Investigating the General Guidelines for Modeling Extra-Dispersed Proportion Data Based on Some Completing Proportion Models

Krishna K. Saha*

Abstract
Proportion data occurring in many applied fields exhibit extra-variation predicted by a simple binomial model. For modeling extra-dispersed proportions, many authors have introduced several alternative extra-dispersed proportion models. With real-life data, a practical problem is deciding how to select one out of a wide variety of candidate models. In this paper, we aim to solve this problem in terms of real-life data occurring in a toxicological study. We discuss the model selection issues using a variety of standard model selection approaches. Moreover, a parametric bootstrap approach of model evaluation using a Mahalanobis squared distance proposed by Allcroft and Glasbey (*Statistical Modelling*, 2003) is applied.

Key Words: Beta-binomial, correlated binomial model, double binomial model, extra-dispersion, toxicological data.

1. Introduction

Proportion data often arise in a wide variety of disciplines. These data often show variation significantly larger or smaller than that predicted by a simple binomial model. This would happen when there is a possible correlation in the occurrence of the events, which indicates that an extension of the simple binomial model is necessary. In studies where the experimental unit is a litter, it has been observed (Weil, 1970) that an inherent characteristic of data from these types of studies is the 'litter effect', i.e., there is a tendency of littermates to respond more alike than animals from different litters. This litter effect is also known as the extra-dispersion or the intra-litter correlation or the intra-class correlation. In some binary-data situations it is interpreted as 'heritability of a dichotomous trait' (see Elston, 1977; Crowder, 1982). For example, a set of toxicological data (Paul, 1982) provided in Table 1 refers to live fetuses in a litter affected by treatment, and the number of live fetuses, for each of four dose groups: control (C), low dose (L), medium dose (M), and high dose (H). The observed variances for all four groups C, L, M, and H are 0.4465, 0.2435, 1.0472, and 0.6186, whereas the respective predicted variances by a binomial model are 0.1465, 0.1617, 0.5100, and 0.2960. The discrepancy between the observed variances and those predicted by the binomial model indicates over-dispersion in the proportion data sets. It is, therefore, important to analyze the extra dispersed proportions by an extended binomial distribution that takes into account the variability shown in the proportion data occurring in biological investigations.

Several over-dispersed models for analyzing proportions have been used by many authors (Lindsey and Altham, 1998; Saha and Paul, 2005). Williams (1975) introduced the beta-binomial model which is a mixture of binomial and beta distributions. Many authors have used this distribution for analyzing extra proportion data (see, for example, Crowder, 1978; Donvan et al., 1994; Gibson and Austin, 1996; Kleinman, 1973; Otake and Prentice, 1984; and Paul and Islam, 1995). Kupper and Haseman (1978) developed the correlated binomial distribution by taking into account the correlation between the siblings

*Department of Mathematical Sciences, Central Connecticut State University, 1615 Stanley Street, New Britain, CT 06050, USA, E-mail: sahakrk@ccsu.edu

Table 1: Toxicological data from Paul (1982). (i) Number of live foetuses affected by treatment. (ii) Total number of foetuses.

Dose Groups																												
Control, C	(i)	1	1	4	0	0	0	0	0	1	0	2	0	5	2	1	2	0	0	1	0	0	0	0	3	2	4	0
	(ii)	12	7	6	6	7	8	10	7	8	6	11	7	8	9	2	7	9	7	11	10	4	8	10	12	8	7	8
Low dose, L	(i)	0	1	1	0	2	0	1	0	1	0	0	3	0	0	1	5	0	0	3								
	(ii)	5	11	7	9	12	8	6	7	6	4	6	9	6	7	5	9	1	6	9								
Medium dose, M	(i)	2	3	2	1	2	3	0	4	0	0	4	0	0	6	6	5	4	1	0	3	6						
	(ii)	4	4	9	8	9	7	8	9	6	4	6	7	3	13	6	8	11	7	6	10	6						
High dose, H	(i)	1	0	1	0	1	0	1	1	2	0	4	1	1	4	2	3	1										
	(ii)	9	10	7	5	4	6	3	8	5	4	4	5	3	8	6	8	6										

in the same litter ignoring the interlitter variation. Altham (1978) proposed the additive generalized binomial model based on Lancaster's definition of no second- or higher- order interaction. This model is identical to the correlated binomial model of Kupper and Haseman (1978). Altham (1978) also developed a two-parameter multiplicative binomial model by drawing an analogy to a model in a 2M contingency table with no second- and higher-order interactions. Efron (1986) introduced a double binomial model obtained based on the double exponential family. Morel and Nagaraj (1993) proposed a finite mixture model for handling the extra variation in the binary outcome data. Paul (1985) derived the correlated beta-binomial model for handling the correlation as well as the extra variation in the binary outcome data. In addition, the zero-inflated binomial model as well as the zero-inflated beta-binomial model can be used to analyze the over-dispersed proportion data (see, Deng and Paul, 2005). Due to its simplicity, many authors have used the beta-binomial distribution for the analysis of over-dispersed proportion data. No work has been done regarding a theoretical comparison for the behavior of these models. Little is known about an application-based comparison of some of the models. Altham (1978) compared the beta binomial, correlated binomial and multiplicative binomial models and preferred to use both the correlated binomial and multiplicative binomial models over the beta-binomial model, whereas Paul (1982) studied the comparison among these three models in terms of the C(alpha) test of Tarone (1979) and concluded that the beta-binomial model is superior to the correlated binomial and the multiplicative binomial models. Saha (2011) extended the comparison study with these three models by adding the double binomial model. Based on the standard goodness-of-fit approaches he showed that no unique model among these four models can be recommended. In this study, we include all eight models that are candidates for the analysis of any real-life over-dispersed proportions occurring in biological investigations.

The purpose of this article is to conduct a comparison study of the well-known competing extra-dispersed proportion models for the analysis of the proportion data occurring in toxicological study described above. In applied fields, one could be wonder the use of the most suitable model in a particular case so we aim to reducing this problem in this study. In addition, we aim to detect the differences among the competing models for proportions.

In the next section, we review all eight competing extra-dispersed proportion models for analyzing proportions. Section 3 discuss the maximum likelihood methods for the estimates of the parameters for these models. The standard model selection approaches as well as a parametric bootstrap approach of model evaluation using a Mahalanobis squared distance proposed by Allcroft and Glasbey (*Statistical Modelling*, 2003) are discussed in Section 4. Section 5 shows whether the researcher in applied fields can really identify the underlying distribution uniquely from toxicological data. A discussion can be found in

Section 6.

2. The Competing Models for Proportion Data

Below we briefly discuss the probability mass functions and their properties of all eight competing parametric models for the over-dispersed proportion data.

2.1 The Binomial Model

The probability mass function of the binomial model is given by

$$f(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

for $y = 0, 1, 2, \ldots, n$, and $0 \leq \pi \leq 1$. The mean and variance of the binomial variable Y are E$(Y) = n\pi$ and var$(Y) = n\pi(1-\pi)$, respectively.

2.2 The Beta-Binomial (BB) Model

The probability mass function of the beta-binomial model is given by

$$f(y|\pi, \phi) = \binom{n}{y} \frac{\prod_{j=0}^{y-1}[(1-\phi)\pi + j\phi] \prod_{j=0}^{n-y-1}[(1-\pi)(1-\phi) + j\phi]}{\prod_{j=0}^{n-1}[(1-\phi) + j\phi]}$$

for $y = 0, 1, 2, \ldots, n$ and $\phi > 0$. The mean and variance of the beta-binomial variable Y are E$(Y) = n\pi$ and var$(Y) = n_i\pi(1-\pi)\{1 + (n_i - 1)\phi\}$, respectively.

2.3 The Correlated Binomial (CB) Model

The probability mass function of the correlated binomial model is given by

$$f(y|\pi, \theta) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \left[1 + \frac{\theta}{2\pi^2(1-\pi)^2}\{(y - n\pi)^2 + y(2\pi - 1) - n\pi^2\}\right],$$

for $y = 0, 1, 2, \ldots, n$. The mean and variance of the correlated-binomial response Y are E$(Y) = n\pi$ and var$(Y) = n_i\pi(1-\pi) + n_i(n_i - 1)\theta$, respectively.

2.4 The Multiplicative Binomial (MB) Model

The probability mass function of the multiplicative binomial model is given by

$$f(y|\pi, \gamma) = \binom{n}{y} \frac{\pi^y (1-\pi)^{n-y} \gamma^{y(n-y)}}{k(\pi, \gamma, n)}$$

for $y = 0, 1, 2, \ldots, n$, and $\gamma > 0$, where $k(\pi, \gamma, n)$ is the intractable factor as

$$k(\pi, \gamma, n) = \sum_{y=0}^{n} \binom{n}{y} \pi^y (1-\pi)^{n-y} \gamma^{y(n-y)}.$$

2.5 The Double Binomial (DB) Model

The probability mass function of the double binomial model is given by

$$f(y|\pi,\psi) \quad = \quad \begin{pmatrix} n \\ y \end{pmatrix} \frac{n^{n\psi}\pi^{y\psi}(1-\pi)^{(n-y)\psi}y^y(n-y)^{n-y}}{n^n y^{y\psi}(n-y)^{(n-y)\psi}c^*(\pi,\psi,n)}$$

for $y = 0, 1, 2, \ldots, n$; $\psi > -1$; and $c(\pi,\psi,n)$ is the intractable factor as

$$c(\pi,\psi,n) = \sum_{y=0}^{n} \begin{pmatrix} n \\ y \end{pmatrix} \frac{n^{n\psi}\pi^{y\psi}(1-\pi)^{(n-y)\psi}y^y(n-y)^{n-y}}{n^n y^{y\psi}(n-y)^{(n-y)\psi}}.$$

2.6 The Finite Mixture (FM) Model

The probability mass function of the finite mixture model is given by

$$f(y|\pi,\nu) \quad = \quad \pi \begin{pmatrix} n \\ y \end{pmatrix} [\nu + (1-\nu)\pi]^y [1-\nu-(1-\nu)\pi]^{n-y} + (1-\pi) \begin{pmatrix} n \\ y \end{pmatrix} [(1-\nu)\pi]^y [1-(1-\nu)\pi]^{n-y}$$

for $y = 0, 1, 2, \ldots, n$ and $0 < \nu < 1$. The mean and variance of the finite mixture response Y are E$(Y) = n\pi$ and var$(Y) = n_i\pi(1-\pi)(1-\nu)$, respectively.

2.7 The Zero-inflated Binomial (ZIB) Model

The probability mass function of the zero-inflated model is given by

$$f(y|\pi,\lambda) = \begin{cases} \lambda + (1-\lambda)(1-\pi)^n & \text{if } y = 0 \\ (1-\lambda) \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y} & \text{if } y = 1, \cdots, n, \end{cases}$$

for $0 < \pi < 1$ and $0 < \lambda < 1$. The mean and variance of the zero inflated binomial response Y are E$(Y) = n\pi(1-\lambda)$ and var$(Y) = n_i\pi(1-\lambda)(1-\pi+n\pi\lambda)$, respectively.

2.8 The Zero-inflated Beta-Binomial (ZIBB) Model

The probability mass function of the zero-inflated beta-binomial model is given by

$$f(y|\pi,\delta,\lambda) = \begin{cases} \lambda + (1-\lambda)\frac{\prod_{r=0}^{n-1}[1-\pi+r\delta]}{\prod_{r=0}^{n-1}[1+r\delta]} & \text{if } y = 0 \\ (1-\lambda) \begin{pmatrix} n \\ y \end{pmatrix} \frac{\prod_{r=0}^{y-1}[\pi+r\delta]\prod_{r=0}^{n-y-1}[1-\pi+r\delta]}{\prod_{r=0}^{n-1}[1+r\delta]} & \text{if } y = 1, \cdots, n, \end{cases}$$

for $0 < \pi < 1$, $\delta > 0$, and $0 < \lambda < 1$. The mean and variance of the zero inflated beta-binomial response Y are E$(Y) = n\pi(1-\lambda)$ and var$(Y) = n_i\pi(1-\lambda)(1-\pi)\frac{1+n\delta}{1+\delta} + \lambda(1-\lambda)n^2\pi^2)$, respectively.

2.9 The Correlated Beta-Binomial (CBB) Model

The probability mass function of the correlated beta-binomial model is given by

$$f(y|\pi,\tau,\omega) \quad = \quad \begin{pmatrix} n \\ y \end{pmatrix} \frac{\prod_{r=0}^{y-1}[\pi+r\tau]\prod_{r=0}^{n-y-1}[1-\pi+r\tau]}{\prod_{r=0}^{n-1}[1+r\tau]} \left\{1 + \frac{\omega}{2}g(y;n,\pi,\tau)\right\}$$

for $y = 0, 1, 2, \ldots, n$ and $0 < \tau < 0$, and

$$g(y;n,\pi,\tau) = \frac{(1-\tau)[(y-n\pi)^2 + y(2\pi-1) - n\pi^2] - n(n-1)\pi(1-\pi)\tau}{\pi(1-\pi) - \tau(1-\tau) - \tau[\tau y^2 - y(n\tau - 2\pi + 1) + n(\tau - \pi)]}.$$

3. Estimation of the Model Parameters

In this section, we discuss the maximum likelihood methods to estimate the model parameters for all the models described above. It can be easily seen from below that the estimates of the dispersion parameters for all models do not have closed-forms, which need to be obtained either by maximizing the log-likelihoods or by solving the estimating equations iteratively. However, for some models the estimates of the proportion parameters do have closed-forms.

3.1 The Maximum BB Likelihood Estimator

Let $Y_1, \ldots, Y_m$ be a random sample from the beta-binomial distribution. Then the log-likelihood, apart from a constant, can be written as

$$l = \sum_{i=1}^{m} \left[\sum_{j=0}^{y_i-1} \ln\{(1-\phi)\pi + j\phi\} + \sum_{j=0}^{n_i-y_i-1} \ln\{(1-\pi)(1-\phi) + j\phi\} - \sum_{j=0}^{n_i-1} \ln\{1-\phi+j\phi\} \right].$$

The maximum likelihood estimates of π and θ can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations:

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{m} \left[\sum_{j=0}^{y_i-1} \frac{1-\phi}{\pi(1-\phi)+j\phi} - \sum_{j=0}^{n_i-y_i-1} \frac{(1-\phi)}{(1-\pi)(1-\phi)+j\phi} \right] = 0, \text{ and}$$

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

(see also Saha and Paul, 2005).

3.2 The Maximum CB Likelihood Estimator

Let $Y_1, \ldots, Y_m$ be a random sample from the correlated binomial distribution. Then the log-likelihood, apart from a constant, can be written as

$$l = \sum_{i=1}^{m} \left[\{y_i \ln \pi + (n_i - y_i) \ln(1-\pi)\} + \ln\left\{1 + \frac{\rho}{2\pi(1-\pi)} h_1(y_i, n_i, \pi)\right\} \right],$$

where

$$h_1(y_i, n_i, \pi) = (y_i - n_i\pi)^2 + y_i(2\pi - 1) - n_i\pi^2.$$

The maximum likelihood estimates of π and ρ can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations:

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{m} \left[\frac{y_i - n_i\pi}{\pi(1-\pi)} + \frac{2\rho(2\pi-1)h_1(y_i,n_i,\pi) - 2\rho\pi(1-\pi)h_2(y_i,n_i,\pi)}{\pi(1-\pi)\{2\pi(1-\pi)+\rho\}h_1(y_i,n_i,\pi)} \right] = 0, \text{ and}$$

$$\frac{\partial l}{\partial \rho} = \sum_{i=1}^{m} \left[\frac{h_1(y_i,n_i,\pi)}{2\pi(1-\pi)+\rho h_1(y_i,n_i,\pi)} \right] = 0,$$

where

$$h_2(y_i, n_i, \pi) = (n_i y_i - n_i^2\pi - y_i + n_i\pi).$$

Note that one should impose the restriction on ρ given in Section 2.3 in order to obtain the valid estimates of the parameters π and ρ.

3.3 The Maximum FM Likelihood Estimator

Let $Y_1, \ldots, Y_m$ be a random sample from the finite mixture distribution. Then the log-likelihood, apart from a constant, can be written as

$$l = \sum_{i=1}^{m} \ln\left[\pi q_1(y_i;\pi,\nu) + (1-\pi)q_2(y_i;\pi,\nu)\right],$$

where

$$\begin{aligned} q_1(y_i;\pi,\nu) &= \binom{n_i}{y_i}[\nu+(1-\nu)\pi]^{y_i}[1-\nu-(1-\nu)\pi]^{n_i-y_i}, \text{ and} \\ q_2(y_i;\pi,\nu) &= \binom{n_i}{y_i}[(1-\nu)\pi]^{y_i}[1-(1-\nu)\pi]^{n_i-y_i}. \end{aligned}$$

The maximum likelihood estimates of π and ν can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations:

$$\begin{aligned} \frac{\partial l}{\partial \pi} &= \sum_{i=1}^{m} \frac{1}{f(y_i|\pi,\nu)} \Bigg[q_1(y_i;\pi,\nu) + \pi(1-\nu)q_1(y_i;\pi,\nu)\left\{\frac{y_i}{\nu+(1-\nu)\pi} - \frac{n_i-y_i}{1-\nu-(1-\nu)\pi}\right\} \\ &\qquad -q_2(y_i;\pi,\nu) + \frac{(1-\pi)[y_i-n_i\pi(1-\nu)]}{\pi[1-\pi(1-\nu)]} q_2(y_i;\pi,\nu)\Bigg] = 0, \text{ and} \\ \frac{\partial l}{\partial \nu} &= \sum_{i=1}^{m} \frac{1}{f(y_i|\pi,\nu)} \Bigg[\pi(1-\pi)q_1(y_i;\pi,\nu)\left\{\frac{y_i}{\nu+\pi(1-\nu)} - \frac{n_i-y_i}{1-[\nu+\pi(1-\nu)]}\right\} \\ &\qquad - \frac{(1-\pi)[y_i-n_i\pi(1-\nu)]}{(1-\nu)[1-\pi(1-\nu)]} q_2(y_i;\pi,\nu)\Bigg] = 0. \end{aligned}$$

3.4 The Maximum ZIB Likelihood Estimator

Let $Y_1, \ldots, Y_m$ be a random sample from the zero inflated binomial distribution. Using $\varphi = \lambda/(1-\lambda)$ the log-likelihood, apart from a constant, can be written as

$$l = \sum_{i=1}^{m} \left[-\ln(1+\varphi) + I_{y_i=0}\ln\{\varphi+(1-\pi)^{n_i}\} + I_{y_i>0}\{y_i\ln\pi + (n_i-y_i)\ln(1-\pi)\}\right].$$

The maximum likelihood estimates of φ and π can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations:

$$\begin{aligned} \frac{\partial l}{\partial \varphi} &= \sum_{i=1}^{m}\left[-\frac{1}{1+\varphi} + \frac{I_{y_i=0}}{\varphi+(1-\pi)^{n_i}}\right] = 0, \text{ and} \\ \frac{\partial l}{\partial \pi} &= \sum_{i=1}^{m}\left[\frac{n_i(1-\pi)^{n_i-1}}{\varphi+(1-\pi)^{n_i}} I_{y_i=0} + \frac{y_i-n_i\pi}{\pi(1-\pi)} I_{y_i>0}\right] = 0. \end{aligned}$$

3.5 The Maximum ZIBB Likelihood Estimator

Let $Y_1, \ldots, Y_m$ be a random sample from the zero inflated beta-binomial distribution. Using $\varphi = \lambda/(1-\lambda)$ the log-likelihood, apart from a constant, can be written as

$$\begin{aligned} l &= \sum_{i=1}^{m}\Bigg[-\ln(1+\varphi) + I_{y_i=0}\ln\{\varphi+s(n_i;\pi,\delta)\} + I_{y_i>0}\sum_{j=0}^{y_i-1}\ln\{\pi+j\delta\} \\ &\qquad + I_{y_i>0}\sum_{j=0}^{n_i-y_i-1}\ln\{1-\pi+j\delta\} - I_{y_i>0}\sum_{j=0}^{n_i-1}\ln\{1+j\delta\}\Bigg], \end{aligned}$$

where

$$s(n_i;\pi,\delta)=\frac{\prod_{r=0}^{n_i-1}[1-\pi+r\delta]}{\prod_{r=0}^{n_i-1}[1+r\delta]}.$$

The maximum likelihood estimates of φ, π, and δ can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations:

$$\begin{aligned}
\frac{\partial l}{\partial\varphi} &= \sum_{i=1}^{m}\left[-\frac{1}{1+\varphi}+\frac{I_{y_i=0}}{\varphi+s(n_i;\pi,\delta)}\right]=0,\\
\frac{\partial l}{\partial\pi} &= \sum_{i=1}^{m}\left[\frac{-I_{y_i=0}}{s(n_i;\pi,\delta)[\varphi+s(n_i;\pi,\delta)]}\sum_{r=0}^{n_i-1}[1-\pi+r\delta]+I_{y_i>0}\sum_{r=0}^{y_i-1}\frac{1}{\pi+r\delta}\right.\\
&\qquad\left.-I_{y_i>0}\sum_{r=0}^{n_i-y_i-1}\frac{1}{1-\pi+r\delta}\right]=0,\ \text{ and}\\
\frac{\partial l}{\partial\delta} &= \sum_{i=1}^{m}\left[\frac{-I_{y_i=0}}{s(n_i;\pi,\delta)[\varphi+s(n_i;\pi,\delta)]}\sum_{r=0}^{n_i-1}[1-\pi+r\delta]+I_{y_i>0}\sum_{r=1}^{y_i-1}\frac{r}{\pi+r\delta}\right.\\
&\qquad\left.+I_{y_i>0}\sum_{r=0}^{n_i-y_i-1}\frac{r}{1-\pi+r\delta}-I_{y_i>0}\sum_{r=0}^{n_i-1}\frac{r}{1+r\delta}\right].
\end{aligned}$$

3.6 The Maximum BCB Likelihood Estimator

Let $Y_1,\ldots, Y_m$ be a random sample from the correlated beta-binomial distribution. Then the log-likelihood, apart from a constant, can be written as

$$l = \sum_{i=1}^{m}\left[\sum_{r=0}^{y_i-1}\ln\{\pi+r\tau\}+\sum_{r=0}^{n_i-y_i-1}\ln\{1-\pi+r\tau\}-\sum_{j=0}^{n_i-1}\ln\{1+r\tau\}+\ln G(y_i,n_i,\pi,\tau,\omega)\right],$$

where

$$G(y_i,n_i,\pi,\tau,\omega)=1+\frac{\omega}{2}g(y_i,n_i,\pi,\tau).$$

The maximum likelihood estimates of π, τ, and ω can be obtained by maximizing l or alternatively by solving the estimating equations:

$$\begin{aligned}
\frac{\partial l}{\partial\pi} &= \sum_{i=1}^{m}\left[\sum_{r=0}^{y_i-1}\frac{1}{\pi+r\tau}-\sum_{r=0}^{n_i-y_i-1}\frac{1}{1-\pi+r\tau}+\frac{\omega}{G(y_i,n_i,\pi,\tau,\omega)}\right.\\
&\qquad\left.\times\left\{\frac{t_1(y_i,n_i,\pi,\tau)}{g_2(y_i,n_i,\pi,\tau)}-\frac{g_1(y_i,n_i,\pi,\tau)t_2(y_i,n_i,\pi,\tau)}{g_2^2(y_i,n_i,\pi,\tau)}\right\}\right]=0,\\
\frac{\partial l}{\partial\tau} &= \sum_{i=1}^{m}\left[\sum_{r=1}^{y_i-1}\frac{r}{\pi+r\tau}+\sum_{r=0}^{n_i-y_i-1}\frac{r}{1-\pi+r\tau}-\sum_{r=0}^{n_i-1}\frac{r}{1+r\tau}+\frac{\omega}{G(y_i,n_i,\pi,\tau,\omega)}\right.\\
&\qquad\left.\times\left\{\frac{u_1(y_i,n_i,\pi,\tau)}{g_2(y_i,n_i,\pi,\tau)}-\frac{g_1(y_i,n_i,\pi,\tau)u_2(y_i,n_i,\pi,\tau)}{g_2^2(y_i,n_i,\pi,\tau)}\right\}\right]=0,\ \text{ and}\\
\frac{\partial l}{\partial\omega} &= \sum_{i=1}^{m}\frac{g(y_i,n_i,\pi,\tau)}{2G(y_i,n_i,\pi,\tau,\omega)}=0,
\end{aligned}$$

simultaneously, where

$$t_1(y_i,n_i,\pi,\tau) = (1-\tau)[n_i(n_i-1)\tau(2\pi-1)-2(y_i-n_i\pi)+2y_i-2n_i\pi]$$

$$
\begin{aligned}
t_2(y_i, n_i, \pi, \tau) &= 1 - 2\pi - \tau(2y_i - n_i) \\
g_1(y_i, n_i, \pi, \tau) &= (1-\tau)[(y - n\pi)^2 + y(2\pi - 1) - n\pi^2] - n(n-1)\pi(1-\pi)\tau \\
g_2(y_i, n_i, \pi, \tau) &= \pi(1-\pi) - \tau(1-\tau) - \tau[\tau y^2 - y(n\tau - 2\pi + 1) + n(\tau - \pi)] \\
u_1(y_i, n_i, \pi, \tau) &= n_i\pi^2 - n_i(n_i - 1)\pi(1-\pi) - (y_i - n_i\pi)^2 - y_i(2\pi - 1) \ \text{ and} \\
u_2(y_i, n_i, \pi, \tau) &= y_i(n_i\tau - 2\pi + 1) - n_i(\tau - \pi) - \tau(y_i^2 - n_i y_i + n_i) - 1 + 2\tau - \tau y_i^2.
\end{aligned}
$$

Note that the maximum likelihood estimates of π, τ, and ω must be obtained using the restriction on ω given in Section 2.9 to avoid the negative estimated probability based on the correlated beta-binomial model.

3.7 The Estimators of the MB Model Parameters

Let $Y_1, \ldots, Y_m$ be a random sample from the multiplicative binomial distribution. Then the log-likelihood, apart from a constant, can be written as

$$
l = \sum_{i=1}^{m} \left[y_i \ln \pi + (n_i - y_i) \ln(1-\pi) + y_i(n_i - y_i) \ln \gamma + \ln k(\pi, \gamma, n_i) \right].
$$

The maximum likelihood estimates of π and γ can be obtained by maximizing l or alternatively by solving the estimating equations:

$$
\begin{aligned}
\frac{\partial l}{\partial \pi} &= \sum_{i=1}^{m} \left[\frac{y_i}{\pi} - \frac{n_i - y_i}{1-\pi} + \frac{1}{k(\pi, \gamma, n_i)} \sum_{y_i=0}^{n_i} f(y_i|\pi, \gamma) \left\{ \frac{y_i}{\pi} - \frac{n_i - y_i}{1-\pi} \right\} \right] = 0, \ \text{ and} \\
\frac{\partial l}{\partial \gamma} &= \sum_{i=1}^{m} \left[\frac{y_i(n_i - y_i)}{\gamma} + \frac{1}{k(\pi, \gamma, n_i)} \sum_{y_i=0}^{n_i} f(y_i|\pi, \gamma) \frac{y_i(n_i - y_i)}{\gamma} \right] = 0.
\end{aligned}
$$

3.8 The Estimators of the DB Model Parameters

Let $Y_1, \ldots, Y_m$ be a random sample from the double binomial distribution. Then the log-likelihood, apart from a constant, can be written as

$$
l = \sum_{i=1}^{m} \left[y_i \psi \ln\left(\frac{\pi}{y_i}\right) + (n_i - y_i)\psi \ln\left(\frac{1-\pi}{n_i - y_i}\right) + \ln c(\pi, \psi, n_i) \right].
$$

The maximum likelihood estimates of π and γ can be obtained by maximizing l or alternatively by solving the estimating equations:

$$
\begin{aligned}
\frac{\partial l}{\partial \pi} &= \sum_{i=1}^{m} \left[\frac{\psi y_i}{\pi} - \frac{(n_i - y_i)\psi}{1-\pi} + \frac{1}{c(\pi, \psi, n_i)} \sum_{y_i=0}^{n_i} f(y_i|\pi, \psi) \left\{ \frac{\psi y_i}{\pi} - \frac{(n_i - y_i)\psi}{1-\pi} \right\} \right] = 0, \ \text{ and} \\
\frac{\partial l}{\partial \psi} &= \sum_{i=1}^{m} \Bigg[n_i \ln n_i + y_i \ln\left(\frac{\pi}{y_i}\right) + (n_i - y_i) \ln\left(\frac{1-\pi}{n_i - y_i}\right) + \frac{1}{c(\pi, \psi, n_i)} \sum_{y_i=0}^{n_i} f(y_i|\pi, \psi) \\
&\qquad \times \left\{ n_i \ln n_i + y_i \ln\left(\frac{\pi}{y_i}\right) + (n_i - y_i) \ln\left(\frac{1-\pi}{n_i - y_i}\right) \right\} \Bigg] = 0.
\end{aligned}
$$

Table 2: The estimates of the parameters and their standard errors for all five competing models for Data in Table 1

Models	Para	Control Group		Low Group		Medium Group		High Group	
		Est	SE	Est	SE	Est	SE	Est	SE
Binomial	π	0.1349	0.0233	0.1353	0.0297	0.3444	0.0387	0.2277	0.0417
BB	π	0.1404	0.0380	0.1272	0.0373	0.3505	0.0678	0.2387	0.0548
	ϕ	0.2148	0.0957	0.1054	0.0813	0.3155	0.1091	0.1132	0.0944
CB	π	0.1376	0.0302	0.1351	0.0368	0.3296	0.0521	0.2387	0.0502
	ϕ	0.1133	0.0346	0.0786	0.0488	0.1269	0.0388	0.1040	0.0802
MB	π	0.3216	0.0594	0.1437	0.0796	0.4281	0.0352	0.3430	0.0635
	γ	0.7980	0.0467	0.9861	0.1181	0.8404	0.0394	0.8172	0.0708
DB	π	0.0633	0.0671	0.1178	0.0481	0.3145	0.0835	0.2145	0.0616
	ψ	-0.7674	0.1535	-0.4773	0.2703	-0.7125	0.1248	-0.4586	0.2324
FM	π	0.1472	0.0392	0.1262	0.0365	0.3480	0.0624	0.2333	0.0535
	ν	0.4754	0.1135	0.3123	0.1238	0.4496	0.0855	0.3375	0.1361
ZIB	λ	0.4429	0.1165	0.3343	0.1816	0.2430	0.1042	0.0981	0.1256
	π	0.2372	0.0462	0.1929	0.0534	0.4301	0.0480	0.2581	0.0569
CBB	π	0.1404	0.0381	0.1164	0.0427	0.3511	0.0686	0.2392	0.0557
	τ	0.3024	0.4651	0.3787	0.3702	0.3665	0.2657	0.2110	0.3281
	ω	-0.0207	0.3195	-0.3384	0.5877	0.0793	0.1533	-0.0643	0.2209

4. The Model Selection Criteria

4.1 Standard Approaches for Model Selection

The standard approaches, such as the likelihood ratio tests, the modified Neyman-Pearson likelihood ratio tests (Cox, 1961), the exponential combinations of competing models (Atkinson, 1970), Akaike's Information Criteria (Akaike, 1973), and Bayesian Information Criteria (Scwarz, 1978) can be usually used to model comparison. Note that these approaches are more applicable when the models being assessed share a common likelihood family, that is, models are nested. Here we briefly review some of the methods as follows.

Lindsey (1974) used the log-likelihood method for model selection criteria. This statistic is measured by $-2logL$, where L is the maximum likelihood for the model. The smaller value of this statistic gives the better model for given data.

Akaike's Information Criteria (AIC) (Akaike, 1973) and Bayesian Information Criteria (BIC) (Schwarz, 1978) are frequently used for the model selection, which are, respectively, given by

$$\text{AIC} = -2log(L) + 2p,$$

and

$$\text{BCI} = -2log(L) + plog(n),$$

where p is the number of parameter estimated and n is the total number of observations. The smaller values of AIC and BIC give the better model for given data.

Table 3: Model selection criteria for eight models for data in Table 1.

Group	Cluster Size	Model	$-2logL$	AIC	BIC
Control	27	BB	77.237	81.237	80.100
		CB	80.909	84.909	83.772
		MB	81.126	85.126	83.989
		DB	76.595	80.595	79.458
		FM	77.443	81.443	80.306
		ZIB	81.264	85.264	84.127
		ZIBB	77.236	83.236	81.531
		CBB	77.233	83.233	81.527
Low	19	BB	46.930	50.930	49.487
		CB	47.434	51.434	49.991
		MB	50.583	54.583	53.140
		DB	48.136	52.136	50.693
		FM	46.882	50.882	49.439
		ZIB	48.229	52.229	50.787
		ZIBB	46.885	52.885	50.722
		CBB	46.545	52.545	50.381
Medium	21	BB	82.014	86.014	84.658
		CB	89.646	93.646	92.290
		MB	89.941	93.941	92.586
		DB	79.202	83.202	81.846
		FM	86.859	90.859	89.504
		ZIB	89.045	93.045	91.689
		ZIBB	81.997	87.997	85.963
		CBB	81.776	87.776	85.467
High	17	BB	52.901	56.901	55.362
		CB	53.014	57.014	55.475
		MB	51.189	55.189	53.650
		DB	52.395	56.395	54.856
		FM	52.798	56.798	55.259
		ZIB	54.698	58.698	57.159
		ZIBB	51.927	57.927	55.618
		CBB	52.818	58.818	56.509

4.2 Parametric Bootstrap for Model Selection

Allcroft and Glasbey (2003) proposed a parametric bootstrap method for model selection based on the observed log-likelihoods and their simulated log-likelihoods using the Mahalanobis squared distances. This method measures the distances between the observed log-likelihoods and their simulated log-likelihoods for all candidate models. The following steps describe how to select the most appropriate model for a given data set:

- Step 1: Fit the candidate models $M_1, M_2, \ldots, M_k$ and save estimates of the model parameters and log-likelihoods for all k models.
- Step 2: Simulate a sample from each fitted model, and refit the candidate models and save their log-likelihoods.
- Step 3: Repeat Step 2, B times and compute the average log-likelihood for each of the k candidate models
- Step 4: Compare log-likelihoods evaluated at original data at Step 1 with log-likelihoods evaluated at the simulated data using the following Mahalanobis squared distances.

Let Λ be the vector of log-likelihoods for the candidate models $M_1, M_2, \ldots, M_k$ at the original data obtained in Step 1. Also, let $\bar{\Delta}_t$ be the vector of average log-likelihoods at the simulated data from the tth candidate model obtained in Step 3. Further, let Σ bet the sample variance-covariance matrix for the simulated log-likelihoods from Step 3. Then the Mahalanobis squared distance for the tth candidate model is obtained by

$$MD_t^2 = (\Lambda - \bar{\Delta}_t)'\Sigma^{-1}(\Lambda - \bar{\Delta}_t), \quad t = 1, \ldots, k,$$

where MD_t^2/k follows approximately F distribution with degrees of freedom k and $B-1$.

4.3 Vuong's Test and Cox's Test

The models also can be compared by the Vuong's test (Vuong, 1989) as well as the Cox's test (Cox, 1961) when the models are non-nested. The Vuong's test statistic uses the Kullback distance between two models. Under the hypothesis that the two models do not differ significantly, the test statistic is defined as

$$LLR(f, g) = \frac{\bar{w} - k}{\sqrt{n\sigma^2}},$$

where $w_i = log(f(y_i, \hat{\theta})) - log(g(y_i, \hat{\eta}))$, and $log(f(y_i, \hat{\theta}))$ and $log(g(y_i, \hat{\eta}))$ are the log-likelihood functions for model f and model g at their maximum, evaluated for sample i. $\bar{w}$ is the mean of the individual log-likelihood functions w_i and σ^2 is defined as

$$\sigma^2 = \frac{1}{n}\sum_{i}^{n}(w_i - \bar{w})^2.$$

To account for the different number of parameters of the models compared the correction term k takes the form $k = 0.5(m_1 - m_2)log(n)$, where m_i $i = 1, 2$ is the number of parameters for model i.

The Cox's test statistics compare the expected value of the likelihood ratio statistic under each of the two non-nested models, and conclude data to be consistent with one, both or neither of the two models. The form of the statistic can be obtained following the equation (48) in Cox (1961), indicating that a larger negative value of the test statistic would lead to the rejection of the model under the null hypothesis, whereas a larger positive value of the test statistic would lead to the acceptance of the model under the null hypothesis.

Table 4: Mahalanobis squared distances

	Control Group		Low Group		Medium Group		High Group	
Models	MSD	P-value	MSD	P-value	MSD	P-value	MSD	P-value
Binomial								
BB	6.7526	0.4613	3.4633	0.7477	25.3841	0.0016	2.8578	0.8952
CB	8.4529	0.3058	3.5026	0.7425	41.2469	0.0000	3.7717	0.8032
MB	8.2612	0.3211	4.8246	0.5691	35.4500	0.0001	4.0924	0.7670
DB	7.4172	0.3953	2.4285	0.8743	33.1036	0.0001	4.0365	0.7734
FM	6.9480	0.4412	3.9742	0.6801	33.8030	0.0001	3.9734	0.7806
ZIB	9.9645	0.2044	3.8508	0.6965	46.5962	0.0000	6.3329	0.5061
CBB	8.2422	0.3226			52.2280	0.0000	3.2918	0.8540

4.4 KL Distance and Jeffreys' Divergence

Kullback-Leibler (KL) distance (Eguchi and Copas, 2006) can be used to measure the discrepancy between the two probability functions. This distance measures the expected value of the log-likelihood ratio with respect to the model itself, that is, it provides the average difference of the contribution to the log-likelihood of any observation, which is defined as

$$K(f,g) = E_f\left[log\left(\frac{f(y,\hat{\theta})}{g(y,\hat{\eta})}\right)\right] = \sum_{i=1}^{n} f(y_i,\hat{\theta})log\left(\frac{f(y_i,\hat{\theta})}{g(y_i,\hat{\eta})}\right)$$

or

$$K(g,f) = E_g\left[log\left(\frac{g(y,\hat{\eta})}{f(y,\hat{\theta})}\right)\right] = \sum_{i=1}^{n} g(y_i,\hat{\eta})log\left(\frac{g(y_i,\hat{\eta})}{f(y_i,\hat{\theta})}\right).$$

Note that, in general $K(f,g) \neq K(g,f)$, that is, KL distance is not symmetric. In this case, one can use the Jeffreys' divergence (Jeffreys, 1998), which measures the difference between the two expectations of the log-likelihood ratio under both models and defined by

$$J(f,g) = \sum_{i=1}^{n}[f(y_i,\hat{\theta}) - g(y_i,\hat{\eta})]log\left(\frac{f(y_i,\hat{\theta})}{g(y_i,\hat{\eta})}\right).$$

This divergence can also be obtained based on the KL distance as

$$J(f,g) = K(f,g) + K(g,f).$$

Small values indicate that the likelihood is the same for the two probability functions.

5. An Illustrative Application

Recall that the motivational example of this paper is to fit the proportion of live foetuses data for each of four dose groups C, L, M, and H using the models described in Section 2. The estimates of the model parameters for the four groups are obtained based on the ML procedures described in Section 3. The ML estimates of the model parameters and their standard errors for all eight completing models are reported in Table 2. Note that the

ML estimates of the parameters for the BB, CB, MB, and DB models and their standard errors are in agreement with those given by Saha (2011). From Table 2 we see that the the dispersion parameters for all eight models are significant, indicating these proportion data seem to be over-dispersed. We first applied the standard approaches described in section 4.1 to select the best model. The results are presented in Table 3. From the results in Table 3 we see that the models DB, FM, DB, and MB are the best to the data in groups C, L, M, and H, respectively. However, some other models fit the data well. For example, for low dose group, the BB model has an acceptable fit. Next, we have fitted the seven models based on the parametric bootstrap approach described in the section 4.2. Here we used $B = 100$. The results of the Mahalanobis squared distances with the associated p-values are reported in Table 3. For the data sets in control, low, and high dose groups, all seven models fit well to the data. For the data in medium group all seven models fail to describe the data. The BB model describes the data better compared to the other models for the data in low and high groups, whereas the DB model fit the data better compared to the other models.

6. Concluding Remarks

In this article, we have carried out a comparison study of eight competing over-dispersed proportion models with the real world data occurring in a toxicological study. It has been shown by many authors (Paul 1982; Pack 1986) the BB model has the superiority for the analysis of the over-dispersed proportion data. In the literature, it is also known that the BB model differs from the CB, the MB, and the DB models, but it was not clear by how much. In addition, no comparison study was conducted for the BB model with other available models such as the FM, the ZIB, the ZIBB, and the CBB models. Clearly, the comparisons were extended in this paper by including all eight models. Although serveral model selection approaches are discussed in the section 4, we only applied the standard approaches as well as the parametric bootstrap approach to the real data analysis to select the best model. From the real data analysis, we have found that no single model fits all data sets well. The standard model selection approaches showed that the double binomial model fits more data sets well, whereas the parametric bootstrap approach for model selection showed that the beta-binomial model describes more data sets well. Therefore, one needs to investigate the performances of the model selection procedures through simulations before drawing any conclusions about the comparisons of these models. We made some progress towards this and will be reported in the future communication. Furthermore, we found from the real data analysis that of all the eight models, the likelihood under the beta-binomial model is the simplest one to maximize. The normalizing constant of the double binomial and the multiplicative binomial models, and the data-dependent bound for the parameters of the correlated binomial and the correlated beta-binomial models, make it difficult to maximize the likelihoods under these models. Therefore, we conclude that the beta-binomial model would be the superior model in terms of computational aspect compared to the other models.

Acknowledgements

This research was partially supported by the CSU-AAUP University Research Grant and the Mathematics Department Endowment Fund. The author would like to thank the Mathematics Department for providing the financial support to present this manuscript at the 2012 Joint Statistical Meetings in San Diego, CA. Part of this work was done while the author was visiting the Department of Statistics at the Texas A & M University, the Department of Mathematics and Statistics at the Concordia University, and the Department of Mathematics and Statistics at the University of Windsor. The author would like to thank these Departments for providing a stimulating environment.

REFERENCES

Allcroft, D.J. and Glasbey, C.A. (2003). A simulation-based method for model evaluation. *Statistical Modelling, An International Journal*, 3, 1-13.

Altham, P.M.E. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, 27, 162-167.

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in *Proceedings of the Second International Symposium on Information Theory*, eds. B, Petrov and F. Czakil, Akademiai Kiado, Budapest, Hungray: pp 267-281.

Atkinson, A.C. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society, Series B*, 32, 323-353.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium*, 1, 105-123.

Crowder, M.J. (1982). On weighted least-squares and some variants. *Surrey University Technical Report in Statistics* 1982; No.13.

Deng, D. and Paul, S.R. (2005). Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*, 15, 257-276.

Efron, B. (1986), Double Exponential Families and Their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, 81, 709-721.

Eguchi, S. and Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, 97, 2034-2040.

Elston, R.C. (1977). Response to query, consultants corner. *Biometrics*, 33, 232-233.

Gibson, G.J. and Austin, E.J. (1996). Fitting and testing spatio-temporal stochastic models with applications in plant pathology. *Plant Pathology*, 45, 172-184.

Jeffreys, H. (1998). *Theory of probability*. New York: The Clarendon Press Oxford University Press.

Kupper, L.L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics*, 34, 69-76.

Lindsey, J.K. (1974), Construction and Comparison of Statistical Models. *Journal of the Royal Statistical Society*, Series B, 36, 41825.

Lindsey, J.K. and Altham, P.M.E. (1978). Analysis of the human sex ratio by using overdispersion models. *Applied Statistics*, 47, 149-157.

Morel. J.G. and Nagaraj, N.K. (1993). A finite-mixture distribution for modeling multinomial extra variation. *Biometrika*, 80, 363-371.

Otake, M. and Prentice, R.L. (1984). The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Rediation Research*, 98, 456-470.

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

Paul, S.R. (1985). A three-parameter generalization of the binomial distribution. *Communications in Statistics: Theory and Methods*, 14, 1497-1506.

Paul, S.R. and Islam, A.S. (1995). Analysis of proportions in the presence of over-/under-dispersion. *Biometrics*, 51, 1400-1411.

Saha, K.K., and Paul, S.R. (2005). Bias Corrected Maximum Likelihood Estimator of the Negative Binomial Dispersion Parameter. *Biometrics*, 61, 179-185.

Saha, K.K. (2011). A comparison study of some competing discrete models for proportions or counts, with applications to biological data. *Journal of the Indian Society of Agricultural Statistics*, 65, 143-153.

Schwarz, F. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, 461-464.

Vuong, Q.H. (1989). Ratio tests for model selection and non-nested hypotheses, *Econometrica*, 57, 307-333.

Weil, C.S. (1970). Selection of valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis reproduction, teratogenesis. *Food and Cosmetic Toxicology*, 8, 177-182.

Williams, D.A. (1975). Analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31, 949-952.