Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Computing the Point-biserial Correlation under Any Underlying Continuous Distribution

Hakan Demirtas & Donald Hedeker

To cite this article: Hakan Demirtas & Donald Hedeker (2016) Computing the Point-biserial Correlation under Any Underlying Continuous Distribution, Communications in Statistics - Simulation and Computation, 45:8, 2744-2751, DOI: 10.1080/03610918.2014.920883

To link to this article: http://dx.doi.org/10.1080/03610918.2014.920883

Accepted author version posted online: 21 Oct 2014.
Published online: 21 Oct 2016.

Communications in Statistics—Simulation and Computation®, 45: 2744–2751, 2016

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.920883

Computing the Point-biserial Correlation under Any Underlying Continuous Distribution

HAKAN DEMIRTAS AND DONALD HEDEKER

Division of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, Illinois, USA

The connection between the point-biserial and biserial correlations is well-established when the underlying distribution is bivariate normal. For many other bivariate distributions, the formula that links these two quantities is not straightforward to derive or does not have a closed form. We propose a simple technique that enables researchers to compute one of these correlations when the other is specified. For this, we take advantage of the constancy of their ratio, which can be easily approximated for any distribution. We illustrate the proposed method using several examples and discuss its extension to the ordinal case. We believe that this approach is potentially useful in stochastic simulation.
.

Keywords Dichotomization; Simulation; Sorting.

Mathematics Subject Classification Primary 62F40; Secondary 62P10

1. Introduction

A correlation between a continuous and a dichotomous variable is known as the point-biserial correlation. In some cases, the dichotomous variable is an artificial dichotomy. An example would be a situation where the dichotomous variable (e.g., obese vs. nonobese) is based on an underlying continuous variable (e.g., body mass index). Other examples include preterm versus term babies based on the gestation duration, high versus low need for cognition, young versus old age, early versus late response time in surveys. A correlation that is computed from a continuous and a dichotomous variable, in which there is an underlying continuous variable for the latter, is referred to as the biserial correlation if it reflects correlation of the two continuous variables (i.e., one is observed and the other is latent). Both the point-biserial and biserial correlations are special cases of Pearson correlation.

If the underlying distribution before dichotomization is bivariate normal, the relationship between the point-biserial and biserial correlations is well-studied (see MacCallum et al., 2002, and references therein). Suppose that X and Y follow a bivariate normal distribution with a correlation of δ_{XY}. If X is dichotomized to produce X_D, then the resulting correlation between X_D and Y can be designated as δ_{X_DY} (point-biserial correlation). The

Received January 15, 2014; Accepted April 24, 2014
Address correspondence to Hakan Demirtas, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL 60612; E-mail: demirtas@uic.edu

effect of dichotomization on δ_{XY} (biserial correlation) is given by

$$\delta_{X_DY} = \delta_{XY}\left(h/\sqrt{pq}\right) \tag{1}$$

where p and $q = 1 - p$ are the proportions of the population above and below the point of dichotomization, respectively, and h is the ordinate of the normal curve at the same point.

However, in many other cases, the formula may not have a closed form or may be difficult to obtain. For example, what if X and/or Y are nonnormal? Or what happens if X is ordinalized rather than dichotomized? Or given p, can one compute δ_{X_DY} when δ_{XY} is specified, and vice versa, under any distributional form of (X, Y)? These questions motivate the current work.

Generalizing this relationship is of interest, because (a) the connection between the point-biserial δ_{X_DY} and biserial δ_{XY} correlations may be difficult to derive in many cases; (b) in simulation studies, it is generally necessary to replicate the characteristics of a real dataset or to reproduce specified trends of synthetic data, and mixed data generation routines are typically based on underlying continuity for binary and ordinal variables of the dataset (Demirtas et al., 2012).

The proposed method has the following salient advantages: (a) it helps in understanding how correlations are related before and after discretization; (b) it works for any underlying distribution for (X, Y), and X and Y do not have to be identically distributed; (c) it only requires a univariate random number generator, thus one does not need to resort to multivariate data generation tools; (d) it works for ordinal-continuous data combinations, and so one can compute the polyserial correlation given the point–polyserial correlation (or vice versa) when the relative proportions of the ordinal categories are specified.

The organization of the article is as follows. In Section 2, we describe the linear relationship between δ_{X_DY} and δ_{XY}, which implies constancy of their ratio, propose a simple technique to compute $\delta_{X_DY}/\delta_{XY}$. In Section 3, we present illustrative examples that involve normal, nonnormal symmetric, right- and left-skewed, flat, u-shaped, and boundary at the mode type of distributions. Section 4 includes concluding remarks, future directions, limitations, and extensions.

2. A Linear Relationship

Suppose that X and Y follow a bivariate normal distribution with a correlation of δ_{XY}. Without loss of generality, we may assume that both X and Y are standardized to have a mean of 0 and a variance of 1. Let X_D be the binary variable resulting from a split on X, $X_D = I(X \geq k)$. Thus, $E[X_D] = p$ and $V[X_D] = pq$, where $q = 1 - p$. The correlation between X_D and X, δ_{X_DX} can be obtained in a simple way, namely, $\delta_{X_DX} = \frac{\text{Cov}[X_D,X]}{\sqrt{V[X_D]V[X]}} = E[X_DX]/\sqrt{pq} = E[X|X \geq k]/\sqrt{pq}$. We can also express the relationship between X and Y via the following linear regression model:

$$Y = \delta_{XY}X + \epsilon, \tag{2}$$

where ϵ is independent of X and Y, and follows $N \sim (0, 1 - \delta_{XY}^2)$. When we generalize it to nonnormal X and/or Y (both centered and scaled), the same relationship can be assumed to hold with the exception that the distribution of ϵ follows a nonnormal distribution. As long as Eq. (2) is valid,

$$\text{Cov}\,[X_D, Y] = \text{Cov}\,[X_D, \delta_{XY}X + \epsilon]$$

$$= \text{Cov}\,[X_D, \delta_{XY} X] + \text{Cov}\,[X_D, \epsilon]$$
$$= \delta_{XY} \text{Cov}\,[X_D, X] + \text{Cov}\,[X_D, \epsilon]\,. \quad (3)$$

Since ϵ is independent of X, it will also be independent of any deterministic function of X such as X_D, and thus Cov$[X_D, \epsilon]$ will be 0. As $E[X] = E[Y] = 0$, $V[X] = V[Y] = 1$, Cov $[X_D, Y] = \delta_{X_D Y}$, and Cov$[X, Y] = \delta_{XY}$, Eq. (3) reduces to

$$\delta_{X_D Y} = \delta_{XY} \delta_{X_D X}. \quad (4)$$

In the bivariate normal case, it is equivalent to Eq. (1). Eq. (4) essentially means that the linear association between X_D and Y is assumed to be fully explained by their mutual association with X. Clearly, these correlations are invariant to location shifts and scaling, X and Y do not have to be centered and scaled; their means and variances can take any finite values. The ratio, $\delta_{X_D Y}/\delta_{XY}$ is equal to $\delta_{X_D X} = E[X_D X]/\sqrt{pq} = E[X|X \geq k]/\sqrt{pq}$. It is a constant given p and the distribution of X. This conditional expectation can be found by integration, however, it could be tedious or analytically intractable. In the following section, we present a straightforward way of computing $\delta_{X_D Y}/\delta_{XY}$ without evaluating a potentially complicated truncated integration. Once the ratio is found, one can compute either correlation (biserial or point-biserial) given the other.

2.1. *Given $\delta_{X_D Y}$ or δ_{XY}, One Can Obtain the Other*

Using the convention in earlier sections, let X and Y be centered and scaled continuous variables in a bivariate setting, and let X_D be the dichotomized version of X, $X_D = I(X \geq k)$. As noted, $\delta_{X_D Y}/\delta_{XY} = c$, where the constant c is a function of characteristics of X and p (proportion of 1's), which in turn determines k (dichotomization threshold). $\delta_{X_D X}$ is always positive when high/low values of X are assigned to 1/0, respectively. This implies that $\delta_{X_D Y}$ and δ_{XY} have identical signs in large enough samples, and $|\delta_{X_D Y}| < |\delta_{XY}|$ as X is more informative than X_D.

The conditional expectation, $E[X|X \geq k]$, is the result of a truncated integration, and can be approximated by the following algorithm:

1. Generate X with a large number of data points (e.g., $N = 100,000$).
2. Dichotomize X to obtain X_D through the specified value of p.
3. Compute the sample correlation, $\delta_{X_D X} = \hat{c}$.
4. Find $\delta_{X_D Y}$ or δ_{XY} by $\delta_{X_D Y}/\delta_{XY} = \hat{c}$.

As the relationship between $\delta_{X_D Y}$ and δ_{XY} is linear, their ratio is a constant given p and the characteristics of X. The distribution of X need not be known or even parametric in order to use this method. We merely need to assume that it is continuous and we are able to generate data from it or from a parametric approximation of it. For example, a multimodal distribution of unknown form may be approximated by a normal mixture, and data can be simulated from that. Alternatively, one could sample X with replacement from an empirical distribution.

2.2. *What If* X *Is Ordinalized Rather Than Dichotomized?*

When X is ordinalized to obtain X_O, the fundamental ideas remain unchanged. As long as the assumptions of Eqs. (2) and (4) are met, the method is equally applicable to the

ordinal case in the context of the relationship between the polyserial (before ordinalization) and point–polyserial (after ordinalization) correlations. In this extension, the key covariance term, $\mathrm{Cov}[X_O X]$, involves far more complicated integrals than in the dichotomous case. Furthermore, even when these integrals can be computed by brute force or numerical integration techniques, the proposed sorting approach makes the problem much less challenging.

3. Illustrative Examples

Our examples are drawn from five bivariate densities, (X, Y) with $\mathrm{Cor}[X, Y] = \delta_{XY}$ whose marginals are given below, representing some key distributional shapes and attributes.

1. *Normal–Normal*: $X \sim N(0, 1)$, $Y \sim N(0, 1)$. The relationship between $\delta_{X_D Y}$ and δ_{XY} is given in Eq. (1).
2. *t-t*: $X \sim t_3(0, 1)$, $Y \sim t_3(0, 1)$. Both X and Y are symmetric, but have heavier tails than normal.
3. *Lognormal–Lognormal*: $X \sim \mathrm{LN}(0, \sigma_1^2)$, $Y \sim \mathrm{LN}(0, \sigma_2^2)$, where $\sigma_1 = \sigma_2 = 0.3$. X and Y are right-skewed.
4. *Uniform–Exponential*: $X \sim U(0, 1)$, $Y \sim \mathrm{Exp}(1)$. X is flat, Y has a boundary at the mode (0).
5. *Beta–Beta*: $X \sim \mathrm{Beta}(\alpha_1, \beta_1)$, $Y \sim \mathrm{Beta}(\alpha_2, \beta_2)$, where $\alpha_1 = 5, \beta_1 = 1, \alpha_2 = \beta_2 = 0.5$. X is left-skewed, Y has a u-shape.

These scenarios were chosen so that the spectrum includes: (a) three identically (1–3), two nonidentically distributed (4–5) pairs; and (b) many major distributional shapes are represented.

To assess how well the method works, we simulated bivariate continuous data from the above distributions with $N = 100,000$ observations. p (the proportion of 1's) and δ_{XY} varied from 0.01 to 0.99 with increments of 0.02, and from -0.75 to $+0.75$ with increments of 0.03, respectively, making the total number of combinations $50 \times 50 = 2,500$. The correlation range was chosen to ensure that the Fréchet–Hoeffding bounds[1] (Fréchet, 1951; Hoeffding, 1940) were not violated. These bounds can be approximated by the sorting algorithm appeared in Demirtas and Hedeker (2011). For the point-biserial correlations, we dichotomized X based on the quantiles of the respective distributions that were implied by p, and computed the empirical point-biserial correlations ($\delta^{\mathrm{emp}}_{X_D Y}$). These were then compared with the outcomes ($\delta^{\mathrm{alg}}_{X_D Y}$) of the algorithm given in Section 2.1 across 2,500 specifications within each distributional setup.

We generated bivariate data in scenarios 1–2 using standard methods, and in scenarios 3–5 via the method of Fleishman polynomials (Demirtas and Hedeker, 2008; Fleishman, 1978; Headrick, 2010; Vale and Maurelli, 1983), which is a moment-matching procedure where any given continuous variable in the system is expressed by the sum of linear combinations of powers of a standard normal variate. This involves solving a set of nonlinear

[1]It is well known that correlations are not bounded between -1 and $+1$ as different upper and/or lower bounds may be imposed by the marginal distributions (Fréchet, 1951; Hoeffding, 1940). Let $\Pi(F, G)$ be the set of all cumulative distribution functions (cdfs) H on R^2 having marginal cdfs F and G. Hoeffding (1940) and Fréchet (1951) proved that in $\Pi(F, G)$, there exist cdfs H_L and H_U, called the lower and upper bounds, having minimum and maximum correlation. For all $(x, y) \in R^2$, $H_L(x, y) = \max[F(x) + G(y) - 1, 0]$ and $H_U(x, y) = \min[F(x), G(y)]$. For any $H \in \Pi(F, G)$ and all $(x, y) \in R^2$, $H_L(x, y) \leq H(x, y) \leq H_U(x, y)$. If δ_L, δ_U, and δ denote the Pearson correlation coefficients for H_L, H_U, and H, respectively, then $\delta_L \leq \delta \leq \delta_U$.

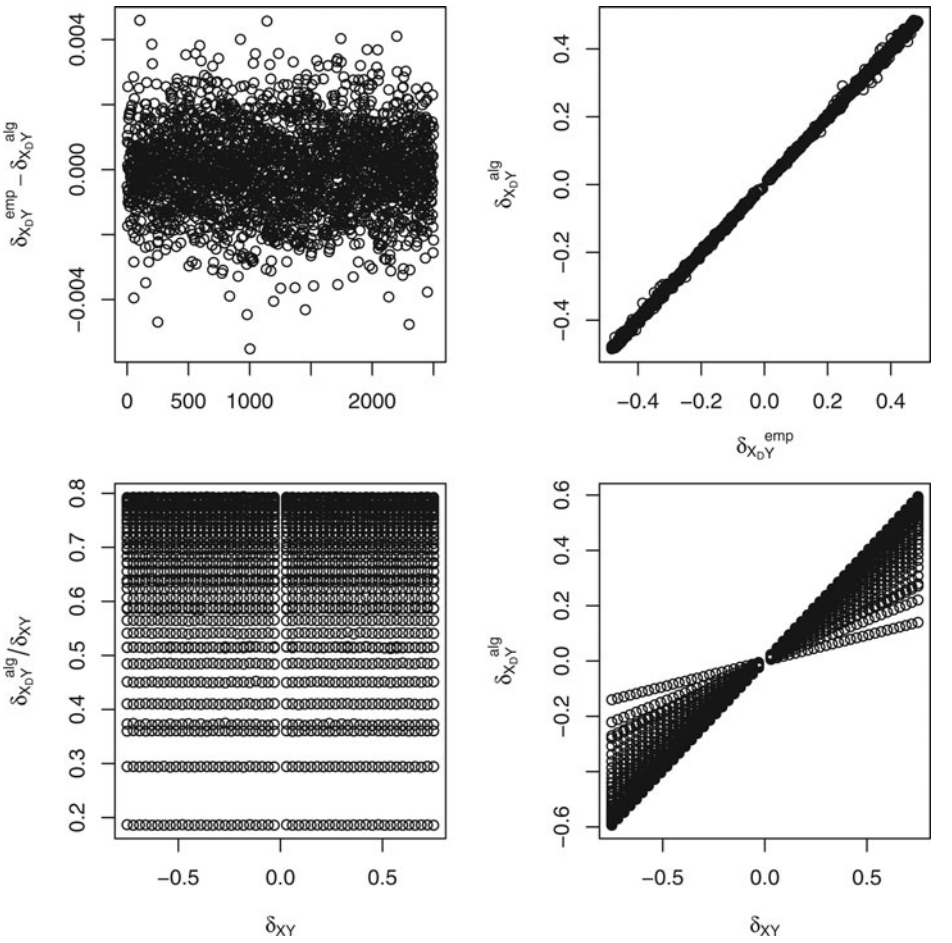

Figure 1. In matrix convention, (1,1) is the scatter plot of deviations in Scenario 1, (1,2) is the plot of $\delta_{X_DY}^{\text{alg}}$ versus δ_{XY}^{emp} in Scenario 2. (2,1) is the plot of $\delta_{X_DY}^{\text{alg}}/\delta_{XY}$ versus δ_{XY} and (2,2) is the plot of $\delta_{X_DY}^{\text{alg}}$ versus δ_{XY} in Scenario 3, respectively, across 2, 500 specifications.

equations via an optimization routine such as Newton–Raphson to compute the coefficients of polynomials marginally, and calculating intermediate correlations among normal variables that form a basis for desired nonnormal continuous distributions to attain the specified correlation structure (Headrick, 2010).

In Fig. 1, which uses all 2,500 combinations, the upper left plot presents the difference $\delta_{X_DY}^{\text{emp}} - \delta_{X_DY}^{\text{alg}}$ in the *Normal–Normal* case. The deviations are negligibly small (between -0.005 and 0.005), indicating that the procedure is working properly. The upper right plot ($\delta_{X_DY}^{\text{alg}}$ versus $\delta_{X_DY}^{\text{emp}}$) shows results from the *t-t* case, which closely resembles an identity line, as one would expect. The lower plots are based on the *Lognormal–Lognormal* scenario. The lower left plot shows the graph of $\delta_{X_DY}^{\text{alg}}/\delta_{XY}$ versus δ_{XY}, which consists of horizontal points for each value of p. The lower right plot presents $\delta_{X_DY}^{\text{alg}}$ versus δ_{XY}. Clearly, each line represents a different value of p, and the corresponding slopes are associated with flat points as in the graph of $\delta_{X_DY}^{\text{alg}}/\delta_{XY}$ versus δ_{XY}. All other scenarios resulted in deviations that are of comparable magnitude and similar relationships between $\delta_{X_DY}^{\text{alg}}$ and δ_{XY} with minor departures.

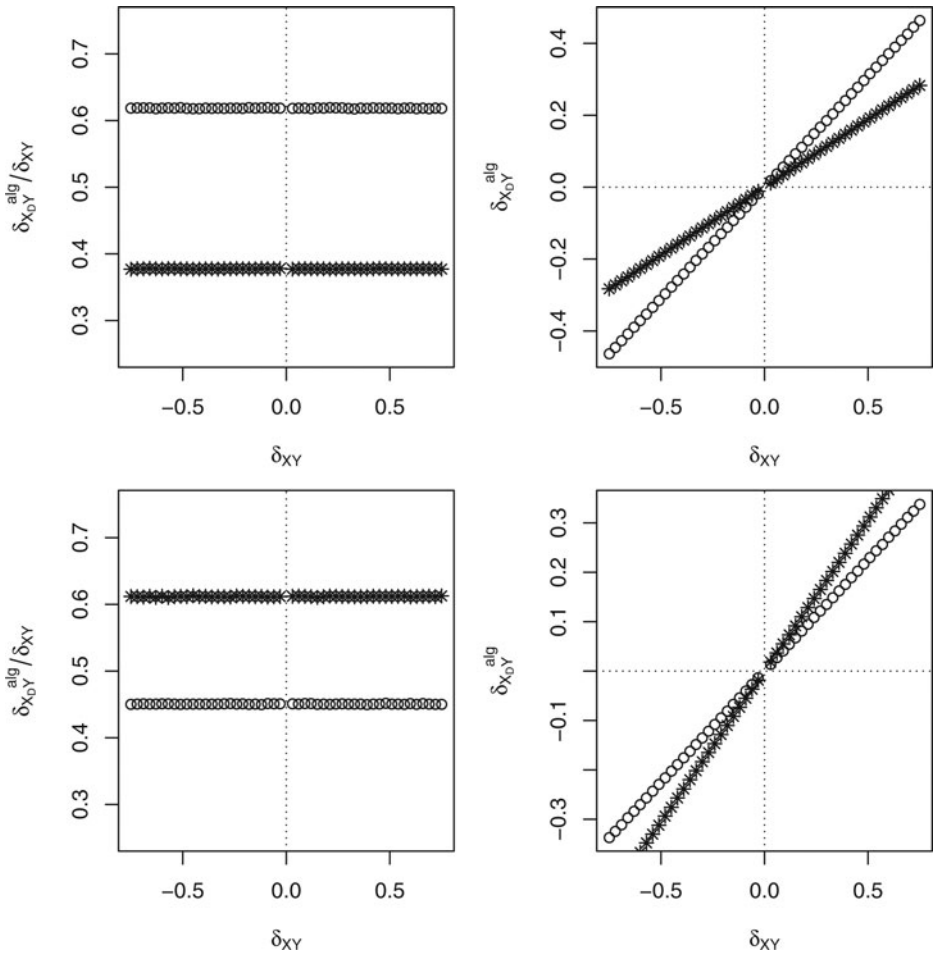

Figure 2. Upper and lower plots of $\delta^{\text{alg}}_{X_DY}/\delta_{XY}$ (left) and $\delta^{\text{alg}}_{X_DY}$ (right) versus δ_{XY} in Scenarios 4 and 5, respectively, for $p = 0.15$ (shown by o) and $p = 0.95$ (shown by *).

Figure 2 presents more specific results in which a small subset of the 2,500 combinations were chosen. The two upper plots are concerned with the *Uniform–Exponential* scenario where $p = 0.15$ and 0.95, and δ_{XY} takes 50 values from -0.75 to $+0.75$ (in Fig. 2, $\delta_{X_DY} = \delta^{\text{alg}}_{X_DY}$). Here, implementation of the algorithm for $p = 0.15$ and 0.95 involved the following steps:

1. $X \sim U(0, 1)$ with $N = 100,000$ data points.
2. Dichotomize X to obtain X_D.
3. Compute the sample correlation, $\delta_{X_DX} = \hat{c}$.
4. Find δ_{X_DY} using $\delta_{X_DY}/\delta_{XY} = \hat{c}$. This was done for all 50 positive values of δ_{XY}.

$\hat{c}$ turned out to be 0.6180271 and 0.3780867 for $p = 0.15$ and $p = 0.95$, respectively. The results for all correlation values when $p = 0.15$ and $p = 0.95$, are represented by open circles (o) and star signs (*), respectively, in Fig. 2. Both trends are in close agreement with what the method suggests with indiscernible deviations, and the slopes in the upper right plot play a confirmatory role.

The bottom plots in Fig. 2 present results from the *Beta–Beta* example. Repeating the same process as above, $\hat{c} = 0.4494789$ and 0.6104856 for $p = 0.15$ and $p = 0.95$, respectively. As can be seen, again, the discrepancies are very small.

For the ordinal case, we generated (X, Y) from a standard bivariate normal distribution with $N = 100,000$ data points and 30 different δ_{XY} values that ranged between -0.75 and $+0.75$ before X is ordinalized into four categories (1, 2, 3, 4) to yield X_O with three sets of corresponding cell probabilities: (0.25, 0.25, 0.25, 0.25), (0.05, 0.45, 0.45, 0.05), and (0.40, 0.10, 0.10, 0.40). Similar to the binary case, we compared $\delta_{X_OY}^{\text{emp}}$ and $\delta_{X_OY}^{\text{alg}}$. The discrepancies were extremely small, ranging from -0.0019545 to 0.0020053, the mean and median were near zero with interquartile range of 0.0010818 across $30 \times 3 = 90$ specifications. Here, $\delta_{X_OY}^{\text{alg}}/\delta_{XY}$ turned out to be 0.9252724, 0.9022056, 0.8613475 for the three sets of proportions in the above order.

4. Discussion

If a real dataset that consists of a set of naturally dichotomous variables and continuously measured or observed variables is at hand, one can compute all possible point-biserial correlations directly. Similarly, a direct calculation of all possible biserial correlations can be performed when the dichotomization thresholds are known and the underlying continuous measurements are accessible. This manuscript is concerned with answering the following intricately interrelated questions in the context of simulation. What should δ_{XY} be in a given bivariate setting in order to obtain a desired δ_{X_DY} after dichotomization? Alternatively, what level of the δ_{X_DY} corresponds to a specified δ_{XY} before dichotomization? What happens if we do ordinalization rather than dichotomization?

The proposed technique works as long as the marginal characteristics, degree of linear association between the two variables, and the proportion parameter are well-defined regardless of the shape of the underlying bivariate continuous density. When the quantities mentioned above are specified, one can find δ_{X_DY} given δ_{XY} and vice versa in a straightforward manner. As mentioned, the variables do not have to be identically distributed, the only necessary operational tool is a univariate random number generator that is capable of generating X.

In the simulation context, where one wants to replicate real data characteristics and/or synthetic trends that are specified, whether the categorical variable represents a true (e.g., sex) or artificial (e.g., high–low cholesterol) dichotomy/polytomy is inconsequential. Mixed data generation routines typically rely on simulating correlated continuous variables before discretization, but what matters is the discrete outcome, which can be treated as true or artificial in subsequent analyses. On a related note, if such a routine is involved with generating multivariate continuous data as an intermediate step, this method could be one of the key algorithmic stages. Furthermore, in conjunction with the published works on joint binary/normal (Demirtas and Doganay, 2012), binary/nonnormal continuous (Demirtas et al., 2012), and multivariate ordinal data generation (Demirtas, 2006), the ideas presented in this manuscript might serve as a milestone for simultaneous ordinal/normal and ordinal/nonnormal data generation schemes. Finally, as correctly pointed out by a reviewer, the method could be very useful in meta-analysis when some studies dichotomize or ordinalize variables and others do not.

A potential limitation of the presented approach is that the linear association between X_D and Y is assumed to be completely explained by their mutual association with X. The other sorts of interdependence between δ_{X_DY} and δ_{XY} due to nonlinear associations

between X and Y will be explored in future work. It is worthwhile to mention that this simple linear relationship does not hold for the phi coefficient (correlation between two discretized variables; in fact the term phi coefficient is reserved for dichotomous variables, but for lack of a better term we also use it for ordinalized variables) and δ_{XY}. Thus, modeling the relationship between the phi coefficient and tetrachoric/polychoric correlation (correlation between two continuous variables before dichotomization/ordinalization) calls for more complicated techniques that we are currently developing. Once these are available, we could more fully understand the nature of discretization in the sense that we would know how the correlation structure is transformed after discretization in simulated settings. Of note, we do not attach any positive or negative connotations to discretization, our attention and the scope of the manuscript is limited to the relationship between δ_{X_DY} and δ_{XY}. Discretization is generally considered a bad idea, but it might have some benefits. For competing views, see MacCallum et al. (2002) and Farrington and Loeber (2000).

One can argue that the contribution of this work in the binary case is not substantial. Once the linearity and constancy arguments are deemed statistically defensible, our approximation method may be perceived as trivial in the presence of today's numerical integration tools. However, we are approximating the ratio of potentially complicated conditional expectation and variance expressions by a simple and flexible computational approach. The advantages become more apparent in the ordinal case, because the integrands could get far more complex than those under dichotomization.

In a nutshell, the proposed algorithm could be a practical and easy-to-implement tool to identify the relationship between the point-biserial and biserial correlations, and has many potential advantages in the random generation world.

References

Demirtas, H. (2006). A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation* 76:1017–1025.

Demirtas, H., Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics* 22:223–236.

Demirtas, H., Hedeker, D. (2008). Multiple imputation under power polynomials. *Communications in Statistics - Simulation and Computation* 37:1682–1695.

Demirtas, H., Hedeker, D. (2011). A practical way for computing approximate lower and upper correlation bounds. *The American Statistician* 65:104–109.

Demirtas, H., Hedeker, D., Mermelstein, J. M. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine* 31:3337–3346.

Farrington, D. P., Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health* 10:100–122.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika* 43:521–532.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon Section A* 14:53–77.

Headrick, T. C. (2010). *Power Method Polynomials and Other Transformations*. Boca Raton, FL: Chapman and Hall/CRC.

Hoeffding, W. (1940). Scale-invariant correlation theory. In: Fisher, N. I., Sen, P. K., eds., *The Collected Works of Wassily Hoeffding*. New York: Springer-Verlag, pp. 57–107.

MacCallum, R. C., Zhang, S., Preacher, K. J., Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods* 7:19–40.

Vale, C. D., Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika* 48:465–471.