Bayesian and Maximum Likelihood Estimation for Correlated Binary data Based on the Double Binomial Distribution

John Appiah Kubi

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
BAYESIAN AND MAXIMUM LIKELIHOOD ESTIMATION FOR CORRELATED BINARY DATA BASED ON THE DOUBLE BINOMIAL DISTRIBUTION

by

John Appiah Kubi

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

August 2021
DEDICATION

My Mom, Madam Agatha Quayson
My Dad, Mr John Appiah-Kubi
My Wife, Abena Adutwumwaah
My Children, Joshua and Ithiel.
My Siblings, Anita, Selina, Fred, Sylvester and Sylvia.
ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to the Lord Almighty for his grace and mercy on my life. Without God, my academic journey would not be possible.

I am immensely grateful to my advisor, Dr Bowman, whose advice, encouragement, enduring patience and indefatigable support throughout this task has been nothing short of amazing. To Dr George, thank you sincerely from the bottom of my heart. You nurtured my interest in Bayesian statistics and guided me in developing a profound understanding of the intricacies of Bayesian statistics, and I am eternally grateful for that.

I also wish to thank the other members of my committee, Dr Deng and Dr Ching-Chi Yang, profusely for their valuable comments, suggestions and extreme patience in the face of numerous impediments. Your inputs made this possible and I am thankful for that.

A special thank you to all the Professors and staff at the Department of Mathematical Sciences, University of Memphis, for their dedication, support and guidance which has enabled me to complete my degree and shaped me into who I am today.

I would like to acknowledge with appreciation, the love and support of my family; my parents John Appiah-Kubi and Agatha Quayson, this vision tarried for so many years yet you waited for it and it has come to pass, and I owe it all to Elohim. To my beloved wife, Abena Adutwumwaah, thank you for believing in me.

Last but not the least, I want to thank my friends who helped me in this journey. The love and encouragement has been worth it. Thank you all and may God replenish whatever you lost for my sake. Thank you once again!
ABSTRACT


In the statistical analysis of binary data, usually the binomial distribution is the most commonly used probability model in many applications. The binomial is a very useful distribution in many practical applications due to its simplicity and having a parameter that has an intuitive meaning. However, it has a unique feature that is the variance depending on the mean, an assumption that does not reflect reality in many practical applications especially in cases where the data exhibits greater variability than predicted by the distribution. The two-parameter double binomial model introduced by Efron (1986) may be considered as a useful alternative to the one-parameter binomial distributions, given that it can account for both overdispersion and underdispersion. In this dissertation, we obtain maximum likelihood estimates for the double binomial distributions. The Bayesian methodology is also considered for estimation procedures and we demonstrate that, under the model assumptions, the fully conditional posterior distributions that are obtained allow for the estimation of posterior distributions via the Gibbs sampling algorithm. A toxicity data set involving the effect of ethylene glycol in mice is analyzed using our considered methodology.
Contents

List of Figures vi
List of Tables vii

1 Introduction 1

2 Literature Review 6
   Developmental Toxicity 6
   Dose Response Relationship 7
   Correlated Binary Data 10
   Overdispersion 11
      Generalized Linear Models 13
   Review of Existing Methods 16
   Approximate Methods 16
      Quasilikelihood 16
      Generalized Estimating Equations 17
   Likelihood Methods 20
      Exchangeable Binary Distributions 21
      Cluster Specific Models 24
   Bayesian Methods 25

3 Efron’s Double Binomial and Bayesian Modeling 27
   Introduction 27
   Double Exponential Families 29
      Example 30
         The Kullback-Leibler Distance 31
   Efron’s Double Binomial Model 31
      Maximum Likelihood Estimators 33
      Bayesian Estimators 35

4 Bayesian and Likelihood Estimation Using Simulated Data Set 38
   Simulation Study 38

5 Application to Teratology 45
   Efron’s Binomial for Teratology 45
   Application to Toxicity Data 49
      Detecting Overdispersion on Toxicity Data 51
      Discussion 51

Bibliography 57
List of Figures

1. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 10 observations

2. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 100 observations

3. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 1000 observations

4. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of Control group.

5. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 0.25 group.

6. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 0.50 group.

7. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 1.00 group.
## List of Tables

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary of maximum likelihood simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>Summary of Bayesian simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.</td>
<td>39</td>
</tr>
<tr>
<td>3</td>
<td>Summary of Beta-binomial MLE simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>Summary of maximum likelihood simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>Summary of Bayesian simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.</td>
<td>41</td>
</tr>
<tr>
<td>6</td>
<td>Summary of Beta-binomial MLE simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>Proportion ($x_i/n_i$) of fetuses with malformation ($x_i$) out of total fetuses ($n_i$).</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>Summary statistics for the toxicity data set of Table 7.</td>
<td>51</td>
</tr>
<tr>
<td>9</td>
<td>Bayesian estimates based on toxicity data set.</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>Maximum likelihood estimates based on toxicity data set.</td>
<td>52</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The difficulty that is encountered when analyzing correlated (or clustered) binary outcomes is common in a number of biomedical research settings. In developmental toxicity, for example, pregnant rats or mice are exposed to one of ordered levels of some potential toxicant and their offspring are examined for malformations or prenatal death. Due to shared environmental and genetic factors, littersmates tend to show greater similarity in their individual probabilities of response. This phenomenon is known as the litter effect and it is present in the areas of ophthalmology, genetics and periodontal studies. Disregarding such inherent dependence can lead to incorrect inferences. Methods that are tailored to account for correlation within groups, or clusters, of observations have been studied extensively in the literature. The cluster effects are accounted for differently by the different approaches. Most commonly used approaches are derived from likelihood-based methods such as the exchangeable model approach, the random effects approach and distributions that jointly model the mean and the intra-cluster correlation or quasi-likelihood methods such as the generalized estimating equations (GEEs). Most of these methods have been non-Bayesian.

In the likelihood based method such as the random effects model, the intra-cluster (litter) correlation is assumed to be induced by a random effect. In this model, the within-cluster responses become conditionally independent and identically distributed given the random effects. Thus, the sum of the responses follows the (conditional) binomial distribution. The proposition by D. Williams (1975) and Kupper and Haseman (1978) to use the beta distribution in modeling the random effects on the response probability led to the now well-known Beta-binomial distribution. Modeling the logit of the response probability to be normally distributed, Stiratelli, N. Laird, and Ware (1984) proposed the logit-normal-binomial distribution. Ochi and R. L. Prentice (1984) modeled the probit of the response probability to be normally distributed, and thus introduced the probit-normal-binomial model. By modeling
the iterated logarithm of the response probability, Conaway (1990) proposed the log-Gamma distribution. Modeling the joint distribution of the multiple binomial responses, Coull and Agresti (2000) used the multivariate logit-normal-binomial. These models are conceptually simple and easily interpreted but require formidable numerical integration.

The joint distribution, in an exchangeable model, is modeled in terms of the marginal probabilities. These marginal probabilities make it possible to model correlations of all orders in an exchangeable model, that is, higher order moments are incorporated into such a model. In contrast to a mixture model where the mixture distribution is chosen arbitrarily, an exchangeable model avoids the use of arbitrary mixing distribution. Rather it makes use of the framework that describes the distinctive nature of the associated exchangeable binary sequence using the elegant de Finetti theorem as the basis. George and Bowman (1995) and Bowman and George (1995) proposed general expressions for the full likelihood of a sum of exchangeable Bernoulli variables through a combinatorial argument. Within this framework, the assumption of exchangeability meant that the joint probability of all outcomes in a cluster does not change irrespective of the permutation of the responses. Under the assumption of exchangeability, George and Kodell (1996) analysed binary data and proposed nonparametric likelihood ratio approaches for testing independence and heterogeneity in dose-related trend. Kuk (2004) proposed a new class of distributions for exchangeable binary data that models the joint success probabilities of all orders using the power family of completely monotone functions. Kuk used the George and Bowman (1995) context to define the families of power functions. The approaches by George and Bowman and Kuk were applied to clinical and developmental toxicity studies. The results from these applications were compared to existing methods. Xu and Prorok (2003) initiated an investigation into the modeling and analysis of exchangeable binary data with random cluster sizes. Stefanescu and Turnbull (2003) modeled the exchangeable binary data with varying cluster sizes using the

GEEs were introduced by Liang and S. L. Zeger (1986) and have been studied extensively since. The only requirement for producing consistent estimates from GEEs is specifying the first moment of the distribution correctly. Correlation between outcomes is accommodated by means of a “working” correlation matrix. Mis-specifying the covariance function produces estimates of model parameters that are still valid because they depend on the first moment. However, the procedure is inefficient for estimating correlations and higher moments (Moore, 1986). The parameters of interest in GEEs are usually the marginal mean and the pairwise associations.

The other approach to analyzing correlated binary data is through the joint modeling of the mean and variance through distributions that generalize the model. Examples include the multiplicative binomial by P. M. Altham (1978), the double binomial by Efron (1986), and the beta binomial by Skellam (1948). The double binomial (Efron, 1986) and the multiplicative binomial (P. M. Altham, 1978) are two-parameter models that belong to the exponential family. The beta-binomial originally could model only overdispersion until R. Prentice (1986) extended and formulated the conditions necessary for modeling underdispersion too. According to Lindsey and P. Altham (1998), the normalizing constants for both the double-binomial and multiplicative binomial are intractable, and that makes them unattractive to practitioners. Dey, Gelfand, and Peng (1997) proposed fitting these double exponential family models within a Bayesian framework. Nott (2006) described nonparametric Bayesian methods for the mean and variance estimation where the response variables are modeled following a distribution in the double exponential family. The mean and dispersion parameters were modeled as additive functions of the predictors.

In this dissertation, we develop a Bayesian approach for jointly modeling the mean and intra-cluster correlation parameters of Efron’s double binomial model. The model of
Efron provides a useful framework for Bayesian inference in the presence of correlated binary data. The proposed framework has many advantages over some commonly used procedures. Because the model is a member of the exponential family, we can bring the full complement of Generalized Linear Models (GLMs) to bare. As a special case of the double exponential family of distributions, the double binomial can accommodate both under and overdispersion. In this way, it provides great flexibility for modeling correlated (or clustered) binary data. The probability mass function is an approximation and so probabilities do not exactly add up to one even though the difference is not that significant for any values of the parameters. The approximate form of the double binomial makes it convenient in applications by its simplicity to computing the likelihoods and obtaining parameter estimates by maximum likelihood and Bayesian methods. The Bayesian framework was chosen because, unlike other methods that rely on asymptotic approximation and thus fare poorly in the presence of small or sparse samples, it also provides an avenue to incorporate prior information about our parameters.

The overall goal is to be able to use the double binomial distribution as a practical and simple alternative for the analysis of correlated binary data in the presence of over and underdispersion. Some of the specific goals include (1) fully working out the properties of the double binomial distribution (2) developing Bayesian and frequentist approaches (3) applying the methods to both simulated and real data sets.

The rest of the dissertation is organized as follows. Chapter 2 describes the dose-response relationship, correlated binary data, overdispersion and reviews the literature on some of the common approaches used in the analysis of correlated binary data.

In Chapter 3, we describe the double binomial distribution introduced by Efron (1986), the maximum likelihood estimation of the parameters, the Bayesian estimation procedures, the expected Fisher’s information matrix, the Kullback-Leibler distance.

We dedicate Chapter 4 to the simulation study conducted to assess the performance of our double binomial model. We show the estimates for our probability of success and
dispersion parameters of our double binomial model for some select values of $\mu$, $\theta$, and $n$. We also have trace plots and histograms to monitor the convergence of our MCMC samples. We shall also simulate data from the beta-binomial distribution for comparison purposes.

In Chapter 5, an application of the model in analyzing the developmental toxicity study of ethylene glycol (EG) in mice is discussed.
Chapter 2
Literature Review
Developmental Toxicity

Developmental toxicity studies are a vital area in the field of toxicology. They deal with the analysis of the outcomes of fetal litters to maternal exposure to toxic agents, which may be considered innocuous for humans. They are carried out to look out for chemicals (substances) that may cause harm or pose a risk to the developing fetus if pregnant women are exposed. Regulatory agencies use the results of such studies to help set guidelines for exposing humans to such poisonous agents.

In developmental toxicity studies, pregnant females are randomly assigned to one of several dose groups corresponding to various exposure levels of the toxicant under study. This may be usually done from the implantation stage to the day before birth. However, they may be exposed throughout the pregnancy. Measurements relating to endpoints on both the animal and fetal litters are recorded. There is an assumption of the existence of a dose-response relationship between the agent level dose and the endpoints. Two important endpoints that measure toxicity are the incidences of fetal malformations and prenatal deaths. Other endpoints such as the fetal weight and length are also important. (Catalano and L. M. Ryan, 1992), however, we will focus on the dose-response modelling of malformations and prenatal deaths observed in fetuses.

It is important to note that the appropriate experimental unit in developmental toxicology experiment is the litter. The litter is often used because of the environment and genetic characteristics shared among littermates. Individual litters naturally form clusters. Measurements for litters are commonly accumulated into a sum or average that represents a data value for the entire litter, e.g. number of deaths or malformations in a litter, or average litter weight. So, the total number of litters make up the study sample size. The group sample sizes are still large enough to provide the study with considerable power and
meaningful estimates. However, this is not the case when we have individual-level data. Individual-level data are usually used for jointly modeling bivariate outcomes such as, fetal weight and malformation. (Catalano and L. M. Ryan, 1992)

**Dose Response Relationship**

Exposure, dose, and dose-relationships are key concepts in toxicology studies. The main idea behind toxicological studies is that after the exposure of an organism to a toxicant, an effect is evoked in the organism. As the dose level go up, the effect will become more evident, until the organism can no longer tolerate the compound. Dose is the amount of chemicals (toxicants) given to, administered to, taken, or absorbed by an organism (Sullivan, Agardy, and Clark, 2005). One of the most important concepts in toxicology is the dose-response relationship. It describes the measure of an organism’s response, as a function of exposure (or doses), to a stimulus after a period of time. It forms the framework on which all hazard assessment testing is done, and dose-response model extrapolations are based. Dose-response relationships are defined using the following (IUPAC,1993)

1. Dose-related effect: Magnitude of a biological change is connected to the dose

2. Dose-response curve: Graph of the association between dose and the fraction of individuals in a population responding with an all-or-none effect.

3. Dose-response relationship: Relationship between dose and the incidence of a defined biological effect in an exposed population.

For dose-response analysis of developmental toxicity data, it is usually recommended to use data from individual fetuses (identified to litter) or individual litters (eg. number of offspring in a litter, proportion of malformed fetuses in a litter) to accurately measure fetal risk and its variability. This is so since responses from fetuses of the same dam are statistically dependent; and the pups from the same litter respond more similarly than pups from different litters because of similar genetic and environment factors. The use of
litter-specific data helps in providing intralitter correlation through modeling methods such as “nested” dichotomous models that depend on a beta-binomial model (Skellam, 1948) of variability, or by quasi-likelihood methods, generalized estimating equations (Liang and S. L. Zeger, 1986) and hierarchical modeling (Fox, K. A. Hogan, and Davis, 2017). It is almost impossible to estimate variances accurately without using litter-specific data. However, it is possible to use summary data, that is, dose-group summaries for the number of malformed or dead fetuses and the number of fetuses, when the design effect has been estimated for each dose group. Past research has shown useful dose-response and trend tests analysis based on design effects estimates using litter-specific data from the same study. The analysis in this case is simplified, however, with the unavailability of litter-specific data, it does not help. Hence, other options that use summary statistics of binomial data, yet can account for the intra-litter correlation are needed.

D. Williams (1975) was the first to propose the use of the beta-binomial distribution in modeling the number of adverse fetal response. The beta-binomial has a desirable feature of having a parameter that models the intralitter correlation. In the context of dose-response modeling, Chen and Kodell (1989) were the first to use the beta-binomial. It was used to fit a monotonic dose-response model to a developmental toxicity dataset. D. L. Hunt and Bowman (2004) also used the beta-binomial to fit a hormetic model to the same data set. A drawback of the beta-binomial in dose-response modeling is the separate modeling of the correlation parameter from the dose-response parameters connected to the parameter equating to response. A desirable alternative to the beta-binomial was introduced by D. Hunt and S. N. Rai (2003). The litter responses conditional on the litter response variation were assumed to follow a binomial distribution while the litter response variation was assumed to be from a normal distribution. The advantage here is that the parameter for the response variation is part of the dose-response model, thus it helps with estimation.

D. Hunt and S. N. Rai (2003) proposed a model where response variation was assumed
to be uniform across dose groups, essentially to restrict the nuisance parameters to be able to compare bias in the estimation of threshold in their model and that of the beta-binomial.

K. Rai and Van Ryzin (1985) modeled the probability of response for an individual fetus as a function of both dose and litter size. Their model had two factors; the first one was defined as the probability of effect on the litter environment and the second was interpreted as the probability of effect on a fetus conditional on the existence of an effect on litter environment. The second probability factor at a given dose level varies with litter size. To account for the observed extrabinomial variation in their original model, they included the litter size variable: for a fixed litter size and dose level, responses were assumed to be binomial. There was also an assumption by K. Rai and Van Ryzin (1985) that litter sizes could be modeled as Poisson distributions with means being nonlinear functions of the dose level. Rai and Ryzin’s model was applied to several developmental toxicity endpoints from the NTP database by Faustman, Wellington, Smith, and C. A. Kimmel (1989). They suggested the model was able to represent the observed dose-response relationships for prenatal death and malformation.

The use of a log-logistic kind of model was proposed by Kupper, Portier, M. D. Hogan, and Yamamoto (1986). They evaluated the model through a simulation study. In such a model, at dose level $d_i$, the logit of the expected response probability, logit ($\mu_i$) was assumed to be a linear function of $\ln (d_i)$. To account for the observed extrabinomial variation of responses in developmental toxicity studies, the observations were assumed to have a beta-binomial distribution. In formulating the beta-binomial distribution, they assumed that the responses among fetuses within a litter are conditionally independent, given the underlying response probability, but the underlying response probability is assumed to be a beta distribution that varies from dam to dam. The beta-binomial distribution resulting from this formulation produces dose group-specific intralitter correlations and shows extrabinomial
variation. In this model, litter size was neither considered as a covariable for probability of effect, nor did it allow for nonzero background rates.

Kodell, Howe, Chen, and Gaylor (1991) proposed a model for developmental toxicity studies in which the variation of responses was assumed to be beta-binomial distribution. In their model, litter size was treated as a possible covariate in the estimation of the probability of response, although their introduction of the litter size into the model differed from the one proposed by Rai and Ryzin. The litter size in Kodell, Howe, Chen, and Gaylor (1991)’s model acted as a modifier of parameters describing background response rate and dose-response slope. This model allowed for nonzero background rates and introduced the ”threshold dose” parameter.

**Correlated Binary Data**

Binary data can have only two values. They are related to, or comprise of, only two possible categories the response measure can take. They either fall into these categories by default (e.g. head/tail, success/failure, yes/no) or they can be formed by dividing continuous data into two categories.(e.g. whether or not the height is less than 6’) (Pendergast, Gange, Newton, Lindstrom, Palta, and Fisher, 1996). Binary data that are obtained through simple random sampling, or under the assumption of independence have their covariances following the binomial model. However, when the data are dependent, researchers must incorporate correlation of responses when estimating parameters, otherwise, they are bound to make incorrect inferences about the regression model parameters (because of incorrect estimation of variances) or biased estimates of the regression coefficients that could lead to incorrect conclusions regarding their research questions (P. Diggle, P. J. Diggle, P. Heagerty, Liang, P. J. Heagerty, S. Zeger, *et al.*, 2002).

Correlated binary data often arise in applications such as longitudinal studies, cluster sample surveys, ophthalmology etc. In ophthalmic research, measurements obtained simultaneously on both eyes of an individual are likely to be more similar due to exposure to
similar environmental and genetic factors compared to measurements obtained on different individuals. In the study of genetically-transmitted diseases, measurements obtained on siblings who belong to the same parents can form a cluster, and are likely to have correlated outcomes. In longitudinal studies, measurements are taken periodically on sampling units. It is highly likely that measurements taken on the same sampling unit are more alike than than that taken on a different sampling unit.

Neglecting the dependency between correlated (or clustered) data can result in the underestimation of standard errors and overstating the significance. This is known as the overdispersion phenomenon where subjects in the same cluster exhibit less variation than expected while between cluster variation is greater than expected.

**Overdispersion**

Analysis of data via one-parameter exponential families such as the binomial and the Poisson imply that the variance is a function of the mean. The variance-mean relation fails in the presence of a very common practical complication known as overdispersion, or more rarely underdispersion. Overdispersion is a phenomenon that arises when the data exhibits more variability than the nominal variance – in this case the nominal binomial variance under the assumed distribution. This “variance discrepancy” (Ehrenberg, 1959) restricts the direct suitability of models involving such distributions. The effect of overdispersion basically depends on how it arises and its degree of incidence. The simplest, and perhaps the most common way overdispersion arises is through the failure of the binomial independence assumption to hold. The violation of the independence assumption is caused via what Efron (1986) refers to as clumped sampling. Households, litters, and neighborhoods are common examples of clustering in a population. In animal litter studies, for example, it is well established that an inherent feature of such data is the so-called litter effect, that is, the tendency for littermates to respond similarly to stimuli than non-littermates. Responses from littermates are often positively correlated.
To illustrate the concept of overdispersion in correlated data, let \( Y = \sum_{i=1}^{m} Z_i \) denote a binomial response where \( Z_1, Z_2, ..., Z_m \) are independent binary variates taking values 0 or 1 such that \( P(Z_i = 1) = \pi \), and \( P(Z_i = 0) = 1 - \pi \). Then

\[
E(Y) = \sum_{i=1}^{m} E(Z_i) = m\pi
\]

\[
Var(Y) = \sum_{i=1}^{m} Var(Z_i) = m\pi(1 - \pi)
\]

If the independence assumption is violated, and the correlation \((z_j, z_k) = \tau > 0\), then

\[
E(Y) = \sum_{i=1}^{m} E(Z_i) = m\pi
\]

\[
Var(Y) = \sum_{i=1}^{m} Var(Z_i) = m\pi(1 - \pi)[1 + \tau(m - 1)]
\]

which leads to overdispersion under the binomial model. It is worth noting that \( \tau < 0 \) signifies underdispersion which is unlikely in practice.

Overdispersion may also arise through the violation of the binomial assumption of constant probability of success from trial to trial. This is a possibility if the population can be subdivided into naturally occurring subunits, for example, in colonies where the probability of success varies from subunit to subunit.

Overdispersion and its effects cannot be overlooked. Many statistical packages have routinely included overdispersion. Failing to account for overdispersion, according to Cox (1983), has two effects. First, summary statistics show greater variation than expected under
the simple model. The second effect is how inefficient it becomes to use appropriate statistics for the single-parameter family.

**Generalized Linear Models**

Before discussing existing methods for the analysis of correlated binary outcomes, we shall review generalized linear models (GLMs) and show how the logistic came into being within the GLM framework.

GLMs are an extension of the ordinary regression models to include responses that have non-normal distributions and modeling functions of the mean. They are made up of three components; we have (1) the random component (2) the systematic component and (3) the link function

**Random component:** The random component consists of a response variable $Y$ with samples $(y_1,..,y_n)$ from a distribution that belongs to the exponential family. The exponential family of distributions has probability density function or mass function given as

$$f(y_i, \theta_i) = a(\theta_i)b(y_i) \exp[y_iQ(\theta_i)]$$

(2.1)

The form of $f$ (and hence that of $a$, $b$, and $Q$) is assumed to be known. The unknowns are $\theta_i$’s which have to be estimated. The $Q(\theta_i)$ is called the natural parameter. If $Q(\theta_i) = \theta$, the exponential family is said to be in canonical form.

**Systematic component:** The systematic component relates some vector $(\eta_1,...,\eta_n)$ to the $p$ explanatory variables through a linear model. We assume the relationship is given by:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

for $i = 1,...,n$. $\beta_1,...,\beta_p$ are unknown and have to be estimated.

**Link function:** The third component of a GLM is the link function that connects the
random component and the systematic component. It is a function \( h \) such that \( \eta_i = h(\mu_i) \) for \( i = 1, \ldots, n \). The function \( h \) is assumed to be known.

If \( \eta_i = h(\mu_i) = \mu_i \), the link function is called the identity link. If \( \eta_i = h(\mu_i) = Q(\theta_i) \), the link function is called the canonical link.

We can extend the GLM to include dispersion exponential families. For the random component of GLMs, we assume that \( Y_i \) has the probability density (or mass) function of the form (2.1). In some cases, it helps to include an additional parameter called the dispersion parameter to model the data accurately. For example, by using the Bernoulli distribution for binary data, we assume that \( Var(Y_i) = p(1 - p) = E(Y_i)(1 - p) \) which may not be the case.

With the addition of the dispersion parameter \( \phi \), it is common for the probability density (or mass) function of \( Y_i \) to be of the form

\[
f(y_i, \theta_i, \phi) = \exp \left[ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]
\]

(2.2)

With \( \phi \) known, (2.2) reduces to the exponential family and we can use all the techniques used for GLMs. However, it is usually unknown, and so we estimate it first and use it for the rest of the GLM procedure.

Formulating the distribution in the form of (2.2) has many useful properties. The most useful is perhaps the mean and variance which are specified as:

\[
E(Y_i) = \mu_i = b'(\theta_i)
\]

(2.3)

\[
Var(Y_i) = b''(\theta_i) a(\phi)
\]

As an example, let \( Y_i \sim \text{Binomial} \left( n_i, \pi_i \right) \), then we can formulate \( Y_i \) as a GLM,
\[ P[Y_i = y_i] = \left( \frac{n_i}{y_i} \right) \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \]

\[ = \exp \left[ \frac{y_i \theta_i - \log[1 + \exp(\theta_i)]}{1/n_i} + \log \left( \frac{n_i}{y_i} \right) \right] \quad (2.4) \]

where

\[ \theta_i = \log \frac{\pi_i}{1 - \pi_i} \]

is the log odds ratio.

Expressing the model in this form has become known as the "logistic regression". Apart from being the canonical parameter for a binomial GLM, modeling the log odds has an advantage over modeling the \( \pi_i \) directly. That is, the odds ratio is unconstrained on the log scales whiles the probabilities are restricted to \([0,1]\).

In GLMs, the analysis is carried out by modeling the canonical parameter, \( \theta \) as a linear function of model parameters, that is, \( \theta = \alpha + \beta x_i \)

The model parameters can be estimated using maximum likelihood techniques. For GLMs, the likelihood equations are

\[ \sum \frac{\partial \mu_i}{\partial \beta} \frac{y_i - \mu_i}{\text{var}(Y_i)} = 0 \quad (2.5) \]

The above likelihood equation, in the context of the logistic regression model, takes the form

\[ \frac{\partial \pi_i}{\partial \beta} \frac{y_i - \pi_i}{\pi_i(1 - \pi_i)} = 0 \quad (2.6) \]

Newton-Raphson and Fisher scoring are some of the methods that can be used to solve equations (2.5) and (2.6).
Review of Existing Methods

As it is well-known, analyzing clustered data using standard univariate methods may not be suitable because units or individuals in the same cluster cannot be treated as functionally independent. The main statistical challenge that arises from analyzing such data is dealing with association among units or individuals within a cluster. A number of methods have been proposed for analyzing correlated binary data, and most of them have been non-Bayesian in nature. These approaches have varied from using a single inflator factor to rescale the covariance matrix to complete likelihood based inferences. We shall generally restrict our attention to some of the common approaches rather than open the door to the large and dynamic literature on this topic presented within the general framework of categorical data models. We shall classify the methods of analyzing clustered data into three broad categories - approximate methods, likelihood-based methods and Bayesian methods.

Approximate methods

Fitting models using maximum likelihood requires the form of distribution of the observations to be specified. In GLM, this means that the response or outcome variable has a specific probability distribution which belongs to the exponential family of distributions. Nelder and Lee (1992), however, noted that for distributions with non-normal errors with GLMs such as the binomial or Poisson in which the errors are a fixed function of the mean, the dispersion parameter cannot vary independently and that restricts the applicability of the GLM to some extent. Under such instances, there is the need for methods that do not depend on exact likelihood. It is worth noting that a quasilikelihood serves as an alternative in situations where the true likelihood of a distribution does not belong to the exponential family of distributions.

Quasilikelihood

This method was first proposed by Wedderburn (1974) for estimating the parameters in the model for the mean via estimating equations. McCullagh et al. (1983) examined the
method extensively by showing the nexus between quasi-likelihood functions, exponential family models and non-linear weighted least squares. Thus, if we know the type of data (discrete, categorical, continuous) at hand, form of the skewness, mean-variance relationship etc, then a quasi-likelihood method can be used for obtaining estimating equations. Few assumptions about the distribution of the response variable, $Y_i$, are required to implement this methodology and so it is useful for different outcomes. Unlike likelihood analysis where the actual distribution form of the response variable must be specified, quasi-likelihood requires only a specification of a relationship between the mean and variance of the observations and between the mean and covariates (S. L. Zeger and Liang, 1986). D. A. Williams (1982) suggested that in order to avoid estimation bias in modeling the extra-binomial interlitter variation via the beta-binomial distribution, the maximum likelihood estimators had to be discarded in favor of the more robust maximum quasi-likelihood estimators. L. Ryan (1992) used quasi-likelihood ideas to account for litter effects. Chen and Li (1994) proposed a quasi-likelihood model under both double beta-binomial and Dirichlet-trinomial variations for modeling the dependency between the proportion of death/resorption and proportion of malformations.

**Generalized Estimating Equations**

Liang and S. L. Zeger (1986) were the first to use GEEs for the analysis of correlated binary data. The method is an extension of the quasilikelihood techniques for dependent data. The GEE approach of analyzing correlated data was to be used in situations where it was appropriate to assume that the marginal mean response follows a generalized linear model in which the main interest lies in the coefficients of the linear model. Under such circumstances, the within-cluster correlation is regarded as a nuisance (Hall and Severini, 1998). Rather than being concerned with the likelihood, the GEE approach relaxes some of the requirements for GLMs. The GEE approach makes three assumptions
1. The marginal mean response, $E(Y_{ij}) = \mu_{ij}$ is linked to the explanatory variables, $x_{ij}$ by a link function, $g(\mu_{ij}) = x_{ij}'\beta$, where $g$ is known. For binary outcomes, the most commonly used link function is the logit function.

2. The marginal variance is linked to the marginal mean according to a known variance function $\text{Var}(Y_{ij}) = V(\mu_{ij})\phi$ where $\phi$ is a scale parameter that can be estimated.

3. The correlation between two observations $Y_{ik}$ and $Y_{il}$ in the same cluster is a function of the marginal means and may be additional parameters $\lambda$, e.g. $\text{Corr}(Y_{ik}, Y_{il}) = \rho(\mu_{ik}, \mu_{il}, \lambda)$

Estimation for GEEs is carried out by making slight adjustments to the likelihood equations in (2.5). The variance portion of the likelihood equation in (2.5) is replaced by a working correlation matrix. Define $D_i$ as a matrix of partial derivatives, $\frac{\partial \pi_i}{\partial \beta}$, $A_i$ as an $n_i \times n_i$ matrix representing the variance of $y_i$, then the likelihood can be written as

$$\sum_{i=1}^{n} D_i^T A_i^{-1}(y_i - \pi_i) = 0$$

(2.7)
In GEEs, we replace the variance matrix, $A_i$, with a working correlation matrix $V_i$ matrix such that (2.7) becomes

$$\sum_{i=1}^{n} D_i^T V_i^{-1} (y_i - \pi_i) = 0$$

where $V_i$ is an $n_i \times n_i$ covariance matrix with the structure

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi$$

Liang and S. L. Zeger (1986) indicated that under mild regularity conditions, estimates obtained using this model are asymptotically consistent and follow a normal distribution assuming the working correlation was specified correctly. Replacing $V_i$ with a robust variance estimate produces consistent estimates of the covariance matrix even when the correlation, $R(\alpha)$, is misspecified. This is what is known as the "sandwich estimator".

One interesting attribute about GEE is that parameters of the correlation structure and that of the model are not necessarily connected. They can be modeled independently. It is possible to change the model without altering the correlation structure and vice versa.

Many different correlation structures can be accommodated by GEEs due to their flexibility (Liang and S. L. Zeger, 1986). Common correlation structures that are available in software packages in R and SAS are independence, exchangeability, autoregressive and unstructured. The independence correlation structure works under the assumption that observations in the same cluster are independent, that is, they are not correlated. In exchangeable correlation structure, correlation between any two observations in the same cluster is assumed to be constant across clusters. The most commonly used correlation structure for repeated measures is the autoregressive. The autoregressive structure works under the assumption that correlation between successive observations is a constant within and across clusters. The unstructured correlation structure puts no assumption on the correlation structure. It works by estimating each component separately. (P. Diggle, P. J. Diggle,
P. Heagerty, Liang, P. J. Heagerty, S. Zeger, et al., 2002). In GEEs, the correlations are constrained to the interval $[-1, 1]$. To break this restriction, Liang, S. L. Zeger, and Qaqish (1992) and Lipsitz, N. M. Laird, and Harrington (1991) suggest that the dependency must be modeled in terms of the odds ratios. Odds ratios are unconstrained and can easily be interpreted compared to correlations. Bowman, Chen, and George (1995) used the generalized estimating equations approach in modeling both the mean responses and the intralitter correlations as functions of dose levels.

**Likelihood Methods**

We consider likelihood-based approaches that are based on the complete representation of the joint probabilities of the binary responses. Fitting models using maximum likelihood requires the form of distribution of the observations to be specified. This means there is a specific data generation mechanism, for example, the data consists of counts of events in a Poisson process or the data are normally distributed with constant variance.

Two of the most commonly used approaches include the exchangeable model approach and the random effects approach. To account for litter effect and overdispersion using the random effects approach, we assume that the intralitter correlation is induced by a random effect that is shared by all the fetuses within the same litter. This random effect can be regarded as a combination of all factors, both genetic and environmental that are shared by litter mates. With this litter-specific random effect, we assume that outcomes of the litter mates are conditionally independent.

The exchangeable model approach is meant to capture symmetry in a problem, that is, symmetry that does not require independence. In this case, the independence assumption is relaxed to exchangeability. Formally, a sequence is exchangeable if its joint probability distribution is a symmetric function of its $n$ arguments. In other words, we can swap or reorder the sequence with no change to its joint distribution. This implies that every independent and identically distributed sequence is exchangeable but not all exchangeable sequences are
independent. They may be identically distributed though. Under exchangeability, we can estimate higher order moments efficiently.

### Exchangeable Binary Distributions

A finite sequence of binary random variables, $B_1, B_2, \ldots, B_m$ is exchangeable if

$$Pr(B_1 = b_1, \ldots, B_m = b_m) = Pr(B_{\pi(1)} = b_1, \ldots, B_{\pi(m)} = b_m)$$

(2.9)

for any $m$, and for every permutation $\pi(1), \pi(2), \ldots, \pi(m)$ of $1, 2, \ldots, m$ where $b_1, b_2, \ldots, b_m \epsilon \{0, 1\}$.

Let $B_1, \ldots, B_m$ be a set of exchangeable binary random variables such that $B_i = 1$ (success) or $0$ (failure), $i = 1, \ldots, m$. Then there are many parameterizations of the joint distribution. We will present a few of these parameterizations.

The Bahadur (1961) representation makes use of the marginal response probability $\pi = P(B_i = 1)$ and correlations of all orders $(\rho_2, \ldots, \rho_m)$ such that

$$\rho_k = E[(B_1 - \pi)(B_2 - \pi)\ldots(B_k - \pi)]/\{[\pi(1 - \pi)]^{k/2}\}$$

where $\rho_k$ is the same for any subset $k$ of $m$ responses $(B_1, \ldots, B_m)$ due to exchangeability. R. L. Prentice (1988) used the Bahadur’s parameterization. Specifically, the pairwise correlation is

$$\rho = \rho_2 = \frac{E(B_1B_2) - \pi^2}{\pi(1 - \pi)}$$

Another form of the joint distribution is the response conditional representation. For this parameterization, the response is given in terms of probabilities $\gamma_0, \ldots, \gamma_{m-1}$. In this case,

$$\gamma_s = P(B_i = 1 \mid \sum_{j=1,j\neq i}^m B_j = s), \quad i = 1, \ldots, m, \quad s = 0, \ldots, m - 1$$

Parameterizing this way produces a class of response conditional models (Pendergast,
These class of response conditional models are used, for example, by Qu George, G. W. Williams, Beck, and Goormastic (1987) and Connolly and Liang (1988). One natural way exchangeability is characterized using this parameterization is that each single \( y \) in the cluster depends of the rest through the sum. In clustering disease within families, Yu and Zelterman (2002) used this parameterization.

Another representation is the cluster sum distribution. In this form, probabilities \( p_0, \ldots, p_m \), such that \( \sum_{i=0}^{m} p_i = 1 \) are used, with \( p_i = Pr(I = i), \ i = 0, \ldots, m \) and \( I = \sum_{i=1}^{m} B_i \). Putting this parameterization in the Bahadur context gives

\[
\pi = \sum_{i=1}^{m} Pr(B_i = 1 | I = i) Pr(I = i) = \frac{1}{m} \sum_{i=0}^{m} ip_i \tag{2.10}
\]

To find the correlation, \( \rho \), we compute \( E(B_i B_j) \) where \( i \neq j \).

\[
E(B_i B_j) = \sum_{i=2}^{m} Pr(B_i = 1, B_j = 1 | I = i) Pr(I = i) = \sum_{i=2}^{m} \frac{i(i - 1)}{m(m - 1)} p_i \tag{2.11}
\]

From equations (2.10) and (2.11), \( \rho \) may be expressed in terms of \( \{p_i\} \).

For the Bowman and George (1995) parameterization, they used joint success probabilities \( 1 = \alpha_0 \geq \alpha_1 \geq \ldots \geq \alpha_m \) such that \( \alpha_k = P(B_1 = 1, \ldots, B_k = 1), \ k = 1, 2, \ldots, m \). Then the joint distribution of \( B_1, \ldots, B_m \), according to Bowman and George (1995), is given by

\[
P(B_1 = b_1, \ldots, B_m = b_m) = \sum_{i=0}^{m-y} (-1)^i \binom{m-y}{i} \alpha_{y+i}, \ y = 0, 1, \ldots, m \tag{2.12}
\]

and

\[
P(Y = y) = \binom{m}{y} \sum_{i=0}^{m-y} (-1)^i \binom{m-y}{i} \alpha_{y+i} \tag{2.13}
\]

where \( Y = B_1 + B_2 + \ldots + B_m \) and \( y = b_1 + b_2 + \ldots + b_m \) is the total number of successes. A correlation, \( \rho_k \) of order \( k \) is given by
\[ \rho_k = \frac{\sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} \alpha_1^{k-i} \alpha_i}{[\alpha_1(1 - \alpha_1)]^{k/2}} \]

If \( B_1, .., B_m \) are independent, it is easy to show that \( \alpha_k = \alpha_1^k \) for \( k = 1, 2, \ldots \) and \( Y \) has a binomial distribution with parameters \( m \) and \( \alpha_1 \), that is \( Y \sim B(m, \alpha_1) \).

The four parameteric representations are similar, and we can alternate between them. Particularly we can make use of the relation

\[ p_i = \binom{m}{y} \sum_{i=0}^{m-y} (-1)^i \binom{m-y}{i} \alpha_{y+i} \]

In the saturated model, which is the most general model, the parameter sets \((\pi, \rho = \rho_2, \rho_3, \ldots, \rho_m), (\alpha_1, \ldots, \alpha_m)\) or \((p_1, \ldots, p_m)\) lie in an \( m \)-dimensional space. In order to make inferences based on information from clusters of different sizes, it is appropriate to use model formulations in which parameters have the same meaning regardless of cluster size. This is known as the "interpretability assumption". In the Bahadur formulation, for example, the marginal mean \( \pi \), and correlation \( \rho_k \) must be same for all clusters of size \( m \geq k \). The situation is similar in the Bowman-George formulation where \( \alpha_k \) must be the same for all clusters of size \( m \geq k \). The situation is, however, different in the case of \( \gamma_s \) and \( p_i \) where the parameters \( \gamma \) and \( p \) depend on the size, \( m \geq k \), of the cluster.

For a given application, there should be a justification of the interpretability assumption. In teratological applications, for example, the marginal mean \( p \) may, contrarily, depend on the cluster size \( m \). This may be so because in a litter where there is competition for food and other things, survival of offspring will largely depend on the size of the cluster. However, in other applications such as familial aggregation studies of disease, the interpretability assumption may be justified. Drawing inferences based on the aggregation of information from clusters of different sizes without the interpretability assumption is difficult and can
lead to the imposition of some restrictions on the model parameters for the effect of cluster size on the joint distribution (Stefanescu and Turnbull, 2003).

In investigating the maximum likelihood estimation for a sample of independent clusters of unequal sizes, Stefanescu and Turnbull (2003) proposed the use of the EM algorithm. They argued that clusters of different sizes were considered as coming from a sample of clusters of equal size, \( M \) where clusters with sizes, \( m \leq M \), have \( M - m \) observations missing at random.

**Cluster Specific Models**

Instead of modeling the marginal response, cluster-specific approach treats the probability distribution of a binary outcome, \( Y_{ij} \) as a function of covariates, \( X_{ij} \) and parameters \( \alpha_i \) specific to each cluster. For instance, we might postulate a linear relationship between a covariate and the log odds of the event that \( Y_{ij} = 1 \). Mathematically, we may write it as

\[
\text{logit}\left[ \Pr(Y_{ij} = 1 \mid x_{ij}) \right] = \alpha_{i,1} + \alpha_{i,2} x_{ij}
\]

where \( \alpha_{i,1} \) and \( \alpha_{i,2} \) are the intercept and slope parameters for the \( ith \) cluster. One of the complications under this model is the growing number of parameters as the clusters increase. Examples of cluster-specific models include the random effects model and the conditional likelihood.

The most popular approach for attaining cluster, or subject-specific, estimates is the random or mixed-effects model. Its popularity comes from its ability to reduce the number of parameters in a cluster specific model. That reduction is carried out under the assumption that the clusters are a random sample from an underlying population of clusters with parameter values following a distribution.

The approach works under the assumption that variation between clusters is as a result of some unobservable variable that is represented by a probability distribution. Given a realization from such a distributions, observations in a cluster are mutually independent.
The simplest random effects model is the random intercept model. In this model, each cluster has its own intercept.

\[
\text{logit}[\Pr(Y_{ij} = 1 \mid Z_i)] = \beta_0 + \beta_1 x_{ij} + Z_i
\]

This model works under the assumption that differences between clusters is attributed to each cluster having its own baseline probability, \(\frac{\exp(\beta_0 + Z_i)}{1 + \exp(\beta_0 + Z_i)}\). If this probability is known, then the odds ratio of a positive response given a covariate \(x\) is a constant across all clusters, \(\exp(\beta_1)\). Allowing covariates to have random effects as well leads to more complicated mixed effects models.

One of the assumptions of the random effects model is the conditional independence within clusters given \(Z_i\), that is, \(Y_i \mid Z_i, x_{ij}, \beta_1\) and \(Y_j \mid Z_j, x_{ij}, \beta_1\) are assumed independent for all \(i \neq j\).

Korn and Whittemore (1979) proposed one of the first random effects logistic regression model. This model was generalized by Gilmour, Anderson, and Rae (1985) into a GLM where the random effects were normally distributed, that is, \(Z_i \sim N(0, \sigma^2)\) with an arbitrary link function. Here, \(\sigma^2\) measures the amount of heterogeneity between clusters. A large \(\sigma^2\) denotes a large amount of heterogeneity between clusters while a small \(\sigma^2\) represents a small amount of heterogeneity (more homogeneous). When \(\sigma^2 = 0\), the model reduces to an ordinary logistic regression.

**Bayesian methods**

Even though there are many sophisticated models and methods that have been devoted to the analysis of clustered binary data; a comprehensive review of the various modeling approaches is provided by Pendergast, Gange, Newton, Lindstrom, Palta, and Fisher (1996), few authors or papers have focused on using Bayesian methodology in analyzing correlated binary data. The development of Bayesian methods for analyzing such data may be important for at least two reasons. First, estimates from maximum likelihood have some desirable
asymptotic properties which make them attractive especially when dealing with large sample sizes. However, when the sample size is small, these ML estimates may not be appropriate. Secondly, the Bayesian framework allows for the incorporation of prior information. It is reasonable to assume that an analyst has such information in the form of past or present data.

One of the simplest, yet versatile Bayesian models used in this regard is the beta-binomial model. In exploring the Bayes theorem, Pearson (1925) suggested the idea of the beta-binomial, yet the formal proposition of the model is credited to Skellam (1948). In situations where the beta-binomial is inapplicable, variants of it have been proposed and applied. Examples can be found in Kupper and Haseman (1978), Morrison and Brockway (1979) and Sabavala and Morrison (1981).

D. Williams (1975) explored the use of the beta-binomial model in analyzing the teratogenic effect of a toxic chemical on the fetuses of pregnant laboratory animals. In his model, he assumed that the binary responses within each litter constituted a Bernoulli trial with a probability parameter. For a treatment group, the probability parameter varied between litters according to a beta distribution. Maximum likelihood estimates of the parameters of the beta distribution for each treatment group was obtained. Using the asymptotic likelihood ratio tests, parameter values from the various treatment groups were compared.
Chapter 3
Efron’s Double Binomial and Bayesian Modeling

Introduction

The issue of overdispersion has received significant attention in the literature. Discrete (count or binary) data are usually found to exhibit greater heterogeneity than what the one-parameter family of models can explain due to a breach in the mean-variance relationship in such a family; sample variance tend to be larger in comparison to that predicted by the models.

To address these problems, we consider a larger set of models, say a two-parameter family, that induce overdispersion when dealing with such data. That is, a two-parameter family with a second parameter that models the variance independent of the mean. In the past, this has been accomplished by mixing a one-parameter family with a two-parameter family resulting in a two-parameter marginal mixture family for the data. Conducting such mixing inflates the model variance at the same value of the mean (Gelfand and Dalal, 1990). For one-parameter exponential family, Shared (1980) showed that this is necessarily the case. Cox (1983) noted that for modest amount of overdispersion, specifying the mixing distribution fully is not required; only the mean and the variance (two parameters) are required. Gelfand and Dalal (1989) argued that the assumption that the one-parameter family is an exponential family enhanced the modeling and inference of overdispersion. They developed a general class of two parameter exponential families which are overdispersed relative to a given one-parameter exponential family. Usually the one-parameter exponential family is mixed with a two-parameter conjugate distribution as in Morris (1982). The overdispersed family of models that results from the mixture has often been awkward to work with since it does not belong to the exponential family.

In modelling overdispersion, Efron (1986) constructed a double exponential family using Hoeffding’s characterization of a one-parameter exponential family via Kullback-Leibler
distance. This family is a two-parameter exponential family with a simple approximate form of the normalizing function. Efron relied on asymptotic arguments on the assumption that each observation was the average of a large number of observations.

In studying strategies for selecting models for overdispersed data, Fitzmaurice (1997), considered an overdispersed model based on Efron (1986)'s exponential family formulation. To account for overdispersion, he proposed simple modifications to the likelihood ratio test (LRT) and other standard model selection criteria such as the AIC and BIC. These adjustments were made to prevent the selection of a model with too many parameters, which can result in the overinterpretation of parameters. More parsimonious models were obtained using these adjustments compared to using the standard model selection criteria.

Lindsey and P. Altham (1998) fitted three overdispersed models in the analysis of the human sex ratio data set collected by Geissler. They included the beta-binomial of Skellam (1948), the multiplicative binomial model of P. M. Altham (1978) and Efron (1986) double binomial model. The probability and the dispersion parameters in each of the three distributions were allowed to vary with family size based on two distinct regression equations. In larger families, all three models showed that the probability of a boy child and the dispersion are greater.

Lee and Nelder (2000) re-analyzed Geissler’s human sex ratio data. The mean and dispersion were jointly modeled. They discussed some of the advantages of using the unnormalized form of Efron’s double exponential family. Results showed that extended quasilikelihood and the unnormalized double exponential family produced similar inference. Under certain conditions of the parameter, they showed that their version of the overdispersed binomial distribution could be seen as a version of the corresponding (independent) overdispersed Poisson distributions.

The double-exponential family is attractive to work with because they can be directly incorporated into the generalized linear model framework and that simplifies inference.
Efron’s model allows the dispersion parameter to be modeled as a function of explanatory variables in a simple way.

**Double Exponential Families**

This section presents some of the basic properties of double exponential families.

We begin with the one-parameter exponential family of densities:

\[
 f(x_i, \mu_i) = \exp\{x_i \eta(\mu_i) - \tau(\mu_i)\} \, dG(x_i),
\]

where \( E(X_i) = \mu_i, \eta(\mu_i) \) is the natural, or canonical parameter (a monotone function of \( \mu_i \)), \( \tau(\mu_i) \) is a normalizing function, \( G(x_i) \) denotes the carrier measure for the exponential family, either Lebesgue or counting measure. Next, suppose \( y \) is an observation that is the average of \( n \) independent observations, \( x_i, ..., x_n \) each of which is of the form (3.1). \( y \) is given as

\[
 y = \frac{\sum_{i=1}^{n} x_i}{n}
\]

Then

\[
 f(y, \mu, n) = \exp[n\{y \eta(\mu) - \tau(\mu)\}]dG_n(y)
\]

where \( E(Y) = \mu, \text{Var}(Y) = V(\mu)/n \)

Given an exponential family (3.2), Efron (1986)’s double-exponential family is defined as

\[
 \tilde{f}(y, \mu, \theta, n) = c(\mu, \theta, n) \theta^{1/2}\{f(y, \mu, n)\}^{\theta}\{f(y, y, n)\}^{1-\theta}dG_n(y)
\]

with parameters \( \mu, \theta, \) and \( n \). The constant \( c(\mu, \theta, n) \) is a normalizing constant chosen to make the density integrate to 1, and is close to 1, that is, \( c(\mu, \theta, n) = 1 + O(n^{-1}) \), hence the family of densities can be approximated by the expression
\[
\tilde{f}(y, \mu, \theta, n) \approx \theta^{1/2} \{ f(y, \mu, n) \}^{\theta} \{ f(y, y, n) \}^{1-\theta} [dG_n(y)] 
\] (3.4)

If \( \theta \) is known, then equation (3.3) reduces to a one-parameter family with \( E(Y) = \mu \), and \( \text{Var}(Y) = V(\mu)/n\theta \). The distribution is overdispersed if \( \theta < 1 \).

The main idea behind double exponential family formulation is the introduction of an additional parameter \( \theta \) which controls the dispersion in the one-parameter family independent of the mean. The two-parameter family behaves just like the one-parameter family except that the sample size changes from \( n \) to \( n\theta \).

**Example**

Let \( X_1, \ldots X_n \) be random variables such that \( X_i \sim N(\mu, \sigma^2) \). Then the average, \( y = \frac{\sum x_i}{n} \sim N(\mu, \sigma^2/n) \), where \( \mu \) is unknown but \( \sigma^2 \) is fixed and known, thus

\[
f(y, \mu, n) = \left( \frac{2\pi \sigma^2}{n} \right)^{-1/2} \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} \right]
\]

Plugging the above into equation (3.3) gives

\[
\tilde{f}(y, \mu, \theta, n) = c(\mu, \theta, n) \left( \frac{2\pi \sigma^2}{n\theta} \right)^{-1/2} \exp \left[ -\frac{n\theta}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right] 
\] (3.5)

With \( c(\mu, \theta, n) \) assumed to be approximately 1, equation (3.5) is the density of \( y \sim N(\mu, \sigma^2/n\theta) \). That is, if we begin with the ordinary family \( N(\mu, \sigma^2/n) \) which has an unknown mean \( \mu \) and a known variance \( \sigma^2/n \), then we will end up with a double family that has an unknown mean \( \mu \) and an unknown variance \( \sigma^2/n\theta \). The double exponential family makes it possible to include an additional parameter, called the dispersion parameter to one-parameter families.
The Kullback-Leibler Distance

**Definition:** The Kullback Leibler (Kullback, 1959) distance is a measure of the discrepancy or difference between two densities \( f_{\theta_1} \) and \( f_{\theta_2} \) defined as

\[
D(\theta_1, \theta_2) = E_{\theta_1} \ln[f_{\theta_1}(Y)/f_{\theta_2}(Y)]
\]  

(3.6)

For our one-parameter family in equation (3.1), the KL distance is thus;

\[
D(\mu_1, \mu_2) = E_{\mu_1} \ln[f(x, \mu_1)/f(x, \mu_2)] \\
= (\eta_1 - \eta_2)\mu_1 - (\tau(\mu_1) - \tau(\mu_2))
\]  

(3.7)

\( D(\mu_1, \mu_2) \) can be thought of as quantifying the distance between the two means. Intuitively, it averages a measure of the discrepancy between two density functions over their support. It is not a distance function in the true sense of the word because it is asymmetric, that is, the KL distance between \( f_{\theta_1} \) and \( f_{\theta_2} \) is not the same as KL distance between \( f_{\theta_2} \) and \( f_{\theta_1} \). However, \( D(\mu_1, \mu_2) \geq 0 \) with equality if and only if, \( f_{\theta_1} \equiv f_{\theta_2} \).

For the general case, \( f(y, \mu, n) \) considered in (3.2),

\[
D_n(\mu_1, \mu_2) = E_{\mu,n} \ln[f(y, \mu_1, n)/f(y, \mu_2, n)] = nD(\mu_1, \mu_2)
\]  

(3.8)

**Efron’s Double Binomial Model**

Let \( y \) be a response variable denoting the sum of \( n \) independent binary random variables (r.v.’s) \( x_1, \ldots, x_n \) with sample size \( n \) and \( \mu \) as the probability parameter. Consider a rescaled binomial random variable \( y \sim \text{Bin}(n, \mu)/n, \mu \in [0, 1] \), Then;

\[
g(y, \mu, n) = \binom{n}{ny} \mu^{ny}(1 - \mu)^{n(1-y)}, \quad y = 0, 1/n, 2/n, \ldots, 1
\]  

(3.9)

Rewriting (3.9) in the one-parameter family form (2.4), we have

\[
g_{\mu,n}(y) = \binom{n}{ny} 2^{-n} e^{-n \left[ y \log \left( \frac{\mu}{1-\mu} \right) + \log \left( \frac{2(1-\mu)}{2} \right) \right]}
\]
where \( \eta = \log[\mu/(1 -\mu)] \), \( \psi = -\log[2(1 -\mu)] \), and \( G_n \) as the discrete distribution with mass \( (\frac{n}{ny})2^{-n} \) at \( y = 0, 1/n, 2/n, \ldots, 1 \). The extreme cases \( \mu = 0 \) and \( \mu = 1 \) have been included for convenience because technically, they do not belong to the exponential family.

Plugging equation (3.9), which is a one-parameter family into equation (3.3), gives us the double binomial distribution (Efron, 1986). Thus, the probability mass function of a random variable \( Y \) that follows a double binomial distribution may be written in the form

\[
\tilde{f}(y, \mu, \theta, n) = c(\mu, \theta, n)\theta^{1/2} \left[ \left( \frac{n}{ny} \right)^{\mu} (1 - \mu)^{n(1-y)} \right]^\theta \left[ \left( \frac{n}{ny} \right)^{\mu} (1 - y)^{n(1-y)} \right]^{1-\theta}
\]

(3.10)

for \( y = 0, 1/n, \ldots, 1 \), where \( \mu > 0 \), \( \theta > 0 \), and \( c(\mu, \theta, n) \) is a normalizing constant that ensures \( P(Y = y) \) sums to unity.

It can be shown that the mean and variance of a double binomial random variable can be approximated by

\[
E(Y) = \mu
\]

\[
Var(Y) = \frac{\mu(1 - \mu)}{n\theta}
\]

(3.11)

Thus, the distribution is overdispersed if \( \theta < 1 \) and underdispersed if \( \theta > 1 \). If \( \theta = 1 \), then the expression (3.10) corresponds to the probability mass function of a random variable that follows binomial distribution with mean \( \mu \) and variance \( \frac{\mu(1-\mu)}{n} \). The double binomial belongs to the two-parameter exponential family of distributions that is generically expressed as

\[
f_{\theta,\mu}(y) = a(y)b(\theta, \mu) \exp[w_1(\theta, \mu)s_1(y) + w_2(\theta, \mu)s_2(y)]
\]

Here we can suppress the \( n \) since the binomial distribution is closed under convolutions and \( g_{\mu, n}(y) \) is the same family for all values of \( n \) where \( a(y) = y^y(1 - y)^{1-y} \) is a function of
\( y, \ b(\theta, \mu) = c(\theta, \mu)\sqrt{\theta}(1 - \mu)^\theta, \ w_1(\theta, \mu) = \theta \log\left(\frac{\mu}{1 - \mu}\right), \) and \( w_2(\theta, \mu) = \theta \) are functions of the parameters \( \theta \) and \( \mu \), and \( s_1 = y, \ s_2 = y \log\left(\frac{1-y}{y}\right) - \log(1-y) \) are functions of \( y \) that does not depend on \( \theta \) and \( \mu \)

**Maximum Likelihood estimators**

Below are some results concerning likelihood analyses of repeated independent observations from a double binomial distribution. We find maximum likelihood estimators (MLEs) for \( \theta \) and \( \mu \) by finding a local maximum of \( \log L(\mu, \theta) \).

Let \( y_1, y_2, ..., y_K \) be independent identically distributed random variables from the double binomial \( \tilde{f}(y, \mu, \theta, n) \) Then the likelihood is given by

\[
L(\theta, \mu \mid y) \propto \prod_{i=1}^{K} \left[ \theta^{1/2} \left( \frac{\mu}{y_i} \right)^{n\theta y_i} \left( \frac{1 - \mu}{1 - y_i} \right)^{n\theta(1-y_i)} \right] \tag{3.12}
\]

Let \( l(\mu, \theta, n) \) be the log-likelihood such that

\[
l(\mu, \theta, n) = \log L(\theta, \mu \mid y) \tag{3.13}
\]

Differentiating the log-likelihood function \( (3.13) \) with respect to \( \theta \) and \( \mu \), we have the following equations

\[
\frac{\partial l}{\partial \mu} = n\theta \left[ \frac{\sum_{i=1}^{K} y_i}{\mu} - \frac{\sum_{i=1}^{K} (1 - y_i)}{1 - \mu} \right]
\]

\[
\frac{\partial l}{\partial \theta} = \frac{K}{2\theta} + n \sum_{i=1}^{K} \left[ y_i \log\left( \frac{\mu}{y_i} \right) + (1 - y_i) \log\left( \frac{1 - \mu}{1 - y_i} \right) \right] = \frac{K}{2\theta} - n \sum_{i=1}^{K} D(y_i, \mu)
\]
From these equations, we obtain the maximum likelihood estimators given by

\[ \hat{\mu} = \frac{1}{K} \sum_{i=1}^{K} y_i = \bar{y} \]

\[ \hat{\theta} = \frac{K}{2n \sum_{i=1}^{K} D(y_i, \bar{y})} \]

As expected, the MLEs are functions of the sufficient statistics

\[ S_1 = \sum_{i=1}^{K} y_i \]

\[ S_2 = \sum_{i=1}^{K} [y_i \log(1/y_i) + (1 - y_i) \log(1/(1-y_i))] \]

for \( \mu \) and \( \theta \).

The inverse of the observed Fisher information matrix, \( I(\mu, \theta) \), give the asymptotic variances of the MLEs. The second partial derivatives of the log likelihood function are given as follows:

\[ \frac{\partial^2 l}{\partial \mu^2} = \frac{n \theta K (\mu^2 - \mu)}{[\mu(1 - \mu)]^2} \]

\[ \frac{\partial^2 l}{\partial \theta^2} = -\frac{K}{2 \theta^2} \]

\[ \frac{\partial^2 l}{\partial \mu \partial \theta} = n \left[ \frac{\sum_{i=1}^{K} y_i}{\mu} - \frac{\sum_{i=1}^{K} (1 - y_i)}{1 - \mu} \right] \]

So, our 2 \times 2 information matrix, \( I(\hat{\mu}, \hat{\theta}) \) is given as

\[ I(\hat{\mu}, \hat{\theta}) = -\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \theta} \\ \frac{\partial^2 l}{\partial \theta \partial \mu} & \frac{\partial^2 l}{\partial \theta^2} \end{pmatrix} = \begin{pmatrix} n \hat{\theta} K / V(\hat{\mu}) & 0 \\ 0 & K / 2 \hat{\theta}^2 \end{pmatrix} \]  \tag{3.14}

The off-diagonal entries of our information matrix are zero. This implies that our MLEs, \( \hat{\mu} \) and \( \hat{\theta} \) are asymptotically independent. The variance-covariance matrix may be approxi-
mated, as we stated earlier, by the inverse of equation (3.14). Thus, the estimated standard errors for \( \hat{\mu} \), and \( \hat{\theta} \), are given by:

\[
\hat{se}(\hat{\mu}) = \sqrt{\frac{\bar{y}(1 - \bar{y})}{n\hat{\theta}K}}
\]

\[
\hat{se}(\hat{\theta}) = \sqrt{\frac{2\hat{\theta}^2}{K}}
\]

Approximate 100(1 - \( \alpha \))% confidence intervals for \( \mu \) and \( \theta \) are given, respectively, as

\[
\hat{\mu} \pm \frac{z_{\alpha/2}}{\sqrt{n\bar{y}(1 - \bar{y})}}
\]

\[
\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{2K}}\frac{2\hat{\theta}^2}{K}
\]

where \( z_{\alpha/2} \) is the upper 100\( \alpha \) - \( th \) percentile of the standard normal distribution.

**Bayesian Estimators**

For Bayesian estimation, we need a prior distribution for the parameters \( \mu \) and \( \theta \) in the double binomial distribution. Since our knowledge on the parameters is scanty, we will use a non-informative prior that is known as the joint Jeffrey’s noninformative prior distribution. This prior distribution is given by the square root of the determinant of the Fisher’s information matrix. We can observe that, we have,
\[ E\left[ -\frac{\partial^2 \log L(\mu, \theta)}{\partial \theta^2} \right] = \frac{K^2 \theta^2}{\mu(1-\mu)} \]

Thus, the joint Jeffrey's prior distribution for \( \mu \) and \( \theta \) is given by

\[
\pi(\mu, \theta) \propto \sqrt{\frac{1}{\theta \mu (1-\mu)}} \quad (\text{3.15})
\]

We combine this joint prior distribution with the likelihood function \( L(\mu, \theta) \) given by (3.12) to yield the joint posterior density given by \( \pi(\mu, \theta|y) \propto \pi(\mu, \theta) \cdot L(\mu, \theta) \). Thus,

\[
\pi(\mu, \theta|y) = k \sqrt{\frac{1}{\theta \mu (1-\mu)}} \prod_{i=1}^{K} \left[ \theta^{1/2} \left( \frac{\mu}{y_i} \right)^{n \theta y_i} \left( \frac{1-\mu}{1-y_i} \right)^{n \theta (1-y_i)} \right]
\]

where \( k \) is the normalizing constant that makes \( \pi(\mu, \theta|y) \) a proper probability density function. The fully conditional distributions of the parameters \( \mu \) and \( \theta \) needed for the Gibbs sampling algorithm are given respectively by

\[
\pi(\mu|\theta, y) \propto \sqrt{\frac{1}{\mu (1-\mu)}} \mu^{n \theta \sum_{i=1}^{K} y_i} (1-\mu)^{n \theta (K - \sum_{i=1}^{K} y_i)}
\]

Thus, \( \pi(\mu|\theta, y) \sim \text{Beta} \left( n \theta \sum_{i=1}^{K} y_i + 1/2, \ n \theta (K - \sum_{i=1}^{K} y_i) + 1/2 \right) \)
\[
\pi(\theta | \mu, y) \propto \sqrt{\frac{1}{\theta}} \theta^{K/2} \mu^{n_\theta} \sum_{i=1}^{K} y_i (1 - \mu)^{n_\theta(K - \sum_{i=1}^{K} y_i)} \prod_{i=1}^{K} y_i^{-n_\theta y_i} (1 - y_i)^{-n_\theta(1 - y_i)}
\]

\[
= \theta^{(k-1)/2} e^{-\theta n \sum_{i=1}^{K} D(y_i, \mu)}
\]

Thus, \(\pi(\theta | \mu, y) \sim \text{Gamma} \left( \frac{K + 1}{2}, n \sum_{i=1}^{K} D(y_i, \mu) \right)\)
Chapter 4
Bayesian and Likelihood Estimation Using Simulated Data Set

Simulation Study

In order to assess the performance of the Bayesian and maximum likelihood estimation methods for the parameters of the double binomial (DB) distribution in the absence of covariates, a brief simulation study was conducted. Random samples from this distribution were generated using the following procedure: The simulated samples were taken to follow double binomial distribution with nominal parameters

1. Start off with nominal parameters \( \mu, \theta \) and \( n \) is the number of trials

2. Compute \( P(Y = y_i) = p_i \) using the probability mass function of the double binomial model for \( y_i = 0, 1/n, 2/n, ..., 1 \) where \( \sum_i p_i \approx 1 \)

3. Generate \( U \) from \( U \sim (0, 1) \)

4. Set \( Y = y_i \) if

\[
\sum_{i=0}^{j-1} p_i \leq U < \sum_{i=0}^{j} p_i
\]

The standard errors and confidence intervals of the parameters were obtained for both the Bayesian and maximum likelihood estimation. We also simulated data from the beta-binomial (BB) for same parameters for purposes of comparison.
Table 1. Summary of maximum likelihood simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.

<table>
<thead>
<tr>
<th>No. of obs</th>
<th>Parameter</th>
<th>Estimate</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.30</td>
<td>(0.202, 0.398)</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.84</td>
<td>(0.104, 1.58)</td>
<td>0.376</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.34</td>
<td>(0.303, 0.379)</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.426, 0.752)</td>
<td>0.083</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.35</td>
<td>(0.342, 0.366)</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.535, 0.638)</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Table 2. Summary of Bayesian simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.

<table>
<thead>
<tr>
<th>No. of obs</th>
<th>Parameter</th>
<th>Estimate</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.30</td>
<td>(0.199, 0.417)</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.85</td>
<td>(0.269, 1.740)</td>
<td>0.380</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.34</td>
<td>(0.303, 0.380)</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.437, 0.760)</td>
<td>0.084</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.35</td>
<td>(0.342, 0.366)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.536, 0.638)</td>
<td>0.027</td>
</tr>
</tbody>
</table>

With the number of trials, $n$ set at 10, the Bayesian and maximum likelihood estimates for the nominal parameters $\mu = 0.37$, and $\theta = 0.56$ are approximately equal for the corresponding number of observations as it is displayed in Tables 1 and 4. As the number of observations increase, the simulated estimates of the parameters seem to improve for both the Bayesian and maximum likelihood estimation except when we had 1000 observations where the estimate for $\theta$ remained the same. For the beta-binomial estimation, there were improvements as well in the estimates as the number of observations increased as depicted in Table 3. The standard errors for both the mean and dispersion parameters also seem to decrease with increasing number of observations in all three estimation methods. This is not unexpected because as the number of observations increase, the simulated parameters
Table 3. Summary of Beta-binomial MLE simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 10$.

<table>
<thead>
<tr>
<th>No. of obs</th>
<th>Parameter</th>
<th>Estimate</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.30</td>
<td>(0.116, 0.548)</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.54</td>
<td>(0.144, 1.639)</td>
<td>0.326</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.38</td>
<td>(0.304, 0.451)</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.68</td>
<td>(0.478, 0.965)</td>
<td>0.122</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.36</td>
<td>(0.337, 0.384)</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.524, 0.656)</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Table 4. Summary of maximum likelihood simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.

<table>
<thead>
<tr>
<th>No. of obs</th>
<th>Parameter</th>
<th>Estimate</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.33</td>
<td>(0.252, 0.398)</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.80</td>
<td>(0.100, 1.49)</td>
<td>0.356</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.36</td>
<td>(0.333, 0.388)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.432, 0.764)</td>
<td>0.085</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.37</td>
<td>(0.357, 0.374)</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.57</td>
<td>(0.523, 0.624)</td>
<td>0.026</td>
</tr>
</tbody>
</table>

are getting closer to the true parameter of the population and so the deviation from the population parameter will decrease.

The estimates seem to improve significantly when the number of trials is increased to 20 as displayed in Tables 4, 5 and 6 for the same parameters defined earlier. The estimates for the Bayesian and maximum likelihood of the double binomial are recorded in Tables 4 and 5. There seem to be one important trend from the results of our simulation study for both the Bayesian and maximum likelihood estimation of the double binomial and that of the maximum likelihood estimation of the beta-binomial distribution. That is, as the number of trials increase, the estimated parameters become closer to the true values. As
Table 5. Summary of Bayesian simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.

<table>
<thead>
<tr>
<th>No. of obs.</th>
<th>Parameter</th>
<th>Bayesian Estimation for Double Binomial</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.33</td>
<td>(0.248, 0.411)</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.80</td>
<td>(0.253, 1.629)</td>
<td>0.359</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.36</td>
<td>(0.334, 0.387)</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.445, 0.777)</td>
<td>0.085</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.37</td>
<td>(0.357, 0.374)</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.57</td>
<td>(0.524, 0.624)</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Table 6. Summary of Beta-binomial MLE simulated data for $\mu = 0.37$, $\theta = 0.56$, and $n = 20$.

<table>
<thead>
<tr>
<th>No. of obs.</th>
<th>Parameter</th>
<th>Maximum Likelihood Estimation for Beta-binomial</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\mu$</td>
<td>0.24</td>
<td>(0.077, 0.498)</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.35</td>
<td>(0.078, 1.125)</td>
<td>0.234</td>
</tr>
<tr>
<td>100</td>
<td>$\mu$</td>
<td>0.36</td>
<td>(0.287, 0.431)</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.53</td>
<td>(0.381, 0.720)</td>
<td>0.086</td>
</tr>
<tr>
<td>1000</td>
<td>$\mu$</td>
<td>0.37</td>
<td>(0.345, 0.390)</td>
<td>0.0112</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.59</td>
<td>(0.542, 0.660)</td>
<td>0.030</td>
</tr>
</tbody>
</table>

expected, there was a general reduction in the standard errors of both $\mu$ and $\theta$ for all three estimation methods using $n = 20$.

In constructing the trace plots, the Markov Chain Monte Carlo (MCMC) simulation was run for 5,000 updates after a burn-in period of 5,000 updates to wipe out the effects of the initial values. We used the posterior means as the point estimates of our Bayesian parameters. Apart from the point estimates that are given by the posterior means, we also constructed the 95% credible intervals. The credible intervals are estimated by the 2.5th and 97.5th centiles of the respective posterior distributions.

Figures 1, 2 and 3 are trace plots and histograms of $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for the 10, 100 and 1000 observations respectively. The trace plots help us observe the convergence of our Gibbs sampling algorithm.
From the histograms in figures 1, 2 and 3, the shape of both parameters look more normal as the number of observations increase.

Figure 1. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 10 observations.
Figure 2. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 100 observations
Figure 3. Trace plots and histograms of the posterior MCMC samples for $\mu = 0.37$, $\theta = 0.56$ and $n = 10$ for 1000 observations
Chapter 5

Application to Teratology

For our teratological experiment, we have $s$ dose groups with the $i$th dose group consisting of $m_i$ litters exposed to dose $d_i$ with $n_{ij}$ denoting the size of the $j$th litter of the $i$th dose group. We define the random variable $X_{ijk}$,

$$X_{ijk} = \begin{cases} 
1 & \text{if the } k\text{th fetus in the } j\text{th litter of the } i\text{th dose is malformed} \\
0 & \text{otherwise}
\end{cases}$$

Let

$$X_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk}$$

denote the number of malformed fetuses in the $j$th litter of the $i$th dose group. Further, for the $j$th litter in the $i$th dose group, we are interested in $Y_{ij} = X_{ij}/n_{ij}$, the proportion of malformed fetuses in the $j$th litter of the $i$th dose group.

Let $\mu_{ij}$ denote the probability of malformation of fetuses in litter $j$ of the $i$th dose group. We consider the model that treats $y_{ij}$ as a rescaled binomial variate with

$$y_{ij} \sim bin(n_{ij}, \mu_{ij})/n_{ij}$$

where $\mu_{ij} = \mu_i$, $i = 1, ..., s$.

In this model, we treat all litters in a particular dose group $i$ as having the same probability of malformation.

Efron’s Binomial for Teratology

For the double binomial where we treat $y_{ij}$ as a rescaled binomial variate, we have
\[
\tilde{f}(y, \mu_i, \theta_i, n_{ij}) = c(\mu_i, \theta_i, n_{ij}) \theta_i^{1/2} \left[ \left( \frac{n_{ij}}{n_{ij}y_{ij}} \right)^{n_{ij}y_{ij}} \mu_i^{n_{ij}y_{ij}} (1 - \mu_i)^{n_{ij}(1 - y_{ij})} \right] \times \left[ \left( \frac{n_{ij}}{n_{ij}y_{ij}} \right)^{y_{ij}n_{ij}y_{ij}} (1 - y_{ij})^{n_{ij}(1 - y_{ij})} \right]^{1 - \theta_i} \tag{5.1}
\]

For a given dose group \(i\) with \(m_i\) litters, the likelihood function is given by

\[
L(\theta_i, \mu_i \mid y_{ij}) \propto \prod_{j=1}^{m_i} \theta_i^{1/2} \left( \frac{\mu_i}{y_{ij}} \right)^{n_{ij}} \left( \frac{1 - \mu_i}{1 - y_{ij}} \right)^{n_{ij}} \tag{5.2}
\]

Let \(l(\mu_i, \theta_i, n_{ij})\) be the log-likelihood such that

\[
l(\mu_i, \theta_i, n_{ij}) = \log L(\theta_i, \mu_i \mid y_{ij}) = \frac{m_i}{2} \log \theta_i + \theta_i \sum_{j=1}^{m_i} n_{ij} y_{ij} \log \mu_i + \theta_i \sum_{j=1}^{m_i} n_{ij} (1 - y_{ij}) \log (1 - \mu_i) - \theta_i \sum_{j=1}^{m_i} n_{ij} y_{ij} - \theta_i \sum_{j=1}^{m_i} n_{ij} (1 - y_{ij}) \log (1 - y_{ij}) \tag{5.3}
\]

Differentiating the log-likelihood function (5.3) with respect to \(\theta_i\) and \(\mu_i\), we have the following equations

\[
\frac{\partial l}{\partial \mu_i} = \theta_i \left[ \sum_{j=1}^{m_i} n_{ij} y_{ij} \mu_i - \sum_{j=1}^{m_i} n_{ij} (1 - y_{ij}) \right]
\]

\[
\frac{\partial l}{\partial \theta_i} = \frac{m_i}{2\theta_i} + \sum_{j=1}^{m_i} n_{ij} \left[ y_{ij} \log \left( \frac{\mu_i}{y_{ij}} \right) + (1 - y_{ij}) \log \left( \frac{1 - \mu_i}{1 - y_{ij}} \right) \right]
\]

\[
= \frac{m_i}{2\theta_i} - \sum_{j=1}^{m_i} n_{ij} \mathcal{D}(y_{ij}, \mu_i)
\]

From these equations, we obtain the maximum likelihood estimators given by
\[ \hat{\mu}_i = \frac{\sum_{j=1}^{m_i} n_{ij} y_{ij}}{\sum_{j=1}^{m_i} n_{ij}} = \bar{y}_i \]

\[ \hat{\theta}_i = \frac{m_i}{2 \sum_{j=1}^{m_i} n_{ij} D(y_{ij}, \bar{y}_i)} \]

The inverse of the observed Fisher information matrix, \( I(\mu_i, \theta_i) \), gives the asymptotic variances of the MLEs. The second partial derivatives of the log likelihood function are given as follows:

\[
\frac{\partial^2 l}{\partial \mu_i^2} = -\theta_i \sum_{j=1}^{m_i} n_{ij} y_{ij} + 2\mu_i \hat{\theta}_i \sum_{j=1}^{m_i} n_{ij} y_{ij} - \mu_i^2 \theta_i \sum_{j=1}^{m_i} n_{ij}^2 \left[ \mu_i (1 - \mu_i) \right] \\
\frac{\partial^2 l}{\partial \theta_i^2} = -\frac{m_i}{2 \theta_i^2} 
\]

So, our \( 2 \times 2 \) information matrix, \( I(\hat{\mu}_i, \hat{\theta}_i) \) is given as

\[
I(\hat{\mu}_i, \hat{\theta}_i) = -\begin{pmatrix}
\frac{\partial^2 l}{\partial \mu_i^2} & \frac{\partial^2 l}{\partial \mu_i \partial \theta_i} \\
\frac{\partial^2 l}{\partial \theta_i \partial \mu_i} & \frac{\partial^2 l}{\partial \theta_i^2}
\end{pmatrix} = \begin{pmatrix}
\hat{\theta}_i \sum_{j=1}^{m_i} n_{ij} / V(\hat{\mu}_i) & 0 \\
0 & m_i / 2 \hat{\theta}_i^2
\end{pmatrix} \quad (5.4)
\]

The off-diagonal entries of our information matrix are zero. This implies that our MLEs, \( \hat{\mu}_i \) and \( \hat{\theta}_i \) are asymptotically independent. The variance-covariance matrix may be approximated, as we stated earlier, by the inverse of equation (5.4). Thus, the estimated standard errors for \( \hat{\mu}_i \) and \( \hat{\theta}_i \) are given by:
\[
\hat{\text{se}}(\hat{\mu}_i) = \sqrt{\frac{\bar{y}_i(1 - \bar{y}_i)}{\hat{\theta}_i \sum_{j=1}^{m} n_{ij}}}
\]

\[
\hat{\text{se}}(\hat{\theta}_i) = \sqrt{\frac{2\hat{\theta}_i^2}{m_i}}
\]

Approximate $100(1 - \alpha)%$ confidence intervals for $\mu_i$ and $\theta_i$ are given, respectively, as

\[
\hat{\mu}_i \pm z_{\alpha/2} \sqrt{\frac{\bar{y}_i(1 - \bar{y}_i)}{\hat{\theta}_i \sum_{j=1}^{m} n_{ij}}}
\]

\[
\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\frac{2\hat{\theta}_i^2}{m_i}}
\]

where $z_{\alpha/2}$ is the upper $100\alpha - \text{th}$ percentile of the standard normal distribution.

For the Bayesian estimation procedure, we do not have information on our parameters $\mu_i$ and $\theta_i$ so we use the noninformative Jeffrey’s prior distribution. Thus, the joint Jeffrey’s prior distribution for $\mu_i$ and $\theta_i$ is given by

\[
\pi(\mu_i, \theta_i) \propto \frac{\theta_i}{\mu_i(1 - \mu_i)}
\]  

\[(5.5)\]

We combine this joint prior distribution with the likelihood function $L(\mu_i, \theta_i)$ given by (5.2) to yield the joint posterior density given by $\pi(\mu_i, \theta_i | y_{ij}) \propto \pi(\mu_i, \theta_i) L(\mu_i, \theta_i)$. Thus,

\[
\pi(\mu_i, \theta_i | y_{ij}) = k \sqrt{\frac{1}{\theta_i \mu_i(1 - \mu_i)}} \prod_{j=1}^{m} \left[ \theta_i^{1/2} \left( \frac{\mu_i}{y_{ij}} \right)^{n_{ij} \theta_i y_{ij}} \left( \frac{1 - \mu_i}{1 - y_{ij}} \right)^{n_{ij} \theta_i (1 - y_{ij})} \right]
\]
where \( k \) is the normalizing constant that makes \( \pi(\mu_i, \theta_i | y_{ij}) \) a proper probability density function. The fully conditional distributions of the parameters \( \mu_i \) and \( \theta_i \) needed for the Gibbs sampling algorithm are given respectively by

\[
\pi(\mu_i | \theta_i, y_{ij}) \propto \sqrt{ \frac{1}{\mu_i (1 - \mu_i)} } \mu_i^{m_i} \left( 1 - \mu_i \right)^{\theta_i \sum_{j=1}^{m_i} n_{ij} y_{ij} - \theta_i n_{ij} (1 - y_{ij})} \]

\[
\pi(\theta_i | \mu_i, y_{ij}) \propto \sqrt{ \frac{1}{\theta_i^{m_i/2}} } \theta_i^{m_i/2} \mu_i^{m_i} \left( 1 - \mu_i \right)^{\theta_i \sum_{j=1}^{m_i} n_{ij} (1 - y_{ij})} \prod_{j=1}^{m_i} y_{ij} - \theta_i n_{ij} y_{ij} (1 - y_{ij})^{1 - y_{ij}} \left( 1 - y_{ij} \right)^{-\theta_i n_{ij} (1 - y_{ij})} \theta_i^{m_i - 1/2} \exp \left( -\theta_i \sum_{j=1}^{m_i} D(y_{ij}, \mu_i) \right) \]

Thus, \( \pi(\mu_i | \theta_i, y_{ij}) \sim \text{Beta} \left( \theta_i \sum_{j=1}^{m_i} n_{ij} y_{ij} + 1/2, \theta_i \sum_{j=1}^{m_i} n_{ij} (1 - y_{ij}) + 1/2 \right) \)

Thus, \( \pi(\theta_i | \mu_i, y_{ij}) \sim \text{Gamma} \left( \frac{m_i + 1}{2}, \sum_{j=1}^{m_i} n_{ij} D(y_{ij}, \mu_i) \right) \)

**Application to Toxicity Data**

The data for our analysis was obtained from a developmental toxicity study of ethylene glycol (EG) in mice conducted through the National Toxicology Program (NTP) (Tyl, Jones-Price, Marr, and C. Kimmel, 1983). The mice were randomized to either control or one of three dose levels (750, 1500 or 3000 mg/kg) of the ethylene glycol (EG). Table 7 gives the proportion \( \frac{x_i}{n_i} \) of fetuses \( x_i \) per litter \( n_i \) that were observed to have malformation, that is, the number of malformed fetuses out of the total fetuses in a litter. Table 8 gives summary statistics of the toxicity data. The summary includes the number of clusters
(litters) in each dose group, the number of fetal malformations in each dose group. From Table 8, the observed rates of fetal malformation increases with dose from 0.045 for the control group to 0.487 at the largest dose group. The cluster (litter) sizes decrease with dose.

Table 7. Proportion \((x_i/n_i)\) of fetuses with malformation \((x_i)\) out of total fetuses \((n_i)\).

<table>
<thead>
<tr>
<th>Dose (mg/kg)</th>
<th>0</th>
<th>750</th>
<th>1500</th>
<th>3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/9</td>
<td>3/11</td>
<td>9/10</td>
<td>5/9</td>
<td></td>
</tr>
<tr>
<td>0/15</td>
<td>4/15</td>
<td>6/12</td>
<td>9/12</td>
<td></td>
</tr>
<tr>
<td>1/13</td>
<td>2/11</td>
<td>5/14</td>
<td>4/12</td>
<td></td>
</tr>
<tr>
<td>1/11</td>
<td>1/13</td>
<td>11/15</td>
<td>7/12</td>
<td></td>
</tr>
<tr>
<td>1/16</td>
<td>2/13</td>
<td>6/12</td>
<td>9/12</td>
<td></td>
</tr>
<tr>
<td>1/16</td>
<td>4/13</td>
<td>7/13</td>
<td>4/9</td>
<td></td>
</tr>
<tr>
<td>0/13</td>
<td>0/15</td>
<td>0/5</td>
<td>7/16</td>
<td></td>
</tr>
<tr>
<td>0/12</td>
<td>0/12</td>
<td>3/12</td>
<td>1/10</td>
<td></td>
</tr>
<tr>
<td>0/11</td>
<td>1/13</td>
<td>8/13</td>
<td>12/14</td>
<td></td>
</tr>
<tr>
<td>0/9</td>
<td>0/14</td>
<td>3/16</td>
<td>12/15</td>
<td></td>
</tr>
<tr>
<td>0/11</td>
<td>3/19</td>
<td>3/13</td>
<td>12/16</td>
<td></td>
</tr>
<tr>
<td>2/15</td>
<td>1/10</td>
<td>10/15</td>
<td>5/12</td>
<td></td>
</tr>
<tr>
<td>0/12</td>
<td>6/14</td>
<td>7/14</td>
<td>5/13</td>
<td></td>
</tr>
<tr>
<td>0/14</td>
<td>2/14</td>
<td>1/13</td>
<td>8/13</td>
<td></td>
</tr>
<tr>
<td>1/15</td>
<td>1/15</td>
<td>3/14</td>
<td>3/15</td>
<td></td>
</tr>
<tr>
<td>0/11</td>
<td>1/13</td>
<td>2/15</td>
<td>9/14</td>
<td></td>
</tr>
<tr>
<td>1/15</td>
<td>5/14</td>
<td>6/9</td>
<td>2/10</td>
<td></td>
</tr>
<tr>
<td>2/12</td>
<td>4/15</td>
<td>2/8</td>
<td>5/11</td>
<td></td>
</tr>
<tr>
<td>2/13</td>
<td>1/15</td>
<td>0/3</td>
<td>5/9</td>
<td></td>
</tr>
<tr>
<td>1/13</td>
<td>2/7</td>
<td>1/7</td>
<td>3/13</td>
<td></td>
</tr>
<tr>
<td>0/16</td>
<td>3/12</td>
<td>0/9</td>
<td>3/12</td>
<td></td>
</tr>
<tr>
<td>1/15</td>
<td>1/9</td>
<td>4/12</td>
<td>2/12</td>
<td></td>
</tr>
<tr>
<td>0/15</td>
<td>1/11</td>
<td>0/12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0/14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8. Summary statistics for the toxicity data set of Table 7.

<table>
<thead>
<tr>
<th>Relative dose level</th>
<th>Control</th>
<th>0.25</th>
<th>0.50</th>
<th>1.00</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters</td>
<td>25</td>
<td>23</td>
<td>23</td>
<td>22</td>
<td>93</td>
</tr>
<tr>
<td>Sample size</td>
<td>333</td>
<td>298</td>
<td>266</td>
<td>271</td>
<td>1168</td>
</tr>
<tr>
<td>Malformation</td>
<td>15</td>
<td>48</td>
<td>97</td>
<td>132</td>
<td>292</td>
</tr>
<tr>
<td>Response rate (%)</td>
<td>0.045</td>
<td>0.161</td>
<td>0.365</td>
<td>0.487</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Detecting Overdispersion on Toxicity Data

The maximum likelihood fit has estimate $\hat{\mu}_i$ equal to the sample proportion of malformations for all fetuses from litters in that dose group. The standard errors for group $i$ is given by $\sqrt{(\hat{\mu}_i \times (1 - \hat{\mu}_i)/\sum_j n_{ij})}$. For litter $j$ in dose group $i$, $n_{ij} \times \hat{\mu}_i$ is a fitted number of malformed fetuses and $n_{ij} \times (1 - \hat{\mu}_i)$ is a fitted number of non-malformed fetuses.

A comparison of these fitted values with the observed counts of malformations and non-malformations in the $N=93$ litters using the Pearson statistic gives $\chi^2 = 177.041$ with degrees of freedom, $df=93 - 4 = 89$, There is considerable evidence of overdispersion, that is,

$$\hat{\phi} = \frac{\chi^2}{N - p} = \frac{177.041}{89} = 1.99$$

Discussion

From Tables 9 and 10, we see that the estimates of the parameters of the double binomial distribution seem equal using both the maximum likelihood and Bayesian estimation methods for our toxicity data. The confidence intervals for the Bayesian estimates and the maximum likelihood estimates in all dose groups have approximately equal length albeit the lower and upper limits differ in some cases. The control group, for example, has $(0.024, 0.066)$ and $(0.486, 1.696)$ as confidence intervals for $\mu$ and $\theta$ respectively while the Bayesian confidence intervals for $\mu$ and $\theta$ are respectively $(0.027, 0.071)$ and $(0.571, 1.771)$. 

51
Table 9. Bayesian estimates based on toxicity data set.

<table>
<thead>
<tr>
<th>Dose Group</th>
<th>Parameter</th>
<th>Bayesian Estimation</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
<td>0.046</td>
<td>(0.027, 0.071)</td>
<td>0.012</td>
</tr>
<tr>
<td>Control</td>
<td>$\theta$</td>
<td>1.09</td>
<td>(0.571, 1.771)</td>
<td>0.305</td>
</tr>
<tr>
<td>0.25</td>
<td>$\mu$</td>
<td>0.162</td>
<td>(0.112, 0.219)</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.661</td>
<td>(0.335, 1.10)</td>
<td>0.194</td>
</tr>
<tr>
<td>0.50</td>
<td>$\mu$</td>
<td>0.367</td>
<td>(0.260, 0.488)</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.282</td>
<td>(0.148, 0.468)</td>
<td>0.082</td>
</tr>
<tr>
<td>1.00</td>
<td>$\mu$</td>
<td>0.487</td>
<td>(0.384, 0.586)</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.370</td>
<td>(0.183, 0.618)</td>
<td>0.112</td>
</tr>
</tbody>
</table>

Table 10. Maximum likelihood estimates based on toxicity data set.

<table>
<thead>
<tr>
<th>Dose Group</th>
<th>Parameter</th>
<th>Maximum Likelihood Estimation</th>
<th>95% C.I</th>
<th>SD of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
<td>0.045</td>
<td>(0.024, 0.066)</td>
<td>0.011</td>
</tr>
<tr>
<td>Control</td>
<td>$\theta$</td>
<td>1.09</td>
<td>(0.486, 1.696)</td>
<td>0.309</td>
</tr>
<tr>
<td>0.25</td>
<td>$\mu$</td>
<td>0.161</td>
<td>(0.110, 0.213)</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.656</td>
<td>(0.277, 1.035)</td>
<td>0.193</td>
</tr>
<tr>
<td>0.50</td>
<td>$\mu$</td>
<td>0.365</td>
<td>(0.255, 0.474)</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.281</td>
<td>(0.148, 0.468)</td>
<td>0.083</td>
</tr>
<tr>
<td>1.00</td>
<td>$\mu$</td>
<td>0.487</td>
<td>(0.389, 0.585)</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>$\theta$</td>
<td>0.372</td>
<td>(0.152, 0.591)</td>
<td>0.112</td>
</tr>
</tbody>
</table>

It can be observed that the lower and upper limits are different, yet the length of the confidence intervals are approximately equal.

With non-informative priors for our Bayesian estimation, it is not surprising that the MLEs may be similar to the Bayesian estimates. With the exception of the control group, all dose groups had a dispersion parameter that was less than 1.

Figures 4, 5, 6 and 7 give the trace plots and the histograms of our posterior MCMC samples for $\mu$ and $\theta$ for the control group, the 0.25 group, the 0.50 group and the 1.00 group respectively. The trace plots are used to monitor the convergence of our samples.
In this dissertation, we have been able to use Efron’s double binomial distribution to model correlated binary data in the presence of overdispersion. Specifically, we modeled jointly the probability of success and the intra-cluster correlation. The parameters for Efron’s model have been determined using maximum likelihood and Bayesian estimation procedures. We have shown the effectiveness of this model in dealing with overdispersed binary data through simulated and real data application.
Figure 5. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 0.25 group.
Figure 6. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 0.50 group.
Figure 7. Trace plots and histograms of the posterior MCMC samples for $\mu$ and $\theta$ of 1.00 group.
Bibliography


