Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 65(2) 2011 143-153

A Comparison Study of Some Competing Discrete Models for Proportions or Counts, with Applications to Biological Data

Krishna K. Saha

Department of Mathematical Sciences, Central Connecticut State University, 1615 Stanley Street, New Britain, CT 06050, USA

Received 04 August 2010; Revised 15 August 2010; Accepted 16 August 2010

SUMMARY

Discrete data such as count data or data in the form of proportions arise in biological investigations and other similar fields. These data often show variation greater or smaller than predicted by the simple probability models such as the Poisson or the binomial model. Several discrete models have been used for modeling the counts or proportions by many authors (see, for example, Byers *et al*. 2003; Consul and Jain, 1973; Efron, 1986; Gibson and Austin, 1996; Kupper and Haseman, 1978; and Saha and Paul 2005). This article reviews briefly several aspects and the properties for some of the most commonly used discrete models for modeling counts or proportions, namely the negative binomial, the generalized Poisson, the double Poisson, the generalized negative binomial, the beta-binomial, the correlated binomial, the multiplicative binomial, and the double binomial models. The maximum likelihood method is outlined for the estimation of the parameters of these models. Comparison studies of these models are considered in light of goodness of fit test as well as model selection criteria through real-life data occurring in agricultural and toxicological fields.

Keywords : Beta-binomial, Biological data, Double binomial, Negative binomial model, Extra dispersion parameter.

1. INTRODUCTION

Discrete data such as count data or data in the form of proportions occur in a wide variety of disciplines. These data often show variation significantly larger or smaller than that predicted by a simple model such as a Poisson or 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 Poisson model is necessary. For example, an agricultural data set given in Table 1 shows that the observed variance exceeds its mean, hence a Poisson model is inappropriate. Several alternative models that take into account the extra Poisson variation have been used by many authors (Greenwood and Yule 1920, Jain and Consul 1971, Consul and Jain 1973, Efron 1986). Greenwood and Yule (1920) introduced the negative binomial model as a mixture of Poisson and gamma distribution. Many authors have used this distribution for analyzing the extra dispersed count data (see, for example, Newbold 1927, Bliss and Fisher 1953, Barnwal and Paul 1988, White and Bennets 1996, and Byers *et al*. 2003). More details about the properties and applications of the negative binomial model is reviewed by Bartko (1961). Jain and Consul (1971) proposed another alternative extra dispersed model, called a generalized negative binomial distribution, by compounding the negative binomial distribution with another parameter which takes into account the variations in the mean and the variance. Consul and Jain (1973) derived the generalized Poisson distribution by approximating the generalized negative binomial distribution using Stirling's formula on the

E-mail address : sahakrk@ccsu.edu

Table 1. Frequency distribution of red mites on apple leaves

Number of adult females (y)	Frequency (f)
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1
Mean ($\bar{Y}$)	1.1467
Variance (S^2)	2.2737

two gamma functions. Efron (1986) develops the double Poisson distribution by using the double exponential family. Holla (1966) also introduced a Poisson-inverse Gaussian distribution as an alternative to the negative binomial model. Sometimes, a Poisson-log-normal distribution can be used as an extra dispersed Poisson model by taking a log-normal distribution as a mixing density. The major disadvantage for some of the distributions is that the probability must be computed numerically. However, among all these distributions, the negative binomial is popular due to its simplicity.

The negative binomial, generalized negative binomial, generalized Poisson, Poisson-inverse Gaussian, and Poisson-log-normal distributions are members of the family of mixed Poisson distributions, whereas the double Poisson distribution is a member of the double exponential family. Not much work has been done on comparing their behavior. Kaas (1995) studied the comparison of the negative binomial, Poisson-inverse Gaussian, and Poisson-log-normal distributions and conclude that NB has lighter tails than the Poisson-inverse Gaussian and Poisson-log-normal distributions. Joe and Zhu (2005) also studied the skewness of the negative binomial and generalized Poisson distributions and showed that the generalized Poisson distribution can be more skewed. Nikoloulopoulos and Karlis (2008) compared the four distributions: the negative binomial, generalized Poisson, Poisson-inverse Gaussian, and Poisson-log-normal distributions. They showed that the negative binomial model differs from the generalized Poisson, Poisson-inverse Gaussian, and Poisson-log-normal models, whereas the generalized Poisson and Poisson-inverse Gaussian distributions behave quite the same. The purpose of this study is to investigate whether these differences are really shown by real-life data. In addition, we include the double Poisson and generalized negative binomial distributions, which were not included in previous studies, along with other distributions in our study.

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

Table 2. 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											

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 provided in Table 2 shows the discrepancy between the observed variances and those predicted by the binomial model, indicating 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 the account the variability shown in the proportion data occurring in biological investigations.

Different extra dispersed models for analyzing proportions have been used by many authors (Altham 1978, Kupper and Haseman 1978, Efron 1986, 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, Klein-man 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 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 analogy with a model in a 2^M contingency table with no second-and higher-order interactions. Efron (1986) introduced what he called a double binomial model from the double exponential family. Due to its simplicity, many authors have used the beta-binomial distribution for the analysis of extra dispersed proportion data. No work has been done about 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. In our comparison study, we include all four models that are candidates for the analysis of any real-life extra dispersed proportions occurring in biological investigations.

The purpose of this article is to conduct a comparison study of some well-known competing discrete models for the analysis of both the count and proportion data occurring in biological fields in light of the well-known model selection criteria as well as the usual goodness of fit test. 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 counts or proportions.

This article is organized as follows. Section 2 reviews some competing models for the analysis of any real-life count data. The five competing models for analyzing proportions are discussed briefly in Section 3. Section 4 introduces the maximum likelihood methods for the estimates of the parameters of the negative binomial and the beta-binomial distributions. The goodness of fit test and the model selection criteria are discussed briefly in Section 5. Section 6 shows whether the researcher in applied fields can really identify the underlying distribution uniquely from agricultural data as well as toxicological data. A discussion can be found in Section 7.

2. THE COMPETING MODELS FOR COUNT DATA

There is a wide range of discrete models for the analysis of extra dispersed count data, for example, see Dean *et al*. 1989, Efron 1986, Hinde 1982 and Jain and Consul 1971. In this section, we review briefly the probability mass functions and their properties of five candidate parametric models for count data below.

2.1 The Poisson Model

A common phenomenon when analyzing data in the form of counts is to assume a Poisson model, which has a probability mass function as

$$f(y \mid \nu) = \frac{e^{-\nu}\nu^{y}}{y!} \qquad (1)$$

for $y = 0, 1, 2, \ldots$, where y is the number of adult females in Table 1. The mean and variance of the Poisson random variable Y are $\mu = \nu$ and $\sigma^2 = \nu$, respectively. Note that the underlying model of counts

would be the Poisson model if data come from a pure random mechanism or if the homogeneity assumption of the population under investigation holds. For many biological studies, these are violated due to the mechanism applied for their investigations. As a result, data obtained from these experiments show extra variability compared to Poisson data. This model is a member of the generalized linear model (GLM) family.

2.2 The Double Poisson (DP) Model

For over-dispersed count data, Efron (1986) proposed the double Poisson model, which has a probability mass function as

$$f(y \mid \nu, \theta) = \frac{\sqrt{(\theta)}\, y^y e^{-(y+\nu\theta)}}{y!\, c(\nu, \theta)} \left(\frac{e\nu}{y} \right)^{y\theta} \qquad (2)$$

for $y = 0,1,2,\ldots$, and $\theta > 0$, where the normalizing constants $c(\nu, \theta)$ can be calculated as

$$c(\nu, \theta) = \sum_{y=0}^{\infty} \frac{\sqrt{(\theta)}\, y^y e^{-(y+\nu\theta)}}{y!} \left(\frac{e\nu}{y} \right)^{y\theta}, \qquad (3)$$

which are intractable so that this model is difficult to fit by standard methods. The mean and the variance of the double Poisson variable Y are $\mu = \nu/c(\nu, \theta)$ and $\sigma^2 = \nu/[\theta c(\nu, \theta)]$, respectively. Setting $\theta = 1$, this model yields a Poisson distribution. Note that this model will be over dispersed for $0 < \theta < 1$ or under dispersed for $\theta > 1$. This model is a member of the double exponential family.

2.3 The Negative Binomial (NB) Model

A popular and convenient model for extra dispersed count data is the negative binomial model. Let Y be a negative binomial random variable with mean ν and dispersion parameter τ. We write $Y \sim NB(\nu, \tau)$, which has a probability mass function as

$$Pr(Y = y \mid \nu, \tau) = \frac{\Gamma(y + \tau^{-1})}{y!\Gamma(\tau^{-1})} \left(\frac{\tau\nu}{1+\tau\nu} \right)^{y} \left(\frac{1}{1+\tau\nu} \right)^{\tau^{-1}}, \qquad (4)$$

for $y = 0, 1, \ldots$; $\nu > 0$; and $\tau > 0$. The mean and variance of the negative binomial variable Y are $\mu = \nu$ and $\sigma^2 = \nu(1 + \tau\nu)$, respectively. The limiting distribution of the $NB(\nu, \tau)$, as $\tau \to 0$, is the Poisson(ν); that is, this model is not over dispersed for $\tau = 0$. Note that this model will be over or under dispersed for $\tau > 0$ or $\tau < 0$, respectively.

2.4 The Generalized Negative Binomial (GNB) Model

Using Lagrange's expansion Jain and Consul (1971) have obtained a generalized negative binomial model, which has the probability mass function

$$f(y \mid c,b,a) = \frac{a^{-1}\Gamma(by + a^{-1})}{y!\Gamma(a^{-1} + \{b-1\}y + 1)} c^y (1-c)^{a^{-1} + (b-1)y} \qquad (5)$$

for $y = 0, 1, 2, \ldots$; $a > 0$, $0 < c < 1$, and $| cb | < 1$. The mean and variance of the generalized negative binomial variable Y are $\mu = \nu$ and $\sigma^2 = \nu(1 - c)/(1 - cb)^2$, with $\nu = c/a(1 - cb)$, respectively. Note that if $b = 1$, this model yields the negative binomial model and if $b = 1$ and $a \to 0$, this model becomes the classical Poisson model.

2.5 The Generalized Poisson (GP) Model

Applying James Stirling's formula to the two gamma functions of a generalized negative binomial model using $\lambda_1 = c/a$ and $\lambda_2 = bc$, Consul and Jain (1973) obtained a generalized Poisson model, which has the probability mass function

$$f(y \mid \lambda_1, \lambda_2) = \frac{\lambda_1(\lambda_1 + y\lambda_2)^{y-1} e^{-(\lambda_1 + y\lambda_2)}}{y!}, \qquad (6)$$

for $y = 0, 1, 2, \ldots$, $\lambda_1 > 0$ and $-1 < \lambda_2 < 1$. The mean and variance of the generalized Poisson variable Y are $\mu = \nu$ and $\sigma^2 = \nu/(1 - \lambda_2)^2$, with $\nu = \lambda_1/(1 - \lambda_2)$, respectively. Note that this model will be over-dispersed or under-dispersed for $0 < \lambda_2 < 1$ or $\lambda_2 > 1$, respectively, and if $\lambda_2 = 0$, this model yields the classical Poisson (λ_1).

3. THE COMPETING MODELS FOR PROPORTION DATA

Different parametric models have been used in a wide range of biological fields for modeling the correlated binary data (see, for example, Altham 1978, Barnwal and Paul 1988, Efron 1986, Kupper and Haseman 1978, and Prentice 1986). We briefly discuss the probability mass functions and their properties of all five competing parametric models for the data in the form of proportions as follows.

3.1 The Binomial Model

The usual method for the analysis of the data in the form of proportions is to assume the binomial model, which has a probability mass function as

$$f(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \tag{7}$$

for $y = 0, 1, 2, \ldots, n$, and $0 \le \pi \le 1$, where n is the number of live foetuses in the litter and y is the number of affected foetuses in Table 2. The mean and variance of the response variable of Y are $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$, respectively. This model is a member of the generalized linear model (GLM) family.

3.2 The Beta-Binomial (BB) Model

The most commonly used model for extra dispersion in binomial data is the beta-binomial model (Williams 1975), which has a probability mass function

$$f(y|\pi, \phi) = \binom{n}{y}\frac{B(\alpha+y, n+\beta-y)}{B(\alpha,\beta)} = \binom{n}{y}\frac{B\left(\frac{\pi}{\theta}+y, n+\frac{1-\pi}{\theta}-y\right)}{B\left(\frac{\pi}{\theta}, \frac{1-\pi}{\theta}\right)} \tag{8}$$

for $y = 0, 1, 2, \ldots, n$, where $B(.)$ is the beta function. Obviously, when $\theta \to 0$, the $BB(\pi, \theta)$ becomes Binomial(π). The mean and variance of the beta binomial response Y are $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)[1 + n\theta]/(1+\theta)$, respectively. The parameter $\rho = \theta/(1+\theta)$ can also be interpreted as the interclass correlation, that is, the relationship between the siblings in the same litter. This distribution will be over-dispersed for $\theta > 0$ and under-dispersed for $\theta < 0$. Note that the limiting distribution of this model, as $\theta \to 0$, is the binomial(n, π).

3.3 The Correlated Binomial (CB) Model

By taking correlation between the siblings in the same litter into account Kupper and Haseman (1978) proposed the correlated binomial model, which has a probability mass function

$$f(y \mid \pi, \phi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}\left[1+\frac{\rho}{2\pi^2(1-\pi)^2}\{(y-n\pi)^2 + y(2\pi-1) - n\pi^2\}\right] \tag{9}$$

This model is identical to the additive generalized-binomial model of Altham (1978). The binomial model corresponds to a special case when $\rho = 0$. This distribution will be under-dispersed or over-dispersed for $\rho < 0$ or $\rho > 0$, respectively.

3.4 The Multiplicative Binomial (MB) Model

By drawing an analogy with a model in a 2^M contingency table with no second-and-higher order interactions, Altham (1978) introduced a two-parameter multiplicative generalization of the binomial model, which has a probability mass function

$$f(y \mid \pi, \gamma) = \binom{n}{y}\frac{\pi^y(1-\pi)^{n-y}\gamma^{\theta y(n-y)}}{k(\pi,\gamma,n)} \tag{10}$$

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

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

For $\gamma = 1$, this model becomes the usual binomial model. This model will be over-dispersed or under-dispersed for $\gamma > 1$ or $\gamma < 1$, respectively. This model is a member of the exponential family.

3.5 The Double Binomial (DB) Model

Efron (1986) developed the double-binomial model for extra dispersed binomial data, which has a probability mass function

$$f(y \mid \pi, \eta) = \binom{n}{y}\frac{n^{n\eta}\pi^{y(\eta+1)}(1-\pi)^{(n-y)(\eta+1)}}{c(\pi,\eta,n)y^{y\eta}(n-y)^{(n-y)\eta}} \tag{11}$$

for $y = 0, 1, 2, \ldots, n$, where $c(\pi, \eta, n)$ is the intractable factor as

$$c(\pi, \eta, n) = \sum_{y=0}^{n}\binom{n}{y}\frac{n^{n\eta}\pi^{y(\eta+1)}(1-\pi)^{(n-y)(\eta+1)}}{y^{y\eta}(n-y)^{(n-y)\eta}} \tag{12}$$

For $\eta = 0$, this model becomes the binomial model. This model will be over-dispersed or under-dispersed for $-1 < \eta < 0$ or $\eta > 0$, respectively. This model is a member of the double exponential family.

4. ESTIMATION OF THE PARAMETERS IN THE MODELS

Fitting an appropriate model for given data, we consider the maximum likelihood method to estimate the parameters in the models. We briefly discuss the

maximum likelihood (ML) methods for the negative binomial and the beta-binomial models in Sections 4.1 and 4.2. We also provide R functions to obtain the ML estimates of the parameters of the negative binomial and the beta-binomial models in Sections 4.3 and 4.4.

4.1 The Maximum NB Likelihood Estimator

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

$$l = \sum_{i=1}^{n}\left[y_i \ln(\nu) - \left(y_i + \frac{1}{\tau} \right)\ln(1+\tau\nu) + \sum_{j=0}^{y_i-1} \ln(1+aj) \right] \tag{13}$$

The ML estimates of ν and τ are then obtained by maximizing l or alternatively, by solving, simultaneously, the estimating equations (see also Piegorsch 1990)

$$U_\nu = \sum_{i=1}^{n}\left[\frac{y_i}{\nu} - \frac{1+\tau y_i}{1+\tau\nu} \right] = 0, \text{ and} \tag{14}$$

$$U_\tau = \sum_{i=1}^{n}\left[\frac{1}{\tau^2}\ln(1+\tau\nu) - \frac{y_i-\nu}{\tau(1+\tau\nu)} - \sum_{j=0}^{y_i-1}\frac{1}{\tau(1+aj)} \right] = 0. \tag{15}$$

Similar to the ML estimates of the NB model, one can easily obtain the maximum likelihood estimates of the parameters involved for other candidate models for count data discussed in Section 2.

4.2 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

$$l = \sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1} \ln\{\pi + j\theta\} + \sum_{j=0}^{n_i-y_i-1} \ln\{1-\pi+j\theta\} - \sum_{j=1}^{n_i-1}\{1+j\theta\} \right] \tag{16}$$

The maximum likelihood estimates of π and θ can be obtained by maximizing l or alternatively, simultaneously, by solving the estimating equations (see also Saha and Paul 2004)

$$\sum_{i=1}^{m}\left[\sum_{j=0}^{y_i-1}\frac{1}{\pi+j\theta} - \sum_{j=0}^{n_i-y_i-1}\frac{1}{1-\pi+j\theta} \right] = 0, \text{ and} \tag{17}$$

$$\sum_{i=1}^{m}\left[\sum_{j=1}^{y_i-1}\frac{j}{\pi+j\theta} + \sum_{j=0}^{n_i-y_i-1}\frac{j}{1-\pi+j\theta} - \sum_{j=0}^{n_i-1}\frac{j}{1+j\theta} \right] = 0. \tag{18}$$

Similar to the ML estimates of the BB model, we can easily obtain the maximum likelihood estimates of the parameters involved for other candidate models for data in the form of proportions discussed in Section 3.

5. THE GOODNESS OF FIT AND MODEL SELECTION CRITERIA

In this section, we consider a few methods to compare the models in Sections 2 and 3 based on the goodness of fit as well as the model selection criteria. Many different techniques have been used by different authors (see, for example, Bliss and Fisher 1953, Cochran 1954, Paul 1982, and White and Bennetts 1996). The following goodness of fit and model selection criteria can be used to determine the better model among many competing parametric models discussed in Sections 2 and 3 for the real-life data occurring in biological investigations.

Bliss and Fisher (1953) and Cochran (1954) used the Pearson's chi-square for the goodness of fit for discrete data, which is given by

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

where O_i and E_i are the observed and expected frequencies for cell i, respectively. This chi-square statistic can be calculated alternatively when data are not in the form of frequency distribution as $\chi^2 = \sum_i r_i^2$, where $r_i = [y_i - E(Y_i)]/\sqrt{\text{var}(Y_i)}$ is the ith residual. The smaller value of this statistic gives the better fitted model for given data.

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

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

Table 3. The comparative statistics for all five competing models for data in Table 1

Model	Parameter	ML Estimate	Standard Error	$\hat{\mu}$	$\hat{\sigma}^2$
Poisson	ν	1.1467	0.0874	1.1467	1.1467
GP	λ_1	0.7787	0.0860	1.1467	2.4866
	λ_2	0.3209	0.0595		
DP	ν	0.8317	0.2004	0.8915	2.62114
	θ	0.3401	0.0891		
NB	ν	1.1467	0.1273	1.1467	2.4299
	τ	0.9760	0.2628		

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 4. Goodness of fit and model selection criteria for five models for data in Table 1.

Model	–2logL	AIC	BIC	χ^2
Poisson	485.62	487.62	487.80	26.65
GP	445.49	449.49	449.84	2.89
DP	444.39	448.39	448.74	2.02
NB	256.17	139.959	260.17	2.49
GNB	443.52	449.52	450.05	2.07

6. APPLICATIONS

To address a comparison study of competing parametric discrete models for counts or proportions described previously, we use two different real-life data sets from biological investigations. The first refers to an agricultural data example where data are counts with extra dispersion, whereas the other treats a toxicological data example where data are in the form of proportions with extra dispersion.

6.1 An Agricultural Example (Count Data)

We now consider the data previously analyzed by Bliss and Fisher (1953) and Clark and Perry (1989). The data consists of counts of the number of European red mites on apple leaves. There were six Macintosh trees which were provided the same spray treatment in a single orchard. Garman (1951) selected 25 leaves at random from each of the six trees and counted the number of adult females on each leaf. The data in the form of frequencies of mites on the 150 leaves are given in Table 1. More details about the study are given in Garman (1951) of The Connecticut Agricultural Experiment Station. The comparative statistics for all competing models are listed in Table 3.

Note that ML estimates of the parameters for all models are obtained based on the R program discussed in Section 4.1. The ratio of variance to mean is d =2.2737/1.1467 = 1.9828, which indicates that the data are quite dispersed. The estimates of the dispersion parameters and estimated variance $\hat{\sigma}^2$ for the GP, DP, NB, and GNB models also indicate that data in Table 1 are significantly dispersed. One more feature of this dataset is the larger zero fraction which is 0.467 and no heavy tail exists because of no extreme values of counts so one could try zero infiated distributions.

Based on the preliminary analysis, we found that extra dispersed models for counts fit these data. Furthermore, we compared the fit of all five different count distributions to these data based on the goodness of fit test and model selection criteria discussed in Section 5, and the results of the models of Poisson, GP, DP, NB, and GNB are reported in Table 4. These results in Table 4 indicate that NB fits the data better than other four models based on all three model selection criteria, whereas DP fits the data better than other four models in terms of goodness of fit test. The fitted models for the data in Table 1 are plotted in Fig. 1. Fig. 1 clearly shows that Poisson model does not fit the data, whereas the other four models fit the data very well. We also see from Fig. 1 that none of the four models shows a clear winner over the others.

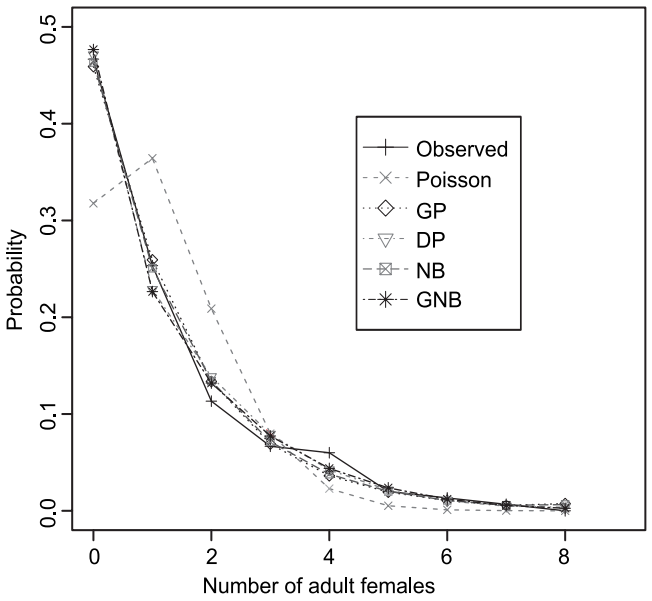

Fig. 1. Fitted Poisson, Generalized Poisson (GP), Double Poisson (DP), negative binomial (NB), and generalized negative binomial (GNB) models for the data in Table 1.

6.2 A Toxicological Example (Proportion Data)

The data in Table 2 refer to live foetuses in a litter affected by treatment, and the number of live foetuses, for each of four dose groups: control (C), low dose (L), medium dose (M), and high dose (H). More details about the study are given in Paul (1982). The estimates of the parameters π and θ for the four groups are obtained using R program given in Section 4.2. The estimates of the parameters and their standard errors for all five completing models are reported in Table 5.

Note that the ML estimates of π and θ for NB model and their standard errors are in agreement with those given by Paul (1982). 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. These results indicate that the data in Table 2 for all groups are quite dispersed. The estimates of the dispersion parameters for BB, CB, MB, and DB models also indicate the extra dispersion of these datasets. As a result, we fit all four extra dispersed models for these data and compared the fit in terms of goodness of fit test and model selection criteria described previously. The results of the fit of binomial, BB, CB, MB, and DB models are given in Table 6. Based on the model selection criteria DB, BB, DB, and MB are, respectively, better fitted models for the treatment groups C, L, M, and H, whereas BB is better

Table 5. The estimates of the parameters and their standard errors for all five competing models for data in Table 2

Treatment	Model	Parameter	ML Estimate	Standard Error
C	Binomial	π	0.1349	0.0233
	BB	π	0.1404	0.0380
		θ	0.2148	0.0957
	CB	π	0.1376	0.0302
		ρ	0.0134	0.0049
	MB	π	0.3216	0.0594
		γ	0.7980	0.0467
	DB	π	0.0621	0.0674
		η	–0.7704	0.1529
L	Binomial	π	0.1353	0.0297
	BB	π	0.1272	0.0373
		θ	0.1054	0.0813
	CB	π	0.1351	0.0370
		ρ	0.0092	0.0066
	MB	π	0.1437	0.0796
		γ	0.9861	0.1181
	DB	π	0.1172	0.0487
		η	–0.4838	0.2687
M	Binomial	π	0.3444	0.0387
	BB	π	0.3505	0.0678
		θ	0.3155	0.1091
	CB	π	0.3296	0.0521
		ρ	0.0280	0.0084
	MB	π	0.4281	0.0352
		γ	0.8404	0.0394
	DB	π	0.3131	0.0849
		η	–0.7177	0.1241
H	Binomial	π	0.2277	0.0417
	BB	π	0.2387	0.0548
		θ	0.1132	0.0944
	CB	π	0.2387	0.0502
		ρ	0.0189	0.0155
	MB	π	0.3430	0.0635
		γ	0.8172	0.0708
	DB	π	0.2131	0.0629
		η	–0.4740	0.2289

Table 6. Goodness of fit and model selection criteria for five models for data in Table 1.

Treatment	# of Litters	Model	–2logL	AIC	BIC	χ^2
C	27	Binomial	91.59	93.59	93.02	66.31
		BB	77.24	81.24	80.10	28.07
		CB	80.91	84.91	83.77	31.52
		MB	81.13	85.13	83.99	30.59
		DB	76.40	80.40	79.26	34.98
L	19	Binomial	50.60	51.60	51.88	29.41
		BB	46.93	50.93	49.49	17.68
		CB	47.43	51.43	49.99	21.35
		MB	50.58	54.58	53.14	27.91
		DB	48.05	52.05	50.61	24.02
M	21	Binomial	99.40	100.40	100.72	55.27
		BB	82.01	86.01	84.66	21.11
		CB	89.65	93.65	92.29	31.21
		MB	89.94	93.94	92.59	29.33
		DB	78.72	82.72	81.36	16.57
H	17	Binomial	55.39	57.39	56.62	28.52
		BB	52.90	56.90	55.36	18.44
		CB	53.01	57.01	55.47	21.33
		MB	51.19	55.19	53.65	20.32
		DB	52.13	56.13	54.59	15.81

fitted model for treatment groups C and L, and DB is better fitted model for treatment groups M and H in terms of goodness of fit test.

7. DISCUSSION

We have conducted a comparison study of some competing discrete models for the analysis of both the count and proportion data occurring in biological fields based on the model selection criteria and the goodness of fit test. We have discussed several aspects of the negative binomial, the generalized negative binomial, and generalized Poisson distributions for modeling the count data. We also have reviewed the properties of the beta-binomial, the correlated binomial, the multiplicative binomial, and the double binomial distributions for modeling the data in the form of proportions. The Pearson's chi-square for the goodness of fit test and the model selection criteria AIC, BIC, and -2logL were used to determine the better model for real life data. From the analysis of a set of agricultural data, we have found that the negative binomial model would be the better model in terms of all model selection criteria considered here, whereas the goodness of fit test suggested that the DP would be an appropriate model. However, the graphical presentation of all four fitted models showed evidence that any of the four models would be appropriate for modeling the agricultural data. For a set of toxicological data analysis, we have found that no unique model among the BB, DB, CB, and MB distributions can be recommended for modeling four treatment groups of datasets. To draw any clear guidelines for the choice of these distributions for modeling counts or proportions occurring in biological fields, further studies are necessary in light of simulation-based methods for model evaluation. Further study in this regard is continuing through a simulation study based on a parametric bootstrap approach of model evaluation using a Mahalanobis squared distance proposed by Allcroft and Glasbey (2003).

REFERENCES

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

Altham, P.M.E. (1978). Two generalizations of the binomial distribution. *Appl. Statist.*, **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.

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

Bartko, J.J. (1961). The negative binomial distribution: A review of properties and applications. *Virginia J. Sci. (New Sr.)*, **12**, 18-37.

Bliss, C.I. and Fisher, R.A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, **9**, 176-200.

Byers, A.L., Allore, H., Gill, T.M. and Peduzzi, P.N. (2003). Application of negative binomial modeling for discrete outcomes: A case study in aging research. *J. Clinical Epidemiology*, **56**, 559-564.

Clark, S.J. and Perry, J.N. (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, **45**, 309-316.

Cochran, W.G. (1954). Some method for strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

Consul, P.C. and Jain, G.C. (1973). A generalization of poisson distribution. *Technometrics*, **15**, 791-799.

Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. *Appl. Statist.*, **27**, 34-37.

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

Dean, A.B., Lawless, J.F. and Willmot, G.E. (1989). A mixed Poisson-Inverse-Gaussian Regression model. *The Canad. J. Statist.*, **17**, 171-181.

Donovan, A.M., Ridout, M.S. and James, D.J. (1994). Assessment of somaclonal variation in apple. II. Rooting ability and shoot proliferation. *J. Hort. Sci.*, **69**, 115-122.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.*, **81**, 709-721.

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

Garman, P. (1951). Original data on European red mite on. apple leaves. Report. Connecticut.

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.

Greenwood, M. and Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc.,* **A93**, 255-279.

Hinde, J. (1982). Compound Poisson Regression Model. *GLIM82: Proc. Internat. Conf. Generalized Linear Model* (ed. R. Gilchrist), 109-121, Springer, Berlin.

Holla, M. (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, **11**, 115-121.

Jain, G.C. and Consul, P.C. (1971). A generalized negative binomial distribution. *SIAM J. Appl. Math.*, **21**, 501-513.

Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical J.*, **47**, 219-229.

Kaas, R. and Hesselager, O. (1995). Ordering claim size distributions and mixed Poisson probabilities. *Insurance: Maths. Eco.*, **17**, 193-201.

Kleinman, J.C. (1973). Proportions with extraneous variance: Single and independent samples. *J. Amer. Statist. Assoc.*, **8**, 46-54.

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. *J. Roy. Statist. Soc.*, **B36**, 418-425.

Newbold, E.M. (1927). Practical applications to the statistics of repeated events particularly of industrial accidents. *J. Roy. Statist. Soc.,* **A90**, 487-547.

Nikoloulopoulos, A.K. and Karlis, D. (2008). On modeling count data: A comparison of some well-known discrete distribution. *J. Statist. Comput. Simul.*, **78**, 437-457.

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

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

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

Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.,* **81**, 321-327.

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

Schwarz, F. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Tarone, R.E. (1979). Testing the goodness of fit of the binomial dictribution. *Biometrika*, **66**, 585-590.

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 Cosmetic Toxicology*, **8**, 177-182.

White, G.C. and Bennetts, R.E. (1996). Analysis of frequency fount data using the negative binomial distribution. *Ecology*, **77**, 2549-2557.

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