Breaking Up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data

David L Streiner, PhD[1]

Researchers often take variables that are measured on a continuum and then break them into categories (for example, above or below some cut-point), either to place subjects into groups or as an outcome measure. In this article, we show that the rationales given for this practice are weak and that categorization results in lost information, reduced power of statistical tests, and increased probability of a Type II error. Dichotomizing a continuous variable is justified only when the distribution of that variable is highly skewed or its relation with another variable is nonlinear.

(Can J Psychiatry 2002;47:262–266)

Key Words: *dichotomizing, data, power, variables*

Those of you who are old enough may remember Neil Sedaka singing "Breaking Up is Hard to Do." If only that were true when it comes to the variables we use in research! Many times (I would say far too many), a researcher uses a continuous measure, such as a depression inventory, as an outcome variable and then dichotomizes it—above or below some cut-point, for example, or the number of people who did and did not show a 50% reduction in their scores from baseline to follow-up (1). Less often, but again far too frequently, researchers may assign patients to different groups by dichotomizing or trichotomizing scores from a continuous scale.

Over the years, several arguments have tried to justify this practice. Perhaps the most common one runs something like this: "Clinicians have to make dichotomous decisions to treat or not to treat, so it makes sense to have a binary outcome." Another rationale that is offered is, "Physicians find it easier to understand the results when they're expressed as proportions or odds ratios. They have difficulty grasping the meaning of beta weights and other indices that emerge when we use continuous variables." In this article, I'll try to show that you pay a very stiff penalty in terms of power or sample size when continuous variables are broken up, with the consequent risk of a Type II error (that is, failing to detect real differences). But before we begin, let me assume the role of a marriage counsellor and see whether the arguments in favour of splitting up are really viable.

The rationale for dichotomizing outcomes because clinical decisions are binary fails on 3 grounds. The primary one is that it confuses measurement with decision making. The purpose of most research is to discover relations—relations between or among variables or between treatment interventions and outcomes. The more accurate the findings, the better the decisions that we can make; that is, the findings come first and the decision making follows. As we will see, findings come more readily and more accurately when we retain the scaling of continuous variables. The second reason is that all the research using the old dichotomy becomes useless if the cut-point changes. For example, the definition of hypertension used to be 160/95 (2). If we defined the outcome of intervention trials dichotomously—with above 160/95 being hypertensive and below being normotensive—then those findings would become useless after the definition changed to 140/90 (3). If we expressed the outcome as a continuum, however, the values of beta coefficients and similar indices showing the effects of various risk and protective factors would not change at all: if we wanted to use statistics such as odds ratios (ORs) or the percentage of patients who improved, it would be a trivial matter to recalculate the results. We have a similar situation in

psychiatry. The diagnosis of antisocial personality disorder (ASP), for example, is a binary one: the person either does or does not satisfy the diagnostic criteria (that is, a certain number of symptoms are present). However, Livesley and others maintain that ASP and many other disorders should actually be seen as a continuum: the more symptoms that are checked off, the more of the trait the person has (4). If the number of symptoms necessary to meet the criteria were to change, as occurred when DSM-IV replaced DSM-III-R, then much previous research using a dichotomous diagnosis would have to be discarded. If the diagnosis were expressed as the number of symptoms present, though, it would be relatively easy to reinterpret the findings using the new criteria.

Finally, whether to hospitalize a patient with suicidal ideation or to discharge a patient with symptoms of schizophrenia may be binary decisions, but many treatments—perhaps most—fall along a continuum involving the dosage or strength of a medication and the number and frequency of therapy sessions.

As for the argument that physicians are more comfortable with statistics based on categorical measures, we are likely dealing with both a base canard that they, like old dogs, cannot learn new tricks and a vicious circle. As long as the belief persists, studies will be designed, analyzed, and reported using proportions and ORs, meaning that physicians will not have the opportunity to become more comfortable with other approaches.

First, I'll give some examples of how dichotomizing can lead us astray, and then I'll use these examples to discuss why this is the case.

Example 1

Let's look at the data in Table 1, which shows scores on a scale for 2 groups, each with 10 subjects. Let's assume that, if we were to dichotomize the scale, we would use a criterion for "caseness" of 15/16: people with scores from 1 to 15 would be considered normal, and those with scores of 16 and over would be defined as cases. The mean for Group 1 is 11.70, and the mean for Group 2 is 16.80. There is slightly more than a 5-point difference between the groups, and the average of the first group is well below the cut-off of 15/16, while the average of the second group is above the cut-point. If we used a *t*-test to compare the groups, we'd find that $t(18) = 2.16$, $P = 0.045$. That is, there is a statistically significant difference between the means. Now, let's dichotomize the results and count the number of people above and below the cut-point in each group. What we'd find is shown in Table 2. Because 2 of the cells have frequencies below 5, we'd use a Fisher's exact test, rather than a chi-square test, and we'd find that the *P* level is 0.057. In other words, the difference is not statistically significant.

Example 2

In the second example, we have 40 subjects, measured on 4 variables, *A* through *D*. If we were to correlate these variables, we'd find the results shown in the upper triangle of Table 3. Of the 6 correlations, 5 are significant at the $P < 0.01$ level. Now, we'll do a median split on each of these variables, so that roughly one-half of the subjects fall above, and one-half below, the cut-point. If we reran the correlations, we would find the results in the lower triangle of the same table. In every case, the correlations are lower—sometimes substantively

Table 1 Data on an outcome measure for 2 groups

Group 1	Group 2
9	16
14	12
13	25
12	14
5	16
5	9
12	10
22	23
13	22
12	21

Table 2 The number of people in each group above and below a cut-point of 15/16

	Group 1	Group 2
15 and below	9	4
16 and below	1	6

Table 3 Correlations among 4 variables[a]

	A	B	C	D
A		0.59[b]	0.56[b]	0.70[b]
B	0.3		0.28	0.84[b]
C	0.16	0.25		0.39
D	0.45[b]	0.55[b]	0.2	

[a]Correlations above the diagonal are for the variables treated as continual; below the diagonal as dichotomized
[b]$P < 0.01$

so—and only 2 of the 6 correlations are significant at the $P < 0.01$ level.

Taking this example a bit further, we can run a regression equation, with *A* as the dependent variable (DV) and *B* through *D* as the predictors. Keeping the variables as continua, we'd find the multiple *R* is 0.767 and $R^2 = 0.588$, which would lead to thoughts of publication and promotion for most people. If we dichotomized the variables, however, we'd find that the multiple *R* is 0.460, with an associated R^2 of 0.211, which might jeopardize that promotion by at least a year. (Purists might say that we should really use a logistic regression with a dichotomous DV. If we did, we'd find the Cox and Snell pseudo-R^2 to be an even more disappointing 0.20.)

Why This Occurs

These examples illustrate 2 points. First, the magnitude of the effects (for example, the differences between groups, the correlations between variables, and the amount of variance explained by the regression) were lower—sometimes dramatically lower—when we took continuous variables and treated them as dichotomies. Second, findings that were significant using continuous variables were sometime not significant when we dichotomized those variables. Let's examine each of these issues separately.

Dichotomizing variables results in a tremendous loss of information. If the values in example 1 were scores on a Beck Depression Inventory (BDI), the possible range would be between 0 and 69. When we dichotomize this scale, we are saying, in essence, that there is no difference between a score of 0 and one of 15 (both would be coded as 1), nor between scores of 16 and 69 (both coded as 2). At the same time, we are making a qualitative difference between scores of 15 and 16. This doesn't seem conceptually logical and ignores the problem of measurement error. As we discussed in previous articles in this series (5,6), every observed score (for example, a numerical value on a questionnaire, a blood level, or the number of diagnostic criteria that are satisfied) is made up of 2 parts: a "true" score, which is never seen, plus some error. The more reliable the scale, the smaller the error and the closer the observed score to the true score. But, since no measurement has a reliability of 1.00 (and this includes lab tests as well as paper-and-pencil ones), every score has some degree of error associated with it. We also assume that the errors are random and have a mean of 0; that is, over a large number of people or over many observations on the same person (or both), the errors will tend to cancel each other out. This means that, if we treat the scores as numbers along a continuum, we may misplace a person to some degree, and this will be reflected in, for example, a lower correlation between the scale score and some other variable. However, because the errors are random with a mean of 0, there will not be any bias in the relation.

But the situation is different when we dichotomize the scale. Now, for people near the cut-point, the measurement error may result not just in a score that's slightly off but in their being misclassified into the wrong group. A person suffering from depression, with a true score of 16 and a relatively small error of –1 point, would end up in the group without depression. Thus, we can see that using a scale as a continuum will present us with some degree of random error (which is inevitable), but dichotomization can easily result in misclassification error.

Another reason dichotomizing variables puts us behind the eight ball is a function of the statistical tests themselves. All statistical procedures can be seen as a ratio between a signal and noise (6). The "signal" is the information that we've captured in the measurement—the difference between group means, the relation between 2 variables, and so forth. The "noise" is the error, usually captured by the differences among the subjects within the same group (when we're comparing means), deviations from a linear relation (in correlational tests), or misclassifications (in procedures such as chi-squared, ORs, and relative risks). As we mentioned, dichotomization results in a loss of information, so that the "signal" is weaker than when we use continua. Not surprisingly, tests based on dichotomous variables are generally less powerful than those based on continuous variables. Suissa (7) determined that a dichotomized outcome is at best only 67% as efficient as a continuous one; that is, if you need 50 subjects in each group to demonstrate statistical significance with a continuous scale, you would need 75 subjects per group to show the same effect after dichotomizing. In fact, though, most clinical scales are split at a clinically important point that doesn't usually correspond to the best place from a statistical point of view, with the result that the efficiency rarely approaches even 67% and may drop to as low at 10% (that is, the required sample size is 10 times as large). Similarly, if the dichotomy is statistically ideal, resulting in one-half the people being in one group and one-half in the other, the correlation of that variable with another one is reduced by about 20%. The more the split deviates from 50–50, the more the correlation is reduced. By the time the division is 90–10, the correlation is reduced by 41% (8).

It's Not All Bad

Up to now, we've treated categorization of a continuum as an unmitigated disaster with no redeeming features. At the risk of appearing to be a Pollyanna who can find positive things to say about the worst situations, there are in fact a few situations

wherein we actually should divide a continuous variable into a dichotomy or an ordinal variable. These, though, are based on statistical considerations; they are not based on clinical considerations or on what is convenient.

Most parametric statistical tests assume that the variables are normally distributed. While we can often get away with variables that deviate from normality to some degree (and, as Micceri has shown, almost all do [9]), there are limits. One of these is found when a variable resembles a J-shaped distribution; that is, most of the subjects clump at one end, and the rest trail off in the opposite direction. This occurs most frequently if there is a "wall," or limit, at one end but not at the other. For example, a population survey may find that most people have had no psychiatric admissions, and a small proportion have had a single admission. Then the numbers trickle off, with a few people having a large number of admissions. There's a lower limit, in that you can't have fewer than 0 admissions, but no upper limit. We can try to transform the variable, but if it's very highly skewed even this won't help. The only solution is to dichotomize (none versus any) or trichotomize it (none, 1 to 2, 3 or more, for example), and treat it as an ordinal variable.

Similarly, we may feel that the relation between 2 variables is not linear. For example, we may suspect that, within the range of low income (say up to $10 000 a year), the actual dollar amount is unimportant, insofar as it buffers against stress, while above a certain amount ($60 000, for example), more money doesn't provide more protection. Within the middle range, however, we may suspect that there is a linear relation. In other words, the relation between income and buffering looks like an elongated S. We can try to model this with a complicated, higher-order equation, but it's often easier to divide income into 3 categories, and again, treat it as if it were an ordinal variable.

Conclusions

Except when the variable deviates considerably from normal, splitting a variable into categories results in lost information, the requirement to use less powerful nonparametric tests, and increased probability of a Type II error. We are most often much further ahead to retain the continuous nature of the variable and analyze the data using the appropriate statistics.

This discussion has focused primarily on taking data that were gathered as continua and then splitting them into categories. The other side of this is that we should gather the data as continua whenever possible. For example, an item on a questionnaire might look like the following:

How old were you on your last birthday?

- ❒ 15–19
- ❒ 20–29
- ❒ 30–39
- ❒ 40–49
- ❒ 50–59
- ❒ 60–65
- ❒ Over 65

It would be better, however, to ask the question, "How old were you on your last birthday?": _____ years.

If you use the first format, you lose fine-grained information, and you're forced to use those categories in all subsequent analyses. With the second format, you can later split age any way you want (although I don't know why you would want to, after all that's been said), and you have all the advantages of a continuous variable. The only possible exception maybe income: people may feel more comfortable reporting it within a range, rather than reporting the exact amount, but the jury is still out on this.

So, in conclusion, the one word of advice about turning continuous variables into dichotomies is—don't!

References

1. Keller MB, McCullough JP, Klein DN, Arnow B, Dunner DL, Gelenberg AJ, and others. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. N Engl J Med 2000;342:1462–70.
2. Arterial Hypertension. Report of WHO Expert Committee. Technical Report Series, No. 628. Geneva: WHO; 1978.
3. Zanchetti A. Guidelines for the management of hypertension: the World Health Organization/International Society of Hypertension view. J Hypertens Suppl 1995;13:S119–S122.
4. Livesley WJ, Schroeder ML, Jackson DL, Jang KL. Categorical distinctions in the study of personality disorder: implications for classification. J Abnorm Psychol 1994;103:6–17.
5. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry 1993;38:140–8.
6. Norman GR, Streiner DL. Biostatistics: the bare essentials. 2nd ed. Toronto: BC Decker; 2000.
7. Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol 1991;44:241–8.
8. Hunter JE, Schmidt FL. Dichotomization of continuous variables: the implications for meta-analysis. J Appl Psychol 1990;75:334–49.
9. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol Bull 1989;105:156–66.

Manuscript received July 2001 and accepted September 2001.

[1]Director, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care; Professor, Department of Psychiatry, University of Toronto, Toronto, Ontario.

Address for correspondence: Dr DL Streiner, Director, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, ON M6A 2E1

e-mail: dstreiner @ klaru-baycrest.on.ca

This is the 21st article in the series on Research Methods in Psychiatry. For previous articles please see Can J Psychiatry 1990;35:616–20, 1991;36:357–62, 1993;38:9–13, 1993;38:140–8, 1994;39:135–40, 1994;39:191–6, 1995;40:60–6, 1995;40:439–44; 1996;41:137–43, 1996;41:491–7, 1996;41:498–502, 1997;42:388–94, 1998;43:173–9, 1998;43:411–5, 1998;43:737–41, 1998;43:837–42, 1999;44:175–9, 2000;45:833–6, 2001;46:72–6, 2002;47:68–75

Résumé: La séparation est pénible : le malaise de la dichotomie des données continues

Les chercheurs prennent souvent des variables qui sont mesurées sur un continuum, puis les divisent en catégories (par exemple, au-dessus ou en-dessous d'un certain point de découpage), soit pour placer les sujets en groupes, soit pour mesurer un résultat. Dans cet article, nous démontrons que les fondements justifiant cette pratique sont faibles et que la catégorisation entraîne une perte d'information, une efficacité réduite des tests statistiques et une probabilité accrue d'une erreur de type II. La dichotomie d'une variable continue ne se justifie que lorsque la distribution de cette variable est très asymétrique ou que sa relation avec une autre variable est non linéaire.