The Cost of Dichotomization

Jacob Cohen
New York University

Assuming bivariate normality with correlation r, dichotomizing one variable at the mean results in the reduction in variance accounted for to $.647r^2$; and dichotomizing both at the mean, to $.405r^2$. These losses, in turn, result in reduction in statistical power equivalent to discarding 38% and 60% of the cases under representative conditions. As dichotomization departs from the mean, the costs in variance accounted for and in power are even larger. Consequences of this practice in measurement applications are considered. These losses may not be quite so large in real data, but since methods are available for making use of all the original scaling information, there is no reason to sustain them.

It is a frequent procedure in the behavioral and social sciences to dichotomize "continuous" (more accurately, graduated) variables. Some researchers follow this practice in order to simplify the data analysis, which it undoubtedly does. The increasing popularity of loglinear models has also promoted this practice among researchers eager for state-of-the-art methodology. Others justify it on the grounds that (1) the variable in question is too crude to warrant the refinement of the original graduated scale and that (2) dichotomization more truly represents the modest measurement content of the variable. The latter is a misconception: In fact, to the measurement error in the original scale, dichotomizing simply adds errors of discreteness. That is, the amount of (unmeasured) true score variance for the cases at each of the two points of the dichotomy is necessarily greater than it would be for cases at each of the multiple points in the original scale.

Humphreys and Fleishman (1974) have criticized dichotomization in the context of its application to individual difference variables in analysis-of-variance designs. Cohen and Cohen (1983) have inveighed against this practice quite generally, arguing that it results in underestimating effect sizes and reducing the power of statistical hypothesis tests. Specifically, they have claimed that dichotomization results in proportions of variance accounted for that are some .64 (or less) as large when it is performed on one of the two variables being correlated, and only .40 (or less) as large when both variables are dichotomized. However, they offered neither reference nor proof.

The Effect on the Population r

This problem can be viewed as the extreme case of what has been called "broad" or "coarse" grouping. Before computers, statistical computation was facilitated by summarizing data arrays as frequency distributions over 10 or 12 intervals with the means or midpoints of the latter used to represent the cases in each interval. To offset the resulting inaccuracy, Peters and Van Voorhis (1940,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 249–253

0146-6216/83/030249-05$1.50

pp. 393–399) showed how correlation coefficients computed from such grouped data could be "corrected for broad categories" and provided a proof and a table of values for effecting such corrections for 2 to 15 intervals under various assumptions.

Assume a bivariate normal population whose product-moment correlation equals r. If one variable, say X, is dichotomized at the mean (or median) so that two equal "broad intervals" result, the observed correlation between the binary X_d and Y will equal $.798r$, which results in X_d accounting for only $(.798^2 =)$ .637 as much Y variance as the original X does. Note that by dichotomizing X, the resulting correlation is the same as the point-biserial correlation.

The .798 value from Peters and Van Voorhis (1940), as noted, holds for equal intervals, hence dichotomization at the mean. An alternate proof can be developed, which proceeds by finding the ratio of the point-biserial r to the biserial r (i.e., the bivariate normal population's r) and results in the general multiplying constant

$$e = h/[p(1 - p)] , \qquad [1]$$

where h is the ordinate of the standard unit normal curve and p is the proportion of the cases in either of the two intervals. For dichotomization at the mean, $h = .3989$, $p = .50$, so $e = .798$, agreeing with the Peters and Van Voorhis (1940) value. As the cut departs from the mean, e decreases. For example, at .5 *SD* from the mean, where $p = .6915$ and $h = .3521$, $e = .762$ and $e^2 = .581$. At 1 *SD*, often used to define "extreme" or "clear-cut" cases, $p = .8413$, $h = .2420$, so $e = .662$ and $e^2 = .439$. Beyond 1 *SD*, the loss in variance accounted for is quite precipitous: At 1.5 *SD*, for example, $e^2 = .269$. Concretely, then, for dichotomization at the mean and at .5, 1.0, and 1.5 *SD* from the mean, an r^2 of .16 becomes an r^2 of X_d with Y (which equals e^2r^2) of .102, .093, .070, and .043, respectively; r^2 is cut in half (i.e., $e^2 = .5$) when $p = .79$ (i.e., at .81 *SD*).

The Effect on the Sample *r* and on the *t* test

A sample from a bivariate normal population cannot be exactly bivariate normal but will approach that form asymptotically. Therefore, where r_s would have been obtained for the sample on the original graduated data, dichotomizing X results approximately in er_s, with the obvious deleterious effect on the value of t when the null hypothesis is tested. Instead of testing the significance of r_s via the standard

$$t = r_s [(n - 2)/(1 - r_s^2)]^{\frac{1}{2}} , \qquad [2]$$

the test is performed on a value approximately er_s (i.e., the point biserial r)[1],

$$t_d = er_s [(n - 2)/(1 - e^2r^2)]^{\frac{1}{2}} , \qquad [3]$$

which, expressed as a proportion of t (i.e., dividing Equation 3 by Equation 2) is

$$q = e [(1 - r_s^2)/(1 - e^2r^2)]^{\frac{1}{2}} , \qquad [4]$$

so that the significance test t_d yields a smaller value than t.

For any given value of r_s, the maximum value of q comes at the maximum value of e, .798 at $p = .50$. Maximally, then, the fraction of t obtained following dichotomization is approximately

$$q_{max} = .798[(1 - r_s^2)/(1 - .637r_s^2)]^{\frac{1}{2}} . \qquad [5]$$

Concretely, for dichotomization at the mean, the proportion of t obtained varies from .78 at $r_s = .2$ to .62 at $r_s = .7$. With unequal cuts, e is no longer at its maximum so the q values are smaller still: e.g., at .5 *SD* from the mean (where $p = .69$ and $e = .762$), the fraction of t that results decreases from .76 to .59 (again as r_s varies from .2 to .7). For large n and/or large r_s, a t that is three-fifths to three-quarters as large as it should be may still be large enough to be statistically significant. However, for the sample sizes and rs that prevail in much of behavioral science, the cost of "simplifying" analyses by dichotomization may well be too dear.

The Effect on Statistical Power and on Effective and Necessary Sample Sizes

The preceding was an indirect demonstration of loss of statistical power caused by dichotomization.

[1]This is identically the value of t (and for the same $df = n - 2$) that would be obtained from a test of the difference between Y means of the two groups of X_d (Cohen & Cohen, 1983).

A direct assessment can proceed by means of the methods given in Cohen (1977, pp. 75–107, 457–458). If a bivariate normal population whose $r = .30$ (operationally defined as a "medium" effect size) is assumed, a test of the hypothesis that $r = 0$ at the two-tailed .05 level performed on a sample of 80 cases has a probability of rejection (power) of .78, i.e., a reasonably good chance. As has been seen, dichotomizing X at the mean results in the reduction of the population correlation to $.798r = .798(.30) = .239$. The power is now reduced to .57, virtually a coin toss. Dichotomization has here produced power of .57 for $n = 80$, but on the original graduated scale where $r = .30$, power of .57 is achieved for $n = 50$ (Cohen, 1977, pp. 92–93). Thus, dichotomizing is equivalent to throwing away 30 of the 80 cases. For $n = 60$, dichotomization reduces power from .65 to .45, the latter obtainable on the graduated scale at $n = 37$, a willful loss of 23 of the 60 cases. For $n = 40$, power is reduced from .47 to .31, which was obtainable on the original scale at $n = 25$, equivalent to discarding 15 cases. The loss in effective sample size when X is dichotomized at the mean is about 38% for these examples and remains at approximately this level over the range $r = .2$ to .5 and for both .01 and .05 level tests.

As the dichotomization departs from the mean, e is reduced from .798 and with it the power and effective sample size. For example, when X is dichotomized at 1 SD,, e is about two-thirds (.6623); the population correlation falls to .20; and at $n = 80$, 60, and 40, power falls to .43, .34, and .24, respectively. These values, in turn, are available on the original scale at ns of 35, 27, and 19, respectively, so that extreme (1 SD) dichotomization has cost the equivalent of 55% of the cases. The effective loss of n remains at about 55% over the range of population r from .2 to .5 and at both the .01 and .05 level at the 1 SD point of cut.

Yet another way to set the price tag on dichotomization is to reverse the above and to determine the *increase* in sample size needed to *offset* dichotomizing. If, again, a population $r = .30$ is assumed, the sample size needed for power to equal .80 for a two-tailed .05 test is 84. For "optimal" dichotomization at the mean, the resulting er of .239 requires 133 cases under the same conditions, an increase in the necessary sample size of 58%. For $r = .20$, the rise is from 193 to 304, and for $r = .40$ from 46 to 74, about the same percentage increase. For dichotomization at 1 SD under the same conditions, the n required increases about 130%.

The Effect of Dichotomizing Both Variables

What happens when both X and Y are dichotomized? This is what is done when the bivariate normal distribution is reduced to a fourfold table. Paralleling the previous argument, the product-moment correlation between X_d and Y_d (each coded, e.g., 1, 0) is the familiar phi coefficient, and its relationship to the original r (effectively, to the tetrachoric r) can be found by following Peters and Van Voorhis (1940, pp. 395–398). With dichotomization of both X and Y at their means, the r between X_d and Y_d equals .637 r. (The constant .637 here is $.798^2$, the result of applying the .798 correction twice.) Expressed in proportion of variance terms, the effect of double dichotomization at the means is to reduce r^2 to ($.637^2 r^2 =$) $.405\ r^2$.

The consequence to power of reducing r to $.637r$ for double dichotomization at the means can be translated, as before, into reduction in effective sample size. As noted above, when the population r is .30, a sample of 80 cases has power = .78 to reject the null hypothesis at the two-tailed .05 level. For the reduced r of $.637(.30) = .191$, power is .39, a level attainable for 32 cases for the original $r = .30$. For $n = 60$, power is reduced from .65 to .31, the latter attainable on the original scale at $n = 25$; for $n = 40$, the reduction in power is from .47 to .21, attainable originally for $n = 16$. Thus, double dichotomization at the mean is equivalent to discarding 60% of the cases here, and this 60% loss holds approximately over the range $r = .20$ to .50 and at both the two-tailed .01 and .05 levels.

It is unnecessary to undertake a proof that with dichotomization away from the mean on X and/or Y, r for the dichotomies is lower still, or to pursue in detail the effects on power. They are clearly grimmer than those seen for dichotomizing one

variable. It is no exaggeration to say that double dichotomization may result in the loss of as much as two-thirds of the proportion of variance that could be accounted for on the original variables, with a resulting loss of power equivalent to throwing away as much as two-thirds of the sample.

It is time to retrieve a hostage left behind in the proof—the assumption of bivariate normality. This is, of course, a convenient fiction: No real data can be exactly normally distributed, since (if for no other reasons) X can neither be literally continuously distributed nor go to infinity. However, there is abundant evidence throughout applied psychometrics and statistics that the failure of the normality assumption, unless extreme, bears only marginally on the validity of the conclusions drawn. It is, of course, possible to construct nonnormal bivariate distributions where dichotomization results in an increase in r over that of the original graduated variables; but these will be characterized by extreme skewness, heteroscedasticity, and curvilinearity, e.g., step functions. Needless to say, such distributions are uncommon in the behavioral and social sciences. When they do occur, dichotomization is a far inferior approach to one that uses all the measurement information in the original graduated data and tackles the curvilinearity directly, e.g., polynomial regression (Cohen & Cohen, 1983, chap. 6).

The accuracy of the three-digit values for the multiplying constants does, of course, depend on the normality assumption. In real data, which lack the exact form or infinite number of gradations of a theoretical normal distribution, they may be 10% to 15% higher. This still implies losses of one-fifth to one-third of r^2 for single and double dichotomization at the mean(s), and more as cuts become more extreme.

Discussion

The reduction in r due to dichotomization is centrally a measurement issue. It seems self-evident that the inevitable effect of dispensing with the score differences within each of the two portions of the distribution, leaving only the distinction between the two, is the loss of a considerable amount of measurement information. It is this loss or degradation of measurement information that produces the drop in r (or, for that matter, any measure of effect size).

This loss from dichotomization should not be confused with the familiar attenuation due to classical random measurement error. Dichotomization results in the systematic loss of measurement information; and the foregoing development requires only the condition of bivariate normality to produce the drop from r to er. Therefore, the dichotomization drop will occur for both true and observed scores. For observed scores, the already measurement-error-attenuated r drops further to er, while a measurement-error-free r' drops to er', with e depending only on the point of cut (Equation 1).

The test construction technology developed a half-century ago frequently employed dichotomization for criteria (e.g., total score became "high-low," final grade became "pass-fail") in the interest of computational efficiency. The test constructors of the precomputer era usually corrected upward the resulting drop to er by determining biserial or tetrachoric correlations. Failure to do so was not serious, since they used the item-criterion correlations primarily to order the items, discarding those with the lowest values or selecting those with the highest. Fully justifiable for that time and purpose, this practice of our grandparents may have left as an unconscious residual among contemporary psychologists and other behavioral scientists a casual readiness to dichotomize that is neither appropriate nor justifiable. Consider some typical examples.

1. Dichotomization in order to apply loglinear models has already been mentioned. A similar practice is to dichotomize on a control variable used for "blocking" (matching) in analysis-of-variance designs. Since blocking is equivalent to partialling, the reduction to er may produce the same insidious kind of distortion that occurs when an unreliable variable is partialled (Cohen & Cohen, 1983, chap. 10). An irony here is that the researcher may dichotomize the blocking variable or control variable *because* it is unreliable, thus making a bad situation worse.
2. Occasionally, it is found that a batch of scaled

items is dichotomized preparatory to a factor analysis. The resulting phi coefficients and the factor loadings they yield are approximately two-thirds as large as the product-moment correlations on the original data would have been and the communalities (however estimated) less than half as large. The standard normal varimax rotation, which is actually performed on communality-adjusted loadings and then scaled back, is quite likely to be distorted because small errors in small communalities can produce large differences in the loadings that actually determine the rotation.

3. The dichotomization of an attitude scale item is frequently encountered in social or political science research, say, one that provides a 4- to 6-point Likert type agree-disagree response scale. The dichotomization may be effected at the middle of the scale, in the interest of simplifying the analysis or display of the results, or (worse) near one end, to sharpen the distinction between extreme cases and the rest of the distribution. A similar practice is endemic in marketing and advertising research, where the "top box," i.e., the most favorable category of a multipoint scale, is distinguished from all the others.
4. Dichotomization is also resorted to frequently in psychiatric research. Symptom or behavior scale items, responded to on a scale of degree (e.g., from "not at all" to "severe") or frequency (e.g., from "never" to "always") are dichotomized at the extreme point, in an effort to assure that the symptom is "actually" present or that only bona fide or "clear-cut" cases are identified. When bivariate distributions involving such scales are plotted, they are almost invariably linear, demonstrating that the dichotomization has indeed produced a loss of measurement information.

Readers can undoubtedly supply additional examples of their own.

To summarize, the cost in the degradation of measurement due to dichotomization is a loss of one-fifth to two-thirds of the variance that may be accounted for on the original variables, and a concomitant loss of power equivalent to that of discarding one-third to two-thirds of the sample. Such losses cannot be justified, given the availability of methods that fully exploit all the original measurement information (Cohen & Cohen, 1983).

References

Cohen, J. *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press, 1977.

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale NJ: Erlbaum, 1983.

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

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

Acknowledgment

The author gratefully acknowledges Patricia Cohen's critical reading and discussion of this article.

Author's Address

Send requests for reprints or further information to Jacob Cohen, Department of Psychology, New York University, 6 Washington Place, New York NY 10003, U.S.A.