Journal of Educational and Behavioral Statistics
Fall 1996, Vol. 21, No. 3, pp. 264–282

Dichotomization, Partial Correlation, and Conditional Independence

András Vargha and **Tamás Rudas**
Loránd Eötvös University, Budapest, Hungary
Harold D. Delaney
University of New Mexico
Scott E. Maxwell
University of Notre Dame

Key words: *attenuation, dichotomization, median splits, partial correlation*

It was recently demonstrated that performing median splits on both of two predictor variables could sometimes result in spurious statistical significance instead of lower power. Not only is the conventional wisdom that dichotomization always lowers power incorrect, but the current article further demonstrates that inflation of apparent effects can also occur in certain cases where only one of two predictor variables is dichotomized. In addition, we show that previously published formulas claiming that correlations are necessarily reduced by bivariate dichotomization are incorrect. While the magnitude of the difference between the correct and incorrect formulas is not great for small or moderate correlations, it is important to correct the misunderstanding of partial correlations that led to the error in the previous derivations. This is done by considering the relationship between partial correlation and conditional independence in the context of dichotomized predictor variables.

A common design in behavioral science research, particularly research focusing on individual differences, involves dichotomizing two continuous variables in order to study their effects on a third variable. Although methodologists have long cautioned against the difficulties that can arise from artificially dichotomizing continuous variables (e.g., Humphreys & Fleishman, 1974), there are plausible explanations for why researchers continue to be attracted to such procedures. One such explanation is that using "median splits" may *seem* to simplify the data analysis for a given study. More

Much of this work resulted from the collaborative efforts of Drs. Vargha, Rudas, and Delaney while Harold Delaney was at Loránd Eötvös University with the support of a Fulbright grant from the U.S. government, and while Tamás Rudas was supported by Grant No. 2513 from the Hungarian National Science Foundation. Dr. Rudas is also affiliated with TARKI, an independent social science research center in Budapest.

The authors wish to express their gratitude to an associate editor and two anonymous reviewers for their detailed and helpful comments.

importantly, dichotomization seems to confer a conceptual benefit in that it permits one to talk about categories or types instead of continua. As an example, "Type A personality" has become standard nomenclature in certain areas, some of which have relevance to education (e.g.,·Tang, 1988), despite the fact that the label may be applied to individuals obtaining any of the wide range of values on the Jenkins Activity Survey observed above the median in a particular study.

There are, however, strong methodological arguments against dichotomizing. Most well known is the fact that artificially dichotomizing a single predictor variable typically results in an underestimation of the magnitude of bivariate relationships and a lowering of statistical power for detecting true effects (Cohen, 1978; Humphreys & Fleishman, 1974; Maxwell, Delaney, & Dill, 1984). As a result, some researchers may have reasoned that they were being statistically conservative by dichotomizing: If an effect is obtained with a dichotomized variable, then the finding must be robust because it was achieved despite the low power. However, Maxwell and Delaney (1993) recently showed that with multiple predictors a somewhat counterintuitive result can occur. Specifically, they present formulas showing that simultaneously dichotomizing two normally distributed predictor variables can dramatically increase the probability of Type I errors in the tests of the predictors' effects.

Maxwell and Delaney (1993) dealt only with the situation where both of two predictor variables are dichotomized, and unfortunately readers may have gotten the impression that these effects are restricted to that situation. Unfortunately as well, some previously published results on the effects of joint dichotomization are incorrect. A major purpose of the current article is to elucidate the true effects of joint dichotomization of two variables, and to examine the implications of the correct formulation for generalizations of Maxwell and Delaney's results.

We begin by deriving correct formulas for the attenuation of correlations that results from dichotomization. We next make comparisons between the correct results and those previously published. After noting that dichotomization involves a quantifiable loss of information, we consider the implications of our derivations for extensions of Maxwell and Delaney's (1993) results under various combinations of dichotomized and continuous variables. Although our focus is on dichotomizations resulting from median splits, we also briefly consider the case where splits are made at points other than the median. Finally, because the errors in the previously published formulas resulted from a misunderstanding of partial correlations that we suspect still persists, we indicate when a partial correlation can and cannot be interpreted as a conditional correlation, or the correlation between two variables at a given level of the third variable. Specifically, we show that even conditional independence does not necessarily imply that the standard partial correlation formula will yield a zero value, and we suggest that other methods may be more appropriate means of addressing substantive questions of interest.

Effects of Dichotomization on Correlations

The question of what happens when both of two normally distributed variables are dichotomized has been considered by a number of authors from Pearson (1900) to the present. One frequently cited source, Peters and Van Voorhis (1940), discusses the problem of dichotomizing one or both of the variables, X_1 and X_2, in a bivariate normal distribution. Peters and Van Voorhis assert that the original correlation ρ between the variables will be reduced to $.798\rho$ if one of the variables is dichotomized and to $.637\rho$ if both are dichotomized. In explaining Peters and Van Voorhis's equations, Cohen (1983) comments, "The constant .637 here is $.798^2$, the result of applying the .798 correction twice" (p. 251)—once for each dichotomization. Although it is plausible that each of the two dichotomizations would have the same effect, this is not quite correct. In fact, the correlation between the two dichotomized variables can be greater than the attenuated correlation resulting from dichotomizing only one of the two variables.

A simple thought problem demonstrates that the .637 multiplier must be wrong: Consider the case where X_1 and X_2 correlate 1.0. The correlation for the dichotomized variables obviously remains at 1.0, in stark contrast to the .637 value that one would have arrived at by applying the .798 multiplier twice.

To derive the correct answer for the joint dichotomization problem, we need to return to the development of the .798 multiplier for the dichotomization of one variable. Let us assume that X_1 and Y follow a bivariate normal distribution. Without loss of generality, we may assume that both X_1 and Y are standardized to have a mean of 0 and a variance of 1. Let X_{1d} be the binary variable resulting from performing a median split on X_1, and let X_{1d} take on the values of -1 and $+1$ depending on whether X_1 is negative or nonnegative, that is,

$$X_{1d} = \begin{cases} 1 \text{ for } X_1 \geq 0 \\ -1 \text{ for } X_1 < 0 \end{cases} . \tag{1}$$

Thus, X_{1d} will also have a mean of 0 and a variance of 1. Hence, the correlation between a normal random variable and the dichotomized form resulting from a median split quickly reduces to the covariance:

$$\rho_{X_1X_{1d}} = \frac{\sigma_{X_1X_{1d}}}{\sigma_{X_1}\sigma_{X_{1d}}} = \frac{\sigma_{X_1X_{1d}}}{1 \cdot 1} = \sigma_{X_1X_{1d}} = \mathscr{E}(X_1 - \mu_{X_1})(X_{1d} - \mu_{X_{1d}}). \tag{2}$$

But since the means of both variables are zero and since the product of X_1 and X_{1d} is just $|X_1|$, we have

$$\rho_{X_1X_{1d}} = \mathscr{E}(X_1 \cdot X_{1d}) = \mathscr{E}(|X_1|). \tag{3}$$

Determining this expected absolute value by integrating $|X_1|$ times the normal density function ϕ yields the .798 multiplier:

$$\begin{aligned}\rho_{X_1X_{1d}} &= \int_{-\infty}^{\infty} |X_1|\phi(X_1)\,dx_1 = \int_{-\infty}^{\infty} |X_1| \frac{1}{\sqrt{2\pi}} e^{-X_1^2/2}\,dx_1 \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} -X_1e^{-X_1^2/2}\,dx_1 + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} X_1e^{-X_1^2/2}\,dx_1 \\ &= \frac{1}{\sqrt{2\pi}}\cdot 1 + \frac{1}{\sqrt{2\pi}}\cdot 1 = \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}} = .798.\end{aligned} \tag{4}$$

To complete the justification of the .798 multiplier, we now show that this correlation of a normal variable with its dichotomous form indicates the factor by which its correlation with other normally distributed variables is reduced. Assuming that X_1 and Y are bivariate normal with a correlation of ρ_{X_1Y}, we may express their relationship via the following model:

$$Y = \rho_{X_1Y}X_1 + \epsilon \tag{5}$$

where ϵ is independent of X_1 and of Y, and is normally distributed with mean of 0 and variance of $1 - \rho^2_{X_1Y}$. Thus, we may express the correlation between the dichotomous X_{1d} variable and the Y variable as follows:

$$\rho_{X_{1d}Y} = \frac{\text{cov}(X_{1d}, Y)}{\sigma_{X_{1d}}\sigma_Y} = \frac{\text{cov}[X_{1d}, (\rho_{X_1Y}X_1 + \epsilon)]}{1 \cdot 1}. \tag{6}$$

Expanding the covariance on the right above, we have

$$\begin{aligned}\text{cov}[X_{1d}, (\rho_{X_1Y}X_1 + \epsilon)] &= \text{cov}(X_{1d}, \rho_{X_1Y}X_1) + \text{cov}(X_{1d}, \epsilon) \\ &= \rho_{X_1Y}\text{cov}(X_{1d}, X_1) + \text{cov}(X_{1d}, \epsilon).\end{aligned} \tag{7}$$

But since ϵ is independent of X_1, it will also be independent of any deterministic function of X_1 such as X_{1d}, and thus the covariance of X_{1d} and ϵ will be 0. And, from Equation 2 we know that the covariance of X_{1d} and X_1 is simply equal to their correlation. So we can write

$$\rho_{X_{1d}Y} = \rho_{X_1Y}\text{cov}(X_{1d}, X_1) = \rho_{X_1Y} \cdot \rho_{X_1X_{1d}} = \rho_{X_1Y}\frac{2}{\sqrt{2\pi}} = .798\rho_{X_1Y}. \tag{8}$$

We can now turn our attention to the development of the central formula of interest, namely, the correlation resulting from doing a bivariate median split on two bivariate normal variables. Cohen's (1983) incorrect formulation is based on Peters and Van Voorhis's (1940, p. 394) fallacious argument. In effect, Peters and Van Voorhis got into trouble by beginning with a wrong premise. They incorrectly assume that the (partial) correlation between two dichotomized variables (say, X_{1d} and X_{2d}) controlling for the continuous form of one of the variables (say, X_1) will necessarily be zero. This in turn rests on their mistaken assertion that this partial correlation is the correlation between the dichotomous variables with the X_1 variable held constant in the sense of examining the X_{1d}, X_{2d} correlation at fixed values of X_1. Their reasoning is that the partial correlation must be zero because X_{1d} is a constant for any particular value of X_1, and "any variable correlated with a constant gives a zero correlation" (Peters & Van Voorhis, 1940, p. 394).

However, in reality, the partial correlation is not necessarily equal to the correlation at a fixed value of the variable being partialed, unless multivariate normality holds, which it obviously cannot in the case of binary variables. In fact, the partial correlation between the dichotomous variables will be nonzero any time the correlation between the two original variables is nonzero.

We demonstrate this by making use of Kendall and Stuart's (1958, p. 350) results for the probability of obtaining particular combinations of observations when median splits are performed in a bivariate normal population. As before, we can express the correlation between two variables as the ratio of their covariance to the product of their standard deviations:

$$\rho_{X_{1d}X_{2d}} = \frac{\mathrm{cov}(X_{1d}, X_{2d})}{\sigma_{X_{1d}} \cdot \sigma_{X_{2d}}} = \frac{\mathscr{E}(X_{1d} \cdot X_{2d}) - \mathscr{E}(X_{1d})\mathscr{E}(X_{2d})}{\sigma_{X_{1d}} \cdot \sigma_{X_{2d}}}. \tag{9}$$

However, if we again let the dichotomous variables take on values of -1 and $+1$ for negative and nonnegative values of the corresponding standardized normal variables, the expected values of the single variables in the numerator above will both be 0, and the standard deviations in the denominator will both be 1. Thus, we can write

$$\rho_{X_{1d}X_{2d}} = \mathscr{E}(X_{1d} \cdot X_{2d}). \tag{10}$$

We can determine this expected value by utilizing expressions for the probability of the four possible combinations of values of X_{1d} and X_{2d}. Specifically, Kendall and Stuart (1958, p. 351) show that the probability of X_1 and X_2 both being above their mean (or equivalently, their median) is

$$Pr(X_1 > 0, X_2 > 0) = \frac{1}{4} + \frac{\arcsin(\rho_{X_1X_2})}{2\pi}. \tag{11}$$

(It should be noted that the arcsine is expressed in radians here and equals 0 when the correlation is 0, and equals $1.5708 = \pi/2$ when the correlation is 1.) But the symmetry of the bivariate normal implies that the probability of both variables being below the mean is the same as in (11):

$$Pr(X_1 < 0, X_2 < 0) = Pr(X_1 > 0, X_2 > 0). \tag{12}$$

And, the probability of one variable being above its mean and the other variable below its mean will be just half the difference between 1 and the sum of the probability for the equal outcome cases:

$$\begin{aligned} Pr(X_1 > 0, X_2 < 0) = Pr(X_1 < 0, X_2 > 0) &= \frac{1}{2}\left[1 - 2\left(\frac{1}{4} + \frac{\arcsin(\rho_{X_1X_2})}{2\pi}\right)\right] \\ &= \frac{1}{4} - \frac{\arcsin(\rho_{X_1X_2})}{2\pi}. \end{aligned} \tag{13}$$

Thus, when the expected value of Equation 10 is computed by multiplying $X_{1d} \cdot X_{2d}$ by the probability of each of the four possible combinations of values, the constants cancel each other out and we are left with

$$\rho_{X_{1d}X_{2d}} = \frac{4 \arcsin (\rho_{X_1X_2})}{2\pi} = (2/\pi) \arcsin (\rho_{X_1X_2}) = .637 \arcsin (\rho_{X_1X_2}). \tag{14}$$

Comparison of Incorrect and Correct Formulas

How do the values one gets with this formula compare with those of the previously published incorrect equation? Because for any nonzero correlation, the arcsine of ρ is always greater than ρ in absolute value, the incorrect formula always provides an underestimate of the magnitude of the correlation. However, as shown in Table 1, if the continuous variables X_1 and X_2 correlate

TABLE 1

The cost of dichotomizing: Reduction in correlations resulting from median splits of bivariate normal variables

	Original value of $\rho_{X_1X_2}$							
Variable(s) dichotomized	.1	.3	.5	.7	.9	.9075	.95	1.0
Both: PVV formula[a]	.0637	.1910	.3138	.4456	.5730	.5777	.6048	.6366
Both: Correct formula[b]	.0638	.1940	.3333	.4936	.7129	.7240	.7978	1.0000
One[c]	.0798	.2394	.3989	.5585	.7180	.7240	.7580	.7979

[a]Values computed using Peters & Van Voorhis's (1940) incorrect formula:

$$\rho_{X_{1d}X_{2d}} = (2/\pi)\rho_{X_1X_2} = .637\rho_{X_1X_2}.$$

[b]Values computed using the correct formula derived in the present article:

$$\rho_{X_{1d}X_{2d}} = (2/\pi)\arcsin(\rho_{X_1X_2}) = .637\arcsin(\rho_{X_1X_2}).$$

[c]Correlations resulting from dichotomizing only one of the two original variables:

$$\rho_{X_{1d}X_2} = \rho_{X_1X_{2d}} = \sqrt{2/\pi}\rho_{X_1X_2} = .798\rho_{X_1X_2}.$$

at .5 or less, the difference in expected correlations of their dichotomous forms X_{1d} and X_{2d} yielded by the incorrect and correct formulas is less than .02.

On the other hand, as the original correlation increases above .5, the difference between the correct and incorrect formulas also increases, until at a $\rho_{X_1X_2}$ value of 1.0 the correct value of the correlation between the two dichotomous forms is one and a half times the .637 others have suggested.

Also shown in the table is the value of .798ρ, that is, the correlation resulting from splitting only one of the two bivariate normal variables at its mean. Although two splits can be better than one, this is usually not the case. For any original correlations below .9075—that is, below the solution of $.637 \arcsin(\rho_{X_1X_2}) = .798\rho_{X_1X_2}$—dichotomizing both variables will result in a lower correlation than dichotomizing only one.

Implications for Extensions of Maxwell and Delaney's Results

Because the consequence of dichotomization (namely, false significance) discussed by Maxwell and Delaney (1993) is different from that usually noted (namely, reduced power), the question arises of whether dichotomizing just one of the two predictors would avoid or reverse the positive bias noted with the bivariate median split. Further, for each combination of continuous or dichotomous predictor variables, the dependent variable could also be dichotomized. Artificially dichotomous dependent variables occur with some regularity in education and psychology—for example, when a criterion is applied to an individual's average score in a course to decide whether he or she passes, or when a cutoff on a personality scale is used to classify an individual as pathological.

In the three-variable situation that we are considering, each variable could be continuous or dichotomized, yielding eight possible combinations, as shown in Table 2, where X_1 and X_2 denote the two predictors and Y denotes

TABLE 2
Eight possible analysis strategies for one continuous criterion and two continuous predictor variables

Case	Y	X_1	X_2
	ANOVA/regression cases		
1	C	C	C
2	C	D	D
3	C	D	C
4	C	C	D
	Categorical cases		
5	D	C	C
6	D	D	D
7	D	D	C
8	D	C	D

Note. C = continuous. D = dichotomized.

the dependent variable. Maxwell and Delaney (1993) considered only the first two cases; here we consider all eight. Readers of the Maxwell and Delaney article may have concluded that spurious statistical significance is a problem only if *both* predictor variables are dichotomized. We show below that dichotomization of one predictor can inflate the apparent effect even more than joint dichotomization. However, perhaps counterintuitively, the effects of dichotomization of each of the two predictor variables individually are not necessarily symmetrical.

As in Maxwell and Delaney (1993), we will assume that the zero-order correlations between all pairs of variables are nonzero in the population, but that the correlation between Y and X_2 controlling for X_1 is zero. Note that $\rho_{X_2 Y \cdot X_1} = 0$ implies that $\rho_{X_2 Y} = \rho_{X_1 X_2} \cdot \rho_{X_1 Y}$, a fact that will be helpful in assessing the effects of dichotomization on the measure of most interest here, namely, the apparent relationship between the second predictor and the criterion once the effects of the first predictor have been removed.

The data analysis strategies suggested by Table 2 could involve several different statistical techniques. Considering first the cases with a continuous dependent variable, Cases 1 and 2 would correspond to a multiple regression analysis and a 2 × 2 analysis of variance, respectively. Cases 3 and 4 would normally be analyzed using analysis of covariance (ANCOVA). In Case 3, the test of X_2 allowing for X_1 would correspond to the test of within-group regression where the dichotomous form of X_1 is the grouping variable and X_2 is the covariate. In Case 4, the principal test of interest would usually be the adjusted group effect, that is, the test of the main effect of the dichotomous X_2 factor controlling for the X_1 covariate.

The situations with a categorical dependent variable might conventionally be analyzed using a discriminant analysis or log-linear routine. Categorical Case 5 would involve an attempt to discriminate between groups derived from the Y variable on the basis of two continuous predictors. For example, Y might be a measure of posttreatment amount of drinking in a study of alcoholism treatment programs. One might derive "problem drinker" and "non–problem drinker" groups from this variable and attempt to discriminate between the groups on the basis of amount of treatment received (X_1) and knowledge of the physical consequences of drinking (X_2). Dichotomizing both or only one of these latter variables yields Categorical Cases 6–8. In Categorical Case 6, all variables are dichotomous and one might perform a log-linear analysis of the resulting 2 × 2 × 2 contingency table. Cases 7 and 8 could be analyzed via discriminant analysis or logistic regression using a combination of discrete and continuous predictors.

Because of the variety of statistical methodologies that could be employed, comparison of the analyses across these cases is potentially difficult. However, as suggested by Rosenthal (1987, pp. 106–107), one can make comparisons readily by using a standardized measure of the size of effect. Here the correlation coefficient will suffice, with the partial correlation between X_2

and Y controlling for X_1 being of most interest, and where each of the three variables is alternately in its original form or in dichotomous form. Difficulties in the valid interpretation of this convenient but misunderstood statistic will be considered in the final section.

The basic conceptual issue in dichotomization, even with the spuriously significant results with bivariate median splits, is the loss of information when one dichotomizes. As an aside, we note that in the univariate case it is possible to quantify precisely the loss of information resulting from dichotomizing. Measures of information were introduced to psychology from communication theory over 30 years ago by cognitive psychologists (e.g., Garner, 1962) and are sometimes referred to as Shannon measures of average information or average uncertainty. Statisticians use essentially the same measures for characterizing distributions, though more commonly referring to the entropy (Pugachev, 1984, p. 113; Rao, 1973) or information content (Hastings & Peacock, 1975, p. 13) of the distribution. A distribution's information content, I, may be defined in bits as

$$I = -\int_{-\infty}^{+\infty} p(x) \log_2 p(x)\, dx. \tag{15}$$

Rao (1973, pp. 162–163) proves that the normal distribution has the maximum information content among all distributions with a given mean and variance, where the random variable varies from $-\infty$ to $+\infty$. In a standard normal distribution having unit variance, this may be expressed (see, e.g., Hastings & Peacock, 1975, p. 96) as

$$I = \log_2 \sqrt{2\pi e} = \log_2 \sqrt{17.079} = \log_2 4.133 = 2.047. \tag{16}$$

In the case of a discrete distribution, the number of bits of information (Garner, 1962, p. 21) is

$$I = -\sum p(x) \log_2 p(x), \tag{17}$$

which in the case of a uniform random variable reduces simply to $\log_2 N$, where N is the number of possible values of the random variable. Thus, in the case of a dichotomized variable with a 50-50 split, we would have

$$I = \log_2 2 = 1.000. \tag{18}$$

Thus, in terms of measured information, such a dichotomized variable has slightly less than half the information of the normally distributed variable from which it was formed.

To return to our primary concern of the impact on apparent effect size in a three-variable situation, we move now to a consideration of the implications of this loss of information for the partial correlation of X_2 and Y controlling for X_1, enumerating results for each of the eight situations listed in Table 2. The formulas for the three-variable situation where the original variables follow a multivariate normal distribution are shown in Table 3. Note that the numbers appearing in the formulas are simply powers of $2/\pi$ expressed to three digits of accuracy—that is, $.798 = (2/\pi)^{1/2}$, $.637 = (2/\pi)^1$, $.508 = (2/\pi)^{3/2}$, and $.405 = (2/\pi)^2$. Also shown are the numerical values for the partial correlation corresponding to one of the examples of Maxwell and Delaney (1993) where $\rho_{X_1Y} = .7$, $\rho_{X_1X_2} = .5$, and $\rho_{X_2Y} = .35$. As shown in Table 3, this set of values results in the true partial correlation of X_2 and Y controlling for X_1 being exactly 0.

Perhaps at first glance the pattern of results seems counterintuitive in that dichotomizing only one of the two predictors (Cases 3 and 4) does not yield results intermediate between those of the original variables (Case 1) and the bivariate median split (Case 2). Instead, dichotomizing only X_1 yields the "worst" results in the sense that the partial correlation is furthest from the "correct" value of 0, and dichotomizing only X_2 results in no apparent loss of information. This pattern is made understandable by realizing that because X_1 is the only conditionally predictive if not the true causal variable here, it is critical that one have full information on its values. Note that while X_1 may indeed be the cause of Y, it may also be that X_1 is simply an indicator of an unobserved latent variable or a correlate of some other variable that is the true cause. In Cases 1 and 4, all of the indirect effects of X_1 through X_2 can be removed. In Case 2, the X_1 effects are not entirely removed from X_2 because some of the information about X_1's values has been lost, but the X_2-Y correlation is suppressed somewhat because of the dichotomization of X_2. Case 3 is the most extreme because the continuous form of X_2 reflects the continuous information in X_1, but the partialing of $X_{1\mathrm{d}}$ removes only some of that information.

The cases where Y is dichotomized are analogous to the four cases where Y is continuous. The same general pattern of zero and nonzero correlations holds, for the same reasons cited above, but the nonzero correlations are suppressed somewhat because of the dichotomization of Y. The one exception is Case 8, where the continuous form of the causal variable is controlled. Because the two other variables are conditionally independent, this partial correlation will be close to zero for moderately sized original correlations. However, it will not be exactly zero because the reduction of the correlation between $X_{2\mathrm{d}}$ and Y_d resulting from their joint dichotomization is not exactly equal to applying the .798 multiplier twice, as is done in the $X_1X_{2\mathrm{d}}$ and X_1Y_d correlations.

Splits at Other Cutting Points

It would be convenient if results could be presented for splits at points other than the median. One can specify the value of a normal variable's

TABLE 3

Effects of median splits with multivariate normal variables on partial correlations

Case	Reduced bivariate correlations	Partial correlation controlling for continuous or dichotomous form of conditionally predictive variable	
		General formula	Numerical example
1 (CCC)	—	$\rho_{X_2Y\cdot X_1} = \dfrac{\rho_{X_2Y} - \rho_{X_1X_2}\rho_{X_1Y}}{\sqrt{1 - \rho^2_{X_1X_2}}\sqrt{1 - \rho^2_{X_1Y}}}$	0.0
2 (CDD)	$\rho_{X_{2d}Y} = .798\rho_{X_2Y}$ $\rho_{X_{1d}X_{2d}} = .637 \text{ arcsin } \rho_{X_1X_2}$ $\rho_{X_{1d}Y} = .798\rho_{X_1Y}$	$\rho_{X_{2d}Y\cdot X_{1d}} = \dfrac{.798\rho_{X_2Y} - .508(\text{arcsin } \rho_{X_1X_2})\rho_{X_1Y}}{\sqrt{1 - .405(\text{arcsin } \rho_{X_1X_2})^2}\sqrt{1 - .637\rho^2_{X_1Y}}}$	.119
3 (CDC)	$\rho_{X_{1d}X_2} = .798\rho_{X_1X_2}$ $\rho_{X_{1d}Y} = .798\rho_{X_1Y}$	$\rho_{X_2Y\cdot X_{1d}} = \dfrac{\rho_{X_2Y} - .637\rho_{X_1X_2}\rho_{X_1Y}}{\sqrt{1 - .637\rho^2_{X_1X_2}}\sqrt{1 - .637\rho^2_{X_1Y}}}$	.167
4 (CCD)	$\rho_{X_{2d}Y} = .798\rho_{X_2Y}$ $\rho_{X_1X_{2d}} = .798\rho_{X_1X_2}$	$\rho_{X_{2d}Y\cdot X_1} = \dfrac{.798\rho_{X_2Y} - .798\rho_{X_1X_2}\rho_{X_1Y}}{\sqrt{1 - .637\rho^2_{X_1X_2}}\sqrt{1 - \rho^2_{X_1Y}}}$	0.0

5 (DCC)	$\rho_{X_2Y_d} = .798\rho_{X_2Y}$ $\rho_{X_1Y_d} = .798\rho_{X_1Y}$	$\rho_{X_2Y_d \cdot X_1} = \dfrac{.798\rho_{X_2Y} - .798\rho_{X_1X_2}\rho_{X_1Y}}{\sqrt{1 - \rho^2_{X_1X_2}}\sqrt{1 - .637\rho^2_{X_1Y}}}$	0.0
6 (DDD)	$\rho_{X_{2d}Y_d} = .637 \arcsin \rho_{X_2Y}$ $\rho_{X_{1d}X_{2d}} = .637 \arcsin \rho_{X_1X_2}$ $\rho_{X_{1d}Y_d} = .637 \arcsin \rho_{X_1Y}$	$\rho_{X_{2d}Y_d \cdot X_{1d}} = \dfrac{.637 \arcsin \rho_{X_2Y} - .405(\arcsin \rho_{X_1X_2})\ (\arcsin \rho_{X_1Y})}{\sqrt{1 - .405(\arcsin \rho_{X_1X_2})^2}\sqrt{1 - .405(\arcsin \rho_{X_1Y})^2}}$	.086
7 (DDC)	$\rho_{X_2Y_d} = .798\rho_{X_2Y}$ $\rho_{X_{1d}X_2} = .798\rho_{X_1X_2}$ $\rho_{X_{1d}Y_d} = .637 \arcsin \rho_{X_1Y}$	$\rho_{X_2Y_d \cdot X_{1d}} = \dfrac{.798\rho_{X_2Y} - .508\rho_{X_1X_2}(\arcsin \rho_{X_1Y})}{\sqrt{1 - .637\rho^2_{X_1X_2}}\sqrt{1 - .405(\arcsin \rho_{X_1Y})^2}}$	.108
8 (DCD)	$\rho_{X_{2d}Y_d} = .637 \arcsin \rho_{X_2Y}$ $\rho_{X_1X_{2d}} = .798\rho_{X_1X_2}$ $\rho_{X_1Y_d} = .798\rho_{X_1Y}$	$\rho_{X_{2d}Y_d \cdot X_1} = \dfrac{.637 \arcsin \rho_{X_2Y} - .637\rho_{X_1X_2}\rho_{X_1Y}}{\sqrt{1 - .637\rho^2_{X_1X_2}}\sqrt{1 - .637\rho^2_{X_1Y}}}$	.006

Note. Letters in parentheses indicate whether Y, X_1, and X_2, respectively, are in continuous (C) or dichotomized (D) form in a given case. In numerical examples, $\rho_{X_1Y} = .7$, $\rho_{X_1X_2} = .5$, $\rho_{X_2Y} = .35$.

correlation with its dichotomized form X_c resulting from splitting at any arbitrary cutting point c, as follows:

$$\rho_{XX_c} = \frac{h_c}{\sqrt{p_c(1 - p_c)}}, \tag{19}$$

where h_c is the ordinate of the standard normal density function at c and p_c is the proportion of cases below c. When $c = 0$—that is, when a median split is performed—ρ_{XX_c} takes on its maximum value of .798.

By a logic similar to that discussed above for median splits (see Equations 5–8), it can be shown that

$$\rho_{X_cY} = \rho_{XX_c}\rho_{XY}. \tag{20}$$

Unfortunately, no closed-form solution for the reduction of a correlation resulting from splitting both variables in a bivariate normal distribution at arbitrary cutting points is known. The tables published by Taylor and Russell (1939), however, provide rather precise numerical results for a wide range of values of ρ_{XY} and p_c, with the tables giving the proportions of cases falling in the various quadrants of the resulting 2 × 2 table. These proportions could then be converted into a Pearson r by the usual formula:

$$r = \frac{ad - bc}{\sqrt{R_1R_2C_1C_2}}, \tag{21}$$

where a, b, c, and d are the proportions in the cells, R_1 and R_2 are the proportions in the rows, and C_1 and C_2 are the proportions in the columns.

The complementary problem of going from a 2 × 2 table of proportions to the estimated correlation in a bivariate normal distribution is the problem of a tetrachoric correlation, which is again solved by numerical approximation, usually involving iterative methods (e.g., Dixon, 1988, p. 547). However, Becker and Clogg (1988) recently supplied formulas for approximating the value of the tetrachoric which do not require iteration.

In all of the above derivations the median has been assumed to be known exactly, whereas in real life we have to estimate it from the data. Naturally, the sample median will often be used as the best estimate of the population median, and the expected reduction in correlations when the true median is used is only approximated when an estimated median is used for the split. More precisely, we can say that whenever a certain function of parameters is equal to another parameter, then replacing the first set of parameters in the function by their maximum likelihood estimates will yield the maximum likelihood estimate of the last parameter. Even so, we can consider the question of whether the correlations with dichotomized variables will change substantially if the estimated median differs somewhat from the true median.

To assess this effect we calculated the value of the multiplier $b = \rho_{XX_c}$ present in Equations 19 and 20 for several cutpoint values and its relative difference from .798, the value valid in the case of a split using the true population median. Our calculations revealed that the change of the b multiplier is quite negligible if the cutpoint (the estimated median) does not differ substantially from the true median value. For example, if the difference between the true median and the estimated median is less than one fourth of the population standard deviation (which happens with probability .99 if the sample size is larger than 106), the value of the correct multiplier will be only about 1% less than .798—that is, the correct multiplier will be .790 or greater. In other words, the multiplier of .798 will be within 1% of the (slightly smaller) exact correction factor that should have been used if it had been known that the split was being made at a point slightly different from the true median. Even if the difference between the estimated and true medians is as large as one half of the standard deviation (which it will be with probability less than .01 if the sample size is greater than 26), the relative change of the multiplier will not exceed 5%. Since the presented formulas for correlations with other variables all include the b multiplier (e.g., Equation 16), those correlations will be proportional to it. Therefore, the robustness of b to departures from the true median will be transmitted to the other correlation values as well, provided we dichotomize only one of X_1 and X_2 (e.g., Cases 3, 4, and 5 in Table 3). It is more difficult to assess precisely the robustness of the reduced correlations based on estimated medians in other cases, particularly since we do not have closed-form solutions for the reduced correlations when both X_1 and X_2 are split at arbitrary cutpoints.

Interpretation of Partial Correlations

From a practical standpoint, the error in the multiplier for correlations in bivariate median splits may not seem of great consequence—given the fact that correlations below .5 are more common in educational and behavioral sciences than those above .5. However, it is important to correct the misinterpretation of partial correlations that contributed to and was carried forward by the previously published erroneous formula.

Peters and Van Voorhis (1940) and others apparently believed that a partial correlation can always be interpreted as a correlation between two variables when a third variable is held constant, but that is not the case in general. Here we are drawing a distinction between the partial correlation between X_2 and Y controlling for X_1, which is the correlation between the residuals of the separate regressions of Y and X_2 on X_1, and the *conditional correlation* between X_2 and Y for a given value of X_1, which concerns only those values of X_2 and Y sharing a particular value of X_1. Although one can always compute such conditional correlations as long as there are multiple observations (not all tied) at a given value of X_1, they are not necessarily related to the partial correlation.

By the Peters and Van Voorhis (1940) logic, the partial correlation of X_{1d} and X_{2d} controlling for X_1 would be 0, because the conditional correlation of X_{1d} and X_{2d} would be 0 at each value of X_1. However, in reality the partial correlation in their situation could be substantial. For example, in a bivariate normal distribution with a .9 correlation between X_1 and X_2, the correlation between X_{1d} and X_{2d} partialing out X_1 is .3335. Thus they were wrong at this point, and in fact in two ways. We can set aside the technical point that if the correlation is defined, as in Equation 9, as the ratio of the covariance of two variables to the product of their standard deviations, then the conditional correlation of a variable and a constant is not strictly zero but undefined because of the impossibility of dividing by zero. More importantly, we move now to the development of our final point: Even if all the conditional correlations between the dichotomous variables were 0, the partial correlation would not necessarily be 0.

A key assumption that allows the partial correlation to be interpretable as the conditional correlation is multivariate normality, which implies that X_1 is univariately normally distributed and the distribution of Y and X_2 is bivariate normal at each value of the conditioning variable X_1. Such an assumption is termed a "primary assumption" by Darlington (1991, p. 110), that is, an assumption "whose violation jeopardizes the very meaning of the parameters under study," as opposed to a secondary assumption, violations of which "merely threaten the accuracy of our inferences about that parameter" (Darlington, 1991, p. 134). Obviously, if one of the two variables in the relationship is binary, their joint distribution cannot be bivariate normal. Thus, the partial correlation would not necessarily be zero even if all the conditional correlations were zero, because in the nonnormal case the partial correlation does not always indicate the values of the conditional correlations.

An alternative perspective on the meaning (or lack thereof) of the partial correlation is provided by the concept of conditional independence. Users of normal-theory-based statistics are aware that, in general, a zero correlation does not imply independence unless the variables do have a bivariate normal distribution. On the other hand, the independence of two variables is known to imply that their correlation will be zero regardless of the shape of their joint distribution.

We suspect that many users of partial correlations would believe that the same applies to partial correlations. That is, it seems very plausible if two variables are conditionally independent at each level of a third variable that their partial correlation controlling for the third variable will be 0. A simple counterexample proves that a nonzero partial correlation does not even imply conditional dependence.

Consider the data shown in Table 4. Suppose that at each of three levels of the variable Z, the joint distributions of dichotomous variables X and Y are given by frequencies that are in proportion to the cell entries in the tables. At each of the three levels of Z, variables X and Y are conditionally

TABLE 4
Example demonstrating that conditional independence does not imply zero partial correlation

		Y		Y		Y	
		−1	1	−1	1	−1	1
X	−1	3	9	3	1	4	4
	1	1	3	9	3	4	4
		$Z = 0$		$Z = 1$		$Z = 2$	

Summary XY table and statistics

		Y	
		−1	1
X	−1	10	14
	1	14	10

$r_{XZ} = -r_{YZ} = .204$
$r_{XY} = -.167$
$r_{XY \cdot Z} = -.130$

independent—and their conditional correlations are zero. Overall, X and Y are negatively correlated at $r_{XY} = -.167$. However, the partial correlation of X and Y controlling for Z, instead of being zero, is similar to the unconditional correlation, that is, $r_{XY \cdot Z} = -.130$. In this case, there is some nonlinear relationship between X and Z, and between Y and Z. When Z is controlled, X and Y are residualized only for their *linear* relationship with Z, and thus the residuals still retain some information about Z. This example illustrates a situation where, although variables are not normally distributed, the partial correlation is well defined but different from the relationship suggested by the patterns of conditional correlations at the various levels of the variable being controlled. This state of affairs is like that in Case 8, considered in a previous section, where the fact that the original causal variable has been controlled assures that the two other variables will be conditionally independent, yet despite this their partial correlation is nonzero (see Table 3).

How then are partial correlations to be interpreted? In the case of dichotomous variables, intuitions derived from the multivariate normal scenario do not necessarily apply. However, it is always true that they are correlations of residuals, where the residuals are computed by regressing the original variables on the control variable(s). For example, in the case of dichotomous variables, the residuals when one controls for the original continuous variable will in fact not be constant because the predictions and hence the residuals will depend on the values of the original variable.

Conclusion

The basic message of this article is threefold. First, bivariate median splits do not always result in lower correlations than splitting only one variable, despite published claims to the contrary. Second, the types of spurious effects discussed by Maxwell and Delaney (1993) will tend to show up any time the causal, or conditionally predictive, variable is controlled by using a dichotomous rather than the original continuous form of the variable. Third, when conditional distributions are not bivariate normal, partial correlations must be interpreted with great caution. Perhaps most disconcerting, nonzero partial correlations do not in general mean that the two variables are conditionally dependent.

In terms of practical implications for researchers, if an investigator chooses to dichotomize, for whatever theoretical or practical reason, our results give guidance about likely effects. But in the usual case where dichotomization is undertaken merely in an effort to simplify the analysis of data, our basic advice, like that of other methodologists, is that dichotomization is a poor strategy. Not only can it lower power, it can also lead to false positives. In general, researchers should want to avoid using a method that may lead to incorrect conclusions about a variable's effect.

We would nonetheless concede that it is possible that the decisions reached in all statistical tests of interest in a given situation may be the same whether the predictors are in dichotomous or continuous form. In the multivariate normal situation, this would occur if relationships between variables were relatively weak, and thus the attenuation or inflation resulting from dichotomization is relatively inconsequential. The two kinds of data analysis could also yield similar results when data are not multivariate normal. This is likely to be the case in designs using extreme groups, that is, where subjects are selected because of being either very high or very low on one of the X variables. Alternatively, the relationship between X and Y may in fact be a step function—for example, if all subjects above a critical value on X obtain one score on Y and all those below the critical value obtain a different score on Y. In any of these cases, if the results of the two sorts of analyses lead to the same conclusions, researchers may want to note that fact in reports of their work but structure their discussion around analyses of the dichotomous form of the variables. Note, however, that in this situation it is the simplicity of the presentation or the facilitation of communication of results, rather than the simplicity of data analysis, that is the guiding principle.

Finally, because investigators want to make inferences concerning the relationship between two variables at given levels of the third, we reiterate that if data are not bivariate normal, then partial correlations will not provide the answer to the question of interest. If conditional correlations or conditional independence are of interest, then those issues should be examined directly.

References

Becker, M. P., & Clogg, C. C. (1988). A note on approximating correlations from odds ratios. *Sociological Methods and Research, 16*, 407–424.

Cohen, J. (1978). Partialed products *are* interactions; partialed powers *are* curve components. *Psychological Bulletin, 85*, 858–866.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

Darlington, R. B. (1991). *Regression and linear models.* New York: McGraw-Hill.

Dixon, W. J. (Ed.) (1988). *BMDP statistical software manual* (Vol. 1). Berkeley: University of California Press.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* New York: Wiley.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical distributions: A handbook for students and practitioners.* London: Butterworths.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-difference variables. *Journal of Educational Psychology, 66*, 464–472.

Kendall, M. G., & Stuart, A. (1958). *The advanced theory of statistics.* London: Griffin.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190.

Maxwell, S. E., Delaney, H. D., & Dill, C. A. (1984). Another look at ANCOVA versus blocking. *Psychological Bulletin, 95*, 136–147.

Pearson, K. (1900). Mathematical contributions to the theory of evolution: VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195A*, 1–47.

Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases.* New York: McGraw-Hill.

Pugachev, V. S. (1984). *Probability theory and mathematical statistics for engineers.* Oxford: Pergamon Press.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Rosenthal, R. (1987). *Judgment studies: Design, analysis and meta-analysis.* Cambridge, England: Cambridge University Press.

Tang, T. L. (1988). Effects of Type A personality and leisure ethic on Chinese college students' leisure activities and academic performance. *The Journal of Social Psychology, 128*, 153–164.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565–578.

Authors

ANDRÁS VARGHA is Associate Professor, Department of General Psychology, Eötvös University, Izabella utca 46, H-1064 Budapest, Hungary; vargha@ludens.elte.hu. He specializes in investigations of the robustness of classical statistical methods to violations of their assumptions.

TAMÁS RUDAS is Senior Lecturer and Head, Statistics Group, Institute of Sociology, Eötvös University, Pollack M. ter 10, H-1088 Budapest, Hungary. He specializes in association models for multivariate data and categorical data analysis.

HAROLD D. DELANEY is Professor and Associate Chair, Department of Psychology, University of New Mexico, Albuquerque, NM 87131. He specializes in applied statistics and individual differences.

SCOTT E. MAXWELL is Professor and Chair, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. He specializes in quantitative psychology.

Received August 8, 1994
Revision received February 28, 1995
Accepted April 10, 1995