1994

Analyze Your Data Optimally Using ODA 1.0

Fred B. Bryant
Loyola University Chicago, fbryant@luc.edu

■ **JACK YURKIEWICZ, Feature Editor, Pace University**

This review of a new and unconventional statistics program, Optimal Data Analysis, is from Dr. Fred Bryant of Loyola University in Chicago. Dr. Bryant was one of the beta testers of the program and has been using it for some time in his research and teaching.

Analyze Your Data Optimally Using ODA 1.0

by Fred B. Bryant, Loyola University

When you make a prediction, would you rather be correct or incorrect? If your answer is 'correct,' then ODA is the appropriate analytic methodology. (Soltysik & Yarnold, ODA Manual, 1993, p. 1)

This bold opening statement from the manual for Optimal Data Analysis (ODA) aptly conveys the wide-ranging scope and versatility of this innovative statistical package. ODA is a statistical paradigm for identifying an optimal discriminant function for assigning observations to categories with theoretically maximum possible accuracy. Although the idea of obtaining an optimal cutting score for classifying observations into categories has been around since the 1950s, the computational ability to solve large-scale problems has evolved only recently.

How does ODA compare to other more traditional statistical packages? There are currently two basic types of general-purpose statistical software packages on the market. The first of these correspond to what can be termed "traditional statistics." These procedures include the commonly used chi-square, *F*, and *t* tests. These procedures are heavily assumption laden (e.g., use of the latter two tests assumes that data are distributed normally and that there are equal variances across independent groups). Examples of this type of software include SAS, SPSS, BMDP, SYSTAT, and others.

The second type of statistical software package offers so-called "exact probability" statistics. These procedures allow one to compute (or estimate for larger problems) permutation probabilities that provide exact probabilities for traditional statistical procedures. Examples of this type of statistical package include RESAMPLING STATS, MATRIX, and SAS/STAT MULTTEST.

These first two approaches to data analysis share some common problems. First, each of them forces you to fit your data (more or less poorly) into a statistical model that makes assumptions about underlying distributions. That is, none of the traditional procedures is specifically tailored to a given application; rather a given application is analyzed using the procedure whose assumptions it violates least. Second, none of the traditional procedures are specifically optimized for accurate forecasting. That is, no traditional procedure specifically identifies a model that explicitly maximizes the percentage accuracy of classification (PAC) that the model achieves. Some traditional procedures, such as F-test or chi-square, do not provide models that can be used for forecasting purposes. Others, such as multiple regression or discriminant analysis, provide classification models; but these models do not explicitly maximize the criterion of achieving theoretically maximum PAC.

ODA does not fall into either of the above categories. It claims to solve the above problems by performing "optimal data analysis." The ODA paradigm is as follows:

1. For any particular data configuration, identify the variable that you wish to predict (this is called the "class variable").
2. Identify the variable(s) that you wish to use to try to predict the class variable (these predictors are called "attributes").

Identify the model(s) that explicitly maximizes the weighted or non-weighted percentage accuracy in classification (or PAC) that is theoretically achievable when using the attribute(s) to predict the class variable.

Compute (or estimate) a permutation probability for the resulting model(s).

Fred B. Bryant is Professor of Psychology at Loyola University Chicago, where he teaches graduate and undergraduate courses in statistics, research methods, and social psychology. After receiving his Ph.D. in social psychology from Northwestern University in 1980, he completed a three-year post-doctoral fellowship in survey research at the University of Michigan's Institute for Social Research. His research interests include structural equation modeling, meta-analysis, and the measurement of emotion.

ODA 1.0 Program for Example 5.4

```
OPEN EX5.4;
OUTPUT EX54.OUT;
FREE CLINCLAS CLINRAT NBMESCOR;
CLASS CLINCLAS;
RETURN CLINRAT;
ATTRIBUTE NBMESCOR;
DIRECTIONAL < 1 2 3 4 5 6;
PRIMARY MAXPAC;
SECONDARY BAL;
MCARLO ITER 10000 TARGET .05 STOP 99.99;
TITLE RETURN-WEIGHTED 6-CATEGORY ODA–INTERVAL DATA;
GO;
```

Figure 1: An annotated version of the ODA input statements used to find an optimal model for predicting clinical performance ratings from standardized test scores (from the ODA 1.0 manual).

ODA's Statistical Capabilities

ODA's elegantly simple paradigm gives rise to an wide array of statistical analyses. Many of these provide what the authors refer to as "optimal analogs" to traditional statistical procedures. For example, consider the following empirical example (5.4) from the ODA manual. Imagine that you wish to test the hypothesis that higher scores on a standardized medical aptitude test predict higher clinical performance ratings (as assessed on a six-point rating scale). Traditionally, one might wish to use one-way analysis of variance (ANOVA) or *F* test to evaluate this directional hypothesis. In this particular case, however, the number of subjects in each level of performance rating is highly imbalanced and their test scores are heterogeneous. Such conditions violate the underlying assumptions of ANOVA.

Alternatively, one can use ODA to test the hypothesis of interest, weighing prior odds to adjust for the imbalanced sample sizes. Figure 1 shows an annotated version of the ODA input statements (taken from the manual) used to test the hypothesis that test scores (NBMESCOR) and performance ratings (CLINCLAS) increase together. Whereas ANOVA yielded only a few weak effects, ODA revealed a highly significant model (with a confidence for p<.05 of 99.99%). Figure 2 shows an annotated version of the output (taken from the manual) for this particular optimal analysis. Note that the model yields an explicit classification rule specifying the relationship between the attribute and the class variable, thus eliminating the need for follow-up contrasts to interpret main effects (as is necessary with ANOVA). The manual illustrates many other popular types of traditional analyses for which ODA provides an optimal analog, including *t* test, correlation, chi-square, kappa, randomized block (and other experimental) designs, cluster analysis, Markov analysis, autocorrelation, item analysis, and the log-linear model.

ODA 1.0 Output for Example 5.4

```
Note: In all classification performance summary tables, the following
abbreviations are used: NP = Number Predicted; PV = Predictive Value;
NA = Number Actual; PAC = Percentage Accuracy in Classification.

Page 1                    ODA 1.0                23:52:56  10-27-1993

RETURN-WEIGHTED 6-CATEGORY ODA—INTERVAL DATA

Input file:   EX5.4
Output file:  EX54.OUT

   Class          Attribute          Weight          Group
   CLINCLAS       NBMESCOR           CLINRAT         OFF

Observations  Classes  Groups  Solution status  Time used (seconds)
    1994         6                 OPTIMAL            1038.9

Hypothesis   Categorical  Priors  Degen   Primary   Secondary
DIRECTIONAL  OFF          OFF     OFF     MAXPAC    BALANCED

ODA model:

IF NBMESCOR <= 385.5 THEN CLINCLAS = 1
IF 385.5 < NBMESCOR <= 401 THEN CLINCLAS = 2
IF 401 < NBMESCOR <= 421 THEN CLINCLAS = 3
IF 421 < NBMESCOR <= 556 THEN CLINCLAS = 4
IF 556 < NBMESCOR <=566 THEN CLINCLAS = 5
IF 566 < NBMESCOR THEN CLINCLAS = 6

Monte Carlo Summary:

Iterations          Estimated p
   200              0.000000

Confidence levels for estimated p:

Desired p    Confidence
p<.001       18.21%
p<.01        86.73%
p<.05        99.99%
p<.10        99.99%

Target p     Confidence
p<.05        99.99%
```

(Continued)

Figure 2: An annotated version of the ODA output for the analysis predicting clinical performance ratings from standardized test scores (from the ODA 1.0 manual).

However, ODA also enables new types of analyses that are not possible using any other software system. For example, ODA can be used to maximize the PAC achieved by commonly used multi-attribute procedures such as multiple regression (e.g., Yarnold & Soltysik, 1991: Refining two-group multivariable classification models using univariate optimal discriminant analysis, *Decision Sciences*, 22, 1158-1164). Consider the following empirical example from the ODA manual.

Medical patients were randomly assigned either to an experimental group (in which their physician advised them to quit smoking and offered them a nicotine substitute) or to a control group (in which the physician did not mention smoking). Patients were also asked whether or not they were willing to make a commitment to stop smoking. These two variables (experimental condition and willingness to make a commitment) were then used as independent variables in a traditional logistic regression to find a model for predicting whether or not people actually quit smoking one week later. The logistic regression model achieved overall classification accuracy = 76.9%; sensitivity for predicting quitters = 76.2%; a sensitivity for predicting continuing smokers = 81.9%; predictive value for quitters = 73.8%; and predictive value for continuing smokers = 78.8%.

ODA was then used to identify an adjusted intercept term (changed from .5 to .13) for the logistic model, which dramatically improved classification accuracy, particularly when predicting people who quit smoking: overall classification accuracy = 91.1%; sensitivity for predicting quitters = 100%; sensitivity for predicting continuing smokers = 61.1%; predictive value for quitters = 89.6%; and predictive value for continuing smokers = 94.8%.

ODA can also be used to optimize the PAC achieved by other commonly used multi-attribute models such as Fisher's linear discriminant function analysis, Smith's quadratic discriminant analysis, and probit analysis, to name a few. In addition, ODA offers multiple-sample analyses, hold-out (cross-generalizability) and/or leave-one-out (jackknife) validity analyses, weighing by prior odds and/or by cost or return, and one- or two-tailed hypothesis testing via Monte Carlo simulation—all for any data configuration.

Furthermore, ODA offers optimal parallel forms, split-half, inter-rater, test-retest, and intraclass reliability analyses, and optimal discriminant, convergent, and construct validation, and much more. Yet, for all its versatility, there are still some types of traditional statistical analyses that ODA cannot as yet handle (e.g., repeated measures MANOVAs, conjoint analysis, MDS).

Compared to the traditional statistical paradigm, I found that the approach of ODA to have many advantages. First is *conceptual clarity*. In the ODA paradigm, for every data configuration there is one precise optimal analysis. In traditional statistics, for a given application, several different

Classification performance summary:

Correct	Incorrect	Return	Overall PAC	Mean PAC across classes	Weighted PAC
645	1349	1932.25	32.35%	69.61%	34.76%

Class CLINCLAS (Actual)	Predicted 1	2	3	4	5	NA	PAC	WTD PAC	RETURN
1	11	9	7	45	0	81	76.92%	13.58%	13.75
2	17	11	17	147	0	228	4.82%	4.82%	19.25
3	11	13	13	152	0	255	5.10%	5.10%	29.25
4	47	18	44	412	0	751	54.86%	54.86%	1133
5	11	5	10	132	11	253	4.35%	4.35%	35.75
6	10	8	8	213	0	426	43.90%	43.90%	701.25
NP	107	64	99	1101	11				
PV	10.28%	17.19%	13.13%	37.42%	100.00%				

Class CLINCLAS	Predicted 6
1	9
2	36
3	66
4	230
5	84
6	187
NP	612
PV	30.56% MEAN PV 34.76%

Figure 2: (continued).

analyses are often feasible, and all are usually "suboptimal" in terms of PAC. Second is *ease of interpretation*. Every ODA analysis provides the same intuitive goodness-of-fit index: PAC. Different traditional statistical procedures, however, provide different goodness-of-fit indices that are both nonintuitive and noncomparable across procedures. Third is *ease of use*. Most ODA analyses require the same basic set of 6-10 commands. Fourth is *maximum accuracy*. Every ODA analysis provides a model that guarantees maximum possible PAC. In contrast, no traditional analysis provides a model that guarantees maximum PAC. Fifth is *"valid Type I error."* ODA provides permutation probabilities and requires no simplifying assumptions: *p* is always valid. Traditional analyses require simplifying assumptions, and *p* is valid only if the assumptions are true for one's data.

Documentation

The ODA manual (hard-bound; 200 pages) is written like a textbook. Using a minimum of formulas, the manual discusses everything you need to know to use ODA. Comprehensive and well-organized, it includes a wealth of references to published empirical examples in a host of literatures and a collection of 30 hypothetical applications in different fields, ranging from astronomy, credit screening, epidemiology, and farming, to personnel selection, target recognition, weather forecasting, and zoology. Both students and educators can use the manual to provide interesting data-driven examples that clearly illustrate how to use ODA. The ODA software also includes a collection of over 60 actual raw-data sets that are analyzed and interpreted in the manual, many including well-annotated input statements and printouts (see Figures 1 and 2). This carefully crafted package is an excellent teaching tool.

Technical Information

ODA is a command-driven software program that may be run in batch mode or interactively. Although ODA has no graphics capabilities, this in no way impairs one's ability to understand the results of analyses (see sample output in Figure 2). The program is very compact, (it comes on one diskette), easy to install, and extremely fast. It requires 640K of RAM and 500K of disk storage, and uses a math co-processor if available. The program currently allows a maximum of approximately 8200 observations for applications involving ordered (e.g., ordinal, interval, or ratio-scale) attributes, and an unlimited number of observations if the data are categorical (e.g., qualitative). Run times for average problems on a 386DX 40-Mhz IBM-class PC with math co-processor range from 1 to 2000 seconds.

Summary

ODA's simplicity lends it an appealing conceptual elegance and makes it exceptionally easy to use. Unlike any other existing statistical system, ODA provides a unifying paradigm for analyzing the full spectrum of data configurations encountered in scientific research.

Pricing Information

A variety of purchase options exist for Optimal Data Analysis 1.0. The regular single copy price is $499. For orders of two or more copies, there is a $100 discount per copy. In addition, academic faculty may subtract $100 from the above prices, and students may subtract $200. Site and network licenses are also available.

Finally, a "classroom" price, available to educational institutions, is $99. This price includes the manual, diskette, and technical support, and requires a minimum order of six copies. ■

Optimal Data Analysis, Inc.
708 W. Bittersweet, Suite 403
Chicago, Illinois 60613
312-528-6092

If you are interested in writing a software review for a future issue of *Decision Line*, please call me at Pace University (212) 346-1908, or e-mail: yurk@pacevm.dac.pace.edu.

***DECISION SCIENCES* JOURNAL**

Decision Sciences Call for Papers

Decision Sciences invites submission of papers for its first research focus on Global Quality Management. A "research focus" is a group of top quality papers, usually five to seven, that concentrate on a topic of special interest to the readership. Effective quality management is a key element for survival in global competition. Most business priorities, such as product and volume flexibility, low price, fast and dependable delivery, and good customer service, are built on a foundation of good quality. A lack of comparative research on quality management across nations has hampered our understanding of the influences of culture, infrastructure, and other systematic sources on quality management. Against this backdrop, *Decision Sciences* aims to publish a timely research focus that contains a balance of theoretical and empirical articles on quality management with a global perspective.

While all papers addressing topics of quality management that have strategic implications for managers facing competition in a global economy are welcome, priority for the limited space will be given to top quality papers that address issues of quality management in the developing and newly developing countries, and to comparative studies of quality practices in Asian and Western countries. The editor of the research focus encourages papers that examine the implications of macro-level influences on quality practices, as well as papers involving a multidisciplinary approach in manufacturing or service settings. Suggested topics include (but are not limited to) the following:

- Global benchmarking for quality practices
- Comparative study of quality practices in Asian and Western nations
- Relationship between quality and other competitive priorities
- Performance measures for quality management
- Implications of new product development for quality management
- Quality issues in the coordination of manufacturing, marketing, and/or R&D
- Human resource development issues in quality management
- Supply chain management and the issue of quality
- Quality management in service industries
- Role of information in quality improvement.

This research focus of *Decision Sciences* will be edited by Dr. Kee Young Kim, College of Business and Economics, Yonsei University, 134 Shinchondong, Sudaemungu, Seoul, 120-749 Korea. Tel: 82(2)361-2500, Fax: 82(2)313-5331. All papers will be reviewed according to the standard *Decision Sciences* review process. If warranted, a complete issue of nine to twelve articles will be devoted to the topic. Papers of high quality that cannot be included in the research focus will be considered for a regular issue of *Decision Sciences*. Information for contributors can be found in any issue of *Decision Sciences*.

The deadline for submissions is October 3, 1994. ■

Send four copies of your submission to:
The Editor, *Decision Sciences*
Research Focus: Global Quality Management
The Ohio State University
College of Business
301 Hagerty Hall
1775 College Road
Columbus, OH 43210-1399