

**University of Illinois at Chicago
School of Public Health
Division of Epidemiology and Biostatistics**

*Technical report#:2007-007
December 2007*

Title: A comparative study on most commonly used
correlated binary data generation methods

Authors: Hakan Demirtas, Donald Hedeker and Kush Kapur

**Affiliation(s): University of Illinois at Chicago, Division of Epidemiology and
Biostatistics—Department of Psychiatry.**

A comparative study on most commonly used correlated binary data generation methods

Hakan Demirtas, Donald Hedeker and Kush Kapur*

December 11, 2007

Abstract

Models for correlated binary data have received considerable interest in the last two decades. Testing and estimation in such models are predicated upon asymptotic theory, and one needs to resort to simulation techniques to assess the small sample behavior of the estimators. To address the need of generating multivariate binary data, several approaches have appeared in the literature. In this article, via a limited simulation study, we evaluate the performance of the three most commonly used methods in terms of the accuracy and precision of the estimates of the mean and association parameters.

Key Words: Correlated binary data; Random number generation; Accuracy; Efficiency; Simulation.

1 Introduction and motivation

In many applied areas of research, one encounters binary observations or measurements made on individual units. Such binary data are typically correlated when measurements are taken repeatedly over time from the same subject or the observations occur in clusters. The multivariate Bernoulli distribution has applications in modeling system reliability, longitudinal clinical trials, and genetic transmission of disease, among other applications.

While the asymptotic properties of regression-type models for binary data are fairly well understood, their small sample properties are not well known. In order to evaluate the

*Hakan Demirtas (e-mail:demirtas@uic.edu) is Assistant Professor and Donald Hedeker is Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 W. Taylor St., Chicago, IL, 60612. Kush Kapur is Research Data Analyst, Center for Health Statistics (MC912), University of Illinois at Chicago, 1601 W. Taylor St., Chicago, IL, 60607.

finite sample properties of the estimators, one needs to conduct a Monte Carlo simulation that requires generating binary random variates having specified mean and correlation structures.

Here, we considered the three most widely used correlated binary data generation techniques (Emrich and Piedmonte, 1991; Park, Park, and Shin, 1996; Lee, 1993), and empirically evaluated their comparative performance in terms of some key bias and efficiency measures across simulated data sets with competing first and second moment specifications.

The organization of the paper is as follows. In Section 2, we describe salient operational characteristics of these approaches for generating correlated binary data. In Section 3, we present a simulation study that is devised to evaluate the performance of the three methods in terms of commonly accepted accuracy and precision measures. Section 4 includes concluding remarks and discussion.

2 Overview

In this section, we elaborate on the three major algorithms that have appeared in the statistical literature for generating correlated binary data.

2.1 Probability integral transformation

Emrich and Piedmonte (1991) introduced a method for generating correlated binary data in which the joint distribution of the binary variables is completely determined by “borrowing” the third and higher order moments from a multivariate normal distribution. Let Y_1, \dots, Y_j represent binary variables such that $E[Y_j] = p_j$ and $Corr(Y_j, Y_k) = \delta_{jk}$, where p_j ($j = 1, \dots, d$) and δ_{jk} ($j = 1, \dots, d-1; k = 2, \dots, d$) are given, and where $d \geq 2$. As Emrich and Piedmonte (1991) noted, δ_{jk} is bounded below by $max\left(-\sqrt{(p_j p_k / q_j q_k)}, -\sqrt{(q_j q_k / p_j p_k)}\right)$ and above by $min\left(\sqrt{(p_j q_k / q_j p_k)}, \sqrt{(q_j p_k / p_j q_k)}\right)$, where $q = 1 - p$. Let $\Phi[x_1, x_2, \rho]$ be the cumulative distribution function for a standard bivariate normal random variable with correlation coefficient ρ . Naturally, $\Phi[x_1, x_2, \rho] = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(z_1, z_2, \rho) dz_1 dz_2$, where $f(z_1, z_2, \rho) = [2\pi(1 - \rho^2)^{1/2}]^{-1} \times exp\left[-(z_1^2 - 2\rho z_1 z_2 + z_2^2)/(2(1 - \rho^2))\right]$. We could generate multivariate

normal outcomes (Z 's) whose correlation parameters are obtained by solving the equation

$$\Phi[z(p_j), z(p_k), \rho_{jk}] = \delta_{jk}(p_j q_j p_k q_k)^{1/2} + p_j p_k, \quad (1)$$

for ρ_{jk} ($j = 1, \dots, d-1; k = 2, \dots, d$) where $z(p)$ denotes the p^{th} quantile of the standard normal distribution. As long as δ_{jk} satisfies the range condition mentioned above, the solution is unique. Repeating this numerical integration process $d(d-1)/2$ times, one can obtain the overall correlation matrix (say Σ) for the d -variate standard normal distribution with mean 0. However, it should be noted that positive-definiteness of Σ cannot be guaranteed. To create dichotomous outcomes (Y_j) from the generated normal outcomes (Z_j), we set $Y_j = 1$ if $Z_j \leq z(p_j)$ and 0 otherwise for $j = 1, \dots, d$. This produces a vector with the desired properties:

$E[Y_j] = P(Y_j = 1) = P(Z_j \leq z(p_j)) = p_j$, and $Cov(Y_j, Y_k) = P(Y_j = 1, Y_k = 1) - p_j p_k = P(Z_j \leq z(p_j), Z_k \leq z(p_k)) - p_j p_k = \Phi[z(p_j), z(p_k), \rho_{jk}] - p_j p_k = \delta_{jk}(p_j q_j p_k q_k)^{1/2}$. Therefore, $Corr(Y_j, Y_k) = Cov(Y_j, Y_k)/(p_j q_j p_k q_k)^{1/2} = \delta_{jk}$ by Equation (1).

2.2 Poisson sums

Park, Park, and Shin (1996) provided a method based on sums of Poisson random variables in which the sums have some terms in common. Briefly, let Z_1, Z_2 , and Z_3 be independent Poisson random variables with nonnegative parameters $\alpha_{11} - \alpha_{12}$, $\alpha_{22} - \alpha_{12}$, and α_{12} , with the convention that a Poisson with parameter 0 is a degenerate random variable equal to 0, and define the random variables X_1 and X_2 as $X_1 = Z_1 + Z_2$ and $X_2 = Z_2 + Z_3$. They further define the binary random variables Y_1 and Y_2 by $Y_i = 1$ if $X_i = 0$, and 0 otherwise, where $i = 1, 2$. Subsequently, they determine the constants, α_{11}, α_{22} , and α_{12} , so that $E(Y_i) = p_i$ and $Corr(Y_1, Y_2) = \delta_{12}$. For arbitrary d , it is easy to see that $\alpha_{ij} = \log\left(1 + \delta_{ij} \sqrt{\frac{(1-p_i)(1-p_j)}{p_i p_j}}\right)$ ($i = 1, \dots, d-1; j = 2, \dots, d$) yields those relations. The details of the procedure are outlined below:

1. Set $k = 0$.
2. Set $k = k + 1$. Let $\beta_k = \alpha_{rs}$ be the smallest positive α_{ij} .

3. If $\alpha_{rr} = 0$ or $\alpha_{ss} = 0$, then stop.
4. Let S_k be the set of all indices, i, j , for which $\alpha_{ij} > 0$. For all $i, j \subseteq S_k$,
set $\alpha_{ij} = \alpha_{ij} - \beta_k$,
5. If not all $\alpha_{ij} = 0$, then go to step 1.
6. Generate k Poisson deviates, Z_j , with parameters β_j . For $i = 1, 2, \dots, d$,
set $Y_i = \sum_{i \in S_j} Z_j$.
7. For $i = 1, 2, \dots, d$, set $Y_i = 1$ if $Z_i = 0$, and 0 otherwise.

2.3 Archimedian copulas

Lee (1993) utilized Archimedian copulas that are a convenient family of distributions having uniform marginals. Within this family, distribution functions are of the form $H(x_1, \dots, x_d) = \phi^{-1}[\sum_{i=1}^d \phi(x_i)]$, where x_i 's are uniformly distributed in $[0, 1]$ and ϕ is a function $[0, 1] \rightarrow [0, \infty]$ having the properties

- $\phi(1) = 0, \phi(0) = \infty$.
- $(-1)^k (\phi^{-1})^{(k)}(x) \geq 0$ for $k = 1, 2, \dots, d$, where $(\phi^{-1})^{(k)}$ is the k^{th} derivative of ϕ^{-1} .

An appropriate set of functions ϕ yielding a complete range of correlations are those given by $\phi_\psi(x) = -\ln\left(\frac{\psi^x - 1}{\psi - 1}\right)$ for $0 < \psi < 1$, and $\phi_\psi(x) = -\ln(x)$ for $\psi = 1$. Here, ψ is the solution to Equation (2) below:

In the bivariate case ($d = 2$), following the previous notation, let $E[Y_j] = p_j$ for $j = 1, 2$, and $Corr(Y_1, Y_2) = \delta_{12}$. The relationship between ψ and the parameter vector, (p_1, p_2, δ_{12}) , is

$$\delta_{12} \sqrt{p_1 p_2 (1 - p_1)(1 - p_2)} + (1 - p_1)(1 - p_2) = \frac{\ln(1 + (\psi^{1-p_1} - 1)(\psi^{1-p_2} - 1)/(\psi - 1))}{\ln(\psi)} \quad (2)$$

Solving for ψ , plugging it in the equation for ϕ for each variable, and employing a linear programming routine that finds each cell probability, allows one to complete the generation

of the correlated binary data. The estimate of ψ does not change with different marginal proportions and correlations in higher dimensions ($d > 2$). Not every combinations of parameter values is possible since a common odds ratio is assumed; in other words, different levels of the mean and correlation parameters are allowed provided that they lead to the same odds ratio and a unique ψ . For deeper conceptual and computational issues, Lee (1993) should be consulted.

2.4 Respective advantages and drawbacks

In what follows, *EP*, *PPS*, and *LEE* stand for the methods proposed by Emrich and Piedmonte (1991), Park, Park, and Shin (1996), and Lee (1993), respectively.

EP involves the solution of a non-linear equation, thereby necessitating numerical integrations. The cumulative distribution function (cdf) of the normal distribution is flat at the extremes, hence the numerical integration solution is not reliable in some regions. Furthermore, the resulting intermediate correlation matrix is not guaranteed to be positive definite. On the positive side, *EP* can accommodate any feasible correlation structure. *LEE* makes a restrictive assumption of common odds ratios, which is unrealistic in many situations. In addition, Equation (2) does not always have a solution for ψ in $(0, 1]$. The complexity of implementation is no more straightforward than *EP* due to non-linear equation solving. *PPS* does not require an iterative fitting method, and is the simplest to execute among the three. However, it only works for non-negative correlations, substantially limiting its usefulness. Since none of the three methods is clearly superior, an empirical assessment is called for. In the next section, we describe our simulation design to assess the comparative performance of the three methods.

3 A Simulation Study

The simulation setup involves bivariate binary data which are generated by specifying the marginal means and Pearson correlations. In fact, for bivariate data these three quantities fully determine the joint distribution, obviating the need to use the above techniques; a

simple uniform random number generator would be adequate to generate bivariate binary data. However, while simple, it should elucidate the key aspects of the relative performances of the three methods for the purpose of comparison. Obviously, the scope of the simulation study could be broader. However, the essential conclusions conveyed in this investigation remained largely unchanged under more complex situations ($d > 2$) that are not reported here.

The marginal expectations (p_1 and p_2) and correlation (δ_{12}) were specified from 0.05 to 0.95 in increments of 0.10, leading to a total of $10 \times 10 \times 10 = 1000$ scenarios. Because *PPS* does not allow for negative correlations, negative correlations were not included in the simulated scenarios. Also, 482 examples out of a total of 1000 scenarios fulfilled the restriction regarding the upper bounds of the correlation imposed by the marginal means (lower bounds are irrelevant since all correlations are positive). A further disqualification occurred in comparisons that involve *LEE* due to the fact that ψ in Equation (2) did not have a root in $(0, 1]$; this eliminated 64 examples. Thus, 418, 482, and 418 scenarios were used in the two-way *PPS* – *LEE*, *PPS* – *EP*, and *EP* – *LEE* comparisons, respectively (Tables 1-3). Naturally, 418 examples were examined in the three-way comparisons (Table 4). For each example, the number of simulated replicates was 500.

3.1 Evaluation Criteria

The evaluation was conducted using the following four quantities:

- Percentage bias (PB): the relative magnitude of the raw bias to the true value of the parameter, $100 \times \left| \frac{E(\hat{\theta}) - \theta}{\theta} \right|$.
- Standardized bias (SB): the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is θ , the standardized bias is $100 \times \frac{|E(\hat{\theta}) - \theta|}{SE(\hat{\theta})}$, where *SE* stands for standard error. (Demirtas, 2004).
- Variance (VAR): the variance of the estimates across $N = 500$ simulated replicates.
- Root-mean-square error (RMSE): an integrated measure of bias and variance. It is considered to be arguably the best criterion for evaluating $\hat{\theta}$ in terms of combined accuracy

Table 1: Comparison of bias and efficiency measures between *PPS* and *LEE* across 418 scenarios (larger is better).

	PB	SB	VAR	RMSE
Parameter	<i>PPS</i> – <i>LEE</i>	<i>PPS</i> – <i>LEE</i>	<i>PPS</i> – <i>LEE</i>	<i>PPS</i> – <i>LEE</i>
p_1	203-215	203-215	217-201	217-201
p_2	199-219	198-220	212-206	215-203
δ_{12}	221-197	222-196	215-203	214-204

Table 2: Comparison of bias and efficiency measures between *PPS* and *EP* across 482 scenarios (larger is better).

	PB	SB	VAR	RMSE
Parameter	<i>PPS</i> – <i>EP</i>	<i>PPS</i> – <i>EP</i>	<i>PPS</i> – <i>EP</i>	<i>PPS</i> – <i>EP</i>
p_1	235-247	237-245	249-233	246-236
p_2	225-257	225-257	253-229	250-232
δ_{12}	307-175	309-173	326-156	336-146

and precision. $RMSE(\hat{\theta})$ is defined as $\sqrt{E_{\theta}[\hat{\theta} - \theta]^2}$.

Under the above specification, *PB* and *SB* are accuracy measures, *VAR* is an efficiency measure, and *RMSE* is a hybrid measure. The reason we use two different bias quantities is that both have relative merits and pitfalls: *SB* depends on the total inherent variability which may be too small or too large, causing misleading interpretations; and *PB* has the assumed true value of the estimand in the denominator which similarly may take extreme values. In our limited experience, it is advisable to consider both accuracy yardsticks simultaneously. For other examples of our evaluation system and its expanded versions, see Demirtas and Schafer (2003), Demirtas (2004, 2005), and Demirtas and Hedeker (2007).

The results for two-way comparisons are presented in Tables 1, 2 and 3. In these tables, the numbers represent frequencies, and larger numbers are associated with smaller percentage and standardized biases, smaller variability and *RMSE*'s. For example in Table 1, the entry for parameter p_1 and *PB*, 203 – 215 means that in 203 out of 418 scenarios,

Table 3: Comparison of bias and efficiency measures between EP and LEE across 418 scenarios (larger is better).

	PB	SB	VAR	RMSE
Parameter	$EP - LEE$	$EP - LEE$	$EP - LEE$	$EP - LEE$
p_1	211-207	211-207	188-230	191-227
p_2	210-208	211-207	216-202	218-200
δ_{12}	144-274	143-275	139-279	128-290

Table 4: Three-way ranks (smaller is better) across 418 scenarios.

	PB	SB	VAR	RMSE
Parameter	$PPS/EP/LEE$	$PPS/EP/LEE$	$PPS/EP/LEE$	$PPS/EP/LEE$
p_1	848/828/832	846/830/832	822/860/826	819/866/823
p_2	860/821/827	861/820/827	823/834/851	824/838/846
δ_{12}	773/952/783	771/954/783	754/994/760	763/973/772

PB for PPS is smaller (hence better) than PB for LEE , and in 215 scenarios, the reverse is true. Therefore, a larger number represents better performance. Table 2 and 3 were also constructed with the same convention. A summary of our findings are given below:

- PPS is more biased than LEE for the marginals (p_1 and p_2) as measured by PB and SB .
- PPS is more precise than LEE as measured by VAR .
- PPS is better than LEE in terms of the $RMSE$ for the marginals.
- PPS is better than LEE in every metric for the correlation (δ_{12}).
- PPS is more biased than EP for the marginals.
- PPS is more precise than EP .
- PPS is better than EP in terms of the $RMSE$ for the marginals.
- PPS is much better than EP in every metric for the correlation.

- *EP* is slightly less biased than *LEE* for the marginals.
- *LEE* is more precise than *EP* for p_1 , and less precise for p_2 .
- *LEE* has lower *RMSE* than *EP* for p_1 , and the opposite is true for p_2 .
- *LEE* is much better than *EP* in every metric for the correlation.

Regarding the marginal proportions, *EP* and *LEE* appear to produce more accurate estimates compared to *PPS*. However, from an efficiency standpoint, the performance of *PPS* is superior to that of *EP* and *LEE*. This is not unexpected considering the historical trade-off between bias and variance. One can see that relative gains in precision using *PPS* outweigh the losses in accuracy in comparison to *EP* and *LEE*, by examining the combined measure of bias and variance, *RMSE*, which is smaller for *PPS* on average. The comparative results of *EP* and *LEE* are a bit ambivalent: *EP* seems to have a slight edge over *LEE* for bias quantities. While variability and *RMSE* favor *LEE* for p_1 , and *EP* for p_2 to a lesser degree; overall, *LEE* appears to perform better than *EP* in terms of precision. As far as the association parameter (δ_{12}) is concerned, *PPS* and *EP* are the clear winner and loser, respectively, with *LEE* taking the second spot in every evaluation metric. The poor performance of *EP* for this parameter stems from the fact that in situations where correlations are close to the upper limits imposed by marginal expectations, the cdf of the bivariate normal is nearly flat, hence normal correlations (ρ_{jk}) in Equation (1) are not well estimated.

For cross-validation purposes, we also performed three-way comparisons (Table 4). For each scenario, ranks of the three methods were obtained with the best, second, and worst performing method assigned values of 1, 2, 3, respectively. Table 4 lists the sum of these ranks across 418 scenarios. In Table 4, smaller numbers correspond to better performance. As one would expect, the results are similar to the results of the pairwise comparisons described above.

We do not intend to attach any negative or positive connotations to any of these methods. The choice between them should be guided by the applied context of the problem,

discipline-specific considerations, and perceived relative importance of accuracy and precision.

4 Concluding remarks

There are a few points that need to be addressed. First of all, one may argue that the presented simulation study is unrealistically simplistic given countless and more complicated scenarios that can be encountered in practice. This argument has certain validity; however, our primary purpose was to make recommendations to applied researchers as to which of these methods should be the preferred approach under differing circumstances via an objective assessment of the performance of the three most commonly utilized correlated binary data generation routines that are typically needed in simulated examples that help to examine the small sample properties in regression type of models that are developed for clustered or longitudinal binary outcomes. Although our simulations revealed no clear winner, the strengths and limitations of the methods were discussed; and potentially useful results were obtained for the relative behavior of accuracy and precision measures. On a related note, one may question the generalizability of the simulation results. Admittedly, this is a somewhat limited experiment. However, within the experiment, many different possible combinations of marginal expectations and correlation (482 to be exact) were examined, and we believe that the scope of the set of scenarios is sufficiently broad. Also, no claims were made in regard to the speed aspect of the implementation since this critically depends upon the choice of compiler, linker, debugger, platform, and software as well as the efficiency of the computer code. Furthermore, in our view, the relative magnitudes of bias and variability with respect to the truth are more useful components of scientific inquiry than computation speed, given today's rather sophisticated computational tools. Finally, we are aware that other methods for correlated binary data generation are available. However, our intent was to evaluate the most popular approaches. A few other pertinent methods that we did not include are described by Gange (1995), Lunn and Davies (1998), and Oman and Zucker (2001).

REFERENCES

- Demirtas H. and Schafer JL. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 22:2553–2575.
- Demirtas H. (2004). Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica* 58:466–482.
- Demirtas, H. (2005). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 24:2345–2363.
- Demirtas, H. and Hedeker, D. (2007). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine* 26:782–799.
- Emrich, J.L. and Piedmonte, M.R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* 45:302–304.
- Gange, S.J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 49:134–138.
- Lee, A.J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician* 47:209–215.
- Lunn, A.D. and Davies, S.J. (1998). A note on generating correlated binary variables. *Biometrika* 85:487–490.
- Oman, S.D. and Zucker, D.M. (2001). Modelling and generating correlated binary variables. *Biometrika* 88:287–290.
- Park, C.G., Park, T. and Shin D.W. (1996). A simple method for generating correlated binary variates. *The American Statistician* 50:306–310.