International Journal of Statistics and Probability; Vol. 2, No. 2; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The McDonald Generalized Beta-Binomial Distribution: A New Binomial Mixture Distribution and Simulation Based Comparison with Its Nested Distributions in Handling Overdispersion

Chandrabose Manoj[1], Pushpa Wijekoon[2] & Roshan D. Yapa[2]

[1] Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka

[2] Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka

Correspondence: Chandrabose Manoj, Post Graduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka. E-mail: cmanoj@live.com

Received: January 3, 2013 Accepted: February 4, 2013 Online Published: February 28, 2013

doi:10.5539/ijsp.v2n2p24 URL: http://dx.doi.org/10.5539/ijsp.v2n2p24

Abstract

The binomial outcome data are widely encountered in many real world applications. The Binomial distribution often fails to model the binomial outcomes since the variance of the observed binomial outcome data exceeds the nominal Binomial distribution variance, a phenomenon known as overdispersion. One way of handling overdispersion is modeling the success probability of the Binomial distribution using a continuous distribution defined on the standard unit interval. The resultant general class of univariate discrete distributions is known as the class of Binomial mixture distributions. The Beta-Binomial (BB) distribution is a prominent member of this class of distributions. The Kumaraswamy-Binomial (KB) distribution is another recent member of this class. In this paper we focus the emphasis on the McDonald's Generalized Beta distribution of the first kind as the mixing distribution and introduce a new Binomial mixture distribution called the McDonald Generalized Beta-Binomial distribution(McGBB). Some theoretical properties of McGBB are discussed. The parameters of the McGBB distribution are estimated via maximum likelihood estimation technique. A real world dataset is modeled by using the new McGBB mixture distribution, and it is shown that this model gives better fit than its nested models. Finally, an extended simulation study is presented to compare the McGBB distribution with its nested distributions in handling overdispersed binomial outcome data.

Keywords: maximum likelihood, simulation, Kumaraswamy, generalized Beta first kind, ANODEV

1. Introduction

It is well known that the discrete random variable Y, the number of successes in n binary trials is generally modeled using the traditional Binomial distribution with the parameters n and p, if the binary trials are identical and independent. The probability of success parameter $(p|0 \leq p \leq 1)$ is usually assumed to be a constant from trial to trial. The mean and the variance of a Binomially distributed random variable is then given by $E(Y) = np$ and $Var(Y) = np(1-p)$. However, in most empirical situations, it has been observed that the actual variance of Y is greater than the assumed Binomial variance. This phenomenon is generally known as *Overdispersion*. "Extra Binomial Variation" and "Binomial Heterogeneity" are some other commonly used terms to label overdispersion. One of the possible reasons for the overdispersion is that the success probability, p, does not remain as a constant from trial to trail but varies itself as a random variable. This leads to treat the success probability as a continuous random variable P, which is bounded between 0 and 1. The resultant class of distributions is known as the class of Binomial Mixture Distributions. A Binomial mixture distribution can be symbolically denoted as $Bin(n, P) \wedge F_P(p)$ where $Bin(n, P)$ represents the Binomial distribution and $F_P(p)$ symbolizes the distribution function of the mixing distribution of the random variable P and the mixing density of P is denoted by $f_P(p)$. If the mean and the variance of the mixing distribution of the random variable P are denoted by $E(P) = n\pi$ and $Var(P) = \sigma^2$ respectively, then it can be shown that the mean and the variance of the Binomial mixture random variable Y are $E(Y) = n\pi$ and $Var(Y) = n\pi(1-\pi) + n(n-1)\sigma^2$ respectively. The additional component in the variance of Y models the overdispersion.

Even though number of possible (Kotz & Van Dorp, 2004; Johnson, Kemp, & Kotz, 2005) univariate continuous distributions defined on the standard unit interval [0,1] are available as the mixing distribution of the success probability random variable P, the Beta distribution $Beta\,(a, b)$, where a and b are the two shape parameters of the Beta distribution, is the most commonly used mixing distribution to model the random variable P due to its ability of accommodating wide range of shapes. Thus the Beta-Binomial (BB) distribution, represented by $Bin\,(n, P) \wedge Beta\,(a, b)$, is considered as a very versatile distribution in modeling overdispersed binomial outcome data in literature. Extensive literatures exist on the study of Beta-Binomial distribution. Theoretical properties, estimation techniques and applications of the Beta-Binomial distribution have been discussed by, for example, Skellam (1948), Chatfield and Goodhardt (1970), Griffiths (1973), Williams (1975), Haseman and Kupper (1977), Paul (1982), Tripathi, Gupta, and Gurland (1994), Gange, Munoz, Saez, and Alonso (1996), Ennis and Bi (1998), and recently Bandyopadhya, Reich, and Slate (2011), not limited but numerous.

Recently Li, Huang, and Zhao (2011) used the Kumaraswamy double bounded (Kumaraswamy, 1980), distribution as the mixing distribution of the Binomial probability of success and obtained a new Binomial Mixture distribution called the Kumaraswamy-Binomial (KB) distribution, which can be represented by $Bin\,(n, P) \wedge Kumaraswamy\,(\zeta, \theta)$, where ζ and θ are the two shape parameters of the Kumaraswamy distribution. The Kumaraswamy distribution which is defined on the standard unit interval [0,1], is also known as Minimax distribution. Analogous to the Beta distribution, the Kumaraswamy distribution also has two parameters and can assume wide variety of shapes (Jones, 2009). Li et al. (2011), have used the Kumaraswamy-Binomial distribution to model overdispersed binomial data and stated that both BB and KB distributions have same flexibility in modeling the overdispersed binomial data. In addition, Li et al. (2011) have compared their distribution with another Binomial mixture distribution proposed by Rodríguez-Avi, Conde-Sánchez, Sáez-Castillo, and Olmo-Jiménez (2007). This distribution is known as the Generalized Beta-Binomial (GBB) distribution, which is developed by generalizing the Gaussian hypergeometric representation of Beta-Binomial distribution. Both Rodríguez-Avi et al. (2007) and Li et al. (2011) have shown that the Generalized Beta-Binomial (GBB) distribution with three additional parameters possesses large flexibility in modeling overdispersion compared to the distributions with only two additional parameters such as Beta-Binomial and Kumaraswamy-Binomial distribution.

The aim of this paper is to propose an alternative Generalized Beta distribution, which has three additional parameters, as the mixing distribution to model the Binomial success probability. The McDonald's Generalized Beta distribution of the first kind (McDonald, 1984, 1995) is considered in our work. It has been shown that this distribution has more flexibility than the Beta and Kumaraswamy distributions in literature (see for example: Alexander, Cordeiro, Ortega, & Sarabia, 2012). We call the distribution which shall be obtained by mixing the McDonald's Generalized Beta distribution of the first kind to the success probability of Binomial distribution in our work as the McDonald Generalized Beta-Binomial distribution (McGBB). It can also be shown that our new mixture distribution includes both Beta-Binomial distribution and Kumaraswamy-Binomial distribution as its nested distributions.

The paper is organized as follows: In section 2, we present a brief review of Beta, Kumaraswamy, McDonald's generalized beta of the first kind, Beta-Binomial and Kumaraswamy-Binomial distributions. The McDonald Generalized Beta-Binomial distribution is developed in section 3 by deriving the probability mass function and moments. The section 4 demonstrates that both BB distribution and KB distribution are nested distributions of McGBB distribution. In section 5, we discuss the parameter estimation of McGBB distribution by means of Maximum Likelihood Estimation technique. An empirical overdispersed binomial dataset is analyzed with our McGBB distribution and its nested distributions, and a comparison study is done in section 6. Then, a simulation study is presented in section 7 to compare the performance of McGBB model with BB model and KB model in handling overdispersed binomial data. Finally, section 8 provides some concluding remarks.

2. Review on Key Ingredients

In this section we briefly outline the three mixing distributions of the random variable P, and the two Binomial mixture distributions, Beta-Binomial distribution and Kumaraswamy-Binomial distribution.

2.1 Beta Distribution

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

$$f_{Beta}\,(p; a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B\,(a, b)};\ 0 \leq p \leq 1\ and\ a, b > 0, \tag{1}$$

where $B\,(a, b)$ denotes a beta function.

2.2 Kumaraswamy Distribution

Let P be a random variable following a Kumaraswamy distribution (Kumaraswamy, 1980), with two shape parameters $\zeta$and ϑ, denoted by $Kumaraswamy\,(\zeta, \theta)$. The probability density function of P is then given by

$$f_{Kum}\,(p;\zeta,\theta) = \zeta\theta p^{\zeta-1}(1-p)^{\theta-1};\ 0 \le p \le 1 \ and\ \zeta,\theta > 0 \tag{2}$$

2.3 McDonald's Generalized Beta Distribution of the First Kind

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

$$f_{GB1}\,(p;\ \alpha,\beta,\gamma) = \frac{\gamma}{B\,(\alpha,\beta)} p^{\alpha\gamma-1}(1-p^{\gamma})^{\beta-1};\ 0 \le p \le 1 \ and\ \alpha,\beta,\gamma > 0. \tag{3}$$

The r^{th} moment of the McDonald's Generalized Beta Distribution of the first kind is given by,

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

It is easy to show that the GB1 distribution is reduced to Beta distribution when $\gamma = 1$, and reduced to Kumaraswamy distribution when $\alpha = 1$.

2.4 Beta-Binomial Distribution and Kumaraswamy-Binomial Distribution

The Beta-Binomial(BB) distribution is obtained from mixing the Binomial probability of success P over a Beta distribution defined in section (2.1). Suppose $Y|p \sim\ Bin\,(n, P)$ and $P \sim\ Beta\,(a, b)$. The probability mass function (PMF) of BB distribution is then given by

$$P_{BB}\,(y) = \frac{B\,(a+y, n+b-y)}{B\,(a,b)};\ y = 0, 1, \ldots, n \ and\ a, b > 0. \tag{5}$$

Likewise, the Kumaraswamy-Binomial distribution (Li et al., 2011) is obtained by mixing the Binomial probability of success P over a Kumaraswamy distribution defined in section (2.2). Suppose $Y|p \sim\ Bin(n, P)$ and $P \sim Kumaraswamy\,(\zeta, \theta)$. The PMF of KB distribution is then given by,

$$P_{KB}\,(y) = \zeta\theta\binom{n}{y}\sum_{i=0}^{\infty}(-1)^i B\,(y+\zeta+\zeta i, n-y+1);\ y = 0, 1, \ldots, n \ and\ \zeta,\theta > 0. \tag{6}$$

3. The McDonald Generalized Beta-Binomial Distribution (McGBB)

In this section we define the McDonald Generalized Beta-Binomial distribution and derive some basic properties of the same distribution. We begin with the definition of developing Binomial mixture distributions.

Generally, a Binomial mixture distribution is obtained through an integration approach. Conditional on p, suppose Y follows a Binomial distribution given by $Bin\,(n, P)$, which is denoted by $Y|p \sim Bin\,(n, P)$. Unconditional probability mass function of the Y can be obtained by evaluating the well-known integral,

$$P_Y\,(y) = \int P_{Y|p}\,(y)\, f_P\,(p|\Theta)\, dp. \tag{7}$$

For $y = 0, 1, \ldots, n$ and Θ is the parameter space of the mixing distribution.

Definition 3.1 A random variable Y is said to have the McDonald Generalized Beta-Binomial (McGBB) distribution with parameters n, α, β and γ if and only if it satisfies the following stochastic representation

$$Y|p \sim Bin\,(n, p) \quad and \quad P \sim GB1\,(\alpha,\beta,\gamma)$$

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

Theorem 3.1 *Let Y be a discrete random variable that follows a McDonald Generalized Beta-Binomial distribution as defined in Definition 3.1. Then the following results hold:*

(i) The probability mass function of McGBB $(n, \alpha, \beta, \gamma)$ *is given by,*

$$P_{McGBB}(y; n, \alpha, \beta, \gamma) = \binom{n}{y} \frac{\gamma}{B(\alpha, \beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta - 1}{i} B(y + \alpha\gamma + \gamma i, n - y + 1), \tag{8}$$

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

(ii) A rearranged probability mass function of McGBB $(n, \alpha, \beta, \gamma)$ *is given by,*

$$P_{McGBB}(y; n, \alpha, \beta, \gamma) = \binom{n}{y} \frac{1}{B(\alpha, \beta)} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma} + \alpha + \frac{j}{\gamma}, \beta\right), \tag{9}$$

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

(iii) The r^{th} *moment of McGBB* $(n, \alpha, \beta, \gamma)$ *is given by,*

$$E(Y^r) = n \frac{B(\alpha + \beta, r/\gamma)}{B(\alpha, r/\gamma)}.$$

Then the mean and variance of McGBB $(n, \alpha, \beta, \gamma)$ *are*

$$E(Y) = n\pi \quad and \quad Var(Y) = n\pi(1 - \pi)\{1 + (n-1)\rho\}, \tag{10}$$

respectively, where

$$\pi = \frac{B\left(\alpha + \beta, {}^1/_\gamma\right)}{B\left(\alpha, {}^1/_\gamma\right)} \quad and \quad \rho = \frac{\left(\frac{B(\alpha+\beta, {}^2/_\gamma)}{B(\alpha, {}^2/_\gamma)}\right) - \left(\frac{B(\alpha+\beta, {}^1/_\gamma)}{B(\alpha, {}^1/_\gamma)}\right)^2}{\left(\frac{B(\alpha+\beta, {}^1/_\gamma)}{B(\alpha, {}^1/_\gamma)}\right) - \left(\frac{B(\alpha+\beta, {}^1/_\gamma)}{B(\alpha, {}^1/_\gamma)}\right)^2}.$$

Here ρ *can be termed as overdispersion parameter of the McGBB distribution.*

Proof. (i) Let$Y|p \sim Bin(n, P)$ and $P \sim GB1(\alpha, \beta, \gamma)$. Then the unconditional PMF of Y can be obtained by using Equation (7) as below,

$$P_Y(y) = \int_{p=0}^{1} P_{Y|p}(y) f_{GB1}(p;\ \alpha, \beta, \gamma)\ dp = \int_{p=0}^{1} \binom{n}{y}\ p^y (1-p)^{n-y} \frac{\gamma}{B(\alpha, \beta)} p^{\alpha\gamma - 1}(1 - p^\gamma)^{\beta - 1}\ dp.$$

By adding the Binomial series representation of $(1 - p^\gamma)^{\beta-1}$ for the above, we get

$$P_Y(y) = \binom{n}{y} \frac{\gamma}{B(\alpha, \beta)} \int_{p=0}^{1} p^{y+\alpha\gamma-1}(1-p)^{n-y} \left(\sum_{i=0}^{\infty} \binom{\beta-1}{i} (-p^\gamma)^i\right) dp.$$

Since the Binomial series is a Power series, it could be integrated term by term and hence we have

$$P_Y(y) = \binom{n}{y} \frac{\gamma}{B(\alpha, \beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} \int_{p=0}^{1} p^{y+\alpha\gamma+\gamma i-1}(1-p)^{n-y} dp.$$

Since $\mathcal{Re}(y + \alpha\gamma + \gamma i) > 0$ and $\mathcal{Re}(n - y) > 0$, [where $\mathcal{Re}(.)$ represents the real part of the number] the inner integral can be represented using a Beta function as,

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

which is the PMF given in Equation (8). Although the above PMF is a valid PMF as it is obtained by means of well-known stochastic compound formula, an infinite series occurs inside the PMF. Therefore it is of interest to know whether the above infinite series can be represented as a finite series. The second part of this theorem rearranges the above PMF as given below.

(ii) Now to obtain the rearranged probability mass function of $McGBB\,(n, \alpha, \beta, \gamma)$, let us begin with,

$$\sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B\,(y+\alpha\gamma+\gamma i, n-y+1) = \int\limits_{p=0}^{1} p^{y+\alpha\gamma-1}\,(1-p)^{n-y}(1-p^{\gamma})^{\beta-1}\,dp.$$

Now consider the binomial series expansion of $(1-p)^{n-y}$ in the above integral. Since $n-y \;\geq 0$ and a positive integer, this series terminates at $n-y$, and can be written in the form $\sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} p^j$. Thus,

$$\sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B\,(y+\alpha\gamma+\gamma i,\; n-y+1) = \int\limits_{p=0}^{1} p^{y+\alpha\gamma-1} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} p^j (1-p^{\gamma})^{\beta-1}\,dp$$

$$= \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} \int\limits_{p=0}^{1} (p^{\gamma})^{\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma}-\frac{1}{\gamma}\right)}(1-p^{\gamma})^{\beta-1}\,dp.$$

Substitute $p = z^{\frac{1}{\gamma}}$, hence $dp = \frac{1}{\gamma} z^{\frac{1}{\gamma}-1} dz$,

$$= \frac{1}{\gamma} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} \int\limits_{z=0}^{1} (z)^{\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma}-1\right)}(1-z)^{\beta-1}\,dz.$$

Again, since $\mathcal{R}e\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma}\right) > 0$ and $\mathcal{R}e\,(\beta) > 0$, the inner integral can be represented using a Beta function and hence,

$$\sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B\,(y+\alpha\gamma+\gamma i,\; n-y+1) = \frac{1}{\gamma} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma},\, \beta\right). \tag{11}$$

Now by inserting Equation (11) in Equation (8) we have,

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

(iii) To obtain the r^{th} moment about zero, mean and variance of $McGBB\,(n, \alpha, \beta, \gamma)$, apply the following well-known identities of probability theory,

- Conditional Expectation Identity $E\,(Y^r) = E_P\,(E\,(Y^r|P))$ and
- Conditional Variance Identity $Var\,(Y) = E_P\,(Var\,(Y|P)) + Var_P\,(E\,(Y|P))$.

Since $Y|p \;\sim Bin\,(n, P)$ it follows that,

$$E\,(Y^r) = E_P\,(E\,(Y^r|P)) = n\;E_P\,(P^r)$$

and

$$Var\,(Y) = E_P\,(nP\,(1-P)) + Var_P\,(nP)$$

Now, by substituting the moments of $P \;\sim\; GB1\,(\alpha, \beta, \gamma)$ given in Equation (4) in the above two identities, the moments of $McGBB\,(n, \alpha, \beta, \gamma)$ stated in Equation (10) can be derived. □

4. Nested Distributions of McGBB

In this section we show that the McGBB distribution can be nested to the BB distribution and KB distribution under specific parameter settings.

Proposition 4.1 *Let $Y \sim McGBB\,(n, \alpha, \beta, \gamma)$, then by setting $\gamma = 1$, we obtain the Beta-Binomial distribution with parameters n, α and β.*

Proof. For $\gamma = 1$, the PMF of McGBB in Equation (8) becomes,

$$P_Y(y) = \binom{n}{y} \frac{1}{B(\alpha,\beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B(y+\alpha+i,\ n-y+1),$$

$$P_Y(y) = \binom{n}{y} \frac{1}{B(\alpha,\beta)} \left(\int_{p=0}^{1} p^{y+\alpha-1}\, (1-p)^{n-y} \sum_{i=0}^{\infty} \binom{\beta-1}{i} (-p)^i \, dp \right),$$

$$P_Y(y) = \binom{n}{y} \frac{1}{B(\alpha,\beta)} \left(\int_{p=0}^{1} p^{y+\alpha-1}\, (1-p)^{n-y+\beta-1} \, dp \right).$$

Since $\mathcal{R}e\,(y+\alpha) > 0$ and $\mathcal{R}e\,(n-y+\beta) > 0$, we have

$$P_Y(y) = \binom{n}{y} \frac{B(y+\alpha, n-y+\beta)}{B(\alpha,\beta)},$$

which is the PMF of Beta-Binomial distribution. □

Proposition 4.2 *Let $Y \sim McGBB(n,\alpha,\beta,\gamma)$, then by setting $\alpha = 1$, we obtain the Kumarasswamy-Binomial distribution with parameters n, β and γ.*

Proof. For $\alpha = 1$, the PMF of $McGBB(n,\alpha,\beta,\gamma)$ in Equation (8) becomes,

$$P_Y(y) = \binom{n}{y} \frac{\gamma}{B(1,\beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B(y+\gamma+\gamma i,\ n-y+1) = \gamma\beta \binom{n}{y} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B(y+\gamma+\gamma i,\ n-y+1),$$

which is the PMF of Kumaraswamy-Binomial distribution. □

Probability Mass Function plots of McGBB for some arbitrary parameter values are shown in Figure 1. Here, a graphical comparison of McGBB is illustrated with its nested mixture distributions, BB and KB, by fixing the common parameters and allowing the additional parameter to vary. These comparison plots indicate that the additional parameter in our new Binomial mixture distribution has more impact on the shape of the PMF compared to that of its nested distributions.

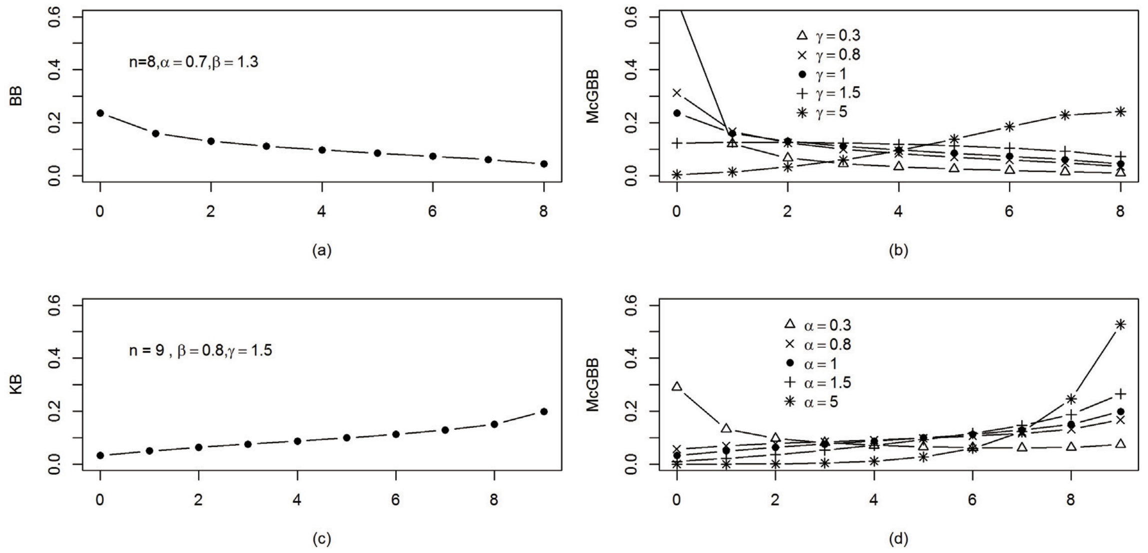

Figure 1. (a) PMF of Beta-Binomial for $n = 8, \alpha = 0.7, \beta = 1.3$ and (b) PMF of of McDonald Generalized Beta-Binomial distribution for $n = 8, \alpha = 0.7, \beta = 1.3$ and some varying values of γ. (c) PMF of Kumaraswamy-Binomial for $n = 9, \beta = 0.8, \gamma = 1.5$ and (d) PMF of McDonald Generalized Beta-Binomial distribution for $n = 9, \beta = 0.8, \gamma = 1.5$ and some varying values of α

5. Estimation of Parameters of McGBB

In this section we consider the Maximum Likelihood Estimation (MLE) of the three unknown parameters of the McGBB distribution. Let $Y = (y_1, y_2....y_N)^T$ be a random sample of size N from a McGBB distribution with unknown parameter vector $\boldsymbol{\Theta} = (\alpha,\beta,\gamma)^T$. Then the log-likelihood function for $\boldsymbol{\Theta}$ can be defined either by considering the probability mass function given in Equation (8) or Equation (9). However, the log-likelihood function in terms of the PMF given in (8) proceeds too many iterations as there is an infinite series that results unnecessary computer time. This is one of the reasons which motivated us to rearrange the PMF of the McGBB distribution from Equation (8) to Equation (9). Thus the log-likelihood function for $\boldsymbol{\Theta}$ can be defined as follows,

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

The Maximum Likelihood Estimates $\hat{\boldsymbol{\Theta}} = \left(\hat{\alpha},\hat{\beta},\hat{\gamma}\right)^T$ can be obtained either by directly maximizing the above log-likelihood function with respect to $\boldsymbol{\Theta}$ or by solving the three simultaneous equations obtained by equating $\mathbf{U}(\boldsymbol{\Theta}) = \mathbf{0}$. The score function $\mathbf{U}(\boldsymbol{\Theta})$ is defined as the gradient of $\ell\ (\boldsymbol{\Theta})$, derived by taking the partial derivatives of $\ell\ (\boldsymbol{\Theta})$ with respect to α,β and γ. The components of the score function $\mathbf{U}(\boldsymbol{\Theta}) = \left(U_\alpha(\boldsymbol{\Theta}), U_\beta(\boldsymbol{\Theta}),\ U_\gamma(\boldsymbol{\Theta})\right)^T$ are given below

$$U_\alpha(\boldsymbol{\Theta}) = \frac{\partial\ell\ (\boldsymbol{\Theta})}{\partial\alpha} = N\left(\psi(\alpha+\beta) - \psi(\alpha)\right) + \sum_{k=0}^{N} \frac{1}{C_k}\left(\sum_{j=0}^{n-y_k} (-1)^j \binom{n-y_k}{j} D_{\alpha k}\right),$$

$$U_\beta(\boldsymbol{\Theta}) = \frac{\partial\ell\ (\boldsymbol{\Theta})}{\partial\beta} = N\left(\psi(\alpha+\beta) - \psi(\beta)\right) + \sum_{k=0}^{N} \frac{1}{C_k}\left(\sum_{j=0}^{n-y_k} (-1)^j \binom{n-y_k}{j} D_{\beta k}\right),$$

$$U_\gamma(\boldsymbol{\Theta}) = \frac{\partial\ell\ (\boldsymbol{\Theta})}{\partial\gamma} = \sum_{k=0}^{N} \frac{1}{C_k}\left(\sum_{j=0}^{n-y_k} (-1)^j \binom{n-y_k}{j} D_{\gamma k}\right),$$

where

$$C_k = \sum_{j=0}^{n-y_k} (-1)^j \binom{n-y_k}{j} B\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma},\ \beta\right),$$

$$D_{\alpha k} = B\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma},\ \beta\right)\left[\psi\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma}\right) - \psi\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma} + \beta\right)\right],$$

$$D_{\beta k} = B\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma},\ \beta\right)\left[\psi(\beta) - \psi\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma} + \beta\right)\right],$$

$$D_{\gamma k} = \left(\frac{y_k+j}{\gamma^2}\right) B\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma},\ \beta\right)\left[\psi\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma} + \beta\right) - \psi\left(\frac{y_k}{\gamma} + \alpha + \frac{j}{\gamma}\right)\right],$$

and $\psi(\cdot)$ is the digamma function.

In particular, the optimization method "Simplex algorithm for minimization" (Nelder, 1965) is used to minimize the user defined negative log-likelihood function with respect to $\boldsymbol{\Theta}$ in our study.

6. Applications of McGBB in Handling Overdispersion

This section demonstrates the superiority of McGBB distribution over its nested binomial mixture distributions BB and KB in handling overdispersed binomial outcome data. We compare the goodness of fit and the Analysis of Deviance(ANODEV) results of the three Binomial mixture models in modeling a real world data that exhibits overdispersion relative to the Binomial distribution. The data for this section is taken from Alanko and Lemmens (1996), which have also been previously used by Rodríguez-Avi et al. (2007) and Li et al. (2011) for similar purposes.

6.1 Data Description

The numbers of alcohol consumption days in two reference weeks are separately self-reported by a randomly selected sample of 399 respondents in the Netherlands in 1983. The number of days an individual consumes alcohol Y, out of n=7 days in a reference week can be treated as a Binomial variable. However, the Binomial success probability p, the probability to consume alcohol on a randomly chosen day in a reference week for an individual, cannot be treated as a constant in this setup since there is a person-to-person variation in the inclination to drink and the drinking behavior. This leads to analyze this data using a Binomial mixture distribution by modeling the random variable P using a continuous distribution bounded in the standard unit interval. Alanko and Lemmens (1996) modeled this data using the Beta-Binomial distribution, Li et al. (2011) approached this data with the Kumaraswamy-Binomial distribution.

6.2 Modeling Results and Discussions

We model the alcohol consumption data by means of McGBB distribution by estimating the Maximum Likelihood Estimates $\hat{\mathbf{\Theta}} = \left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right)^T$ as described in section 5. The MLEs $\hat{a}$ and $\hat{b}$ of BB model are taken as starting values for α and β and the initial value of γ is taken as 1 in the numerical iterative procedures.

Table 1. presents the modeling results of BB model (Alanko & Lemmens, 1996), KB model (Li et al., 2011) and new McGBB model. MLEs, Log-Likelihood values, χ^2 statistics. Degrees of Freedoms (DF) and the corresponding p-values of the three models for both weeks are reported.

Table 1. BB, KB and McGBB modeling results of alcohol consumption data

Number of Drinking Days	Observed Frequency (Week 1)	Expected BB Frequency (Week 1)	Expected KB Frequency (Week 1)	Expected McGBB Frequency (Week 1)	Observed Frequency (Week 2)	Expected BB Frequency (Week 2)	Expected KB Frequency (Week 2)	Expected McGBB Frequency (Week 2)
0	47	54.6	54.30	51.29	42	47.9	47.72	45.92
1	54	42.0	41.74	45.67	47	42.9	42.81	45.13
2	43	38.9	38.86	43.17	54	41.9	41.98	44.75
3	40	38.5	38.70	41.61	40	42.5	42.57	44.50
4	40	40.1	40.38	40.52	49	44.3	44.45	44.35
5	41	44.0	44.41	40.01	40	47.8	48.00	44.51
6	39	53.1	53.51	41.83	43	54.9	55.02	46.57
7	95	87.8	87.10	94.90	84	76.7	76.45	83.26
Total	399	399	399	399	399	399	399	399
χ^2		9.6	9.9837	2.162		9.7	9.8632	4.0036
DF		5	5	4		5	5	4
p-value		0.086	0.0757	0.706		0.082	0.0792	0.4055
Maximum		$\hat{a} = 0.722$	$\hat{\zeta} = 0.700$	$\hat{\alpha} = 0.037$		$\hat{a} = 0.857$	$\hat{\zeta} = 0.850$	$\hat{\alpha} = 0.037$
Likelihood		$\hat{b} = 0.581$	$\hat{\theta} = 0.590$	$\hat{\beta} = 0.195$		$\hat{b} = 0.700$	$\hat{\theta} = 0.710$	$\hat{\beta} = 0.278$
Estimates				$\hat{\gamma} = 24.023$				$\hat{\gamma} = 26.350$
Log-Likelihood		-813.4571	-813.6939	-809.68		-821.3922	-821.4543	-818.48

As can be seen in Table 1, the p-values in the Chi-Square goodness of fit tests for both Beta-Binomial (0.086 for week 1 and 0.082 for week 2) and Kumaraswamy-Binomial (0.0757 for week 1 and 0.0792 for week 2) models are noticeably small. Further, there are considerably large discrepancies between the expected frequencies obtained by means of both Beta-Binomial models and Kumaraswamy-Binomial models and the actual observed frequencies. Also, Li et al. (2011) noted that both Beta-Binomial distribution and Kumaraswamy-Binomial distribution, which have only two additional parameters, have the same flexibility in modeling this consumption data. Thus we conclude that both Beta-Binomial and Kumaraswamy-Binomial modeling approaches are not very satisfactory in analyzing this data. On the other hand, a Generalized Beta-Binomial(GBB) distribution proposed and developed by Rodríguez-Avi et al. (2007) possesses great flexibility in modeling this data. Since Rodríguez-Avi's GBB distribution was developed in a different context, we do not pay much attention on that distribution at present.

However, the newly proposed McGBB distribution provides an admirable fit to the alcohol consumption data compared to its nested distributions. The discrepancy between the observed frequencies and the expected frequencies is much reduced in McGBB model over BB and KB models. For example, the observed number of respondents who consume alcohol in all seven days of a week is 95 over the week 1 and 84 over the week 2; the BB model provides the expected frequencies 87.8 and 76.7 for this; the KB model provides a similar 87.1 and 76.45 for this; While, as anticipated, the McGBB model results 94.90 and 83.26 for expected number of respondents who consume alcohol in all seven days of a week over the week 1 and week 2 respectively. The lesser discrepancies between the observed and expected frequencies in the McGBB model also results a substantial decrease in the χ^2 goodness of fit test statistic and accordingly larger p-values(0.7060 for week 1 and 0.4055 for week 2). This indicates that for any standard statistical significance level McGBB distribution models this data well whereas BB and KB distributions fail to fit.

Moreover, ANODEV results which compare McGBB model with its nested models in fitting this alcohol consumption data are presented in Table 2.

Table 2. ANODEV to compare McGBB models with nested models in fitting alcohol consumption data

Hypothesis	Comparison	Week	Deviance Difference	DF	p-values
1. $H_0 : \gamma = 1\ vs\ H_1 : \gamma \neq 1$	*McGBB vs BB*	Week 1	7.5456	1	0.006016
		Week 2	5.8248	1	0.015800
2. $H_0 : \alpha = 1\ vs\ H_1 : \alpha \neq 1$	*McGBB vs KB*	Week 1	8.0195	1	0.004628
		Week 2	5.9502	1	0.014720

The p-values in Table 2. indicate that both BB and KB models are significantly rejected in favor of the McGBB model to fit this alcohol consumption data. Therefore, based on these results, we conclude that the proposed McGBB distribution provides better fit to model this data than its nested BB and KB distributions.

7. Simulation Study

In this section, we present a Monte Carlo Simulation study conducted to investigate the performance of McGBB distribution with its nested BB and KB distributions in modeling simulated overdispersed binomial outcome data under varying degrees of overdispersion.

7.1 Generation of Overdispersed Binomial Variates

In general, random generation from the Beta-Binomial distribution is used as a standard method to simulate overdispersed Binomial variables (See for example: Ennis & Bi, 1998). However, since our present study focuses on comparing three Binomial mixture distributions including the Beta-Binomial distribution in handling overdispersed Binomial data, there is a suspicion that perhaps the results may be influenced towards Beta-Binomial distribution. Therefore an alternative algorithm proposed by Ahn and Chen (1995) is used to simulate overdispersed Binomial variables. The algorithm developed by Ahn and Chen (1995) to generate overdispersed Binomial variables for specified mean and variance from an underlying multivariate normal distribution is simplified using equal correlation structure and briefly outlined in subsection 7.1.1.

7.1.1 Representation of Overdispersed Binomial Variables

Let the overdispersed binomial random variable Y_i be the sum of the correlated binary variables $X_{i1},\ X_{i2},\ldots,X_{in}$ where n is the number of trials and $i = 1, 2, \ldots, N$. Suppose, $E\left(X_{ij}\right) = \pi$, $Var\left(X_{ij}\right) = \pi\ (1-\pi)$ and $corr\left(X_{ij}, X_{ik}\right) = \rho$, for $j \neq k$. Then the mean and the variance of Y_i is given by

$$E\left(Y_i\right) = n\pi \quad and \quad Var\left(Y_i\right) = n\pi\left(1-\pi\right)\left(1+\left(n-1\right)\rho\right). \tag{12}$$

The parameter ρ is the overdispersion (also, the intracluster correlation coefficient) parameter of the overdispersed Binomial random variable. This explains that when $\rho \to 0$ this distribution reduces to Binomial distribution and $\rho \to 1$ results severe overdispersion. The means and the variances of the three Binomial mixture distributions under consideration can be represented similar to Equation (12), nevertheless, the expressions of π and ρ depend on the mixing distribution. Such a representation of the mean and the variance of the McGBB distribution is given in Equation (10).

7.1.2 Algorithm to Generate Overdispersed Binomial Random Variable

Step 1: Solve the following equation for a given n, π and ρ,

$$\Phi\left[z\left(\pi\right),\ z\left(\pi\right), \delta\right] = \pi\left(1-\pi\right)\rho + \ \pi^2.$$

For δ , where $\Phi\left[z\left(\pi\right),\ z\left(\pi\right), \delta\right]$ is the cumulative distribution function of the standard bivariate normal random variable with correlation coefficient δ , and $z\left(\pi\right)$ denotes the π^{th} quantile of the standard normal distribution.

Step 2: Generate $n-$dimensional multivariate normal random variables, $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in})^T$ with mean $\mathbf{0}$ and constant correlation matrix $\sum_i$ for $i = 1, 2, \ldots, N$, where the elements of $[\sum_i]_{lm}$ are δ for $l \neq m$.

Step 3: Now, for each $j = 1, 2, \ldots, n$ define

$$X_{ij} = \begin{cases} 1; & if\ Z_{ij} < z(\pi) \\ 0; & otherwise. \end{cases}$$

Then, it can be showed that the random variable $Y_i = \sum_{j=1}^{n} X_{ij}$ is overdispersed relative to the Binomial distribution.

The reason for the above follows from the fact that $E\left(X_{ij}\right) = P\left\{X_{ij} = 1\right\} = P\left\{Z_{ij} < z(\pi)\right\} = \pi$ and

$$corr\left(X_{ij}, X_{ik}\right) = \frac{cov\left(X_{ij}, X_{ik}\right)}{\sqrt{Var\left(X_{ij}\right)\ Var\ \left(X_{ij}\right)}} = \frac{E\left(X_{ij}X_{ik}\right) - E\left(X_{ij}\right)^2}{\sqrt{Var\left(X_{ij}\right)\ Var\ \left(X_{ij}\right)}} = \frac{P\left\{Z_{ij} < z(\pi),\ \ Z_{ik} < z(\pi)\ \right\} - \pi^2}{\pi\ (1-\pi)} = \rho.$$

Thus, it is apparent that the correlated binary variables generated by this algorithm encompass an overdispersed Binomial random variable which is characterized in Equation (12).

7.2 Simulation Design

In a simulation study like this, determining the values of the parameters to be used to generate required data is indeed a challenging task. In the alcohol consumption data presented in the previous section the McGBB estimate of π for both week 1 and week 2 is around 0.5 while the similar estimate of overdispersion parameter ρ for both week 1 and week 2 is around 0.4. Also, the number of trails in this data is 7 and the total number of observations is 399. Even though we can generate overdispersed data for different ρ values by keeping the other parameters π, n and N fixed as those in this data, in the present study we have conducted an extended simulation to clearly understand the complete scope of the problem. Consequently, in this simulation study, three π values are chosen as 0.1 (extreme), 0.5 (middle) and 0.75 (moderate), [here the range of success probabilities from 0-0.5 is adequate as the other half of the possible π values can be included by modeling the failure probability, however without loss of generality, we include $\pi = 0.75$ as a moderate value instead of $\pi = 0.25$]; two n values are chosen as 5(lower) and 10(greater); two N values are selected as 20 (small number of observations) and 500 (large number of observations). Further, 10 overdispersion parameter ρ values are picked from 0.05 to 0.9 in the increasing order since the main objective of this simulation study is understanding the behavior and comparing the performance of the three Binomial mixture distributions discussed above for different degrees of overdispersion. Thus, we run a total of $3 \times 2 \times 2 \times 10 = 120$ factorial combinations of the four factors $\pi = \{0.1, 0.5, 0.75\}$, $n = \{5, 10\}$, $N = \{20, 500\}$ and $\rho = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The value of $\rho = 0.05$ is chosen just to observe the performance of the mixture models when there is almost no overdispersion.

For each of the 120 parameter combinations, we then simulate 1000 overdispersed Binomial datasets (total of 120,000 datasets) using the algorithm described above. Starting from the data generation, the entire simulation study (including the model fitting, evaluation and comparison presented below) is programmed and performed using the open source statistical software R and RStudio: an integrated development environment for R.

The types of simulated data herein may arise in many real world applications. For instance, the surveys of consumption of a product or service for a small time frame, like the one described in section 6; or any other types of behavior reporting in a short retrospective time period, such as the consumer purchasing behavior investigated by Chatfield and Goodhardt (1970). Moreover, other types of data such as plant decease incidence data modeled using Beta-Binomial distribution by Hughes and Madden (1993).

7.3 Model Evaluation Procedures

For each of 120,000 datasets the maximum likelihood estimates of the three Binomial mixture distributions (McGBB, BB and KB) are obtained along with the Log-Likelihood values and the AIC values of the model fit. Then, by comparing the observed frequencies with the expected frequencies obtained by means of each of the three estimated Binomial mixture models, the Chi-Square goodness of fit test statistics, the number of degrees of freedom and the associated p-values to test that the data is consistent with the distribution specified under the null hypothesis are also obtained. The model evaluations are done in two procedures as stated below.

Procedure 1:

The three Binomial mixture models are evaluated individually for each of 120 parameter combinations. This is done graphically by comparing the boxplots of the calculated AIC values and also for an inferential discussion, the percentage of p-values that significantly reject the Binomial mixture model under consideration at 5% significance level are reported and evaluated. These measures are also used to compare across the models for each set of parameter combinations.

Procedure 2:

Pairwise comparisons are done for McGBB model with its nested BB and KB models for each of the 120,000 simulated datasets. Here, we perform Analysis of Deviance (ANODEV) by calculating the deviance difference using the Log-Likelihood values to determine whether the complex McGBB model with an additional parameter provides a significantly better fit compared to its nested models. Again, the percentage of p-values that significantly rejects the simple model over the complex McGBB model at 5% significance level are reported for each of the 120 parameters combinations.

7.4 Results of the Simulation Study

First we compare the individual evaluation of the three Binomial mixture models by means of the methods stated above in Procedure 1. The percentage of the significant p-values which leads to reject the model under consideration at 5% significance level, out of the 1000 simulated datasets for each of the 120 parameters combinations are presented in Table 3.

It can be seen from these numbers that, regardless of the π and n values, the rejection percentage of all three Binomial mixture models increases with increasing number of observations for two extremes of degree of overdispersion [$\rho = 0.05, 0.1$ and $\rho = 0.9$]. This large scale increase in the rejection percentage of all three Binomial mixture models with N, also continues for extreme π values from $\rho = 0.2$ to $\rho = 0.5$. The same continues in the KB models even for $\rho = 0.6, 0.7$ and 0.8. Moreover, the KB model fails to fit any of the dataset out of the 1000 simulated datasets for high overdispersion, extreme probability and large number of observations irrespective of the number of trials, which is a major drawback of KB model. The significant p-value rejection percentage does not increase with N in a considerable magnitude for the BB model and the new McGBB model for middle and moderate π values, from $\rho = 0.2$ to $\rho = 0.7$ despite the changes in the n values.

The across comparison of all three Binomial mixture models for each of the parameter combinations suggests that there are no any large differences in the rejection percentages between the three models for most of the parameter combinations except in the case of high overdispersion, in which KB model fails to fit very high number of simulated datasets compared to BB and McGBB models. Further, for low overdispersion values, at some particular parameter combinations both the BB and KB models result a very serious rejection percentage(for example, more than 20% the simulated datasets are rejected by BB models at $\{\pi = 0.1, n = 10, N = 500, \rho \leq 0.1\}$ whereas the newly proposed McGBB models have a less rejection percentage). Even though there are few configurations in which the McGBB model results a slightly high rejection percentage than BB and KB models, this is not a serious issue as BB and KB are nested distributions of McGBB distribution. Besides, it should be noted there may be few simulated datasets which cannot be modeled using all three Binomial mixture models considered herein. By its nature the low probability of success can result binomial outcome datasets that contain excessive amount of zeros, known as zero-inflated data. Relatively high rejection percentage(>10%) of all three Binomial mixture models at $\{\pi = 0.1, n = 10, N = 500, \rho \leq 0.2\}$ is due to this zero-inflation feature.

In addition, the graphical comparison of the distribution of AIC values plotted by calculating the AIC values of all model fits for each of the parameter combinations (Selected boxplots of the distributions of AIC values are given in Appendix A.), does not indicate any noteworthy differences between the three models via the distributions of AIC values, except in the high overdispersion values in which the KB models tend to give large AIC values.

Table 3. The percentage of the significant p-values at 5% significance level which leads to reject the model under consideration out of the 1000 simulated datasets

			$\rho = 0.05$			$\rho = 0.1$			$\rho = 0.2$		
π	n	N	BB	KB	McGBB	BB	KB	McGBB	BB	KB	McGBB
0.1	5	20	0.6	0.6	1.5	0.7	0.7	1.8	1.1	1.1	3.5
0.1	5	500	14.2	9.9	12.3	8.4	7.5	11.6	7.8	6.6	9.6
0.1	10	20	0.4	0.3	0.5	1.2	0.8	1.4	0.8	1.0	1.4
0.1	10	500	21.0	15.9	11.2	25.3	22.5	18.0	17.1	10.5	11.8
0.5	5	20	3.9	4.7	9.1	2.6	2.7	8.6	4.6	4.7	9.5
0.5	5	500	12.9	6.6	14.3	16.3	6.2	10.8	5.3	5.5	8.0
0.5	10	20	4.2	3.4	5.6	3.5	3.4	4.7	4.0	3.7	6.0
0.5	10	500	18.1	5.8	8.8	10.6	6.7	6.8	5.1	4.8	6.0
0.75	5	20	2.6	3.4	6.7	2.9	3.2	7.9	3.9	3.9	7.9
0.75	5	500	6.5	5.9	12.0	4.9	4.8	11.1	6.1	6.1	11.5
0.75	10	20	1.8	1.9	2.7	2.3	2.6	4.1	5.2	5.4	6.2
0.75	10	500	7.5	6.4	9.8	6.7	5.4	9.4	7.5	7.5	11.9
			$\rho = 0.3$			$\rho = 0.3$			$\rho = 0.5$		
π	n	N	BB	KB	McGBB	BB	KB	McGBB	BB	KB	McGBB
0.1	5	20	2.6	2.4	6.0	3.4	3.2	6.7	3.4	3.0	6.5
0.1	5	500	9.4	7.6	12.5	7.3	5.8	11.9	6.8	7.1	12.1
0.1	10	20	2.5	2.8	3.4	2.3	2.3	3.3	2.9	2.5	4.4
0.1	10	500	12.9	7.6	10.8	8.4	6.6	9.5	6.4	7.2	7.5
0.5	5	20	5.0	5.0	9.8	4.8	4.8	9.7	3.6	3.5	6.5
0.5	5	500	6.3	6.3	12.1	5.7	5.7	9.5	5.0	5.1	6.5
0.5	10	20	4.7	4.7	6.5	4.5	4.6	6.7	4.7	4.6	5.2
0.5	10	500	6.0	5.8	8.6	5.4	5.5	7.3	5.1	5.5	5.5
0.75	5	20	3.9	4.2	8.6	4.2	4.3	7.8	4.8	5.0	8.0
0.75	5	500	5.6	5.8	10.3	4.5	4.8	9.1	4.9	4.8	8.2
0.75	10	20	4.1	4.3	5.9	4.4	4.4	5.2	4.2	4.3	4.1
0.75	10	500	6.5	7.6	10.3	6.2	6.3	8.5	5.9	5.5	7.4

Table 3. Continued

			$\rho = 0.6$			$\rho = 0.7$			$\rho = 0.8$			$\rho = 0.9$		
π	n	N	BB	KB	McGBB	BB	KB	McGBB	BB	KB	McGBB	BB	KB	McGBB
0.1	5	20	4.1	3.7	7.1	4.3	4.5	7.5	2.9	1.9	5.0	0.7	2.7	1.8
0.1	5	500	6.1	8.4	8.1	5.7	10.6	7.4	5.3	5.9	7.0	5.2	NA	10.9
0.1	10	20	3.8	3.6	5.4	4.6	4.7	5.1	2.9	1.2	3.7	1.8	2.8	2.6
0.1	10	500	5.7	7.4	4.5	4.9	10.0	5.1	5.2	5.2	5.5	5.2	NA	8.0
0.5	5	20	4.3	4.6	7.7	3.2	3.7	6.8	1.9	4.4	5.9	0.5	2.6	2.6
0.5	5	500	5.1	6.2	6.4	5.7	10.7	6.5	5.2	6.2	6.2	4.6	42.9	5.5
0.5	10	20	4.6	4.8	6.1	4.4	4.5	4.9	4.2	4.1	3.5	3.9	2.7	3.1
0.5	10	500	5.4	7.7	5.9	7.3	12.8	7.5	7.2	7.8	7.0	8.4	43.9	7.6
0.75	5	20	4.0	4.6	7.2	4.5	5.4	6.3	2.9	3.7	4.0	2.0	5.4	2.3
0.75	5	500	4.1	4.5	6.0	5.6	7.0	7.0	4.5	5.1	5.5	4.9	11.8	7.5
0.75	10	20	4.8	5.2	4.5	4.6	5.2	3.8	4.5	3.6	3.1	3.9	8.1	2.0
0.75	10	500	5.8	4.9	6.3	5.0	6.5	5.6	6.7	6.6	6.1	6.6	17.6	7.3

Next, we compare the Analysis of Deviance comparison results that are performed to determine the superiority of complex McGBB models in favor of its nested simpler models in fitting overdispersed binomial outcome data. Table 4. presents the results of Procedure 2. described in section 7.3 for ANODEV comparison of McGBB over KB for the first set of hypotheses stated in Table 2. Correspondingly, Table 5. contains the similar ANODEV comparison of McGBB over KB for the second set of hypotheses stated in the Table 2.

Table 4. Percentage of significant p-values in the ANODEV of McGBB over BB

			Overdispersion Parameter ρ									
π	n	N	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	5	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	5	500	6.0	0.7	0.0	0.3	0.1	0.2	2.8	4.2	2.5	0.4
0.1	10	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	10	500	12.9	14.5	2.0	0.6	0.0	1.7	8.2	7.2	4.1	0.3
0.5	5	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.5	5	500	7.3	11.8	1.7	0.2	1.3	3.5	2.8	3.2	3.4	3.7
0.5	10	20	0.0	0.0	0.0	0.0	0.2	0.4	0.9	1.4	0.6	0.0
0.5	10	500	13.6	7.5	2.9	1.2	1.5	2.2	2.9	4.7	4.9	4.3
0.75	5	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.75	5	500	0.0	0.0	0.0	0.3	0.9	1.1	1.5	1.5	1.9	2.1
0.75	10	20	0.0	0.0	0.0	0.0	0.5	0.9	1.4	1.0	0.3	0.0
0.75	10	500	0.0	0.2	0.0	0.0	0.3	0.7	1.5	4.2	3.8	3.9

Table 5. Percentage of significant p-values in the ANODEV of McGBB over KB

			Overdispersion Parameter ρ									
π	n	N	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	5	20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.1	5	500	2.3	0.4	0.4	0.9	0.2	0.6	8.2	18.4	21.4	NA
0.1	10	20	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.0
0.1	10	500	10.9	15.1	1.0	0.5	0.3	4.4	21.9	31.3	39.3	NA
0.5	5	20	0.2	0.1	0.0	0.0	0.0	0.1	0.3	0.5	0.5	0.0
0.5	5	500	0.6	0.3	1.3	0.2	1.3	4.9	8.1	14.1	21.0	69.5
0.5	10	20	0.1	0.0	0.0	0.0	0.2	0.7	2.0	3.7	5.2	4.9
0.5	10	500	1.1	1.7	1.5	0.8	1.6	5.6	9.9	24.0	42.1	87.6
0.75	5	20	0.9	0.2	0.0	0.0	0.0	0.0	0.2	0.2	0.1	0.4
0.75	5	500	0.0	0.0	0.0	0.3	0.8	1.4	2.8	7.5	11.9	18.1
0.75	10	20	0.0	0.0	0.0	0.0	0.5	1.2	2.3	2.8	2.8	2.8
0.75	10	500	0.0	0.1	0.0	0.0	0.2	0.9	3.1	8.7	21.3	44.6

In Table 4, it is clear that, there are many cells with 0's which indicate that the BB model is adequate to fit all datasets simulated at those parameter combinations. In particular, this suggests that when number of observations (N) is small, regardless of other parameter combinations, the BB model adequately fits the overdispersed data with any degree of overdispersion. On the other hand, when the number of observations is large and for extreme and middle π values irrespective of n values, noticeably high percentage of simulated datasets are favoring the complex McGBB model over the simpler BB model for low overdispersion values ($\rho \leq 0.1$). Again, for the high overdispersion values ($\rho \geq 0.6$) a few percentage of simulated datasets are favoring the complex McGBB model over the simpler BB model when N is large.

Finally, it can be seen from Table 5, that ANODEV comparisons of McGBB vs KB also result many parameter combinations with zero percentage of significant p-values indicating that KB is a minimum adequate model in favor of McGBB model at those parameter settings. Nevertheless, for high overdispersion values ($\rho \geq 0.6$), the simpler KB models are extensively rejected in favor of the complex McGBB models for large N values. For example, at the parameter combination $\{\pi = 0.5\ ,\ n = 10,\ N = 500\ ,\ \rho = 0.9\}$, 87.6% of the simulated datasets are rejected favoring the complex McGBB model over the simpler KB model which is exceptionally high.

Our overall findings of the simulation study are that even though the BB model and the KB model are minimum adequate models compared to complex McGBB model when N is small regardless of the other parameter combinations, which is not true in general. In particular, when N is large there are considerable number of simulated datasets which cannot be modeled either by means of BB or KB models but the proposed McGBB possibly models those datasets well. Note that for many simulated datasets, we observe that McGBB model outperforms BB model for very low degree of overdispersion ($\rho \leq 0.1$) when $N = 500$ and $\pi = 0.1$ and 0.5, as well as McGBB model outperforms KB model for relatively high degree of overdispersion($\rho \geq 0.6$) when $N = 500$ and all π values considered and, also when $\rho \leq 0.1$, $N = 500$, $\pi = 0.1$ and $n = 10$.

8. Concluding Remarks

In the present study, we propose a new three parameter Binomial Mixture distribution, namely the McDonald Generalized Beta-Binomial (McGBB) distribution, by mixing the McDonald's Generalized Beta distribution of the first kind to the success probability of the Binomial distribution. We present two different parameter arrangements of the probability mass function of the McGBB distribution one containing an infinite series and the other with a finite series. The central moments of the McGBB distribution are also obtained along with the mean and variance of the McGBB distribution. The additional parameter in the McGBB distribution allows accommodating wide range of shapes in addition to the shapes that are accommodated by its nested Binomial mixture distributions. The parameters of the McGBB distribution are estimated by maximum likelihood estimation technique. The main objective of our study is comparing the new McGBB mixture distribution with its nested distributions in handling Binomial overdispersion. This objective is achieved by means of a real data and an extended simulation study. The results of the real data shows that the new McGBB mixture model provides a better fit and good improvement in the goodness of fit tests than BB and KB models. From the results of the simulation study, it is also evident that the McGBB model is superior to its nested models for some parameter combinations. Hence the proposed McGBB distribution has a great potential for handling Binomial overdispersion.

Acknowledgements

The authors are grateful to the editor and two reviewers for their helpful comments and suggestions that improved the article.We are also gratefully acknowledge the facilities provided by the Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.

References

Ahn, H., & Chen, J. J. (1995). Generation of over-dispersed and under-dispersed binomial variates. *Journal of Computational and Graphical Statistics, 4*(1), 55-64. http://dx.doi.org/10.1080/10618600.1995.10474665

Alanko, T., & Lemmens, P. H. (1996). Response effects in consumption surveys: an application of the beta-binomial model to self-reported drinking frequencies. *Journal of Official Statistics, 12*(3), 253-273.

Alexander, C., Cordeiro, G. M., Ortega, E. M., & Sarabia, J. M. (2012). Generalized beta-generated distributions. *Computational Statistics and Data Analysis, 56*(2), 1880-1897. http://dx.doi.org/10.1016/j.bbr.2011.03.031

Bandyopadhyay, D., Reich, B. J., & Slate, E. H. (2011). A spatial beta-binomial model for clustered count data on dental caries. *Statistical Methods in Medical Research, 20*(2), 85-102. http://dx.doi.org/10.1177/0962280210372453

Chatfield, C., & Goodhardt, G. J. (1970). The beta-binomial model for consumer purchasing behaviour. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 19*(3), 240-250. http://dx.doi.org/10.2307/2346328

Crowder, M. J. (1978). Beta-binomial Anova for proportions. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 27*(1), 34-37. http://dx.doi.org/10.2307/2346223

Ennis, D. M., & Bi, J. (1998). The beta-binomial model: accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies, 13*(4), 389-412. http://dx.doi.org/10.1111/j.1745-459X.1998.tb00097.x

Gange, S. J., Munoz, A., Saez, M., & Alonso, J. (1996). Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 45*(3), 371-382. http://dx.doi.org/10.2307/2986094

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics, 29*(4), 637-648. http://dx.doi.org/10.2307/2529131

Haseman, J. K., & Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics, 35*(1), 281-293. http://dx.doi.org/10.2307/2529950

Hughes, G., & Madden, L. V. (1993). Using the beta-binomial distribution to describe aggregated patterns of disease incidence. *Phytopathology, 83*(7), 759-763. http://dx.doi.org/10.1094/Phyto-83-759

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (Vol. 444). Hoboken, NJ: Wiley-Interscience.

Jones, M. C. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology, 6*(1), 70-81. http://dx.doi.org/10.1016/j.stamet.2008.04.001

Kotz, S., & Van Dorp, J. R. (2004). *Beyond beta: Other continuous families of distributions with bounded support and applications*. New Jersey, USA: World Scientific Publishing Company Incorporated.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology, 46*(1), 79-88. http://dx.doi.org/10.1016/0022-1694(80)90036-0

Li, X. H., Huang, Y. Y., & Zhao, X. Y. (2011). The Kumaraswamy Binomial Distribution. *Chinese Journal of Applied Probability and Statistics, 27*(5), 511-521.

Lindsey, J. K., & Altham, P. M. E. (1998). Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 47*(1), 149-157. http://dx.doi.org/10.1111/1467-9876.00103

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society, 52*(3), 647-663. http://dx.doi.org/10.2307/1913469

McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics, 66*(1-2), 133-152. http://dx.doi.org/10.1016/0304-4076(94)01612-4

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313. http://dx.doi.org/10.1093/comjnl/7.4.308

Paul, S. R. (1982). Analysis of proportions of affected foetuses in teratological experiments. *Biometrics, 38*(2), 361-370. http://dx.doi.org/10.2307/2530450

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association, 81*(394), 321-327. http://dx.doi.org/10.1080/01621459.1986.10478275

R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing (Version 2.10.0), Vienna, Austria. Retrieved from http://www.R-project.org/

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., & Olmo-Jiménez, M. J. (2007). A generalization of the beta-binomial distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 56*(1), 51-61. http://dx.doi.org/10.1111/j.1467-9876.2007.00564.x

RStudio. (2012). RStudio: Integrated development environment for R (Version 0.95.256), Boston, MA, USA. Retrieved from http://www.rstudio.org/

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological), 10*(2), 257-261. http://dx.doi.org/10.2307/2983779

Tripathi, R. C., Gupta, R. C., & Gurland, J. (1994). Estimation of parameters in the beta binomial model. *Annals of the Institute of Statistical Mathematics, 46*(2), 317-331. http://dx.doi.org/10.1007/BF01720588

Williams, D. A. (1975). 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics, 31*(4), 949-952. http://dx.doi.org/10.2307/2529820

Appendix A: Boxplots of the Distribution AIC Values with Varying Degrees of Overdispersion for All Three Binomial Mixture Models

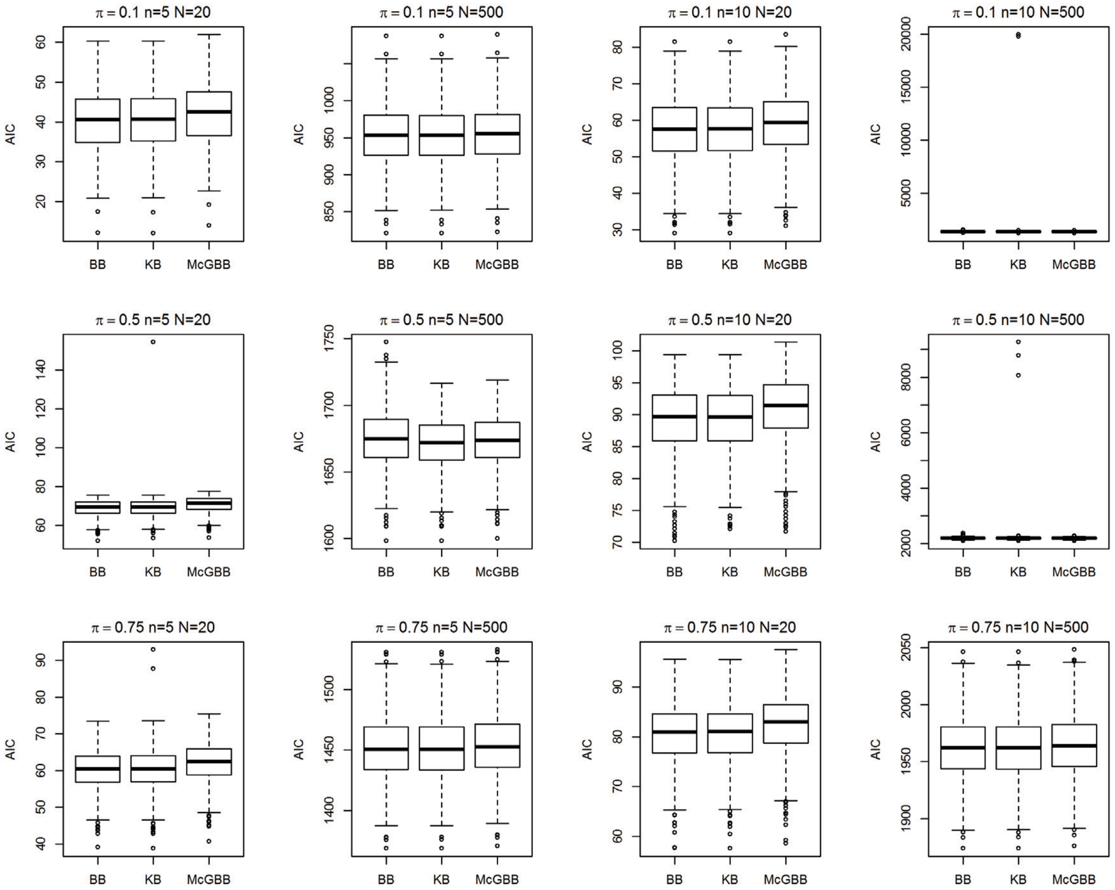

Figure A1. Boxplots of the distribution of AIC values when $\rho = 0.1$

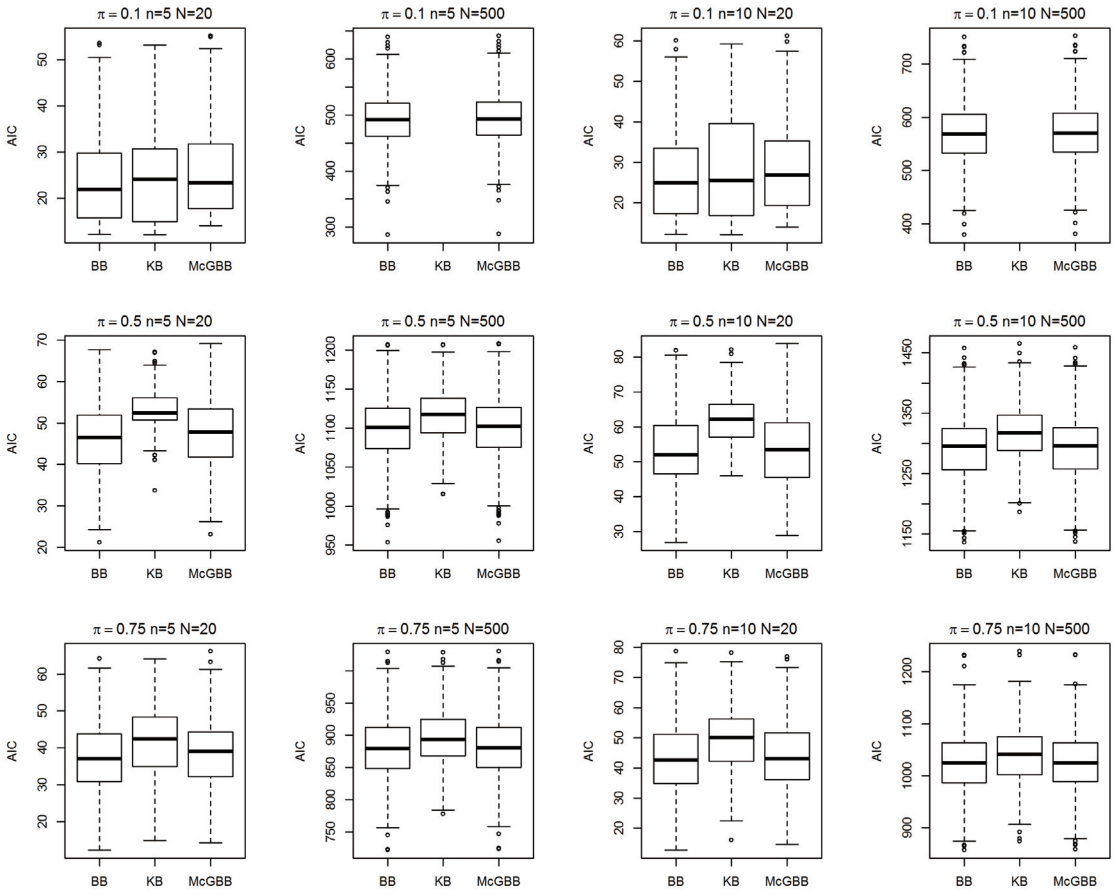

Figure A2. Boxplots of the distribution of AIC values when $\rho = 0.9$

*** Comparison of AIC values via boxplots is presented only for selected parameter combinations since these comparisons do not indicate any noteworthy differences between BB, KB and McGBB models except in the high overdispersion values.

Appendix B: R Codes

- The R function that calculates McGBB probabilities is as follows

```
dMcGBB<- function(x, a, b,c, n) {
  j <- 0:(n-x)
  term<-sum(((-1)**j)*(choose(n-x,j))*(beta(((x/c)+a+(j/c)),b)))
  return(choose(n,x)*(1/beta(a,b))*term)
}
```

- The R function that construct Negative Log-likelihood of the McGBB is as follows

```
McGBBNegLL<-function(x,a,b,c,fre,n){
  density<-c()
  for( i in 0:n){
    j <- 0:(n-i)
    term<-sum(((-1)**j)*(choose(n-i,j))*(beta(((i/c)+a+(j/c)),b)))
    vector.density<-choose(n,i)*(1/beta(a,b))*term
    density[i+1]<-vector.density
  }
  McGBBLL<-sum(fre*log(density))
  return(-McGBBLL)
}
```

Arguments
x = values of the variable
fre = vector of frequencies
n = number of trials
a,b,c = Three parameters of the McGBB Distribution

Optimization functions available in R, such as optim{stats} , mle{stats4} or mle2{bbmle}, can be used to estimate the parameters of the McGBB distribution by minimizing the negative log-likelihood constructed above.

Note: More details on R programming in estimating the parameters of the new McGBB distribution can be obtained from the first author (cmanoj@live.com).