

Some New Results on Overdispersed Models for the Human Sex Ratio Data

Lawal HB*

Department of Statistics and Mathematical Sciences, Kwara State University, Malete, Kwara State, Nigeria

*Corresponding author: Lawal HB, Department of Statistics and Mathematical Sciences, Kwara State University, Malete, Kwara State, Nigeria, E-mail: bayo.lawal@kwasu.edu.ng

Citation: Lawal HB (2018) Some New Results on Overdispersed Models for the Human Sex Ratio Data. J Biostat Biometric App 3(2): 205

Received Date: March 03, 2018 Accepted Date: July 06, 2018 Published Date: July 09, 2018

Summary

We have presented in this paper alternative computation to arrive at results obtained in Lindsey and Altham (1998) on the use of the beta-binomial, the double binomial and the multiplicative binomial models to the Human Sex Ratio data. Our procedure estimates fewer parameters and thus the goodness-of-fit test statistics computed here are based on more *d.f.* The results obtained, using SAS PROC NLMIXED agree with those presented in the Lindsey *et al.* paper. Further, we extend the study to two additional distributions, namely, the Com-Poisson binomial and the correlated binomial distributions. Our results give the estimated probabilities under all the models considered here and we are able to compute the success probability under the multiplicative binomial model for the global data.

Keywords: Beta-binomial; Multiplicative Binomial; Double Binomial; Com-Poisson; Correlated Binomial; Success Probability

Introduction

Lindsey and Altham (1998) re-analyzed the human sex ratio data collected by Geissler (1889) [1,2]. The data relate to the distribution of the sexes of children in families in Saxony during 1876-1885. A reproduction of the data is also presented in Edwards (1958) for distribution of boys of size 2-13 [3]. We have in this paper further provided a presentation of the data arranged by family size and corresponding frequencies. The data has been arranged so that for instance the data used by Sokal and Rohlf (1969), p.80 can now be found in row F12 for $n=12$. Similarly, that used by Fisher (1958) for a family size of $n=8$ can be found in row F08 for $n=8$ [4]. The number of children in each family therefore varies from 1 to 12, and is denoted by n in this paper. A detailed description of the data can be found in Edwards (1958) [3].

Questions of interest from previous studies are succinctly stated in Lindsey and Altham and include 'whether the probability that the child is a boy varies among families and whether it can vary over time within a family'. As argued in Lindsey and Altham, because the Geissler study results in an aggregated data, both questions may not be distinguishable and that they will inherently manifest themselves as overdispersion with respect to the binomial model. As presented in Lindsey and Altham, we will look at whether variability changes with family size.

In this study, we shall compare the behaviors of the binomial model (BN) with those of five other models, namely, the beta-binomial (BB) Skellam (1948); the multiplicative binomial (MBM) Altham (1978), Elamir (2013); the Com-Poisson Binomial (CPM) Borges *et al.* (2013), the double binomial (DBM) Lindsey and Altham (1995) and the correlated binomial Kupper and Haseman (1978) distributions [5-10].

Lindsey & Altham (1998), employed the beta-binomial, the multiplicative and the double binomial models as overdispersed models for the data in Table 1. In their paper, the forms of the multiplicative and double binomial models employed were characterized with intractable normalizing constants $c(n,\psi)$ and $c(n,\pi)$ respectively. Consequently, these models were implemented in Lindsey and Altham (1998) by utilizing a generalized linear model with a Poisson distribution and log link to the frequency data. This approach has earlier being similarly employed in Lindsey and Altham (1995) [1,9]. This approach, which employs joint sufficient statistics in both distributions was proposed in Lindsey and Mersch (1992) [11]. Thus, both distributions are fitted using a Poisson regression model having sufficient statistics from both distributions as explanatory variables with the frequencies being the mean dependent variables. For the Double binomial model (DBM), the sufficient statistics are $y \log(y)$ and $(n-y)\log(n-y)$. Similar sufficient statistics for the multiplicative binomial model (MBM) are y and $y^{(n-y)}$ with the offset being $z = \log\left(\frac{n}{y}\right)$ for both models. However, this approach is not efficient. Their approach fits the following model:

$$\log\left(\frac{\pi_n}{1-\pi_n}\right) = lr = \beta_0 + n\beta_1 \tag{1}$$

Girls	n	Boys												
		0	1	2	3	4	5	6	7	8	9	10	11	12
F01	1	114609	108719	-	-	-	-	-	-	-	-	-	-	-
F02	2	47819	89213	42860	-	-	-	-	-	-	-	-	-	-
F03	3	20540	57179	53789	17395	-	-	-	-	-	-	-	-	-
F04	4	8628	31611	44793	28101	7004	-	-	-	-	-	-	-	-
F05	5	3666	16340	30175	28630	13740	2839	-	-	-	-	-	-	-
F06	6	1579	7908	17332	22221	15700	6233	1096	-	-	-	-	-	-
F07	7	631	3725	9547	14479	13972	8171	2719	436	-	-	-	-	-
F08	8	264	1655	4948	8498	10263	7603	3951	1152	161	-	-	-	-
F09	9	90	713	2418	4757	6436	5917	3895	1776	432	66	-	-	-
F10	10	30	287	1027	2309	3470	3878	3072	1783	722	151	30	-	-
F11	11	24	93	492	1077	1801	2310	2161	1540	837	275	72	8	-
F12	12	7	45	181	478	829	1112	1343	1033	670	286	104	24	3

Table 1: Distribution of Boys in family sizes 2 to 12

resulting in, $\pi_n = 1/(1 + \exp(-lr))$. Here, we see that the logit of the probability of “success” is a linear function of family size n . This approach is very similar to those employed in Lefkopoulou *et al.* (1989), where,

$$\text{logit}(\pi_n) = \beta_0 + \beta X \tag{2}$$

The model in (1) is employed to model the probability of success for each of the distributions considered in this study, the exception being the Com-Poisson binomial model (CPM) described later in this paper [12].

Further, the dispersion parameters for each of the models are expressed as $\exp(a_0 + na_1), n = 2, 3, \dots, 12$. The exception here being the correlated binomial (CBM) and the Com-Poisson binomial, because the dispersion parameters θ_i and ν_i in both models can take positive and negative values.

The formulation of the success probability model and the dispersion parameter models are as formulated in Lindsey and Altham (1998) [1]. However, because of the joint sufficient statistics’ approach employed in their paper, each model so formulated is based on thirteen parameters. In this study, the same models will be fitted, but with only four parameters, giving us more degrees of freedom for inference. Our models generate perfectly, the results in Lindsey and Altham (1998), even though we have employed only four parameters for each distribution [1]. We also examine the following variations of these models, for $n=2, 3, \dots, 12$ family sizes.

(a) The model that assumes that for all n , we have constant probability of ‘success’ and constant dispersion parameter (ϕ is that for the double binomial for instance). That is,

$$\text{logit}(\pi_n) = \beta_0, \text{ and } \phi_n = a_0$$

(b) The model that assumes in this case (Case II) that:

$$\text{logit}(\pi_n) = \beta_0, \text{ and } \phi_n = a_0 + na_1$$

(c) In case III, we assume that:

$$\text{logit}(\pi_n) = \beta_0 + n\beta_1, \text{ and } \phi_n = a_0$$

(d) Case IV: This is the general case, where,

$$\text{logit}(\pi_n) = \beta_0 + n\beta_1, \text{ and } \phi_n = a_0 + na_1$$

Since the data is binary, a binomial model applied to the entire data (excluding family size of one) gives an estimated probability of males 'success' to be 0.5150. The dispersion parameter estimate from this model ($X^2/d.f.$) is 9117.9513 which is very much greater than 1, indicating a very strong over-dispersion in the data. Thus, the binomial model would not fit these data and we thus consider in the following sections, the five over-dispersed probability distributions employed in this study.

The Multiplicative Binomial Model-MBM

Lovison (1998) proposed an alternative form of the two-parameter exponential family generalization of the binomial distribution first introduced by Altham (1978) which itself was based on the original Cox (1978) representation as [6,13,14]:

$$f(y) = \frac{\binom{n}{y} \psi^y (1-\psi)^{n-y} \omega^{y(n-y)}}{\sum_{j=0}^n \binom{n}{j} \psi^j (1-\psi)^{n-j} \omega^{j(n-j)}}, y = 0, 1, 2, \dots, n \quad (3)$$

Where $0 < \psi < 1$ and $\omega > 0$. When $\omega = 1$ the distribution reduces to the binomial with $\pi = \psi$. If $\omega = 1$, $n \rightarrow \infty$, and $\psi \rightarrow 0$, then $n\psi \rightarrow \mu$ and the MBD reduces to Poisson (μ).

Elamir (2013) presented an elegant characteristics of the multiplicative binomial distribution including its four central moments. His treatment includes generation of random data from the distribution as well as the likelihood profiles and several examples.

Following Elamir (2013), the probability π of success for the Bernoulli trial, can be computed from the following expression in (4) as [7]:

$$\pi = \psi \frac{\kappa_{n-1}(\psi, \omega)}{\kappa_n(\psi, \omega)} \quad (4)$$

Where:

$$\kappa_{n-1}(\psi, \omega) = \sum_{k=0}^{n-a} \binom{n-a}{y} \psi^y (1-\psi)^{n-a-y} \omega^{(y+a)(n-a-y)} \quad (5)$$

with π defined as in (4), ψ therefore can be defined as the probability of success weighted by the intra-units association measure ω which measures the dependence among the binary responses of the n units. Thus if $\omega = 1$, then $\pi = \psi$ and we have independence among the units. However, if $\omega \neq 1$, then, $\pi \neq \psi$ and the units are not independent.

The mean and variance of the MBM are given respectively as:

$$E(Y) = n\pi_1 \quad (6a)$$

$$Var(Y) = n\pi_1 + n(n-1)\pi_2 - n^2\pi_1^2 \quad (6b)$$

Where:

$$\pi_i = \psi^i \frac{\kappa_{n-i}(\psi, \omega)}{\kappa_n(\psi, \omega)} \quad (7)$$

and the $k(\cdot)$ are as defined previously in (5).

The Beta-Binomial Distribution

The Beta-Binomial distribution Skellam (1948) has the probability distribution den-sity function [5]:

$$f(y; \alpha, \beta) = \binom{n}{y} \frac{B(\alpha + y, \beta + n - y)}{B(\alpha, \beta)}, y = 0, 1, \dots, n \tag{8}$$

Where

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

and n is the family size $n=2, 3, \dots, 12$. Following Lindsey and Altham (1998), the model in (8) can be parameterized as [1]:

$$f(y; \alpha, \beta) = \binom{n}{y} \frac{B(\pi e^\psi + y, (1 - \pi) e^\psi + n - y)}{B(\pi e^\psi, (1 - \pi) e^\psi)} \tag{9}$$

Where $\psi < \infty$ and $\rho^2 = 1/(1 + e^\psi)$.

The Com-Poisson Binomial (CPB) Model

The probability density function for the Com-Poisson Binomial distribution is given by:

$$f(y | n, p, \nu) = \frac{\binom{n}{y}^{\nu} \pi^y (1 - \pi)^{n-y}}{\sum_{k=0}^n \binom{n}{k}^{\nu} \pi^k (1 - \pi)^{n-k}}, y = 0, 1, \dots, n \tag{10}$$

With $\pi \in (0, 1)$ and $\nu \in R$. If $\nu = 1$, the model reduces to the binomial distribution and values of $\nu > 1$ indicate underdispersion, while values of $\nu < 1$ similarly indicate overdispersion with respect to the binomial distribution.

An approximation to the CPB distribution in the limit $n \rightarrow \infty$ and with $\lambda = n^{\nu} p$ is given in Shmueli *et al.* (2005) [15]. Following Borges *et al.* (2013), if we let θ_n be defined as [8]:

$$\theta_n = \frac{\pi_n}{1 - \pi_n} \tag{11}$$

and dividing both the denominator and numerator of the expression in (10) by a factor of $(1 - \pi)^m (m!)^{\nu}$, we thus have:

$$f(y | n, \theta, \nu) = \frac{\theta^y}{(y!)^{\nu}} \frac{1}{Z(\theta, \nu)}, y = 0, 1, \dots, \theta > 0, \nu \geq 0, \tag{12}$$

where the normalizing term is defined as:

$$Z = \sum_{j=0}^n \frac{\theta^j}{[j!(n - j)!]^{\nu}} \tag{13}$$

The various properties of the CPB or the Com-Poisson have been presented in various papers Borges *et al.* (2013), Shmueli *et al.* (2005), and Kadane *et al.* (2006) applied the CPB to the number of killings in rural Norway [8,15,16]. In this study, with θ defined as $\theta_n = \alpha_0 + n\alpha_1$ for the family sizes, then the estimated probabilities are obtained as $\hat{\pi}_n = \frac{\hat{\theta}_n}{1 + \hat{\theta}_n}$.

The Double Binomial Model (DBM)

In Feirer *et al.* (2013), the double binomial distribution was presented, having the pdf form [17]:

$$f(y; \pi, \phi) = \frac{\binom{n}{y} [y^y (n-y)^{n-y}]^{1-\phi} [\pi / 1 - \pi]^{y\phi}}{\sum_{j=0}^n \binom{n}{j} [j^j (n-j)^{n-j}]^{1-\phi} [\pi / 1 - \pi]^{j\phi}} \quad (14)$$

For $y = 0, 1, \dots, n$, and following Feirer *et al.* (2013), its log-likelihood is presented in (16c).

Correlated Binomial Model

A fetal death model proposed by Kupper and Haseman (1978), which is a generalization of the binomial model has the form [10]:

$$f(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y} \left\{ 1 + \frac{\theta_i}{2p^2(1-p)^2} [(y-np)^2 + y(2p-1) - np^2] \right\} \quad (15)$$

In (15), θ_i is the covariance between any two responses within the same litter. The model above allows for negative intra-litter correlation, as distinct from the other models considered that only allow for positive dispersion parameters.

Because of the bounds on θ_i for this model, it is modeled here as $\theta_i = b_0 + n_i b_p$, $i=2,3, \dots, 12$ rather than taking its exponent.

Methodology

For a single observation, the log-likelihoods for the multiplicative binomial, beta- binomial, double binomial, the Com-Poisson binomial, and the correlated binomial are displayed in expressions (16a) to (16e) respectively.

$$LL1 = \log \binom{n}{y} + y \log(\psi) + y(n-y) \log \omega - \log \left[\sum_{j=0}^n \binom{n}{j} \psi^j (1-\psi)^{n-j} \omega^{j(n-j)} \right] \quad (16a)$$

$$LL2 = \log \binom{n}{y} + \log[\Gamma(\phi\pi + y)] + \log[\Gamma(\phi(1-\pi) + n - y)] - \log[\Gamma(n + \phi)] - \log[\Gamma(\phi\pi)] - \log[\Gamma(\phi(1-\pi))] + \log[\Gamma(\phi)] \quad (16b)$$

$$LL3 = \log \binom{n}{y} + (1-\phi) \left[y \log y + (n-y) \log(n-y) + y\phi \log \left(\frac{\pi}{1-\pi} \right) \right] - \log \left[\sum_{j=0}^n \binom{n}{j} (j^j (n-j)^{n-j})^{1-\phi} \left(\frac{\pi}{1-\pi} \right)^{j\phi} \right] \quad (16c)$$

$$LL4 = y_i \log \theta_i - v [y_i!(n-y_i)!] - \log \left[\sum_{j=0}^n \frac{\theta_i}{[j!(n-1)!]^v} \right] \quad (16d)$$

$$LL5 = \log \binom{n}{y} + y \log(p) + (n-y) \log(1-p) + \log \left[1 + \frac{\theta_i}{2p^2(1-p)^2} \{(y-np)^2 + y(2p-1) - np^2\} \right] \quad (16e)$$

Maximum-likelihood estimations of the above models are carried out with PROC NLMIXED in SAS, which minimizes the function $-LL(y, \Theta)$ over the parameter space numerically. The default integral approximations of the marginal likelihood in PROC NLMIXED is the Adaptive Gaussian Quadrature as defined in Pinheiro and Bates (1995) and the optimization algorithm adopted in this study is the Newton-Raphson iterative procedure (NEWRAP). Convergence is often a major problem here and the choice of starting values is very crucial [18]. For each of the cases considered here, we are able to achieve rapid convergence for all the distributions.

Results

As discussed earlier, following Lindsey & Altham, we model the probability, πn as a function of family size n , viz $\text{logit}(\pi_n) = \beta_0 + n\beta_1$. Similarly, the corresponding dispersion parameters are also modeled as: $\log(\omega) = \alpha_0 + n\alpha_1$. Thus, for the multiplicative binomial, for instance, $\omega = \exp(lp)$, and both are functions of family size n . Further, for each model, we are estimating only four parameters where for instance, for the beta-binomial model for example, we are estimating only $\{\alpha_0, \alpha_1, \beta_0, \beta_1\}$ as compared to 13 parameters that were estimated in Lindsey and Altham paper.

In Table 2 are presented the parameter estimates and their standard errors (in parentheses) under the five models for cases I to IV [19-21]. For the full models (Case IV), we see that only four parameters are estimated for each model in contrast to Lindsey and Altham (1998) thirteen parameters [1].

Parameters	I	II	III	IV
Beta-Binomial Model				
$\hat{\alpha}_0$	4.5817 (0.0362)	5.3977 (0.1248)	4.5826 (0.0362)	5.3954 (0.1248)
$\hat{\alpha}_1$	-	-0.1100(0.0143)	-	-0.1096 (0.0143)
$\hat{\beta}_0$	0.0600(0.0011)	0.0599 (0.0011)	0.0497 (0.0026)	0.0498 (0.0026)
$\hat{\beta}_1$	-	-	0.0018 (0.0004)	0.0018 (0.0004)
-2LL	4180.8	4118.6	4162.4	4100.8
X ²	194.0652	126.3841	174.8058	108.4632
G ²	187.6311	125.5495	169.2988	107.7504
Multiplicative Binomial Model				
$\hat{\alpha}_0$	-0.0196 (0.0007)	-0.0055(0.0020)	-0.0195(0.0007)	-0.0058 (0.0020)
$\hat{\alpha}_1$	-	-0.0019 (0.0003)	-	-0.0019 (0.0003)
$\hat{\beta}_0$	0.0573(0.0011)	0.0573(0.0011)	-0.0507(0.0025)	0.0522 (0.0025)
$\hat{\beta}_1$	-	-	0.0011(0.0004)	0.0009 (0.0004)
-2LL	4159.4	4103.2	4151.2	4098.4
X ²	171.3320	110.7833	162.8512	106.0609
G ²	166.2496	110.1267	158.1647	105.3662
Double-Binomial Model				
$\hat{\alpha}_0$	-0.0318(0.0014)	0.0256 (0.0032)	-0.0317(0.0014)	0.0256 (0.0032)
$\hat{\alpha}_1$	-	-0.0123 (0.0006)	-	-0.0123 (0.0006)
$\hat{\beta}_0$	0.0601 (0.0011)	0.0598(0.001)	0.0499 (0.0026)	0.0498 (0.0026)
$\hat{\beta}_1$	-	-	0.0018 (0.0004)	0.0018 (0.0004)
-2LL	4525.0	4143.4	4506.8	4125.6
X ²	562.2668	152.0558	541.8158	133.5381
G ²	531.9262	150.2028	513.7150	132.4454

Parameters	I	II	III	IV
Com-Poisson Binomial Model-CPM				
$\hat{\alpha}_0$	0.9477 (0.0021)	1.0256(0.0050)	0.9479 (0.0021)	1.0247 (0.0051)
$\hat{\alpha}_1$	-	-0.0141 (0.0008)	-	-0.0140 (0.0006)
$\hat{\beta}_0$	1.0595 (0.0011)	1.0589 (0.0011)	1.0499 (0.0027)	1.0533 (0.0027)
$\hat{\beta}_1$	-	-	0.0017 (0.0004)	0.0010 (0.0004)
-2LL	4396.8	4110.8	4381.6	4105.6
X^2	424.1706	118.6071	407.2423	113.1593
G^2	403.7245	117.7763	388.4260	112.4623
Correlated Binomial Model-CBM				
$\hat{\alpha}_0$	0.0025 (0.0001)	0.0005 (0.0003)	0.0025 (0.0001)	0.0497 (0.0026)
$\hat{\alpha}_1$	-	0.0003 (0.00004)	-	0.0018 (0.0004)
$\hat{\beta}_0$	0.0560 (0.0011)	0.0599 (0.0011)	0.0497 (0.0026)	0.0005 (0.0003)
$\hat{\beta}_1$	-	-	0.0018 (0.0004)	0.0003 (0.00004)
-2LL	4185.6	4123.8	4167.2	4105.8
X^2	200.9393	133.6860	181.5540	115.5038
G^2	192.5759	130.6794	174.1174	112.7036
$\hat{\pi}$	0.5150			
θ	0.0025			

Table 2: Parameter Estimates for the five Models

The $-2LL$ values for the beta-binomial in case IV were 2174100.8 but we have reported this simply as 4100.8. That is, we are reporting only the last four digits. In each case, the values of 217 had been truncated for brevity only. What is obvious from Table 2, is that, irrespective of the model being employed, the data is better modeled with a variable dispersion parameter, that is, cases II and IV. In other words, models that assume uniform dispersion parameter (within a given distribution here) are not satisfac-

tory for the data. The $-2LL$ and goodness-of-fit test statistics, the Pearson's $X^2 = \sum_{i=1}^{88} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i}$ and the likelihood-ratio test $G^2 = -2 \sum_{i=1}^{88} y_i \log \left(\frac{y_i}{\hat{m}_i} \right)$ all indicate that the models in cases II and IV are much better. Further, of the two parsimonious models in

cases II and IV, model IV which is equivalent to Lindsay and Altham (1998) model is most parsimonious. For instance, for the beta-binomial model, the difference in G^2 for the two models is $125.5495 - 107.7504 = 17.7991$ on 1 d.f., (G^2 rather than X^2 is used here because of its partitioning property over X^2) which clearly indicates that β_1 is important in the model, given that α_0 , α_1 and β_0 are already in the model. Consequently, we would focus our attention on models in case IV for all the distributions.

For these five models, the best three models in case IV that best fit our data are in the order (a) Multiplicative binomial, (b) The beta-binomial and (c) the Com-Poisson binomial distributions. However, we present in Table 3, the estimated probabilities for each family size under each of these models.

In Table 3 are presented the estimated probabilities from each of these models. We present the computation of the first probabilities for family size $n=6$ as an example. For the beta-binomial, this equal $\hat{\pi}_6 = 1 / [1 + \exp(-0.0498 - 0.0018 * 6)] = 0.5151$.

Similarly, under the multiplicative binomial model, $\hat{\psi}_6 = 1 / [1 + \exp(-0.05219 - (0.000864 * 6))] = 0.51434$.

Family Size	$\hat{\pi}_n$				
	BBD	MBD	DDM	CPB	CBM
2	0.5133	0.5135	0.5133	0.5134	0.5133
3	0.5138	0.5139	0.5138	0.5137	0.5138
4	0.5142	0.5142	0.5142	0.5139	0.5142
5	0.5147	0.5146	0.5147	0.5141	0.5147
6	0.5151	0.5150	0.5151	0.5143	0.5151
7	0.5156	0.5154	0.5155	0.5146	0.5156
8	0.5160	0.5159	0.5160	0.5148	0.5160
9	0.5165	0.5165	0.5164	0.5150	0.5165
10	0.5169	0.5171	0.5169	0.5153	0.5169
11	0.5173	0.5177	0.5173	0.5155	0.5174
12	0.5178	0.5184	0.5177	0.5157	0.5178

Table 3: Estimated success probabilities under the five models

However, this is not the required probability. The required probability, π_6^1 can be computed using (7) as:

$$\pi_6^1 = \hat{\psi} \left(\frac{K_n}{K_{(n-1)}} \right) = 0.51434 \left(\frac{0.88240}{0.88132} \right) = 0.51497$$

We present in Table 4, the calculated values necessary to compute π_1, π_2 and the variance under the multiplicative binomial model. The table also includes computed estimated variances as well as the dispersion parameter ω . We shall refer to this table again later. Clearly, the beta-binomial, the double binomial and the correlated give estimated global family size probabilities that are very similar. Of the three models here, the beta-binomial fits best and we see that the estimated probabilities under the BB agree with those presented in Lindsey and Altham (1998), even though we have employed only four parameters to fit the models [1]. The Com-Poisson binomial seems to be underestimating the global probabilities slightly, but seems to fit the model better than the DBM and the CBM.

Similarly, the estimated π_n^1 obtained (first probability) under the multiplicative binomial model are very consistent with those from the BB. The questions posed in Linsey and Altham regarding the good fit of the multiplicative binomial model, namely:

- (a) Does it (i.e. the multiplicative model) have some biological significance similar to the idea for the beta-binomial model of the Bernoulli probability varying across the population according to a beta distribution?
- (b) Can more be said about the parameters than the empirical statement that the probability and dispersion both increase with family size?

n	K_n	K_{n-1}	K_{n-2}	$\hat{\psi}$	$\hat{\pi}_1$	$\hat{\pi}_2$	var	$\hat{\omega}_n$
2	0.99526	0.99538	1.00000	0.51348	0.51354	0.26492	1.0295	0.99050
3	0.98313	0.98343	0.98905	0.51369	0.51385	0.26547	2.3422	0.98868
4	0.96124	0.96177	0.96838	0.51391	0.51419	0.26606	4.1919	0.98686
5	0.92770	0.92849	0.93603	0.51412	0.51456	0.26670	6.5829	0.98503
6	0.88132	0.88240	0.89076	0.51434	0.51497	0.26738	9.5200	0.98321
7	0.82189	0.82327	0.83227	0.51456	0.51542	0.26811	13.0090	0.98139
8	0.75033	0.75198	0.76141	0.51477	0.51591	0.26891	17.0567	0.97956
9	0.66872	0.67062	0.68020	0.51499	0.51645	0.26977	21.6707	0.97774
10	0.58022	0.58229	0.59173	0.51520	0.51705	0.27070	26.8602	0.97592
11	0.48876	0.49093	0.49992	0.51542	0.51771	0.27172	32.6362	0.97409
12	0.39862	0.40079	0.40907	0.51564	0.51844	0.27285	39.0116	0.97227

Table 4: Multiplicative Model necessary Values

Based on our ability to now be able to compute the central moments of the multiplicative Binomial model (Elamir (2013)), the ψ in our model specification in (3) is not necessarily equivalent to the desired π_n , our ability to now be able to compute the $\pi_n^i, i=1,2,3,4$ and in particular, the case when $i = 1$ which corresponds to our Bernoulli or BB parameter π_n from expressions in (7) answers the question posed in (a) [7]. Thus, π^1 is considered to be the 'probability of success while ψ is the probability of success weighted by the intra-association measure ω given the dependence among the binary responses of the units.

As to the second question, the results of $\hat{\omega}_n$ in Table 5, indicate that $\hat{\omega}_n < 1, \forall i$ and decreases as the family size increases. Thus, while $\hat{\omega}_2 = 0.99050$, that for $\hat{\omega}_{12} = 0.97227$. These indicate decreasing variability as family size increases. On the other hand, for the beta-binomial, both probability and dispersion parameter increase with family size. For values of n small, estimates of ω are very close to 1, indicating closeness to being independent observations and can be appropriately modeled with the binomial distribution.

Also, in Table 5, the computed dispersion parameters for the double binomial, ϕ and the Com-Poisson binomial, v , indicate decreasing dispersion parameters as the family size increases. For both the beta and correlated binomial, the dispersion parameters however, increase with family size.

Family Size n	Dispersion Estimated Parameters				
	BB	MBD	DDM	CPB	CBM
	$\hat{\rho}^2$	$\hat{\omega}_i$	$\hat{\phi}_i$	\hat{v}_i	$\hat{\theta}_i$
2	0.0056	0.9905	1.0011	0.9968	0.0011
3	0.0063	0.9887	0.9889	0.9828	0.0014
4	0.0070	0.9869	0.9769	0.9688	0.0016
5	0.0078	0.9850	0.9650	0.9549	0.0019
6	0.0087	0.9832	0.9532	0.9409	0.0022
7	0.0097	0.9814	0.9416	0.9270	0.0025
8	0.0108	0.9796	0.9301	0.9130	0.0028
9	0.0120	0.9777	0.9188	0.8990	0.0030
10	0.0134	0.9759	0.9076	0.8851	0.0033
11	0.0149	0.9741	0.8966	0.8711	0.0036
12	0.0166	0.9723	0.8856	0.8572	0.0039

Table 5: Dispersion Estimated Parameters

Goodness-of-fit Statistics

We present in this section results of goodness-of-fit tests for all the models considered in this study. From the global model for each distribution, and a given estimated set of parameters, the probabilities of $Y=0,1,\dots,n$ and corresponding expected values are computed from the log-likelihood equations. For example, for the multiplicative model, this is, for a given family size n , the LL1 is as presented earlier in (16a). Hence, the corresponding probabilities are computed for $Y=0,1,\dots,n$ as $\hat{p}_y = P(Y=y) = \exp(LL1)$ and the corresponding expected values are $\hat{m}_y = N\hat{p}_y$, where N for the family of size 12 for instance is 6115. We give in Table 6, the computed probabilities, the expected values, and contributions to the likelihood-ratio test statistic G^2 and Pearson's X^2 for the multiplicative binomial for $n=2, 3, 4$ and $n=12$. Here, G^2 and X^2 are defined as

$$G^2 = 2 \sum_{i=1}^{88} y_i \log \left(\frac{y_i}{\hat{m}_i} \right); \quad X^2 = \sum_{i=1}^{88} \frac{(y_i - \hat{m}_i)^2}{\hat{m}_i}$$

where $\hat{m}_i = N\hat{p}_i$ are the expected values as defined above for each family size n .

From Table 6, our computations for the G^2 and X^2 exclude the case when $n=1$, that is, family size of 1. Thus, we have the computations over 88 observations only. Note that for each n , the cumulative sum of probabilities (designated here by $\sum p_y$) sums to 1 for each family size n . We are re-assured that for a given n these probabilities sum to one. We have presented the results for $n=2, 3, 4$ and $n=12$ only. The sample sizes for $n=2, 3, 4, 12$ are respectively 179,892; 148,903; 120,137; and 6,115. The expected values under each N are provided under the column labeled \hat{m}_y . This approach is employed for all the models considered in this study and in Table 7 are the resulting GOFs for all the five distributions under case IV.

Table 7 gives the GOF for the five models (in comparison to the binomial model) and results here demonstrate that while the multiplicative binomial model is best, it is however not much better than the beta-binomial model. However, since the multiplicative model is not much difficult to model than the beta-binomial, it would be better to actually employ it. Our results in other studies regarding the two models suggest that the MBD may be a better choice. Clearly, both the BBD and the MBD show considerable improvements in model fits than the binomial model.

n	y	\hat{p}_y	$\sum \hat{p}_y$	\hat{m}_y	$\sum X^2$	$\sum G^2$	$\hat{\omega}_n$
2	0	0.2378	0.2378	42783.3261	0.1374	153.4851	0.9905
2	1	0.4973	0.7351	89453.4146	0.7835	-326.6974	0.9905
2	2	0.2649	1.0000	47655.2593	1.3462	1.3460	0.9905
3	0	0.1170	0.1170	17418.3948	1.3776	-45.4121	0.9887
3	1	0.3624	0.4793	53957.9978	1.9069	-382.8778	0.9887
3	2	0.3828	0.8621	56996.6365	2.4904	-17.5679	0.9887
3	3	0.1379	1.0000	20529.9709	2.4953	2.4951	0.9887
4	0	0.0581	0.0581	6977.4982	2.5959	55.5992	0.9869
4	1	0.2361	0.2941	28359.9063	4.9596	-459.8425	0.9869
4	2	0.3694	0.6636	44383.7008	8.7340	362.5188	0.9869
4	3	0.2639	0.9274	31698.7478	8.9769	187.2663	0.9869
4	4	0.0726	1.0000	8717.1469	9.8886	9.8873	0.9869
.
.
.
12	0	0.0004	0.0004	2.5561	90.9736	91.2014	0.9723
12	1	0.0039	0.0043	23.9718	90.9736	91.2578	0.9723
12	2	0.0178	0.0222	108.9954	91.2026	81.4994	0.9723
12	3	0.0520	0.0741	317.7147	94.3684	21.3468	0.9723
12	4	0.1081	0.1823	661.2655	94.4838	38.9307	0.9723
12	5	0.1693	0.3516	1035.2798	94.4888	34.3761	0.9723
12	6	0.2044	0.5560	1250.1793	101.3803	226.7445	0.9723
12	7	0.1919	0.7479	1173.2732	104.5803	107.4552	0.9723
12	8	0.1389	0.8868	849.2952	105.0653	67.3537	0.9723
12	9	0.0756	0.9624	462.4465	105.5884	98.9780	0.9723
12	10	0.0294	0.9918	179.7935	105.5965	101.3991	0.9723
12	11	0.0073	0.9991	44.8135	105.5973	101.7729	0.9723
12	12	0.0009	1.0000	5.4154	106.0609	105.3662	0.9723
12	12	0.0009	1.0000	5.4154	106.0609	105.3662	0.9723

Table 6: Bayo Lawal

Model	d.f	G ²	X ²
Binomial	85	3551.857	3807.590
Beta-B	83	107.7504	108.4632
MBM	83	105.3662	106.0609
DBM	83	132.4454	133.5381
Com-PB	83	112.4623	113.1593
CBM	83	112.7036	115.5038

Table 7: Models' GOF test statistics

Conclusion

Our results here agree with those in Lindsey and Altham (1998) that the "multi- plicative model fitted somewhat better than did the beta-binomial" [1]. Our procedure here estimates fewer parameters than those employed in Lindsey and Altham (1998) and the optimization techniques open to us with SAS PROC NLMIXED, makes the use of this procedure much easier to use [1]. The implementation and use of more models in this study gives a broader perspective on models used for data having binary outcomes, especially as are being applied to the Human Sex Ratio Data. The SAS programs employed for this analysis are available from the author.

References

1. Lindsey JK, Altham PME (1998) Analysis of the human sex ratio by using overdispersion models. *Appl Statist* 47: 149-57.
2. Geissler A (1889) Beitrage zur Frage des Geschlechtsverhältnisses der Geborenen. *Z Königl Sächs Statist Bur* 35: 1-24.
3. Edwards AWF (1958) An analysis of Geissler's data on human sex ratio. *Ann Hum Genet* 23: 6-15.
4. Sokal RR, Rohlf FJ (1969) *Biometry: the Principles and Practice of Statistics in Biological Research*. San Francisco: Freeman.
5. Skellam JG (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as a variable between sets of trials. *J Roy Stat Soc B* 10: 257-61.
6. Altham PM (1978) Two generalizations of the binomial distribution. *Appl Statist* 27: 162-7.
7. Elamir EAH (2013) Multiplicative-binomial distribution: Some results and characterization, inference and random data generation. *J Stat Theory Appl* 12: 92-105.
8. Borges P, Rodrigues J, Balakrishnan N (2013) A Com-Poisson type generalizations of the Binomial distribution and its properties and applications. *Stat probab lett* 87: 158-66.
9. Lindsey JK, Altham, PK (1995) Varieties of over- and underdispersion models for binary data.
10. Kupper LL, Haseman JK (1978) The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* 34: 69-76.
11. Lindsey JK, Mersch G (1992) Fitting and comparing probability distributions with log linear models. *Comput Stat Data Anal* 13: 373-84.
12. Lefkopoulou M, Moore D, Ryan L (1989) The Analysis of multiple correlated binary outcomes: Applications to Rodent Teratology Experiment. *J Am Stat Ass* 84: 810-5.
13. Lovison G (1998) An alternative representation of Altham's multiplicative-binomial distribution. *Stat Probab lett* 36: 415-20.
14. Cox DR (1978) The analysis of multivariate binary data. *Appl Stat* 21: 113-20.
15. Shmueli G, Minka T, Borle J, Boatwright P (2005) A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *J Roy Stat Soc Series C* 54: 127-42.
16. Kadane J, Shmueli G, Minka G, Borle T, Boatwright P (2006) Conjugate analysis of the Conway Maxwell Poisson distribution. *Bayesian Anal* 1: 363-74.
17. Feirer V, Friedl H, Hirn U (2013) Modeling over- and underdispersed frequencies of successful ink transmissions onto paper. *J Appl Stat* 40: 626-43.
18. Pinheiro JC, Bates DM (1995) Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *J Comput Graph Stat* 4: 12-35.
19. Efron B (1986) Double exponential families and their use in generalized linear regression. *J Amer Stat Assoc* 81: 709-21.
20. Fisher RA (1958) *Statistical Methods for Research Workers*. (13th Edn) Edinburgh: Oliver and Boyd.
21. Williams DA (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 31: 945-51.

Submit your next manuscript to Annex Publishers and benefit from:

- ▶ Easy online submission process
- ▶ Rapid peer review process
- ▶ Online article availability soon after acceptance for Publication
- ▶ Open access: articles available free online
- ▶ More accessibility of the articles to the readers/researchers within the field
- ▶ Better discount on subsequent article submission

Submit your manuscript at

<http://www.annexpublishers.com/paper-submission.php>