

Generation of Over-Dispersed and Under-Dispersed Binomial Variates

Hongshik Ahn and James J. Chen

Division of Biometry and Risk Assessment
National Center for Toxicological Research
Food and Drug Administration
Jefferson, Arkansas 72079

Abstract

This paper proposes an algorithm for generating over-dispersed and under-dispersed binomial variates with specified mean and variance. The over-dispersed/under-dispersed distributions were derived from correlated binary variables with an underlying continuous multivariate distribution. Different multivariate distributions or different correlation matrices resulted in different over-dispersed (or under-dispersed) distributions. The over-dispersed binomial distributions generated from three different correlation matrices of a multivariate normal were compared with the beta-binomial distribution for various mean and over-dispersion parameters by quantile-quantile (Q-Q) plots. The two distributions appear to be similar. The under-dispersed binomial distribution was simulated to model an example data set which exhibits under-dispersed binomial variation.

KEY WORDS: Beta-binomial; Correlated binary; Intra-cluster correlation; Monte Carlo; Teratology.

1. INTRODUCTION

There has been a great deal of interest in quasi-likelihood and generalized estimating equation methods for analysis of data with over-dispersion (Williams 1982; Lefkopoulou, Moore and Ryan 1989) and under-dispersion (Brooks, James and Gray 1991). Data with over-dispersion or under-dispersion are common in toxicology, epidemiology, and many other biological research areas. In

standard teratology bioassays, the fetus responses (presence or absence of a malformation type) in a litter commonly exhibit extra-binomial variations in the dosed group animals. On the other hand, Brooks *et al.* (1991) showed that the number of males (or females) per pig litter exhibits significant sub-binomial dispersion. The classical parametric (prior-posterior) modeling of over-dispersion data is by mixing distributions. The beta-binomial model assumes that the responses within a litter (cluster) follow an independent Bernoulli process, and the Bernoulli parameter itself is a random variable varies from litter to litter according to a beta distribution. A problem with this approach is that it cannot model the under-dispersion data, since the variance for the distribution of the Bernoulli parameter must be non-negative. Alternatively, an over-dispersion or under-dispersion can be modeled by an implicit intra-cluster correlation structure among the Bernoulli responses within each cluster. The intra-cluster correlations are positive for the fetus response in a teratology study to represent an over-dispersion. The intra-cluster correlations are negative for the response being a male or a female, such as in the data of Brooks *et al.* (1991), to represent an under-dispersion.

The main purpose of this paper is to present an algorithm for generating over-dispersed and under-dispersed binomial distributions with specified mean and variance values in the framework of quasi-likelihood (or generalized estimating equation) method. Generation of different over-dispersed/under-dispersed binomial variates for the specified mean and variance using different correlation structures or different underlying multivariate distributions is discussed. This algorithm will provide an alternative to the beta-binomial for studying properties of estimation or test statistics of the quasi-likelihood and/or parametric likelihood approaches. For example, an over-dispersed distribution can be simulated to study the robustness of the maximum likelihood estimate or likelihood ratio test of the beta-binomial. Moreover, it also provides a tool to generate under-dispersion samples to describe data with under-dispersion.

2. OVER-DISPERSED AND UNDER-DISPERSED BINOMIAL DISTRIBUTIONS

Let Y_{i1}, \dots, Y_{in_i} be binary variables from a cluster of size $n_i, i = 1, \dots, m$. Assume that $E(Y_{ij}) = \mu$, $\text{Var}(Y_{ij}) = \mu(1 - \mu)$ and $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{ijk}$, for $j \neq k$. Let R_i denote the correlation matrix, then the (jk) th element of R_i is ρ_{ijk} . Let $X_i = Y_{i1} + \dots + Y_{in_i}$, then

$$E(X_i) = n_i\mu \text{ and } \text{Var}(X_i) = n_i\mu(1 - \mu)(1 + \psi_i), \quad (1)$$

where $\psi_i = n_i^{-1} \sum \sum_{j \neq k} \rho_{ijk}$. The parameters μ and ψ_i are the mean and intra-cluster correlation (over-dispersion or under-dispersion) parameter of the distribution, respectively. If $\psi_i > 0$, X_i is an over-dispersed binomial, if $\psi_i < 0$, X_i is an under-dispersed binomial, and if $\psi_i = 0$, X_i reduces to a binomial.

We now propose a procedure for generating over-dispersed and under-dispersed binomial distributions. This procedure is based on the method of generating multivariate binary data proposed by Emrich and Piedmonte (1991). For given n_i , μ , and ρ_{ijk} , the first step of the procedure is to solve the equation

$$F[z(\mu), z(\mu), \delta_{ijk}] = \mu(1 - \mu)\rho_{ijk} + \mu^2 \quad (2)$$

for δ_{ijk} , where $F(x_1, x_2, \delta_{ijk})$ is the cumulative distribution of a continuous bivariate random variable with mean 0 and correlation coefficient δ_{ijk} , and $z(\mu)$ denotes the μ th quantile of the distribution. Emrich and Piedmonte (1991) used the cumulative distribution function Φ of a standard bivariate normal random variable for F . When ρ_{ijk} is equal to zero, the bivariate normal distribution becomes a joint distribution of two independent standard normals and δ_{ijk} becomes zero. The bisection method can be applied to solve equation (2). The second step is to generate an n_i -dimensional random variable of the chosen distribution $Z_i = (Z_{i1}, \dots, Z_{in_i})'$ with mean 0 and

correlation matrix Σ_i , where $[\Sigma_i]_{jk} = \delta_{ijk}$ for $j \neq k$. The final step is that for $j = 1, \dots, n_i$, one sets $Y_{ij} = 1$ if $Z_{ij} \leq z(\mu)$ and sets $Y_{ij} = 0$ otherwise. The sum $X_i = \sum_{j=1}^{n_i} Y_{ij}$ has the desired distribution, since $E[Y_{ij}] = P\{Y_{ij} = 1\} = P\{Z_{ij} \leq z(\mu)\} = \mu$ and

$$\begin{aligned} \text{Corr}(Y_{ij}, Y_{ik}) &= [P\{Y_{ij} = 1, Y_{ik} = 1\} - \mu^2] / \{\mu(1 - \mu)\} \\ &= [P\{Z_{ij} \leq z(\mu), Z_{ik} \leq z(\mu)\} - \mu^2] / \{\mu(1 - \mu)\} \\ &= [F\{z(\mu), z(\mu), \delta_{ijk}\} - \mu^2] / \{\mu(1 - \mu)\} = \rho_{ijk}. \end{aligned}$$

For fixed n_i , the distribution X_i has the mean and variance given in (1); conversely, for given mean and variance in (1), any continuous n_i -variate distribution with an intra-cluster correlation structure R_i such that $\psi_i = n_i^{-1} \sum \sum_{j \neq k} \rho_{ijk}$ can be used to generate X_i . The n_i -variate distributions can be multivariate normal, Student-t, log-normal, etc, and the correlation structures can be equicorrelated, tridiagonal, and auto-correlated, etc. A simple multivariate distribution is the normal with an equal correlation. This is the correlated-probit model considered by Ochi and Prentice (1984).

The higher moments of X_i can be obtained by using the relationship $E(Y_1 \cdots Y_r) = P\{Z_1 \leq z(\mu), \dots, Z_r \leq z(\mu)\}$. For example, the third central moment is

$$\begin{aligned} E[(X_i - n_i \mu)^3] &= \sum_{j=1}^{n_i} E[(Y_{ij} - \mu)^3] + 3 \sum_{j \neq k} E[(Y_{ij} - \mu)^2 (Y_{ik} - \mu)] \\ &\quad + \sum_{j \neq k, j \neq l, k \neq l} E[(Y_{ij} - \mu)(Y_{ik} - \mu)(Y_{il} - \mu)]. \end{aligned}$$

Here, $E[(Y_{ij} - \mu)^3] = \mu(1 - \mu)(1 - 2\mu)$, $E[(Y_{ij} - \mu)^2 (Y_{ik} - \mu)] = \mu(1 - \mu)(1 - 2\mu)\rho_{ijk}$ and $E[(Y_{ij} - \mu)(Y_{ik} - \mu)(Y_{il} - \mu)] = F(z(\mu), z(\mu), z(\mu); \Sigma_{ijkl}) - \mu^2\{(1 - \mu)(\rho_{ijk} + \rho_{ijl} + \rho_{ikl}) + \mu\}$, where $F(x_1, x_2, x_3; \Sigma_{ijkl})$ is a standard trivariate cumulative distribution function, and the (1, 2)th element

of Σ_{ijkl} is δ_{ijk} , the (1, 3)th element is δ_{ijl} and the (2, 3)th element is δ_{ikl} . Hence,

$$\begin{aligned}
\mathbb{E}[(X_i - n_i\mu)^3] &= \mu(1 - \mu)(1 - 2\mu)[n_i + 3 \sum_{j \neq k} \sum \rho_{ijk}] + \sum_{j \neq k, j \neq l, k \neq l} \sum \sum F(z(\mu), z(\mu), z(\mu); \Sigma_{ijkl}) \\
&\quad - (n_i - 2)\mu^2[(1 - \mu)(\sum_{j \neq k} \sum \rho_{ijk} + \sum_{j \neq l} \sum \rho_{ijl} + \sum_{k \neq l} \sum \rho_{ikl}) + n_i(n_i - 1)\mu] \\
&= n_i\mu(1 - \mu)(1 - 2\mu)[1 + 3\psi_i] + \sum_{j \neq k, j \neq l, k \neq l} \sum \sum F(z(\mu), z(\mu), z(\mu); \Sigma_{ijkl}) \\
&\quad - n_i(n_i - 2)\mu^2[3(1 - \mu)\psi_i + (n_i - 1)\mu]. \tag{3}
\end{aligned}$$

It can be seen that a different distribution F or a different correlation structure R_i (Σ_i) will give a different third moment, in general (More discussions about the distribution are given later).

A common approach to derive an over-dispersed binomial distribution is to suppose that X_i is a binomial(n_i, p_i) random variable, where p_i itself is a random variable with expectation μ . Then the mean and variance of X_i are $\mathbb{E}(X_i) = n_i\mu$ and $\text{Var}(X_i) = n_i\mu(1 - \mu) + n_i(n_i - 1)\text{Var}(p_i)$. The variance $\text{Var}(p_i)$ represents the over-dispersion component. If the random variable p_i is beta(α_i, β_i), then $\text{Var}(p_i) = \{1/(\alpha_i + \beta_i + 1)\}\mu(1 - \mu)$; and X_i becomes a beta-binomial. If the random variable p_i has a degenerate distribution with probability 1 at a single point (or $\alpha_i \rightarrow \infty$ and $\beta_i \rightarrow \infty$), then $\text{Var}(p_i) = 0$; and X_i becomes the binomial. The mean and variance of the beta-binomial X_i are

$$\mathbb{E}(X_i) = n_i\mu \text{ and } \text{Var}(X_i) = n_i\mu(1 - \mu)\{1 + (n_i - 1)\rho_i\}, \tag{4}$$

where $\mu = \alpha_i/(\alpha_i + \beta_i)$ and $\rho_i = 1/(\alpha_i + \beta_i + 1)$. Let $\psi_i = (n_i - 1)\rho_i$, then the variance of the beta-binomial distribution can be represented by an equal correlation model with $[R_i]_{jk} = \rho_i$ for $j \neq k$. The third central moment of the beta-binomial is

$$\mathbb{E}[(X_i - n_i\mu)^3] = n_i\mu[1 + \frac{\{(1 - \mu)\psi_i + (n_i - 1)\mu\}\{(1 - 2\mu)\psi_i + (n_i - 1)(2\mu - 3)\}}{(n_i - 1)\{\psi_i + (n_i - 1)\}}]. \tag{5}$$

The third moment of the beta-binomial is different from that of the over-dispersed binomial distribution in (3). For example, if $n_i = 2$, (3) becomes $2\mu(1 - \mu)(1 - 2\mu)(1 + 3\psi_i)$ and (5) becomes $2\mu(1 - \mu)(1 - 2\mu)(1 - \psi_i)^2/(1 + \psi_i)$, where $\psi_i = \rho_i (= \rho_{i12})$.

3. SIMULATIONS

A Monte Carlo simulation was conducted to compare the beta-binomial distribution with over-dispersed binomial distributions generated from the three commonly used correlation models; equal correlation (exchangeable model), tridiagonal (one-dependent model), and auto-correlation (first-order autoregressive model) matrices. The beta-binomial distribution was used since it has been used as the standard procedure to simulate over-dispersion samples. The Q-Q plots of the two distributions were examined for various values of μ and ρ . Five thousand samples were generated in each plot.

The under-dispersed binomial samples were simulated to model the number of males in a pig litter based on the data presented by Brooks *et al.* (1991). The actual data were compared with the simulated data generated using the tridiagonal and equal correlation structures.

In the simulation, the multivariate normal distribution was chosen for the n_i -dimensional continuous multivariate distribution F , since it is the most commonly available cumulative distribution function. If the correlations are positive and equal, the following simple method based on Stuart (1958) is recommended. Let U_{i0}, \dots, U_{in_i} be $n_i + 1$ independent standard normal random variables and define $Z_{ij} = (U_{ij} - aU_{i0})/(a^2 + 1)$, where $a = \sqrt{\delta/(1 - \delta)}$, then $Z_i = (Z_{i1}, \dots, Z_{in_i})'$ is an n_i -dimensional multivariate normal with mean 0 and correlation matrix Σ_i . Thus, this method only requires generating $n_i + 1$ independent univariate normal random numbers. For this method, the correlation δ should be non-negative since $\delta = a^2/(a^2 + 1)$.

3.1 Over-dispersed binomial distribution

First, the beta-binomial was compared with the over-dispersed binomial using the equal correlation structure, $[R_i]_{jk} = \rho$ for $j \neq k$. For each sample, an n_i was randomly generated using the relative frequency distribution from actual developmental toxicity experimental data given by Haseman and Hogan (1975). The beta-binomial random sample was generated by a two-stage procedure. For given μ and ρ , a probability p_i was generated from a beta distribution with the parameters $\alpha_i = \mu(1 - \rho)/\rho$ and $\beta_i = (1 - \mu)(1 - \rho)/\rho$ from equation (4). The number of successes (or failures) X_i was generated from the binomial distribution with the parameters n_i and p_i . The IMSL subroutines were used to generate normal, bivariate normal, binomial, and beta random numbers. However, the beta random numbers from the IMSL subroutine were not acceptable when ρ is close to 1. Therefore, S (Becker *et al.* 1988) was used to generate beta-binomial random numbers for $\rho \geq .9$. Figure 1 shows the Q-Q plots of the over-dispersed binomial distribution and the beta-binomial distribution for selected values of μ and ρ . The two distributions look alike when the correlation ρ is not very large, especially when the value of μ is .5. See Figure 1 (a), (b), (e) and (f). When ρ is close to 1, most of the points are either 0 or 1, and there are not many points between 0 and 1 due to high intra-cluster correlation. The small differences in frequency at the extremes, 0 and 1, greatly affect the shape of the Q-Q plot. Therefore, the two distributions appeared to be a little different in Figure 1 (c) and (d).

Next, the beta-binomial was compared with three over-dispersed binomials generated from the equal correlation, tridiagonal, and autocorrelation structures. In this comparison, the n_i was chosen randomly from $\{5, 6, 7, 8, 9, 10\}$. The correlation for the autocorrelated matrix was given as

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{ijk} = (.7)^{|k-j|}.$$

Note that the over-dispersion parameter was $\psi_i = \sum \sum_{j \neq k} \rho_{ijk}/n_i$. The correlation in the equal correlation matrix was $\rho_i = \psi_i/(n_i - 1)$, and the nonzero correlation in the tridiagonal matrix was $\rho_i = n_i \psi_i / \{2(n_i - 1)\}$. The parameters of the beta-binomial distribution were $\alpha_i = \mu(\kappa - 1)$ and $\beta_i = (1 - \mu)(\kappa - 1)$, where $\kappa = (n_i - 1)/\psi_i$ by equating (1) and (4).

Figure 2 shows the Q-Q plots of the three over-dispersed binomial versus the beta-binomial distributions. Figure 2 (a), (c) and (e) were generated with $\mu = .5$, and (b), (d) and (f) were generated with $\mu = .1$. The plots show that the three over-dispersed binomial and beta-binomial seem to be close when the value of μ is .5. The two distributions appear to be different for $\mu = .1$. The quantiles in Figure 2 are more discrete than those in Figure 1, since n_i ranges from 1 to 20 in Figure 1, but 5 to 10 in Figure 1.

The cluster size n_i is a constant in many experiments, e.g., repeated measure studies. For n_i fixed at 10, the Q-Q plots look similar to Figure 2. These plots are not shown.

3.2 Under-dispersed binomial distribution

Brooks *et al.* (1991) presented some samples of pig litters which appeared to have “fewer unisexual and more sex-balanced litters than expected under a binomial model”, and proposed modeling the number of males with sub-binomial variations. They presented a maximum likelihood estimation of the probability of males μ , and intra-litter correlation ρ using the equal correlation and tridiagonal models (structures). An example data set which contained a total of 1838 males with litter sizes range from 5 to 11 was analyzed. The estimates of μ and ρ are $\hat{\mu} = .4861$ and $\hat{\rho} = -.0196$ for the equal correlation, and $\hat{\mu} = .4863$ and $\hat{\rho} = -.0885$, for the tridiagonal. Brooks *et al.* (1991) concluded that both the models gave significantly better fits than the binomial model.

The simulation mimics the structures of the example (pig) data with the parameter estimates from the results of Brooks *et al.* (1991). Figure 3 shows the Q-Q plots of our under-dispersed

binomial data with the pig data. Figure 3 (a) is for the equal correlation and Figure 3 (b) is for the tridiagonal correlation structure. Both plots show that the simulated data are very close to the real data; the procedure seems to identify the sub-binomial variation satisfactorily.

4. DISCUSSION

The distribution $X_i = Y_{i1} + \dots + Y_{in_i}$ is derived by summing the n_i identical but dependent binary variables, Y_{i1}, \dots, Y_{in_i} . Bahadur (1961) showed that the density function of X_i can be expressed as the product of a binomial density and a correction factor,

$$f(\mathbf{x}_i) = C(n_i, \mu) \left[1 + \sum_{j < k} E(W_{ij}W_{ik})w_{ij}w_{ik} + \sum_{j < k < l} E(W_{ij}W_{ik}W_{il})w_{ij}w_{ik}w_{il} + E(W_{ij1} \dots W_{ijn_i})w_{ij1} \dots w_{ijn_i} \right], \quad (6)$$

where $C(n_i, \mu)$ is the density of the binomial distribution with parameters n_i and μ , and $W_{ij} = (Y_{ij} - \mu) / \{\mu(1 - \mu)\}^{1/2}$. Equation (6) provides a general formula to characterize the distribution of X_i . For the X_i in this paper, the probability density can be computed from the relationship $E(Y_1 \dots Y_r) = P\{Z_1 \leq z(\mu), \dots, Z_r \leq z(\mu)\}$, for $i = 1, \dots, n_i$. This would require a numerical integration with respect to the distribution F . Kupper and Haseman (1978) derived a correlated-binomial distribution from (6) by assuming all the correlations are higher than the order of two are zero. Altham (1978) proposed two generalizations of a binomial model; incidently, the Altham additive generalized binomial model is identical to the Kupper and Haseman correlated-binomial. These parametric distributions including the beta-binomial were proposed by these authors to model the binomial data with under-dispersion or over-dispersion. All the distributions discussed here have the same first two moments which determine a two-parameter binomial generalization model, e.g. beta-binomial; but the third and higher moments of the distributions are different from

each other in general.

The quasi-likelihood approach models mean and variance without specification of the full distribution. For example, Moore (1987) assumed $\text{Var}(X_i) = n_i\mu(1-\mu)\{1 + (n_i - 1)\rho\mu^{\xi-1}(1-\mu)^{\xi-1}\}$ to model chromosome aberration data from Hiroshima. The distribution can be generated by letting $\text{Corr}(Y_{ij}, Y_{ik}) = \rho\mu^{\xi-1}(1-\mu)^{\xi-1}$. Lefkopoulou *et al.* (1989) used $\text{Var}(X_i) = n_i\mu(1-\mu)(1+\rho)$ to model the litter effect from teratological data. Their distribution can be generated by letting $\text{Corr}(Y_{ij}, Y_{ik}) = \rho/(n_i - 1)$. This paper will provide an opportunity to evaluate the quasi-likelihood methodologies which use only the mean and variance of a distribution.

Finally, to generate an under-dispersed binomial distribution, the correlation matrix Σ_i of the given multivariate distribution should be positive definite. For instance, if we use the multivariate normal with common correlation, δ_i needs to be greater than $-1/(n_i - 1)$. Figure 4 shows the relationship between δ_i and ρ_i in equation (2) for selected values of μ . It can be seen that δ_i is smaller (larger) than ρ_i for negative (positive) correlation. Therefore, ρ_i should be much greater than $-1/(n_i - 1)$ to ensure that δ_i is greater than the lower bound $-1/(n_i - 1)$.

ACKNOWLEDGEMENT

The authors wish to thank the associate editor and the referees for their valuable comments that improved the paper.

REFERENCES

- Altham, P. M. E. (1978), "Two Generalizations of the Binomial Distribution," *Applied Statistics*, **27**, 162-167.
- Bahadur, R. R. (1961), "A Representation of the Joint Distribution of Responses to n Dichotomous items," In *Studies in Item Analysis and Prediction*, H. Solomon (ed.), Stan-

ford University Press, Stanford, California.

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, California: Wadsworth.

Brooks, R. J., James, W. H., and Gray, E. (1991), "Modeling Sub-Binomial Variation in the Frequency of Sex Combinations in Litters of Pigs," *Biometrics*, **47**, 403-417.

Emrich, L. J., and Piedmonte, M. R. (1991), "A Method for Generating High-Dimensional Multivariate Binary Variates," *The American Statistician*, **45**, 302-304.

Haseman, J. K., and Hogan, M. D. (1975), "Selection of the Experimental Unit in Teratology Studies," *Teratology*, **12**, 165-172.

Kupper, L. L., and Haseman, J. K. (1978), "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics*, **34**, 69-76.

Lefkopoulou, M., Moore, D., and Ryan, L. (1989), "The Analysis of Multiple Correlated Binary Outcomes: Application to Rodent Teratology Experiments," *Journal of the American Statistical Association*, **84**, 810-815.

Moore, D. F. (1987), "Modeling the Extraneous Variance in the Presence of Extra-Binomial Variation," *Applied Statistics*, **36**, 8-14.

Ochi, Y., and Prentice, R. L. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, **71**, 531-543.

Stuart, A. (1958), "Equally Correlated Variates and the Multinormal Integral," *Journal of the Royal Statistical Society, B* **20**, 373-378.

Williams, D. A. (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics*, **31**, 144-148.

Figure 1. Q-Q plots of the over-dispersed binomial (y -axis) versus beta-binomial (x -axis) distribution for common intra-cluster correlation.

Figure 2. Q-Q plots of the over-dispersed binomial (y -axis) versus beta-binomial (x -axis) distribution for different correlation structures.

Figure 3. Q-Q plots of the under-dispersed binomial distribution (y -axis) versus the pig data.

Figure 4. Plots of δ (y -axis) versus ρ (x -axis) in Equation (2).