Dichotomization: How bad is it?

Thomas R. Knapp
©
2013

Introduction

I love percentages (Knapp, 2010). I love them so much I'm tempted to turn every statistical problem into an analysis of percentages. But is that wise, especially if you have to dichotomize continuous variables in order to do it? Probably not (see, for example, Cohen, 1983; Streiner, 2002; MacCallum, et al., 2002; Owen & Froman, 2005). But the more important question is: How much do you lose by so doing? What follows is an attempt to compare "undichotomized" variables with dichotomized variables, with special attention given to situations where the relationship between two variables is of primary concern.

An example

In conjunction with the high school Advanced Placement Program in Statistics, Bullard (n.d.) gathered data on 866 major league baseball players, including their heights (x) and their weights (y). Here is the Minitab scatter plot for the relationship between those two variables for this "population" of 866 persons:

```
         -
     300+
         -                                          *
 y       -                                    *  *
         -                             *            *         *
         -                      *         2   2     *         *   *
     250+                *     2   7  5  +   5  *  2
         -                  *  4   7  +  +   6  4      2
         -            2   3  2  7   +  +  +   +  7      2  *  *
         -         2      7  +  +   +  +  +   +  4      *
         -         *  3   7  +  +   +  +  +   6  *
     200+      3      +   +  +  +   +  +  4   3
         -     2  5  +   +  +  +   +  9  2   3  *
         -  2   *  9  +   +  +  7   5  *  *   *
         -  *   *  5  5   4  4  4   2
         -            *   3  2  2
     150+      *     *                *
         -
         -
           --------+---------+---------+---------+---------+--------x
                69.0      72.0      75.0      78.0      81.0
```

That's a nice elliptical plot, for which the the "ordinary" Pearson r correlation is .609.

I asked Minitab to dichotomize the two variables at their medians and calculate the Pearson correlation between the dichotomized height and the dichotomized weight (this is sometimes called a phi coefficient). The result was a correlation of .455.

As you can see, the correlation between the original heights and weights is greater than the correlation between the dichotomized heights and weights. Intuitively that is to be expected, because you're throwing away potentially useful information by dichotomizing.

I then carried out what I like to call "a poor man's Monte Carlo simulation". I asked Minitab to draw 30 random samples each of size 30 from the population of the 866 original heights and weights, and 30 other random samples from the population of the 866 dichotomized heights and weights (sampling was "without replacement" within sample and "with replacement" between samples). For each of those 60 samples I also asked Minitab to calculate the correlation between height and weight, and to summarize the data. Here's what I got:

```
                  original    dichotomy
Size of sample       30          30
Number of samples    30          30
Mean r            .6275       .4374
Median r          .6580       .4440
SD (std. error)   .0960       .1973
Minimum r         .4240       .0000
Maximum r         .7570       .7360
```

Not only are the correlations lower for the dichotomized variables, but the standard error is higher, meaning that the dichotomization would produce wider confidence intervals and lower power for significance tests regarding the correlation between height and weight. Therefore, case closed? Don't ever dichotomize? Well, not quite.

First of all, the above demonstration is not a proof. Maybe the dichotomized variables don't "work" as well as the original variables for this dataset only. (Irwin & McClelland, 2003, do provide a proof of the decrease in predictability for a special case.) Secondly, although it's nice to have high correlations between variables in order to be able to predict one from the other, the primary objective of research is to seek out truth, and not necessarily to maximize predictability. (The dichotomized version of a variable might even be the more valid way to measure the construct of interest!). Finally, there are some known situations where the reverse is true, and there are some frequency distributions of continuous variables that are so strange they cry out for dichotomization. Read on.

Some counter-examples

In their critique of dichotomization, Owen and Froman cite a study by Fraley and Spieker (2003) in which the correlation between dichotomized variables was higher than the correlation between the original continuous variables. Maxwell and Delaney (1993) showed that the interaction effect of two dichotomized variables in an ANOVA could be greater than the effect of their continuous counterparts in a multiple regression analysis. And while I was playing around with the baseball data I had Minitab take a few random samples of size 30 each and calculate the correlation between height and weight for the undichotomized and the dichotomized variables for the same sample. For one of them I got a correlation of .495 for continuous heights and weights and a correlation of .535 for their associated dichotomies. Here is the scatterplot:

```
        -
        -                                                  *
     240+                              *
        -                                   *
wt     -                                                  *    *
        -
        -                                        *
     220+                         2              *    *
        -                         *         2                   *
        -                                   *
        -                         *                   *
        -                    *                        *
     200+     *              *
        -                    *    2
        -                    *         *
        -                              2
        -
     180+                                   *
        -                              *
          ----+---------+---------+---------+---------+---------+--ht
           68.0      70.0      72.0      74.0      76.0      78.0
```

And here is the contingency table:

		wt	
		0	1
ht	1	3	11
	0	12	4

A few outliers "destroyed" the correlation for the original variables (.609 in the population), while all of those 1,1 and 0, 0 combinations "enhanced" the correlation for the dichotomized variables (.455 in the population). It can happen. That's one of the vagaries of sampling.

Strange frequency distributions

In their article, MacCallum, et al. (2002) acknowledged that dichotomization might be justified for the frequency distribution of number of cigarettes smoked per day, with spikes at 0,10, 20, and 40, and lots of holes in-between multiples of 10. I displayed an actual such distribution in my percentages book (Knapp, 2010). I also displayed a similarly strange distribution for the number of cards successfully played in the solitaire game of Klondike. Both of those distributions were strong candidates for dichotomization.

Age

If there ever is a variable that is subject to dichotomization more than age is, I don't know what that variable might be. When people are interested in a research question such as "What is the relationship between age and political affiliation, more often than not they choose to either break up the age range into groupings such as "Generation X", "Baby Boomers", and the like, or dichotomize it completely into "young" and "old" by cutting somewhere.

Chen, Cohen, and Chen (2006) have shown that not only do you lose information by dichotomizing age when it is an independent variable (when else?!), but one of the statistics of greatest use in a field such as epidemiology, the odds ratio, turns out to be biased: The further the cutpoint is from the median of the continuous distribution, the greater the bias, with the net effect that the odds ratio is artificially larger. As indicated above, although it's nice to get high correlations between variables, including high odds ratios, the quest should be for truth, not necessarily for predictability.

So does that mean that age should never be dichotomized? Again, not quite. Chen, Cohen, and Chen admit that there are some situations where age dichotomization is defensible, e.g., if subjects are intentionally recruited in age groups that are hypothesized to differ on some dependent variable.

Those terrible Likert-type scales

I don't know about you, but I hate the 4, 5, 6, or 7-point ordinal scales that permeate research in the social sciences. If they can't be avoided in the first place (by using interval scales rather than ordinal scales or by using approaches that are tailor-made for ordinal variables...see, for example, Agresti, 2010), then they certainly should be dichotomized. Do we really need the extra sensitivity provided by, say, the typical "Strongly agree", "Agree", "Undecided", "Disagree", "Strongly Disagree" scales for measuring opinions? Isn't a simple "Agree" vs. "Disagree" sufficient?

For a Likert-type scale with an even number of scale points I suggest dichotomizing into "low" (0) and "high" (1) groups by slicing in the center of the scale. If it has an odd number of scale points I suggest dichotomizing by slicing through the middle category, randomly allocating half of the observations in that category to "low" and the other half to "high". There; isn't that easy?

A final comment

I can't resist ending this paper with a quotation from a blogger who was seeking statistical help (name of blogger and site to remain anonymous) and asked the following question: "Can anyone tell me how to dichotomize a variable into thirds?" Oy.

References

Agresti, A. (2010). The analysis of ordinal categorical data (2nd. ed.). New York: Wiley.

Bullard, F. (n.d.) Excel file of data for 866 Major League Baseball players. http://apcentral.collegeboard.com/apc/public/repository/bullard_MLB_list.xls

Chen, H., Cohen, P., & Chen, S. (2006). Biased odds ratios from dichotomization of age. Statistics in Medicine, 26, 3487-3497.

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

Fraley, R.C., & Spieker, S.J. (2003). Are infant attachment patterns continuously or categorically distributed? A taxometric analysis of strange situation behavior. Developmental Psychology, 39, 387–404.

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

Knapp, T.R. (2010). Percentages: The most useful statistics ever invented. Accessible free of charge at both of my websites (www.tomswebpage.net & www.statlit.org/knapp.htm).

MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7 (1), 19-40.

Maxwell, S.E., & Delaney, H.D. (1993). Bivariate median splits and spurious statistical significance. Psychological Bulletin, 113 (1), 181-190.

Owen, S.V., & Froman, R.D. (2005). Why carve up your continuous data? Research in Nursing & Health, 28, 496-503.

Streiner, D.L. (2002). Breaking up is hard to do: The heartbreak of dichotomizing continuous data. Canadian Journal of Psychiatry, 47, 262-266.