Optimal Data Analysis
Vol. 2, Release 2 (October 14, 2013), 69-70

2155-0182/13/$3.00

Analysis Involving Categorical Attributes Having Many Response Categories

Paul R. Yarnold, Ph.D., and Fred B. Bryant, Ph.D.
Optimal Data Analysis, LLC Loyola University of Chicago

Attributes measured on a categorical response scale are common in the literature. Categorical scales for attributes such as, for example, political affiliation, ethnic origin, marital status, state of residence, or diagnosis may consist of many qualitative response categories. Such disorganized variables rarely appear in multivariable models: some effects are missed in analysis due to inadequate statistical power for the many categories, and some findings are dismissed due to inability of the investigator to recognize the dimension(s) underlying segmented categories. This research note recommends that such multi-categorical attributes are replaced by a new set of attributes created via content analysis. In this approach observations are scored on new dimensions all theoretically motivated to predict the class variable. The methodology is illustrated using a hypothetical example in the field of investment realty.

Research Problem and Solution (Dr. Yarnold)

Imagine a study investigated American homeowner satisfaction (satisfied or dissatisfied) as the class variable. It is common knowledge that an important multi-categorical attribute in this application is "*location, location, location*". Examples of location attributes include region or state in a national study, county or municipality in a state study, or neighborhood or street in a municipal study. For exposition, let's consider a national study. With so many categories (50), it is unlikely that a statistically or conceptually compelling model would be identified if state was used as a multi-categorical attribute.

On the other hand, data at the state level exist on many attributes that may potentially influence homeowner satisfaction. For example, measures of annual average per-capita savings; crude mortality rate; low winter temperature; high summer temperature; beachfront acreage; public school academic performance rank; non-federal taxes; number of children per household; crime rate; etcetera. Many such measures are reported using real-number, interval, integer, or ordinal scales.[1] In this manner the likely non-interpretable 50-category variable is instead transformed into a panel of easily interpreted, theoretically justifiable ordered attributes.

2155-0182/13/$3.00

Comment
(Dr. Bryant)

The insight here serves to improve the quality of analyses of categorical variables. Your recommendation to recode multinomial categorical variables into a smaller number of substantively meaningful, theoretically relevant dimensions is exactly the way we were trained on my postdoc to work with nominal variables consisting of a large number of categories.

At the University of Michigan's Survey Research Center, we were taught to begin by coding open-ended qualitative variables into the smallest-grained categories we could (thereby maximizing the number of categories), but not to analyze these basic "starting point" categorical variables. Instead, for analysis, we were trained to recode these fine-grained variables into multiple sets of smaller, conceptually-grounded categories, with each set reflecting a potentially important conceptual dimension of interest. For example, instead of analyzing the earliest memory based on all of the categories that were originally coded, we might instead content analyze these initial categories to form new coding schemes, each of which reflected a different conceptual concern that consisted of a smaller number of categories. The full list of all of the specific earliest memories might, for example, be recoded in terms of focus of attention (self, others, world), valence (positive, neutral, negative), social situation (alone versus with others), setting (indoors versus outdoors), etc. Some initial coding schemes had a hundred or more categories, from which dozens of interesting and conceptually meaningful recoded variables, each with a much smaller number of categories, were formed.

It's a great idea that well-trained survey researchers routinely use to maximize the utility and conceptual focus of open-ended responses. Your insight is that this same approach can be applied to any type of categorical measure, even if it is a closed-ended one.

References

[1]Yarnold PR, Soltysik RC (2004). *Optimal Data Analysis: Guidebook with software for Windows*, APA Books.

Author Notes

E-mail: Journal@OptimalDataAnalysis.com.

Mail: Optimal Data Analysis, LLC
1220 Rosecrans St., #330
San Diego, CA 92106