East Carolina University
Department of Psychology

Dichotomizing Continuous Variables: A Bad Idea

I posted the following query on the EDSTAT-L list early in 2003:

When interested in the relationship between two continuous variables, some researchers will dichotomize one of them prior to analysis. I generally discourage such dichotomization, but the practice is common. A colleague asked me today about the practice of dichotomizing by a median split (top half versus bottom half) versus the practice of using only the tails (bottom third versus top third, for example). That is, if you are going to dichotomize a continuous subject variable and compare the resulting two groups on a second continuous variable, even though that is not generally a good idea, is it more useful (less destructive) to use a median split (upper half vs lower half) or to compare the tails (such as upper third versus lower third)?" I suggested to my colleague that this would depend, in part, on the form of the relationship between the two continuous variables (not necessarily strictly linear), and reminded him that throwing out the middle of the distribution would reduce *N* and thus might reduce power too. I vaguely recall having read an article or two on this matter long ago (not the recent articles on why not to dichotomize, but rather on how best to do it if you feel you must), but cannot put my finger on the article(s). Can any of you all?

Here are some of the interesting responses I got:

Dennis Roberts quickly made several comments disparaging the practice of such dichotomization, including:

- Why toss away information from the data?
- If you use top 1/3 and bottom 1/3 ... you are also throwing data away ... which is worse than just lowering the information value of it.

David Howell noted:

- There is an excellent paper on median splits by MacCallum et al. in *Psychological Methods, 2002, 7,* 19-40. [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.] KLW: This is required reading for my students.
- There is also an equally good paper by Julie Irwin and Gary McClelland in a marketing journal.
- Both papers agree with Karl's advice. "When you think about a median split, DON'T."

Gary McClelland added much detail:

Consider the model Y = a + *b* X + error. A key component of calculating the standard error of the estimate of *b* and its confidence interval is *N* * V(X). Tradeoffs between *N* and the Variance of X are exact. We can use this to examine the effect of (a) splitting X at its median or (b) using only the upper and lower 3rds of the distribution.

Note that no matter what the distribution of X, the usual regression provides an unbiased least-squares estimate of the coefficient *b*. In particular if we split the observations on the basis of X, to

compute the mean of Y, if we also compute the mean of X within each subgroup and use that as the predictor values in a regression, we will still get an unbiased estimate of *b*, but with a different standard error. Comparing the standard error for the continuous X and the split X allows an examination of the effects of splitting.

(a) **Median split**. Let's assume a standard normal distribution for illustration. If we split at the median, this will also be splitting at the mean. The mean value of X in the lower half of the distribution is -Sqrt[2/Pi] = -.8 and the mean for the top half of the distribution is Sqrt[2/Pi] = .8. The new variance is 2/Pi = .636. All the components of estimating the standard error of *b* will be the same for both the continuous and the split model except

N V(X) = *N* (for the standard normal)
in the continuous model will be replaced by
N .636V(X) = **.636** *N* (for the standard normal)
This is the same proportion by which the r^2 will be reduced and this value has appeared in numerous articles criticizing the splitting of data.

(b) For the case of using the **upper 1/3 and lower 1/3 of cases**. For a standard normal distribution the mean of the lower 1/3 of the values of X is -1.09 and the mean for the upper half is then +1.09.

So the variance is 1.09^2 = 1.19. But we've also lost 1/3 of our cases, hence, the term ***N* V(X) = *N*** is replaced in the thirds model by **(2/3) *N* 1.19 V(X) =** .79 ***N***

Thus, in terms of the standard error and the confidence interval width, using only the top 1/3 and bottom 1/3 of the data is not as destructive as median splits, but it is still a bad idea.

Furthermore, for modest sizes of *N*, the loss of 1/3 of the degrees of freedom might substantially increase the value of the critical *t*. In other words, **the thirds model will have substantially less statistical power.**

The message, repeated in numerous methodological articles, and well known by Pearson in 1900 is that (a) throwing away information about your variable is never a good idea and (b) throwing away observations in the middle of the distribution is never a good idea.

I've always thought a physicist considering the not uncommon practice in the social sciences of doing median splits or using discrete cutoffs of a continuous variable would think our practices are crazy and unscientific. Last fall I had the opportunity to observe a confirmation of my hypothesis when a Nobel-prize winning physicist sat on the honors committee of one our psychology students. She was studying reading disability and, as is not uncommon in that field, defined the reading disabled as those below the 10th percentile.

The physicist gently but firmly pointed out that surely that was a bad idea and that it would obviously be better to leave a continuous variable as a continuous variable.

So, resist the temptation to split. Leave your continuous variables continuous.

Useful reading:

- Irwin, J.R., & McClelland, G.H. (2003)
- MacCallum R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002).

both provide a lot of the earlier references. I know of no published article using statistical arguments to support splitting continuous variables.

gary
gary.mcclelland@colorado.edu

and, later,

If the relationship is nonlinear, then dividing into two groups, whether the extreme third tails or median splits, precludes any possibility of detecting the nonlinearity. Furthermore, Maxwell & Delaney (*Psych Bulletin*, 1993, *113*, 181-190) demonstrate that obscuring nonlinearity in that way can produce a spurious interaction. Why anyone continues to split data after that article is beyond me, but subsequent articles like the recent MacCallum et al. article in *Psych Methods* (indeed a gem) remain necessary. Irwin & McClelland (*Journal of Marketing Research*, 2003) squashes another false belief that perhaps median splits are a good idea when the predictor variables are very skewed, non-normal distributions. Even in those situations, splitting the data remains a bad idea.

Gary has also provided a nice visual showing the effect of dichotomization -- check it out at http://psych.colorado.edu/~mcclella/MedianSplit/

Dale Glaser added:

Cohen's oft-cited article: Cohen, J. (1983) The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.

Werner W. Wittmann [wittmann@tnt.psychologie.uni-mannheim.de] contributed:

Both strategies are bare nonsense ("blanker Unsinn" in the language of your ancestors). Others have said that already, but I have a different reason from a psychometric stance. Median splits lead to an underestimation of the variable's *SD* one splits. That restriction of range leads to an underestimate of the effect size *r*. **Using the upper and the lower thirds leads to enhancement of range (the *SD* of the variable split gets larger than using all data points) and to an overestimate of the effect size *r*.** (BTW:The calculation of Gary McClelland in his last posting saying that "the extreme thirds will reduce the expected r^2 to 79% of what it would have been" therefore must be wrong, probably due to not using the continuous information of all remaining data points). The multiplier, say S, which is biasing the estimate is basically a function of the quotient of the *SD* of the restricted/enhanced *SD* in relationship to the original SD and the correlation r (don't have the exact equation handy, but it can be found in Hunter & Schmidt's textbook about meta-analysis or in the good old Gulliksen).

Who wants to have confidence intervals around estimates, which one knows are being biased right from the start?

The latter strategy was and unfortunately is still very popular with experimental psychologists, because the extreme group (high and low tail) strategy leads to higher effect sizes, which often lead, despite the loss of degrees of freedom, to significant results, and the beloved significance stars in their papers. Hans Eysenck loved that strategy and used it often. Later others using the full distributions found that his results could not be replicated or resulted in much lower effect sizes, no wonder given these psychometric facts.

This phenomenon can also used to explain why qualitatively oriented researchers often do not believe our quantitative results after using better measurement. Experienced practitioners in educational, clinical and other settings often contrast a couple of extreme cases against each other and get thus the impression of a large effect (in terms of Cohen's *d* or *r*). One can demonstrate with contrasting a couple of cases above + 2-3 *SD*'s against a couple of cases below - 2-3 *SD*'s that the biasing factor might be larger than 4.This means where the true effect might be a small one only (r = .10) the practitioner's impression is related to a medium to large sized one r >.40, which cannot be generalized.(I' ve published about that phenomenon recently, but in German only)

So your colleague and others should always use all the information given, whenever it is available.

Steve Simon was less critical of dichotomization than were others:

There's a trade-off here. By removing the middle third, you increase the separation of the two groups, which is good, while at the same time reducing the sample size, which is bad. Usually the trade-off is good.

It's not too hard to show that the loss of information is related to the correlation between the original variable and a new variable which equals -1, 0, or +1 depending on which third of the data you are in. For most data sets, this is slightly better than the correlation between the original variable and a new variable which equals -1 for the first half and +1 for the second half.

Tukey came up with a simple regression fit that involved removing the middle third of the data. So you have some precedent for this approach.

I would not be as critical as some of the others on the list. Sometimes a categorical variable is easier to interpret. A lot of dietary research, for example, looks at the highest quintile of fat consumption and compares it to the lowest quintile. I can visualize those two groups pretty well. Furthermore, categorization mitigates some of the problems caused by measurement error.

If I were doing it myself, I would almost never dichotomize. But I wouldn't be too upset if someone else did it, especially if the data set was already quite large.

Steve Simon, ssimon@cmh.edu, Standard Disclaimer.

Related Article: Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178-192.

The Extreme Groups Approach (EGA) involves the investigation of the relationship between continuous variable X and criterion variable Y. Although X is continuous, the researcher elects to obtain data on only those cases which have high or low values of X. Assuming Y is also continuous, Y is then correlated with those extreme values of X. When the relationship between X and Y is linear, this method can actually be more powerful than correlating the full range of X with Y, holding sample size constant. Note that if you obtain data on the full range of X and then throw out the middle scores, you are not holding sample size constant and are likely to lose power by the loss of cases. Past research has indicated that power is likely to be greatest if you select those cases in the upper and lower quartiles of X. Preacher et al. remind us that power is not everything. Unless we are uninterested in the relationship between Y and intermediate levels of X, it seems more sensible to relate the full range of X to Y.

The authors show that EGA will result in upwardly biased estimates of the size of the effect (association between full range of X and Y), which can be, but are not likely to be, adjusted to remove (some of) such bias – after all, what researcher wants to do more arithmetic just to make her findings appear less impressive – only the honest researcher who isn't all that interested in getting published.

Apparently some researchers try to justify EGA by arguing that it enhances the reliability of their measurement of X by eliminating the less reliable measurements in the middle of the distribution. It is not, however, generally true that measurements in the middle will be less reliable (quite the contrary is expected), and any apparent increase in the reliability of the measurement of X is an artifact of EGA.

After scolding colleagues for dichotomizing a continuous variable, I have sometimes been told that the dichotomization is justified because it estimates an underlying dichotomous characteristic

(Type A versus Type B personality, for example). Preacher et al. and I have serious doubts about such reasoning.

EGA is sometimes justified because the researcher is interesting in studying interaction or moderator effects and is not aware that these can be studied without categorizing continuous variables. I must confess that I have done this myself, not because of my ignorance, but because I have learned that my audience typically cannot understand interactions between continuous variables. For example, I found that ethical idealism (continuous) moderated the relationship between misanthropy and support for animal rights, but I then dichotomized idealism for the analysis that I presented.

So, when can EGA be justified? In exploratory research, especially where data are expensive, EGA may be justified as a method for determining whether or not there exists any relationship between X and Y (and hopefully not a quadratic one). Dichotomization of X should not generally be part of EGA, but may be a last resort transformation to meet the normality assumption of the correlation analysis used to relate X to Y -- but other transformations and analyses may well be superior. When one's primary interest is in demonstrating an interaction, one may deliberately oversample extreme scores, which should increase power, but will also overestimate the size of the interaction effect in the population of interest.

References

Irwin, J.R., & McClelland, G.H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, *40*, 366-371.

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.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178-192.

Return to Wuensch's Stat Help Page

Karl L. Wuensch, Dept. of Psychology, East Carolina Univ., Greenville, NC USA, March, 2021.