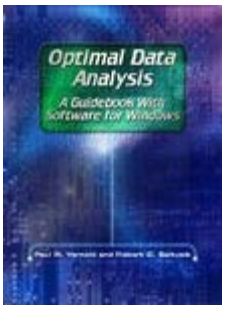
PsycCRITIQUES
February 2, 2005, Vol. 50, No. 5, Article 5



How to Make the Best of Your Data

A Review of

Optimal Data Analysis: A Guidebook With Software for Windows
by Paul R. Yarnold and Robert C. Soltysik
Washington, DC: American Psychological Association, 2004. 287 pp. ISBN 1-55798-981-8.
$99.95, paperback

doi: 10.1037/050066

Reviewed by
Fred B. Bryant

“When you make a prediction, would you rather be correct or incorrect? If your answer is'correct,' then ODA [optimal data analysis] is the appropriate analytic methodology…” (p. xi). With this audacious, provocative opening statement, Paul Yarnold and Robert Soltysik set the stage for a revolutionary paradigm shift in statistical analysis. It is obvious that no one could rationally answer “no” to such a question—who would rather be incorrect than correct when making predictions? Why else would one make predictions in the first place, except in hopes of being as accurate as possible? Yet, as Yarnold and Soltysik emphasize, to make predictions using any statistical approach other than ODA (pronounced “oh-dah,” with the accent on the “oh”) is to employ a paradigm that does not intentionally maximize the accuracy of predictions.

If you think that prediction is a highly specialized type of statistical task in which you should only be occasionally interested, then think again. As this book emphasizes, making predictions is what most statistical analyses are all about, and most statistical procedures can easily be recast in terms of classification problems for which one either tests hypotheses about or seeks to find the most accurate ways of predicting critical outcomes. Although this idea is not new, it is radical and nontraditional, and it holds the promise of improving the quality of quantitative research in every discipline if researchers are open to it.

Of course, as the authors note, one might indirectly stumble on the most accurate set of predictive rules in trying to maximize (a) a variance ratio (as when analyzing data with *t* test, analysis and multiple analysis of variance, correlation, or multiple regression) or (b) a likelihood function (as when analyzing data with chi-square, logistic regression, log-linear analysis, or structural equation modeling). But how would one know that this particular statistical solution offered maximum predictive accuracy? For no statistical technique but ODA is designed to make predictions that explicitly and directly optimize classification accuracy.

The Elegant Simplicity of ODA

ODA's statistical approach is shockingly simple. Imagine you want to predict whether or not a manuscript you submit for editorial review will be accepted for publication. On the basis of the notion that less is more, you believe that accepted manuscripts tend to be shorter than manuscripts that are rejected for publication. To test this a priori directional hypothesis, you look back at all of the manuscripts you have ever submitted for editorial review and record their number of pages. You also note whether or not each manuscript was accepted for publication. In ODA terminology, the number of pages of each manuscript is the independent variable or *attribute* (i.e., a random variable that can attain two or more levels), and whether the manuscript was accepted (coded as 1) or rejected (coded as a 0) is the dependent variable or *class variable* (i.e., a random variable that can attain two or more levels).

To conduct an optimal data analysis, the ODA software would begin by arranging all of the manuscripts (i.e., observations) along a continuum formed by page length, with each manuscript represented by a 0 or 1 depending on its publication status. ODA would then examine all possible *cutpoints* along the continuum (i.e., midpoints between two successive observations that have different values on the class variable) and would separately evaluate the classification performance achieved across all observations, using each cutpoint that conforms to the directional hypothesis (i.e., for which the lower score on the page-length continuum is associated with acceptance and the higher score on the page-length continuum is associated with rejection). The final ODA model would consist of the cutpoint that matches the directional hypothesis and produces the greatest overall percentage of accurate predictions across both categories of the class variable. For example, the optimal model might be, "If page length ≤ 25.5, then predict the manuscript is accepted for publication; otherwise, predict the manuscript is rejected." This particular model would be considered optimal because no other cutpoint consistent with the directional hypothesis could achieve a greater overall percentage of classification accuracy with these data.

The Unique Versatility of ODA

Although this particular example illustrates the elegant simplicity of the ODA paradigm, it fails to convey the extraordinary versatility of ODA. Attributes need not be continuous, and class variables need not be binary. Indeed, ODA can analyze attributes and class variables that are measured using ratio, interval, ordinal, or nominal scales and that are either dichotomous or polychotomous. This capability enables one to use ODA to analyze most data configurations for which traditional parametric or nonparametric statistical tests are commonly employed—including t test, analysis of variance, chi-square, log-linear analysis, logistic regression, repeated-measures analyses, time-series analysis, and Markov modeling. ODA can also be used to conduct a variety of reliability analyses, including interrater, intrarater, parallel forms, split-half, and temporal reliability, as well as analyses of convergent and discriminant validity and single-case "*N*-of-1" designs. Besides testing directional hypotheses, ODA can also find optimal classification models testing nondirectional hypotheses in which the user does not specify beforehand whether higher or lower values of the attribute are associated with a particular level of the class variable.

Even more impressive, this book is filled with examples (many drawn from classic statistics texts) illustrating the use of ODA to analyze data configuration for which no other statistical method exists. ODA is truly a cornucopia of statistical delights. Perhaps most striking of all is the fact that every one of these analyses can be accomplished with the same basic set of seven or so simple syntax commands, using the same intuitively appealing index of goodness-of-fit, namely percentage accuracy in classification (PAC). This integrative model-building heuristic makes the ODA paradigm an all-encompassing statistical framework unlike any other. Its unparalleled versatility makes ODA a must-have tool for quantitative researchers.

ODA as an Empirical Research Tool

Researchers will be interested to know that ODA also estimates Type I error and effect size for use in inferential hypothesis testing. The software computes *p* values for optimal classification models using a distribution-free permutation probability method known as Fisher's randomization procedure (Bradley, 1968). When the number of permutations is too large to obtain an exact probability, the program relies on Monte Carlo methodology to estimate it. The authors report the results of roughly two billion Monte Carlo experiments verifying the accuracy, stability, and convergence properties of this approach to estimating the exact probability of known ODA distributions. To estimate the size of observed effects, ODA reports a standardized index of effect strength (ES) that expresses the strength of classification accuracy on a scale ranging from 0 (the classification performance expected by chance) to 100 (perfect classification accuracy). The authors provide the following rules of thumb for judging effect strength: ES values less than 25% are *weak,* between 25% and 50% are *moderate,* between 50% and 75% are *relatively strong,* between 75% and 90% are strong, and above 90% are *very strong.*

Additional strong points include (a) the option of weighting outcomes by either prior odds or the subjective importance of a particular attribute (e.g., financial return), (b) an integrated jack-knife resampling procedure for assessing expected

model stability, (c) the capability of directly assessing model generalizability across multiple samples, and (d) an easy-to-use appendix for implementing sequentially rejective Bonferroni adjustments that control experiment-wise Type I error in a way that maximizes statistical power. What impresses me most about ODA is its power, versatility, and conceptual clarity. ODA is the most extraordinary statistical tool that I've encountered in 30 years of research in the social sciences.

Strengths of the Guidebook and Software

Yarnold and Soltysik's instructional guidebook to ODA is written in a simple, straightforward, reader-friendly style that effectively introduces one to the logic and mechanics of the general ODA paradigm using numerous well-annotated, applied examples with a minimum of formulas. In a refreshingly engaging approach to statistics, the authors rely on crystal-clear worked examples and narrative explanations of optimal classification modeling in action to engage the reader and encourage the use of ODA in research. The guidebook includes a CD-ROM containing the ODA software along with 62 example data sets and 50 syntax files linked to the various worked examples in the book. The book's greatest strengths are its clarity, comprehensiveness, and creative insights into the use of statistics in research.

Not only is the book easy to read, but the ODA software is exceptionally easy to use. The syntax-driven analyses can be executed using a Windows-based editor included with the software or from a DOS prompt. Only a handful of simple commands are required to specify output files, data format, missing data, attributes, class variables, the continuous versus categorical nature of measured variables, and directional or nondirectional hypotheses and to run jack-knife validity analyses and assess Type I error rate. Compared with other syntax-driven programs, such as LISREL (which uses Greek letters to represent parameter matrices) and EQS (which requires users to specify regression equations for each measured and latent variable), ODA is a breeze to learn. At $99.95, the software is a steal, especially compared with other competitive packages such as SPSS, EQS, and LISREL, which sell for 4–6 times that. By pricing the book and software so low, the authors and publisher aim to make ODA accessible to all.

Obstacles to the Use of ODA

Readers may well begin, as I did, with a considerable amount of skepticism toward ODA—it simply sounds too good to be true. Since the dawn of time, human beings have tried to make predictions about the world around them. From hunting to farming, warfare to hostage negotiation, gambling to prospecting, psychological assessment to weather forecasting, the goal has always been the same—to maximize the percentage of predictions that are correct. If ODA is superior in its predictive accuracy and is always optimal, then why would humans not have discovered this statistical paradigm earlier? Surely someone would have thought of this approach before now.

In fact, someone did. The notion of obtaining an optimal cutting score for classifying observations into categories was around for hundreds of years before Otis Dudley Duncan (Duncan, Ohlin, Reiss, & Stanton, 1953) and Paul Meehl (Meehl & Rosen, 1955) first developed the idea systematically during the 1950s. However, the computational brute force necessary to find an optimal cut score by hand rendered this procedure impractical for solving large-scale, real-world problems. With the advent of modern-day computers, however, this limitation vanished, and the availability of lightning-fast processing speed paved the way for the development of ODA. In other words, ODA would still be just a statistician's pipedream were it not for the computer technology that exists at this point in history. With the spread of this new technology, researchers have begun using optimal cut scores in a variety of different disciplines, including operations research, computer science, neural networks, medicine, and biology. Thus, ODA is riding a technology-based wave of expansion in the world of statistics.

It would be easy to get the impression that ODA has only recently emerged and is still wet behind the ears, but nothing could be further from the truth. On the contrary, Yarnold and Soltysik have devoted nearly 30 years of work to the creation of ODA, beginning in 1976 with a mixed-integer program called MIP and eventually developing a DOS-based ODA program in 1993 (Soltysik & Yarnold, 1993). The current version of ODA represents the life's work of two dedicated world-class scholars and is hardly a flash in the pan.

Researchers may also hesitate, as I did initially, to analyze their data using a nontraditional statistical technique that is not well known. Journal editors and reviewers alike are often suspicious of novel methodological or statistical approaches, and they generally prefer to publish articles that follow traditional scientific paradigms. However, rest assured—there is now an established and growing body of published research employing ODA as the primary statistical method of analysis. To support the book, the guidebook publisher (the American Psychological Association) has provided a Web site at which users will find a list of theoretical and applied journal publications featuring ODA, including research in the areas of clinical psychology, personality, psychiatry, criminal justice, family studies, and medicine (http://www.apa.org/books/resources/yarnoldsoltysik/).

Nevertheless, it is clear that shifts in the dominant scientific paradigm require time to evolve (Kuhn, 1970). Despite their reputation for progressive open-mindedness (Chang, 2000), scientists do not readily discard prevailing frameworks in favor of revolutionary approaches, and it is clear that ODA is revolutionary. Indeed, both the general linear model and structural equation modeling met tremendous resistance from the research community when these statistical methods first appeared on the scene. Yet, these two statistical approaches are now mainstream and well established. As was the case for these statistical paradigms, will it take a quarter century for ODA to break through the inevitable resistance to its use? One would hope not, given ODA's tremendous potential for the advancement of science. But only time will tell.

Weaknesses of the Guidebook and Software

No book is without its shortcomings, and as exceptional as it is, the ODA guidebook is no exception to this rule. Although readers do not need specialized knowledge of statistics to read the book, a working knowledge of quantitative analysis is definitely helpful. In this regard, the authors link ODA to a wealth of existing statistical approaches and explicate in detail its connections to numerous technical concepts, such as Type I error, prior odds, reliability, validity, sensitivity, specificity, and bootstrapping, to name just a few. Prior experience with these concepts would certainly be useful, although it is not absolutely necessary in order to learn to use ODA effectively.

In addition, some important statistical topics are conspicuous in their absence from the book. In particular, there is no explicit treatment of the concept of statistical power or power analysis in relation to ODA. Nor do the authors show the reader how ODA could be used to conduct such popular multivariate statistical procedures as multilevel regression modeling, survival analysis, meta-analysis, and missing data imputation. The authors might also have noted the kinds of traditional statistical analyses for which ODA would be inappropriate or more unwieldy to use, such as canonical correlation, factor analysis, or multidimensional scaling. Nevertheless, no one book can do it all. Despite these limitations, the guidebook is exceptionally comprehensive and well written and is likely to set a new standard for statistics texts in the years to come.

The ODA software also has several restrictions. For example, ODA accepts only ASCII files for data input, as was the case for initial versions of LISREL and EQS structural equation modeling software. This limitation requires users to convert their data file from whatever software they use to enter, check, and clean the data (e.g., SPSS, SAS, Excel) into a text file containing numbers, spaces, or tabs, with no other formatting information. The authors provide detailed instructions for how to create ASCII data sets for analysis by ODA, including how to convert SPSS, SAS, and Excel data files into ASCII format. However, it would be advantageous if ODA were capable of reading and automatically converting files from these other software packages without requiring users to reformat the data on their own (which, although not difficult, is probably the hardest step in running ODA). Perhaps the authors will offer this feature in an upgraded version of the ODA software, as has been done with LISREL and EQS.

As is the case for most statistics software, ODA also limits the number of observations users can analyze. Specifically, the largest sample size allowable is 65,535, although this limit is unlikely to affect most researchers. Users with larger samples can segment the total sample into smaller subsamples and use ODA to assess the cross- sample generalizability of optimal models.

The Future of ODA

The guidebook concludes with some intriguing descriptions of new ODA- based software packages currently in development. For example, the authors describe a forthcoming MultiODA program that finds a linear combination of attributes that provides optimal classification accuracy from among a set of potential predictors—a procedure they term *optimal discriminant analysis.* The authors also explain how to use ODA to construct optimal classification tree models consisting of multiple attributes that combine through a set of nonlinear branches to discriminate observations from different classes optimally—a technique that typically yields greater classification accuracy than linear approaches. The authors report they are in the process of developing an automated ODA-based classification tree program. In the meantime, ODA users can use the current software to construct such tree models branch by branch; ODA syntax can be used to include or exclude subsets of the sample iteratively. Be forewarned—what the guidebook contains is likely to whet your appetite and make you hungry for more.

Conclusion

Why struggle to make your data conform to the restrictive assumptions of traditional statistical tests, such as the assumptions of distributional normality, homogeneity of variance, and the absence of skewness? To meet these assumptions, researchers often transform their data, employing a host of tortuous mathematical alterations involving logarithmic, square root, reciprocal, or arc sine transformations; Winsorizing the data; or excluding outliers, with the goal of stabilizing variances and improving the shape of distributions (Hoaglin, Mosteller, & Tukey, 1983). When expected cell frequencies fall below five, researchers typically adopt a correction factor to improve the accuracy of chi-square analyses. When all else fails, researchers often fall back on the convenient concept of robustness to reassure themselves and others of the negligible consequences of violating the assumptions of statistical tests.

Instead, why not use a statistical procedure that is free from distributional assumptions and that, by design, conforms to the nature of the data? Why not use a statistical method that always finds the optimal model that maximizes classification accuracy? Why not use ODA? After all, when you make a prediction, would you rather be correct or incorrect?

References

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

Chang, L. (Ed.). (2000). *Scientists at work: Profiles of today's ground-breaking scientists*. New York: McGraw-Hill.

Duncan, O. D., Ohlin, L. E., Reiss, A. J., & Stanton, H. R. (1953). Formal devices for making selection decisions. *American Journal of Sociology*, *58*, 573–584.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Meehl, P. E., & Rosen, A. (1955). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Soltysik, R. C., & Yarnold, P. R. (1993). *ODA 1*. Chicago: Optimal Data Analysis.

©2012 American Psychological Association | PDF documents require Adobe Acrobat Reader | Terms and Conditions