TESTING FOR HOMOGENEITY OF PROPORTIONS USING NEW McDONALD GENERALIZED BETA-BINOMIAL DISTRIBUTION

BICHANGA LAWRENCE AREBA

A Thesis Submitted to the Graduate School in Partial Fulfilment of the Requirements for the Award of a Master of Science Degree in Statistics of Egerton University

EGERTON UNIVERSITY

MAY, 2016

DECLARATION AND RECOMMENDATION

DECLARATION

This thesis is my work and has not been submitted or presented in part or whole for examination in any institution.

Signature: _____ Date:_____

Lawrence Areba Bichanga
SM12/3347/12

RECOMMENDATION

This research thesis has been submitted for examination with our approval as university supervisors

Signature:_____ Date:_____

Prof. Ali Salim Islam
Mathematics department
Egerton University

Signature: _____ Date:_____

Dr. Orawo Luke Akong'o
Mathematics department
Egerton University

COPY RIGHT

DEDICATION

To

My dad Robert Bichanga, Mum Alice Kerubo, Wife Ann Moraa and Son Ivan Bichanga.

ACKNOWLEDGEMENT

First I would like to thank Almighty God for graciously letting me persue my ambition when all was bleak and for granting me good health, sane mind, strength and sufficient grace to see me through this study. Thank you and Glory be to you Lord forever and ever! Secondly, wish to express my deep appreciation to my supervisors proffesor Ali Salim Islam and Doctor Luke Orawo for their guidance and encouragement during the preparation of this thesis. I would like to thank the Department of Mathematics for giving me financial support in terms of Teaching Assistanship and the Government of Kenya for providing me with research funds during the duration of my graduate studies. Without this financial support it would not have been possible to complete my studies through the National Council of Science, Technology and Inovation (NACOSTI). To my lecturers in the department, thank you all for your support, humble and friendly environment you provided. To my collegues Albert, Timothy, Japar, Charles, Jenniffer and Veronicah; we've been through challenges and we've learnt a lot and maintained the friendship and have encouraged one another to move on even when all seemed difficult to bear. Thank you for your support. I must thank all my wonderful friends in Egerton University Mathematics department for they were generous with their time and recollection. I would like to thank my parents Alice and Robert for their guidance and support in my education throughout my life . My brothers and sisters, David, Shem, Rhoda, Abel and Pauline, thank you for the financial and moral support you accorded me. I would like to thank my wife Ann , son Ivan and all the family members thank you for the moral support, understanding, prayers and the encouragement you gave to me during the whole study period. To my best friend Mr. Dennis Oyunge and Uncle Joseph Ondiba; thank you for their financial support and affection they gave me. To the Ngata Bridge S.D.A church members and friends; thanks for every contribution you made .

Thank you and God Bless you all!

ABSTRACT

Testing for homogeneity of proportions in handling over-dispersion is employed in toxicology, teratology, consumers purchasing behavior, alcohol drinking behavior, in studies of dental caries in children and other similar fields. An important inference problem of interest is to compare proportions of certain characteristic in several groups. However, these proportions often exhibit variation greater than predicted by a simple binomial model. In real world applications, the binomial outcome data are widely encountered and the binomial distribution often fails to test homogeneity of proportions due to over-dispersion. The binomial proportion is assigned a continuous distribution defined on the standard unit interval as one way of handling over-dispersion in the test for homogeneity of proportions. The new McDonald Generalized Beta-Binomial distribution (McGBB) with three shape parameters has been shown to give better fit to binomial outcome data than the Kumaraswamy-Binomial (KB) distribution and Beta-Binomial (BB) distribution based on both simulated data and real data sets and hence considered in this work. This thesis considered derivation of the $C(\alpha)$ tests based on Quasi-likelihood (QL) and Extended Quasi-likelihood (EQL) estimating functions using the new McGBB distribution which have not been done in testing homogeneity of the proportions. Simulation was done by using R package and also real data was used to calculate p-values for both $C(\alpha)$ tests and LR test. The size and power of a test was compared for the simulated data and showed that $C(\alpha)$ tests maintained nominal level well and had higher power than LR test. The comparison of p-values for real data showed that $C(\alpha)$ tests had smaller p-values than LR test hence $C(\alpha)$ tests were preffered since they require estimates only under the null hypothesis. Thus, this thesis has provided a better tests ($C(\alpha)$ tests) based on Quasi-likelihood and Extended Quasi-likelihood estimating functions for testing homogeneity of proportions in presence of overdispersion using the new McGBB distribution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

BB	Beta-Binomial
EQL	Extended Quasi-likelihood
KB	Kumaraswamy-Binomial
LNB	Logit Normal Binomial distribution
LR	Likelihood Ratio
McGBB	McDonald Generalized Beta-Binomial
PMF	Probability Mass Function
PNB	Probit Normal Binomial distribution
QL	Quasi-likelihood

LIST OF SYMBOLS

$B(a,b)$	Beta function
$G\ B1$	McDonald Generalized Beta-Binomial distribution of the first kind
π	Proportions
ρ	Over-dispersion parameter
Θ	Parameter space
$\psi\ (.)$	Digamma function
$\psi'(.)$	Trigamma function

CHAPTER ONE

INTRODUCTION

1.1 Background information

A number of parametric (BB model, the correlated binomial model, the new McGBB model, the additive and multiplicative binomial models) and non-parametric (Quasi-likelihood, extended quasi-likelihood, pseudo-likelihood and those based on optimal quadratic equations) procedures are available for testing homogeneity of the proportions. Of these, the LR test based on the BB model has found prominence in literature. In a recent study conducted (Chandrabose, M., Pushpa, W., and Roshan D) it was evident that the new McGBB distribution is superior to the BB distribution in handling over-dispersion.

This study developed $C(\alpha)$ tests for testing homogeneity of the proportions in presence of over-dispersion. $C(\alpha)$ test is based on the residual of a regression of the score function for the parameter(s) of interest on the nuisance parameters and it has been shown to be asymptotically equivalent to the likelihood ratio test and to test using maximum likelihood estimates (Moran, 1970; Cox and Hinkley, 1974). $C(\alpha)$ test has been widely used as a test statistic (Neyman and Scott, 1966; Moran, 1973; Paul, 1982; Tarone, 1985; Barnwal and Paul, 1988; Boos, 1992; Paul and Islam, 1992; Islam,1994). The advantages of $C(\alpha)$ test are; it require estimates only under the null hypothesis, it often produces a statistic which is simple to calculate, it has been found useful for detecting over-dispersion in binomial and poisson data (Paul *et al.,* 1989; Dean and Lawless, 1990). It also often maintains at least approximately, a pre-assigned level of significance (Bartoo and Puri, 1967). It is locally asymptotically most powerful test (Bühler and Puri, 1966; Moran, 1970).

Two versions of $C(\alpha)$ tests have been developed for testing the significance of added variables in over-dispersed poisson regression and quasi-likelihood model. One version was calculated from the usual model based on covariance matrix and the other version was based on the empirical covariance matrix that has asymptotic justification. $C(\alpha)$ tests developed which were applicable to more general semi-parametric models were robust to misspecification of mean

and variance relation. When the sample size was sufficiently large, the empirical one performed well and in small samples, the performance was not good as the model-based test.

Wald test and likelihood ratio test have potential drawbacks in that wald test required estimates of the parameters only under the alternative hypothesis and LR test required estimates of the parameters under both null and alternative hypothesis. Thus, this study derived $C(\alpha)$ tests, and computed size and power of $C(\alpha)$ test statistics and LR test using the simulated data and p-value using real data. Finally, the performance of $C(\alpha)$ tests and LR test in terms of p-value, size and power of the tests was compared in testing homogeneity of proportions.

1.2 Statement of the problem

$C(\alpha)$ test is based on the residual of a regression of the score function for the parameter(s) of interest on the nuisance parameters. $C(\alpha)$ tests are preferred to LR test because $C(\alpha)$ test has been found to be useful for detecting over-dispersion in binomial and poisson data. $C(\alpha)$ test derived based on the extended Beta-Binomial model hold nominal level well, but do not produce simple forms and also may not be robust when data is from a different distribution. $C(\alpha)$ tests derived based on a quasi-likelihood and Extended quasi-likelihood, required estimates only under the null hypothesis and it often maintains, at least approximately, a pre-assigned level of significance. It is locally asymptotically most powerful and it often produces a statistic which is simple to calculate. Therefore, due to this merits the aim of this study was to derive $C(\alpha)$ tests for testing homogeneity of proportions based on the QL and EQL using the McGBB distribution which had not been done and are consistent interms of size and power.

1.3 Objectives

1.3.1 General objective

To test homogeneity of the proportions in presence of over-dispersion based on the new Mcdonald Generalized Beta-Binomial Distribution.

1.3.2 Specific objectives

(i) To derive $C(\alpha)$ tests for testing homogeneity of proportions in presence of over dispersion based on the quasi-likelihood and Extended quasi-likelihood using the new McGBB distribution.

(ii) To compare the performance in terms of p-values of the $C(\alpha)$ tests and LR test for testing the homogeneity of proportions in presence of over-dispersion using real data based on the new McGBB distribution.

(iii) To compare the performance in terms of size and power of the LR test and $C(\alpha)$ tests for testing the homogeneity of proportions in presence of over-dispersion through simulation using the new McGBB distribution.

1.4 Assumptions

(i) The first two moments of the binomial response with some unknown dispersion of the semi-parametric procedure is assumed in deriving $C(\alpha)$ test.

(ii) This study assumed that the data is binomial over-dispersed.

(iii) The study assumed that the proportions are homogeneous.

1.5 Justification

There has been considerable interest in the derivation of a test statistic that is effective for testing the homogeneity of the proportions in presence of over dispersion. In the studies conducted previously, researchers have found that $C(\alpha)$ tests are more efficient than likelihood ratio test and wald test in the testing the homogeneity of the proportions. $C(\alpha)$ tests derived based on a quasi-likelihood and Extended quasi-likelihood, requires estimates only under the null hypothesis and it often maintains, at least approximately, a pre-assigned level of significance. It is locally asymptotically most powerful test and it often produces a statistic which is simple to calculate. The potential drawbacks of the LR and Wald tests is that the LR test requires estimates of the parameters under both the null and alternative hypothesis and the Wald test requires estimates of the parameters only under the alternative hypothesis. The new McGBB distribution has proved to perform better than Beta-Binomial distribution in modeling over-dispersed data. Thus, this study provided the researchers with a better tests based on the quasi-likelihood and Extended quasi-

likelihood using the new McGBB distribution for testing homogeneity of the proportions in handling over-dispersion which has not been done.

CHAPTER TWO

LITERATURE REVIEW

2.2 Overview of testing homogeneity of proportions

Comparison of homogeneity of proportions of a certain characteristic in several groups is an important problem that arise in toxicology, teratology, consumers purchasing behavior, alcohol drinking behavior, in studies of dental caries in children and other similar fields. Data can be described as follows. Suppose that there are T treatment groups and that the i-th group has k_i litters. The proportion responding in the j-th litter of i-th group is $\frac{y_{ij}}{n_{ij}}, j = 1,\ldots,k_i; i = 1,...,T$. where n_{ij} is the number of trials and y_{ij} is the number of successes in n_{ij} binary trials. A data set with T groups can be represented as in Table 1:

Table 1: Data set with T groups

Groups	Proportions
1	$\frac{y_{11}}{n_{11}}$ $\frac{y_{12}}{n_{12}}$ . . . $\frac{y_{1j}}{n_{1j}}$ . . . $\frac{y_{1k_1}}{n_{1k_1}}$
2	$\frac{y_{21}}{n_{21}}$ $\frac{y_{22}}{n_{22}}$ . . . $\frac{y_{2j}}{n_{2j}}$ . . . $\frac{y_{2k_2}}{n_{2k_2}}$
.	. . . .
.	. . . .
.	. . . .
i	$\frac{y_{i1}}{n_{i1}}$ $\frac{y_{i2}}{n_{i2}}$ . . . $\frac{y_{ij}}{n_{ij}}$ . . . $\frac{y_{ik_i}}{n_{ik_i}}$
.	. . . .
.	. . . .
.	. . . .
T	$\frac{y_{T1}}{n_{T1}}$ $\frac{y_{T2}}{n_{T2}}$ . . . $\frac{y_{Tj}}{n_{Tj}}$ . . . $\frac{y_{Tk_T}}{n_{Tk_T}}$

These proportions often exhibit extra variation that cannot be explained by a simple binomial distribution. In the analysis of these proportions, interest is generally in the estimation of the mean or the regression parameters. Therefore, a special case is to compare the proportion in a control group with that in a treatment group. The LR test and some non-parametric procedures have been developed for testing homogeneity of the proportions.

$C(\alpha)$ test have been widely used (Neyman and Scott, 1966; Moran, 1973; Paul, 1982; Tarone, 1985; Barnwal and Paul, 1988; Boos, 1992; Paul and Islam, 199; Islam, 1994) because of its merits compared to LR test and wald test which have potential drawbacks as discussed previously in section (1.1).

Accordind to Neyman and Scott(1966), $C(\alpha)$ test was used for testing for an unusual distribution of Rare Variants**.** $C(\alpha)$ test statistic as a novel approach for testing for the presence of mixture of effects across a set of rare variants. Unlike existing LR and Wald test, $C(\alpha)$, by testing the variance rather than the mean, maintains consistent power when the target set contains both risk and protective variants. Through simulations and analysis of case/control data, Neyman and Scott(1966) demonstrated that $C(\alpha)$ test maintains good power relative to existing LR and Wald test methods that assess the burden of rare variants in individuals.

Bartoo and Puri (1967) developed a $C(\alpha)$ test and showed that, it maintains at least approximately, a pre-assigned level of significance and it is locally asymptotically most powerful test for composite hypothesis under the null hypothesis for the case where observable random variable $\left(Y_{nk},\ k=1,...,n\right)$ are independently but not necessary identically distributed.

Paul and Islam (1992) discussed on LR test and $C(\alpha)$ tests based on QL and EQL estimating functions using BB distribution in testing homogeneity of proportios in presence of overdispersion. Through simulation and use of real data, the results showed that $C(\alpha)$ test based on quasi-likelihood and Extended quasi-likelihood using BB distribution proved to perform better than LR test in that it holds nominal level quite well and had higher empirical powers (Paul and Islam, 1992). $C(\alpha)$ test also showed that it maintains nominal level most effectively and had higher power than LR test when data came from a different distribution. LR test showed liberal behaviour when data was from a different distributon. i.e LR test could not maintain the

nominal level. Chandbrose *et al*.(2013) proved that the new McGBB distribution is superior to the BB distribution in handling over-dispersion. Hence, the new McGBB distribution was considered in this study.

The new McGBB distribution is a parametric model obtained by mixing McDonald Generalized Beta-Binomial of the first kind and binomial success probability p of binomial distribution (Chandbrose *et al*., 2013). The new McGBB distribution has three shape parameters which makes it flexible in handling over-dispersion than the BB distribution. Thus, this study considered derivation of the $C(\alpha)$ tests based on QL and EQL using the new McGBB distribution which had not been done. This study showed that, the derived $C(\alpha)$ tests performs better than LR test (i.e hold nominal level well and have higher power) for testing homogeneity of the proportions in presence of over-dispersion using McGBB distribution which had not been done.

2.2 $\sqrt{m}$ consistent estimators

Definition: Let $\left\{\hat{\theta}_m\right\}, m = 1, 2, \ldots$ be a sequence of estimators of θ. If the quantity $\left|\hat{\theta}_m - \theta\right|\sqrt{m}$ remains bounded in probability as $m \to \infty$, then the sequence of estimates $\hat{\theta}_m$ is called $\sqrt{m}$ consistent estimators.

By using Chebyshev's inequality, for given $\varepsilon > 0$, $P\left(\left|\hat{\theta}_m - \theta\right|\sqrt{m} \le \varepsilon\right) \ge 1 - \dfrac{\operatorname{var}\left(\hat{\theta}_m\right) m}{\varepsilon^2}$. Then by asymptotic properties of mle and $\operatorname{var}\left(\hat{\theta}_m\right)$ tends to zero as $m \to \infty$, i.e. $\operatorname{var}\left(\hat{\theta}_m\right) = o\left(\dfrac{1}{m}\right)$. Thus, mle is $\sqrt{m}$ - consistent.

2.3 Beta distribution

Let P be a random variable following a Beta distribution with two shape parameters a and b denoted by $B(a,b)$. The probability density of P is given by

$$f(p;a,b) = \frac{P^{a-1}(1-P)^{b-1}}{B(a,b)};\ 0 \le P \le 1 \text{ and } a,b > 0$$

where $B(a,b) = \dfrac{\Gamma a \Gamma b}{\Gamma(a+b)}$ denotes a beta function. (1)

2.4 Beta-Binomial distribution

The BB distribution is obtained from mixing the binomial probability of success P over a Beta distribution defined in (2.3). If $Y/P \sim Bin(n, p)$ and $P \sim Beta(a,b)$, then the PMF of BB distribution is given by

$$P(y) = \binom{n}{y} \frac{B(y+a, n-y+b)}{B(a,b)};\ y = 0,1,\ldots,n \text{ and } a,b > 0. \tag{2}$$

2.5 McDonald Generalized Beta-Binomial Distribution of the first kind

Let P be a random variable following the McDonald's Generalized Beta-Binomial distribution of the first kind (GB1) with three shape parameters α, β and γ (McDonald, 1984; McDonald and Xu, 1995). The probability density function of P is then given by

$$f(p;\alpha,\beta,\gamma) = \frac{\gamma}{B(\alpha,\beta)} p^{\alpha\gamma-1}(1-p^{\gamma})^{\beta-1};\ 0 \le p \le 1 \text{ and } \alpha,\beta,\gamma > 0. \tag{3}$$

The s^{th} moment of the McDonald Generalized Beta-Binomial distribution of the first kind is given by

$$E\left(P^s\right) = \frac{B\left(\alpha+\beta, \frac{s}{\gamma}\right)}{B\left(\alpha, \frac{s}{\gamma}\right)}. \tag{4}$$

The McGBB distribution of the first kind reduces to BB distribution when $\gamma = 1$ (Chandrabose *et al.*, 2013)

2.6 The new McDonald Generalized Beta-Binomial distribution

Generally, a Binomial mixture distribution is obtained through an integration approach. Conditional on *p*, suppose *Y* follows a Binomial distribution given by *Bin* (*n*, *P*), which is denoted by *Y*/*p* ~ *Bin* (*n*, *P*). Unconditional probability mass function of the *Y* can be obtained by evaluating the integral

$$P_Y(y) = \int P_{Y/p} f_P(p/\Theta)\, dp. \tag{5}$$

A random variable *Y* is said to have the new McGBB distribution with parameters *n*, α, β and γ if and only if it satisfies the following stochastic representation

Y|*p* ~ *Bin* (*n*, *p*) and *P* ~ *GB*1 (α, β, γ)

where α, β and γ and are positive real numbers. We denote this distribution as $Y \sim McGBB(n, \alpha, \beta, \gamma)$. Some basic properties of *McGBB*(*n*, α, β, γ) are given below.

2.6.1 Properties of the new McDonald Generalized Beta-Binomial distribution

Let *Y* be a discrete random variable that follows the new McGBB distribution then the following basic properties of the new $McGBB(\alpha, \beta, \gamma)$ distribution holds (Chandrabose *et al*., 2013):

1. The probability mass function of the new $McGBB(\alpha, \beta, \gamma)$ distribution is given by,

$$P(y;\alpha,\beta,\gamma) = \binom{n}{y} \frac{\gamma}{B(\alpha,\beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B(y+\alpha\gamma+\gamma i, n-y+1) \qquad (6)$$

   where $y = 0,1,\ldots,n$ and $\alpha, \beta, \gamma > 0$.

2. A rearranged probability mass function of the new $McGBB(\alpha, \beta, \gamma)$ distribution is given by,

$$P(y;\alpha,\beta,\gamma) = \binom{n}{y} \frac{\gamma}{B(\alpha,\beta)} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma}, \beta\right) \qquad (7)$$

   where $y = 0,\ldots,n$ and $\alpha, \beta, \gamma > 0$.

3. The S^{th} moment of the new $McGBB(\alpha, \beta, \gamma)$ distribution is given by,

$$E\left(Y^s\right) = n \frac{B\left(\alpha+\beta, \frac{s}{\gamma}\right)}{B\left(\alpha, \frac{s}{\gamma}\right)}.$$

Then the mean and variance of the new $McGBB(\alpha, \beta, \gamma)$ distribution are given by,

$E(Y) = n\pi$ and $\operatorname{var}(Y) = n\pi(1-\pi)\{1+(n-1)\rho\}$, respectively where

$$\pi = \frac{B\left(\alpha+\beta,\frac{1}{\gamma}\right)}{B\left(\alpha,\frac{1}{\gamma}\right)} \text{ and } \rho = \frac{\left(\frac{B\left(\alpha+\beta,\frac{2}{\gamma}\right)}{B\left(\alpha,\frac{2}{\gamma}\right)}\right) - \left(\frac{B\left(\alpha+\beta,\frac{1}{\gamma}\right)}{B\left(\alpha,\frac{1}{\gamma}\right)}\right)^2}{\left(\frac{B\left(\alpha+\beta,\frac{1}{\gamma}\right)}{B\left(\alpha,\frac{1}{\gamma}\right)}\right) - \left(\frac{B\left(\alpha+\beta,\frac{1}{\gamma}\right)}{B\left(\alpha,\frac{1}{\gamma}\right)}\right)^2} \tag{8}$$

where ρ is the overdispersion parameter of the new McGBB distribution.

The new McGBB distribution has been shown to be more flexible than BB distribution (Alexander *et al*., 2012).

2.7 Likelihood Ratio Test

Suppose $Y = (Y_1,...,Y_m)$ is a random sample of size m taken from a disribution with pdf $f(y;\lambda)$, where $\lambda = (\theta,\phi)' = (\theta_1,...,\theta_k,\phi_1,...,\phi_s)$ is a $k+s$ component vector. Then the likelihood can be given as $L(Y_1,...,Y_m,\lambda)$. It is of interest to test the null hypothesis $H_0 : \theta = \theta_0 = (\theta_{10},...,\theta_{k0})'$ treating $\phi = (\phi_1,...,\phi_s)'$ as a nuisance parameter. The likelihood ratio for testing H_0 is defined as

$$\Delta = \frac{L\left(Y_1,...,Y_m,\theta_0,\hat{\phi}\right)}{L\left(Y_1,...,Y_m,\hat{\theta},\hat{\phi}\right)}$$

Then the log likelihood ratio test is given by $LR = -2\ell n\Delta = 2(l_1 - l_0)$ where l_0 is the maximum log-likelihood function under H_0 and l_1 is the maximum log-likelihood function under alternative hypothesis. Under the null hypothesis H_0, for large m, the statistic LR is distributed approximately as a chi-square with k degrees of freedom.

2.8 $C(\alpha)$ Test

Suppose $X = (X_1, X_2, ..., X_n)'$ is a random sample of size n taken from a distribution with pdf $f(X;\theta)$. It is of interest to test $H_0 : \theta = \theta_0$ verses $H_1 : \theta = \theta_1$. Let l be the log-likelihood of the data. The partial derivatives evaluated at $\theta = \theta_0 = (\theta_{10}, ..., \theta_{k0})'$ are

$$\psi = \left.\frac{\partial l}{\partial \theta}\right|_{\theta=\theta_0} = \left.\left[\frac{\partial l}{\partial \theta_1}, ..., \frac{\partial l}{\partial \theta_k}\right]'\right|_{\theta=\theta_0} \text{ and}$$

$$\gamma = \left.\frac{\partial l}{\partial \phi}\right|_{\theta=\theta_0} = \left.\left[\frac{\partial l}{\partial \phi_1}, ..., \frac{\partial l}{\partial \phi_s}\right]'\right|_{\theta=\theta_0}$$

Cramer (1946) has shown that under the null hypothesis and mild regularity conditions, $\left(\frac{\partial l}{\partial \theta}, \frac{\partial l}{\partial \phi}\right)$ follows a multivariate normal distribution with mean vector 0 and variance covariance matrix I^{-1}, where $I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$ is the Expected Fisher information with elements

$I_{11} = E\left(\left.\frac{-\partial^2 l}{\partial\theta\partial\theta'}\right|_{\theta=\theta_0}\right)$, $I_{12} = E\left(\left.\frac{-\partial^2 l}{\partial\theta\partial\phi}\right|_{\theta=\theta_0}\right)$ and $I_{22} = E\left(\left.\frac{-\partial^2 l}{\partial\phi\partial\phi'}\right|_{\theta=\theta_0}\right)$.

Define $S = \frac{\partial l}{\partial \theta} - B\frac{\partial l}{\partial \phi}$, where B is the partial regression coefficient matrix obtained by regressing $\frac{\partial l}{\partial \theta}$ on $\frac{\partial l}{\partial \phi}$. From Bartlett (1953), $B = I_{12}I_{22}^{-1}$ and the variance-covariance matrix of S is $I_{11.2}$ where $I_{11.2} = I_{11} - I_{12}I_{22}^{-1}I_{21}$. Thus S is multivariate normal with mean vector 0 and variance-covariance matrix $I_{11.2}$, i.e. $S \sim MN(0, I_{11.2})$. Hence following Neyman (1959) $S'I_{11.2}^{-1}S \sim \chi^2_{(k)}$. Moran (1970) suggested that when θ and unknown nuisance parameter ϕ is replaced by $\sqrt{m}$ consistent estimator obtained from the data. Following the Neyman (1959)

procedure, $\chi^2_{C(\alpha)} = \hat{S}' \hat{I}^{-1}_{11.2} \hat{S}$ which is asymptotically distributed with k degrees of freedom. If the nuisance parameter ϕ is replaced by its maximum likelihood estimate (mle) $\hat{\phi}$, then the score function S reduces to ψ. The $C(\alpha)$ statistic reduces to $\hat{\psi}' \hat{I}^{-1}_{11.2} \psi$ (Rao, 1947). The null hypothesis H_0 is rejected if the computed value of $C(\alpha) \geq \chi^2_{\alpha,r}$. where r is the degrees of freedom.

CHAPTER THREE

METHODS

3.1 Data collection

This study is conducted on the real data (Alanko and Lemmens, 1996). The data is based on the numbers of alcohol consumption days in two reference weeks which are separately self-reported by a randomly selected sample of 399 respondents in the Netherlands in 1983. The number of days an individual consumes alcohol Y, out of $n = 7$ days in a reference week can be treated as a binomial variable. The probability P, to consume alcohol on a randomly chosen day in a reference week for an individual cannot be treated as a constant in this setup since there is a person-to-person variation in the drinking behavior and to drink. This leads to analyzing this data using a Binomial mixture distribution by testing homogeneity of proportions for the random variable *P* using a continuous distribution bounded in the standard unit interval. This data have also been previously used by (Rodríguez-Avi *et al*., 2007; Li *et al.,* 2011; Chandrabose *et al*., 2013).

3.2 Simulation

Simulation study is conducted to investigate the performance of LR test and $C(\alpha)$ test statistics in terms of size and power for testing homogeneity of the proportions in handling over-dispersion. The simulated data was generated based on the new McGBB distribution. The developed algorithm was used to generate overdispersed Binomial variables (Ahn and Chen, 1995). In the simulation study, empirical levels were calculated based on 1000 replications for each combination of varying values of $\alpha_1 = \alpha_2 = 0.10$, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45,

0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95 and 1.00 and for values $\beta_1 = \beta_2 = \beta = 0.30$ and $\gamma_1 = \gamma_2 = \gamma = 1$ parameters were chosen. For power, varying values of $\alpha_2 =$ 0.22, 0.24, 0.26, 0.28, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.70, 0.71, 0.72, 0.73, 0.74 and 0.75 was considered. For each value of α_2, empirical powers were calculated for $\beta_1 = \beta_2 = \beta = 0.70$ and $\gamma_1 = \gamma_2 = \gamma = 1$. The open source statistical software R (version 3.1.1) was used in this study to simulate data.

3.3 Data analysis

The size and power of both LR test and $C(\alpha)$ tests were obtained using simulated data and p-values using real data. Since this study intended to determine better based on QL and EQL using the new McGBB distribution, the performance of LR test and $C(\alpha)$ tests was analysed using both real data and the simulated data. Analysis for both the real data and simulated data was done using a R statistical package. Analysis was such that, programs for computation of both size and power for LR test and $C(\alpha)$ tests were developed. Comparison was done based on the p-values, size and power computed for both the LR test and $C(\alpha)$ tests. Based on the results, recommendations was given on the better test that was appropriate for testing homogeneity of the proportions in presence of over dispersion parameter.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter displays the derivation and findings for LR test and $C(\alpha)$ tests interms of p-value for real data and size and power for the simulated data. A detailed discussion of the results is given based on observation on the tables displayed. The results are given interms of p-value for real data and size and power for the simulated data. Tables 2, 3 and 4 display results for p-value, size and power respectively. Results for p-values were obtained using the Rodríguez-Avi *et al*. (2007) real data set and are displayed in table 2. The displayed figure 1 and 2 display the size and power respectively for both LR and $C(\alpha)$ tests for the simulated data. The procedures for the displayed results can be replicated by running the subroutines displayed in the appendices.

4.2 Derivation of $C(\alpha)$ tests for testing the homogeneity of the proportions

4.2.1 The $C(\alpha)$ test statistic based on the quasi-likelihood(C_{QL})

The Quasi-likelihood (Wedderburn, 1974) is based on the knowledge of the first two moments of the random variable $z = \frac{y}{n}$ where y is the number of successes in *n* binary trials, n is the number of trials and ϕ is the over-dispersion parameter.

$$E(z) = \pi,\ \text{var}(z) = \frac{\pi(1-\pi)}{n}\{1+(n-1)\phi\}, 0 \le \pi \le 1 \ \text{ and } \left(\frac{-1}{n-1}\right) < \phi < 1.$$

This specification of mean and variance coincides with those based on the new McGBB model. The Quasi-likelihood for an observation z with the above mean and variance is given by $Q(z,\pi,\phi) = \int_z^{\pi} \frac{(z-t)n}{t(1-t)\{1+(n-1)\phi\}}dt,$ integration by partial fraction becomes

$$Q(z,\pi,\phi)=\sum_{i=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[y\log\left(\frac{\pi}{z}\right)+(n-y)\log\left\{\frac{(1-\pi)}{(1-z)}\right\}\right] \tag{9}$$

where $\pi=\dfrac{B\left(\alpha+\beta,\dfrac{1}{\gamma}\right)}{B\left(\alpha,\dfrac{1}{\gamma}\right)}$ and $\hat{\phi}=\dfrac{\left(\dfrac{\mathrm{var}(y)}{\bar{y}\left(1-\dfrac{\bar{y}}{n}\right)}-1\right)}{(n-1)}$

Define $\lambda=(\lambda_1,\lambda_2,\lambda_3)=(\alpha,\beta,\gamma)$. Then let $\Psi_i=\dfrac{\partial Q}{\partial\alpha}$, $i=1,...,T-1$,

and $\varphi_k=\dfrac{\partial Q}{\partial\lambda_k}$, $k=1,2,3$.

To make things simple we assume homogeneity of proportions and under this assumption we wish to test the hypothesis $H_0:\pi_1=...=\pi_T$ against $H_1:\ \pi_1\neq...\neq\pi_T$. Now, let $\hat{\lambda}$ be some $\sqrt{m}$ consistent estimator of λ under the null hypothesis. Then the $C(\alpha)$ test is based on $S_i(\hat{\lambda})=\Psi_i(\hat{\lambda})-\beta_{1i}\varphi_1(\hat{\lambda})-\beta_{2i}\varphi_2(\hat{\lambda})-\beta_{3i}\varphi_3(\hat{\lambda})$, $i=1,...,T-1$ where β_{1i}, β_{2i} and β_{3i} are the partial regression coefficients of Ψ_i and φ_1, Ψ_i and φ_2 and Ψ_i and φ_3 respectively. The variance-covariance matrix of $S(\hat{\lambda})=\left\{S_1(\hat{\lambda}),...,S_{T-1}(\hat{\lambda})\right\}'$ is $D-AB^{-1}A'$ and the regression coefficients $\beta=(\beta_1,\beta_2,\beta_3)=AB^{-1}$ where $\beta_1=(\beta_{11},...,\beta_{1T-1})$, $\beta_2=(\beta_{21},...,\beta_{2T-1})$, $\beta_3=(\beta_{31},...,\beta_{3T-1})$.

$$D_{it}=E\left(\frac{-\partial Q}{\partial\lambda_i\partial\lambda_t}\right),\ i,t=1,...,T-1.$$

$$A_{ik}=E\left(\frac{\partial Q}{\partial\lambda_i\partial\lambda_k}\right)\qquad\begin{matrix}i=1,...,T-1,\\ k=1,2,3.\end{matrix}$$

$$B_{ks}=E\left(\frac{-\partial Q}{\partial\lambda_k\partial\lambda_s}\right),\quad k,s=1,2,3.$$

Using $\hat{\lambda}$ in S, A, B and D, the $C(\alpha)$ test is given by $S'\left(D - AB^{-1}A'\right)^{-1} S$, which is approximately distributed as chi-square with $T-1$ degrees of freedom.

Using the Quasi log-likelihood (9) and taking partial derivatives, we obtain

$$\frac{\partial Q}{\partial \alpha} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(y - \frac{\pi(n-y)}{1-\pi} \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha + \frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{10}$$

$$\frac{\partial Q}{\partial \beta} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(y - \frac{\pi(n-y)}{1-\pi} \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{11}$$

$$\frac{\partial Q}{\partial \gamma} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{y}{\gamma^2} - \frac{\pi(n-y)}{\gamma^2(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \right] \tag{12}$$

Maximum quasi-likelihood estimates of $\hat{\lambda} = \left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right)$. are obtained by equating (10), (11) and (12) to zero and solving simultaneously. Denote the estimates by $\hat{\lambda}_{QL}$. The second derivatives of Q are given below

$$\frac{\partial^2 Q}{\partial \alpha^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(y - \frac{\pi(n-y)}{1-\pi} \right) \left(\psi'(\alpha+\beta) + \psi'\left(\alpha+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) - \left(\frac{(n-y)\pi}{(1-\pi)^2} \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{13}$$

$$\frac{\partial^2 Q}{\partial \beta^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(y - \frac{\pi(n-y)}{1-\pi} \right) \left(\psi'(\alpha+\beta) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) - \left(\frac{(n-y)\pi}{(1-\pi)^2} \right) \times \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{14}$$

$$\frac{\partial^2 Q}{\partial \gamma^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{y}{\gamma^4} - \frac{\pi(n-y)}{\gamma^4(1-\pi)} \right) \left(\psi'\left(\alpha+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) - \left(\frac{(n-y)\pi}{\gamma^4(1-\pi)^2} \right) \times \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 - \left(\frac{2ny}{\gamma^3} - \frac{2\pi(n-y)}{\gamma^3(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{15}$$

$$\frac{\partial^2 Q}{\partial\alpha\partial\beta}=\sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(y-\frac{\pi(n-y)}{1-\pi}\right)\left(\psi'(\alpha+\beta)-\psi'\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)-\left(\frac{\pi(n-y)}{(1-\pi)^2}\right)\times\right.$$
$$\left.\left(\psi(\alpha+\beta)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\left(\psi(\alpha+\beta)+\psi\left(\alpha+\frac{1}{\gamma}\right)-\psi(\alpha)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] \quad (16)$$

$$\frac{\partial^2 Q}{\partial\alpha\partial\gamma}=\sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(\frac{y}{\gamma^2}-\frac{\pi(n-y)}{\gamma^2(1-\pi)}\right)\left(\psi'\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)-\left(\frac{\pi(n-y)}{\gamma^2(1-\pi)^2}\right)\times\right.$$
$$\left.\left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)-\psi\left(\alpha+\frac{1}{\gamma}\right)\right)\left(\psi(\alpha+\beta)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] \quad (17)$$

$$\frac{\partial^2 Q}{\partial\beta\partial\gamma}=\sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(\frac{y}{\gamma^2}-\frac{\pi(n-y)}{\gamma^2(1-\pi)}\right)\left(\psi'\left(\alpha+\beta+\frac{1}{\gamma}\right)-\psi'\left(\alpha+\frac{1}{\gamma}\right)\right)-\left(\frac{\pi(n-y)}{\gamma^2(1-\pi)^2}\right)\times\right.$$
$$\left.\left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)-\psi\left(\alpha+\frac{1}{\gamma}\right)\right)\left(\psi(\alpha+\beta)+\psi\left(\alpha+\frac{1}{\gamma}\right)-\psi(\alpha)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] \quad (18)$$

where $\psi(.)$ and $\psi'(.)$ are digamma and trigamma functions respectively.

Expectations of the minus the second derivatives are given below,

$$D_{11}=B_{11}=A_{11}=\sum_{y=0}^{n}\frac{n}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi}{1-\pi}\right)\left(\psi(\alpha+\beta)+\psi\left(\alpha+\frac{1}{\gamma}\right)-\psi(\alpha)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)^2\right] \quad (19)$$

$$B_{22}=\sum_{y=0}^{n}\frac{n}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi}{1-\pi}\right)\left(\psi(\alpha+\beta)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)^2\right] \quad (20)$$

$$B_{33}=\sum_{y=0}^{n}\frac{n}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi}{\gamma^4(1-\pi)}\right)\left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)-\psi\left(\alpha+\frac{1}{\gamma}\right)\right)^2\right] \quad (21)$$

$$A_{12}=B_{12}=\sum_{y=0}^{n}\frac{n}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi}{1-\pi}\right)\left(\psi(\alpha+\beta)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\times\right.$$
$$\left.\left(\psi(\alpha+\beta)+\psi\left(\alpha+\frac{1}{\gamma}\right)-\psi(\alpha)-\psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] \quad (22)$$

$$A_{13} = B_{13} = \sum_{y=0}^{n} \frac{n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{\gamma^2(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \times \right.$$
$$\left. \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{23}$$

$$B_{23} = \sum_{y=0}^{n} \frac{n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{\gamma^2(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{24}$$

Denote the Quasi-likelihood estimates of $\lambda = (\alpha, \beta, \gamma)$ by $\hat{\lambda}_{QL}$. If $\hat{\lambda}_{QL}$ is used in S, A, B and D, which is $\sqrt{m}$ consistent estimates of λ under the null hypothesis, then $S\left(\hat{\lambda}_{QL}\right) = \Psi\left(\hat{\lambda}_{QL}\right)$. Then the quasi-likelihood score or the $C(\alpha)$ test is

$$C_{QL} = \Psi'\left(D - AB^{-1}A'\right)^{-1}\Psi. \tag{25}$$

4.2.2 The $C(\alpha)$ test statistic based on the Extended quasi-likelihood(C_{EQL})

The extended quasi-likelihood(Nelder and Pregibon, 1987) can be used for the simultaneous estimation of the $\lambda = (\alpha, \beta, \gamma)$. The extended quasi-log-likelihood for an observation z with mean and variance specified is

$$Q^{+}(z, \pi, \phi) = -\frac{1}{2}\log(2k) - \frac{1}{2}\log\left[\frac{z(1-z)\{1-(n-1)\phi\}}{n} \right] + \int_{z}^{\pi} \frac{(z-t)n}{t(1-t)\{1+(n-1)\phi\}} dt. \tag{26}$$

The Extended quasi-log-likelihood for the data under consideration, then is

$$Q^{+}(z, \pi, \phi) = C - \frac{1}{2}\sum_{y=0}^{n} \left[\log\{1+(n-1)\phi\} + \frac{2}{\{1+(n-1)\phi\}} \left[y\log\left(\frac{\pi}{z}\right) + (n-y)\log\left(\frac{1-\pi}{1-z}\right) \right] \right] \tag{27}$$

where C is term not involving the parameters. Define $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (\alpha, \beta, \gamma)$. Then let $\Psi_i = \frac{\partial Q^{+}}{\partial \alpha}$, $i = 1, \ldots, T-1$ and $\varphi_k = \frac{\partial Q^{+}}{\partial \lambda_k}$, $k = 1, 2, 3$.

We wish to test the hypothesis $H_0 : \pi_1 = ... = \pi_T$ against $H_1 : \pi_1 \neq ... \neq \pi_T$. Now, let $\hat{\lambda}$ be some $\sqrt{m}$ consistent estimator of λ under the null hypothesis. Then the $C(\alpha)$ test is based on $S_i(\hat{\lambda}) = \Psi_i(\hat{\lambda}) - \beta_{1i}\varphi_1(\hat{\lambda}) - \beta_{2i}\varphi_2(\hat{\lambda}) - \beta_{3i}\varphi_3(\hat{\lambda}),\ i = 1,...,T-1$ where β_{1i}, β_{2i} and β_{3i} are the partial regression coefficients of Ψ_i and φ_1, Ψ_i and φ_2 and Ψ_i and φ_3 respectively. The variance-covariance matrix of $S(\hat{\lambda}) = \left\{S_1(\hat{\lambda}),...,S_{T-1}(\hat{\lambda})\right\}'$ is $D - AB^{-1}A'$ and the regression coefficients $\beta = (\beta_1, \beta_2, \beta_3) = AB^{-1}$ where $\beta_1 = (\beta_{11},...,\beta_{1T-1})$, $\beta_2 = (\beta_{21},...,\beta_{2T-1})$, $\beta_3 = (\beta_{31},...,\beta_{3T-1})$.

$$D_{it} = E\left(\frac{-\partial Q^+}{\partial\lambda_i\partial\lambda_t}\right),\ i,t = 1,...,T-1$$

$$A_{ik} = E\left(\frac{\partial Q^+}{\partial\lambda_i\partial\lambda_k}\right) \qquad \begin{matrix} i = 1,...,T-1 \\ k = 1,2,3 \end{matrix}$$

$$B_{ks} = E\left(\frac{-\partial Q^+}{\partial\lambda_k\partial\lambda_s}\right),\quad k,s = 1,2,3.$$

Using $\hat{\lambda}$ in S, A, B and D, the $C(\alpha)$ test is given by $S'\left(D - AB^{-1}A'\right)^{-1}S$, which is approximately distributed as chi-square with $T-1$ degrees of freedom.

The unbiased estimating equations for $\lambda = (\alpha, \beta, \gamma)$. obtained from Q^+ are

$$\frac{\partial Q^+}{\partial\alpha} = \sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi(n-y)}{1-\pi} - y\right)\left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] = 0 \tag{28}$$

$$\frac{\partial Q^+}{\partial\beta} = \sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi(n-y)}{1-\pi} - y\right)\left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right)\right)\right] = 0 \tag{29}$$

$$\frac{\partial Q^+}{\partial\gamma} = \sum_{y=0}^{n}\frac{1}{\{1+(n-1)\phi\}}\left[\left(\frac{\pi(n-y)}{\gamma^2(1-\pi)} - \frac{y}{\gamma^2}\right)\left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right)\right)\right] = 0 \tag{30}$$

Maximum extended quasi-likelihood estimates of $\lambda = (\alpha, \beta, \gamma)$. are obtained by solving (28), (29) and (30) simultaneously. Denote the estimates by $\hat{\lambda}_{EQL}$ The second derivatives of Q^+ are given below

$$\frac{\partial^2 Q^+}{\partial \alpha^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{1-\pi} - y \right) \left(\psi'(\alpha+\beta) + \psi'\left(\alpha+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) + \left(\frac{(n-y)\pi}{(1-\pi)^2} \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{31}$$

$$\frac{\partial^2 Q^+}{\partial \beta^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{1-\pi} - y \right) \left(\psi'(\alpha+\beta) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) + \left(\frac{(n-y)\pi}{(1-\pi)^2} \right) \times \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{32}$$

$$\frac{\partial^2 Q^+}{\partial \gamma^2} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{\gamma^4(1-\pi)} - \frac{y}{\gamma^4} \right) \left(\psi'\left(\alpha+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) - \left(\frac{(n-y)\pi}{\gamma^4(1-\pi)^2} \right) \times \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 + \left(\frac{2ny}{\gamma^3} - \frac{2\pi(n-y)}{\gamma^3(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{33}$$

$$\frac{\partial^2 Q^+}{\partial \alpha \partial \beta} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{1-\pi} - y \right) \left(\psi'(\alpha+\beta) - \psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) + \left(\frac{\pi(n-y)}{(1-\pi)^2} \right) \times \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{34}$$

$$\frac{\partial^2 Q^+}{\partial \alpha \partial \gamma} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{\gamma^2(1-\pi)} - \frac{y}{\gamma^2} \right) \left(\psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) + \left(\frac{\pi(n-y)}{\gamma^2(1-\pi)^2} \right) \times \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{35}$$

$$\frac{\partial^2 Q^+}{\partial \beta \partial \gamma} = \sum_{y=0}^{n} \frac{1}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi(n-y)}{\gamma^2(1-\pi)} - \frac{y}{\gamma^2} \right) \left(\psi'\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi'\left(\alpha+\frac{1}{\gamma}\right) \right) + \left(\frac{\pi(n-y)}{\gamma^2(1-\pi)^2} \right) \times \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{36}$$

Expectations of the minus the second derivatives are given below,

$$D_{11} = A_{11} = B_{11} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{1-\pi} \right) \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{37}$$

$$B_{22} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{1-\pi} \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right)^2 \right] \tag{38}$$

$$B_{33} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{\gamma^4(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right)^2 \right] \tag{39}$$

$$A_{12} = B_{12} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{1-\pi} \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \times \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{40}$$

$$A_{13} = B_{13} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{\gamma^2(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \times \left(\psi(\alpha+\beta) + \psi\left(\alpha+\frac{1}{\gamma}\right) - \psi(\alpha) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{41}$$

$$B_{23} = \sum_{y=0}^{n} \frac{-n}{\{1+(n-1)\phi\}} \left[\left(\frac{\pi}{\gamma^2(1-\pi)} \right) \left(\psi\left(\alpha+\beta+\frac{1}{\gamma}\right) - \psi\left(\alpha+\frac{1}{\gamma}\right) \right) \left(\psi(\alpha+\beta) - \psi\left(\alpha+\beta+\frac{1}{\gamma}\right) \right) \right] \tag{42}$$

Denote the Extended Quasi-likelihood estimates of $\lambda = (\alpha, \beta, \gamma)$ by $\hat{\lambda}_{EQL}$. If $\hat{\lambda}_{EQL}$ is used in S, A, B and D, which is $\sqrt{m}$ consistent estimates of λ under the null hypothesis, then $S\left(\hat{\lambda}_{EQL}\right) = \Psi\left(\hat{\lambda}_{EQL}\right)$. Then the quasi-likelihood score or the $C(\alpha)$ test is

$$C_{EQL} = \Psi'\left(D - AB^{-1}A'\right)^{-1}\Psi. \tag{43}$$

Note that under the null hypothesis the parameters α, β and γ are common across groups. In this study we consider $T = 2$ groups. So the estimation of α, β and γ from the T groups can be considered to be estimation from a single group consisting of the combined data in

the T groups. The critical region (Rejection region) is obtained if the null hypothesis H_0 is rejected if the computed value of LR, C_{QL} and $C_{EQL} \geq \chi^2_{\alpha,r}$. where r is the degrees of freedom.

4.3 Likelihood Ratio test

We consider that $\frac{y_{ij}}{n_{ij}} \sim McGBB(n, \alpha, \beta, \gamma)$ for $j = 1, \ldots, n$ and $i = 1, \ldots, T$. Our interest is to test

$$H_0 : \pi_1 = \ldots = \pi_T$$

$$H_1 : \pi_1 \neq \ldots \neq \pi_T$$

The new McGBB log-likelihood is given by;

$$l(\Theta) = \sum_{k=0}^{N} \log \binom{n}{y_k} + \sum_{k=0}^{N} \log \left(\sum_{j=0}^{n-y_k} (-1)^j \binom{n-y_k}{j} B\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma}, \beta \right) \right) \tag{44}$$

The likelihood ratio test is defined as

$$LR = 2(l_1 - l_0) \tag{45}$$

where l_0 is the maximum log-likelihood function under the null hypothesis and l_1 is the maximum log-likelihood function under the alternative hypothesis. The LR statistics under the null hypothesis is distributed asymptotically chi-square with $(T-1)$ degrees of freedom.

4.4 Summary

The new McGBB distribution proved to perform better than BB distribution in modeling over-dispersed binomial data and also gave a better fit to model over-dispersed data. Therefore, the aim of this study was to derive better test statistics ($C(\alpha)$ test) for testing homogeneity of proportions based on the QL and EQL using the McGBB distribution which had not been done. Size and power of $C(\alpha)$ test statistics and LR test was computed using the simulated data and p-values using real data set, i.e. alcohol consumption data. Finally, the performance of $C(\alpha)$ tests and LR test in terms of p-values, size and power of the tests in testing homogeneity of

proportions were compared to determine a better test which had not been done using the McGBB distribution.

4.5 Real Data Results

The computation of p-values for LR and $C(\alpha)$ tests of alcohol data is given in table 2 below.

Table 2: Testing homogeneity of proportions results of alcohol consumption data

Number of Drinking Days	0	1	2	3	4	5	6	7	Total
Observed frequency(Week 1)	47	54	43	40	40	41	39	95	399
Observed frequency(Week 2)	42	47	54	40	49	40	43	84	399
LR test		1.59581							
DF		3							
P-value		0.33966							
C_{QL} test		0.25335							
DF		3							
P-value		0.03145							
C_{EQL} test		0.25435							
DF		3							
P-value		0.03163							

The data set in Table 2 was used by Alanko and Lemmens (1996), Rodriguez-Avil *et al* (2007) and Chandrabose *et al*.(2013) in the study of handling over-dispersion. It shows the number of days an individual consumes alcohol in two reference weeks which are separately self-reported by a randomly selected sample of 399 respondents in the Netherlands in 1983. The number of days an individual consumes alcohol Y, out of $n = 7$ days. For this data sets, the p-values for LR test, C_{QL} test and C_{EQL} test are given in Table 2. As summarized in table 2, the results for LR test value is 1.59581 and $C(\alpha)$ tests values are $C_{QL} = 0.25335$ and $C_{QL} = 0.25435$ of alcohol consumption data. The p-value in the LR test (0.33966) is noticeably larger than those

for both C_{QL} test (0.03145) and C_{EQL} test (0.03163). This indicates that for any good test should have the smallest p-value. Therefore, based on these results, this study concludes that the proposed $C(\alpha)$ tests provide better test to test for homogeneity of proportions in presence of over-dispersion than LR test. The C_{QL} test is the best having the smallest p-value (0.03145).

4.3 Empirical levels (Size)

The computation of Empirical levels results and graph comparison for LR test and $C(\alpha)$ tests for the simulated data sets is represented in table 3 and figure 1. Empirical level is the probability of rejecting the null hypothesis when the null hypothesis is true.

Table 3: Empirical levels ; $\alpha = 0.05$; based on 1000 simulated data sets for $\beta_1 = \beta_2 = \beta = 0.3$, $\gamma_1 = \gamma_2 = \gamma = 1$ and $\alpha_1 = \alpha_2$ varied

Estimated Empirical levels			
$\alpha_1 = \alpha_2$ varied	LR Test	C_{EQL} Test	C_{QL} Test
0.10	0.044	0.046	0.048
0.15	0.046	0.044	0.046
0.20	0.042	0.042	0.046
0.25	0.029	0.048	0.048
0.30	0.038	0.045	0.041
0.35	0.027	0.039	0.044
0.40	0.022	0.048	0.049
0.45	0.023	0.046	0.050
0.50	0.038	0.057	0.052
0.55	0.099	0.043	0.049
0.60	0.129	0.046	0.049
0.65	0.085	0.054	0.045
0.70	0.101	0.038	0.044
0.75	0.076	0.047	0.050
0.80	0.058	0.049	0.049
0.85	0.073	0.052	0.050
0.90	0.058	0.042	0.041
0.95	0.051	0.056	0.049
1.00	0.065	0.054	0.048

Generally, the results given in table 3 show that, for all varying values of $\alpha_1 = \alpha_2$ the $C(\alpha)$ tests C_{QL} and C_{EQL} show conservative behaviour i.e the empirical levels are closer to 0.05 unlike the LR test. For small $\alpha_1 = \alpha_2$ ($\alpha_1 = \alpha_2 = 0.10$, 0.15 and 0.20) the LR and the $C(\alpha)$ tests show some conservative behavior which is near 0.05, otherwise all the test statistic produce

consistent empirical levels close to the nominal level (0.05). At $\alpha_1 = \alpha_2$ ($\alpha_1 = \alpha_2 = 0.55$, 0.60, 0.65, 0.70 and 0.75), the LR test is not consistent and hence it show nonconservative behavior and produce empirical levels(0.099, 0.129, 0.085, 0.101 and 0.076) that are far away to the nominal level, this is because LR test require estimates on both the null and alternative hypothesis. The $C(\alpha)$ tests C_{QL} and C_{EQL} are consistent and produce empirical levels very close to the nominal level while LR test show nonconservative behavior hence the $C(\alpha)$ tests are preferred since they require estimates only under the null hypothesis. The performance of the $C(\alpha)$ test is better than LR test in that it is consistent, holds nominal level quite well and also has a simple form. Hence, a good test should have empirical levels being 0.05 or very close to 0.05.

Figure 1 below represent the graph comparison on empirical level for LR, C_{QL} and C_{EQL} test for $\alpha_1 = \alpha_2$ varied

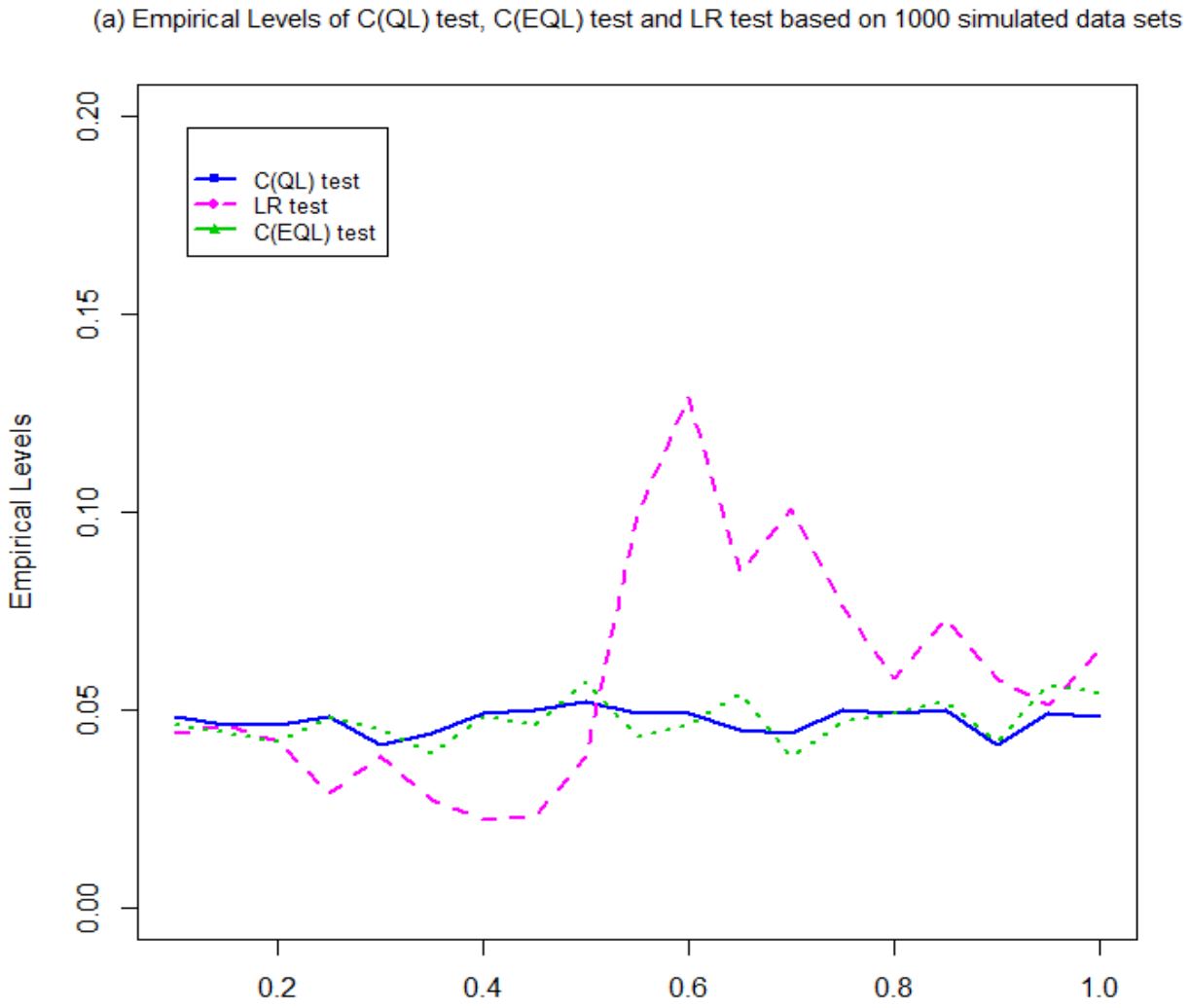

Figure 1: Plot of empirical level comparison for C_{QL} test, C_{EQL} test and LR test under McGBB model for varied $\alpha_1 = \alpha_2$ and for values of $\beta_1 = \beta_2 = \beta = 0.30$ and $\gamma_1 = \gamma_2 = \gamma = 1$ for all procedures.

From figure1, is evident that the C_{QL} and C_{EQL} test are consistent and tries to maintain the nominal level (0.05) well for all varying values of $\alpha_1 = \alpha_2$ and show conservative behavior. In general, LR test fails to maintain the empirical level near 0.05.

4.4 Empirical Power

This subsection represents the computation of Empirical Power results and graph comparison for LR test and $C(\alpha)$ tests for the simulated data sets as given in table 4 and figure 2 respectively. Empirical power is the probability of correctly rejecting the null hypothesis.

Table 4: Empirical Power; $\alpha = 0.05$; based on 1000 simulated data sets for $\alpha_1 = 0.20$, $\beta_1 = \beta_2 = \beta = 0.70$, $\gamma_1 = \gamma_2 = \gamma = 1$ and α_2 varied

	Estimated Empirical Power		
α_2 varied	LR Test	C_{EQL} Test	C_{QL} Test
0.22	0.058	0.156	0.165
0.24	0.071	0.189	0.247
0.26	0.108	0.231	0.340
0.28	0.125	0.285	0.381
0.30	0.147	0.341	0.452
0.35	0.270	0.505	0.554
0.40	0.350	0.581	0.672
0.45	0.522	0.662	0.748
0.50	0.664	0.716	0.803
0.55	0.698	0.775	0.856
0.60	0.857	0.824	0.886
0.61	0.879	0.919	0.895
0.62	0.910	0.932	0.966
0.63	0.925	0.952	0.974
0.64	0.942	0.953	0.975
0.65	0.956	0.964	0.975
0.66	0.968	0.967	0.977
0.67	0.970	0.972	0.979
0.68	0.981	0.977	0.986
0.69	0.982	0.983	0.989
0.70	0.987	0.992	0.992
0.71	0.989	0.992	0.993
0.72	0.990	0.994	0.995
0.73	0.992	0.998	0.996
0.74	0.996	0.998	0.998
0.75	0.998	0.999	0.999

Generally, for the results given in table 4, for α_2 ($\alpha_2 = 0.22$, 0.24, 0.26, 0.28, 0.30, 0.35, 0.40, 0.45 and 0.50), the power of the LR test is to some extent smaller than those of the other two tests C_{QL} and C_{EQL}. $C(\alpha)$ tests show higher power than LR test hence they are consistent and better tests preferable, as they require estimates of the parameters only under the null hypothesis. C_{QL} is the best for all varying values of α_2 with the highest empirical power. The studies carried out on $C(\alpha)$ test for example the study by Paul and Islam (1992) proved the superiority of $C(\alpha)$ test in testing for homogeneity of proportions in presence of over-dispersion than LR test using BB distribution. They showed that, $C(\alpha)$ tests based on QL and EQL estimating function were consistent and had higher power than LR test but C_{QL} test was the best. Thus, this study echo this findings and show that $C(\alpha)$ tests based on QL and EQL estimating functions are superior than LR test since they hold the nominal level well have higher empirical power.

Figure 2 represent the graph comparison on empirical power for LR, C_{QL} and C_{EQL} test for α_2 varied.

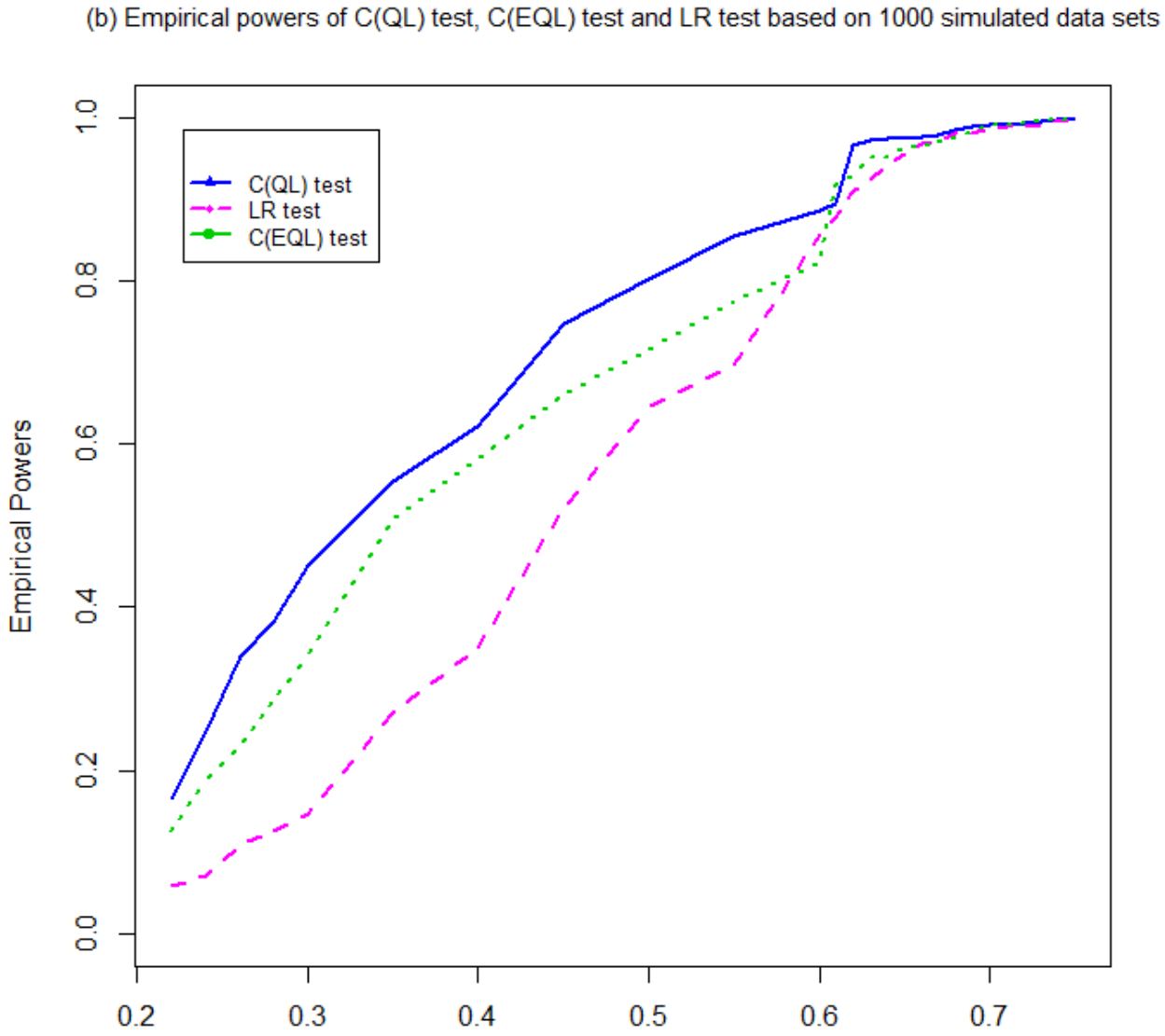

Figure 2: Plot of empirical power comparison for C_{QL} test, C_{EQL} test and LR test under McGBB model for varied α_2 and for values of $\alpha_1 = 0.20$, $\beta_1 = \beta_2 = \beta = 0.70$ and $\gamma_1 = \gamma_2 = \gamma = 1$ for all procedures.

From figure 2, is evident that the C_{QL} and C_{EQL} test are consistent for all varying of α_2 with higher empirical power than LR test hence $C(\alpha)$ tests are preffered. C_{QL} test have highest empirical power and consistent for all varying value of α_2. As the varying value of α_2 increases

for fixed value of $\alpha_1 = 0.20,$ the empirical power increases for all the test. C_{QL} test shows to be consistent and having the highest power for varying values of α_2 hence the best test.

Maximum likelihood estimates (mle's) of the parameters under the null and alternative hypothesis were obtained by maximizing log-likelihood of McGBB distribution (44) using the R package subroutine. The quasi-likelihood and Extended quasi-likelihood estimates of the parameters under the null hypothesis were obtained by maximizing the Quasi log-likelihood (9) and Extended Quasi log-likelihood (27) using the R package subroutine.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Introduction

This section gives the summary of the findings of research and conclusion of the study. The conclusion is given based on each specific objective given in this study. Recommendations for further study and areas of application of the study based on the results are also given.

5.2 Summary and Conclusion

In this work, the $C(\alpha)$ tests based on Quasi-likelihood and Extended Quasi-likelihood estimating functions procedures have been derived for testing homogeneity of proportions. Performance evaluation measures empirical level (size) and the empirical power were obtained for the simulated data. The comparison of the tests based on empirical level and empirical power is as given in table 3 and 4 respectively. Based on the results from the table 3 and 4 and figure 1 and 2, $C(\alpha)$ tests performs better and are consitent than LR test since they holds nominal level quite well and have higher empirical power. The C_{QL} test is the best since it shows the highest conservative behavior and the highest empirical power. As for the real data set i.e. alcohol consumption, it can be seen in Table 2, the p-value in the LR test is noticeably larger than those for both C_{QL} and C_{EQL} tests. Hence $C(\alpha)$ tests are consistent and preferable.

5.3 Recommendation and Further Research

This study has investigated the size and power of LR test and $C(\alpha)$ tests based on quasi-likelihood and Extended quasi-likelihood using the new McGBB distribution. The derived $C(\alpha)$ tests have proved to maintain the nominal level well and have higher power than the LR test. Hence, $C(\alpha)$ tests are recommended for testing homogeneity of proportions in presence of over-dispersion.

Future research may consider robustness study for the size and power of this three test. i.e. LR, C_{QL} and C_{EQL} test when data come from other over/under-dispersed binomial distribution such as Probit Normal Binomial (PNB) distribution and Logit Normal Binomial

(LNB) distribution. Secondly, the performance of $C(\alpha)$ test for testing homogeneity of proportions under equal dispersion parameter was investigated using a simulation study for two groups. A simulation study is suggested to investigate the effect of unequal dispersion on inferences concerning the proportions for two groups. Thirdly, a simulation study is suggested to investigate the effect of equal or unequal dispersion on inferences concerning the proportions for more than two groups.

5.4 Application

This work has much application in family studies, where it can be used to measure the degree of intra-family resemblance with respect to blood group, weight, height and also in the investigation of hereditability traits that are either continuous or discontinuous between generations (e.g, prostate cancer patient and eye defects within family trees). In a medical setup, we may want to model the percentage of patients who have successfully undergone a particular medication procedure. We may want to assess whether the success proportions are equal among a number of hospitals. Given the existence of some un-predetermined excess variation among the different hospitals, the information obtained would have a lot on policy implications.

This work may also be applied in surveys of consumption of a product or services for a small time frame, like one described in section 3.1; or any other types of behavior reporting in a short restrospective time period, such as the consumer purchasing behavior on products or services.

This work may also be applied in the agricultural set-up. For example, in Kenya, bee farming can be improved based on the knowledge from this distribution. One may access forage preferences in the different kinds of bees (bees that live in hives, ant-holes and tree barks in forests). We may be interested in investigating the behavior of bees among different colour of flowers and modeling the pattern of visitation as a random movement. This will be a test that will be used to advise farmers on the colour of flowers to plant depending on the kinds of bees reared in their farms.

REFERENCES

Ahn, H., and Chen, J. J. (1995). Generation of over-dispersed and under-dispersed binomial variates. *Journal of Computational and Graphical Statistics,* **4**: 55-64.

Alanko, T., and Lemmens, P. H. (1996). Response effects in consumption surveys: an application of the beta-binomial model to self-reported drinking frequencies. *Journal of Offic*ial *Statistics,* **12**: 253-273.

Alexander, C., Cordeiro, G. M., Ortega, E. M., and Sarabia, J. M. (2012). Generalized beta-generated distributions. *Computational Statistics and Data Analysis,* **56**: 1880-1897.

Barnwal, R.K., and Paul, S.R. (1988). Analysis of one-way layout of count data with negative binomial variation. *Biometrika,* **75**: 215-222

Bartoo, J.B., and Puri, P.S. (1967). On optimal asympotic tests of composite statistical hypothesis. *Ann. Math, Statist.,* **38**: 1845-1852.

Boos, D.D. (1992). On the generalized score tests. *The American Statistician,* **46**: 327-333.

Bühlur, W. J., and Puri, P.S. (1966). On optimal asymptotic tests of composite hypothesis with several constraints. Z. *Wahrsch,* **5**: 71-88.

Chandrabose, M., Pushpa, W., and Roshan D. (2013). The McDonald Generalized Beta-Binomial distribution: A New Binomial Mixture Distribution and simulation based comparison with its nested distributions in handling overdispersion. *International journal of statistics and probability*; **2**: 213 - 223.

Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Cramer, H. (1946). *Mathematical methods of statistics.* Prenceton University Press.

Dean, C., and Lawless, J.F. (1990). Tests for detecting over-dispersion in poisson regression models. *J. Am. Statist. Assoc***., 84**: 467-472.

Li, X. H., Huang, Y. Y., and Zhao, X. Y. (2011). The Kumaraswamy Binomial Distribution. *Chinese Journal of Applied Probability and Statistics,* **27**: 511-521.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society,* **52**: 647-663.

McDonald, J. B., and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics,* **66**: 133-152.

Moran, P.A.P. (1970). On asymptotically optimal tests of composite hypotheses tests. *Biometrika,* **57**: 45-75.

Moran, P.A.P. (1973). Asymptotic properties of homogeneity tests. *Biometrika*, **60**: 79-85.

Nelder, J.A. and Pregibon, D.(1987). An extended quasi-likelihood function. *Biometrika*, **74**: 221-232.

Neyman, J. (1959). Optimal asymptotic test for composite hypotheses. In probability and statistics, ed. U. Grenander, 213-234. The Harold Cramer Volume. Uppsala: Almgrist Wiksell.

Neyman, J., and Scott, E.L. (1966). On the use of C(α) optimal tests of composite hypotheses. *Bull. Int. Statist. Inst.*, **41**: 477-497.

Paul, S. R. (1982). Analysis of proportions of affected foetuses in teratological experiments. *Biometrics*, **38**: 361-370.

Paul, S. R., and Islam, A. S. (1992). C(α) tests for homogeneity of proportions in taxicology in presence of Beta-Binomial over-dispersion. Accepted in David Williams Volume: *Statistics in Toxicology.* Oxford University Press.

Islam, A. S. (1994). Parametric and semi-parametric models for the analysis of proportions in presence of over/under dispersion. PhD thesis, University of Windsor

Paul, S. R., Liang, K. Y., and Self, S. A (1989). On testing departure from the bino mial and multinomial assumptions. *Biometrics*, **45**: 231-236.

Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Proc. Camb. Phil. Soc.*, **44**: 50-57

Rodr´ıguez-Avi, J., Conde-S´anchez, A., S´aez-Castillo, A. J., and Olmo-Jim´enez, M. J. (2007). A generalization of the beta-binomial distribution. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **56**: 51-61.

Tarone, R.E. (1985). On heterogeneity tests based on efficient scores. *Biometrika*, **72**: 91-95.

Wedderburn, R. W. M. (1974). Quasi-likelihood function, generalized linear models and the gauss-newton method. *Biometrika*, **61**: 439-447.

APPENDICES

Appendix A: Size and Power of the LR test

```
############### simulation of McGBB distribution ##############
dMcGBB<-function(x){
                    j<-0:(n-x)
        term<-sum(((-1)^j)*(choose(n-x,j))*(beta(((x/c)+a+(j/c)),b)))
          return(choose(n,x)*(1/beta(a,b))*term)
              }
G<-1000
jj<-0
for (ii in 1:G){
matfre<-matrix(0,3,n+1)
matpar<-matrix(c(0.8,0.5,1,0.4,0.3,1),2,byrow=TRUE)
      for (l in 1:2){
                          a<-matpar[l,1]
                          b<-matpar[l,2]
                          c<-matpar[l,3]
      pj<-rep(0,7)
            for (k in 1:8)  pj[k]<-dMcGBB(k-1)
            qj<-c(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)
            N<-500
            Xobs<-rep(0,N)
            for (i in 1:N){
      u2<-4; d<-2
      while(u2>d){
         u1<-runif(1); u2<-runif(1)
         y<-trunc(8*u1)+1
```

```
            d<-pj[y]/0.22}
            Xobs[i]<-y-1
                  }
                        matfre[l,]<-table(Xobs)
                  }
                  matfre[3,]<-matfre[1,]+matfre[2,]
f1<-matfre[1,]
f2<-matfre[2,]
f3<-matfre[3,]
matfreq<-matrix(c(f1,f2,f3),3, byrow=TRUE)
matest<-matrix(0,3,3)
################ McGBB Negative Log Likelihood ##############
library(bbmle)
alpha<-0.05
MGenBetaBinNLL<-function(a,b,c,fre,n){
            density<-c()
            for( i in 0:n){
            j <- 0:(n-i)
                  term<-sum(((-1)**j)*(choose(n-i,j))*(beta(((i/c)+a+(j/c)),b)))
                  vector.density<-choose(n,i)*(1/beta(a,b))*term
                  density[i+1]<-vector.density
                  }

            MGBBLL<-sum(fre*log(density))
            return(-MGBBLL)
            }
      veclogs<-rep(0,3)
```

```
for (k in 1:3){
    frequency<-matfreq[k,]
  #### Estimates of the parameters of McDonald Generalized BETA-BINOMIAL ########
    est<-mle2(MGenBetaBinNLL, start=list(a=0.7,b=0.5,c=1), data=list(fre=frequency, n=7))
    a1<-summary(est)@coef["a","Estimate"]
    b1<-summary(est)@coef["b","Estimate"]
    c1<-summary(est)@coef["c","Estimate"]
    matest[k,]<-c(a1,b1,c1)
    veclogs[k]<-MGenBetaBinNLL(a1,b1,c1,frequency,n=7)}
lambda<--2*(veclogs[1]+veclogs[2]-veclogs[3])
if (lambda>=qchisq(1-alpha,3)){R<-1}else{R<-0}
if (R==0){jj<-jj}else{jj<-jj+1}
print(ii)}
jj/G
```

Appendix B: Size and Power of $C(\alpha)$ test based on Quasi-likelihood

```
####simulation of McGBB distribution
dMcGBB<-function(x){
                    j<-0:(n-x)
          term<-sum(((-1)^j)*(choose(n-x,j))*(beta(((x/c)+a+(j/c)),b)))
          return(choose(n,x)*(1/beta(a,b))*term)
              }
G<-1000
jj<-0
for (ii in 1:G){
matfre<-matrix(0,3,n+1)
matpar<-matrix(c(0.2,0.7,1,0.22,0.7,1),2,byrow=TRUE)
        for (l in 1:2){
                              a<-matpar[l,1]
```

```
                    b<-matpar[l,2]
                    c<-matpar[l,3]
    pj<-rep(0,7)
        for (k in 1:8)  pj[k]<-dMcGBB(k-1)
        qj<-c(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)
        N<-200
        Xobs<-rep(0,N)
        for (i in 1:N){
    u2<-4; d<-2
    while(u2>d){
      u1<-runif(1); u2<-runif(1)
      y<-trunc(8*u1)+1
      d<-pj[y]/0.22}
      Xobs[i]<-y-1
          }
                    matfre[l,]<-table(Xobs)
              }
              matfre[3,]<-matfre[1,]+matfre[2,]
f1<-matfre[1,]
f2<-matfre[2,]
f3<-matfre[3,]
matfreq<-matrix(c(f1,f2,f3),3, byrow=TRUE)
# Quasi Negative Log Likelihood
library(bbmle)
alpha<-0.05
QLNLL<-function(a,b,c){
        n<-7
        pii<-(beta(a+b,1/c))/(beta(a,1/c))
        y0<-rep(0:7,f3)
        ybar<-mean(y0)
        aa<-which(y0==0 | y0==7)
```

```
                y<-y0[-aa]
                phi<-(var(y0)/(ybar*(1-ybar/n))-1)/(n-1)
                Z<-y/n
            term<-y*log(pii/Z)+(n-y)*log((1-pii)/(1-Z))
            QL<-sum((1/(1+(n-1)*phi))*term)
                return(-QL)
                }
#### Estimates of the par of Quasi-Likelihood
   est<-mle2(QLNLL, start=list(a=0.037,b=0.195,c=5))
#### C-alpha test declarations
   a2<-summary(est)@coef["a","Estimate"]
   b2<-summary(est)@coef["b","Estimate"]
   c2<-summary(est)@coef["c","Estimate"]
## C(alpha) test based on Quasi-likelihood
                n<-7
                pii<-(beta(a+b,1/c))/(beta(a,1/c))
                y<-rep(0:7,f3)
                ybar<-mean(y)
                phi<-(var(y0)/(ybar*(1-ybar/n))-1)/(n-1)
                S<-ψ
                D<-B11
                A<-matrix(c(A11, A12, A13), 1, byrow=”TRUE” )
        B<-matrix(c(B11, B12, B13, B21, B22, B23, B31, B32, B33), 3, byrow=”TRUE”)
        lambda<-t(S)*solve(D-A*solve(B)*t(A))*(S)
if (lambda>=qchisq(1-alpha,3)){R<-1}else{R<-0}
if (R==0){jj<-jj}else{jj<-jj+1}
print(ii)}
jj/G
```

Appendix C: Size and Power of $C(\alpha)$ test based on Extended Quasi-likelihood

```
#### simulation of McGBB distribution
dMcGBB<-function(x){
                    j<-0:(n-x)
          term<-sum(((-1)^j)*(choose(n-x,j))*(beta(((x/c)+a+(j/c)),b)))
          return(choose(n,x)*(1/beta(a,b))*term)
              }
G<-1000
jj<-0
for (ii in 1:G){
matfre<-matrix(0,3,n+1)
matpar<-matrix(c(0.2,0.7,1,0.22,0.7,1),2,byrow=TRUE)
      for (l in 1:2){
                              a<-matpar[l,1]
                              b<-matpar[l,2]
                              c<-matpar[l,3]
      pj<-rep(0,7)
            for (k in 1:8)  pj[k]<-dMcGBB(k-1)
            qj<-c(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)
            N<-200
            Xobs<-rep(0,N)
            for (i in 1:N){
      u2<-4; d<-2
      while(u2>d){
         u1<-runif(1); u2<-runif(1)
         y<-trunc(8*u1)+1
         d<-pj[y]/0.22}
         Xobs[i]<-y-1
                }
                              matfre[l,]<-table(Xobs)
                    }
```

```
                    matfre[3,]<-matfre[1,]+matfre[2,]
f1<-matfre[1,]
f2<-matfre[2,]
f3<-matfre[3,]
matfreq<-matrix(c(f1,f2,f3),3, byrow=TRUE)
# Extended Quasi Negative Log Likelihood
   library(bbmle)
alpha<-0.05
EQLNLL<-function(a,b,c){
          n<-7
          pii<-(beta(a+b,1/c))/(beta(a,1/c))
          y0<-rep(0:7,f3)
          ybar<-mean(y0)
          aa<-which(y0==0 | y0==7)
          y<-y0[-aa]
          phi<-(var(y0)/(ybar*(1-ybar/n))-1)/(n-1)
          Z<-y/n
        term<-(2/(1+(n-1)*phi))*(y*log(pii/Z)+(n-y)*log((1-pii)/(1-Z)))
        EQL<-(1/2)*sum(log(1+(n-1)*phi)+term)
          return(-EQL)
          }
  # Estimates of the par of Quasi-Likelihood
   est<-mle2(EQLNLL, start=list(a=0.037,b=0.195,c=24))
   a2<-summary(est)@coef["a","Estimate"]
   b2<-summary(est)@coef["b","Estimate"]
   c2<-summary(est)@coef["c","Estimate"]
## C(alpha) test based on Extended Quasi-likelihood
          n<-7
          pii<-(beta(a+b,1/c))/(beta(a,1/c))
          y<-rep(0:7,f3)
          ybar<-mean(y)
```

```
        phi<-(var(y)/(ybar*(1-ybar/n))-1)/(n-1)
        S<-ψ
        D<-B11
        A<-matrix(c(A11, A12, A13), 1, byrow=”TRUE” )
    B<-matrix(c(B11, B12, B13, B21, B22, B23, B31, B32, B33), 3, byrow=”TRUE”)
    lambda<-t(S)*solve(D-A*solve(B)*t(A))*(S)
if (lambda>=qchisq(1-alpha,3)){R<-1}else{R<-0}
if (R==0){jj<-jj}else{jj<-jj+1}
print(ii)}
jj/G
```

Graphs for Empirical levels comparison for $C(\alpha)$ tests and LR test

```
require(stats)
xlab.names<-expression(alpha)
main.names<-expression(paste("(a) Empirical Levels of C(QL) test, C(EQL) test and LR test based on 1000 simulated data sets"))
win.graph()
par(mfrow=c(1,2))
ylim1<-seq(0,0.2,0.05)
values<-c(0.1, 0.15, 0.20, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1)
EMPIRICAL.LEVELS1<-c(0.048, 0.046, 0.046, 0.048, 0.041, 0.044, 0.049, 0.050, 0.052, 0.049, 0.049, 0.045, 0.044, 0.050, 0.049, 0.050, 0.041, 0.049, 0.048)
EMPIRICAL.LEVELS2<-c(0.044, 0.046, 0.042, 0.029, 0.038, 0.027, 0.022, 0.023, 0.038, 0.099, 0.129, 0.085, 0.101, 0.076, 0.058, 0.073, 0.058, 0.051, 0.065)
EMPIRICAL.LEVELS3<-c(0.046, 0.044, 0.042, 0.048, 0.045, 0.039, 0.048, 0.046, 0.057, 0.043, 0.046, 0.054, 0.038, 0.047, 0.049, 0.052, 0.042, 0.056, 0.054)
plot(values,EMPIRICAL.LEVELS1,col=4,lwd=2,lty=1,type="l",pch=15,xlim=c(0.1,1),ylim=c(0,0.2),font.main=3, cex.main=0.9,xlab=xlab.names,ylab="Empirical Levels", main=main.names)
lines(values,EMPIRICAL.LEVELS2,col=6,lwd=2,lty=2,type="l",pch=16,xlim=c(0.1,1),ylim=c(0,0.2),xlab=xlab.names,ylab="Empirical Levels")
```

```
lines(values,EMPIRICAL.LEVELS3,col=11,lwd=2,lty=3,type="l",pch=17,xlim=c(0.1,1),ylim=c(0,0.2),xlab=xlab.names,ylab="Empirical Levels")
legend("topleft",inset=.05,col = c(4, 6, 11),lwd=2,pch=c(15,16,17),lty=c(1,2),cex=0.8, title="", c("C(QL) test","LR test","C(EQL) test"),horiz=F)
```

Graphs for Empirical powers comparison for $C(\alpha)$ tests and LR test

```
xlab.names<-expression(alpha)
main.names<-expression(paste("(b) Empirical powers of C(QL) test, C(EQL) test and LR test based on 1000 simulated data sets"))
values1<-c(0.22, 0.24, 0.26, 0.28, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.70, 0.71, 0.72, 0.73, 0.74, 0.75)
EMPIRICAL.POWERS1<-c(0.165, 0.247, 0.34, 0.381, 0.452, 0.554, 0.622, 0.748, 0.803, 0.856, 0.886, 0.895, 0.966, 0.974, 0.975, 0.975, 0.977, 0.979, 0.986, 0.989, 0.992, 0.993, 0.995, 0.996, 0.998, 0.998)
EMPIRICAL.POWERS2<-c(0.058, 0.071, 0.108, 0.125, 0.147, 0.270, 0.350, 0.522, 0.644, 0.698, 0.857, 0.879, 0.910, 0.925, 0.942, 0.956, 0.968, 0.970, 0.981, 0.982, 0.987, 0.989, 0.990, 0.992, 0.996, 0.998)
EMPIRICAL.POWERS3<-c(0.126, 0.189, 0.231, 0.285, 0.341, 0.508, 0.581, 0.662, 0.716, 0.775, 0.824, 0.919, 0.932, 0.952, 0.953, 0.964, 0.967, 0.972, 0.977, 0.983, 0.992, 0.992, 0.994, 0.998, 0.998, 0.999)
plot(values1,EMPIRICAL.POWERS1,col=4,lwd=2,lty=1,type="l",pch=17,xlim=c(0.22,0.75),ylim=c(0,1),font.main=3,cex.main=0.9,xlab=xlab.names,ylab="EmpiricalPowers",main=main.names)
lines(values1,EMPIRICAL.POWERS2,col=6,lwd=2,lty=2,type="l",pch=18,xlim=c(0.3,0.75),ylim=c(0.15,1),xlab=xlab.names,ylab="Empirical Powers")
lines(values1,EMPIRICAL.POWERS3,col=11,lwd=2,lty=3,type="l",pch=19,xlim=c(0.3,0.75),ylim=c(0.15,1),xlab=xlab.names,ylab="Empirical Powers")
legend("topleft",inset=.05,col = c(4, 6, 11),lwd=2,pch=c(17,18, 19),lty=c(1,2),cex=0.8, title="", c("C(QL) test","LR test","C(EQL) test"),horiz=F)
```