INTERVAL ESTIMATION FOR THE BETA-BINOMIAL DISPERSION PARAMETER

SIELE DICKSON CHERUIYOT

A Thesis Submitted to the Graduate School in Partial Fulfillment for the Requirements of the award of Master of Science Degree in Statistics of Egerton University

EGERTON UNIVERSITY

May 2012

DECLARATION AND RECOMMENDATION

Declaration

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

Signature: …….. Date: ……..

Siele Dickson Cheruiyot
SM12/ 2541/ 09

Recommendations

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

Signature: …….. Date: ……..

Dr. Ali Salim Islam
Egerton University

Signature: …….. Date: ……..

Dr. Orawo Luke Akongo
Egerton University

COPY RIGHT

DEDICATION

To

My dad Samson and mum Catherine.

ACKNOWLEDGEMENT

First I would like to thank the Lord Almighty for bestowing upon me a good health, sane mind, strength and sufficient grace to see me through this study. Thank you and Glory be to You Lord!

Secondly, I pass my regards to my supervisors; Dr. Ali S. Islam and Dr. Orawo L. Akongo for their guidance and relentless support that enabled me come up with this project work. To Dr. Ali; thank you for your encouragement and believing in me every step we made and achievements we attained. I wouldn't forget to thank the Chairman, Department of Mathematics, Dr. J. W. Mwangi for providing the necessary research materials I needed for this work. To my lecturers in the department, thank you all for your support and the humble and friendly environment you provided. To my colleagues; we've been through challenges that we've had to break up and make up. We've learnt a lot and maintained the friendship and have encouraged one another move on even when all seemed difficult to bear. Thank you for your support.

To my beloved family; mum and dad, you've been my encouragement all the way. Thank you for the moral and financial support you have accorded me. You guided me and led me along the right part of life and supported me attain what I want in life. My brothers and sisters, Nelly, Lillian, Kennedy, Titus, Oliviah, Allan and Lornah, you have also supported me whenever I needed your assistance. Thank you and God bless you.

To my dear wife Susan; thank you for loving me and persevering with me in the hard times we've had. We should appreciate all those who stood with us along the way to this successful achievement. My sons; Carlos and Collins, you have made me a proud father. Strive to achieve better than me. God bless and protect you always.

To my church members and friends; thanks for every contribution you made in one way or another.

I also would not forget to thank the Kenya National Council of Science and Technology for their financial support in this work. Your financial support is a good investment towards meaningful research.

Thank you and God bless you all!

ABSTRACT

The dispersion parameter in proportions occurring in toxicology, biology, clinical medicine and epidemiology is important in making inference regarding the regression parameters on the mean. Most data in form of proportions often exhibit excess variation (extra-dispersion). This can arise when the data is from different sub-populations (clusters) or when the assumption of independence is violated. The Beta-Binomial distribution has been applied to model over-dispersion in binary responses in clustered samples. This parametric procedure involves numerical methods of finding MLEs. Many authors have also proposed among other non-parametric procedures, the Quasi-likelihood and Method of Moments for estimation of the over-dispersion parameter. However, much literature focuses discussion on point estimation only. Interval estimation for the over-dispersion parameter in proportions is yet to be done. In this thesis, estimates for the construction of asymptotic confidence interval for the over-dispersion parameter based on the Beta-Binomial distribution, Method of Moments and Quasi-Likelihood procedures were first derived using: the likelihood function in the case of MLE and the quadratic estimating equations for Quasi-likelihood procedure and the Method of Moments. We then apply Monte' Carlo simulation technique to perform bootstrapping procedures for the case of equal and un-equal cluster sizes. The asymptotic coverage probabilities with the lengths of confidence intervals were computed for small and large cluster samples. It is apparent from simulation results that confidence interval lengths reduce with the increase in the mean response probability or increase in the cluster size. The asymptotic CIs based on these three estimators have coverage below the nominal coverage probability (0.95). This shows that these confidence intervals are completely inadequate.
Moreover, when the over-dispersion parameter is small, the resulting coverage probabilities are high. These coverage probabilities decrease as the over-dispersion parameter exceeds 0.3. It was observed that when the cluster size is greater than or equal to 40 , the over-dispersion parameter estimate performs well in the coverage probabilities except when this parameter is greater than 0.2. It is concluded that bootstrapping technique reduces the width of confidence intervals and improves coverage probabilities significantly for the case of unequal cluster sizes in over-dispersed data. An example of real biological proportions data was presented to demonstrate the above results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BB	Beta-Binomial
CI	Confidence Interval
DEQL	Double Extended Quasi-Likelihood
EQL	Extended Quasi-Likelihood
MLE	Maximum Likelihood Estimator
MME	Method of Moments Estimator
SE	Standard Error

CHAPTER ONE
INTRODUCTION

1.1 Background Information

When conducting studies on discrete data, one common assumption is that the population under consideration is homogeneous. Under this assumption, one may try and use a single density function to model the population. However, sometimes using a single probability distribution function may lead to incorrect results. This may be due to the fact that the population is heterogeneous. Heterogeneity comes from the inherent characteristics possessed by the population. This may be brought about by clustering of the population under study. These inherent characteristics may vary within the population. The characteristic is known as over-dispersion. One common attempt to overcome this problem is to use a mixture of distributions, such as the negative binomial and the beta-binomial distributions for data presented in counts and proportions respectively.

In toxicological studies, the number of occurrences of a certain kind in a litter of fetuses which may be death, malformation or mental disorder is recorded. The experimental unit is not the individual fetus but rather the litter itself because the mother is administered with the treatment and observation is made on the fetus. The mean parameter differs across litters even in the same treatment group and follows a beta distribution. Although each fetus within the same litter is assumed to have the same mean, the probability varies across litters and the experimental units for each count variables Y_i are observed.

Owing to its simplicity, many researchers have resorted to the use of the negative binomial distribution. Extensive work has been done on point estimation in terms of bias and efficiency and the test for presence of over-dispersion for both counts and proportions data. Saha (2010) has done exhaustive study on the dispersion parameter for counts data and obtained improved confidence intervals using profile likelihood. Studies on the small sample coverage probabilities of four different approaches; the Wald statistics, profile likelihood, hybrid profile variance and parametric bootstrap based on estimates of the dispersion parameter were performed. Convergence of parameter estimates is faster since it belongs to the class of exponential family.

Paul and Islam (1998) have derived the C(α) tests for testing homogeneity of proportions in presence of over-dispersion. They used the method of moments, quasi-likelihood, extended quasi-likelihood and maximum likelihood among other procedures to estimate the mean of

proportions data for equal dispersion parameters. Lee (2003) extended this work to derive the C(α) test for such data in presence of unequal dispersion parameter and in the process gave estimates of both the mean and dispersion parameters. Paul (2009) estimated the mean and regression parameters in proportions data and has used the above estimation methods to study the efficiency of these estimates based on the two parameters (the mean and dispersion parameters). None of the above authors has performed studies on confidence intervals for the over-dispersion parameter for binary over-dispersed proportions data using any of the above estimation procedures. This study addresses the problem of construction of bootstrap confidence intervals for the over-dispersion parameter and in the process, variance functions for the parameter estimates are derived and used in the estimation of the small and large sample coverage probabilities of the CIs obtained based on the Inagaki (1973) results.

1.2 Statement of the problem

Many simulation studies have examined the bias and efficiency of different estimators of the over-dispersion parameter in over-dispersed proportions data. Point estimation of the parameters for the BB distribution has also been done using; MLE, MME, and quasi-likelihood approaches. No studies have so far been made on the construction and accuracy of the confidence intervals for the over-dispersion parameter in over-dispersed proportions data.. This study is therefore geared towards obtaining confidence intervals for the over-dispersion parameter in over-dispersed binary response data in proportions based on the above three estimation procedures. This will be done by performing a Large scale Monte' Carlo simulation technique with bootstrapping. Bootstrapping is an important technique in the construction of confidence intervals with high coverage probabilities if applied on over-dispersed data. Data arising from equal sample sized groups was used and generalizations were made for unequal sample sized groups. Furthermore, estimation was done for the corresponding coverage probabilities of the CIs based on the above methods for the purpose of performing comparative studies for the same.

1.3 Objectives

1.3.1 General objective

The overall objective of this study was to construct bootstrap confidence intervals for the over-dispersion parameter based on: MLE, MME, and EQL.

1.3.2 Specific objectives

1. To derive estimates for the mean and variance of the over-dispersion parameter for over-dispersed binary response data in proportions using; MLE, MME and EQL.
2. To derive expressions for the construction of bootstrap CIs based on the above estimation procedures for over-dispersed binary response data.
3. To estimate coverage probabilities for the obtained bootstrap CIs by simulation approach.
4. To perform a comparative study on the above three methods based on widths and coverage probabilities of the bootstrapped confidence intervals.

1.4 Assumptions

i. This study assumes that the data is over-dispersed.
ii. The data was collected by simple random sampling among litters in order to avoid bias. The intra-litter proportions were un-equal.
iii. The within litter successes are correlated whereas the between litter successes are independent and identically distributed.

1.5 Justification

The over-dispersion parameter is important for making inference regarding the regression parameters on the mean of over-dispersed binary response data in proportions. A confidence interval is used to determine the accuracy of a point estimate. We measure the strength of a confidence interval based on its width and its coverage probability on the parameter estimate. Extensive work has been done on point estimation of the mean and the over-dispersion parameter by Paul and Islam (1998) and Paul (2009). No work has been done on interval estimation for over-dispersed data in proportions. Bootstrapping has been known

to provide improved CIs and is useful even when very little is known about the underlying distributions.

The aim of this study is to derive bootstrap confidence interval estimates for the over-dispersion parameter by a large scale Monte' Carlo simulation approach. The widths of CIs will also be presented. The results displayed are useful in the analysis of toxicological data. This will have an impact on policy making.

1.6 Expected outputs

The following outputs were derived from this work:

(i) A procedure for obtaining bootstrap CI estimates for the over-dispersion parameter using computer simulation approach was developed.

(ii) A paper for publication in referred journals.

(iii) Master of Science degree in statistics.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

The dispersion parameter in proportions occurring in toxicology, biology, clinical medicine, epidemiology and other similar fields is important in making inference regarding the logistic regression parameters on the mean of over-dispersed data, (Paul and Islam, 1998, Saha, 2008). In studies where the experimental unit is a litter, it has been observed that due to the genetic similarity and the same treatment conditions, fetuses within the same litter tend to behave more similarly than those from different litters. As a consequence, littermates are likely to be dependent. This tendency of littermates to respond more alike than animals from different litters is called the 'litter effect'. This litter effect is also known as the extra-dispersion (over/under-dispersion) or the intra-litter correlation or the intra-class correlation.

A common approach to modeling the extra-binomial data is to assume that the proportions for the binomial probability are random. The beta-binomial model (Williams, 1961, Crowder, 1978) assumes that the responses within a cluster follow an independent Bernoulli process, and the Bernoulli parameter, itself a random variable, varies from cluster to cluster according to a beta distribution. One difficulty with the use of the likelihood-based beta-binomial model is the bias and instability of its MLE's, (Yamamoto and Yanagimoto, 1994). The beta-binomial approach implicitly assumes an underlying variance-covariance structure for the intra-cluster correlation induced by the random effects. It is well known that MLEs based on BB distribution may be biased when the sample size n or the total Fisher information is small (Paul and Islam, 1998). The bias is usually ignored in practice; the justification being that it is small compared with the standard errors. Williams (1975) proposed using the quasi-likelihood model, which assumed only the first two moments (mean and variance) for the distribution of the Bernoulli parameter, as an alternative to the beta-binomial. The quasi-likelihood approach is more robust and involves easier computations than the maximum likelihood method under the beta-binomial. The other estimator is the double extended quasi-likelihood estimator. Lee and Nelder (2001) compare the efficiencies of; DEQL, MLE, MME and EQL based on the BB model and using simulated data from the BB distribution they conclude that the DEQL estimator has high efficiency.

In many situations the dispersion parameter or the intra-class correlation parameter may be of interest in its own right. Estimation of the dispersion parameter is important for

making inference regarding the logistic regression parameters, (Saha and Paul 2009). Much study has been done to estimate the mean and the over-dispersion parameters.

For the Beta-Binomial distribution, marginal or conditional estimation of the dispersion parameter is difficult. Joint estimation of the regression (mean) and the dispersion parameters was therefore considered. We take a parametric model, the beta-binomial or the extended beta-binomial model to allow over-dispersion as well as under-dispersion to obtain maximum likelihood estimates of the parameters. If the assumption of the parametric model is accurate, maximum likelihood estimates are known from classical theory to be consistent, asymptotically normal, and efficient. However, maximum likelihood procedure may produce inefficient or biased estimates when the assumed parametric model does not fit the data well.

Alternatively, more robust estimates such as moment estimates of Klein-man (1973), quasi-likelihood estimates of Moore and Tsiatis (1991), extended quasi-likelihood estimates of Nelder and Pregibon (1987), the Gaussian likelihood estimates of Crowder (1985), estimates based on the pseudo-likelihood estimating equations of Davidian and Carrol (1987) and estimates based on quadratic estimating functions of Crowder (1987) and Godambe and Thompson (1989) can be considered. Paul and Islam (1998) studied six such estimates and compared the small and large sample efficiency and bias properties of these estimates with the maximum likelihood estimates.

The method of moments, which requires assumptions on the form of the mean and variance, is an alternative approach to over dispersed data. Klein-man (1973) presented a moment method for proportions from a single sample. The quasi-likelihood method for over dispersion may also be considered as a moment method (Wedderburn, 1974).

Saha and Paul (2009) obtained bootstrap confidence intervals for the over-dispersion dispersion parameter using the negative binomial model. The bootstrap procedures improved the coverage probabilities of the CIs obtained.

For the three approaches that were considered in construction of confidence intervals, of interest were the mean and variance estimates for the over-dispersion parameter. Construction of confidence intervals then followed and comparison based on coverage probabilities of the CIs under the three estimating procedures was done.

2.2 Over-dispersion

Let x_i be the number of successes in n_i trials. The respective proportions (Y_i) of successes in this case are random each with probability p_i, $i = 1,...,n$.

The following first and second moments hold for the probabilities p_i.

$$E(p_i) = \pi_i \text{ and } \text{var}(p_i) = \phi\ \pi_i(1-\pi_i)\ ,\ \phi \geq 0$$

by conditional expectation, it can be proved that;

$$E(Y_i) = n_i\pi_i \text{ and}$$

$$\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\left[1+\ n_i - 1\ \phi\right]$$

For ϕ =0, it implies that there is no over-dispersion and thus we have the binomial mean and variance.

Over-dispersion corresponds to unexpected heterogeneity in the outcomes of a toxicity test. It occurs when the variance of the response data is greater than the nominal variance. Discrete data which come in the form of counts or proportions often display greater variability than would be predicted by simply fitting binomial or Poisson models. Over-dispersion has been seen to occur when the population is clustered. Clusters in this case, are in form of house-holds, litters and colonies which vary in size. The elements of these clusters each possess a random variable within itself leading to an extra variability within the clusters. The dispersion parameter thus depends on the cluster size and on the variability of proportions from cluster to cluster.

For data in form of counts and proportions, parameter estimates are distribution sensitive and may lead to incorrect statistical inference when the assumed distribution is incorrect, Sudhir (2009). In the context of confidence intervals, the estimated standard error of the parameter estimates in the analysis will be too small and thus will provide confidence intervals that are too narrow with very low coverage probabilities. An approach to dealing with this problem is to specify parametric models that accommodate over-dispersion and collapse to simpler models when over-dispersion is not present. Haseman and Kupper (1979) proposed several methods to be used with the binomial family. This has found much prominence in the literature. This entail: scaling the standard errors by the dispersion statistics, using robust variance estimators and application of mixed model techniques.

A common way to account for the over-dispersion is to assume that the intra-litter correlation is induced by some random effect shared by all the elements within the same

cluster. This random effect can be looked upon as the combined effect of all factors, both genetic and environmental, that are shared by the littermates. Given this litter specific random effect, the outcomes of the littermates are assumed to be conditionally independent. The use of a beta distribution to model the random effects results in the beta-binomial distribution (Williams, 1975; Haseman and Kupper, 1979).

2.3 The Beta-Binomial Distribution

This is a mixture of the beta and binomial distributions, an extension of the binomial model. It assumes that the proportions in the binomial distribution have probabilities that are random according to a beta distribution with parameters α and β (Williams 1975, Crowder 1978). This distribution was developed to fit over-dispersed data, Paul and Islam (1998). Basic theoretical properties of this distribution have been discussed by Skellam (1948) and Kleinman (1973).

The beta-binomial distribution has been adopted by among others, Robert and Bailer (2009), to model over-dispersed binary response data in aquatic toxicology. The beta-binomial model assumes that the responses within a cluster follow an independent Bernoulli process. The Bernoulli parameter is itself a random variable that varies from cluster to cluster according to a beta distribution. Jiaxin and Lord (2007) have used this model to analyze car-crash data due to its robustness and ability to handle small mean values.

The flexibility that the Beta-Binomial exhibits is as a result of the two-parameter nature of the Beta-binomial distribution. It is dependent upon how well the Beta distribution can represent the population of π_i's. It can exhibit more plasticity than the one parameter binomial. The α and β parameters determine the shape of the BB distribution. Taking values on the interval [1,0], the distribution is uni-modal if α >1 and β >1. If both α and β are 1, then the beta distribution is equivalent to the continuous uniform distribution on that interval. If only one of these parameters are less than one, then the distribution is J-shaped or reverse J-shaped. If both are less than one, then it is U-shaped. The Beta-Binomial also has the flexibility to model the correlation of observations from the same individual that the binomial distribution does not possess, Moore (1986).

The suitability of the Beta-Binomial distribution to account for extra-dispersion is demonstrated as follows.

Let x_i be the number of successes observed in n_i clusters. Then, p_i be the proportion of the successes. By conditional expectation;

$$\begin{aligned} Var(y_i) &= E\ Var(y_i / p_i)\ + Var\ E(y_i / p_i) \\ &= E\ n_i p_i (1 - p_i)\ + Var(n_i p_i) \\ &= n_i \left[E(p_i) - E(p_i^2) \right] + n_i^2 Var(p_i) \\ &= n_i \pi_i (1 - \pi_i) + n_i (n_i - 1) Var(p_i) \end{aligned}$$

The first part, $n_i p_i (1 - p_i)$, in the final expression is variance of the binomial distribution. The next part represents the extra variability in the model which varies with sample (cluster) size and takes a value zero when there is no over-dispersion ($\phi = 0$).

The beta-binomial distribution is derived as follows;

Let, $X_1, \ldots, X_n$ be the number of successes from different clusters under study. They are independent and identically distributed random variables. Then consider the binomial probability function $f(x / p) = \binom{n}{x} p^x (1 - p)^{n-x}$ whose probability of success p is randomly distributed with parameters α and β.

The proportion p arises from the re-parameterized beta distribution $g(p)$ with parameters π and ϕ. The marginal distribution of x is,

$$f(x) = \int \binom{n}{x} p^x (1 - p)^{n-x} g(p) dp \qquad (2.1)$$

The beta-binomial distribution thus has Bayesian probability function,

$$P(Y_i = \frac{x_i}{n_i}) = \binom{n_i}{x_i} \frac{B(\alpha + x_i, n - x_i + \beta)}{B(\alpha, \beta)} \qquad (2.2)$$

where, $B(\alpha, \beta) = \frac{\lceil (\alpha + \beta)}{\lceil \alpha \lceil \beta}$

We re-parameterize the mean and the over-dispersion parameters in terms of α and β as; $\pi = \frac{\alpha}{\alpha + \beta}$ and $\phi = \frac{1}{1 + \alpha + \beta}$ respectively.

We thus have the probability function for the Beta-Binomial distribution given as;

$$P\left(Y_i = \frac{x_i}{n_i}\right) = \binom{n_i}{x_i} \frac{\prod_{r=0}^{x_i-1}(1-\phi)\pi_i + r\phi \prod_{r=0}^{n_i-x_i-1}((1-\phi)(1-\pi_i)+r\phi)}{\prod_{r=0}^{n_i-1}(1+\phi(r-1))} \quad (2.3)$$

$$0 \le \pi_i \le 1 \text{ and } \max\left(\frac{-1}{n_i - 1}\right) \le \phi \le 1$$

The mean structure $\pi = \pi_i$ is given by the logistic model $\pi_i \ x_i, \beta \ = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

This conditional parameter variation on observed Y_i is given by:

$$Var(Y_i) = Var\ E\ Y_i/P_i\ + E\ Var\ Y_i/P_i$$

$$Var(Y_i) = Var\left(E\left(\frac{x_i}{n_i}\right)\right) + E\left(Var\left[\left(\frac{x_i}{n_i}\right)\right]\right), \text{ where;}$$

$$E\left(\frac{x_i}{n_i}\right) = n_i\alpha\ \ \alpha+\beta\ ^{-1} = n\pi_i(1-\pi_i)$$

and

$$Var\left(E\left(\frac{x_i}{n_i}\right)\right) = n_i^2 Var\ \ \pi_i\ = n_i^2\phi\ \ \pi_i(1-\pi_i)$$

Thus, $Var\ \ Y_i\ = n_i^2 Var(\pi_i) + n_i E(\pi_i) - n_i E\ \ \pi_i^2$

$$= n_i(E(\pi_i) - Var(\pi_i) - E(\pi_i)^2 + n_i Var(\pi_i))$$

$$= n_i\phi\pi_i(1-\pi_i) + n_i(n_i\pi_i) - n_i(\phi\pi_i(1-\pi_i) + n_i^2\pi_i^2)$$

$$= n_i\pi_i\ \ 1-\pi_i\ \ \ 1+\ \ n_i-1\ \ \phi \quad (2.4)$$

where, $\phi = \frac{\pi_i}{1+\pi_i}$, and $E(\pi_i^2) = Var(\pi_i) + E(\pi_i)$

The parameter π_i and ϕ are not orthogonal except when ϕ =0.

This is the extended beta-binomial model which takes into consideration ϕ as positive or negative based on whether the data is over-dispersed or under-dispersed.

2.4 Bootstrapping

Resampling methods involve the use of many samples, each taken from a single sample that was taken from the population of interest. Inference based on resampling makes use of the conditional sampling distribution of a new sample (the "resample") drawn from a given sample. Resampling methods therefore can be useful even when very little is known about the underlying distribution.

A basic idea in bootstrap resampling is that, because the observed sample contains most of the available information about the underlying population, the observed sample can be considered to be the population; hence, the distribution of any relevant statistic can be simulated by using random samples from the "population" consisting of the original sample. Bootstrapping improves the coverage probabilities of confidence intervals if applied to over-dispersed data that is in form of proportions, Saha and Sen (2009).

2.5 The Inagaki results

Denote the unbiased estimating equations obtained by the method of moments and other semi-parametric procedures by $u_1, u_2, ..., u_k$ and u_{k+1}, where u_j, $j = 1, 2, ..., k$ represent unbiased estimates for β_j and u_{k+1} represents the unbiased estimating equation for ϕ.

Let $\hat{\lambda}$ be an estimate for $\lambda = (\beta_1, \beta_2, ..., \beta_k, \phi)$. Using the method of moments or any semi-parametric procedure, the Inagaki (1973) result obtained under the usual regularity conditions, such as the finite dimensional parameter space, the expected values are continuously differentiable, is given by

$$Var(\hat{\lambda}) = A(\hat{\lambda})^{-1} B(\hat{\lambda}) \left[A(\hat{\lambda})^{-1} \right]^T \tag{2.5}$$

where A and B are square matrices of order $k+1$ with entries.

$$A_{j,s} = E\left(-\frac{\partial U_j}{\partial \beta_s} \right) \qquad A_{j,k+1} = E\left(-\frac{\partial U_j}{\partial \phi} \right)$$

$$A_{k+1,j} = E\left(-\frac{\partial U_{k+1}}{\partial \beta_j} \right) \qquad A_{k+1,k+1} = E\left(-\frac{\partial U_{k+1}}{\partial \phi} \right)$$

$$\beta_{j,s} = E\ U_j U_s \qquad \beta_{j,k+1} = \beta_{k+1,j} = E\ U_j U_{k+1}$$

$$\beta_{k+1,k+1} = E\ U_{k+1}^2$$

This is for all, $j, s = 1, 2, ..., k$

2.6 Parameter estimation

2.6.1 Maximum Likelihood Estimation

Maximum likelihood estimators possess the asymptotic properties; consistency, asymptotic normality and asymptotic efficiency. Saha and Paul (2005) noted that both MLE's and MME's are $\sqrt{m}$ consistent estimators of π_i and ϕ.

The elements of the score function in the study; testing for homogeneity of proportions with equal dispersion parameters have been derived (Paul and Islam, 1998).

Saha and Paul (2005) have performed studies on the maximum likelihood estimator using the extended Beta-Binomial (BB) model to analyze over/under-dispersed proportions data. They then estimated this parameter by maximum likelihood estimation (MLE) procedure. They concurred with Williams (1975) in observing that the MLEs may be biased when the sample size n or the total Fisher information is small. This bias is usually ignored in practice, the justification being, it is small compared with the standard errors. Williams (1975) went further to derive a bias correlated Maximum likelihood estimator (BCML), with the view of reducing errors due to bias.

Studies have shown that the maximum likelihood estimators do not converge earlier than expected for non-exponential family of distributions like the Beta-Binomial. Paul and Islam (1998) proposed solving equations (2.6) and (2.7) simultaneously for $\hat{\beta}$ and $\hat{\phi}$.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{x_i - 1} \frac{1-\phi}{\left[1-\phi \;\; \pi_i + r\phi \right]} - \sum_{r=0}^{n_i - x_i - 1} \frac{1-\phi}{\left[1-\phi \;\; 1-\pi_i \;\; + r\phi \right]} \right\} \partial_{ij} \;\; \beta \;\; = 0 \qquad (2.6)$$

and

$$\frac{\partial l}{\partial \phi_{k+1}} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{x_i - 1} \frac{r - \pi_i}{\left[1-\phi \;\; \pi_i + r\phi \right]} + \sum_{r=0}^{n_i - x_i - 1} \frac{r - (1-\pi_i)}{\left[1-\phi \;\; 1-\pi_i \;\; + r\phi \right]} - \sum_{r=0}^{n_i - 1} \frac{r-1}{1-\phi \;\; \pi_i + r\phi} \right\} = 0 \qquad (2.7)$$

where, $\partial_{ij} \;\; \beta \;\; = \frac{\partial \pi_i}{\partial \beta_j} = \pi_i \;\; 1-\pi_i \;\; x_{ij}$

It's worth noting that when there are no covariate-independent samples, $\partial_{ij} \;\; \beta \;\; = 1.$

The estimates of the variance estimator are functions of elements obtained in the score function derived by, Paul and Islam (1998) and Saha, (2008).

The maximum likelihood estimates are hereby denoted by $\hat{\lambda}_{ml}$.

2.6.2 Quasi-Likelihood Estimation

Let, $U = U(u, y) = \dfrac{Y - u}{\sigma^2 \operatorname{var}(u)}$ (2.8)

The quasi-likelihood expression is given by

$Q(u, y) = \int_y^u \dfrac{y - t}{\sigma^2 v_{(t)}} dt$, with $E(u) = 0$, $\operatorname{var}(u) = \dfrac{1}{\sigma^2 v_{(u)}}$ and $-E\left[\dfrac{\partial U}{\partial u}\right] = \dfrac{1}{\sigma^2 \operatorname{var}(u)}$

This likelihood function was first introduced by Nelder and Wedderburn (1972). It is used to draw inferences from experiments in which there are sufficient information to construct the likelihood function. The quasi-likelihood does not assume the probability distribution of unobserved heterogeneity which causes over-dispersion. Its estimation is based on the iteratively re-weighed least squares (IRLS) algorithm, which only requires a relationship between conditional mean and variance instead of its full conditional distribution. This feature was noted by Wedderburn (1974).

The main idea behind quasi-likelihood method is to avoid a fully specified distribution for the response variable when one is uncertain about the random mechanism by which the data were generated. Xu (2008) recommended quasi-likelihood with a common intra-litter correlation parameter be used in the analysis of clustered binary data when the number of litters is small.

The assumption of independent observations warrants that $v(u)$ must be diagonal. The quasi-likelihood method does not require full assumption on the distribution, and the estimates of the dose response coefficients are generally consistent and asymptotically normal even if the structure of the within litter correlation is mis-specified, (Mc. Cullagh and Nelder, 1989).

Our quasi-likelihood in this case is based on the knowledge of the first two moments of the random variable, $Z_i = \dfrac{Y_i}{n_i}$.

By virtue of independence between samples, the quasi-likelihood with the above means and variance is given by $Q = \sum_{i=1}^{n} Q\ z_i, \pi_i, \phi$.

where, $Q\ z_i, \pi_i, \phi\ = \int_{z_i}^{\pi_i} \dfrac{z_i - \pi_i\ n_i}{\pi_i\ 1 - \pi_i\ \ 1 + \ n_i - 1\ \phi} \partial \pi_i$ (2.9)

From the expression $\sum_i n_i\ z_i - \pi_i$ two estimating equations arise; given any value of ϕ the unbiased estimating equation for β_j is

$$U_j\ \beta,\phi\ = \frac{\partial Q}{\partial \beta_j} = \sum_{i=1}^{m} \frac{z_i - \pi_i\ \ n_i \partial_{ij}(\beta)}{\pi_i\ \ 1-\pi_i\ \ \ 1+\ \ n_i - 1\ \ \phi} = 0 \tag{2.10}$$

Paul (2009) suggested that there was no such estimating equation that exists for ϕ. This therefore calls for an unbiased estimating equation that can be obtained by using the method of moments when the k β parameters are estimated. The estimating equation is given by,

$$U_{k+1}\ \beta,\phi\ = \sum_{i=1}^{m} \frac{z_i - \pi_i\ ^2\ n_i}{\pi_i\ \ 1-\pi_i\ \ \ 1+\ \ n_i - 1\ \ \phi} - m - k = 0 \tag{2.11}$$

Equations (2.10) and (2.11) are solved simultaneously to obtain estimates of β and ϕ. These estimates are denoted by $\hat{\lambda}_{Ql}$

2.6.3 Method of Moments Estimation

The moment estimates for the parameters π_i and ϕ can be obtained by equating the first two sample moments to the corresponding population moments. Lee (2003) derived expressions for the mean and over-dispersion parameter estimates as follows:

$$\hat{\mu} = \bar{y} \text{ and } \hat{\phi} = \frac{s^2 - \hat{\mu}}{\hat{\mu}^2}$$

where $s^2 = \sum_{i=1}^{n}\ y_i - \bar{y}\ ^2 / (n-1)$ and $\hat{\mu}$ is the overall mean.

Let $\hat{p} = \sum_{i=1}^{k} \frac{w_i Y_i}{w}$, where $w_i = {n_i}/{1+\phi(n_i - 1)}$ represents a set of weights and w is the sum of all the weights. Y_i is the proportion associated with the number of successes in a given cluster. Also, let $S = \sum_{i=1}^{k} w_i (p_i - \hat{p})^2$. Then using the method of moments, estimates of π and ϕ are respectively,

$$\hat{\bar{\pi}} = \hat{p} \text{ and} \tag{2.12}$$

$$\hat{\phi}_{mme} = \frac{S - \hat{p}\hat{q}\left[\sum_{i=1}^{n} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]}{\hat{p}\hat{q}\left[\sum_{i=1}^{n} w_i \left(1 - \frac{w_i}{w}\right) - \sum_{i=1}^{n} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]} \tag{2.13}$$

In terms of actual data observed from different clusters, let $p_i = \frac{x_i}{n_i}$, i = 1,2, ... k, where i indexes the i^{th} cluster sampled, x_i is the number of successes recorded in the i^{th} cluster and

n_i is the respective sample size of the i^{th} cluster. The n_is here may be un-equal in toxicological studies. w_i is a function of the unknown parameter ϕ and in this case, we let $w_i = n_i / \text{var}(p_i)$ such that we can obtain an initial approximation of estimates of π and ϕ. In cases where ϕ estimates are negative, they are to be set to zero. The sum of all weights w =1. The unbiased estimating equation is given by the expression;

$$\sum_{i=1}^{m} n_i (z_i - \pi_i) = 0$$

This is generalized to the regression situation as

$$U_j(\beta_j, \phi) = \sum_{i=1}^{m} n_i (z_i - \pi_i) \partial_{ij}(\beta) = 0 \tag{2.14}$$

$$U_{k+1}(\beta, \phi) = \sum_{i=1}^{m} n_i^2 \left[z_i - \pi_i \right]^2 - \pi_i (1 - \pi_i)(1 + (n_i - 1)\phi) \tag{2.15}$$

where $\pi_i = \pi_i(x_{i,}\beta) = \dfrac{e^{x_i \beta}}{1 + e^{x_i \beta}}$ and

$$\partial_{ij} \left[\beta \right] = \pi_i (1 - \pi_i) x_{ij}$$

2.7 Interval estimation of the dispersion parameter

In order to obtain confidence intervals, we need to find the mean and the asymptotic variances for the over-dispersion parameter, by MLE, MME and EQL. Paul and Islam (1998) verified that the asymptotic variance of the variance estimator is based on the expectation of the Fisher information. From results of Inagaki (1973), the estimators, let *t = MM, ML* and QL under conditions similar to those for which standard MLE asymptotics hold, are consistent and asymptotically normal (as $m \to \infty$) with covariance matrix,

$$Var(\hat{\phi}_t) = \left[A(\hat{\phi}_t) \right]^{-1} B(\hat{\phi}_t) \left[\left[A(\hat{\phi}_t) \right]^{-1} \right]^T$$

The asymptotic variances of the estimates for the three methods based on the beta-binomial model were obtained. The data that was used was simulated from a Beta-binomial distribution with both α and β parameters taking the value one. We thus obtained proportions data that was over-dispersed and thus avoided misleading inference.

2.7.1 Maximum Likelihood Estimation

By the asymptotic theory of the ML estimates, it can be shown that $\hat{\phi}_{ml} \sim N(\phi, \text{var}\,\hat{\phi}_{ml})$ as $n \to \infty$, (Saha and Sen, 2009). The asymptotic confidence interval of ϕ based on ML method is given by:

$$\hat{\phi}_{ml} \pm Z_{\frac{\alpha}{2}} \sqrt{\text{var}\,\hat{\phi}_{ml}}$$

where $Z_{\alpha/2}$ is the upper $100(1-\alpha)\%$ quantile of the standard normal distribution. $Var(\hat{\phi}_{ml})$ is obtained from the diagonal element of the variance-covariance matrix obtained.

2.7.2 Method of Moments

Following the Inagaki (1973) result, it can be shown that $\hat{\phi}_{mm}$ is consistent and asymptotically normal with $\hat{\phi}_{mm} \sim N(\phi, \text{var}\,\hat{\phi}_{mm})$ as $n \to \infty$. The corresponding $100(1-\alpha)\%$ confidence interval for $\hat{\phi}_{mm}$ is given by,

$$\hat{\phi}_{mm} \pm Z_{\frac{\alpha}{2}} \sqrt{\text{var}\,\hat{\phi}_{mm}}$$

We replace $\hat{\pi}$ by its moment estimate and ϕ by $\hat{\phi}_{mm}$, with

$$Var(\hat{\phi}_{mm}) = A(\hat{\phi}_{mm})^{-1} B(\hat{\phi}_{mm}) \left[A(\hat{\phi}_{mm})^{-1} \right]^T$$

2.7.3 Quasi-Likelihood

Quasi-likelihood estimators are also found to be consistent and asymptotically normal, with

$$\hat{\phi}_{eql} \sim N(\phi, \text{var}\,\hat{\phi}_{eql}) \text{ as } n \to \infty$$

Consequently, the $100(1-\alpha)\%$ confidence interval of ϕ based on the quasi-likelihood estimator is given by.

$$\hat{\phi}_{eql} \pm Z_{\frac{\alpha}{2}} \sqrt{\text{var}\,\hat{\phi}_{eql}}$$

where,

$$Var(\hat{\phi}_{Ql}) = A(\hat{\phi}_{Ql})^{-1} B(\hat{\phi}_{Ql}) \left[A(\hat{\phi}_{Ql})^{-1} \right]^T$$

CHAPTER THREE
METHODS

3.1 Introduction

In this chapter we describe a simulation study for assessing the performance of the confidence intervals for the over-dispersion parameter based on MLE, MME and EQL. We first derive moment estimates and find the first and second derivatives of the estimating equations that will be used in the construction of the hessian matrix. This is then integrated into the subroutines (as shown in the appendix) during simulation in order to obtain results on lengths of CIs (in parenthesis) and percentage coverage probabilities that are displayed in chapter four.

3.2 Simulation study

The Beta-binomial distribution assesses the variability in estimates of the parameters (π and ϕ) when clustered data is collected. This distribution accounts for the extra variability within the clusters and allows for the creation of confidence intervals. The data was simulated for varying π's and fixed ϕ parameters. Data simulated from the Beta-Binomial distribution was then used with the three estimation procedures to generate asymptotic results of coverage probabilities, relative lengths of the confidence intervals and asymptotic bootstrap confidence intervals based on the three estimation procedures under study.

We now present expressions for the variance functions used in the estimation of confidence intervals based on MLE, MME and EQL. The bootstrap confidence intervals thus constructed are based upon asymptotic Normality. This is done in line with the Inagaki (1973) results.

In order to evaluate the accuracy of the beta-binomial CI, a large scale Monte Carlo simulation was conducted. Since there are many parameters involved (n, k, π, ϕ), a theoretical evaluation is difficult to conduct. Therefore, the simulation approach is adopted to study the coverage probability and the width of the CI of the over-dispersion parameter. We chose arbitrary values of π and ϕ that meet the conditions, $0 \le \phi \le 1$ and $0 \le \pi_i \le 1$. For $n_i = 5$, situations were considered where the sample sizes (K) equals 20, 30, 40, 60, 100 and 200, the underlying mean response probability π equals 0.1, 0.2, 0.3, 0.4, 0.6, and 0.7 and the

over-dispersion parameter, ϕ, takes one of five values 0.1, 0.2, 0.3, 0.4 and 0.6. These parameter values were used by Paul (2009).

The simulation study was used to assess the performances of the CIs for the over-dispersion parameter ϕ based on MME, MLE, and EQL. All programming was implemented using R-language, version 2.12.1. This was integrated with its add-on packages, *emdbook*, *sensR* and *bbmle*. The data, the respective number of successes per cluster, was simulated from the beta-binomial distribution arbitrarily varying the size of each cluster. The random numbers are generated in two steps. First, the probabilities of success P_i were generated from a Beta (1,1) distribution. Next, the successes $x_i's$ were generated from a Binomial (5, P_i). The simulated data is termed as the sample and is assumed to contain some information about the underlying population. This sample is considered to be the population that can be resampled from. The bootstrap sample is simulated by using the random sample from the "population" consisting of the original sample. This is done for a thousand times.

For various combinations of π, ϕ, K and n, we simulated beta-binomial random variables using Uniform (0,1) from which estimation of the parameters, π and ϕ was performed for the three procedures above. A total combination of the factors; π, ϕ and k was used giving a total of 180 combinations.

Construction of bootstrap confidence intervals using the three estimation procedures then followed. The derivatives for the variance functions were used for estimation in Maximum Likelihood Estimation, the Method of Moments and Quasi-Likelihood method. Bootstrapping technique was applied in the construction of confidence intervals. Graphical presentations of these confidence intervals were displayed. For each combination of the parameters, 1000 valid samples were generated to compute the coverage probability of the bootstrap CIs, which equals the number of times the CIs by each method, contained the true value of $\phi/1000$ and these were reported in tabular form.

Small cluster sizes generated a high percentage of invalid samples. This situation was also observed when the mean parameter was outside the interval [0.3, 0.7]. Whenever the generated samples produced invalid samples, the sample was discarded and a new one generated till we obtained a total of 1000 valid samples. In addition to this, based on this interval, we selected samples whose mean parameter estimates were contained in the interval above for reasons of consistency of our confidence interval lengths. If the values of the mean

parameter fell outside the range, a new sample with the parameter contained in the interval was obtained.

To display the property of litter effect (tendency of litter-mates to behave in a similar manner), the confidence intervals selected was free of negatives, i.e., the CI should consist of only the positive set. When the confidence interval obtained contained a negative lower interval boundary, simulations for the same combination was repeated. If the sample still produced such CI up to a maximum of 10 times, the lower bound of the CI was approximated to zero. This problem featured especially for small cluster size groups, 20 and 30.

Finally, an illustration of our findings was shown by computing confidence intervals using the real binary data, Potthoff and Whittinghill (1966) data (Example 6.5), which was used by Paul (2009) in the study on efficiency of these estimators.

3.2 Moment estimates

The asymptotic variance-covariance matrix for the estimator ϕ is obtained by using the Inagaki (1973) result. Based on the assumption that the Kurtosis and skewness of our beta-binomial distribution is unknown, we have derived moment estimators and obtained:

$$\mu_{1i} = 0$$

$$\mu_{2i} = \pi_i(1-\pi_i)\{1+(n_i-1)\phi\}/n_i$$

$$\mu_{3i} = \mu_{2i}(1-2\pi_i)\ \ 1+(2n_i-1)\phi\ \ /n_i(1+\phi)$$

$$\mu_{4i} = \mu_{2i}\left[\frac{1+(2n_i-1)\phi\quad 1+(3n_i-1)\phi\ \ (1-3\pi_i(1-\pi_i))}{1-\phi}+(n_i-1)(\phi+3n_i\mu_{2i})\right]\left(\frac{(1-\phi)}{1+\phi\quad 1+2\phi\ \ n_i^2}\right)$$

Based on the forms of moments above we derive the entries of the variance covariance matrix whose entries are derived from the Inagaki (1973) results.

3.2.1 Maximum Likelihood Estimation

Maximum likelihood methods are frequently used in statistical applications. The basic underlying idea for ML estimation is to find the parameter value most likely to have produced the observed data. For a variety of distributions such as the Binomial, maximum likelihood estimates can be found in closed form. However, for the Beta-binomial distribution, no closed form solution exists. Consequently, it is necessary to use numerical computation to estimate π and ϕ. We can obtain standard errors for ML estimates based on the quantity of the log-likelihood at $\hat{\phi}$. This method accounts for the extraneous variability in the cluster and allows for the creation of confidence intervals under certain regularity conditions (conditions of the Inagaki (1973) result and the delta method). The data must be distribution specific to avoid producing biased estimates.

The beta-binomial model function is represented by equation (2.3). We thus find the log-likelihood to be of the form;

$$\log l = C + \sum_{r=0}^{x_i-1} (1-\pi)\phi + r\phi \; + \sum_{r=0}^{n_i-x_i-1} (1-\pi)(1-\phi) + r\phi \; - \sum_{r=0}^{n_i-1} (1+r\phi)$$

Two likelihood equations arise from this log-likelihood;

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{x_i-1} \frac{1-\phi}{\left[\, 1-\phi \;\; \pi_i + r\phi \,\right]} - \sum_{r=0}^{n_i-x_i-1} \frac{1-\phi}{\left[\, 1-\phi \;\; 1-\pi_i \;\; + r\phi \,\right]} \right\} \partial_{ij} \;\; \beta \;\; = 0$$

and

$$\frac{\partial l}{\partial \phi_{k+1}} = \sum_{i=1}^{m} \left\{ \sum_{r=0}^{x_i-1} \frac{r-\pi_i}{\left[\, 1-\phi \;\; \pi_i + r\phi \,\right]} + \sum_{r=0}^{n_i-x_i-1} \frac{r-(1-\pi_i)}{\left[\, 1-\phi \;\; 1-\pi_i \;\; + r\phi \,\right]} - \sum_{r=0}^{n_i-1} \frac{r-1}{1-\phi \;\; \pi_i + r\phi} \right\} = 0$$

where;

$$\partial_{ij} \;\; \beta \;\; = \frac{\partial \pi_i}{\partial \beta_j} = \pi_i \;\; 1-\pi_i \;\; x_{ij}$$

When there are no covariate-independent samples, $\partial_{ij} \;\; \beta \;\; = 1.$

These equations are solved simultaneously by numerical methods to yield estimates of π and ϕ respectively. Taking second derivatives of the equations above with respect to these parameters, we obtain the elements of the variance covariance matrix **I**.

The asymptotic variance –covariance matrix of MLE's is obtained by inverting the expected Fisher information matrix, where,

$$I = \begin{pmatrix} I_{BB} & I_{B\theta} \\ I_{\theta B} & I_{\theta\theta} \end{pmatrix} \qquad I_{BB} = E\left(-\frac{\partial^2 l}{\partial\beta_j \partial\beta_s} \right)_{(kxk)}$$

$$I_{B\theta} = I_{\theta B}^{'} = E\left(-\frac{\partial^2 l}{\partial\beta_j \partial\theta} \right)_{kx1} \qquad I_{\theta\theta} = E\left(-\frac{\partial^2 l}{\partial^2\theta} \right)$$

With the assumption of no covariate structure, the second derivatives of the beta-binomial log-likelihood l are obtained as

$$-\frac{\partial^2 l}{\partial\pi_j \partial\pi_s} = \left[\sum_{r=0}^{y_i-1} \frac{1}{(\pi_i + r\theta)^2} + \sum_{r=0}^{n_i-y_i-1} \frac{1}{1-\pi_i + r\theta^{\ 2}} \right] \qquad j,\ s=1,2,\ldots,k$$

$$-\frac{\partial^2 l}{\partial\pi_j \partial\theta} = \left[\sum_{r=0}^{y_i-1} \frac{r}{(\ 1-\phi\ \ \pi_i + r\phi)^2} - \sum_{r=0}^{n_i-y_i-1} \frac{r}{(1-\phi_i)(1-\pi_i) + r\theta^{\ 2}} \right] \qquad j=1,\ 2,\ldots,k$$

$$-\frac{\partial^2 l}{\partial\phi^2} = \sum_{i=1}^{m} \left[\sum_{r=0}^{y_i-1} -\frac{(r-\pi_i)^2}{((1-\phi)\pi_i + r\phi)^2} + \sum_{r=0}^{n_i-y_i-1} \frac{[r-(1-\pi_i)]^2}{(1-\phi)(1-\pi_i) + r\phi^{\ 2}} \right] - \sum_{i=1}^{m} \sum_{r=0}^{n_i-1} \frac{(r-1)^2}{((1-\phi) + r\phi)^2}$$

Entries for the fisher information matrix are obtained as

$$E\left(-\frac{\partial^2 l}{\partial\pi_j \partial\pi_S} \right) = \sum_{i=1}^{m} \left[\sum_{r=0}^{y_i-1} \frac{P\ \ Y_i > r}{(\pi_i + r\theta)^2} + \sum_{r=0}^{n_i-1} \frac{P\ \ Y_i > n_i - r}{1-\pi_i + r\theta^{\ 2}} \right]$$

$$E\left(-\frac{\partial^2 l}{\partial\pi_j \partial\phi} \right) = \sum_{i=1}^{m} \left[\frac{1}{\phi} \left(\sum_{r=0}^{n_i-1} \frac{-(1-\phi)\pi_i P\ \ Y_i > r}{((1-\phi)\pi_i + r\theta)^2} + \sum_{r=0}^{n_i-1} \frac{1-\pi_i\ \ (1-\phi)P\ \ Y_i > n_i - r}{(1-\phi)(1-\pi_i) + r\phi} \right) \right]$$

$$E\left(-\frac{\partial^2 l}{\partial\phi^2} \right) = \sum_{i=1}^{m} \frac{1}{\phi^2} \left[\sum_{r=0}^{y_i-1} \frac{[(1-\phi)\pi_i]^2 P\ \ Y_i > r}{((1-\phi)\pi_i + r\phi)^2} + \sum_{r=0}^{n_i-y_i-1} \frac{[(1-\phi)(1-\pi_i)]^2 P\ \ Y_i < n_i - r}{(1-\phi)(1-\pi_i) + r\phi^{\ 2}} + \sum_{r=0}^{n_i-1} \frac{(1-\phi)^2 1}{((1-\phi) + r\phi)^2} \right]$$

$\text{var}(\hat{\theta})$ is the corresponding diagonal element of $\left[\ \pi, \phi\ \right]^{-1}$

Then by the Delta method (Kendall and Stuart, 1986), the asymptotic variance of $\hat{\theta}$ is obtained as $\text{var}(\hat{\theta}) = \dfrac{\text{var}(\hat{\phi})}{1+\hat{\phi}^{\ 4}}$

Thus, the over-dispersion parameter is distributed as, $N\left(\hat{\phi}, \frac{\text{var}(\hat{\phi})}{1+\hat{\phi}^{4}}\right)$

Computations using these results are presented in chapter four.

3.1.2 Method of Moments

We give two estimating equations for the moment estimators of the over-dispersion parameter.

$$U_j(\beta_j,\phi)=\sum_{i=1}^{m} n_i(z_i-\pi_i)\partial_{ij}\beta=0 \quad \text{and}$$

$$U_{k+1}(\beta,\phi)=\sum_{i=1}^{m} n_i^2 \; z_i-\pi_i \;^2-n_i\pi_i(1-\pi_i)(1+(n_i-1)\phi) \;=0$$

We propose the adoption of the point estimation procedure using the weighted least squares method presented by Saha (2008). We then proceeded to find the expressions for the variance functions based on the Inagaki (1973) results.

The entries for the A $\hat{\lambda}$ and B $\hat{\lambda}$ are;

$$\text{A}_{j,s}=E\left(-\frac{\partial U_j}{\partial\beta_s}\right)=\sum_{i=1}^{m} n_i(1-2\pi_i)x_{ij}$$

$$\text{A}_{j,k+1}=E\left(-\frac{\partial U_j}{\partial\phi}\right)=0$$

$$\text{A}_{k+1,j}=E\left(-\frac{\partial U_{k+1}}{\partial\beta_j}\right)=\sum_{i=1}^{m} n_i \;\; 1-2\pi_i \quad 1+(n_i-1)\phi$$

$$\text{A}_{k+1,k+1}=E\left(-\frac{\partial U_{k+1}}{\partial\phi}\right)=\sum_{i=1}^{m} n_i(\pi_i(1-\pi_i)) \;\; n_i-1$$

$$B_{j,s}=E(U_jU_s)=\sum_{i=1}^{m} n_i\pi_i^3 \;\; 1-\pi_i \;_i^{\;3}(1+(n_i-1)\phi)x_{ij}x_{is}$$

$$\mathbf{B}_{k+1,j}=E(U_{k+1}U_j)=\sum_{i=1}^{m} n_i\pi_i \;\; 1-\pi_i \quad 1+ \;\; n_i-1 \;\; \phi \quad 1+ \;\; 2n_i-1 \;\; \phi \;\; \frac{1-2\pi_i}{1+\phi}\, x_{ij}$$

$$B_{k+1,k+1} = E(\mathrm{U}_{k+1}^2) = \sum_{i=1}^{m} n_i^4(\mu_4 - \mu_2^2)$$

Thus, variance estimates are found by using the variance covariance matrix;

$$\mathrm{var}(\hat{\lambda}_{MM}) = \begin{pmatrix} A_{js} & \underline{0} \\ A_{j,k+1} & A_{k+1,k+1} \end{pmatrix}^{-1} \begin{pmatrix} B_{j,s} & B_{k+1,j} \\ B_{j,k+1} & B_{k+1,k+1} \end{pmatrix} \left(\begin{pmatrix} A_{js} & \underline{0} \\ A_{j,k+1} & A_{k+1,k+1} \end{pmatrix}^{-1} \right)^T \text{ where}$$

$$A = \begin{pmatrix} A_{js} & \underline{0} \\ A_{j,k+1} & A_{k+1,k+1} \end{pmatrix}$$

$$\mathrm{B} = \begin{pmatrix} B_{j,s} & B_{k+1,j} \\ B_{j,k+1} & B_{k+1,k+1} \end{pmatrix}$$

3.2.3 Quasi-Likelihood method

Two estimating equations are involved,

$$U_j\ \beta,\phi\ = \frac{\partial Q}{\partial \beta_j} = \sum_{i=1}^{m} \frac{z_i - \pi_i\ n_i \partial_{ij}(\beta)}{\pi_i\ 1-\pi_i\ \ 1+\ n_i - 1\ \phi} = 0 \quad \text{and}$$

$$U_{k+1}\ \beta,\phi\ = \sum_{i=1}^{m} \frac{z_i - \pi_i\ ^2\, n_i}{\pi_i\ 1-\pi_i\ \ 1+\ n_i - 1\ \phi} - m - k = 0$$

We solve the two equations simultaneously and denote estimates of β and ϕ by $\hat{\lambda}_{Ql}$.

Based on these estimating equations, the Inagaki (1973) results give the expressions for the asymptotic variances of the quasi-likelihood estimate whose elements, A $\hat{\lambda}$ and $B\ \hat{\lambda}$ are components of the Hessian matrix. We derived these expressions and obtained:

$$\mathrm{A}_{j,s} = E\left(-\frac{\partial U_j}{\partial \beta_s} \right) = \sum_{i=1}^{m} \frac{n_i x_{ij}}{1+(n_i - 1)\phi}$$

$$\mathrm{A}_{j,k+1} = E\left(-\frac{\partial U_j}{\partial \phi} \right) = 0$$

$$\mathrm{A}_{k+1,j} = E\left(-\frac{\partial U_{k+1}}{\partial \beta_j}\right) = \sum_{i=1}^{m} \frac{1-2\pi_i \;\; (1+(n_i-1)\phi)}{n_i}\left[w_i(1-{}^{w_i}\!/_{w}) - \frac{1}{\mu_{2i}} \right]$$

$$\mathrm{A}_{k+1,k+1} = E\left(-\frac{\partial U_{k+1}}{\partial \phi}\right) = \sum_{i=1}^{m}\left\{\left\{\sum_{i=1}^{m}\frac{w_i}{n_i}\left(1-\frac{w_i}{w}\right)\right\}\pi_i(1-\pi_i)(n_i-1) - \frac{n_i-1}{1+(n_i-1)\phi}\right\}$$

$$B_{j,s} = E(U_j U_s) = \sum_{i=1}^{m} \frac{n_i\pi_i \;\; 1-\pi_i \;\; x_{ij}x_{is}}{(1+(n_i-1)\phi)}$$

$$B_{k+1,j} = B_{j,k+1} = E(U_{k+1}U_j) = \sum_{i=1}^{m} \frac{1-2\pi_i \;\; 1+ \;\; 2n_i-1 \;\; \phi}{n_i \;\; 1+\phi}$$

$$B_{k+1,k+1} = E(\mathrm{U}_{k+1}^2) = \sum_{i=1}^{m}\left\{\frac{\mu_4}{\mu_2^2} - 2\left(\sum_{i=1}^{m} w_i(1-\frac{w_i}{w})\right)\mu_2 n_i + \left(\sum_{i=1}^{m} w_i(1-\frac{w_i}{w})\right)^2 \mu_2^2 n_i^2\right\}$$

The entries of matrices A and B are;

$$\mathrm{A}=\begin{pmatrix} A & \underline{0} \\ a' & a_{k+1} \end{pmatrix} \qquad\qquad \mathrm{B}=\begin{pmatrix} A & b \\ b' & b_{k+1} \end{pmatrix}$$

Thus, substituting the above matrices into the asymptotic variance equation yields the expression below: $\mathrm{var}(\hat{\lambda}_{Qm}) = \begin{pmatrix} A & \underline{0} \\ a' & a_{k+1} \end{pmatrix}^{-1}\begin{pmatrix} A & b \\ b' & b_{k+1} \end{pmatrix}\left(\begin{pmatrix} A & \underline{0} \\ a' & a_{k+1} \end{pmatrix}^{-1}\right)^T$

The A's and B's in this case are the first and second derivatives of the estimating equations respectively.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, we display our findings on coverage probability estimates and the widths of bootstrap confidence intervals (in parenthesis). A detailed discussion of the results is given based on observation on the tables displayed for MLE, MME and Quasi-likelihood estimators. Tables (1), (2) and (3) display results for MLE, MME and Quasi-likelihood procedures respectively. The displayed graphs (1) and (2) display the widths of CI for data that is not bootstrapped and when bootstrapping technique is applied respectively. Moreover, results of point and interval estimation for the over-dispersion parameter using the Potthoff and Whittinghill (1966) data are displayed in table 4. CI widths for the same are displayed in graph (3).

For various combinations of the over-dispersion and mean parameters, the procedures displayed in the appendix are carried out. The results displayed can be replicated by running the subroutines displayed in the appendix.

4.2 Coverage probability estimates

Tables 1-3 show coverage probabilities for the over-dispersion parameter based on 1000 samples for the three procedures under investigation. It appears from the simulation results that the asymptotic CIs based on these three estimators have coverage below the nominal coverage probability. The resulting confidence interval, with nominal coverage probability 0.95, has low coverage probabilities. This shows that these confidence intervals are completely inadequate.

Generally, the coverage probabilities decrease with increase in the over-dispersion parameter estimate. Optimization was difficult for bootstrapped successes owing to the small number of successes selected. This led to a high number of invalid samples especially for small cluster sizes. This brought about the need to obtain several bootstrap simulates that avoided repetitive successes that were beyond three. i.e., if one of the successes obtained was repeated three or more times, this was discarded and a new bootstrap procedure performed. In each table we present coverage probabilities alongside the lengths of confidence intervals in parenthesis.

Examining Table 1, the following was observed based on the estimated coverage probabilities and the lengths of the bootstrap CI; MLE procedure is known for the biasedness of the over-dispersion parameter estimates. This estimate had small deviations from the ones based on MME and EQL procedures. Despite this, the differences in the lengths of the confidence intervals for different simulation studies were more or less the same.

Computation time was the greatest drawback when producing simulation results in Table 1. This led to less number of simulations performed. Re-estimation of coverage probabilities based on valid estimates of π and ϕ using MLE for each bootstrap sample is time consuming. To save on the computation and estimation time, simulation was performed without re-estimation of ϕ and π, on production of a first valid estimate of the parameters. Generally, the length of time that was spent by the program to estimate the coverage probabilities was higher than time spent for MME and EQL. This was much dependent of the cluster size. Large cluster sizes needed less time to estimate the coverage probabilities.

Out of the total 180 cases the percentage of samples that were close to the nominal level (95 %) was around 60%. These samples are largely concentrated to this interval when our cluster size was 40 or greater than 40 ($K \geq 40$), ϕ was less than 0.4 and π = [0.3, 0.4]. A cluster size of 20 worked well for small values of ϕ and the worst situations happened at ϕ = 0.6 for all cluster sizes. A feature that was noticed here is that more time is used for estimation of coverage probabilities when we use small clusters. The time that is used for both simulation and estimation is much reduced for cases when the cluster sizes are large. This is attributed to the fact that the probability of producing invalid samples reduces with an increase in cluster size.

Examining the average length of the bootstrap CI, we found that for all conditions fixed, the average length of CI tends to decrease as either the cluster size increases or when the mean parameter increases.

TABLE 1

MAXIMUM LIKELIHOOD METHOD

Percentage coverage probabilities and widths of bootstrap CIs (in parenthesis)for several combinations of:

Cluster sizes; 20, 30, 40, 60, 100 and 200, π = 0.1, 0.2, 0.3, 0.4, 0.6 and 0.7, ϕ = 0.1, 0.2, 0.3, 0.4 and 0.6.

π	0.1					0.2					0.3				
ϕ K	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20	93.3 (1.07)	94.5 (1.05)	97.5 (1.01)	98.1 (0.91)	53.4 (0.79)	93.9 (0.94)	93.0 (0.94)	90.1 (0.89)	86.9 (0.75)	56.5 (0.64)	95.6 (0.87)	93.4 (0.87)	87.1 (0.80)	73.2 (0.60)	52.3 (0.49)
30	92.6 (0.67)	94.5 (0.96)	91.3 (0.89)	87.6 (0.81)	53.1 (0.67)	94.4 (0.78)	92.3 (0.62)	87.9 (0.64)	79.9 (0.59)	54.9 (0.45)	95.7 (0.78)	92.0 (0.61)	89.8 (0.56)	74.3 (0.49)	46.9 (0.34)
40	97.6 (0.84)	94.5 (0.84)	90.6 (0.81)	81.8 (0.70)	58.4 (0.63)	95.8 (0.66)	93.8 (0.67)	87.6 (0.65)	72.9 (0.51)	59.9 (0.49)	91.9 (0.62)	92.1 (0.63)	90.0 (0.59)	74.9 (0.45)	56.7 (0.32)
60	97.7 (0.63)	94.8 (0.59)	89.8 (0.56)	77.7 (0.52)	53.5 (0.50)	96.9 (0.55)	93.1 (0.44)	92.1 (0.46)	71.4 (0.39)	56.7 (0.37)	93.9 (0.56)	92.9 (0.42)	85.8 (0.41)	72.2 (0.34)	49.9 (0.37)
100	97.8 (0.49)	94.4 (0.53)	86.1 (0.54)	72.0 (0.43)	53.4 (0.42)	90.9 (0.42)	93.1 (0.44)	93.2 (0.44)	78.9 (0.34)	56.3 (0.38)	93.9 (0.40)	94.0 (0.40)	94.3 (0.39)	70.9 (0.26)	50.6 (0.32)
200	90.9 (0.34)	92.0 (0.38)	89.9 (0.40)	72.4 (0.30)	56.2 (0.28)	94.7 (0.30)	93.9 (0.31)	94.9 (0.31)	78.1 (0.23)	56.0 (0.31)	96.0 (0.28)	94.5 (0.29)	96.0 (0.27)	74.9 (0.19)	58.2 (0.20)

Cont'- Table 1

π	0.4					0.6					0.7				
ϕ K	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20	92.9 (0.85)	93.1 (0.84)	87.3 (0.75)	70.1 (0.55)	39.8 (0.46)	91.9 (0.85)	98.0 (0.86)	92.1 (0.61)	78.6 (0.61)	58.6 (0.60)	94.1 (0.87)	92.4 (0.87)	92.1 (0.83)	84.3 (0.69)	60.1 (0.60)
30	93.4 (0.76)	92.9 (0.69)	90.1 (0.57)	71.1 (0.46)	44.3 (0.41)	90.8 (0.60)	96.3 (0.58)	92.6 (0.52)	88.5 (0.73)	64.2 (0.54)	93.2 (0.66)	92.6 (0.62)	92.4 (0.69)	87.8 (0.76)	58.9 (0.39)
40	94.9 (0.61)	93.0 (0.61)	92.3 (0.56)	74.3 (0.41)	50.6 (0.41)	94.8 (0.61)	93.8 (0.61)	92.9 (0.57)	85.9 (0.46)	54.9 (0.41)	92.9 (0.62)	92.1 (0.63)	91.7 (0.61)	90.0 (0.52)	58.0 (0.47)
60	96.7 (0.57)	91.5 (0.54)	91.5 (0.43)	70.6 (0.42)	45.9 (0.38)	96.5 (0.48)	93.4 (0.46)	89.9 (0.45)	86.0 (0.49)	45.1 (0.47)	97.2 (0.48)	91.9 (0.58)	92.9 (0.51)	90.1 (0.60)	54.6 (0.53)
100	96.8 (0.39)	94.4 (0.39)	95.6 (0.36)	76.7 (0.26)	53.2 (0.25)	96.0 (0.39)	94.0 (0.39)	94.9 (0.37)	85.4 (0.28)	53.1 (0.29)	93.9 (0.40)	94.0 (0.40)	94.3 (0.39)	93.9 (0.34)	60.2 (0.31)
200	94.1 (0.28)	95.0 (0.27)	93.9 (0.26)	81.4 (0.19)	54.7 (0.21)	94.7 (0.28)	95.0 (0.27)	94.9 (0.26)	90.1 (0.21)	50.1 (0.24)	95.0 (0.28)	95.5 (0.29)	94.3 (0.27)	94.9 (0.24)	62.1 (0.23)

TABLE 2

METHOD OF MOMENTS

Percentage coverage probabilities and widths of bootstrap CIs (in parenthesis) for several combinations of:

Cluster sizes: 20, 30, 40, 60, 100 and 200. π = 0.1, 0.2, 0.3, 0.4, 0.6 and 0.7, ϕ = 0.1, 0.2, 0.3, 0.4 and 0.6.

π		0.1					0.2					0.3				
K	ϕ	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20		96.0 (0.45)	91.7 (0.52)	80.8 (0.57)	68.7 (0.68)	21.3 (0.60)	97.3 (0.58)	91.4 (0.70)	83.8 (0.79)	66.2 (0.56)	23.9 (0.59)	96.9 (0.58)	91.1 (0.65)	81.0 (0.70)	65.8 (0.52)	22.1 (0.56)
30		96.5 (0.46)	90.7 (0.54)	81.8 (0.54)	64.5 (0.58)	19.8 (0.62)	97.4 (0.56)	90.5 (0.58)	83.5 (0.53)	64.5 (0.52)	18.5 (0.58)	97.1 (0.51)	91.1 (0.51)	82.7 (0.60)	63.1 (0.57)	19.5 (0.55)
40		96.8 (0.65)	91.9 (0.57)	79.6 (0.56)	64.2 (0.80)	16.4 (0.51)	98.0 (0.60)	91.7 (0.55)	80.4 (0.55)	61.1 (0.57)	15.4 (0.65)	97.6 (0.63)	92.7 (0.53)	79.5 (0.53)	61.8 (0.61)	18.8 (0.56)
60		96.8 (0.37)	91.3 (0.26)	79.3 (0.65)	56.3 (0.60)	15.5 (0.58)	97.7 (0.43)	89.3 (0.52)	79.4 (0.59)	59.4 (0.67)	16.2 (0.65)	97.5 (0.51)	90.9 (0.52)	78.4 (0.45)	60.8 (0.53)	12.3 (0.50)
100		97.5 (0.27)	91.4 (0.40)	79.9 (0.48)	56.2 (0.47)	14.3 (0.43)	96.6 (0.40)	92.3 (0.54)	78.4 (0.57)	57.5 (0.46)	14.0 (0.46)	97.7 (0.69)	91.0 (0.46)	78.5 (0.54)	58.6 (0.49)	11.9 (0.49)
200		97.1 (0.38)	91.3 (0.49)	78.2 (0.50)	52.9 (0.48)	9.9 (0.53)	96.7 (0.36)	91.3 (0.50)	79.2 (0.68)	57.0 (0.43)	9.7 (0.44)	97.7 (0.54)	91.0 (0.55)	78.7 (0.42)	56.5 (0.66)	12.6 (0.42)

cont'-table 2

π		0.4					0.6					0.7				
K	ϕ	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20		96.8 (0.61)	93.6 (0.57)	79.4 (0.59)	68.9 (0.54)	22.5 (0.60)	95.6 (0.59)	92.3 (0.61)	81.5 (0.65)	67.2 (0.56)	24.4 (0.56)	97.8 (0.63)	91.2 (0.57)	82.2 (0.57)	66.8 (0.62)	23.7 (0.70)
30		97.1 (0.56)	91.4 (0.61)	79.5 (0.71)	65.9 (0.60)	20.0 (0.57)	97.4 (0.69)	90.6 (0.54)	81.8 (0.65)	60.8 (0.60)	19.3 (0.58)	97.6 (0.52)	90.5 (0.73)	80.0 (0.54)	64.2 (0.66)	18.6 (0.62)
40		97.7 (0.54)	92.2 (0.56)	80.3 (0.60)	63.5 (0.55)	14.5 (0.51)	98.0 (0.54)	90.6 (0.54)	79.5 (0.49)	63.6 (0.63)	18.8 (0.49)	97.9 (0.55)	90.3 (0.52)	79.3 (0.62)	61.6 (0.61)	17.9 (0.67)
60		97.6 (0.44)	91.2 (0.55)	76.3 (0.53)	58.3 (0.57)	12.1 (0.44)	98.4 (0.60)	91.3 (0.51)	79.7 (0.48)	60.7 (0.45)	12.5 (0.53)	97.7 (0.58)	92.1 (0.67)	78.5 (0.55)	57.7 (0.53)	16.3 (0.48)
100		98.0 (0.49)	89.9 (0.50)	76.1 (0.66)	61.5 (0.46)	13.1 (0.55)	97.2 (0.49)	89.2 (0.56)	78.1 (0.52)	57.8 (0.39)	12.3 (0.49)	97.9 (0.44)	90.4 (0.63)	77.9 (0.46)	60.5 (0.42)	14.5 (0.58)
200		96.8 (0.61)	90.6 (0.50)	77.0 (0.46)	54.9 (0.57)	10.9 (0.37)	97.4 (0.48)	90.5 (0.63)	78.7 (0.50)	56.3 (0.51)	12.2 (0.52)	97.0 (0.42)	92.4 (0.47)	75.0 (0.38)	55.7 (0.42)	9.9 (0.48)

Table 2 shows observed coverage probability estimates for the over-dispersion parameter based on the method of moments.

Generally, coverage probabilities are good for over-dispersion parameters 0.1, 0.2, 0.3 and 0.4. These coverage probabilities reduce as over-dispersion increases. Higher over-dispersion parameters perform poorly in terms of coverage probabilities. For data simulated from the BB distribution, lower coverage probabilities are attained as the over dispersion parameter tends to 1. The percentage of invalid samples are high for small cluster sizes and reduced to near zero percent as the cluster sizes increased beyond a sample size of 100. The mean parameter lies in the interval [0.3, 0.7]. The over-dispersion parameter lied in the interval [0.05, 0.35]. Generally, good coverage probabilities are attained for the over-dispersion parameter between 0.1 and 0.2. The bootstrap confidence interval lengths reduce as the mean parameter estimate increases. In contrary to small cluster sizes, the lengths of bootstrap confidence intervals are generally larger than the lengths for small cluster sizes. In our case, when we have five successes, the lengths of the bootstrap CIs are smaller. Optimal coverage probabilities are obtained when the cluster size is 40.

In the study, parameter estimates were seen to be consistent and less biased as compared to the case of MLE procedure. The difference in the lengths of confidence intervals are seen to be smaller for all the combinations studied.

The length of time used by the program for simulation was approximately 10 times shorter than the time spent by the MLE procedure. This was much dependent on the cluster size. The larger the cluster size, the lesser the time needed for the estimation of coverage probabilities.

TABLE 3

QUASI-LIKELIHOOD

Percentage coverage probabilities and widths of bootstrap CIs (in parenthesis) for several combinations of:

Cluster sizes; 20, 30, 40, 60, 100 and 200. π = 0.1, 0.2, 0.3, 0.4, 0.6 and 0.7, ϕ = 0.1, 0.2, 0.3, 0.4 and 0.6.

π		0.1					0.2					0.3				
K	ϕ	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20		97.8 (0.64)	92.1 (0.65)	82.6 (0.63)	69.0 (0.61)	26.4 (0.58)	97.6 (0.59)	91.7 (0.71)	81.8 (0.59)	70.2 (0.59)	27.5 (0.60)	97.3 (0.60)	92.6 (0.66)	82.5 (0.63)	67.2 (0.62)	25.8 (0.57)
30		97.2 (0.65)	91.5 (0.53)	82.8 (0.53)	63.6 (0.52)	23.1 (0.61)	97.0 (0.52)	92.3 (0.59)	79.8 (0.59)	66.3 (0.52)	20.9 (0.63)	97.7 (0.62)	92.5 (0.55)	80.1 (0.56)	65.4 (0.57)	23.7 (0.54)
40		97.5 (0.57)	90.8 (0.50)	80.0 (0.50)	62.5 (0.50)	20.2 (0.58)	98.0 (0.53)	91.2 (0.73)	81.2 (0.54)	64.7 (0.47)	18.9 (0.51)	97.8 (0.56)	92.4 (0.56)	79.7 (0.51)	63.6 (0.48)	19.5 (0.50)
60		97.8 (0.54)	91.6 (0.57)	79.0 (0.49)	61.6 (0.51)	16.5 (0.58)	96.5 (0.45)	91.7 (0.57)	79.8 (0.55)	62.4 (0.52)	16.7 (0.45)	97.3 (0.48)	92.5 (0.44)	78.6 (0.49)	63.5 (0.52)	16.0 (0.52)
100		97.0 (0.41)	92.0 (0.43)	74.9 (0.38)	61.2 (0.52)	15.7 (0.45)	98.0 (0.45)	92.3 (0.49)	78.6 (0.43)	61.7 (0.43)	17.3 (0.50)	97.6 (0.52)	90.6 (0.53)	76.3 (0.46)	59.6 (0.45)	15.4 (0.48)
200		97.6 (0.43)	90.9 (0.46)	80.6 (0.40)	59.6 (0.46)	14.4 (0.41)	97.0 (0.46)	91.0 (0.53)	79.6 (0.45)	58.5 (0.54)	12.3 (0.38)	97.8 (0.56)	93.2 (0.53)	76.9 (0.44)	58.3 (0.48)	14.8 (0.38)

Cont'-Table 3

π		0.4					0.6					0.7				
K	ϕ	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6	0.1	0.2	0.3	0.4	0.6
20		97.9 (0.58)	90.8 (0.57)	80.8 (0.63)	64.2 (0.57)	27.1 (0.56)	97.1 (0.60)	93.4 (0.56)	84.1 (0.56)	66.8 (0.60)	25.7 (0.59)	98.2 (0.71)	92.7 (0.63)	82.2 (0.59)	68.6 (0.58)	26.6 (0.70)
30		98.0 (0.58)	92.7 (0.58)	81.6 (0.57)	63.9 (0.54)	24.3 (0.56)	97.2 (0.54)	92.2 (0.60)	82.1 (0.51)	65.1 (0.55)	21.5 (0.54)	97.2 (0.62)	91.9 (0.56)	83.5 (0.57)	67.2 (0.61)	22.4 (0.56)
40		97.9 (0.65)	92.2 (0.50)	81.5 (0.56)	65.7 (0.54)	18.1 (0.53)	97.2 (0.58)	92.4 (0.51)	80.7 (0.51)	64.2 (0.48)	20.4 (0.52)	97.7 (0.59)	92.7 (0.72)	80.6 (0.58)	64.7 (0.53)	21.3 (0.47)
60		98.0 (0.59)	91.5 (0.57)	78.6 (0.52)	60.6 (0.47)	16.5 (0.47)	97.7 (0.58)	90.5 (0.53)	79.3 (0.47)	62.0 (0.45)	16.8 (0.47)	96.3 (0.48)	93.8 (0.53)	78.3 (0.50)	60.1 (0.51)	16.2 (0.46)
100		98.2 (0.49)	91.7 (0.43)	76.3 (0.46)	59.2 (0.47)	16.1 (0.47)	97.1 (0.67)	94.0 (0.53)	76.7 (0.44)	59.2 (0.43)	17.0 (0.41)	98.2 (0.53)	91.5 (0.44)	78.8 (0.44)	60.6 (0.45)	15.9 (0.45)
200		98.1 (0.48)	90.8 (0.47)	76.9 (0.44)	61.4 (0.43)	12.6 (0.43)	97.2 (0.57)	92.5 (0.48)	79.0 (0.41)	57.5 (0.45)	13.7 (0.45)	98.1 (0.48)	90.8 (0.48)	76.8 (0.48)	59.3 (0.48)	14.9 (0.46)

Displayed on Table 3 are coverage probability estimates based on the quasi-likelihood procedure.

Similar to MME procedure, data was simulated from the beta-binomial distribution. EQL coverage probabilities are good for over-dispersion parameters 0.1, 0.2 and 0.3. These coverage probabilities also reduce with an increase in the over-dispersion parameter estimate. When the over-dispersion parameter is large, the CIs perform poorly in their coverage probabilities. The percentage of invalid samples is also high for small cluster sizes and reduces to near zero when the cluster sizes increased beyond a sample size of 100. The mean parameter lies in the interval [0.3, 0.7]. The over-dispersion parameter lied in the interval [0.05, 0.35]. Generally for the mean parameter values 0.1, 0.2 and 0.3, good coverage probabilities are attained for the over-dispersion parameter values, 0.1 and 0.2. The bootstrap confidence intervals are reduced for the above mean intervals.

Contrary to small cluster sizes, the lengths of bootstrap confidence intervals are generally larger than the lengths for small cluster sizes. In our case, when we have five successes, the lengths of the bootstrap CIs are smaller. Optimal coverage probabilities are obtained when the cluster size is 40.

Parameter estimates were also seen to be less biased like in the case of MME procedure. The lengths of coverage probabilities differ slightly with the case of MME but greatly with the case of MLE.

Computation time was less of a problem in this case. This time was equal irrespective of the cluster size.

4.3 Bootstrap confidence intervals

When we use equal sample size groups to estimate the bootstrap confidence intervals, the shift in the coverage probabilities is the same for the three procedures when different resamples are used. When different resampling is done for the same data set, the difference between the lengths of the CIs for the three procedures above is not significant. The lengths of the MLE confidence interval are generally less than the CI s for the MMEs and EQL procedures. Moreover, a shift to the lower or upper side is simultaneous for all resampling procedures. For equal cluster sizes, the lengths of bootstrap CI do not differ much. This is observed in table (1). There is also simultaneous increase and decrease in the lengths of CIs.

Based on the fig (3), it is shown that for unequal sample sized groups, bootstrap confidence intervals are generally shorter in MLEs as compared to the other estimators.

Confidence interval lengths for MME and EQL procedures are shorter than in the case of equal sample size clusters (see fig (2)). This is much dependent on the sample selected after the bootstrap resampling is done. Generally, CIs of MLEs are much shorter while CIs of MMEs are longer in most bootstrapped samples.

Based on the example given by Paul (2009), Potthoff and Whittinghill (1966) data on Example 6.5, it was observed that parameter estimates were biased for MLE and unbiased for MME and QLE procedures. Applying bootstrap technique reduces the biasness of these estimates.

Subsequent bootstrapping on the same data set tends to produce small differences in the lengths of the obtained bootstrap confidence interval. Not all these subsequent bootstrap CI are the same since the samples picked may be different. The bootstrap CI is dependent on the bootstrap sample picked. For a specified sample picked, an increase or decrease in the length of confidence intervals is the same for all the three procedures above.

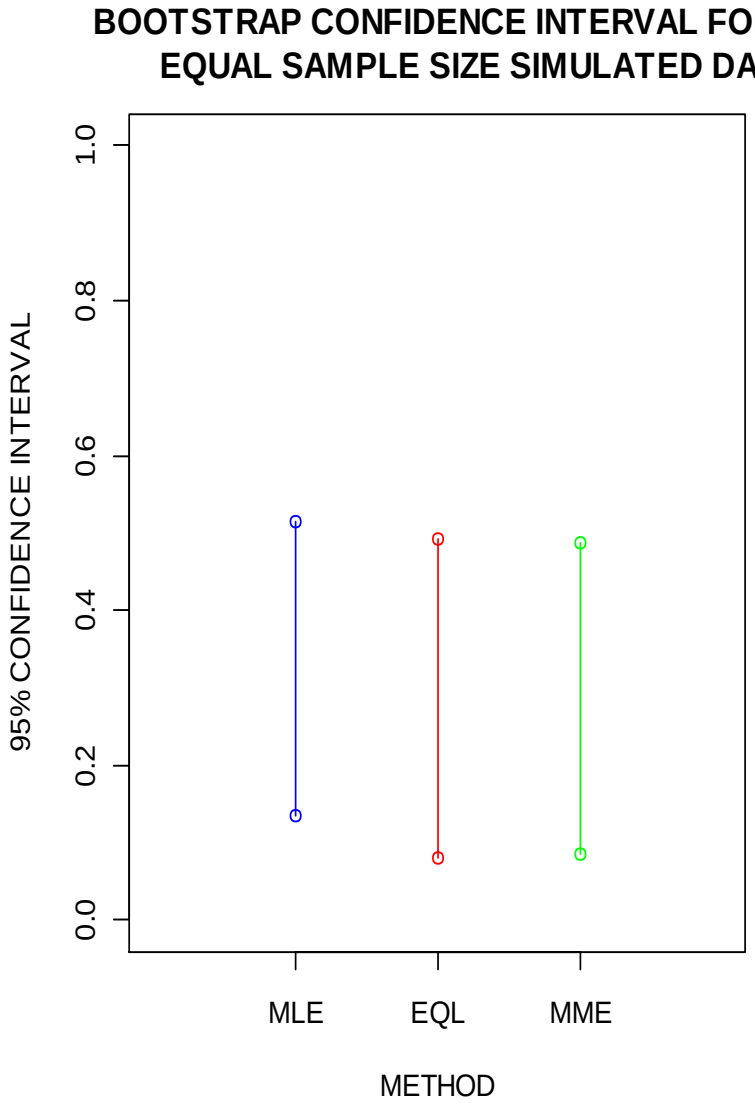
When there is no bootstrapping done on the data set, the confidence interval widths for the three procedures above are large. It therefore surfaces that when bootstrap technique is applied to over-dispersed data in proportions, the confidence interval length generally reduces with the increase in the over-dispersion parameter.

Example (Table 4)

y	7	1	4	3	5	3	0	11	3	0	10	3	0	4	2	2	3	5	2	1	2	3	1	1	4	5	3	3	5	1	1	3	4	0	1	2
n1	11	1	6	7	8	6	2	19	4	2	15	6	6	10	8	4	5	6	6	4	12	8	4	5	5	6	4	10	8	11	4	4	4	2	2	3

The data set above was used by Sudhir (2009), (Table 6.5), in the study of efficiency of parameter estimates based on, MME, EQL, MLE, DEQL and GL. It shows the number of cross-over off-springs in M=36families from Potthoff and Whittinghill (1966). y= the number of ++ off-springs, n=total cross-over off-springs.

	Data type	**Ordinary CI**					**Bootstrapped CI**				
	Method	**parameter**		**Confidence interval of ϕ**		**Length of CI**	**Average Parameter estimates**		**Bootstrap Confidence interval ϕ**		**Length of Bootstrap CI**
Data set		ϕ	π	**Lower**	**Upper**		ϕ	π	**lower**	**upper**	
One	MLE	0.0949	0.4728	-0.0136	0.2035	0.217	0.07949	0.45801	-0.0247	0.1837	0.2085
	EQL	0.0915	0.4742	-0.0271	0.2101	0.238	0.07963	0.4504	-0.03166	0.1909	0.2262
	MME	0.0915	0.4743	-0.0272	0.2102	0.237	0.07963	0.4505	-0.03047	0.1898	0.2202


BOOTSTRAP CONFIDENCE INTERVAL FOR
EQUAL SAMPLE SIZE SIMULATED DATA
95% CONFIDENCE INTERVAL
0.0
0.2
0.4
0.6
0.8
1.0
MLE
EQL
MME
METHOD

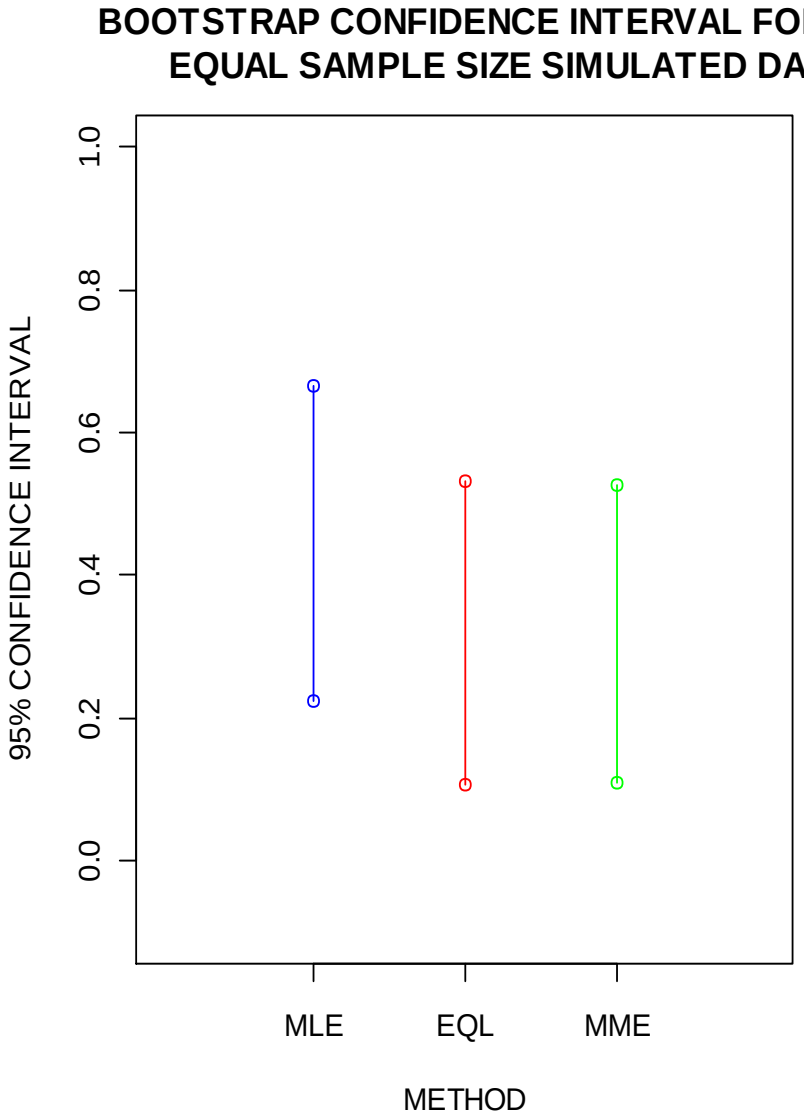

BOOTSTRAP CONFIDENCE INTERVAL FOR
EQUAL SAMPLE SIZE SIMULATED DATA
95% CONFIDENCE INTERVAL
0.0
0.2
0.4
0.6
0.8
1.0
MLE
EQL
MME
METHOD

Figure 1

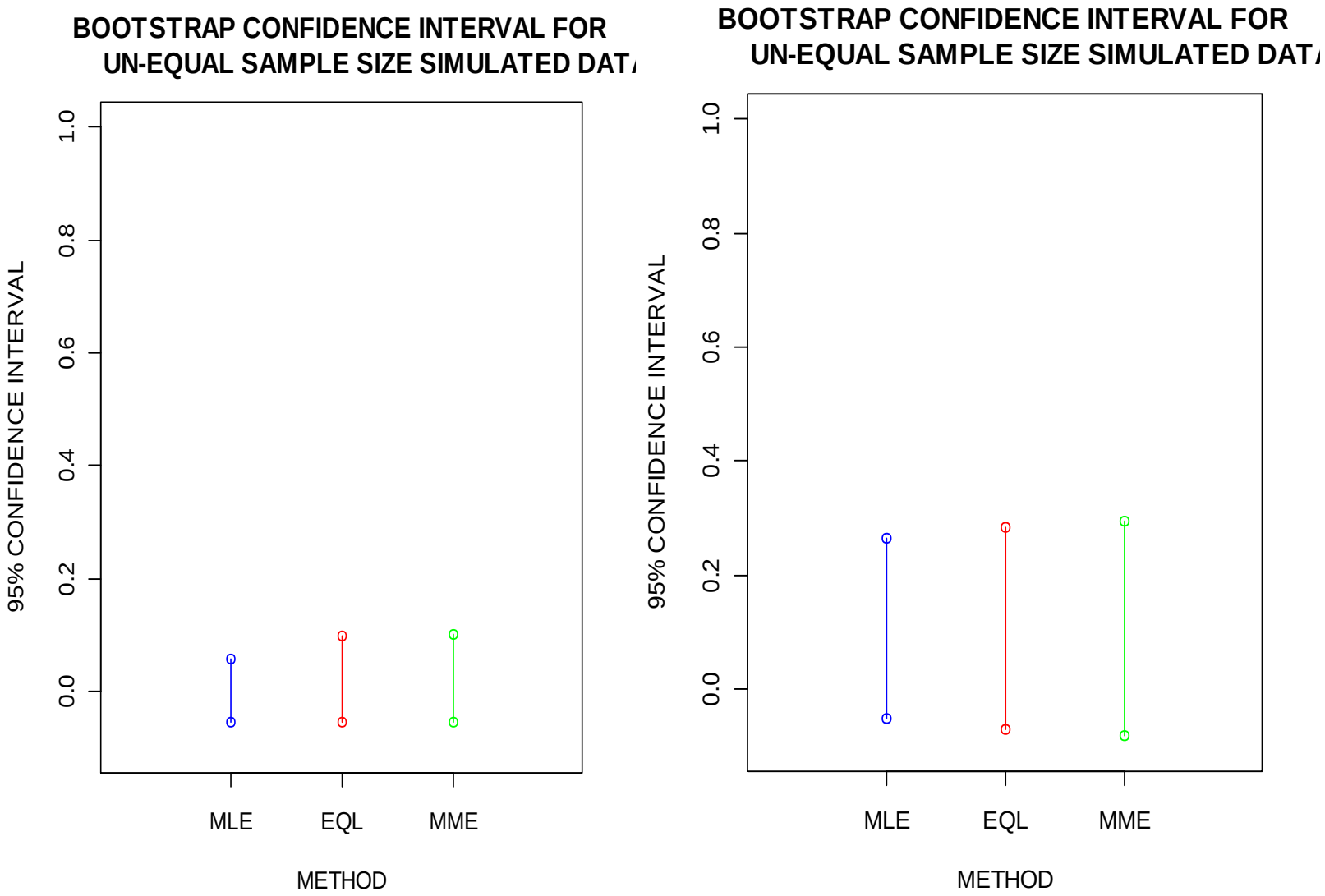
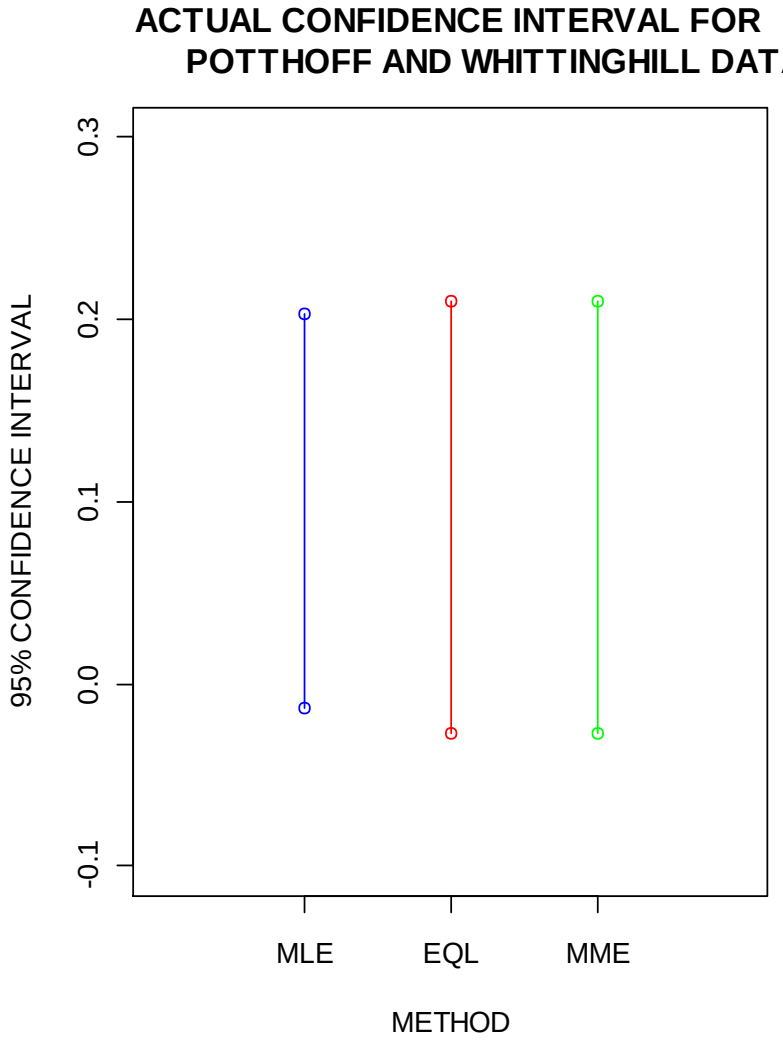

BOOTSTRAP CONFIDENCE INTERVAL FOR
UN-EQUAL SAMPLE SIZE SIMULATED DAT
95% CONFIDENCE INTERVAL
0.0
0.2
0.4
0.6
0.8
1.0
MLE
EQL
MME
METHOD
BOOTSTRAP CONFIDENCE INTERVAL FOR
UN-EQUAL SAMPLE SIZE SIMULATED DAT
95% CONFIDENCE INTERVAL
0.0
0.2
0.4
0.6
0.8
1.0
MLE
EQL
MME
METHOD

Figure 2


ACTUAL CONFIDENCE INTERVAL FOR
POTTHOFF AND WHITTINGHILL DATA
95% CONFIDENCE INTERVAL
0.3
0.2
0.1
0.0
-0.1
MLE
EQL
MME
METHOD

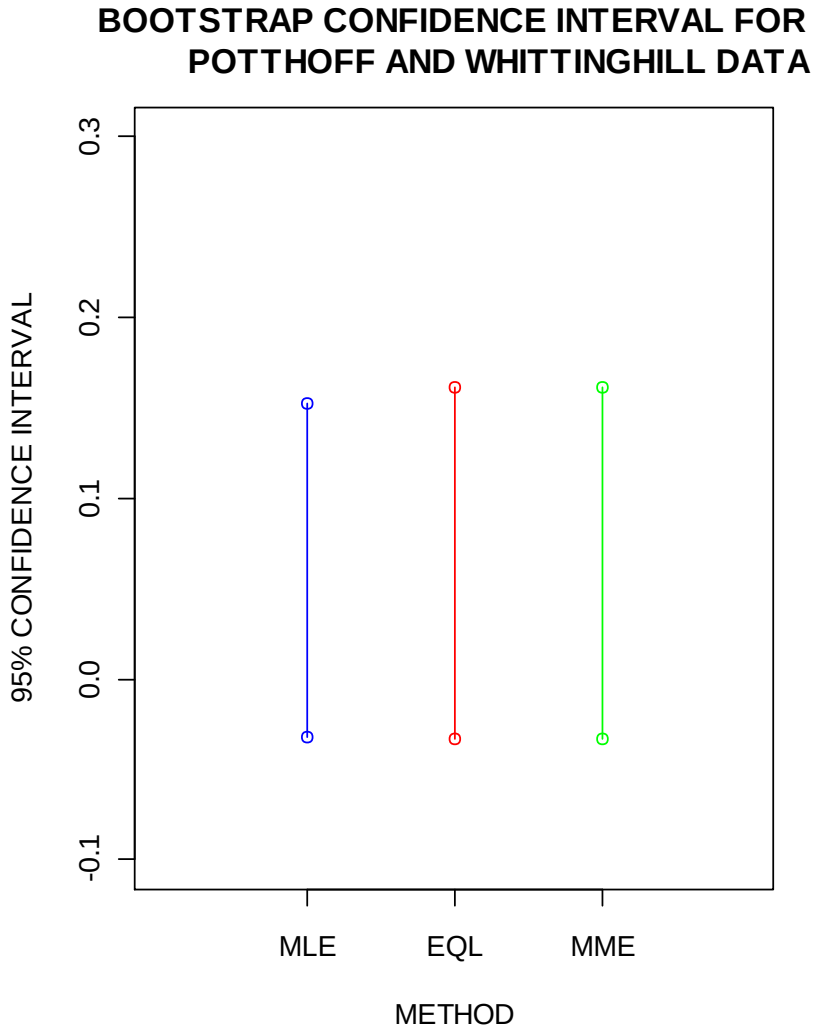

BOOTSTRAP CONFIDENCE INTERVAL FOR
POTTHOFF AND WHITTINGHILL DATA
95% CONFIDENCE INTERVAL
0.3
0.2
0.1
0.0
-0.1
MLE
EQL
MME
METHOD

Figure 3

CHAPTER FIVE
SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Introduction

This section gives the summary of the findings of research, the limitations of the study, recommendations for further study and areas of application of study based on the results.

5.2 Summary and Conclusion

In this work, we have derived the variance functions based on MLE and the estimating equations for both MME and MLE procedures. We examined the performance of the confidence intervals using Monte' Carlo simulation technique with bootstrapping and making observations on the widths of CIs obtained after bootstrapping. Results show that high coverage probabilities are obtained when a population has large clusters and when the over-dispersion parameter is small. The over-dispersion parameter in this case is less than an approximate of 0.4 for optimal results. Large samples should be used in order to make bootstrapping effective and efficient. Bootstrap procedure improves the coverage probabilities of confidence intervals and reduces the width of Confidence Intervals.

From the simulation study, we found out that as the cluster size increases, the chance of obtaining valid samples increases. Based on this fact, we recommend that the sampled cluster size should be at least 40. In addition to this, the over-dispersion parameter should be at-most approximately 0.4 with mean parameter lying within the interval [0.3, 0.7].

The limitation of this study is that it is restricted to the number of successes n_i =5. For effective bootstrapping, the number of successes should be at least ten to avoid repetitive samples that would yield high number of invalid samples. Furthermore, we have assumed that the over-dispersion parameter is constant which is not usually the case.

5.3 Further Research

In this study, much has been done for the case of a constant over-dispersion parameter. Further research recommends that the over-dispersion parameter should not be fixed since not all clusters are homogeneous.

From the results in tables (1), (2) and (3), we observe that the CIs are good since the widths of bootstrap CIs are shorter with higher coverage probabilities when the over-dispersion parameter is small. Though the CIs are good, they are inadequate since most coverage probabilities are below the nominal level (95 %). There is need to improve the coverage probabilities by using profile likelihood which involves the reduction of the effect of the nuisance parameter.

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 probabilities 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 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

Crowder, M .J. (1987). On linear quadratic estimating functions, *Biometrika,* 74: 591-597

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions, *applied statistics*, 27: 43-37.

Crowder, M. J. (1985). Gaussian estimation for correlated binomial data, *Journal of the Royal Statistical Society*, Series B, 47, 229–237.

Davidian, M., and Carrol, R. J (1987). Variance function estimation*; journal of the American statistical association.*82 1079-1091

Godambe, V. P., and Thompson, M. E. (1989). An extension of Quasi-likelihood estimation, J*ournal of statistical planning and inference*, 22: 137-152.

Haseman, J. K., and Kupper, L. L. (1979). "Analysis of dichotomous response data from certain toxicological experiments, *Biometrics* 35: 281-293

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating functions and the maximum likelihood estimator, *Annals of the institute of statistical mathematics*, 25:1-26.

Jian, X., and Lord, D. (2007). Investigating the application of beta-binomial Models in highway safety, *Texas A&M University*.

Kendall, M. and Stuart A. (1986). *The Advanced Theory of Statistics, Volume* 1: Distribution Theory. Macmillan.

Kleinman, J. C. (1973). Proportions with extraneous variance: single and independent samples. *Journal of American Statistical Association*, 68:46-54.

Lee, S. (2003). Analysis of Binary data in one way layout, *Biometrics,* 45:195-206.

Lee, X., and Nelder, J. A. (2001). Hierarchical generalized linear models: synthesis of generalized linear models, random effect models and structured dispersions. *Biometrics*, 88: 987-1006.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

Moore, D. F. (1986). Asymptotic properties of moment estimators for Over-dispersed Counts and Proportions, *Biometrics*, 73: 583-588.

Moore, D. F., and Tsiatis, M. (1991). Robust estimation of the standard error in moment methods for extra binomial/extra Poisson variation, *Biometrika* 47: 383-401

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

Nelder J. A., and Wedderburn, R. W. (1972). Generalized linear models, *J. Roy statistical society. Ser.* A 135: 370-384.

Paul, S. R. (2009). Estimation of regression and dispersion parameters in the analysis of proportions, *Canadian journal of statistics* 283-303.

Paul, S. R., and Islam, A. S. (1998). Joint estimation of the mean and dispersion parameters in the analysis of proportions: a comparison of efficiency and bias, *The Canadian Journal of Statistics, **26:*** 83–94.

Potthoff, R.F. and Whittinghill, M. (1966). Testing for Homogeneity. The Binomial and Multinomial Distribution. *Biometrika*, 53, 167-182.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81: 321-327.

Robert, B. N., and Bailer, A. J. (2009). Comparing methods for analyzing over-dispersed binary data in aquatic toxicology, *Environmental toxicology and Chemistry*, 28: 997-1006.

Saha, K. K, Sen, D., and Bilisoly, R. (2009). Asymptotic confidence interval for the over-dispersion parameter, with applications to biological count data, *JSM Proceedings of the American Statistical Association, Biometrics.*

Saha, K. K. (2008). Semi-parametric estimation for the dispersion parameter in the analysis of over or under dispersed count data. *Journal of Applied Statistics*, 35, 1383-1397.

Saha, K. K. (2010). Interval Estimation of the dispersion parameter in the analysis of one way layout of count data, *statistics in medicine*. *(wileyonlinelibrary.com*) DOI: 10.1002.

Saha, K. K., and Paul, S. R. (2009). Interval estimation of the negative binomial dispersion parameter. *Journal of Statistical Planning and Inference*, under review.

Saha, K. K., and Paul, S.R. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61, 179-185.

Saha, K. K., and Sen, D. (2009). Improved confidence interval for the dispersion parameter in count data using profile likelihood, *Technical Report No. 4/09*, September 2009.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J Royal Statistical Society, Series B*, 10:257-261.

Sudhir, R. P. (2009). Quadratic estimating equations for the estimation of regression and dispersion parameters in the analysis of proportions, *The Indian Journal of Statistics* 63 1:43–55.

Wedderburn, R. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method, *Biometrics*, 61**:** 439–447.

Williams, D. A. (1961). Extra-binomial variation in logistic linear models, *applied statistics*, 31: 144-148.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31: 946-952.

Xu, Y. Y. (2008). Pooling over-dispersed binomial data to estimate event rate, *BMC Medical Research Methodology*, 1471-2288-8-58.

Yamamoto, E., and Yanagimoto, T. (1994). Statistical Methods for the beta-binomial Model in Teratology, *Environmental Health Perspectives Supplements* Vol. 102 1:25-31.

APPENDIX
##UN-EQUALSAMPLESIZEPLOTS

##MAXIMUM LIKELIHOOD ESTIMATION

```
y<-c(7,1,4,3,5,3,0,11,3,0,10,3,0,4,2,2,3,5,2,1,2,3,1,1,4,5,3,
     3,5,1,1,3,4,0,1,2)
n1<-c(11,1,6,7,8,6,2,19,4,2,15,6,6,10,8,4,5,6,6,4,12,8,4,5,5,
      6,4,10,8,11,4,4,4,2,2,3)
#numberofbootstrapsamples
nBoot=1000
#bootstraparray
B=array(0,dim=c(nBoot,2))
#bootstraploop
for(i in 1:nBoot){

     rn=sample(1:length(y),replace=T)#conditioningntoappearwit
hxvalues
     x<-y[rn]
     n<-n1[rn]
}
for(i in 1:length(n)){
dat<-data.frame(x,n)
bb   <-   betabin(dat,   corrected   =   FALSE,   method   =
"twoAFC",Hessian=True)
summary(bb)
vcov(bb)
coef(bb)
ciML<-c(coef(bb)[2]-
qnorm(0.975)*sqrt(vcov(bb)[4]),coef(bb)[2]+qnorm(0.975)*sqrt(v
cov(bb)[4]))}
length1<-ciML[2]-ciML[1]
```

##METHOD OF MOMENTS ESTIMATION

```
for(i in 1:length(n)){
pi<-x/n
```

```
xbar<-mean(x)
var<-sum(x^2)-n*(pi^2)
wi<-n/var
w<-sum(wi)
py<-sum(wi*pi)/w
s<-sum(wi*(pi-py)^2)
fn1<-sum((wi/n)*(1-wi))
fn2<-sum((wi)*(1-wi))-sum((wi/n)*(1-wi))
p<-py
q<-(1-py)
phi1<-(s-((p*q)*fn1))/(p*q*fn2)
phi<-phi1/(1+phi1)
mu2<-((py*(1-py))*(1+(n-1)*phi)/n)
mu3<-(mu2*(1-2*py)*(1+(2*n-1)*phi)/(n*(1+phi)))
fna<-(1+(2*n-1)*phi)*(1+(3*n-1)*phi)*(1-3*py*(1-
py))*mu2/((1+phi)*(1+2*phi)*n^2)
fnb<-(n-1)*(phi+3*n*mu2)*(mu2*(1-phi))/((1+phi)*(1+2*phi)*n^2)
mu4<-fna+fnb
Ajs<-sum(n*(1-2*py)*x)
Ajk<-0
Akj<-sum(n*(1+(n-1)*phi)*(1-2*py))
Akk<-sum(n*(py*(1-py))*(n-1))
Bjs<-sum(n*(1+(n-1)*phi)*x*(py*(1-py))^3)
Bjk<-sum(n*(py*(1-py))*(1+(n-1)*phi)*(1+(2*n-1)*phi)*(1-
2*py)/(1+phi))
Bkj<-Bjk
Bkk<-sum((mu4-(mu2)^2)*n^4)
A<-matrix(c(Ajs,Ajk,Akj,Akk),2,2,byrow=T)
B<-matrix(c(Bjs,Bjk,Bkj,Bkk),2,2,byrow=T)
mat1<-(solve(A))%*%(B)%*%(t(solve(A)))
ciMM<-c(phi-
qnorm(0.975)*sqrt(mat1[4]),phi+qnorm(0.975)*sqrt(mat1[4]))}
length2<-(ciMM[2]-ciMM[1])
phi
```

```
py
```

##QUASILIKELIHOOD

```
for(i in 1:length(n)){
pi<-x/n
xbar<-mean(x)
var<-sum(x^2)-n*(pi^2)
wi<-n/var
w<-sum(wi)
py<-sum(wi*pi)/w
s<-sum(wi*(pi-py)^2)
fn1<-sum((wi/n)*(1-wi))
fn2<-sum((wi)*(1-wi))-sum((wi/n)*(1-wi))
p<-py
q<-(1-py)
phi1<-(s-((p*q)*fn1))/(p*q*fn2)
phi<-phi1/(1+phi1)
mu2<-py*(1-py)*(1+(n-1)*phi)/n
mu3=mu2*(1-2*py)*(1+(2*n-1)*phi)/(n*(1+phi))
fna<-(1+(2*n-1)*phi)*(1+(3*n-1)*phi)*(1-3*py*(1-
py))*mu2/((1+phi)*(1+2*phi)*n^2)
fnb<-(n-1)*(phi+3*n*mu2)*(mu2*(1-phi))/((1+phi)*(1+2*phi)*n^2)
mu4<-fna+fnb
Ajs<-sum(n*x/(1+(n-1)*phi))
Ajk<-0
Akj<-sum(n*(1+(n-1)*phi)*(1-2*py))
Akk<-sum(n*(n-1)*(py*(1-py)))
Bjs<-sum(n*(py*(1-py))*x/(1+(n-1)*phi))
Bkj<-sum(n*(1-2*py)*(1+(n-1)*phi)/(1+phi))
Bjk<-Bkj
Bkk<-sum(n^4*(mu4-mu2^2))
A<-matrix(c(Ajs,Ajk,Akj,Akk),2,2,byrow=T)
B<-matrix(c(Bjs,Bjk,Bkj,Bkk),2,2,byrow=T)
mat1<-(solve(A))%*%(B)%*%(t(solve(A)))
```

```
ciQL<-c(phi-
qnorm(0.975)*sqrt(mat1[4]),phi+qnorm(0.975)*sqrt(mat1[4]))}

length1<-ciML[2]-ciML[1]
length2<-ciQL[2]-ciQL[1]
length3<-ciMM[2]-ciMM[1]
ciML
ciQL
ciMM
length1
length2
length3
coef(bb)
phi
py
```

##Bootstrap Confidence Interval Plots

```
plot.window(xlim=c(0,1),ylim=c(0,1))
plot.new()
plot.window(xlim=c(0,4),ylim=c(-0.1,.3))
axis(1,1:3,c("MLE","EQL","MME"))
axis(2)
segments(1,ciML[1],1,ciML[2],"blue")
points(c(1,1),c(ciML[1],ciML[2]),col="blue")
segments(2,ciQL[1],2,ciQL[2],"red")
points(c(2,2),c(ciQL[1],ciQL[2]),col="red")
segments(3,ciMM[1],3,ciMM[2],"green")
points(c(3,3),c(ciMM[1],ciMM[2]),col="green")
title(main="ACTUAL CONFIDENCE INTERVAL")
title(xlab="METHOD")
title(ylab="95% CONFIDENCE INTERVAL")
box()
```

COVERAGE PROBABILITIES FOR THE CI FOR THE OVERDISPERSION PARAMETER

##MAXIMUM LIKELIHOOD ESTIMATION

```
M=1000
count<-0
for(i in 1:M){
y<-
     rbetabinom(n=10,size=100,prob=0.1,theta=0.1,shape1=1,shap
     e2=1)
m<-rep(100,length(x))
#numberofbootstrapsamples
nBoot=1000
#bootstraparray
B=array(0,dim=c(nBoot,2))
#bootstraploop
for(i in 1:nBoot){
x=sample(y,replace=T)}
for(i in 1:length(m)){
dat <- data.frame(x,m)
bb<-betabin(dat,     corrected     =     FALSE,     method     =
     "twoAFC",Hessian=True)
summary(bb)
vcov(bb)
coef(bb)
ciML<-c(coef(bb)[2]-
     qnorm(0.975)*sqrt(vcov(bb)[4]),coef(bb)[2]+qnorm(0.975)*s
     qrt(vcov(bb)[4]))}
if(ciML[1]<=0.1   &   ciML[2]>=0.1){count<-count+1}else{count<-
     count}
}
count
ciML
length<-ciML[2]-ciML[1]
length
```

```
##QUASILIKELIHOOD METHOD
M=1000
count<-0
for(i in 1:M){
y<-
     rbetabinom(n=5,size=200,prob=0.1,theta=0.1,shape1=1,shape
     2=1)
#numberofbootstrapsamples
nBoot=1000
#bootstraparray
B=array(0,dim=c(nBoot,2))
#bootstraploop
for(i in 1:nBoot){
x=sample(y,replace=T)}
n<-rep(200,length(x))
for(i in 1:length(n)){
pi<-x/n
xbar<-mean(x)
var<-sum(x^2)-n*(pi^2)
wi<-n/var
w<-sum(wi)
py<-sum(wi*pi)/w
s<-sum(wi*(pi-py)^2)
fn1<-sum((wi/n)*(1-wi))
fn2<-sum((wi)*(1-wi))-sum((wi/n)*(1-wi))
p<-py
q<-(1-py)
phi1<-(s-((p*q)*fn1))/(p*q*fn2)
phi<-phi1/(1+phi1)
mu2<-py*(1-py)*(1+(n-1)*phi)/n
mu3=mu2*(1-2*py)*(1+(2*n-1)*phi)/(n*(1+phi))
fna<-(1+(2*n-1)*phi)*(1+(3*n-1)*phi)*(1-3*py*(1-
     py))*mu2/((1+phi)*(1+2*phi)*n^2)
fnb<-(n-1)*(phi+3*n*mu2)*(mu2*(1-phi))/((1+phi)*(1+2*phi)*n^2)
```

```
mu4<-fna+fnb
Ajs<-sum(n*x/(1+(n-1)*phi))
Ajk<-0
Akj<-sum(n*(1+(n-1)*phi)*(1-2*py))
Akk<-sum(n*(n-1)*(py*(1-py)))
Bjs<-sum(n*(py*(1-py))*x/(1+(n-1)*phi))
Bkj<-sum(n*(1-2*py)*(1+(n-1)*phi)/(1+phi))
Bjk<-Bkj
Bkk<-sum(n^4*(mu4-mu2^2))
A<-matrix(c(Ajs,Ajk,Akj,Akk),2,2,byrow=T)
B<-matrix(c(Bjs,Bjk,Bkj,Bkk),2,2,byrow=T)
mat1<-(solve(A))%*%(B)%*%(t(solve(A)))
ciQL<-c(phi-1.96*sqrt(mat1[4]),phi+1.96*sqrt(mat1[4]))}
if(ciQL[1]<=0.1   &   ciQL[2]>=0.1){count<-count+1}else{count<-
     count}
}
count
length<-ciQL[2]-ciQL[1]
length
##c(ciQL[1],ciQL[2])
```

##METHODOFMOMENTS

```
M=1000
count<-0
for(i in 1:M){
y<-
rbetabinom(n=5,size=200,prob=0.3,theta=0.1,shape1=1,shape2=1)
#numberofbootstrapsamples
nBoot=1000
#bootstraparray
B=array(0,dim=c(nBoot,2))
#bootstraploop
for(i in 1:nBoot){
x=sample(y,replace=T)}
```

```
n<-rep(200,length(x))
for(i in 1:length(n)){
pi<-x/n
xbar<-mean(x)
var<-sum(x^2)-n*(pi^2)
wi<-n/var
w<-sum(wi)
py<-sum(wi*pi)/w
s<-sum(wi*(pi-py)^2)
fn1<-sum((wi/n)*(1-wi))
fn2<-sum((wi)*(1-wi))-sum((wi/n)*(1-wi))
p<-py
q<-(1-py)
phi1<-(s-((p*q)*fn1))/(p*q*fn2)
phi<-phi1/(1+phi1)
mu2<-((py*(1-py))*(1+(n-1)*phi)/n)
mu3<-(mu2*(1-2*py)*(1+(2*n-1)*phi)/(n*(1+phi)))
fna<-(1+(2*n-1)*phi)*(1+(3*n-1)*phi)*(1-3*py*(1-
     py))*mu2/((1+phi)*(1+2*phi)*n^2)
fnb<-(n-1)*(phi+3*n*mu2)*(mu2*(1-phi))/((1+phi)*(1+2*phi)*n^2)
mu4<-fna+fnb
Ajs<-sum(n*(1-2*py)*x)
Ajk<-0
Akj<-sum(n*(1+(n-1)*phi)*(1-2*py))
Akk<-sum(n*(py*(1-py))*(n-1))
Bjs<-sum(n*(1+(n-1)*phi)*x*(py*(1-py))^3)
Bjk<-sum(n*(py*(1-py))*(1+(n-1)*phi)*(1+(2*n-1)*phi)*(1-
     2*py)/(1+phi))
Bkj<-Bjk
Bkk<-sum((mu4-(mu2)^2)*n^4)
A<-matrix(c(Ajs,Ajk,Akj,Akk),2,2,byrow=T)
B<-matrix(c(Bjs,Bjk,Bkj,Bkk),2,2,byrow=T)
mat1<-(solve(A))%*%(B)%*%(t(solve(A)))
ciMM<-c(phi-1.96*sqrt(mat1[4]),phi+1.96*sqrt(mat1[4]))}
```

```
if(ciMM[1]<=0.2   &   ciMM[2]>=0.2){count<-count+1}else{count<-
     count}
}
count
length3<-ciMM[2]-ciMM[1]
length
phi
py
```