University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2016 *Paper* 46

Simulating Longer Vectors of Correlated Binary Random Variables via Multinomial Sampling

Justine Shults*

*Division of Biostatistics, Univeristy of Pennsylvania Perelman School of Medicine, jshults@mail.med.upenn.edu

Simulating Longer Vectors of Correlated Binary Random Variables via Multinomial Sampling

Justine Shults

Abstract

The ability to simulate correlated binary data is important for sample size calculation and comparison of methods for analysis of clustered and longitudinal data with dichotomous outcomes. One available approach for simulating length n vectors of dichotomous random variables is to sample from the multinomial distribution of all possible length n permutations of zeros and ones. However, the multinomial sampling method has only been implemented in general form (without ?rst making restrictive assumptions) for vectors of length 2 and 3, because specifying the multinomial distribution is very challenging for longer vectors. I overcome this di?culty by presenting an algorithm for simulating correlated binary data via multinomial sampling that can be easily applied to directly compute the multinomial distribution for any n. I demonstrate the approach to simulate vectors of length 4 and 8 in an assessment of power during the planning phases of a study and to assess the choice of working correlation structure in an analysis with generalized estimating equations.

Simulating Longer Vectors of Correlated Binary Random Variables via Multinomial Sampling

J. Shults[a*]

[a]*Department of Biostatistics, University of Pennsylvania, PA 19104, USA*

(*v4.0 released January 2015*)

The ability to simulate correlated binary data is important for sample size calculation and comparison of methods for analysis of clustered and longitudinal data with dichotomous outcomes. One available approach for simulating length n vectors of dichotomous random variables is to sample from the multinomial distribution of all possible length n permutations of zeros and ones. However, the multinomial sampling method has only been implemented in general form (without first making restrictive assumptions) for vectors of length 2 and 3, because specifying the multinomial distribution is very challenging for longer vectors. I overcome this difficulty by presenting an algorithm for simulating correlated binary data via multinomial sampling that can be easily applied to directly compute the multinomial distribution for any n. I demonstrate the approach to simulate vectors of length 4 and 8 in an assessment of power during the planning phases of a study and to assess the choice of working correlation structure in an analysis with generalized estimating equations.

Keywords: binary random variables; dichotomous; generalized estimating equations; multinomial sampling; simulation

1. Introduction

Methods to simulate realizations of dependent variables with specified marginal means and pairwise correlations are useful to assess semi-parametric approaches such as generalized estimating equations (GEE) [1], which only require models for the first two moments of the distribution of the outcome variable. Continuous variables can be simulated via the multivariate normal distribution that is determined by its mean and covariance matrix. In contrast, dependent Bernoulli random variables present a greater simulation challenge, due to the lack of an equally general and flexible equivalent of the normal distribution for discrete data.

Quite a few useful methods have been proposed, but how best to simulate correlated binary data remains an active area of research in the statistical literature. [2] reviewed methods for the simulation of correlated binary data, including an approach by [3] that allows for unstructured correlations and non-stationary data. [4] offered a flexible method for simulation that is somewhat complex because it involves the solution of non-linear equations via numerical integration. [5] proposed an approach that relies on the association among the random variables resulting from their sharing some common components that induce correlation. Recently, [6] compared the approach of [3] with a method based on the multivariate probit distribution.

*Corresponding author. Email: jshults@mail.med.upenn.edu

This manuscript considers the simulation of Bernoulli random variables with specified marginal means and pairwise correlations via the multinomial sampling approach that considers "k-variate binary data as a multinomial distribution with 2^k possible outcomes" [7]. For paired data, the multinomial sampling approach that was proposed by [7] is a special case of the simulation method proposed for bivariate binomial data by [8, p.751]. More recently, [9] showed that multinomial sampling was comparable with the method of [4], which [9] referred to as the gold-standard approach. (However, [9] also acknowledged that they did not make comparisons with the approach of [3].)

Because the multinomial sampling approach involves complete specification of the underlying distribution of all possible permutations of zeros and ones, its use will ensure that a valid multivariate parent distribution exists that is compatible with particular specified marginal means and covariance matrix. The lack of a compatible parent is not typically a concern for continuous variables, because the multivariate normal distribution is a possible valid parent, even if it does not fit the data well. However, for discrete random variables it is known that for a given marginal mean, covariance matrix pair, it is not guaranteed that a valid distribution exists with the specified mean and covariance matrix [10]. As discussed by [11], although not fully specified, the parent distribution provides an estimation framework and probabilistic basis for semi-parametric methods such as GEE or the quasi-least squares approach [12–14, QLS] that is in the framework of GEE.

From a practical perspective, [15] cautioned that the additional constraints necessary to ensure a valid parent distribution are especially important to assess during the planning phases of a study. For example, consider sample size calculation with a standard formula, such as the one provided in [16, p. 167]. If means and correlations for which there is no valid parent distribution are used, the formula will provide results, but they will be invalid, and no warning will be provided. Assessing power with a method that simulates from a compatible parent distribution will ensure that the results are sound. [17] suggested that a *severe* violation of constraints during the analysis could be used as a rule-out criterion in the selection of a working correlation structure to describe the pattern of association in the data.

Although it is useful for assessing the parent distribution, the multinomial sampling approach has not been implemented for vectors of length four or more, without first making simplifying assumptions such as the first-order Markov property [18, p. 13] or the exchangeability condition [7, Section 5]. As explained by [9] (who simulated vectors of length 2 and 3), "the CDF for establishing decision rules becomes complicated for cases of four or more repeated measures. While not impossible, constructing higher order joint probabilities can be computationally challenging." To overcome the difficulty described by [9], this paper presents an algorithm for constructing higher order joint probabilities that is straightforward to implement. I describe the method in Section 2 and demonstrate its application in Section 3, in an assessment of power and selection of a working correlation structure for GEE.

2. Methods

2.1. *Notation and assumptions*

Let $\underline{Y}_n = (Y_1, \ldots, Y_n)$ be an $n \times 1$ vector of Bernoulli random variables Y_j with marginal means $E(Y_j) = P(Y_j = 1) = p_j$ and variances $Var(Y_j) = p_j\, q_j$, where $q_j = 1 - p_j$ $(j = 1, \ldots, n)$. Let $R_n = \text{Corr}(\underline{Y}_n)$ be the $n \times n$ correlation matrix of $\underline{Y}_n$, with $(j,k)^{th}$

entry,

$$\begin{aligned} R_n[j,k] &= \rho_{jk} && (1)\\ &= \frac{E(Y_jY_k) - p_jp_k}{\sqrt{p_jp_kq_jq_k}} && (2)\\ &= \frac{p_{jk} - p_jp_k}{\sqrt{p_jp_kq_jq_k}}, && (3) \end{aligned}$$

where $p_{jk} = E(Y_jY_k) = P(Y_j = 1, Y_k = 1)$.

Let $mod(x, y)$ represent the modulus of x with respect to y, so that

$$mod(x, y) = x - y\,\text{floor}(x/y), \tag{4}$$

where floor(x/y) is the unique integer n such that $n \leq x/y < n + 1$. Let $B_n(i)$ be the length n binary representation of $i-1$ that is expressed as an $n \times 1$ vector $(i = 1, \ldots, 2^n)$. For example, $B_3(2) = \big(1, 0, 0\big)'$ is the length 3 binary representation of 1 because $1 = 1 \times 2^0 + 0 \times 2^1 + 0 \times 2^2$. Furthermore, let $P_n(i) = P(\underline{Y}_n = B_n(i))$ $(i = 1, \ldots, 2^n)$.

2.2. *Construction of multinomial distribution for Bernoulli vectors of length n*

Kang and Jung [7] proposed an algorithm for simulating dependent Bernoulli random variables $\underline{Y}_n$ via multinomial sampling that, with slightly different notation, can be described as follows. To simulate a sample of size m, alternate m times between the following two steps. *Step One:* Simulate a value U from a uniform $(0, 1)$ distribution. *Step Two:* Select sequence $B_n(i)$ if $Z_n(i-1) \leq U < Z_n(i)$, where $Z_n(0) = 0$ and $Z_n(i) = \sum_{j=0}^{i} P_n(j)$ for $i = 1, \ldots 2^n$. The probability that $B_n(i)$ is selected in Step Two is the length of the interval $[Z_n(i-1), Z_n(i)\rangle = Z_n(i)$ - $Z_n(i-1) = P_n(i)$ $(i = 1, \ldots, 2^n)$.

This algorithm is straightforward to implement for $n = 2$ because the marginal means p_1, p_2 and the pairwise correlation ρ_{12} can be used to easily construct the well-known bivariate Bernoulli distribution, for which $P_2(1) = q_1\,q_2 + \rho_{12}\sqrt{p_1q_1p_2q_2}$; $P_2(2) = p_1q_2 - \rho_{12}\sqrt{p_1q_1p_2q_2}$; $P_2(3) = q_1p_2 - \rho_{12}\sqrt{p_1q_1p_2q_2}$; and $P_2(4) = p_1p_2 + \rho_{12}\sqrt{p_1q_1p_2q_2}$. [7] noted that we also require $0 \leq P_n(i) \leq 1$ $(i = 1, \ldots, 2^n)$; simple algebra can be used to show that these inequalities will be satisfied for $n = 2$ as long as the pairwise correlations satisfy the following constraints that were provided by Prentice [19, p. 46]:

$$\rho_{12L} < \rho_{12} < \rho_{12U}, \tag{5}$$

where $\rho_{12L} = \max\left[-\sqrt{\frac{p_1p_2}{q_1q_2}}, -\sqrt{\frac{q_1q_2}{p_1p_2}}\right]$ and $\rho_{12U} = \min\left[\sqrt{\frac{p_2q_1}{p_1q_2}}, \sqrt{\frac{p_1q_2}{p_2q_1}}\right]$.

However, as noted earlier, it becomes increasingly difficult to directly obtain expressions for the $P_n(i)$ for larger n. I therefore developed an easy to program algorithm for constructing a consistent system of equations with a unique solution that contains the probabilities $P_n(i)$ of all possible length n permutations of zeros and ones. To achieve this, I first note that $B_n(j)$ can be expressed as follows:

$$B_n(j) = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{nj} \end{pmatrix}, \tag{6}$$

where y_{ij} is $mod\,[w_n(i,j), 2]$ and $w_n(i,j)$ is the integer part of $(2^n - j)/2^{i-1} + 1$. Then, the set of all possible length n permutations of zeros and ones is given by S_n, where $S_n = \{B_n(j), j = 1, \ldots, 2^n\}$. For example, if $n = 3$, then

$$S_3 = \{B_3(1), \ldots, B_3(8)\} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

The following theorem then provides the basis for an approach to construct the multinomial distribution for any n.

THEOREM 2.1 *Define $P_{rob} = (P_n(1), \ldots, P_n(2^n))'$, where $P_n(i)$ is the probability that $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ is the binary representation of $i - 1$. Then, P_{rob} can be expressed as*

$$P_{rob} = \left(X'X\right)^{-1} X'\,E, \tag{7}$$

where X and E are defined as follows.

$$X_{2^n \times 2^n} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \\ \underline{One}_{2^n} \end{pmatrix} \quad and \quad E_{2^n \times 1} = \begin{pmatrix} E_1 \\ \vdots \\ E_n \\ 1 \end{pmatrix}, \ where \tag{8}$$

$$X_1 = \left(B_n(1)\ B_n(2) \ldots B_n(2^n)\right), \tag{9}$$

$B_n(i)$ are defined using (6), $\underline{One}_{2^n}$ is a 1×2^n is row vector of ones,

$$E_1 = \left(p_1\ p_2 \ldots p_n\right)', \tag{10}$$

and X_j and E_j are constructed using the following algorithm ($j = 2, \ldots, n$):

- *Let* **num** *= 1.*
- *Let $i_1 = 1$.*
- *While $i_1 \leq n - j + 1$ {*
- *Let $i_2 = i_1 + 1$.*
- *While $i_2 \leq n - j + 2$ {*
- $\vdots$
- *Let $i_j = i_{j-1} + 1$.*
- *While $i_j \leq n - j + j = n$ {*

- *Row* **num** *of* X_j *is*

$$r_n(i_1) \cdot r_n(i_2) \cdot \ldots r_n(i_j) = \left(\prod_{w=1}^{j} y_{i_w\,1} \ldots, \prod_{w=1}^{j} y_{i_w\,2^n} \right), \tag{11}$$

the row vector obtained by element-wise multiplication of $r_n(i_1)$ *through* $r_n(i_j)$*, where* $r_n(a)$ *is the* a^{th} *row of matrix* X_1*.*
- *The corresponding element* **num** *of* E_j *is*

$$E(\prod_{w=1}^{j} Y_{i_w}) = P(Y_{i_1} = 1, \ldots, Y_{i_j} = 1) = p_{i_1 \ldots i_j}. \tag{12}$$

- *Let* **num** = **num** + *1.*
- *Let* $i_j = i_j$ + *1.* }
- ⋮
- *Let* $i_2 = i_2 + 1$ }
- *Let* $i_1 = i_1 + 1$ }.

Proof. To prove that $X\,Prob = E$, I first note that the product of row i of matrix X_1 and $Prob$ is $\sum_{j=1}^{2^n} y_{ij} P_n(j) = E(Y_i) = P(Y_i = 1) = p_i$ $(i = 1, \ldots, n)$. Similarly, the product of row **num** of matrix X_j in (11) and $Prob$ is $\sum_{k=1}^{2^n} \prod_{w=1}^{j} y_{i_w\,k} P_n(k) = E(\prod_{w=1}^{j} Y_{i_w}) = P(Y_{i_1} = 1, \ldots, Y_{i_j} = 1) = p_{i_1 \ldots i_j}$ $(j = 2, \ldots, n)$. In addition, the product of the last row of X and $Prob$ is $\sum_{i=1}^{2^n} P_n(i) = 1$. Next, the dimension of matrix X is $\sum_{j=1}^{n} \binom{n}{j} + 1 = \sum_{j=0}^{n} \binom{n}{j} = 2^n$ by 2^n. The rank of X is 2^n because X_1 has full column rank. Matrix X is therefore a full-rank square matrix, and is invertible. Equation $X\,Prob = E$ therefore has a unique explicit solution $P_{rob} = (X'X)^{-1} X'\,E$. ■

In words, matrix X was obtained by first stacking matrices X_1 through X_n, where the rows of X_j contain the $\binom{n}{j}$ rows $(j = 1)$ or element-wise products $(j = 2, \ldots, n)$ of matrix $\left(B_n(1)\ B_n(2) \ldots B_n(2^n)\right)$ that contains all possible realizations of $\underline{Y}_n$. Note that X_n will be the element-wise product of all rows of X_1, and will therefore be a 1 by 2^n row vector with all elements equal to 0, except for element $(1, 2^n)$ that equals 1. The final row of X is then a 1 by 2^n row vector of ones.

The probabilities in (7) satisfy $Prob[i] = P_n(i)$ for which $\sum_{i=1}^{2^n} P_n(i) = 1$. However, as was the case for $n = 2$, additional constraints must be satisfied to ensure that the $P_n(i)$ take value in $(0, 1)$. To achieve $0 \leq P_n(i) \leq 1$ $(i = 1, \ldots, 2^n)$ for a specified value for E, it is sufficient to ensure that each element of E_j satisfies the constraints necessary to ensure non-negative probabilities for a vector of length j $(j = 1, \ldots, n)$. For example, for $n = 4$, we first check that the constraints are satisfied for $(p_1,\ p_2,\ p_3,\ p_4)$, and then for $(p_{12},\ p_{13},\ p_{14},\ p_{23},\ p_{24},\ p_{34})$, and then for $(p_{123},\ p_{124},\ p_{134},\ p_{234})$, and finally, for p_{1234}.

Since $\sum_{i=1}^{2^n} P_n(i) = 1$, we can find the constraints on $p_{1\,\ldots\,n}$ by solving $0 \leq P_n(i)$ for $p_{1\,\ldots\,n}$ $(i = 1, \ldots, 2^n)$. The $P_n(i)$ are linear combinations of the elements of E that each have a coefficient of $+1$ or -1 for $p_{1\,\ldots\,n}$. As a result, it is trivial to solve the equations directly for smaller n, or by using the following algorithm that can be applied using software that performs symbolic computations:

- Let $j = 1$ and $k = 1$.
- For $i = 1, \ldots, 2^n$:
  - Let $\delta = P_n(i) - sol + p_{1\ldots n}$, where sol is the solution of $P_n(i) = 0$ for $p_{1\ldots n}$.
  - If $\delta = 0$ then Upper$(k) = sol$ and then $k = k + 1$.
  - Else Lower$(j) = sol$ and then $j = j + 1$.

The overall constraints for $p_{1\ldots n}$ will then be

$$(L_{1,\ldots,n}, U_{1,\ldots,n}) = \left(\max_{j\in\{1,\ldots,J\}} \{\text{Lower}(j)\}, \min_{k\in\{1,\ldots,K\}} \{\text{Upper}(k)\} \right),$$

where J and K are the maximum values of j and k, respectively. This algorithm can also be applied to obtain the constraints for $p_{i_1\ldots i_w}$, by substituting i_j for j $(j = 1, \ldots, w)$ in the expressions for the constraints for $p_{1\ldots w}$.

I next demonstrate the use of Theorem 2.1 to obtain the $Prob[i] = P_n(i)$ in (7), and obtain constraints sufficient to ensure $0 \le P_n(i)$ $(i = 1, \ldots, 2^n)$, for $n = 1, \ldots, 4$. Results for $n = 5$ are provided in Appendix B. Text files that contain summarized results for $n = 2, \ldots, 12$ are provided in the Supplemental Material, as described in Section 5.

2.2.1. For $n = 1$:

$$X = \begin{pmatrix} X_1 \\ \underline{One}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} p_1 \\ 1 \end{pmatrix}. \tag{13}$$

Then,

$$\begin{aligned} P_{rob} &= (P_1(1), P_1(2))' && (14)\\ &= \left(X'X\right)^{-1} X'E && (15)\\ &= (1 - p_1, p_1)'. && (16) \end{aligned}$$

Solving $0 \le P_1(i)$ for p_1 $(i = 1, 2)$ yields the known constraint $0 \le p_1 \le 1$.

2.2.2. For $n = 2$:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \underline{One}_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} p_1 \\ p_2 \\ p_{12} \\ 1 \end{pmatrix},$$

where

$$p_{12} = p_1 p_2 + \rho_{12}\sqrt{p_2 p_2 q_1 q_2}. \tag{17}$$

Then,

$$\begin{aligned} P_{rob} &= (P_2(1),\, P_2(2),\, P_2(3),\, P_2(4))' && (18)\\ &= \left(X'X\right)^{-1} X'E && (19)\\ &= (p_{12} - p_2 - p_1 + 1,\, p_1 - p_{12},\, p_2 - p_{12},\, p_{12})'. && (20) \end{aligned}$$

Solving $0 \leq P_2(i)$ for p_{12} $(i = 1, 2, 3, 4)$ yields the known constraints

$$\max\{0,\, p_1 + p_2 - 1\} \leq p_{12} \leq \min\{p_1,\, p_2,\}\,, \tag{21}$$

or solving for ρ_{12} yields the Prentice constraints (5) for the pairwise correlation.

2.2.3. For $n = 3$:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \underline{One}_8 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_{12} \\ p_{13} \\ p_{23} \\ p_{123} \\ 1 \end{pmatrix}. \tag{22}$$

Then,

$$\begin{aligned} P_{rob} &= (P_3(1), \ldots\, P_3(8))' \\ &= \left(X'X\right)^{-1} X'E \\ &= (p_{12} - p_2 - p_3 - p_1 + p_{13} + p_{23} - p_{123} + 1,\, p_1 - p_{12} - p_{13} + p_{123},\, p_2 - p_{12} - p_{23} + p_{123}, \\ &\quad p_{12} - p_{123},\, p_3 - p_{13} - p_{23} + p_{123},\, p_{13} - p_{123},\, p_{23} - p_{123},\, p_{123})'. \end{aligned} \tag{23}$$

Solving $0 \leq P_3(i)$ for p_{123} $(i = 1, \ldots, 8)$ yields the known constraints

$$L_{123} \leq p_{123} \leq U_{123}, \tag{24}$$

where $L_{123} = \max\{0, p_{12} + p_{13} - p_1, p_{12} + p_{23} - p_2, p_{13} + p_{23} - p_3\}$ and $U_{123} = \min\{p_{13},\, p_{12},\, p_{23},\, 1 + p_{12} - p_2 - p_3 - p_1 + p_{13} + p_{23}\}$. See Appendix A for discussion of prior work for $n = 3$.

2.2.4. For $n = 4$:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \underline{One}_{16} \end{pmatrix} = \begin{pmatrix} 0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1 \\ 0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \\ 0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1 \\ 0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1 \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_{12} \\ p_{13} \\ p_{14} \\ p_{23} \\ p_{24} \\ p_{34} \\ p_{123} \\ p_{124} \\ p_{134} \\ p_{234} \\ p_{1234} \\ 1 \end{pmatrix}.$$

Then,

$$\begin{aligned} P_{rob} &= (P_4(1), \ldots P_4(16))' \\ &= \left(X'X\right)^{-1} X'E \\ &= p_{12} - p_2 - p_3 - p_4 - p_1 + p_{13} + p_{14} + p_{23} + p_{24} + p_{34} - p_{123} - p_{124} - p_{134} - p_{234} + p_{1234} + 1, \\ &\quad p_1 - p_{12} - p_{13} - p_{14} + p_{123} + p_{124} + p_{134} - p_{1234}, p_2 - p_{12} - p_{23} - p_{24} + p_{123} + p_{124} + p_{234} \\ &\quad -p_{1234}, p_{12} - p_{123} - p_{124} + p_{1234}, p_3 - p_{13} - p_{23} - p_{34} + p_{123} + p_{134} + p_{234} - p_{1234}, p_{13} \\ &\quad -p_{123} - p_{134} + p_{1234}, p_{23} - p_{123} - p_{234} + p_{1234}, p_{123} - p_{1234}, p_4 - p_{14} - p_{24} - p_{34} + p_{124} + \\ &\quad p_{134} + p_{234} - p_{1234}, p_{14} - p_{124} - p_{134} + p_{1234}, p_{24} - p_{124} - p_{234} + p_{1234}, p_{124} - p_{1234}, \\ &\quad p_{34} - p_{134} - p_{234} + p_{1234}, p_{134} - p_{1234}, p_{234} - p_{1234}, p_{1234})'. \end{aligned}$$

Solving $0 \leq P_4(i)$ for p_{1234} $(i = 1, \ldots, 16)$ yields the constraints

$$L_{1234} \leq p_{1234} \leq U_{1234}, \tag{25}$$

where $L_{1234} = \max\{\ p_1 + p_2 + p_3 + p_4 - p_{12} - p_{13} - p_{14} - p_{23} - p_{24} - p_{34} + p_{123} + p_{124} + p_{134} + p_{234} - 1,\ p_{123} - p_{12} + p_{124},\ p_{123} - p_{13} + p_{134},\ p_{123} - p_{23} + p_{234},\ p_{124} - p_{14} + p_{134},\ p_{124} - p_{24} + p_{234},\ p_{134} - p_{34} + p_{234},\ 0\ \}$ and $U_{1234} = \min\{\ p_1 - p_{12} - p_{13} - p_{14} + p_{123} + p_{124} + p_{134},\ p_2 - p_{12} - p_{23} - p_{24} + p_{123} + p_{124} + p_{234},\ p_3 - p_{13} - p_{23} - p_{34} + p_{123} + p_{134} + p_{234},\ p_{123},\ p_4 - p_{14} - p_{24} - p_{34} + p_{124} + p_{134} + p_{234},\ p_{124},\ p_{134},\ p_{234},\ \}$. As mentioned earlier, these results for $n = 4$ (and for $n > 4$) are new.

3. Demonstration

Here I demonstrate the multinomial sampling approach to simulate vectors of length 4 and 8, for assumed values of the marginal means p_i and pairwise correlations ρ_{jk}.

3.1. *Vectors of length* 4

For this brief demonstration, I consider a model that could be used in a cross-sectional study to compare the probability of positive response between two equally sized groups (treatment A versus B) of clusters of size four. I assume that the marginal means $E(Y_{ij}) = p_{ij}$ can be expressed as $p_{ij} = p_i$, where

$$\text{logit}\,(p_i) = \beta_0 x_{i1} + \beta_1 x_{i2}; \tag{26}$$

$x_{i1} = 1$ and x_{i2} is an indicator variable for treatment A, which equals 1 for the treatment A group ($i = 1, \ldots, m/2$) and 0 for treatment B ($i = m/2 + 1, \ldots, m$). For this model, $\text{logit}\,(p_1) = \beta_0$ for subjects receiving treatment A and $\text{logit}\,(p_2) = \beta_0 + \beta_1$ for subjects receiving treatment B. For specified (p_1,p_2), $(\beta_0, \beta_1) = (logit(p_1), logit(p_2)$ - $logit(p_1))$.

I also assume that the pairwise correlations are equal within clusters, which is a plausible (and common) assumption for clustered data in cross-sectional studies. The correlation matrix for each cluster is therefore a 4×4 exchangeable (equicorrelated) structure, with parameter α that must take value in $(-1/(n-1), 1) = (-1/3, 1)$ to yield a positive definite correlation matrix. The constraints (5) necessary to achieve valid bivariate Bernoulli distributions are $(-0.2195, 1)$, so that α must take value in $(-0.2195, 1)$ for this example.

I assessed the power to test the hypothesis that the probability of response is equal between the two groups, when GEE was implemented in Stata 14.0 with a two-sided test of the hypothesis that β_1 is 0. The power with type-one error of 0.05 was estimated as the proportion of 10000 simulation runs that resulted in a p-value less than 0.05 based on Wald's test, with application of a "sandwich" covariance matrix for estimation of the covariance matrix of $\hat{\beta}$. For this demonstration, my goal was to determine the sample size necessary to achieve 80 % power to detect a difference between 25 % versus 15 % response in the low versus high income groups, respectively.

Table 1 displays the values of $(Z_4(i-1), Z_4(i))$ $(i = 1, \ldots, 16)$ used to simulate dependent Bernoulli random variables $\underline{Y}_4$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2, when $\alpha = .80$. (Tables for $\alpha = 0$ and $\alpha = 0.40$ are provided in Table C1 and Table C2 in Appendix C.) I obtained the values in Table 1 using the expressions for P_{rob} that are provided in Section 2.2.4. The values for p_{jk} (j,$k \in \{1, 2, 3, 4\}$; $j \neq k$) were determined by the assumed values for p_1, p_2, and α. (See Section 2.2.2.) I then specified the values of p_{ijk} as $p_{ijk} = L_{ijk} + w\,(U_{ijk} - L_{ijk})$ (i, j,$k \in \{1, 2, 3, 4\}$; $i \neq j$; $j \neq k$; $i \neq k$), for $w = 0.90$ and (L_{ijk}, U_{ijk}) as defined in Section 2.2.3. Finally, I specified $p_{1234} = L_{1234} + w\,(U_{1234} - L_{1234})$, for $w = 0.90$ and (L_{1234}, U_{1234}) as defined in Section 2.2.4.

Table 2 displays the simulation results. Table 2 indicates that when $\alpha = 0$, a sample size of 240 per group ($m/2 = 60$ clusters of size 4 per group) will achieve approximately 79 % power to detect a difference between $p_1 = 0.15$ and $p_2 = 0.25$ in a GEE analysis. This is similar to the 78 % power that would be achieved for a two-group Chi-Square test with a 0.05 two-sided significance level (per nQuery Advisor statistical software). As the value of α increases, the sample size required to achieve approximately 80 % power increases. For example, for $\alpha = 0.80$, a sample size of 820 per group ($m/2 = 210$ clusters of size 4 per group) is required to achieve 80 % power. The results were very similar for $w = 0.30$ and $w = 0.60$.

Table 1. Values to simulate dependent Bernoulli random variables $\underline{Y}_4$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2, for $p_1 = .15$, $p_2 = .25$, $\alpha = .80$, and $w = 0.90$. Simulation results are provided in Table 2.

				Group 1				Group 2		
Y_1	Y_2	Y_3	Y_4	$i-1$[a]	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$
0	0	0	0	0	.780895	0	.780895	.648375	0	.648375
1	0	0	0	1	.020655	.780895	.80155	.030375	.648375	.67875
0	1	0	0	2	.020655	.80155	.822205	.030375	.67875	.709125
1	1	0	0	3	.002295	.822205	.8245	.003375	.709125	.7125
0	0	1	0	4	.020655	.8245	.845155	.030375	.7125	.742875
1	0	1	0	5	.002295	.845155	.84745	.003375	.742875	.74625
0	1	1	0	6	.002295	.84745	.849745	.003375	.74625	.749625
1	1	1	0	7	.000255	.849745	.85	.000375	.749625	.75
0	0	0	1	8	.020655	.85	.870655	.030375	.75	.780375
1	0	0	1	9	.002295	.870655	.87295	.003375	.780375	.78375
0	1	0	1	10	.002295	.87295	.875245	.003375	.78375	.787125
1	1	0	1	11	.000255	.875245	.8755	.000375	.787125	.7875
0	0	1	1	12	.002295	.8755	.877795	.003375	.7875	.790875
1	0	1	1	13	.000255	.877795	.87805	.000375	.790875	.79125
0	1	1	1	14	.000255	.87805	.878305	.000375	.79125	.791625
1	1	1	1	15	.121695	.878305	1	.208375	.791625	1

[a] Each realization of $(Y_1, Y_2, Y_3, Y_4) = B_4(i) =$ the binary representation of $i-1$.

Table 2. Simulated power (Pow) in the test of equality of the probability of response (p_1 versus p_2) between equally sized groups ($m/2$) of clusters of size 4, for varying levels of α, when $p_1 = .15$, $p_2 = .25$ and $w = 0.90$.

p_1	p_2	OR	α	w	%[a]	$m/2$	Pow
.15	.25	1.89	0	.9	100	30	.53
.15	.25	1.89	.4	.9	100	30	.27
.15	.25	1.89	.8	.9	100	30	.18
.15	.25	1.89	0	.9	100	60	.79
.15	.25	1.89	.4	.9	100	60	.45
.15	.25	1.89	.8	.9	100	60	.31
.15	.25	1.89	0	.9	100	90	.92
.15	.25	1.89	.4	.9	100	90	.61
.15	.25	1.89	.8	.9	100	90	.44
.15	.25	1.89	0	.9	100	120	.97
.15	.25	1.89	.4	.9	100	120	.74
.15	.25	1.89	.8	.9	100	120	.55
.15	.25	1.89	0	.9	100	150	.99
.15	.25	1.89	.4	.9	100	150	.84
.15	.25	1.89	.8	.9	100	150	.65
.15	.25	1.89	0	.9	100	180	1
.15	.25	1.89	.4	.9	100	180	.89
.15	.25	1.89	.8	.9	100	180	.73
.15	.25	1.89	0	.9	100	210	1
.15	.25	1.89	.4	.9	100	210	.93
.15	.25	1.89	.8	.9	100	210	.8
.15	.25	1.89	0	.9	100	240	1
.15	.25	1.89	.4	.9	100	240	.96
.15	.25	1.89	.8	.9	100	240	.85

[a] The percentage of simulation runs that resulted in convergence for GEE.

3.2. *Vectors of length* 8

In [17, 20, 21] we used simulations to demonstrate that a severe violation of the Prentice constraints could be used as a rule-out criterion when assessing working correlation structures for GEE or QLS analysis of longitudinal binary data. As is appropriate for a short demonstration, here I replicate some of the simulations. The results are similar to our earlier findings, but for slightly different assumptions.

I consider a model that could be used to compare the change over time in an 8 study period in the probability of positive response between two equally sized groups of patients who had been randomized to one of two treatment groups (A versus B). I assume that the marginal means $\mathrm{E}(Y_{ij}) = p_{ij}$ satisfy

$$\text{logit}\,(p_{ij}) = \beta_0 x_{i1} + \beta_1 x_{i2} + \beta_1 x_{i3} + \beta_3 x_{i4}, \tag{27}$$

where $x_{i1} = 1$; $x_{i2} \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ indicates month of measurement; x_{i3} is an indicator variable that equals 1 for treatment group A ($i = 1, \ldots, m/2$); and x_{i4} is an interaction term that was constructed as the product of x_{i2} and x_{i3}. If the regression coefficient β_3 differs significantly from zero, this indicates that the change over time in the probability of response differs significantly between the two treatment groups. I assumed that $(\beta_0, \beta_1, \beta_2, \beta_3)' = (0, -.1 - .01.13)$, which yields probabilities that increase linearly (in the logit) from 0.512 and 0.505 at month 1 to 0.858 and 0.703 at month 8, in groups A and B, respectively. I considered group sizes 15, 30, 60, and 120.

For this model, I assumed that the true correlation structure is first-order autoregressive [AR(1)], so that $Corr(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$. The AR(1) structure is plausible when we expect measurements on a subject to be more similar, and therefore to be more highly correlated, when they are measured more closely together in time. The correlation matrix for each subject is therefore an 8×8 AR(1) structure, with parameter α that must take value in $(-1, 1)$ to yield a positive definite correlation matrix.

For the AR(1) structure and, more generally, for any product correlation structure of the form $Corr(Y_{ij}, Y_{ij+w}) = \prod_{z=1}^{w-1} Corr(Y_{iz}, Y_{iz+1})$, satisfaction of the Prentice constraints (5) for adjacent Y_{ij}, Y_{ij+1} ($j = 1, \ldots, n_i - 1$) is sufficient to ensure satisfaction of the Prentice constraints for all Y_{ij}, Y_{ik} [20, Theorem 7.1, p.146]. The Prentice constraints (5) are $(-0.449, 0.942)$ and $(-0.187, 0.882)$ for groups A and B respectively, so that α must take value in $(-0.187, 0.882)$ for this example.

Previous authors simulated longer length binary vectors with an AR(1) structure, by assuming first-order antedependence and linearity of the expectations of the conditional distributions. They simulated from the distribution of Y_1 and then from the distribution of Y_j given Y_{j-1} ($j = 1, \ldots, n$) [3, 6, 22–24]. Also see [25, p.14] for a very brief review. [18, p. 13] also assumed the first-order Markov property and linearity of the conditional expectations, but fully specified the distribution of all length n possible permutations of zeros and ones, and used multinomial sampling [7] for the simulations.

For this short demonstration, I compared results for the distribution that stems from assumed first-order antedependence and linear conditional expectations (MARK1) (so that the true correlation structure is AR(1)), with a different distribution that has the same specified marginal means and AR(1) structure. For assumed p_{ij} and α, I first used [24, (5)] to obtain the MARK1 distribution for each group. Next, to modify the MARK1 distributions, I first used the explicit expressions that are provided in the Supplemental Material (see Section 5) to obtain the constraints $(L_{12345678}, U_{12345678})$ for $p_{12345678}$. I then changed the value of $p_{12345678}$ to $L_{12345678} + .99(U_{12345678} - L_{12345678})$. I then used the explicit expressions for the $P_8(i)$ ($i = 1, \ldots, 256$) that are provided in the Supplemental Material to obtain a new distribution based on the modified value of $p_{12345678}$. Tables D1 and D2 provide a partial listing $i = 1, \ldots, 50$ of the distributions for $\alpha = 0.20$. The complete distributions are provided in Supplemental Material (see Section 5). In each of 1000 simulation runs, I used this approach to simulate data for model (27), for a true AR(1) structure with $\alpha = -0.10, 0, 0.20, 0.40, 0.80, 0.88$ and group sizes 15, 30, 60, and 120. The results were almost identical for the MARK1 and non-MARK1 distributions, so only the figures for the non-MARK1 distributions are shown here.

In each simulation run I correctly specified model (27) and the correct AR(1) correlation structure in a QLS analysis. In addition, I correctly specified model (27), but incorrectly specified the true AR(1) structure as exchangeable. The QLS models converged in 100 % of simulation runs for the AR(1) and exchangeable working structures, except for ($\alpha = 0.8$ and group-size 15) and ($\alpha = 0.88$ and group-sizes 15 or 30); the % convergence was 99.8 for each of these scenarios.

Figure 1 displays box-plots of the QLS estimates $\widehat{\alpha}$ along with their estimated Prentice constraints (5) versus the true values of α. To improve readability, the Prentice lower and upper values of the constraints are plotted slightly to the left and right, respectively, of the true value of α. Figure 3.2 shows what we have observed earlier [17, 20, 21], that violation of the Prentice constraints is unlikely unless the true value of α is very close to the upper or lower boundary value for the Prentice constraint. Furthermore, the *severity* of violation decreases with increasing group size.

To allow for more carefully assessment of the concept of *severity* of violation of the Prentice constraints, figure 2 displays box-plots of $\hat{\alpha}$ and the estimated upper constraint for $\alpha = 0.80$ and 0.88. These graphs show that as the sample sizes increase, the estimates of the correlation and of upper Prentice constraint move closer to their true values. As a result, the distance between $\widehat{\alpha}$ and the estimated upper value of the Prentice constraint decreases as the group sizes increase. The average distance for simulation runs that resulted in a violation of constraints (number of simulation runs that resulted in violation) was 0.027 (503), 0.018 (497), 0.013 (463), 0.008 (401) when $\alpha = 0.88$, for group sizes 15, 30,60, and 120, respectively. Although it was not shown on the graphs, the same behavior was observed at the lower constraints. The average distance for simulation runs that resulted in a violation of constraints (number of simulation runs that resulted in violation) was 0.044 (185), 0.027 (78), 0.014 (27), 0.006 (5) when α = -0.10, for group sizes 15, 30,60, and 120, respectively. What we observed earlier for MARK1 distributions therefore also holds true for this non-MARK1 distrbution. Severe violation of the constraints is unlikely for a moderate sample size, when the true correlation structure is correctly specified as AR(1).

Figure 3 then displays the same triplets (estimated lower constraint, $\widehat{\alpha}$, estimated upper constraint) versus the true value of α, when the working correlation structure is misspecified as exchangeable. Horizontal solid lines are again displayed at the true Prentice constraints, but horizontal dashed lines are only displayed at the limiting values of $\hat{\alpha}$, when the limiting values exceed the Prentice constraints. The limiting value of $\hat{\alpha}$ when the true AR(1) correlation structure is misspecified as exchangeable was provided earlier; for example see [21, (9) and (10) of Online Appendix C]. Figure 3 shows that a severe violation of constraints is likely to occur for larger values of α when the true AR(1) structure is misspecified as exchangeable.

To allow for more careful assessment of the severity of violation of constraints, figure 4 then displays box-plots of $\hat{\alpha}$ and the estimated upper constraint for $\alpha = 0.80$ and 0.88. These graphs show that as the group sizes increase, the severity of violation of constraints increases. The average distance for simulation runs that resulted in a violation of constraints (number of simulation runs that resulted in violation) was 0.027 (503), 0.018 (497), 0.013 (463), 0.008 (401) when $\alpha = 0.88$, for group sizes 15, 30,60, and 120, respectively. What we observed earlier for MARK1 distributions therefore again holds true for this non-MARK1 distrbution. Severe violation of the constraints is likely for larger values of the correlation, and for increasing sample sizes, when the true AR(1) structure is misspecified as exchangeable. The rule-out criterion [17] could be applied in situations like these, to remove working correlation structure from consideration, when its application results in a severe violation of constraints for QLS or GEE.

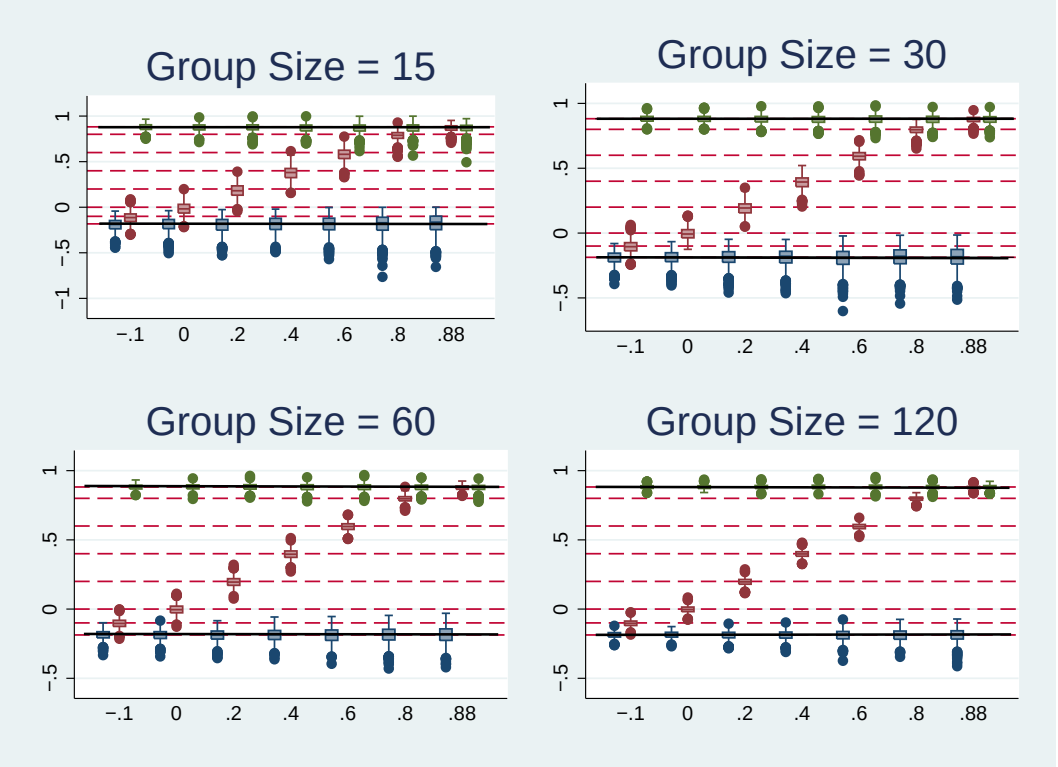

Figure 1. Box-plots of the estimated correlations $\widehat{\alpha}$ along with their estimated Prentice constraints versus the true values of α. The Prentice lower and upper values of the constraints are plotted slightly to the left and right, respectively, of the true value of α. The estimates were obtained in a QLS analysis, when model 27 and the AR(1) structure were correctly specified. Solid horizontal lines are displayed at the true upper and lower values of the Prentice constraints, and dashed horizontal lines are displayed at $-0.10, 0, 0.20, 0.40, 0.80, 0.88$, the true values of α. Looking from left to right within each graph, we see the triplet (estimated lower constraint,$\widehat{\alpha}$, estimated upper constraint) plotted above each true value of α. The figures show that as the group sizes increase, the estimates become closer to their true values.

4. Discussion

Simulation via multinomial sampling requires the user to specify the full likelihood of the length n correlated binary variables. Specifying the likelihood is helpful to ensure the existence of a valid parent distribution, to avoid unknowingly specifying impossible parameter values in a sample size formula, or incorrectly basing results of a GEE or QLS analysis on estimates that are not compatible with any valid parent distribution. However, the multinomial sampling approach was only implemented in previous papers for vectors of length 2 or 3, unless simplifying assumptions such as assuming the first-order Markov property were made first. I therefore proved a theorem that can be used to obtain expressions for the probabilities of all possible length n permutations of zeros and ones. I also provided an algorithm that can be easily implemented in a program such as MATLAB to obtain the constraints that must be satisfied by the probability that $Y_1, \ldots, Y_n$ all take value one. I implemented my approach to obtain explicit expressions of the probabilities and constraints for $n = 2, \ldots, 12$; these results are provided in Supplemental Material and will be updated to include results for larger values of n.

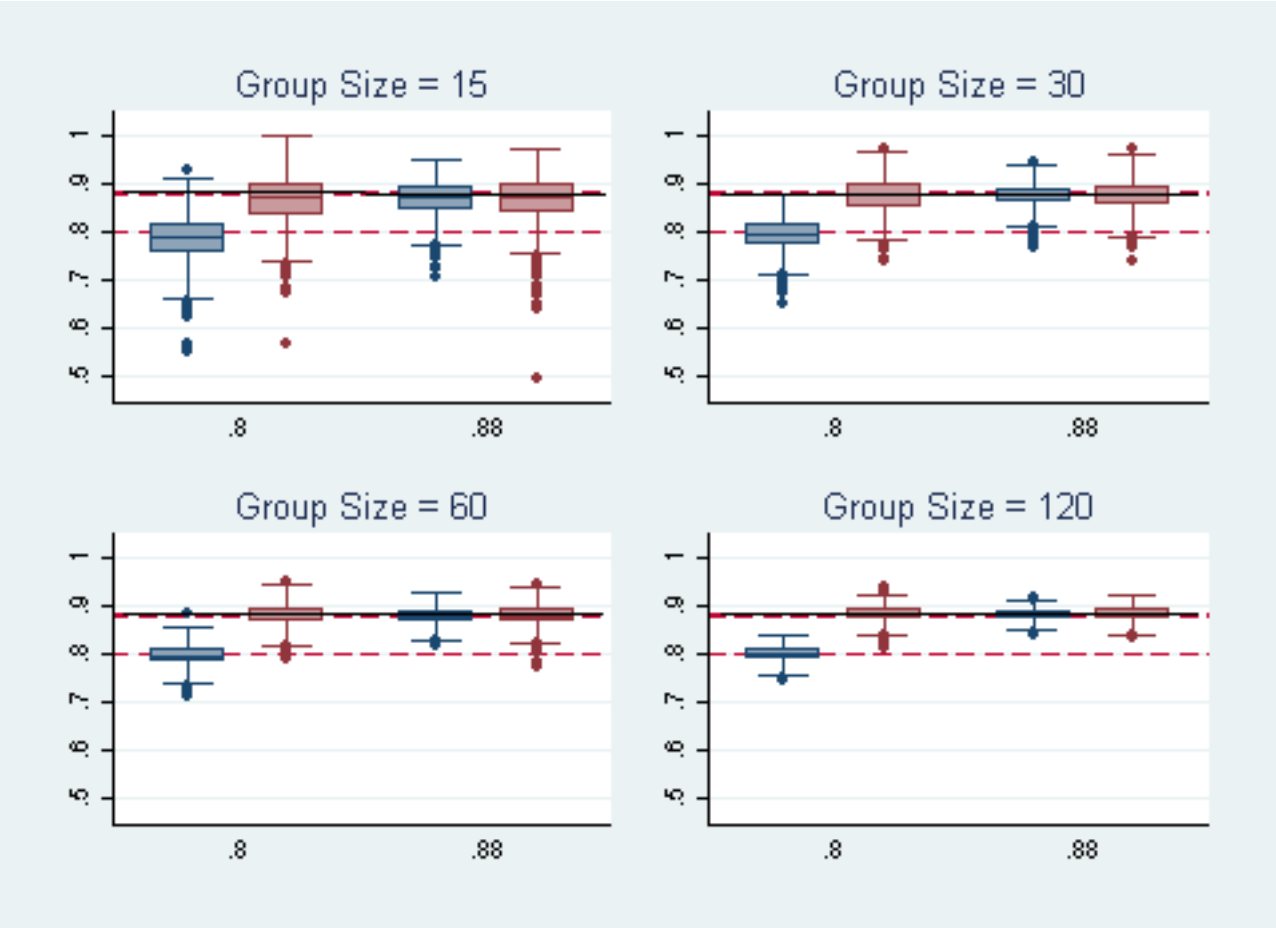

Figure 2. Box-plots of the estimated correlations $\widehat{\alpha}$ along with their estimated **upper** Prentice constraints versus the true values of α, for $\alpha = 0.80$ and 0.88. These values were also displayed in Figure 1 but are shown here to allow for greater examination of the violation of the Prentice constraints, which tends to occur when α is close to a boundary value. Looking from left to right within each graph, we see the pair ($\widehat{\alpha}$, estimated upper constraint) plotted above each true value of α, for $\alpha = 0.80$ and α= 0.88.

5. Supplemental material

(Supplemental files are available under Downloadable Files at https://dbe.med.upenn.edu/biostat-research/JustineShults .)

The supplemental materials include the following:

(1) Files that contain explicit expressions for the probabilities $P_{rob}(i)$ $(i = 1, \ldots, 2^n)$ and lower level constraints $L_{1\ldots n}$ and upper level constraints $U_{1\ldots n}$. For each n, three files were written, to organize the expressions and facilitate their use in programs. The files Probn.txt, Lowern.txt,Uppern.txt provide expressions for the probabilities, lower, and upper values for the constraints $(n = 2, \ldots, 13)$.

(2) Full listing of the values (for $i - 1 = 0$ through 255) to simulate dependent Bernoulli random variables $\underline{Y}_8$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2 for $\alpha = -0.10, 0.20, 0.40, 0.60, 0.80, 0.88$. The listings are provided in the files DemonstrationN8w.txt where w $= -1, 0, 20, 40, 60, 80, 88$ for $\alpha = -.10, 0, 0.20, 0.40, 0.60, 0.88$, respectively. Corresponding files for distributions obtained under an assumption of the first order Markov property are named DemonstrationN8FM1w.txt, with the same definitions for w.

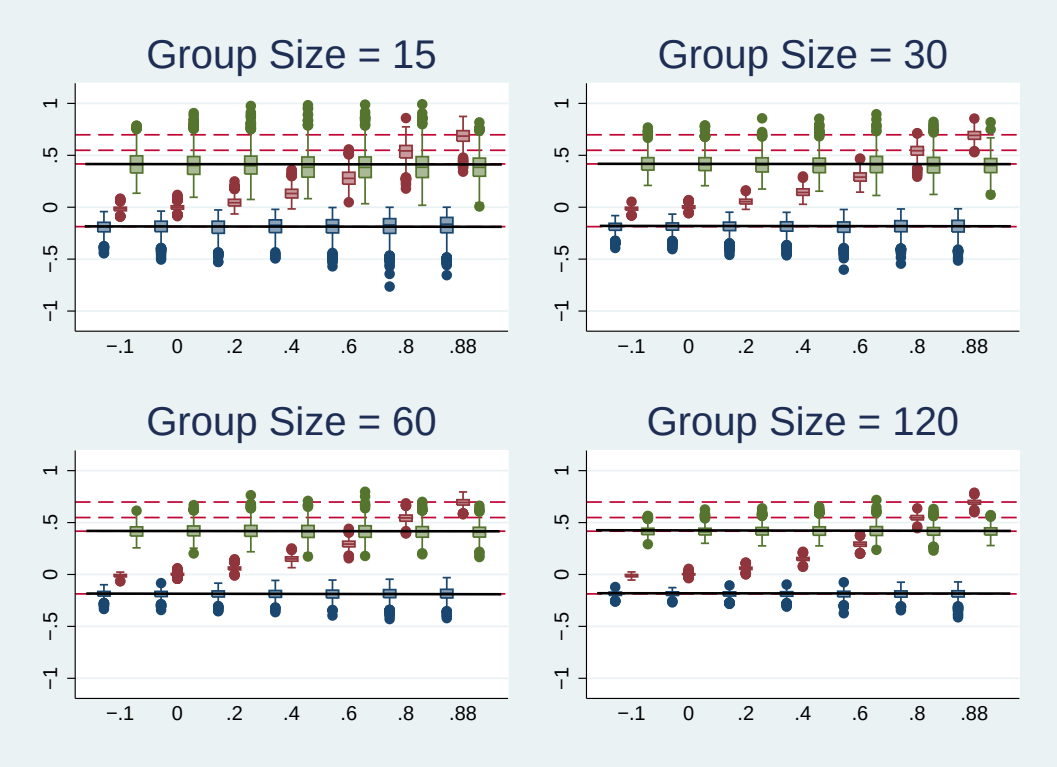

Figure 3. Box-plots of the estimated correlations $\widehat{\alpha}$ along with their estimated Prentice constraints versus the true values of α. The Prentice lower and upper values of the constraints are plotted slightly to the left and right, respectively, of the true value of α. The estimates were obtained in a QLS analysis, when model (27) was correctly specified, but the model for the correlation was incorrect (exchangeable). Solid horizontal lines are displayed at the true upper and lower values of the Prentice constraints, and dashed horizontal lines are displayed at the limiting values of the $\widehat{\alpha}$ when the limiting values exceed the Prentice constraints. Looking from left to right within each graph, we see the triplet (estimated lower constraint,$\widehat{\alpha}$, estimated upper constraint) plotted above each true value of α. The figures show that as the group sizes increase, the estimates become closer to their true values. This means that for increasing group sizes, we tend to have more severe violation of constraints for larger values of α, when the limiting values of $\widehat{\alpha}$ violate the Prentice constraints.

References

[1] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.

[2] Farrell PJ, Rogers-Stewart K. Methods for generating longitudinally correlated binary data. Int Stat Rev. 2008; 76: 28–38.

[3] Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika. 2003; 90: 455–63.

[4] Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. Am Stat. 1991; 45: 302–304.

[5] Al-Osh MA, Lee SJ. A simple approach for generating correlated binary variables. J Stat Comput Sim. 2001; 70: 231-255.

[6] Preisser JS, Qaqish BF. A comparison of methods for simulating correlated binary variables with specified marginal means and correlations. J Stat Comput Sim. 2014; 84 (11): 2441-2452.

[7] Kang SH, Jung SH. Generating correlated binary variables with complete specification of the joint distribution. Biometrical J. 2001; 43 (3): 1521-4036. Available from: http://www.researchgate.net/profile/Seung_Ho_Kang2/publication/229675001_

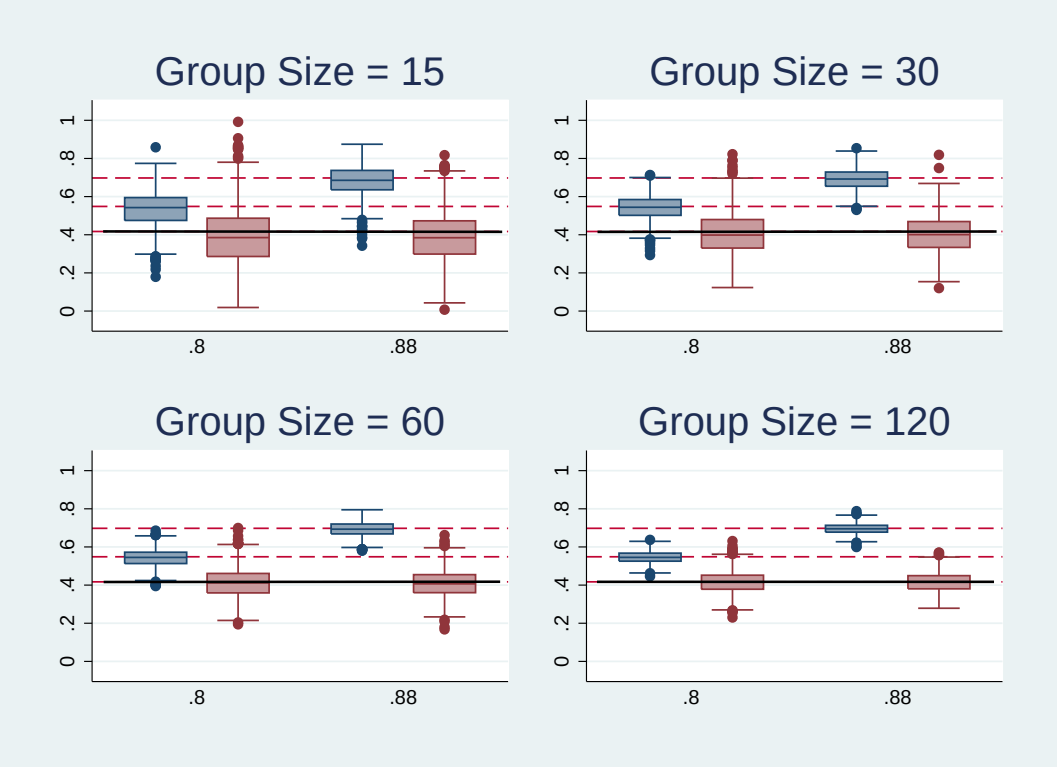

Figure 4. Box-plots of the estimated correlations $\widehat{\alpha}$ along with their estimated **upper** Prentice constraints versus the true values of α, for $\alpha = 0.80$ and 0.88. These values were also displayed in Figure 3 but are shown here to allow for greater examination of the violation of the Prentice constraints, which can be severe for larger values of α, when the true AR(1) structure is misspecified as exchangeable. Looking from left to right within each graph, we see the pair ($\widehat{\alpha}$, estimated upper constraint) plotted above each true value of α, for $\alpha = 0.80$ and α= 0.88.

Generating_Correlated_Binary_Variables_with_Complete_Specification_of_the_Joint_Distribution/links/53d306370cf2a7fbb2e9cb62.pdf

[8] Hamdan MA, Nasro MO. Maximum likelihood estimation of the parameters of the bivariate binomial distribution. Commun Stat Theory. 1986; 15(3): 747-754.

[9] Haynes ME, Sabo RT, Chaganty NR. Simulating dependent binary variables through multinomial sampling. J Stat Comput Sim. 2015; 0(0): 1-14.

[10] Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. Biometrika. 2006; 93(1): 197206.

[11] Molenberghs G, Kenward MG. Semi-parametric marginal models for hierarchical data and their corresponding full models. Comput Stat Data An. 2010; 54: 585–597.

[12] Chaganty NR. An alternative approach to the analysis of longi- tudinal data via generalized estimating equations. J Statist Plann Inference. 1997; 63: 39–54.

[13] Shults J, Chaganty NR. Analysis of serially correlated data using quasi-least squares. Biometrics. 1998; 54: 1622–1630.

[14] Chaganty NR, Shults J. On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. J Statist Plann Inference. 1999; 76: 127–144.

[15] Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Stat Med. 1998; 17: 1643 -1658.

[16] Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of longitudinal data (2nd ed.). 2002; Oxford: Oxford University Press.

[17] Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T. A comparison of several approaches for choosing between working correlation structures in generalized esti-

mating equation analysis of longitudinal binary data. Stat Med. 2009; 28: 2338-2355.
[18] Shults J, Sun W, Tu X, Amsterdam J. On the violation of bounds for the correlation in generalized estimating equation analysis of binary data from longitudinal trials. UPenn Biostatistics Working Papers. Working Paper 8. 2006. Available from: http://biostats.bepress.com/upennbiostat/art8
[19] Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics. 1988; 44: 1033–1048.
[20] Shults J, Hilbe J. Quasi-least squares regression. New York (NY): Chapman and Hall: CRC Monographs on Statistics and Applied Probability; 2014.
[21] Shults J, Guerra M W. (2014), A note on implementation of decaying product correlation structures for quasi-least squares. Stat Med. 2014; 33: 33983404.
[22] Jung SH, Ahn WW. Sample size for a two-group comparison of repeated binary measurements using GEE. Stat Med. 2005; 24: 2583–2596.
[23] Guerra MW, Shults J. A note on the simulation of overdispersed random variables with specified marginal means and product correlations. Am Stat. 2014; 68(2):104–107.
[24] Guerra MW, Shults J, Amsterdam J, Ten-Have T. The analysis of binary longitudinal data with time-dependent covariates. Stat Med. 2012; 31(10):931-948.
[25] Guerra M, Shults J. On the simulation of longitudinal discrete data with specified marginal means and first-order antedependence. UPenn Biostatistics Working Papers. Working Paper 35. 2013. Available from: http://biostats.bepress.com/upennbiostat/art35
99

Appendix A. Prior Work for $n = 3$:

As mentioned earlier, the multinomial sampling approach has only been implemented without simplifying assumptions, for $n \leq 3$. Here I describe the example provided in [7] for $n = 3$. [7] assumed $(p_1, p_2, p_3) = (0.9,0.7,0.5,0.2)$ and $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.2, 0.3, 0.4)$. Noting that specified values for the marginal means and pairwise correlations are not sufficient to determine a particular distribution for $n = 3$, [7] assumed $P_3(1) = 0.055$ and reported that they obtained a solution with Mathematica. Their code was not printed with the paper and is no longer available (per personal correspondence with the authors). However, we can obtain their reported probability density function by first using the bivariate Bernoulli distribution to obtain $p_{12} = 0.65749545$, $p_{13} = 0.495$, and $p_{23} =$ Then, using (23), $p_{123} = p_{12}$ - p_2 - p_3 - p_1 + p_{13} + p_{23} + 1 - $P_3(1) = 0.43914697$. Next, substituting the assumed values for $(p_1, p_2, p_3, p_{12}, p_{13}, p_{23}, p_{123})$ into (23) yields the following values that agree, after rounding, with those provided in [7, Table 1]:

$$P_3(1) = 0.055 \quad P_3(2) = 0.186652 \quad P_3(3) = 0.04 \quad P_3(4) = 0.218348$$
$$P_3(5) = 0.002495 \quad P_3(6) = 0.055853 \quad P_3(7) = 0.002505 \quad P_3(8) = 0.439147$$

Note that there was a typographical error in [7, Table 1] because the authors wrote $(p_1, p_2, p_3) = (0.1, 0.3, 0.5)$ which is actually $(1 - p_1, 1 - p_2, 1 - p_3)$ for this example.

[7] presumably chose a value for $P_3(1)$ that yields a valid probability density function (all non-negative probabilities) by trial and error. In contrast, [9] provided expressions for $P_3(i)$ $(i = 1, \ldots, 8)$ and suggested selecting a value for p_{123} that is the midpoint of its constraints, although they noted that any value within the constraints could be used. For this example, the midpoint of $(L_{123}, U_{123}) = (0.43665151, 0.44165151)$ is 0.43915151, which agrees to five digits with the value for $P_3(8) = P_{123}$ that is provided in the table above. Selecting the midpoint therefore yields a distribution that agrees, after rounding, with the one provided in [7, Table 1, p. 266].

Appendix B. Results for $n = 5$:

$$X = \left(X_1', X_2', X_3', X_4', X_5', \underline{One}_{32}'\right)' =$$

$$\begin{pmatrix}
0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1\\
0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1&1\\
0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1\\
0&0&0&0&0&1&0&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0\\
0&0&0&0&0&0&1&1&0&0&0&0&1&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&1&1&0&0&1&1&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
\end{pmatrix}$$

and $E = (\ p_1, p_2, p_3, p_4, p_5, p_{12}, p_{13}, p_{14}, p_{15}, p_{23}, p_{24}, p_{25}, p_{34}, p_{35}, p_{45}, p_{123}, p_{124}, p_{125}, p_{134}, p_{135}, p_{145}, p_{234}, p_{235}, p_{245}, p_{345}, p_{1234}, p_{1235}, p_{1245}, p_{1345}, p_{2345}, p_{12345}, 1\ )'$. Then, $P_{rob} = (P_5(1), \ldots P_5(32))' = (\ p_{12} - p_2 - p_3 - p_4 - p_5 - p_1 + p_{13} + p_{14} + p_{15} + p_{23} + p_{24} + p_{25} + p_{34} + p_{35} + p_{45} - p_{123} - p_{124} - p_{125} - p_{134} - p_{135} - p_{145} - p_{234} - p_{235} - p_{245} - p_{345} + p_{1234} + p_{1235} + p_{1245} + p_{1345} + p_{2345} - p_{12345} + 1, p_1 - p_{12} - p_{13} - p_{14} - p_{15} + p_{123} + p_{124} + p_{125} + p_{134} + p_{135} + p_{145} - p_{1234} - p_{1235} - p_{1245} - p_{1345} + p_{12345}, p_2 - p_{12} - p_{23} - p_{24} - p_{25} + p_{123} + p_{124} + p_{125} + p_{234} + p_{235} + p_{245} - p_{1234} - p_{1235} - p_{1245} - p_{2345} + p_{12345}, p_{12} - p_{123} - p_{124} - p_{125} + p_{1234} + p_{1235} + p_{1245} - p_{12345}, p_3 - p_{13} - p_{23} - p_{34} - p_{35} + p_{123} + p_{134} + p_{135} + p_{234} + p_{235} + p_{345} - p_{1234} - p_{1235} - p_{1345} - p_{2345} + p_{12345}, p_{13} - p_{123} - p_{134} - p_{135} + p_{1234} + p_{1235} + p_{1345} - p_{12345}, p_{23} - p_{123} - p_{234} - p_{235} + p_{1234} + p_{1235} + p_{2345} - p_{12345}, p_{123} - p_{1234} - p_{1235} + p_{12345}, p_4 - p_{14} - p_{24} - p_{34} - p_{45} + p_{124} + p_{134} + p_{145} + p_{234} + p_{245} + p_{345} - p_{1234} - p_{1245} - p_{1345} - p_{2345} + p_{12345}, p_{14} - p_{124} - p_{134} - p_{145} + p_{1234} + p_{1245} + p_{1345} - p_{12345}, p_{24} - p_{124} - p_{234} - p_{245} + p_{1234} + p_{1245} + p_{2345} - p_{12345}, p_{124} - p_{1234} - p_{1245} + p_{12345}, p_{34} - p_{134} - p_{234} - p_{345} + p_{1234} + p_{1345} + p_{2345} - p_{12345}, p_{134} - p_{1234} - p_{1345} + p_{12345}, p_{234} - p_{1234} - p_{2345} + p_{12345}, p_{1234} - p_{12345}, p_5 - p_{15} - p_{25} - p_{35} - p_{45} + p_{125} + p_{135} + p_{145} + p_{235} + p_{245} + p_{345} - p_{1235} - p_{1245} - p_{1345} - p_{2345} + p_{12345}, p_{15} - p_{125} - p_{135} - p_{145} + p_{1235} + p_{1245} + p_{1345} - p_{12345}, p_{25} - p_{125} - p_{235} - p_{245} + p_{1235} + p_{1245} + p_{2345} - p_{12345}, p_{125} - p_{1235} - p_{1245} + p_{12345}, p_{35} - p_{135} - p_{235} - p_{345} + p_{1235} + p_{1345} + p_{2345} - p_{12345}, p_{135} - p_{1235} - p_{1345} + p_{12345}, p_{235} - p_{1235} - p_{2345} + p_{12345}, p_{1235} - p_{12345}, p_{45} - p_{145} - p_{245} - p_{345} + p_{1245} + p_{1345} + p_{2345} - p_{12345}, p_{145} - p_{1245} - p_{1345} + p_{12345}, p_{245} - p_{1245} - p_{2345} + p_{12345}, p_{1245} - p_{12345}, p_{345} - p_{1345} - p_{2345} + p_{12345}, p_{1345} - p_{12345}, p_{2345} - p_{12345}, p_{12345}\ \}$. Solving $0 \leq P_5(i)$ for p_{12345} $(i = 1, \ldots, 32)$ yields the constraints

$$L_{12345} \leq p_{12345} \leq U_{12345}, \tag{B1}$$

where $L_{12345} = \max\{\ p_{12} - p_1 + p_{13} + p_{14} + p_{15} - p_{123} - p_{124} - p_{125} - p_{134} - p_{135} - p_{145} + p_{1234} + p_{1235} + p_{1245} + p_{1345}, p_{12} - p_2 + p_{23} + p_{24} + p_{25} - p_{123} - p_{124} - p_{125} - p_{234} - p_{235}$

Table C1. Values to simulate dependent Bernoulli random variables $\underline{Y}_4$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2, for $p_1 = .15$, $p_2 = .25$, $\alpha = 0$, and $w = 0.90$. Simulation results are provided in Table 2.

				Group 1				Group 2		
Y_1	Y_2	Y_3	Y_4	$i-1$[a]	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$
0	0	0	0	0	.474025	0	.474025	.205625	0	.205625
1	0	0	0	1	.123225	.474025	.59725	.175625	.205625	.38125
0	1	0	0	2	.123225	.59725	.720475	.175625	.38125	.556875
1	1	0	0	3	.002025	.720475	.7225	.005625	.556875	.5625
0	0	1	0	4	.123225	.7225	.845725	.175625	.5625	.738125
1	0	1	0	5	.002025	.845725	.84775	.005625	.738125	.74375
0	1	1	0	6	.002025	.84775	.849775	.005625	.74375	.749375
1	1	1	0	7	.000225	.849775	.85	.000625	.749375	.75
0	0	0	1	8	.123225	.85	.973225	.175625	.75	.925625
1	0	0	1	9	.002025	.973225	.97525	.005625	.925625	.93125
0	1	0	1	10	.002025	.97525	.977275	.005625	.93125	.936875
1	1	0	1	11	.000225	.977275	.9775	.000625	.936875	.9375
0	0	1	1	12	.002025	.9775	.979525	.005625	.9375	.943125
1	0	1	1	13	.000225	.979525	.97975	.000625	.943125	.94375
0	1	1	1	14	.000225	.97975	.979975	.000625	.94375	.944375
1	1	1	1	15	.020025	.979975	1	.055625	.944375	1

[a] Each realization of $(Y_1, Y_2, Y_3, Y_4) = B_4(i) =$ the binary representation of $i-1$.

- p_{245} + p_{1234} + p_{1235} + p_{1245} + p_{2345}, p_{13} - p_3 + p_{23} + p_{34} + p_{35} - p_{123} - p_{134} - p_{135} - p_{234} - p_{235} - p_{345} + p_{1234} + p_{1235} + p_{1345} + p_{2345}, p_{1234} - p_{123} + p_{1235}, p_{14} - p_4 + p_{24} + p_{34} + p_{45} - p_{124} - p_{134} - p_{145} - p_{234} - p_{245} - p_{345} + p_{1234} + p_{1245} + p_{1345} + p_{2345}, p_{1234} - p_{124} + p_{1245}, p_{1234} - p_{134} + p_{1345}, p_{1234} - p_{234} + p_{2345}, p_{15} - p_5 + p_{25} + p_{35} + p_{45} - p_{125} - p_{135} - p_{145} - p_{235} - p_{245} - p_{345} + p_{1235} + p_{1245} + p_{1345} + p_{2345}, p_{1235} - p_{125} + p_{1245}, p_{1235} - p_{135} + p_{1345}, p_{1235} - p_{235} + p_{2345}, p_{1245} - p_{145} + p_{1345}, p_{1245} - p_{245} + p_{2345}, p_{1345} - p_{345} + p_{2345}, 0 } and $U_{12345} = \min\{$ p_{12} - p_2 - p_3 - p_4 - p_5 - p_1 + p_{13} + p_{14} + p_{15} + p_{23} + p_{24} + p_{25} + p_{34} + p_{35} + p_{45} - p_{123} - p_{124} - p_{125} - p_{134} - p_{135} - p_{145} - p_{234} - p_{235} - p_{245} - p_{345} + p_{1234} + p_{1235} + p_{1245} + p_{1345} + p_{2345} + 1, p_{12} - p_{123} - p_{124} - p_{125} + p_{1234} + p_{1235} + p_{1245}, p_{13} - p_{123} - p_{134} - p_{135} + p_{1234} + p_{1235} + p_{1345}, p_{23} - p_{123} - p_{234} - p_{235} + p_{1234} + p_{1235} + p_{2345}, p_{14} - p_{124} - p_{134} - p_{145} + p_{1234} + p_{1245} + p_{1345}, p_{24} - p_{124} - p_{234} - p_{245} + p_{1234} + p_{1245} + p_{2345}, p_{34} - p_{134} - p_{234} - p_{345} + p_{1234} + p_{1345} + p_{2345}, p_{1234}, p_{15} - p_{125} - p_{135} - p_{145} + p_{1235} + p_{1245} + p_{1345}, p_{25} - p_{125} - p_{235} - p_{245} + p_{1235} + p_{1245} + p_{2345}, p_{35} - p_{135} - p_{235} - p_{345} + p_{1235} + p_{1345} + p_{2345}, p_{1235}, p_{45} - p_{145} - p_{245} - p_{345} + p_{1245} + p_{1345} + p_{2345}, p_{1245}, p_{1345}, p_{2345} }. Expressions for the probabilities and constraints also are provided in text files in the on-line supplemental materials, as described in Section 5.

Appendix C. Demonstration for $n = 4$: Tables for $\alpha = 0$ and $\alpha = 0.40$

Appendix D. Tables for Demonstration for $n = 8$:

Table C2. Values to simulate dependent Bernoulli random variables $\underline{Y}_4$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2, for $p_1 = .15$, $p_2 = .25$, $\alpha = 0.40$, and $w = 0.90$. Simulation results are provided in Table 2.

				Group 1				Group 2		
Y_1	Y_2	Y_3	Y_4	$i-1$[a]	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$	$P_4(i-1)$	$Z_4(i-1)$	$Z_4(i)$
0	0	0	0	0	.641815	0	.641815	.445125	0	.445125
1	0	0	0	1	.062535	.641815	.70435	.091125	.445125	.53625
0	1	0	0	2	.062535	.70435	.766885	.091125	.53625	.627375
1	1	0	0	3	.006615	.766885	.7735	.010125	.627375	.6375
0	0	1	0	4	.062535	.7735	.836035	.091125	.6375	.728625
1	0	1	0	5	.006615	.836035	.84265	.010125	.728625	.73875
0	1	1	0	6	.006615	.84265	.849265	.010125	.73875	.748875
1	1	1	0	7	.000735	.849265	.85	.001125	.748875	.75
0	0	0	1	8	.062535	.85	.912535	.091125	.75	.841125
1	0	0	1	9	.006615	.912535	.91915	.010125	.841125	.85125
0	1	0	1	10	.006615	.91915	.925765	.010125	.85125	.861375
1	1	0	1	11	.000735	.925765	.9265	.001125	.861375	.8625
0	0	1	1	12	.006615	.9265	.933115	.010125	.8625	.872625
1	0	1	1	13	.000735	.933115	.93385	.001125	.872625	.87375
0	1	1	1	14	.000735	.93385	.934585	.001125	.87375	.874875
1	1	1	1	15	.065415	.934585	1	.125125	.874875	1

[a] Each realization of $(Y_1, Y_2, Y_3, Y_4) = B_4(i) =$ the binary representation of $i-1$.

Table D1. Partial listing of the values (for $i-1=0$ through 68) to simulate dependent Bernoulli random variables $\underline{Y}_8$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2 when $\alpha = 0.20$. The full listing of values is provided in Supplemental Material, as described in Section 5.

									Group 1			Group 2		
Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	$i-1$[a]	$P_8(i-1)$	$Z_8(i-1)$	$Z_8(i)$	$P_8(i-1)$	$Z_8(i-1)$	$Z_8(i)$
0	0	0	0	0	0	0	0	0	.00415331	0	.00415331	.00087353	0	.00087353
1	0	0	0	0	0	0	0	1	.00174934	.00415331	.00590265	.00026287	.00087353	.0011364
0	1	0	0	0	0	0	0	2	.00112827	.00590265	.00703092	.00017716	.0011364	.00131356
1	1	0	0	0	0	0	0	3	.0031696	.00703092	.01020053	.00073127	.00131356	.00204483
0	0	1	0	0	0	0	0	4	.00125872	.01020053	.01145924	.00022878	.00204483	.00227361
1	0	1	0	0	0	0	0	5	.00182505	.01145924	.01328429	.00046089	.00227361	.0027345
0	1	1	0	0	0	0	0	6	.00264183	.01328429	.01592612	.00070072	.0027345	.00343522
1	1	1	0	0	0	0	0	7	.00244413	.01592612	.01837026	.00057757	.00343522	.00401279
0	0	0	1	0	0	0	0	8	.00138963	.01837026	.01975989	.0002765	.00401279	.00428929
1	0	0	1	0	0	0	0	9	.00191151	.01975989	.0216714	.00049222	.00428929	.00478152
0	1	0	1	0	0	0	0	10	.00156417	.0216714	.02323557	.00043424	.00478152	.00521576
1	1	0	1	0	0	0	0	11	.00083948	.02323557	.02407505	.00018027	.00521576	.00539603
0	0	1	1	0	0	0	0	12	.00297164	.02407505	.02704668	.00086615	.00539603	.00626218
1	0	1	1	0	0	0	0	13	.0009689	.02704668	.02801558	.00025803	.00626218	.00652021
0	1	1	1	0	0	0	0	14	.00201261	.02801558	.03002819	.00064896	.00652021	.00716917
1	1	1	1	0	0	0	0	15	.0044864	.03002819	.03451459	.00143469	.00716917	.00860386
0	0	0	0	1	0	0	0	16	.00151907	.03451459	.03603366	.00031615	.00860386	.00892001
1	0	0	0	1	0	0	0	17	.001997	.03603366	.03803066	.00051826	.00892001	.00943827
0	1	0	0	1	0	0	0	18	.00162704	.03803066	.0396577	.00045532	.00943827	.00989359
1	1	0	0	1	0	0	0	19	.0009331	.0396577	.0405908	.0002117	.00989359	.01010528
0	0	1	0	1	0	0	0	20	.00170475	.0405908	.04229555	.00049322	.01010528	.01059851
1	0	1	0	1	0	0	0	21	.00013218	.04229555	.04242773	.00001317	.01059851	.01061167
0	1	1	0	1	0	0	0	22	.00061872	.04242773	.04304645	.00018927	.01061167	.01080094
1	1	1	0	1	0	0	0	23	.00241087	.04304645	.04545732	.00074932	.01080094	.01155026
0	0	0	1	1	0	0	0	24	.00333895	.04545732	.04879627	.0010562	.01155026	.01260647
1	0	0	1	1	0	0	0	25	.00121149	.04879627	.05000777	.00038282	.01260647	.01298928
0	1	0	1	1	0	0	0	26	.0007327	.05000777	.05074047	.00027428	.01298928	.01326356
1	1	0	1	1	0	0	0	27	.0025806	.05074047	.05332107	.00087607	.01326356	.01413963
0	0	1	1	1	0	0	0	28	.00267282	.05332107	.05599389	.00108278	.01413963	.01522241
1	0	1	1	1	0	0	0	29	.002759	.05599389	.05875288	.00102162	.01522241	.01624402
0	1	1	1	1	0	0	0	30	.0041977	.05875288	.06295058	.00175343	.01624402	.01799745
1	1	1	1	1	0	0	0	31	.00476084	.06295058	.06771142	.00214707	.01799745	.02014453
0	0	0	0	0	1	0	0	32	.00164486	.06771142	.06935628	.00034367	.02014453	.0204882
1	0	0	0	0	1	0	0	33	.00208008	.06935628	.07143636	.00053633	.0204882	.02102453
0	1	0	0	0	1	0	0	34	.00168815	.07143636	.07312451	.00046995	.02102453	.02149448
1	1	0	0	0	1	0	0	35	.00102408	.07312451	.07414859	.00023351	.02149448	.02172799
0	0	1	0	0	1	0	0	36	.00177047	.07414859	.07591905	.00050992	.02172799	.02223791
1	0	1	0	0	1	0	0	37	.00017558	.07591905	.07609464	.00002414	.02223791	.02226205
0	1	1	0	0	1	0	0	38	.00069102	.07609464	.07678566	.00020986	.02226205	.02247191
1	1	1	0	0	1	0	0	39	.00251854	.07678566	.0793042	.00078002	.02247191	.02325193
0	0	0	1	0	1	0	0	40	.00185308	.0793042	.08115728	.00054688	.02325193	.02379881
1	0	0	1	0	1	0	0	41	.00023014	.08115728	.08138742	.0000484	.02379881	.02384721
0	1	0	1	0	1	0	0	42	.00001095	.08138742	.08139837	3.501e-06	.02384721	.02385071
1	1	0	1	0	1	0	0	43	.0015059	.08139837	.08290427	.00047236	.02385071	.02432308
0	0	1	1	0	1	0	0	44	.00089915	.08290427	.08380342	.00033796	.02432308	.02466103
1	0	1	1	0	1	0	0	45	.00158757	.08380342	.085391	.00053257	.02466103	.02519361
0	1	1	1	0	1	0	0	46	.00224622	.085391	.08763722	.0008353	.02519361	.02602891
1	1	1	1	0	1	0	0	47	.00185506	.08763722	.08949228	.00077822	.02602891	.02680714
0	0	0	0	1	1	0	0	48	.00374309	.08949228	.09323537	.00125893	.02680714	.02806607
1	0	0	0	1	1	0	0	49	.00147841	.09323537	.09471378	.00051593	.02806607	.028582
0	1	0	0	1	1	0	0	50	.00092901	.09471378	.09564278	.00038205	.028582	.02896405

[a] Each realization of $(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8) = B_8(i) =$ the binary representation of $i-1$.

Table D2. Partial listing of the values (for $i-1=0$ through 50) to simulate dependent Bernoulli random variables $\underline{Y}_8$ within each group according to the algorithm of Kang and Jung [7] described in Section 2.2. **These values were obtained under an assumption of first-order antedependence** for $\alpha = 0.20$. The full listing of values is provided in Supplemental Material, as described in Section 5.

									Group 1			Group 2		
Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	$i-1$[a]	$P_8(i-1)$	$Z_8(i-1)$	$Z_8(i)$	$P_8(i-1)$	$Z_8(i-1)$	$Z_8(i)$
0	0	0	0	0	0	0	0	0	.00355485	0	.00355485	.00068599	0	.00068599
1	0	0	0	0	0	0	0	1	.00234781	.00355485	.00590265	.00045041	.00068599	.0011364
0	1	0	0	0	0	0	0	2	.00172674	.00590265	.00762939	.0003647	.0011364	.0015011
1	1	0	0	0	0	0	0	3	.00257114	.00762939	.01020053	.00054373	.0015011	.00204483
0	0	1	0	0	0	0	0	4	.00185718	.01020053	.01205771	.00041632	.00204483	.00246115
1	0	1	0	0	0	0	0	5	.00122658	.01205771	.01328429	.00027335	.00246115	.0027345
0	1	1	0	0	0	0	0	6	.00204337	.01328429	.01532766	.00051318	.0027345	.00324768
1	1	1	0	0	0	0	0	7	.0030426	.01532766	.01837026	.00076511	.00324768	.00401279
0	0	0	1	0	0	0	0	8	.0019881	.01837026	.02035835	.00046404	.00401279	.00447683
1	0	0	1	0	0	0	0	9	.00131304	.02035835	.0216714	.00030469	.00447683	.00478152
0	1	0	1	0	0	0	0	10	.0009657	.0216714	.0226371	.0002467	.00478152	.00502822
1	1	0	1	0	0	0	0	11	.00143794	.0226371	.02407505	.00036781	.00502822	.00539603
0	0	1	1	0	0	0	0	12	.00237317	.02407505	.02644821	.00067861	.00539603	.00607464
1	0	1	1	0	0	0	0	13	.00156736	.02644821	.02801558	.00044557	.00607464	.00652021
0	1	1	1	0	0	0	0	14	.00261108	.02801558	.03062666	.0008365	.00652021	.00735671
1	1	1	1	0	0	0	0	15	.00388793	.03062666	.03451459	.00124715	.00735671	.00860386
0	0	0	0	1	0	0	0	16	.00211754	.03451459	.03663213	.00050369	.00860386	.00910755
1	0	0	0	1	0	0	0	17	.00139853	.03663213	.03803066	.00033072	.00910755	.00943827
0	1	0	0	1	0	0	0	18	.00102858	.03803066	.03905924	.00026778	.00943827	.00970605
1	1	0	0	1	0	0	0	19	.00153157	.03905924	.0405908	.00039924	.00970605	.01010528
0	0	1	0	1	0	0	0	20	.00110628	.0405908	.04169708	.00030568	.01010528	.01041097
1	0	1	0	1	0	0	0	21	.00073065	.04169708	.04242773	.00020071	.01041097	.01061167
0	1	1	0	1	0	0	0	22	.00121718	.04242773	.04364491	.00037681	.01061167	.01098848
1	1	1	0	1	0	0	0	23	.00181241	.04364491	.04545732	.00056179	.01098848	.01155026
0	0	0	1	1	0	0	0	24	.00274049	.04545732	.04819781	.00086866	.01155026	.01241893
1	0	0	1	1	0	0	0	25	.00180996	.04819781	.05000777	.00057036	.01241893	.01298928
0	1	0	1	1	0	0	0	26	.00133117	.05000777	.05133894	.00046181	.01298928	.0134511
1	1	0	1	1	0	0	0	27	.00198213	.05133894	.05332107	.00068853	.0134511	.01413963
0	0	1	1	1	0	0	0	28	.00327129	.05332107	.05659235	.00127032	.01413963	.01540995
1	0	1	1	1	0	0	0	29	.00216053	.05659235	.05875288	.00083408	.01540995	.01624402
0	1	1	1	1	0	0	0	30	.00359923	.05875288	.06235211	.00156589	.01624402	.01780991
1	1	1	1	1	0	0	0	31	.00535931	.06235211	.06771142	.00233461	.01780991	.02014453
0	0	0	0	0	1	0	0	32	.00224333	.06771142	.06995475	.00053121	.02014453	.02067574
1	0	0	0	0	1	0	0	33	.00148161	.06995475	.07143636	.00034879	.02067574	.02102453
0	1	0	0	0	1	0	0	34	.00108968	.07143636	.07252604	.00028241	.02102453	.02130694
1	1	0	0	0	1	0	0	35	.00162255	.07252604	.07414859	.00042105	.02130694	.02172799
0	0	1	0	0	1	0	0	36	.001172	.07414859	.07532059	.00032238	.02172799	.02205038
1	0	1	0	0	1	0	0	37	.00077405	.07532059	.07609464	.00021167	.02205038	.02226205
0	1	1	0	0	1	0	0	38	.00128949	.07609464	.07738413	.0003974	.02226205	.02265945
1	1	1	0	0	1	0	0	39	.00192007	.07738413	.0793042	.00059248	.02265945	.02325193
0	0	0	1	0	1	0	0	40	.00125461	.0793042	.08055881	.00035934	.02325193	.02361127
1	0	0	1	0	1	0	0	41	.00082861	.08055881	.08138742	.00023594	.02361127	.02384721
0	1	0	1	0	1	0	0	42	.00060942	.08138742	.08199684	.00019104	.02384721	.02403825
1	1	0	1	0	1	0	0	43	.00090743	.08199684	.08290427	.00028482	.02403825	.02432308
0	0	1	1	0	1	0	0	44	.00149762	.08290427	.08440189	.0005255	.02432308	.02484857
1	0	1	1	0	1	0	0	45	.0009891	.08440189	.085391	.00034504	.02484857	.02519361
0	1	1	1	0	1	0	0	46	.00164775	.085391	.08703875	.00064777	.02519361	.02584137
1	1	1	1	0	1	0	0	47	.00245353	.08703875	.08949228	.00096576	.02584137	.02680714
0	0	0	0	1	1	0	0	48	.00314463	.08949228	.0926369	.00107139	.02680714	.02787853
1	0	0	0	1	1	0	0	49	.00207687	.0926369	.09471378	.00070347	.02787853	.028582
0	1	0	0	1	1	0	0	50	.00152748	.09471378	.09624125	.00056959	.028582	.02915159

[a] Each realization of $(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8) = B_8(i)$ = the binary representation of $i-1$.