International Journal of Behavioral Medicine (2019) 26:329–330
https://doi.org/10.1007/s12529-019-09790-7

COMMENTARY

Statistical Guideline #1. *Avoid Creating Categorical Variables from Continuous Variables*

Suzanne C. Segerstrom [1]

Published online: 28 May 2019

The Statistics Guru

Frequently, editors and reviewers see continuous data artificially categorized for statistical analysis—for example, above or below a clinical cutoff or the sample median. The first statistical guideline for *IJBM* is to keep continuous variables as continuous variables whenever possible.

A continuous variable contains the most amount of information about a data point. Categories, particularly dichotomization, lose much of this information. Consider the example of a normally distributed variable with a mean of 50 and a range 0–100. If I dichotomize the variable at 50, creating 2 categories ("high" and "low"), then I essentially claim that two data points with values 1 and 49 are more alike than two data points with values 49 and 51. That claim is almost certainly false. Furthermore, dichotomization leads to underestimates of effect size and loss of measurement reliability [1]. Why does this practice persist?

One candidate is the influence of the medical model. Medicine dichotomizes people: you are either hypertensive or not, obese or not, diabetic or not. This practice was sensible when medicine dealt mostly with infectious disease, because infected or not is a true dichotomy. (Also, famously, you cannot be a little bit pregnant.) A dichotomous diagnosis does provide clinical guidance on whether to treat or not. However, in research and in statistical models, there is rarely a good reason to treat blood pressure or BMI or blood glucose as dichotomous. The same is true for psychological variables. Even psychiatric disorders may be best conceptualized as continuous [2]. Finally, median, quartile, or other sample-specific splits create the additional problem of idiosyncrasy. Two samples are unlikely to have exactly the same median even when drawn from the same population. Sample-specific splits therefore work against the goal of cumulative science.

Perhaps one does want information about a specific level of, for example, scores on the Beck Depression Inventory, which has cutoff scores for minimal, mild, moderate, and severe depression. To test the BDI as an explanatory variable for C-reactive protein (CRP), for example, one could create four categories of BDI scores and use ANOVA to test for differences. However, to estimate the levels of CRP at particular levels of the BDI, one need only re-center the continuous BDI variable around the desired levels and examine the intercept term. For example, redefining x (BDI score) as x—20 (the cutoff score for moderate depression) will give a result in which the intercept is the predicted CRP value when BDI equals 20. A particular advantage of this approach is that different studies can report comparable results, thereby working toward cumulative science. Some people may not be comfortable with using continuous variables to test interactions; however, this is also a fairly simple matter of appropriately centering the explanatory variables and creating product terms between them. Aiken and West's book, *Multiple Regression: Testing and Interpreting Interactions* provides excellent instructions [3].

Note that the guideline applies to the problematic habit of *creating* fewer levels or categories than were measured. Sometimes categories are unavoidable, even when the underlying construct is continuous. Education or income data may

From the Editors: This is the first column from the Statistics Guru. The Statistics Guru will appear in every issue. In these columns, we briefly discuss appropriate ways to analyze and present data in the journal. As such, the Statistics Guru can be seen both as an editorial *amuse bouche* and a set of guidelines for reporting data in the *International Journal of Behavioral Medicine.* If you have ideas for a column, please email the Statistical Editor, Suzanne Segerstrom at segerstrom@uky.edu.

✉ Suzanne C. Segerstrom
segerstrom@uky.edu

[1] Department of Psychology, University of Kentucky, 125 Kastle Hall, Lexington, KY 40506-0044, USA

have been collected as ordinal data (e.g., indicating education as "some college" or income between $10,000 and $25,000) rather than continuous years or dollars. Single-item Likert scales with few response options (i.e., < 6; [4]) might be better treated as ordinal (1 < 2 < 3) than as continuous (this possibility should be considered on a case-by-case basis for the specific item). When data are not continuous per se, appropriate analyses are available for ordinal or categorical models (e.g., [5]).

MacCallum et al. [1] conclude that "In common usage, dichotomization is typically carried out without apparent justification and without serious awareness or regard for its consequences. Such ad hoc procedures are simply inappropriate and incur substantial costs" (p. 38). This practice is to be avoided.

References

1. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7:19–40.
2. Widiger TA, Samuel DB. Diagnostic categories or dimensions? A question for the diagnostic and statistical manual of mental disorders – fifth edition. J Abnorm Psychol. 2005;114:494–504.
3. Aiken LS, West SG. Multiple regression: testing and interpreting interactions. Thousand Oaks: Sage; 1991.
4. Simms LJ, Zelazny K, Williams TF, Bernstein L. Does the number of response options matter? Psychometric perspectives using personality questionnaire data. Psychol Assess. 2019;31:557–66.
5. Bürkner PC, Vuorre M. Ordinal regression models in psychology: a tutorial. Adv Methods Pract Psychol Sci. 2019;2:77–101.