R., (1986). *Algorithm 643. FEXACT : A FORTRAN subroutine for Fisher's Exact test on Unordered $r \times c$ contingency tables.* ACM Transactions on Mathematical Software 12, 154–161.

[43] Moore D. S., (1986). *Tests of Chi-squared Type. In Goodness-of-Fit Techniques (editors R. B. D'Augustino and M. A. Stephens).* 63–95. New York, Marcel Dekker.

[44] Morris C., (1975). *Central Limit Theorems for Multinomial Sums.* Annals of Statistics 3, 165–188.

[45] Nass C. A. G., (1959). *The χ^2 Tests for Small Expectations in Contingency Tables, with Special Reference to Accidents and Absenteeism.* Biometrika, 46, 365–85.

[46] Neyman J., (1949). *Contribution to the Theory of the χ^2 Test.* Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, 239–273.

[47] Neyman J., and Pearson E. S., (1931). *Further Notes on the χ^2 Distribution.* Biometrika 22, 298–305.

[48] McCullagh P., (1986). *The Conditional Distribution of Goodness-of-fit Statistics for Discrete Data.* Journal of the American Statistical Association 81, 104–107.

[49] Pearson K., (1900). *On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables is*

Such That it can be Reasonably Supposed to Have Arisen From Random Sampling. Philosophy Magazine 50, 157–172.

[50] Pearson K., (1932). *Experimental Discussion of the (χ^2, P) Test for Goodness-of-fit.* Biometrika 24, 351–81.

[51] Radlow R., and Alf E. F., (1975). *An Alternate Multinomial Assessment of the Accuracy of the χ^2 Test of Goodness of Fit.* Journal of the American Statistical Association 70, 811–813.

[52] Read T. R. C., and Cressie N. A. C., (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer-Verlag.

[53] Roscoe J. T., and Byars J. A., (1971). *An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic.* Journal of the American Statistical Association, 66, 755–59.

[54] Rudas T., (1986). *A Monte Carlo comparison of the Small Sample Behaviour of the Pearson, the Likelihood Ratio and the Cressie-Read Statistics.* Journal of Statistical Computation and Simulation 24, 107–120.

[55] SAS Institute Inc., (1988). *SAS/STAT User's Guide, Release 6.03 Edition.* Cary, NC : SAS Institute Inc.

[56] Searle S. R., (1971). *Linear Models.* John Wiley and Sons Inc.

[57] Shanaway M. R., (1936). *An Illustration of the Accuracy of the χ^2 Approximation.* Biometrika, 28, 179–87.

[58] Simonoff J. S., (1982). *The Application of Penalized Likelihood to the Smoothing of Large Sparse Multinomials.* Proceedings of the Statistical Computing Section, American Statistical Association, 180–183.

[59] Simonoff J. S., (1983). *A penalty Function Approach to Smoothing Large Sparse Contingency Tables.* Annals of Statistics, 11, 208–218.

[60] Simonoff J. S., (1985). *An Improved Goodness-of-fit Statistic For Sparse Multinomials.* Journal of the American Statistical Association, 80, 671–677.

[61] Simonoff J. S., (1986). *Jackknifing and Bootstrapping Goodness-of-fit Statistics in Sparse Multinomials.* Journal of the American Statistical Association, 81, 1005–1011.

[62] Simonoff J. S., (1987). *Probability Estimation Via Smoothing In Sparse Contingency Tables With Ordered Categories.* Statistics and Probability Letters, 5, 55–63.

[63] Slakter M. J. (1966). *Comparative Validity of the Chi-Square and Two Modified Chi-Square Goodness-of-Fit Tests for Small but Equal Expected Frequencies.* Biometrika, 53, 619–23.

[64] Steck G.P., (1957). *Limit Theorems for Conditional Distributions.* University of California Publications in Statistics, 2, 12, 237–284.

[65] Tate M. W., and Hyer L. A., (1973). *Inaccuracy of the X^2 Test of Goodness of Fit When Expected Frequencies Are Small.* Journal of the American Statistical Association 68, 836–841.

[66] Van der Waerden B. L., (1957). *Mathematisch Statistik.* Berlin : Springer-Verlag.

[67] Wakimoto K., Odaka Y., and Kang L. (1987). *Testing the Goodness of Fit of the Multinomial Distribution based on Graphical Representation.* Computational Statistics and Data Analysis 5, 137–147.

[68] West E.N. and Kempthorne O., (1971). *A Comparison of the* Chi^2 *and Likelihood Ratio Tests for Composite Alternatives.* Journal of Statistical Computation and Simulation 1, 1–33.

[69] Yarnold J. K., (1970). *The Minimum Expectation In X^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution.* Journal of the American Statistical Association 65, 864–886.

[70] Yule G. U., (1911). *An Introduction to the Theory of Statistics.* Griffin.

[71] Zelterman D., (1984). *Approximating the Distribution of Goodness of Tests for Discrete Data.* Computational Statistics and Data Analysis 2, 207–214.

[72] Zelterman D., (1986). *The Log-likelihood Ratio for Sparse Multinomial Mixtures.* Statistics and Probability Letters 4, 95–99.

[73] Zelterman D., (1987). *Goodness-of-fit Tests for Large Sparse Multinomial Distributions.* Journal of the American Statistical Association 82, 624–629.

Appendix

Program 1

The source code below appeared in Crowther and Joubert (1988). It was further adapted for the power-divergence statistic. The following program uses the IML loglinear option and analyses data from Agresti (1996, p.. 186) which is quoted in table 4 of chapter 7. The data refers to the 1991 General Social Survey which looked at the relationship between job satisfaction (S) and income (I) grouped according to gender (G).

```
proc iml worksize= 100;
options pagesize=200;
* FREQUENCY VECTOR(factor that changes slowest first );
 x= { 1, 3, 11, 2,
       2, 3, 17, 3,
       0, 1,  8, 5,
       0, 2,  4, 2,
       1, 1,  2, 1,
       0, 3,  5, 1,
       0, 0,  7, 3,
       0, 1,  9, 6};
  xr=nrow(x);
  x=x<>J(xr,1,1e-8);
*******************************************************************************
* The above line is inserted to adjust the zero cell frequencies which occur in
 the matrix, of size 32, to some positive value. In this case, cells with zero
  frequency are adjusted to 10^-8;
* The line can be omitted if no cells with zero frequency occur.;
* NUMBER OF VARIABLES;
*------------; nf=3;
* NAME OF VARIABLES;
*---------; name={"g.","i.","s."};
*--------------------------------------------------;
* POWER SERIES;
*---------; POW= 0.66667;
*--------------------------------------------------;
k=j(6,1,0);
```

```
* NUMBER OF LEVELS FOR EACH FACTOR  (MAX 6 FACTORS);
*---------; k[1,]=2  ;k[2,]=4  ;k[3,]=4  ; k[4,]=0  ;k[5,]=0 ;k[6,]=0 ;
*------------------------------------------------------------;
* SPECIFICATION OF HYPOTHESIS MATRIX AH ;
* index vector nh of the hypothesis in the order (lambdas set to zero)
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  ens.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*-----------;  nh={8 };
*-----;G1={ 0  0 };
*------------------------------------------------------------;
* CONSTRUCTION OF DESIGN MATRIX A;
reset nolog;
reset fw=10;
c=k[1,];
een=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k[1];
do i=2 to nf;
c=k[i,];
een=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
een1=j(e,1,1);
A1=A@een;
Y1=een1@Y;
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k[i,]*e;
end;
A=j(e,1,1)||A;
*------------------------------------------------------------;
vg=1;
do i=1 to nf;
vg=vg//((k[i,]-1)*vg);
end;
kol=cusum(vg);
nrh=nrow(nh);
```

```
ii=nh[1,]-1;iii=nh[1,];
a1=kol[ii,]+1;a2=kol[iii,];
AH=A[,a1:a2];
do i=2 to nrh;
ii=nh[i,]$-$1;iii=nh[i,];
a1=kol[ii,]+1;a2=kol[iii,];
AH=AH||A[,a1:a2];
end;
* CONSTRUCTION OF THE INDEX VECTOR;
tyd=name[1,1];
naam1={"" }//tyd;
do i=2 to nf;
naam1=naam1//concat(naam1,name[i,1]);
end;
naam1=rowcatc(naam1);
nn=nrow(naam1);
index={"mu"};
do i=2 to nn;
tyd=naam1[i,1];
index=index//repeat(tyd,vg[i,1]);
end;
*----------------------------------------------------------------;
* HYPOTHESIS MATRIX WITH STRUCTURE;
sg=sum(g1*g1‘);
if sg^=0 then AH=AH*G1‘;
*----------------------------------------------------------------;
free Y A1 Y1 tyd naam1 ;
A=inv(A‘*A)*A‘;
lambda=A*log(x);
    x1=1/x;
varl=A*(x1#A‘);
stdl=sqrt(vecdiag(varl));
standl=lambda/stdl;
free varl stdl ;
gx=AH‘*log(x);
   m=x;
   gm=gx;
   itr=0;
   diff=1;
   do  while (diff> 0.000001);
   m1=m;
   mi=1/m;
   m=m-AH*inv(AH'*(mi#AH))*gm;
m=m<>J(xr,1,1e-10);
*------- The above line is omitted when the data contains no zero cells ;
   gm=AH'*log(m);
   diff=sqrt((m-m1)'*(m-m1));
   itr=itr+1;
   end;
lambdah=A*log(m);
varlh=vecdiag(A*(mi#A'))-vecdiag(A*(mi#AH)*inv(AH'*(mi#AH))*AH'*(mi#A'));
vecvar=varlh<>J(e,1,1E-10);
stdlh=sqrt(vecvar);
standlh=lambdah/stdlh;
vgh=ncol(ah);
X2=(x-m)'*(mi#(x-m));
G2=2*x'*log(x/m);
K2ft=4*(sqrt(x)-sqrt(m))'*(sqrt(x)-sqrt(m));
Wald=gx'*inv(AH'*(x1#AH))*gx;
xdmpl=(x/m)##pow;
sbr=xdmpl-J(xr,1,1);
pdivser=2*(x'*sbr)/(pow*(pow+1));
vec=x2||G2||K2ft||Wald||pdivser;
prob=J(1,5,1)-probchi(vec,vgh);
 resid=(x-m)/sqrt(m);
*output;
vec1={"Pearson" "LR" "F-T" "Wald" "POWDIV"};
R={"Chi^2" "Df" "Prob"};
Test=vec//J(1,5,vgh)//prob;
nrt=xr/k[nf];
x=shape(x,nrt);
m=shape(m,nrt);
resid =shape(resid,nrt);
print "------------LOG.IML-----------------";
print"number of iterations=" itr;
print" " ;
```

```
print x[format=7.1]
       m[format=12.6] ;
print" ";
print  index lambda[format=12.6] standl[format=12.6]
       lambdah [format=12.6] standlh[format=12.6];
print" ";
print "Chi-squared statistics with p-values";
print Test[rowname=R colname=vec1 format=12.6];
print" ";
print "Standardized Residuals";
print resid[format=12.6];
```

Program 2

This program uses PROC CATMOD of the SAS system for the data in Table 4 of chapter 7.

```
data gis;
input g i s f;
If f=0 then f+0.00000001;
cards;
1 1 1 1
1 1 2 3
1 1 3 11
1 1 4  2
1 2 1  2
1 2 2  3
1 2 3  17
1 2 4   3
1 3 1   0
1 3 2   1
1 3 3   8
1 3 4   5
1 4 1   0
1 4 2   2
1 4 3   4
1 4 4   2
2 1 1   1
```

```
2 1 2   1
2 1 3   2
2 1 4   1
2 2 1   0
2 2 2   3
2 2 3   5
2 2 4   1
2 3 1   0
2 3 2   0
2 3 3   7
2 3 4   3
2 4 1   0
2 4 2   1
2 4 3   9
2 4 4   6
;
proc catmod;
weight f;
model g*i*s= _response_/pred=freq ml;
* The "ml" statement  can be replaced by the WLS option;
* The addcell option can also be inserted at this stage;
loglin g i s  g*i g*s;
quit;
```