Optimal Data Analysis
Vol. 7 (January 15, 2018), 3-9

2155-0182/10/$3.00

Minimize Usage of Binary Measurement Scales in Rigorous Classical Research

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Dichotomous measurement scales are likely insufficiently granular to empower breakthrough scientific discoveries for classical phenomena.

As used herein, "classical" refers to phenomena associated with observations which, unlike electrons, may manifest variable intrinsic characteristics.[1] For example, in serial research using rat strains specifically bred for experimentation[2], or using humans[3], ipsative transformation is used to eliminate "nuisance variability" attributable to baseline differences—and thereby attain non-confounded conclusions.[4] Even data acquired in *N*-of-1, single-case longitudinal research require ipsative transformation to avoid confounding.[5-7]

This note addresses the use of binary scales to measure classical class variables and/or attributes (together, "variables"). A class variable is a measure of the phenomenon which is being modeled, predicted or discriminated. An attribute is a measure of another phenomenon which is being used in an effort to accurately model, predict or discriminate the class variable (in legacy statistics[8,9] these are called dependent and independent variables, respectively).[10] The least granular experimental design uses binary indicators for both class variable and attribute.

Binary Measurement Scales

Binary scales putatively measure phenomena reflecting two distinct states, such as alive *vs*. dead. Considered from the perspective of making accurate predictions, two implicit assumptions regarding a variable, which apply irrespective of scale granularity, are that sample observations are (a) homogeneous within, and (b) heterogeneous between states. As violations of these assumptions increase, accurate prediction of the class categories becomes increasingly difficult: this is particularly salient for binary variables due to minimal granularity.[1]

Variables studied in the literature often masquerade as being qualitative, but reflect one or more underlying quantitative dimensions.[11] For example consider the variable, died *vs*. lived. Did all who died experience a uniform death, or was death heterogeneous (in *The Princess Bride*, Miracle Max drops the arm of dead Westley, saying: "I've seen worse"[12])? And, do all who lived have identical health—are they equivalently not-dead? If the research design uses an arbitrary (theoretically unmotivated) follow-up period, such as 30-day survival, are those who die on day 31 substantively or theoretically different than those who die on day 30, or on day 32 or day 27?

Color is an example of a variable often used as a prototypical example of a qualitative variable. Underlying the cornucopia of labels used to distinguish different colors is the quantitative dimension, wavelength.

2155-0182/10/$3.00

Binary scales are also constructed from inherently quantitative phenomena, such as age: <65 (class 0) *vs*. ≥65 (class 1) years, for example. An embarrassment of criteria are used to parse ordered variables into categories.[12-15]

Examples

Following are three examples typically analyzed by legacy statistical methods discriminating *two* pooled groups (class 0 *vs*. class 1), *vs*. ODA and CTA which can identify different types of both class categories—defined in terms of attribute profiles identified by the model.[16]

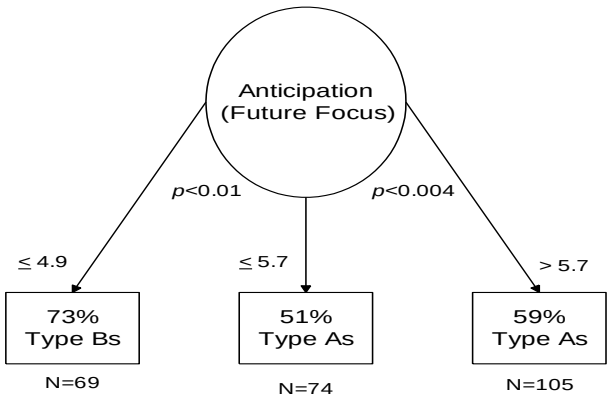
Example 1: Binary Class Variable, Ordered Attribute, Linear CTA Model

A study compared self-ratings by Type B (class 0) *vs*. Type A (class 1) undergraduates regarding their capacity to enjoy positive outcomes using anticipation (the ordered attribute).[17] It was hypothesized that, compared to Type Bs, Type As perceive themselves as being more able to savor through anticipation due to their greater goal orientation. Anticipation scores were compared between Type As *vs*. Type Bs by Student's *t*-test, but no effect emerged ($p<0.23$).

Figure 1 presents the CTA model that emerged—indicating a linear effect. Type Bs are more likely (3:1 odds) than Type As to score at lowest levels (≤4.9) on the Anticipation dimension (this cut-point represents the 28nd percentile on this dimension for the sample); Type As and Bs are comparably likely (1:1 odds) to score at intermediate levels (≤5.7); and Type As are more likely (3:2 odds) to score at highest levels (>5.7, the 58th percentile) on Anticipation. For this model ESS=24.1, a borderline moderate effect.[1]

The authors stated (p. 29): "The three-endpoint parse that emerged in the CTA model reveals the hypothesized A-B difference in capacity to anticipate exists at the lower and upper range of the subscale, but not in the middle range. While more Bs than As fall in the lower range, and more As than Bs fall in the upper range, As and Bs are equally distributed in the mid-range of the subscale. The CTA model not only confirms the *a priori* hypothesis, but it also pinpoints the specific levels of anticipation at which the predicted A-B differences emerge."

Figure 1: CTA Model Discriminating Type As *vs*. Type Bs



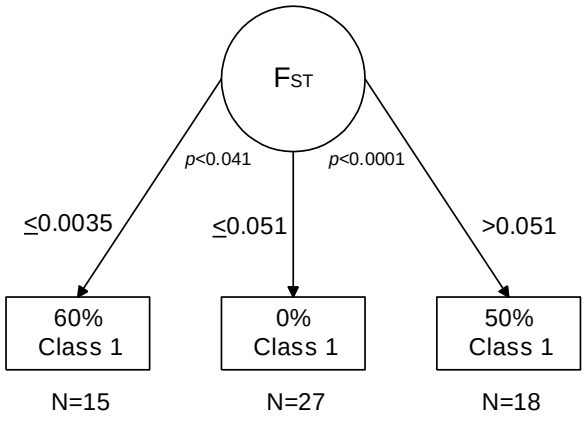
Example 2: Binary Class Variable, Ordered Attribute, Non-Linear ODA Model

A study investigated the exploratory hypothesis that protein polymorphisms (class=0) may have either lower or higher FST values (ordered attribute) than anonymous DNA polymorphisms (class=1). Evaluating this hypothesis using the Kruskal Wallace test, the null hypothesis that FST of DNA and protein polymorphisms have the same mean ranks is not rejected.[18]

Figure 2 presents the ODA model that emerged in simulation (sample size was tripled to increase statistical power)—indicating a non-linear effect. As seen, 40% of protein polymorphisms have lower FST, 50% have higher FST, and none have intermediate levels of FST, compared to DNA polymorphisms: ESS=64.3, a relatively strong effect.[1] As in the prior example, the ODA model supports the exploratory hypothesis and pinpoints specific FST levels at which the predicted protein *vs*. DNA polymorphism differences emerge.

2155-0182/10/$3.00

Figure 2: Simulated ODA Model Discriminating Protein *vs*. DNA Polymorphisms

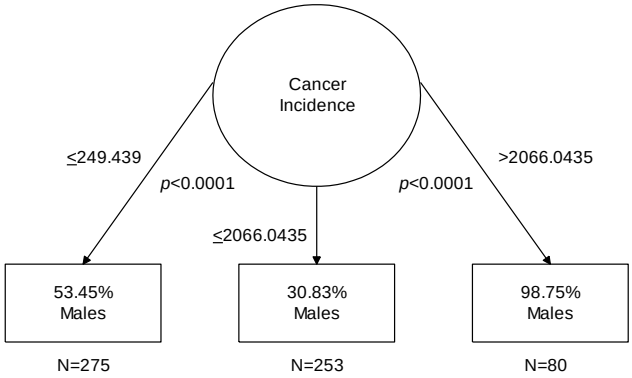


Example 3: Binary Class Variable, Ordered Attribute, Non-Linear CTA Model

Data are from the Surveillance, Epidemiology, and End Results (SEER) Program that collects and publishes cancer incidence and survival data so as to assemble and report estimates of cancer incidence, survival, mortality, other measures of the cancer burden, and patterns of care, in the USA.[19] In this example cancer incidence rate (number of new cancers occurring in a specified population in one year, expressed as number of new cancers for every 100,000 population at risk) is parsed by gender to identify discrete patient strata differing in cancer incidence. No legacy statistical analysis was reported.

CTA identified three types of male strata: reading from left to right across model endpoints, males had low, moderate, and high cancer incidence strata (Figure 3). And, CTA also identified low and moderate incidence female strata. The strata with lowest cancer incidence (≤0.249439%) was comparably represented by females and males, and it comprised (275/608) 45.2% of the total sample; the strata with highest incidence (>2.0660435%) was dominated (98.8%) by males, 13.2% of the total sample; and the strata having intermediate incidence (0.249440% to 2.0660435%) was predominantly females (69.2%), comprising 41.6% of the total sample. Here ESS=31.9, a moderate effect. Akin to prior examples, CTA pinpoints the exact cancer incidence rates for which specific gender-based differences exist.

Figure 3: CTA Model Parsing Cancer Incidence Rate Across All Sites by Gender



Discussion

Classical phenomena occasionally involve observations that, functioning nominally, can only exist in one of two possible states. For example, an electric circuit that either closes (coded as 1) or remains open (coded as 0) in response to an experimental stimulus—which itself is either "on" (coded as 1) or "off" (coded as 0)—for each trial in a serial *N*-of-1 reliability study.

However, such designs are uncommon in social or health science literatures. Prototypically, binary class variables such as gender are studied by discriminating pooled groups of males *vs*. females on one or more attributes. Parallel methods are used to evaluate interventions, discriminating pooled groups of treatment *vs*. control subjects on the attributes. In such research the groups are prototypically compared using rigid linear statistical methods which contrast the (assumed homogeneous) class=0 *vs*. (assumed homogeneous) class=1 samples.

One route to obtaining greater predictive accuracy for applications featuring binary class variables and/or attributes involves the use of modern statistical methods. In the present study, Student's *t*-test failed to discriminate A/B

2155-0182/10/$3.00

Types, and Kruskal Wallace analysis failed to discriminate polymorphisms. In contrast, using ODA and CTA to compare groups, more than one homogeneous type of the class categories was identified.

For instance, in the first example a linear CTA model emerged which identified three groups of Type A subjects: a group (from Figure 1, 0.27% x 69) of 19 Type As having a low future focus (<28th percentile); a group of 62 Type As with a high future focus (>58th percentile; and a group of 38 Type As with intermediate future focus. Likewise, the CTA model identified 50 Type Bs with a low future focus; 43 Type Bs with a high future focus; and 36 Type Bs with an intermediate future focus.

In the second example a non-linear ODA model identified two groups of anonymous DNA polymorphisms, half (N=0.6x15=9) of which had FST $\leq$ 0.0035; and half (N=0.5x18= 9) of which had FST > 0.051. The ODA model also identified three groups of protein polymorphisms: six (14.3%) with FST $\leq$ 0.0035; nine (21.4%) with FST > 0.051; and 27 (64.3%) with intermediate FST values.

Finally, in the third example a non-linear CTA model identified three male groups with a low, intermediate, or high overall cancer incidence rate, and two female groups with a low or intermediate cancer incidence rate.

In all three examples ODA and CTA models identified different types of both class categories (A/B types, polymorphisms, and genders). Identifying sample strata which are homogeneous within, heterogeneous between strata is necessary to avoid Simpson's paradox: pooling disparate groups (as done by rigid linear methods) can hide, exaggerate, or reverse the findings that would be obtained if analyses were separately conducted for individual groups.[4,20]

In all three examples only one ODA or CTA model was identified, in part attributable to modest sample sizes and associated modest statistical power, and in part due to effect sizes suggesting mediocre measurement or failure to support the theory.[1] For applications involving larger samples, better measures having stronger reliability and validity characteristics, and/or stronger theories, it is possible that more than one statistically viable model exists for the sample. Some viable models may involve only main effects, others include interactions. Accidental omission of a crucial term from a model is a misspecification error, however evaluating all possible linear models may be infeasible for applications with numerous attributes.[21] Moreover, linear models employing many binary attributes are difficult to visualize and unwieldy to apply in practice; susceptible to computational instability resulting from multicollinearity; and are too complex to be theoretically feasible.[22]

In contrast, CTA identifies all statistically viable models that exist for the sample—providing a comprehensive sensitivity analysis that prevents model misspecification.[23,24] CTA also enables testing and comparing competing theoretical specifications, including model features such as attribute entry order, threshold values, and jackknife reliability constraints.[25]

A second avenue to increasing predictive accuracy for applications using binary variables is follow-up research focusing on advancing conceptual specificity and measurement granularity in theoretically promising directions. For example, A/B Type may be may deconstructed into its constituent dimensions: omnipresent time urgency and hard-driving dominance rank among the most prevalent symptoms of Type A behavior using gold-standard clinical interview assessment, and among the most generalizable factors identified on questionnaire measures.[26-38] Similarly-focused research may be undertaken to advance understanding regarding promising (statistically significant, with moderate or better effect strength), theoretically motivated dimensions of gender[11] and savoring.[39]

Sometimes one's only option is to use (archival) binary measures, for example in early stages of a new line of research—when little is known and exploratory directions have to be

2155-0182/10/$3.00

identified. In such instances recall (and use as a developmental guide) the implicit assumptions that, with respect to the attribute(s), the groups to be discriminated should be homogeneous within and heterogeneous between group.

In light of this perspective, the category "other" should never be used as an option on a measurement scale, because it implies "everything else in the Universe", and combining disparate groups (e.g., everything else in the universe) gives rise to paradoxical confounding, as was discussed earlier.

Conclusion

Linear statistical models are susceptible to paradoxical confounding attributable to heterogeneity in either group being discriminated. This limitation may be overcome by using modern statistical methods which can identify solutions involving subtypes of the groups. Nevertheless, binary variables are likely insufficiently theoretically specific and empirically granular to empower breakthrough scientific discoveries for classical phenomena in social or health sciences.

References

[1]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[2]Yarnold PR, Soltysik RC (2013). Ipsative transformations are essential in the analysis of serial data. *Optimal Data Analysis, 2*, 94-97.

[3]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

[4]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56*, 430-442.

[5]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis, 1*, 67-100.

[6]Yarnold PR (2013). Comparing responses to dichotomous attributes in single-case designs. *Optimal Data Analysis, 2*, 154-156.

[7]Yarnold PR (2013). Ascertaining an individual patient's symptom dominance hierarchy: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis, 2*, 159-171.

[8]Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books.

[9]Grimm LG, Yarnold PR (2000). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books.

[10]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books.

[11]Yarnold PR, Bryant FB (2013). Analysis involving categorical attributes having many categories. *Optimal Data Analysis, 2*, 69-70.

[12]http://www.imdb.com/character/ch0003789/quotes

[13]Yarnold PR (2010). Aggregated *vs*. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis, 1*, 46-49.

[14]Yarnold PR (2010). UniODA *vs*. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis, 1*, 62-65.

[15]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis, 3*, 19.

[16]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis, 6*, 26-42.

2155-0182/10/$3.00

[17]Bryant FB, Yarnold PR (2014). Finding joy in the past, present, and future: The relationship between Type A behavior and savoring beliefs among college undergraduates. *Optimal Data Analysis*, *3*, 36-41.

[18]Yarnold PR (2015). An example of nonlinear UniODA. *Optimal Data Analysis*, *4*, 124-128.

[19]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

[20]Bryant FB, Siegel EKB (2010). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v*. Illinois Bell. *Optimal Data Analysis*, *1*, 176-198.

[21]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190.

[22]Yarnold PR, Linden A. (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[23]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[24]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12744

[25]Linden A, Yarnold PR (2017). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12848

[26]Yarnold PR, Grimm LG (1982). Time urgency among coronary-prone individuals. *Journal of Abnormal Psychology*, *91*, 175-177.

[27]Grimm LG, Yarnold PR (1984). Performance standards and the Type A behavior pattern. *Cognitive Therapy and Research*, *8*, 59-66.

[28]Yarnold PR, Mueser KT (1984). Time urgency of Type A individuals: Two replications. *Perceptual and Motor Skills*, *59*, 334.

[29]Yarnold PR, Mueser KT, Grimm LG (1985). Interpersonal dominance of Type As in group discussions. *Journal of Abnormal Psychology*, *94*, 233-236.

[30]Yarnold PR, Grimm LG (1986). Interpersonal dominance and coronary-prone behavior. *Journal of Research in Personality*, *20*, 420-433.

[31]Yarnold PR, Grimm LG, Lyons JS (1987). The Wiggins Interpersonal Behavior Circle and the Type A behavior pattern. *Journal of Research in Personality*, *21*, 185-196.

[32]Mueser KT, Yarnold PR, Bryant FB (1987). Type A behavior and time urgency: Perception of time adjectives. *British Journal of Medical Psychology*, *60*, 267-269.

[33]Yarnold PR, Grimm LG (1988). Interpersonal dominance of Type As and Type Bs during involved group discussions. *Journal of Applied Social Psychology*, *18*, 787-795.

[34]Yarnold PR, Grimm LG (1988). Interpersonal dominance and coronary-prone behavior: Reply to Ray. *Journal of Research in Personality*, *22*, 254-258.

[35]Yarnold PR, Bryant FB (1988). A note on measurement issues in Type A research: Let's not throw out the baby with the bath water. *Journal of Personality Assessment*, *52*, 410-419.

[36]Yarnold PR, Mueser KT, Lyons JS (1988). Type A behavior, accountability, and work rate in small groups. *Journal of Research in Personality*, *22*, 353-360.

2155-0182/10/$3.00

[37]Yarnold PR, Bryant FB, Litsas F (1989). Type A behavior and psychological androgyny among Greek college students. *European Journal of Personality*, *3*, 249-268.

[38]Yarnold PR, Bryant FB (1994). A measurement model of the Type A Self-Rating Inventory. *Journal of Personality Assessment*, *62*, 102-115.

[39]Bryant FB, Veroff J (2007). *Savoring: A new model of positive experience*. Mahweh, NJ: Erlbaum

Author Notes

This statistical article is exempt from IRB review. No conflict of interest was reported.