

Optimal Data Analysis, LLC

Maximizing Predictive Accuracy

Paul R. Yarnold, Robert C. Soltysk

2016

6348 N. Milwaukee Avenue, Chicago, Illinois 60646

Maximizing Predictive Accuracy

Paul R. Yarnold and Robert C. Soltysik

**Optimal Data Analysis, LLC
Chicago, IL**

Copyright @ 2016 by Optimal Data Analysis, LLC. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

Published by:

Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
www.ODAJournal.com

To order:

Online: www.ODAJournal.com/resources/
Optimal Data Analysis, LLC
Book Department
6348 N. Milwaukee Ave., #163
Chicago, IL 60646

Typeset in Calibri by Optimal Data Analysis, LLC, Chicago, IL

Printer: <https://www.diggypod.com/>
Cover Designer: Optimal Data Analysis, LLC, Chicago, IL
Technical/Production Editor: Paul R. Yarnold, Ph.D.

The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of Optimal Data Analysis, LLC—bear in mind, and suffice it to say, that new discoveries in this field are constantly emerging.

Library of Congress Cataloging-in-Publication Data

Yarnold, Paul R.

Maximizing predictive accuracy / Paul R. Yarnold and Robert C. Soltysik.—1st ed.

p. cm.

Includes bibliographical references and index.

ISBN-10: 0-692-70092-7

ISBN-13: 978-0-692-70092-1

DOI: 10.13140/RG.2.1.1368.3286

Printed in the United States of America

First Edition

First Printing

My creation for my Creator -- Paul R. Yarnold

For Samuel -- Robert C. Soltysik

Contents

Preface xi

Acknowledgments xii

Introduction

Chapter 1	<i>Pragmatic Considerations</i>	1
	Learning, Publishing, and Teaching ODA	2
	Obtaining Research Funding	5
	Commercial Applications	6
	Improving Science	6
Chapter 2	<i>Fundamental Concepts</i>	9
	The UniODA Algorithm	9
	Establishing Statistical Reliability	11
	Criterion for Statistical Significance	14
	Assessing Classification Accuracy	15
	The CTA Algorithm	17
	Data Transformations	21
	Weighting Observations	22
	Pre-Processing Data	22
	Normative Standardization	25
	Ipsative Standardization	25
	Interactive Transformation	28
	Two Common Mistakes	28
	Evaluating Model Reproducibility	32
	Leave-One-Out (Jackknife) Analysis	32
	Hold-Out Analysis	33
	Multisample Generalizability Analysis	38
	Simpson's Paradox	39
Chapter 3	<i>Methodological Matters</i>	40
	Measurement Scales	40
	Class Variables	41
	Attributes	42
	Weights	43
	Precision	43
	Adaptability	48
	Instrumentation	52
	Statistical Power Analysis	53
	Parametric Approach	54
	Exact Minimum Precision Approach	56
	Model Geometry and Sample Size	59

Data Set Design	60
Initial Issues to Resolve	61
Exporting a Source Data File	62
Quality Assurance	64
Running ODA Software Using DOS Prompt	64
Treatment of Missing Data	65
The Role of Residuals	66
Reporting Analytic Findings	66
Descriptive Statistics	67
UniODA Findings	67
Multiattribute ODA Findings	71

Bivariate Methods

Chapter 4	<i>UniODA with Categorical Attributes</i>	72
	Bowker's Test for Symmetry	72
	Bray-Curtis Dissimilarity Index	74
	Chi-Square	75
	Not Chi-Square	78
	Cochren's Q Test	81
	Cohen's Kappa	84
	Log-Linear Model	86
	Logistic Regression	87
	Markov Processes	88
	McNemar's Test for Correlated Proportions	90
	Turnover Table	92
	Consecutive Codes Can Repeat	92
	Consecutive Codes Can't Repeat	93
	Multisample Analysis	97
Chapter 5	<i>UniODA with Ordered Attributes</i>	99
	Kendall's Coefficient of Concordance	99
	Kruskal-Wallace Test	100
	Mann-Whitney U Test	102
	One-Way Analysis of Variance	105
	All Possible Comparisons	105
	Optimal Range Test	107
	Polychoric Correlation	110
	Reliability Analysis	111
	Inter-Rater Reliability Analysis	111
	Inter-Method Reliability Analysis	117
	Paradoxical Confounding in Reliability Assessment	119
	Split-Half Reliability	119
	Test-Retest (Temporal) Reliability	120
	Repeated-Measures	121
	ROC Analysis	128

	Student's <i>t</i> -Test	129
	Between Subjects	129
	Within Subjects	131
	Validity Analysis	134
	Construct Validity	134
	Convergent and Discriminant Validity	137
<i>Linear Multiattribute Methods</i>		
Chapter 6	<i>Optimized General-Linear Models</i>	142
	OLS Regression Analysis	142
	Regression Toward the Mean	142
	Regression Away From the Mean	144
	Optimizing Regression Models	155
	Analysis of Variance	156
	Linear Discriminant Function	156
Chapter 7	<i>Optimized Maximum-Likelihood Models</i>	158
	Log-Linear Model	158
	Probit Model	158
	Logistic Regression	159
	Categorical Attributes Having Many Levels	160
	Categorical Attributes May Overwhelm Linear Models	160
Chapter 8	<i>Explicitly Optimal Linear Models</i>	173
	MIP45 Mixed Integer Programming Formulation	173
	Resolving Classification Gaps and Ambiguities	175
	Weighted Classification	176
	Adding Nonlinear Terms as Attributes	177
	Optimal Attribute Subset Selection	177
	Aggregation of Duplicate Observations	177
	WARMACK Search Algorithm	179
	Multicategorical Class Variables	180
	MultiODA Research in the ODA Laboratory	180
	Special-Purpose MultiODA Models	180
	Boolean ODA	180
	Exact ODA	181
	Tau ODA	181
	Template ODA	182
	Unit-Coefficient Models	182
Chapter 9	<i>Identifying and Ameliorating Statistical Confounding</i>	183
	Confounding by Covariates	183
	Partial UniODA	183
	Unconstrained Covariates	185
	Exploratory Methods	187
	Multiple Confounders	196

Marginal Structural Models	197
Confounding by Combining Groups	198
“Junk Science” in the Courtroom	204
Confounding by Combining Time Periods	205
Measuring Atmospheric Circulation Patterns	207
Unconfounded Measurement of Major Modes	210
Qualitative Interpretation of Ipsative Modes	214
Predicting Temperature Anomalies	216
Predicting Precipitation Anomalies	217
Predicting Export of Arctic Sea Ice	217
Confounding in Single-Case Series	219
Comparing Two Serial Ratings	219
Interactive Measurement	222

Non-Linear Multiattribute Methods

Chapter 10	<i>Hierarchically Optimal Classification Tree Analysis</i>	231
	Obtaining an HO-CTA Model	231
	Determining the Minimum <i>N</i> for HO-CTA Model Endpoints	231
	Growing the HO-CTA Model	232
	Pruning the Fully-Grown HO-CTA Model to Ensure Maximum-Accuracy	239
	Forward HO-CTA	242
	Reverse HO-CTA	242
	Multiple Regression Analysis	243
	Motivating Reverse HO-CTA	243
	Defining Attributes	243
	Obtaining the Model	244
	HO-CTA in Applied Research	246
	Clinical Medicine	247
	Psychosocial Aspects of Medicine	249
	Allied Health Disciplines	249
	Conclusion	250
Chapter 11	<i>Enumerated Optimal Classification Tree Analysis</i>	251
	Obtaining an EO-CTA Model	251
	Context of the Exposition	251
	Determining the Minimum <i>N</i> for EO-CTA Model Endpoints	251
	Obtaining the EO-CTA Model	252
	HO-CTA versus EO-CTA Models	254
	In-Hospital Mortality from <i>Pneumocystis cariini</i> Pneumonia (PCP)	254
	Psychosocial Adaptation in Early Adolescence	256
	Person-Environment Fit Theory and Freshman Attrition	257
	Parsing Attributes	258
	Modeling Moderating Effects	259
	Enigma	260

Chapter 12	<i>Globally Optimal Statistical Analysis</i>	261
	Defining a Theoretically Ideal Statistical Model	261
	Comparing Empirical and Theoretically Ideal GO-CTA Models	262
	Novometric Theory	263
	Axiom 1: Statistical Sample	263
	Axiom 2: Structural Decomposition Analysis	263
	Axiom 3: Descendant Family	264
	Axiom 4: Model Reproducibility	264
	Novometric Theory and Quantum Mechanics	264
	Exact Discrete Confidence Intervals	266
	Model Endpoint Redundancy	267
	Obtaining a GO-CTA Model: Binary Class Variable, One Attribute	268
	Obtaining a GO-CTA Model: Binary Class Variable, Multiple Attributes	289
	MDSA without SDA	290
	MDSA with SDA	292
	Obtaining a GO-CTA Model: Unrestricted Class Variable and Attribute(s)	293
	Analysis Involving Missing Data	293
	The Best is Yet to Come	295
Appendix A:	UniODA and MegaODA Command Syntax	298
	Running ODA Software	305
Appendix B:	MegaODA Time Trials	306
	Study 1: Identifying ns Effects	305
	Relatively Large Samples	305
	Big Data Samples	307
	Study 2: Identifying Statistically Significant Effects	308
	Study 3: Binary Designs	310
Appendix C:	CTA Command Syntax	312
	Interpreting Automated CTA Software Output	316
Appendix D:	Troubleshooting ODA Software	319
Appendix E:	Weather Prediction Results	320
Chapter References		338
Alphabetical References		362
Index		388
About the Authors		396

Preface

In 1977 we began collaborative full-time scientific investigation of the application of mathematical optimization in the context of the statistical classification problem, and we continue this pursuit today. Our research gave rise to the optimal data analysis (ODA) statistical paradigm, which identifies models that explicitly optimize (maximize) predictive accuracy for every unique application. This book discusses the ODA paradigm beginning with its genesis, and concluding with state-of-the-art methods in use today.

“A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.” [Max Planck: *Wissenschaftliche Selbstbiographie. Mit einem Bildnis und der von Max von Laue gehaltenen Traueransprache*. Johann Ambrosius Barth Verlag (Leipzig 1948), p. 22, as translated in *Scientific Autobiography and Other Papers*, trans. F. Gaynor (New York, 1949), pp. 33–34 (as cited in T. S. Kuhn, *The Structure of Scientific Revolutions*).]

Acknowledgements

We appreciate support given to us by our family and friends, colleagues and teachers. Colleagues who were particularly instrumental in helping to bring this work to completion are Fred B. Bryant, Ph.D., and Ariel Linden, DrPH.

“New scientific ideas never spring from a communal body, however organized, but rather from the head of an individually inspired researcher who struggles with his problems in lonely thought and unites all his thought on one single point which is his whole world for the moment.” [From address by Max Planck, the father of Quantum Mechanics, on the 25th anniversary of the Kaiser-Wilhelm Gesellschaft (January 1936), as quoted in *Surviving the Swastika : Scientific Research in Nazi Germany* (1993) ISBN 0-19-507010-0.]

Maximizing Predictive Accuracy

Chapter 1

Pragmatic Considerations

It is natural to wonder, and important to understand, why this new statistical paradigm is called the *optimal data analysis* (ODA) paradigm. The answer is that the word “optimal” represents corner-stone nomenclature (jargon) in the field of mathematical optimization, and ODA harnesses techniques from this field to identify statistical models that *explicitly* maximize (optimize) classification accuracy obtained for the training sample used in model development.¹⁻⁴ Mathematically, classification accuracy is the *objective function* that is maximized by ODA models, rather than, for example, variance or the value of the likelihood function—objective functions maximized by the General Linear Model (GLM) and the Maximum-Likelihood (ML) paradigms, respectively.^{5,6} By definition, it is impossible for an alternative model to yield greater classification accuracy than an ODA model in a given application, as it is likewise impossible for an alternative model to explain more variance in a given application than a GLM model—because the GLM paradigm maximizes (optimizes) explained variation. In this context “optimal” implies a statistical model explicitly achieves the theoretical maximum attainable classification accuracy for a given application.

It is also natural to wonder, and crucial to understand, what is meant by *classification accuracy*. The answer is clear if one imagines creating a statistical model that is capable of accurately classifying (i.e., “predicting”) observations from different groups. If the predicted group and the actual group for an observation are the same, then a point is scored. If the predicted and actual group membership are different for an observation, then no point is scored. In every specific application the ODA model explicitly maximizes the overall number of points obtained. Observations may also be weighted, in which case the optimal (most accurate, greatest number of points) *weighted solution* is identified.¹⁻³

An *optimal solution* exists for every unique combination of sample, data geometry, and hypothesis. If a statistical *model* explicitly identifies the most accurate possible solution for a specific application, then the model is optimal. If a statistical *methodology* explicitly obtains the most accurate possible solution, then the methodology is optimal. This language is definitional—it is the meaning of the word *optimal* in this context. Any statistical method that fails to explicitly prove an optimal (maximum accuracy) solution is defined as *suboptimal*. Statistical methods that explicitly seek maximum accuracy by design (i.e., by formulation), but that fail to explicitly prove optimality, are known as *heuristic methods*. Neither GLM nor ML represent heuristic methods, because neither paradigm seeks maximum accuracy by formulation—rather, they seek variance or value of the likelihood function.

To be *explicitly optimal*, a method must be specifically formulated to find the most accurate solution possible for a given application. Obtaining an optimal model requires the use of various methods such as mathematical programming, linear algebra, integer programming, and so forth. Each unique application represents a unique optimization problem. For every unique statistical application the optimal solution is obtained using mathematical optimization methods—including, of course, using specifically-engineered UniODA, MegaODA, and CTA software.

To be thorough we note that preceding the discovery of the ODA paradigm, researchers in fields such as operations research, computer science, and systems engineering created linear discrimination algorithms—collectively known as *optimal discriminant analysis* models—that explicitly maximized classification accuracy for a training sample: the genesis of the acronym ODA. Researchers representing a vast domain of substantive areas, and 145 countries, are presently learning the ODA paradigm: our laboratory thus increasingly characterizes ODA procedures in the context of being exact maximum-accuracy statistical methods, in the hopes of maximizing conceptual clarity.

Learning, Publishing, and Teaching ODA

Learning the ODA paradigm is accomplished most easily by *working this book*—the most comprehensive exposition of the ODA statistical paradigm available. Working this book necessitates the use of statistical software. In tandem with this book, the UniODA, MegaODA and CTA statistical analysis software systems were designed to enable motivated researchers to learn, conduct, interpret, and disseminate different types of ODA analyses with maximum efficiency. We recommend a careful, sequential reading of each chapter, and when it is possible solving every example problem with ODA software: constructing a proper data set, operating software correctly, and interpreting statistical findings are crucial analytic skills that should be practiced and honed. We recommend running variations of each problem—with versus without weighting by prior odds, specifying a (non)directional hypothesis, using half of the sample as a hold-out group, and so forth—in order to become familiar with the operations and effects of software parameter settings in the context of scientific hypothesis testing. When data used in an example in the book aren't provided, obtaining a data set possessing parallel geometry is required to work the book: while simulated data are fine for this purpose, effort expended in collecting actual data may pay dividends in the form of discoveries, publications, and possibly in the form of ever-more-difficult-to-obtain research funding.

Publishing using the ODA paradigm is simplified by working this book—which discusses many of the most widely, frequently reported statistical designs used in empirical science. As each type of example problem is mastered, identify a new parallel data set and replicate the analysis using the new data. It is a wise idea to begin with very simple designs. Such analyses are easily publishable in a myriad of methodological and substantive journals in scientific and engineering subdisciplines (such articles are cited herein). If you have analyzed data using legacy statistical methods, the same statistical hypothesis can be evaluated for those data via ODA, corresponding results compared side-by-side, and findings published in substantively appropriate methodological subspecialty journals (such articles are cited herein). Once UniODA and CTA are mastered all empirical journals represent publication opportunities, because models obtained using these methods are more accurate and parsimonious than models otherwise obtained.

Peer-reviewed articles that focus on the ODA paradigm, currently being read by scientists from 130 countries, are available at the open-access eJournal *Optimal Data Analysis* (<http://ODAJournal.com>), and articles on ODA published in other journals by many independent labs (including the ODA laboratory) are listed in the Publications tab at the eJournal website. The ODA eJournal is also an excellent choice for disseminating original articles on ODA: there are no minimum or maximum article length restrictions; authors receive a timely, prescriptive, expert review; there is minimal lag-time in publication of accepted manuscripts; and there is no publication or access fee.

Teaching the ODA paradigm served as a primary motivation for writing this book. Although we've presented continuing education courses, seminars, and invited colloquia on ODA, we haven't yet taught university courses that focus exclusively on the ODA paradigm, and are unaware of others yet teaching such courses. We believe this is due in large part to the absence of a comprehensive book on this rapidly expanding new statistical paradigm. This book can serve as the text for a one-year academic course for graduate and advanced undergraduate students who completed standard program-required statistics courses. For semester-based programs the first course can include Chapters 1-7, and the second course Chapters 8-13. For trimester-based programs the first course can cover Chapters 1-5, the second course Chapters 6-9, and the third course Chapters 10-13. A laboratory course should in principle accompany the lecture class, and students should be encouraged to publish their laboratory projects (analyses of data sets consistent with lecture topics).

Some evidence suggests that even minimal exposure to the ODA paradigm may sufficiently intrigue academic faculty and students to motivate exploration and utilization of ODA methods. For example, in 1996 Paul Yarnold, Ph.D., presented a colloquium on ODA in the Department of Psychology of Loyola University Chicago. His best friend, Fred Bryant, PhD, a Professor of Psychology at Loyola, was the moderator. In 2010 Fred wrote an invited article for the ODA eJournal about the Loyola experience with ODA. His report⁷ is reproduced here.

Abstract: This article traces the origins and development of the use of optimal data analysis (ODA) within the Department of Psychology at Loyola University Chicago over the past 17 years. An initial set of ODA-based articles by Loyola faculty laid the groundwork for a sustained upsurge in the use of ODA among graduate students which has lasted for more than a decade and a half. These student projects subsequently fueled an increase in ODA-based publications by other Loyola Psychology faculty, who directly supervised the various student projects. Thus, ODA initially trickled down from faculty to students, but later grew up in the opposite direction. The most frequent use of ODA in Loyola's Psychology Department has been to conduct classification tree analysis, with less common uses of ODA including optimal discriminant analysis and the iterative structural decomposition of transition tables. As more Loyola Psychology graduate students find academic jobs and continue using ODA in their research, we expect that they will replicate the Loyola experience in these new academic settings.

When you discover a new tool that you believe is superior to other tools you've used before, naturally you want not only to use the new tool, but also to tell others about it so they can enjoy its benefits too. Such has been the case in the Department of Psychology at Loyola University Chicago since early 1993, when the first version of Optimal Data Analysis (ODA) 1.0 for DOS became publicly available. The purpose of this brief article is to describe the 17-year process through which the use of ODA sprang up, took hold, and spread among graduate students and faculty in Loyola's Psychology Department.

The Early Days of ODA at Loyola: I have known Paul Yarnold and Rob Soltysik since they first began working on the problem of optimal classification in the early 1980s. I served as a beta-tester for both the original DOS-based¹ and more recent Windows-based² versions of the ODA software. In late 1992, I cheered from the sidelines as Paul and Rob put the finishing touches on ODA 1.0 for DOS. And when ODA 1.0 for DOS appeared in print, I wrote the first published review of the new software³ and began using ODA in my research. Later I also published the first review of ODA for Windows.⁴

Having fallen in love with the power, versatility, and elegance of ODA, I began publishing research articles using ODA as a statistical tool in 1994.¹⁰ I first directly collaborated with departmental colleagues to use ODA in 1996, in publishing an article using optimal discriminant analysis as an alternative to Student's *t* test with two Loyola clinicians in the *Journal of Consulting and Clinical Psychology*.¹¹ At the same time, I continued publishing ODA-based research on my own, and began extolling the capabilities of the new ODA software to my graduate students. Interestingly, it was the graduate students, rather than the faculty, who more eagerly embraced ODA as a statistical tool in their research.

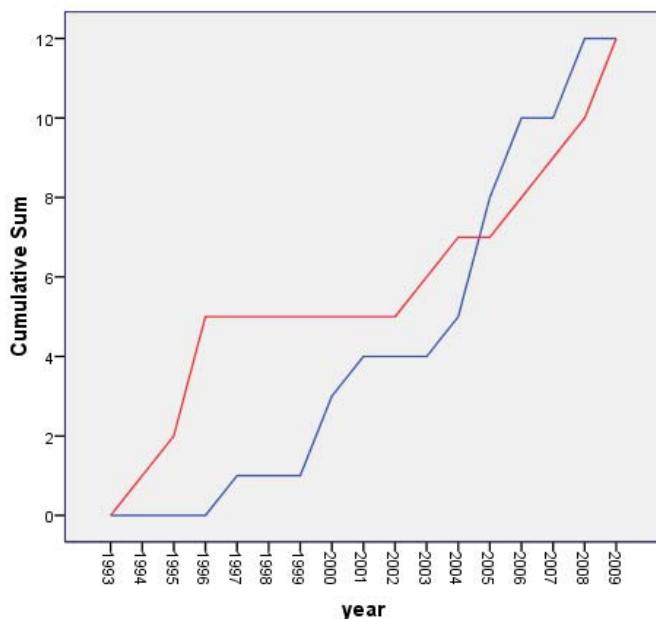
How Loyola Researchers Have Used ODA: At Loyola, researchers have used ODA in multiple ways to address a wide variety of different research questions in clinical psychology, social psychology, neuropsychology, behavioral medicine, and biochemistry. Figure 1.1 illustrates the cumulative number of faculty publications (shown in red) and graduate student projects (shown in blue) from 1993 to 2009.

Note the patterns that emerge across the 17-year span. The Loyola Psychology Department's experience with ODA originated in the early publications by department faculty. This initial set of articles laid the groundwork for a sustained upsurge in the use of ODA by Loyola graduate students over more than a decade and a half. These graduate student projects subsequently fueled the increase in ODA-based publications by other Loyola Psychology faculty, who directly supervised the various student projects. Although ODA initially trickled down from faculty to students, it later grew up in the opposite direction.

Classification Tree Analysis: By far, the most frequent use of ODA at Loyola has been to conduct multiattribute classification tree analysis (CTA). For example, Loyola graduate students have used CTA to identify predictive models for discriminating students who drop out versus return to college following the first year¹², children's emotional responsiveness versus unresponsiveness during psychotherapy¹³, child molesters versus non-molesters¹⁴, positive versus nonpositive adaptation to childhood¹⁵, convicted juvenile delinquents versus non-delinquent youth¹⁶, positive versus negative morbidity and mortality outcomes following bone marrow transplant¹⁷, high versus low effect sizes in a meta-analysis of methodological and intervention characteristics associated with primary prevention programs for children and adolescents¹⁸, engaging versus not engaging in risky sexual behavior among minority adolescents¹⁹ and adult male homosexuals²⁰, high versus low social competence among children with spina bifida²¹, and state mental health care agency decisions to commit children to residential treatment versus foster homes.²² In addition, department faculty and graduate students have jointly published journal articles using CTA to predict early sexual debut among adolescents²³, positive adaptation to childhood²⁴, psychiat-

ric hospital admission decisions for children in foster care²⁵, malingering in forensic neuropsychological examinations²⁶, change in job status following traumatic brain injury²⁷, and clinically significant sexual concerns in a child welfare population.²⁸

Figure 1.1: Loyola Psychology Department Publications and Dissertations/Theses Using ODA



Optimal Discriminant Analysis: The next most common use of ODA in Loyola's Psychology Department has been to conduct optimal discriminant analysis, as an exact-probability alternative to parametric discriminant analysis or Student's *t* test. For example, Loyola faculty publications have used ODA in this fashion to discriminate Type As versus Type Bs using the Type A Self-Rating Inventory¹⁰ and the Students Jenkins Activity Survey²⁹, males versus females in self-ratings of affective intensity³⁰, high-versus low-quality child therapy sessions based on therapist discourse¹¹, and physicians versus undergraduates in levels of sympathy and empathy.³¹ Layden et al. used this form of discriminant analysis to identify an optimal cut-score for using psychiatric ratings to assess toxicity in patients undergoing lithium treatment for bipolar depression.³²

Iterative Structural Decomposition: Another Loyola dissertation in clinical psychology used ODA to conduct an analysis for which no alternative statistical test exists. In this particular project, the student had couples discuss an area of disagreement in their marriage for 15 minutes, and then used an established interaction scoring system to code these interactions. Based on existing theory, the student predicted that couples having only one depressed spouse would engage in the following sequence of behaviors: (a) depressive behavior, followed by (b) spouse's supportive behavior, followed by (c) more depressive behavior, followed by (d) spouse's incongruent behavior, followed by (e) angry/defensive behavior, followed finally by (f) spouse's critical/rejecting behavior. Following procedures outlined by Yarnold and Soltysik² (pp. 209-222), the data were organized into transition tables representing the frequencies of various verbal exchanges between spouses over time. Supporting the hypothesized temporal model, an iterative structural decomposition of the transition tables revealed that the data conformed to the predicted sequence of behaviors significantly more than would be expected by chance alone.³³ It is unclear how one would test the hypothesized behavioral sequence using any other inferential statistical tool.

The Future of ODA in Psychology: If the past is any indication of the future, then ODA has a bright future, not only at Loyola but elsewhere. The recent availability of ODA-based software that automatically constructs classification tree models is likely to accelerate the use of CTA across a wider variety of research disciplines. In the future, enumerated CTA models may well replace traditional hierarchically-optimal CTA models, particularly given the superior classification accuracy of the former. The automated CTA software also offers the ability to analyze class variables that have more than two levels, thereby enabling new forms of nonlinear optimal regression modeling. We can foresee a vast array of new applications for CTA, including meta-analysis, cross-cultural tests of similarities and differences, mediation and moderation models, and optimal path analysis.

Obviously, it is relatively easy to export the Loyola Experience with ODA to other universities. All that is needed is a faculty member to lay the groundwork through an initial set of ODA-based publications, along with graduate students who are seeking to analyze data for their dissertation or master's thesis, or for research presentations and publications. As more Loyola Psychology graduates find academic jobs and continue to use ODA in their research, we expect that they will replicate the Loyola experience in these new academic settings.

I close by noting an unanticipated aspect of the Loyola experience with ODA. Namely, some of the graduate students who have used ODA in their dissertation research have later had the opportunity to teach introductory statistics in psychology at the undergraduate level, both at Loyola and at other colleges and universities. Naturally, these graduate instructors have taught their students about ODA and its statistical advantages, and these undergraduates are now approaching faculty members in psychology at Loyola and elsewhere to supervise independent research projects and honors theses that use ODA. Once again, the process of learning has come full circle, as the students themselves become teachers and disseminate statistical methods to students, faculty, and beyond.

Obtaining Research Funding

Since beginning collaborative study of maximum-accuracy discriminant classifiers in 1979, this research focus has been our primary vocation. Although our research funding derives from multiple sources, we devoted most of our attention to theoretical and applied research so our research funding to date has derived primarily from research grants. As of this writing we have received competitive grant funding for ODA-based statistical collaboration on projects totaling more than \$14 million (USD) in direct costs, from numerous public and private research institutions including the National Fund for Medical Education; Northwestern Memorial Foundation; National Center for Supercomputing Applications; National Science Foundation; Public Health Service—including the Health Care and Policy Research, Drug Abuse Health Services Research, and HIV/AIDS Agencies; National Institute of Health; National Institute on Disability and Rehabilitation Research; AHRQ Small Research Grant Program; American Academy of Allergy, Asthma and Immunology; Department of Veteran's Affairs; US Department of Health and Human Services; American Cancer Society; NASA; and SBIR-DARPA.

Our funded research areas included predicting physician resource utilization and quality of care; cascade iatrogenesis in older patients hospitalized for high-risk illnesses; optimal resident selection using mixed integer programming; psychological influences on physician practice decisions; assessing functional status of the elderly via microcomputer; prospective outcome study of intermittent claudication; health risk appraisal for Medicare patients; multicity study of quality of care for HIV-related PCP; quality of care for bacterial pneumonia in patients with HIV; development of a colorectal cancer outcomes database; chronic prostatitis collaborative clinical research; quality of life assessment for women with and at increased risk for the development of ovarian cancer; study of instruments for measuring satisfaction with care in prostate cancer; spinal injury risk assessment for thromboembolism; patient safety and system errors reduction; improving asthma care in the low income elderly; case-controlled study of TTP incidence rates and risk factors; improving asthma care for the elderly; research for adverse drug-events and health (RADAR); the impact of health literacy on racial differences in cancer stage at presentation; development and validation of a simulator-based pediatric emergency medicine curriculum for emergency care provid-

ers; evaluation of adverse drug events during cancer drug clinical trials; PSA rising as a national concern in prostate cancer care; and novel multiple myeloma drugs.

We are aware of at least three other independent research laboratories (one each in DC, FL, and MA) that recently received extramural funding for applied research using ODA statistical methods.

Commercial Applications

One avenue for commercial application of ODA methods is statistical consulting. The consulting that is offered by the ODA laboratory involves design, analysis, and report generation for individual researchers, and for private and public institutions. Our consulting thus far has primarily involved applications in the fields of aviation, business administration, criminal justice, environmental sustainability, human resources (e.g., personnel selection; patient satisfaction; psychological testing), and pharmaceuticals. We have colleagues who use ODA in their consulting in the fields of trial law, clinical sciences (neurology, psychiatry, psychology), education, social services, and nursing.

A second avenue is private instruction in ODA methods offered to individual researchers, public institutions, and private and public corporations: tutoring, seminars, colloquia, continuing education courses, and adjunct college courses are examples of mentoring opportunities.

A third avenue for commercial application of ODA methods is business development, of which one modality is becoming a new purveyor of goods or services for sale. In this domain the ODA laboratory uses optimal methods to create special-purpose black-box integrated systems. For example, fibromyalgia (FM) is a chronic illness without medical cure with prevalence estimated as high as twelve million Americans—primarily women of child-bearing age, although children, elderly, and men are also affected. The ODA laboratory designed, constructed, alpha³⁴ and beta³⁵ tested, and clinically evaluated a unique, proprietary, interactive, user-friendly, web-hosted, evidence-based, single-case, algorithmic approach to FM self-management. The commercial release of this system is anticipated in 2016.

Finally, a second modality of business development is partnering with an established purveyor of goods or services for sale, particularly as regards creating maximum-accuracy decision-making algorithms. Many of the examples presented in this book have obvious implications for such business opportunities: disease-staging models are critical in the insurance business; loan default models are crucial in financial businesses; symptom reduction and functional status improvement models are critical in rehabilitation businesses; job placement models are crucial in employment agencies; models of customer purchasing behavior are critical in advertising, marketing, and production businesses; and so forth.

In this modality of business development, for more than two decades the ODA laboratory has conducted research on the measurement and modeling of an individual's satisfaction ratings in a variety of contexts. This book presents models of patient satisfaction with medical care received, all of which are qualitatively superior to suboptimal alternative models. The utility of ODA has been clearly demonstrated in the field of patient satisfaction with medical care received, but the methods employed in this context aren't specific to this particular application. Indeed, the methods are expected to cross-generalize to all areas of customer satisfaction. To ascertain the total potential market for maximum-accuracy customer satisfaction algorithms, we recently undertook a survey of the institutions in the United States that are active in the field of assessing, monitoring, and modeling customer satisfaction (we omitted the social relationship “person-matching” business). Table 1.1 presents a summary of the findings of our survey: as seen, this is an enormous field of application and a rich opportunity for big business.

Improving Science

Researchers across scientific disciplines tend to use the same “tried and trusted” statistical methods in their work, even though much of what is reported in the literature reflects relatively weak, and often non-reproducible findings.³⁶ We believe this is attributable in large part to ubiquitous use of linear statistical models (LSMs) to analyze sample data, representing one of the greatest challenges to the validity and reproducibility of empirical findings reported in the literature.³⁷

Table 1.1: Maximum Satisfaction Analytics Marketplace

Marketplace	Estimated Number of Organizations
Health	<ul style="list-style-type: none"> • 5,754 hospitals (2012 AHA) • 15,245 nursing facilities (2004 Managed Care Digest/Institutional) • 14,000 drug rehabilitation centers (2011 US Drug Rehabilitation Center) • 25,750 outpatient care centers having 590,000 employees (2002 US Economic Census) • 14,500 home health care agencies, 1,500,000 patients (2012 CDCP) • 186,004 professionally active dentists (2009 ADA) • 29,960 health clubs, 51,400,000 members (2012 IHRSA) • 160,490 personal fitness trainers (2012 USBLS)
Education	<ul style="list-style-type: none"> • 7,000 universities and colleges, 15 million students (2012 BrainTrack) • 200 law schools (2012 ABA) • 137 medical schools (2012 AAMC) • 3,660 postsecondary vocational/technical institutions (2003 GOOGLE) • 33,366 private elementary and secondary schools, 4.7 million students (2010 NCES) • 600,000 music, art, dance, and finishing schools (2012 MANTA)
Finance	<ul style="list-style-type: none"> • 5,453 banks (2012 FDC) • 7,535 credit unions (2012 CUNA) • 914,342 financial services companies (2012 MANTA)
Insurance	<ul style="list-style-type: none"> • 452,291 insurance agencies (2012 MANTA)
Human Resources	<ul style="list-style-type: none"> • 6,164 HR consulting companies (2012 MANTA) • 38,555 HR departments (2012 MANTA) • 66,233 HR managers (2012 MANTA)
Travel	<ul style="list-style-type: none"> • 142,067 motels and hotels (2012 MANTA) • 24,818 resorts (2012 MANTA) • 12,037 marinas (2012 MANTA) • 944,390 restaurants (2012 MANTA) • 14,760 tour operators (2012 MANTA)
Professional Service	<ul style="list-style-type: none"> • 17.75 million employees in professional/business services supersector (2012 BLS) • 47,563 law firms (2010 ABA), 805,928 law companies (2012 MANTA) • 132,682 accounting firms (2012 MANTA)
Others	<ul style="list-style-type: none"> • 684,662 retail firms, 30,737 with multiple establishments (2007 NRF) • 425,500 food service/drinking firms, 14,849 with multiple establishments (2007 NRF) • 255,297 grocery stores (2012 MANTA) • 205,020 automobile repair companies (2012 MANTA) • 250,000 hair and nail salons and spas (2010 SBDC) • 116,840 child care facilities, 16,491,000 children (2012 CCCUS) • 298,000 nannies and au-pairs (2010 Google answers)

An omnipresent challenge is the ability of *data to comply* with crucial assumptions underlying LSMs. Violation of crucial assumption(s) underlying any method is problematic because such violation undermines the internal and external validity of obtained findings. Automatically extorting the virtue of “robustness” begs the question of how “incorrect” can something be, and still be considered “correct”?^{5,6} Another inherent limitation of legacy LSMs is inaccuracy: as will be demonstrated, most legacy models are only capable of accurately predicting values that lie at or close to the sample mean (GLM models) or mode (ML models).

A method was sought to eradicate both of these issues for the simplest case of a multiattribute application—a binary class variable and two or more ordered attributes (multicategorical attributes and class variables are inherently problematic for all LSM formulations). Accordingly, the exact maximum-accuracy LSM called MultiODA was created. MultiODA requires no distributional assumptions and it explicitly maximizes classification accuracy: not only does MultiODA routinely obtain more accurate and more parsimonious models than legacy LSM methods, MultiODA also identifies models in applications for

which legacy methods find nothing.^{38,39} MultiODA models with *unit-weight* beta coefficients (i.e., 0, 1, and -1) are more accurate and more parsimonious than legacy LSMSs with *continuous* beta coefficients.⁴⁰

However, the most important omnipresent challenge for all LSMSs—including explicitly optimal LSMSs—is circumventing Simpson’s Paradox, a phenomenon whereby pooling (combining) of data from different groups (e.g., genders) and/or time periods (e.g., before and after an intervention) produces spurious confounding. Paradoxical confounding is a focus of much discussion herein.

Explicitly optimal (maximum accuracy) *classification tree analysis* (CTA) was developed in order to eradicate shortcomings of LSMSs, in particular the problem of paradoxical confounding. Research using hierarchical and enumerated CTA approaches routinely obtained the strongest statistical classification models reported in many research domains. Most recently, algorithms were discovered that identify the globally optimal model for any combination of sample, data geometry, and experimental hypothesis. This book concludes by demonstrating how—vis-à-vis identification of valid, accurate, and non-confounded models—novometry eliminates challenges to the validity of empirical findings and offers promise of increasing the accuracy, efficiency, and reproducibility of programmatic scientific research.

It may be most important to note that the ODA paradigm is so new that the overwhelming majority of the possible uses and methods have not yet been developed. Nevertheless, many of the ODA methods already discovered are able to answer questions which no parametric paradigm-based method can address. As researchers begin to experiment with new data geometries and structural hypotheses, it is highly likely that many new theoretical advances are forthcoming.

Chapter 2

Fundamental Concepts

Considered from an analytical gestalt perspective, the fundamental objective of the ODA statistical paradigm is the discovery of transparently intuitive, accurate, parsimonious statistical models that explicitly maximize training classification accuracy normed against chance, and that hold-up in cross-generalizability validity analyses.

The UniODA Algorithm

The UniODA algorithm is the elemental unit of the ODA paradigm. There are two forms of a UniODA model: one for applications involving an ordered attribute, the other for applications having a categorical attribute. For applications with a *two-category* class variable, and an *ordered* attribute, the UniODA model has the following form—which includes four defining features:

If observation's *Score* on Attribute > *Threshold*, then Predict *Class* = 1.

In context this implies that the following is also true:

If observation's *Score* on Attribute ≤ *Threshold*, then Predict *Class* = 0.

The *class variable* must have at least two levels: male versus female, true versus false, success versus failure, dead versus alive, experimental versus control, or "A" versus "B". Known as a "dependent variable" in GLM and ML paradigms, the class variable is what the UniODA model tries to classify, model, discriminate, or predict.¹⁻³ More granular class variables—having three, four, or more levels, and even ordered class variables, may also be analyzed. However, when introducing or learning a new statistical paradigm it is a good idea to start with a simple design and subsequently add complexity.

Known as an "independent variable" in GLM and ML classification methods, the *attribute* also must have at least two levels. The UniODA model uses the attribute to predict the class variable. The lowest level of precision possible for an ordered attribute is a binary measurement scale: for example, in a study investigating effects of testosterone an imprecise measure of the latter is gender because male status is generally associated with a higher testosterone level than is female status.

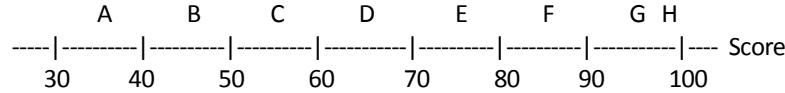
Working in tandem, the *direction* (analogous to the "sign" of a correlation coefficient) and the *optimal threshold* together define how scores on the attribute are used to predict an observation's class membership status. As a means of illustrating this synergy, imagine a salesperson desires a fast, easy, and relatively unobtrusive method to predict if a customer who walks into the dealership showroom will buy a car. Based on previous experience, the salesperson believes that people who buy cars generally say they are looking to buy. To test this *a priori* hypothesis, the next eight customers walking into the showroom were greeted by the salesperson and asked to respond to the question: "On a scale from 0% to 100%, what is the chance that you are planning to buy a new car today?" After a standard sales session with each customer, the salesperson noted if the customer bought (coded as 1) or didn't buy (0) a car. In this example the score on the 100-point scale is the attribute, and the sales outcome (0 versus 1) is the class variable. The direction and threshold features of the UniODA model will determine how best to use the attribute to predict the class variable with *maximum possible overall classification accuracy*. In order to illustrate how this is accomplished, imagine the salesperson obtained the hypothetical results presented in Table 2.1.

Table 2.1: Hypothetical Data used for Demonstrating How to Obtain a UniODA Model

<u>Customer ID</u>	Customer's Self-Rated Likelihood of Buying a Car Today		<u>Sales Outcome</u>
A	35		0
B	45		0
C	55		1
D	65		0
E	75		0
F	85		1
G	95		1
H	99		1

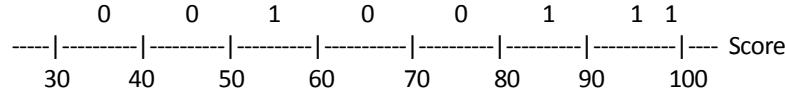
The first step in obtaining the UniODA model involves reorganizing the data along a continuum: letters identify the eight customers:

Figure 2.1: Stratifying Observations Along a Continuum



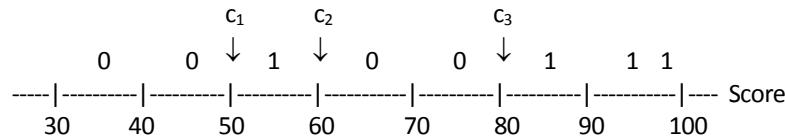
The second step involves replacing the letters, by substituting the code “1” for customers who bought a car, and the code “0” for customers who didn’t buy a car:

Figure 2.2: Representing Observations using Class Category Codes



Brute force will be used to identify the combination of optimal threshold and direction that maximizes overall classification accuracy. A *cutpoint* is a point on the continuum midway between successive observations from a different class. The three different cutpoints for the present example are indicated using arrows below: cutpoint $c_1 = (45 + 55) / 2 = 50$; cutpoint $c_2 = (55 + 65) / 2 = 60$; and cutpoint $c_3 = (75 + 85) / 2 = 80$.

Figure 2.3: Identifying All Possible Cutpoints



Direction refers to the manner in which cutpoints are used to classify observations. There are two directions: *greater than* (observations with scores $>$ cutpoint are predicted to be from class 1; observations with scores \leq cutpoint are predicted to be from class 0); and *less than* (observations with scores \leq cutpoint are predicted to be from class 1; observations with scores $>$ cutpoint are predicted to be from class 0). In the present application, because the salesperson hypothesized that customers who buy (coded as 1) would score at *higher levels* on the attribute than customers who don’t buy (coded as 0), it is only necessary to consider the greater than direction. If instead the salesperson had hypothesized that people who buy would score at lower levels on the attribute than customers who don’t buy, then it would only be necessary to examine the less than direction. If the hypothesis was non-directional then both directions would have to be searched.

Identification of the UniODA model requires evaluating classification performance achieved using all three cutpoints with the greater than direction. The model for cutpoint c_1 and the greater than direction is: if score > 50 then predict that class = 1; otherwise (i.e., if score ≤ 50) predict that class = 0. With this model two observations (D and E) actually from class 0 are misclassified as being from class 1; all four observations (C, F, G, and H) actually from class 1 are correctly classified as being from class 1; and two observations (A and B) actually from class 0 are correctly classified as being from class 0 (there are no class 1 observations with scores ≤ 50). For this sample, this model correctly classified 2 of 4 observations actually from class 0, and 4 of 4 observations from class 1. Because the objective value that is being maximized is overall classification accuracy, the model performance statistic is the *overall percentage accurate classification*: $Overall\ PAC = (\text{number of correctly classified observations} / \text{number of observations classified}) \times 100\% = 6 / 8 \times 100\% = 75\%$. The number of misclassified observations (here, two) is called the *optimal value*.³

Similarly, for c_2 and the greater than direction the model is: if score > 60 then predict class = 1; otherwise predict class = 0. This model misclassifies one observation (C) actually from class 1 as being from class 0, and misclassifies two observations (D and E) actually from class 0 as being from class 1, for an *optimal value* of three misclassifications: $Overall\ PAC = 5 / 8 \times 100\% = 62.5\%$.

Finally, for c_3 and the greater than direction the model is: if score > 80 then predict class = 1; otherwise predict class = 0. This model yields an *optimal value* of one, misclassifying one observation (C) actually from class 1 as being from class 0: $PAC = 7 / 8 \times 100\% = 87.5\%$. This combination of cutpoint and direction yields greatest *Overall PAC*, so the UniODA model is: if score > 80 then predict class = 1; otherwise predict class = 0. The confirmatory UniODA model is consistent with the salesperson's *a priori* hypothesis that people who say that they are thinking about buying a car (indicated by high scores on the attribute) are the best prospects for a car sale: indeed, all observations with scores greater than 80 in this example bought a car.

No other combination of cutpoint and the greater than direction can return a *greater value* for *Overall PAC* than is obtained by UniODA. However, it is sometimes possible for UniODA to identify more than one optimal solution—that is, more than one model that attains the identical highest-observed level of *PAC*. UniODA³ and MegaODA⁴⁻⁶ software enables one to specify primary and secondary selection heuristics for choosing a model if multiple optimal solutions occur. Available selection heuristics are summarized in Appendix A, and detailed discussion concerning the nature and use of these heuristics is available elsewhere.³

For applications involving a two-category class variable and a *categorical attribute*, the UniODA model has the following form—which includes four defining features:

If observation's *Score* on Attribute = {*Category List*}, then Predict *Class* = 1.

In context this implies that the following is also true:

If observation's *Score* on Attribute = {*Other Category List*}, then Predict *Class* = 0.

As for applications involving ordered attributes, in applications with categorical attributes the class variable and attribute must both have at least two levels. Rather than having an optimal threshold and a direction indicator like a UniODA model for an ordered attribute, for a categorical attribute the UniODA model instead has two categorical assignments: one set of attribute categories is assigned to (i.e., predicts membership in) class 1, and the second set of attribute categories is assigned to class 0. For the training sample, UniODA identifies attribute category list assignments that explicitly maximize *Overall PAC*. Again, *a priori* selection heuristics are used if multiple optimal solutions are obtained.³

Establishing Statistical Reliability

Exact one- and two-tailed UniODA distributions of *optimal values* for two-category discrimination of random ordered attributes were discovered for applications maximizing *unweighted Overall PAC*. For directional applications exact p for any *Overall PAC* can be computed for any N : computation time for the recursive, closed-form solution is linear in N .^{7,8} For non-directional (two-tailed) applications no closed-form solution for the distribution of *optimal values* has been discovered, and the enumerable open-form

solution⁹ is computationally intractable for $N > 30$. The one-tailed solution can be used to determine the two-tailed distribution if $PAC \geq 75\%$, but other methods are needed to obtain exact p for the two-tailed distribution if $PAC < 75\%$.⁷

Accordingly, ODA (UniODA, MegaODA, and CTA¹⁰) software uses Fisher's randomization procedure to compute exact p in weighted and unweighted applications.^{3,11} In applications for which the number of class variable shuffles needed to obtain a problem-specific exact permutation p is computationally intractable, Monte Carlo (MC) simulation is used to estimate p that would be obtained using Fisher's procedure. To perform this simulation, UniODA is conducted to obtain the actual *optimal value*. Next, simulation involving a user-specified number of MC experiments is conducted: in each experiment the class category memberships of observations are randomly shuffled, UniODA is conducted, and the MC *optimal value* is stored. The estimated exact p is the proportion of experiments in which MC *optimal value* \geq actual *optimal value*.

MC p is an estimate of the exact permutation p , and the accuracy of this estimate increases as the number of MC experiments conducted increases. Although one can't determine the accuracy of a given MC p for a specific number of MC experiments, one may determine the likelihood that an obtained MC p is less than a "target" p value. This likelihood is expressed in terms of confidence levels. For example, how confident is one that exact p for a specific application is less than the target p (e.g., 0.05), if MC $p \leq 0.04$ for 100 experiments? Clearly this confidence is less than would be the case if MC $p \leq 0.04$ for 1,000,000 experiments. The confidence that MC $p <$ target p increases with an increasing number of MC experiments. The method used to compute confidence levels by ODA software involves integrating the beta function by the method of partial fractions.^{12,13} Confidence levels reported by ODA software are rounded down to the nearest 0.01%, except where 100% confidence is reported, in which case the computed value exceeds $1 - e^{-28}$. We recommend that a confidence level of 99.9% or higher be used in actual applications.³

We conducted three MC simulation studies to investigate precision and convergence properties of MC methodology when employed to estimate known exact p in non-directional UniODA involving two balanced classes.¹⁴ The first study investigated the accuracy of simulated p -values. One million MC experiments were run for each balanced design of $N \leq 30$. A design is balanced if the number of class 1 and class 0 observations is identical for even N , or differs by one for odd N . In every MC experiment the attribute was a uniform random number between 0 and 1. For even N experiments the first $N / 2$ observations were assigned to class 1, and the rest were assigned to class 0. For odd N experiments the first $(N - 1) / 2$ observations were assigned to class 1, and the rest to class 0. For each experiment the *optimal value* (the number of misclassifications) was determined and stored. For each N the estimated UniODA distribution was cumulated after 10^6 experiments were completed. To compare estimated and known distributions, cumulative $p > 0.001$ were rounded up to the nearest thousandth and cumulative $p < 0.001$ were rounded up on the second significant digit. The results demonstrated that MC experiments accurately estimated known exact UniODA statistical distributions. Over all N and possible *optimal values*, 170 of 238 (71.4%) estimated cumulative probabilities were identical to the exact value; 237 of 238 (99.6%) of the estimates were within ± 0.001 of the exact value; and all estimates were within ± 0.002 of the exact probability. The estimated cumulative probabilities were most accurate when the exact probability was small. For *optimal values* with associated exact cumulative probabilities of $0.05 \leq p < 0.001$, 45 of 50 (90%) of the estimated probabilities were identical to the corresponding exact probability; 49 of 50 (98%) estimated probabilities were within ± 0.001 of the exact probability; and all of the estimated probabilities were within ± 0.002 of the exact probability. MC experiments also yielded accurate estimates of exact cumulative probabilities for statistically marginal ($0.05 \leq p < 0.10$) effects: 13 of 15 (86.7%) of estimated cumulative probabilities were identical to their corresponding exact values, and all estimated probabilities were within ± 0.001 of the exact probability.

Cumulating 10^6 MC experiments for a given N returns an accurate approximation of a UniODA distribution, however the computational cost is high. Accordingly, the second study studied convergence properties of MC methodology, and sought to determine the number of MC experiments that is sufficient to obtain stable, accurate estimates of known exact UniODA *optimal value* distributions. MC experiments were designed and data generated as in Study 1. For each N between 3 and 30 inclusive, 10^5 experiments were run in successive blocks of 1,000 experiments, and the UniODA distribution was cumulated at each block. Thus, 100 UniODA distributions were estimated for each N : the first based on 1,000 experiments, the second based on 2,000 experiments, and the 100th based on 10^5 experiments. About half (56.9%) of

the estimated p 's converged to their final value (at the end of the study) within 20,000 experiments, and most (86.3 percent) converged to their final value within 70,000 experiments. After 10^5 experiments were completed, every estimated p in the range $0.001 \leq p < 0.10$ was identical to the corresponding estimated p based on 10^6 experiments (Study 1). Consistent with the first study, known UniODA distributions were accurately modeled using MC methodology. For probabilities in the range $0.001 \leq p < 0.05$: 35 of 50 (70%) estimated cumulative probabilities were identical to corresponding exact values; 49 of 50 (98%) estimated probabilities were ± 0.001 of exact; and all estimated probabilities were ± 0.002 of the exact value. Thus, the UniODA cumulative probabilities estimated using 100,000 MC experiments are only modestly less accurate than corresponding probabilities estimated using one million MC experiments: the maximum observed estimation error (± 0.002) is small and rarely occurs.

The final study investigated convergence properties for balanced two-category non-directional UniODA for increasing N . MC experiments were designed and data generated as in Study 1. For all N between 1,000 and 8,000 inclusive, in steps of 1,000, a total of 10^5 MC experiments were run. Tabled for the indicated value of p and N are the *optimal value* and the corresponding *Overall PAC* (Table 2.2, top and bottom row, respectively). The *optimal value* is the maximum number of misclassifications possible to still achieve the p value. For example, for $N = 1,000$ observations and $p < 0.001$ a maximum of 438 misclassifications can be made, corresponding to 562 correct classifications, and thus to $Overall\ PAC = (562 / 1,000) \times 100\% = 56.2\%$. And, for $p < 0.05$ a maximum of 457 misclassifications are possible, corresponding to $Overall\ PAC = (543 / 1,000) \times 100\% = 53.9\%$. For $N = 5,000$ and $p < 0.01$, a maximum of 2,384 misclassifications are possible, corresponding to $Overall\ PAC = [(5,000 - 2,384) / 5,000] \times 100\% = 52.3\%$. In balanced designs with as few as 1,000 observations, a UniODA model performing marginally better than an unbiased coin flip yields classification accuracy sufficient to achieve $p < 0.001$. Therefore, as N increases in magnitude, the value of p as an index of model performance rapidly diminishes to trivial levels.

Table 2.2: Maximum *Optimal Value* for 2-Tail p in Balanced 2-Category UniODA

N	Two-Tail $p <$			
	0.001	0.01	0.05	0.10
1,000	438 56.2	448 55.2	457 54.3	461 53.9
2,000	912 54.4	927 53.6	939 53.1	945 52.8
3,000	1393 53.6	1411 53.0	1425 52.5	1433 52.2
4,000	1876 53.1	1896 52.6	1913 52.2	1922 52.0
5,000	2361 52.8	2384 52.3	2403 51.9	2413 51.7
6,000	2849 52.5	2874 52.1	2894 51.8	2905 51.6
7,000	3336 52.3	3364 51.9	3386 51.6	3397 51.5
8,000	3825 52.2	3853 51.8	3878 51.5	3890 51.4

Our research experience investigating analytic/MC computation/estimation of weighted/unweighted, (non)directional, (im)balanced UniODA *optimal value* distributions, for random attributes measured using a variety of metrics, yielded three insights.

First, a *precision dimension* may be used to exactly describe the metric underlying any attribute. This dimension is bounded at the polar extremes by binary data (least precise) and “continuous” data (most precise). As exact distribution theory can be derived for the poles of the precision dimension, so too can exact distribution theory be derived for any specific attribute measurement metric. For example, if an attribute is measured using a 7-point Likert-type scale, then exact distribution theory may be derived that assumes a 7-point Likert scale. If an attribute is measured using a categorical scale having four categories, exact distribution theory may be derived that assumes a four-category categorical scale. Because it is possible to derive distribution theory that assumes that the specific measure metric actually used in a given application was in fact used, distribution theory for UniODA can be based strictly on structural and configurational features of a problem, and distribution theory will always be exact for a specific application.^{15,16}

Second, UniODA clearly represents a powerful alternative to some of the most popular legacy statistical methods. For example, Student’s *t*-test is commonly used to analyze data consisting of a binary class variable and a “continuous” attribute, and chi-square analysis is commonly used to analyze data consisting of a binary class variable and a binary attribute. Of course, UniODA can also be used, and exact statistical distributions derived for, designs located anywhere on the precision dimension—at either pole or anywhere between.

Third, only when the class variable and the attribute both reach their theoretical minimum level of measurement precision (i.e., in a *binary* application) do the exact statistical distributions for one- and two-tailed UniODA and Fisher’s exact test converge.

Criterion for Statistical Significance

In addition to determining how to obtain exact *p* for a test of a statistical hypothesis, it is crucial to define the level of *p* that is used to establish “statistical significance” for a given application. What heuristic decision-making algorithm should be used to decide whether to accept or reject the null hypothesis that the class categories can’t be discriminated on the basis of the attribute, given the observed *p*-value? What “target value” of *p* should be used in a given application as the criterion for determining whether or not the level of classification accuracy achieved by an optimal model exceeded the level of classification accuracy expected by chance? An excellent discussion of this issue, presented in the context of the ODA paradigm and including a detailed expository demonstration, is available elsewhere.³ A “necessary and sufficient” summary of this issue is presented here.

The *generalized criterion* is appropriate for a study in which exactly one test of a statistical hypothesis is conducted. Although ubiquitous in the literature, conventional definitions of *generalized* ($p < 0.05$) and *marginal* ($0.05 < p < 0.10$) statistical significance are arbitrary, and different criteria may be appropriate for different applications.¹⁷ For example, a relatively conservative criterion (e.g., $p < 0.005$) may be appropriate in applications with extreme statistical power (e.g., large samples), or when committing a Type I error (i.e., falsely concluding an effect occurred) is costly or otherwise undesirable. A more liberal criterion (e.g., $p < 0.10$) may be appropriate in applications with relatively low statistical power (e.g., small samples), or when committing a Type II error (i.e., falsely concluding no effect occurred) is costly or otherwise undesirable.

Ryan¹⁸ distinguished between the *error rate per comparison* (the probability that any comparisons—that is, any of the tests of statistical hypotheses—conducted within a single study will be incorrectly considered to be statistically significant) and the *error rate per experiment* (the expected number of such errors per study). The problem with using the generalized criterion to evaluate multiple statistical tests conducted within a single study is that the actual resulting criterion for the study (the *experimentwise p*) is higher (more liberal) than the generalized (*per-comparison*) criterion: an effect known as *alpha inflation*. It is therefore necessary to adjust the criterion for establishing experimentwise statistical significance on the basis of the total number of tests of statistical hypotheses that are conducted in the study.

There are two multiple comparison strategies. When conducting *all possible comparisons* all of the different statistical hypotheses that it is possible to test for a given application are tested, often in the context of exploratory research. When conducting *planned comparisons*, a theoretically circumscribed subset of all possible comparisons are evaluated, often in the context of confirmatory research. In the ODA paradigm a *sequentially-rejective Sidak Bonferroni-type multiple-comparisons procedure* is used to control the experimentwise *p* when

multiple comparisons are performed (see Chapter 10).³ Sequentially-rejective Bonferroni-type procedures ensure the desired experimentwise p , and provide greater statistical power than non-sequential alternatives.¹⁹ As is true of Dunn's procedure, the use of Sidak's per-comparison criterion ensures the desired experimentwise p : however, if the number of contrasts is greater than one, then the per-comparison p by Sidak's procedure will be greater (more liberal) than the corresponding per-comparison p by Dunn's procedure.^{20,21}

Finally, until this point multiple comparisons were *unit weighted*—the theoretical importance of each comparison was treated as being equivalent. However, it is sometimes the case that only a subset of all the tests of statistical hypotheses evaluated within a given study are of greatest importance to the central hypothesis. In such circumstances one may partition experimentwise p in accordance with the perceived importance of the tests in the context of the study: a procedure known as an ensemble-adjusted^{22,23} or ordered^{17,24} Bonferroni-type procedure, or as *alpha splitting*²¹. Of course, this is strictly an *a priori* methodology.^{17,21}

Assessing Classification Accuracy

How best to summarize classification performance achieved by a statistical model for a training sample has been widely discussed.²⁵⁻²⁸ Assessment of model performance based only on statistical significance considerations is inappropriate for evaluating classification and forecasting models.²⁹⁻³³ Instead, attention should focus on the ability of the model to achieve clinically^{34,35} or other context-specific ecologically significant³⁶⁻³⁹ levels of predictive performance.

Called the “simple matching coefficient” in numerical taxonomy⁴⁰, *Overall PAC* is not the only nor is it necessarily the best way to conceptualize model performance.^{15,41,42} For example, imagine a study evaluating a rapid method of screening people for an imminent fatal illness, using a sample with 999,990 healthy and 10 fatally ill observations. If a model were to predict that all of the observations were healthy, the model would yield $999,990/1,000,000 \times 100\% = 99.999\%$ *Overall PAC*. Thus, even though this model yields nearly perfect *Overall PAC*, every single fatally ill observation is misclassified.

In applied research, statistical analysis involving markedly skewed data isn't a rare phenomenon: this is the indubitable scenario facing investigators of rare diseases, adverse drug reactions, bank failures, and emerging technologies, for example. Generally speaking, applied researchers are typically primarily interested in predicting outlying cases: the most likely to die, to be arrested, to graduate, to evolve into a professional athlete, to return the most profit on a loan, to be dissatisfied, and so forth. Skewed data are the norm, not the exception in research involving rare or outlying phenomena. Because it ignores the classification performance achieved for different classes, *Overall PAC* is clearly inappropriate for use as an omnibus measure of model classification accuracy in such applications.

Two widely-reported molecular (not omnibus) indices of classification accuracy are the models *sensitivity* (probability that an observation from class category c will be classified into c) and *predictive value* (probability that an observation classified into c is a member of c). Intuitively these measures assess the ability of a theory (as reflected by the model) to *explain*⁴³ and to *predict*⁴⁴ the observed phenomena, respectively. Sensitivity assesses the ability of the model to discriminate observations from the different class categories: it is an index of the *descriptive* utility of the model, reflecting its proficiency in correctly characterizing different class categories. Predictive value assesses the ability of the model to make correct classifications (point predictions) of observations into class categories: it is an index of the *prognostic* utility of a model, reflecting its proficiency in correctly predicting the class membership status of individual observations.⁴⁵

To facilitate clarity of exposition, please imagine that the classification results in Table 2.3 were obtained by a model: as seen in the *confusion table*, the actual class status of an observation is given in the rows, and the class status of an observation as predicted by the model is given in the columns. As seen, the sensitivity of the model for class category 0 = $(20 / 25) \times 100\% = 80\%$, and the sensitivity of the model for class 1 = $(15 / 25) \times 100\% = 60\%$. The model correctly classified 80% of the actual class 0 observations, and 60% of the actual class 1 observations.

In this application the mean sensitivity across classes—the *Mean PAC*—is $(80\% + 60\%) / 2 = 70\%$. The use of weighting by prior odds (i.e., weight all n_c observations in class category c by the value $1 / n_c$) explicitly maximizes the *Mean PAC* achieved by an ODA model, and is analogous to the use of *antecedent*

probability or *base rate* in Fisher's discriminant analysis.⁴⁵⁻⁴⁹ This strategy ensures that no class category dominates another simply due to differential numbers of observations in the categories.

Table 2.3: Hypothetical Confusion Table for Discussion of Classification Performance Measures

<u>Actual Class Membership</u>	<u>Predicted Class Membership</u>		Row Marginal
	Class 0	Class 1	
Class 0	20	5	25
Class 1	10	15	25
Column Marginal	30	20	

Mean PAC can be used to create an omnibus measure of model classification accuracy that, in contrast to overall *PAC*, is sensitive to the classification accuracy that is achieved for the different class categories. For an application with $C \geq 2$ class categories, a *Mean PAC* of $(1 / C) \times 100\%$ is expected by chance under the null hypothesis that the attribute is uniform random (i.e., all data are equally likely).^{8,50} This enables the computation of a standardized index of effect strength normed against chance, known as the *effect strength for sensitivity* or *ESS*—that may be used to directly contrast different ODA models across *structural* (number of class categories, attribute metrics) and *configural* (relative class category sample size imbalance, total sample size, hypothesis directionality) differences. First, compute $C^* = 100 / C$. Next, compute *ESS* using the following interactive transformation of *Mean PAC*:

$$ESS = (Mean\ PAC - C^*) / (100 - C^*) \times 100\%.$$

On this normed effect strength index, 0 represents the theoretical lower bound (i.e., the level of classification accuracy that is expected by chance), and 100 represents the theoretical upper bound (i.e., perfect, errorless classification). For the hypothetical example $ESS = (70 - 50) / (100 - 50) \times 100\%$, or 40%. Thus, the model achieves 40% of the theoretical possible improvement in classification accuracy that it is possible to attain above and beyond (i.e., after removing) the effect of chance. Readers are encouraged to verify that the three UniODA models represented by cutpoints c_1 , c_2 , and c_3 in Figure 2.3 have corresponding *ESS* values of 50, 30, and 75, respectively.

In Table 2.3, the predictive value of the model for class category 0 = $(20 / 30) \times 100\% = 66.7\%$, and the predictive value of the model for class 1 = $(15 / 20) \times 100\% = 75\%$. Thus, the model was correct 66.7% of the time it predicted an observation was from class category 0, and 75% of the time it predicted an observation was from class category 1. The mean predictive value across classes—the *Mean PV*—is $(66.7\% + 75\%) / 2 = 70.85\%$. As done for *Mean PAC*, *Mean PV* can be used to create an omnibus measure of model predictive value that is sensitive to the predictive value for the different class categories:

$$ESP = (Mean\ PV - C^*) / (100 - C^*) \times 100\%.$$

On this normed effect strength index, 0 represents the theoretical lower bound (i.e., the predictive value that is expected by chance), and 100 represents the theoretical upper bound (i.e., perfect, errorless point predictions). For the hypothetical example $ESP = (70.85 - 50) / (100 - 50) \times 100\%$, or 41.7%. Thus, the model yields point predictions that reflect 41.7% of the theoretical possible improvement in predictive value that it is possible to attain above and beyond (i.e., after removing) the effect of chance. Readers are encouraged to verify that the three UniODA models represented by cutpoints c_1 , c_2 , and c_3 in Figure 2.3 have corresponding *ESP* values of 67, 27, and 80, respectively.

In contrast to sensitivity, predictive value (*PV*) is influenced by the base rate of the class category c in the population, and by the false-positive (*FP*) rate (i.e., the probability that the model will classify an observation into class category c when the observation is *not* actually a member of c).⁵¹ Thus, the utility of a model should be assessed for different base rates.⁵² An *efficient* model has a *PV* that is greater than the base rate for c in the population of interest.⁵³ For example, if it is known that 53% of the population are in

c , then a model should yield a PV for c that is greater than 53%. Otherwise, the decision-maker is better off not using the model, and instead guessing that every observation will have a 0.53 probability of being a member of c . For any given base rate (b), model efficiency for class category c is defined as:

$$[(PV \text{ for class } c) \times (b)] / [(PV \text{ for class } c) \times (b) + (FP \text{ for class } c) \times (1 - b)],$$

where $FP = 1 - PV$ for class c . To demonstrate an *efficiency analysis* we compute model efficiency for the hypothetical example, for classifications into class category 1 (PV for class category 1 = 0.75; $FP = 1 - PV = 0.25$) and b ranging between 0.1 and 1 in increments of 0.1 (by definition, when $b = 0$, model efficiency = 0). As seen in Table 2.4, for the hypothetical example the model efficiency exceeds all base rates except zero and one: therefore, the model clearly provides superior PV versus chance for classifications into class category 1, especially for base rates between 0.3 and 0.5. Readers are encouraged to conduct efficiency analysis for classifications made into class category 0.

Table 2.4: Efficiency Analysis for Hypothetical Example

b	$1 - b$	$PV \times b$	$FP \times b$	Model Efficiency for $c = 1$	Efficiency - b
.1	.9	.075	.225	.075 / (.075 + .225) = .25	.15
.2	.8	.15	.2	.15 / (.15 + .2) = .4286	.2286
.3	.7	.225	.175	.225 / (.225 + .175) = .5625	.2625
.4	.6	.3	.15	.3 / (.3 + .15) = .6667	.2667
.5	.5	.375	.125	.375 / (.375 + .125) = .75	.25
.6	.4	.45	.1	.45 / (.45 + .1) = .8182	.2182
.7	.3	.525	.075	.525 / (.525 + .075) = .875	.175
.8	.2	.6	.05	.6 / (.6 + .05) = .9231	.1231
.9	.1	.675	.025	.675 / (.675 + .025) = .9643	.0643
1	0	.75	0	.75 / (.75 + 0) = 1	0

Monte Carlo-based research we conducted investigating distributions of ESS and ESP statistics obtained by UniODA used to analyze random data suggested the following heuristic “rule-of-thumb” for characterizing the relative strength of an effect.³ For both of these statistics: a *relatively weak* effect is (ESS or ESP) $value < 25\%$; a *moderate* effect is $25\% \leq value < 50\%$; a *relatively strong* effect is $50\% \leq value < 75\%$; a *strong* effect is $75\% \leq value < 90\%$; and a *very strong* effect is $value \geq 90\%$.

The CTA Algorithm

Methodology underlying the development of hierarchically, enumerated, and globally optimal (maximum-accuracy) classification tree analysis (CTA) models is thoroughly discussed ahead. Expressed in a nutshell, recursively-derived CTA models chain series of UniODA models across monotonically diminishing sample strata in order to identify the configuration that explicitly maximizes training ESS . Here we introduce the motivation for CTA models, their geometric representation, and their functionality in addressing statistical hypotheses.¹⁰

Multiple issues associated with the use of a linear model (LM) motivate the use of a non-linear model (NLM). A pragmatic issue is sample conservation and related protection of statistical power: an observation that is missing data on any of the independent variables used in a LM is dropped from the analysis sample, whereas an observation is retained by a NLM if data are available for the attributes that are actually used to classify the observation. Several theoretical difficulties associated with the use of LMs remain unanswered. For example, although parametric LMs require data to meet underlying assumptions to ensure the validity of findings, there is no statistical test for assessing whether the data are multivariate normally distributed, and residual normality is impossible to test if the sample is small due to insufficient statistical power, and if the sample is adequate then departures from normality must occur as a result of

regression toward the mean. A synergistic problem is that LMs are less parsimonious than NLMs, and LMs yield lower levels of classification (predictive) accuracy than NLMs. LMs are cumbersome to use in tests of moderation, mediation, and confounding, whereas NLMs are inherently engineered to identify moderating, mediating, and confounding effects. And, of course, there is the rationale that there is nothing to lose but everything to gain in using a NLM in statistical analysis: (a) if in reality the true effect is linear, then both the LM and the NLM will identify the true effect; however (b) if in reality the true effect is non-linear, then only the NLM model will identify the true effect (in this circumstance the LM may still discover a statistically reliable effect, but the effect will be spurious and reflect an example of paradoxical confounding). Then there is the observation that few phenomena in nature seem to reflect fundamentally linear processes: if even the ballistic path of a bullet requires 6 DOF modeling to obtain an accurate NLM, what hope is there that poorly conceived, poorly measured, competitive interactive social systems involving feedback reflect an underlying linear phenomenon?

We have long argued that a primary theoretical motivation to avoid using LMs is their inherent nature, reflected by the expression: "one size fits all." This is in contrast to the nature of NLMs, reflected by the expression: "different strokes for different folks." These different perspectives are attributable to three assumptions needed to justify the use of a LM, regardless of its derivation. First, LMs assume that the attributes in the model are important for every observation in the sample. In contrast, with NLMs different attribute sets can be used with different partitions of the sample: one set of attributes is used for classifying one partition of the sample; another set of attributes for classifying a different partition; and so forth. Second, LMs assume the model attributes have identical direction of influence (positively or negatively predictive) for every observation. In contrast, with NLMs an attribute may predict class category 1 for one partition of the sample, versus category 0 for a different partition. Third, LMs assume that the attributes in the model have the same coefficient value (decision weight) for all sample observations. In contrast, in NLMs the coefficient (i.e., optimal coefficient) for an attribute may assume two different values for two different sample partitions: for example, 0.2 and -1.8, respectively.

An example of a CTA model is provided in Figure 2.4 to promote expositional clarity. The first CTA model published was an exploratory study discriminating geriatric (at least 65 years of age) versus non-geriatric adult ambulatory medical patients on the basis of self-reported well-being.⁵⁴ Forty geriatric and 85 non-geriatric patients completed a survey assessing five functional status dimensions [Basic Activities; Intermediate Activities; Mental Health (absence of depression); Social Activity; and Quality of Social Interaction], and five single-item measures of health satisfaction, physical limitations, and quantity of social interaction.

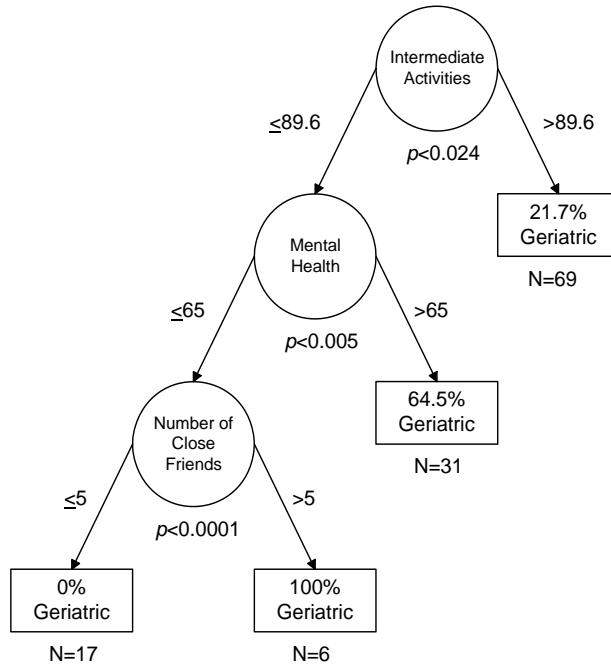
At first glance a CTA model may appear similar to results obtained by decision analysis (DA), because both methods depict findings using tree-like representations. As seen, CTA models initiate with a *root node*, from which two or more *branches* emanate and lead to other *nodes*: branches indicate pathways through the tree, and all branches ultimately terminate in model *endpoints* representing unique sample strata. The CTA algorithm determines the attribute subset that predicts the outcome (class variable) with maximum *ESS*. While CTA identifies the specific decision-making strategy that yields maximum accuracy in predicting *a specific outcome*, DA estimates the valence and likelihood associated with *all possible decision-making strategies and outcomes*.

In the schematic illustration of the CTA model, circles represent nodes (attributes), arrows indicate branches, and rectangles represent model endpoints—unique patient strata. Numbers (or words, when attributes are categorical) adjacent to branches indicate the value of the *optimal threshold* (or the optimal *category assignment*) for the node. Numbers beneath nodes are generalized (per-comparison) *p*'s (all generalized *p*'s in the model must meet the criterion for experimentwise $p < 0.05$). The number of observations classified into each endpoint (strata *N*) is given, and the percentage of class category 1 (here, geriatric) observations is indicated inside the rectangle representing the endpoint.

Using CTA models to classify individual observations is straightforward. Consider a hypothetical person having an Intermediate Activities score = 85, a Mental Health score = 64, and 7 close friends. Starting with the first node, since the person's Intermediate Activities score is ≤ 89.6 , the left branch is appropriate. At the second node the left branch is again appropriate because the person's Mental Health score is ≤ 65 . Finally, at the third node the right branch is appropriate since the person has more than 5 close friends. The person is classified into the corresponding model endpoint: as seen, all six observations classified into this model endpoint were geriatric. In this sample the probability of being geriatric in this endpoint is $p_{geriatric} = 1$, and for prognostic purposes the

probability of being geriatric in this endpoint is $p_{geriatric} \geq 6/7$. In this example, had the patient instead reported 5 or fewer close friends, then the left-hand endpoint would be appropriate, with $p_{geriatric} = 0$ (and $p_{geriatric} \leq 1/18$).

Figure 2.4: CTA Model Discriminating Geriatric vs. Non-Geriatric Ambulatory Medical Patients



Some intuitive aspects of CTA models are immediately obvious. For example, the CTA model “coefficients” are optimal thresholds or category descriptions expressed in their natural measurement units. Using a CTA model sample stratification unfolds in a “flow” process which is easily visualized across attributes in the schematic illustration of the model. The manner in which CTA handles observations with missing data is also intuitive: while LMs drop observations missing data on any attributes included in the model, CTA only drops observations that are missing data on attributes used in their classification. In the present example, imagine an observation having an Intermediate Activities score > 89.6 , but missing data on number of close friends and/or on Mental Health. Using a LM the observation would be dropped, but using CTA the observation would be classified.

Staging tables are an alternative representation of CTA models that are useful for assigning “severity” or “propensity” scores (weights) to observations (Table 2.5). The rows of the staging table are the CTA model endpoints reorganized in increasing order of percent of class 1 (geriatric) membership. Stage is an *ordinal index* of (here, geriatric) propensity and $p_{geriatric}$ is a more precise ordinal index: increasing values on either index indicates increasing propensity. Compared to Stage 1 (with p_{Odds} set at $\leq 1/18$, or 0.056), $p_{geriatric}$ is approximately 4-times higher in Stage 2, 12-times higher in Stage 3, and 15-times higher in Stage 4 (with p_{Odds} set at $\geq 6/7$, or 0.857).

To use the staging table to obtain (geriatric) propensity for a given observation, simply evaluate the fit between the observation’s data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor that is *exactly true* for the observation undergoing staging. Consider the hypothetical person discussed earlier. Stage 1 does not fit because the person has more than five close friends. Stage 2 does not fit because the person’s Intermediate Activities score is ≤ 89.6 . Stage 3 does not fit because the person’s Mental Health score is ≤ 65 . Through the process of elimination, Stage 4 must be appropriate: because the person has an Intermediate Activities score ≤ 89.6 , Mental Health score ≤ 65 , and > 5 close friends, Stage 4 clearly fits the data of this hypothetical person.

Table 2.5: Staging Table for Predicting Geriatric Status

Stage	Intermediate Activities	Mental Health	Close Friends	N	$p_{geriatric}$	Odds	p_{Odds}
1	≤ 89.6	≤ 65	≤ 5	17	0	$\leq 1:17$	≤ 0.056
2	> 89.6	-----	-----	69	0.217	2:7	0.222
3	≤ 89.6	> 65	-----	31	0.645	7:4	0.636
4	≤ 89.6	≤ 65	> 5	6	1	$\geq 6:1$	≤ 0.857

Note: Increasing scores on Intermediate Activities indicate increasing adaptability, and on Mental Health indicate decreasing depression.

Table 2.6 presents the confusion table for the CTA model. The CTA model generally achieved relatively strong classification accuracy. For the CTA model: $Overall\ PAC = [(71 + 26) / (71 + 11 + 15 + 26)] \times 100\% = 78.9\%$; model sensitivity for Geriatric = $(26 / 41) \times 100\% = 63.4\%$; model sensitivity for Non-Geriatric = $(71 / 82) \times 100\% = 86.6\%$; Mean PAC = $(63.4\% + 86.6\%) / 2 = 75.0\%$; ESS = $(75 - 50) / (100 - 50) \times 100\% = 50.0\%$ (a relatively strong effect); model predictive value for Geriatric = $(26 / 37) \times 100\% = 70.3\%$; model predictive value for Non-Geriatric = $(71 / 86) \times 100\% = 82.6\%$; Mean PV = $(70.3\% + 82.6\%) / 2 = 76.45\%$; ESP = $(76.45 - 50) / (100 - 50) \times 100\% = 52.9\%$ (a relatively strong effect).

Table 2.6: Confusion Table for the Geriatric CTA Model

<u>Actual Class</u>	<u>Predicted Class</u>		Marginals
	Non-Geriatric	Geriatric	
Non-Geriatric	71	11	82
Geriatric	15	26	41
Marginals	86	37	123

In this example a logistic regression model discriminating (non)geriatric patients identified two patient strata: (a) relatively active, depressed non-geriatric people; and (b) relatively inactive, non-depressed geriatric people.⁵⁴ In contrast, the CTA model identified four patient strata: (a) patients scoring > 89.6 on Intermediate Activities were primarily (78.3%) relatively active non-geriatric adults; (b) patients scoring ≤ 89.6 on Intermediate Activities, and at high levels (> 65) on Mental Health, were largely (64.5%) relatively inactive, non-depressed geriatric adults; (c) all patients scoring ≤ 89.6 on Intermediate Activities and ≤ 65 on Mental health, with fewer than six close friends, were inactive, depressed, socially isolated non-geriatric adults, primarily young depressed women; and (d) all patients scoring ≤ 89.6 on Intermediate Activities and ≤ 65 on Mental health, with more than five close friends, were inactive, depressed, socially-connected geriatric adults. As seen in Figure 2.5, such findings are intuitively communicated using a pie-chart.

It is also informative to evaluate the attributes loading in the CTA model in terms of their importance in the prediction-making process. Conceptually analogous to the R^2 statistic in regression analysis—that indicates the percentage of the variance in the class (independent) variable explained by attributes (dependent measures) in the model, an *Attribute Importance in Discrimination* (AID) analysis indicates the percentage of the sample for which class membership classifications were influenced by an observation's score on the attributes (Table 2.7).

Figure 2.5: Pie-Chart Illustrating Distribution of Total Sample in Four CTA-Identified Strata

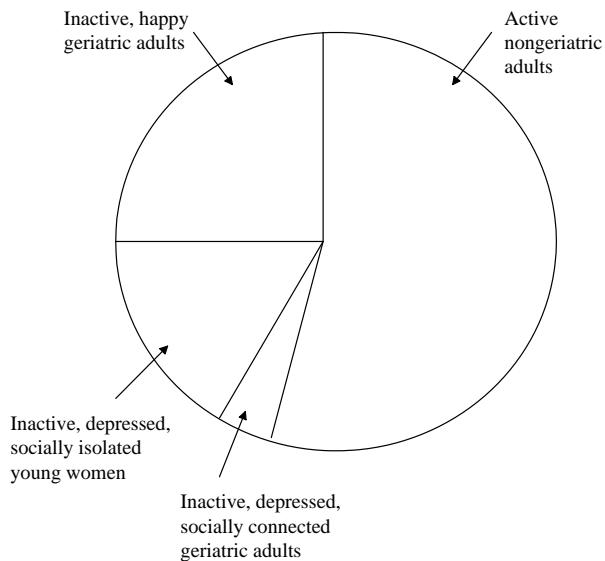


Table 2.7: AID Analysis for CTA Example

Attribute	Percent of Sample Evaluated in Part on the Basis of the Attribute	
<hr/>		
Intermediate Activities	123/123	100.0%
Mental Health	54/123	43.9%
Number of Close Friends	23/123	18.7%

Only the root attribute is involved in the classification decisions for all observations in the sample. Easily seen in Figure 2.4, Mental Health was involved in classification decisions for all of the observations except for those classified on the right-hand side of the root attribute: $123 - 69 = 54$ observations. Mental Health thus influenced classification decisions for $100\% \times 54 / 123 = 43.9\%$ of the total sample. Also easily seen in Figure 2.4, Number of Close Friends influenced classification decisions for $100\% \times 23 / 123 = 18.7\%$ of the total sample.

Data Transformations

Observation is sometimes insufficient to reveal a phenomenon. For example, the observation may be imprecise or unreliable when considered from a measurement perspective. Such limitations are widely recognized and have amassed a large literature. Less widely discussed is that even if a raw observation has excellent psychometric characteristics, a transformation may be required to bring the phenomenon of interest clearly into theoretical focus (this *doesn't* refer to application of transformations to attributes in an attempt to satisfy distributional assumptions underlying statistical methods and thereby obtain a valid Type I error rate: ODA makes no distributional assumptions and provides exact p -values).

Weighting Observations

Discussed earlier in this Chapter, unless there is an *a priori* reason to maximize *Overall PAC* (which isn't normed against chance), our laboratory weights observations by prior odds so the ODA model explicitly maximizes *Mean PAC* (which is normed against chance using the *ESS* statistic). We haven't yet extensively studied applications maximizing *Overall PAC*, but we are intrigued by the potential applicability of this strategy in minimizing the digital information footprint in applications requiring miniaturization, for example micro- and mini-transmissions with satellites and roving devices exploring deep space and celestial objects. The issue isn't identifying a statistically reliable model, but rather using the smallest number of rules to achieve perfect *Overall PAC*. Depending on the application, for some media finding a solution that achieves "close" to perfect *Overall PAC* may suffice.

Some areas of research naturally weight individual observations, such as finance or weather forecasting. Consider an application predicting daily price movement of a stock. An ODA model identified without weighting would maximize ability to correctly predict daily direction of movement in stock price. Even though the model might be correct (for example) 85% of the time in predicting the direction of daily movement in stock price, trading using this strategy may lose money if the model misclassified the days on which the price of the stock changed the most. However, specification of a weight indicating the dollar value of the change in stock price would identify an ODA model that maximized the amount of dollars correctly predicted: although the percentage of correct predictions might decrease, the model would nevertheless maximize trading profit.³ Similarly, an unweighted model of daily precipitation would maximize accuracy in predicting if measurable precipitation will occur, whereas a model weighting days by amount of precipitation would maximize accuracy in predicting the amount of precipitation expected to occur.⁵⁵

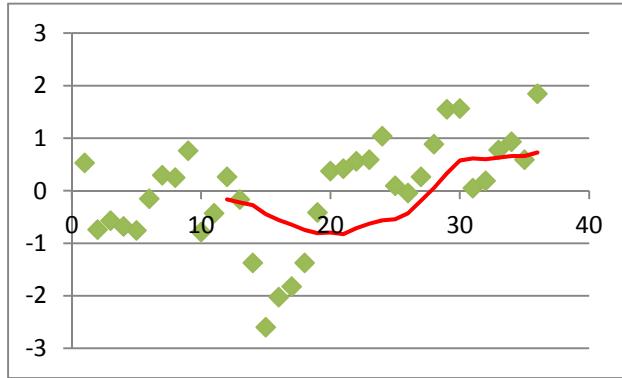
Due to the potential importance of weights in accurately reflecting phenomena being modeled, all ODA analyses can be weighted. When the phenomenon being modeled is properly framed by using weights, then the use of weights in experimental design and statistical analysis is crucial, and failure to use weights may limit both the empirically attained and theoretically attainable *ESS*, as well as reduce the cross-generalizability (reproducibility) of the model when applied to independent random samples.

Pre-Processing Data

ODA has been employed in a myriad of disciplines, including in a multitude of medical, psychological, and educational applications involving serial recordings of an outcome measure. In this context, pre-processing bridges the difference between the sample-focused methods of administrators and researchers, and the individual patient-centered focus of providers.⁵⁶ This approach to evaluating treatment effectiveness first determines whether each observation has a reliably greater posttest score, reliably greater baseline score, or statistically comparable baseline and posttest scores. Once the outcome for every observation has been determined, then groups of observations experiencing the same outcome are constructed, and UniODA-based comparisons of outcomes between, for example, control versus intervention groups can be made in terms of the relative proportion of observations showing reliable improvement. In CTA-based applications comparisons between the different patient strata in model endpoints can be made in terms of the relative proportion of observations showing reliable improvement. Of course, the pre-processing of serial data isn't limited to the study of clinical interventions. The following example demonstrates the use of an exact (UniODA) and a parametric (*N*-of-1 classical test theory-based *z* test) in pre-processing a series.⁵⁷

Published monthly by the Survey Research Center of the University of Michigan, the Index of Consumer Sentiment (ICS) is widely followed, and one of its factors (the Index of Consumer Expectations) is used in the Leading Indicator Composite Index published by the US Department of Commerce, Bureau of Economic Analysis.⁵⁸ This example investigates the trajectory of the ICS over three recent years, evaluating the statistical significance of month-over-month and year-over-year changes. These analyses define a longitudinal series of class variables which may be modeled temporally using time-lagged single-(UniODA) and multiple- (CTA) attribute ODA methods. Figure 2.6 presents a line plot of the ipsative ICS series (discussed later in this Chapter) over a recent 36-month period (Mean ICS = 72.6, SD = 6.5).

Figure 2.6: Scatter Plot of Ipsative ICS Score: Recent 36-Month Period



Year-Over-Year Changes: UniODA is used to assess each of the 13 year-over-year comparisons that exist for data in Figure 2.6. The analysis is called a *forward-stepping little jiffy* with a bin width of 24 months: the first half (more dated) of the ipsative ICS scores in each analysis are statistically compared with the second (more recent) ipsative ICS values. The results of these analyses are summarized in Table 2.8: changes indicated as UP were statistically significant at the generalized criterion; **bold** indicates that changes were significant at the experimentwise criterion.

Table 2.8: Summary of UniODA Analysis of 13 Year-Over-Year Changes in Ipsative ICS Score

Year Ending	Change in Annual z_{ICS}	ESS	ESP
May 2013	Up	58.3	70.6
April 2013	Up	58.3	70.6
March 2013	Up	66.7	75.0
February 2013	Up	75.0	80.0
January 2013	Up	75.0	80.0
December 2012	Up	75.0	80.0
November 2012	Up	58.3	70.6
October 2012	None	50.0	51.1
September 2012	None	41.7	42.0
August 2012	None	33.3	33.3
July 2012	None	33.3	44.4
June 2012	None	41.7	63.2
May 2012	None	41.7	63.2

As seen the ipsative ICS series did not have a statistically significant ($p < 0.05$) year-over-year change from May through October of 2012, and during this period the accuracy of the UniODA model (indexed by *ESS*) was in the moderate range with the exception of October—which met the minimum criterion for a relatively strong effect. In November of 2012 the first year-over-year increase in ipsative ICS score in the series occurred (a relatively strong effect), but it was statistically significant only at the generalized criterion. In the following three months, December of 2012 through February of 2013, statistically significant increases occurred ($p < 0.05$, experimentwise criterion), and all three models exactly met the criterion for a very strong effect. Finally, the most recent three months continue to show sustained, relatively strong year-over-year increases in ipsative ICS score, that are statistically reliable when considered at the generalized criterion.

Professional traders are more interested in recent trajectory than historical trends, so the second exploration of the recent ipsative ICS series examines successive year-over-year comparisons for data in Figure 2.6, starting by using UniODA to compare the most recent 12-month period with the preceding 12-month period, stepping backwards in time one month at a time.⁵⁹ When this is done presently the first, most current analysis—comparing the year ending May 2013 versus the prior year—was statistically significant: estimated exact $p < 0.028$; $ESS = 58.3$; $ESP = 70.6$. The second analysis, for the next-most-recent comparison involving year ending April 2013 versus the prior year, was not statistically significant at the experimentwise criterion. Using this approach, focusing on the most recent changes, only the most recent change had experimentwise $p < 0.05$.

Month-Over-Month Changes: While the year-over-year changes in the ipsative CSI series are of interest to longer-term investors, short-term traders focus on more recent, more granular time horizons. For example, were the index updated every hour using a different random set of respondents, then hourly changes in the index would be of greatest interest to short-term traders.

Comparing one month versus another month statistically is not possible using ODA methods.⁶⁰ However, statistical methods have been developed on the basis of classical test theory which are used for analyzing data from a single-case “ N -of-1” series involving one or more attributes, and two or more measurement periods.⁶¹⁻⁶⁴ In the N -of-1 method, the ipsative z-score for the i^{th} measurement is subtracted from the ipsative z-score for the following $i+1^{\text{th}}$ measurement: if the difference is positive then the more recent ($i+1^{\text{th}}$) measurement was greater than the less recent (i^{th}) measurement; if the difference is negative then the opposite is true; and if the difference is zero then the two measurements were identical. The absolute value of the difference between the two ipsative z-scores is compared against a *critical difference* (CD) score, which is a function of the lag-1 autocorrelation coefficient [ACF(1)] for the data in the series; the number of inter-score comparisons which are being conducted (J); and the z-score corresponding to the desired experimentwise Type I error (p) level, taking into consideration if analyses are one- or two-tailed (for one-tailed $p < 0.05$, $z = 1.64$; for two-tailed $p < 0.05$, $z = 1.96$). CD is computed as $CD = z \times (J \times [1 - ACF(1)])^{1/2}$.

In the present application $ACF(1) = 0.742$ ($p < 0.0001$), and a total of 35 month-over-month two-tailed comparisons are to be evaluated. Thus $CD = 1.96 \times (35 \times (1 - 0.742))^{1/2} = 5.89$. The CD score is massive due to the large number of tests of statistical hypotheses (35) that are conducted. None of the month-to-month absolute differences in ipsative ICS score were as large as the CD score, indicating the absence of any statistically significant effects at the experimentwise criterion. If one instead used the generalized “per-comparison” criterion, the value 1 is used in the formula for CD rather than 35 (indicating one test of a statistical hypothesis), and $CD = 1.96 \times (1 \times (1 - 0.742))^{1/2} = 0.996$. Presently, six of the month-over-month differences were as large or were larger than this CD score, indicating the presence of a statistically reliable month-over-month change at the generalized criterion. The six significant month-over-month changes are listed in Table 2.9.

Table 2.9: Month-Over-Month Differences in Ipsative ICS Score with Generalized $p < 0.05$

Month i	Month $i + 1$	$Z_{i+1} - Z_i$
1 (6/1/2010)	2 (7/1/2010)	-1.27
9 (2/1/2011)	10 (3/1/2011)	-1.55
13 (6/1/2011)	14 (7/1/2011)	-1.21
14 (7/1/2011)	15 (8/1/2011)	-1.22
30 (11/1/2012)	31 (12/1/2012)	-1.52
35 (4/1/2013)	36 (5/1/2013)	1.25

As seen, five of the six statistically significant changes involved a precipitous decline in the ipsative ICS value. In Figure 2.6 the six significant month-over-month changes are readily seen. The five significant monthly declines occurred after months 1, 9, 13, 14 and 29. The only statistically significant rise in the ipsative ICS score over this series occurred in the most recent measurement, after week 35.

Normative Standardization

Normatively standardized (z) scores are widely used in the literature. The computational formula for a normative z score is: $(\text{observation's score} - \text{mean score}) / \text{standard deviation (SD)}$, where z_N , *mean* and *SD* are based on data for the *sample of observations*. Conceptually z_N measures the magnitude of an observation's score relative to the population of scores *for all observations* (i.e., the population). A common use of normative z scores is equating units of measurement across independent groups. An example of the use, and of the importance of normative z scores in this context is presented later in this Chapter, in the example demonstrating assessing model hold-out validity.

Ipsative Standardization

Ipsatively standardized (z) scores are rarely reported in the literature. The computational formula for an *ipsatively standardized z score* is: $(\text{observation's score} - \text{mean score}) / \text{standard deviation (SD)}$, where z_i , *mean* and *SD* are based on the data *from the observation*: conceptually z_i measures the magnitude of any observation's score relative to all scores in the population of scores *for the observation*.³ Ipsative standardization is crucial in analysis of data from designs involving multiple data recordings—an experimental method widely used across quantitative scientific disciplines, and called time-series, repeated measures, clinical trial, test-retest, longitudinal, prospective, pre-post, AB, or, generally, a serial design. In a serial design each observation is assessed on the same measure or set of measures on two or more test sessions that are spaced by a theoretically meaningful time span. Two different types of serial statistical designs are the *N-of-1 single-case* design where data are collected for a single observation assessed across time, and the *sample-based* design in which data are collected and combined for a sample of two or more observations assessed across time.

Discussion presented ahead illustrates how to analyze *N-of-1* serial data using UniODA and CTA (Chapter 5, *t*-Test, Within Subjects), and demonstrates how failing to ipsatively standardize serial data prior to analysis can induce paradoxical confounding in *N-of-1* series involving ordered or categorical attributes (Chapter 9, Confounding in Single-Case Series).

As an example of the use of ipsative standardization in a sample-based serial design, consider a study of the molecular level of a substance present in samples of blood collected at four theoretically meaningful time points for $N = 12$ rats.⁶⁵ The objective is to assess whether the combined (sample) data changed across time in an organized or disorganized manner. Data are presented in Table 2.10 separately by observation (rats are dummy-coded as 1 through 12) and by testing period (dummy-coded as 1 through 4), and include raw scores (*Raw*) normative (z_N) and ipsative (z_i) standard scores. A scatterplot of the *Raw* scores (Figure 2.7) reveals enormous variation in the *Raw* scores among observations in every session.

Figure 2.7: Observations' *Raw Scores* Across Four Serial Measurements

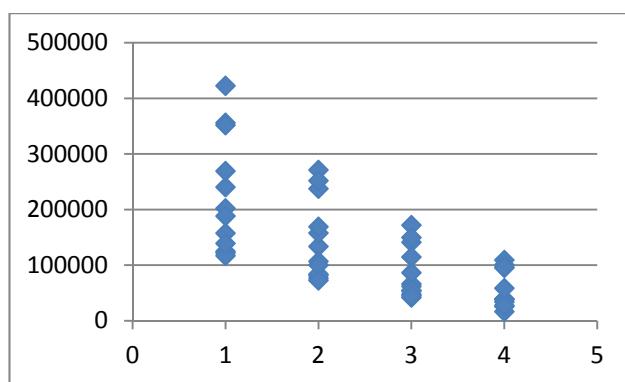


Table 2.10: Raw, z_N and z_I Data by Observation (Rat) and Test Session

Observation	Test Session	Raw Score	z_N Score	z_I Score
1	1	124820	-0.950838618	1.3251421085
	2	81560	-0.879555332	0.2094230035
	3	47820	-0.837290734	-0.660765684
	4	39560	-0.510339148	-0.873799428
2	1	122040	-0.977369006	1.2652444035
	2	76500	-0.949258542	0.2672495973
	3	41980	-0.963754323	-0.489245368
	4	16700	-1.185151183	-1.043248633
3	1	422960	1.8944023744	1.3095198891
	2	271380	1.7352793294	0.2002261326
	3	172240	1.8569900367	-0.52530087
	4	109500	1.554243257	-0.984445152
4	1	117189	-1.023663579	1.1939232206
	2	84200	-0.84318844	0.3555004892
	3	53580	-0.712559523	-0.422713583
	4	25880	-0.914163673	-1.12671521
5	1	356400	1.2591999847	1.2884876019
	2	237660	1.2707749316	0.2582181214
	3	141380	1.1887252514	-0.577173032
	4	96160	1.1604553062	-0.969532691
6	1	352000	1.2172094421	1.2380738368
	2	252120	1.4699663193	0.3438745829
	3	150040	1.3762551619	-0.57002069
	4	100680	1.2938827079	-1.01192773
7	1	202400	-0.210469006	1.2579331645
	2	133800	-0.159931674	0.3196312993
	3	66546	-0.431784371	-0.600260153
	4	38980	-0.527460363	-0.977304311
8	1	157640	-0.637627343	1.2612165217
	2	107080	-0.528008706	0.2992023807
	3	62240	-0.525029613	-0.5539763
	4	38460	-0.542810418	-1.006442603
9	1	188540	-0.342739215	1.372310766
	2	100500	-0.61865043	0.0865315946
	3	54920	-0.683542193	-0.579140993
	4	34340	-0.664430085	-0.879701367
10	1	139300	-0.81265156	1.3883732411
	2	72740	-1.001053813	0.0305988813
	3	45280	-0.892293733	-0.529564639
	4	27640	-0.86220964	-0.889407483
11	1	240780	0.1558030453	1.3706827988
	2	158200	0.1761865732	0.090386616
	3	115000	0.61747363	-0.579373557
	4	95500	1.1409725441	-0.881695858
12	1	269380	0.4287415721	1.3028529781
	2	169180	0.3274397846	0.244473081
	3	86800	0.0068104107	-0.625679972
	4	58780	0.0570225026	-0.921646087

The first normative standardization uses the sample *mean* (128,300) and *SD* (93,574) computed across all four sessions. As seen in Figure 2.8, this produces a scatterplot closely resembling the results obtained for *Raw* data. Note that the scale of the vertical axis is now indexed in standard units.

Figure 2.8: Observations' z_N Scores (Based on Grand *Mean* and *SD*) Across Four Serial Measurements

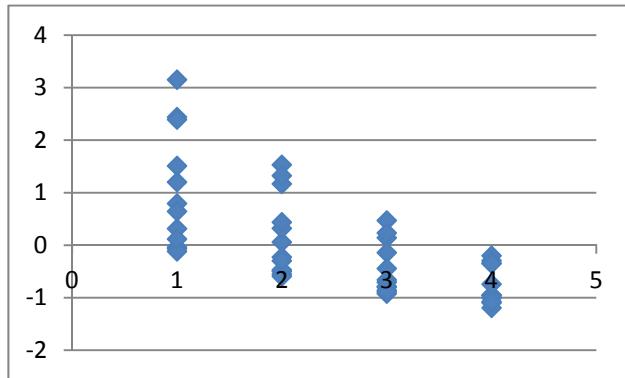
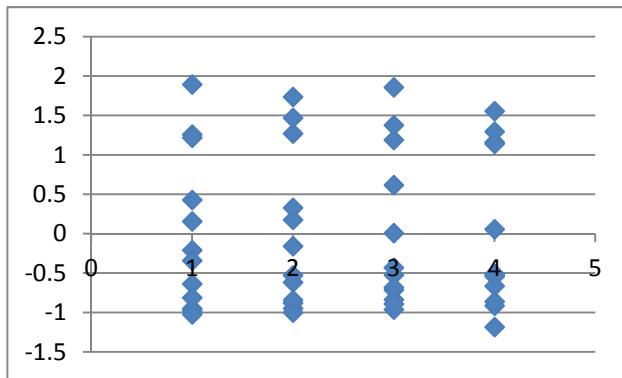


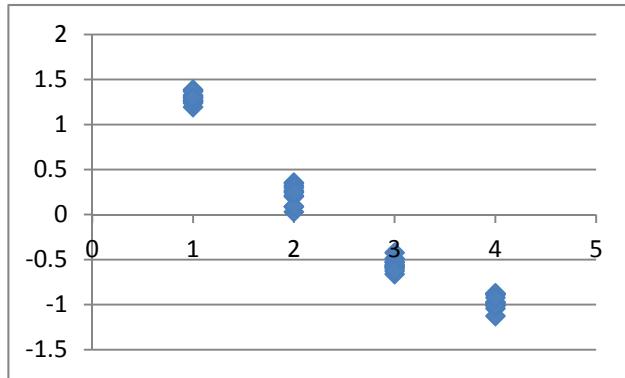
Figure 2.9 shows z_N scores computed with the mean and *SD* for each session. Note that this method *increases* the relative dispersion between the observations as a result of equating the *SD*—which is now 1 for all four test sessions. And, this method invalidates statistical comparison between the testing sessions by equating the means—that are now 0 for every session.

Figure 2.9: Observations' z_N (Computed using Session *Mean* and *SD*) Across Four Serial Measurements



Finally, Figure 2.10 presents the scatterplot of observations' z_I scores measured across four test sessions, which clearly shows ipsative standardization of the *Raw* data eliminated the variability between observations attributable to “base-rate” differences in the rats’ individual *mean* and *SD* parameters. When data for individual observations are viewed in the context of their own personal base-rate—that is, in terms of their own *mean* and *SD*, then as is seen in Figure 2.10, the molecules in the rat blood behaved notably uniformly.

Figure 2.10: Observations' z-Scores Across Four Serial Measurements



Interactive Transformation

In an interactive transformation values on the attribute or weight are transformed into a scale where, as an arbitrary example, 0 represents the theoretical minimum score (*MIN*) that it is possible to attain (an absolute scale) or that was observed (a relative scale), and 1 represents the theoretical maximum score (*MAX*) that it is possible to attain (absolute scale) or that was observed (relative scale).^{66,67} For any raw score on the attribute or weight, denoted as *X*, interactive standardization is accomplished using the following formula:

$$X_{\text{Interactive}} = 1 - (\text{MAX} - X) / (\text{MAX} - \text{MIN}),$$

where $X_{\text{Interactive}}$ is the value of *X* on the 0 to 1 interactive scale. If, for example, an observation had a raw score of 2.5 on an attribute or weight for which the theoretical minimum score was 1 and the theoretical maximum score was 3, the corresponding interactive score would be $1 - (3 - 2.5) / (3 - 1)$, or 0.75.

Two Common Mistakes

In some applications *a priori* data transformations are essential to properly represent the spirit of a theory and most precisely address the phenomenon under investigation. Data transformations discussed above are “statistically-motivated” in this context, and it is thus somewhat remarkable that most are relatively sparsely seen in the literature. In contrast, the literature is replete with examples of non-statistically-motivated data transformations that are commonly performed on ordered and categorical attributes in an effort to “simplify” data collection, force data to conform to assumptions underlying the validity of suboptimal statistical methods, and “simplify” the interpretation of statistical models. Unlike motivated transformations selected on the basis of theoretical considerations, non-motivated transformations contrived on the basis of pragmatic considerations can unwittingly mask underlying effects, limit statistical power, reduce model classification accuracy, and induce paradoxical confounding.

Don’t Parse Ordered Attributes: Within the limits of the particular measurement methodology used in a scientific investigation, raw data reflect the precise numerical representation of the observed phenomenon. In contrast, modifying the measurements vis-à-vis a statistically unmotivated or “arbitrary” parsing of raw data is *not* an exact numerical reflection of the observed phenomenon. Nevertheless, the arbitrary (“eyeball”) parsing became a standard practice with the introduction of digital scanners that required the respondent’s answers (e.g., to survey items or test questions) to be indicated using standardized scan sheets. Each item on the survey or test could have a maximum of ten possible responses: the response to each item was indicated by making a mark (filling a small “bubble” using a #2 graphite pencil). For example, to assess age, an item on a scan sheet would offer options: bubble 1 = 10 to 19 years; bubble 2 = 20 to 29 years; bubble 3 = 30 to 39 years; and so forth. This technology triggered a tsunami of research

involving arbitrarily-parsed attributes across science. The following example illustrates how arbitrary parsing can reduce the accuracy obtained by a statistical model.⁶⁸ Imagine that ten randomly selected attendees of an art auction volunteered information on whether they purchased an item (yes=1, no=0), and their age in years. Table 2.11 presents hypothetical data for this imaginary field study.

Table 2.11: Art Purchasing and Age (Hypothetical Data)

Was Item Purchased	Age	Recoded Age
0	23	2
0	25	2
0	27	2
0	31	3
0	33	3
1	35	3
1	39	3
1	41	4
1	45	4
1	49	4

UniODA can be used to predict whether or not an item was purchased as a function of customer age. Here the UniODA model is: if Age \leq 34 years then predict NO purchase; otherwise predict a purchase occurs. This model correctly classifies every subject in the sample ($p < 0.008$; $ESS = 100$), and the results indicate that all people older than 34 years of age purchased an art item.

Imagine age was arbitrarily parsed into categories: ages 20-29 coded as “2”; 30-39 as “3”; and 40-49 as “4” (Table 2.11). For these data the UniODA model is: if Age \leq 2 (20-29 years) then predict NO purchase; otherwise predict that a purchase occurs. This model misclassified two people (ages 31 and 33), and this result *wasn’t* statistically significant, $p < 0.29$, $ESS = 60$. Clearly, parsing an attribute may mask optimal threshold scores that would otherwise provide greater accuracy.

There are situations in which a manual parse is appropriate, such as when the optimal threshold is well-established. For example, in emergency medicine a widely-accepted empirically-based definition of fever is a temperature $\geq 101.5^{\circ}\text{Farenheit}$: thus, if a desired attribute is a binary indicator of whether an observation had a fever, this operational definition could be used to create the desired binary indicator. The application of *a priori* theoretically-based scoring algorithms to define attributes is also appropriate: theoretical findings should be compared with findings obtained in exploratory analysis that specifically optimizes attribute thresholds for the sample. Finally, there are circumstances in which a manual parse is necessary. For example, consider a study investigating the use of plasma D-Dimer as a screening test for deep vein thrombosis (DVT) for 105 non-ambulatory rehabilitation patients experiencing recent ischemic or hemorrhagic stroke.⁶⁹ Using raw D-Dimer data UniODA was unable to identify a model that provided perfect sensitivity for predicting DVT, so a threshold value that correctly classified all patients positive for DVT was defined based on the sample data and statistically evaluated using Fisher’s permutation method.

Don’t Parse Categorical Attributes: Early instances of arbitrary parsing occurred for multicategorical attributes such as marital status, ethnicity, or vocation that were evaluated using chi-square analysis. Such attributes typically feature disproportionate and often sparse representation of some categories. As a means of avoiding markedly imbalanced marginal distributions and satisfying the minimum expectation assumption, various categories were combined. Even though this practice reduces statistical power (vis-à-vis decreased measurement precision) and can induce paradoxical confounding (discussed ahead), it persists today. For example, many studies examining ethnic category include a category named “other”.

Multivariable parametric linear methods such as logistic regression analysis, discriminant analysis, or multiple regression analysis directly incorporate binary categorical attributes into the model. However, for multicategorical attributes having more than two levels each level must first be individually

dummy-coded, then one level must be selected for use as a *reference category* and omitted from analysis. For example, imagine that a study involved three ethnic categories: Navajo, Sumatran, and Inuit. Readyng this attribute for linear analysis requires creating three new binary attributes: for example, [a] Navajo (1) vs. others (0); [b] Sumatran (1) vs. others (0); and [c] Inuit (1) vs. others (0). Only two of these dummy-coded variables can be used as attributes in analysis, and one's choice can mask an effect depending on which class is selected as reference category. As an increasing number of polychotomous attributes are used, the associated design matrix rapidly becomes massive, increasing the prevalence of sparse or empty cells and the likelihood of highly imbalanced marginal distributions.⁷⁰ In addition to possibly masking actual effects, inducing numerical instability, failing assumptions underlying the validity of p , identifying overdetermined models, and being antithetical to the axiom of parsimony, when using computer-intensive methods such as CTA a larger number of attributes increases memory and time resources needed to obtain an optimal solution. In contrast, UniODA and CTA use aggregated multicategorical attributes. For example, in the present example one “ethnicity” attribute with three levels (rather than three different attributes each with two levels) requires coding: Navajo (1), Sumatran (2), or Inuit (3).

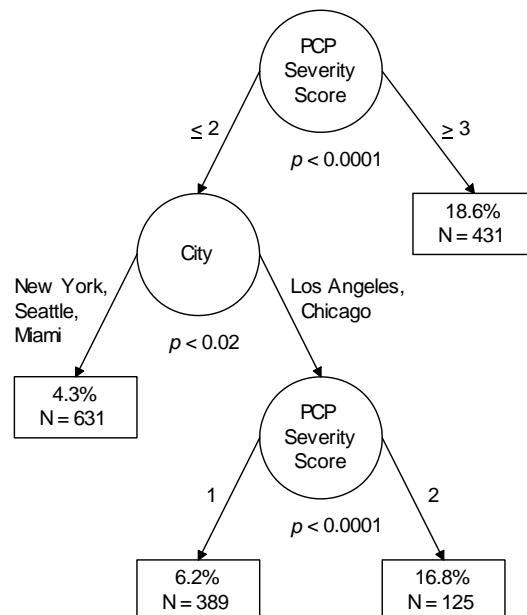
As an example of the use of aggregated attributes in bivariate ODA analyses, consider an application involving predicting use of mechanical ventilation for hospitalized patients with *Pneumocystis cariini* pneumonia (PCP).⁷¹

The UniODA example contrasts intubation rate for a total sample of 1,211 hospitalized PCP patients in Chicago, Los Angeles, Miami, New York, and Seattle. Analysis used an aggregated city attribute, with cities arbitrarily coded using dummy-codes of 1-5. The UniODA model was: if city = Los Angeles or Chicago then predict a higher ventilation rate; otherwise predict a lower ventilation rate. This model correctly classified 54.9% of 1,418 non-ventilated, and 61.9% of 147 ventilated patients, yielding a relatively weak $ESS = 16.8$ ($p < 0.0006$), that was stable in jackknife validity analysis.

In the original research from which the example was drawn, ventilation was modeled by logistic regression analysis: predictive factors that emerged included a PCP severity score developed previously by CTA, location (Los Angeles vs. others), ethnicity (African-American), and cytological confirmation of PCP diagnosis.⁷² For the present exposition the same attributes (severity score, city, and ethnicity) used for logistic regression were modeled using enumerated CTA (see Chapter 11) with a minimum endpoint strata size of $N = 25$ to ensure minimally adequate statistical power (see Chapter 3).

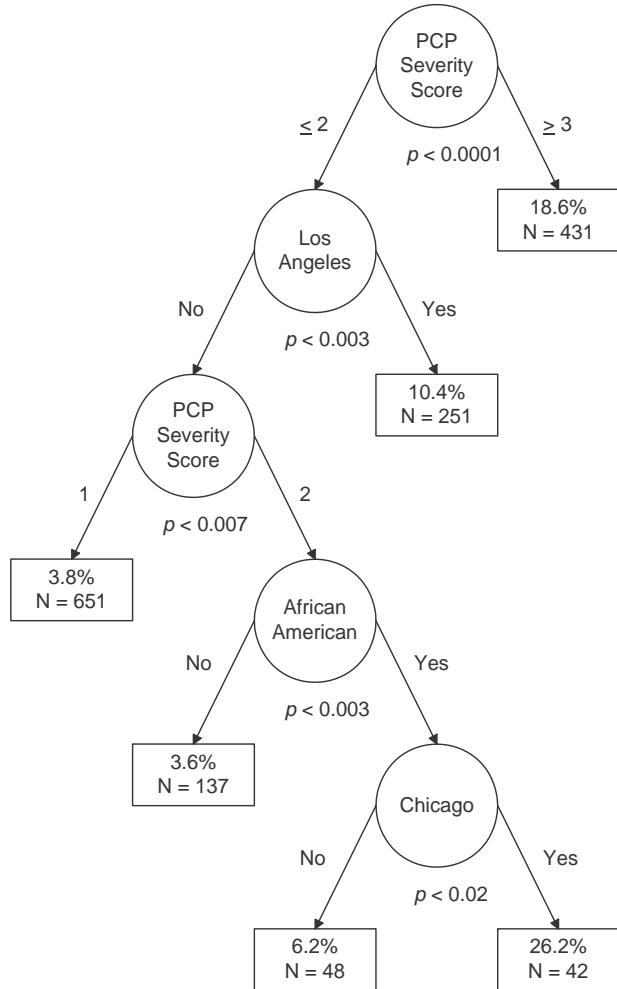
The first analysis used aggregated race and city attributes. The “aggregated attributes” model (Figure 2.11) correctly classified 66.4% of intubated and 68.1% of non-intubated patients: $ESS = 34.5$.

Figure 2.11: CTA Intubation Model using Aggregated Race and City Attributes



The second analysis used individually dummy-coded race and city attributes, although unlike linear models which require omission of a reference attribute from analysis, with CTA all of the binary (dummy-coded) attributes compete for admission to the model: rather than guessing if the “correct” reference category was eliminated, CTA identifies the model that explicitly maximizes *Mean ESS*. The “separately coded attributes” model (Figure 2.12) correctly classified 78.0% of intubated and 57.0% of non-intubated patients, yielding *ESS* = 35.0.

Figure 2.12: CTA Intubation Model using Disaggregated Dummy-Coded Race and City Attributes



In the ODA paradigm, model *parsimony* is quantified as $\text{efficiency} = \text{model ESS} / \text{number of strata}$ (see Chapter 12). Presently the aggregated attributes model used four endpoints (unique sample strata) to achieve $\text{efficiency} = 34.5 / 4 = 8.625$ *ESS* units-per-strata. The disaggregated attributes model used six endpoints to achieve $\text{efficiency} = 35 / 6 = 5.833$ *ESS* units-per-strata. Thus, the aggregated attributes model is 47.9% more efficient (parsimonious) than the disaggregated attributes model. Note that an advantage of parsimonious models is that by having fewer endpoints into which observations are partitioned, the minimum endpoint *Ns* are typically larger. Here, the minimum endpoint size for the aggregated attributes model (*N* = 125) is nearly three times greater than is the case for the disaggregated attributes model (*N* = 42): estimates for the former model are more likely to cross-generalize to smaller independent samples.

Evaluating Model Reproducibility

Sensitivity, predictive value, Overall PAC, ESS, ESP, and efficiency statistics are used to evaluate the classification performance achieved by a statistical model, regardless of how the model was developed. Because classification accuracy is the objective function explicitly maximized by the ODA paradigm, by definition models developed using the ODA paradigm can achieve the highest-possible levels of *Overall PAC* and *ESS* for a given sample, data geometry, and hypothesis. However, in most applications attaining maximum possible accuracy comes at the cost of reduced parsimony reflected by many model endpoints, and corresponding small endpoint *Ns*. If the analysis is a “one-off” specifically focused on a singular application, or conducted for rare, important data, then maximal-possible specificity may be desirable. However, in applications presumed to reflect universal (within context) applicability to parallel designs, cross-generalizability (reproducibility, replication) is the true measure of success. This perspective of classification performance directed the development of the ODA paradigm since inception, and multiple validity analysis methods can be used for every ODA model.

Identifying accurate, parsimonious, cross-generalizable classification models is in part the reward for ceaseless effort in this regard: living a statistical classification way of life, so to speak. Birth begins with selection of appropriate measurement methodologies for attributes and class variables, and the drafting of *a priori* hypotheses whenever possible. Development continues in the training analysis in which the optimal model is identified: minimum strata (endpoint) sample sizes are specified to ensure adequate statistical power; a sequentially-rejective multiple-comparisons method is used to guarantee the desired experimentwise Type I error rate; and CTA software enables the operator to enforce the constraint that only models with stable (unchanged) classification performance in leave-one-out validity analysis (see below) are permissible. Maturity is reached in validity analysis, and two different cross-generalizability methodologies are offered by UniODA and MegaODA software: hold-out methodology that is widely used in the literature, and a unique generalizability algorithm that identifies the ODA model that—when it is simultaneously and independently applied to two or more independent samples—maximizes the lowest *ESS* (or *Overall PAC*) that is attained across samples.

Leave-One-Out (Jackknife) Analysis

To what degree is the classification performance obtained by an ODA model—developed using a training sample—indicative or representative of the classification performance expected by using the ODA model to make classifications for an independent sample of observations? To address this issue the *one-sample jackknife* or *split-half* procedure, which randomly splits a data sample into two groups, was developed as a method for determining bias and standard error of the estimate for statistical procedures.⁷³⁻⁸¹

The *leave-one-out* (LOO) procedure is an extreme variation of the one-sample jackknife in which each observation constitutes a hold-out validity sample of size $N = 1$. The rationale underlying LOO validity analysis is that training classification performance is an optimistic *upper-bound* estimate of the true cross-generalizability of the model. This is because of the problem of *overfitting*: the training model capitalizes on chance errors that occur in the training sample so as to maximize classification performance specifically for the training sample. Because independent random samples don’t share the idiosyncrasies of the training sample, classification performance obtained using the training model to classify observations in independent samples will therefore generally be lower than was obtained in training.

To illustrate how LOO validity analysis proceeds, imagine an application with ten observations, each indexed by a unique integer between 1 and 10, inclusive: observations are referred to as observation 1, observation 2, etcetera. LOO is an iterative procedure, involving the same number of iterations as the number of observations in the sample: in this example the LOO procedure will require ten iterations.

In iteration number 1: (a) remove (i.e., hold-out) observation 1 from the sample; (b) obtain a ODA model for the subsample consisting of observations 2 - 10; (c) use the resulting model to classify observation 1; and (d) store the result for later tabulation.

In iteration number 2: (a) hold-out observation 2; (b) obtain a ODA model for the subsample consisting of observation 1 and observations 3 - 10; (c) use the resulting model to classify observation 2; and (d) store the result for later tabulation.

In iteration i : (a) hold-out observation i ; (b) obtain an ODA model for the subsample consisting of all observations except for observation i ; (c) use the resulting model to classify observation i ; and (d) store the result for tabulation. This procedure is continued until all observations are held-out and classified by an ODA model obtained using a sample that didn't include the observation being classified.

Once iterating is completed, the ten LOO classifications are tabulated using a confusion table, and the LOO classification performance of the model is assessed. If LOO classification performance is lower than training classification performance, this suggests that if the model is used to classify an independent random sample, then the classification performance yielded by the model may be lower than was obtained in training analysis. In contrast, if LOO and training classification performance are the same, this suggests that model performance may cross-generalize if the model is used to classify an independent random sample. Examples of LOO analysis are presented throughout this book.

Hold-Out Analysis

A widely-used, straightforward method of estimating the classification error rate of a predictive model, known as the hold-out, one-sample jackknife, cross-generalizability, replication validity, or the cross-validation procedure, involves applying the model to classify an independent random sample and assessing if the findings from training analysis replicate. In large-sample applications some researchers randomly split the total sample into halves, one for use as the training sample to develop the model, and the other for use as the hold-out validity sample ("K-fold" variations of this method involving more than two partitions of the total sample have also been developed). To estimate hold-out validity, a model developed using a training sample is used to classify observations in one or more independent hold-out samples: the classification error rate for the hold-out sample(s) is used as the estimated hold-out classification error rate for the model.^{82,83} Worked examples of hold-out analysis performed via UniODA for an application involving a directional hypothesis, a multicategorical class variable and attribute, and two independent samples, and for another application involving a directional hypothesis, a multicategorical class variable, an ordered attribute, and five independent samples, are available elsewhere.³

Hold-out methodology is demonstrated here with data from a study using information available prior to patient hospital admission to create a staging system for categorizing in-hospital mortality risk of HIV-associated community-acquired pneumonia (CAP).^{84,85} Data were acquired via retrospective review of medical records of 1,415 patients with HIV-associated CAP, hospitalized in 1995-1997 at 86 hospitals in seven metropolitan areas. The sample was randomly halved, and one half-sample was randomly selected as the training sample and used to develop the CTA model (Figure 2.13). As seen in Table 2.12 the model accurately predicted the mortality status of 68% of patients who lived and 80% who died ($ESS = 48$), and it was accurate in 97% of the cases predicted to live and 20% predicted to die ($ESP = 17$).

Figure 2.13: Training Sample CTA Model

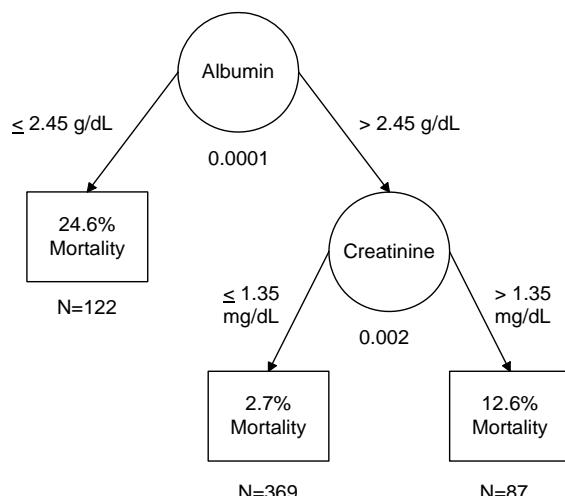


Table 2.12: Confusion Table for Training Model

		Patient Predicted Status		
		Alive	Dead	
Patient Actual Status	Alive	359	168	68.1%
	Dead	10	41	80.4%
		97.3%	19.6%	

The other half-sample is used as a *hold-out sample* and is employed to assess hold-out validity of the training CTA model. Validity analysis is conducted using the holdout function provided in UniODA and MegaODA software. Using this procedure every model node is individually evaluated, starting at the root and working down model branches. Any discrepancies which may emerge between training and hold-out results will be discovered at their inception in the tree model.

Raw Score Method: Hold-out validity analyses are typically conducted using data recorded in their original metric (i.e., raw scores), which therefore is how this exposition begins. Table 2.13 gives descriptive statistics for the model attributes separately by sample.

Table 2.13: Descriptive Statistics for Model Attributes, Separately by Sample

Attribute	Sample	N	Mean	SD
Albumin	Training	580	3.02	0.82
	Hold-Out	541	3.09	0.79
Creatinine	Training	717	1.37	1.73
	Hold-Out	674	1.32	1.82

Note: Units are g/dL for albumin, and mg/dL for creatinine.

In the first step of the procedure the root node was evaluated using the following UniODA / MegaODA syntax: training.dat and holdout.dat are the training and hold-out datasets, and “creat” is creatinine.

```
OPEN training.dat;
OUTPUT holdout.out;
VARS mortal albumin creat;
CLASS mortal;
ATTR albumin;
MC ITER 25000;
HOLDOUT holdout.dat;
GO;
```

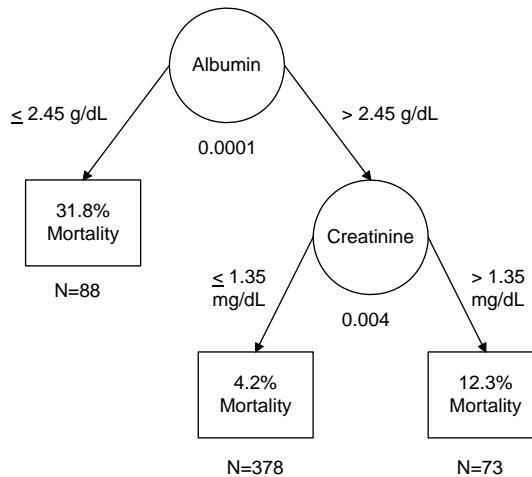
From the output the UniODA model for the root attribute of the CTA model is: if albumin ≤ 2.45 g/dL then predict the patient died; otherwise predict that the patient died ($p < 0.0001$, $ESS = 41.4$). Also from the output, application of this model to holdout data replicated training results: $p < 0.0001$, $ESS = 40.5$. Hold-out p is one-tailed: the null hypothesis is that the training model will not replicate when it is used to classify observations in the hold-out sample.

The second and final attribute was then evaluated by adding three lines of code:

```
IN albumin>2.45;
ATTR creat;
GO;
```

Program output for this analysis includes the UniODA model for the second attribute of the CTA model: if creatinine ≤ 1.35 mg/dL then predict the patient lived. The output also gives performance data for exactly this model applied to holdout data, and results indicate this node was replicated: $p < 0.004$, $ESS = 21.0$. The overall hold-out validity performance ($ESS = 44.3$) was 8.8% lower than was achieved for the training sample. Hold-out findings for the CTA model are illustrated in Figure 2.14.

Figure 2.14: Classifying the Hold-Out Sample Using the Training Model and Raw Data



Summarized in Table 2.14, the CTA model accurately predicted mortality status of 74% of patients who lived (9% more accurate than the training model) and of 70% of patients who died (13% less accurate than training model). When the model predicted that a patient would live it was accurate in 96% of cases (2% less accurate than the training model), and in 23% of cases predicted to perish (17% more accurate than the training model).

Table 2.14: Confusion Table for Application of Training Model to Raw Hold-Out Data

		Patient Predicted Status		
		Alive	Dead	
Patient Actual Status	Alive	362	124	74.5%
	Dead	16	37	69.8%
		95.8%	23.0%	

It is natural to wonder what would occur in a split-half analysis if the training and validity data sets were switched, and data originally used in training were instead used in validation, and *vice versa*. Just discussed, this issue motivated development of a family of bootstrap methods that involve iterative resampling of jackknife half-samples. The *leave-one-out* method is an efficient error estimation procedure available in ODA software, in which each observation is a hold-out validity sample of size $N = 1$.

A potentially serious problem involves a hold-out sample for which there is a significant mean difference in the attributes entering a CTA model as compared with the training data. Imagine that scores on an attribute are much higher in the training sample than in the hold-out sample. In this case, the cut-points obtained by CTA on an attribute for the training sample may likewise be much too high (or too low) to be *equivalently representative* in the hold-out sample. In this circumstance, transforming the original metric into a new sample-equivalent isometric, such as normative z scores, is necessary for a successful analysis. Between-sample mean differences noted presently were not extreme, suggesting that potential gain in hold-out validity arising from separate standardization may be limited, in particular because the *ESS* for training and hold-out models based on raw data are relatively comparable. However, classification decline in hold-out analysis suggests the possibility of a gain in *ESS*. This method is demonstrated next.

Normative z-Score Method: Before initiating validity analysis it is first necessary to normatively standardize the attributes, albumin and creatinine. This was done for observations in the training sample using parameters (*Mean*, *SD*) obtained for the training sample, and for observations in the hold-out sample using parameters obtained for the hold-out sample (see Table 2.13). For clarity, three significant digits are used in syntax reported here than were actually used in software code, in which eight significant digits were employed to minimize round-off error. The minimally sufficient number of significant digits to use in syntax is easily determined when standardizing attributes, as being the number of digits required to produce a sample having *Mean* = 0, and *SD* = 1. In the first step of the procedure the root node was evaluated with the following UniODA code: *ztrain.dat* and *zholdout.dat* are training and hold-out datasets, *zalbumin* is *z*, for albumin, and *zcreat* is *z*, for creatinine.

```
OPEN ztrain.dat;
OUTPUT zexample.out;
VARS mortal zalbumin zcreat;
CLASS mortal;
ATTR zalbumin;
MC ITER 25000;
HOLDOUT zholdout.dat;
GO;
```

Program output for this analysis gives the UniODA model for the root attribute of the CTA model: if *zalbumin* \leq -0.695 g/dL then predict the patient died ($p < 0.0001$, *ESS* = 41.4). The standardized cutpoint may also be computed: standard cutpoint = $(2.45 - 3.02) / 0.82$. Program output also reports performance data for this model applied to holdout data, and results indicate that the root attribute was replicated ($p < 0.0001$). However, using an equally representative (based on standardized data) cut-point resulted in 11.5% greater overall accuracy at the root node (*ESS* = 45.2) compared with analysis using raw scores.

The second and final attribute was then evaluated by adding three lines of syntax:

```
IN zalbumin>-0.695;
ATTR zcreat;
GO;
```

Program output for this analysis includes the UniODA model for the second attribute of the CTA model: if *zcreat* \leq -0.0091 mg/dL [i.e., $(1.35 - 1.37) / 1.73$] then predict the patient lived. This is exactly the same finding as was derived for raw data for the training sample, because the raw and standardized cutpoints have precisely the same relative value for training data. The output also gives performance data for exactly this model applied to holdout data, and results indicate this node was replicated: $p < 0.04$, but the *ESS* = 18.6 is 11.3% lower than the corresponding effect for raw data at this node.

Overall training performance of the CTA model was identical with raw and standardized data. In contrast, hold-out validity performance obtained with standardized data (*ESS* = 45.9) was 3.7% greater than was achieved using raw data. The hold-out validity findings obtained for standard data are illustrated in Figure 2.15, and summarized in Table 2.15.

Figure 2.15: Classifying the Hold-Out Sample Using the Training Model and Standardized Data

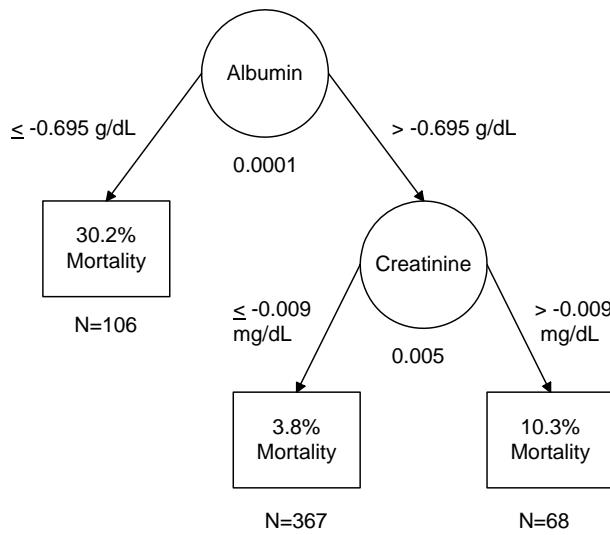


Table 2.15: Confusion Table for Application of Training Model for Standardized Data to Hold-Out Sample

		Patient Predicted Status		
		Alive	Dead	
Patient Actual Status	Alive	353	135	72.3%
	Dead	14	39	73.6%
		96.2%	22.4%	

Compared to hold-out results obtained using raw data, when using normatively standardized data the CTA model accurately predicted the mortality status of 72.3% of patients who lived (3.0% less accurate) and 73.6% of patients who died (5.4% more accurate). And, when the model predicted that a patient would live it was accurate in 96.2% of cases (0.4% more accurate), and in 22.4% of the cases predicted to perish (2.6% less accurate).

Separate standardization of attributes in training and all hold-out samples is a recommended practice, and is a safeguard against unwitting induction of Simpson's paradox (discussed ahead). If model statistics are desired in their raw units of measure, standardized units may easily be reconverted. The use of standard scores provides information about relative magnitude of model cut-points. For example, for creatinine the cut-point was virtually zero, which indicates a value approximating the sample Mean. For raw data, the albumin cutpoint of -0.70 falls beneath the sample Mean, representing the 24th percentile if the sample is normally distributed.

If a training model node fails to replicate (i.e., $p > 0.05$) in the hold-out sample, then in an effort to retain the training model geometry vis-à-vis relaxing parameter (threshold) estimates, UniODA is used to identify an adjusted threshold that is optimized for the hold-out data. Similar parameter relaxation may be required from the failed node through model endpoints. If the relaxed model fails to replicate then it is concluded that the model identified in training analysis *didn't* cross-generalize to the hold-out sample. In this circumstance, multisample generalizability analysis (see below) may identify a cross-generalizable model that meets the researcher's *a priori* minimum criterion for minimally acceptable performance. All else failing, exploratory CTA comparing the training and hold-out samples should be conducted in order to characterize the nature of the inter-sample differences.

Multisample Generalizability Analysis

A multisample application involves two or more independent samples for which all observations have data on the same set of class variables, attributes, and weights.

ODA is used in multisample applications to identify *differences* between samples. For example, in clinical research comparing control and intervention groups, treating group (sample) as a class variable enables UniODA and CTA to identify the attributes that differentiate the intervention and control groups.

ODA is also used in multisample applications to assess *similarities* between samples: by randomly selecting one sample to develop a model, and validating the resulting “training” model using “hold-out” samples, absolute (raw data) and relative (normative z-score data) cross-generalizability of the model can be assessed. Identification of a model having strong training accuracy, that cross-generalizes to hold-out samples, is more likely in designs having a directional hypothesis and using the same measurement scale (e.g., a binary scale; a 5-point Likert-type scale; etc.) for both class variable and attribute, than in designs involving a non-directional hypothesis and/or using class and attribute measurement scales of differing precision.³ As the magnitude of between-sample heterogeneity increases so does the impact of choice of training sample, and for applications with a moderate or weak effect strength a sensitivity analysis that permutes the selection of training sample is appropriate. As an extreme example, imagine a multisample application in which the identical ODA model achieves perfect accuracy for every sample except one—for which the ODA model yields the same level of accuracy that is expected by chance. Clearly, selecting the latter sample for training analysis would compromise the study. This example begs the question of how best to select the training sample in the context of multisample hold-out validity assessment.

Developed in response to this issue, the *multisample generalizability algorithm* (*Gen*) identifies the model that, when simultaneously and independently applied to each sample within a set of two or more independent samples, explicitly maximizes the minimum accuracy (i.e., weighted or non-weighted *Overall PAC* or *ESS*) that is obtained by the model for any of the samples.³ Combining Gen and prior odds weighting yields a model that maximizes minimum *Mean PAC* achieved for any sample; combining Gen and return weighting yields a model that maximizes minimum return-weighted *Overall PAC* achieved by the model for any sample; and combining Gen with prior odds and return-weighting yields a model that maximizes minimum return-weighted *Mean PAC* achieved by the model for any sample.

The Gen algorithm offers several unique advantages in statistical analysis involving multisample designs. Gen identifies the best model for a multisample application, in the sense that the model ensures that the minimum classification performance achieved for any sample is maximized. Gen represents a straightforward, objective means of assessing if the best ODA model generalizes for a set of independent samples. If the effect strength for the worst-classified sample is *lower* than the *a priori* criterion for successful generalizability, *or* if the corresponding *p* achieved for the worst-classified sample is *greater* than the *a priori* criterion for statistical significance, then no single model can achieve satisfactory classification performance for every sample. In contrast, if the *worst* performance achieved across samples *exceeds* the minimum *a priori* standards for acceptable accuracy, *and* if *p* for the weakest effect is statistically reliable, then the model achieved satisfactory performance for all samples. Finally, the finding that a Gen model is not satisfactory in the worst case implies that the model fails to apply across all of the samples: this suggests the possibility of paradoxical confounding if data are pooled (see below).

Detailed discussion of selection heuristics specifically available for Gen models in ODA software (Appendix A); distinctions between and examples of structurally invariant *fixed* (directional hypothesis, symmetric class and attribute precision) versus undetermined *room-to-vary* (non-directional hypothesis, asymmetric class and attribute precision) designs; and worked examples of Gen analyses for applications such as cross-cultural comparisons of associations, assessing generalizability of factors influencing patient satisfaction across different divisions of medicine, analysis of randomized block designs, and optimizing the ESS of multiple suboptimal multiattribute linear models, are available elsewhere.³ Multiple examples of Gen models are also presented in this book.

Simpson's Paradox

Confounding attributable to Simpson's Paradox⁸⁶ ranks among the greatest threats to the validity and reproducibility of findings reported in the literature.⁸⁷ Paradoxical confounding is discussed in Chapter 9, and in examples in other Chapters. The nature of the paradox is that statistical analysis of data conducted separately for individual samples, groups, or time periods may produce results that contradict the findings of the same statistical analysis conducted using pooled data for the samples, groups, or time periods. Paradoxical confounding can suggest invalid conclusions and/or mask actual differences existing between samples, groups, or time periods⁸⁸⁻⁹⁰ in applications involving categorical⁹¹⁻⁹⁷ and/or ordinal⁹⁸ data. Left unchecked there are many opportunities for paradoxical confounding to occur. For example, for a fundamentally simple design involving two ordered attributes (X and Y), and two groups (A and B), 21 different circumstances induce some form of paradoxical confounding.⁹⁸

To illustrate one of these 21 forms of paradoxical confounding, imagine that groups (samples, time periods) A and B both have two observations. For group A the first observation's scores are $X = 2$, $Y = 0$; the second observation's scores are $X = 0$, $Y = 2$; and $r_{XY} = -1.0$. For group B the first observation's scores are $X = 8$, $Y = 6$; the second observation's scores are $X = 6$, $Y = 8$; and $r_{XY} = -1.0$. Thus, when groups A and B are assessed independently, X and Y are perfectly negatively correlated. However, when data of groups A and B are pooled to construct a combined (total) sample consisting of all four observations, $r_{XY} = +0.8$. In this circumstance the remedy for Simpson's paradox is normative standardization of X and Y separately by sample: standardizing data in samples A and B produces $z_X = 0.71$, $z_Y = -.71$ for the first observation; and $z_X = -0.71$, $z_Y = 0.71$ for the second observation. For the pooled standardized data, $r_{XY} = -1.0$.

Normative standardization by group does not address all 21 circumstances. For example, in the example imagine that X and Y were perfectly positively correlated in sample C, and perfectly negatively correlated in sample D. If data were normatively standardized separately by sample and then combined, $r_{XY} = 0$. To meaningfully combine data from multiple groups, samples or time periods, it is not sufficient to normatively standardize the data separately by sample, but it is also necessary to verify that the relationship between X and Y is homogeneous across samples.⁹⁸⁻¹⁰⁰

Paradoxical confounding is a primary concern for all linear statistical models, regardless of their underlying methodological derivation (see Chapter 9). Forthcoming Chapters demonstrate how CTA used with appropriately transformed data circumvents paradoxical confounding: indeed, attributes that induce paradoxical confounding using linear models are often identified as nodes in CTA models, and in this context they are conceptually isomorphic with the construct of a "moderating variable".

Chapter 3

Methodological Matters

This Chapter discusses methodological matters that are central to every empirical investigation: selecting the measurement scales to use for class variable, attribute and weight; establishing the appropriate study sample size; preparing an ASCII data set for analysis; conduction the analysis and managing the associated data, program, and output files; and reporting the findings.

Measurement Scales

The objective of analysis conducted in the ODA paradigm is identification of the most accurate, parsimonious, and cross-generalizable model that is possible for a given sample, data geometry, and hypothesis. Discussion here focuses on sample and data geometry, and a visualization exercise will motivate conceptualization of the role of measurement in the description and modeling of classical phenomena.

A statistical design defines a space. Imagine the simplest statistical space, involving one binary class variable (dimension X in the space), and one binary attribute (dimension Y). The units on X are 0 and 1 (the two class categories), and the units on Y are also 0 and 1 (the two attribute categories). Neither the class variable nor the attribute have any hypothesized order with respect to each other, so the design is exploratory. The sample consists of a collection of N observations, with each observation represented as a circle (class 0 observations as open circles, class 1 observations as filled circles) at their measured position in the X by Y space. This two-dimensional, minimal precision, non-directional design is comparatively easy to visualize. Let's take one little step and make the design three-dimensional: add dimension Z—time, that is used to illustrate repeated measurements. Visually connect successive recordings for each individual observation using a line segment between circles, and visualize how the pattern of empty and filled dots could possibly change in different ways within and across time.

Make the attribute mult categorical. Rather than simply looking at empty and filled circles on units 0 and 1 across time, now observations from the two class categories are scattered among more than two measurements units across time. Imagine a mult categorical attribute having three possible levels. Visually connecting the corresponding measurements across time by individual, visualize the types of changes that could occur. There is a lot more space to visualize than was true for the binary attribute.

We continue to pursue more precise measurement, moving to ordinal scores such as Likert-type scores. See how the empty and filled circles are distributed on the Likert scale and how they vary over dimension Z. Visualizing this space is increasingly difficult as the number of measurement levels of the attribute increases. In the limit, imagine the most precise measurement scale possible for an attribute, real numbers—temperature, mass, molarity, frequency, time, monetary value, and so forth. In this design the empty and filled circles are distributed in a space that is wide (scale) and deep (recordings). Visualizing how an observation's scores vary within and across recordings and class category is much more difficult for a design involving repeated real numbers than for a binary design involving a single recording.

Now make the class variable mult categorical, first with three class categories. Observations are no longer empty and filled circles, rather the circles are different colors: empty for class 0, blue for class 1, red for class 2, and so forth. Visualizing the design is becoming more difficult. Now imagine a class variable with ten categories. In the limit imagine that the class variable is a real number. While this design is relatively complex, we have hardly yet begun to conceptually concoct!

Next start adding more attributes: hyperdimensions A, B, C, and so forth. Some of the attributes are categorical, some are ordered. Visualize the observations, located in this hyperdimensional space—their scores represented by circles, each circle the same size. Now weight the observations, and the size of the circles changes—the larger the size, the greater the weight. Now imagine and visualize the effect of

“measurement slop” in the system. For example, imagine the class variable is imperfectly precise (e.g., a measure of “ethnicity”), or is imperfectly valid in the sense that any of the class categories can unwittingly include groups (defined on the basis of other variables) that induce paradoxical confounding when they are combined (CTA disentangles the different types of groups). Further imagine that the attributes are imperfectly reliable, imperfectly valid indicators of putative underlying constructs. Thus, visualize a larger circle in single attribute space (or a hypersphere in multiattribute space) envelopes individual observation scores, reflecting a 95% confidence region for the actual location of the observation in the design space. If a directional hypothesis is being tested then it is not necessary to look at all of the differences between class categories and attributes but rather only at a subset of the differences, yet as the reader likely has some difficulty in clearly visualizing, this remains a complex optimization problem.

A statistical model is intended to make sense of these data, to succinctly discriminate between class categories in this design space. Identifying a model that makes a little bit of sense (i.e., returns a relatively weak effect strength) of the data is often relatively simple to accomplish, especially for a large sample, but is not of great interest in terms of functionality in solving a problem. Finding a model that makes a great deal of sense (returns a relatively strong effect strength) of data is difficult, especially for a large sample, but this is the ultimate objective. How can this ultimate objective be obtained?

Class Variables

The first step in designing a statistical analysis involves defining what the statistical model will be used to predict or compare. In the ODA paradigm this “target variable” is called a *class variable* (in legacy statistical paradigms it is called a “dependent variable”). A class variable is any random variable that may attain two or more categories or levels which represent the phenomena that are to be predicted. Examples of class variables include biological sex (female, male), experimental condition (intervention, control), or the direction of the daily price change for a commodity futures contract (lower, unchanged, higher). The *category level* of a class variable is the number of different levels that it is possible for the class variable to attain: mortality status and experimental condition are both two-category class variables, and direction of commodity price change is a three-category class variable.

Ideally the class categories should represent qualitatively distinct phenomena, conditions or states. For example, experimental condition involves two qualitatively different categories controlled by the investigator (it is traditional to conduct an analysis to determine if the experimental manipulation was successfully performed for each observation in the intervention arm of the study). Biological sex also involves two qualitatively different categories, and is a relatively stable class variable. It should be noted that while it is possible for programmatic experimental studies to control potentially confounding factors, this is not true for observational studies that are conducted in the “real world.” For example, combining all of the males in a study into one category and all of the females into another category, and comparing the two categories—implies that the males are homogeneous, the females are homogeneous, and the males and females are heterogeneous (the alternative hypothesis). Optimal non-linear multiattribute methods can ameliorate this issue—which otherwise “sets the stage” for paradoxical confounding. However, class variables defined as “phenomenon X” versus “all other phenomena” are likely analytically toxic because the latter group represents a paradoxical soup of unknown constitution.

In contrast to class variables involving qualitatively distinct categories, class variables that are defined on the basis of quantitative criteria are less ideally suitable. For example, in the commodity market late-breaking news can sometimes produce rapid price swings, and had the market been open for only a few more moments on a particular day, then the closing price might have been positive rather than negative, or vice versa. A widely studied class variable is mortality status: imagine prospectively following a cohort for a year (a common arbitrary time-unit selection), at which time each patient was classified as being alive or not. It is possible that if the study had been continued for one more day, then some patients classified as being alive would have been otherwise classified. A conceptually similar widely-studied class variable is age category: much research compares geriatric (≥ 65 years of age) and non-geriatric (< 65 years) groups of people. Imagine some subjects in the study will turn 65 years of age within a month: how much younger than 65 years of age should the threshold be to avoid this issue? The greater the instability or unreliability in a class variable, the more “fuzzy” the class variable becomes. Instability in the class variable that occurs near the defining threshold (e.g., 65 years of age) theoretically limits the upper bound of accuracy that an ODA model can achieve. Discussed ahead, for some quantitatively-defined class variables pre-processing data may be used to minimize definitional fuzziness, and aggregated confusion

tables may be used to assess if enhanced reliability (precision) of measurement of class variable and/or of attribute(s) will yield enhanced classification accuracy.

Finally, pragmatically speaking, it is a good idea for those learning the ODA paradigm to initially study binary class variables. In some applications it is possible that adding an intermediate “undecided” category might enhance model performance in those instances in which some subjects can’t be reliably classified into either type of the dichotomy (Loretta J. Stalans, Ph.D., personal communication). And, it is an excellent idea for those learning the ODA paradigm to experiment using data with which they are personally familiar, including all of one’s prior studies for which data are available. Familiarity with the data will simplify conducting ODA and interpreting resulting models, and thereby provide an excellent basis for objective comparison of legacy versus optimal statistical methods: such studies are readily published in high impact-factor peer-reviewed methodological journals in every substantive area of empirical science involving classical data. It is also possible, indeed when using a rich data set it is likely, that ODA will yield more accurate, parsimonious, and cross-generalizable models for the data than have been identified previously: such studies are readily published in high impact-factor peer-reviewed substantive journals in all areas of empirical science involving classical data.

Several decision analysis foci support the efficacy of starting small, the first and most important of which is desire to find a meaningful model—the foundation of programmatic research. Recall the visualization exercise. In extremely complex models there are many opportunities for measurement slop to reduce data quality and enlarge the diameter of 95% confidence hyperspheres enveloping individual observations at their measured locations in design space: the larger these fuzzy spheres, the lower the likelihood of identifying a high-quality model that will cross-generalize to independent samples. On the opposite side of the coin, the greater its strength (accuracy) and parsimony (efficiency), the greater the likelihood the model will cross-generalize to independent random samples—the actualization of reproducible research findings.

Proponents of granular class variables cite increased precision and greater statistical power as being their unqualified advantages, but this is not necessarily accurate. For example, discussed later in this Chapter, if the use of groups reflecting extreme scores (these are, after all, the scores and the groups the applied researcher seeks to explain, as opposed to the mean score and the “mean group”—if it exists) dramatically increases the reliability of group assignments, and the resulting classification accuracy of the model discriminating the groups, then the design with more extreme groups will have greater statistical power. In the visualization exercise, the open and filled circles will be more clearly separated in design space—versus less clearly separated as would be the case for weaker effect strength.

Another pragmatic issue favoring class variables with fewer categories is solution speed. ODA software is remarkably fast for problems involving binary class variables, but ODA is computationally intensive and problems become much more difficult to solve as the number of category levels increases (Appendix B).

The final pragmatic issue is the availability of commercially-available software for conducting UniODA and CTA analysis: the former allows up to ten class categories, the latter allows only two class categories. While the algorithm for globally-optimal CTA involving an unrestricted (real number) class variable is discovered and several analyses have been conducted, these methods aren’t yet available outside of the ODA laboratory.

Attributes

The second step in designing a statistical analysis involves defining the variable or set of variables to be used in an effort to predict the class variable. In the ODA paradigm the predictor variables are known as *attributes* (versus “independent variables” in legacy statistical paradigms). An attribute is any random variable that can attain two or more levels. In the ODA paradigm it is necessary to specify all of the categorical attributes consisting of two or more qualitatively distinct, *unordered* categories. The minimal precision level possible for an ordered attribute is two categories. For example, in a confirmatory study using two two-category variables, predicting poor (versus good) health status as a function of low (versus high) income, both the class variable (health status) and the attribute (income) are ordered in the context of the *a priori* hypothesis.

One of the most prevalent ordered scales in the literature are “Likert-type” ordinal categorical scales consisting of a small number of qualitative categories that are ordered with respect to an underlying theoretical factor.^{1,2} For example, as a means of rating their level of satisfaction with medical care received, a patient might be asked to select among the following descriptors: “very dissatisfied” (coded as 1), “somewhat dissatisfied” (2); “neutral” (3); “somewhat satisfied” (4) and “very satisfied” (5). The scale is designed to ask independent raters to

identify their internal state using a scale that is universally applicable: respondents putatively understand the difference between satisfaction and dissatisfaction, and between “very” and “somewhat.” For those for whom these distinctions are impossible to decipher, the mid-point is provided. And, as it turns out, Likert-type scales are actually much more sensitive than they seem, when they are examined using the correct analytic lens (see Chapter 9, Confounding in Single-Case Series).

Criticism of these measurement scales is common, typically based on the failure of scores obtained using these scales to satisfy the assumptions required by legacy statistical methods. However, as demonstrated in many examples in this book, such scales are often among the most productive attributes in ODA models—for which there are no distributional assumptions for data to violate. In the ODA laboratory we suggest the use of real-number measures whenever possible. For phenomena for which validated real-number scales are not available, we favor three-category Likert-type scales (e.g., no, unsure, yes) to assess phenomena with which the observations are unfamiliar (e.g., new or uncommon conceptual domains), and five- or seven-point scales when the rated phenomena are familiar. We avoid the use of percentage or ten-point ratings, given the national cliché of “giving 110% effort”, “scoring 11 on a 10-point scale”, and so forth. And, we have seen no evidence that randomly selected observations are able to rate exceptionally common phenomena accurately, for example the amount of time spent waiting to be seen by a physician.³

Weights

The third step of any ODA analysis involves specification of appropriate weights, used to enable an ODA model to explicitly maximize the desired objective function.

As discussed in Chapter 2, one type of weights is prior odds (antecedent probability). For example, imagine predicting whether a hospitalized patient with influenza will survive. If an ODA model for predicting mortality due to influenza failed to consider the base rate for mortality among hospitalized influenza patients, the model might overestimate the number of patients who died (most cases of influenza are not fatal).

The second type of weight is a quantitative assessment of the value or importance of the attribute to the modeler. For example, an investment model for daily trading in a stock would weight daily observations by the dollar value of the change in stock price. Were an ODA model constructed for this application without considering the return, the model would maximize ability to correctly predict the direction of movement of the stock. In the absence of a return weight the model might be correct, for example, 85% of the time that it predicted the movement in stock price—and yet still lose money because the model misclassified the days on which the price of the stock changed the most. However, specification of a return weight would identify an ODA model that maximized the amount of dollars correctly predicted: although overall predictive accuracy might decrease (e.g., say, to 40% correct predictions), the model would maximize accurate prediction of the times that the stock value changed substantially, and thereby maximize profit.

An interesting idea that hasn’t been addressed empirically is the use of an interactive transformation to standardize weights across applications: for example, allow a minimum weight of 1 and a maximum weight of 2 for every observation in the sample.

Precision

Imagine that, in the development of a funding proposal, a pilot study is conducted to obtain a preliminary estimate of the effect strength achieved by using an attribute to predict a class variable (discussed ahead, this estimate is used in statistical power analysis to determine the appropriate sample size for proposed research). Further imagine the preliminary data reveal a statistically reliable ($p < 0.05$) but relatively weak ($ESS < 25$) effect. Because a relatively weak effect means that a large, expensive sample is needed in order to demonstrate an ecologically marginal result, it is important to determine if the underlying relationship is inherently weak (i.e., is theoretically trivial or meaningless), or if increasing the precision by which the attribute and/or the class variable are assessed (i.e., adequate measurement) will reveal a stronger effect. In this context, a heuristic procedure involving examining the so-called aggregated confusion tables (ACTs) that exist for an application may aid in interpreting UniODA findings as well as in evaluating the potential for increasing effect strength vis-à-vis more precise measurement of the ordered class variables and/or the attributes that are used to represent underlying theoretical constructs.

ACTs have two columns and two rows for measurement scales having an *even* number of (class or response) categories, and three columns and three rows for scales having an *odd* number of response categories. If one exists, then the midpoint category (e.g., 2 on a 3-point scale; 3 on a 5-point scale) forms the second (middle) row and column of the ACT, and represents a “neutral” or “undecided” response. Measurement scales having an even number of response categories have no middle value in the ACT. To standardize ACTs across scale, entries *lower* than the midpoint are summed and entered on the left-hand side of the ACT, and entries *higher* than the midpoint are summed and entered on the right-hand side: the entries equal to the midpoint are ignored.

Two demonstrations of the use of ACTs are presented below: the first is a test of the confirmatory hypothesis, evaluated for a single sample, that men have a higher income than women (ACT simulation is used on the attribute); the second is a test of the exploratory hypothesis, evaluated for a training sample and validated for a hold-out validity sample, that mental focus and barometric pressure are related (ACT simulation is used on the class variable).⁴

Discriminating Sex using Income: Data were obtained from a convenience sample of 416 adult ambulatory patients waiting to be seen in general internal medicine clinic at a private hospital in Chicago. The binary class variable “sex” indicated if the patient was female (dummy-coded as 2; $n = 324$) or male (dummy-coded as 1; $n = 92$). The ordered attribute “income” was a 7-point scale with 1 used to indicate up to \$10,000 per year, 2 used to indicate $\leq \$20,000$, 3 for $\leq \$30,000$; 4 for $\leq \$40,000$; 5 for $\leq \$50,000$; 6 for $\leq \$60,000$; and 7 used to indicate more than \$60,000 per year (as is discussed in Chapter 2, this scale was designed for scan forms: actual annual income is the preferred more accurate measure). Descriptive statistics for income were as follows. For men: *mean* = 3.24; *SD* = 1.98; *median* = 3; *skewness* = 0.55; *kurtosis* = -0.68; and *CV* = 61.1. For women: *mean* = 2.88; *SD* = 1.64; *median* = 2; *skewness* = 0.76; *kurtosis* = 0; and *CV* = 56.9. The first analysis tested the *a priori* hypothesis that men have higher income than women using the following UniODA and MegaODA syntax [the DIR command specifies the directional hypothesis that women (2) have lower (<) income than men (1)]:

```
VARS sex income;
CLASS sex;
ATTR income;
DIR < 2 1;
MCARLO ITER 10000;
GO;
```

Summarized in Tables 3.1 and 3.2, the findings support the *a priori* hypothesis that men have a higher income than women: *ESS* is effect strength for sensitivity (0 = chance, 100 = perfect classification); *ESP* is effect strength for predictive value (0 = chance, 100 = errorless prediction); and the confidence for estimated exact generalized $p < 0.05$ is > 99.999%.

As was hypothesized, men had a significantly higher income than women, but the classification accuracy (*ESS* = 11.7) and prognostic accuracy (*ESP* = 13.3) of the model were relatively weak. While the model made accurate overall classifications (83.6%) and point predictions (80.4%) for females, the corresponding performance for males was poor (28.3% and 32.9%, respectively).

The ACT heuristic involves computing subsets of confusion tables including observations scoring successively further from the model decision threshold, thereby increasing the reliability of and thus the discriminability between class categories. A conceptually related method commonly used in personality research involves limiting the sample to observations having relatively extreme scores on a measured factor, thus increasing the reliability of group designations based on the measured factor.⁵⁻⁷

Table 3.1: *Confirmatory UniODA Model Performance: Discriminating Gender using Income*

<i>Model Sensitivity</i>						
Actual Gender	N	Number Correctly Classified	Sensitivity (% Accuracy)	ESS	p <	
1 (Male)	92	26	28.3	11.7	0.040	
2 (Female)	324	271	83.6			
<i>Model Predictive Value (PV)</i>						
<i>UniODA Model</i>		PV (% of Correct Predictions)				
Income	Predicted Gender	N	Correct Predictions	ESP		
≤40	2 (Female)	337	80.4	13.3		
>40	1 (Male)	79	32.9			

Table 3.2: Confusion Table for Confirmatory UniODA Model Discriminating Gender using INCOME

		Patient Predicted Gender		
		Male	Female	
Patient Actual Gender	Male	26	66	28.3%
	Female	53	271	83.6%
		32.9%	80.4%	

Here the procedure begins by dropping patients in the two attribute levels in the middle of the scale (i.e., <\$40,000 and <\$50,000) from the sample, and then assessing the resulting model performance. Table 3.3 shows that results ($ESS = 13.3$; $ESP = 19.7$) were similar to findings achieved for the total sample, although model sensitivity and predictive value for females increased marginally. A total of 288 patients were included in the ACT, representing 69.2% of the total sample.

Table 3.3: ACT Confusion Table for UniODA Model *Excluding* Patients Scoring <\$40,000 and <\$50,000

		Patient Predicted Gender		
		Male	Female	
Patient Actual Gender	Male	13	44	22.8%
	Female	22	209	90.5%
		37.1%	82.6%	

The next ACT increases the reliability of group membership further by dropping patients scoring at the response levels next-closest to the midpoint: here, <\$30,000 and <\$60,000. Table 3.4 reveals little change: $ESS = 9.5$, $ESP = 22.2$. A total of 222 patients were included in the ACT, 53.4% of the total sample.

Table 3.4: ACT Confusion Table for UniODA Model Also *Excluding* Patients Scoring <\$30,000 and <\$60,000

		Patient Predicted Gender		
		Male	Female	
Patient Actual Gender	Male	7	39	15.2%
	Female	10	166	94.3%
		41.2%	81.0%	

The final ACT increases the reliability of group membership to the most extreme possible level, by including only the most extreme response categories (<\$10,000 and >\$60,000): as seen in Table 3.5, the model classification performance diminished (*ESS* = 3 .6, *ESP* = 15.0). A total of 105 patients were included in the ACT, 25.2% of the total sample. Thus, although there is statistical support for the *a priori* hypotheses that men have a greater income than women, the effect is very weak and no evidence suggests that increasing the precision of income measurement will improve *ESS* or *ESP*. In fact, model performance degraded when only the most extreme income scores were used in analysis.

Table 3.5: ACT Confusion Table for UniODA Model *Including* Patients Scoring <\$10,000 and >\$60,000

		Patient Predicted Gender		
		Male	Female	
Patient Actual Gender	Male	2	25	7.4%
	Female	3	75	96.2%
		40.0%	75.0%	

Predicting Mental Focus via GHA: Data were abstracted with permission from a computer log containing 297 sequential daily entries made by an anonymous patient with fibromyalgia (FM) using an intelligent health diary. The diary collects and analyzes *N*-of-1 (single-case) information in order to provide individual FM patients with prospective alerts about upcoming bad and good symptom periods. Increasing experience teaches patients to interact with the dairy in an individually-tailored manner, and generated alerts become increasingly sensitive as more patient data are obtained. Classifications are most accurate when the class variable and attributes are stable over time. Because lowest accuracy levels are expected under conditions in which antecedent attributes (e.g., weather) change randomly, a lower-bound estimate of the accuracy of the UniODA model (i.e., of system alerts) was obtained vis-à-vis a simulation study that randomly assigned the patient's daily data into either a training (*N* = 164) or a hold-out validity (*N* = 133) sample. It was necessary to ensure there were a sufficient number of responses in every rating category for both samples. Accordingly, the total of 297 responses included two focus ratings of 10% and 18 of 20% that were combined with 24 focus ratings of 30% to construct a new lower-end category, focus \leq 30%. And, seven focus ratings of 90% were combined with 22 focus ratings of 80% to construct a new higher-end category, focus \geq 80%. Not presented here, sensitivity analysis indicated that disentangling the polar ends of this scale primarily served to increase *p*.

The present study tested the exploratory alternative hypothesis that mental focus is predicted by atmospheric pressure. The ordered attribute was atmospheric pressure (500 mb GHA) assessed on a ratio scale (GHA), and the ordered class variable was patient's 10-point Likert-type rating of the percent of maximum possible mental focus available in the prior 24 hours (focus).⁸⁻¹⁰ Focus and GHA were synchronized by recording date. For the GHA *training* data (train.dat): *mean* = 5,575; *SD* = 164; *median* = 5,565; *skewness* = -0.11; *kurtosis* = -0.46; and *CV* = 3.0. For the GHA *hold-out* data (holdout.dat): *mean* = 5,575; *SD* = 151; *median* = 5,568; *skewness* = 0.13; *kurtosis* = -0.82; and *CV* = 2.7. For the focus *training* data (train.dat): *mean* = 6.45; *SD* = 1.47; *median* = 6; *skewness* = 0.10; *kurtosis* = -0.88; and *CV* = 22.7. And, for

focus *hold-out* data (*holdout.dat*): *mean* = 6.32; *SD* = 1.33; *median* = 6; *skewness* = 0.03; *kurtosis* = -0.66; and *CV* = 21.1. The exploratory hypothesis was tested using the following UniODA and MegaODA syntax:

```
OPEN train.dat;
OUTPUT example.out;
VARS focus GHA;
CLASS focus;
ATTR GHA;
MCARLO ITER 500;
HOLDOUT holdout.dat;
GO;
```

Inspection of the UniODA model reveals a non-linear model reflecting local regression in low- and high-symptom categories (see Chapter 5, Reliability Analysis). Considered over the symptom domain, low atmospheric pressure ($\leq 5,668$ 500 mb GHA) was associated with high symptom levels (30% to 50% of maximum mental focus), and higher pressure ($> 5,668$ 500 mb GHA) with low symptom levels (60% to 80% of maximum mental focus). As seen in Table 3.6 the *ESS* and *ESP* statistics suggest moderate accuracy that was statistically reliable: based on 300 MC experiments completed in 1.21 CPU hours by UniODA software running on a 3 GHz Intel Pentium D microcomputer, estimated exact $p < 1 / 300 < 0.0022$ (confidence for generalized $p < 0.05 > 99.99\%$) for the non-directional alternative hypothesis that GHA predicts mental focus. Model predictive values are weak for predicted ratings $\leq 80\%$ of maximum; strong for predicted ratings $\leq 30\%$ of maximum; and moderate for the other categories. Model sensitivities are weak for classified ratings $\leq 50\%$ and $\leq 70\%$ of maximum; moderate for ratings $\leq 40\%$ of maximum; and strong for the other categories—notably (and fortunately in the present context) for the scale extremes.

Table 3.6: UniODA Model *Training* Performance: Predicting Mental Focus using GHA

<i>Model Sensitivity</i>					
Actual Symptom Level	N	Number Correctly Classified	Sensitivity (% Accuracy)	ESS	p <
3 (30% of Maximum)	24	16	66.7	27.0	0.0033
4 (40% of Maximum)	18	5	27.8		
5 (50% of Maximum)	38	4	10.5		
6 (60% of Maximum)	39	20	51.3		
7 (70% of Maximum)	26	4	15.4		
8 (80% of Maximum)	19	12	63.2		

Model Predictive Value (PV)

<u>UniODA Model</u>		PV (% of Correct Predictions)		
GHA	Predicted Mental Focus	N	Correct Predictions)	ESP
≤ 5496	6 (60% of Maximum)	51	39.2	28.7
≤ 5636	8 (80% of Maximum)	53	22.6	
≤ 5668	7 (70% of Maximum)	9	44.4	
≤ 5692	5 (50% of Maximum)	9	44.4	
≤ 5760	4 (40% of Maximum)	16	31.2	
> 5760	3 (30% of Maximum)	26	61.5	

Table 3.7 summarizes findings when the UniODA model identified for the training sample was used to classify the observations in the hold-out validity sample. Results indicate statistically significant support ($p < 0.01$) for the *a priori* hypothesis that the training model using GHA to predict mental focus will cross-generalize to hold-out sample. However, the *ESS* and *ESP* statistics both indicate that the effect is relatively weak. The only effects that cross-generalized strongly from the training analysis to the hold-out analysis were the accurate classifications of the patient's best (80% of maximum) and worst (30% of maximum) days for mental focus, and the accurate point predictions made into the worst days for mental focus (30% of maximum).

Table 3.7: UniODA Model *Hold-Out* Performance: Predicting Mental Focus using GHA

<i>Model Sensitivity</i>						
Actual Symptom Level	N	Number Correctly Classified	Sensitivity (% Accuracy)	ESS	p <	
3 (30% of Maximum)	20	14	70.0	16.6	0.01	
4 (40% of Maximum)	18	3	16.7			
5 (50% of Maximum)	31	3	9.7			
6 (60% of Maximum)	30	11	36.7			
7 (70% of Maximum)	24	0	0			
8 (80% of Maximum)	10	5	50.0			

<i>Model Predictive Value (PV)</i>						
<i>UniODA Model</i>			PV (% of Correct Predictions)			
GHA	Predicted Mental Focus	N	Correct Predictions	ESP		
≤5496	6 (60% of Maximum)	45	24.4	19.8		
≤5636	8 (80% of Maximum)	45	11.1			
≤5668	7 (70% of Maximum)	5	0			
≤5692	5 (50% of Maximum)	4	75.0			
≤5760	4 (40% of Maximum)	12	25.0			
>5760	3 (30% of Maximum)	22	63.6			

All possible ACTs were examined next. As seen in Table 3.8, the more extreme the mental focus ratings (i.e., the class categories), the stronger the sensitivity and the predictive value of the model. Should a scoring strategy (i.e., a class variable category definition) reflected in one of the ACT models be selected for use in study design or funding proposal development, then the hold-out estimates of effect strength should be used in statistical power analysis (discussed next).

Adaptability

UniODA is a highly adaptable algorithm, and Chapters 4 (categorical attributes) and 5 (ordered attributes) demonstrate how UniODA identifies superior solutions in a diverse palate of applications that otherwise require a small army of legacy statistical methods (none of which explicitly identify optimal solutions) in order to be addressed. In contrast to suboptimal models, UniODA explicitly maximizes the (weighted) classification accuracy (whether or not accuracy is normed against chance) that is achieved for a sample, and UniODA models are often more parsimonious, particularly in applications involving multicategorical class variables or attributes.¹¹⁻¹³ UniODA sometimes identifies simple linear relationships: for example, a model predicting gender on the basis of serum testosterone level typically identifies an optimal threshold

Table 3.8: All Possible Confusion Tables for GHA and Mental Focus Example:
Training ($N = 164$) and Hold-Out ($N = 133$) Analyses

<u>Actual</u>	<u>Predicted</u> Mental Focus					
	30%	40%	50%	60%	70%	80%
30%	16	1	2	0	1	4
	14	3	0	0	1	2
40%	4	5	0	4	0	5
	4	3	0	6	0	5
50%	6	6	4	12	3	7
	4	1	3	11	2	10
60%	0	2	2	20	1	14
	0	4	0	11	2	13
70%	0	0	1	10	4	11
	0	1	1	12	0	10
80%	0	2	0	5	0	12
	0	0	0	5	0	5

<u>Actual</u>	<u>Predicted</u> Mental Focus		<u>Actual</u>	<u>Predicted</u> Mental Focus		<u>Actual</u>	<u>Predicted</u> Mental Focus	
	30-50	60-80		30-40	70-80		30	80
30-50	44	36	30-40	26	10	30	16	4
	32	37		24	8		14	2
60-80	7	77	70-80	2	15	80	0	12
	6	58		2	24		0	5
ESP=54.4 ESP=45.3			ESP=63.4 ESP=57.5			ESP=75.0 ESP=71.3		
ESS=46.7 ESS=37.0			ESS=64.5 ESS=63.2			ESS=80.0 ESS=87.5		

Note: **Bold** entries are for **training** model, other entries are for hold-out model. Training ACT excluding the 50 and 60 categories classified $N = 53$ (32.3% of the sample), and excluding categories 40-70 classified $N = 32$ (38.7% of the sample). Hold-out ACT excluding categories 50 and 60 classified $N = 58$ (43.6% of the sample), and excluding categories 40-70 classified $N = 32$ (38.7% of the sample).

that most males within a sample will exceed, and most females within a sample will not exceed. However, UniODA is also used to model complex, dynamic, non-linear applications such as occur in the analysis of Markov processes, turnover tables, group dynamics networks, or protein-bonding geometries.¹¹ Chapter 5 (Reliability Analysis) reviews recent investigations of inter-rater and inter-device reliability of emergency medicine triage data, and of the paradoxical confounding that occurs when triage data are combined for multiple pairs of raters, that involve the use of nonlinear UniODA models. Indeed, novometric theory—a set of ODA algorithms that explicitly identifies the globally optimal (maximum-accuracy) model for every unique combination of hypothesis, data geometry, and sample—suggests that for conditions involving adequate statistical power it is *likely* that *most* classical phenomena are *best* modeled using nonlinear UniODA models (see Chapter 12).

This is a remarkable domain of capability for a simple algorithm that explicitly ties statistical hypotheses to empirical data without requiring data to conform with any pre-conceived notions regarding underlying parent distributions, the functional form of model residuals, and so forth.^{14,15} Nevertheless, the adaptability of the UniODA algorithm does not end here, with what may be described as “*analysis by theoretical design*”. Sometimes scientists “get lucky,” observing a phenomenon that at first seems to be a curiosity, but that subsequent investigation clarifies as a discovery having scientific and/or commercial

merit: what may be described as “*analysis by serendipity*”. The highly adaptable UniODA algorithm is well-suited for modeling linear and nonlinear phenomena, whether pre-conceived or stumbled-upon.

As an example¹⁶ of how UniODA can be used to model a new empirical phenomenon, consider McDonald’s¹⁷ description of the analytic question underlying data used here (Table 3.9): “The biological question was whether protein polymorphisms (class = 0) would have generally lower or higher F_{ST} values than anonymous DNA polymorphisms (class = 1)”.

Table 3.9: Data Sorted by F_{ST}

Class	F_{ST}
1	-0.006
1	-0.005
0	-0.005
0	-0.002
1	0.003
0	0.004
0	0.006
0	0.015
0	0.016
0	0.016
0	0.024
0	0.041
0	0.044
0	0.049
1	0.053
0	0.058
0	0.066
1	0.095
1	0.160
0	0.163

Comparing the classes via the Kruskal-Wallace test¹⁷: “For the example data, the mean rank for DNA (class 1) is 10.08 and the mean rank for protein (class 0) is 10.68, $H = 0.043$, there is 1 degree of freedom, and the p value is 0.84. The null hypothesis that the F_{ST} of DNA and protein polymorphisms have the same mean ranks is not rejected.” A complementary conclusion is obtained by conducting bivariate UniODA to compare the two classes. For an application having an *ordered* attribute, both the exploratory and the confirmatory bivariate UniODA model have the following ordinal form: if subject’s score \leq (or \geq) threshold then predict class = 0; otherwise predict class = 1. Exploratory UniODA identifies the specific combination of threshold and direction that explicitly maximizes classification accuracy normed against chance, and confirmatory UniODA finds the threshold (or evaluates a specified threshold) for an *a priori* specified direction, and assesses the level of classification accuracy normed against chance. Here the exploratory UniODA model was: if $F_{ST} \leq 0.0035$ then predict class = 1; otherwise predict class = 0. While this model was not statistically reliable ($p < 0.55$) due to weak statistical power, the obtained $ESS = 35.7$ represents a moderate effect.

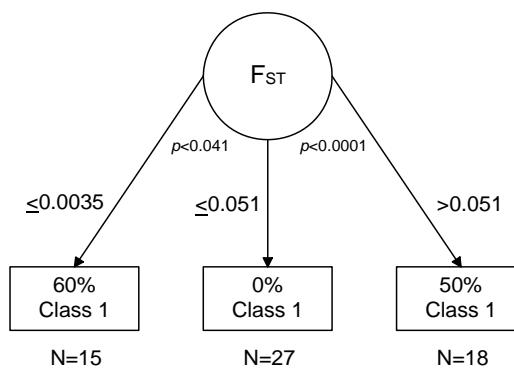
However, using the ODA paradigm any structural hypothesis can be tested using an explicitly optimized (i.e., yielding maximum-accuracy for the sample) statistical classification model.^{18,19} In the present study for example, the biological question can be restated as whether protein polymorphisms (class = 0) generally have *either lower or higher* F_{ST} values compared to anonymous DNA polymorphisms (class = 1). Parsing of the data in the context of this new analytic question is illustrated in Table 3.10. As seen, for this new question the threshold values are defined as the values above which (at the low end, strata C) and beneath which (at the high end, strata A) no members of class = 1 are found. By definition, observations between the upper and lower thresholds (strata B) are thus all members of class = 0.

Table 3.10: Parsed Data Sorted by F_{ST}

Class	F_{ST}	
1	-0.006	
1	-0.005	
0	A	-0.005
0		-0.002
1	0.003	
<hr/>		
0	0.004	
0	0.006	
0	0.015	
0	B	0.016
0		0.016
0	0.024	
0	0.041	
0	0.044	
0	0.049	
<hr/>		
1	0.053	
0	0.058	
0	C	0.066
1	0.095	
1	0.160	
0	0.163	

Manual identification and evaluation of these thresholds via UniODA is straightforward. First, test the directional hypothesis that class = 1 observations have lower F_{ST} values than class = 0 observations: this identifies the threshold value $[(0.003 + 0.004) / 2] = 0.0035$, that separates strata A from strata B and C. Second, test the directional hypothesis that class = 1 observations have greater F_{ST} values than class = 0 observations: this identifies the threshold $[(0.049 + 0.053) / 2] = 0.051$ that separates strata C from strata A and B. However, attributable to less-than-perfect classification accuracy in concert with the small sample and associated inadequate statistical power, these analyses were unrevealing: exact p 's < 0.29 and 0.44, ESS 's = 35.7 and 28.6, respectively. However, for identical results in a sample *three times as large*, the globally optimal UniODA model (Chapter 12) seen in Figure 3.1, for which $ESS = 64.3$ (a relatively strong effect), would emerge:

Figure 3.1: Globally Optimal UniODA Model Addressing the Restated Biological Question for $3N$ Sample



Nevertheless, the sample is what it is, so what can be done in light of the paucity of statistical power? Cross-classification findings obtained by applying the nonlinear UniODA model (if Strata = A or C then predict class = 1; if Strata = B then predict class = 0) to classify the sample are given in Table 3-11.

Table 3.11: Confusion Table for the Nonlinear UniODA Model Applied to Classify the Sample

		Predicted Class	
		0	1
Actual	0	9	0
	1	5	6

UniODA applied to this result indicates a relatively strong ($ESS = 54.6$) and statistically reliable (exact directional $p < 0.0120$, non-directional $p < 0.0141$) effect. The ESS for this approach is lower than was obtained (64.3) by the model in Figure 3.1, because the latter separates subjects having extreme F_{ST} values into two groups, rather than combining them into a single group as in Table 3.11. Combining different groups can induce Simpson's paradox: the effect here is to reduce the estimated effect.²⁰ The importance of replication in such research can't be overstated.

Instrumentation

How can a researcher efficiently determine if established methods and/or instruments are available for measuring a given construct, if/how scores on the measures have been validated, how the measuring instruments may be obtained, and how to score the instruments and interpret the scores? To obtain measurement information of this ilk, the ODA laboratory employs a resource known as the *Health and Psychosocial Instruments* (HaPI) data base, which provides information on an enormous, diverse collection of measurement instruments used in many fields.²¹ Our colleague and friend (an expert in measurement) Fred Bryant, Ph.D., kindly wrote the following invited description of the HaPI data base.

"Available in libraries around the world, the HaPI contains a wide range of records that provide comprehensive, accurate information about measurement instruments across diverse disciplines and professions. The HaPI database identifies measurement tools relevant for investigations and explorations in the fields of medicine, nursing, public health, psychology, social work, communication, sociology, and organizational behavior/human resources. The reference librarian at any academic library should be able to tell you how to obtain access to the HaPI. If this is not the case, or if you don't have access to an academic library, then here is the link to the website for Behavioral Measurement Database Services (BMDS)--the company that produces the HaPI data base: <http://bmdshapi.com/>

This extensive collection of records describes measurement instruments from peer-reviewed scholarly journals, books, technical reports, and test publishers' catalogs. HaPI contains information about a variety of measurement instruments, such as Questionnaires, Surveys, Interviews, Tests, Checklists, Rating Scales, and Coding Schemes. HaPI records provide descriptive information about instruments, such as Title, Acronym, Authors, Language, Index Terms, and References. Some records also include Abstracts, Sample Items, Number of Questions, Subscales, Reliability, and other information. HaPI provides detailed abstracts summarizing measurement instruments, as well as citations to (a) the study that originally developed a given measure (primary source), as well as (b) later published research studies in the behavioral and medical sciences literature that have used this same measure (secondary sources).

The records in the HaPI data base also indicate when a given primary source includes a copy of the actual items for an instrument in the original article; and they indicate when an electronic copy of the particular instrument and scoring instructions can be obtained (free of charge) from the producer of the data base.²² If an electronic or hard copy of a particular instrument is not available directly from BMDS, then users can ask BMDS to contact the original author of the instrument to secure a copy of the instrument, along with scoring instructions and permission for use; BMDS will then attempt to contact the

author and will send the instrument and scoring instructions to the user free of charge, if the author agrees to provide them.

Unlike the HaPI data base, “open access” resources that consist of electronic copies of measurement instruments will necessarily be limited in the number of instruments they include. This is because open access data bases of instruments cannot include any instruments for which journal publishers hold the copyright. These copyrighted instruments include any measure whose items have been printed in a table or appendix in the original published article. In contrast, the HaPI data base is not limited to instruments for which journal publishers do not own the copyright, but rather includes a wider range of access to information about measurement instruments (more than 200,000 records) than any other source in the world. And the HaPI data base is continually being updated with new primary and secondary sources. Why restrict yourself only to instruments that are immediately available for download from open access data bases? When searching for the most appropriate measurement instruments, quick access should not take priority over thoroughness and a careful and systematic scrutiny of all measurement options.

Nor do open access data bases allow you to find and track the use of specific measurements instruments in the published research literature over time. Through its provision of secondary sources, HaPI enables users to find published studies that have used particular instruments in a wide variety of different settings, with different populations, in different languages, in abbreviated or modified forms, in different research designs, and with a host of different independent or dependent variables. Having access to this information enables researchers to make better decisions about which instruments to use in their particular research projects.

HaPI also offers a powerful and versatile Boolean search capability, with which users can combine key terms and descriptors to locate the ideal measurement instrument to meet their particular needs. The HaPI database can be used to find alternative versions of existing instruments (e.g., original vs. short forms; state vs. trait forms; adult vs. child versions), available translations of instruments, and multiple scoring frameworks for a given instrument. The flexibility of combinatory searching (e.g., optimism ‘and’ trait ‘and’ English ‘and’ children) offers far greater power and efficiency in finding measurement tools than do open access data bases or printed sources. With this extraordinary tool, you can find, for example, a short form measure of a particular construct validated for use with children in a school setting in a given language—a task that would be like searching for a needle in a hay stack using any other data base. In my opinion, based on 40 years of experience in conducting social research, there is no better resource than the Health and Psychosocial Instruments (HaPI) data base for finding the most suitable measurement instruments to use in research.”

Statistical Power Analysis

A mission-critical task in planning a study is determining the number of observations to include in the study. The first approach to determining the appropriate sample size for the study involves conducting a statistical power analysis, and two different methods for performing this analysis are described below. However, for large samples this approach will produce models that are not parsimonious (Table 2.2). Therefore, the second approach is the heuristic of establishing an *a priori* minimum strata N (all UniODA and CTA models terminate in two or more endpoints that represent sample strata) corresponding to some percentage of the study total N (all ODA software allows specification of the minimum strata sample size). For example, by setting the minimum strata N to be 10% of the overall study sample, resulting statistical models will be less susceptible to overfitting, and thus are more likely to cross-generalize to independent random samples having comparable or smaller sample size.

In our laboratory the smallest multi-observation samples ($N = 8$) for which a statistically reliable effect was obtained by UniODA both involved a confirmatory hypotheses with an ordered attribute and a binary class variable. The studies were a prospective evaluation of cost savings in asthma management occurring after treatment²³ and an analysis of the relationship between oil leverage and gross domestic product.²⁴ For an ordered attribute the smallest sample size for which it is theoretically possible to obtain a generalized (“per-comparison”) *two-tailed* $p < 0.05$ is $N = 8$ —half ($N = 4$) from each of the two compared groups²⁵ (perfect classification yields $p < 0.029$), and the smallest sample size for which it is theoretically possible to obtain generalized *one-tailed* $p < 0.05$ is $N = 6$ —again half ($N = 3$) from each of two compared

groups²⁶ (perfect classification yields $p < 0.05$). Such minute samples offer minute statistical power.

Parametric Approach

To facilitate comparison with parametric GLM and ML methods^{14,15} the first exploration of statistical power (1- β) obtained by UniODA assumed a normally-distributed attribute.²⁷ Here the direction of classification assignment is fixed, corresponding to the evaluation of a one-tailed null hypothesis, and class membership is balanced ($n_0 = n_1$).²⁶ Tables 3.12 and 3.13 summarize the results of Monte Carlo analyses for $\alpha = 0.01$ and 0.05 , in which 100,000 sets of normal deviates were generated for each of two classes.²⁸ Power is represented by the proportion of rejections of the null hypothesis for each experiment at the corresponding level of α . Its complement β (Type II error) is the proportion of acceptances of these hypotheses. Table entries correspond to 14 different values of Cohen's d , increasing values of which cause increasing separation between the observations comprising each class.²⁹ These Tables may be used to estimate the sample size required to achieve a specific level of power for a given effect. For example, in a study with $n_0 = n_1 = 60$, for power of at least 90%, an effect having Cohen's $d = 0.8$ is required for $\alpha = 0.01$ (Table 3.12), versus Cohen's $d = 0.7$ for $\alpha = 0.05$ (Table 3.13).

Table 3.12: Statistical Power of Optimal Discrimination with One Normal Attribute and Two Balanced Classes, Obtained from Monte Carlo Analysis: One-Tailed Test, $\alpha = 0.01$

Cohen's d															
n_1	.2	.3	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2	1.3	1.4	1.5	
5	.008	.010	.015	.021	.028	.036	.045	.059	.074	.091	.112	.136	.161	.192	
6	.002	.004	.005	.008	.011	.015	.022	.029	.037	.051	.063	.080	.101	.125	
7	.009	.014	.020	.029	.039	.054	.073	.093	.121	.151	.188	.231	.273	.322	
8	.021	.031	.044	.062	.084	.112	.147	.188	.232	.285	.340	.398	.461	.524	
9	.008	.013	.019	.029	.043	.062	.084	.114	.147	.191	.241	.294	.353	.416	
10	.016	.025	.038	.055	.079	.110	.147	.195	.246	.306	.371	.441	.509	.583	
11	.006	.011	.017	.028	.043	.064	.087	.123	.165	.213	.273	.338	.408	.477	
12	.011	.019	.031	.049	.071	.103	.143	.193	.252	.317	.391	.469	.546	.621	
13	.019	.031	.047	.074	.109	.152	.209	.273	.345	.424	.508	.588	.666	.739	
14	.028	.044	.070	.103	.150	.209	.275	.354	.437	.523	.610	.691	.760	.825	
15	.013	.023	.037	.061	.094	.137	.191	.262	.336	.423	.509	.599	.681	.756	
16	.019	.033	.053	.084	.130	.186	.249	.336	.421	.514	.606	.689	.767	.831	
17	.026	.044	.073	.114	.166	.236	.316	.404	.502	.596	.689	.767	.834	.889	
18	.014	.024	.042	.070	.110	.163	.230	.312	.404	.504	.599	.690	.772	.840	
19	.017	.033	.056	.092	.142	.207	.287	.377	.481	.586	.678	.763	.835	.889	
20	.024	.042	.073	.118	.178	.254	.346	.449	.554	.654	.745	.822	.884	.927	
21	.030	.055	.093	.145	.218	.304	.404	.513	.620	.717	.800	.867	.917	.951	
22	.016	.031	.055	.097	.153	.225	.314	.419	.530	.633	.735	.813	.878	.925	
23	.021	.039	.071	.117	.181	.270	.368	.481	.592	.696	.788	.859	.913	.949	
24	.026	.049	.088	.144	.220	.315	.421	.540	.652	.749	.833	.895	.939	.966	
25	.032	.059	.104	.170	.257	.361	.476	.592	.703	.798	.872	.923	.956	.978	
30	.038	.075	.133	.215	.322	.448	.579	.699	.803	.882	.934	.967	.984	.993	
40	.044	.092	.172	.286	.430	.579	.719	.830	.911	.958	.982	.993	.998	.999	
50	.046	.103	.199	.337	.505	.672	.809	.902	.958	.984	.995	.998	.999		
60	.067	.153	.289	.466	.651	.803	.908	.965	.989	.997	.999				
70	.091	.202	.370	.570	.754	.885	.956	.988	.997	.999					
80	.079	.193	.371	.583	.778	.904	.967	.991	.998	.999					
90	.096	.232	.437	.661	.840	.943	.985	.997	.999						
100	.112	.269	.498	.727	.888	.966	.993	.998	.999						

Table 3.13: Statistical Power of Optimal Discrimination with One Normal Attribute and Two Balanced Classes, Obtained from Monte Carlo Analysis: One-Tailed Test, $\alpha = 0.05$

n_1	Cohen's d													
	.2	.3	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2	1.3	1.4	1.5
4	.024	.031	.040	.053	.063	.082	.098	.117	.139	.166	.194	.222	.256	.293
5	.066	.086	.107	.138	.166	.200	.237	.284	.325	.372	.424	.474	.524	.576
6	.025	.034	.047	.063	.084	.105	.135	.166	.201	.246	.285	.337	.388	.438
7	.052	.070	.093	.124	.156	.198	.244	.294	.348	.406	.465	.530	.587	.643
8	.085	.115	.149	.194	.241	.297	.360	.422	.487	.557	.618	.680	.739	.787
9	.038	.054	.077	.104	.142	.185	.234	.293	.354	.421	.490	.558	.628	.687
10	.059	.083	.117	.157	.209	.265	.328	.401	.472	.545	.616	.687	.749	.802
11	.081	.116	.162	.214	.274	.346	.421	.500	.578	.652	.721	.784	.839	.881
12	.110	.152	.209	.271	.341	.425	.508	.588	.670	.739	.804	.857	.897	.931
13	.056	.084	.122	.171	.235	.303	.384	.467	.551	.635	.711	.778	.838	.884
14	.074	.111	.159	.219	.291	.374	.458	.552	.636	.713	.786	.847	.892	.928
15	.093	.140	.196	.270	.348	.440	.531	.626	.706	.784	.844	.894	.931	.957
16	.117	.171	.236	.319	.413	.505	.596	.690	.770	.837	.890	.928	.955	.973
17	.064	.101	.153	.217	.297	.390	.488	.583	.677	.757	.829	.884	.926	.956
18	.081	.125	.184	.261	.350	.449	.549	.647	.737	.815	.875	.920	.951	.973
19	.096	.148	.219	.304	.400	.501	.606	.704	.789	.857	.907	.944	.969	.983
20	.113	.174	.252	.346	.451	.560	.661	.754	.833	.892	.934	.962	.980	.990
21	.131	.199	.287	.387	.498	.608	.707	.798	.868	.918	.952	.973	.987	.993
22	.079	.132	.201	.290	.394	.502	.612	.716	.804	.872	.924	.955	.977	.988
23	.092	.150	.231	.326	.434	.553	.664	.763	.843	.903	.943	.969	.984	.992
24	.108	.172	.260	.365	.481	.598	.706	.801	.873	.924	.959	.979	.990	.995
25	.121	.194	.285	.400	.523	.640	.749	.832	.898	.944	.970	.985	.993	.997
30	.125	.206	.315	.440	.570	.698	.805	.883	.936	.967	.985	.994	.997	.999
40	.185	.304	.451	.604	.746	.853	.924	.966	.986	.995	.998	.999		
50	.167	.294	.452	.627	.775	.885	.949	.980	.994	.998	.999			
60	.201	.359	.546	.723	.858	.937	.979	.994	.998	.999				
70	.233	.410	.614	.790	.908	.967	.990	.998	.999					
80	.255	.454	.670	.840	.938	.983	.995	.999						
90	.275	.493	.714	.877	.959	.990	.998	.999						
100	.294	.524	.756	.904	.973	.994	.999							

Table 3.14 represents the values of ESS corresponding to tabled values of Cohen's d , computed in the same manner as the parameters in the previous tables ($n_1 = n_0$). Table 3.14 enables an investigator to evaluate statistical power in experimental designs, using the more appropriate and intuitive ESS index of effect strength. Colors are used to indicate the effect strength represented by ESS values in Table 3.14: *blue* indicates relatively weak effects; *green* indicates moderate effects; *red* indicates relatively strong effects; and *purple* indicates strong effects. The “saw-tooth” behavior of the power values with increasing n_1 is due to the discrete nature of both the sample and the exact UniODA distributions of *optimal values*.²⁷

Table 3.14 may be used to obtain Cohen's d on the basis of expected ESS . Imagine an investigator wishes to evaluate statistical power for a confirmatory (one-tailed) problem involving a single binary class variable, a single ordered attribute, and $n_1 = n_2 = 100$. For this problem, imagine an effect of moderate strength is anticipated. Since a moderate-strength ESS is defined as $0.25 \leq ESS < 0.50$, the lower end of this domain indicates Cohen's $d = 0.5$, and the upper end of the domain indicates Cohen's $d = 1.3$.

Table 3.14: Median *ESS* of Optimal Discrimination with One Normal Attribute and Two Balanced Classes,
Obtained from Monte Carlo Analysis, Corresponding to Cohen's *d* Values: One-Tailed UniODA

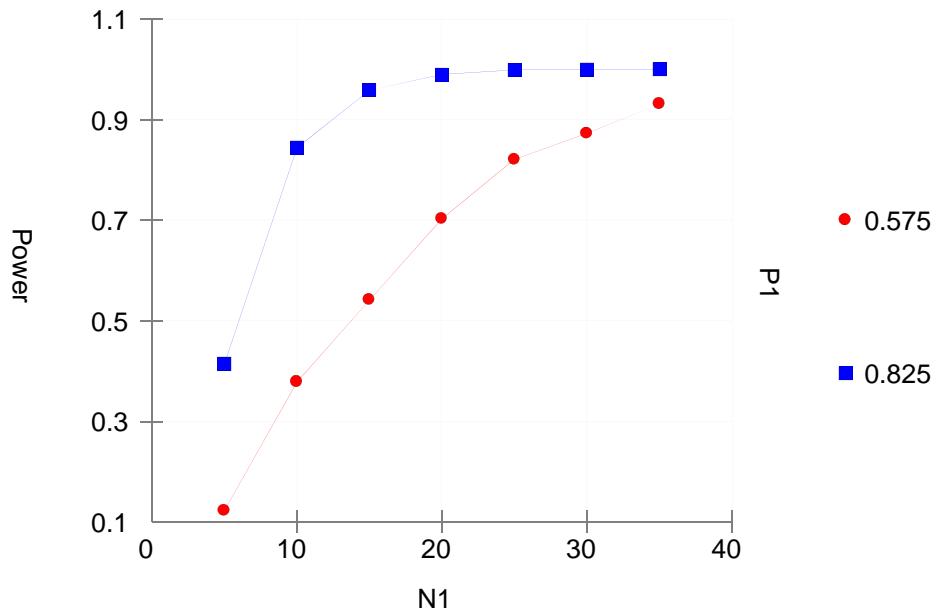
n_1	Cohen's <i>d</i>														
	.2	.3	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2	1.3	1.4	1.5	
5	.400	.400	.400	.400	.400	.600	.600	.600	.600	.600	.600	.600	.800	.800	
6	.333	.333	.333	.500	.500	.500	.500	.500	.500	.667	.667	.667	.667	.667	
7	.286	.286	.429	.429	.429	.429	.571	.571	.571	.571	.571	.714	.714	.714	
8	.250	.375	.375	.375	.375	.500	.500	.500	.500	.625	.625	.625	.625	.750	
9	.333	.333	.333	.333	.444	.444	.444	.556	.556	.556	.556	.667	.667	.667	
10	.300	.300	.300	.400	.400	.400	.500	.500	.500	.600	.600	.600	.700	.700	
11	.273	.273	.364	.364	.364	.455	.455	.545	.545	.545	.545	.636	.636	.636	
12	.250	.333	.333	.333	.417	.417	.500	.500	.500	.583	.583	.583	.667	.667	
13	.231	.308	.308	.385	.385	.462	.462	.462	.538	.538	.615	.615	.615	.692	
14	.286	.286	.286	.357	.357	.429	.429	.500	.500	.571	.571	.643	.643	.643	
15	.267	.267	.333	.333	.400	.400	.467	.467	.533	.533	.600	.600	.667	.667	
16	.250	.250	.312	.312	.375	.437	.437	.500	.500	.562	.562	.625	.625	.625	
17	.235	.294	.294	.353	.353	.412	.412	.471	.529	.529	.588	.588	.647	.647	
18	.222	.278	.278	.333	.389	.389	.444	.444	.500	.556	.556	.611	.611	.667	
19	.211	.263	.316	.316	.368	.421	.421	.474	.474	.526	.579	.579	.632	.632	
20	.250	.250	.300	.350	.350	.400	.450	.450	.500	.500	.550	.600	.600	.650	
21	.238	.238	.286	.333	.333	.381	.429	.476	.476	.524	.571	.571	.619	.619	
22	.227	.273	.273	.318	.364	.409	.409	.455	.500	.500	.545	.590	.590	.636	
23	.217	.261	.304	.304	.348	.391	.435	.435	.478	.522	.565	.565	.609	.652	
24	.208	.250	.292	.333	.333	.375	.417	.458	.500	.500	.542	.583	.583	.625	
25	.200	.240	.280	.320	.360	.400	.400	.440	.480	.520	.560	.560	.600	.640	
30	.200	.233	.267	.300	.333	.367	.400	.433	.467	.500	.533	.567	.600	.633	
40	.175	.225	.250	.275	.325	.350	.400	.425	.450	.500	.525	.550	.575	.600	
50	.180	.200	.240	.280	.320	.340	.380	.420	.440	.480	.520	.540	.580	.600	
60	.167	.200	.233	.267	.300	.333	.367	.400	.433	.467	.500	.533	.567	.600	
70	.157	.186	.229	.257	.300	.329	.371	.400	.443	.471	.500	.529	.557	.586	
80	.150	.187	.225	.262	.287	.325	.362	.400	.437	.462	.500	.525	.562	.587	
90	.144	.178	.222	.256	.289	.322	.356	.400	.433	.467	.500	.522	.556	.589	
100	.140	.180	.210	.250	.290	.320	.360	.390	.430	.460	.490	.520	.550	.580	

Exact Minimum Precision Approach

Discussed in Chapter 2, pre-processing data is a statistically motivated method of creating a measurement scale free of paradoxical confounding. For example, using *N*-of-1 methods to assess if each participant in a weight reduction program achieved a statistically reliable level of weight loss between baseline and end-of-study, and using the processed variable as a class variable (e.g., contrasting groups of participants who did versus didn't achieve statistically reliable personal weight loss over the course of the study) or as an attribute. This Chapter discussed how for some applications it is necessary to manually create a threshold on an attribute, for example when a minimum level of accuracy is required for a class category³⁰, or when a new nonlinear structure is being modeled.¹⁶ In all these cases the number of levels of the transformed variable is reduced to two—the minimum possible number of measurement levels (i.e., the minimum precision level) possible for a variable. Recall from Chapter 2 that when a class variable and attribute both reach their theoretical minimum level of measurement precision (i.e., a *binary* application) the exact statistical distributions for one- and two-tailed UniODA and Fisher's exact test converge. Therefore, a minimum-precision estimate of statistical power for an ODA model can be obtained via analysis of Fisher's exact test. All UniODA and CTA models terminate in endpoints that represent sample strata, and the most granular comparison involves comparing two strata. Three examples demonstrate this power analysis methodology.

Figure 3.2 presents power curves for a simulation involving one confirmatory bivariate test with generalized $p < 0.05$ ($\beta = 0.20$) for a balanced application (i.e., the sample sizes in the two groups are identical), for an effect strength in the middle of the *moderate* range ($p_1 = 0.575$, $p_0 = 0.20$; $ESS = 37.5$; shown in red) and in the middle of the *relatively strong* range ($p_1 = 0.825$, $p_0 = 0.20$; $ESS = 62.5$; shown in blue), for a design having between 5 and 35 observations in each class category.

Figure 3-2: Statistical Power Simulation, One Bivariate Test, Generalized $p < 0.05$



For a moderate effect 25 observations per class category (i.e., per group) are needed to obtain 80% power for a confirmatory test with generalized $p < 0.05$, and 32 observations per class category are needed to obtain 90% power. For a relatively strong effect these sample sizes are 9 and 12, respectively.

The second simulation involves four exploratory bivariate tests of statistical hypotheses (i.e., four UniODA analyses) with experimentwise $p < 0.05$ ensured using a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure. For the four effects that are being evaluated, the effect with lowest observed p (i.e., the “most statistically significant” effect) has a Sidak criterion of generalized $p < 0.01275$ ($\beta = 0.051$) for a balanced application (Chapter 10 has a worked example of this procedure). Figure 3.3 presents associated power curves for this application and effect strengths in the middle of the *moderate* range (shown in red) or in the middle of the *relatively strong* range (shown in blue), for a design having between 10 and 45 observations in each class category. As seen, for the effect with the smallest p -value, a moderate effect necessitates 35 observations per class category to obtain 80% power for an exploratory test with experimentwise $p < 0.05$, and 45 observations per class category are needed for 90% power (70 and 90 observations in all, respectively). For a relatively strong effect these sample sizes are 13 (total of 26) and 17 (total of 34), respectively.

The final simulation repeats the prior simulation, but fixes the sample at 45 observations for one class category in order to demonstrate the effect of imbalance in class category sample sizes on statistical power. In Figure 3.4, for the effect with the smallest p -value, a moderate effect (red) needs a total of $45 + 35 = 80$ observations to reach 80% power for an exploratory test and experimentwise $p < 0.05$, and a total of $45 + 45 = 90$ observations for 90% power. For a relatively strong effect (blue) both sample sizes are $45 + 10 = 55$ (the strength of cross-generalizability diminishes as endpoint samples become very small).

Figure 3-3: Statistical Power Simulation, Four Bivariate Tests, Experimentwise $p < 0.05$

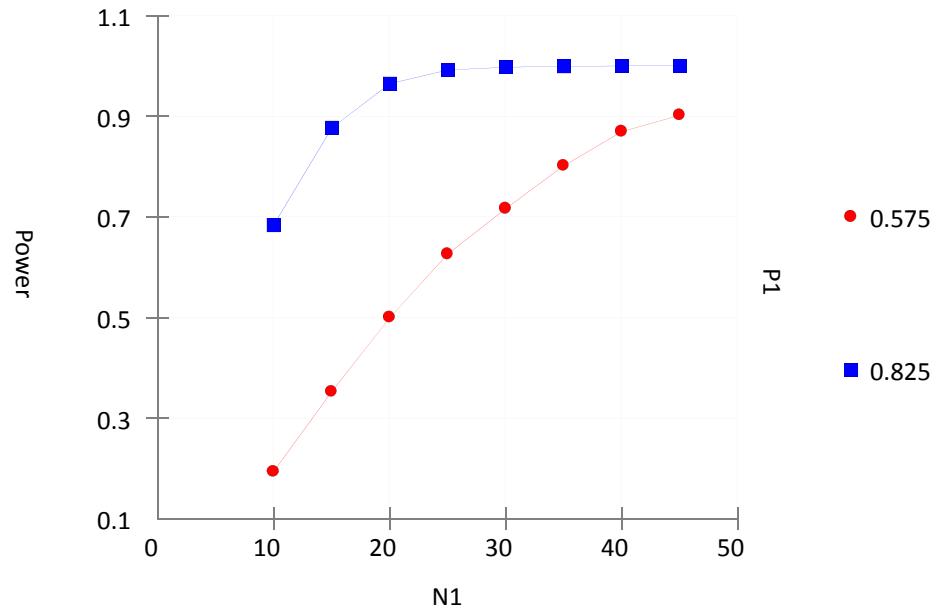
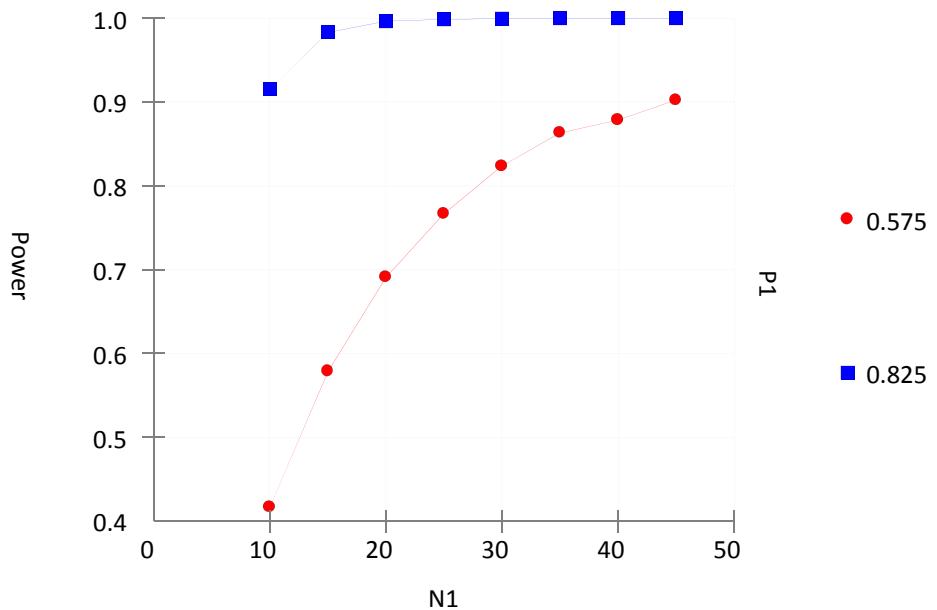


Figure 3-4: Statistical Power Simulation, Four Bivariate Tests,
Experimentwise $p < 0.05$, Imbalanced Sample Sizes

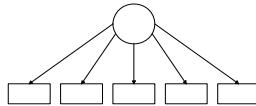


Model Geometry and Sample Size

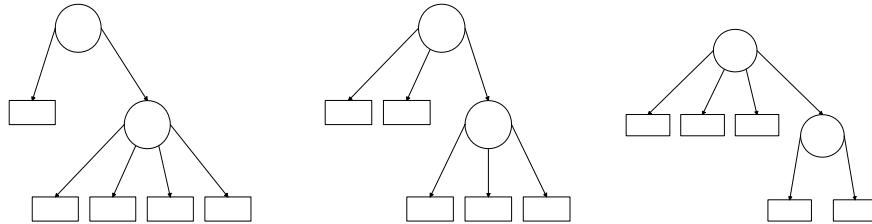
Having determined a conservative (minimum precision) sample size for a model endpoint, the next task is determining the overall sample size for the ODA model being simulated. Since the number of endpoints in a model is equal to the total number of UniODA analyses conducted plus one, in the present example four tests of statistical hypotheses implies five model endpoints. Computing the overall N is complicated by the geometry of the ODA model having five endpoints that emerges in analysis. Illustrated in Figure 3.5, nine different model geometries—circles represent model nodes, and rectangles represent model endpoints—identify five sample strata (symmetric models are not unique structures).

Figure 3.5: Possible ODA Model Geometries Involving Five Endpoints

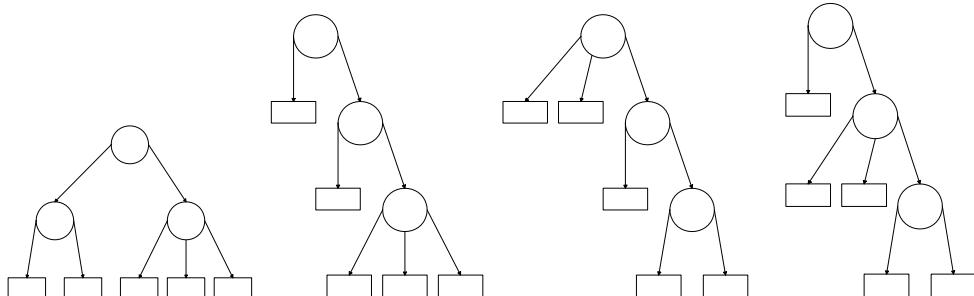
One Node, Five Strata ODA Model



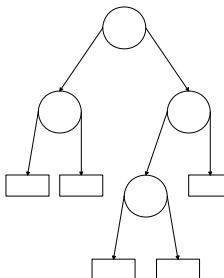
Two Node, Five Strata ODA Models



Three Node, Five Strata ODA Models



Four Node, Five Strata ODA Model



Computing the minimum sample size for the overall model is a dynamic problem: even if the estimated effect strength (moderate or relatively strong) is held constant across model nodes (for models having more than one node), as the remaining number of tests of statistical hypotheses decreases, the N required to attain the targeted power also decreases due to increasing Sidak- criterion p -value (Chapters 2, 10). In most published multi-node, multi-strata optimal models, endpoint N generally decreases with increasing depth of the endpoint in the model: thus, the minimum overall sample size increases in multi-node models as the number of endpoints close to the root node increases.

As of this writing, no algorithmic solution to general power analysis has been developed. Rather, each statistical power analysis is a specific simulation. Discussed in Chapter 12, adequate statistical power is axiomatic to appropriate statistical analysis. A pragmatic approach of assessing whether an ODA model has adequate statistical power involves computing power for every statistical test that is conducted. A pragmatic approach of ensuring that an ODA model has adequate statistical power involves continuing to collect data until a model converges to a solution having adequate statistical power. Fortunately, globally optimal models (see Chapter 12) rarely involve many nodes, endpoints, or tests of statistical hypotheses, so statistical power analysis is greatly simplified.

As a means of demonstrating the importance of parsimony in statistical modeling, the reader is encouraged to identify all of the ODA model geometries that exist by adding one more endpoint than is already expanded in Figure 3.5.

Data Set Design

The mantra “*garbage in, garbage out*” was chanted incessantly by teacher and student alike when computers were first becoming available in public and private research institutions in the United States. In the early days of computing the operators and users were technically savvy—many were engineers, so the *absolute importance of data quality* was universally understood. Today it is estimated that 90% of the spreadsheets and associated data bases of all S&P 500 corporations contain errors.³¹ In our experience spanning four decades of research in academia, this estimate also represents the proportion of data sets we receive that contained data and measurement errors.

As an analytic methodology ODA is extremely particular about data because explicitly optimal computational operations are performed using actual data values, rather than using primarily (massaged) summary statistics as is the case for parametric methods. Fortunately, because ODA models are expressed in the original units of measurement (rather than in terms of model coefficients), erroneous data values are often easy to identify as one examines analysis output from ODA software. If ever one is attempting to run a program using UniODA, MegaODA, or CTA software—and the analysis fails to execute, if the control syntax is correct then the problem *is* in the data (for troubleshooting ODA software see Appendix D).

As a means of promoting professional data file quality standards, ODA software only reads ASCII files, also known as “flat” or “text” files. Because using an error-free data set is crucial for a successful and a reproducible research study, Dr.’s Fred Bryant and Patrick Harrison wrote an invited article³² reproduced below for the eJournal *Optimal Data Analysis*, concerning how to construct an ASCII data file that is ready for statistical analysis.

Abstract: UniODA and CTA software require an ASCII (unformatted text) file as input data. Arguably the most difficult task an operator faces in conducting analyses is converting the original data file from (a) whatever software package was used to enter the data, into (b) an ASCII file for analysis. This article first highlights critical issues concerning missing data, variable labels, and variable types that users must address in order to convert their data into an ASCII file for analysis using ODA software. Specific steps needed to convert a data set from its original file-type into a space-delimited ASCII file are then discussed. The process of converting data into ASCII files for use as input data is illustrated for three leading statistical software packages: SPSS, SAS, and STATISTICA.

ODA-based software such as UniODA and CTA requires only a few easy-to-understand control commands to conduct powerful, accurate non-linear modeling. Ironically, given the simplicity of ODA software syntax, the most difficult task for users to complete in conducting analyses is often the creation

of an ASCII data file for ODA software to analyze. But, with a little forethought and attention to detail, this critical task is simple and straightforward.

Researchers often use statistics software such as the *Statistical Package for the Social Sciences* (SPSS), *Statistical Analysis System* (SAS), or *STATISTICA*, for example, to enter and process raw data. In contrast, ODA software requires a delimited ASCII file (i.e., an unformatted text file) as input data. Typically, spaces, tabs or commas are used to separate the data entries in an ASCII file. These delimiters enable ODA software to read input data in free form without requiring operators to specify formatting, making it easier to implement analysis. This paper explains how to create ASCII files which use spaces as delimiters separating data entries.

Initial Issues to Resolve

Before converting data from the original file type into an ASCII file, three basic issues must be resolved: (1) the handling of missing data; (2) the creation of variable labels having usable format; and (3) the transformation of alphabetic string variables into a quantified form which CTA can analyze.

Missing Data: An important point to keep in mind is the fact that ODA software will not treat a blank space in an ASCII data file as a missing value, but instead will skip over a blank space and use the next numeric value in the data set to stand in for the missing value. Therefore, *before* converting their data into an ASCII file for ODA analysis, users must replace any system-missing “blank” values (e.g., a “.” in SPSS or SAS) with a numeric value designating a missing response. This conversion can easily be accomplished by using the original statistical software package to recode each variable so as to replace whatever values were originally used to indicate missing data with a chosen, global missing-value indicator (e.g., -999). Note, however, that this task must be done *before* one converts the original data set into an ASCII file.

Observed values are often missing for at least some cases on one or more variables in a data set. Researchers typically use blank spaces to indicate missing values in the original data file or use one or more specific numeric values (e.g., -9, 99, 999) to designate missing responses for different variables. Before creating an ASCII file, users must first convert the value(s) which are being used to represent missing values for each variable—such as blank spaces or numeric values—into a *single numeric value* to be used to designate *all missing values* for every variable in the data set.

Yarnold and Soltysik¹¹ emphasized the importance of specifying missing values when creating ASCII files for analysis by ODA: “A very important point that one cannot overlook is that *all system missing data must be changed to a specified missing numeric value* prior to analysis via ODA” (p. 55). A popular choice for a universal missing-value indicator is -999. Of course, using -999 as a marker for missing values assumes that this is not a valid response for any variables included in the data set. If the value -999 is a valid response for any variable in the data set, then a value which is not a valid response should be selected instead.

Saving Variable Labels in Usable Format: In UniODA and CTA software, variable labels may be no longer than eight characters, so users of general-purpose statistical software, such as SPSS, SAS or STATISTICA, should ensure that the variable names consist of no more than eight characters before exporting their data set for ODA. Otherwise, ODA software will be unable to read the names of variables having more than eight characters and will produce an error message. Alternatively, users can export variable names longer than eight characters, and use a text file editor to truncate variable names to a maximum of eight characters in the exported ASCII file. However, changing variable names to a maximum of eight characters in the original source data file makes it easier to verify the accuracy of the exported ASCII data file, when comparing descriptive statistics from the original and exported data sets.

Transforming Alphabetic Variables into Numeric Form: Some variables in the original data set may consist of alphabetic or “string” values, rather than numbers. Examples of such alphabetic variables are gender (e.g., “male” or “female”), religious affiliation (e.g., “Catholic,” “Protestant,” “Jewish,” “Buddhist,” “Muslim,” or “none”), or ethnicity (e.g., “White,” “Black,” “Hispanic,” “Asian,” or “Other”). To analyze such string variables in ODA, users must first recode each alphabetic value of the variable into a numeric value (e.g., “female” = 0, “male” = 1; “White” = 1, “Black” = 2, “Hispanic” = 3, “Asian” = 4, “Other” = 5). After converting all alphabetic values to numeric values, users should save the data file,

carefully noting which variable is the class variable, which attributes are ordered, and which attributes are categorical (string variables typically reflect the latter). *The ASCII data file to be analyzed by UniODA and CTA software should contain only numeric values, delimited by spaces.*

Additional Considerations: To streamline the analysis as much as possible, it is recommended that users delete any unnecessary variables from the original data file before exporting the data as an ASCII file. Thus, users should eliminate any variable which is neither a class variable nor an attribute in the analysis (e.g., ID number). Excluding unused variables will make the ASCII data file as small as possible and will minimize the time required to obtain final CTA results. Alternatively, one can choose to export only a subset of the variables from the full data set when constructing an ASCII data set.

For the UniODA and CTA programs to access the data file, users should assign the exported ASCII file a name that is no more than 8 characters, followed by a dot and 3 characters (e.g., CTA_RUN1.DAT). Finally, if applicable, users should cut and paste the variable labels from the first line of the exported ASCII data file into a separate ASCII file, to serve as part of the VARIABLES control command in the syntax file for the UniODA or CTA programs.

Exporting a Source Data File

Driven by both pull-down menus and syntax, **SPSS** is perhaps the most commonly used statistical program in academia. Imagine an SPSS data file (ODAdata.sav) containing 20 variables, 12 of which are to be exported into a space-delimited ASCII data file (ASCIIdat.dat) for analysis by ODA software.

After opening *ODAdata.sav* in SPSS, the SAVE TRANSLATE command may be used to convert the active SPSS data file into a space-delimited ASCII data file, as follows:

```
SAVE TRANSLATE OUTFILE='C:\Documents and Settings\localuser\Desktop\ASCIIdat.dat'  
/TYPE=CSV  
/MAP  
/REPLACE  
/FIELDNAMES  
/TEXTOPTIONS DELIMITER=' '  
/KEEP=v1 v2 v3 v4 v5 v6 v7  
v8 v9 v10 v11 v12.
```

Here, the subcommand: *OUTFILE='C:\ Documents and Settings\localuser\Desktop\ ASCIIdat.dat'* is used to instruct SPSS to save the exported ASCII file (which we have named ASCIIdat.dat) to the Windows desktop. Users should alter this subcommand to specify the correct path to the folder on their hard drive where they wish to save the ASCII file.

The */TYPE=CSV* subcommand specifies that the exported data file will be in text-file (ASCII) format.

The */MAP* subcommand displays in the SPSS output a list of the variables and the number of cases exported in the ASCII data file.

The */REPLACE* subcommand gives SPSS permission to overwrite an existing ASCII file of the same name. Because the default is not to overwrite an existing ASCII file, SAVE TRANSLATE will not overwrite an existing file without an explicit REPLACE subcommand. If users wanted to prevent the possibility of overwriting an existing ASCII file, then they could omit the */REPLACE* subcommand.

The */FIELDNAMES* subcommand is used to instruct SPSS to write variable names separated by a delimiter (see below) in the first row of the ASCII data file. As noted earlier, before implementing CTA, users should cut and paste the variable labels from the first line of the exported ASCII data file into a separate ASCII file, to serve as part of the VARIABLES control command in the syntax file for ODA programs.

The */TEXTOPTIONS DELIMITER=' '* subcommand instructs SPSS to employ a blank space (empty column) to delimit or separate variable names and data values in the exported ASCII file.

The */KEEP* subcommand may be used to export to the ASCII data file either: (a) all of the variables in the active SPSS data (by specifying */KEEP=ALL*); or (b) a subset of the variables in the in the

active SPSS data (by specifying `/KEEP=<variable names separated by spaces>`, as in the example above). Also, the `/KEEP` subcommand may be used to change the order in which the variables appear in the ASCII data file, by using a particular order to list these variables in the `/KEEP` subcommand.

SAS is another popular statistical software program, which is widely used in business analytic settings for operations research, data mining, and predictive modeling. Imagine a SAS data file `ODAdata.dat` containing 20 variables, 12 of which are to be exported into a space-delimited ASCII data file (`ASCIIidat.dat`) for analysis by ODA software.

In SAS, the `PUT` command may be used to convert the active SAS data file into an ASCII data file, as follows:

```
DATA ODAdata2;
SET ODAdata;
FILE 'C:\Documents and Settings\localuser\Desktop\ASCIIidat.dat';
PUT v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12;
RUN;
```

The `DATA` command begins the process of data restructuring in SAS. The command `DATA ODAdata2` instructs SAS not to overwrite the active SAS data set (i.e., `ODAdata`), but to give the new, restructured data set the name `ODAdata2` (later changed to `ASCIIidat.dat` using the `FILE` command).

The `SET` command reads all variables and observations from the SAS input data set.

The `FILE` command renames the restructured data set and writes the contents of the active data set to an external ASCII file.

The `PUT` command outputs the listed variables to the ASCII data specified in the `FILE` command.

The `RUN` command has SAS process the set of commands listed in the syntax file.

Note that the above SAS commands do *not* output the variable names to the first line of the ASCII data file. However, users can use the following commands to enter variable names on line 1 of the ASCII file (followed by space-delimited data):

```
DATA ODAdata2 (keep= v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12);
SET ODAdata;
FORMAT v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 10.6;
PROC EXPORT DATA=ODAdata2 OUTFILE ='C:\Documents and Settings\localuser\Desktop\ASCIIidat.dat'
DBMS=DLM REPLACE;
RUN;
```

Note that this method requires that a `FORMAT` command is used to prevent rounding of exported values, as indicated. The `FORMAT` command uses the value of "10.6" to tell the SAS program to allot a total of 10 spaces with 6 decimal points for each exported variable.

STATISTICA is another popular data analysis program, commonly used in healthcare, financial services, insurance, and consumer product industries. Imagine a STATISTICA data file named `ODAdata.sta` containing 20 variables, 12 of which are to be exported into a space-delimited ASCII data file (`ASCIIidat.dat`) for analysis by ODA software.

With STATISTICA the Windows drop-down menu may be used to export the active data set into a comma-delimited ASCII data file by first opening the data file (`ODAdata.sta`). To export only 12 of the 20 variables in the data set, first delete the variables which will not be exported. Then click on "Save As..." under the File command on the top left-hand side of the main Data Editor screen. In the Save Data As window, click on the down arrow to the right of the "Save as type" box, and select "Text file (*.txt)." Users should then specify the name of the ASCII output file in the "File name box" using the *.txt extension (e.g., `ODAdata.txt`) and the location in which to save this file, and click on the Save command. STATISTICA will respond by warning users that the data file "may contain features that will be lost when saved as text" and asking them if they "want to export the Spreadsheet in this format." Users should click on "Yes."

STATISTICA will also display a smaller window giving users the option of specifying the particular “field separator” to use as a delimiter in the ASCII file (users should click on “Space”), and writing the variable names separated by the delimiter in the first row of the ASCII data file. Finally, to create the ASCII space-delimited data file, users should click on the Save command. As noted earlier, before implementing ODA software, users should cut and paste the variable labels from the first line of the exported ASCII data file into a separate ASCII file, to serve as part of the VARIABLES command in the syntax file for ODA programs.

Quality Assurance

Before running UniODA or CTA, it is essential first to check the accuracy of the ASCII data file by comparing it to the original data set. To check the accuracy of the exported ASCII file in relation to the original (source) data file, follow the following six steps.

First, run descriptive statistics on the variables in the original data file, using the statistical software employed to enter the raw data originally (e.g., SAS, SPSS, etc.). Second, replace all blanks and other missing data values with a valid value, such as -999, which will be used to designate missing values in ODA software. Third, export the original (source) data file into an ASCII format, making sure to export the variable labels on the first line of the ASCII data file. We recommend exporting data as a space-delimited ASCII file. Fourth, import the exported ASCII file back into the original statistical software (e.g., in SPSS use the “Read Text Data” option beneath the “File” drop-down menu). Fifth, after importing the exported ASCII data file, use the statistical software to designate values of -999 as missing. Finally, run descriptive statistics, and compare the results for equivalence with the initial set of descriptive statistics.

If the first and second sets of descriptive statistics are identical, then one can be confident of having accurately exported the original data set into an ASCII format. If the two sets of descriptive statistics based on the original and imported ASCII data do not match perfectly, then pinpoint the source of the problem and repeat the process until perfect correspondence is obtained. Although there are countless mistakes one can make when converting an original data set into a space-delimited ASCII data file, the most common errors include forgetting to change blank data entries to a specific missing numeric value, forgetting to convert alphabetic string variables into numbers, exporting variable names that exceed the maximum of 8 characters, and failing to export all of the variables from the original data set that one wishes to analyze.

Running ODA Software Using DOS Prompt

There are two methods for running ODA software. The first method is an intuitive, easy-to-use integrated editing application for text files created by Alan Phillips, known as the “Programmers File Editor” or PFE. With PFE an operator can write and edit ODA scripts and data files, execute analyses, and view outputs from multiple runs. The second method for running ODA software is using the command line editor in the MS-DOS command prompt window (thoroughly described in Windows documentation). When in the MS-DOS command prompt window, execute the ODA program by using the command: ODA *filename*. If a UniODA or MegaODA program named “ex51.cmd” is ready to run, simply enter: ODA ex51.cmd. Or, if a CTA program named “ex51.cmd” is ready to run, enter: CTA ex51.cmd. Which system an operator prefers to use is a matter of personal preference: among the authors Paul prefers command prompt window, and Rob prefers PFE. Detailed explanation of how to use PFE to run ODA software¹¹ is available elsewhere, so discussion below suggests an efficient means of operating ODA software using the DOS command prompt.

For ODA analysis a master folder was created on the c: drive, named ODA. In the master folder are four files: copies of MegaODA and CTA executable software and corresponding “template” programs. Also in the master folder is a directory for every different major research area: AIDS, Cancer, Ecology, and so forth. Inside the subdirectory I have a folder for every specific project. Table 3.15 provides an example of this directory structure for one research area.

Table 3.15: Directory Structure for ODA Analysis Folder

```

C:\ODA
    UniODA.EXE
    CTA.EXE
    UniODA.PGM
    CTA.PGM
    C:\ODA\AIDS
        C:\ODA\AIDS\PCP
            C:\ODA\AIDS\PCP\SURVIVAL
                UniODA.EXE
                CTA.EXE
                UniODA.PGM
                CTA.PGM
                PCPSURV.DAT
            C:\ODA\AIDS\PCP\TREATMENT
        C:\ODA\CANCER
    
```

Every project folder (e.g., AIDS PCP SURVIVAL) has at least five files: UniODA (UniODA.EXE) and CTA (CTA.EXE) *software*, the *data set* for the project (PCPSURV.DAT), and UniODA (UniODA.PGM) and CTA (CTA.PGM) *programs*. While the data set is unique, the software and programs are common to every analysis. The four common files are kept in the ODA master folder. When a new study is initiated, a folder is created for the project, and the four common programs are copied from the master folder. The UniODA (and MegaODA) and CTA programs are templates: each contains all of the commands for the respective software. If a command needs to be altered, then it is altered. If a command is not used then it can be disabled with a leading asterisk (see Appendix A) or deleted. In this way no commands are accidentally omitted or misspecified. When conducting analysis it is extremely simple to remember the names of the UniODA and CTA program files using this convention. The first author uses “P.OUT” as the name of every program output file, as a means of essentially eliminating the need to recall program executable, program and output file names. Using the DOS “hot key” (up arrow) to automate analysis processing, with practice and a quad-core microcomputer it is relatively simple and a bit fun to manage four concurrent analyses.

Treatment of Missing Data

Would it be a good idea to impute data crucial for proper flight-worthiness evaluation of a spacecraft, or for safe operation analysis for a nuclear reactor? Is science less worthy of restriction to actual fact? The policy of the ODA laboratory is to avoid data imputation in empirical applications.

To some extent the use of non-linear maximum-accuracy methods (CTA models) mitigates the role of missing data. CTA only drops observations missing data on the attributes used in their actual classification—that is, on the branch of the tree model on which the observation resides. In contrast, linear models drop all observations having any missing data on any of the attributes used in the model. In addition, the CTA approach enables a researcher to determine which attributes having too many missing values to be included in the model might be able to improve model performance, if sufficient *authentic* (not simulated) data are collected (see Chapter 12).

If artificial data are to be used, then it is crucial to perform a sensitivity analysis to determine how variations of the artificial data influence/change results as random error of different structure and magnitude is added to the artificial values. Comparison of residual values for authentic and artificial data might be revealing, and should also be performed.

The Role of Residuals

Stated succinctly, the analysis of model residual values focuses on assessing the *invalidity* of estimated Type I error rates for parametric methods, versus on deducing ways to improve the *validity* of maximum-accuracy methods.

Analysis of residuals—the difference between the predicted and actual values of *observations* with respect to the attribute (dependent variable)—is important in assessing the validity of parametric statistical methods.^{14,15} In particular, a crucial assumption is that the residuals are normally distributed: failing this assumption threatens the validity of Type I error estimates. This is an important limitation because residuals are greatest for absolutely extreme values of the attribute for general linear model-based methods (e.g., ordinary least-squares regression), and for the smallest class category for maximum-likelihood-based methods (e.g., logistic regression analysis).^{13,33}

Residual values are also an integral part of structural equation modeling (SEM), which uses a “fitting function” to obtain parameter estimates that minimize the size of the residuals between the elements of the *observed* covariance matrix based on the set of measured variables being analyzed (S) and the elements of the *predicted* covariance matrix implied by the parameter estimates in the model (Σ). The most commonly used method of estimation in SEM is maximum-likelihood, which finds parameter estimates that maximize the likelihood that the fitted residuals ($S - \Sigma$) are due to chance.^{34,35} In SEM, the overall size of residuals is used to assess a structural model’s goodness-of-fit to the data (e.g., via a chi-square value testing the statistical significance of the size of fitted residuals, or descriptive fit indices reflecting the average size of residuals); individual elements in the matrix of fitted residuals can be inspected to identify specific relationships between measured variables that the model explains poorly; and the model can be modified to include additional estimated parameters to improve its fit to the data. Note that this statistical method does NOT address the residuals associated with individual *observations*.

In contrast, in the optimal (maximum-accuracy) data analysis (ODA) paradigm no distributional assumptions underlie theoretical distributions of optima, so the validity of the Type I error rate is never in doubt (see Chapter 2). However, in the ODA paradigm the analysis of residuals is arguably the most important aspect of an analysis—in terms of assessing ways in which prediction of observations’ actual class categories can be improved. Residuals tell one what remains to be explained. The ultimate objective is to eliminate all such errors—that is, to correctly classify all of the observations in the sample.

Compared to suboptimal methods, the ability of residuals to indicate ways to improve statistical models is a major benefit of the UniODA and CTA algorithms. Discussed in detail in Chapter 12, model endpoints that are homogeneous (i.e., all or most observations in the endpoint have the same class category) are well explained and leave little room for further model improvement. In contrast, endpoints that are heterogeneous are poorly explained, and leave much room for improvement. When an endpoint has a large N , and is heterogeneous, it is an appropriate area in which to work to improve overall model performance—and thus understanding of the phenomenon. It also is clear in the latter case that none of the measured attributes used to find the model thus far will help in this regard—or else those attributes would be included in the model. Clues to the characteristic nature of the observations in the targeted strata are garnered by content analysis of attributes (and their cut-point values) defining the endpoint. This not only paves the way toward fastest improvement in model performance, but it indicates what the subject inclusion criteria for future research should be (i.e., observations classified into the targeted endpoint), thereby providing “bread crumbs” pointing the way to new conceiving new theory, new hypotheses, and new attributes to study. In a word, residuals lie at the heart of the matter.

Reporting Analytic Findings

As the number of researchers using a new statistical method (and the province of disciplines their work represents) increases, opportunity for and likelihood of development of disparate traditions for reporting analytic findings also increases. Establishing minimum standards for reporting the analytic results acquired vis-à-vis the new method empowers researchers to understand fundamental statistical results of any study that uses the new method. Accordingly, the minimum information required to understand the structure and performance of an ODA model is considered here. Expressed in a nutshell, the statistical results of

empirical articles should generally unfold in two or three stages. If only the UniODA algorithm is used in the article, then the stages are the presentation and interpretation of descriptive statistics separately by class variable category, followed by the UniODA results. If an optimal multiattribute analysis is also conducted, then the third stage of the statistical results is the presentation and interpretation of structure and performance of the multiattribute ODA model.

Descriptive Statistics

It is customary for the results section of an empirical investigation to begin by presenting and interpreting descriptive statistics for study attributes, separately by class variable category (e.g., for the class variable gender, the categories are male and female). Every descriptive dimension is a potential attribute, and assessing whether or not a particular dimension discriminates the class variable requires variability in the dimension that (hopefully) occurs within and across studies. Summary statistics generally *shouldn't* be presented for *any* pooled class categories, except in the context of understanding paradoxical confounding that occurs in the application (for Simpson's Paradox see Chapters 2, 5, and 9). Statistical moments that are commonly and widely reported in the literature include the *mean*, *SD*, *CV*, *median*, *skewness*, and *kurtosis*. Although not directly relevant to ODA analyses, these indices are useful in characterizing the representativeness of a sample relative to samples of conceptually related investigations: this information constitutes data for meta-analytic research after a sufficient sample of findings accumulates. Information about distributions of class-variable-relevant descriptive dimensions occurring in observational (field) studies may help to estimate the expected limits of cross-generalizability for a specific statistical model.

UniODA Findings

Tables 3.1 and 3.2 and associated text provide an example of the minimum information that is needed to understand the classification performance achieved by a simple UniODA model involving a binary class variable and a single ordered attribute. Tables 3.6-3.8 demonstrate parallel information for a more complex application involving a multicategorical class variable. In Table 3.8 the grand confusion table gives results for training as the top value in the row, and results for hold-out as the bottom value in the row. For applications that (also) report LOO analysis (an estimate of potential cross-generalizability), the jackknife result should be entered beneath the training results, followed by the hold-out results. In cases in which a more compressed presentation of the findings is appropriate or necessary, for example word and/or table limitations, a template for compressed presentation of UniODA results that is used in the ODA laboratory is shown in Table 3.16 .

Table 3.16: Example of Compressed Presentation of UniODA Findings

Attribute	UniODA Model	N	% Class 1	p <	ESS	ESP
Categorical	If Attribute = (<i>Category List</i>) then predict <i>Class</i> = 0;	--	--	----	----	----
			Training Values LOO Values Hold-Out Values			
Ordered	Otherwise predict <i>Class</i> = 1		Training Values LOO Values Hold-Out Values			
	If <i>Attribute</i> ≤ <i>Threshold</i> then predict <i>Class</i> = 0;		Training Values LOO Values Hold-Out Values			
	Otherwise predict <i>Class</i> = 1		Training Values LOO Values Hold-Out Values			

As an example of this prescription regarding dissemination of descriptive statistics and UniODA results (particularly apropos for areas of science and/or journals that are new to legacy and/or optimal statistical methods), consider the following report of the statistical findings for a study assessing the comparative strength of three versatile knots widely used in big-game fishing, presented to a diverse audience of professionals in the sport fishing industry.³⁶

Knot breaking point was assessed using the following methodology. A Shimano spring scale accurate to $\frac{1}{2}$ pound (used to calibrate reel drags) was utilized to assess the pounds of force required to induce knot failure (a sliding pointer indicates maximum force). The free end of monofilament line was tied directly to the scale, the other end was the top shot on a Penn 113HN fishing reel with drag cinched closed, mounted on a reel seat and secured in a rod holder. On each trial, after tying a knot and confirming it was perfectly tied, the study engineer pulled the scale until the knot broke, recorded the force at knot failure, and reset the scale pointer. Pulls were steady and made using a 90-degree angle of attack, and knot testing order was determined by coin flip. The *Uni* and *San Diego* knots were both tied using seven turns: loops were tight with no overlap, cinching was steady with no frictional heating, and tag ends were cut at $\frac{1}{4}$ -inch.³⁷ The fishing line used was 30-50-pound-test P-Line monofilament.

Statistical analysis was used to compare the breaking strength of *Uni* versus *San Diego* knots tied in 30-, 40- and 50-pound-test monofilament line. Ten of both types of knot were tied in three classes of pound-test line, for a total of 60 knots. Table 3.17 summarizes the results of analyses for each of three different line pound-test categories.

Table 3.17: Knot Breaking Strength: Descriptive Statistics and UniODA Analysis Results

	30-Pound-Test Line		40-Pound-Test Line		50-Pound-Test Line	
Moment	Uni	San Diego	Uni	San Diego	Uni	San Diego
<i>N</i>	10	10	10	10	10	10
<i>Median</i>	24.5	27	33.5	34.5	41.5	41
<i>Mean</i>	26.3	27.6	33.2	36.0	39.7	41.0
<i>SEM</i>	1.75	1.38	1.28	1.54	2.58	0.82
95% CI, <i>Mean</i>	23 - 30	25 – 30	31 – 36	33 – 39	35 – 45	39 - 43
<i>SD</i>	5.54	4.35	4.05	4.88	8.17	2.58
95% CI, <i>Knot</i>	15 – 37	19 – 36	25 – 41	26 – 46	23 – 56	36 - 46
<i>Minimum</i>	21	22	27	30	28	37
<i>Maximum</i>	38	38	39	44	49	45

UniODA Model	If knot breaks at ≤ 23.5 pounds, then predict <i>Uni</i> knot	If knot breaks at ≤ 29.5 pounds, then predict <i>Uni</i> knot	If knot breaks at ≤ 33.5 pounds, then predict <i>Uni</i> knot
<i>p</i> <	0.36, 0.18	0.77, 0.99	0.71, 0.99
<i>ESS</i>	40.0, 30.0	30.0, 30.0	30.0, 0.0
<i>ESP</i>	47.6, 33.0	58.8, 30.0	58.8, n/a

Actual Knot	Predicted Knot		Predicted Knot		Predicted Knot	
	Uni	San Diego	Uni	San Diego	Uni	San Diego
<i>Uni</i>	9	1	10	0	10	0
	8	2	7	3	10	0
<i>San Diego</i>	5	5	7	3	7	3
	5	5	7	3	10	0

The first row in the top section of Table 3.17 (*N*) gives the number of knots tied. The second row (*Median*) is the median number of pounds of force required to induce a knot failure. If observed breaking

strengths for a given knot/line combination are sorted from lowest to highest, then the median value is the number in the middle (mid-point) of the sorted list. Median values differ only slightly (1%-3%) for knots made using 40-50-pound-test line, and modestly (10%) for knots made using 30-pound-test line. More notable is the consistent failure of either purported “100% knot” to achieve line class: median values are 82%-90% of line rating for 30-pound-test; 84%-86% for 40-pound-test; and 82%-83% for 50-pound-test line.

The third row (*Mean*) provides the mean pounds of force required to induce knot failure. Consistent with the findings for medians, mean knot strength differed modestly (3% to 8%; the San Diego knot was consistently stronger), and means were notably lower than was the rated line strength (the deficit increased as line test-class increased): 88%-92% of rating for 30-pound-test; 83%-90% for 40-pound-test; and 79%-82% for 50-pound-test line.

The standard error of the mean (*SEM*) in row four is an estimate of the standard deviation of observed mean knot strength, and it is used to estimate the range within which the *mean knot strength can vary*, if this experiment is repeated. The expected mean variability is given in row five (95% CI, *Mean*): a 95% confidence interval for the mean is the range within which the mean knot strength is expected to fall 95% of the times that this experiment is repeated. Only for 30-pound-test line does the upper-bound of the expected range include line-class rating: upper-bounds were 90%-98% of rating for 40-pound-test line, and 86%-90% for 50-pound-test line. The lower bounds of the expected range were 77%-83% of the rating for 30-pound-test line, 78%-82% for 40-pound-test line, and 70%-78% for 50-pound-test line.

Row six (*SD*) is the standard deviation, a measure of degree of difference between mean and individual measurements of knot strength, used to estimate the range within which the *strength of individual knots can vary*, if this experiment is repeated. Expected variability of individual knots is provided in the seventh row (95% CI, *Knot*) as the range within which the strength of 95% of the individual knots is expected to fall if this experiment is repeated. Results parallel findings for means, but are more extreme. Upper bounds of the 95% confidence interval exceeded line-class rating of 30-pound-test line by 20%-23%, and 40-pound-test line by 1%-15%. The upper bound for Uni knots tied in 50-pound-test line exceeded line-class rating by 12%, but it was 92% of line-class rating for San Diego knots. Lower bounds were 50%-63% of line-class rating for 30-pound-test line, 62%-65% for 40-pound-test line, and 46%-72% for 50-pound-test line.

The eighth (*Minimum*) and ninth (*Maximum*) rows provide the strongest and weakest observed knot strengths, respectively. Values in these rows all were within the 95% confidence interval for knots in 40- and 50-pound-test line, but the strongest Uni and San Diego knots (38 pounds) tied in 30-pound-test line exceeded the upper 95% bound.

In Table 3.17 the first row in the middle section gives the UniODA model. The Uni knot is predicted to be weakest for all three line-class categories, and to break at levels significantly lower than rated line-class strength: model thresholds are 78% (30-pound line), 74% (40-pound line), and 67% (50-pound line) of the rated line-class strength.

The second row in the middle section gives two *p*-values. The bold *p* on the left is for the UniODA model used with the actual data—this is called the “training model.” The second *p* on the right is for a “leave-one-out” (LOO) validity analysis performed to estimate what *p* would be if the UniODA model was used to classify an independent random sample of knots tested in another study. As seen, differences between knots were not statistically reliable for training or LOO analysis (all *p*'s >> 0.05).

The third and fourth rows in the middle section also both give two accuracy measures: the bold values on the left are for the training analysis, and values on the right are for LOO analysis.

The *ESS* “accuracy” index (row three) measures the extent to which the UniODA model is able to differentiate the overall sample of knots: on this index 0 represents the level of predictive accuracy that is expected by chance (i.e., the effect of chance is “factored out” so that different models can be compared using a universal index), and 100 represents perfect prediction: models with $25 \leq ESS < 50$ are considered to be of moderate strength. As seen, all training models achieved moderate overall accuracy. In LOO analysis the overall accuracy declined for 30-pound class line; was stable for 40-pound class line; and was zero—the same level of accuracy that is expected by chance—for 50-pound class line.

The *ESP* “prediction” index (row four) measures the accuracy of the model in classifying the individual knots (i.e., predicting whether a given knot was a tied using a Uni or a San Diego): on this index 0 =

represents accuracy in making point predictions expected by chance, and 100 = perfectly accurate point predictions. As seen, all training models yielded moderate (30-pound class line) to relatively strong (the 40- and 50-pound line class categories) point prediction accuracy. In LOO analysis the point prediction accuracy declined to moderate levels for 30- and 40-pound pound class line, but was incomputable for 50-pound class line because no knots were predicted to be San Diego knots (this induces division by zero in the computational formula for *ESP*).

And, in Table 3.17 the bottom section gives the so-called “confusion table” that summarizes the classification performance of the UniODA model for each of the three line-class categories. In each table bold values (on top) are for the training analysis, the other values (on bottom) are for LOO analysis. As seen, for the 30-pound class line, in training analysis the accuracy for the ten actual Uni knots was $9 / (9 + 1) = 9 / 10 = 90\%$, versus $5 / (5 + 5) = 5 / 10 = 50\%$ (the level of accuracy expected by chance) for the actual San Diego knots. In LOO analysis accuracy for Uni knots fell to 80%, but it was stable (at a chance level) for San Diego knots. And, in training analysis the model was correct $9 / (9 + 5) = 9 / 14 = 64.3\%$ of the times that a knot was predicted to be a Uni, and $5 / (5 + 1) = 5 / 6 = 83.3\%$ of the times that a knot was predicted to be a San Diego. In LOO analysis these values diminished to 61.5% and 71.4%, respectively.

For 40-pound class line, in training analysis the accuracy for the ten actual Uni knots was perfect (100%), versus worse than expected by chance (30%) for the actual San Diego knots. In LOO analysis accuracy for Uni knots fell to 70%, but it was stable (at a worse-than-chance level) for San Diego knots. In training analysis the model was correct 58.8% of the times that a knot was predicted to be a Uni, and 100% of the times a knot was predicted to be a San Diego. In LOO analysis these values both diminished to 50% (chance).

Finally, for 50-pound class line, in training analysis the accuracy was exactly same as for 40-pound class line. In LOO analysis the accuracy for Uni knots remained perfect (100%), but accuracy fell to 0% for San Diego knots. In training analysis the model was correct 58.8% of the times that a knot was predicted to be a Uni, and 100% of the times a knot was predicted to be a San Diego. In LOO analysis accuracy fell to 50% (chance) for knots predicted to be a Uni, and no knots were predicted to be a San Diego.

It is not surprising to learn the strength of Uni and San Diego knots differs *moderately* (training analysis) and *unreliably* (*p* and LOO analysis). These knots are similar in configuration and the way they cinch down when tightened, and both are rated 100% as being knots, meaning they presumably will break at the breaking strength of the line—and thus won’t diminish the integrity of the rigging. Tied and fished by experts, both knots are extremely reliable. There are typically too few trials (and likely few anglers attempt) to detect the modest difference on a single fishing trip. However, the experimental comparison showed that the San Diego knot has moderately greater strength, so a sufficiently large sample of knots will reveal a statistically significant difference. Considered at the individual level, selecting the Uni knot may increase the number of large fish lost over the course of a lifetime for casual anglers, perhaps over the course of a season for active anglers. But for anglers hunting large, fast, powerful fish, if for food, competition, commerce, or record, it is illogical to lose advantage without cause—thus the San Diego is the knot of choice.

In contrast, it was surprising to observe how substantially lower than rated line-class the strength of both types of knots routinely fell. Rated line-class fundamentally influences angler decision-making on reel, drag setting, rod, hook and bait selection—and thus on the targetable species. For example, hooking a tuna fish on gear appropriate for much smaller and weaker species is a low-probability catch.

Replication/extension of this research is warranted that compares different monofilament brands, and collects data on a larger number of knots tied by multiple technical experts, so as to assess generalizability across brand and person. It also is important to measure or estimate actual targeted species fighting behavior: for example, head shakes and sudden acceleration both create force spikes ($F = m \times a$). This may be done via simulation or *in vivo*, for example by attaching a hand line to a spring scale. Such mission-critical information regarding anticipated operational force levels is necessary to ensure that appropriate equipment and rigging is selected.

Multiatribute ODA Findings

Regardless of whether derived by the general linear model (e.g., multiple regression analysis, analysis of variance) or maximum-likelihood (e.g., log-linear model, logistic regression analysis) statistical paradigm, established protocols exist for presenting the results of parametric methods.^{14,15} UniODA can explicitly maximize the (weighted) *ESS* or *Overall PAC* achieved by any parametric linear model (Chapters 6 and 7). Thus, when reporting an optimized suboptimal linear model, information on the nature and performance of the UniODA model is also presented.^{11,33,38}

Optimal multiatribute linear models, known as MultiODA models, don't require optimization vis-à-vis UniODA, because MultiODA models explicitly maximize the (weighted) *ESS* or *Overall PAC* achieved in the training sample (Chapter 8). Nevertheless, information concerning the structure and performance of a MultiODA model parallels that which is presented for optimized suboptimal multiatribute linear models.

Finally, for statistical models obtained via optimal non-linear classification-tree analysis (CTA) that explicitly maximizes (weighted) *ESS* or *Overall PAC* obtained for the training sample, omnibus performance is assessed in the same manner as UniODA model (Chapter 2). Protocols for presenting the results of CTA models are suggested in Chapters 10, 11, and 12.

Chapter 4

UniODA with Categorical Attributes

Chapter 4 illustrates how UniODA, a highly adaptable algorithm, identifies superior solutions in a diverse palate of applications involving categorical attributes, that otherwise require a small army of legacy statistical methods to address. Perhaps the most significant advantage of using ODA in statistical analysis, rather than a host of legacy methods, is that only ODA explicitly maximizes (weighted) classification accuracy and provides a forecasting model for every application. Not only do legacy methods fail to explicitly maximize forecasting accuracy, but many, such as chi-square, also fail to provide a forecasting model. No matter what the structure of a particular data configuration might be—for example, the number of class levels, attribute metrics, or class sample-size imbalances, classification performance of every ODA model is summarized using normed measures of model sensitivity (*ESS*) and predictive value (*ESP*) for which 0 represents the level of classification performance expected for the application by chance, and 100 represents perfect, errorless classification. No such intuitive, universal index can be used to compare the effect strength of different legacy statistical methods. Whereas legacy methods require the data to meet the assumptions underlying the method, with ODA distribution theory is exact for every design: ODA models must exactly conform to the data, rather than vice versa. Therefore, with ODA a *single methodology* may be *optimally applied* to analyze a *host of problems*, whereas with the legacy approach a *host of methods* may be *suboptimally applied* to analyze a *single problem*. To illustrate the flexibility and power of ODA as a general statistical analysis paradigm, a variety of different common data configurations are analyzed by legacy methods versus by optimal methods, and the results are compared.

Bowker's Test for Symmetry

Identical to McNemar's test for correlated proportions for 2x2 tables, Bowker's test for symmetry is used with square tables involving more than two categories.¹ For both of these tests the null hypothesis is that the cell proportions are symmetric: that is, $p_{ij} = p_{ji}$ for all pairs of table cells. Both tests are inherently two-tailed because the alternative hypothesis is non-directional. Bowker's test is chi-square asymptotic-based: it ignores empty cell pairs, and the minimum expectation² must be satisfied. The present exposition compares results achieved by Bowker's test versus by an iterative UniODA-based procedure that is known as structural decomposition analysis.³ Both of these methods are employed to model stability and change in region of residence (North East, Midwest, South, West) for data collected by the US Bureau of the Census in 1980 and 1985 (Table 4.1).

Table 4.1: Region of Residence

		In 1985		
In 1980	North East	Midwest	East	South
North East	11,607	100	366	124
Midwest	87	13,677	515	302
East	172	225	17,819	270
South	63	176	286	10,192

Data were first analyzed by Bowker's method using log-linear models testing for independence, quasi independence, and quasi symmetry: all of these legacy approaches failed to achieve satisfactory fit and thus were statistically untenable.⁴

The first step of the structural decomposition procedure tested the *a priori* hypothesis that the region of residence was stable between 1980 (treated as the attribute) and 1985 (treated as the class variable).⁵ The MegaODA and UniODA syntax used to conduct this analysis is:

```

OPEN DATA;                                11607 100 366 124
OUTPUT example.out;                      87 13677 515 302
CATEGORICAL ON;                           172 225 17819 270
TABLE 4;                                  63 176 286 10192
CLASS COL;                                END DATA;
DIRECTIONAL < 1 2 3 4;                   GO;
MCARLO ITER 10000;
DATA;
```

In a binary application in which the class variable and attribute both assume only two levels the exact Type I error rate converges for UniODA and Fisher's exact test, and both the UniODA and MegaODA software systems report exact (one- or two-tailed) p for Fisher's exact test for binary applications.³ For the analysis of 2 x 2 tables, Monte Carlo (MC) simulation is thus not needed to estimate exact p . For designs larger than 2 x 2, UniODA solutions have an underlying exact statistical distribution for every combination of sample, data structure, and hypothesis. The p -value associated with an observed level of classification performance for a 2 x 3 (or larger) design is estimated using MC simulation: in syntax provided above the MCARLO command specifies that 10,000 iterations of Fisher's randomization procedure are conducted.

The *stability* model specifying that data fall into the major diagonal of the table was statistically significant ($p < 0.0001$), and had a very strong $ESS = 93.7$ —indicating that the model achieved 93.7% of the classification accuracy that is possible to attain above what is achieved by chance alone. This model correctly classified 53,295 (95.2%) of the total of 55,981 observations: the overwhelming majority of the residences didn't change regions from 1980 to 1985, and the resulting small minority of residences that did change regions are troublesome to model using asymptotic statistical methods.³

In step two of the structural decomposition procedure an exploratory model of regional change in residence (marginal dissymmetry) was sought: the directional hypothesis was eliminated and the table cells successfully modeled in step one were set to zero in the UniODA data table. MegaODA and UniODA syntax used to conduct this analysis is:

```

OPEN DATA;                                DATA;
OUTPUT example.out;                      0 100 366 124
CATEGORICAL ON;                           87 0 515 302
TABLE 4;                                  172 225 0 270
CLASS COL;                                63 176 286 0
MCARLO ITER 25000 TARGET .001 STOP 99.99; END DATA;
                                         GO;
```

The resulting model identified the movement of 172 residences from the East region to the North East, versus the 2.1-fold greater movement of 366 residences from the North East region to the East; as well as the movement of 176 residences from the South region to the Midwest, versus the 1.7-fold greater movement of 302 residences from the Midwest to the South. This result was statistically significant ($p < 0.0001$), but represented a relatively weak effect ($ESS = 21.1$).

Step three of the structural decomposition procedure sought to identify a second marginal dissymmetry model of *residence relocation*. As seen below, table cells successfully modeled in the first two steps were set equal to zero: the same MegaODA and UniODA syntax used to conduct the prior analysis was used for the present analysis, using the following data table:

0	100	0	124
87	0	515	0
0	225	0	270
63	0	286	0

The resulting model identified the movement of 63 residences from the South to the North East, versus the 2.0-fold greater movement of 124 residences from the North East to the South; as well as the movement of 225 residences from the East to the Midwest, versus the 2.3-fold greater movement of 535 residences from the Midwest to the East. This result was statistically significant ($p < 0.0001$), and was a moderate effect ($ESS = 35.7$).

A fourth UniODA statistical analysis is not conducted: for a categorical design with C class categories and C non-empty cells, classification accuracy and ESS will always be perfect. Nevertheless, a final exploratory model can be identified without ascertaining Type I error or the ESS statistic: as seen in the UniODA data table below, all cells already successfully modeled were set to zero.

0	100	0	0
87	0	0	0
0	0	0	270
0	0	286	0

The resulting model identified the movement of 87 residences from the Midwest to the Northeast, versus the comparable 1.1-fold greater movement of 100 residences from the North East to the Midwest; as well as the movement of 270 residences from the East to the South, versus the comparable 1.1-fold greater movement of 286 residences from the South to the East.

Together the four UniODA models correctly classified all the data in the original table. The three UniODA models having statistical significance ascertained together correctly classified 55,238 (98.7%) of the total of 55,981 observations in the table, yielding an overall ESS statistic of 98.2. The initial *a priori* analysis showed that the overwhelming effect was that region of residence was stable between 1980 and 1985. Nevertheless two exploratory analyses identified eight specific statistically reliable instances of marginal dissymmetry. The final UniODA model revealed that the four residual cells in the table reflected marginal symmetry vis-à-vis comparable proportional changes.

Bray-Curtis Dissimilarity Index

The Bray-Curtis dissimilarity index (BCDI) is a widely-used index of the degree or magnitude of difference (i.e., the *compositional dissimilarity*) in the number or count of different categories between two samples. When used with raw count data the BCDI is obtained by computing the sum of the absolute differences between the counts across categories, and dividing this value by the sum of the abundances of the counts across categories. The BCDI is bounded at the extremes by 0 if the samples are identical, and by 1 if the samples are completely disjoint (i.e., for each category the count is zero for one sample and non-zero for the other sample). The result is conventionally multiplied by 100 and expressed in terms of a percentage. Subtracting this value from 100 yields a measure of the *similarity* between the two samples, that is called the Bray-Curtis Index (BCI). For example, for data presented in Table 4.2 the BCDI is computed as 56.8%.⁶

Table 4.2: Number (Count) of Five Ecological Categories for Two Sampling Sites

Ecological Category	Sampling Site	
	S29	S30
A	11	24
B	0	37
C	7	5
D	8	18
E	0	1

The BCDI index begs the answer to three questions: (a) whether the observed between-sample difference is statistically reliable; (b) which categories discriminate the two samples, and in what manner; and (c) would a similar finding occur if the analysis was repeated for an independent random sample? It is straightforward to demonstrate that UniODA may be used to evaluate the inter-sample differences in this example⁷ in a manner that addresses all three questions, by utilizing the following UniODA and MegaODA software syntax (site was dummy-coded as S29 = 1, S30 = 2; ecological category was dummy-coded as A = 1; B = 2; C = 3; D = 4; and E = 5):

OPEN bray.txt;	ATTR category;
OUTPUT bray.OUT;	CAT category;
VARS site category;	MC ITER 25000;
CLASS site;	LOO;
	GO;

For the data in Table 4.2 the UniODA model is: if sample = S29 then predict ecological category = A, C, or D; otherwise if sample = S30 then predict ecological category = B or E: this addresses question (b) above. For this model $p < 0.0002$ [addressing question (a)]. Here, $ESS = 44.7$ indicates a moderate effect. The model correctly classified 26 / 26 (100%) of the counts from sample S29, and 38 / 85 (44.7%) of the counts from sample S30. When the model predicted that the sample was S29 it was correct for 26 / 73 (35.6%) of the counts, and when the model predicted that the sample was S30 it was correct for 38 / 38 (100%) of the counts. To address question (c) a leave-one-out (LOO) validity analysis was performed: the classification performance diminished marginally ($p < 0.0001$, $ESS = 43.5$) suggesting this finding is likely to cross-generalize to an independent random sample.

Chi-Square

Discussed in Chapter 3, less-precise categorical data typically yield comparatively lower levels of statistical power than more-precise ordered data, but categorical data are sometimes more informative than ordered data.^{8,9} For example, serum cholesterol is measured on an integer scale typically ranging between 25 and 1500 mg/dL. However, rather than evaluating specific values, most physicians assess if a patient's cholesterol is "high" (> 240 mg/dL), because an intervention is appropriate for patients with high cholesterol. Therefore, in applications in which cholesterol is used as an attribute in modeling physician behavior, the binary variable "high" versus "not high" cholesterol may be the most appropriate index of a patient's cholesterol level, particularly in multi-sample research involving hold-out or generalizability validity analyses. In this respect physicians behave similarly to a thermostat, automatic transmission, or computer-based stock index arbitrage system—systems for which surpassing threshold values triggers a system response. It is possible that, as is the case for a thermostat, such forms of "automatic" behavior become unstable (less reliable) as the threshold value is approached. This in turn suggests the relevance of fuzzy set theory¹⁰ in such situations, and initial study of the use of fuzzy logic in ODA is promising.¹¹

Ubiquitous in the literature, empirical designs with (multi)categorical class variable and attribute are analyzed using chi-square analysis, an approximate statistic used for assessing the degree of statistical independence between the categorical variables.¹² Chi-square and methods based on chi-square should

not be used if the expected value for any cell is less than five, the “minimum expectation” for chi-square-based analyses.^{2,13,14} It is simple to create a test problem in which UniODA yields perfect intergroup discriminability, and for which chi-square is an invalid test statistic. Imagine an application with $N = 6$ in which, as hypothesized, three class 0 observations score at level 0 on a binary attribute, and three class 1 observations score at level 1. Chi-square should not be used to analyze these data because the minimum expectation assumption is violated: the expected value is < 5 for all four cells in the design, and thus chi-square is a biased test statistic. Imbalance in the class category N s also reduces power of chi-square to detect statistical dependence, and threatens the validity of estimated Type I error rates.^{15,16} An inherent limitation of chi-square is that testing directional hypotheses is impossible. In contrast, none of these issues exist for UniODA. For example, for the test problem the directional hypothesis that attribute values coded as 0 are predictive of class category 0 yields errorless classification and has exact $p < 0.05$.

Failing to meet the minimum expectation assumption and imbalanced (skewed) class category N s are common issues in applications involving multicategorical (also called polychotomous) attributes.¹⁷ It is equally easy to create a test data set for such a design, for which chi-square and methods based on chi-square are invalid. Imagine an application having class categories A, B, and C: $N_A = 3$ observations score in category X on a three-category categorical attribute; $N_B = 3$ observations score in category Y; and $N_C = 3$ observations score in category Z. Due to the small sample and sparse cross-classification table, methods based on chi-square are biased and therefore inappropriate. In contrast, using UniODA to conduct a non-directional test of the alternative hypothesis that the class variable can be predicted (discriminated) on the basis of the attribute, $ESS = 100$, $p < 0.01$.

A frequently employed legacy method based on chi-square and used in the statistical analysis of polychotomous data is the log-linear model: recursive partitioning and iterative proportional fitting are employed in an effort to identify combinations of square and/or rectangular subtables meeting the assumption of independence, or of quasi-independence.¹⁶⁻¹⁸ Structural and empirical zeroes (i.e., cells in the table which by definition do not exist, versus cells which theoretically exist but are empirically empty, respectively) complicate the mechanics underlying analysis as well as the interpretation of findings.¹⁹ The ODA approach is much more straightforward.

As an example of an analysis involving a multicategorical class variable and attribute, and a non-directional hypothesis, consider data in Table 4.3 on congressional voting on the 1836 Pinckney Gag rule, which had historical implications regarding antislavery petitions. Analyzed vis-à-vis the log-linear model the data are difficult to interpret as a result of imbalanced marginal distributions.¹⁷

Table 4.3: Congressional Voting on the 1836 Pinckney Gag Rule

	<u>Yea</u>	<u>Abstain</u>	<u>Nay</u>
<u>North</u>	61	12	60
<u>Border</u>	17	6	1
<u>South</u>	39	22	7

For UniODA the non-directional alternative hypothesis is that vote can be discriminated on the basis of region of country, and the null hypothesis is that this is not true. Data were entered into an ASCII file (Chapter 3) in free tabular format (Appendix A): columns indicated vote (column 1 = yea; 2 = abstain; 3 = nay), and rows indicated region of country (row 1 = north; 2 = border; 3 = south). UniODA and MegaODA software syntax used to conduct this analysis is:

```

OPEN pinckney.dat;
OUTPUT pinckney.out;
CATEGORICAL ON;
TABLE 3;

CLASS COL;
MCARLO ITER 10000;
GO;
```

The resulting priors-weighted UniODA model (if Region = North predict Vote = Nay; if Region = Border predict Vote = Yea; if Region = South predict Vote = Abstain) was statistically reliable ($p < 0.0001$), and returned a moderate effect ($ESS = 28.8$, $ESP = 24.2$). The model had strong sensitivity (88.2%) and moderate predictive value (45.1%) in classification of Nay votes; moderate sensitivity (55.0%) and weak predictive value (32.4%) in classification of Abstain votes; and weak sensitivity (14.5%) but relatively strong predictive value (70.8%) in classification of Yea votes.

This UniODA model identifies the primary voting pattern maximizing ESS for the total sample. However, this primary pattern still misclassifies $(61 + 39) = 100$ Yea voters, $(12 + 6) = 18$ Abstain voters, and $(1 + 7) = 8$ Nay voters. The structural decomposition methodology demonstrated in the example for Bowker's test can be used to extract a second UniODA model in an effort to explain the "residual" (misclassified observations) minority voting pattern. To accomplish this analysis presently the voters correctly classified by the primary UniODA model are removed from the data (Table 4.4) and the UniODA / MegaODA program is rerun for the reduced sample.

Table 4.4: Congressional Voting on the 1836 Pinckney Gag Rule, Primary Voters Eliminated

	<u>Yea</u>	<u>Abstain</u>	<u>Nay</u>
<u>North</u>	61	12	0
<u>Border</u>	0	6	1
<u>South</u>	39	0	7

The resulting non-directional priors-weighted UniODA model (if Region = North then predict Vote = Yea; if Region = Border then predict Vote = Abstain; if Region = South then predict Vote = Nay) was statistically reliable ($p < 0.0003$), and returned a moderate effect ($ESS = 40.9$, $ESP = 42.2$). The model had relatively strong sensitivity (61.0%) and strong predictive value (83.6%) in classification of Yea votes; weak (chance) sensitivity (33.3%) but strong predictive value (85.7%) in classification of Abstain votes; and strong sensitivity (87.5%) but very weak predictive value (15.2%) in classification of Nay votes. As seen, the greatest proportion and number of residual observations represent voters in the South, and the smallest proportion and number of unexplained observations represent voters in the Bordering states.

As an example of a test of a directional hypothesis, consider data on the political affiliation status of 1,852 high school students and their parents (Table 4.5). The seven different political affiliations and their dummy-codes in the original analysis¹⁶ included: strong Democrat (1); Democrat (2); Independent-Democrat (3); Independent (4); strong Republican (5); Republican (6); and Independent-Republican (7). Notice that if the category order and associated codes had instead been Independent-Republican (5), Republican (6), strong Republican (7), then the scale would be ordered, ranging from strong Democrat (1) to strong Republican (7), with Independent (4) located in the center of the scale.

Table 4.5: Political Affiliation of Parents and Children

Student	Parents						
	1	2	3	4	5	6	7
1	180	108	30	20	2	5	3
2	147	167	39	30	10	38	17
3	63	78	38	30	14	30	14
4	33	49	32	50	17	42	14
5	9	13	14	23	17	35	45
6	16	29	14	23	17	92	61
7	9	13	4	10	9	35	64

Data were entered into an ASCII file using free tabular format: rows were the (original) dummy-coded score for the student's (i.e., child's) political affiliation, and columns were the corresponding score for parents' political affiliation. The directional alternative hypothesis that is tested first is that the family members have the same political affiliations: the null hypothesis is that this is not true. Student political affiliation is treated as being the multicategorical class variable and parent political affiliation is treated as the polychotomous attribute. When the number of class categories is the same as the number of attribute categories, and a directional hypothesis is specified, LOO analysis is superfluous. Following is the UniODA and MegaODA syntax required to perform this analysis.

```
OPEN politics.dat;
OUTPUT politics.out;
CATEGORICAL ON;
TABLE 7;
CLASS ROW;
DIRECTIONAL < 1 2 3 4 5 6 7;
MCARLO ITER 10000;
GO;
```

This directional model was statistically reliable ($p < 0.0001$), but classification performance was relatively weak in ecological terms: $ESS = 19.4$; $ESP = 17.9$. Thus, consistent with the *a priori* hypothesis, there is a statistically significant tendency but ecologically modest tendency for students to have the same political affiliation as their parents. Other patterns of changes in parent and child political affiliation may be identified in exploratory analysis using the structural decomposition procedure as was illustrated in the prior example.

Not Chi-Square

Do categorical attributes consisting of inherently unordered categories truly exist for classical data? In the first chi-square example voting behavior ranging from Yea to Nay appears to reflect a manifest measure of the underlying *ordered* latent construct of agreement (level of support). In the second chi-square example political affiliation may be measured as an ordered (bi-polar) attribute ranging between strong Democrat and strong Republican. Biological sex (female versus male) is an imperfect ordered attribute with respect to many possible class variables: (non)support of women's rights; (non)development of breast cancer; or (non)commitment of violent crimes against people, for example. Often cited as an inherently categorical variable, color is ordered in terms of underlying frequency characteristics. Apples and oranges are easily discriminated. In the limit, for an N -of-1 design, any stimulus on any dimension may be ordered in terms of relative prevalence or personal preference.

Consisting of a relatively small number of graduated levels of the attribute, ordinal scales may be the most broadly employed type of measurement scale in all of science. Ordinal categorical scales consist of a relatively small number of qualitative categories ordered with respect to some theoretical factor. For example, at the conclusion of a clinical trial patients are classified as being worse, unchanged, or better: the three qualitative categories are worse, unchanged, and better; the latent factor is quality of clinical outcome; and the categories are ordered from lowest (worse) to highest (better) with respect to quality of clinical outcome. The most widely-used ordinal categorical scales are Likert-type scales, usually having between three and ten levels. Data obtained by categorical ordinal measurement scales are inappropriate for statistical analysis using legacy procedures. In the clinical outcome example, the metric underlying the attribute fails to meet the criteria for analysis by chi-square (nominal data) or t -test (continuous data), so neither method can validly be used to evaluate whether, for example, two therapies (or a therapy versus a placebo/control condition) had different clinical outcomes (Chapter 5 discusses legacy non-parametric methods used in analysis of small-domain ordinal attributes).

Perhaps this mismatch between measurement scale and the statistical assumptions that underlie well-known legacy methods helps to explain the extremely common error of treating *categorical ordinal*

data as if they are *categorical* data. Similarly, attributes yielding highly skewed or multimodal distributions are often first separated into a few ordinal categories using arbitrary criteria (Chapter 2 discusses this very common error), which are then treated as categorical data. Several examples of such analyses compared with the appropriate UniODA analysis are presented.^{20,21}

Smile Production in Infants: A study of smile production in 10-month-old infants compared type of smile and inter-glance interval for attentive versus inattentive mothers (Table 4.6). Inter-glance interval (in seconds) was arbitrarily split into five levels that were considered to be categorical so that statistical analysis via chi-square could be conducted, violating the minimum expectation assumption.² Analysis via chi-square was unrevealing: “The distributions of inter-glance intervals preceding smiles in the Attentive and Inattentive conditions were not reliably different from one another or from the distributions for non-smiling glances in each condition. Furthermore, distributions of inter-glance intervals preceding S→M smiles were also the same in the Attentive and Inattentive conditions. Finally, the distributions of all inter-glance intervals did not differ in the two conditions” (p. 48).²²

Table 4.6: Inter-Glance Interval Preceding Anticipatory Smiles to Mother (S→M), Smiles During Glances (M→S), and Non-Smiling Glances to Mother (No Smile), for Attentive and Inattentive Mothers²²

Attentive Mother Condition				Inattentive Mother Condition			
Inter-glance Interval	Infant Smile Status			Inter-glance Interval	Infant Smile Status		
	S→M	M→S	No Smile		S→M	M→S	No Smile
≤ 5 secs	8	2	10	≤ 5 secs	1	0	9
6-15 secs	8	7	16	6-15 secs	6	1	15
16-30 secs	13	4	16	16-30 secs	1	3	14
31-60 secs	5	8	10	31-60 secs	3	2	17
> 60 secs	4	4	8	> 60 secs	5	4	15

UniODA was conducted for the data in Table 4.6. The class variable was mother’s attention status (0 = inattentive, 1 = attentive). Dummy-coded attributes were infant’s smile status (1 = S→M, 2 = M→S, 3 = No Smile) treated as a categorical scale (it is unclear if using No Smile as the second rather than the third category would create a theoretically cogent ordered attribute assessing the latent construct “the origin of attention” ranging from mother to infant), and inter-glance interval (1 = ≤ 5 secs, 2 = 6 - 15 secs, 3 = 16 - 30 secs, 4 = 31 - 60 secs, 5 = > 60 secs) treated as an ordered scale. The analysis was performed by using the following UniODA and MegaODA syntax:

OPEN DATA;	0 3 5 (repeated 15 times)
OUTPUT smile.out;	END;
VARS mother smile interval;	CLASS mother;
DATA;	ATTR smile interval;
1 1 1 (repeated 8 times)	CAT smile;
1 1 2 (repeated 8 times)	MCARLO ITER 25000;
1 1 3 (repeated 13 times)	LOO;
(etcetera)	GO;
0 3 3 (repeated 14 times)	EX smile=3;
0 3 4 (repeated 17 times)	GO;

For infant smile status the UniODA model was: if Smile = 3 (No Smile) predict class = Inattentive; otherwise predict class = Attentive. The LOO-stable model yielded relatively weak *ESS* = 24.1, *p*<0.0003. A UniODA-based range test^{23,24} (see Chapter 5) comparing the two types of smiles was unrevealing: *ESS* = 1.2, *p* < 0.99. For infant inter-glance interval the UniODA model was: if Interval ≤ 3.5 (≤ 30 secs) predict class = Attentive; otherwise predict class = Inattentive. The LOO-stable model yielded relatively weak *ESS*

$\chi^2 = 16.2$, $p < 0.011$. No multiattribute model was possible: CTA only included infant smile status in the model. In summary, while chi-square analysis found no statistically reliable effects, UniODA discovered that infants smile less often, with greater inter-glance intervals, with inattentive mothers.

Plaintiff Gender and Age: A study evaluated whether age discriminates the plaintiff (i.e., wife or husband) in divorce actions (Table 4.7).

Table 4.7: Plaintiff Age in a Divorce Action

	Age	<25	25-34	35-44	>44
Husband		8	8	6	16
Wife		18	48	22	10

Note: Tabled are frequency counts.²⁵

Age was arbitrarily parsed to create an ordinal scale: < 25, 25 - 34, 35 - 44, and > 44 (coded as 1 - 4, respectively). All possible pairwise comparisons between age categories involving six different 2 x 2 chi-square tests were conducted.²⁶ Pairwise comparisons of the > 44 category with the 25 - 34 and the 35 - 44 age categories were statistically reliable (generalized $p < 0.05$). Analysis via chi-square thus indicated that a greater proportion of husband plaintiffs falls in the > 44 age category, and a greater proportion of wife plaintiffs falls in the 25 - 34 and 35 - 44 age categories.

UniODA was conducted for the data in Table 4.7. The class variable was “plaintif” (ODA software allows variable names of between one and eight characters), with 1 = Husband, and 2 = Wife. The analysis was performed by using the following UniODA and MegaODA syntax:

```

OPEN DATA;                                1 3 (repeated 6 times)
OUTPUT divorce.out;                      1 2 (repeated 8 times)
VARS gender age;                          1 1 (repeated 8 times)
DATA;                                     END;
2 4 (repeated 10 times);                 CLASS gender;
2 3 (repeated 22 times);                 ATTR age;
2 2 (repeated 48 times);                 MCARLO ITER 25000;
2 1 (repeated 18 times);                 LOO;
1 4 (repeated 16 times);                 GO;

```

The resulting UniODA model was: if age \leq 35-44 then predict that plaintiff = wife, otherwise predict that plaintiff = husband. The LOO-stable model yielded moderate $ESS = 31.9$ ($p < 0.0001$): 88 (89.8%) of 98 women were classified correctly, versus only 16 (42.1%) of 38 men.

Outcomes of Marital Therapy. An investigation reported four-year termination outcomes of two different types of therapy for unhappily married couples (Table 4.8).

Table 4.8: Outcomes of Marital Therapies

Type of Therapy	Divorced	No Change	Improved
Insight	3	22	4
Behavior	12	13	1

Note: Tabled are frequency counts.

The expected value is less than three for both entries in the right-most column of the data table, invalidating the use of chi-square.² An omnibus chi-square statistic was reported for the 2×3 table ($p < 0.01$) and an eyeball analysis of the omnibus effect was rendered²⁷: “a significantly higher percentage of (behavior therapy couples) had experienced divorce.”

These data were analyzed by UniODA in the same manner as the prior example. The model was: if outcome = divorced then predict that therapy = behavior, otherwise predict that therapy = insight. The model correctly classified 90% of the observations in insight therapy and 46% in behavior therapy, yielding moderate $ESS = 35.9$, $p < 0.006$. These statistical findings suggest that a significantly higher percentage of the insight therapy couples had experienced unchanged or improved marital relationships.

Strength of Gender Differences: A study reported the frequencies of five arbitrary categories of Cohen's d measure of effect strength for representative studies of gender differences, versus for studies of other effects in the field of psychology (Table 4.9). The omnibus chi-square statistic for the 2×5 table was given ($p < 0.0001$), and then an eyeball analysis concluded: “more gender differences fall in the close-to-zero category than other effects in psychology.”

Table 4.9: Cohen's d by Type of Study²⁸

Type of Study	≤ 0.1	≤ 0.35	≤ 0.65	≤ 1.0	> 1.0
Gender	43	60	46	17	5
Other	17	89	116	60	20

Note: Tabled are frequency counts.

For these data the exploratory hypothesis that type of study could be discriminated on the basis of effect strength was tested using priors-weighted UniODA, paralleling analyses in prior examples. The model was: if $d \leq 0.35$ then predict gender study; otherwise, predict non-gender study. Thus, relative to other areas, gender studies have disproportionately more effect sizes in the close-to-zero (≤ 0.1) and the next-to-close-to-zero (0.11 - 0.35) categories. Correctly classifying 60.2% of the gender difference studies, and 64.9% of other studies, the model yielded a moderate, LOO-stable $ESS = 25.1$, $p < 0.0001$.

Cochran's Q Test

Cochran's Q test is a popular statistical analysis approach in two-way *randomized block designs* in which the response variable may assume only two possible outcomes—for example, success or failure, correct or incorrect, or satisfied versus dissatisfied. Q is a non-parametric extension of McNemar's test for experimental designs having three or more matched sets of frequencies or counts. The null hypothesis is that the treatments (matched sets) are equally effective, and the non-directional alternative hypothesis is that there is a difference in effectiveness among treatments. If a statistically significant omnibus effect results then pairwise comparisons using McNemar's test are conducted to specify the exact effect.²⁹

Comparing Success of Alternatives: In this example UniODA and Q are compared in a small sample application.³⁰ Summarized in Table 4.10, 12 subjects each performed 3 tasks, and the outcome of each task is either success or failure.³¹ Here, $Q = 8.67$, $df = 2$, $p < 0.013$: the null-hypothesis is rejected so it is concluded that the proportion of success in at least two tasks are significantly different. Multiple comparisons indicated that the success proportion between tasks 2 and 3 was significantly different ($p < 0.05$).

Table 4.10: Outcome of Alternative Tasks

Task	Success	Failure
1	8	4
2	3	9
3	10	2

Exploratory (non-directional) UniODA analysis was conducted for this application, modeling success or failure (a two-category class variable called “group”) as a function of three different tasks (a categorical attribute having three levels and called “task”). The MegaODA and UniODA syntax used to conduct this analysis is:

```
OPEN example1.txt;
OUTPUT example1.out;
VARS group task;
CLASS group;
ATTR task;
CAT task;
MCARLO ITER 25000;
GO;
```

The UniODA model was: if task = 1 or 3 then predict success; if task = 2 predict failure. This model correctly classified 18 of 21 (85.7%) successes, and 9 of 15 (60.0%) failures: this moderate accuracy level ($ESS = 45.7$) was statistically significant, $p < 0.027$. A second UniODA analysis was conducted in an effort to discriminate tasks 1 and 3 (an optimal multiple range test) using the following additional syntax, but was statistically unproductive ($p < 0.65$):

```
EX task=2;GO;
```

Evaluating Success Rate in Web Usability Testing: Imagine that a sample of observations independently performs the same set of three or more tasks, and that performance of each observation on each task is scored as being either a success or a failure. Clearly the domain of phenomena for which this design may be useful is enormous. The research considered here is a web usability application, where “success rate” is defined as the proportion of a sample that successfully completes a particular task. For each subject, and for each task, success is coded as a 1, and failure is coded as a 0 (Table 4.11).

For these data $Q = 13.33$, $df = 5$, $p < 0.05$: the null hypothesis is thus rejected and it is concluded that the proportion of success in at least two tasks are significantly different.³² Multiple comparisons were not conducted: for an application involving a total of nine non-zero entries in a design with four categories it is impossible to satisfy the minimum expectation for the chi-square approximation.²

Table 4.11: Performance at Six Tasks (0=Failure; 1=Success)

		<u>Participant Number</u>				
		Task	1	2	3	4
	1	1	0	0	1	1
	2	2	0	0	0	0
	3	3	1	0	1	1
	4	4	0	0	1	0
	5	5	0	0	0	0
	6	6	1	0	1	1

Exploratory UniODA was conducted for these data³³ modeling success or failure (class variable) as a function of six different tasks (a categorical attribute having six levels), with the following syntax:

OPEN example2.txt;	ATTR task;
OUTPUT example2.out;	CAT task;
VARS task outcome;	MCARLO ITER 25000;
CLASS outcome;	GO;

The UniODA model was: if task = 1, 3, or 6 then predict success; if task = 2, 4, or 5 predict failure. This model correctly classified 8 of 9 (88.9%) successes, and 11 of 15 (73.3%) failures: this accuracy level was relatively strong ($ESS = 62.2$) but statistically marginal ($p < 0.085$) due to the small sample and weak statistical power. Additional UniODA analyses (a range test) to discriminate between tasks 1, 3 and 6, and 3 between 2, 4, and 5, were not run due to the marginal omnibus statistical test.

Reptile Display by Store and Holiday: This example compares UniODA versus Cochran's Q test in the analysis of reptile display behavior (lizards *and* versus *or* snakes) for 12 pet shops, collected at times corresponding to four widely-celebrated US holiday seasons (Table 4.12). For these data, $Q = 13.29$, $df = 3$, $p < 0.05$. The null-hypothesis is thus rejected and it is concluded that the proportion of dual reptile display in at least two of the holidays are significantly different.³⁴ Multiple comparisons were not conducted: for an analysis based on chi-square—involved a total of 19 non-zero entries in a design with three degrees of freedom and a strong distributional skew—it is impossible to satisfy the minimum expectation for the chi-square approximation.²

Table 4.12: Reptile Display by Store and Holiday: 0=Snakes *or* Lizards; 1=Snakes *and* Lizards

Store	Valentine's Day	July 4 th	Halloween	Christmas
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	1	1	1	1
5	1	0	0	1
6	0	1	0	1
7	1	0	0	1
8	0	0	0	1
9	0	1	0	0
10	0	0	0	0
11	1	0	0	1
12	0	0	1	1

Exploratory (non-directional) UniODA analysis was conducted for these data, modeling partial versus total reptile display (class variable) as a function of four different holidays (categorical attribute having four levels).³⁵ The MegaODA and UniODA syntax used to conduct this analysis is:

```

OPEN example3.txt;
OUTPUT example3.out;
VARS holiday display;
CLASS display;
ATTR holiday;
CAT holiday;
MCARLO ITER 10000;
GO;
```

The UniODA model was: if holiday = Christmas then predict snakes *and* lizards; otherwise predict snakes *or* lizards. This model correctly classified 27 of 29 (93.1%) mono-reptile data, and 10 of 19 (52.6%) duo-reptile data. This moderate level of accuracy ($ESS = 45.7$) was statistically significant ($p < 0.0078$). A second UniODA analysis attempted to discriminate the three holidays other than Christmas (an optimal range test) using the following additional syntax, but was statistically unproductive ($p < 0.89$, $ESS = 14.7$):

```

EX day=4;
GO;
```

Cohen's Kappa

Commonly used to assess inter-rater agreement for binary designs, kappa has the same test statistic as chi-square applied to a 2 x 2 table.³⁶⁻³⁸ An arbitrarily-weighted kappa statistic is computed for ordered attributes (see Chapter 5, Reliability Analysis), and for applications with more than two raters generalized kappa is problematic when only few raters are available.³⁹⁻⁴¹

Assessing Long-Term Stability: The kappa statistic was used to evaluate the temporal reliability of survey- and Structured Interview (SI)-based assessments of Type A Behavior (TAB) for a 27-year follow-up investigation involving 1,180 surviving participants in the Western Collaborative Group Study. Table 4.13 presents the 27-year test-retest cross-classification table obtained for TAB assessments (Type A or Type B) based on the SI. As seen, 32% of Type As became Type Bs, and 45% of Type Bs became Type As. Here kappa = 0.24, reflecting “moderate” reliability (estimated $p < 0.0001$).⁴²

Table 4.13: Stability of SI-Based TAB Assessments Over a 27-Year Retest Interval

<i>Second Rating</i>	<i>Initial Rating</i>	
	Type A	Type B
Type A	372	284
Type B	175	349

Note: Tabled are frequency counts.

A confirmatory UniODA analysis was conducted for these data to test the *a priori* hypothesis that TAB assessments are consistent across time—that is, that ratings fell into the major diagonal running from the upper left-hand corner to the lower right-hand corner of the test-retest cross-classification table.⁴³ The analysis was accomplished using the following UniODA and MegaODA syntax (Monte Carlo simulation is not conducted because this is a binary application and thus the exact p is computed; see Chapter 2):

OPEN DATA;	DIRECTIONAL < 1 2;
CATEGORICAL ON;	DATA;
OUTPUT tab.out;	372 284
TABLE 2;	175 349
CLASS ROW;	END DATA;
	GO;

The UniODA model yielded $ESS = 23.3$ (exact $p < 0.0001$), indicating relatively weak long-term temporal reliability normed against chance.

Table 4.14 gives the 27-year test-retest cross-classification table for TAB survey self-assessments. As seen, 56% of Type As became Type Bs, and 3% of Type Bs became Type As: for these data kappa = 0.39 reflecting “fair” reliability (estimated $p < 0.0001$).

Table 4.14: Stability of Survey-Based TAB Assessments Over a 27-Year Retest Interval

<i>Second Rating</i>	<i>Initial Rating</i>	
	Type A	Type B
Type A	180	8
Type B	227	272

Note: Tabled are frequency counts.

A confirmatory UniODA model (conducted using the same software syntax used in the preceding example, but substituting the data in Table 4.13) achieved relatively strong $ESS = 50.2$ (exact $p < 0.0001$), indicating relatively strong, statistically significant test-retest reliability: instability of Type As underscores the success of efforts to modify TAB behavior at the study outset.

Finally, Table 4.15 gives the *parallel-forms reliability* cross-classification table for the SI and the survey at study intake and at study follow-up. For intake data $\kappa = 0.16$, and for follow-up data $\kappa = 0.11$: both results indicate “low” inter-method concordance and have estimated p ’s < 0.01 .

Table 4.15: Agreement of SI- and Survey-Based TAB Assessments

Study Intake			Study Follow-Up		
<u>Survey Rating</u>	<i>SI-Based Rating</i>		<u>Survey Rating</u>	<i>SI-Based Rating</i>	
	Type A	Type B		Type A	Type B
Type A	209	95	Type A	118	67
Type B	199	185	Type B	243	243

The confirmatory UniODA model achieved weak effects for intake ($ESS = 16.9$, exact $p < 0.0001$) and follow-up ($ESS=13.8$, exact $p < 0.002$) data, indicating statistically reliable but relatively weak parallel-forms reliability.

Concordance of Clinician and Patient Assessments: Weighted kappa was used to evaluate the concordance between clinician and patient ratings of the patient’s *physical* and *mental* functioning, made on 4-category ordinal scales, for a consecutive sample of 166 outpatients with rheumatoid arthritis. Table 4.16 gives the inter-rater cross-classification table for physical health ratings.

Table 4.16: Agreement Matrix Comparing Clinician and Patient Pairings of Patient’s *Physical* Health Status

<i>Patient Rating</i>	<i>Clinician Rating</i>			
	Complete	Adequate	Limited	Incapacitated
Complete	11	12	0	0
Adequate	12	65	28	0
Limited	0	13	21	3
Incapacitated	0	0	0	1

Note: Tabled are frequency counts.

As a measure of concordance weighted kappa was computed and reported to be 0.39: there are no standards for evaluating ecological significance of this estimate. Here estimated $p < 0.0001$, however the validity of this estimate is called into question as the minimum expectation assumption⁴⁴ is violated.⁴⁴

Demonstrated in prior examples, confirmatory UniODA was used to test the *a priori* hypothesis that clinician and patient ratings agree, and thus fall into the major diagonal running from the upper left-hand corner to the lower right-hand corner of the inter-rater cross-classification table.⁴⁵ This analysis was conducted using the following UniODA and MegaODA software syntax:

OPEN DATA;	DATA;
OUTPUT wkappa.out;	1 1 (note: repeated 11 times)
VARS doctor patient;	2 1 (12 times)
CLASS patient;	etcetera
ATTR doctor;	4 4
DIRECTIONAL < 1 2 3 4;	END DATA;
MCARLO ITER 25000;	GO;

The UniODA model achieved relatively strong $ESS = 55.5$, exact $p < 0.0001$. Off-diagonal entries are disagreements between clinician and patient: structural decomposition can be used to determine if disagreements are patterned or random (see Chapter 5, Reliability Analysis).

Table 4.17 presents the inter-rater cross-classification table for mental health ratings. For these data, weighted kappa was reported as 0.30, with estimated $p < 0.0001$. Here, the confirmatory UniODA model yielded moderate $ESS = 43.3$, exact $p < 0.0001$.

Table 4.17: Agreement Matrix Comparing Clinician and Patient Pairings of Patient's *Mental Health Status*

<i>Patient Rating</i>	<i>Clinician Rating</i>			
	<u>Complete</u>	<u>Adequate</u>	<u>Limited</u>	<u>Incapacitated</u>
Complete	46	28	2	1
Adequate	19	40	6	0
Limited	2	18	2	1
Incapacitated	0	0	0	1

Note: Tabled are frequency counts.

Log-Linear Model

Examples comparing UniODA and log-linear model have already been presented, and more examples are presented in later sections of this book. Table 4.18 presents data for an application having a binary class variable (gender), an ordinal attribute (academic rank), and two testing periods separated by six years.

Table 4.18: Number of Faculty by Academic Rank, Gender, and Year

Rank	1978		1984	
	Male	Female	Male	Female
1	45	28	39	28
2	176	21	114	27
3	144	6	171	18
4	127	2	121	5

Note: Tabled are frequency counts.

Log-linear analysis was used to model the relative odds of men versus women at each academic rank level and across time.⁴⁶ The original analysis also included additional putative determinants of rank (unavailable for this example), including academic degree, publication level and age. Age and publication level were arbitrarily split into three categories, and degree into two categories, as a means of limiting the number of cells in the design matrix (see Chapter 7). The use of five predictor variables necessitated the use of too many interaction terms, so the three variables were combined to make a polychotomous variable with 18 levels. The researchers concluded that: "...none of the direct discrimination values differ significantly" (p. 383). Furthermore, all estimates obtained by collapsed contingency (CC) table odds ratio analysis fell outside of the range of odds estimated by other methods, indicating paradoxical confounding: "...underestimation is much more severe for the odds ratio CC derived from collapsing fitted subtables, further underlining problems associated with collapsing across a non-independent variable" (p. 384).

UniODA was employed to assess if males and females (class variable) can be discriminated on the basis of the academic rank measure (ordered attribute) in a consistent manner over time (generalizability variable). Faculty rank (1 = Instructor, 2 = Assistant Professor; 3 = Associate Professor; 4 = Professor), year (1 = 1978, 2 = 1984), and gender (1 = male, 2 = female) were dummy-coded, and analysis was conducted using the following UniODA and MegaODA software syntax:

OPEN DATA;	3 0 1 (repeated 6 times)	4 1 2 (repeated 121 times)
OUTPUT academic.out;	4 1 1 (repeated 127 times)	4 0 2 (repeated 5 times)
VARS rank gender year;	4 0 1 (repeated 2 times)	END;
DATA;	1 1 2 (repeated 39 times)	CLASS gender;
1 1 1 (repeated 45 times)	1 0 2 (repeated 28 times)	ATTR rank;
1 0 1 (repeated 28 times)	2 1 2 (repeated 114 times)	GEN year;
2 1 1 (repeated 176 times)	2 0 2 (repeated 27 times)	MCARLO ITER 25000;
2 0 1 (repeated 21 times)	3 1 2 (repeated 171 times)	LOO;
3 1 1 (repeated 144 times)	3 0 2 (repeated 18 times)	GO;

The resulting UniODA multisample (Gen) model was: if Academic Rank ≤ 2 (Assistant Professor) then predict Gender = Female (77.0% correct), otherwise predict that Gender = Male (60.1% correct). The omnibus test was statistically significant ($p < 0.0001$), and the effect was of moderate strength ($ESS = 37.1$), indicating that the LOO-stable model generalized over year. Applying this model to the 1978 data, females were 86.0% correctly classified and males were 55.1% correctly predicted: $p < 0.0001$, $ESS = 41.1$. Applying the model to the 1984 data, females were 70.5% correctly predicted and males were 65.6% correctly predicted: $p < 0.0001$, $ESS = 36.1$. The omnibus performance values were inside the domain defined by 1978 and 1984 values, indicating the absence of paradoxical confounding (see Chapter 9).

Multisample (Gen) UniODA found moderate evidence of gender discrimination: a greater proportion of females are Instructors or Assistant Professors, and of males are (Associate) Professors, compared to what is expected by chance. Eyeball analysis suggests that the strength of the effect may be diminishing in time, because the percent of females classified correctly by the model, and ESS , fell in 1984. Relative to 1978, in 1984 the number of male professors fell 4.6% while the number of women in this rank increased by 150%. The rank of Associate Professor saw a 18.8% gain in males, compared to a 200% increase in females. There were 35.2% fewer male Assistant Professors compared with a 28.5% gain for females, and while male Instructors diminished by 13.3%, there was no change in this rank for females. Considered together these results suggest that not only is the relative standing of women increasing, but so too is the relative number of women on the faculty.

Logistic Regression

Logistic regression analysis⁴⁷ (LRA) is widely used in the analysis of applications with a binary class variable and multiple attributes (see Chapter 7). However, LRA is also employed to evaluate individual attributes. Table 4.19 presents findings of a prospective population study examining the effect of arbitrarily-defined serum cholesterol level on coronary heart disease and mortality for middle-aged diabetic men.⁴⁸

LRA was used to test the linear trend over quintiles, and yielded an estimated $p < 0.02$. However, in Table 4.19 the expected value for the cell containing two entries is 4.7, which is lower than the assumed minimum expectation for this design.²

Table 4.19: Number of Diabetic Men Developing Coronary Heart Disease, or Dying During Follow-Up, by Quintiles of Serum Cholesterol

Serum Cholesterol	Heart Disease	NO Heart Disease
$\leq 5.5 \text{ mmol/l}$	3	53
5.6–6.1 mmol/l	2	34
6.2–6.6 mmol/l	6	33
6.7–7.3 mmol/l	5	48
$> 7.3 \text{ mmol/l}$	15	38

Note: Tabled are frequency counts.

It was hypothesized that serum cholesterol level is positively predictive of the development of heart disease and mortality, so a directional UniODA analysis was conducted on these data.⁴⁹ Heart disease status was dummy-coded as 0 = negative, 1 = positive; serum cholesterol level was indicated as quintile (1 = lowest, 5 = highest); and UniODA and MegaODA software syntax was run (see next page). The UniODA model was: if serum cholesterol ≤ 7.3 mmol/l predict NO heart disease; otherwise predict positive for heart disease. In this model only the highest quintile of serum cholesterol level was positively predictive of disease and mortality (see Table 4.19). The model correctly classified 168 of 206 (82%) men without heart disease, and 15 of 31 (48%) men who were positive for heart disease. The model was correct 91% of the time it predicted no disease, and 28% of the time that heart disease was predicted to occur. Classification performance was stable in LOO analysis, so findings are expected to cross-generalize if the model is used to classify an independent random sample.

```

OPEN DATA;
OUTPUT coronary.out;
VARS disease serum;
DATA;
1 1 (repeated 3 times)
1 2 (repeated 2 times)
1 3 (repeated 6 times)
1 4 (repeated 5 times)
1 5 (repeated 15 times)
0 1 (repeated 53 times)
0 2 (repeated 34 times)
0 3 (repeated 33 times)
0 4 (repeated 48 times)
0 5 (repeated 38 times)
END;
CLASS disease;
ATTR serum;
DIR < 0 1;
MCARLO ITER 25000;
LOO;
GO;
```

Markov Processes

Markov models are widely used in the study of sequential processes in the sciences.⁵⁰⁻⁵⁹ In this method every unique state that it is possible to observe in a specific application is assigned a unique identification code. The raw data for a Markov model (that will be pre-processed prior to statistical analysis) constitute a continuous consecutive sequence of coded events. For example, imagine modeling the sequential daily change in the closing price (in any currency units) of a stock between a beginning and final day (d_1 and d_e , respectively). The closing price of the stock the day before d_1 (indicated as day d_0) serves as the reference point to begin the sequence: *change in price* for $d_1 = d_1 - d_0$. If $d_1 > 0$ then stock price relative to the day before increased (Up); if $d_1 < 0$ then stock price relative to the day before decreased (Down); and if $d_1 = 0$ then stock price relative to the day before was unchanged (Same). This process ends when every possible sequential pairwise difference score is computed (the final comparison is $d_e = d_e - d_{e-1}$): at the conclusion of this process there are $d_e - 1$ sequential pairwise differences scores, each coded as being Up, Down, or Same. For exposition, imagine the following ten-day coded sequence was obtained, starting with d_1 and ending with day d_{10} : Up, Up, Same, Up, Up, Down, Down, Same, Up, Up.

Once data in the sequence have been coded, the next step involves creating a transition table in which the number of rows and of columns equal the number of uniquely coded states. In this example the transition table has three ordered rows and three ordered columns (categorical states may also be used): specifically, Up, Same, and Down. Rows of the transition table indicate the state at step i in the sequence, and columns represent the state at step $i + 1$. The transition table is created by evaluating all consecutive pairs of i and $i + 1$ states, placing a tally in the cell of the transition table that describes their relation. Here on d_1 (state i) the price of the stock went Up, and on d_2 (state $i + 1$) the price of the stock also went Up: for this comparison a tally would be indicated in the transition table cell corresponding to Row = Up, Column = Up. Evaluating, tallying and cumulating the results of nine consecutive pairwise comparisons for the present example, the following transition table (Table 4.20) emerged:

Table 4.20: State Transition Table for Hypothetical Stock Daily Price Change Modeling Application

		Price Change at Day $i+1$		
Price Change at Day i		Up	Same	Down
Up	3	1	1	
Same	2	0	0	
Down	0	1	1	

In legacy Markov analyses the transition table is transformed into a *transition matrix* by dividing the number in each cell of the table by the sum of all of the numbers in the table. For example, for the cell (Row = Up, Column = Up), the observed frequency for the cell (3) would be divided by the total number of events recorded in the table (9), resulting in a *state transition probability* for this cell of 3 / 9, or 0.333. Markov models attempt to find structure in transition matrices. Transition probabilities may be zero for theoretical reasons that render the cell logically impossible (a *structural zero*), or because that is simply how the data happened to occur (an *empirical zero*). Legacy statistical methods have difficulty when the transition matrix is sparse—that is, when there are many small probabilities.

Discussed in Chapter 2, in designs involving an ordered class variable the use of return weights is often necessary to address the desired functionality of the model. For example, a model of the data in Table 4.20 would facilitate accurate prediction of the *direction* of daily price change. If observations were weighted by the absolute value of the change in price change then the model would maximize accuracy in predicting the *amount of change* in currency units—that is, profit derived from trading the stock based on the model. Comparative use of raw absolute change scores versus interactively transformed (e.g., to a scale ranging from 1 to 2) absolute change scores in this context has not yet been investigated.

As an example of the use of ODA to identify structure underlying transition tables, consider data concerning 22 stratigraphic sections from the Wasatch and Uinta Mountains for which a modified Markov model was used to determine whether order of different types of rock in carbonate units is random. Data consisted of a transition table for seven different types of rock (states) and a total of 514 state transitions from one type of rock to another. Rock types were: bioclastic grainstone to packstone (dummy-coded 1); pelletoidal grainstone to packstone (2); whole-fossil wackestone (3); pelletoidal to evaporitic wackestone to mudstone (4); mixed terrigenous carbonate (5); bioturbated sandstone (6); and cross-bedded sandstone (7). Note that the major diagonal of the transition table consists of structural zeros: rock A can't undergo transition to rock A, rock B can't undergo transition to rock B, and so forth. The sparse transition table (24 of the 49 cells had fewer than five observed events) was analyzed using a log-linear analysis: "...a stepwise selection procedure was used for identification of positive and negative departures from facies transition frequencies expected under a quasi-independent model" (p. 588).

The log-linear analysis identified eleven cells in the transition table with a difference between the observed and expected frequencies that was "statistically significant" (generalized $p < 0.088$). Each cell is a specific transition of the form, rock type i is followed by rock type j . For these eleven cells two patterns of transitions were identified by eyeball examination: one pattern for transitions observed more often than expected by chance (assessed as the difference between the observed and expected cell frequency), and the other pattern for transitions observed less often than expected by chance. The more often seen pattern was: A leads to C, which in turn leads to D and to B, and D leads to B and to E, and then B leads to E and also to D. The less often seen pattern was: B leads to A, and both A and C lead to both F and G. Considered together, these two patterns and eleven assignment rules correctly identified 178 of the total of 514 events in the transition table: *Overall PAC = 34.6%*.^{5,60}

Discussed in detail elsewhere³ these data were analyzed via nondirectional categorical UniODA: rows corresponded to the rock at state i (the multicategorical class variable), and columns corresponded to the rock at state $i + 1$ (the multicategorical attribute). Weighting by prior odds was used because there were different base rates for rocks of different types, and LOO validity analysis was used to assess model stability. Analysis was accomplished using the following UniODA and MegaODA software syntax:

```

OPEN rocks.dat;
CLASS ROW;
OUTPUT rocks.out;
LOO;
CATEGORICAL ON;
MC ITER 10000;
TABLE 7;
GO;

```

The resulting UniODA model identified one pattern (versus two patterns for the log-linear model) and eight (versus eleven) assignment rules that together correctly classified 217 (versus 178) of the total of 514 events in the transition table: *Overall PAC = 42.2% (ESS = 32.0, p < 0.0001* under the null hypothesis that rocks at state i can't be discriminated on the basis of rocks at state $i + 1$). In the flowchart notation below, unless otherwise notated it is assumed end (right-most) component in the sequence refeeds into the most prevalent (left-most) component of the sequence, and repeats. The flowchart seen in Figure 4.1 suggests a geologically sensible one-dimensional sequential process underlying the state transition table, reflecting the effect of gradually fluctuating depths (energy levels) on deposition in a subaqueous environment (Jack C. Yarnold, Ph.D., personal communication)..

Figure 4.1: UniODA Model of Rock State Transitions

Whole-fossil wackestone	→ Bioclastic grainstone	→ Bioturbated sandstone	→ Cross-bedded sandstone	→ Mixed terrigenous carbonate	→ Pelletoidal grainstone to packstone	→ Pelletoidal to evaporitic wackestone to mudstone
----------------------------	----------------------------	----------------------------	-----------------------------	-------------------------------------	---	---

McNemar's Test for Correlated Proportions

McNemar's test is applied to a 2x2 contingency table to assess the statistical significance of the difference between two correlated proportions: for example when the proportions are based on the same sample of subjects, or on matched-pair samples. The test assesses whether row and column marginal frequencies are equal: the null hypothesis of "marginal homogeneity" states that marginal probabilities for each outcome are identical. If 25 or more observations fall into the minor diagonal (the lower-left-hand and upper-right-hand cells) of the table, then the test has a chi-square distribution with one degree of freedom. If the chi-square test is statistically significant, then the null hypothesis is rejected in favor of the alternative hypothesis that the marginal proportions are significantly different from each other.⁶¹ If either of the cell entries in the minor diagonal is "small" then an exact binomial test is used to evaluate imbalance in the minor diagonal cells.⁶²

This example assesses whether a drug effects a disease. In Table 4.21, counts of individuals with a diagnosis of disease (present or absent) before treatment are given in rows, and counts of individuals with a diagnosis of disease after treatment (present or absent) are presented in columns. This is called a matched pairs design because the same subjects are included in before-and-after measurements. Note that 59 / 92 (64.1%) of observations without disease before treatment had disease after treatment, and 121 / 222 (54.5%) of observations with the disease before treatment were without disease following treatment. The McNemar test statistic is chi-square ($N = 314, df = 1$) = 21.35, $p < 0.00001$: it was concluded that: "...the test provides strong evidence to reject the null hypothesis of no treatment effect."⁶³

Table 4.21: Disease Diagnosis

		After Treatment	
		Present	Absent
<u>Before Treatment</u>	Present	101	121
	Absent	59	33

UniODA and MegaODA syntax used in this application⁶⁴ is:

OPEN DATA;	DATA;
OUTPUT disease.out;	101 121
CATEGORICAL ON;	59 33
TABLE 2;	END DATA;
CLASS COL;	GO;

The UniODA model was: if disease is present before treatment, predict that disease is absent after treatment (78.6% accurate classification, 54.5% predictive value); if disease is absent before treatment, predict that disease is present after treatment (36.9% accurate classification, 64.1% predictive value). This finding was statistically significant ($p < 0.0034$), but $ESS = 15.4$ indicates that this relatively weak effect represented only 15.4% of the theoretical gain in classification accuracy that it is possible to achieve beyond chance.

If either cell entry in the minor diagonal is “small” then an exact binomial test is used to evaluate imbalance between minor diagonal cells. This is illustrated for this second example (Table 4.22) assessing if a drug has an effect on a disease using asymptotic and alternative variations of McNemar’s test.⁶³

Table 4.22: Disease Diagnosis

		After Treatment	
<u>Before Treatment</u>		Present	Absent
Present	Present	59	6
	Absent	16	80

Notice that 16 / 96 or 16.7% of observations without disease before treatment had disease after treatment, while 6 / 65 or 9.2% of observations with disease before treatment were without disease after treatment. For these data the exact binomial test gives $p < 0.053$; McNemar’s test with continuity correction gives chi-square ($N = 161$, $df = 1$) = 3.68, $p < 0.055$; asymptotic McNemar’s test gives chi-square ($N = 161$, $df = 1$) = 4.55, $p < 0.033$; and the mid-P McNemar’s test gives $p < 0.035$. It was concluded that: “...the McNemar’s test and mid-P version provide stronger evidence for a statistically significant treatment effect in this...example.”⁶³

UniODA and MegaODA syntax used in this application is identical to the prior analysis, with the data substituted for the present example. The UniODA model was: if disease is present before treatment then predict disease is present after treatment (93.0% accurate classification, 83.3% predictive value), and if disease is absent before treatment then predict disease is absent after treatment (78.7% accurate classification, 90.8% predictive value). Note that a greater percentage (and more) of the observations without disease acquired the disease after treatment, as compared with observations with disease who were without disease after treatment. The UniODA finding was statistically significant ($p < 0.06 \times 10^{-20}$), and $ESS = 71.7$ reveals that this is a relatively strong effect.

A theoretical limitation on the level of scientific rigor that is possible vis-à-vis McNemar’s method is the use of inherently exploratory chi-square to test confirmatory hypotheses. A related pragmatic issue with reliance on exploratory hypothesis testing is associated reduced statistical power, particularly when the minimum expectation for chi-square is violated, calling the validity of estimated p into question.² For small samples UniODA can identify statistically reliable models that achieve strong effects. For example, in the following small sample matched-pairs demonstration a test is conducted before and after treatment for a sample of 20 patients.⁶³ Counts of individuals with a given test result (either positive or negative) before treatment are indicated in rows, and of individuals with a given test result following treatment (positive or negative) are indicated in columns (Table 4.23). The null hypothesis is there is no difference in the marginal proportions before versus after treatment.

As seen, 1 / 11 (9.1%) of the observations with a negative test result before treatment had a positive test result after treatment, and 3 / 9 (33.3%) of the observations with a positive test result before treatment had a negative test result after treatment. Because the number of discordant pairs ($3 + 1 = 4$) in

the minor diagonal is less than 25, Type I error for McNemar's two-tailed test is based on the cumulative binomial distribution: here $p < 0.63$, failing to reject the null hypothesis.⁶³

Table 4.23: Disease Diagnosis

		After Treatment	
		Present	Absent
<i>Before</i> Treatment	Present	6	1
	Absent	1	10

UniODA and MegaODA syntax used in this application is identical to the prior analysis, with the data substituted for the present example. The UniODA model obtained was: if the test is positive before treatment then predict the test is positive after treatment (86.7% accurate classification, 66.7% predictive value), and if the test is negative before treatment then predict the test is negative following treatment (76.9% accurate classification, 90.9% predictive value). The UniODA finding was statistically significant ($p < 0.017$), and $ESS = 62.6$ reveals the relatively strong effect yielded 62.6% of the theoretical gain in classification accuracy that it is possible to achieve beyond chance.

Turnover Table

Used to summarize the behavior of an attribute assessed for a sample of observations that is measured at two points in time, a turnover table is a special case of a state transition table in which the rows indicate level on the attribute at the first recording (i.e., the state at time $i - 1$), and columns indicate level on the attribute at the second recording (i.e., the state at time i).^{3,17,19,65} If observation responses to the attribute are completely consistent at the two measurements then all of the data will fall into the major diagonal of the table, and the test-retest reliability is perfect: in this circumstance the turnover table will necessarily be sparse and thus problematic for analysis via legacy statistical methods. As an observation's responses begin to vary across measurements the temporal reliability of the variable decreases. Analysis of turnover tables thus requires assessing the *stability* (responding on the major diagonal) and *instability* (off-diagonal responding) of an attribute. Often analyzed using log-linear modeling, analysis is problematic when the turnover table is sparse, or has structural or empirical zeros, or imbalanced marginal distributions.^{17,19,65-67} Because such data are usually inappropriate for analysis by legacy statistical methods, researchers often simply report eyeball-based summaries of the cross-classified data.

Consecutive Codes Can Repeat

As an example of the use of UniODA to evaluate a turnover table for which consecutive codes are allowed to repeat, consider data on temporal stability of learning styles.⁶⁸ The experiential learning model theorizes that learning involves a four-stage cycle beginning with concrete experience, followed in turn by observation/reflection, assimilation of concepts/construction of generalizations, and finally by new interactions with the world.⁶⁹ While individual learning styles are hypothesized to be relatively stable, change (e.g., attributable to development) is also expected. A measure of learning style called the LSI-1985 scores observations on two independent dimensions: concrete experience/abstract conceptualization and active experimentation/reflective observation.⁶⁹ These two dimensions are crossed and the quadrants represent four theoretical learning styles: Accommodator, Diverger, Assimilator, and Converger. In this example the LSI-1985 was administered at the beginning and again at the end of the semester (i.e., 10-week inter-test separation) to a sample of 149 undergraduate students (Table 4.24).⁶⁸

Because the table is sparse (25% of the cells have four or fewer observations) these data were not amenable to legacy statistical analysis, so a qualitative analysis was given: "As seen in the diagonal entries ... the largest effect is the stability (42.8% to 50.0%) of learning styles for each of the four styles. The direction of change indicates that Assimilators who change tend toward Accommodators (21.1%), Accommodators tend toward Diversers (28.6%), Convergers tend toward Assimilators (21.9%), and Diversers tend toward Convergers (21.8%)" (pp. 531-532).⁶⁸

Table 4.24: Temporal Stability of Four Theoretical Learning Styles

		<u>Time 2</u>			
<u>Time 1</u>		Assimilator	Accommodator	Converger	Diverger
Assimilator	26	12	10	9	
Accommodator	4	12	4	8	
Converger	7	3	16	6	
Diverger	4	6	7	15	

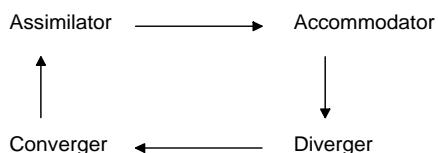
UniODA analysis of stability in turnover tables involves testing the alternative hypothesis that an observation has the same score at the first and second measurements: the data will fall into the major diagonal of the turnover table. The directional analysis was accomplished using the following UniODA and MegaODA software syntax:

```
OPEN learn.dat;
OUTPUT learn.out;
CATEGORICAL ON;
DEGEN ON;
TABLE 4;
```

The directional UniODA model testing the stability hypothesis correctly classified 69 of the total of 149 observations (*Overall PAC = 46.3%*): the model identified a statistically significant ($p < 0.0001$) level of temporal stability of moderate strength ($ESS = 27.5$). Nevertheless, the majority of the entries in the turnover table (53.7%) are misclassified by the stability model. Do these residual observations reflect random error, or does a statistically reliable non-linear transition profile underlie the residual data?

A second, exploratory UniODA analysis was conducted to determine if any statistically reliable sequential structure underlies off-diagonal elements of the turnover table. This was accomplished using structural decomposition analysis: all correctly classified observations (falling in the major diagonal of the table) were deleted from the data set, DIRECTIONAL was turned off, and the program was executed. The UniODA model was: if type at time 1 = Assimilator, predict type at time 2 = Accommodator; if type at time 1 = Accommodator, predict type at time 2 = Diverger; if type at time 1 = Converger, predict type at time 2 = Assimilator; and if type at time 1 = Diverger, predict type at time 2 = Converger. This model correctly classified 34 (42.5%) of the 80 off-diagonal entries in the turnover table: this result was statistically significant ($p < 0.0053$) but relatively weak ($ESS = 24.3$). A schematic illustration of the transition profile identified by this model of reliable change is given in Figure 4.2: arrows indicate transition directions. The pair of UniODA analyses presented here exactly statistically motivate the reported eyeball analysis.⁶⁸

Figure 4.2: Secondary Non-Linear Pattern of Temporal Changes in Learning Style

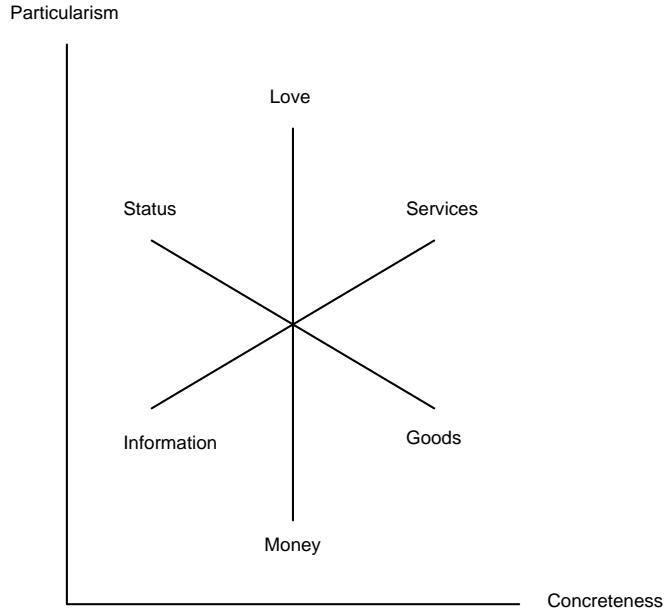


Consecutive Codes Can't Repeat

As an example of the use of UniODA to analyze a turnover table where consecutive codes aren't allowed to repeat (i.e., the major diagonal of the table is structural zeros), consider data on exchange of material and psychological resources.⁷⁰ Theoretically, six resource categories (love, services, goods, money, information, status) are cyclically ordered in the two-dimensional space created by crossing the latent factors of *particularism* and *concreteness*. Categories near each other in the cyclic ordering (e.g., love and ser-

vices) are more similar than further separated categories (e.g., love and goods). In the schematic illustration of Foa's dissimilarity hypothesis (Figure 4.3), lines indicate the hypothesized most *dissimilar* pairings of resource categories.

Figure 4.3: Hypothesized Most *Dissimilar* Resource Categories



To test the dissimilarity hypothesis, 37 people received three messages for each resource category: separately for each received message a person returned a message selected from an accompanying deck, that was judged to be most unlike the message received. In the deck all categories were represented except the category from which the message was received: the diagonal was thus structural zeros. In this application the receipt of the message may be conceived as having occurred at time 1, and return of the message may be conceived as having occurred at time 2: stimulus (attribute) and response (class variable), respectively. Table 4.25 gives the resulting turnover table.

Table 4.25: Dissimilarity of Resource Categories

Message Received (Time 1)	Message Returned (Time 2)					
	Love	Status	Information	Money	Goods	Services
Love	0	5	21	48	29	8
Status	4	0	19	27	30	31
Information	20	11	0	20	25	35
Money	56	10	21	0	4	20
Goods	42	18	27	6	0	18
Services	12	20	37	26	16	0

Although row marginals were constrained to be equivalent, column marginals were and are not. In addition to the structural zeros in the diagonal, matters were complicated by two empirical zeros involving love and money, and two sparse cells with three or fewer entries. Accordingly, eyeball analysis of the turnover table was presented⁷⁰: "...the highest frequency occurred in the cell three steps removed from the diagonal with a decrease as one approached the diagonal from either direction" (p. 347).

Testing the dissimilarity hypothesis using UniODA involves attempting to classify the message returned (second testing) on the basis of message received (first testing). In the present context it is hypothesized that the pairings (Figure 4.3) are most dissimilar. Thus, if a love (row = 1) message is received, then money (column = 4) is the hypothesized least similar response (and 4 is the first code in the DIR command). If a status (row = 2) message is received, then a goods response (column = 5) is hypothesized (and 5 is the second code in DIR). Finally, if a services (row = 6) message is received, then an information response (column = 3) is hypothesized (and 3 is the last, or sixth, code in DIR). Prior odds weighting is appropriate due to class (column) category sample size imbalance. The analysis was accomplished using the following UniODA and MegaODA software syntax:

```

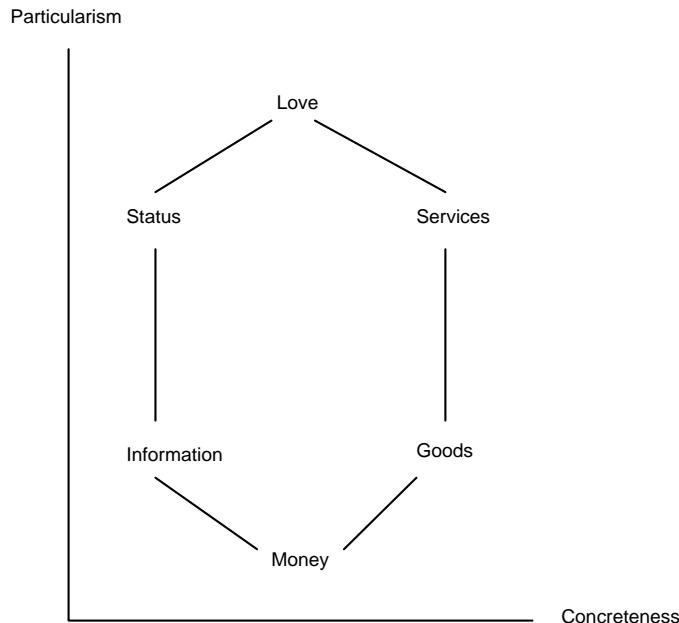
OPEN dissim.dat;                                CLASS COL;
OUTPUT dissim.out;                             DIRECTIONAL < 4 5 6 1 2 3;
CATEGORICAL ON;                               MC ITER 10000;
TABLE 6;                                     GO;

```

The results suggest statistically significant but ecologically weak support for the dissimilarity hypothesis. The *a priori* UniODA model for the dissimilarity hypothesis correctly classified 224 (33.6%) of the total of 666 entries: this result was statistically significant ($p < 0.0001$) but weak in theoretical ($ESS = 19.5$) and practical ($ESP = 20.4$) terms.

As a second, complementary example of the use of UniODA to evaluate a directional nonlinear hypothesis for a turnover table in which consecutive codes may not repeat (the major diagonal consists of structural zeros), consider the *similarity* hypothesis: in Figure 4.4 lines indicate the resource categories that are hypothesized to be the most similar.

Figure 4.4: Hypothesized Most *Similar* Resource Categories



To test the similarity hypothesis, 37 people received three messages for each resource category: separately for each received message a person returned a message selected from an accompanying deck, judged to be most like the message received. In the deck all categories were represented except the category from which the message was received. The resulting turnover table is given in Table 4.26. The row

marginals were constrained to be equivalent, but the column marginals were and are not. In addition to structural zeros, three cells have five or fewer entries. Due to these analytic issues, an eyeball analysis of the turnover table was given⁷⁰: “With a few exceptions, the highest frequencies in each row or column are in the two cells bordering the main diagonal. Frequencies in the cells two steps removed from the diagonal are lower and the lowest frequency is in the cell which is three steps removed, and thus the most distant from the diagonal” (p. 346).

Table 4.26: Similarity of Resource Categories

Message Received <u>(Time 1)</u>	<u>Message Returned (Time 2)</u>					
	Love	Status	Information	Money	Goods	Services
Love	0	5	21	48	29	8
Status	4	0	19	27	30	31
Information	20	11	0	20	25	35
Money	56	10	21	0	4	20
Goods	42	18	27	6	0	18
Services	12	20	37	26	16	0

Two directional alternative hypotheses are implied by the similarity hypothesis: one for similarity assessed in the counter-clockwise direction, and another for similarity assessed in the clockwise direction (Figure 4.4). Consider the counter-clockwise hypothesis. To understand how to specify this hypothesis via the DIR command, start with the first type of message received. As indicated in Table 4.26, love—coded as 1—is the first type of message received; status—coded as 2—is the second type of message received; and services—coded as 6—is the sixth type of message received. In Figure 4.4, travelling in the counter-clockwise direction: love (1) is followed by status (2) so the first parameter code in the DIR command is “2”; status (2) is followed by information (3) so the second parameter code in DIR is “3”; and services (6) is followed by love (1) so the sixth parameter code in DIR is “1”. The following UniODA and MegaODA syntax was employed to evaluate the counter-clockwise hypothesis:

```
OPEN counter.dat;
OUTPUT counter.out;
CATEGORICAL ON;
TABLE 7;
CLASS COL;
DIRECTIONAL < 2 3 4 5 6 1;
MCARLO ITER 10000;
GO;
```

For this directional hypothesis the classification performance was weak in practical terms (*ESS* = 20.6), but statistically significant ($p < 0.0001$).

The second half of the similarity hypothesis involves the clockwise direction. Here love (1) is followed by services (6) so the first DIR parameter code is “6”; status (2) is followed by love (1) so the second DIR parameter code is “1”; and services (6) is followed by goods (5) so the sixth DIR parameter code is “5”. The following UniODA and MegaODA syntax was substituted into the above script in order to evaluate the clockwise hypothesis for the pruned table:

```
OPEN clock.dat;
OUTPUT clock.out;
DIRECTIONAL < 6 1 2 3 4 5;
GO;
```

This directional hypothesis returned *Overall PAC* = 51.0%: classification accuracy was moderate (*ESS* = 37.4), and statistically significant ($p < 0.0001$). Considered together these two *a priori* models explained $(233 + 221) / 666 \times 100\%$, or 68.2% of the total number of events in the original table.

Multisample Analysis

Assessing stability and change in turnover tables is also straightforward for multisample applications.³ To demonstrate this methodology consider five turnover tables used to report the responses of 445 people completing six consecutive monthly interviews regarding their voting intentions for the 1940 presidential election (consecutive codes were allowed to repeat). Relatively small samples and relatively strong test-retest stability created sparse cells and marginal imbalance for all five transition tables (Table 4.27).

Table 4.27: Voting Intentions Across Time⁷¹

May	Republican	June			Stability		Change	
		Democrat	Undecided	ESS	p	ESS	p	
Republican	125	5	16	75.8	.0001	34.7	.0001	
Democrat	7	106	15					
Undecided	11	18	142					
June	Republican	July			Stability		Change	
		Democrat	Undecided	ESS	p	ESS	p	
Republican	124	3	16	76.9	.0001	23.2	.0053	
Democrat	6	109	14					
Undecided	22	9	142					
July	Republican	August			Stability		Change	
		Democrat	Undecided	ESS	p	ESS	p	
Republican	146	2	4	71.6	.0001	33.0	.0001	
Democrat	6	111	4					
Undecided	40	36	96					
August	Republican	September			Stability		Change	
		Democrat	Undecided	ESS	p	ESS	p	
Republican	184	1	7	85.1	.0001	41.4	.0001	
Democrat	4	140	5					
Undecided	10	12	82					
September	Republican	October			Stability		Change	
		Democrat	Undecided	ESS	p	ESS	p	
Republican	192	1	5	85.3	.0001	30.4	.0001	
Democrat	2	146	5					
Undecided	11	12	71					

Stationarity of transition probabilities in a first-order Markov chain was assessed using log-linear models.⁷¹ Analyses suggested that the May-June and June-July tables are similar to each other and different from the (similar to each other) August-September and September-October tables. The July-August table, which was different than the other four tables, reflected the time period during which the Demo-

cratic convention was held, and also that the only significant difference between the July-August table and the August-September and September-October tables involved the undecided category.

A multisample (Gen) UniODA approach may be used to assess if stability and change phenomena generalize across turnover tables. It is hypothesized that voting intentions are primarily stable, so data will fall in the major diagonal of each turnover table. Thus, in the first step of the analysis, a single directional model specifying that voting intention at time $t + 1$ will equal voting intention at time t is simultaneously imposed on each individual turnover table. Analysis was run using the following UniODA and MegaODA software syntax:

```

OPEN voting.dat;                                GEN TABLE 5;
OUTPUT voting.out;                             DIRECTIONAL < 1 2 3;
CATEGORICAL ON;                               MC ITER 10000;
TABLE 3;                                     GO;
CLASS COL;
```

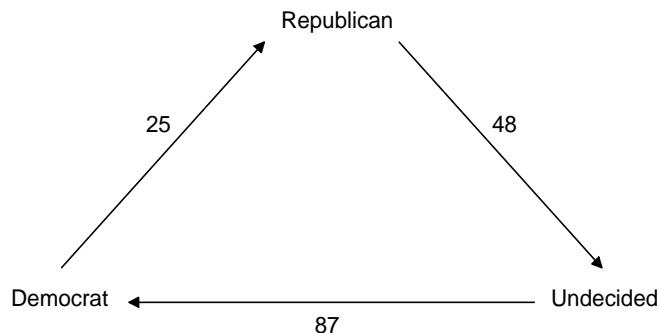
As seen in Table 4.27, for each turnover table the stability hypothesis was statistically significant (experimentwise $p < 0.05$) and yielded relatively strong or strong effects: clearly, the data in each table primarily reflect stable voting intentions. To assess change, the tables were pruned by setting the diagonal elements equal to zero, and then a nondirectional analysis was conducted using the following ODA script.

```

OPEN change.dat;                                CLASS COL;
OUTPUT change.out;                            GEN TABLE 5;
CATEGORICAL ON;                               MC ITER 10000;
TABLE 3;                                     GO;
```

Summarized in Table 4.27, this model was statistically significant (experimentwise $p < 0.05$) and yielded moderate total effect strength for every sample. The resulting Gen UniODA model is illustrated in Figure 4.5 (movement occurs in a clockwise direction). As seen, the primary direction of change was from Undecided to Democrat, followed by from Republican to Undecided: transitions favored the Democratic party. Model classification performance was stable in LOO validity analysis for all turnover tables except July-August. In the July-August table, model LOO sensitivity was stable for the Democrat and Republican classes, but degraded for the Undecided category. Considered in whole the findings suggest that all five turnover tables shared a common transition profile, but data were unstable for those people who were undecided during the July-August time period.

Figure 4.5: Transitions Identified in the Gen UniODA Model



Chapter 5

UniODA with Ordered Attributes

Chapter 5 demonstrates how UniODA, a highly adaptable algorithm, identifies the most accurate possible solutions in a cornucopia of applications involving ordered attributes, that otherwise require a small army of legacy statistical methods (none of which explicitly identify optimal solutions) in order to be addressed.

Kendall's Coefficient of Concordance (W)

Kendall's coefficient of concordance W is a non-parametric statistic used to assess agreement in rankings of multiple stimuli made by multiple raters. A normalization of the test statistic for Friedman's non-parametric alternative to ANOVA with repeated measures, W ranges from 0 (no agreement) to 1 (complete agreement). W is typically used in applications with three or more raters, whereas Cohen's Kappa is typically used to assess agreement for a pair of ratings. W is not a correlation coefficient but it is linearly related to the mean Spearman's rank correlation coefficient obtained for all pairs of rankings. When computing W tied rankings are replaced by the mean of the ranks that would have been assigned if no ties had occurred, serving to reduce the magnitude of W . When there are many ties (a qualitative assessment) a correction is employed. Statistical significance of W is commonly assessed using chi-square, but when this methodology is invalid because the minimum expectation is small, bootstrap assessment is appropriate.¹⁻³

Data for the present example involved rankings of eight movies made by seven raters.³ For this example $W = 0.635$ (mean Spearman $r = 0.574$), suggesting moderate levels of inter-rater agreement. Chi-square indicated statistically marginal rejection of the null hypothesis of no agreement among raters: $p < 0.0591$. However, because the expected values for ratings of movies C (4.6), G (1.7) and H (1.9) were all less than five, chi-square is an invalid approximation in this application.⁴

The theoretical perspective underlying the maximum-accuracy approach is that if raters are in complete agreement regarding the relative rankings of the movies then the UniODA model will perfectly discriminate ($ESS = 100$) movie (a multicategorical class variable; movies were dummy-coded as 1-8) on the basis of ranking (an ordered attribute). In contrast, if movie ranking is a random variable then the UniODA model will be unable to discriminate ($ESS = 0$) movie on the basis of rank. Analysis was conducted using the following UniODA and MegaODA software syntax:

```
OPEN movies.txt          ATTR rank;  
OUTPUT movies.out        MCARLO ITER 1000 TARGET .05 STOP 95.0 STOPUP 95.5;  
VARS movie rank;         GO;  
CLASS movie;
```

The following UniODA model was obtained⁵: if Rank ≤ 1.5 then Movie = H; if $1.5 < \text{Rank} \leq 2.25$ then Movie = G; if $2.25 < \text{Rank} \leq 2.75$ then Movie = D; if $2.75 < \text{Rank} \leq 3.5$ then Movie = F; if $3.5 < \text{Rank} \leq 4.25$ then Movie = C; if $4.25 < \text{Rank} \leq 5.25$ then Movie = E; if $5.25 < \text{Rank} \leq 7.25$ then Movie = B; and if $7.25 < \text{Rank}$ then Movie = A. This model yielded moderate $ESS = 34.7$ and $ESP = 36.6$, and this level of classification accuracy resulted in rejecting the null hypothesis of no agreement among raters: $p < 0.024$

(300 Monte Carlo experiments produced 98.4% confidence for generalized $p < 0.05$). Not illustrated due to lack of requisite data, UniODA can conduct an interesting weighted analysis in this application. If raters provide an estimate of their certainty for each ranking on an ordered index (e.g., Likert-type rating), then a weighted UniODA model of *certainty-adjusted concordance* can be identified. In applications involving more than ten raters, or in which detailed analysis of concordance between pairs of raters is desired, alternative UniODA-based methods for assessing inter-rater reliability discussed ahead are used.

Kruskal-Wallace Test

The Kruskal-Wallace test is a non-parametric one-way analysis of variance (ANOVA) conducted on ranks, an extension of the Mann-Whitney U test for applications involving more than two independent groups.⁶ The null hypothesis of the Kruskal-Wallace test is that the mean ranks of the groups are the same, and the alternative hypothesis is that at least one population mean rank of one group is different from the population mean rank of at least one other group. The Kruskal-Wallace test doesn't identify the pairs of groups for which differences exist, or the directionality of the differences. If an identically shaped and scaled distribution is assumed for all groups, the null hypothesis for the Kruskal-Wallace test is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group. If the null hypothesis is rejected then pairwise comparisons must be conducted in order to identify the underlying differences.⁷

In the first example the Kruskal-Wallace test and UniODA are employed to compare corn yield produced by four different farming methods (Table 5.1).

Table 5.1: Independent Samples of Corn Yield by Farming Method

Method			
1	2	3	4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
84			

Analyzed by the Kruskal-Wallace test⁸, after adjustment for ties, $T = 25.63$, $p < 0.0001$. From this omnibus test it is concluded that at least one farming method has a significantly higher mean rank score than at least one of the other farming methods. Reported in Table 5.2, all pairwise comparisons were conducted in an effort to identify the precise profile of mean rank differences among the farming methods.

Exploratory (non-directional) analysis via UniODA was conducted next⁹, comparing corn yield data (the ordered attribute) between four different farming methods (the class variable with four categorical levels, dummy-coded as 1-4), using the following UniODA and MegaODA software syntax:

OPEN corn.txt;	ATTR yield;
OUTPUT corn.out;	MCARLO ITER 10000;
VARS method yield;	LOO;
CLASS method;	GO;

Table 5.2: p for Pairwise Comparisons Conducted to Disentangle the Omnibus Kruskal-Wallace Finding

Farming Methods <u>Compared</u>	<i>Multiple-Comparison Methodology</i>	
	Dwass-Steel-Chritchlow-Fligner	Conover-Inman
1, 2	0.1529	0.0078
1, 3	0.0782	0.0044
1, 4	0.0029	0.0001
2, 3	0.0048	0.0001
2, 4	0.0044	0.0001
3, 4	0.0063	0.0001

The UniODA model was: if yield ≤ 82 predict method = 4; or if yield ≤ 88 predict method = 2; or if yield ≤ 92 predict method = 1; or if yield > 92 predict method = 3. The model correctly classified 6 of 9 (66.7%) of the samples for farming method 1; 5 of 10 (50.0%) of the method 2 samples; 6 of 7 (85.7%) of the method 3 samples; and all 8 (100%) of the method 4 samples. This level of accuracy was statistically significant ($p < 0.0001$; 99.3% certainty for $p < 0.01$), and considered ecologically this level of accuracy was relatively strong ($ESS = 66.5$). Accuracy declined marginally in LOO validity analysis to $ESS = 59.4$ (accuracy declined for methods 2 and 3), suggesting the finding may cross-generalize with marginally lower accuracy to independent random samples.

UniODA was next used to conduct all six possible pairwise comparisons between pairs of farming methods so as to ascertain the specific nature of the omnibus effect (an alternative method involves conducting an optimal range test, discussed under One-Way ANOVA). This analysis was accomplished by appending the following UniODA and MegaODA software syntax at the end of the current syntax:

```
EX method=1;EX method=2;GO; EX method=2;EX method=3;GO;
EX method=1;EX method=3;GO; EX method=2;EX method=4;GO;
EX method=1;EX method=4;GO; EX method=3;EX method=4;GO;
```

The Type I error rate for each UniODA-based pairwise comparison is presented as the right-most column in Table 5.2, and the UniODA models obtained in pairwise comparisons are presented in Table 5.3.

Table 5.3: UniODA-Based Pairwise Comparisons of Corn Yield Between Four Farming Methods

Farming Methods <u>Compared</u>	UniODA Model	ESS	ESP	p <
1, 2	Yield $\leq 88 \rightarrow$ Method 2	49	52	0.107
1, 3	Yield $\leq 92 \rightarrow$ Method 1	63	62	0.044
1, 4	Yield $\leq 82 \rightarrow$ Method 4	100	100	0.0001
2, 3	Yield $\leq 92 \rightarrow$ Method 2	86	91	0.0015
2, 4	Yield $\leq 82 \rightarrow$ Method 4	90	89	0.0003
3, 4	Yield $\leq 86 \rightarrow$ Method 4	100	100	0.0005

Two perfect models and two nearly-perfect models emerged that help to interpret the omnibus finding. The strongest effect is that farming method 4 has statistically and ecologically *lower* corn yield than the other three methods.

In the second example UniODA and the Kruskal-Wallace test are used to compare the dominance rankings (1=most dominant, 27=most submissive) of 27 free-ranging domestic dogs observed in the outskirts of Rome (Table 5.4). Statistical comparison of dominance rankings of male versus female dogs made using the Kruskal-Wallace test revealed: “The mean rank for males (11.1) is lower than the mean rank for females (17.7), and the difference is significant ($H = 4.61$, df = 1; $p < 0.032$).”¹⁰

Table 5.4: Rankings of Male and Female Dogs

<u>Male Dogs (1)</u>	<u>Female Dogs (0)</u>
1, 2, 3, 4, 5, 6,	7, 8, 9, 10
11, 12, 13, 14,	15, 16,
17, 18, 29, 20, 21	22, 23, 24, 25, 26, 27

UniODA was conducted next, modeling dog gender (class variable) as a function of dominance ranking (ordered attribute).¹¹ First the *a priori* hypothesis was tested that male dogs are more dominant, and thus had lower rankings on the dominance scale used presently than the female dogs:

OPEN dogs.txt;	ATTR rank;
OUTPUT dogs.out;	DIR < 1 0;
VARS gender rank;	LOO;
CLASS gender;	MCARLO ITER 10,000;
	GO;

The UniODA model was: if ranking ≤ 21 then predict male dog; otherwise predict female dog. The model correctly classified all 15 (100%) male dogs, but only 6 of 12 (50%) female dogs. This performance was statistically significant ($p < 0.031$) and relatively strong ($ESS = 50.0$) LOO-stable effect.

For expository purposes a non-directional UniODA analysis was conducted by commenting-out the DIR command (Appendix A) and re-running the analysis. Although the discriminant threshold and thus the classification performance of directional and non-directional models was identical, performance for the exploratory model was *not* statistically significant: $p < 0.052$.

Mann-Whitney U Test

The U test is also known as the Mann–Whitney–Wilcoxon test, the Wilcoxon rank-sum test, the Wilcoxon–Mann–Whitney test, and Kendall’s S , and in the presence of ties U is equivalent to a chi-square test for trend. U is a non-parametric test having the null hypothesis that two populations are the same with respect to the sampled values obtained on an ordered variable. A non-directional (exploratory, two-sided) alternative hypothesis is that one population tends to have different values than the other, and a directional (confirmatory, one-sided) alternative hypothesis is that one population tends to have larger (or to have smaller) values than the other. U is more efficient than t -test for non-normal distributions, and is nearly as efficient for normal distributions. Assumptions underlying use of the U test include random samples drawn from populations; observations are independent of each other; responses are ordinal (i.e. one can say, of any two observations, which is greater); distributions of both groups are equal under the null hypothesis, so the probability of an observation from one population exceeding an observation from the second population is symmetric between populations; and under the alternative hypothesis, the probability of an observation from one population exceeding an observation from the other population (after the exclusion of ties) is not 0.5 (an exploratory hypothesis) or is significantly greater (or less) than 0.5 (a confirmatory hypothesis). The theta statistic used as an index of effect size is equivalent to the area under the receiver operating characteristic (ROC) curve.^{12,13}

Petal Width and Sunlight: In this example U and UniODA are both used to compare cm at the widest point (attribute) of bramble bush leaves growing in full sunlight versus in the shade (class variable): U was judged to be appropriate because sample sizes are too small to assess if they reflect normally distributed data.¹⁴ In Table 5.5, the identical petal width values have the same rank and are scored as the mean of the corresponding rank values. Findings of analysis using U with these data indicated: “Reject the null hypothesis if the smallest value of U_1 (58.5) or U_2 (5.5) is below U_{crit} . In this case U_2 is below 13 so we reject the null hypothesis and accept the alternative hypothesis. The difference between the size of the bramble leaves in the light and the dark is significant for $p > 0.05$.¹⁴ Thus the difference found in pedal size for the two light conditions is unlikely to have occurred by chance. Inspecting medians suggests that

shade leaves are larger than sunlight leaves. However it was initially predicted there would be some kind of difference between the sizes of the two types of leaves, not that shade leaves would be larger than sunlight leaves. Having conducted a non-directional, two-tailed test, strictly speaking all one can conclude from it is that the two types of leaves differ in ranked width.

Table 5.5: Petal Width and Sunlight: Raw and Rank Score

Petal Width	N_{Sunlight}	N_{Shade}	Rank Score
4.1	1		1
4.5	1		2
4.8	1		3
5.1	2		4.5
5.3	1		6
5.5	1	3	8.5
5.9		1	11
6.0	1		12
6.3		1	13
6.5		1	14
6.8		1	15
7.2		1	16

UniODA analysis requires the restructured data matrix shown in Table 5.6.

Table 5.6: Sunlight and Petal Width: UniODA Data Set

Class (Sunlight Group)	Petal Width	Rank Score
1	4.1	1
1	4.5	2
1	4.8	3
1	5.1	4.5
1	5.1	4.5
1	5.3	6
0	5.5	8.5
0	5.5	8.5
0	5.5	8.5
1	5.5	8.5
0	5.9	11
1	6.0	12
0	6.3	13
0	6.5	14
0	6.8	15
0	7.2	16

Petal width was first compared between the groups¹⁵ using the following UniODA and MgaODA software syntax:

```
OPEN leaf.txt;                                ATTR width rank;
OUTPUT leaf.out;                               MC ITER 25,000;
VARS group width rank;                        LOO;
CLASS group;                                  GO;
```

The UniODA model was: if width \leq 5.4 cm predict group = sunlight; otherwise predict that group = shade. This model correctly classified all 8 (100%) of the leaves from the shade group and 6 of 8 (75%) of the leaves from the sunshine group. This level of accuracy was statistically significant ($p < 0.012$), and corresponded to a very strong effect ($ESS = 75.0$). Model performance was stable in LOO validity analysis, suggesting the finding may cross-generalize if the model is used to classify an independent random sample of bramble bush leaves.

Rank score was compared between the two groups next. The UniODA model was: if rank score \leq 7.25 predict group = sunlight; otherwise predict group = Shade. This model correctly classified all 8 (100%) of the leaves from the shade group and 6 of 8 (75%) of the leaves from the sunshine group. This level of accuracy was statistically significant ($p < 0.012$), and corresponded to a very strong effect ($ESS = 75.0$). The stable LOO validity analysis performance suggests that the model may cross-generalize if used to classify an independent random sample of bramble bush leaves.

The performance of the UniODA models for the width and rank score attributes was isomorphic because *UniODA is invariant over a monotonic transformation of the attribute*. Thus, in these applications, when using UniODA no transformation of raw data into ranks is necessary, as is required by U .

Comparative Effectiveness of Laxatives: In this example U and UniODA are used to compare the effectiveness rating—a Likert-type score ranging from 0="very ineffective" to 10="very effective" (treated as the attribute) of two different types of laxatives (the class variable): identical effectiveness ratings have the same rank and were scored the mean of the corresponding rank values (Table 5.7).

Table 5.7: Laxative Type, Effectiveness Rating, and Corresponding Rank Score

Type	Rating	Rank
1	3	3
1	4	4
1	2	1.5
1	6	7.5
1	2	1.5
1	5	5.5
2	9	11
2	7	9
2	5	5.5
2	10	12
2	6	7.5
2	8	10

Findings of analysis using U with these data indicated that: "The difference that we have found between the ratings for the two laxatives is unlikely to have occurred by chance ($p < 0.05$). It looks as if participants' assessments of the laxative's effectiveness do indeed differ. Inspection of the medians sug-

gests one brand is rated as being more effective than the other brand. However we initially predicted that there would be *some kind of difference* between the two laxatives, and not which brand would be better than the other brand: we have therefore conducted a non-directional, two-tailed test, and strictly speaking all we can conclude from it is that the two laxatives differ in rated effectiveness" (p. 5).¹⁶

Exploratory ("two-tailed") analysis via UniODA was conducted next, comparing the effectiveness ratings of the two types of laxatives.¹⁷ The exploratory ("two-tailed") UniODA program used to find and evaluate the optimal models for rating and rank data was obtained by using the following MegaODA and UniODA software syntax:

```
OPEN laxative.dat; ATTR rating rank;
OUTPUT laxative.out; MC ITER 25000;
VARS type rating rank; LOO;
CLASS type; GO;
```

The UniODA model was: if rating ≤ 4.5 then predict that type = 1; otherwise predict type = 2. This model correctly classified all 6 (100%) of type 2 laxatives and 4 of 6 (66.7%) of type 1 laxatives. This level of accuracy was NOT statistically significant ($p < 0.29$), though the effect was relatively strong ($ESS = 0.67$). Model performance declined in LOO validity analysis (4 / 6 of type 2 laxatives were correctly classified, $ESS = 33.3$), suggesting that the finding is unlikely to cross-generalize with comparable strength if the model is used to classify an independent random sample.

Rank score was compared between the two groups next. The UniODA model was: if rank score ≤ 4.75 then predict that type = 1; otherwise predict that type = 2. This model yielded identical classification accuracy for each laxative type, ESS , Type I error, and LOO performance as was obtained for the analysis of effectiveness ratings: UniODA is invariant over a monotonic transformation of the attribute.

A directional analysis was conducted next. The confirmatory UniODA program was identical to the exploratory model, except that a directional hypothesis was specified (DIR < 1 2;). Laxative type 2 was hypothesized to have greatest effectiveness ratings and rankings for illustrative purposes. Stable ESS was achieved in LOO analysis, revealing a statistically marginal effect ($p < 0.073$). U indicated that the median effectiveness ranking of laxative types 1 and 2 differed, but the absence of a cross-generalizability analysis failed to reveal that this effect is unlikely to hold-up in an attempted replication.

One-Way Analysis of Variance

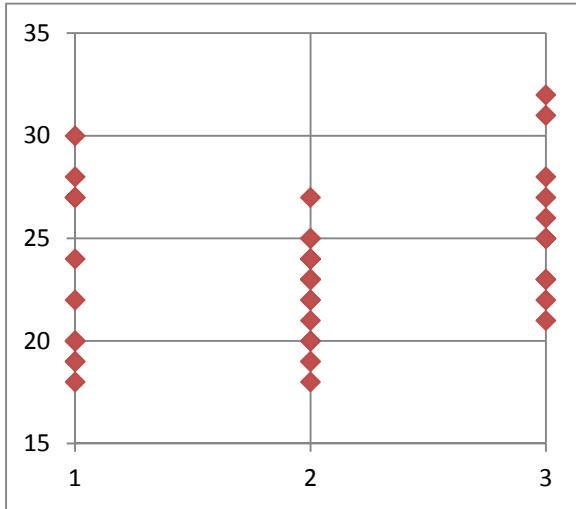
Among the most popular conventional statistical methods, Student's t -test is used to compare the means of two groups on a single dependent measure assessed on a continuous scale. If three or more groups are compared, t -test is generalized to one-way analysis of variance (ANOVA). If the F statistic for the omnibus effect is statistically reliable, then all pairwise comparisons or a more efficient multiple range test is used to ascertain the exact nature of class category differences.

All Possible Comparisons

The first example uses ANOVA and UniODA to compare three methods commonly employed to connect sections of fishing line widely used in big-game sport fishing. This study compared strength of the Double Uni knot used to attach 40-pound-test monofilament line to: a) 50-pound-test solid spectra; b) 60-pound-test hollow spectra; and c) 65-pound-test solid spectra (common reel-backing selections).¹⁸ The three different combinations were respectively dummy-coded as classes 1, 2, and 3. Knot breaking strength was assessed using methodology described elsewhere¹⁹ in which the free end of monofilament line was pulled with increasing force until the connection failed, at which point pounds of force (attribute) was recorded.

Figure 5.1 helps visualize interclass differences: compared to class 3 (40-to-65) few class 2 (40-to-60) data points are 25 pounds or greater, indicating these two class categories are discriminable. A similar but weaker pattern exists for the comparison of classes 1 (40-to-50) and 3 (40-to-65).

Figure 5.1: Knot Strength Data by Class



Eyeball analysis suggests classes 1 (40-to-50) and 2 (40-to-60) involve similar mean pounds of force (Table 5.7 gives descriptive statistics by class category), and although the mean pounds of force for class 3 (40-to-65) is greater, the inter-class mean differences are smaller than the SDs, and the number of knots evaluated is modest, so if a statistically reliable effect is identified then it will be weak.

Table 5.7: Descriptive Statistics: Pounds of Force Required to Break Line-to-Line Connections

	40-to-50	40-to-60	40-to-65
<i>N</i>	14	20	13
<i>Mean</i>	23.4	22.5	25.6
<i>SD</i>	4.13	2.35	3.25
<i>Median</i>	23	23	25
<i>Skewness</i>	0.13	-0.32	0.72
<i>Kurtosis</i>	-1.72	-0.44	0.08

Evaluated by one-way ANOVA, the omnibus effect of class was statistically significant at the generalized criterion: $F(2, 44) = 3.8, p < 0.032$, and $R^2 = 14.6$ indicates that the GLM model accounts for one-seventh of the variation in knot strength observed between classes. The omnibus effect was disentangled using different multiple-comparisons procedures (MCPs), starting with planned contrasts. Performing all possible pairwise comparisons adds three tests of statistical hypotheses (the Sidak criterion for experimentwise $p < 0.05$ with a total of four tests of statistical hypotheses is $p < 0.0128$). Pairwise contrasts found that class 1 and 2 means were not significantly different ($t = 0.8, p < 0.42$); class 1 and 3 means differed marginally at the generalized criterion ($t = 1.8, p < 0.09$); and class 2 and 3 means differed significantly at the experimentwise criterion ($t = 2.7, p < 0.0092$). In summary, analysis using one-way ANOVA and all possible pairwise comparisons to disentangle the statistically reliable omnibus effect indicated that class 3 knots required greater mean force to break than class 1 and class 2 knots, which had statistically comparable mean knot strength. In contrast, MCPs examining overlapping 95% confidence intervals (t -test, Sheffe, studentized maximum modulus, Sidak t -test) found no significant pairwise effects. All range test-based MCPs (Sidak t -tests, Sheffe test, LSD t -test, Ryan-Einot-Gabriel-Welsch multiple range test, Duncan multiple range test, Tukey's studentized HSD range test) indicated that class 3 means are greater than class 2 means, with class 1 means intermediate and not significantly different than class 2 or class 3 means. MCPs report p but don't address the strength of observed differences.²⁰

While ascertaining the appropriate MCP for a given specific application is an issue for parametric analyses, no such ambiguity exists in the ODA paradigm. UniODA was used to conduct an omnibus comparison of the three types of line connections, and all possible pairwise comparisons, using the following UniODA and MegaODA software syntax:

```

OPEN connect.dat;                                MC ITER 20000 TARGET .05 SIDAK 4 STOP 99.9;
OUTPUT connect.out;                             GO;
VARS connect pounds;                           EX connect=3;GO;
CLASS connect;                                 EX connect=1;GO;
ATTR pounds;                                   EX connect=2;GO;

```

The omnibus effect of class was statistically significant at the generalized criterion ($p < 0.015$), and a moderate effect was identified ($ESS = 36.0$; $ESP = 35.6$). The UniODA model was: if pounds ≤ 20 then predict class = 40-to-50; otherwise if $20 < \text{pounds} \leq 24.5$ then predict class = 40-to-60; and if pounds > 24.5 then predict class = 40-to-65. The resulting confusion table is presented in Table 5.8.

Table 5.8: Omnibus ODA Model Confusion Table: Classes Indicated by Strongest Line Strength (Pounds)

		Predicted Class			
		50	60	65	
Actual	50	5	2	6	42.9%
	60	5	12	3	60.0%
Class		65	0	4	9
		54.6%		66.7%	50.0%

It is next necessary to ascertain which pairwise comparison(s) contributed to the omnibus effect. Only the pairwise comparison involving the strongest connections was statistically reliable (generalized $p < 0.022$). The UniODA model was: if Pounds < 24.5 predict the 40-to-60 connection; otherwise predict the 40-to-65 connection ($ESS = 42.8$, $ESP = 36.2$).

Optimal Range Test

The second example uses ANOVA and UniODA to evaluate differences between $N = 377$ white (coded as 1), $N = 378$ African American (coded as 2), and $N = 257$ Hispanic (coded as 3) patients with HIV-associated *Pneumocystis carinii* pneumonia (PCP) on two laboratory tests known to predict PCP outcomes. Data for a random sample of patients with PCP included the three-category class variable *race*, and two attributes: *albumin* (g/dl), and the alveolar-arterial oxygen difference or *AAO2* (mm Hg).²¹ Table 5.9 gives descriptive data for these attributes separately by race.

One-Way ANOVA: Eyeball analysis suggests that White and African American patients obtain essentially the same mean score on both labs. For the Hispanic patients the mean *AAO2* score is lower and the mean albumin score is higher than for the other patients, but mean differences are much smaller than the SD so any statistically significant effect which may be identified will be weak. Evaluated using the GLM paradigm the exploratory hypothesis that race categories have different mean albumin and *AAO2* scores is tested via one-way ANOVA (not presented, the initial GLM analysis in this application involves conducting a multivariate *t*-test, which if statistically significant is subsequently followed by analysis to disentangle the multivariate effect²⁰).

For *albumin* the omnibus effect of race was significant at the generalized criterion: $F(2, 1,009) = 3.7$, $p < 0.026$. The extremely weak $R^2 = 0.73$ reveals the ANOVA model accounts for three-quarters of one percent in the total variation in albumin observed between classes. The omnibus effect was disentangled by multiple comparisons procedures (MCPs). Most MCPs comparing 95% CIs (Sheffe, *t*-test, studentized

maximum modulus, Sidak *t*-test) had no significant pairwise effects. Gabriel CIs found the white patients had higher mean albumin than Hispanic patients, with African American patients in-between and statistically comparable to both other groups. This latter pattern was identified by range test-based MCPs tried, including Sidak *t*-tests, Sheffe's test, LSD *t*-test, Duncan multiple range test, Ryan-Einot-Gabriel-Welsch multiple range test, and Tukey's studentized HSD range test and maximum modulus.

Table 5.9: Descriptive Statistics for Laboratory Data by Race

Albumin, g/dl	African			AAO2, mm Hg		
	<u>White</u>	<u>American</u>	<u>Hispanic</u>	<u>White</u>	<u>African</u>	<u>Hispanic</u>
<i>N</i>	377	378	257	<i>N</i>	377	378
<i>Mean</i>	2.91	2.95	3.06	<i>Mean</i>	50.5	50.0
<i>SD</i>	0.64	0.73	0.82	<i>SD</i>	27.8	28.7
<i>Median</i>	2.90	3.00	3.10	<i>Median</i>	46.0	46.6
<i>CV</i>	22.0	24.6	26.8	<i>CV</i>	55.1	57.4
<i>Skewness</i>	0.15	1.46	1.26	<i>Skewness</i>	1.84	1.85
<i>Kurtosis</i>	1.26	14.0	10.0	<i>Kurtosis</i>	5.96	5.37

For AAO2 the omnibus race effect was statistically significant at the experimentwise criterion: $F(2, 1,009) = 5.7, p < 0.0033$. The extremely weak $R^2 = 1.1$ indicates that the GLM model accounts for one-eighty-ninth of the total variation in AAO2 observed between classes. MCPs comparing 95% CIs were split: studentized maximum modulus, Sidak *t*-test and Sheffe's CIs reported no statistically significant pairwise comparisons, while Gabriel and *t*-test CIs both found (White = African American) > Hispanic. This same pattern was found by range test-based MCPs including Sidak *t*-tests, Sheffe's test, LSD *t*-test, Duncan and Ryan-Einot-Gabriel-Welsch multiple range tests, Tukey's studentized HSD range test, and the studentized maximum modulus range test.

Range Test: Using the UniODA algorithm, no equivocation regarding the “correct” statistical test to use is required, because every ODA analysis conforms perfectly to the analytic task-at-hand. Figure 5.2 and Figure 5.3 help visualize the interclass differences in albumin and AAO2, respectively.²²

Figure 5.2: Albumin Data by Class

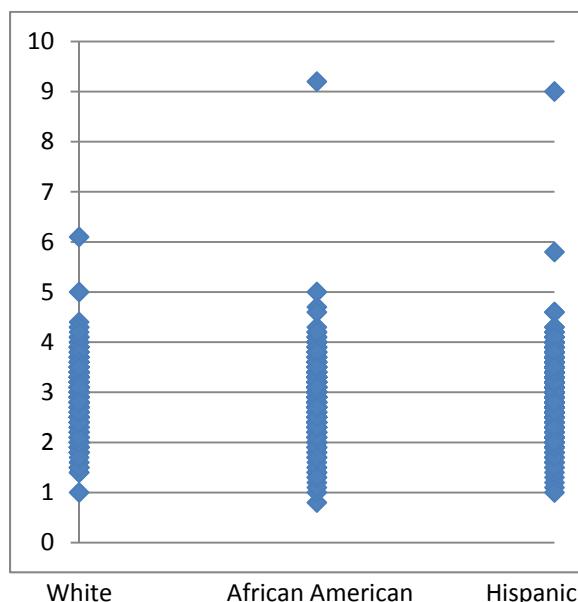


Figure 5.3: AAO2 Data by Class

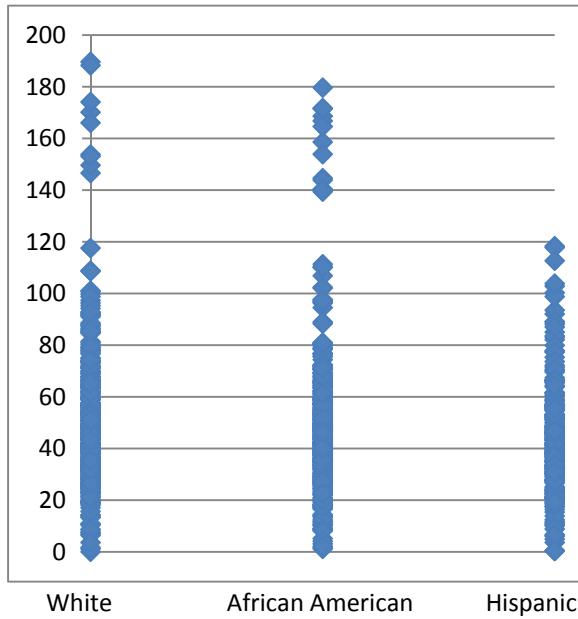


Figure 5.2 shows a high degree of overlap between classes. Figure 5.3 similarly shows substantial distributional overlap: a minority of White and African American patients have AAO2 values greater than 120 mm Hg, but no Hispanic patients have AAO2 values exceeding this threshold. UniODA and MegaODA software syntax used to compare the race groups on both attributes using a non-directional analysis was:

<pre>OPEN race.dat; OUTPUT race.out; VARS race albumin aao2; CLASS race;</pre>	<pre>ATTR albumin aao2; MC ITER 25000; GO;</pre>
--	--

For *albumin* the omnibus class effect was statistically significant at the generalized criterion ($p < 0.016$), but it reflected a weak effect ($ESS = 8.1$; $ESP = 14.7$). The UniODA model was: if $\text{albumin} \leq 1.45 \text{ g/dl}$ then predict African American; otherwise if $1.45 \text{ g/dl} < \text{albumin} \leq 3.45 \text{ g/dl}$ then predict White; and finally if $\text{albumin} > 3.45 \text{ g/dl}$ then predict Hispanic. The corresponding confusion table is given in Table 5.10: note that $100\% - 2.4\% = 97.6\%$ of African American patients were misclassified, and $(307 + 291 + 169) / 1,102$ or 69.6% of observations were predicted to be White.

Table 5.10: Omnibus UniODA Model Confusion Table for Albumin

		Predicted Class			
		W	AA	H	
Actual Class	W	307	3	67	81.4%
	AA	291	9	78	2.4%
	H	169	5	83	32.3%
		40.0%	52.9%	36.4%	

The omnibus UniODA effect is symbolically represented as $AA \leq W \leq H$: yet unresolved analytic ambiguity is if either sign should be a strict (in)equality. Step One of the optimal range test begins at the left-hand side (or right-hand side—this decision is immaterial to the solution) of this model, and the pairwise comparison $AA \leq W$ is conducted. Here, $p < 0.82$, so the symbolic notation is updated: $(AA = W) \leq H$.

The last step of this optimal range test involves resolving this last remaining equality. This is done by combining the indiscriminable class categories AA and W and conducting a directional UniODA analysis testing if class H is greater than the combined class AA and W. Here, $p < 0.0006$ ($ESS = 13.9$, $ESP = 14.2$) so symbolic representation is finalized: $(AA = W) < H$. Albumin levels of African American and White patients are statistically comparable and significantly lower than the albumin levels observed for Hispanic patients. The final UniODA model was: if albumin ≤ 3.45 g/dl predict African American and White, otherwise predict Hispanic: Table 5.11 gives the confusion table: as seen, the final model does well classifying and predicting combined African American and White patients (sensitivity = 80.8%, predictive value = 77.8%).

Table 5.11: Final UniODA Model Confusion Table for Albumin

		<i>Predicted Class</i>		80.8%
		<u>AA+W</u>	H	
<i>Actual Class</i>	<u>AA+W</u>	610	145	80.8%
	H	174	83	
		77.8%	36.4%	

For AAO2 the omnibus class effect was not statistically significant, $p < 0.15$, and UniODA revealed a weak effect: $ESS = 7.6$; $ESP = 14.7$. More examples of optimal range tests conducted in complex one-way designs, some involving repeated measurements, are presented later in this book.

Polychoric Correlation

Widely employed to assess (inter-rater) agreement between ordered-categorical data such as Likert-type ratings, polychoric correlation estimates what the Pearson correlation (see Chapter 6) would be if ratings were made using a continuous scale. The assumptions underlying the validity of this method include that the latent trait (T) on which ratings are based is continuous and normally distributed; the rating errors are normally distributed; the variance of rating errors is homogeneous across levels of T; and the rating errors are independent between raters and cases. If these assumptions are met then the value of the polychoric correlation is interpreted as a Pearson correlation.^{23,24}

Prior research used the number of lambs born to 227 ewes on two consecutive years to illustrate polychoric correlation, conceptualizing the number of lambs that are born as being a continuous, normally distributed indicator of ewe fertility. For these data (Table 5.12) the polychoric correlation was 0.42, but the G^2 statistic indicated that the underlying assumptions were not satisfied so this wasn't a valid method for assessing agreement in this example.^{24,25}

Table 5.12: Number of Lambs Born to 227 Ewes Over Two Years

<i>Lambs Born in 1953</i>			
<i>Lambs Born in 1952</i>	<u>Zero</u>	<u>One</u>	<u>Two</u>
Zero	58	52	1
One	26	58	3
Two	8	12	9

The UniODA and MegaODA software syntax employed to test the exploratory hypothesis that the number of lambs born to ewes is related across year (the null hypothesis is this is not true) was:

```

OPEN lambs.dat;
ATTR number;
OUTPUT lambs.out;
MC ITER 25000;
VARS year number;
GO;
CLASS year;
```

The exploratory UniODA model was: if number of lambs born in 1952 = zero, then predict that the number of lambs born in 1953 = zero; if lambs born in 1952 = one, then predict that lambs born in 1953 = one; and if lambs born in 1952 = three, then predict lambs born in 1953 = three. This linear model yielded moderate accuracy ($ESS = 25.0$) that was statistically reliable ($p < 0.0001$).²⁶ If desired, structural decomposition analysis may be conducted to determine whether statistically reliable structure underlies the off-diagonal table elements, as was done in prior applications.

Reliability Analysis

In general the field of psychometrics addresses reliability (consistency) and validity (truth) characteristics of empirical measurements.²⁷⁻³⁶ Discussion here focuses on reliability, but neither psychometric facet can be considered independently of the other. For both of these psychometric characteristics the concept of measurement precision lies at the heart of the matter: the level of precision that a measurement method can attain limits the potential reliability of its empirical measurements. Only in theoretical measurement utopia is perfectly precise measurement via perfectly valid methods possible. In this utopia, independent observers measuring an attribute of an unchanging object will always obtain the identical score. However, in reality as empirical measurements become increasingly unreliable, scores for an unchanging stimulus increasingly begin to vary within and between observers. Many procedures for assessing the reliability of a set of scores derived using a measurement methodology have been derived, and all provide a reliability coefficient (r_{tt}) that serves as a summary estimate of the stability, consistency, or precision of the set of measurements. The theoretical maximum value r_{tt} can attain (perfect reliability) is 1, and the theoretical minimum value r_{tt} can attain is 0 (the absence of reliability).^{37,38} Many approaches of estimating reliability of empirical measures share a common theoretical basis derived from classical test theory that assumes that an *observed score* is the sum of the true value of the attribute in the subject of measurement (*true score*) plus the unmeasured sources of variability (*error score*): $observed\ score = true\ score + error\ score$. In this approach it is assumed that if a distribution of observed scores is obtained for a single observation, the distribution of error scores will have a mean of zero, be normally distributed, and be independent of the magnitude of the observed scores. For a random sample of observations, the reliability of a measurement methodology is defined as the ratio of the variance of the true scores for the observations divided by the total variance (the sum of the variance of true scores plus the variance of error scores).^{33,39-41}

Inter-Rater Reliability Analysis

Evaluation of inter-rater agreement is a widely, frequently reported method for assessing the reliability of empirical measurements.^{35,36,42-45} In a study designed to collect data for inter-rater reliability analysis, two or more independent raters rate every observation in a (usually small) sample on one or more attributes.

As an example of an inter-rater reliability study for an attribute having an ordinal response scale, consider data from an efficacy study of neuroleptic dosage maintenance and family treatment for schizophrenia.⁴⁶ To estimate inter-rater reliability two psychiatrists (A and B) independently rated the identical videotaped psychopathology interviews of ten randomly selected patients on “unchanging facial expression,” using a categorical ordinal scale ranging from 1 (behavior not present) to 5 (behavior present with extreme severity). The data (tabled are the number of patients) are presented in Table 5.13.

Table 5.13: Five-Point Ratings of Unchanging Facial Expression of Ten Patients, Made by Two Psychiatrists

Psychiatrist		<u>B</u>				
		<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
A	<u>1</u>	3				
	<u>2</u>		2			
	<u>3</u>			1		
	<u>4</u>				1	1
	<u>5</u>				2	

Using UniODA to evaluate inter-psychiatrist agreement, the directional alternative hypothesis is that ratings made by psychiatrists A and B are consistent and thus fall in the major diagonal of the cross-classification table: the null hypothesis is that this is not true. This *a priori* hypothesis is defined in UniODA (by the *directional* command) as shown in Figure 5.4.

Figure 5.4: A UniODA Confirmatory Hypothesis: Ordinal Ratings Agree for an Independent Pair of Raters

<u>Rater-A</u>	<u>Rater-B</u>
1	→ 1
2	→ 2
3	→ 3
4	→ 4
5	→ 5

For example, if Rater-A assigns a triage code of 1 to an ED patient, the UniODA model predicts Rater-B likewise assigns a triage code of 1 to the patient; if Rater-A assigns a triage code of 2 to a patient, the model predicts Rater-B likewise assigns a triage code of 2 to the patient; and so forth. UniODA and MegaODA software syntax used to conduct this analysis (LOO analysis is superfluous, and the decision regarding which rater to select to serve as the class variable is arbitrary in this application) was:

```
OPEN facial.dat;
OUTPUT facial.OUT;
VARS a b;
CLASS a;
ATTRIBUTE b;
DIRECTIONAL < 1 2 3 4 5;
MCARLO ITER 10000;
GO;
```

The classification performance of the UniODA model was statistically significant ($p < 0.0001$) and strong ($ESS = 87.5$). Model sensitivity was 100% for all response levels except for moderate-severe ratings: inter-rater agreement was less than perfect only for one of the three patients rated by psychiatrist B as manifesting extreme unchanging facial expression. The findings reveal that ratings at the low prevalence pole of this measure are highly reliable, and indicate that further rater training, if it is conducted, should emphasize cases in which the symptom is present at severe to extreme levels.

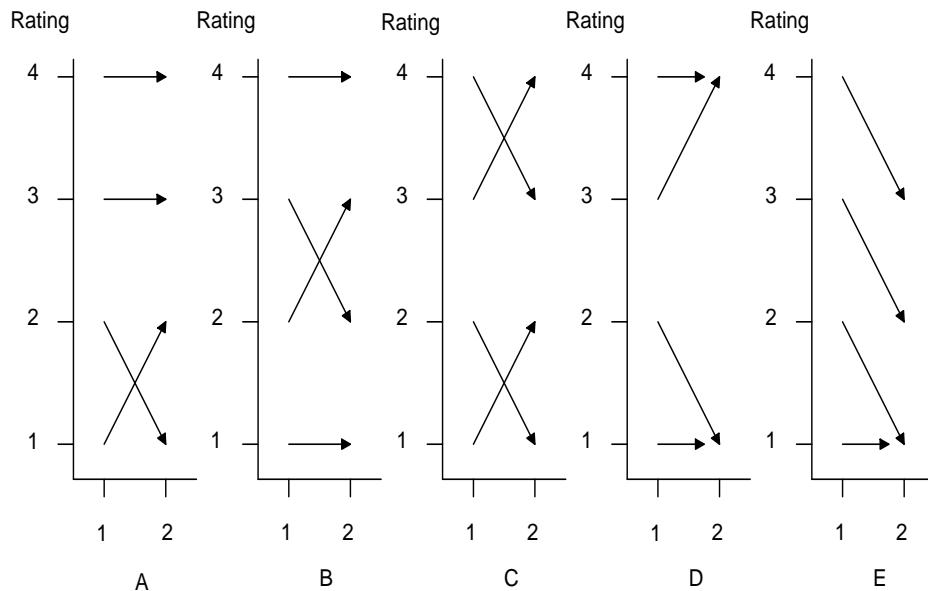
Figure 5.4 isn't the only "reliability" structure that can be hypothesized, or that has been identified in applied research.³⁷ For example, Figure 5.5 illustrates five *non-linear* UniODA models indicating patterns of *reliable bias*.³⁷ Any of these nonlinear structures may underlie data, regardless of whether temporal, inter-rater, split-half, or parallel forms reliability is being assessed. These nonlinear models may be specified *a priori*, or they may be identified in exploratory analysis.

The five different patterns of bias models are indexed beneath the horizontal axis of each plot, respectively indicated from left-to-right as A-F. Ratings made by two independent raters, or made by a single rater (or observation) at two points in time, are indexed beneath each horizontal axis as 1 and 2. Each plot illustrates the relationship between independent ratings of identical stimuli made at the first and second rating vis-à-vis arrows inside the body of the plot. For example, in the first pattern of nonlinear model (prototype A) illustrated in Figure 5.5, the attribute is reliable (stable) at higher values: ratings of 4 made by rater 1 (or made at time 1) are related to (i.e., predict) ratings of 4 made by rater 2 (or made at time 2), and the same is true for ratings of 3. However, in this first pattern the attribute demonstrates local regression at lower values: ratings of 2 made by rater 1 (or made at time 1) are related to (i.e., predict) ratings of 1 made by rater 2 (or made at time 2), and ratings of 1 made by rater 1 (or made at time 1) are related to (i.e., predict) ratings of 2 made by rater 2 (or made at time 2). A symmetric model is possible for some prototypes in Figure 5.5. In prototype A for example, a symmetric pattern involves an attribute that is reliable at lower values but that regresses at higher values.

In the second pattern of nonlinear model illustrated an attribute is reliable at extreme values but demonstrates regression at intermediate values (prototype B in Figure 5.5). This pattern may be seen if a class variable or attribute is constructed by a procedure that assigns observations into different categories

on the basis of their location relative to the mean, median, or any location along a continuum. Typically the greatest likelihood instability occurs near the discriminant threshold(s): in this example the threshold might be 2.5 units on the indicated scale, for example.

Figure 5.5: Five Non-Linear Reliable Bias Structures



In the third pattern of nonlinear model illustrated an attribute may demonstrate local regression throughout its range (prototype C in Figure 5.5). Scores tend to be positively related over the domain of the attribute. However, although lower values (1 and 2) at first testing are associated with lower values at the second testing, and higher values (3 and 4) at first testing are associated with higher values at second testing, there nevertheless is local instability over the range of the attribute. In an application in which the first analysis is a directional stability model, prototype C can be hypothesized to underlie off-diagonal data in a structural decomposition analysis that is designed to identify secondary bias, if it exists.

The last two nonlinear patterns shown in Figure 5.5 are both degenerate models having at least one missing response category value: for example, in prototype D the response categories 2 and 3 were both empty at the second testing. Prototype D illustrates polarization: extreme scores at the first testing remain stable at the second testing, and intermediate scores at the first testing become more extreme (polarized) at the second testing. The symmetric pattern whereby extreme scores at the first testing are less extreme at the second testing is also possible.

Finally, prototype E illustrates a consistent difference in measurement sensitivity between the two ratings, as well as consequent range restriction occurring at the low end of the measurement scale. Disjoint ratings having this pattern may occur as a result of differences in rating criteria existing between raters, changes in rating criteria or successful intervention occurring between ratings, or diminished sensitivity attributable to adjusted cognitive schema occurring with experience (i.e., “rater drift”).

The following example evaluates inter-observer reliability for two different emergency medicine triage algorithms, both of which classify patients into one of five ordinal categories. Ten triage nurses that were previously trained in using the 5-point Canadian Emergency Department (ED) Triage and Acuity Scale (CTAS) were randomized into one of two conditions: (1) five of the nurses (Rater-1 through Rater-5) were trained in use of the 5-point Emergency Severity Index (ESI) Version 3 triage algorithm; (2) the other five nurses (Rater-6 through Rater-10) were instead given refresher training in the CTAS triage algorithm. All training sessions required three hours. Each nurse independently assigned triage scores using either the ESI or the CTAS (as per their assigned condition) to each of 200 case scenarios abstracted from prospectively collected local ED cases. Quadratically-weighted kappa applied to the data produced reliability coef-

ficients exceeding 0.90, reflecting nearly perfect inter-rater reliability.⁴⁸ UniODA was used to test the directional hypothesis that triage codes (integers from 1 to 5) assigned to a sample of patients by two independent nurses are consistent (the null hypothesis is that this is not true).⁴⁹

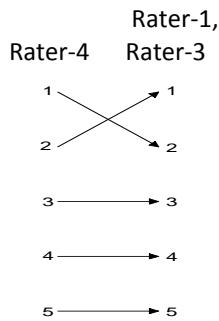
Inter-Observer Reliability of Scores on the ESI: First, separately for each unique rater pairing, UniODA was used to evaluate the *a priori* hypothesis that the triage codes of the raters agree (Figure 5.4).

The *a priori* UniODA model was consistent with the data of four of the total of ten unique rater pairings. Strongest inter-observer reliability was obtained for the (Rater-1, Rater-2) pairing, for which the overall agreement was 61.5%, and the *a priori* UniODA model achieved $ESS = 59.9$ ($p < 0.0001$), indicating relatively strong inter-observer agreement. Consistent performance was obtained in LOO validity analysis, so these results are expected to cross-generalize to an independent random sample of ED patients. The second-strongest inter-observer reliability was obtained for the (Rater-1, Rater-3) pairing: overall agreement was 61.5%, and the *a priori* model achieved $ESS = 47.9$ ($p < 0.0001$), indicating moderate inter-observer agreement. In LOO analysis overall agreement dropped to 61.0% and ESS fell to 35.4, so reduced inter-observer reliability is expected if this pair assesses an independent random patient sample. The third-strongest inter-observer reliability occurred for the (Rater-2, Rater-3) pairing for which the overall agreement = 59.5%, and the *a priori* model achieved $ESS = 45.3$ ($p < 0.0001$), reflecting moderate inter-observer agreement. Overall agreement dropped to 59.0%, and ESS fell to 37.0 in LOO analysis. Finally, weakest inter-observer reliability was obtained for the (Rater-2, Rater-4) pairing, for which the overall agreement was 57.8%, and the *a priori* model achieved $ESS = 39.4$ ($p < 0.0001$): moderate inter-observer agreement. Overall agreement fell to 57.3%, and ESS fell to 31.1 in jackknife analysis.

For all six remaining pairings the *a priori* model was untenable—no UniODA model was possible for the *a priori* hypothesis given the actual data. Thus, for these remaining pairings the *a priori* hypothesis was dropped and an exploratory UniODA analysis was conducted.

For two pairings the UniODA model (an example of prototype A in Figure 5.5) illustrated in Figure 5.6 was identified. For the (Rater-1, Rater-4) pair, overall agreement = 38.7%, $ESS = 36.6$ (moderate inter-observer reliability; $p < 0.0001$; stable in LOO analysis). For the (Rater-3, Rater-4) pair, overall agreement = 38.2%, $ESS = 33.9$ (moderate inter-observer reliability; $p < 0.0005$; there were too few class 1 patients to conduct a LOO analysis). As seen, if Rater-4 assigns a triage code of 1 to an ED patient, the UniODA model predicts Rater-1 and Rater-3 assign a triage code of 2 to the patient; if Rater-4 assigns a triage code of 2 to a patient, then the model predicts Rater-1 and Rater-3 assign a triage code of 1 to the patient.

Figure 5.6: Exploratory UniODA Model for the (Rater-1, Rater-4) and (Rater-3, Rater-4) Pairings



For the remaining four rater pairings—all of which involve Rater-5, no UniODA model was possible that included all five triage levels. Thus, exploratory UniODA was allowed to forego the use of one or more class categories (triage codes) in the model: known as a *degenerate solution*, this methodology is used to identify an optimal model in applications involving sparse or missing class categories.

The strongest exploratory degenerate model was obtained for the (Rater-1, Rater-5) pair (overall agreement = 35.5%, $ESS = 32.8$, $p < 0.0002$, stable in LOO analysis), and it identifies collapsing (*local compression*) in the most serious cases (triage codes < 3), but consistent assignments for codes ≥ 3 . The model is degenerate since no predictions of triage code 2 are made for Rater-1, Rater-3 or Rater-4 (Figure 5.7).

Figure 5.7: Exploratory Degenerate UniODA Model for the (Rater-1, Rater-5), (Rater-3, Rater-5), and (Rater-4, Rater-5) Pairings

Rater-5	Rater-1, Rater-3, Rater-4
1,2	→ 1
3	→ 3
4	→ 4
5	→ 5

The second-strongest exploratory degenerate model was obtained for the (Rater-2, Rater-5) pair: overall agreement = 25.5%, *ESS* = 32.5 (moderate agreement), $p < 0.0001$ (Figure 5.8). Overall agreement diminished to 22.0%, and *ESS* to 20.8 (relatively weak agreement), in LOO analysis. There is local compression for this pair for the more serious and the less serious cases—indicating polarization (prototype D in Figure 5.5, with the middle code uncompressed).

Figure 5.8: Exploratory Degenerate UniODA Model for (Rater-2, Rater-5) Pairing

Rater-5	Rater-2
1,2	→ 1
3	→ 3
4,5	→ 5

The third-strongest exploratory degenerate model occurred for the (Rater-3, Rater-5) pair: overall agreement = 34.0%, *ESS* = 29.3 (moderate agreement), $p < 0.03$ (while this is statistically significant at the generalized criterion, it *isn't* statistically significant at the experimentwise criterion). The UniODA model is illustrated in Figure 5.7. LOO analysis was not possible as the required minimum of two observations per class category wasn't met.

Finally, the weakest exploratory degenerate model was obtained for the (Rater-4, Rater-5) pair: overall agreement = 29.6%, *ESS* = 28.6 (moderate agreement), $p < 0.05$ (while this is statistically significant at the generalized criterion, it *isn't* statistically significant at the experimentwise criterion). The UniODA model is illustrated in Figure 5.7. LOO analysis was not possible because of sparse data.

Table 5.14 summarizes the *ESS* achieved by the UniODA model for the training analyses involving the ESI triage ratings.

Table 5.14: ESI Inter-Observer ESS Results

	<u>Rater-2</u>	<u>Rater-3</u>	<u>Rater-4</u>	<u>Rater-5</u>
Rater-1	59.9	47.9	38.7*	32.8**
Rater-2		45.3	39.4	32.5**
Rater-3			33.9*	29.3**
Rater-4				28.6**

Note: Tabled is ESS for training analysis predicting rating of one rater given rating of the other rater. Entries with an asterisk were obtained by an exploratory model, and entries with two asterisks were obtained by an exploratory degenerate model.

These findings clearly demonstrate that the inter-observer agreement observed for ESI scores is far from perfect. For only one of the ten rater pairs was agreement relatively strong, yielding 59.9% of the gain in agreement that it is theoretically possible to attain above what is expected by chance. For only

four of ten rater pairings was the *a priori* UniODA model even feasible, and the other six models indicated patterns of *consistent disagreement*. All but one of ten models achieved mediocre ESS, and six models identified reliable inconsistencies such as compression, omission, and regression. Results for two models weren't statistically significant at the experimentwise criterion. It is thus concluded that prior kappa- and quadratically-weighted-kappa-based estimates suggesting nearly perfect inter-observer reliability of ESI triage scores, are untenable.

Inter-Observer Reliability of Scores on the CTAS: For each unique rater pairing UniODA was used to evaluate the *a priori* hypothesis that CTAS triage codes of the raters agree (Figure 5.4), and the *a priori* UniODA model was consistent with the data of all ten unique rater pairs.

Strongest inter-observer reliability was obtained for the (Rater-7, Rater-9) pair, for which overall agreement was 52.5%, and the *a priori* UniODA model achieved $ESS = 53.6$ ($p < 0.0001$), indicating relatively strong inter-observer agreement. Consistent performance was obtained in LOO analysis.

The second-strongest inter-observer reliability was obtained for the (Rater-8, Rater-10) pairing, for which overall agreement was 48.7%, and the *a priori* UniODA model achieved $ESS = 52.2$ ($p < 0.0001$), indicating relatively strong inter-observer agreement. Sparse data prevented LOO analysis.

Third-strongest inter-observer reliability was obtained for the (Rater-8, Rater-9) pair, for which overall agreement was 48.2%, and the *a priori* UniODA model achieved $ESS = 45.9$ ($p < 0.0001$), indicating moderate inter-observer agreement. Sparse data prevented LOO analysis.

The fourth-strongest inter-observer reliability was obtained for the (Rater-6, Rater-9) pair, for which overall agreement was 45.5%, and the *a priori* UniODA model achieved $ESS = 43.5$ ($p < 0.0001$), indicating moderate inter-observer agreement. Consistent performance was obtained in LOO analysis.

Fifth-strongest inter-observer reliability was obtained for the (Rater-7, Rater-8) pairing, for which overall agreement was 51.3%, and the *a priori* UniODA model achieved $ESS = 40.4$ ($p < 0.0001$; moderate agreement). In LOO analysis overall agreement fell to 50.8% and ESS fell to 27.9 (moderate agreement).

Sixth-strongest inter-observer reliability was obtained for the (Rater-7, Rater-10) pairing: overall agreement = 49.5%; $ESS = 34.0$ ($p < 0.0001$; moderate agreement). Overall agreement fell to 49.0%, and ESS fell to 21.5 (relatively weak agreement) in LOO analysis.

The seventh-strongest inter-observer reliability was obtained for the (Rater-6, Rater-7) pairing, for which overall agreement was 42.0%, and the *a priori* UniODA model achieved $ESS = 31.1$ ($p < 0.0001$), indicating moderate inter-observer agreement. Consistent performance was obtained in LOO analysis.

The eighth-strongest inter-observer reliability was obtained for the (Rater-9, Rater-10) pairing: overall agreement = 49.0%; $ESS = 29.9$ (moderate agreement); $p < 0.0001$. In LOO analysis the overall agreement fell to 48.5% and ESS fell to 26.8 (moderate agreement).

Ninth-strongest inter-observer reliability was obtained for the (Rater-6, Rater-8) pairing: overall agreement = 44.7%; $ESS = 25.4$ (moderate agreement); $p < 0.0001$. Consistent performance was obtained in LOO analysis.

Finally, the weakest *a priori* model was obtained for the (Rater-6, Rater-10) pair: overall agreement = 37.0%; $ESS = 22.8$ (relatively weak agreement); $p < 0.0001$. Overall agreement fell to 36.5%, and ESS fell to 17.8 (relatively weak agreement) in LOO analysis.

Table 5.15 summarizes the ESS achieved by the UniODA model for the training analyses involving the CTAS triage ratings.

Table 5.15: CTAS Inter-Observer ESS Results

	Rater-7	Rater -8	Rater-9	Rater-10
Rater-6	31.1	25.4	43.5	22.8
Rater-7		40.4	53.6	34.0
Rater-8			48.2	52.2
Rater-9				29.9

As was discovered in the analysis of ESI-based triage codes, findings obtained in the analysis of CTAS-based triage codes clearly demonstrate that inter-observer agreement is far from perfect. Although two of the ten models identified relatively strong inter-observer agreement, another model was relatively

weak, and two models failed the experimentwise criterion for statistical significance. However, in contrast to the exploratory/degenerate UniODA models required in analysis of ESI-based triage ratings, the *a priori* hypothesis was successfully tested for all ten CTAS-based rater pairings. It is concluded that prior kappa-based estimates indicating almost perfect inter-observer reliability of CTAS triage scores are untenable.

Comparing Inter-Observer Reliability of ESI and CTAS Triage Codes: *ESS* values achieved by the inter-observer UniODA models were compared between the ESI (Table 5.14) and the CTAS (Table 5.15) via UniODA. Triage algorithm was treated as a binary class variable, and *ESS* as an ordered attribute (no *a priori* hypothesis was specified). The model achieved $ESS = 20.0$, $p > 0.99$, indicating that the *ESS* values of the inter-observer reliability models didn't discriminate triage algorithm: the inter-observer reliabilities achieved for the ESI and the CTAS were comparably mediocre.

The number of UniODA models which were consistent with the *a priori* hypothesis was compared between the ESI and the CTAS using UniODA. Triage algorithm was treated as the binary class variable, and whether or not the *a priori* model fit the triage data for the pair was treated as the binary attribute (no directional hypothesis was specified). The model yielded $ESS = 60.0$ (relatively strong effect), $p < 0.011$ (statistically significant at the generalized criterion, but *not significant* at the experimentwise criterion). The CTAS inter-observer models were significantly more consistent with the *a priori* hypothesis versus ESI inter-observer models.

In Table 5.14 it is readily apparent that Rater-4 and Rater-5 are using the ESI algorithm in a different manner than the other three raters. It is interesting that although all raters had prior experience using the CTAS to triage emergency patients, versus no prior experience using the ESI, no difference was found in level of inter-observer agreement (assessed as *ESS*) between algorithms. Instead the difference was manifest in terms of the number of models that supported the *a priori* hypothesis. The effect of insufficient experience using the ESI triage algorithm was therefore the observed omission, compression, polarization, and regression anomalies identified in the ESI ratings, but not found in the CTAS ratings.

Inter-Method Reliability Analysis

Viewed from the perspective of classical test theory, parallel forms are alternative equivalent forms of a measuring instrument which, although constituted by different sets of items, measure exactly the same latent construct. An individual should receive identical scores on parallel forms, and samples should have identical means and variances on scores of parallel forms. In this view, any intra-individual differences in scores on parallel forms are attributable to random error. To estimate parallel forms reliability the parallel forms are typically administered to a sample of observations and the correlation between the two sets of scores (the *equivalence coefficient*) is used as the estimated parallel forms reliability.^{37,49,50} The square root of the equivalence coefficient estimates the upper bound of the correlation between the instrument and any other measure.²⁹ Inter-method congruence was evaluated presently using generalizability theory: "The two triage scales appear to be in moderate agreement with one another, as indicated by an inter-test generalizability of 0.58" (p. 243).⁴⁷

Using UniODA to assess parallel-forms (also called inter-method) agreement, the confirmatory alternative hypothesis is that ratings made by forms (methods) A and B are consistent and thus fall in the major diagonal of the cross-classification table: the null hypothesis is that this is not true.⁵¹ This *a priori* hypothesis is defined in UniODA (by the *directional* command) for the present application as in Figure 5.9.

Figure 5.9: Illustration of UniODA *a priori* Hypothesis that Two Triage Coding Algorithms Agree

<u>CTAS</u>	→	<u>ESI</u>
1	→	1
2	→	2
3	→	3
4	→	4
5	→	5

As illustrated, if the CTAS assigns a triage code of 1 to an ED patient, the UniODA model predicts the ESI likewise assigns a triage code of 1 to the patient; if the CTAS assigns a triage code of 2 to a patient, the model predicts the ESI likewise assigns a triage code of 2 to the patient; and so forth. For each unique rater pairing involving different algorithms, UniODA evaluated the *a priori* hypothesis that the triage codes agreed. This model fit the data of 13 of the 25 unique rater pairings, as summarized in Table 5.16.

Table 5.16: Rater Pairings Consistent with the *a priori* Hypothesis

ESI	CTAS	ESS	
		Training	LOO
1	1	35.0	
	2	46.7	
	3	27.3	14.8
	4	46.5	
	5	33.8	21.3
2	1	32.2	
	2	47.2	
	3	25.5	17.1
	4	39.6	
	5	36.0	27.7
3	1	41.9	
	2	49.8	
	4	46.5	

In Table 5.16 the LOO agreement was stable unless provided, and all $p < 0.0001$ were statistically significant at the experimentwise criterion.

For all 12 remaining pairings the *a priori* model was untenable—no UniODA model was possible for the *a priori* hypothesis given the actual data. Thus, for these remaining pairings the *a priori* hypothesis was dropped and an exploratory UniODA analysis was conducted. For seven pairs the exploratory UniODA model illustrated in Figure 5.6 was identified (substitute “method” for “rater” in the Figure caption). Table 5.17 summarizes the findings of these analyses: LOO analysis was not possible for these models due to sparse data for one or more triage codes, and an asterisk indicates that p was statistically significant at the generalized (per-comparison) criterion, but *wasn’t* statistically significant at the experimentwise criterion.

Table 5.17: Rater Pairings Consistent with the Exploratory UniODA Model in Figure 5.6

ESI	CTAS		
Rater	Rater	ESS	$p <$
3	3	33.5	0.0008
	5	28.5	0.04*
4	1	33.2	0.007
	2	35.7	0.0008
	3	33.6	0.003
	4	40.4	0.0001
	5	34.6	0.002

For the remaining five rater pairs no UniODA model was possible that used all five triage levels. The exploratory degenerate UniODA model illustrated in Figure 5.7 emerged for all remaining pairs (substitute “method” for “rater” in the Figure caption). Table 5.18 summarizes the findings of these analyses (all $p < 0.002$).

Table 5.18: Rater Pairings Consistent with the Exploratory Degenerate UniODA Model in Figure 5-7

Rater		ESS	
ESI	CTAS	Training	LOO
5	1	23.6	
	2	32.6	
	3	25.0	
	4	31.5	
	5	35.4	21.6

The findings indicate moderate inter-method agreement for ESI- and CTAS-based triage codes. No rater pair achieved relatively strong agreement, one pair returned relatively weak agreement, and 24 pairs evidenced moderate levels of agreement. The *a priori* UniODA model was feasible for 13 rater pairs, but 12 rater pairs revealed patterns of *consistent disagreement*. Five of the exploratory models identified compression inconsistencies, and seven models identified local regression for the most serious emergency medicine cases. The finding for one rater pair wasn't statistically significant at the experimentwise criterion. As expected, the weakest training *ESS* obtained for inter-method analyses (23.6) was comparable to the weakest training *ESS* values obtained for inter-observer analyses (22.8 for CTAS ratings; 28.6 for ESI ratings). Also as expected, the strongest *ESS* observed for inter-method UniODA models (49.8) was lower than the strongest *ESS* observed for inter-observer models for ESI (59.9) or CTAS (53.6) triage ratings.

Paradoxical Confounding in Reliability Assessment

A recent study conducted a meta-analysis of psychometric properties of patient triage scores assigned using the ESI.⁵² Five studies included in the meta-analysis provided a pooled inter-rater reliability table indicating overall agreement between all unique pairs of raters in the study. In the meta-analysis these data were integrated into an inter-rater reliability table summarizing the agreement between all unique pairs of raters in all five studies. The pooled data indicated 35 / 44 (79.5%) ratings of triage code 1 were consistent; as were 730 / 868 (84.1%) triage code 2 ratings; 610 / 795 (76.7%) triage code 3 ratings; 587 / 767 (76.5%) triage code 4 ratings; and 489 / 646 (75.7%) triage code 5 ratings. For these data the *a priori* hypothesis that ratings between pairs of raters were consistent was evaluated using UniODA and revealed relatively strong inter-rater agreement: *ESS* = 73.2, $p < 0.0001$.

One of the five studies included in the pooled inter-rater reliability table (the study used in the prior examples) involved all expert raters, and reported the highest weighted kappa for ESI ratings in the literature. However, UniODA analysis of the data in that study indicated that the *a priori* hypothesis was only tenable for four (40%) of the total of ten unique inter-rater pairs: for the remaining six inter-rater pairs exploratory non-linear models reflecting various manifestations and magnitudes of *disagreement* were identified.⁵³ For these ten pairs of raters the *ESS* values obtained ranged between 59.9 (relatively strong agreement) and 28.6 (moderate agreement). Since the *ESS* for the combined data falls outside of this range, findings obtained for pooled ratings thus clearly indicate paradoxical confounding (see Chapter 9). These findings raise the concern that all inter-rater (and parallel-forms) reliability estimates based on pooled data reported for *all instruments discussed in the literature* are susceptible to paradoxical confounding, and likely serve to over-estimate the empirical reliability.

Split-Half Reliability

Rather than using two different measuring tools to assess parallel-forms methodology, to assess split-half reliability (an estimate of the parallel-forms reliability) only one measuring tool is used. In the most common procedure for obtaining split-halves (the *adjusted split-half method*), after a multi-item measurement tool is administered to a sample the test items are scored and sorted by descending variance. Initially one split-half consists of even-numbered items in the sorted list, and the other split-half consists of the odd-numbered items. Items are interchanged between halves until mean and variance of the halves is as similar as possible, to satisfy

theoretical assumptions underlying the split-half methodology (the identical assumptions underlying the parallel-forms methodology). To obtain the split-half reliability by legacy psychometric methods a Pearson correlation is computed between split-half scores: after correction for attenuation via the Spearman-Brown prophecy formula the result is used as the estimated split-half reliability of the combined measuring tool.^{35,54-58}

To demonstrate evaluation of split-half reliability in an application involving a polychotomous (rather than ordered) attribute, consider data concerning assessment of psychological androgyny.⁵⁹ A total of 68 male undergraduates completed a self-report androgyny measure assessing two dimensions—instrumentality (I) and expressiveness (E), each using 20 items. To form split-halves, I items were randomly divided into two adjusted split-halves each having 10 items (I_1, I_2), and so were E items (E_1, E_2). Separately for each pair of corresponding split-halves $[(I_1, E_1), (I_2, E_2)]$, each undergraduate was classified into one of four mutually exclusive and exhaustive polychotomous categories reflecting conceptually distinct typologies: androgynous (dummy-coded as 1); instrumentally-typed (2); expressively-typed (3); and undifferentiated (4). Finally, illustrated in Table 5.19, a 4×4 contingency table was created by crossing typology assigned by one split-half (rows) and typology assigned by the other split-half (columns).

Table 5.19: Androgyny Typology by Split-Half

Split-Half #1	Split-Half #2			
	1	2	3	4
<u>1</u>	11	2	2	0
<u>2</u>	3	11	0	4
<u>3</u>	2	0	13	4
<u>4</u>	1	5	1	9

To evaluate the split-half reliability of this four-category assessment procedure via UniODA, data were analyzed under the directional alternative hypothesis that observations classified as being type t ($t = 1, 2, 3$, or 4) by split-half 1 (the four-category class variable) would be likewise classified as type t by split-half 2 (the polychotomous attribute). The UniODA and MegaODA software syntax used to conduct this analysis was:

```

OPEN andro.dat;
CLASS ROW;
OUTPUT andro.out;
DIRECTIONAL < 1 2 3 4;
MCARLO ITER 10000;
TABLE 4;
MCARLO ITER 10000;
CATEGORICAL ON;
GO;
```

The classification performance of the UniODA model is statistically significant ($p < 0.0001$), and it reflects relatively strong inter-rater reliability ($ESS = 53.2$). The model is robust: all of the performance indices indicate a relatively strong effect.

Test-Retest (Temporal) Reliability

The effectiveness of a measurement methodology is dependent not only upon its validity and reliability but also upon its responsiveness, defined in terms of the ability of the methodology to detect minimal yet clinically important differences that occur over time or treatments.^{60,61} Thus, it is quite natural to think of an instrument as being reliable even though it may give different results for a single observation that has been measured at two or more points in time.³⁵ Low responsivity that reduces the accuracy of ratings can be induced by halo biases (ratings that are consistent across—and fail to distinguish among—evaluation dimensions), leniency and severity biases (ratings that are higher or lower than is actually warranted by performance, respectively), and central tendency and range restriction biases (failure to use the full range of the rating scale).^{62,63} It is also important to consider the possibility of reactivity bias that is induced by the measurement methodology, whereby the act of measuring a phenomenon induces changes in the state of the objects of measurement.³⁶ Assuming that such challenges to the ability of an instrument to measure change over time can be addressed, it is a very common practice among researchers to employ a

measuring instrument to assess a sample of observations twice, and then to conceptualize the correlation of the two sets of scores estimating the test-retest (temporal) reliability of the instrument.³⁵⁻³⁷ It is posited that responses are correlated over time because they measure the same true score.²⁹ Widely used, test-retest methodology has been criticized on statistical (lack of independent groups) and methodological (test items are sampled from a population of items; memory confounds measurement) grounds.^{29,64}

The use of UniODA to assess test-retest reliability of scores is demonstrated presently for a study investigating temporal stability of affective (emotional) experience. A sample of 160 undergraduates twice completed a self-report survey assessing a variety of emotions, with a two-week retest interval. Survey items were single-word descriptors of different dimensions of affective experience. People completing the survey indicate the degree to which the affect described by each item constitutes an accurate description of their current state of mind using a 5-point categorical ordinal scale (0 = not at all accurate, 4 = very accurate).⁶⁵ Four items were selected for this exposition: two of the items reflected negative affect (peeved, lonely) and two reflected positive affect (cheerful, friendly); two items reflected comparatively temporary, rapidly-changing emotional states (peeved, cheerful) and two reflected relatively stable, slow-to-change dispositional traits (lonely, friendly).³⁷ Because undergraduates received two scores on each of these four items, the last character of the software syntax name for each item is either 1 or 2, indicating the score is from the first or second testing, respectively. The directional hypothesis evaluated is that responses to items are consistent over the two recordings (structural decomposition analysis may be used to determine if alternative statistically reliable temporal patterns underlie these data). UniODA and MegaODA software syntax used to accomplish these analyses was:

OPEN emotion.dat;	MICARLO ITER 10000;
OUTPUT emotion.out;	CLASS peeved2;ATTR peeved1;GO;
VARS peeved1 peeved2 lonely1 lonely2 cheer1	CLASS lonely2;ATTR lonely1;GO;
cheer2 friend1 friend2;	CLASS cheer2;ATTR cheer1;GO;
DIRECTIONAL < 0 1 2 3 4;	CLASS friend2;ATTR friend1;GO;

Considering first the findings for the trait-like dispositions, analysis revealed moderate temporal reliability for scores on lonely ($ESS = 39.3, p < 0.0001$) and friendly ($ESS = 29.9, p < 0.0003$): sensitivity was highest at the poles (i.e., for codes 0 and 4) for these traits. In contrast, no UniODA model was identified for either of the two state-like transient emotions—peeved and cheerful. This implies that, for these two attributes, at least one rating category level is empty (e.g., no undergraduate answered these items using one of the response options), and/or that the directional hypothesis is untenable given actual temporal structure underlying the data. Presently, none of the attribute categories are empty, therefore the linear directional hypothesis tested is untenable and a (degenerate) nonlinear model should be evaluated.

Repeated Measures Designs

This book addresses many different types of repeated-measures designs such as Markov processes, turnover tables, reliability and bias models, single-case series, intervention studies, and weather forecasting, for example. The example presented here, involving *statistical map-making*, was selected because it represents a productive substantive area for application of established UniODA methodologies, and a fertile empirical field for the development of new UniODA methodologies.

The *Imago Mundi* illustrating Babylon on the Euphrates River is the earliest known map, dating to the 6th to 9th century BCE.⁶⁶ Today maps are ubiquitous, used to illustrate everything from the wealth of nations to different cuts of beef. Some maps, such as physical and topographic maps, display relatively stable phenomena. In comparison, road or resource maps require more frequent modifications to remain accurate, because phenomena they portray change more quickly than geological phenomena change. At the other end of this continuum are maps of dynamic phenomena such as weather, spread of infectious disease, results of political polling, or acreage consumed by a forest fire—phenomena that require real-time updates to effectively augment real-time decision-making. Unless the nature of change is pre-determined, any phenomenon that changes its state or value is a random variable (unchanging or stable phe-

nomena are called constants). Some maps are used to display the findings of statistical analysis of random variables. *Statistical maps* usually give exploratory (two-tailed) results: between-group comparisons for a given point in time, or within-group comparisons for consecutive measurements. Statistical methods commonly used to construct such maps include 95% confidence intervals and *t*-tests.^{67,68}

In this example the use of UniODA in making a statistical map reporting the findings of confirmatory statistical analyses is demonstrated by comparing the annual crude mortality rate in counties of North Dakota, before versus after large-scale commercial usage of toxic chemicals and biocides in the environment began there in 1998. The extraction of shale oil and natural gas is dramatically improved by combining directional drilling and hydraulic fracturing: commonly known by the moniker *fracking*, horizontal slickwater fracturing breaks (fractures) rock using pressurized liquid.⁶⁹ Commercialized in 1998 fracking has inspired global boomtown growth, however there is concern over possible short- and long-term human⁷⁰ and animal⁷¹ health effects of air and water contamination attributable to additives in the fracking liquid. Fracking a single well typically uses four million gallons of water containing 80 tons of toxic chemicals and biocides including agents such as benzene and benzene derivatives, glycol-ethers, toluene, ethanol, naphthalene, and methylene chloride, some of which are known as carcinogenic.⁷² A study of 353 common fracking chemicals (one-third of the known fracking chemicals) found that exposure to 75% of the chemicals affect skin, eyes, and other sensory organs; 52% affect the nervous system; 40% affect the immune and kidney systems; and 46% affect the cardiovascular system and blood.⁷³

Untreated *produced water* ("brine") that is brought to the surface along with the oil or gas has created massive and yet unresolved environmental problems in countries other than the US.⁷⁴ Brine includes fracking liquid injected into the well, and water naturally trapped underground that contains the chemical composition of the geologic formation, including naturally occurring radioactive material or NORM.^{75,76} Following injection, approximately one quarter of the total produced water generally comes to surface within two months, and the balance in 15-20 years: this *flowback* is then treated for subsequent reuse, or transported to deep wells for injection. The largest volume byproduct and waste stream in the field of energy exploration and production, an estimated one million US oil/gas wells generate 2.4 billion gallons of produced water *daily*.⁷⁷ In addition to the composition (or "formula") of the fracking liquid, factors influencing physical and chemical properties of produced water include geographic location, geological formation, and the type of hydrocarbon product produced.⁷⁵ Other public health hazards unrelated to large-scale commercial use of toxic chemicals and biocides also occur as a typical consequence of boomtown development: increases in social problems such as mental health problems, substance abuse, prostitution, traffic and other accidents, unreliable medical care, crime, and water, food and other resource shortages, are inevitable.⁷⁸ And, the drilling profession is inherently hazardous, with workplace accidents in the US resulting in 10 to 20 deaths annually for oil and gas extraction.⁷⁹

Data on the annual crude death rate are available for 1937 - 2005 in North Dakota.⁸⁰ Presently these data were compared before versus after wide-scale use of toxic chemicals in the environment began in North Dakota in 1998, testing the *a priori* hypothesis that crude mortality rate increased after 1997: the class variable was *frack* ($\leq 1997 = 0$; $\geq 1998 = 1$), and the ordered attribute was annual crude mortality rate.⁸¹ Analysis was conducted using the following UniODA and MegaODA software syntax:

```
OPEN fracking.dat; DIR < 0 1;
OUTPUT fracking.out; MCARLO ITER 25000;
VARS frack rate; LOO;
CLASS frack; GO;
ATTR rate;
```

A statistically reliable ($p < 0.0001$), ecologically strong (ESS = 50.0), LOO-stable UniODA model emerged: if the annual crude death rate is $\leq 8.9\%$ predict the year was 1997 or earlier, otherwise predict the year was 1998 or more recent. The model correctly predicted 77% of the years before 1998, and 88% of the years after 1997.

The identical analysis was individually repeated separately for each county, and the findings are summarized in Table 5.20.

Table 5.20: Summary Findings of Separate UniODA Analyses Conducted for Every North Dakota County

County	Mortality Rate Cut-Point	Training <i>p</i> <	Training ESS	LOO <i>p</i> <	LOO ESS	Mean Mortality Rate Before 1998	Mean Mortality Rate After 1998	Mean Annual % Increase in Mortality Rate
Adams	11.8	0.0001	80	0.0005	66	9.6	14.3	6.0
Barnes	10.9	0.005	57	0.25	20	10.5	12.5	2.5
Benson	9.6	0.61	17	0.84	-11	10.1	9.7	-0.5
Billings	6.8	0.36	28	0.56	10	5.3	6.0	1.9
Bottineau	11.2	0.008	55	0.13	28	10.4	12.5	2.5
Bowman	12.9	0.0001	78	0.0002	65	10.1	13.9	4.8
Burke	8.0	0.97	2	0.92	-12	10.8	9.6	-1.4
Burleigh	7.6	0.0004	67	0.002	53	6.6	8.3	3.1
Cass	5.9	0.94	5	0.97	-9	7.6	6.6	-1.6
Cavalier	9.6	0.04	46	0.08	33	10.1	11.9	2.2
Dickey	12.8	0.02	52	0.05	38	10.6	13.5	3.4
Divide	15.0	0.001	64	0.003	51	11.2	16.0	5.4
Dunn	9.4	0.003	59	0.01	46	7.8	10.5	4.3
Eddy	15.6	0.002	62	0.005	49	11.3	15.9	5.1
Emmons	11.1	0.0006	68	0.03	42	8.6	12.6	5.7
Foster	15.4	0.0001	84	0.0001	70	10.5	15.4	5.8
Golden Valley	6.1	0.93	5	0.97	-9	9.6	8.7	-1.2
Grand Forks	10.4	0.74	12	0.22	11	7.2	6.8	-0.7
Grant	11.0	0.0001	87	0.0001	74	8.4	14.4	8.9
Griggs	12.0	0.002	62	0.02	50	11.1	15.2	4.6
Hettinger	10.5	0.0001	82	0.0004	68	8.6	12.6	5.8
Kidder	11.6	0.0001	83	0.0001	71	8.3	12.8	6.7
Lamoure	11.3	0.03	47	0.02	45	9.9	11.8	2.4
Logan	10.9	0.0001	71	0.003	57	8.5	12.4	5.8
McHenry	9.6	0.26	28	0.62	2	10.0	10.5	0.7
McIntosh	16.4	0.0002	71	0.02	45	10.9	17.5	7.5
McKenzie	10.6	0.07	42	0.05	29	8.4	9.7	1.9
McLean	11.2	0.0001	89	0.0001	76	9.3	13.4	5.5
Mercer	8.0	0.008	56	0.03	42	7.6	9.4	3.0
Morton	9.3	0.002	62	0.005	49	8.3	9.4	1.7
Mountrail	11.5	0.03	48	0.07	36	10.8	12.9	2.4
Nelson	16.7	0.003	63	0.009	50	13.4	17.6	3.8
Oliver	7.0	0.29	27	0.31	15	5.7	6.4	1.7
Pembina	8.8	0.18	33	0.26	19	10.6	11.1	0.6
Pierce	12.5	0.0001	87	0.0001	85	9.5	15.0	7.2
Ramsey	12.2	0.0005	67	0.005	54	10.2	12.6	3.0
Ransom	13.9	0.0001	75	0.001	63	11.8	15.4	3.8
Renville	10.7	0.09	40	0.35	13	9.6	10.7	1.4
Richland	7.4	0.62	16	0.74	2	8.9	8.6	-0.4
Rolette	8.6	0.33	25	0.69	-1	9.0	9.9	1.1
Sargent	7.5	0.91	5	0.84	-7	9.7	8.4	-1.6
Sheridan	9.9	0.09	38	0.17	24	8.5	9.5	1.4
Sioux	15.1	0.73	12	0.22	11	9.1	8.5	-0.9
Slope	9.1	0.38	33	0.58	7	6.5	7.4	1.8
Stark	8.6	0.0001	90	0.0001	78	7.6	9.4	2.9
Steele	12.9	0.55	18	0.48	6	9.9	9.6	-0.3
Stutsman	9.8	0.0001	82	0.0002	69	8.9	10.8	2.7
Towner	12.4	0.009	55	0.02	43	10.2	12.6	2.8
Traill	10.5	0.07	41	0.38	14	11.1	12.7	1.8
Walsh	11.5	0.006	58	0.03	44	10.4	12.4	2.4
Ward	7.6	0.002	62	0.02	50	7.4	8.5	1.8
Wells	13.3	0.002	62	0.007	48	10.6	14.2	4.2
Williams	9.0	0.02	53	0.15	26	8.8	9.8	1.5

In the second column of Table 5.20 the mortality rate cut-point, determined by UniODA, is the value which most accurately separates years before versus after 1998. Cut-points with low values indicate the county has a low annual crude mortality rate, and cut-points with high values indicate the county has a high rate. For the US the current annual crude mortality rate—the number of deaths per 1,000 people—is 8.0, so as seen, 43 of 53 (81%) cut-points exceed the national rate.⁸²

The third column (training $p <$) is Type I error for the *a priori* hypothesis, for the UniODA model. Values indicated in **bold** are statistically significant ($p < 0.05$) at the experimentwise criterion. There were statistically significant effects at the experimentwise criterion for 16 counties, and at the generalized criterion for 18 counties. Because 2.65 statistically significant models are anticipated under the null hypothesis these results reflect a 13-fold greater effect than is expected by chance.

Simply because something rarely occurs by chance doesn't imply that if chance does strike and something does happen—that it happens well. An event may indeed be rare, and still not be very good in an absolute sense. The fourth column, training *ESS*, is a normed index of effect strength: as seen, of the 53 counties 11 (21%) had strong effects, 20 (38%) had relatively strong effects, 13 (25%) had moderate effects, and 9 (17%) had relatively weak effects.

Simply because something happens now is no guarantee that it will happen again in the future. LOO validity analysis gives an upper-bound estimate of expected cross-generalizability of a model when it is applied to classify an independent random sample. In this application LOO analysis estimates the generalizability of the model across time—for predicting future annual crude mortality rates.

In column five (LOO $p <$) **bold** values are statistically significant at experimentwise $p < 0.05$. Ten counties had statistically significant effects at the experimentwise criterion, and another 20 counties at the generalized criterion, reflecting an 11-fold greater rate than expected by chance.

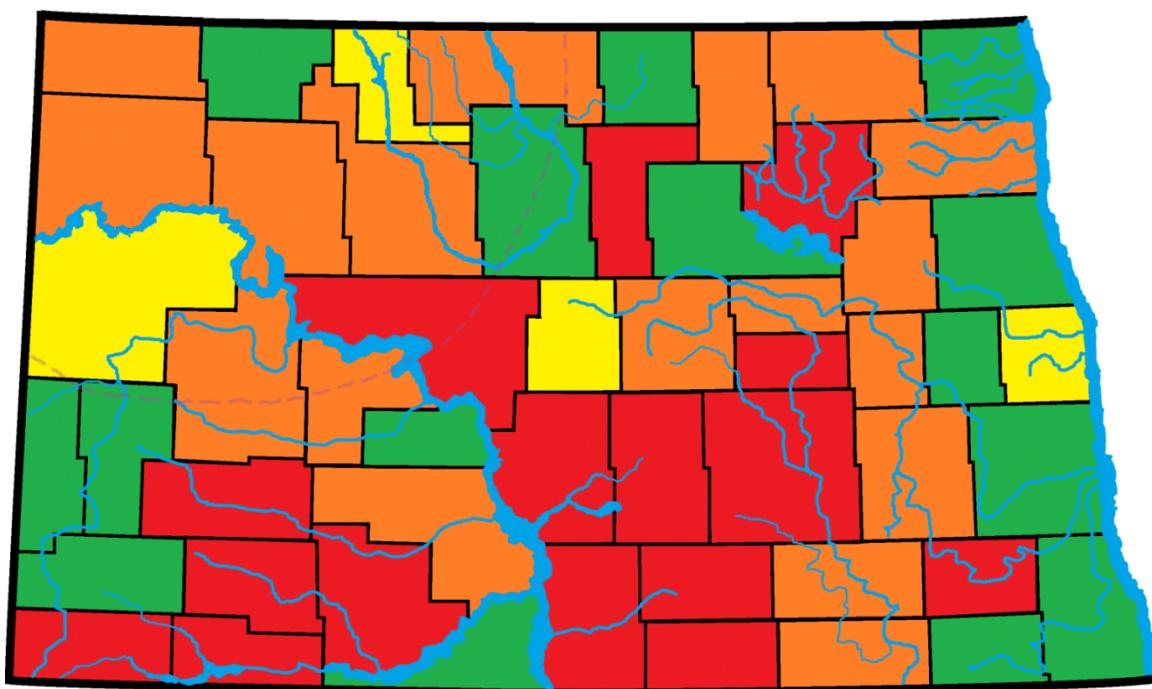
Column six gives LOO *ESS* for the *a priori* hypothesis. Of 53 counties: 3 (6%) had strong effects; 15 (28%) had relatively strong effects; 16 (30%) had moderate effects; 13 (25%) had relatively weak effects; and models of 6 (11%) counties with *negative ESS* values are expected to yield accuracy in future predictions that is lower than anticipated by chance.

The seventh and eighth columns give the mean annual crude mortality before versus after 1998, respectively. Consistent with the findings regarding high cut-point values, mean crude mortality rate in 43 of 53 (81%) counties exceeded the current national average before 1998, as did 48 (91%) counties after 1997. The final column is the annual mean percentage increase in crude mortality rate for the eight years of data since 1997, that were available for analysis. The annual mean crude mortality increase *exceeds 5% per year* for 12 of 53 (23%) counties.

Figure 5.10 presents a map illustrating findings reported in the third column of Table 5.20. The map indicates statistical significance by color: red, orange and yellow respectively reflect experimentwise, generalized and marginal significance, with no significance indicated by green. Also illustrated (not to scale) is the Bakken formation, indicated using a light, broken purple line (see the fourth county from left-hand-side at top of map, and third county from top of map on left-hand-side). Waterways are not drawn to scale. As hypothesized, data clearly revealed a significantly higher annual crude death rate in North Dakota after commercialization of toxic fracking liquid occurred in 1998. It is interesting that a cluster of counties having exceptionally strong effects is seen in the lower west and lower center areas of the state, where there is a west wind component for 9 out of 12 months, and a north wind component for 9 of 12 months. The prevailing wind is pointed directly toward these counties from the producing area, and the affected counties are also down river from the energy-producing region.⁸³

In the absence of data which correspond to the mortality rate series, it wasn't possible to adjust rates for age, or to rule-out an aging population as a competing hypothesis. Data indicate that since 2000 the population (more than 90% white) of North Dakota is aging, and illicit drug use has been increasing as well.⁸⁴ Interestingly, in some primary producing regions the median age *decreased* as young people came to pursue boom-town opportunities. Anecdotal reports credit influx of young people, combined with frontier culture (long hard work; surplus money, scarce, and inadequate resources—including health and medical care; and social as well as geographic isolation), as an explanation underlying noted increases in the numbers of assaults, homicides, and vehicular accidents.

Figure 5-10: Statistical Reliability of Increase in Annual Crude Mortality Rate After 1997

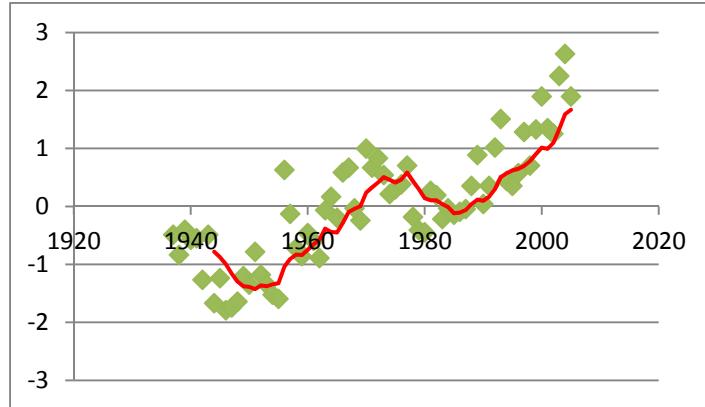


Powerful improvements of the present map are possible. By presenting the results of only one statistical analysis the present map is static, like a snapshot. However, the findings of successive analyses can be integrated and used to illustrate changes occurring across time, creating a dynamic map like a motion picture. Digital maps can be interactive, with pointers used to select a county or counties, and to operate menu-driven information systems. Serial data available for additional attributes should be easily selectable for UniODA (illustrated presently) and CTA analysis, and between-group and within-subject (illustrated presently) analyses should both be available. Information content of the map can be increased (and new phenomena discovered) if multiple attributes are available for analysis, because the accuracy of statistical models increases if multiple predictors of the class variable are identified. Thus, every county should be simultaneously coded on multiple attributes, rather than on only one attribute (presently, outcome of the test of the *a priori* hypothesis). A dynamic multivariate display would help to identify and understand interactions among variables and time. Some applications which might be fruitfully addressed by these methods include political polls; crime surveys; weather phenomena; agricultural and mining inventories; spread of infectious disease; power consumption; sports team performance; and economic series such as money market inflows, consumer debt, unemployment, number of jobs created, consumer sentiment, retail sales, durable goods orders, manufacturing/purchasing indices, inflation, major market indices, and leading indicators, for example.

Longitudinal measurement is widely practiced by people in all facets of real-world phenomena: an athlete wishes to know how many training sessions are needed to achieve a significant performance increase; a musician wishes to assess how many on-stage performances are required before a significant increase in media attention is attracted; and a medical patient wishes to understand how many units of medication are needed before significant symptom improvement is realized. UniODA is easily adapted to statistically evaluate a longitudinal series by starting at the series beginning and traveling forward in time, or starting at the series end and travelling backward in time. The following example illustrates backwards travel in a time-ordered series, to determine when recently observed statistically significant increases in annual crude mortality rate occurred in McLean County, North Dakota.⁸⁵

Figure 5.11 shows the annual crude mortality rate (ACMR) over time for McLean County in North Dakota—an active area of productive oil and gas fields. The 8-year forward-moving-average is illustrated because data for a total of eight “post-commercialization” years are available for analysis.

Figure 5.11: 8-Year Forward-Moving-Average Ipsative ACMR Data for McLean County, North Dakota



As seen, ipsative ACMR began its most recent rise in 1984. The first test of the *a priori* hypothesis compared post-commercialization data (1998 - 2005; Class = 1) against *all* prior data (1937 - 1997; Class = 0). In contrast, here the number of pre-event data points used to test the *a priori* hypothesis is the same as the number of post-event data points used, in order to equate the time horizon on both sides of the cutpoint used to define class categories. The first analysis compared eight years of post-commercialization ipsatively-standardized ACMR data (1998 - 2005; class = 1) versus the preceding eight years (1990 - 1997; class = 0), testing the confirmatory hypothesis that class 1 observations have higher ACMR values. Analysis was accomplished using the following UniODA and MegaODA software syntax:

```
OPEN acmr.dat;                                ATTR acmr;
OUTPUT acmr.out;                               DIR < 0 1;
VARS class acmr;                             MC ITER 25000;
CLASS class;                                 GO;
```

Data are presented, and UniODA findings are summarized, for four backward-stepping analyses in Table 5.21. The UniODA model for the first analysis is: if ipsative ACMR ≤ 0.64 then predict class = 0 (1990 - 1997); otherwise predict class = 1 (1998 - 2005). This model had relatively strong sensitivity in classifying actual year correctly (*ESS* = 62.5), and strong predictive value in making accurate classifications into each class category (*ESP* = 72.7). Looking backwards in time, the *most recent statistically significant increase* in the McLean County ipsative ACMR series occurred when comparing eight post-commercialization years (1998 - 2005) versus eight preceding years (1990 - 1997).

This backward-stepping procedure continues until the first *non-statistically-significant* result is obtained. Table 5.21 provides the sequentially-rejective Sidak criterion for experimentwise $p < 0.05$ for each step in this procedure, as well as the actual p obtained for the model. The procedure was completed in the third step: the model p (0.043) exceeds the target p (0.01696), so the *a priori* hypothesis is rejected.

In the context of the objective of the study, looking backwards in time the second most recent statistically significant increase in the McLean County ipsative ACMR series occurred when comparing data from 1997 - 2004 versus 1989 - 1996. Because the following (third) test was not statistically reliable under the experimentwise criterion, the second effect was the *earliest*, or the *initial* statistically significant increase in ipsative ACMR data in the recent series.

Because the bin-width of the present study is small ($N = 16$ observations), the statistical power available for testing the *a priori* hypothesis is limited. In this case, especially if data are expensive and/or

rare, the use of the multiple comparisons criterion may be relaxed, and the generalized criterion of per-comparison $p < 0.05$ can be used instead to gauge statistical significance. This also is illustrated in Table 5.21. Note that in the third analysis the estimated p (0.043) is less than the estimated p (0.044) in the first analysis, but neither of these is smaller than the critical Sidak value for the second comparison. Had the generalized criterion been used rather than the experimentwise criterion from the beginning (step 1) of the analysis then the null hypothesis would have been rejected in this step. However, in the fourth analysis p is clearly not statistically significant.

Table 5.21: Data and Findings for Backward-Stepping Analysis

Year	ACMR	Class Variable Coding			
		Analysis 1	Analysis 2	Analysis 3	Analysis 4
1987	-0.04				0
1988	0.35			0	0
1989	0.89		0	0	0
1990	0.04	0	0	0	0
1991	0.35	0	0	0	0
1992	1.02	0	0	0	0
1993	1.51	0	0	0	0
1994	0.44	0	0	0	0
1995	0.35	0	0	0	1
1996	0.58	0	0	1	1
1997	1.29	0	1	1	1
1998	0.71	1	1	1	1
1999	1.33	1	1	1	1
2000	1.90	1	1	1	1
2001	1.35	1	1	1	1
2002	1.25	1	1	1	1
2003	2.25	1	1	1	
2004	2.63	1	1		
2005	1.90	1			
Model Cut-Point		0.64	1.14	0.51	0.51
Sidak p		0.05	0.02533	0.01696	0.01275
Model p		0.044	0.0092	0.043	0.13
<i>ESS</i>		62.5	75.0	62.5	50.0
<i>ESP</i>		72.7	75.0	72.3	53.3

In the context of the objective of the analysis, the remaining questions have been answered: looking backwards in time, the second most recent statistically significant increase in the McLean County ipsative ACMR series occurred comparing data from 1997-2004 versus 1989-1996. The earliest, or initial statistically significant increase in ipsative ACMR data in the recent series occurred comparing data from 1996-2003 versus 1988-1995. The comparison of the 1995-2002 and 1987-1994 data was *not* statistically significant. This “little Jiffy” procedure may be used in a conceptually-parallel manner to conduct analysis in a forward-stepping mode, initiating from the beginning of the series.

ROC Analysis

Receiver operator characteristic (ROC) analysis is used to discriminate a dichotomous class variable by identifying an “optimal” discriminant cutpoint on an ordered attribute.⁸⁶ The *ROC curve* is a plot that displays *sensitivity* (“true positive rate”) on the ordinate, and displays $1 - \text{specificity}$ (“false positive rate”) on the abscissa, for all possible threshold values (cut-points) that separate class 0 and class 1 observations in the sample. The total area under the ROC curve (*AUC*) is used as an index of the discriminative validity of scores on the attribute: the greater the *AUC* value, the greater the ability of scores on the attribute to discriminate the two class categories for the sample (*AUC* isn’t normed against chance). In ROC analysis the distance d between the point representing perfect classification and any point on the ROC curve is: $d = \sqrt{[(1 - s_n)^2 + (1 - s_p)^2]}$, where $s_n = \text{sensitivity}$ and $s_p = \text{specificity}$. In ROC analysis the optimal cutpoint for discriminating the class categories is defined as the threshold value associated with the minimum value of d . In UniODA the optimal threshold is defined as the cutpoint maximizing the value $(s_n + s_p) / 2$, that yields the maximum possible *ESS* for the sample. Because $\sqrt{[(1 - s_n)^2 + (1 - s_p)^2]}$ and $(s_n + s_p) / 2$ aren’t isomorphic, ROC analysis and UniODA mustn’t identify identical optimal discriminant threshold values for a sample.

These competing methods are illustrated for an application predicting Cesarean delivery (class variable) based on duration of membrane rupture (attribute) for 166 hospitalized women^{87,88} (Table 5.22).

Table 5.22: Computing the Optimal Cut-Point Value by ROC Analysis versus UniODA

	Cutpoint	Sensitivity	1 – Specificity	Specificity	Distance	ESS
ROC	0.00	1.000	1.0000	0.0000	1.00000	0
	0.63	1.000	0.9760	0.0240	0.97600	2.40
	0.88	1.000	0.9690	0.0310	0.96900	3.10
	1.25	1.000	0.8900	0.1100	0.89000	11.00
	1.75	1.000	0.8660	0.1340	0.86600	13.40
	2.13	1.000	0.8190	0.1810	0.81900	18.10
	2.38	1.000	0.8110	0.1890	0.81100	18.90
	2.75	1.000	0.7800	0.2200	0.78000	22.00
	3.25	1.000	0.7170	0.2830	0.71700	28.30
	3.75	1.000	0.7090	0.2910	0.70900	29.10
	4.50	1.000	0.6460	0.3540	0.64600	35.40
	5.13	0.971	0.5830	0.4170	0.58372	38.80
	5.38	0.971	0.5750	0.4250	0.57573	39.60
	5.75	0.971	0.5510	0.4490	0.55176	42.00
	6.25	0.914	0.3780	0.6220	0.38766	53.60
	6.75	0.914	0.3460	0.6540	0.35653	56.80
	7.13	0.857	0.2910	0.7090	0.32424	56.60
	7.38	0.857	0.2830	0.7170	0.31708	57.40
	7.75	0.857	0.2760	0.7240	0.31085	58.10
	8.25	0.800	0.1892	0.8108	0.27531	61.08
ODA	8.75	0.800	0.1810	0.8190	0.26974	61.90
	9.25	0.743	0.1100	0.8900	0.27955	63.30
	9.75	0.743	0.1020	0.8980	0.27650	64.10
	10.25	0.543	0.0390	0.9610	0.45866	50.40
	10.75	0.543	0.0310	0.9690	0.45805	51.20
	11.50	0.457	0.0240	0.9760	0.54353	43.30
	12.50	0.400	0.0080	0.9920	0.60005	39.20
	13.50	0.343	0.0000	1.0000	0.65700	34.30
	14.50	0.286	0.0000	1.0000	0.71400	28.60
	15.50	0.257	0.0000	1.0000	0.74300	25.70
	16.50	0.200	0.0000	1.0000	0.80000	20.00
	17.50	0.171	0.0000	1.0000	0.82900	17.10
	18.50	0.143	0.0000	1.0000	0.85700	14.30
	19.50	0.114	0.0000	1.0000	0.88600	11.40
	20.25	0.057	0.0000	1.0000	0.94300	5.70
	21.50	0.000	0.0000	1.0000	1.00000	0

Table 5.22 gives every possible cutpoint value (ranging from 0 to 21.5) separating a class 0 and class 1 observation in the sample, and the corresponding values of sensitivity, 1-specificity, specificity, the ROC distance measure d , and the UniODA normed accuracy measure ESS . As seen, the minimum distance d is 0.27 (shown in **bold**), corresponding to an optimal cutpoint of 8.75 for ROC analysis. The optimal ESS value is 64.1 (shown in **bold**), corresponding to an optimal cutpoint of 9.75 for UniODA. ESS is 61.9 for the ROC cutpoint, yielding 3.4% lower classification performance than obtained using UniODA. These results demonstrate that ESS achieved using this ROC analysis approach is *not* explicitly optimal for the sample.

Differential costs of both types of *miscalifications* are important in some applications—such as the diagnosis of disease, for example. Indeed, the same point may also be raised regarding the differential benefit of both types of *correct* classifications. In the ODA paradigm every application and hypothesis (not only ROC analysis) can be weighted using any quantitative index (e.g., desirability, valence, threat, fear, cost, return, price, distance, time, mass). Different weights may be assigned to different class categories: for example, class 0 observations can be overweighted using increasingly large weights (e.g., successive integers) until the desired level of specificity is attained. The same procedure may be used to overweight class 1 (positive) observations and thereby maximize sensitivity. Weights can also be applied individually to each observation in the study—since not all people have the same weighting priorities. And, a set of numerical weights for each observation can be multiplied, in order to model the overall interactive effect of the profile of weights. For applications involving multiple weights for multiple outcomes, multi-criteria decision-making analysis may help to identify the omnibus optimal solution.⁸⁹

Student's *t*-Test

Among the most frequently reported statistical tests, Student's *t*-test is a legacy procedure for analyzing data involving a binary class variable and an ordered (assumed to be continuous) attribute. It is simple to construct a hypothetical problem for which *t*-test fails to find a significant intergroup mean difference on the attribute, while UniODA detects nearly perfect intergroup discriminability: imagine that ten class A observations score a value of 0 on the attribute; nine class B observations score a value of 1, and a tenth class B observation scores a value of -9. Because the mean difference on the attribute between groups is zero, *t*-test would conclude that the groups can't be discriminated on the basis of the attribute. But, with UniODA, *Overall PAC* = 95% and ESS = 90 indicate early perfect intergroup discriminability.

Between-Subjects

Consider a study comparing the number of migraine attacks experienced in a clinical trial of two alternative treatments, for a sample of 67 patients (Table 5.23).

Table 5.23: Data for Clinical Trial of Two Migraine Treatments

Number of Attacks	Treatment 1	Treatment 2
0	13	5
1	9	13
2	4	6
3	2	1
4	1	2
5	1	3
6	3	3
7	0	1

Note: Tabled are frequency counts.

These data were statistically analyzed using a variety of legacy methods.⁹⁰ Student's *t*-test was used to compare the mean number of attacks between treatments, but the result wasn't statistically significant ($p < 0.14$). The *t*-test was employed again to compare the data after modification using a square

root transformation ($p < 0.06$), and then a log transformation ($p < 0.07$): t still failed to identify a reliable mean difference, and crucial assumptions underlying the validity of the t -test could not be met. The non-parametric Mann-Whitney U test yielded $p < 0.07$. A normal test assuming a Poisson distribution was tried next, and it identified a statistically significant effect ($p < 0.04$), but the Poisson assumption was untenable. It was decided to try Fisher's exact test so a method to parse (pre-process) the number of migraine attacks into a binary indicator was needed: "...discretizing the data at some point, probably between no attacks and one or more" (p. 242): for an arbitrary cut-off between 0 and 1, $p < 0.022$.

In contrast to conventional statistical methods, with UniODA one has no concerns about parent distributions (p is always exact), or about where to parse an ordered attribute (UniODA always maximizes accuracy). Analysis was performed with the following UniODA and MegaODA software syntax (the class variable was treatment group coded as 1 or 2; the ordered attribute was number of migraine attacks)⁹¹:

OPEN DATA;	2 1 (repeated 13 times)
OUTPUT migraine.out;	2 2 (repeated 6 times)
VARS group attacks;	2 3
DATA;	2 4 (repeated 2 times)
1 0 (repeated 13 times)	2 5 (repeated 3 times)
1 1 (repeated 9 times)	2 6 (repeated 3 times)
1 2 (repeated 4 times)	2 7
1 3 (repeated 2 times)	END;
1 4	CLASS group;
1 5	ATTR attacks;
1 6 (repeated 3 times)	MCARLO ITER 25000;
2 0 (repeated 5 times)	LOO;
	GO;

Consistent with the eyeball parse, the UniODA model identified was: if Number of Attacks > 0 predict class = Treatment 2; otherwise predict class = Treatment 1. The relatively weak $ESS = 24.7$ was marginally significant, $p < 0.085$. The model correctly classified 13 (39%) of the 33 patients actually in treatment 1, and 29 (85%) of 34 patients actually in treatment 2. The model was correct 72% of the time it predicted a patient was in treatment 1, and 59% of the time it predicted a patient was in treatment 2. The stable LOO classification performance was statistically significant ($p < 0.022$) because in LOO analysis a UniODA model is applied directionally.

As a second example consider a study predicting susceptibility to sudden cardiac death (SCD) as a function of heart rate variability (HRV). To test the *a priori* hypothesis that people with depressed HRV are more susceptible to SCD, a continuous measure of HRV known as the Singer score was computed for each person (higher Singer scores indicate lower HRV).³⁷ Comparing mean Singer score between SCD groups using t -test failed to identify a statistically significant difference.⁹² Using UniODA in this application the *a priori* alternative hypothesis is that people who are susceptible to SCD (class variable) should have higher Singer scores (continuous attribute) than people who aren't susceptible to SCD: the null hypothesis is that this is not true. Data entered in free format were the class code *scd* (susceptible = 1; not susceptible = 0) and the continuous attribute *Singer*. UniODA and MegaODA software syntax that performed analysis was:

OPEN singer.dat;	ATTRIBUTE singer;
OUTPUT singer.out;	DIRECTIONAL > 1 0;
VARS scd singer;	MCARLO ITER 25000;
CLASS scd;	GO;

The resulting ODA model was: if Singer $\leq .0796$ then predict the person is not susceptible to SCD, otherwise predict the person is susceptible to SCD. Classification performance was stable in LOO analysis,

was relatively strong ($ESS = 50.0$), and was statistically significant ($p < 0.0037$). Consistent with the *a priori* hypothesis, higher Singer scores, indicative of lower HRV, are predictive of increased susceptibility to SCD.

Within-Subjects

Our interest in the application discussed here derives from findings of content analysis of free-form comments, voluntarily recorded by 339 American patients with fibromyalgia (FM) over a 14-week period, who used our web-based, interactive, self-monitoring and feedback system (SMART) during its alpha test as a behavioral CAM intervention for FM.^{93,94} Of 2,215 discrete comments made in total, 1,732 (78%) involved symptoms, and 244 (11%)—*one of nine comments made*—mentioned weather-related phenomena, often remarking that worse weather exacerbated symptoms, or that improved weather reduced symptoms. A total of *twenty-two independent comments* stated that specific weather events triggered *clinically significant symptom flares*.

A high prevalence of weather-related symptoms exists even for people who are well-adapted to challenging climatic conditions: in a study of cold-related complaints of 8,723 Finns aged 25-64 years, for example, 75% reported decreased mental or physical performance, and 33% reported musculoskeletal pain.⁹⁵ It is intuitive that extreme weather is especially challenging to people with FM, as diminished concentration and physical ability, and musculoskeletal pain, are *prevalent* symptoms of FM. Concurrent research conducted in the USA studied the relationship between changes in weather and symptoms of FM patients. A survey of 94 patients found prevalent reporting of modulation of aches and pains by weather factors, especially among young patients.⁹⁶ A study of 84 patients found subjects believed weather predominantly affected their musculoskeletal symptoms, and that higher weather sensitivity is associated with greater functional impairment and psychological distress.⁹⁷ And, an internet survey of 2,596 patients reported the most common aggravating factors for symptoms were weather changes, emotional distress, insomnia, and strenuous activity.⁹⁸ International research findings have been consistent. For example, 17 patients in Argentina completed surveys assessing the presence and features of spontaneous daily pain occurring over a one-year period, and same-day barometric pressure and temperature were significant correlates of pain ratings.⁹⁹ A retrospective cross-sectional study of 955 rheumatic patients in Portugal reported FM patients were strongly influenced by weather change.¹⁰⁰ And, nearer the equator where meteorological variation is lower, a clinical interview of 15 female patients in Brazil revealed that climate variation was uniformly considered to be a trigger event, as well as a modulating factor, for pain.¹⁰¹

Prospective research also examined the relationship between weather and symptoms of FM patients. In a one-month study of pain ratings and changes in daily weather conditions in Israel, pain was significantly related to barometric pressure for 11 patients.¹⁰² Longitudinal survey assessment of seasonal symptoms was conducted for 1,424 patients with rheumatic disease in the USA, and were associated with seasonal weather differences measured for periods up to 24 years; number and severity of weather-related symptoms were elevated in patients with FM.¹⁰³ Finally, daily pain ratings of 55 female patients in Norway were recorded for 28 days, and related to weather parameters and a composite weather variable using time series analysis: no association between same- or prior-day weather and pain was found, but *post hoc* analysis found patients with less than ten years of symptoms had significantly greater weather sensitivity than patients with longer illness.¹⁰⁴ In light of the personal, family and societal costs of FM,¹⁰⁵ the abundant qualitative evidence that weather change plays an important precipitating and modulating role in FM symptom flares, and the paucity of actionable findings, further study in this area is clearly warranted. This example thus considers the appropriate statistical methodology for assessing the relationship between weather and individual symptoms.

Atmospheric pressure was assessed as 500 mb geopotential height anomaly (GHA) measured in meters.¹⁰⁶ This height is proportional to the mean temperature of the air column extending from a point on the Earth's surface to approximately 18,000 feet: the 500 mb GHA is the amount above or below mean height for that point and time. GHA is more appropriate than barometric pressure presently because of the broader geographic expanse of the features that GHA defines, given imprecise information available concerning the location of subjects and the time their symptoms occurred.

Patient symptom ratings were obtained using the SMART system: individuals rate their condition (maximum frequency is once daily) across time on ten prevalent FM symptoms using 10-point Likert-type

scales (mean = 3.5 entries/patient/week).⁹⁴ Review of the database yielded a total of 11 individual patient records meeting three inclusion criteria: symptom ratings unequivocally organized by entry date; weather conditions unequivocally identified on entry date; and statistical power analysis mandated a minimum of 48 ratings.⁹³ Table 5.24 gives the Kendall tau b correlation coefficient (and p) for GHA and symptom, both assessed on the same day, *separately by patient*. Tau was selected because the distributional assumptions underlying parametric methods were unsupported empirically.

Table 5.24: Single-Case Statistical Analysis of the Relationship between GHA and Physical Symptoms for 11 FM Patients

Patient #	1	2	3	4	5	6	7	8	9	10	11
State of Residence	ID	AZ	OR	CA	MN	TX	OH	TN	TN	CA	PA
Days Reported	99	98	89	79	75	62	59	56	52	52	51
<i>Pain</i>	0.09 0.31	0.14 0.08	-0.02 0.84	0.11 0.20	-0.20 0.007	-0.06 0.51	-0.01 0.91	-0.15 0.17	-0.18 0.10	0.09 0.40	-0.05 0.61
<i>Stiffness</i>	0.10 0.24	-0.01 0.87	-0.08 0.32	0.22 0.008	-0.17 0.022	-0.01 0.93	0.01 0.93	0.12 0.91	-0.18 0.08	0.05 0.68	-0.14 0.19
<i>Fatigue</i>	-0.02 0.78	-0.01 0.92	-0.07 0.39	0.06 0.51	-0.20 0.008	-0.05 0.60	-0.20 0.05	0.09 0.36	-0.12 0.25	0.01 0.99	0.08 0.46
<i>Mental Focus</i>	-0.06 0.48	0.17 0.04	0.17 0.03	0.09 0.31	-0.27 0.001	0.07 0.48	-0.13 0.20	0.19 0.08	-0.17 0.11	0.06 0.60	-0.01 0.93
<i>Memory</i>	-0.03 0.72	0.10 0.23	0.16 0.05	-0.03 0.73	-0.24 0.002	0.10 0.29	-0.13 0.21	0.12 0.27	-0.13 0.19	0.01 0.99	0.03 0.78
<i>Anxiety</i>	-0.27 0.002	0.01 0.96	0.08 0.32	0.12 0.18	-0.19 0.018	0.11 0.25	0.01 0.94	-0.04 0.73	0.01 0.94	0.04 0.73	0.06 0.61
<i>Depression</i>	-0.30 0.001	0.21 0.01	-0.07 0.42	-0.07 0.39	-0.38 0.001	0.10 0.31	-0.01 0.99	0.11 0.33	-0.11 0.30	-0.01 0.97	-0.05 0.65
<i>Gastro-intestinal</i>	-0.14 0.07	-0.04 0.63	0.06 0.49	0.06 0.47	0.05 0.55	0.07 0.49	-0.01 0.90	-0.15 0.17	-0.14 0.19	-0.22 0.05	0.02 0.83
<i>Sleep Difficulties</i>	-0.17 0.042	0.06 0.47	0.11 0.21	-0.01 0.90	-0.36 0.001	0.04 0.68	-0.10 0.33	0.18 0.09	-0.10 0.33	0.07 0.53	-0.04 0.73

Note: Tabled separately by patient (columns) and symptoms (rows) is Kendall Tau b coefficient (top value in cell) for same-day GHA and the indicated symptom, and the corresponding p value (bottom value). State of residence and number of data ratings (days) are indicated in the column heading for each patient. Results in **bold** are statistically reliable (generalized $p \leq 0.05$). For 99 tests of statistical hypotheses, 4.95 "statistically significant" ($p \leq 0.05$) effects are expected by chance: the total of 18 reliable effects observed is 3.6 times greater than is expected by chance.

Considering the inherent non-stationary non-linearity of both daily symptoms and GHA across time, tau found a surprisingly high number of statistically reliable, ecologically non-trivial associations of symptoms and weather. As expected, the number of reliable coefficients decreased with sample size: statistical power is reduced, and smaller N yield lower variability, limiting the maximum magnitude that correlation indices may attain.⁴⁵ Seven of 11 patients (64%) had at least one reliable association, and 9 of 11 patients (82%) had at least one marginal association. Patient 5 showed extreme reactivity to GHA (every symptom except gastrointestinal was associated with GHA), and patients 6 and 11 had no GHA reactivity. Comparing symptoms, mental focus and depression were most reactive (three patients each had statisti-

cally reliable associations). Associations were generally strongest for depression, and gastrointestinal difficulties were least reactive to GHA. For patient 5 the coefficient for sleep difficulties was twice as large as for any other patient for this symptom.

Results show *GHA influenced symptom(s) of most people studied*. Results also demonstrate that people are not uniform in their responses to weather, but rather that one's reaction to weather change is *intrapersonal* and *idiosyncratic*. For example, compare the reliable *negative* depression coefficient found for patient 1, and the reliable *positive* depression coefficient identified for patient 2.

It is important to note that if the depression data for patients 1 and 2 are combined the resulting coefficient is tau = -0.04, $p \leq 0.49$. Combining the data of these two patients eliminates both effects found for depression—masking effects which in reality existed: clearly, combining *intrapersonal* data can induce Simpson's Paradox (see Chapter 9). *The current standard operating methodology in research design and statistical analysis involving symptom data involves combining subjects into groups, a convention that can obviously induce paradoxical confounding*.

Patients sometimes recorded recurring symptoms (e.g., migraine headache) in their SMART comment log. Patient (MN, 75) was selected to demonstrate the use of ODA in this application. For this patient, when a headache was reported the record (day) was coded positive for headache: all other records were coded negative for headache. First, UniODA was used to predict if the patient was positive or negative for headache on a given day (class variable), based on the GHA value for that day (ordered attribute). Figure 5.12 illustrates the resulting model.¹⁰⁷

Figure 5.12: Headache UniODA Model

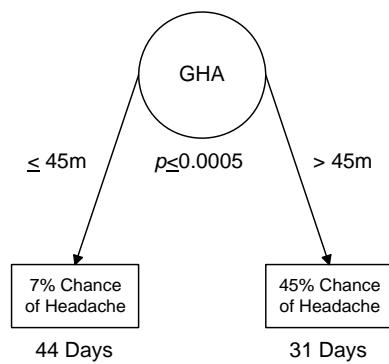
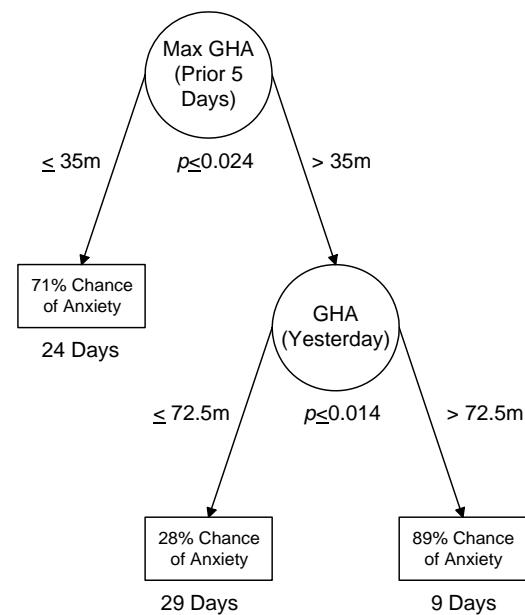


Figure 5.13: Anxiety CTA Model



In Figure 5.12 the attribute is indicated by the node (circle); the cutpoint that is used to predict headache status (identified by UniODA, this threshold maximizes predictive accuracy) is indicated next to arrows (pathways through the model); the exact Type I error rate (p value) is given beneath the attribute; and model endpoints (sample strata) are indicated using rectangles. Endpoints give the likelihood that the patient will report a headache, and the number of days represented by the endpoint. There is more than a 6-fold difference in the likelihood of a migraine for this patient, based simply on whether the GHA value for the day is 45m or less (7% likelihood of headache), versus greater than 45m (45% likelihood).

Classification tree analysis (CTA) chains UniODA models together in order to create multiattribute non-linear models (see Chapters 10-12). In the prior (UniODA) example the patient was *highly reactive* to GHA in tau analysis. In contrast, daily anxiety self-ratings of patient (TX, 62)—who was *nonreactive* to GHA in tau analysis, was selected to illustrate CTA. When this patient recorded anxiety greater than median for

her time series, the record (day) was coded *positive for anxiety*. All other records were coded *negative for anxiety*. CTA was used to predict if the patient was positive or negative for anxiety on a given day (class variable): weights reflected increasing deviation from median. Attributes available for predicting anxiety for any day in the patient's time-series were GHA value for the day; GHA value on the prior day ("yesterday") as well as two, three and four days ago; and minimum and maximum GHA over the prior five days.

Figure 5.13 presents the model that was obtained for the patient by enumerated CTA (Chapter 11): as seen, when maximum GHA (prior 5 days) is $\leq 35m$ there is a 71% likelihood that the patient will report high anxiety; when maximum GHA (prior 5 days) is $> 35m$ and GHA yesterday is $> 72.5m$ there is a 89% likelihood the patient will report high anxiety; and if GHA yesterday is $\leq 72.5m$ there is only a 28% likelihood the patient will report high anxiety. Because this is a *U-shaped* association—anxiety increases as GHA deviates from median, it is thus not surprising that tau—a linear model, failed to identify the effect. In LOO analysis the CTA model correctly classified 25 of 33 (76%) high anxiety days, and 21 of 29 (72%) low-anxiety days: weighted *ESS* = 48, nearly satisfying the criterion for a relatively strong effect. For this patient the CTA model provides greater than a 3-fold improvement in accurate forecasting of anxiety based on the value of GHA.

These methods may be used in single-case statistical analysis of *any serial measure* in relation to *any binary outcome*.

Validity Analysis

One meaning of the term "validity" as used herein refers to reproducibility of the estimated classification performance achieved for a sample by an ODA model.¹⁰⁸ Discussed in Chapter 2 and illustrated in many examples, ODA software (UniODA, MegaODA, CTA) facilitates estimating potential cross-generalizability of ODA models vis-à-vis LOO (jackknife) and hold-out (independent random samples are used to test the training model) functionalities. Also discussed in Chapter 2 and demonstrated using many examples, the use of the Gen (multisample) algorithm ensures that the maximum minimum classification performance (i.e., validity estimate) is obtained when using one model independently with multiple samples measured on identical variables.

A second meaning of the term validity used herein derives from the area of psychometrics, and it speaks to the appropriateness of a label or name used to describe a measure (i.e., class variable, attribute, weight, and corresponding measurement scales) by the researcher.^{27,35-37,109} Consider the attribute head circumference: if labeled as a measure to select hat size, no one will dispute the validity of that assertion; if labeled as a measure of intelligence (e.g., to test the *a priori* hypothesis that, as a proxy for brain size, circumference is a proxy for intelligence), many will dispute the validity of the assertion.³⁷

Construct Validity

Construct validity is concerned with the consistency and strength with which a measure relates to other measures to which, according to theory it should relate.^{29,34,110-112} As Magnusson³⁵ described: "We begin from a logically defined variable that is included as a logical construct in a system of constructs, in which all of the concepts logically belong, and where all of the relationships are explained by a theory. From this theory certain practical consequences can be derived about the outcome of the test under certain conditions. These consequences can be tested: if the result is what was expected in a series of such tests, then the test is said to have construct validity for the variable tested" (p. 130).

Type A Behavior (TAB) and Coronary Artery Disease (CAD): The first example of construct validity analysis using UniODA features an application with an ordered binary class variable and attribute. The study examined the relationship TAB and CAD: TAB is theorized to predict cardiovascular disease, CAD is a form of cardiovascular disease, so TAB should therefore predict CAD. This test of the validity of the TAB construct was accomplished by administering a TAB survey to a sample of male patients undergoing coronary angiography.¹¹³ Based on their scores patients were classified as being either Type A or Type B (low versus high scores on the survey, respectively). Patients were also independently classified as having severe CAD defined as ≥ 2 coronary arteries obstructed by at least 50%, or mild CAD defined as ≤ 1 coro-

nary arteries obstructed by at least 50% (discussed in Chapter 2, such arbitrary parsing of attributes can reduce model accuracy and induce paradoxical confounding, so this design illustrates the use of a suboptimal *measurement* methodology). Using UniODA the directional alternative hypothesis is that compared to Type Bs, Type As (binary class variable, coded as “0” and “1”, respectively) should have more severe CAD (binary attribute): the null hypothesis is that A/B Type can’t be discriminated on the basis of CAD severity (mild was coded as 0, and severe as 1). The data are presented in Table 5.25.

Table 5.25: TAB and CAD

A/B Type	<u>Mild CAD</u>	<u>Severe CAD</u>
Type A	12	33
Type B	24	19

The *a priori* hypothesis that TAB is related to severe CAD was evaluated via the following UniODA and MegaODA software syntax:

```

OPEN tab.dat;
ATTR cad
OUTPUT tab.out;
DIRECTIONAL < 0 1;
VARS tab cad;
MCARLO ITER 25000;
CLASS tab;
GO;

```

The *a priori* hypothesis yielded a statistically significant ($p < 0.004$), moderately strong effect ($ESS = 29.2$), supporting the construct validity of TAB as a precursor of cardiovascular disease endpoints.

Emergency Department Triage Coding and Hospital Admission: The second example of construct validity analysis using UniODA features an application in which hospital admission status, Emergency Severity Index (ESI) Version 3 triage score, and binary indicators of whether lab work or radiological examinations were completed in the Emergency Department (ED), were available for 160,471 patients seen over a three-year period in the ED of a leading community teaching hospital in Toronto.¹¹⁴ A hierarchically-optimal CTA model (see Chapter 10) was manually-constructed by chaining successive UniODA models, in order to determine if the triage score successfully partitioned the sample into strata representing significantly different likelihoods of patients being admitted to hospital. Lower triage codes (indicating greater severity), and the use of laboratory and radiological tests, are all hypothesized to increase the likelihood of hospital admission. Figure 5.14 illustrates the hierarchically-optimal CTA model for these data. For the overall model $ESS = 63.8$, a relatively strong effect: the model correctly classified 80.2% of admitted patients and 83.6% of non-admitted patients, and was correct 60.1% of the time it predicted a patient would be admitted and 93.2% of the time it predicted a patient wouldn’t be admitted.

As hypothesized, ESI score stratified the sample into three strata: ESI scores < 3 were associated with highest likelihood of admission; ESI scores = 3 were associated with intermediate likelihood of admission; and ESI scores > 3 were associated with the lowest likelihood of admission. For each of these three strata, a binary indicator of whether laboratory work was completed in the ED was the initial attribute, followed by a binary indicator of whether radiological work was completed in the ED.

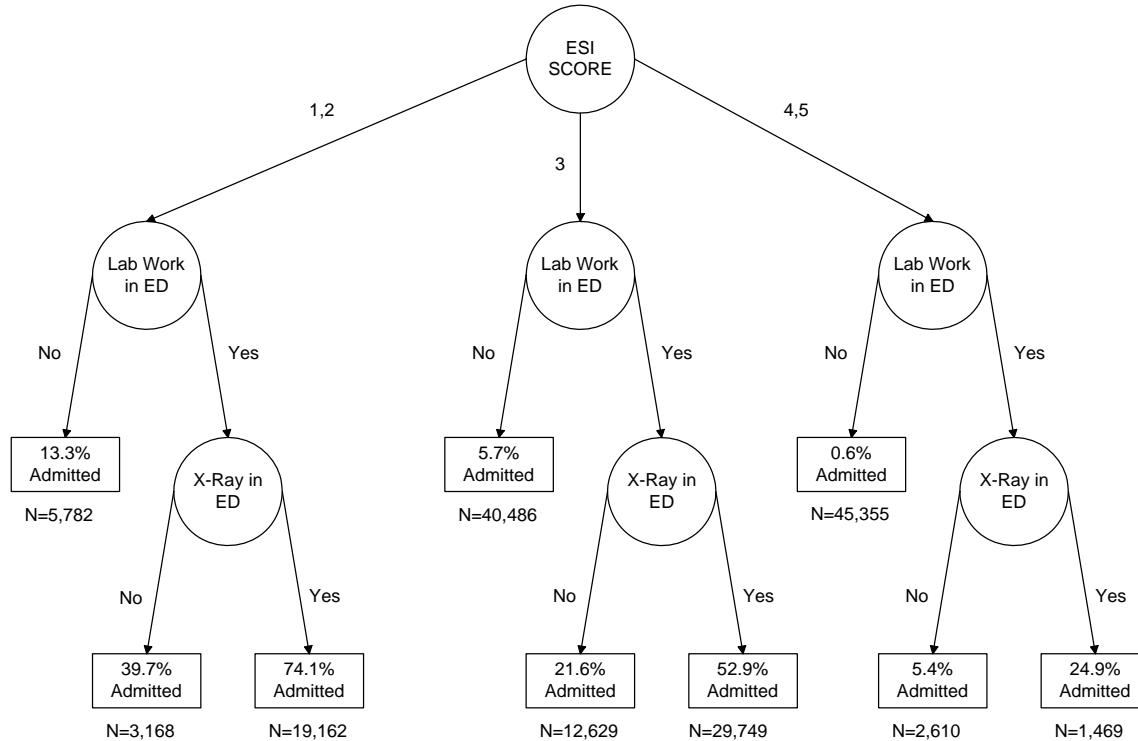
If no laboratory work was completed in the ED then likelihood of admission is 2.3 times greater for patients with ESI scores of 1 or 2 versus scores of 3; 10.4 times greater for patients with ESI scores of 3 versus scores of 4 or 5; and 24.2 times greater for patients with ESI scores of 1 or 2 versus scores of 4 or 5.

If lab work was completed in the ED, but no radiological work was completed, then the likelihood of admission is 3.0 times greater (versus patients with no lab work) for patients with ESI scores of 1 or 2; 3.8 times greater for patients with ESI scores of 3; and 9.8 times greater for patients with ESI scores of 4 or 5. The likelihood of admission is 1.8 times greater for patients with ESI scores of 1 or 2 versus 3; 4.0 times greater for patients with ESI scores of 3 versus scores of 4 or 5; and 7.4 times greater for patients with ESI scores of 1 or 2 versus scores of 4 or 5.

Finally, if radiological work was also completed in the ED, the likelihood of admission is 1.9 times greater (versus patients with no radiological work) for patients with ESI scores of 1 or 2; 2.4 times greater for patients with ESI scores of 3; and 4.6 times greater for patients with ESI scores of 4 or 5. And, the like-

likelihood of admission is 1.4 times greater for patients with ESI scores of 1 or 2 versus scores of 3; 2.1 times greater for patients with ESI scores of 3 versus scores of 4 or 5; and 3.0 times greater for patients with ESI scores of 1 or 2 versus 4 or 5.

Figure 5.14: Hierarchical CTA Model Predicting Hospital Admission from the ED (all $p < 0.0001$)



The CTA model was used to construct a staging table for predicting hospital admission: in Table 5.26 dashes indicate that the attribute isn't included in the corresponding branch and endpoint of the CTA model); N is the number of patients with the indicated attribute profile; p_{Admit} is the empirical probability that patients in an endpoint were admitted to the hospital; and Odds is p_{Admit} expressed as approximate odds of hospital admission. The CTA model provides $0.741 / 0.0055 = 134.7$ -fold stratification in the likelihood of patient hospital admission.

Table 5.26: Staging Table Based on Hierarchically-Optimal CTA Admission Model

Stage	ESI	Labs	X-Ray	N	p_{Admit}	Odds
1	4,5	No	---	45,355	0.0055	1:170
2	4,5	Yes	No	2,610	0.054	1:17
3	3	No	---	40,486	0.057	1:17
4	1,2	No	---	5,782	0.13	2:13
5	3	Yes	No	12,629	0.22	2:7
6	4,5	Yes	Yes	1,469	0.25	1:3
7	1,2	Yes	No	3,168	0.40	2:3
8	3	Yes	Yes	29,749	0.53	8:7
9	1,2	Yes	Yes	19,162	0.74	3:1

For the three ESI strata identified by the CTA model the *absolute magnitude* of the difference between patients with no lab work, versus patients with lab and radiological work completed in the ED, is greatest for *lower* ESI scores. For ESI scores of 1 and 2 the absolute difference is 74.1% - 13.3% = 60.8%. For scores of 3 the absolute difference is 47.2%, and for scores of 4 or 5 absolute difference is 24.3%. In contrast, the *relative magnitude* of the difference between patients with no lab work, versus patients with lab and radiological work completed in the ED, is greatest for *higher* ESI scores. For ESI scores of 1 and 2 the relative difference is 74.1% / 13.3% or 5.6. For ESI scores of 3 the relative difference is 9.3, and for scores of 4 or 5 the relative difference is 41.5.

Convergent and Discriminant Validity

Campbell and Fiske¹¹⁵ describe an experimental methodology involving the *multitrait-multimethod matrix* that represents a comprehensive nomological framework for assessing construct as well as other types of validity. They argue that, in the process of assessing the validity of a construct, although it is important to demonstrate that a measure relates to other measures to which it theoretically should relate, this isn't a sufficient test of validity. Indeed, it is also important to demonstrate that a measure does *not* relate to other measures to which it theoretically shouldn't relate: for example, because the latter type of measure falls outside of the system within which validity is being assessed. Investigations involving measures that theoretically should be related are focusing on convergent validity, and investigations involving measures that theoretically should be unrelated are focusing on discriminant validity.³⁷ To demonstrate the use of UniODA in assessing convergent and discriminant validity, two examples involving a topic of substantive focus in the ODA laboratory for more than two decades—Type A Behavior (TAB)—are presented.

TAB and Savoring: Much work has studied the consequences of TAB, characterized by a strong achievement orientation, hard-driving competitiveness, speed-impatience, and hostility in response to threat to personal control over salient outcomes, versus Type B behavior (TBB) characterized by a relaxed, easy-going orientation and lower levels of competitiveness, impatience, and hostility.^{116,117} Investigating differences in the characteristic styles through which Type As and Bs savor positive outcomes, research has found that Type As are less likely than Type Bs to look back on positive events afterwards in order to store memories for later recall—a past-focused savoring response that might undermine the ability to savor positive outcomes retrospectively.¹¹⁸ More recent research has, on the one hand, identified cognitive and behavioral response among Type As that dampen their enjoyment of ongoing positive events—in particular, less counting of blessings, less memory building, and more “kill joy” fault-finding.^{119,120} On the other hand, research has also found that Type As, relative to Type Bs, report higher levels of self-congratulation (i.e., telling oneself how proud one is and how impressed others are) in response to achievement-related outcomes—a present-focused savoring strategy that amplifies enjoyment.¹²⁰ Concerning future-focused savoring, one might expect Type As’ greater achievement orientation, relative to Type Bs, to be associated with a greater capacity to derive pleasure through the anticipation of goal attainment.

Accordingly, this study compared Type As’ and Bs’ generalized beliefs regarding their capacity to enjoy positive outcomes through reminiscence, savoring the moment, and anticipation. We tested the *a priori* hypotheses that, compared to Type Bs, Type As perceive themselves as being less able to savor via reminiscence due to their reluctance look back to store memories, and more able to savor via anticipation due to their greater goal orientation. An exploratory analysis addresses differences between As and Bs on savoring the moment, because there is no compelling reason to hypothesize that As and Bs will differ in any systematic manner on this measure.

The sample was drawn from a large pool of college undergraduates who completed a battery of questionnaires.¹²⁰ TAB was assessed using the short form of the Jenkins Activity Survey for Students.¹²¹⁻¹²⁶ Normative guidelines were followed to maximize reliability of assignments into A/B categories, yielding an analysis sample having 131 extreme Type B and 117 extreme Type A college undergraduates.¹²⁷⁻¹²⁹ The savoring belief subscales were assessed using the Savoring Beliefs Inventory (SBI).¹³⁰ The 24-item SBI has separate subscales assessing perceived capacity to savor positive outcomes through reminiscing, enjoying the moment, and anticipating, and scores on the SBI have been shown to have good internal consistency and test-retest reliability, as well as strong convergent, discriminant, and predictive validity, among both younger and older adults.^{130,131}

Table 5.27 presents descriptive statistics for the three savoring belief subscales separately by A/B Type. For expository purposes, and to provide data for meta-analysis, means on the three subscales were compared between A/B Types using Student's *t*-test. No statistically reliable effect emerged for scores on reminiscence [$t(244) = 1.2, p < 0.25$], savor the moment [$t(246) = 0.7, p < 0.49$], or anticipation [$t(246) = 1.2, p < 0.23$] subscales.

Table 5.27: Descriptive Statistics for Savoring Belief Subscales, by A/B Type

Savoring Belief Subscale	A/B Type	Mean	SD	Median
Reminiscence (Past Focus)	B	5.8	0.80	5.8
	A	5.9	0.89	6.1
Savor the Moment (Present Focus)	B	5.4	0.93	5.5
	A	5.5	1.10	5.6
Anticipation (Future Focus)	B	5.3	0.90	5.4
	A	5.5	1.09	5.8

Note: There was one missing value for each A/B type on Reminiscence

UniODA analysis was conducted treating A/B type as the class variable and the three savoring subscale scores as ordered attributes.¹³² For reminiscence a statistically reliable, ecologically weak effect emerged ($p < 0.04, ESS = 16.6$), that was stable in LOO validity analysis ($p < 0.007$). The UniODA model was: if reminiscence ≤ 5.93 (53rd percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals that Type As had significantly higher reminiscence scores than Type Bs. The model correctly classified 56% of the Type Bs, and 61% of the Type As. The model was correct 62% of the time a prediction of Type B was made, and 55% of the time a prediction of Type A was made.

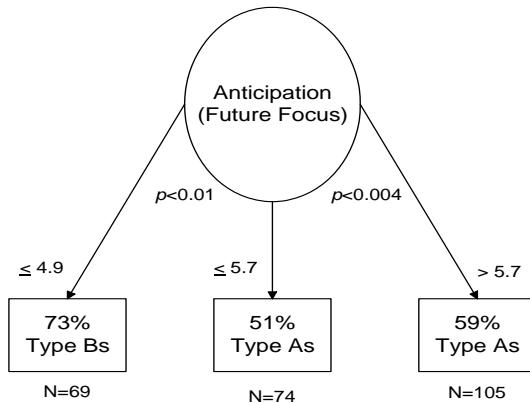
For savor the moment a statistically marginal, ecologically weak effect emerged ($p < 0.08, ESS = 14.7$), that was stable in LOO analysis ($p < 0.005$). The UniODA model was: if savor the moment ≤ 6.19 (77th percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals that the Type As had marginally higher savor the moment scores compared to the Type Bs. The model correctly classified 84% of the Type Bs, and 31% of the Type As. The model was correct 58% of the time that a prediction of Type B was made, and 63% of the time that a prediction of Type A was made.

Finally, for anticipation a statistically reliable, ecologically weak effect emerged ($p < 0.003, ESS = 20.2$), that was stable in LOO analysis ($p < 0.002$). The UniODA model was: if anticipation ≤ 5.69 (58th percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals that the Type As had significantly higher anticipation scores compared to the Type Bs. The model correctly classified 67% of the Type Bs, and 53% of the Type As. The model was correct 62% of the time that a prediction of Type B was made, and 59% of the time that a prediction of Type A was made.

UniODA analyses were next iteratively applied to the data to create the enumerated-optimal CTA model (see Chapter 11) illustrated in Figure 5.15: the model optimally discriminates A/B Type treating reminiscence, savor the moment and anticipation subscale scores, and gender, as possible attributes.

Only the anticipation subscale emerged as a statistically significant attribute in the CTA model that identified a three-endpoint parse. In the model, extreme Type B undergraduates are substantially *more likely* (3:1 odds) than extreme Type As to score at *lowest levels* on the anticipation dimension of savoring beliefs: the cut-point 4.9 represents the 28nd percentile on this dimension for the sample. And, while A/B Types are *comparably likely* to score at *intermediate levels* on anticipation (1:1 odds), Type As are modestly *more likely* (3:2 odds) to score at *highest levels* on anticipation: the cut-point 5.7 represents the 58th percentile on this dimension for the sample.

Figure 5.15: CTA Model Discriminating A/B Type Using Three Savoring Belief Dimensions



Taken in sum the CTA model reveals Type Bs are substantially more likely to score in the lowest 30% of the scores on anticipation, while Type As are modestly more likely to score in the highest 60% of the scores. The $ESS = 24.1$ achieved by the model was at the boundary separating relatively weak versus moderate effect strength: the model correctly classified 41% of Type As and 83% of Type Bs in the sample, and it was correct 73% of the time it predicted an observation was Type B and 56% of the time an observation was predicted to be Type A.

The results reveal an interesting pattern of differences between Type As and Type Bs in terms of their perceived ability to savor positive experiences retrospectively, concurrently, and prospectively. Concerning past-focused savoring, Type As reported a *greater* capacity than Type Bs to derive enjoyment by reminiscing about positive memories, contrary to the *a priori* hypothesis. Concerning present-focused savoring, there was only a marginally significant A-B difference in the perceived capacity to savor the moment. Concerning future-focused savoring, UniODA analysis revealed that Type As perceived higher capacity to derive enjoyment through anticipation relative to Type Bs, and CTA analysis revealed specific thresholds of anticipation subscale scores that reliably discriminated As and Bs. In particular, significantly more Type Bs and fewer Type As scored below the 28th percentile on anticipation, and significantly more Type As and fewer Type Bs score above the 58th percentile on anticipation; whereas As and Bs were equally likely to fall between the 28th and 58th percentile on anticipation. Thus, while UniODA analysis is consistent with the *a priori* hypothesis, CTA analysis provides strong evidence to support the *a priori* hypothesis. In sum, Type As, relative to Type Bs, believe they are more capable of enjoying positive memories through reminiscence and marginally more capable of enjoying positive moments; and are less likely to report a lower capacity (< 28th percentile) and more likely to report a higher capacity (> 58th percentile) to derive joy through anticipation.

The difference between the results obtained by the UniODA and CTA analyses of anticipation for As and Bs highlights the potential benefit of considering nonlinear effects in testing research hypotheses. The UniODA model reflects the optimal threshold on anticipation that produces highest possible accuracy in classifying As and Bs when selecting a single cut-point to predict TAB on the basis of anticipation. The CTA model, in contrast, represents the combination of reminiscence, savoring the moment, and anticipation subscale scores that produces the highest possible accuracy in classifying As and Bs. The three-end-point parse that emerged in the CTA model reveals that the hypothesized A-B difference in the capacity to anticipate exists at the lower and upper range of the Anticipation subscale, but not in the middle range of the subscale. Whereas 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. Thus, the CTA model not only confirms the *a priori* hypothesis, but also pinpoints the specific levels of anticipation at which the predicted A-B differences emerge. Clearly, researchers would be wise to evaluate the possibility of nonlinear effects in testing bivariate relationships, in order to avoid missing important and informative research conclusions. CTA is the only statistical methodology available which is capable of identifying *explicitly optimal* parsed models such as the model obtained presently.

Prior research tested the *a priori* hypothesis that TAB undermines enjoyment of *leisure time*, and that this effect is mediated by savoring responses which hamper enjoyment.¹¹⁹ Findings suggested that the hypothesized A-B differences in savoring reflect differences in perfectionism, but not in time urgency. This second example of the use of ODA models in assessing convergent and discriminant validity uses the same sample to compare 117 extreme Type A and 131 extreme B undergraduates on ten dimensions of savoring assessed for a *performance-related* stimulus. Described earlier, classification of subjects into the extreme A/B categories was made based on normative recommendations. Subjects completed the Ways of Savoring Checklist (WOSC), a 60-item survey assessing types of savoring responses and strategies, and providing scores on ten dimensions of savoring.¹³¹ The WOSC was completed twice: once using one's most recent vacation as the target stimulus, and again using one's most recent grade on a test as the target stimulus.¹³¹ There was no relationship between A/B Type and gender ($p < 0.63$, $ESS = 2.9$). Findings for the vacation enjoyment (leisure-related) stimuli had $p > 0.08$, $ESS \leq 14.6$ (data not presented).

Table 5.28 gives UniODA findings for the ten dimensions of savoring for the test grade (performance-related) stimulus.¹³³ Only a relatively weak effect of self-congratulation was statistically reliable: extreme Type A undergraduates are more likely to score at higher levels on this dimension (the cut-point value of 4.79 corresponds to the 62nd percentile in the sample) versus extreme Type B undergraduates.

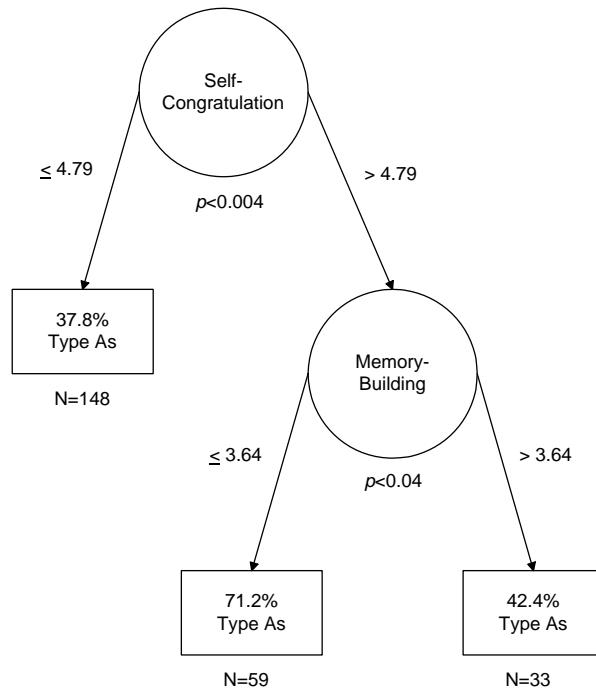
Table 5.28: UniODA Relationships of Type A Behavior and Savoring: Test Grade Stimulus

Savoring Dimension	UniODA Cut-Point	Percent of Type A's			
		N	p<	ESS	
Sharing with Others	≤ 3.92	127	40.2	0.12	13.8
	> 3.92	113	54.0		
Memory Building	≤ 2.64	148	40.5	0.09	14.6*
	> 2.64	93	55.9		
Self-Congratulation	≤ 4.79	148	37.8	0.003	21.9*
	> 4.79	92	60.9		
Temporal Awareness	≤ 3.64	154	42.2	0.37	10.4
	> 3.64	88	53.4		
Behavioral Expression	≤ 1.38	69	37.7	0.36	10.3*
	> 1.38	175	50.3		
Sensory-Perceptual Sharpening	≤ 1.62	27	53.1	0.90	5.6
	> 1.62	207	45.1		
Absorption	≤ 2.25	99	37.4	0.12	14.1
	> 2.25	141	51.8		
Comparing	≤ 3.70	166	42.2	0.19	12.5*
	> 3.70	76	56.6		
Counting Blessings	≤ 3.50	93	37.7	0.07	14.8*
	> 3.50	152	53.3		
Kill-Joy Thinking	≤ 2.79	160	43.8	0.66	7.8*
	> 2.79	82	52.4		

Note: Total N varies due to missing values. An asterisk (*) indicates that the model performance was stable in LOO analysis.

Figure 5.16 presents the enumerated optimal CTA model obtained by using the ten savoring dimensions as potential attributes to predict A/B status. Note that the effect for memory-building *wasn't* statistically reliable in total-sample analysis, but memory-building *was* statistically reliable for the sub-partition of undergraduates scoring at higher levels on self-congratulation.

Figure 5.16: CTA Model Discriminating Type As and Bs on Ten Savoring Dimensions: Test Grade Stimulus



Type A undergraduates are modestly *less likely* (2:3 odds) than Type Bs to score at *lower levels* on the self-congratulation dimension of savoring (the CTA model cut-point reflects the 62nd percentile on this dimension for the sample). And, among those undergraduates scoring at higher levels on self-congratulation, Type As are modestly *less likely* (2:3 odds) than are Type Bs to score at *higher levels*, and Type As are substantially *more likely* (7:3 odds) to score at *lower levels* on the memory-building dimension of savoring versus Type Bs (the CTA cut-point reflects the 82nd percentile on this dimension for the sample). The model correctly classified 86.7% of the Type Bs, and 37.5% of the Type As. The model was correct 61.3% of the time it was predicted that an observation was a Type B, and 71.2% of the time that it was predicted an observation was a Type A. Overall the CTA model achieved *ESS* = 24.2, a borderline moderate effect.

Considered as a whole this model reveals that extreme Type As are most likely to score in the *highest* quintile on self-congratulation, and in the *lowest* three quintiles on memory-building. Extreme Type A undergraduates focus strongly on how proud they are and how impressed others must be, but are moderately or less involved in actively storing positive memories for later recall, or in reminiscing about prior positive events. The current findings are consistent with prior research on Type A behavior and reminiscence that found Type As are less likely than Type Bs to store details of positive events for later recall.¹¹⁸ Type As' tendency to avoid building memories of personal achievements may stem from their impatience to move on to new opportunities, or from their reluctance to spend time encoding memories at the expense of striving toward future accomplishments.¹¹⁸

Chapter 6

Optimized General Linear Models

Chapter 6 discusses how to explicitly maximize the classification performance (*ESS*) achieved by widely-used GLM models representing the general linear model (GLM) paradigm, including ordinary least-squares (OLS) regression analysis, factorial analysis of variance (ANOVA), and discriminant function analysis.^{1,2}

OLS Regression Analysis

OLS regression analysis³ is among the most broadly used of statistical methods in all of empirical science, so it may be common knowledge that four principal assumptions justify using linear regression models for purposes of inference or prediction:

- (1) linearity and additivity of the relationship between dependent and independent variables:
 - (a) the expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed;
 - (b) the slope of that line does not depend on the values of the other variables;
 - (c) the effects of different independent variables on the expected value of the dependent variable are additive;
- (2) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data);
- (3) homoscedasticity (constant variance) of the errors:
 - (a) versus time (in the case of time series data);
 - (b) versus the predictions;
 - (c) versus any independent variable; and
- (4) normality of the error distribution.

If these assumptions are violated then the forecasts, confidence intervals, and scientific insights yielded by a regression model may (at best) be inefficient or (at worst) seriously biased or misleading.⁴⁻⁶ While the use of distribution-free methods can yield valid p for the sample (e.g., Chapter 2), the issue of whether the *underlying phenomenon being modeled is inherently linear* is an entirely different matter. If it is hypothesized that the underlying phenomenon is inherently linear, then a linear model is appropriate: a linear model that explicitly achieves (weighted) maximum accuracy (Chapter 8) may be preferred unless a less accurate solution is explicitly desired. In contrast, if the underlying phenomenon is *not* hypothesized to be inherently linear, then a *nonlinear* model that explicitly achieves maximum (weighted) accuracy is appropriate unless a less accurate nonlinear model is desired. Here nonlinear does *not* mean a convex or concave curve, but rather an *interaction in which the relationships between variables depend on the levels of one or more moderator variables* (violating OLS assumption 1.c). If the underlying phenomenon actually

is linear then a *maximum-accuracy nonlinear model* will identify the maximum-accuracy linear solution, but no *linear* model can identify maximum-accuracy (weighted) *nonlinear* solutions (Chapters 10-12).

Regression **Toward** the Mean

Linear regression models fail to predict extreme values in the training sample accurately unless the linear association is nearly perfect due to a phenomenon called *regression toward the mean*.⁷ Linear regression models predict values that are near the mean better than they predict values that deviate from the mean: the more extreme that a given value is relative to the mean, the worse it's predicted by regression models (i.e., the greater the absolute magnitude of the residual value created by subtracting actual and predicted values for the observation). This is demonstrated experimentally via simulation involving one independent and one dependent variable (class variable and attribute, respectively)—representing the simplest case, and a small sample of $N = 15$ observations.⁸ First, 500 artificial datasets were constructed, each by randomly shuffling an array consisting of integers 1-15 inclusive, and then pairing this array with another randomly shuffled array of integers 1-15, thereby creating a set of 15 randomly paired data values. The simplest case of a regression model, the Pearson product-moment correlation (r) was computed for each of these 15 randomly paired data values. The first 100 datasets that were identified with associated $0.5 \leq r < 0.6$ were saved, as were the first 100 datasets with $0.6 \leq r < 0.7$, $0.7 \leq r < 0.8$, $0.8 \leq r < 0.9$, and $0.9 \leq r$. Then, separately for each data set, a regression model (r) was used to compute the predicted score (\hat{y}_i) for each observation (y_i). Predicted scores were rounded to the nearest integer, \hat{y}_{i-INT} . For each block of 100 data sets the percentage of the actual y_i values (i.e., of integers 1 through 15, inclusive) accurately predicted (i.e., $y_i = \hat{y}_{i-INT}$) was computed. The results of this simulation are given in Table 6.1, in which y_i values are indicated as the Target. For example, if $0.7 \leq r < 0.8$, the value "4" was correctly predicted 3% of the time over the 100 different analyses; "5" was correctly predicted 40% of the time, "6" was correctly 100% of the time; and "3" was never predicted correctly. Regression models associated with correlations ranging in magnitude between 0.8 ($R^2 = 0.64$) and 0.9 ($R^2 = 0.81$) were unable to predict *any* of the most extreme target integers (1 - 3, 13 - 15) constituting 40% of the most extreme values in the sample.

Table 6.1: Probability of Correctly Predicting Target Values via Regression: N = 100 Data Sets per Column

Target	<u>Correlation Domain</u>				
	$.5 \leq r < .6$	$.6 \leq r < .7$	$.7 \leq r < .8$	$.8 \leq r < .9$	$.9 \leq r$
1					.13
2					.42
3					.85
4				.03	1
5				.40	1
6			.33	1	1
7	.91	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10			.37	1	1
11				.40	1
12				.03	1
13					1
14					.42
15					.18
ESS	13.6	14.3	19.3	34.7	78.6

Note: Missing entries indicate $p = 0$.

Findings of this simulation indicate that for $.5 \leq r < .8$, the only target values predicted accurately lie within 0.5 to 1.5 units of the mean (7.5) of the 15 response values, and only when r approaches 1.0 will prediction of extreme values attain ecologically meaningful levels. This is problematic because the F test used to assess statistical significance of regression model parameters involves dividing the sum of squares for the model by the sum of squares for error.⁵ As r approaches one, sum of squares for error approaches zero, and resulting division by zero induces numerical instability called “*bouncing betas*”. Thus, before a linear regression model can achieve excellent predictive validity for extreme values, numerical instability creates algorithm failure. As real-world samples increase in size they may become dominated by mediocre values, and exhibit characteristics such as bipolarity, multimodality, and skew—violating assumptions underlying regression and working against the accurate prediction of extreme values.

Regression Away From the Mean

Prior work has shown that an effective way to increase the classification accuracy achieved by suboptimal linear multivariable models is to use UniODA to adjust the cutpoints that are used to interpret predicted scores (\hat{y}_i).⁹⁻¹¹ Many applications use dependent (class) variables measured using Likert-type ratings with ten or fewer discrete ordered response categories.^{11,12} For such applications, the optimization process is straightforward: obtain the suboptimal linear model; create a dataset having values for y_i (class variable) and \hat{y}_i (attribute) for every observation; use UniODA to predict y_i as a function of \hat{y}_i ; and assess classification performance.⁹⁻¹¹ Following are examples of this procedure used to explicitly maximize the ESS of multiple regression models with class variables measured using 7-, 5-, and 10-point scales, respectively.

Modeling Vacation Enjoyment: The first example involves five multiple regression predictors of a categorical ordinal dependent measure for a sample of $N = 787$ college undergraduates.¹³ The dependent variable was how much the respondent enjoyed their last vacation, rated using a 7-point Likert-type scale (1=“very little”, 7=“a great deal”). The independent variables were five scales selected from the Ways of Savoring Checklist (WOSC), a self-report measure of strategies that people use to regulate positive feelings: sharing with others (SWO), self-congratulation (SC), temporal awareness (TA), counting blessings (CB), and kill-joy thinking (KJT). Mean scores on these scales range between 1 and 7, with higher values indicating greater use of the particular savoring strategy.¹³

The omnibus linear multiple regression model, $\hat{y}_i = 4.01 + 0.23 * \text{SWO} + 0.15 * \text{SC} + 0.14 * \text{TA} + 0.14 * \text{CB} - 0.39 * \text{KJT}$, was statistically reliable [$F(5,781) = 147.1$, $p < 0.0001$, $R^2 = 0.70$], and each of the independent variables had a statistically reliable independent contribution (p 's < 0.0001). Table 6.2 shows cutpoints used on \hat{y}_i to obtain the predicted “target” value, for each value of the dependent variable: first for standard regression analysis (i.e., the indicated cutpoint strategy is used for all such 7-point Likert-type scales), and then for optimized regression analysis (cutpoints were optimized for the particular data in this example, and thus vary across applications). For example, for $\hat{y}_i = 4.12$, the predicted value of the dependent variable is 4 for standard regression analysis, and 2 for optimized regression analysis (see Table 6.2).

Table 6.2: Cutpoints on \hat{y}_i Defining Predicted Target for Standard and Optimized Multiple Regression

Target	Standard	Optimized
1	$\hat{y}_i < 1.5$	$\hat{y}_i < 3.7$
2	$1.5 \leq \hat{y}_i < 2.5$	$3.7 \leq \hat{y}_i < 4.6$
3	$2.5 \leq \hat{y}_i < 3.5$	$4.6 \leq \hat{y}_i < 4.7$
4	$3.5 \leq \hat{y}_i < 4.5$	$4.7 \leq \hat{y}_i < 5.3$
5	$4.5 \leq \hat{y}_i < 5.5$	$5.3 \leq \hat{y}_i < 5.8$
6	$5.5 \leq \hat{y}_i < 6.5$	$5.8 \leq \hat{y}_i < 6.5$
7	$5.5 < \hat{y}_i$	$6.5 < \hat{y}_i$

Table 6.3 gives classification results using standard versus optimized regression to predict the Target value of the dependent measure, y_i . Given for both models for every Target value is the total num-

ber of correctly predicted Target values divided by the total number of instances of the Target value in the sample, and the sensitivity (accuracy) of the model for each Target value of the dependent measure.

Table 6.3: Predicting Actual Target Values by Standard and Optimized Multiple Regression

Target	Standard		Optimized	
	Regression		Regression	
1	0/12	0.0%	8/12	66.7%
2	0/13	0.0%	11/13	84.6%
3	1/20	5.0%	1/20	5.0%
4	11/36	30.6%	13/36	36.1%
5	31/79	39.2%	29/79	36.7%
6	113/183	61.7%	84/183	45.9%
7	254/444	57.2%	270/444	60.8%
<i>ESS</i>		15.8		39.3

Note: Greatest sensitivity obtained for each Target Value is indicated in **bold**.

As seen, standard regression achieved 0% success in correctly predicting either of the two lowest response scale (Target) values (1, 2), as compared with 66.7% or greater success for optimized regression. Optimized regression also predicted the highest response scale value (7) more accurately than standard regression. Standard regression achieved greater success in predicting the Target values 5 and 6, both of which were close to the sample mean score (6.2) on the dependent variable. Overall, accuracy achieved by standard regression (*ESS* = 15.8) corresponds to a weak effect, versus a moderate effect (*ESS* = 39.3) for optimized regression, which yielded a 149% boost in classification performance. The UniODA solution was obtained in 8.1 CPU minutes running UniODA software on a 3 GHz Intel Pentium D microcomputer.

Optimistic Benefit-Finding in the Face of Adversity: The next example involves three multiple regression predictors of a 5-point outcome measure. The dependent variable was optimistic benefit-finding, or the positive cognitive reappraisal of adversity (item #11 from the Life Orientation Test): “I’m a believer that every cloud has a silver lining” (0 = strongly disagree; 1 = disagree; 2 = neutral; 3 = agree; 5 = strongly disagree).¹⁴ Three independent measures were used to predict this outcome for a sample of $N = 774$ college undergraduates: the savoring the moment (SM) subscale of the Savoring Beliefs Inventory¹⁵; an index of self-esteem (SE) adapted from Rosenberg¹⁶; and the positive affectivity (PA) subscale of the Affect Intensity Measure.¹⁷ The linear multiple regression model, $\hat{y}_i = 0.14 + 0.15 * \text{SM} + 0.21 * \text{SE} + 0.24 * \text{PA}$, was statistically reliable [$F(3,770) = 37.2, p < 0.0001, R^2 = 0.36$], and each independent variable had a reliable independent contribution (all p 's < 0.0001). Table 6.4 presents the cutpoints used on \hat{y}_i to obtain the predicted Target value, for all five values of the dependent variable, first for standard regression analysis and then for optimized regression. For example, for $\hat{y}_i = 2.31$ the predicted value of the dependent variable is 2 for standard regression analysis, and 1 for optimized regression analysis.

Table 6.4: Cutpoints on \hat{y}_i Defining Predicted Target for Standard and Optimized Multiple Regression

Target	Standard	Optimized
0	$\hat{y}_i < 0.5$	$\hat{y}_i < 2.27$
1	$0.5 \leq \hat{y}_i < 1.5$	$2.27 \leq \hat{y}_i < 2.38$
2	$1.5 \leq \hat{y}_i < 2.5$	$2.38 \leq \hat{y}_i < 2.55$
3	$1.5 \leq \hat{y}_i < 3.5$	$2.55 \leq \hat{y}_i < 2.59$
4	$3.5 < \hat{y}_i$	$2.59 < \hat{y}_i$

Table 6.5 gives classification results using standard versus optimized regression to predict the Target value of the dependent measure, y_i . Given for both models for every Target value is the total number of correctly predicted Target values divided by the total number of instances of the Target value in the sample, and the sensitivity (accuracy) of the model for each Target value of the dependent measure.

Table 6.5: Predicting Actual Target Values by Standard and Optimized Multiple Regression

Target	Standard		Optimized	
	Regression		Regression	
0	0/16	0.0%	9/16	56.2%
1	1/61	1.6%	15/61	24.6%
2	146/316	46.2%	76/316	24.1%
3	166/269	61.7%	21/269	7.8%
4	0/112	0.0%	84/112	75.0%
<i>ESS</i>		2.4		21.9

Note: Greatest sensitivity obtained for each Target Value is indicated in **bold**.

While standard regression achieved 0% success in correctly predicting either of the most extreme Target values, optimized regression had 56% success predicting the lowest Target value (0) and 75% success predicting the greatest Target value (4). Optimized regression also predicted the next-most extreme Target value (1) with 25% success, versus less than 2% for standard regression. In contrast, standard regression classified Target values 2 and 3—bordering the mean Target value of 2.5—more accurately than optimized regression. Overall, the accuracy achieved by standard regression corresponds to a minuscule effect ($ESS = 2.4$), while optimized regression yielded a weak effect ($ESS = 21.9$), reflecting a 812% boost in classification performance versus standard regression. The UniODA solution was obtained in 1.5 CPU minutes, running MegaODA software on a 3 GHz Intel Pentium D microcomputer.

Looking Forward to Receiving a Good Grade: The third example involves four multiple regression predictors of a 10-point outcome measure. The dependent (class) variable was how much students ($N = 629$ college undergraduates) who received a good grade on an academic test or paper looked forward to the outcome beforehand, as assessed by a modified version of item #6 from the Ways of Savoring Checklist (“To what extent did you look forward to the last time you got a good grade on a test or paper?”), that uses a 10-point Likert-type response scale (1 = “not much”; 10 = “a great deal”).¹³ The independent variables, that also used 10-point Likert-type response scales, included how often respondents receive good grades (frequency: 1 = “it does not happen very often”, 10 = “it happens very often”), the degree to which they expected the good grade to occur (expectancy: 1 = “I did not expect it to happen”, 10 = “I expected it to happen”), how long their experience of getting the good grade lasted (duration: 1 = “it lasted a short time”, 10 = “it lasted a long time”), and how desirable the good grade was (desirability: 1 = “worse thing that could happen”, 10 = “best thing that could happen”). The linear multiple regression model, $\hat{y}_i = 0.42 + 0.12 * \text{frequency} + 0.15 * \text{expectancy} + 0.19 * \text{duration} + 0.31 * \text{desirability}$, was statistically significant [$F(4,624) = 19.2$, $p < 0.0001$, $R^2 = 0.33$], and all independent variables had a reliable independent contribution (p 's < 0.0001).

Table 6.6 presents the cutpoints used on \hat{y}_i to obtain the predicted target value, for each of the ten values of the dependent variable, first for standard regression analysis (the indicated cutpoint strategy is used for all such 10-point Likert-type scales) and then for optimized regression analysis (cutpoints were optimized for the particular data in this example, and vary across applications). For example, for $\hat{y}_i = 4.49$, the predicted value of the dependent variable is 4 for standard regression analysis, and 1 for optimized regression analysis.

Table 6.6: Cutpoints on \hat{y}_i Defining Predicted Target for Standard and Optimized Multiple Regression

Target	Standard	Optimized
1	$\hat{y}_i < 1.5$	$\hat{y}_i < 4.97$
2	$1.5 \leq \hat{y}_i < 2.5$	$4.97 \leq \hat{y}_i < 5.47$
3	$2.5 \leq \hat{y}_i < 3.5$	$5.47 \leq \hat{y}_i < 5.94$
4	$3.5 \leq \hat{y}_i < 4.5$	$5.94 \leq \hat{y}_i < 5.97$
5	$4.5 \leq \hat{y}_i < 5.5$	$5.97 \leq \hat{y}_i < 6.31$
6	$5.5 \leq \hat{y}_i < 6.5$	$6.31 \leq \hat{y}_i < 6.50$
7	$6.5 \leq \hat{y}_i < 6.5$	$6.50 \leq \hat{y}_i < 6.56$
8	$7.5 \leq \hat{y}_i < 6.5$	$6.56 \leq \hat{y}_i < 7.02$
9	$8.5 \leq \hat{y}_i < 6.5$	$7.02 \leq \hat{y}_i < 7.78$
10	$9.5 < \hat{y}_i$	$7.78 < \hat{y}_i$

Table 6.7 gives classification results using standard versus optimized regression to predict the Target value of the dependent measure, y_i . Given for both models for every Target value is the total number of correctly predicted Target values divided by the total number of instances of the Target value in the sample, and the sensitivity (accuracy) of the model for each Target value of the dependent measure.

Table 6.7: Predicting Actual Target Values by Standard and Optimized Multiple Regression

Target	Standard		Optimized	
	Regression		Regression	
1	0/42	0.0%	14/42	33.3%
2	0/45	0.0%	17/45	37.8%
3	1/52	1.9%	17/52	32.7%
4	4/60	6.7%	3/60	5.0%
5	21/70	30.0%	14/70	20.0%
6	44/92	47.8%	15/92	16.3%
7	19/89	21.4%	7/89	7.9%
8	0/69	0.0%	14/69	20.3%
9	0/60	0.0%	14/60	23.3%
10	0/50	0.0%	2/50	4.0%
ESS		0.86		11.18

Note: Greatest sensitivity indicated in **bold**.

Consistent with the prior examples, standard regression achieved greatest accuracy when predicting target values (4 - 7) near the sample mean (5.8). Also consistent across examples, standard regression correctly predicted few—presently 1 of the 318 most extreme target values (1 - 3, 8 - 10), for a paltry 0.3% success rate. In contrast, optimized regression consistently correctly predicted many more—here 78 (24.5%) of these most extreme values. Overall, the accuracy achieved by standard regression was negligible: $ESS = 0.9$. Optimized regression managed a weak effect ($ESS = 11.2$), representing a 1,144% boost in classification performance compared to standard regression. The UniODA solution was obtained in 22.2 CPU hours, running UniODA software on a 3 GHz Intel Pentium D microcomputer.

The middle value (center) of an ordinal categorical scale with an odd number of response options represents (by design) an uncertainty, equivalency, or indecision zone. For example, when assessing satisfaction (see Chapter 1) using a 7-point Likert-type scale it is common practice for values below the mid-

point to reflect increasingly dissatisfied responses, and for values above midpoint to represent increasingly satisfied responses. Decision-makers are typically primarily interested in predicting the extreme values on their measures, versus the values that lie near the middle of the scale. The final example of the use of UniODA to explicitly maximize *ESS* yielded by a regression model demonstrates the procedure for an application involving predicting patient satisfaction (if we had data involving weights for individual observations then there would also have been an example of maximizing weighted *ESS*). In addition, the following example conducts UniODA in order to understand the maximum accuracy that it is possible to attain for each attribute considered separately, and to establish a frame of reference for comparing the *ESS* attained by the standard and optimized regression models; aggregated confusion tables (see Chapter 3, Precision) are used to assess potential discriminatory power of the optimized regression model; and LOO analysis is conducted on the optimized models to assess their potential cross-generalizability.

In an effort to better understand patient ratings of overall satisfaction with care received in the Emergency Department (ED), the present study evaluates how to maximize satisfaction at every stage of care delivery in a patient's journey through the ED: this extends prior research¹⁸ assessing the influence of the registration process on overall satisfaction ratings, by also assessing the influence of patient interactions with nurses, doctors and technicians seen in the ED visit. Data were obtained from patients receiving care in the ED of a private Midwestern hospital, who were mailed, completed, and returned a satisfaction survey.¹⁹ The survey obtained ratings of *technicians*, *nurses*, and *doctors* seen by the patient during the ER visit. Ratings on all study variables were made using categorical ordinal Likert-type scales: 1 = very poor, 2 = poor, 3 = fair, 4 = good, and 5 = very good. Descriptive statistics for all study measures are presented in Table 6.8. As seen, all means exceeded scale midpoint value of 3 due to negatively skewed distributions, and modest variability indicates relatively homogeneous responding.

Table 6.8: Descriptive Statistics: All Study Measures

<i>Technicians (N = 535)</i>	Mean	SD	Median	Skewness	Kurtosis	CV
Overall satisfaction	4.14	1.15	5	-1.37	0.99	27.9
How well blood was taken	4.21	1.10	5	-1.60	1.96	26.2
Courtesy of person taking blood	4.36	0.94	5	-1.88	3.75	21.5
Waiting time in X-Ray	4.21	1.06	5	-1.49	1.70	25.2
Courtesy of X-Ray technologist	4.43	0.89	5	-1.92	4.01	20.0
<i>Nurses (N = 1,800)</i>						
Overall satisfaction	4.16	1.08	4	-1.40	1.38	25.9
Courtesy	4.37	0.87	5	-1.78	3.70	19.9
Took your problem seriously	4.31	0.94	5	-1.63	2.69	21.8
Attention paid to you	4.11	1.02	4	-1.20	1.07	24.9
Concern to keep you informed	4.04	1.09	4	-1.11	0.61	27.1
Concern for your privacy	4.12	1.01	4	-1.24	1.26	24.5
Technical skill	4.33	0.88	5	-1.70	3.50	20.4
<i>Doctors (N = 1,806)</i>						
Overall satisfaction	4.16	1.07	4	-1.38	1.34	25.9
Wait time to see doctor	3.74	1.21	4	-0.80	-0.23	32.3
Courtesy	4.43	0.84	5	-1.85	3.92	19.0
Took your problem seriously	4.39	0.91	5	-1.87	3.65	20.8
Concern for your comfort	4.30	0.93	5	-1.59	2.60	21.7
Explanation of test/ treatment	4.30	0.98	5	-1.48	1.90	22.9
Explanation of illness or injury	4.18	1.02	4	-1.34	1.36	24.4
Advice about self-care	4.22	1.00	5	-1.43	1.77	23.8

Before conducting planned regression analysis, UniODA was used to obtain a basic understanding of the ability of ratings of care-providers to predict overall satisfaction. Table 6.9 summarizes findings of analyses assessing the ability of the rated technician attributes to predict overall satisfaction ratings. All four technician ratings were statistically reliable predictors of overall satisfaction, and achieved moderate accuracy in training analysis. With the exception of blood-taking skill, *ESS* decreased in LOO analysis (this is indicated if second values of *p* and *ESS* are presented for a LOO-unstable attribute): this is interesting in the sense that all ratings of doctors were LOO-stable, as were most ratings of nurses (see ahead).

Table 6.9: UniODA Accuracy Predicting Overall Satisfaction: Technicians

Attribute	Predicted Rating	<i>N</i>	Predictive		
			Value (%)	<i>p</i> <	<i>ESS</i>
Phlebotomist skill	1	31	51.6	0.001	32.7
	2	46	19.6		
	3	16	50.0		
	4	160	50.6		
	5	282	75.5		
Phlebotomist courtesy	1	18	77.8	0.001	32.1
	2	8	25.0	0.001	30.6
	3	40	37.5		
	4	165	52.7		
	5	304	77.0		
X-Ray wait time	1	23	60.9	0.001	26.7
	2	56	14.3	0.001	19.6
	3	21	9.5		
	4	155	49.0		
	5	280	77.1		
X-Ray technician courtesy	1	13	92.3	0.001	27.6
	2	8	12.5	0.001	26.9
	3	42	28.6		
	4	147	55.1		
	5	325	73.8		

Note: UniODA models tested the exploratory hypothesis that ratings of the attribute predict ratings of overall satisfaction. *N* is number of patients having the indicated predicted rating. If classification accuracy declined in LOO analysis, then LOO *ESS* and *p* are given beneath the training results.

Table 6.10 summarizes findings of parallel analyses assessing the ability of rated nurse attributes to predict overall satisfaction ratings, and Table 6.11 summarizes findings for rated doctor attributes. LOO performance was lowest for ratings of X-Ray wait time: this is not surprising, as wait times for procedures, and to be seen by the doctor, are unreliable in the ED due to the need to triage real-time cases.

The next step of the analysis involved obtaining \hat{y}_i for each type of service provider. For *technicians* the regression model was: $\hat{y}_i = 0.092 + 0.26 * \text{phlebotomist skill} + 0.37 * \text{phlebotomist courtesy} + 0.14 * \text{X-Ray waiting time} + 0.17 * \text{X-Ray technician courtesy}$. This model was statistically significant [$F(4, 530) = 127.1, p < 0.0001, R^2 = 0.49$], and all independent variables made a statistically significant independent contribution to overall R^2 (p 's < 0.03). For *nurses* the regression model was: $\hat{y}_i = 0.45 + 0.35 * \text{took problem seriously} + 0.23 * \text{attention paid to patient} + 0.20 * \text{kept patient informed} + 0.11 * \text{technical skill}$. This model was statistically significant [$F(4, 1795) = 589.1, p < 0.0001, R^2 = 0.49$], and all independent variables made a statistically significant independent contribution to overall R^2 (p 's < 0.002). Finally, for *doctors* the regression model was: $\hat{y}_i = -0.23 + 0.25 * \text{wait time} + 0.24 * \text{took problem seriously} + 0.21 *$

concern for comfort + 0.12 * test/treatment explanation + 0.23 * home self-care advice. The model was statistically significant [$F(5,1800) = 789.2$, $p < 0.0001$, $R^2 = 0.69$], and all included independent variables made a statistically significant independent contribution to overall R^2 (p 's < 0.0001).

Table 6.10: UniODA Accuracy Predicting Overall Satisfaction: Nurses

<u>Attribute</u>	<u>Predicted Rating</u>	<u>N</u>	<u>Predictive Value (%)</u>	<u>p <</u>	<u>ESS</u>
Courtesy	1	42	71.4	0.001	28.7
	2	25	20.0		
	3	144	39.6		
	4	599	53.8		
	5	990	75.4		
Took your problem seriously	1	50	76.0	0.001	34.2
	2	45	28.9		
	3	168	37.5		
	4	578	56.1		
	5	959	77.2		
Attention paid to you	1	60	65.0	0.001	38.2
	2	78	25.6		
	3	263	33.5		
	4	608	54.8		
	5	791	84.1		
Kept you informed	1	77	55.8	0.001	36.4
	2	98	22.4		
	3	286	27.6		
	4	555	51.7		
	5	784	82.8		
Concern for your privacy	1	62	58.1	0.001	32.8
	2	61	23.0		
	3	270	28.2		
	4	612	51.8		
	5	795	82.0		
Technical skill	1	48	62.5	0.001	28.5
	2	17	29.4		
	3	164	33.5		
	4	640	52.7		
	5	931	78.0		

Note: See Note to Table 6.9.

Table 6.12 presents the cutpoints used on \hat{y}_i to obtain predicted ratings for all five values of the dependent measure. The first set of cutpoints given is for standard regression-based classification based on \hat{y}_i for 5-point Likert-type scales. Also shown are cutpoints for optimized regression models for all three service types, explicitly optimized for each sample via UniODA. For example, for $\hat{y}_i = 3.02$ the dependent measure is predicted to be 3 by the standard and optimized doctor MRA models, and is predicted to be 2 by the optimized nurse and technician regression models.

Table 6.11: UniODA Accuracy Predicting Overall Satisfaction: Doctors

<u>Attribute</u>	<u>Predicted Rating</u>	<u>N</u>	<u>Predictive Value (%)</u>	<u>p <</u>	<u>ESS</u>
Wait time	1	136	34.6	0.001	30.1
	2	152	12.5		
	3	341	20.5		
	4	595	43.4		
	5	582	86.2		
Courtesy	1	31	77.4	0.001	27.9
	2	122	23.8		
	3	36	8.3		
	4	545	55.8		
	5	1072	74.1		
Took your problem seriously	1	48	70.8	0.001	33.3
	2	41	29.3		
	3	127	44.9		
	4	528	59.5		
	5	1062	75.9		
Concern for your comfort	1	47	76.6	0.001	34.4
	2	173	19.6		
	3	46	32.6		
	4	599	55.1		
	5	941	78.8		
Test/treatment information	1	47	76.6	0.001	34.6
	2	76	26.3		
	3	177	31.6		
	4	568	56.0		
	5	938	79.2		
Illness/injury information	1	57	61.4	0.001	33.8
	2	79	20.2		
	3	224	32.6		
	4	562	54.8		
	5	884	80.1		
Self-care information	1	61	59.0	0.001	33.9
	2	220	15.4		
	3	57	33.3		
	4	560	56.4		
	5	908	80.3		

Note: See Note to Table 6.9.

Table 6.12: Model \hat{y}_i Cutpoints used to Obtain Predicted Overall Satisfaction Rating

Predicted Rating	Standard Regression	Optimized Regression		
		Nurses	Technicians	Doctors
1	$\hat{y}_i < 1.5$	$\hat{y}_i < 2.48$	$\hat{y}_i < 2.65$	$\hat{y}_i < 2.02$
2	$1.5 \leq \hat{y}_i < 2.5$	$2.48 \leq \hat{y}_i < 3.34$	$2.65 \leq \hat{y}_i < 3.39$	$2.02 \leq \hat{y}_i < 2.71$
3	$2.5 \leq \hat{y}_i < 3.5$	$3.34 \leq \hat{y}_i < 3.77$	$3.39 \leq \hat{y}_i < 3.68$	$2.71 \leq \hat{y}_i < 3.73$
4	$3.5 \leq \hat{y}_i < 4.5$	$3.77 \leq \hat{y}_i < 4.66$	$3.68 \leq \hat{y}_i < 4.49$	$3.73 \leq \hat{y}_i < 4.45$
5	$4.5 < \hat{y}_i$	$4.66 < \hat{y}_i$	$4.49 < \hat{y}_i$	$4.45 < \hat{y}_i$

Figure 6.1 illustrates these four sets of cutpoints. Rather than use a “one-set-fits-all” template to predict overall satisfaction ratings on the basis of \hat{y}_i as done in the standard regression approach, UniODA optimizes the set of cut-points to explicitly yield maximum classification accuracy for the sample. Here the optimized doctor model is most similar to the standard regression template, sacrificing some \hat{y}_i domain that the standard model uses to predict ratings of 2 (thinner yellow band), to use more of the \hat{y}_i domain to predict ratings of 1 (wider green band). The optimized technician and standard regression models are least similar: the technician model devotes a small portion of the \hat{y}_i domain to predict neutral ratings of 3.

Conceptually it is clear why OLS regression fails to explicitly achieve maximum possible *ESS* for a sample of data. *Theoretically*, regression is formulated to maximize the proportion of variance in (here) overall satisfaction ratings explained by a linear model, and the proportion of variance explained is not isomorphic with classification accuracy: the objective function that all ODA models explicitly maximize. *Empirically*, the combined effect of a distributional skew and the use of an ordinal discrete measurement scale—both of which result in violation of assumptions underlying OLS regression, cause the regression classification template to be overly conservative for estimates of the overall satisfaction ratings used by a minority of the sample presently. In general the inability of the regression template to adapt to conform to actual sample data is a clear disadvantage as regards accurate classification. *Paradoxically*, it is unlikely that the restrictive assumption underlying all linear models—that all independent variables apply equally to all the observations in the sample, are satisfied in the present research. Indeed, there is evidence that use of a general linear model likely induced Simpson’s Paradox presently, because *ESS* achieved using a single rating—*amount of attention paid to the patient*—to predict overall satisfaction via UniODA (38.2, in both training and LOO analysis) was greater than *ESS* for the regression model (34.6 in training, LOO not available) *involving four ratings*—one of which was *amount of attention paid to the patient*.

Table 6.13 summarizes classification results obtained using standard versus optimized regression models to predict patient overall satisfaction ratings. Shown for the standardized and optimized models, for all three service areas, and for all five overall satisfaction ratings, is the total number of correct predictions of the given rating divided by total number of instances of the given rating in the sample, and the corresponding sensitivity (accuracy) of the regression model for each actual rating. Considered across all three service areas, standard regression always achieved greater accuracy than optimized regression when predicting responses of 3 (fair), and optimized regression always achieved greatest accuracy in predicting the dissatisfied responses of 1 (very poor), or 2 (poor). Overall, accuracy yielded by all training models reflected moderate effects except for the optimized doctor regression model, which met the minimal criterion for a relatively strong effect. Nevertheless, in training analysis the *ESS* of the optimized model was greater than *ESS* of the standard model by 21.4% for the nurse model, 16.8% for the technician model, and 16.2% for the doctor model. The LOO performance for the optimized technician and nurse models exceeded the training performance of the corresponding standard regression models.

Table 6.14 gives confusion tables for the standard and optimized regression models, that clearly show the dominant negative skew: both the actual and predicted overall satisfaction ratings are primarily either 4 or 5.

Figure 6.1

Model-Specific Cutpoint-Based \hat{y}_i Domains Yielding Predicted Overall Satisfaction Ratings of 1 (Green), 2 (Yellow), 3 (Blue), 4 (Red), and 5 (Black)

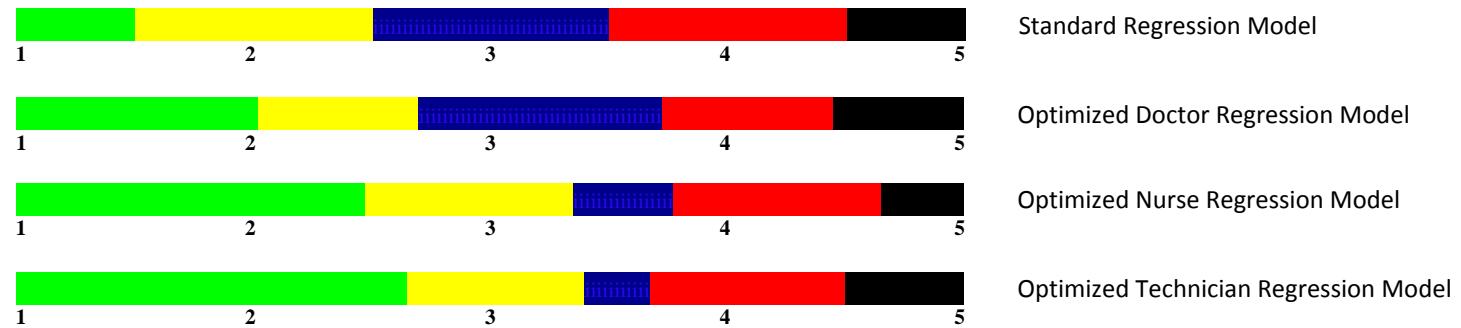


Table 6.13: Predicting Actual Overall Satisfaction Ratings by Standardized versus Optimized Regression Analysis

Actual Rating	Nurses				Technicians				Doctors			
	Standard		Optimized		Standard		Optimized		Standard		Optimized	
1	28/81	34.6%	51/81	63.0%	12/29	41.4%	17/29	58.6%	32/79	40.5%	47/79	59.5%
2	15/83	18.1%	39/83	47.0%	3/33	9.1%	13/33	39.4%	26/82	31.7%	34/82	41.5%
3	81/192	42.2%	45/192	23.4%	17/50	34.0%	7/50	23.4%	109/203	53.7%	122/203	60.1%
4	377/559	67.4%	350/559	62.6%	108/145	74.5%	101/145	69.7%	382/555	68.8%	330/555	59.5%
5	680/885	76.8%	636/885	71.9%	213/278	76.6%	213/278	76.6%	715/887	80.6%	736/887	83.0%
ESS	34.6		42.0		33.9		39.6		43.8		50.9	
LOO	34.7				35.0							

Note: Tabled for each *Actual* (overall satisfaction) *Rating* is the number of correctly predicted ratings (numerator); the total number of times each rating was used in the sample (denominator); and the percentage accuracy or *sensitivity* obtained in classifying each rating category. Greatest sensitivity obtained for each category is indicated in red. LOO analysis: was unavailable for regression in the SAS regression software we used; is superfluous for standard regression because the thresholds are fixed; and was abandoned for the optimized doctors model after failing to solve in two CPU days, running UniODA software on a 3 GHz Intel Pentium D microcomputer.

Table 6.14: Confusion and Aggregated-Confusion Tables for the **Optimized** and Standard Regression Models of Overall Satisfaction Ratings

		<u>Nurses</u>					<u>Technicians</u>					<u>Doctors</u>				
Actual Rating		Predicted Rating					Predicted Rating					Predicted Rating				
Rating	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	
1	51	16	4	5	5	17	3	2	4	3	47	16	12	4	0	
	28	23	19	6	5	12	4	5	5	3	32	27	16	4	0	
2	15	39	12	17	0	3	13	1	13	3	5	34	29	13	1	
	1	15	42	24	1	0	3	13	14	3	4	26	33	18	1	
3	18	65	45	55	9	9	11	7	18	5	6	22	122	40	13	
	3	15	81	81	12	0	5	17	23	5	2	12	109	68	12	
4	4	45	76	350	84	0	7	12	101	25	2	4	115	330	104	
	0	4	62	377	116	0	0	12	108	25	1	3	73	382	96	
5	2	9	15	223	636	1	4	4	56	213	0	1	8	142	736	
	2	0	13	190	680	0	0	6	59	213	0	0	5	167	715	
		1 or 2		4 or 5		1 or 2		4 or 5		1 or 2		4 or 5				
1 or 2		121		27		36		23		102		18				
		67		36		19		25		89		23				
4 or 5		60		1293		12		395		7		1312				
		6		1363		0		405		4		1360				
ESS=77.3					ESS=64.6					ESS=58.1					ESS=84.5	ESS=79.2

Note: **Bold** entries are for the optimized model.

Examination of aggregated confusion tables for the models (Table 6.14) indicates that performance of the standard and optimized regression models was approximately comparable for patients predicted to score 4 or 5 (satisfied) across service areas. In contrast, the optimized technician (48 versus 19) and nurse (181 versus 73) models were much more likely than standard regression to accurately predict scores of 1 or 2 (dissatisfied), although the difference was marginal for the doctor model (109 versus 93)—for which the classification template was the most similar to the standard regression model (Figure 6.1). If the phenomena assessed by items in the scale were actionable in real-time, and if a potent intervention existed to address such phenomena in real-time, then the optimized nurse and technician MRA models could be used to prevent substantially more cases of dissatisfaction, as compared with the standard regression model.

Optimizing Regression Models

SAS™ syntax for creating the requisite data set, and UniODA™ and MegaODA™ syntax for maximizing the *ESS* obtained by a multiple regression analysis (MRA)-based model involving a categorical ordinal (Likert-type) dependent (class) variable having ten or fewer response options, is relatively straightforward.²⁰ Here it is assumed that the MRA-based model is already identified.

The first step involves creating a new data set. Imagine that the original data are in a file called *original.dat* with *N* rows and 4 space-delimited variables: DV and IV1 - IV3. The SAS™ code given below reads this file; computes \hat{Y}_i (*Ypred*) for every observation using a hypothetical MRA- based equation; computes \hat{Y}_{i-INT} (*MRApred*) via standard MRA cutpoints; sorts data by DV and *MRApred* and prints a report of *N* for every combination (for table preparation); and saves a space-delimited ASCII analysis file with DV and *Ypred*, called *maximize.dat*.

```

data q;
infile 'c:\original.dat';
input DV IV1 IV2 IV3;
Ypred=0.45-.12*IV1+.66*IV2-.1*IV3;
if Ypred<=1.5 then MRApred=1;
if Ypred>1.5 and Ypred<=2.5 then
  MRApred=2;
if Ypred>2.5 and Ypred<=3.5 then
  MRApred=3;
if Ypred>3.5 and Ypred<=4.5 then
  MRApred=4;
if Ypred>4.5 then DVpred=5;
proc sort;by DV MRApred;
proc means n;var DV;by DV MRApred;
data q2;
set q;
file 'c:\maximize.dat';
put DV Ypred;
run;
```

The second step involves maximizing *ESS* of the MRA model via UniODA. This is accomplished by running the following UniODA or MegaODA software syntax, in order to predict actual score (DV, the class variable) using predicted score (*Ypred*, the attribute):

```

OPEN maximize.dat;
VARS DV Ypred;
CLASS DV;
ATTR Ypred;
```

DIR < 1 2 3 4 5
MCARLO ITER 25000;
GO;

The *a priori* hypothesis specified by the DIR command indicates *Ypred* is hypothesized to be lower for DV scores of 1 than for 2, lower for DV scores of 2 than for 3, etcetera. Finally, imagine that cutpoints identified by UniODA for classifying predicted score (1 - 5) were, in order: 0.5, 2.1, 2.9, and 3.4. The SAS™ code below assigns observations optimal predicted scores (called *DVpred*) using these cutpoints; sorts the data by actual (DV) and predicted (*DVpred*) score; and prints a report of *N* for every combination of actual and optimal predicted score (used in table preparation).

```

if Ypred<=.5 then DVpred=1;
if Ypred>.5 and Ypred<=2.1 then DVpred=2;
if Ypred>2.1 and Ypred<=2.9 then DVpred=3;
if Ypred>2.9 and Ypred<=3.4 then DVpred=4;
if Ypred>3.4 then DVpred=5;
proc sort;by DV DVpred;
proc means n;var DV;by DV DVpred;
run;

```

For regression models involving class variables and attributes that all have more than ten levels, how can ESS be explicitly maximized? The most straightforward option is divide the dependent measure into ten categories (e.g., deciles, each of size = domain / 10), and treat these as ten ordered class variable categories. However, combining groups in an arbitrary manner can induce Simpson's paradox (Chapter 9), which must be assessed if this option is selected. Alternative strategies providing more granular solutions have been suggested, but haven't yet been investigated in the ODA laboratory.²¹

Analysis of Variance

When they are properly specified the analysis of variance (ANOVA) and multiple regression analysis (MRA) methods identify identical p and R^2 results for a given application, although ANOVA is typically used if the independent variables (attributes) can assume only a few levels and/or are categorical, and for specific types of experimental designs.^{1-5,22-24}

One-way ANOVA involving a single class (dependent) variable with ten or fewer (un)ordered class categories is discussed in Chapter 5, and the procedures discussed above for optimizing regression models in which the class variable has more than ten levels may be adapted for one-way ANOVA. The simplest ANOVA design has two class-categories, in which case data are usually analyzed using t -test (Chapter 5).

A main-effects ANOVA model is a regression model having an intercept and a beta coefficient for each attribute (independent variable) used in the model (all regression models include an error term that induces indeterminacy as it approaches zero⁵). Shown earlier in this Chapter, combining different groups reflected by different levels on attributes (e.g., rich and poor, young and old, Republican and Democrat) can induce paradoxical confounding (Chapter 9): for example, in the cited example the regression model using four attributes yielded lower ESS than a UniODA model involving only one of the four attributes. In the ODA paradigm main-effects designs are analyzed using CTA models that eliminate these issues, as is discussed in the final four Chapters of this book.

A factorial ANOVA model is a main-effects ANOVA model that also includes interaction terms. For example, imagine a design involving two attributes, X1 and X2. A main effects ANOVA model includes an interaction term and terms (beta coefficients) for X1 and for X2, and a factorial model also includes a term for the product of the attributes: X1 * X2: two-factor products are known as *first-order interactions*; three-factor products (X1 * X2 * X3) are known as *second-order products*, and so forth. The higher the order of interaction the greater the likelihood of collinearity, and the more difficult it becomes to interpret the nature and conceptual meaning of the interaction.^{5,24} Factorial models are linear because the underlying regression model is linear in its coefficients. In ODA, CTA is used to analyze this data geometry.

Finally, a factorial ANOVA application involving one or more putative confounding variables that are to be statistically controlled (i.e., their effect is removed prior to evaluating other attributes) is known as an analysis of covariance (ANCOVA, or ANACOVA).²⁴ The identification and eradication of confounding in the ODA paradigm is discussed in Chapter 9.

Linear Discriminant Function

Fisher's linear discriminant analysis (FLDA) is a widely-used suboptimal multivariable classification method in the GLM paradigm, that requires data to conform to the standard assumptions underlying multivariate linear parametric statistical procedures.²⁴⁻²⁷ FLDA gives an equation (*response function*) whereby values on the attributes are combined into a single predicted response function score (\hat{y}_i) for each observation. To make a classification, an observation's \hat{y}_i score is compared against a fixed (template) criterion: if $\hat{y}_i \leq 0$, then predict class = 0; otherwise predict class = 1. The identical UniODA-based optimization procedure used to maximize the ESS yielded by a standard (template) regression model may be used to maximize the

ESS achieved by an FLDA model, by identifying a new combination of *cutpoint* and *direction* that explicitly maximizes (weighted) *ESS* for the training sample:

1. obtain the suboptimal model;
2. using the suboptimal model, obtain \hat{y}_i for each observation;
3. perform priors-weighted UniODA (class = group) using \hat{y}_i as the attribute;
4. use the resulting UniODA model to make classifications and compute training *ESS*; and
5. conduct LOO analysis to evaluate potential cross-generalizability of the optimized FLDA model.

If, in addition to the training sample, at least one hold-out validity sample is also available, then:

6. use the optimized FLDA model to classify observations in the hold-out sample(s) and compute the hold-out *ESS*.

Worked examples of this procedure are available.⁹⁻¹¹ Research investigating the performance of UniODA-based optimization in improving the training and hold-out classification *ESS* of FLDA (and also of logistic regression analysis) is encouraging. A meta-analysis of the results of UniODA-based optimization of FLDA (and logistic regression) models for 15 data sets representing a variety of substantive areas and application configurations (i.e., combinations of sample size, class category sample size balance, number and metric of attributes) found that UniODA-based optimization yielded a mean increase of 5.5% in *ESS* in training analysis, and 4.8% in hold-out validity analysis.¹⁰ Since this meta-analysis was conducted more studies reported enhanced classification performance using this methodology.²⁸⁻³² Chapter 8 covers linear discriminant models that explicitly maximize (weighted) classification accuracy.

Chapter 7

Optimized Maximum-Likelihood Models

Chapter 7 discusses how to explicitly maximize the classification performance (*ESS*) achieved by popular models representing the maximum-likelihood (ML) paradigm, including the log-linear model, probit analysis, and logistic regression analysis.^{1,2}

Log-Linear Model

Chapters 3 and 4 presented examples of applications in which the log-linear model was unable to identify a model that fit the empirical data, whereas UniODA identified a statistically reliable model of moderate strength. Exploratory log-linear analysis involves evaluating the goodness-of-fit and residual distributions for all possible linear models.³ For example, for a cross-classification application involving three attributes (A, B, C), the series of log-linear models that may be evaluated include: (1) intercept only; (2) intercept and main effect of A; (3) intercept and main effect of B; (4) intercept and main effect of C; (5) intercept and main effects of A and B; (6) intercept and main effects of A and C; (7) intercept and main effects of B and C; (8) intercept and main effect of A, B, and C; (9) intercept and main effects of A and B, and interaction of A and B; (10) intercept and main effects of A and C, and A * C; (11) intercept and main effects of B and C, and B * C; (12) intercept and main effects of A, B, and C, and A * B; (13) intercept and main effects of A, B, and C, and A * C; (14) intercept and main effects of A, B, and C, and B * C; (15) intercept and main effects of A, B, and C, and A * B and A * C; (16) intercept and main effects of A, B, and C, and A * B and B * C; (17) intercept and main effects of A, B, and C, and A * C and B * C; (18) intercept and main effects of A, B, and C, and A * B, A * C, and B * C; (19) intercept and main effects of A, B, and C, and A * B, A * C, B * C, and A * B * C. The functional form of the model parallels the equation for a linear model (“formula”) seen in OLS regression analysis and FLDA (Chapter 6), as well as Probit and logistic regression models (discussed ahead). Thus, the same UniODA procedure described in Chapter 6 may be used to explicitly maximize the *ESS* each log-linear model achieves for the sample. This approach runs the risk of inducing paradoxical confounding (as is true for all liner models), data must be consistent with the assumptions required by the method, and there is no guarantee that the maximum-accuracy solution will be identified—particularly if the maximum-accuracy solution is non-linear. Marginal imbalances and sparse tables cause assumption violations, and can induce numerical instability in parameter estimation.³ An example of these issues and of the use of CTA to mitigate them is discussed ahead under logistic regression analysis.

Probit Model

Probit analysis (PA) is gaining in popularity, for example in political science research seeking accurate models of court decision-making.⁴⁻¹⁰ For applications having a binary class variable and two or more attributes, PA allows assessment of the independent relationship between class variable and attribute. Parameter estimates are obtained by maximum-likelihood, and indicate the amount of change in the cumulative normal probability function that is associated with a one-unit change in the attribute value. Assessing the goodness-of-fit of PA models is traditionally accomplished using R^2 and chi-square analysis, but this has been criticized.¹¹ The primary objective of all classification models is the ability to make accu-

rate classifications. Although PA doesn't explicitly maximize classification accuracy, the *ESS* yielded by PA models may be maximized by optimizing the models decision-making criterion, exactly as is accomplished for the other suboptimal linear multiattribute models covered in Chapters 6 and 7.

To illustrate this method consider the asylum-related appeals to the federal courts covering the period of 1980-1987, constituting 137 cases.^{5,12} The class variable indicated if aliens won ($N = 59$) or lost ($N = 78$) their appeal. Six binary attributes included if: any organizations were involved in the appeal; the alien was from a country hostile to the USA; the alien was from Europe; the court was located in the Western USA; a high percentage of the judges involved in the appeal were appointed by a Democratic President; and whether there was a high level of immigrant-flow into the circuit. The resulting PA model correctly predicted 71.2% of the wins and 55.1% of losses, resulting in *ESS* = 26.4. UniODA was then used to optimize the model: the adjusted decision criterion for the PA model was: if $\hat{y}_i > 0.025$ predict class = 1 (win); otherwise predict class=0 (loss). The optimized PA model correctly predicted 64.4% of the wins and 71.8% of losses yielding *ESS* = 36.2, a 37% increase in accuracy.

Logistic Regression

Logistic regression analysis (LRA) is a very widely-used suboptimal ML linear multiattribute classification method, the "method of choice" when the class (dependent) variable is binary, though LRA is also used in applications having more than two class categories (multinomial LRA).¹³⁻¹⁸ As was the case for other linear suboptimal multiattribute models covered in this and the prior Chapter, LRA gives an equation (response function) whereby values on the attributes are combined into a single predicted response function score (\hat{y}_i) for each observation. To make a classification, an observation's \hat{y}_i score is compared against a fixed (template) criterion: if $\hat{y}_i \leq 0.5$, then predict class = 0; otherwise predict class = 1. The identical UniODA-based optimization procedure used to maximize the *ESS* yielded by an FLDA model (Chapter 6) is used to maximize the *ESS* achieved by a LRA model, by identifying a new combination of *cutpoint* and *direction* that explicitly maximizes (weighted) *ESS* for the training sample.

Worked examples of this procedure are available.¹⁹⁻²¹ Research investigating the performance of UniODA-based optimization in improving the training and hold-out classification *ESS* of LRA (and FLDA) is motivating. Meta-analysis of the results of UniODA-based optimization of LRA (and FLDA) models for 15 data sets representing a variety of substantive areas and application configurations (i.e., combinations of sample size, class category sample size imbalance, number and metric of attributes) found UniODA-based optimization had mean *ESS* increase of 5.5% in training analysis, and 4.8% in hold-out validity analysis.²⁰

Given the popularity of the linear methods discussed here and in the prior Chapter, researchers who learn of the boosted performance achieved using UniODA-based (weighted) *ESS* optimization may be motivated to investigate its use in a host of different applications. As a means of facilitating such research, here Dr. Fred Bryant explains the steps involved and provides the SPSS™ syntax needed to run two-group LRA using *SPSS 17 for Windows™*, and output to an ASCII space-delimited data file the binary class variable and predicted probability of group membership (i.e., \hat{y}_i) from an SPSS-conducted LRA.²²

1. Obtain an SPSS data set containing a binary class variable (e.g., sex), along with categorical (e.g., city1, city2, city3, colorA, colorB, colorC) and continuous (e.g., age) attributes. Missing data should be indicated with a value (e.g., -9) in the SPSS data set.

2. Open the SPSS data set, and run the following syntax file, which saves predicted probability of group membership as a variable named PRED_1 in the active SPSS data file.

```
LOGISTIC REGRESSION VARIABLES sex /SAVE=PRED  
/METHOD=ENTER age raceA raceH city2 city3 /CLASSPLOT  
/CONTRAST (city3)=Indicator /PRINT=GOODFIT  
/CONTRAST (city2)=Indicator /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20)  
/CONTRAST (colorA)=Indicator CUT(0.5).  
/CONTRAST (colorC)=Indicator
```

3. If desired, in Variable View, edit the SPSS data file to rename PRE_1 as "Iryhat," for example, to reflect "logistic regression \hat{y}_i ".

4. From the drop-down SPSS Windows menu, select Transform, Recode into Same Variable, and change the value of "system missing" (blank) to -9 (or value used) for the PRE_1 (Iryhat) variable. Then resave the SPSS data set.

5. Run the following SPSS syntax to write a space-delimited ASCII data file named "Iryhat.dat" that contains a code for the class variable (e.g., sex) and the predicted probability of group membership (e.g., Iryhat):

```
FORMATS sex (f4.0).          /1 sex Iryhat.  
FORMATS Iryhat (f13.8).       execute.  
write outfile='c:\Iryhat.dat' records=1
```

6. Locate the file "Iryhat.dat" in the root folder for the c:\ drive, and move this file to the ODA directory for analysis.

Categorical Attributes Having Many Levels

Attributes measured on a categorical response scale are ubiquitous in the literature. Categorical scales for attributes such as political affiliation, ethnic origin, marital status, state of residence, or diagnosis may consist of many qualitative response categories. For example, imagine a hypothetical study using homeowner satisfaction (satisfied or dissatisfied) as the class variable. It is a cliché that an important multi-categorical attribute in this application is "*location, location, location*". Examples of location categories are region or state in a national study, county or municipality in a state study, or neighborhood or street in a municipal study. For exposition, consider a national study: with 50 categories it is unlikely that a statistically or conceptually compelling model would be identified if *state* was employed as a multi-categorical attribute. However, data at the state level are available on many attributes that may potentially influence homeowner satisfaction, such as measures of annual average per-capita savings, crude mortality rate, low winter temperature, high summer temperature, beachfront acreage, public school academic performance rank, local taxes, number of children per household, and crime rate for example. Such measures are reported using real-number, interval, ordinal, or nominal scales. In this way the likely non-interpretable 50-category variable is transformed into a panel of easily interpreted, theoretically justifiable attributes.²³

Categorical Attributes May Overwhelm Linear Models

Many empirical studies in the literature involve categorical attributes (independent variables). Sometimes studies report only two categorical variables: a class variable and an attribute in the ODA paradigm, or an independent and dependent variable in legacy statistical paradigms. When the categorical variables have response scales with a different number of categories, the resulting data geometry is called a *rectangular categorical design* (RCD). As demonstrated in earlier examples marginal imbalance and sparse cells occur commonly in RCDs and can induce interpretive and conceptual difficulties for chi-square, and numerical instability for log-linear and other linear parametric multiattribute methods based on chi-square.^{1-3,17,24-26}

In multiattribute designs categorical attributes using a binary response scale may be directly incorporated into linear parametric models, but multicategorical attributes must be disaggregated (see Chapter 2, Two Common Mistakes) and require specification of a reference group—which can reduce ESS and parsimony, and may induce paradoxical confounding—identifying phantom effects and masking actual effects (see Chapter 2; demonstrated ahead).

Using UniODA and CTA the analysis of RCDs straightforward.²⁷ These methods are illustrated for an example involving randomly selected patients with *Pneumocystis carinii* pneumonia (PCP).²⁸ Four categorical variables used in analysis include patient status and gender (each with two categories), city of residence (seven categories), and type of health insurance (ten categories). Examination of the cross-tabulations of these variables makes it obvious why conventional statistical methods such as chi-square

analysis, LRA, and log-linear analysis are inappropriate for and are easily overwhelmed by such designs. The reader is encouraged to experiment using parametric methods with the analytic problems that are presented below, in order to gain a first-hand understanding of the complexities that are involved using parametric methods, and the simple, transparent, intuitive, accurate, and reproducible models that are effortlessly obtained via ODA.

Data for a random sample of 1,568 PCP patients are *gender* (male = 1, female = 2), *status* (alive = 0, died = 1), *city* of residence (Los Angeles or LA = 1, Chicago = 2, New York or NY = 3, Seattle = 4, Miami = 5, Nashville = 6, Phoenix = 7); and type of insurance or *insure* (1 = Medicaid, 2 = Medicare, 3 = unused because $N = 0$, 4 = fee for service, 5 = PPO, 6 = POS, 7 = managed care, 8 = HMO, 9 = private non-HMO, 10 = self-pay, 11 = charitable organization).²⁸ First, UniODA is conducted for each of the six different bivariate pairings of these four variables.

Status and Gender: Table 7.1 presents the 2 x 2 cross-tabulation of status and gender.

Table 7.1: Status and Gender

	Females	Males
Alive	278	937
Died	23	100

The non-directional hypothesis that women and men had a different mortality rate was tested by running the following UniODA and MegaODA software syntax: MC simulation wasn't used to estimate exact p because (only) for binary designs without return weights, p obtained by the UniODA randomization algorithm and Fisher's exact test are isomorphic):

```
VARS gender status city insure;
CATEGORICAL gender;
CLASS status;
GO;
ATTR gender;
```

The UniODA model was: if gender = female, the predict status = alive; otherwise predict status = died. The model was not statistically significant ($p > 0.10$) and achieved negligible accuracy ($ESS = 4.2$) and predictive value ($ESP = 2.0$). There is no evidence that men and women had different mortality rates.

Status and City: Table 7.2 presents the 2 x 7 cross-tabulation of status and city.

Table 7.2: Status and City

	Alive	Died
LA	190	12
Chicago	250	15
NY	316	44
Seattle	165	17
Miami	133	16
Nashville	95	10
Phoenix	66	9

The exploratory hypothesis that the cities had different mortality rates was tested by running the following appended UniODA and MegaODA software syntax:

```
ATTR city;CAT city;MC ITER 10000;GO;
```

The resulting model was: if city = LA or Chicago predict status = alive; otherwise predict status = died. The model was statistically significant ($p < 0.035$), with weak accuracy ($ESS = 14.3$) and negligible predictive value ($ESP = 5.2$). Table 7.3 presents the resulting confusion table. Findings thus far may be symbolically indicated as: (LA, Chicago) > Rest; where parentheses indicate it hasn't yet been determined

if embedded cities are significantly different on status; > indicates a significantly greater proportion of living patients; and Rest indicates all other cities in the sample.

Table 7.3: Confusion Table for UniODA Model Predicting Status Based on City

		Predicted Status		
		Alive	Died	
Actual Status	Alive	440	775	36.2%
	Died	27	96	78.1%
		94.2%	11.0%	

The second step of this *optimal range-test* procedure involves two comparisons, one between LA and Chicago, and another between the other five cities (Chapter 2 discusses control of experimentwise p). The UniODA and MegaODA syntax for the first test is appended as follows:

EX city>2;MC ITER 10000;GO;

The resulting model was not statistically significant ($p > 0.10$), with minute accuracy ($ESS = 1.3$) and predictive value ($ESP = 0.3$), so the symbolic representation of the effect thus far remains unchanged. The UniODA and MegaODA syntax for the second test is replaced as follows:

EX city<3;GO;

The resulting model was not statistically significant ($p > 0.10$), with negligible accuracy ($ESS = 5.9$) and minute predictive value ($ESP = 2.3$), so the symbolic representation of the effect is complete. There is evidence that the mortality rate is comparable in LA and Chicago, and is significantly lower than the (comparable) mortality rates in NY, Seattle, Miami, Nashville, and Phoenix.

Conducting all-possible comparisons (e.g., via chi-square) for pairs of these seven cities requires running and integrating $(7 \times 6) / 2 = 21$ analyses, with final runs using a SIDAK criterion for 21 tests. In contrast the optimal range-test procedure required 3 tests. The SIDAK criterion for 3 versus 21 tests is target $p < 0.017$ and $p < 0.0025$, respectively (see Chapter 2).²¹

Status and Insurance: Table 7.4 gives the 2x10 cross-tabulation of status and insurance.

Table 7.4: Status and Insurance

	Alive	Died
Medicaid	127	16
Medicare	68	6
Fee for Service	78	4
PPO	92	8
POS	471	49
Managed Care	32	5
HMO	92	10
Private non-HMO	52	5
Self-Pay	38	5
Charitable Group	165	15

The exploratory hypothesis that different types of insurance had different mortality rates was tested by appending and running the following UniODA or MegaODA syntax:

ATTR insure;CAT insure;GO;

The resulting UniODA model was: if insurance = Medicare, fee for service, PPO, private non-HMO, or charitable group, predict status = alive; for all other insurance categories predict status = died. The model was not statistically significant ($p > 0.10$), with negligible accuracy ($ESS = 6.6$) and predictive value ($ESP = 2.4$). There thus is no evidence that different types of insurance are associated with different mortality rates. To conduct all-possible comparisons for pairs of these ten insurance categories requires running and integrating $(10 \times 9) / 2 = 45$ analyses, with final runs using a SIDAK criterion for 45 tests of target $p < 0.00114$. In contrast the UniODA range-test procedure used one test, with target $p < 0.05$.

Gender and City: Table 7.5 gives the 2x7 cross-tabulation of gender and city.

Table 7.5: Gender and City

	Females	Males
LA	18	184
Chicago	41	224
NY	116	244
Seattle	78	104
Miami	11	138
Nashville	25	80
Phoenix	12	63

The exploratory hypothesis that women and men were distributed differently in different cities was tested by appending and running the following UniODA or MegaODA syntax:

CLASS gender;ATTR city;CAT city;GO;

The UniODA model was: if city = LA, Chicago, Miami, or Phoenix, predict gender = male; for all other cities predict gender = female. The model was statistically significant ($p < 0.0001$), with moderate accuracy ($ESS = 31.5$) and weak predictive value ($ESP = 22.0$). Table 7.6 is the resulting confusion table.

Table 7.6: Confusion Table for UniODA Model Predicting Gender Based on City

		Predicted Gender		Actual Gender	Male	Female	58.7%
		Male	Female				
Actual Gender	Male	609	428				
	Female	82	219				
		88.1%	33.8%				

Findings thus far are symbolically indicated with respect to proportion of females as: (NY, Seattle, Nashville) > (LA, Chicago, Miami, Phoenix). The second step of optimal range-test procedure involves two comparisons, the first between NY, Seattle, and Nashville, and the second between LA, Chicago, Miami, and Phoenix. The first test was run by appending the following UniODA or MegaODA syntax:

EX city=3;EX city=4;EX city=6;GO;

The resulting UniODA model was: if city = LA or Miami predict gender = male; if city = Chicago or Phoenix predict gender = female. The model was statistically significant ($p < 0.0081$), with weak accuracy ($ESS = 17.5$) and negligible predictive value ($ESP = 7.3$): Table 7.7 gives the resulting confusion table.

Table 7.7: Confusion Table for UniODA Model, City and Gender: LA, Chicago, Miami, Phoenix

		Predicted Gender		
		Male	Female	
Actual Gender	Male	322	287	52.9%
	Female	29	53	64.6%
		91.7%	15.6%	

The symbolic representation of the effect thus far is: (NY, Seattle, Nashville) > (Chicago, Phoenix) > (LA, Miami). The second test was run by appending the following UniODA or MegaODA syntax:

```
EX city<3;EX city=5;EX city=7;GO;
```

The resulting UniODA model was: if city = NY or Nashville, predict gender = male; if city = Seattle, predict gender = female. The model was statistically significant (experimentwise $p < 0.05$), but it has weak accuracy ($ESS = 11.3$) and predictive value ($ESP = 12.5$): Table 7.8 gives the confusion table.

Table 7.8: Confusion Table for UniODA Model, City and Gender: NY, Seattle, Nashville

		Predicted Gender		
		Male	Female	
Actual Gender	Male	324	104	75.7%
	Female	141	78	35.6%
		69.7%	42.9%	

The symbolic representation of the effect thus far is: Seattle > (NY, Nashville) > (Chicago, Phoenix) > (LA, Miami). The third step of this optimal range-test procedure involves three comparisons, one for each set of parentheses remaining in the symbolic representation. The UniODA and MegaODA syntax used for the first test is:

```
EX city=1;EX city=2;EX city=4;EX city=5;EX city=7;GO;
```

The resulting model was not statistically significant ($p > 0.10$), with negligible accuracy ($ESS = 7.0$) and predictive value ($ESP = 8.4$), and so the symbolic representation thus far remains unchanged. Similar results were obtained for the second and third tests so the symbolic representation is complete.

Obtaining the confusion table for the final UniODA model requires integrating confusion tables for the two halves of the analysis. Adding corresponding entries in confusion tables for the first (Table 7.7) and second (Table 7.8) analyses creates the integrated table in Table 7.9: $ESS = 5.8$, $ESP = 4.3$.

Table 7.9: Confusion Table for Final UniODA Model Predicting Gender Based on City

		Predicted Gender		
		Male	Female	
Actual Gender	Male	646	391	62.3%
	Female	170	131	43.5%
		79.2%	25.1%	

The proportion of females in the sample is significantly greater in Seattle than in NY or Nashville (that are statistically comparable), which have a significantly greater proportion of female patients in the sample than Chicago or Phoenix (that are statistically comparable), which have a significantly greater pro-

portion of female patients than LA or Miami. To conduct all-possible comparisons for pairs of seven cities requires running and integrating 21 analyses, with final runs using a SIDAK criterion of target $p < 0.00244$. In contrast the optimal range-test procedure used six tests for target $p < 0.008513$.

Gender and Insurance: Table 7.10 is the 2x7 cross-tabulation of gender and insurance.

Table 7.10: Gender and Insurance

	Females	Males
Medicaid	18	125
Medicare	17	57
Fee for Service	12	70
PPO	8	92
POS	152	368
Managed Care	8	29
HMO	25	77
Private non-HMO	12	45
Self-Pay	14	29
Charitable Group	35	145

The exploratory hypothesis that women and men had different types of insurance coverage was tested using the following appended UniODA and MegaODA syntax:

CLASS gender;ATTR insure;CAT insure;GO;

The resulting UniODA model was: if insurance = Medicaid, fee for service, PPO, managed care, private non-HMO, or charitable group, then predict gender = male; if insurance = Medicare, POS, HMO, or self-pay, predict gender = female. The model was statistically significant ($p < 0.0001$), with weak accuracy ($ESS = 17.9$) and predictive value ($ESP = 12.6$). Table 7.11 presents the resulting confusion table.

Table 7.11: Confusion Table for UniODA Model Predicting Gender Based on Insurance

		Predicted Gender		
		Male	Female	
Actual Gender	Male	506	531	48.8%
	Female	93	208	69.1%
		84.5%	28.2%	

The findings thus far are symbolically indicated with respect to proportion of females as: (Medicare, POS, HMO, self-pay) > (Medicaid, fee for service, PPO, managed care, private non-HMO, charitable group). The second step of this optimal range-test procedure involves two comparisons, one comparison for each set of parentheses: UniODA and MegaODA syntax for the first test is appended as follows:

EX insure=1;EX insure=4;EX insure=5;EX insure=7;EX insure=9;EX insure=11;GO;

The resulting UniODA model was not statistically significant ($p > 0.10$), with negligible accuracy ($ESS = 5.0$) and predictive value ($ESP = 5.6$), so symbolic representation remains unchanged. UniODA and MegaODA code for the second test is appended as follows:

EX insure=2;EX insure=6;EX insure=8;EX insure=10;GO;

The resulting UniODA model was: if insurance = Medicaid, fee for service, or PPO, predict gender = male; if insurance = managed care, private non-HMO, or charitable group, predict gender = female. The

model was not statistically significant at the experimentwise criterion, however it met the generalized “per-comparison” criterion for $p < 0.05$: $ESS = 15.9$, $ESP = 8.4$. The symbolic notation is thus complete, unless it is decided to include the effect significant at the generalized criterion, in which case final symbolic notation would be: (Medicare, POS, HMO, self-pay) > (Medicaid, fee for service, PPO) > (managed care, private non-HMO, charitable group). Females in the sample are *most* (comparably) likely to have Medicare, POS, HMO, or self-pay health coverage; significantly (comparably) *less* likely to have Medicaid, fee for service, or PPO health coverage; and significantly (comparably) *least* likely to have managed care, private non-HMO, or charitable group health coverage. The UniODA range-test involved three (experimentwise criterion) or five (generalized criterion) tests of statistical hypotheses, versus 45 needed for all possible comparisons.

City and Insurance: The final univariate analysis, Table 7.12 is the 7×10 cross-tabulation of city and insurance. Cell entries indicated in **bold** are very small: as seen, analysis by chi-square, LRA, log-linear model, and other ML-based methods is inappropriate because the minimum expectation is too small in too many cells.^{1,2,29} The exploratory UniODA model “CLASS city; ATTR insure;) was statistically significant using 1,000 MC experiments ($p < 0.001$), and achieved moderate accuracy ($ESS = 39.0$) and predictive value ($ESP = 41.5$). However, the model was degenerate since no observations were predicted to reside in Seattle (in Table 7.12 insurance categories that predict a given class category are indicated via **red N**).

Table 7.12: City and Insurance

	LA	Chi	NY	Sea	Mia	Nas	Pho
Mcaid	43	23	61	5	8	0	3
Mcare	23	15	5	19	8	0	4
FFS	36	16	1	7	15	2	5
PPO	11	26	29	7	19	1	7
POS	68	64	257	73	58	0	0
MCare	0	0	0	0	0	0	37
HMO	2	0	0	0	1	102	0
nHMO	5	19	2	20	0	0	11
Self	7	19	3	5	4	0	5
Charity	9	83	2	46	37	0	3

Based on the present analysis, type of health insurance coverage is not comparably represented across cities. No procedure has yet been developed to disentangle effects (degenerate or not) in *super-categorical designs* involving two or more multicategorical attributes each with response scales consisting of three or more categories: thus, disentangling such effects constitutes an application-specific enterprise. Has the reader tried to analyze these data by chi-square, LRA, log-linear model, or other legacy methods?

Multiattribute Analysis: Exposition turns to multiattribute analyses in purely categorical designs: the analysis will treat gender as the class variable, and status, city and insurance as possible attributes.

Compared with the analytically troublesome data in Table 7.12, the cross-tabulation results given in Table 7.13 might well be described as “the end of the linear statistical analysis world” because the data cannot meet the assumptions required by parametric methods.^{1,2,24-26} In Table 7.13 cell entries indicated in **red** are very small and render analysis by chi-square, LRA, log-linear model, probit, and other methods based on chi-square unsuitable because the expected value is too small in too many cells.²⁹ Recalling the visualization exercise in Chapter 3, can the reader mentally “see” the data geometry for this application?

Computing the total number of cells in a *cross-tabulation* of all categorical data as seen in Table 7.13 requires obtaining the product of the number of response categories for all variables. Here status and gender both have 2 response categories, city has 7, and insurance has 10, so a total of $2 \times 2 \times 7 \times 10 = 280$ cells exist in Table 7.13. If observations were distributed uniformly in the cells (the opposite is in fact true), then on average 5.6 observations would exist in every cell of the cross-tabulation table. As a means of maximizing perspective concerning the complexities involved in modeling such data, it is recommended that readers attempt to analyze these data using various suboptimal methods.

Table 7.13: Distribution of Four Cross-Tabulated Categorical Variables

<u>GENDER</u>	<u>STATUS</u>	<u>CITY</u>	<u>INSURANCE</u>	<u>N</u>			<u>Local Charity</u>	<u>0</u>	
Female	Alive	LA	Medicaid	0		Phoenix	Medicaid	0	
			Medicare	2			Medicare	0	
			Fee for Service	2			Fee for Service	0	
			PPO	0			PPO	0	
			POS	8			POS	0	
			Managed Care	0			Managed Care	8	
			HMO	0			HMO	0	
			Private non-HMO	2			Private non-HMO	1	
			Self Pay	1			Self Pay	1	
			Local Charity	1			Local Charity	1	
		Chicago	Medicaid	0		Died	LA	Medicaid	0
			Medicare	3		Medicare	0		
			Fee for Service	2		Fee for Service	1		
			PPO	1		PPO	0		
			POS	14		POS	1		
			Managed Care	0		Managed Care	0		
			HMO	0		HMO	0		
			Private non-HMO	1		Private non-HMO	0		
			Self Pay	7		Self Pay	0		
			Local Charity	12		Local Charity	1		
		NY	Medicaid	15		Chicago	Medicaid	0	
			Medicare	3			Medicare	0	
			Fee for Service	1			Fee for Service	0	
			PPO	4			PPO	0	
			POS	79			POS	0	
			Managed Care	0			Managed Care	0	
			HMO	0			HMO	0	
			Private non-HMO	1			Private non-HMO	0	
			Self Pay	1			Self Pay	0	
			Local Charity	1			Local Charity	0	
		Seattle	Medicaid	0		NY	Medicaid	1	
			Medicare	8			Medicare	0	
			Fee for Service	3			Fee for Service	0	
			PPO	2			PPO	0	
			POS	33			POS	9	
			Managed Care	0			Managed Care	0	
			HMO	0			HMO	0	
			Private non-HMO	5			Private non-HMO	0	
			Self Pay	2			Self Pay	1	
			Local Charity	19			Local Charity	0	
		Miami	Medicaid	1		Seattle	Medicaid	0	
			Medicare	0			Medicare	1	
			Fee for Service	2			Fee for Service	0	
			PPO	1			PPO	0	
			POS	5			POS	3	
			Managed Care	0			Managed Care	0	
			HMO	0			HMO	0	
			Private non-HMO	0			Private non-HMO	2	
			Self Pay	1			Self Pay	0	
			Local Charity	0			Local Charity	0	
		Nashville	Medicaid	0		Miami	Medicaid	1	
			Medicare	0			Medicare	0	
			Fee for Service	0			Fee for Service	0	
			PPO	0			PPO	0	
			POS	0			POS	0	
			Managed Care	23			Managed Care	0	
			HMO	0			HMO	0	
			Private non-HMO	0			Private non-HMO	0	
			Self Pay	0			Self Pay	0	

			Local Charity	0		Medicare	8
Nashville			Medicaid	0		Fee for Service	12
			Medicare	0		PPO	14
			Fee for Service	0		POS	51
			PPO	0		Managed Care	0
			POS	0		HMO	0
			Managed Care	0		Private non-HMO	0
			HMO	2		Self Pay	0
			Private non-HMO	0		Local Charity	32
			Self Pay	0	Nashville	Medicaid	0
			Local Charity	0		Medicare	2
Phoenix			Medicaid	0		Fee for Service	1
			Medicare	0		PPO	0
			Fee for Service	0		POS	0
			PPO	0		Managed Care	69
			POS	0		HMO	0
			Managed Care	0		Private non-HMO	0
			HMO	0		Self Pay	0
			Private non-HMO	0		Local Charity	0
			Self Pay	0	Phoenix	Medicaid	2
			Local Charity	0		Medicare	4
Male	Alive	LA	Medicaid	38		Fee for Service	3
			Medicare	19		PPO	7
			Fee for Service	32		POS	0
			PPO	11		Managed Care	24
			POS	57		HMO	0
			Managed Care	0		Private non-HMO	8
			HMO	0		Self Pay	4
			Private non-HMO	3		Local Charity	2
			Self Pay	6	Died	Medicaid	5
			Local Charity	8		Medicare	2
Chicago			Medicaid	21		Fee for Service	1
			Medicare	12		PPO	0
			Fee for Service	14		POS	2
			PPO	22		Managed Care	0
			POS	47		HMO	0
			Managed Care	0		Private non-HMO	0
			HMO	0		Self Pay	0
			Private non-HMO	17		Local Charity	0
			Self Pay	11	Chicago	Medicaid	0
			Local Charity	66		Medicare	0
NY			Medicaid	39		Fee for Service	0
			Medicare	1		PPO	3
			Fee for Service	0		POS	3
			PPO	24		Managed Care	0
			POS	4		HMO	0
			Managed Care	0		Private non-HMO	1
			HMO	0		Self Pay	1
			Private non-HMO	1		Local Charity	4
			Self Pay	1	NY	Medicaid	6
			Local Charity	1		Medicare	1
Seattle			Medicaid	5		Fee for Service	0
			Medicare	8		PPO	1
			Fee for Service	4		POS	25
			PPO	5		Managed Care	0
			POS	33		HMO	0
			Managed Care	0		Private non-HMO	0
			HMO	0		Self Pay	0
			Private non-HMO	13		Local Charity	0
			Self Pay	3	Seattle	Medicaid	2
			Local Charity	22		Medicare	0
Miami			Medicaid	6		Fee for Service	0
						PPO	0

	POS	4		Fee for Service	0
	Managed Care	0		PPO	0
	HMO	0		POS	0
	Private non-HMO	0		Managed Care	0
	Self Pay	0		HMO	8
	Local Charity	5		Private non-HMO	7
Miami	Medicaid	0		Self Pay	0
	Medicare	0		Local Charity	0
	Fee for Service	1	Phoenix	Medicaid	1
	PPO	4		Medicare	0
	POS	2		Fee for Service	1
	Managed Care	0		PPO	0
	HMO	0		POS	0
	Private non-HMO	0		Managed Care	5
	Self Pay	3		HMO	0
	Local Charity	5		Private non-HMO	2
Nashville	Medicaid	0		Self Pay	0
	Medicare	0		Local Charity	0

When a linear analysis is conducted all of the categorical attributes with three or more response categories are reduced to a set of one-fewer binary dummy-coded indicator variables than there are of response options for the categorical scale.^{3,30-33} Presently, city would be reduced to 6 binary indicators, and insurance to 9. To predict status or gender using the indicator variables instead of original city and insurance, implies a design matrix with 2 [gender or status] \times $(2 \times 2 \times 2 \times 2 \times 2 \times 2)$ [city] \times $(2 \times 2 \times 2)$ [insurance] = $2 \times 64 \times 512$, or a total of 65,536 cells ($2^1 \times 2^6 \times 2^9 = 2^{16}$). Not only would the cross-classification table be *long* (each cell constitutes a row in the table), it would be *wide*. To display this table would require 18 columns in Table 7.13, instead of the 5 columns used presently. If the observations were uniformly distributed in the cells, then on average 0.024 observations would fall in every cell of the table. Equivalently, there would be one observation for every 41.7 cells: a sparsely-populated table. This analytic nightmare happens with only three categorical attributes included in the design. Perusal of any empirical journal reporting linear models for dichotomous class variables (i.e., dependent measures) will likely reveal that many studies include numerous such attributes (i.e., independent variables) in their design.

An inherent, immitigable issue with of *all* suboptimal methods is their inability to explicitly (by formulation) maximize the (weighted) classification accuracy (*Overall PAC* or *ESS*) that is obtained by the model for a sample. Any model that explicitly returns maximum accuracy for a sample is called an *optimal* or a *maximum-accuracy* model, and any model that fails to explicitly yield maximum *ESS*—but that is specifically engineered to seek maximum-accuracy solutions for a sample—is known as a *heuristic* maximum-accuracy model.²¹ Inherent, immitigable issues for *all linear methods* are small cell *N*, empirical and structural zeros (cells having *N* = 0) in the design matrix, and non-normality of the data and residuals.³⁴

Predicting Patient Gender: Enumerated CTA (Chapter 11) was used to predict patient status (the class variable) with gender, city, and insurance treated as being categorical attributes, using the following CTA software³⁵ syntax:

```
VARS gender status city insure;
CLASS gender;
ATTR status city insure;
CATEGORICAL status city insure;
MC ITER 5000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
GO;
```

Using three nodes, a 4-strata partition of the sample was identified by CTA (Figure 7.1), yielding moderate accuracy (*ESS* = 32.3) and weak predictive value (*ESP* = 23.1). As is seen, CTA models initiate with a *root node*, from which two or more *branches* emanate and lead to other *nodes*: branches indicate pathways through the tree, and all branches terminate in model *endpoints*. CTA models are highly intuitive: model “coefficients” are cutpoints or category descriptions expressed in their natural measurement units, and sample stratification unfolds in a flow process which is easily visualized across model attributes. Numbers (ordered attributes) or words (categorical attributes) adjacent to branches give the value of the

cutpoint (category) for the node. Numbers under nodes give the *experimentwise* or *generalized p* for the node (the latter is typically reported). The number of observations classified into each endpoint is indicated beneath the endpoint, and the percentage of targeted (presently, female) observations is given inside the rectangle representing the endpoint. Using CTA models to classify individual observations is also straightforward. Imagine a hypothetical person on managed care living in LA. Starting at the root node, since the person lives in LA the left branch is appropriate. At the second node the right branch is appropriate because the person has managed care. Finally, at the third node the left branch is appropriate since the person is from LA. The person is thus classified into the corresponding model end-point: as seen, 11.7% of the observations classified into this model endpoint were females. Note that endpoints represent sample strata identified by the CTA model. The probability of being female for this endpoint is $p_{female} \leq 0.117$: if instead the person had lived in Chicago, then the right-hand endpoint would be appropriate, with $p_{female} \leq 0.240$.

Figure 7.1: CTA Model Predicting Gender

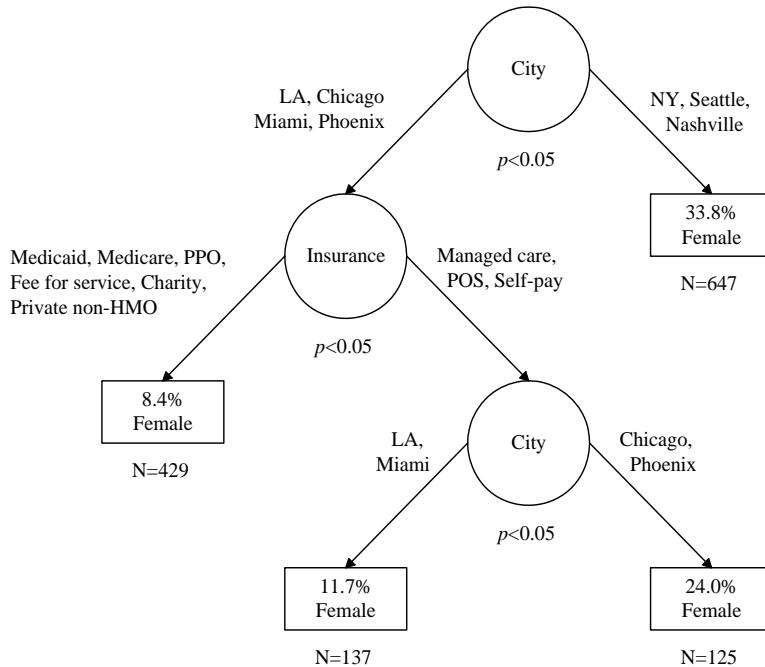


Table 7.14 presents the confusion table for the overall model.

Table 7.14: Confusion Table for CTA Model Predicting Gender

		Predicted Gender		
Actual Gender		Male	Female	
		Male	Female	
Male	Male	514	523	49.6%
	Female	52	249	82.7%
		90.8%	32.2%	

The CTA model accurately classified 82.7% of the women in the sample (and thus is a reasonably good theoretical representation of women), and was accurate 90.8% of the times it predicted that a given observation was male. Males presented with a more complex profile than females: sensitivity achieved for men by the CTA model was virtually the same as is expected by chance (50%) in this application.

There are similarities and differences between UniODA and CTA findings in this example. When predicting patient gender UniODA found no effect for status: likewise, status didn't enter the CTA model.

UniODA identified significant effects for city (based on the optimal range-test: $ESS = 5.8$; $ESP = 4.3$) and for insurance (based on the experimentwise optimal range-test: $ESS = 17.9$; $ESP = 12.6$). Both of these attributes entered the CTA model (this pattern does not always occur). City emerged as the most influential attribute in the model, involved in the classification decisions for all (100%) of the observations in the sample, and insurance was involved in the classification of $N = 1,568 - 647 = 921$ observations (see Figure 7.1)—corresponding to 58.7% of the total sample.

The CTA-model-based order of cities with respect to percent of females in the model endpoints is: (LA, Miami) < (Chicago, Phoenix) < (NY, Seattle, Nashville). This is identical to the UniODA model in the second step of the range test (Table 7.7), but it *isn't* the final model that was identified by UniODA for city (Table 7.8): the additional reduction that occurred in UniODA would serve to reduce the overall ESS of the CTA model. This same argument suggests that the UniODA model identified in an optimal range test that has the highest ESS should be selected (see Chapter 12).

Considering the insurance groupings parameterizing CTA model branches (Figure 7.1), UniODA and CTA model left-hand branches shared Medicare, and right-hand branches shared managed care: the other insurance types were all on the *opposite* branch, and the CTA model did not include HMO in the roster of insurance categories (HMO was not an insurance category for the cities in the left-hand branch of the CTA model emanating from the root node). This illustrates the difference between UniODA and CTA: the former finds the optimal (maximum ESS) solution for the sample considering one attribute at a time in isolation of all other attributes, while the latter finds the optimal (maximum ESS) solution for the sample considering all attributes simultaneously in conjunction with one another.

Table 7.15 presents the staging table³⁵ for the CTA model (see Chapter 2, The CTA Algorithm).

Table 7.15: Staging Table for CTA Model Results

<u>Stage</u>	<u>City</u>	<u>Insure</u>	<u>City</u>	<u>N</u>	<u>p_{female}</u>	<u>Odds_{female}</u>
1	LA, Chicago, Miami, Phoenix	Medicaid, Medicare, Fee for Service, PPO, Private non-HMO, Charitable group	---	429	0.084	1:11
2	LA, Chicago, Miami, Phoenix	POS, Managed care, Self-pay	LA, Miami	137	0.117	1:8
3	LA, Chicago, Miami, Phoenix	POS, Managed care, Self-pay	Chicago, Phoenix	125	0.240	1:3
4	NY, Seattle, Nashville	---	---	647	0.338	2:3

A staging table is an intuitive alternative representation of a CTA model that is useful for defining “propensity” scores (weights) for observations that are based on the findings of the CTA model. The rows of the staging table represent stages having increasing propensity with respect to the class variable (here, female gender): the stages are simply the model endpoints reorganized in increasing order of percent of class 1 (female) membership. Stage is thus an *ordinal index of propensity*, and p_{female} is a *continuous index of propensity*: increasing values on either index indicates increasing propensity. Compared to Stage 1, the p_{female} is 1.4-times greater in Stage 2; 2.9-times greater in Stage 3; and 4.0-times greater in Stage 4. For an enhanced perspective the reader may wish to construct a staging table representing the findings obtained using a parametric linear model in this application.

To use the table to stage a given observation, simply evaluate the fit between the observation’s data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor that is *exactly true* for the data of the observation undergoing staging. For example, consider the hypothetical person discussed earlier living in LA with managed care. Starting with Stage 1, city is appropriate, but insurance does not include managed care. Moving to Stage 2, city is appropriate (LA),

insurance is appropriate (managed care), and the second city column is appropriate (LA): the person is thus classified as Stage 2 along with 136 other people in the sample. The Stage 2 patient strata is 11.7% female: odds of being female in Stage 2 are approximately 1:8. If the numerator of the presented odds is one, the denominator of the presented odds is $(1 / p_{female}) - 1$. For example, for Stage 1 p_{female} is 0.0884, so denominator = $(1 / 0.0884) - 1 = 11.905 - 1$, or 10.91: in Table 7.15 the odds for Stage 1 are given as 1:11.

The CTA model achieved greater overall *ESS* and *ESP* than any of the UniODA models; obtained greater sensitivity in accurately classifying the actual women in the sample than UniODA models; and was only surpassed in ability to make accurate classifications of observations as being women by one UniODA model (Table 7.7). The CTA model segmented four sample strata: this level of discriminant gradation was only achieved by the UniODA model for predicting gender based on city. It should be noted that while *ESS* and *ESP* index the overall strength of the model, model *efficiency* (*ESS* / number of strata identified by the model) adjusts classification performance to reflect relative complexity—the opposite of parsimony (see Chapter 12). Mentioned earlier, it may be argued that the optimal model for discriminating gender based on city using UniODA was the initial model with two endpoints, for which *ESS* = 31.5 and *ESP* = 22.0 (Table 7.6). This is the strongest of the UniODA models in this analysis and the most parsimonious. Having two strata (endpoints) the efficiency statistics for this UniODA model are 15.8 (*ESS*) and 11.0 (*ESP*): these are 95% and 90% *greater* than were achieved using the CTA model, respectively.

Chapter 8

Explicitly Optimal Linear Multiattribute Models

Preceding awareness of the extent and prevalence of paradoxical confounding¹ of linear statistical models (Chapter 9), and before the CTA algorithm² was discovered, several laboratories studied *explicitly optimal linear models* for applications involving a binary class variable and two or more ordinal and/or binary attributes: a maximum-accuracy analogue of FLDA (Chapter 6) or LRA (Chapter 7) models.³⁻⁷ Linear multiattribute ODA models are called *MultiODA* models.⁸ Although UniODA models can be identified for giant samples, MultiODA models can be computationally intractable for tiny samples, even using the fastest computers. Two comparatively fast methods of solving MultiODA problems were developed in the ODA laboratory: MIP45 is a mixed integer formulation, and WARMACK is a special-purpose search algorithm.

MIP45 Mixed Integer Programming Formulation

Mixed-integer linear programming formulations for binary class variable MultiODA models require the estimation of the value of a parameter, M , that is commonly defined⁴ as “a prohibitively large number.” In application if the guesstimated M is too low then suboptimal solutions occur, and excessively large values of M decrease computational efficiency and may introduce numerical (round-off) error.³ Discussed below, the MIP45 goal programming formulation⁹ eliminates this problem by establishing a lower bound for M .

In a two-group linear MultiODA problem with p attributes and m observations, a set of m row vectors \mathbf{a}_i is given, the components of which are $p = n-1$ observed values and a dummy value of unity. Each observation i is a member of either class 0 or class 1. A weight vector \mathbf{x} is determined so that i is predicted to belong to class 0 when $\mathbf{a}_i \mathbf{x} < 0$, or to class 1 when $\mathbf{a}_i \mathbf{x} > 0$. Observation i is considered to be correctly classified if its predicted class membership is the same as its actual class membership, and misclassified otherwise. Solutions of interest yield maximum classification accuracy, that is, minimize the number of misclassified observations. This is achieved by determining \mathbf{x}^* that satisfy the maximum number of inequalities in the system:

$$\begin{aligned}\mathbf{a}_i \mathbf{x} &< 0 \text{ for observations in class 0,} \\ \mathbf{a}_i \mathbf{x} &> 0 \text{ for observations in class 1.} \end{aligned}\quad (1)$$

This problem may be formulated as a mixed-integer linear programming model. To accomplish this, the strict inequalities in (1) are replaced with $\mathbf{a}_i \mathbf{x} \leq -\varepsilon$ or $\mathbf{a}_i \mathbf{x} \geq \varepsilon$, where $\varepsilon \geq 0$. This is necessary due to the inability of simplex-based algorithms for mixed-integer programming to handle strict inequalities (mixed-integer techniques based upon interior-point algorithms¹⁰ may not suffer this limitation). Letting ε be strictly positive removes the ambiguity in the classification status of observations i for which $\mathbf{a}_i \mathbf{x} = 0$, but also introduces the possibility of a classification gap. It will be shown that there are conditions under which ambiguities can be removed for $\varepsilon = 0$. Consider the following model:

$$\text{MIP45: } z = \min \sum_{i=1}^m d_i \quad (2)$$

subject to

$$\sum_{j=1}^n a_{ij} (x_j^+ - x_j^-) - M_i d_i \leq -\varepsilon, i \in I_0 \quad (3)$$

$$\sum_{j=1}^n a_{ij} (x_j^+ - x_j^-) + M_i d_i \geq \varepsilon, i \in I_1 \quad (4)$$

$$\sum_{j=1}^n (x_j^+ + x_j^-) = 1 \quad (5)$$

$$x_j^+ - g_j \leq 0, j = 1, \dots, n \quad (6)$$

$$x_j^- + g_j \leq 1, j = 1, \dots, n \quad (7)$$

$$x_j^+, x_j^- \geq 0, j = 1, \dots, n \quad (8)$$

$$g_j \in \{0, 1\}, j = 1, \dots, n \quad (9)$$

$$d_i \in \{0, 1\}, i = 1, \dots, m \quad (10)$$

where

a_{ij} is the j th component of observation \mathbf{a}_i

I_0 is the set of observations belonging to class 0

I_1 is the set of observations belonging to class 1

$$M_i = \max_j |a_{ij}| + \varepsilon \quad (11)$$

z is the number of misclassified observations.

The weight vector \mathbf{x} is obtained by

$$x_j = x_j^+ - x_j^-, j = 1, \dots, n. \quad (12)$$

Since constraints (6) and (7) ensure that not more than one of the x_j^+ and x_j^- are positive for any j , one can think of these values as the "positive" and "negative" parts of x_j , respectively. Note that $g_j = 1$ if $x_j > 0$ and $g_j = 0$ when $x_j < 0$. Also note that the g_j , along with (6), (7), and (9), may be dropped when $\varepsilon > 0$.

Constraint (5) normalizes \mathbf{x} so that

$$\sum_{j=1}^n |x_j| = 1; \quad (13)$$

that is, the sum of the absolute values of the discriminant weights is constrained to equal one. This normalization prevents the trivial solution $\mathbf{x} = \mathbf{0}$ (when $\varepsilon > 0$), and allows establishment of a lower bound for the M_i . It is necessary for the M_i to be large enough to force compliance of the constraints (3) and (4). This is accomplished by (11). To see this, consider constraint (4). Since $\sum_j |x_j| = 1$, it is clear that

j

$$\mathbf{a} \cdot \mathbf{x} \geq - \max_j |a_{ij}| \quad (14)$$

and

$$\mathbf{a} \cdot \mathbf{x} + \max_j |a_{ij}| + \varepsilon \geq \varepsilon \quad (15)$$

Therefore, when $d_i = 1$,

$$\mathbf{a} \cdot \mathbf{x} + M_i d_i \geq \varepsilon. \quad (16)$$

It is because the normalization (5) requires that all optimal weight vectors \mathbf{x}^* lie on a 45° properly rotated hypercube centered at the origin, that this formulation is referred to as MIP45. It may be the case that more than one solution for \mathbf{d} may be optimal for a problem. This corresponds to the existence of multiple optimal dichotomies of predicted class membership. It is also generally true that a solution space for \mathbf{x} of positive volume exists for each dichotomy. The issue of selecting among optimal \mathbf{x}^* may be addressed by a number of methods, such as linear programming⁵ and *a priori* decision heuristics.¹¹

Resolving Classification Gaps and Ambiguities

In the above formulation, at least $n - 1$ of the $\mathbf{a} \cdot \mathbf{x}^*$ are at zero when $\varepsilon = 0$ is specified. From (1), it is seen that the criterion of strict separation of the classes should be met. An optimal value $z^* > 0$ in the solution of the following linear program guarantees that this separation is maintained.

$$\text{LP: } \max z = y$$

subject to

$$\sum_{j=1}^n a_{ij} (b_j^+ - b_j^-) + y \leq 0, \quad i \in I_0 \text{ and } \mathbf{a} \cdot \mathbf{x}^* \leq 0 \quad (17)$$

$$\sum_{j=1}^n a_{ij} (b_j^+ - b_j^-) - y \geq 0, \quad i \in I_1 \text{ and } \mathbf{a} \cdot \mathbf{x}^* \geq 0 \quad (18)$$

$$\sum_{j=1}^n (b_j^+ - b_j^-) = 1 \quad (19)$$

$$b_j^+, b_j^-, y \geq 0 \quad (20)$$

$$b_j = b_j^+ + b_j^- . \quad (21)$$

This LP may be executed for each optimal dichotomy. If $z^* > 0$ is obtained, \mathbf{b}^* is a new discriminant vector which optimizes criterion (1). Otherwise, ambiguity remains in the classification status of observations for which $\mathbf{a} \cdot \mathbf{b}^* = 0$: such observations should not be classified.

The advantage of establishing a lower bound for M is illustrated with an example involving discriminating between excellent versus less than excellent medical residents using information obtained during their application for residency training. Rating applicants for residency training is a difficult, time-intensive decision-making task, so a linear discriminant classifier that successfully predicts resident performance might be of great interest and utility to admissions committees. The sample was $m = 49$ resi-

dents enrolled in a three-year internal medicine residency program.¹¹ The clinical performance (class) variable was based on the mean rating on an explicit 10-point scale made by residents' supervisors: a mean rating of nine or greater on this scale indicated "excellent" (or better) clinical performance (class = 1, $m_1 = 27$), and a mean rating of less than nine indicated less than excellent clinical performance (class = 0; $m_0 = 22$). The $n - 1 = 3$ application information variables (attributes) included medical board scores, faculty evaluations (a composite measure reflecting ratings of letters of recommendation and medical school grading system), and academic distinction (a composite measure reflecting honors attained in medical school and medical school status).

Computer resources required to solve this problem by the MIP45 versus the Stam and Joachimsthaler⁷ formulation was compared (other prior formulations were slower). For MIP45, ε was set at 0. For Stam and Joachimsthaler values of 1, 10, 100, and 1000 were used for M , and ε was set at one.¹³ All formulations were solved on an IBM 3090/600 computer running SAS/OR.¹⁴ As seen in Table 8.1, except when $M = 1$, Stam and Joachimsthaler required more computational effort (CPU time, pivots, and integer branches) than MIP45. Using $M = 1$, $\varepsilon = 1$ in Stam and Joachimsthaler resulted in a useless solution, and using $M = 10$ or 100 resulted in suboptimal solutions of (3). Since a decision-maker using $M = 10$ or $M = 100$ would have no direct evidence that these solutions were suboptimal, it would be unclear whether the solution attained by any unbounded formulation using $M = 1000$ was optimal. In contrast, since the value of z^* attained in LP was positive, a decision-maker using MIP45 to solve this problem would be certain that the solution was unambiguously optimal: obviously a clear advantage.

TABLE 8.1: Computational Resources Needed by MIP45 versus Stam and Joachimsthaler to Solve a MultiODA Problem with 49 Observations and 3 Attributes, via SAS/OR run on an IBM 3090/600 Computer

Formulation	M	ε	Objective	CPU	Integer	
			Value	Seconds	Branches	Pivots
Stam	1	1	29	1.1	0	31
Stam	10	1	17	131.8	8,629	36,607
Stam	100	1	15	276.7	19,755	89,564
Stam	1000	1	14	268.4	14,549	57,351
MIP45	LB	0	14	48.0	2,896	15,333

Note: For MIP45 the M_i were set at their lower bounds (LB). For solutions yielding the optimal value of 14 misclassifications the model coefficients for board scores and faculty evaluation were positive, and the coefficient for academic distinction was negative. For MIP45, $z^* = .00439$.

Weighted Classification

Rather than weighting each observation equally, consider weighting each case in (2) by a positive scalar c_i . This is significant for two reasons. First, the c_i may represent the cost of misclassifying observation i . In this case an optimal solution would minimize the cost of misclassification (or, equivalently, maximize the return of correct classification) for the sample. Second, the c_i may represent factors which balance the number of class 0 and class 1 observations when these are not equal. In this case an optimal solution would maximize the number of correct classifications weighted by population membership in each class. An example would be $c_i = 1 / m_0$ for observations in class 0, and $c_i = 1 / m_1$ for observations in class 1, where m_0 and m_1 are the number of observations in categories 0 and 1, respectively. This latter weighting scheme is particularly useful in badly imbalanced applications for which $m_0 \gg m_1$, or vice versa: use of such "priors weights" forces the model to classify observations from both classes accurately, and inhibits the identification of degenerate models which classify all observations into a single class category.

Adding Nonlinear Terms as Attributes

Here the notion of maximum pattern classification accuracy achieved by separating hyperplanes is generalized to sets of nonlinear separating surfaces. For example, consider quadratic surfaces in p -measurement space of the form:

$$\sum_j a_{ij} x_j + \sum_{k \leq p} \sum_{l \leq k} a_{ik} a_{il} x_k x_l + a_{in} x_n \quad (22)$$

for all i . The MultiODA solution can be attained by augmenting the a_j and x in the MIP45 model by the interaction terms in (22). This solution produces a weight vector \mathbf{x} which yields the minimum number of misclassifications achievable by a quadratic separating surface. This process may be applied to any nonlinear discriminant function which is linear in the parameters of the measurement space.

Optimal Attribute Subset Selection

In the foregoing derivation all p attributes are included in the MultiODA model. However, one may wish to select a subset of $k < p$ attributes for the application of the model. For example, imagine an application involving 50 observations and 10 attributes. In an effort to avoid over-fitting and identify a model that may generalize if used to classify independent random samples, one may wish to maintain a minimum observation-to-attribute ratio of 10-to-1 (a rule-of-thumb for some parametric methods such as principal components analysis and LRA¹⁵): in this example a maximum of five of the ten potential attributes may be used. Of all of the possible 5-attribute models in this application, which model yields maximum accuracy? Optimal attribute subset selection methodology is incorporated in the MIP45 model by defining n zero-one variables q_j and including the following constraints:

$$\underline{x}_j - q_j \leq 0, j = 1, \dots, n, \quad (23)$$

$$g_j + q_j \leq 1, j = 1, \dots, n, \quad (24)$$

and

$$\sum_{j=1}^n g_j + \sum_{j=1}^n q_j = k. \quad (25)$$

In an optimal solution to such a MultiODA model, measurement j is selected for inclusion only if $g_j + q_j = 1$. The number of misclassifications obtained is the fewest achievable in any k -dimensional subspace of the original p -dimensional measurement space.

Aggregation of Duplicate Observations

If duplicate observations occur in the data set (i.e., two or more observations have the same value for every attribute measurement), the following procedure may be used to aggregate the duplicate observations into a single observation, reducing the size of the overall problem. The resulting problem is equivalent to the original one, with m' observations, and objective value $z + v$.

1. $m' := m : s_0 = 0 : s_1 = 0 : v := 0$
2. **for each** $i = 1, \dots, m'$
3. **for each** $j < i$
4. **if** $a_i = a_j$ **then**
5. **if** $i \in I_0$ **then** $s_0 := s_0 + c_i$ **else** $s_1 := s_1 + c_i$

```

6.      remove observation  $i$  from list :  $m' := m' - 1$ 
7.      end if
8.      next  $j, i$ 
9.      for each  $i = 1, \dots, m'$ 
10.     if  $s_0 > s_1$  then
11.        $w_j := s_0 - s_1 : v := v + s_1$ 
12.     else if  $s_1 > s_0$  then
13.        $w_j := s_1 - s_0 : v := v + s_0$ 
14.     else
15.        $v := v + s_0$  : remove observation  $i$  from list :  $m' := m' - 1$ 
16.     end if
17.   next  $i$ 

```

This procedure is particularly useful when α_j is a zero-one vector (i.e., every attribute is binary; see also Asparoukhov and Stam¹⁶). Here all the patterns lie on the vertices of the p -dimensional unit hypercube. If more than one pattern lies on some vertex, then by using the above procedure we may obtain a weighted MIP45 model equivalent to the original model, but with fewer constraints. If the number of original patterns m is large relative to the number of attributes p , a significant reduction in the size of the model may be obtained. For example, regardless of the value of m , if $p=8$ then there are no more than $2^8 = 256$ constraints of type (6) in the model. Since the number of constraints is independent of m , extremely large problems may be solved with this procedure, provided p is moderately small.

To illustrate the potential solution efficiency achieved by using this special purpose algorithm for problems involving entirely binary data, 30 MC experiments were conducted. Because each experiment involved five binary attributes the total possible number of different profiles was $2^5 = 32$. Values on every attribute were determined separately for each observation on the basis of a random uniform number between 0 and 1: numbers < 0.5 were assigned the value of 0, and numbers ≥ 0.5 were assigned the value of 1. Five balanced ($m_0 = m_1$) experiments were run for each sample size of 50, 100, 10^3 , 10^4 , 10^5 , and 10^6 total observations. All formulations were solved on an IBM 3090/300 computer running SAS/OR. As seen in Table 8.2, as the number of observations increased: (a) the number of distinct profiles increased toward its theoretical upper bound (the theoretical upper bound was achieved in all of the problems involving 10^6 observations, and in four of the five problems involving 10^5 observations); (b) the misclassification rate increased towards its theoretical upper bound (i.e., for a balanced design with an even number of observations, the theoretical upper bound for the number of misclassifications is one less than one-half of the total number of observations); and (c) the mean number of CPU seconds required to solve the problem was approximately twenty seconds for problems with 10^3 or more total observations.

MIP45 solves two problems common to prior goal programming formulations of two-group MultiODA: M is automatically set at its lower bound, and it is possible to determine if classification gaps or ambiguities exist. Collateral benefits of MIP45 include its greater computational efficiency and solution speed relative to prior formulations, particularly for applications involving binary attributes. The present research contrasted the computational characteristics of the MIP45 formulation of the MultiODA problem to the formulation of Joachimsthaler and Stam (Table 1). Other mixed-integer programming formulations should be considered: for example, Rubin¹⁷ developed a decomposition technique to solve the MultiODA problem; Silva and Stam¹⁸ developed a partitioning method for MultiODA that was reported to compare favorably with MIP45; Pfetsch¹⁹ developed a technique to optimize irreducible inconsistent subsystems (IIS) of linear inequalities in order to determine a maximum feasible subsystem of these inequalities; and Bremner and Chen²⁰ developed a mixed integer programming formulation for the halfspace depth problem that uses IIS cuts in a branch-and-cut algorithm.

TABLE 8.2: Results of Monte Carlo (MC) Experiments for Binary Data: Five Random Attributes

<u>Number of Observations</u>	<u>Number of Profiles</u>	<u>Number (%) of Misclassifications</u>	<u>CPU Seconds</u>
50	19	14 (28%)	4.5
50	23	13 (26%)	5.5
50	20	16 (32%)	8.4
50	23	14 (28%)	12.5
50	21	12 (24%)	1.7
100	30	27 (27%)	14.2
100	26	34 (34%)	9.0
100	25	43 (43%)	10.5
100	24	37 (37%)	7.5
100	20	33 (33%)	3.0
1000	32	432 (43%)	17.8
1000	30	445 (44%)	25.1
1000	31	449 (45%)	16.4
1000	31	460 (46%)	24.3
1000	31	454 (45%)	19.2
10000	29	4870 (49%)	12.3
10000	31	4838 (48%)	23.8
10000	31	4842 (48%)	24.9
10000	29	4828 (48%)	11.9
10000	31	4839 (48%)	9.2
100000	32	49545 (50%)	14.5
100000	32	49532 (50%)	21.6
100000	31	49526 (50%)	6.3
100000	32	49475 (49%)	25.2
100000	32	49376 (49%)	16.8
1000000	32	498331 (50%)	24.3
1000000	32	498759 (50%)	17.2
1000000	32	498450 (50%)	32.5
1000000	32	497861 (50%)	4.5
1000000	32	498837 (50%)	16.8

WARMACK Search Algorithm

A second approach to obtaining fast solutions to MultiODA problems involves the so-called WARMACK adaptation of a fast search algorithm initially developed by Warmack and Gonzalez (hence the origin of the name used to refer to the method).^{21,22} In MC-based research conducted in the ODA laboratory the WARMACK algorithm obtained an order of magnitude or greater reduction in computation time versus MIP45: for example, problems with two attributes and 700 observations were solved in less than one CPU minute on an IBM 3090/600 computer. This is also true for problems involving three attributes and 200

observations, or four attributes and 100 observations: the number of attributes exerts greater influence on computation time than number of observations or relative discriminability of the data.²¹

Multicategorical Class Variables

Two procedures may be used to solve multicategorical problems involving more than two class categories using MIP45 (see also Loucopoulos and Pavur²³, Adem and Gochety²⁴) or WARMACK. If there are $k > 2$ class categories, the first method is to determine the ODA solution obtained with $k - 1$ separating surfaces in parallel with each other. From a computational standpoint, this is equivalent to adding an extra attribute for each additional class. The second method involves the determination of k different discriminant functions: an observation is assigned to the class for which the maximum value is obtained over these functions. Given p original attributes, this is equivalent to a MultiODA problem with p times k attributes.

MultiODA Research in the ODA Laboratory

Limited substantive application of MultiODA conducted by the ODA laboratory occurred in the field of medicine. In the first study three attributes (age and two measures of heart rate variability) were treated as attributes and used to predict sudden cardiac death (binary class variable) for a sample of 45 patients (exact p is estimated by Fisher's randomization procedure).⁸ The MultiODA model outperformed an LRA model on overall PAC, sensitivity, specificity, and positive and negative predictive value: the worst result by MultiODA in *LOO validity analysis* exceeded the *best* performance result by LRA in *training analysis*. A second project reanalyzed data obtained from previously published studies originally analyzed using linear parametric methods: in study one *Overall PAC* obtained in training by MultiODA was 73.5% versus 69.9% by FLDA; in study two an LRA model employed three attributes to obtain 76.1% training and 79.4% hold-out *Overall PAC*, and identical performance was achieved using a two-attribute MultiODA model; in study three the *Overall PAC* obtained in training by MultiODA was 87.5% versus 82.5% by FLDA; and in study four MultiODA identified a two-attribute model that achieved 93.3% *Overall PAC*, while LRA was unable to identify a statistically significant effect.²⁵ Interesting technical research involving MultiODA that the ODA laboratory hasn't yet explored is optimal receiver operator characteristic²⁶⁻²⁹ models incorporating ideas from fuzzy set theory.^{30,31} Interesting theoretical research involving MultiODA not yet explored is the use of two- or three-attribute MultiODA models, rather than individual attributes, within CTA model nodes.

Special-Purpose MultiODA Models

The flexibility of ODA methodology lends itself to explicitly ideal special-purpose classification applications for which no alternative or corresponding parametric statistical procedures can compare. The number of different ODA models that can be created is limitless due to the inherently infinite number of possible unique classification applications.³² For example, below are some specialized ODA models that may be of great utility across a variety of applications.

Boolean ODA

The ODA objective function of minimum classification error may be applied to classification problems with purely logical attributes.³³ In this case the decision rule involved in the assignment of an observation to a class category is a Boolean function of logical attributes measured for that observation: we wish to find a Boolean function with at most t terms that minimizes the number of classification errors. Alternatively, we may look for a function with at most m misclassifications that minimizes the number of logical terms. Either way, these problems can be formulated as integer programs, or solved by exhaustive enumeration.

For example, consider an application in which two emergency medicine physicians independently diagnosed if 51 patients with hip trauma had a bony abnormality. Each physician rated the patients as normal or abnormal based on the PPP test (a measure of sound conduction), and as normal or abnormal based on visual inspection. The presence versus absence of bony abnormality (class variable) was determined radiographically. Boolean ODA identified a single optimal model that yielded 96% *Overall PAC* (ver-

sus 75% via logistic regression): if *either physician rates either attribute as abnormal*, then classify the observation as abnormal; otherwise classify the observation as normal.

Exact ODA

In some problems observations are available for which class membership is unknown. Often, exactly o of these observations are to be acted upon in some manner. The initial phase of the Exact ODA approach to this problem involves the partitioning of the observations into two sets: the decision set—which consists of observations with unknown class membership, and the evaluation set—which consists of observations with known class membership. To illustrate this procedure, consider the problem of selecting j job applicants from a pool of applicants. The attributes may reflect measures of previous employment experience and of skills required to perform the job task. The evaluation set is comprised of previously hired individuals who have been measured on these attributes. Each individual in the evaluation set is weighted by a performance index: here, by a measure of job performance. The decision set is comprised of the pool of job applicants who have been measured on the attributes, j of whom are to be selected for employment. Exact ODA finds a solution that maximizes the weighted number of inequalities in the evaluation set, such that exactly j inequalities in the decision set are satisfied. As another example, consider the problem of selecting prisoners to be released under a court mandate which requires that exactly p must be released, due to overcrowding. In this example the decision set is the current population of prisoners, and the evaluation set are those prisoners who previously have been released. The performance index, which is to be minimized, is a measure of mayhem produced by the previously released prisoners. Additional obvious applications of Exact ODA lie in the areas of market research (direct mail) and investment.

Tau ODA

Another fruitful area of investigation relates to the use of MultiODA in the analysis of data that have been sorted into ordered categories. The motivation underlying this procedure involves maximizing goodness-of-fit between the actual and predicted category assignments. Widely used for comparison of two ranked sequences Kendall's tau is a similarity index computed between two variables (e.g., a class variable and an attribute), that is proportional to the number of satisfied inequalities between paired observations.³⁴ Tau ODA is a MultiODA procedure that *explicitly maximizes the value of Kendall's tau* on the basis of a linear model of the attributes. For example, consider the problem of ranking residency applicants for a program in General Internal Medicine. For each of 41 applicants, three attributes were measured: board scores, average scores obtained from ratings of letters of recommendation, and overall performance on four desirability indices that were scored by a committee. The consensus ranking of the applicants was obtained, and then Tau ODA was used to identify the optimal prediction model. An optimal Kendall's tau of 0.641 was obtained. The optimal model was then applied to the set of applicants for 1989, from which a tau of 0.456 was obtained.

Another promising application of Tau ODA involves a metric-free approach to multiple linear regression: it differs from classical regression in that the criterion involved is the preservation of the rank order of the observations on the dependent measure, rather than the sum of squared errors between the values on the dependent measure and the predicted values.³⁵ This is accomplished using a linear model that maximizes Kendall's tau. Because in this case the rank order of the predicted values (rather than the magnitude of the deviations) is of primary interest, this approach should theoretically be less sensitive to the presence of outliers in the training data than would be the case in classical regression. The first step in the analysis involves sorting the observations by their values on the dependent measure. Next, a system of homogeneous strict linear inequalities between all pairs of observations is established (pairs of observations with tied values on the dependent measure have their associated inequalities dropped). Tau ODA is then applied to this inequality system in order to find the optimal predicted rank sequence. The resulting solution is then adjusted as follows. A ray in the interior of the optimal cone is found by solving a linear program which maximizes the minimum distance to the facets forming the boundary of the cone (this step guarantees that the inequalities are satisfied strictly). A univariate L1 (absolute deviation) or L2

(squared deviation) regression is then performed, to a point on the ray that optimizes either the L1 or L2 error criterion. The solution of this univariate regression problem is the optimal model.

The ODA laboratory investigated the properties of this TauODA regression procedure using MC experiments. In all, 400 experiments were performed for each of seven models: in each experiment there were three attributes and 30 observations. Both training and validity data sets were generated, half of which were correlated standard normal data, and the other half normal data with outliers (obtained by multiplying the value of the dependent variable by 10 for 25% of the observations). Each data set was analyzed with classical regression, Tau ODA regression, and L1- and L2-weighted Tau ODA regression. For each Tau ODA regression, both the L1 and L2 adjustments were performed. The problems were solved using the WARMACK algorithm run on a 50 MHz 486 microcomputer. For all problems, the values of R^2 , Kendall's τ , and L1- and L2-weighted τ were recorded. The most striking result was found in the analysis of 3-way interactions with R^2 as the dependent variable. When the seven models were trained without outliers they performed similarly, with a moderate drop-off in R^2 when outliers were present in the validity set. When the models were trained with outliers, however, the L1-adjusted Tau ODA models outperformed the others in validity sets both with and without outliers. For the case in which outliers were present in the training set and not in the validity set, L1-adjusted models outperformed the others by a huge margin. In addition, these models showed stable performance across all data sets trained with outliers. Thus, the Tau ODA regression technique with L1 adjustment is a promising alternative to classical multiple regression, especially in the case when outliers may exist in the dependent measure.

Template ODA

Another interesting MultiODA application involves the design of optimal templates. In this situation, an individual is given a list of questions, along with a set of possible responses for each question, one of which is to be selected as the individual's answer to the question (e.g., each question is answered by "filling in" a circle corresponding to a selected answer on a paper or digital recording sheet). The actual class membership status of each observation is known. The objective of this procedure is to produce a template, that is, a series of "holes" on an opaque sheet: by laying the template over an answer sheet and counting the number of filled-in circles, a discriminant score is produced for an individual. This score is then compared to the cutpoint obtained by the model in order to assign class membership to the individual: this assignment minimizes the number of classification errors for the training sample. The ODA laboratory formulated Template ODA as a pure integer problem for an application involving a 38-item questionnaire (each item answered as "true" or "false") completed by 107 employees of a corporation, 70 of whom were known to be desirable workers, and 37 of whom were known to be undesirable workers. For this problem, Template ODA identified a template which resulted in 74.8% *Overall PAC*, requiring 26 CPU minutes to solve on an IBM 3090/600 computer running SAS/OR.

Unit and Integer Coefficient Models

UniODA can be used to solve MultiODA problems in which the discriminant coefficients are constrained to take on a small set of values. For example in a problem involving a attributes, discriminant coefficients restricted to the values 0, 1, or -1, and an unconstrained threshold coefficient (i.e., cutpoint), all optimal solutions may be found by solving $(3^a / 2) - 1$ UniODA problems. More granular coefficients such as integer scales may also be used as model coefficients: for problems involving k possible coefficient values and p attributes, $k^p / 2$ UniODA analyses are solved. If k and p are relatively small then computation is not an issue due to the fast speed of UniODA. As an example of this method consider an application involving predicting the in-hospital mortality status of patients receiving cardiopulmonary resuscitation.³⁶ Unit-weighted MultiODA outperformed prior scoring schemes, and yielded greater accuracy than LRA in training and hold-out validity analysis using only half as many attributes.

Chapter 9

Identifying and Ameliorating Statistical Confounding

In some applications a researcher is aware of the presence of one or more confounding covariates that, if left unchecked, threaten the validity of conclusions based on statistical analysis. Unfortunately, the vast majority of the empirical literature utterly ignores the possibility that one or more confounding variables threaten the validity of statistical conclusions. The former situation is addressed first.

Confounding by Covariates

Viewed conceptually, partial UniODA is an optimal (maximum-accuracy) analogue of GLM methods such as partial correlation, analysis of covariance (ANCOVA), backward and stepwise multiple regression, and hierarchical linear model, all of which are used to remove *variance* that is attributable to a confounder (or covariate) prior to assessing the relationship between the class variable and attribute.^{1,2} In applications with multiple confounders, partial UniODA is conducted iteratively with each iteration removing surviving observations correctly classified by the current covariate for the current reduced sample. This structural decomposition method is analogous to principal components analysis³ except that the former maximizes *accuracy*, and the latter maximizes *variance*.⁴ In the present context, structural decomposition serves as an optimal analogue to hierarchical and “last-to-enter” multiple regression approaches.^{1,2}

Partial UniODA

Partial UniODA is a multi-step structural decomposition procedure that may be used to obtain a statistical model that maximizes the classification accuracy (normed versus chance) that is achieved for a sample by using an attribute to classify observations’ actual class categories after first eliminating (“controlling for”) the effect of one or more covariates.⁵ In the first step of partial UniODA all of the observations that are correctly classified by a UniODA model treating the (first) covariate as an attribute are dropped from the sample: the surviving observations in the reduced sample weren’t correctly predicted by the confounder. Every covariate that is to be controlled is treated in the same manner, each evaluated using the current reduced sample: order of entry does not influence the final result of this procedure. The final step of the partial UniODA procedure assesses the non-confounded relationship between the attribute and the class variable using the final reduced sample with all covariates controlled.

The following demonstration of the use of partial UniODA to obtain a non-confounded bivariate model uses data from a study of factors that influence the self-rated likelihood of a discharged Emergency Department (ED) patient recommending the ED to others.⁶ Research has reported that waiting time to see the physician—which in part is a function of case-mix and not subject to physician control—is consistently a moderate-to-strong predictor of patient satisfaction and recommendation ratings.⁶⁻⁹ In this application the research objective is to assess which if any (and how) different dimensions of putatively controllable physician behavior accurately predict if patients report they are likely to recommend the ED to others, independently of waiting time to see the physician.

The setting of the study was an 800-bed urban university-based level 1 Trauma center having an annual census of 48,000 patients.⁶ Patients were mailed a survey assessing satisfaction with care received in the ED after one week post-discharge. The survey elicited ratings of the likelihood of recommending the

ED to others, and of satisfaction with aspects of administration, nurse, physician, laboratory, and family and friend care. A 17% return rate achieved over a six-month period yielded 2,109 surveys with a patient's self-rating of the likelihood that they will recommend the ED to others. This rating was assessed using a five-point Likert-type scale on which scores of 3 (*fair*, $N = 239$) indicate *ambivalence*; and 4 (*good*, $N = 584$) reflect *likely to recommend*, and was treated as the class variable. Satisfaction ratings of aspects of care received from physicians (p1 = waiting time; p2 = courtesy; p3 = took patient's problem seriously; p4 = concern for comfort; p5 = explanation of test/treatment; p6 = explanation of illness/injury) were used as attributes. Satisfaction items were completed using five-point Likert-type scales: 1 = *very poor* satisfaction, 2 = *poor*, 3 = *fair*, 4 = *good* and 5 = *very good* satisfaction.

Confounded Association: For the total sample of $N = 823$ patients the first analysis obtains the raw bivariate relationships between the class variable ("recom"), and patient ratings of satisfaction with waiting time (p1—the confounding variable) and five aspects of physician behavior (p2 - p6). This analysis was accomplished using the following UniODA and MegaODA software syntax:

```

OPEN recom.dat;                                ATTR p1 to p6;
OUTPUT recom.out;                             MISSING all (-9);
VARS recom p1 to p6;                         MC ITER 10000;
CLASS recom;                                  GO;

```

As expected the UniODA model for p1 (waiting time, the confounding variable) was statistically significant ($p < 0.0001$) and yielded a moderate level of accuracy: $ESS = 30.8\%$. The UniODA model was: if wait time rating ≤ 3 (*fair*) predict recommendation = 3 (*ambivalent*), otherwise predict recommendation = 4 (*recommend ED*). Table 9.1 presents the confusion table.

Table 9.1: UniODA Analysis of Recommendation Rating as a Function of Waiting Time

		Predicted Recommendation	
		3	4
Actual	3	173	65
	4	241	334

Raw analysis also found that all five measures of physician behavior were statistically significant predictors of patient recommendation rating (p 's < 0.0001): all attributes generated the same UniODA model (direction and threshold values) that was obtained for waiting time. ESS values indicated moderate effects for explanation of test/treatment ($ESS = 28.1$) and explanation of illness/injury (31.3), and relatively weak effects for courtesy (20.3), took patient's problem seriously (24.5), and concern for patient's comfort (22.6). It should be noted that the effect strength of associations involving physician behaviors is lower than obtained for waiting time, except for the measure of explanation of illness/injury.

The confounder, waiting time, is associated with all five measures of physician behavior: this was assessed using the following UniODA and MegaODA software syntax:

```

OPEN recom.dat;                                ATTR p2 to p6;
OUTPUT recom.out;                             MISSING all (-9);
VARS recom p1 to p6;                         MC ITER 10000;
CLASS p1;                                   GO;

```

Models for all five ratings of physician behavior were statistically reliable: the collinear effect was relatively strong for ratings of courtesy ($p < 0.0001$, $ESS = 50.0$), and moderate for ratings of explanation of test/treatment ($p < 0.0001$, $ESS = 39.4$), took patient's problem seriously ($p < 0.0001$, $ESS = 38.8$), explanation of illness/injury ($p < 0.0001$, $ESS = 37.0$), and concern for comfort ($p < 0.001$, $ESS = 28.5$). The strong association of the confounding variable to patient recommendation rating, and also to the various rated

aspects of physician behavior, underscores the importance of controlling waiting time to assess the independent (partial) association of ratings of physician behavior and patient recommendations.

Non-Confounded Association: As seen in Table 9.1, when the waiting time model predicted a recommend rating of 3 a total of 241 observations were misclassified, and when a recommend rating of 4 was predicted a total of 65 observations were misclassified. This total of 306 “residual observations” are the portion of the original sample that remains after statistically “controlling for” or “partialing out” (i.e., eliminating) the effect of waiting time from the sample data. Using the reduced sample of 306 patients (recom2.dat) the second analysis identifies the non-confounded bivariate relationships between the class variable and patient ratings of satisfaction with five aspects of physician behavior. Analysis was accomplished via the following UniODA and MegaODA software syntax:

```
OPEN recom2.dat; ATTR p2 to p6;
OUTPUT recom.out; MISSING all (-9);
VARS recom p1 to p6; MC ITER 10000;
CLASS recom; GO;
```

Non-confounded analysis found no statistically reliable association between recommend rating and physician courtesy ($p < 0.46$) or concern for comfort ($p < 0.29$). However, ratings of if the physician took the patient’s problem seriously ($ESS = 16.7$), and explanation of test/treatment ($ESS = 24.3$), and of illness/injury ($ESS = 19.8$) were statistically significant ($p < 0.0001$), relatively weak predictors of patient recommendation rating, after statistically eliminating the effect of the confounding variable of waiting time from the sample data. These results also provide evidence supporting the incremental validity of physician behavior in taking patients’ problems seriously and in explaining tests, treatments, illnesses, and injuries, as predictors of patient satisfaction over and above the effects of waiting time.

Unconstrained Covariates

Discussed in Chapter 2, creating arbitrary definitions of attribute measurement scales or of class-variable categories can yield diminished effects and misleading conclusions and thus should be avoided.

Similarly, the common practice in legacy statistical covariate analysis of forcing covariates into the model first is ill-advised in the ODA paradigm: if forced entry is used then the resulting model should be compared against an exploratory model that doesn’t constrain the order-of-entry. Rather, in ODA—for structural decomposition and for CTA, a covariate is treated as an ordinary attribute that must compete with other eligible attributes for selection into the model based on accuracy assessed as ESS .

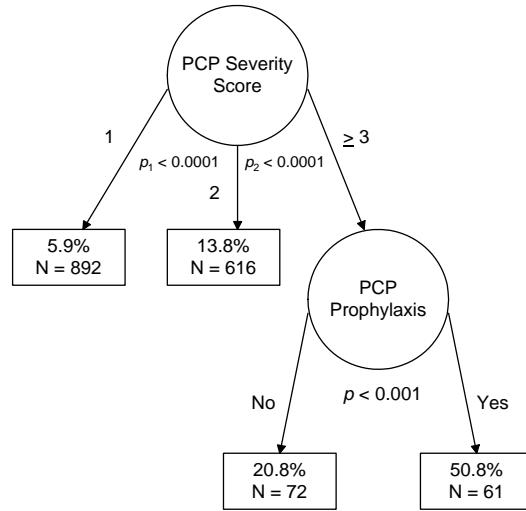
Analysis involving constrained versus unconstrained covariates is illustrated using an application involving predicting patient in-hospital mortality.¹⁰ A study of 1,641 patients hospitalized for *Pneumocystis cariini* pneumonia (PCP) used logistic regression analysis (LRA) to model in-hospital mortality: after forcing a measure of severity-of-illness into the model first, PCP prophylaxis was the only attribute significantly associated with lower hospital survival.¹¹ CTA models treated recommendation as a binary class variable, and p1 through p6 as possible attributes.

Emulating the LRA procedure, the first analysis forced p1 (waiting time) to serve as the root node using the following CTA software¹² syntax (the minimum endpoint denominator of $N = 25$ was used based on a statistical power analysis; see Chapter 2, Appendix C):

```
OPEN recom.dat; FORCENODE 1 p1;
OUTPUT recom.out; MC ITER 5000 CUTOFF .05 STOP 99.9;
VARS recom p1 to p6; PRUNE .05;
CLASS recom; MINDENOM 25;
ATTR p2 to p6; GO;
MISSING all (-9);
```

This analysis yielded the CTA model illustrated in Figure 9.1. This model returned relatively weak gain versus chance in the accurate prediction of mortality status: 97.9% of 1,457 living and 16.8% of 184 deceased patients were correctly classified: $ESS = 14.8$. Though the CTA model is relatively weak, the right-most endpoint indicates the combination of a PCP severity score ≥ 3 , and PCP prophylaxis, accurately predicted nearly 51% mortality for 61 patients. Discussed in Chapter 12, for an application in which it is important to identify particularly vulnerable strata, all possible different CTA models should be examined in hopes of discovering one or more of such fruitful branches.¹³

Figure 9.1: Predicting In-Hospital Mortality, Covariate Forced to Enter the CTA Model First



For exposition the second CTA analysis used only the same two attributes that were identified by the forced-entry model (Figure 9.1). After deleting the FORCENODE command and rerunning the program, the enumerated-optimal CTA model (see Chapter 11) in Figure 9.2 was obtained. As seen, the enumerated CTA model has robust endpoint denominators; correctly classified 67.9% of the 1,457 living and 61.4% of the 184 dead patients; and obtained moderate strength ($ESS = 29.4$).

Figure 9.2: Predicting In-Hospital Mortality, Covariate Entry Order Unconstrained

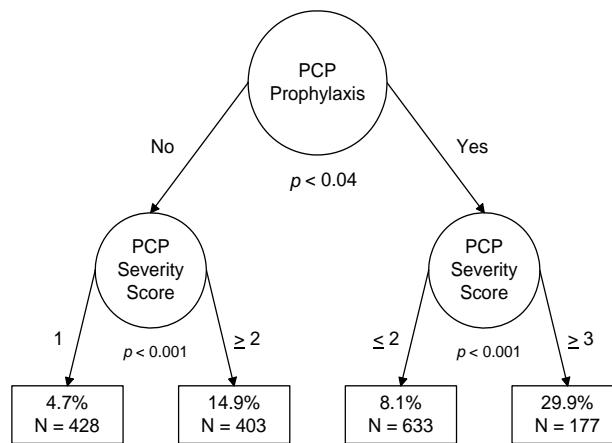


Table 9.2 provides the staging table for the enumerated CTA model, that is used for predicting in-hospital mortality from PCP. Table rows are model endpoints reorganized in increasing order of percent of class 1 (“dead”) membership. Stage is an *ordinal index* indicating increasing severity of illness, and p_{death} is

a continuous index of disease severity. The 1st and 4th strata reflect a 6.4-fold difference in likelihood of dying in-hospital: compared to Stage 1, p_{death} is approximately two times higher in Stage 2, three times higher in Stage 3, and six times higher in Stage 4.

Table 9.2: Staging Table for Predicting In-Hospital Mortality From PCP

Stage	PCP Prophylaxis	Severity			
		Score	N	p_{death}	Odds
1	No	1	428	0.047	1:20
2	Yes	≤ 2	633	0.081	1:11
3	No	≥ 2	403	0.149	1:6
4	Yes	≥ 3	177	0.299	3:7

Though identical attributes were used by the two CTA models and the original LRA, the attributes were arranged in different geometries in the different models. Specific imposition of attribute entry or of sequence order in CTA, or in any chained optimal analysis, should be performed on the basis of theory to directly address *a priori* hypotheses. The present example clearly indicates the need for caution regarding unchecked adherence to legacy methodological traditions that may actually impede progress achieved by emerging technologies.

Exploratory Methods

Partial UniODA is a conservative methodology unless the putative confounder is statistically benign (i.e., is unrelated to the attribute): because UniODA models explicitly maximize classification accuracy, associated reduction in sample size that occurs as correctly classified observations are eliminated from the sample rapidly reduces statistical power. The present example is continued, used to compare different analytic approaches to evaluating the role of a confounding variable in the maximum-accuracy modeling of a class variable using a set of attributes.¹⁴

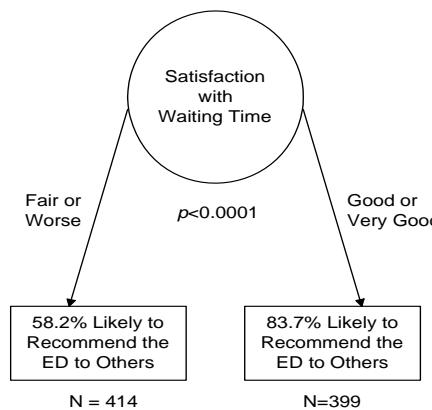
Simple Bivariate Associations of Attributes and Class Variable: Simple bivariate UniODA effects were consistent with respect to model *direction* (increasing satisfaction with patient-care and with waiting time predicts increasing likelihood of recommending the ED) and *optimal threshold* (critical rating value). For every simple effect the UniODA model was: if patient rating on the attribute or covariate is ≤ 3 (fair or worse) then predict patient likelihood to recommend the ED = 3 (ambivalent); otherwise, if patient rating on the attribute or covariate is > 3 (good or very good) then predict patient likelihood to recommend the ED = 4 (likely to recommend ED to others). Simple bivariate associations between patient self-rated likelihood of recommending the ED to others, and patient ratings of satisfaction with waiting time and physician patient-care behaviors, are summarized in the first column of Table 9.3. Discussed previously, moderate, statistically significant relationships emerged between likelihood of recommending the ED and patient satisfaction with waiting time (Figure 9.3, Table 9.1), and with physician explanations of test/treatment and illness/injury. Relatively weak, statistically significant relationships emerged between the likelihood of recommending the ED and the other physician patient-care behaviors.

As seen in Table 9.1, a total of $(173 + 65) = 238$ patients were ambivalent, and of these a total of 173 (72.7%) were correctly predicted by the UniODA model. And, a total of $(241 + 334) = 575$ patients were likely to recommend the ED to others, and of these a total of 334 (58.1%) were correctly predicted by the UniODA model. The percentage of observations of a given class category correctly predicted by the model is called the sensitivity of the model for the given class category. Table 9.3 reports the number of observations in each class category (N_3 , N_4), and the corresponding sensitivities for each class category. Table 9.1 also reports the UniODA model predictive value for each class category. Seen in Figure 9.3 the UniODA model predicted that 414 observations had likelihood = 3 [173 (41.8%) were correctly classified] and that 399 observations had likelihood=4 [334 (83.7%) were correctly classified]. Table 9.3 also gives the predictive values for each class category, after model sensitivities.

Table 9.3: Empirical Comparison of Alternative Strategies for Controlling a Single Confounding Variable in the ODA Paradigm

Attribute	Bivariate Association with Self-Rated Likelihood of Recommending ED to Others		Alternative Strategies of Controlling a Confounding Variable in ODA (See Text for Detailed Explanation and Discussion)		
			Treat the Confounding Variable as an Attribute in an HO-CTA Model	Treat the Confounding Variable as an Attribute in an EO-CTA Model	Treat the Confounding Variable as an Attribute in a GO-CTA Model
Waiting Time to See the Physician (confounder)	$p<0.0001$ <i>ESS=30.78, D=4.5</i> $N_3=238$ (72.7, 41.8) $N_4=575$ (58.1, 83.7)	Discard Observations Correctly Predicted by the Confounding Variable	$p's<0.0091$ <i>ESS=37.60, D=8.3</i> $N_3=238$ (68.9, 47.7) $N_4=575$ (68.7, 84.2)	$p's<0.0093$ <i>ESS=39.66, D=7.6</i> $N_3=236$ (72.9, 47.4) $N_4=575$ (66.8, 85.7)	$p's<0.0001$ <i>ESS=39.49, D=4.6</i> $N_3=236$ (56.4, 57.8) $N_4=575$ (83.1, 82.3)
Physician Courtesy	$p<0.0001$ <i>ESS=20.26, D=7.9</i> $N_3=236$ (26.3, 63.9) $N_4=582$ (94.0, 75.9)	$p<0.46$ <i>ESS=6.73, D=27.7</i> $N_3=65$ (46.2, 24.0) $N_4=241$ (60.6, 80.7)			
Physician Took Patient's Problem Seriously	$p<0.0001$ <i>ESS=24.83, D=6.1</i> $N_3=237$ (31.2, 66.7) $N_4=579$ (93.6, 76.9)	$p<0.015$ <i>ESS=16.67, D=10.0</i> $N_3=65$ (24.6, 45.7) $N_4=239$ (92.1, 81.8)	$p's<0.0001$ <i>ESS=40.91, D=5.8</i> $N_3=238$ (55.0, 61.8) $N_4=573$ (85.9, 82.1)	$p's<0.0001$ <i>ESS=44.33, D=3.8</i> $N_3=237$ (61.6, 59.6) $N_4=573$ (82.7, 83.9)	EO-CTA and GO-CTA models were identical
Physician Concern For Patient's Comfort	$p<0.0001$ <i>ESS=22.62, D=6.8</i> $N_3=236$ (34.3, 54.4) $N_4=581$ (88.3, 76.8)	$p<0.29$ <i>ESS=8.08, D=22.8</i> $N_3=65$ (23.1, 29.4) $N_4=240$ (85.0, 80.3)	$p's<0.031$ <i>ESS=40.16, D=7.5</i> $N_3=238$ (61.8, 54.2) $N_4=574$ (78.4, 83.2)	HO-CTA and EO-CTA models were identical	$p's<0.0001$ <i>ESS=39.84, D=4.5</i> $N_3=236$ (61.4, 53.9) $N_4=574$ (78.4, 83.2)
Physician Explanation of Test/Treatment	$p<0.0001$ <i>ESS=28.11, D=5.1</i> $N_3=237$ (41.4, 56.3) $N_4=574$ (86.8, 78.2)	$p<0.0006$ <i>ESS=24.26, D=6.2</i> $N_3=65$ (36.9, 44.4) $N_4=237$ (87.3, 83.5)	$p's<0.0001$ <i>ESS=43.36, D=6.5</i> $N_3=237$ (66.2, 54.7) $N_4=568$ (77.1, 84.6)	HO-CTA and EO-CTA models were identical	$p's<0.0001$ <i>ESS=43.36, D=3.9</i> $N_3=237$ (66.2, 54.7) $N_4=568$ (77.1, 84.6)
Physician Explanation of Illness/Injury	$p<0.0001$ <i>ESS=31.27, D=4.4</i> $N_3=233$ (48.5, 53.6) $N_4=569$ (82.8, 79.7)	$p<0.0028$ <i>ESS=19.83, D=8.1</i> $N_3=64$ (39.1, 35.7) $N_4=234$ (80.8, 82.9)	$p's<0.0001$ <i>ESS=43.54, D=3.9</i> $N_3=233$ (70.0, 52.2) $N_4=564$ (73.6, 85.6)	HO-CTA, EO-CTA, and GO-CTA models were identical	HO-CTA, EO-CTA, and GO-CTA models were identical

Figure 9.3: UniODA Model Predicting Likelihood of Recommending ED using Satisfaction with Wait Time



Overall the UniODA model for waiting time correctly classified 3 of 4 patients who in reality were ambivalent, and 3 of 5 patients who in reality were likely to recommend the ED. When the model predicted a patient was ambivalent it was correct for 2 of 5 patients, and when it predicted a patient was likely to recommend the ED it was correct for 7 of 8 patients.

Examination of all simple bivariate models reveals that all UniODA models obtained for physician patient-care ratings achieved sensitivity greater than 75% for accurate classification of the patients who in reality were likely to recommend the ED to others, and they all yielded a predictive value $> 75\%$ in making accurate predictions regarding those patients who are likely to recommend the ED to others. Only the UniODA model obtained for satisfaction with waiting time yielded a sensitivity of approximately 75% for accurate classification of patients who in reality were ambivalent. Considered as a whole these results suggest fair or worse waiting times increase the likelihood of an ambivalent recommendation, whereas good or very good physician patient-care behaviors increase the likelihood of a positive recommendation.

In novometric theory (see Chapter 12) an ideal statistical model is conceptualized as a perfectly accurate, maximally parsimonious model for a given application. For any given application, the empirical model that is closest to the theoretically ideal model (in accuracy-by-parsimony space) for the application is defined as being the globally optimal empirical model for the application. D , the distance of an empirical model from the theoretically ideal model for a given application, is a function of both ESS and parsimony: D is defined as the number of additional effects having ESS equivalent to the mean ESS of effects already in the model, that are still needed to obtain the theoretically ideal model for the application. As seen, the simple bivariate UniODA model for waiting time is second-closest to a theoretically ideal model versus all attributes except for physician explanation of patient illness/injury ($D = 4.5$ and 4.4 , respectively).

Analysis by Partial UniODA: Associations between likelihood of recommending the ED to others and the five physician patient-care evaluations—with the effect of waiting time eliminated using partial UniODA, are summarized in the second column of Table 9.3. Models with $p < 0.05$ had the same direction and optimal threshold that was identified in simple bivariate analysis. Performance for models involving physician courtesy and physician concern for patient comfort were not statistically significant. Only the performance of the model involving physician explanation of test/treatment was statistically significant at the experimentwise criterion (Sidak $p < 0.05$).

Of the five physician patient-care attributes, the performance of the non-confounded model for the physician's explanation of test/treatment was influenced the least by the confounder, yielding 10.9% lower sensitivity for class category 3 (ambivalent) and 0.6% greater sensitivity for class category 4 (likely to recommend the ED to others) versus the simple bivariate model. In contrast, the performance of the non-confounded model for physician courtesy was influenced the most by the confounder, yielding 75.7% greater sensitivity for class category 3 (the improved sensitivity of 46.2% was nevertheless 3.8% less than is expected by chance), and 35.5% lower sensitivity for class category 4 (for this model $ESS = 6.7$, a very weak effect). Non-confounded models for the other three attributes all had marginally lower sensitivity for class category 4, and substantially lower sensitivity for class category 3: the values respectively were

3.7% and 32.7% lower for the model of physician's concern for the patient's comfort; 1.6% and 19.4% lower for the model of physician took the patient's problem seriously; and 2.4% and 19.4% lower for the model of physician's explanation of illness/injury. For the three non-confounded models with generalized $p < 0.05$ there was a mean reduction of 1.1% in sensitivity for patients likely to recommend the ED, and of 17.2% in sensitivity for ambivalent patients. As stated earlier, the model for waiting time best predicted ambivalence as a function of dissatisfaction waiting times, and when this effect was eliminated from the models for the physician ratings the effect was to reduce the accuracy with which the adjusted models predict ambivalence.

As discussed earlier, forced entry of covariates into CTA models can adversely affect classification accuracy. Presently the reduction in normed accuracy occurring for models corrected for confounding via partial UniODA is evident in the D statistic indicating distance of the empirical model from a theoretically ideal model. Compared to the simple bivariate model, D for the corresponding non-confounded partial UniODA model was greater by 21.6% (physician explanation of test/treatment) to 250.6% (physician courtesy). As seen in Table 9.3, for every attribute except for physician explanation of test/treatment, D for the partial UniODA model was the greatest distance that was reported for the attribute.

In addition to reducing normed accuracy, partial UniODA also greatly diminished the sample size available for assessing the non-confounded relationships. Even though the ESS of the confounder used to predict patient rating of likelihood to recommend the ED to others only fell into the moderate range, the UniODA model for waiting time nevertheless correctly classified almost two-thirds of the observations, leaving only 37% of the original sample available for assessing non-confounded relationship of the patient likelihood rating and rated dimensions of physician behavior.

Quantitative challenges to (sub)optimal partial methods—reduction in ESS and also in statistical power—represent significant statistical engineering issues. However, the *qualitative* challenge to partial methods—the ecological utility of the findings—calls into question their *theoretical significance*. In the present context, for example, after eliminating the effect of waiting time, patient ratings of physician's explanation of the patient's illness/injury predict if a patient is likely to recommend the ED (sensitivity = 82.8), but fail to predict (at a level exceeding accuracy expected by chance) if a patient is ambivalent in this respect (sensitivity = 48.5). Pragmatically, how is this laboratory finding translated into clinical care? Real-world differences in patient satisfaction with waiting time exist, and cannot simply be eliminated from consideration in the clinical setting. What do inter-patient differences in satisfaction with waiting time imply for the validity of the finding that illness/injury explanation affects patient recommendations of the ED? Viewed from a translational perspective, rather than eliminating confounding or moderating factors from the analysis, a more actionable approach involves including the confounders or moderators in analysis and assessing how and to what extent they influence patient decision-making.

Analysis Using CTA: The quintessential statistical methodology for studying moderation in multi-attribute applications, CTA chains successive UniODA models to create a nonlinear multiattribute model that explicitly maximizes ESS for the sample, data geometry (class variable, attributes, and corresponding measurement scales), and hypothesis under investigation. Three modalities of CTA have been developed. The first mode is hierarchically-optimal CTA (HO-CTA): it enters the attribute yielding greatest ESS at every step of the analysis (Chapter 10). The second mode is enumerated-optimal CTA (EO-CTA): it enumerates the first three nodes of the CTA model to identify the model yielding maximum ESS (Chapter 11). The final mode is globally-optimal CTA (GO-CTA): it identifies the CTA model that yields the best (lowest value of D) combination of accuracy and parsimony for a given sample, data geometry, and hypothesis (Chapter 12).

While analyses presented below are non-directional, it is also possible to use parallel methods to analyze confirmatory hypotheses. For example, CTA software allows operators to force attributes into any desired location in the tree model.¹² In the present context it is possible to force the confounding variable (waiting time) to enter the CTA model at the root node: this is conceptually consistent with widely-used GLM methods such as hierarchical linear models or backward-stepping multiple regression analysis.^{1,2} For the present exploratory analyzes this was not done—the CTA algorithm identified the models presented. However, because waiting time had a greater ESS than all physician patient-care behavior ratings except for physician explanation of illness/injury, except for the latter attribute, waiting time was selected as the initial variable in the HO-CTA analyses. Schematic illustrations of UniODA and CTA models follow the same conventions and are subject to complementary interpretations. All three CTA modes are used to evaluate

the relationship between recommendation (class variable), waiting time (confounder), and each separate facet of rated physician patient-care behavior.

Physician Courtesy: Figure 9.4 presents the HO-CTA model obtained by using patient ratings of the physician's concern for the patient's comfort, and patient satisfaction with waiting time, as attributes: the model uses two attributes to identify five distinct patient strata. Waiting time is the root variable: 2 of 5 patients who are dissatisfied with waiting time are likely to recommend the ED to others, compared to 7 of 8 patients who are satisfied. For patients who are ambivalent about waiting time, physician courtesy is important: 2 of 5 patients ambivalent about or dissatisfied with physician courtesy are likely to recommend the ED, versus 7 of 10 patients rating physician courtesy as good, and 8 of 9 patients rating courtesy as very good (this latter difference is not statistically significant at the experimentwise criterion). Evaluated quantitatively, although the D statistic of 8.3 for this model (Table 9.3) is greater than $D = 27.7$ for the partial UniODA model, the HO-CTA model nevertheless has weak efficiency = ESS / number of strata = $40.16 / 5 = 8.03$. Evaluated qualitatively, the model is difficult to translate into clinical practice: an ambivalent satisfaction rating for waiting time is difficult to ascertain in a clinical setting.

Figure 9.4: HO-CTA Model: Physician Courtesy

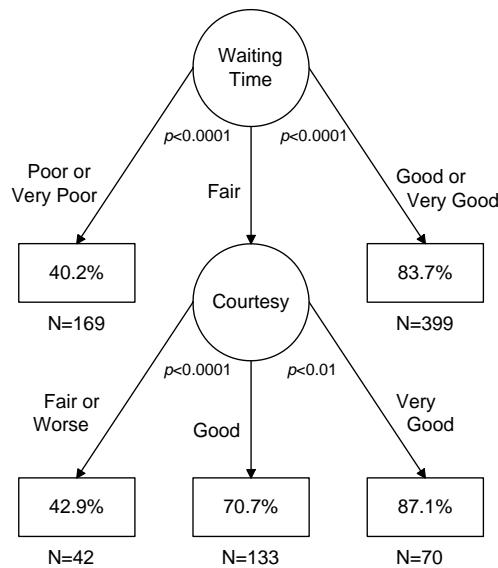


Figure 9.5: EO-CTA Model: Physician Courtesy

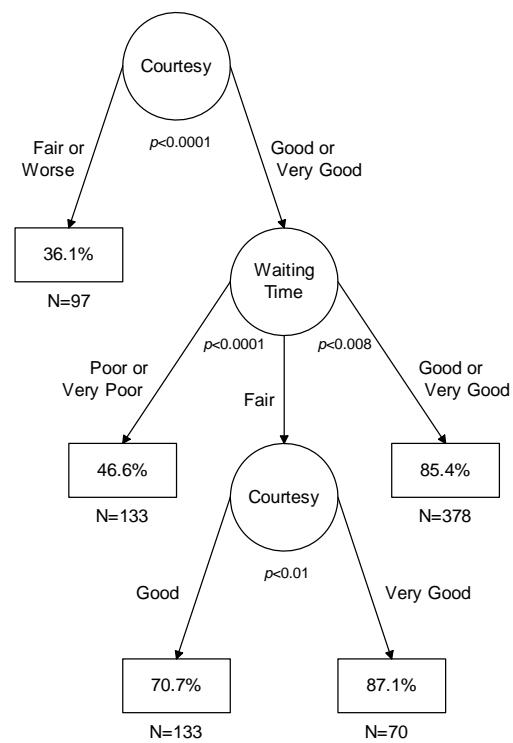
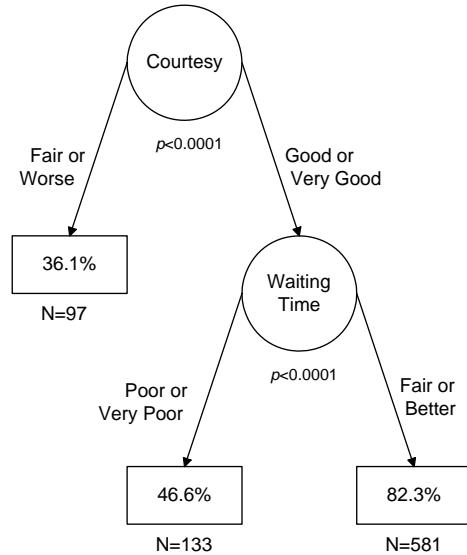


Figure 9.6 presents the GO-CTA model for this application. Quantitatively the GO-CTA model has a smaller D statistic (4.6) than all other models in this application; it is least complex (most parsimonious) of all of the CTA models (only three strata are identified); it achieved ESS nearly as strong as was achieved by the more complex EO-CTA model; its efficiency (13.16) is stronger than was achieved by other models; and both Type I error rates are statistically significant at the experimentwise criterion. Weaknesses of the model include low sensitivity for classifying ambivalence that is only marginally greater than expected by chance (Table 9.3), and there is substantial room for improvement in efficiency. Qualitatively the model is promising. The primary factor influencing the likelihood that a patient will recommend the ED to others is the actionable physician behavior of expressing courtesy to the patient: interventions may be employed to address deficiencies and to establish a baseline level of competency in this regard.^{15,16} For physician-patient interactions rated as reflecting good or very good courtesy, the secondary factor affecting a posi-

tive recommendation is waiting time, which can be ambivalent or better and still motivate a positive outcome from 7 of 8 patients. Satisfaction with waiting time may be achieved as a synergy between nurse management of patient waiting time expectations^{7,8} and information systems developed to help health care workers to keep abreast of patient actual waiting times.¹⁷

Figure 9.6: GO-CTA Model: Physician Courtesy



Physician Takes Patient's Problem Seriously: Figure 9.7 presents the HO-CTA model obtained by using satisfaction with waiting time, and patient ratings of satisfaction with the degree to which the physician approached the patient's problem seriously, as attributes.

Figure 9.7: HO-CTA Model:
Physician Took Patient's Problem Seriously

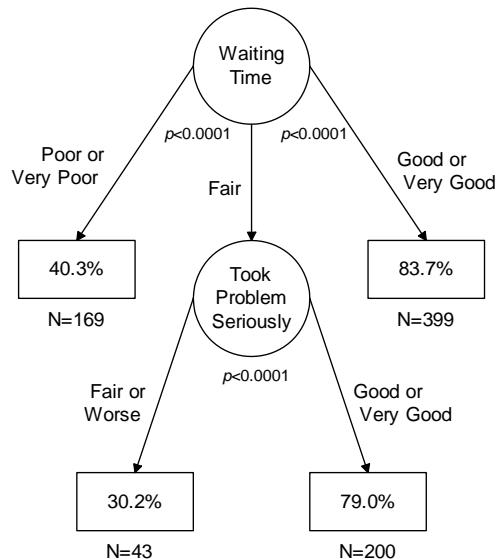
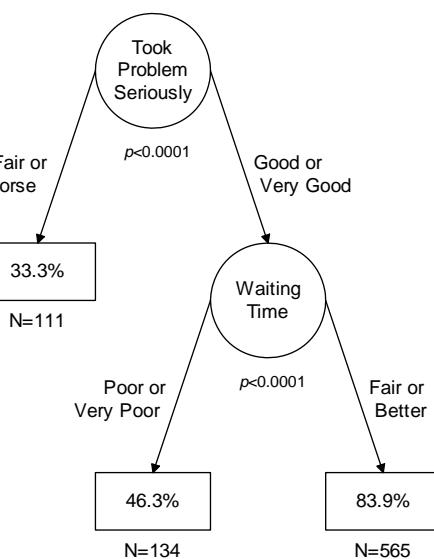


Figure 9.8: EO-CTA and GO-CTA Models:
Physician Took Patient's Problem Seriously



As seen, this model identifies four patient strata. Waiting time is the root variable: 2 of 5 patients who are dissatisfied with waiting time are likely to recommend the ED to others, compared to 7 of 8

patients who are satisfied. For patients ambivalent about waiting time, the perceived degree to which the physician approached the patient's problem is important: 3 of 10 patients ambivalent about or dissatisfied with physician serious problem-solving are likely to recommend the ED, versus 8 of 10 patients satisfied with their perception of the physician's serious problem-solving approach (all effects were statistically significant at the experimentwise criterion). Evaluated quantitatively, the D statistic of 5.8 for this model (Table 9.3) is superior to the partial UniODA model ($D = 10.0$), yet the HO-CTA model has relatively weak efficiency of 10.23. Qualitatively the model is difficult to translate into clinical practice because, again, an ambivalent satisfaction rating for waiting time is difficult to ascertain in a clinical setting.

Figure 9.8 presents EO-CTA and GO-CTA models—which were identical in this application. These models have the same structure and optimal thresholds as the GO-CTA model for physician courtesy. The D statistic of 3.8 is the lowest identified in this example. Little research addresses non-serious physician treatment of a patient's problem. It is reported that cultural influences and embedded cultural implications of some diseases (e.g., sexual, mental illness, obesity, alcoholism, drug addiction) may render some physicians a poor match for some patients.¹⁸ Poorly understood diseases, such as fibromyalgia¹⁹ and hypermobility syndrome²⁰, can frustrate patients and caregivers. The present finding, and the paucity of research in this area, suggests that refining this theoretical construct and its measurement are warranted.

Physician Concern for Patient's Comfort: Figure 9.9 presents the HO-CTA and EO-CTA models—identical in this application—obtained using satisfaction with waiting time, and patient ratings of satisfaction with physician concern for patient comfort, as the attributes. As seen the model identifies five patient strata: the D statistic indicates a mediocre effect, the efficiency of the model (8.03) is weak, not all Type I error rates are statistically significant at the experimentwise criterion, and the model is difficult to translate into clinical practice.

Figure 9.9: HO-CTA and EO-CTA Models:
Concern for Patient's Comfort

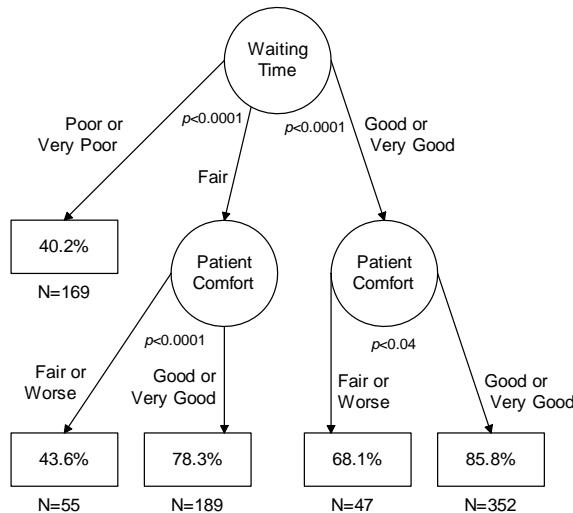
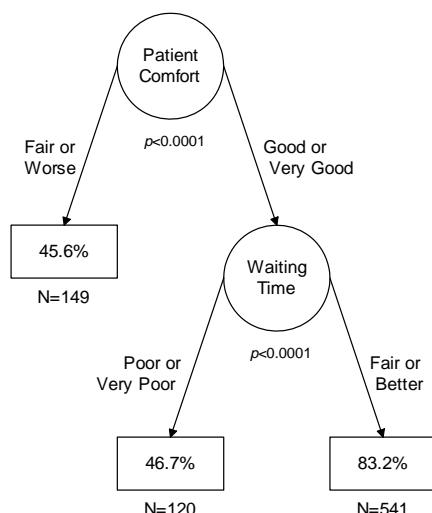


Figure 9.10: GO-CTA Model:
Concern for Patient's Comfort



The GO-CTA model obtained in this application is presented in Figure 9.10. This model has the same structure and optimal thresholds as the GO-CTA model for physician courtesy. The D statistic of 4.5 is relatively low, but would be lower if the accuracy reflected by the two left-most endpoints was greater than the level of classification accuracy expected by chance. Comfort, specifically pain management, is a widely-reported correlate of patient satisfaction.²¹

Physician Explanation of Test/Treatment: Figure 9.11 presents the identical HO-CTA and EO-CTA models obtained using satisfaction with waiting time, and patient ratings of satisfaction with physician explanation of the test/treatment, as attributes. The model uses four attributes (the most complex model

reported presently) to identify five patient strata: the quality ($D = 8.67$) and efficiency (8.67) of the model are mediocre, and the model is exceedingly difficult to translate into clinical practice.

Figure 9.11: HO-CTA and EO-CTA Models:
Explanation of Test/Treatment

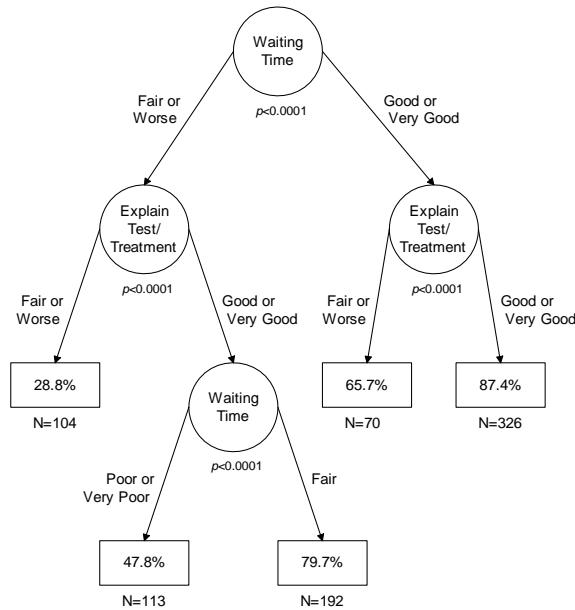
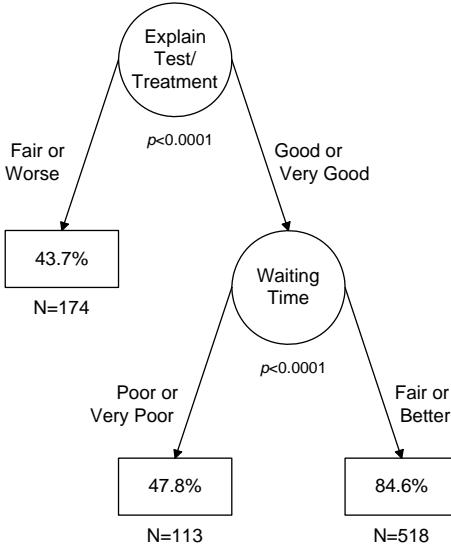


Figure 9.12: GO-CTA Model:
Explanation of Test/Treatment



The GO-CTA model obtained for this application is presented in Figure 9.12. Consistent with the GO-CTA model identified for physician serious problem-solving orientation, the D statistic of 3.9 is low—indicative of a powerful model. Consistent with the GO-CTA model identified for physician concern for patient comfort, D would have been lower if the two left-most endpoints had surpassed the classification accuracy expected by chance. Research examining this attribute in the context of patient recommendation of the ED hasn't been reported, however physician "explanation" has been reported as secondary to physician interpersonal skills in predicting hospital recommendations of patients being treated for stroke, diabetes mellitus, Caesarean section, or appendectomy.²²

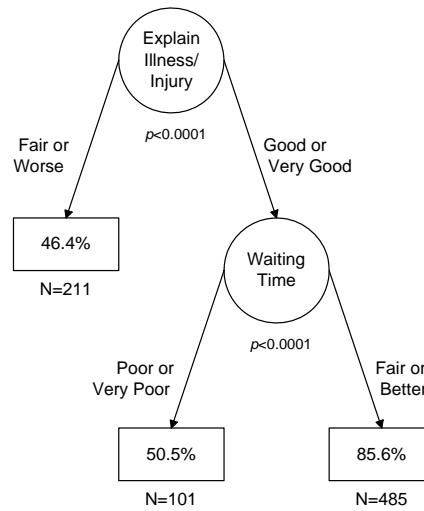
Physician Explanation of Illness/Injury: Finally, Figure 9.13 gives the identical HO-CTA, EO-CTA, and GO-CTA models obtained using satisfaction with waiting time, and patient ratings of satisfaction with physician explanation of illness/injury, as the attributes. Consistent with the GO-CTA model identified for physician explanation of illness/injury, the D statistic of 3.9 is low—indicative of a powerful model, and D would have been lower if the two left-most endpoints had yielded classification accuracy that was greater than is expected by chance.

The consistency of all five of the GO-CTA models that were identified presently is striking—all had identical structure, including optimal threshold values. While this result is to be anticipated if the ratings of physician patient-care behaviors are strongly associated, this was not the case: while directional (confirmatory) and non-directional (exploratory) models of inter-rating association were statistically significant at the experimentwise criterion, the *ESS* statistics ranged between 34.1 and 63.2—that is, between moderate to relatively strong effects. Nevertheless the GO-CTA model obtained by using waiting time *and all five of the physician patient-care behaviors as attributes* was identical to the GO-CTA model obtained by using only waiting time and rating of whether the physician took the patient's problem seriously (Figure 9.8). Clearly, in order to better understand patient recommendation of the ED, superior measures of the current constructs, and/or additional, presently unmeasured attributes are needed.

Another aspect of consistency between the five GO-CTA models is the pattern of the findings. For every model the right-most endpoint (reflecting satisfaction with physician behavior, and absence of dis-

satisfaction with waiting time) is strongly homogeneous—at least 4 of 5 observations in the endpoint are consistent in reporting being likely to recommend the ED to others. For each model the middle endpoint (reflecting satisfaction with physician behavior, and dissatisfaction with waiting time) was the least homogeneous—with half of the observations reporting being likely to recommend the ED to others. The left-most endpoint ranged between moderately homogeneous (for ratings of physician courtesy, and of serious problem-solving orientation) to heterogeneous (for ratings of concern for patient comfort, and of explanation of test/treatment and injury/illness). In a theoretically ideal classification model, all endpoints are perfectly homogeneous, and classification accuracy is perfect (Chapter 12). Clearly, therefore, the two left-most endpoints identify the patient strata for which the GO-CTA models require additional, presently unmeasured attributes, to achieve substantially improved accuracy and thereby reduce D .

Figure 9.13: HO-CTA, EO-CTA, and GO-CTA Models: Explanation of Illness/Injury



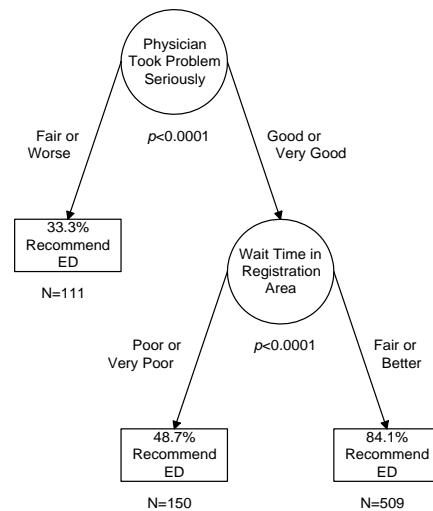
Substantively the two strongest GO-CTA models identified—based on courtesy and serious problem-solving orientation—are conceptually consistent with theoretical constructs known as *expressive* and *instrumental* predispositions, respectively.²³ Popularized by the study of psychological androgyny—a “personality” typology defined as a behavioral repertoire consisting of many instrumental (concern with getting the job done) and expressive (concern for the well-being of others) capabilities—these behavioral dimensions have also been identified in research in fields including management (production- and employee-centered focus, respectively), leadership (initiating structure—or task completion focus, and consideration—or psychological closeness of leader and subordinate, respectively), and conflict resolution (assertiveness—concern with one’s own needs, and cooperation—concern with others’ needs, respectively), among others.²³ These dimensions may be measured for physicians using brief self-rating instruments.²⁴ Scores on instruments assessing these dimensions possess good reliability properties, and some evidence suggests that measures of these constructs are culturally cross-generalizable.^{25,26} An androgynous behavioral repertoire has been shown to be related to lower reliance on technology to solve complex and difficult decision-making tasks among both new and also experienced physicians.^{27,28} An androgynous predisposition in physicians has also been shown to be related to psychological empathy—reflecting cognitive understanding of (versus sympathy—reflecting an emotional reaction to) the condition of a patient.^{29,30} Consistent with results obtained for androgyny, an empathic orientation has been shown to be related to lower utilization of technology when solving complex, difficult decision-making tasks among physicians.³¹ Research is warranted comparing satisfaction and quality-of-care of both patients and their attending physicians, when both members of the dyad have complementary homeostatic preferences for instrumental and expressive aspects of patient-care. If training physicians to detect the preference of the patient for these independent dimensions proves to be difficult and/or unsuccessful, it is possible that an

efficient pre-screening of new patients will enable administrators to assign patients-physician dyads that are primed for optimal outcomes with respect to desired and delivered patterns of patient-care.³²

Multiple Confounders

The example is continued to demonstrate a CTA-based approach to assessing the effect of two or more confounding variables on the estimated association of a class variable and attribute(s). Here, as in the prior analyses, patient self-ratings of the likelihood they will recommend an ED to others (class variable) are modeled on the basis of five patient-rated dimensions of physician patient-care behavior (attributes). However, in addition to patient ratings of their satisfaction with waiting time spent in the *treatment area*, patient ratings were also available for their satisfaction with waiting time in the *registration area*, before going to wait in the treatment area (assessed using the identical 5-point Likert-type scale). The GO-CTA model obtained by including the second confounder in the analysis is presented in Figure 9.14.

Figure 9.14: GO-CTA Model Predicting Self-Rated Likelihood of Recommending the ED to Others



Overall classification performance of this GO-CTA model is summarized in Table 9.4.

Table 9.4: Confusion Table for GO-CTA Model Predicting Likelihood of Recommending the ED to Others

		Predicted Patient Rating	
		3	4
Actual Patient	3	151	81
	4	110	428

For this model $D = 3.7$, which is superior to (lower than) $D = 3.8$ for the parallel model (including discriminant thresholds) developed earlier using waiting time in the treatment area in the bottom node (Figure 9.8). When a GO-CTA model was attempted using only the two confounding waiting times as the potential attributes, no multiattribute model was identified: the best model was a UniODA involving a single threshold on waiting time in the treatment area ($D = 4.5$).

In the model the root node involves ratings of some of the last (most recent) interactions of the patient and “the ED”: the patient-doctor interaction is reminiscent of a recency effect. And, the bottom node involves ratings of some of the first (earliest) interactions of the patient and “the ED”: waiting time in the registration area is reminiscent of a primacy effect.

For the left-most endpoint, 2 in 3 patients are ambivalent about recommending the ED, and for the right-most endpoint, 7 in 8 patients are likely to recommend the ED to others. Neither of these strata

is perfectly homogeneous: by definition, perfect predictive accuracy requires perfect homogeneity within all sample strata identified by the model (Chapter 12). The middle endpoint is most heterogeneous, with 1 in 2 patients ambivalent about recommending the ED to others. This middle strata represents 150 / 770 or 19.5% of the overall sample. The greatest opportunity for meaningful increase in model accuracy (i.e., the biggest decrease in D) lies in improving homogeneity in the middle strata.

Marginal Structural Models

Statistical analysis of a treatment effect in the context of a known confounder via the *marginal structural model* (MSM) approach is well-described³⁴ and illustrated using the data presented in Table 9.5. For each observation the status of the confounder and of the treatment are coded as present ("1") or absent ("0"). The observed data also are measured on a nominal scale (N is the number of observations in every cell in the design). For this application MSM analysis found: "the causal risk difference, risk ratio, and odds ratio are -0.32, 0.50, and 0.26" (Robins³⁴, p. 558).

Table 9.5: Observed Dichotomous Data from Point-Treatment Study, Stratified by Measured Confounder

Confounder	Data	Treatment	N
1	1	1	108
1	1	0	24
1	0	1	252
1	0	0	16
0	1	1	20
0	1	0	40
0	0	1	30
0	0	0	10

Note: Adapted from Robins, *et al.*³⁴

Figure 9.15 is the GO-CTA model obtained for these data using treatment as a class variable and the observed data and confounder as possible attributes.³⁵ Among 100 observations with the confounder *absent* (left-most endpoint), the model correctly predicts 50 (50.0%) were in the treatment condition: for this strata the observed data didn't improve model accuracy. However, among 400 observations with the confounder *present*, an observed response is associated with percent of strata in the treatment condition.

Figure 9.15: GO-CTA Model Relating Treatment and Observed Dichotomous Data: Known Confounder

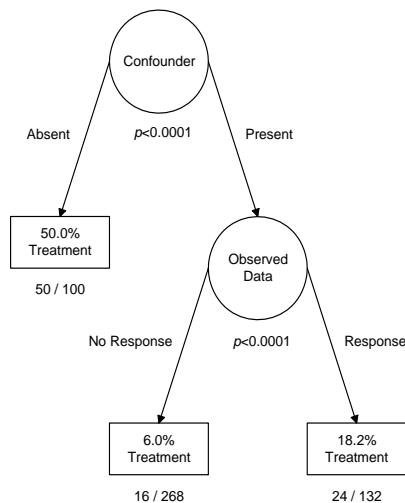


Table 9.6 presents the confusion table for this model.

Table 9.6: Confusion Table for GO-CTA Model Relating Treatment and Observed Data: Known Confounder

		Predicted Treatment Condition	
		0	1
Actual Treatment Condition	0	74	16
	1	158	252

The model correctly classified 82.2% of observations who in reality were in the control condition, and 61.5% of observations who in reality were in the treatment condition: $ESS = 43.7\%$. The model was correct 31.9% of the time that it predicted an observation was from the control condition, and 94.0% of the time that it predicted an observation was from the treatment condition: $ESP = 25.9\%$.

Confounding by Combining Groups

Knowing or suspecting that a confounder threatens the validity of one's statistical analysis and research conclusions is one matter. However, not knowing or suspecting that paradoxical confounding threatens the validity of one's statistical analysis and research conclusions is an entirely different matter. *Simpson's Paradox may be the single greatest threat to the validity of quantitative analysis in all empirical science.*³⁶ Paradoxical confounding can occur when the data from two or more samples, groups or time periods are combined into a single sample: under such conditions the results obtained when analyzing the combined data may be different than when analyzing individual data sets separately.³⁷

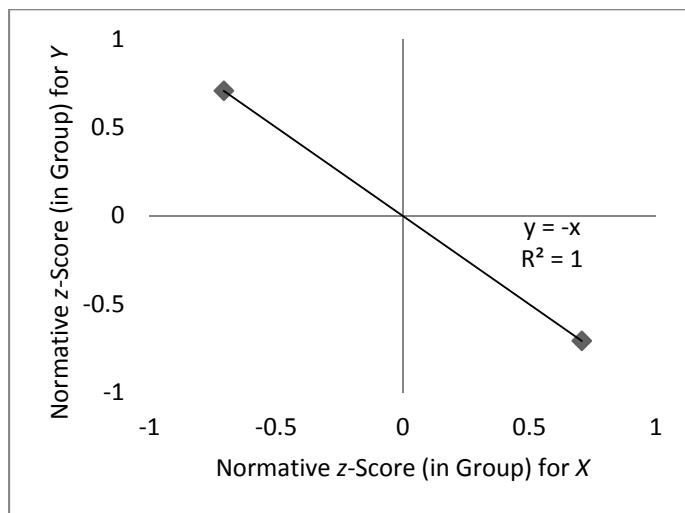
Table 9.7: Hypothetical Raw and Normatively Standardized Data for Ten Groups

Group	Observation	X	Z _X	Y	Z _Y
1	1	0	-0.707	1	0.707
	2	1	0.707	0	-0.707
2	3	1	-0.707	2	0.707
	4	2	0.707	1	-0.707
3	5	2	-0.707	3	0.707
	6	3	0.707	2	-0.707
4	7	3	-0.707	4	0.707
	8	4	0.707	3	-0.707
5	9	4	-0.707	5	0.707
	10	5	0.707	4	-0.707
6	11	5	-0.707	6	0.707
	12	6	0.707	5	-0.707
7	13	6	-0.707	7	0.707
	14	7	0.707	6	-0.707
8	15	7	-0.707	8	0.707
	16	8	0.707	7	-0.707
9	17	8	-0.707	9	0.707
	18	9	0.707	8	-0.707
10	19	9	-0.707	10	0.707
	20	10	0.707	9	-0.707

A hypothetical example is used to illustrate paradoxical confounding for a simple design involving assessing the linear correlation r_{XY} between two ordered attributes, X and Y. Presented in Table 9.7, data on X and Y are available for two observations in each of ten different groups (samples, time periods). The ten groups could be, for example, observations of different income levels, disease stage, geographic area, ethnicity, political orientation, age, or any other such *subject factor*. The ten groups could be people who were experiencing different levels of an emotional response such as (dis)satisfaction or pain—assessed by a Likert-type scale. These could also be ten sequential measurements for a sample or an individual series.

In this hypothetical example when data are normatively standardized separately by group, the low ($z = -0.707$) and high ($z = 0.707$) score in each group is identical. Figure 9.16 illustrates the regression model (r_{XY}) obtained for each group when standardized data are analyzed separately by group.

Figure 9.16: Scatterplot for Hypothetical Normatively Standardized (by Group) Data



In each of the individual groups the raw and normative z-scores are perfectly correlated ($r_{XY} = 1$): this is indicated in **bold** in the major diagonal of Table 9.8. Seen above the diagonal, combining data that were normatively standardized separately by group correctly yields $r_{XY} = -1$ actually underlying the data, for every possible combination of two different groups.

Table 9.8: r_{XY} for Two Combined Groups: Raw versus Normatively Standardized Data

Raw Data	Normative z-Score Data									
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
<u>G1</u>	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
<u>G2</u>	0	1	-1	-1	-1	-1	-1	-1	-1	-1
<u>G3</u>	0.60	0	1	-1	-1	-1	-1	-1	-1	-1
<u>G4</u>	0.80	0.60	0	1	-1	-1	-1	-1	-1	-1
<u>G5</u>	0.88	0.80	0.60	0	1	-1	-1	-1	-1	-1
<u>G6</u>	0.92	0.88	0.80	0.60	0	1	-1	-1	-1	-1
<u>G7</u>	0.95	0.92	0.88	0.80	0.60	0	1	-1	-1	-1
<u>G8</u>	0.96	0.95	0.92	0.88	0.80	0.60	0	1	-1	-1
<u>G9</u>	0.97	0.96	0.95	0.92	0.88	0.80	0.60	0	1	-1
<u>G10</u>	0.98	0.97	0.96	0.95	0.92	0.88	0.80	0.60	0	1

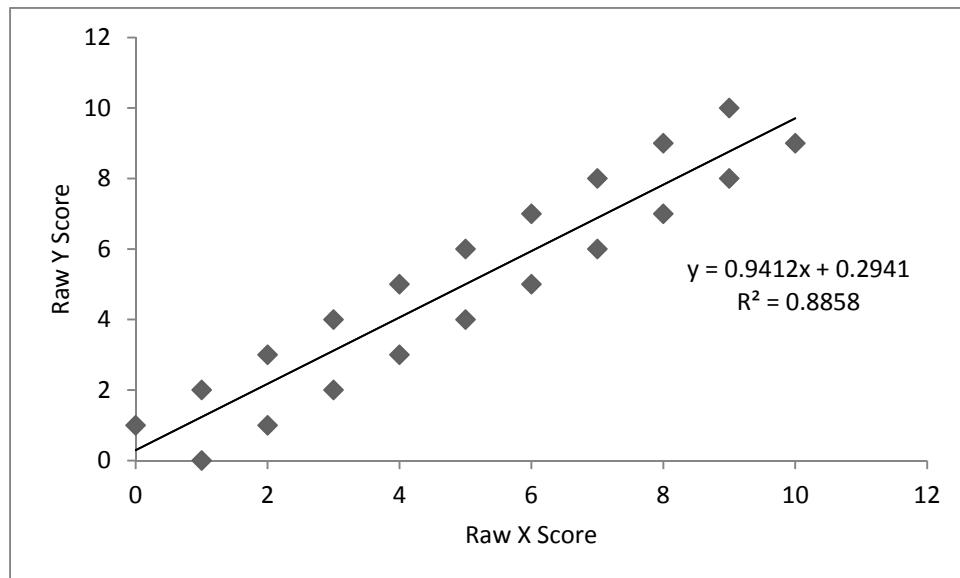
In contrast, in this example, combining raw data of two or more different groups *always induces paradoxical confounding*. In Table 9.8 the entry within the cell directly underneath each cell of the major diagonal is the r_{XY} obtained for every combination of data from group G_i and group G_{i+1} : in every case $r_{XY} = 0$. For all other entries beneath the diagonal, $r_{XY} > 0$: the greater the distance of the cell (in the Table) from the diagonal, the greater the value of r_{XY} . Table 9.9 gives the r_{XY} obtained for group G1; for groups G1 and G2; for the first three groups; and finally for the combination of all ten groups—by using raw data versus normative z-scores computed separately by group.

Table 9.9: Confounding Increases for (Combined) Group(s) as the Domain of X and Y Increases

r_{XY} for (Combined) Group(s)		
<u>Group(s)</u>	<u>Raw Data</u>	<u>Normative z-Score</u>
G1	-1	-1
G1-G2	0	-1
G1-G3	0.45	-1
G1-G4	0.67	-1
G1-G5	0.78	-1
G1-G6	0.84	-1
G1-G7	0.88	-1
G1-G8	0.91	-1
G1-G9	0.93	-1
G1-G10	0.94	-1

Figure 9.17 is a scatterplot of raw data of all groups combined into a single sample to estimate r_{XY} , as well as the corresponding regression model. Although in reality each of the ten pairs of raw scores that are plotted yields a *perfect negative* relationship (see Figure 9.16), nevertheless here the paradox is compelling—manifest as an *extremely strong positive* relationship.

Figure 9.17: Scatter Plot for Hypothetical Raw Data



This particular manifestation of Simpson's paradox is *not* the only data geometry (configuration) that induces confounding when estimating r_{XY} . Normative standardization performed separately by group will *not* always circumvent paradoxical confounding. As a means of illustrating other susceptible data configurations, symbolic representation illustrated in Figure 9.18 is used: an arrow having a positive slope (A) indicates a group (sample) with underlying $r_{XY} > 0$; an arrow having a zero slope (B) indicates a group with no underlying linear effect ($r_{XY} = 0$); and an arrow having a negative slope (C) indicates a group with underlying $r_{XY} < 0$. If desired an ellipse could be used to surround the individual data points (or to indicate 95% confidence boundaries) for A and C, and a circle used for the data of B.

Figure 9.18: Symbolic Representation of Sample Actual r_{XY}

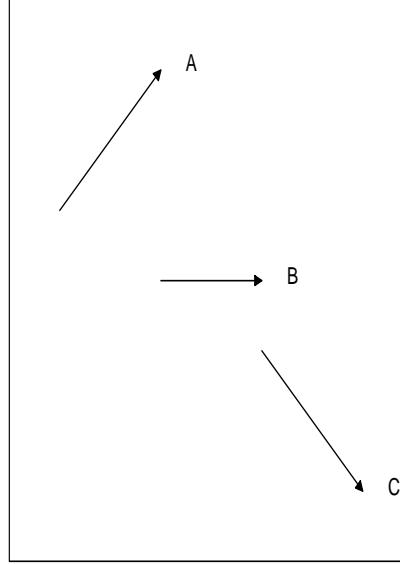
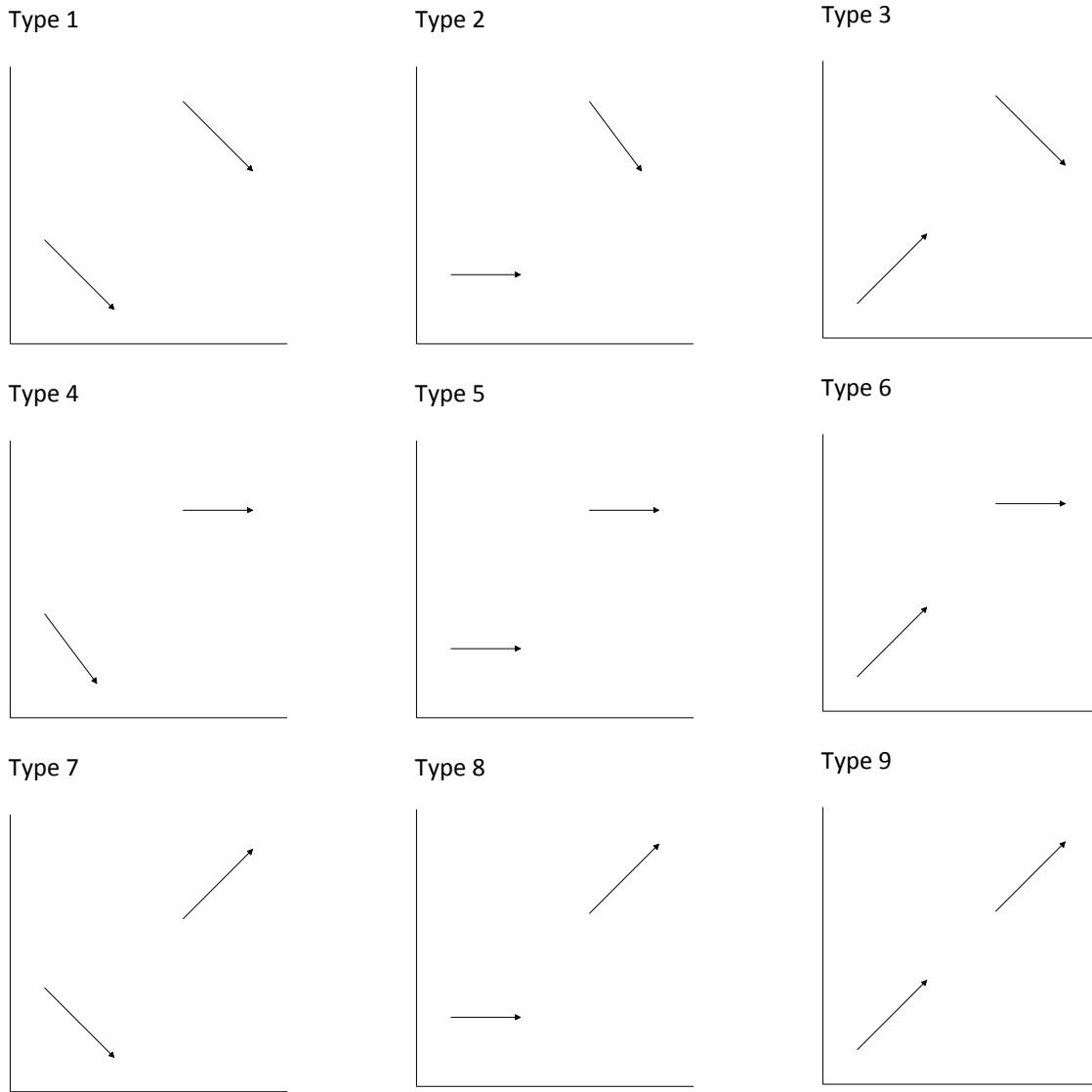


Figure 9.19 illustrates nine different data configurations that can occur for two or more groups, that can induce paradoxical *positive* correlations if data are combined. The most extreme confounding in this configuration family is indicated as Type 1, and was the focus of the example just provided. In Type 1 configurations, if r_{XY} in the groups is comparable then normative standardization of X and Y conducted separately by group will circumvent confounding. Identification of a positive r_{XY} for configurations with at least one group having an underlying $r_{XY} = 0$ would be paradoxical (Types 2, 4, 5, 6, and 8), and identification of a positive r_{XY} for the configurations with at least one group having an underlying $r_{XY} < 0$ would be paradoxical (Types 1, 2, 3, 4, and 7). In configuration Type 9, although $r_{XY} > 0$ for both groups, r_{XY} obtained for the combined sample may be higher or lower than is obtained for any group separately, depending upon the magnitude of score overlap between groups.³⁷ All these configurations involve groups that have different mean values on X and/or Y : the greater the difference—the less overlap of the distributions, the greater the magnitude of the confounding (Table 9.9).

Figure 9.20 illustrates nine different data configurations that can occur for two or more groups, that can induce paradoxical *negative* correlations if data are combined. The most extreme confounding in this configuration family is indicated as Type 18: if r_{XY} in the groups is comparable then normative standardization of X and Y conducted separately by group will circumvent confounding. Identification of a negative r_{XY} for configurations with at least one group having an underlying $r_{XY} = 0$ would be paradoxical (Types 11, 13, 14, 15, and 17), and identification of a negative r_{XY} for the configurations with at least one group having an underlying $r_{XY} > 0$ would be paradoxical (Types 12, 15, 16, 17, and 18). In configuration Type 10, although $r_{XY} < 0$ for both groups, r_{XY} obtained for the combined sample may be higher or lower than is obtained for any group separately, depending upon the magnitude of score overlap between groups.³⁷ These configurations involve groups with different means on X and/or Y : the greater the difference (i.e., the less overlap of the distributions) the greater the magnitude of the confounding (Table 9.9).

Of the 18 different data configuration patterns that are illustrated in Figure 9.18 and Figure 9.19, only Type 1, 9, 10, and 18 configurations may be combined if: (a) r_{XY} in all constituent groups is statistically comparable, and (b) data are normatively standardized separately by group prior to being combined.

Figure 9.19: Nine Two—Group Configurations that Induce Paradoxical Positive Relationships



In applications with two or more groups (samples, testings, time periods), even if distributions of X and Y overlap this doesn't obviate possible paradoxical confounding: three different data configurations that induce confounding in this circumstance are presented in Figure 9.21.

In contrast to this set of 21 data configurations that can induce paradoxical confounding if data of two or more groups are combined, Figure 9.22 illustrates three homogeneous data configurations that reflect either positive, negative, or null relationships (recall Chapter 2 discussion of pre-processing data).

Figure 9.20: Nine Two—Group Configurations that Induce Paradoxical *Negative* Relationships

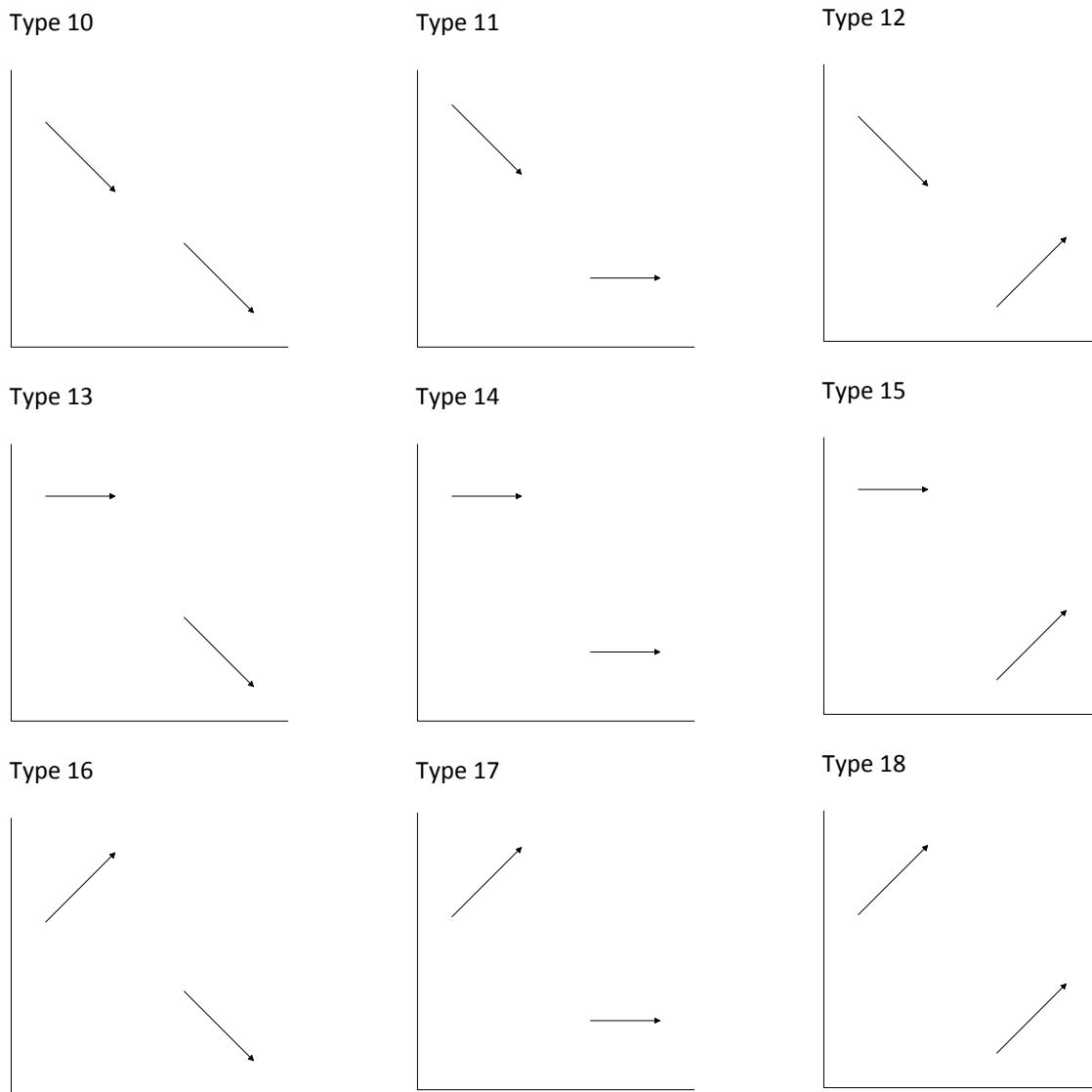


Figure 9.21: Three Two—Group Configurations that Induce Paradoxical Findings

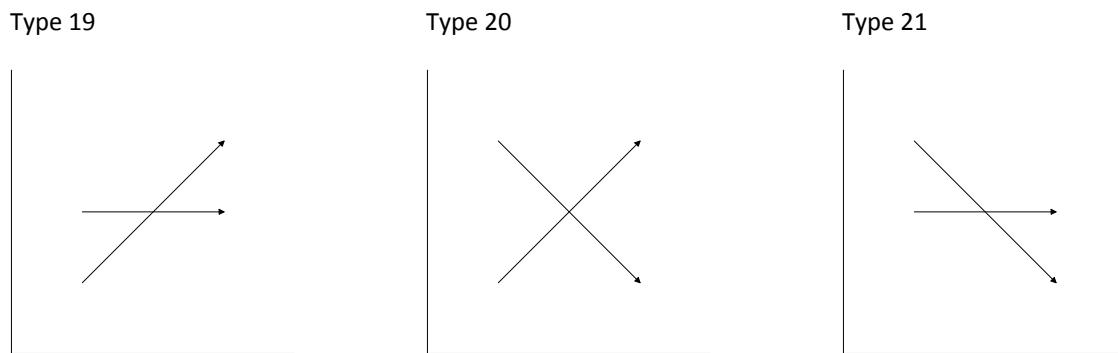
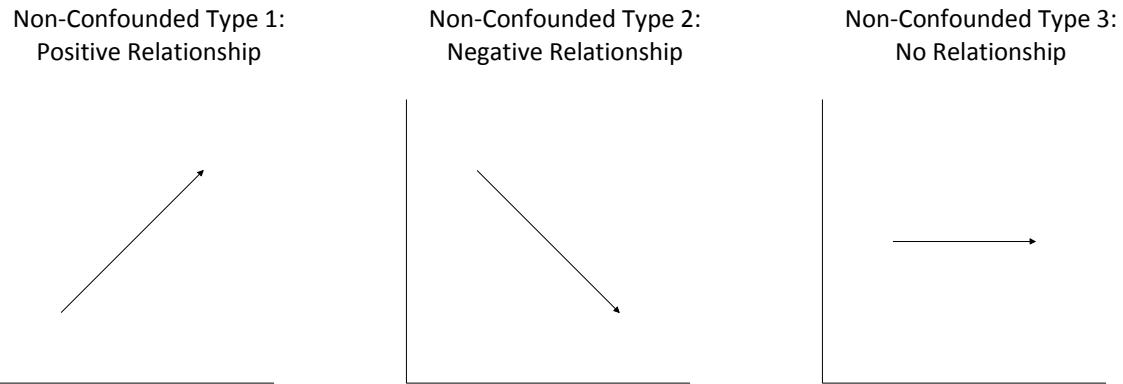
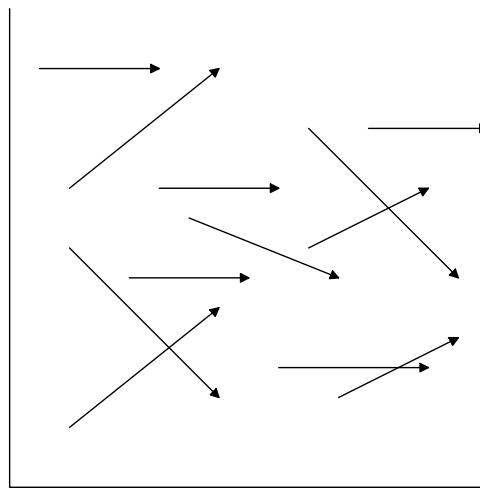


Figure 9.22: Three Non-Confounded Two-Group Configurations



In reality the analytic situation is more complex than is presented in these idealized situations. Except under conditions of strict experimental control, empirical samples typically reflect many different groups. For example, in research involving convenience samples of humans many possible confounders exist in “the sample”: age, wealth, political views, gender, marital status, intelligence, personality factors, knowledge, experience, interest (passion), emotional status at time of testing, and so forth. For research investigating X and Y , a more realistic representation of the analytic problem confronting linear models is illustrated in Figure 9.23.

Figure 9.23: Typical Empirical Two-Group Data Configuration



The complexity of the linear approach presented here is surprising given the relative simplicity of the problem—identifying the relationship between two ordered attributes. Mental imagery will facilitate an enhanced understanding of the true complexity involved in using linear models. Imagine an application involving a dependent (class) variable Y and two independent variables (attributes) X and Z . Dimension Z is added to Figure 9.23 at a right angle from the centroid, forming a three-dimensional cube: the arrows are now surrounded by three-dimensional ellipses (effects) or spheres (no effects). Many applied articles use four, five or more attributes: inherent complexity of such hyperdimensional problems makes visualization challenging, and avoidance of paradoxical confounding—difficult to impossible.

“Junk Science” in the Courtroom

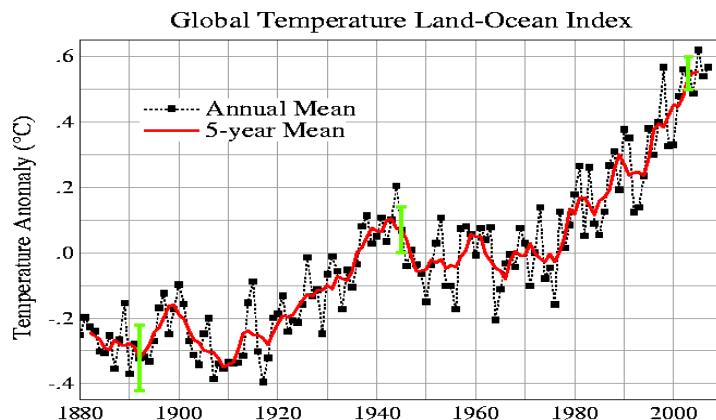
Applied statistics is increasingly used by decision-makers in a wide panorama of areas, in an effort to exert control over resources and to maximize return. If measurement or statistical analysis in such applications is compromised, and if decision-makers act upon corresponding incorrect conclusions in a manner that

induces unintended harm, then legal remedies can result. As an example of such a scenario, consider an article reviewing a recent federal court case in which a standardized test of cognitive ability was ruled invalid and discriminatory for use in hiring Latinos. Reflecting standard use practice, spurious statistical arguments exploit language in the current Uniform Guidelines for evaluating the fairness and validity of personnel selection tests. These issues include how to avoid capitalizing on chance; how to define what constitutes “a measure” of job performance; how to evaluate the meaningfulness of group differences in performance measures; and how to combine data from different sex, race, or ethnic subgroups when computing validity coefficients for the pooled, total sample. Pursuant to the Uniform Guidelines’ standard for unfairness, if one (ethnic) group scores higher on an employment test then the test is deemed “unfair” if this difference is not reflected in a measure of job performance. Although studies validating selection may survive the unfairness test, such data are vulnerable to bias and manipulation unless appropriate statistical procedures are used. Benefits (greater clarity and precision) and potential costs (loss of legal precedent) of revising the Uniform Guidelines to address these issues, legal procedures to limit the use of “junk science” in the courtroom, and the need to reevaluate validity generalization in light of Simpson’s “false correlation” paradox are considered elsewhere.³⁸

Confounding by Combining Time Periods

Simpson’s Paradox threatens the validity of quantitative atmospheric science because nonstationarity is prevalent in longitudinal data series used in atmospheric science, such as temperature or pressure—and nonstationarity can induce Simpson’s Paradox.³⁹ For example, global surface temperature data clearly are nonstationary: in Figure 9.24, anomalies are computed relative to the period 1951–1980.

Figure 9.24: Mean Global Temperature Land-Ocean Index Anomaly by Year⁴⁰



Analysis was restricted to the time period that is the focus of most current quantitative atmospheric science, beginning in the year 1948. Eyeball analysis of Figure 9.24 suggests a relatively flat trajectory (“stationary series”) through 1976, versus a steadily increasing trajectory (“non-stationary series”) across subsequent years. Regression analyses modeling temperature anomaly (dependent measure) as a function of year (independent measure), separately by month, are summarized In Table 9.10: findings confirm the eyeball observations, and establish the generalizability of the phenomenon to a time period more granular than is afforded by annual measurements. Tabled for each model is the intercept as well as the value of the *t*-test for the two-tailed hypothesis that the value of the intercept is zero, and the associated Type I error rate. For every model, in every month, the intercept is *not* significantly different than zero for the stationary series, but *is* significantly different than zero for the nonstationary and combined series. Also tabled for each model is the slope (regression beta weight) and the value of the *t*-test for the two-tailed hypothesis that the value of the slope is zero, and the associated Type I error rate. Consistent with findings for intercept, for every model, in every month, the slope is *not* significantly different than

zero for the stationary series, but *is* significantly different than zero for the nonstationary and combined series. Finally, Table 9.10 provides the percent of variance in temperature that is explained by the regression model as a function of year (R^2), and p for the regression model. If model performance for the combined sample lies outside performance results for samples considered individually, then paradoxical confounding exists: this is indicated in red.

Table 9.10: Regression Modeling of Temperature Anomaly using Year, Separately by Month:
Evidence of Paradoxical Confounding

Month	Time Period	Intercept, t , p				Slope, t , p				R^2 , p	
January	Stationary	559.3	0.8	0.45		-0.29	-0.8	0.46		2.1	0.45
	Non-Stationary	-3239.1	-5.3	0.0001		1.64	5.3	0.0001		49.4	0.0001
	Combined	-2114.5	-7.9	0.0001		1.08	8.0	0.0001		52.2	0.0001
February	Stationary	-140.0	-0.2	0.87		0.07	0.2	0.87		1.0	0.87
	Non-Stationary	-3842.6	-5.5	0.0001		1.95	5.6	0.0001		51.6	0.0001
	Combined	-2451.3	-8.4	0.0001		1.25	8.5	0.0001		55.3	0.0001
March	Stationary	-550.5	-0.8	0.46		0.28	0.8	0.46		2.1	0.46
	Non-Stationary	-3374.5	-5.9	0.0001		1.71	5.9	0.0001		54.9	0.0001
	Combined	-2451.8	-10.0	0.0001		1.25	10.1	0.0001		63.8	0.0001
April	Stationary	-229.4	-0.4	0.71		0.12	0.4	0.72		0.5	0.72
	Non-Stationary	-3216.2	-7.1	0.0001		1.63	7.1	0.0001		63.7	0.0001
	Combined	-2159.7	-10.3	0.0001		1.10	10.4	0.0001		65.0	0.0001
May	Stationary	-197.5	-0.3	0.75		0.10	0.3	0.75		0.4	0.75
	Non-Stationary	-2590.9	-4.9	0.0001		1.31	4.9	0.0001		45.4	0.0001
	Combined	-1845.2	-8.6	0.0001		0.94	8.7	0.0001		56.7	0.0001
June	Stationary	-145.7	-0.3	0.75		0.07	0.3	0.75		0.4	0.75
	Non-Stationary	-3291.0	-6.3	0.0001		1.67	6.4	0.0001		58.3	0.0001
	Combined	-1918.6	-9.7	0.0001		0.98	9.7	0.0001		62.0	0.0001
July	Stationary	-111.3	-0.3	0.78		0.06	0.3	0.79		0.3	0.79
	Non-Stationary	-2841.5	-4.7	0.0001		1.44	4.8	0.0001		43.8	0.0001
	Combined	-1937.1	-9.5	0.0001		0.99	9.6	0.0001		61.3	0.0001
August	Stationary	203.2	0.4	0.73		-0.10	-0.4	0.73		0.5	0.73
	Non-Stationary	-3492.9	-6.5	0.0001		1.77	6.6	0.0001		60.0	0.0001
	Combined	-1933.3	-8.5	0.0001		0.98	8.6	0.0001		55.8	0.0001
September	Stationary	3.9	0.0	0.99		-0.01	-0.0	0.99		0.1	0.99
	Non-Stationary	-3359.2	-6.3	0.0001		1.70	6.4	0.0001		58.4	0.0001
	Combined	-1888.2	-8.8	0.0001		0.96	8.8	0.0001		57.3	0.0001
October	Stationary	298.4	0.6	0.58		-0.15	-0.6	0.58		1.2	0.58
	Non-Stationary	-4082.0	-8.5	0.0001		2.06	8.5	0.0001		71.4	0.0001
	Combined	-1920.6	-8.5	0.0001		0.98	8.5	0.0001		55.7	0.0001
November	Stationary	-253.9	-0.5	.062		0.13	0.5	0.62		0.9	0.62
	Non-Stationary	-3719.7	-6.1	0.0001		1.88	6.1	0.0001		56.3	0.0001
	Combined	-2056.9	-9.1	0.0001		1.05	9.1	0.0001		58.9	0.0001
December	Stationary	41.4	0.1	0.95		-0.02	-0.7	0.95		0.1	0.95
	Non-Stationary	-3076.1	-5.0	0.0001		1.56	5.1	0.0001		45.1	0.0001
	Combined	-1998.4	-8.2	0.0001		1.02	8.3	0.0001		54.2	0.0001

Note: Stationary = 1948 – 1976; Non-Stationary = 1977 – 2007; Combined = 1948 - 2007.

This exercise demonstrates that temperature does *not* increase between 1948 and 1976, but it *does* increase thereafter; that fundamentally different “statistical infrastructure” (i.e., regression models) underlie the stationary and nonstationary series; and that combining data from these two series typically results in paradoxical confounding. What is the nature of this confounding? In the hypothetical example involving combining groups the nature of the confounding is one of “direction”: the result for the com-

bined sample was opposite in direction to results obtained for individual samples. For temperature data (Table 9.10) the effect of confounding is one of “magnitude”: the finding for the combined sample is in the same direction (indicating increase over time) as the finding for the nonstationary series, but the model for the combined sample misestimates the magnitude of the effect. For any month, compared to the nonstationary series, the model for the combined sample has intercept and slope coefficients with lower absolute values: models for the combined data thus underestimate the rate of change in temperature for the nonstationary series. *If Simpson’s Paradox confounds fundamental data, then models using those confounded data also are confounded.*

Measuring Atmospheric Circulation Patterns

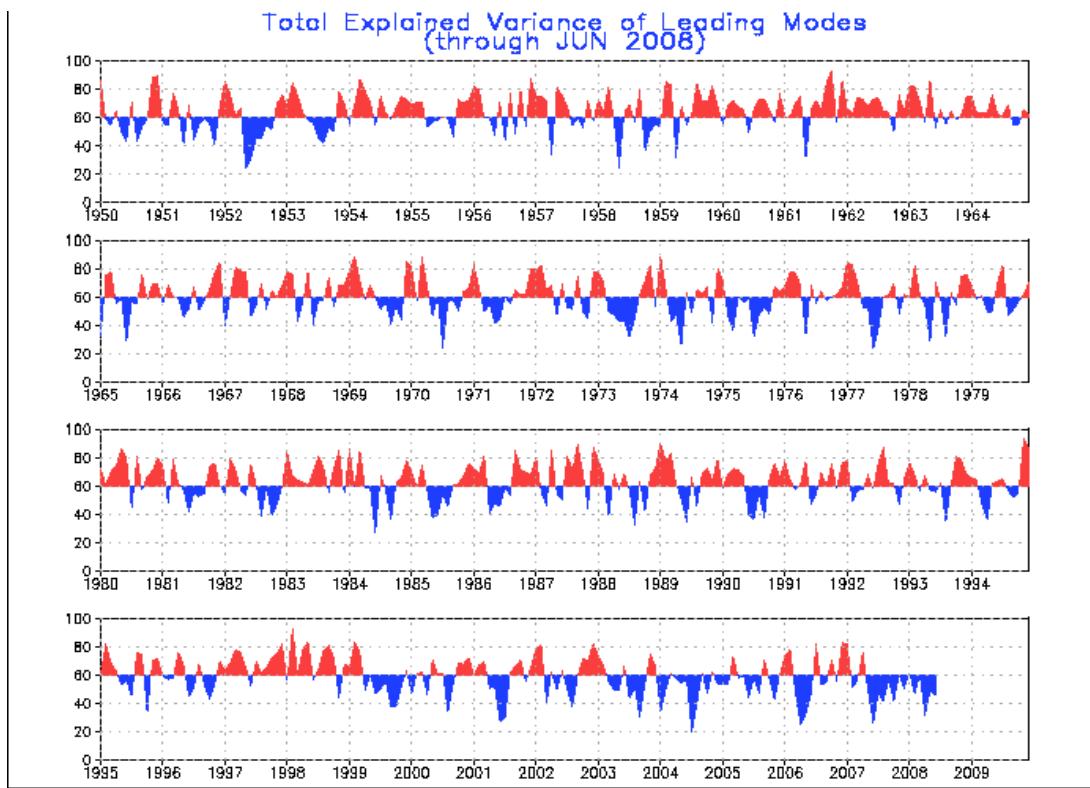
Seminal research conducted by Barnston and Livezey⁴¹ used orthogonally rotated principal components analysis (PCA) of monthly mean 700 mb geopotential heights to identify the major modes of northern hemisphere upper-air variability. They used combined data from the years 1950 through 1984: measurements were taken on a 358-point grid covering latitudes from 20°N to 85°N, and ten “robust” modes (components) were identified which persisted throughout the year. The Climate Prediction Center (CPC) performed a similar analysis of northern hemisphere 500 mb heights using data from 1950 to 2000: ten modes were identified and used to compute the values of the teleconnection indices.⁴² Table 9.11 gives the ten modes of upper-air variability determined by the CPC analysis.

Table 9.11: Ten Modes of Upper-Air Variability Determined by the CPC Analysis

CPC Mode	Abbreviation	Description
1	NAO	North Atlantic Oscillation
2	EA	East Atlantic Pattern
3	WP	West Pacific Pattern
4	EP/NP	East Pacific / North Pacific Pattern
5	PNA	Pacific / North American Pattern
6	EA/WR	East Atlantic/West Russia Pattern
7	SCA	Scandinavia Pattern
8	TNH	Tropical / Northern Hemisphere Pattern
9	POL	Polar/ Eurasia Pattern
10	PT	Pacific Transition Pattern

Figure 9.25 gives the total variance in 500 mb height data that is explained by these ten modes each year. In the Figure, blue shading indicates levels of explained variation that fall below the mean. In 2003 the combined sample includes an equal number of data points from stationary (1950-1976) and nonstationary (1977-2003) series, but data from the nonstationary series dominate the combined sample by 2004. Extrapolation of earlier results suggests that increasing domination will accelerate paradoxical confounding and resulting underestimation of magnitude of effect. Note that after 2003, *performance of the quantitative model used to identify major modes of northern hemisphere upper-air variability has never been lower.*

Figure 9.25: Variance in 500mb Height Data Explained by 10 CPC Modes, by Year



It is straightforward to show that this accelerating failure of the current state-of-the-art is in part attributable to paradoxical confounding. We obtained January 500 mb geopotential height data from 1948-2007 from the NCEP/NCAR Reanalysis dataset, for the full 379-point grid used in research cited earlier, separating the data into stationary (1948-1976) versus nonstationary (1977-2007) series.⁴³ We replicated prior varimax-rotated, ten-extracted-factor PCA of 500 mb height data (Table 9.12). The principal component column indicates successive eigenvector (mode). For Sample, S is the stationary series, NS the non-stationary series, and C the combined S and NS data. Eigenvalue is given for each sample and mode, as is corresponding percent of total variance explained by the mode. For example, the first mode for the stationary series had an eigenvalue of 68.1, thus explaining 18.0% of the total variance of 379 measurements of 500 mb heights. Indicated using red, paradoxical confounding exists when the eigenvalue for the C sample falls outside of the domain defined by the S and NS samples. Note that 80% of the modes clearly reveal paradoxical confounding: in every case except mode number 2 the effect was underestimation of explained variation.

Table 9.12 also provides the *cumulative* percent of total variance (of 379 variables) explained by the modes for each sample, across successive modes. Indicated using blue, paradoxical confounding exists when the cumulative value of this performance index for the C sample falls outside of the domain defined by the S and NS samples. All factors clearly reveal paradoxical confounding, and the effect was always underestimation of explained variation.

Table 9.12: Replication of Prior Analysis of January 500 mb Geopotential Height Data, Separately by Series

Principal Component	Sample	Eigenvalue	Percent of Variance	Cumulative Percent Variance
1	S	68.1	18.0	18.0
	NS	75.3	19.9	19.9
	C	63.3	16.7	16.7
2	S	58.0	15.3	33.3
	NS	50.2	13.3	33.1
	C	60.0	15.8	32.5
3	S	42.0	11.1	44.4
	NS	39.1	10.3	43.4
	C	32.4	8.6	41.1
4	S	37.4	9.9	54.2
	NS	34.2	9.0	52.5
	C	29.5	7.8	48.9
5	S	24.8	6.5	60.8
	NS	27.3	7.2	59.7
	C	27.0	7.1	56.0
6	S	23.9	6.3	67.1
	NS	22.7	6.0	65.7
	C	21.0	5.5	61.5
7	S	18.6	4.9	72.0
	NS	19.6	5.2	70.8
	C	18.1	4.8	66.3
8	S	16.1	4.2	76.2
	NS	15.4	4.1	74.9
	C	13.4	3.5	69.8
9	S	13.7	3.6	79.8
	NS	15.3	4.0	78.9
	C	12.5	3.3	73.1
10	S	13.2	3.5	83.3
	NS	11.0	2.9	81.8
	C	11.4	3.0	76.2

In addition to examining the omnibus performance results of the current ten-mode solution, it is instructive to examine the internal measurement properties of the individual modes. If the structure underlying the modes (reflected by the relationship of the 379 measurements of 500 mb heights to the mode score) is parallel, then the mode scores for the S, NS and C samples will be internally consistent (i.e., measure the same underlying construct), and a one-factor PCA of the three mode scores should explain most of the variation (theoretical maximum = 100%), coefficient *Alpha* (positively related to the mean item-total correlation and the number of measures in the index) for the resulting factor score should be high (theoretical maximum = 1.0), and the root-mean-squared-residual, or RMSR (an index of the average error in estimating the actual inter-measure correlation based on the mode structure) of the resulting factor score should be low (theoretical minimum = 0). Given in Table 9.13, the ten confounded current

modes have poor internal measurement properties even by social science standards—for example, for personality surveys with modes measured using a fraction as many measures.³

Table 9.13: Internal Measurement Properties of Ten CPC Modes

Principal Component	Eigenvalue	Percent of Variance	Alpha	RMSR
1	1.89	63.3	0.710	0.2772
2	1.82	60.5	0.674	0.2913
3	2.22	74.1	0.825	0.1749
4	1.71	57.1	0.625	0.2744
5	1.54	51.4	0.527	0.2771
6	1.42	47.2	0.440	0.1812
7	1.45	48.5	0.469	0.3011
8	1.96	65.2	0.734	0.1805
9	1.63	54.2	0.577	0.2293
10	1.56	52.0	0.539	0.2404

Empirical results clearly demonstrate that current state-of-the-art models of modes of northern hemisphere upper-air variability are confounded by Simpson's paradox, underestimate model performance and phenomenon effect strength, and produce modes having poor measurement properties. Because data for only one month were used in this demonstration, these analyses represent a “best case scenario.” Prior research first smoothed data over successive three month periods prior to conducting PCA: because the reliability of a composite exceeds the reliability of the constituents, smoothed scores will result in lower volatility (i.e., less extreme outliers) and weaker inter-measure correlations, eigenvalues, and measurement properties.

Theoretical consideration of current state-of-the-art models of modes also is not compelling. First, current modes are *non-granular*: postulating that a total of only ten modes underlie northern hemisphere upper-air variability is relatively simplistic compared with complexity underlying many large natural systems. Second, current modes are *nonparsimonious*, because computing an omnibus mode score requires (in the scoring formula) the use of all geopotential height measures. Third, low parsimony makes current mode scores robust: because many constituents (grid locations) are included in the scoring formula, positive changes in some constituents are offset by negative changes in others, so mode scores are *insensitive*. Finally, by formulation PCA is designed to produce *linear* models (modes), yet the present results failed to reveal strong linear modes as indicated by modest eigenvalues: there is therefore discordance between methodology (PCA), data (paradoxically confounded), method (how PCA was conducted), and objective (identifying psychometrically sound measures of major modes of northern hemisphere upper-air variability).

Unconfounded Measurement of Major Modes

Theoretical and empirical limitations of the original solution motivated development of a new methodology for identifying superior modes, which eliminates problems discussed earlier. Our method, a search algorithm, constitutes a theoretical shift in the way teleconnections are conceptualized. The theoretical shift necessitates an *ipsative* standardization of geopotential height data prior to conducting PCA.⁴⁴ The application of our algorithm involved searching for homogeneous spatial areas within which geopotential height measurements are highly related. Constraints included that independent application of PCA to the S, NS and C samples yields comparable, excellent macroperformance (strong eigenvalues) and internal measurement properties across samples, and that mode constituents are physically contiguous. Manually applied to January data the algorithm yielded 46 new modes summarized in Table 9.14 (labels used are nominal placeholders), ordered by percent of variance explained (i.e., decreasing linearity) for the sta-

tionary sample. For Sample, S = stationary, NS = nonstationary, and C = combined S and NS data. M is the number of geopotential height measures (grid locations) constituting the mode. Eigen indicates the eigenvalue of the mode for a one-factor PCA solution, and Var is the associated variance explained (100% \times Eigen / M). The theoretical upper-bound for internal consistency is Alpha = 1, and the theoretical lower-bound for root-mean-square-error is RMSR = 0. Finally, cumulative total eigenvalue, number of height measures, and total variance explained are also provided across successive modes.

Table 9.14: Principal Components Analysis of Unconfounded January 500 mb
Geopotential Height Data, Separately by Series

Mode	Sample	M	Eigen	Var	Alpha	RMSR	Cumulative Totals		
							Eigen	M	Var
J	S	3	2.866	95.5	.977	.0331	2.866	3	95.5
	NS		2.831	94.4	.970	.0386	2.831		94.4
	C		2.844	94.8	.973	.0364	2.844		94.8
H	S	3	2.840	94.7	.972	.0412	5.706	6	95.1
	NS		2.819	94.0	.968	.0471	5.650		94.2
	C		2.827	94.2	.969	.0445	5.671		94.5
PP	S	3	2.826	94.2	.969	.0360	8.532	9	94.8
	NS		2.641	88.0	.932	.0685	8.291		92.1
	C		2.761	92.0	.957	.0476	8.432		93.7
MM	S	3	2.803	93.4	.965	.0337	11.335	12	94.5
	NS		2.743	91.4	.953	.0433	11.034		92.0
	C		2.773	92.4	.959	.0380	11.205		93.4
P	S	4	3.731	93.3	.976	.0404	15.066	16	94.2
	NS		3.575	89.4	.960	.0608	14.609		91.3
	C		3.651	91.3	.968	.0499	14.856		92.8
L	S	3	2.795	93.2	.963	.0558	17.861	19	94.0
	NS		2.729	91.0	.950	.0735	17.338		91.3
	C		2.790	93.0	.962	.0568	17.646		92.9
NN	S	3	2.793	93.1	.963	.0406	20.654	22	93.9
	NS		2.676	89.2	.939	.0562	20.014		91.0
	C		2.748	91.6	.954	.0464	20.394		92.7
M	S	4	3.724	93.1	.975	.0416	24.378	26	93.8
	NS		3.551	88.8	.958	.0603	23.565		90.6
	C		3.604	90.1	.963	.0575	23.998		92.3
Q	S	3	2.789	93.0	.962	.0541	27.167	29	93.7
	NS		2.613	87.1	.926	.0992	26.178		90.3
	C		2.707	90.2	.946	.0750	26.705		92.1
YY	S	3	2.788	92.9	.962	.0411	29.955	32	93.6
	NS		2.663	88.8	.937	.0566	28.841		90.1
	C		2.729	91.0	.950	.0474	29.434		92.0
I	S	3	2.785	92.8	.961	.0511	32.740	35	93.5
	NS		2.725	90.8	.950	.0720	31.566		90.2
	C		2.755	91.8	.955	.0612	32.189		92.0
CC	S	3	2.775	92.5	.960	.0492	35.515	38	93.5
	NS		2.653	88.4	.935	.0677	34.219		90.1
	C		2.717	90.6	.948	.0577	34.906		91.9

G	S	3	2.773	92.5	.959	.0586	38.288	41	93.4
	NS		2.802	93.4	.965	.0540	37.021		90.3
	C		2.788	92.9	.962	.0563	37.694		91.9
K	S	3	2.773	92.4	.959	.0561	41.061	44	93.3
	NS		2.672	89.1	.939	.0875	39.693		90.2
	C		2.703	90.1	.945	.0764	40.397		91.8
JJ	S	6	5.544	92.4	.984	.0348	46.605	50	93.2
	NS		5.236	87.3	.971	.0685	44.929		90.0
	C		5.360	89.3	.976	.0568	45.757		91.5
WW	S	3	2.770	92.3	.959	.0547	49.375	53	93.2
	NS		2.675	89.2	.939	.0581	47.604		89.8
	C		2.722	90.7	.949	.0483	48.479		91.5
R	S	3	2.769	92.3	.958	.0617	52.144	56	93.1
	NS		2.869	95.6	.977	.0358	50.473		90.1
	C		2.843	94.8	.972	.0422	51.322		91.6
O	S	3	2.764	92.1	.957	.0646	54.908	59	93.1
	NS		2.864	95.5	.976	.0373	53.337		90.4
	C		2.828	94.2	.970	.0468	54.150		91.8
XX	S	3	2.763	92.1	.957	.0453	57.671	62	93.0
	NS		2.730	91.0	.951	.0498	56.067		90.4
	C		2.744	91.5	.953	.0474	56.894		91.8
T	S	3	2.756	91.9	.956	.0613	60.427	65	93.0
	NS		2.694	89.8	.943	.0801	58.761		90.4
	C		2.715	90.5	.948	.0731	59.609		91.7
F	S	5	4.585	91.7	.977	.0437	65.012	70	92.9
	NS		4.393	87.9	.965	.0742	63.154		90.2
	C		4.471	89.4	.970	.0612	64.080		91.5
EE	S	3	2.749	91.6	.954	.0426	67.761	73	92.8
	NS		2.529	84.3	.907	.0898	65.683		90.0
	C		2.658	88.6	.936	.0608	66.738		91.4
2	S	3	2.743	91.4	.953	.0609	70.504	76	92.8
	NS		2.599	86.6	.923	.0844	68.282		89.8
	C		2.627	87.6	.929	.0824	69.365		91.3
B	S	6	5.472	91.2	.981	.0535	75.976	82	92.7
	NS		5.352	89.2	.976	.0773	73.634		90.0
	C		5.399	90.0	.978	.0654	74.764		91.2
ZZ	S	3	2.727	90.9	.950	.0464	78.703	85	92.6
	NS		2.787	92.9	.962	.0422	76.421		89.9
	C		2.738	91.3	.952	.0450	77.502		91.2
E	S	4	3.634	90.8	.966	.0511	82.337	89	92.5
	NS		3.526	88.2	.955	.0697	79.947		89.8
	C		3.567	89.2	.960	.0606	81.069		91.1
RR	S	3	2.723	90.8	.949	.0555	85.060	92	92.5
	NS		2.611	87.0	.925	.0813	82.558		89.7
	C		2.658	88.6	.936	.0694	83.727		91.0
D	S	3	2.721	90.7	.949	.0782	87.781	95	92.4
	NS		2.807	93.6	.966	.0521	85.365		89.9
	C		2.724	90.8	.949	.0758	86.451		91.0

C	S	4	3.605	90.1	.964	.0566	91.386	99	92.3
	NS		3.667	91.7	.970	.0500	89.032		89.9
	C		3.637	90.9	.967	.0537	90.088		91.0
U	S	3	2.703	90.1	.945	.0648	94.089	102	92.2
	NS		2.746	91.5	.954	.0624	91.778		90.0
	C		2.727	90.9	.950	.0631	92.815		91.0
LL	S	3	2.695	89.8	.943	.0603	96.784	105	92.2
	NS		2.599	86.6	.923	.0832	94.377		90.0
	C		2.680	89.3	.940	.0646	95.495		90.9
TT	S	3	2.687	89.6	.942	.0565	99.471	108	92.1
	NS		2.840	94.7	.972	.0271	97.217		90.0
	C		2.780	93.3	.964	.0345	98.275		91.0
V	S	3	2.687	89.6	.942	.0845	102.158	111	92.0
	NS		2.659	88.6	.936	.0922	99.876		90.0
	C		2.662	88.7	.937	.0914	100.937		90.9
HH	S	3	2.683	89.4	.941	.0567	104.841	114	92.0
	NS		2.567	85.6	.916	.0994	102.443		89.9
	C		2.615	87.2	.926	.0797	103.552		90.8
UU	S	3	2.681	89.4	.941	.0536	107.522	117	91.9
	NS		2.638	87.9	.931	.0757	105.081		89.8
	C		2.667	88.9	.938	.0623	106.219		90.8
GG	S	3	2.675	89.2	.939	.0627	110.197	120	91.8
	NS		2.723	90.8	.949	.0540	107.804		89.8
	C		2.714	90.5	.947	.0525	108.933		90.8
1	S	3	2.673	89.1	.939	.0603	112.870	123	91.8
	NS		2.771	92.4	.959	.0438	110.575		89.9
	C		2.747	91.6	.954	.0473	111.680		90.8
II	S	3	2.672	89.1	.939	.0578	115.542	126	91.7
	NS		2.745	91.5	.954	.0427	113.320		89.9
	C		2.706	90.2	.946	.0502	114.386		90.8
DD	S	4	3.562	89.1	.959	.0616	119.104	130	91.6
	NS		3.540	88.5	.957	.0588	116.860		89.9
	C		3.547	88.7	.957	.0595	117.933		90.7
VV	S	3	2.656	88.5	.935	.0715	121.760	133	91.5
	NS		2.728	90.9	.950	.0479	119.588		89.9
	C		2.717	90.6	.948	.0547	120.650		90.7
Y	S	3	2.652	88.4	.934	.0941	124.412	136	91.5
	NS		2.791	93.0	.962	.0555	122.379		90.0
	C		2.680	89.3	.940	.0864	123.330		90.7
3	S	4	3.530	88.3	.956	.0733	127.942	140	91.4
	NS		3.623	90.6	.965	.0539	126.002		90.0
	C		3.559	89.0	.959	.0671	126.889		90.6
FF	S	3	2.646	88.2	.933	.0987	130.588	143	91.3
	NS		2.701	90.0	.945	.0835	128.703		90.0
	C		2.649	88.3	.934	.0972	129.538		90.6
A	S	5	4.357	87.1	.963	.0777	134.945	148	91.2
	NS		4.538	90.8	.975	.0706	133.241		90.0
	C		4.414	88.3	.967	.0770	133.952		90.5

SS	S	3	2.603	86.8	.924	.0778	137.548	151	91.1
	NS		2.755	91.8	.955	.0471	135.996		90.1
	C		2.652	88.4	.934	.0676	136.604		90.5
BB	S	4	3.473	86.8	.949	.0656	141.021	155	91.0
	NS		3.612	90.3	.964	.0566	139.608		90.1
	C		3.514	87.8	.954	.0645	140.118		90.4

There is no evidence of paradoxical confounding (performance results for C always fall between results for S and NS), and the percentage of variance explained, *Alpha*, and RMSR meet psychometric criteria for “good to excellent” fit for exploratory PCA models.³ We also examined internal measurement properties of the individual modes via one-factor PCA of the three sample scores (S, NS, C), and analysis revealed virtually perfect measurement: for every mode, percent of total variance (of M measures) explained > 99.9%; *Alpha* > 0.99, and RMSR ≤ 0.0002 . We attempted to model the original ten modes using the new 46 modes, and vice versa, using multiple regression analysis, but no satisfactory models were identified: the original ten modes and the new 46 modes are *not* related to each other.

Considered together these findings clearly show that the 46 new and unique modes eliminate every *empirical* problem identified for the original ten modes: there is no evidence of Simpson’s paradox (S and NS data may be combined without inducing confounding); model performance and phenomenon effect strength are not erroneously misestimated (estimates from all samples are convergent); and mode scores exhibit ideal measurement properties. The new modes also address all *theoretical* concerns identified for the original ten modes: granularity increased 4.6-fold; the new modes are parsimonious (factor weighting coefficients are all approximately one in absolute magnitude, each grid location appears on only one mode); mode scores are sensitive (composed of six or fewer strongly related grid locations, small changes in geopotential heights are easily detectable); and the modes are extremely well-modeled by PCA, representing a set of nearly perfectly linear measures.

Qualitative Interpretation of Ipsative Modes

Figure 9.26 locates the ipsative modes on a polar projection map of the northern hemisphere. The PCA-derived CPC modes of upper-air variability listed in Table 9.14 are consistent with the modes identified in the original principal components analysis⁴¹ of 700 mb height data, and have counterparts in the ipsative modes developed presently.

The first mode, North Atlantic Oscillation (NAO), had strong positive coefficients for grid points over Greenland, corresponding to ipsative mode U. NAO also had strong negative coefficients for grid points in the North Atlantic, west of the Azores (ipsative mode VV); Manchuria (ipsative mode H); and the central plains of the US (between ipsative factors EE and 1).

The second mode, East Atlantic Pattern (EA), had strong positive coefficients for grid points over North Africa (ipsative mode DD), and in the Atlantic east of Cuba (ipsative mode F). EA also had strong negative coefficients for grid points in the North Atlantic, east of Labrador and south of Greenland (ipsative mode FF).

The West Pacific Pattern (WP) had strong positive coefficients for grid points in the Philippine Sea (ipsative mode D), and strong negative coefficients for grid points east of Kamchatka (ipsative mode ZZ).

The East Pacific/North Pacific Pattern (EP/NP) had strong positive coefficients for grid points over southeast Alaska (between ipsative modes GG and 2). EP/NP also had strong negative coefficients for grid points in the North Pacific south of the Aleutian Islands (ipsative mode TT), and near James Bay in Canada (ipsative mode M).

The Pacific/North American Pattern (PNA) had strong positive coefficients for grid points west of Hawaii (ipsative mode A), and in the Pacific Northwest of the US (ipsative mode LL). PNA also had strong negative coefficients for grid points in the North Pacific southwest of the Aleutian Islands (ipsative mode O), and over the southeast US (ipsative mode EE).

The East Atlantic/West Russia Pattern (EA/WR) had strong positive coefficients for grid points near England (between ipsative factors II and UU), and in Siberia north of Manchuria (ipsative mode G). EA/WR had strong negative coefficients for grid points northeast of the Caspian Sea (ipsative mode JJ).

The Scandinavian Pattern (SCA) had strong positive coefficients for grid points in Central Russia (between ipsative modes G and P), and in the North Atlantic, northwest of Spain (ipsative mode WW). SCA also had strong negative coefficients for grid points near Finland (between ipsative modes XX and JJ).

The Tropical/Northern Hemisphere Pattern (TNH) had strong positive coefficients for grid points in the North Pacific west of the Pacific Northwest of the US (ipsative mode SS), and near the Bahamas (ipsative mode MM). TNH also had strong negative coefficients for grid points near James Bay in Canada (ipsative mode M).

The Polar/Eurasia Pattern (POL) had strong positive coefficients for grid points in eastern Mongolia (near ipsative modes G and H), and strong negative coefficients for grid points in the Arctic Ocean north of eastern Siberia (ipsative mode HH).

Finally, the Pacific Transition Pattern (PT)—which did not materialize in either of the original PCA for the month of January, had for the month of September strong positive coefficients for grid points over the northern plains of the US (ipsative mode 1), and west of Hawaii (ipsative mode A). PT also had strong negative coefficients for grid points in the North Pacific south of Alaska (ipsative mode C), and over the eastern US (ipsative mode V).

Figure 9.26: Polar projection Map of the Ipsative Modes



Predicting Temperature Anomalies

To determine whether predictive validity is augmented by nonconfounded measurement, we assessed if statistical models using the 46 newly discovered (*vs.* original ten) modes of northern hemisphere upper-air variability produce more accurate temperature forecasting. We used weighted CTA to predict if the mean temperature in January, February, and March fell above or below the median temperature for the years 1950 - 2007, for 48 contiguous US states. Weights were determined by sorting the observations by monthly mean temperature, and adding 1.5 for every position above or below the median. *WESS* is a standardized measure of weighted *ESS*, on which 0 is the level of weighted predictive accuracy expected by chance, and 100 represents errorless (perfect) weighted predictive accuracy. Weighted CTA was performed using three sets of attributes: *ipsative modes* (46 modes discovered presently); *published normative modes* obtained from the CPC, with PT omitted due to inactivity in January; and *computed normative modes* obtained from our replication of the CPC analysis using only January data.

The findings of these analyses are summarized in Table 9.15 (see Appendix E). Tabled are modes (see Table 9.14 for coding) emerging with $p < 0.05$ in the weighted CTA model. A dash (-) indicates that no solution was identified having $p < 0.05$ for any mode; a missing row indicates no solution was identified for any data type (ipsative, published, or computed); and an asterisk (*) indicates that results obtained for the indicated modes were identical to findings for the ipsative modes. Models derived using the ipsative modes to predict temperature anomalies in the United States convincingly and broadly outperformed the corresponding models derived with normative modes when considered from the perspective of predictive accuracy, and quantified using the WESS index:

- For a given state and month (corresponding to individual rows in Table 9.15), the ipsative mode model yielded the greatest WESS 117 times (91.4%), versus 5 and 6 (3.9% and 4.7%) times for published and computed normative mode models, respectively.
- In January the ipsative mode models always achieved greater WESS than the corresponding normative mode models. In February the ipsative mode models almost always (93.2% of the time) achieved greatest WESS (44 states had models based on February data), and even as the data aged substantially—for March, ipsative models usually (78.1% of the time) achieved greatest WESS (32 states had models using March data).
- For January data, using ipsative modes, all 48 states had CTA models with $\text{WESS} \geq 90\%$, versus two states with CTA models involving published normative modes, and one state CTA model involving computed normative modes. For February data, using ipsative modes, a dozen states had CTA models with $\text{WESS} \geq 90\%$ (and three for March data), versus none for normative modes.
- We statistically contrasted the WESS of each pair of these three sets of factors. If no model was found, WESS was assumed to be zero. ODA was used to determine which set of modes was better at predicting whether or not the mean temperature of the states exceeded the median. The PTMP procedure⁴⁵ was used to estimate the exact Type I error of each contrast. Analyses indicated that ipsative mode models had significantly greater WESS than the published or computed normative mode models for all three months (p 's < 0.0001), and that normative models could never reliably be discriminated from each other by WESS (p 's > 0.17).
- As a test of cross-sample generalizability we also evaluated a larger field of northern hemisphere data. In the crutem3v dataset are 217 locations which have no missing data for January, February or March, for the years 1948 - 2007. As a test of cross-method generalizability, temperature predictions for each location and month were obtained using stepwise multiple regression analysis: the independent variables were the January data, and ipsative, published raw, or computed raw modes were used as dependent variables. The R^2 value for each model was determined: if no model was found, R^2 was assumed to be zero. Statistical comparison via the PTMP procedure showed that ipsative modes clearly outperformed the other modes (p 's < 0.0001). Computed raw

modes outperformed published raw modes in all cases: contrasts were statistically significant for January and February (p 's < 0.0001), but not March (p < 0.27).

Predicting Precipitation Anomalies

As a second investigation of predictive validity we assessed if the statistical models using ipsative modes produce more accurate precipitation forecasting. We used weighted CTA to predict if mean precipitation in January, February, and March fell above or below the median precipitation for the years 1950-2007, for 48 contiguous US states. As for temperature modeling, the weighted CTA algorithm was performed using three sets of attributes: the 46 newly discovered ipsative modes; published normative modes (obtained from the CPC, with PT omitted due to inactivity in January); and computed normative modes (obtained from our replication of CPC analysis using only January data). Findings of these analyses are summarized in Table 9.16 (see Appendix E). Tabled are modes emerging with p < 0.05 in the weighted CTA model. The weights were determined by the same method as was used in predicting temperature anomalies, but total monthly precipitation was used for the sort and median.

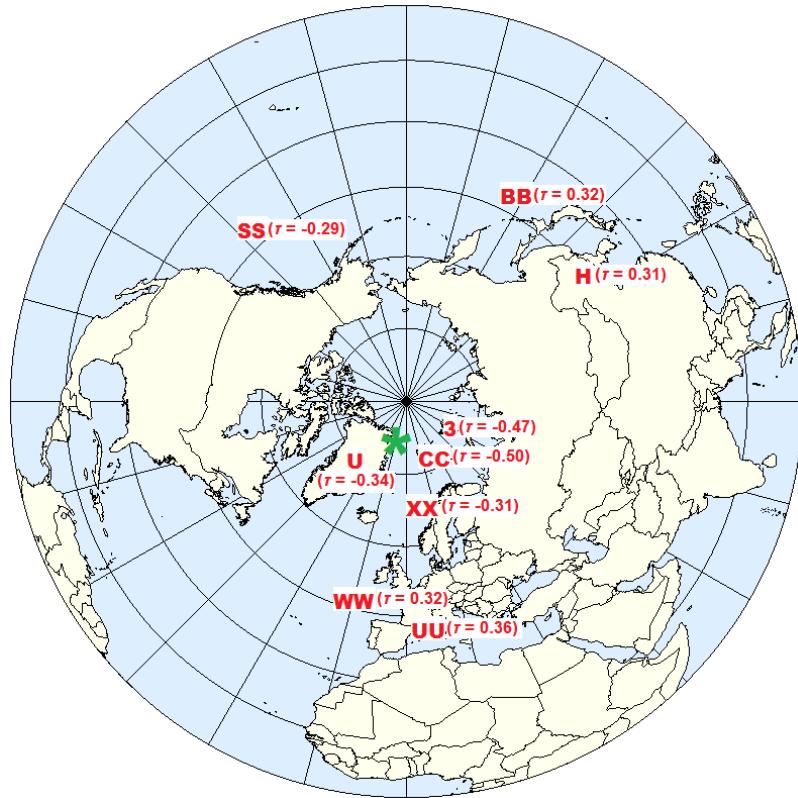
As when modeling temperature anomalies, models derived using ipsative modes to predict precipitation anomalies in the United States convincingly and broadly outperformed the corresponding models derived by normative modes, when considered from the perspective of predictive accuracy:

- For a given state and month (corresponding to individual rows in Table 9.16), the ipsative mode model yielded the greatest WESS 126 times (92.6%), versus 5 (3.7%) times each for the published and computed normative mode models.
- In January, ipsative mode models achieved greater WESS than corresponding normative mode models 91.3% of the time (46 states had models based on January data). Similarly, in February the ipsative mode models almost always (93.3% of the time) achieved greatest WESS (45 states had models based on February data), and even as data aged substantially—in March, ipsative models almost always (93.5% of the time) achieved greatest WESS (46 states had models based on March data).
- Using ipsative modes, for January data 12 states had CTA models with $\text{WESS} \geq 90\%$, as did 6 states for February data and 4 states for March data. Zero normative mode models achieved this level of WESS in any month modeled.
- We statistically contrasted the WESS of each pair of these three sets of modes. If no model was found, then WESS was assumed to be zero. We used ODA to determine which set of modes was better at predicting whether the mean precipitation of the states exceeded the median, or not. The PTMP procedure was used to estimate the exact Type I error for each contrast. Analyses of January data (March and February had comparatively sparse data) indicated that the ipsative mode model had significantly greater WESS than the normative mode models (p 's < 0.0002), but computed and published raw modes were indiscriminable (p < 0.15).

Predicting Export of Arctic Sea Ice

The export of Arctic sea ice through the Fram Strait off northeast Greenland is an important factor in the freshwater balance of the North Atlantic Ocean, and affects the North Atlantic thermohaline circulation. The January monthly ice export at fluxgate a of the Fram Strait⁴⁶ was studied using the ipsative modes. Data consisted of sea ice area flux for the years 1979 - 2002. Kendall's τ_{ab} statistic was used to determine the correlation of modes with ice export, and the significant associations are shown in Figure 9.27. Negative associations were found with ipsative modes U (over Greenland), CC (near Svalbard), 3 (near Franz Josef Land), XX (off the coast of northern Norway), and SS (eastern Pacific Ocean). Positive associations were found with ipsative modes UU (Mediterranean Sea south of France), WW (North Atlantic Ocean northwest of Spain), H (over Manchuria), and BB (east of Japan).

Figure 9.27: Ipsative modes and Kendall's *Tau b* Coefficients with Statistically Significant ($p < 0.05$) Associations with Ice Export at Fram Strait Fluxgate a , Indicated as *

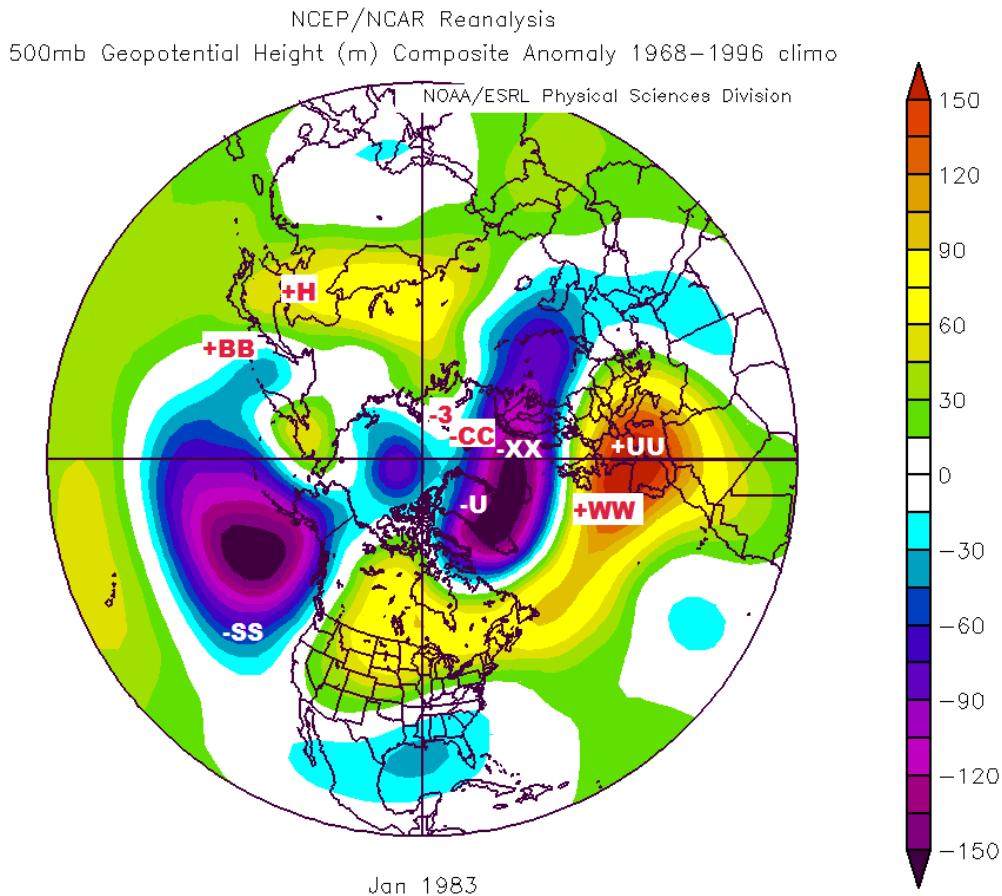


An example of a pattern with high sea ice export is illustrated in Figure 9.28. The 500 mb pattern in January 1983 yielded the maximal ice export for any January in the years of 1979 - 2002. Low 500 mb heights extend from Greenland to Scandinavia and western Russia, and another area of low heights is found off of the Pacific coast of the USA. Areas of high 500 mb heights are seen over southwest Europe and the western Mediterranean Sea, and over Mongolia and northeast China.

Recent research⁴⁷ reported no correlation between SLP-based NAO and Arctic wintertime sea ice export over 1958 - 1977, and a positive correlation of 0.7 over 1978 - 1997. An eastern shift in the NAO centers of variability was suggested to explain this phenomenon. However, for the 500 mb level, ipsative mode U was a stable center over Greenland, for both sets of years, 1948 - 1976 and 1977 - 2007. Mode U represents the northern center of the NAO dipole at the 500 mb level. Mode II (near Iceland) was also a stable center, coincident with the northern center of surface-level winter NAO variability: this does not support the idea of a shift at 500 mb. Furthermore, factors XX, CC and 3, located in this region, were stable in both eras and reliably associated with sea ice movement. Mode 3 is coincident with the surface center of variability in the Kara Sea, previously found to be associated with sea ice export variability.⁴⁸

Preliminary results using unconfounded climatic data in atmospheric prediction are very positive. An important extension of the present research is obtaining GHA modes for all months of the year. Further evaluation of optimal statistical methods used with unconfounded climatic data is warranted. Future research should use these data in applications such as, for example: predicting the ontogenesis, intensity, and path of hurricanes⁴⁹, and the ontogenesis, intensity, and location of sudden stratospheric warmings^{50,51}; modeling of seasonal energy consumption and management of climate risk for energy firms⁵²; forecasting and understanding the ENSO cycle (El Niño)⁵³, and development and evaluation of numerical weather prediction models.⁵⁴

Figure 9.28: 500 mb GHA for January 1983, which Entailed the Maximal January Ice Export for 1979-2002:
Ipsative modes are Prefixed by the Sign of their Associated Kendall's Tau_b Coefficient



Confounding in Single-Case Series

Consider an application in which a single individual (e.g., medical patient, job trainee, athlete) or “object of investigation” (e.g., financial algorithm, river, musical group) is assessed on one or more attributes on a longitudinal series of measurement (testing) sessions: measurement 1, measurement 2, etcetera, ending at measurement N . Each measurement records the value of the attribute(s): for attribute A the value at the initial measurement is denoted as A_1 , at the i th measurement as A_i , and at the final measurement as A_N . In many applications, such as in the weather forecasting example just presented, the study focus is on predicting the direction and magnitude of measurement-to-measurement changes in A_i values (modeling price changes of equities is another excellent example of such an application). In other applications the study focus is on comparing two or more attributes across the time span investigated.

Comparing Two Serial Ratings

Often at the advice of their physician, patients managing chronic disease such as fibromyalgia or arthritis, or undergoing therapy in rehabilitation medicine or oncology, record weekly, daily, and real-time ratings of their physical (pain), mental (memory) and emotional (depression) symptoms. Similarly, often at the advice of their coaches, athletes in a variety of group and customized fitness programs record ratings of physical (vigor), mental (concentration) and emotional (anxiety) states at the beginning and/or the end of

training sessions. In such applications symptoms and states are typically assessed using an ordered Likert-type scale with between three and eleven possible response categories.

This example illustrates the use of UniODA to compare distributions of symptoms or states using data obtained across multiple measurements in an individual series.⁵⁵ Data were abstracted with permission from a computer log containing 297 sequential entries by an anonymous patient with fibromyalgia (FM) voluntarily participating in the prospective clinical trial of a web-based self-monitoring and symptom management system. On each session users rated their state with respect to the symptoms most often reported by FM patients, including two of the most prevalent FM symptoms—pain and fatigue. Symptoms were rated via an 11-point Likert-Type scale ranging from 0 (“not at all bothersome”) to 10 (“extremely bothersome”).^{56,57} Analysis is conducted to determine if the patient rated these two symptoms similarly: that is, if these two symptoms were experienced with comparable intensity, or if ratings of one symptom exceed or dominate the ratings of the other symptom. Analyses comparing ratings on pain and fatigue are first conducted using raw data, and a second time using data ipsatively standardized into z-scores using mean and standard deviation (SD) computed for the individual’s data.

Analysis of Raw Scores: For *pain*: mean = 4.60; SD = 1.52; median = 5; skewness = 0.20; kurtosis = -0.53; and CV = 33.2. For *fatigue*: mean = 6.38; SD = 1.44; median = 6; skewness = -0.03; kurtosis = -0.59; and CV = 22.6. Table 9.15 provides the pain and fatigue rating distributions for this patient.

Table 9.15: Distributions of the Patient’s Raw Pain and Fatigue Ratings

Response Category	Pain	Fatigue
2	26	
3	48	5
4	74	25
5	65	50
6	49	85
7	27	60
8	7	49
9	1	23

Note: Response categories 0, 1 and 10 weren’t used.

A new data set (*new.txt*) was constructed with $2N$ observations ($2 \times 297 = 594$), each forming a row in *new.txt*. First, all 297 *pain* ratings were copied to *new.txt* (which at this point has 297 rows). Next, beneath these, all 297 *fatigue* ratings are copied (*new.txt* now has 594 rows). The number “0” (used as an arbitrary class category dummy-code) delimited by a space is appended to the beginning of each of the first (top) set of 297 rows, and the number “1” (arbitrary dummy-code) delimited by a space is appended to the beginning of each of the second (bottom) set of 297 rows. Using *new.txt* as input, the exploratory hypothesis that the raw pain and fatigue measures were rated differentially by the patient is tested using the following UniODA and MegaODA software syntax:

OPEN new.txt;	ATTR rating;
OUTPUT example.out;	MCARLO ITER 25000;
VARS class rating;	LOO;
CLASS class;	GO;

The UniODA model was: if rating is ≤ 5 predict that the attribute is pain, otherwise predict the attribute is fatigue ($p < 0.0001$). As seen in Table 9.15, $26 + 48 + 74 + 65 = 213$ (71.7%) of 297 pain ratings were correctly predicted, as were $85 + 60 + 49 + 23 = 217$ (73.1%) of 297 fatigue ratings. Table 9.16 gives the confusion table for the model: classification performance was moderate ($ESS = 44.78$, $ESP = 44.79$), and was stable in LOO analysis.

Table 9.16: Confusion Table for UniODA Model Discriminating Raw Pain and Fatigue Ratings

		Predicted Symptom		
Actual Symptom	Pain	Fatigue		
	213	84	71.7%	
Symptom	Fatigue	80	217	73.1%
		72.7%	72.1%	

In summary the comparison of the serial raw pain and fatigue ratings made by this FM patient suggests that the latter dominate the former –that is, are “more bothersome”, and also suggests that this finding may cross-generalize to another period (series) for this patient in the future.

Analysis of Ipsative z-Scores: Ipsative standardization of raw data into z-scores utilizes the mean and SD computed for the data from an observation (not a sample of observations), and is appropriate for analysis of serial data as a means of eliminating variability attributable to “base-rate” differences between observations that introduce noise into the data (see Chapter 2). Analyses that were performed on the raw data are thus repeated here after the pain and fatigue ratings are ipsatively standardized. For z_{pain} : mean = 0; SD = 1; median = 0.265; skewness = 0.20; kurtosis = -0.53; CV = 0. For z_{fatigue} : mean = 0; SD = 1; median = -0.261; skewness = -0.03; kurtosis = -0.59; CV = 0. Table 9.17 presents z_{pain} and z_{fatigue} rating distributions.

Table 9.17: Distributions of the Patient’s Ipsative z_{pain} and z_{fatigue} Ratings

Response <u>Category</u>	<u>z_{pain}</u>	<u>z_{fatigue}</u>
-2.34		5
-1.70	26	
-1.65		25
-1.05	48	
-0.95		50
-0.39	74	
-0.26		85
0.27	65	
0.43		60
0.92	49	
1.13		49
1.58	27	
1.82		23
2.23	7	
2.89	1	

Comparison of Table 9.15 and Table 9.17 reveals that “identical” rating scales, such as Likert-type scales, are clearly *not* identical if considered statistically from the perspective of the individuals who are using the scales to rate their personal experience. This is in some extent due to ambiguity in the cognitive labels used to give meaning to the numerical options on the scale, ambiguity in the target of the rating, the changing nature of the symptoms over time, and the inherent differences in intensity and variability of rated symptoms. Such complexity is ignored in analysis raw data, but serves as theoretical motivation for the use of “ipsatized” data.^{44,58}

The analysis is conducted as before: construct a new data set (*new.txt*) having $2N$ observations (here, 594), each forming a row in *new.txt*. Copy all 297 z_{pain} ratings to *new.txt* (at this point *new.txt* has 297 rows), then beneath these copy all 297 z_{fatigue} ratings (*new.txt* now has 594 rows). Add the number “0” (an arbitrary class category dummy-code) delimited by a space to the beginning of each of the first

(top) set of 297 rows, and then likewise add the number “1” (arbitrary dummy-code) delimited by a space to the beginning of each of the second (bottom) set of 297 rows. Using new.txt as input, the exploratory hypothesis that z_{pain} and z_{fatigue} were rated differentially is tested by running the same UniODA code used for raw score analysis but making the following substitutions:

```
VARS class zrating;
```

```
ATTR zrating;
```

The UniODA model was: if ipsative z-score ≤ -0.33 (0.33 SD lower than the mean rating for the individual) then predict the attribute is *pain*; otherwise predict *fatigue* ($p < 0.0001$). As found for raw data, UniODA revealed that standardized fatigue ratings dominated standardized pain ratings. Table 9.18 gives the confusion table for this model: classification performance was relatively weak ($ESS = 22.9$, $ESP = 24.2$), and was stable in LOO analysis.

Table 9.18: Confusion Table for UniODA Model Discriminating Ipsative Pain and Fatigue Ratings

		Predicted Symptom		
Actual	z_{pain}	z_{pain}	z_{fatigue}	
		148	149	49.8%
Symptom	z_{fatigue}	80	217	73.1%
		64.9%	59.3%	

A standardized response scale such as a Likert-type scale may be provided to and used by an individual to rate two or more attributes, but that doesn't imply that the scale has the same psychological meaning when applied by the individual to rate the different attributes. Comparison of confusion tables for analysis of raw versus ipsative data presently reveals that the UniODA models correctly classified the identical subset of the highest (z)fatigue ratings: severe fatigue was clearly more bothersome than severe pain for this patient across the period of time studied. Analysis of raw data (reflecting meaning defined by the researcher and conveyed vis-à-vis the specific categorical rating options) revealed that moderate and weak levels of pain can be discriminated from corresponding levels of fatigue. Analysis of the ipsative data (reflecting symptom intensity relative to the experience of the individual) reveals that moderate and weak levels of pain and fatigue can't be discriminated at better than a chance level on the basis of the patient's use of the categorical rating options available on the scale. Because the objective of measurement in the present context is to assess the individual's perceived internal status, and because ipsative data reflect the perspective borne by experience of the individual, attributes should be ipsatively standardized before being compared against each other for an individual, as well as before agglomeration for sample-based analyses (see Chapter 2).

Interactive Measurement

We extend the preceding example examining an *N*-of-1 single-case series of a patient with FM, for a series consisting of 297 sequential daily ratings of a profile of nine common FM symptoms. UniODA is used to identify the *symptom dominance hierarchy* (most to least severe symptom) for this patient. The findings reveal a crucial difference between: (a) the meaning of numbers and labels that constitute the categorical response scale as *defined* and *interpreted* by the *investigator* (raw data), and (b) the meaning of numbers and labels as *perceived* and *applied* by an *individual* to quantify personal internal status (ipsative z-scores). The latter meaning is the purpose of the scale (valid measurement requires that subject perception and response scale interact) and reflects it the motivation underlying measurement.⁵⁹

Analysis of Raw Scores: Distributions and descriptive statistics for raw rating data for the series are presented in Table 9.19 and Table 9.20, respectively. Eyeball examination of Table 9.19 suggests the patient rated different symptoms using different ranges on the response scale, and Table 9.20 shows that the anxiety, depression, and gastrointestinal rating distributions are positively skewed. Symptoms having

many relatively high (more severe) ratings are stiffness, fatigue, and memory issues, and to a lesser extent sleep and concentration issues.

Table 9.19: Raw Score Distributions for Nine Symptoms: *Investigator's Perspective*

<u>Rating</u>	<u>Pain</u>	<u>Stiffness</u>	<u>Fatigue</u>	<u>Concentration</u>	<u>Memory</u>	<u>Anxiety</u>	<u>Depression</u>	<u>GI</u>	<u>Sleep</u>
0						74	97	233	1
1		1		2		173	85	44	4
2	26	11		18		32	43	8	12
3	48	25	5	24	23	11	37	3	61
4	74	56	25	36	48	6	22	5	85
5	65	66	50	69	69	1	8	1	54
6	49	71	85	69	80		2	2	40
7	27	42	60	50	51		2	1	22
8	7	18	49	22	22		1		18
9	1	7	23	7	4				
10									

Note: Tabled are frequencies. GI=gastrointestinal.

Table 9.20: Raw Score Descriptive Statistics for Nine Symptoms: *Investigator's Perspective*

<u>Symptom</u>	<u>Mean</u>	<u>SD</u>	<u>Median</u>	<u>Skewness</u>	<u>Kurtosis</u>	<u>CV</u>	<u>SEM</u>
Pain	4.60	1.52	5	0.20	-0.53	33	0.088
Stiffness	5.32	1.60	5	-0.01	-0.30	30	0.093
Fatigue	6.38	1.44	6	-0.03	-0.59	23	0.084
Concentration	5.39	1.70	5	-0.25	-0.33	32	0.099
Memory	5.57	1.41	6	0.05	-0.56	25	0.082
Anxiety	1.01	0.86	1	1.42	3.32	86	0.050
Depression	1.49	1.55	1	1.17	1.26	104	0.090
Gastrointestinal	0.39	1.03	0	4.34	24.1	267	0.060
Sleep	5.57	1.62	5	0.28	-0.03	29	0.094

Note: $N = 297$. CV = coefficient of variation. SEM = standard error of the mean.

Table 9.21 summarizes results of all possible comparisons between pairs of attributes conducted using UniODA. As seen, a total of 30 of 36 pairwise comparisons met the criterion for experimentwise $p < 0.05$, which is 16.7 times more than are expected by chance. Of these 30 statistically significant effects, 19 (63%) were strong ($ESS > 50$): strongest effects occurred for the comparisons involving anxiety, depression and GI issues.

A symptom dominance hierarchy (SDH) is constructed by mapping all 36 pairwise comparisons. Because the SDH is linear the mapping is like solving a 36-piece puzzle having two “ends” rather than four “corners.” It is a sound strategy when solving this linear puzzle to examine the pairwise comparison table (Table 9.21) for attributes associated with the strongest effects (greatest pairwise differences) all of which are in the same direction (the attribute is always on one side of the inequality). In Table 9.21 the fatigue ratings are significantly greater than all other symptoms, so fatigue forms the high (most severe) end of the SDH. At this point the SDH is mapped as shown.

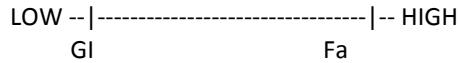
LOW -----|-- HIGH
Fa

Table 9.21: Raw Score Summary of UniODA Comparisons of All Pairs of Patient Symptoms: *Investigator's Perspective*

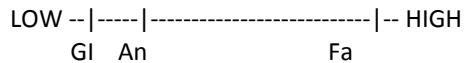
	<u>Stiffness</u>	<u>Fatigue</u>	<u>Concentration</u>	<u>Memory</u>	<u>Anxiety</u>	<u>Depression</u>	<u>Gastrointestinal</u>	<u>Sleep</u>
<u>Pain</u>	St > Pa 18.5, 19.2	Fa > Pa 44.8, 44.8	Co > Pa 22.9, 24.2	Me > Pa 25.9, 27.8	Pa > An 85.2, 85.2	Pa > De 67.0, 68.6	Pa > Ga 93.3, 93.7	SI > Pa 23.6, 25.0
<u>Stiffness</u>		Fa > St 26.6, 27.6	St = Co 4.4, 5.3	St = Me 7.4, 9.3	St > An 89.9, 89.9	St > De 75.8, 75.8	St > Ga 92.9, 93.3	St = SI 6.7, 20.4
<u>Fatigue</u>			Fa > Co 23.2, 24.5	Fa > Me 20.2, 21.7	Fa > An 96.0, 96.0	Fa > De 86.5, 87.4	Fa > Ga 96.0, 96.1	Fa > SI 28.0, 28.9
<u>Concentration</u>				Co = Me 7.1, 17.7	Co > An 87.2, 87.2	Co > De 73.4, 73.5	Co > Ga 92.6, 92.9	Co = SI 9.1, 24.7
<u>Memory</u>					Me > An 93.9, 94.3	Me > De 80.5, 80.6	Me > Ga 96.0, 96.1	Me = SI 7.7, 7.8
<u>Anxiety</u>						De > An 21.9, 27.3	An > Ga 53.5, 53.6	SI > An 92.3, 92.4
<u>Depression</u>							De > Ga 45.8, 46.4	SI > De 82.5, 82.8
<u>Gastrointestinal</u>								SI > Ga 94.3, 94.3

Note: All effects for which an *inequality* is provided have $p < 0.05$ at the experimentwise criterion; effects indicated using an *equality* (=) are *not* statistically significant. ESS and ESP values (the left and right entries in the second row of every table cell, respectively) shaded in grey indicate weak effects; green indicates moderate effects; red indicates strong effects; and blue indicates very strong effects.

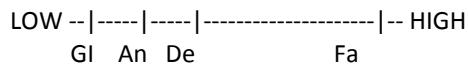
In the pairwise comparisons table with fatigue-related entries eliminated, GI ratings are significantly lower than other ratings, so GI forms the low (least severe) end of the SDH, now mapped as shown.



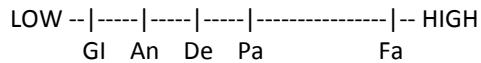
In the pairwise comparisons table with fatigue- and GI-related entries eliminated, anxiety ratings are significantly lower than other ratings: the SDH is now mapped as shown.



Depression is next to follow this pattern: the SDH is now mapped as shown.

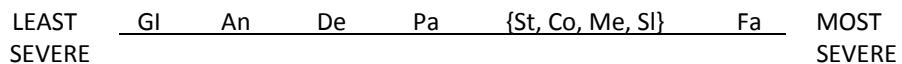


Pain is next to follow this pattern: the SDH is now mapped as shown.



Remaining symptom ratings represented in unexplained cells of the pairwise comparisons table are significantly greater than pain ratings, but comparisons between ratings of remaining symptoms are not statistically significant. The final SDH for raw score data is mapped as illustrated in Figure 9.28.

Figure 9.28: Symptom Dominance Hierarchy for Patient: Raw Score Analysis



From the investigator's perspective an identical rating on the response scale indicates the same level of severity for the patient (and for all patients in sample-based studies), for all rated symptoms. For example, a rating of 4 is taken to indicate the same level of severity if the patient rates stiffness as if the patient rates every other symptom. Seen in this manner fatigue is the patient's dominant, most severe symptom. Significantly less severe than fatigue ratings are a statistically indistinguishable cluster of four symptoms: stiffness, concentration, memory, and sleep issues. Severity ratings for this cluster are significantly greater than pain ratings, which are greater than depression ratings, which are significantly greater than anxiety ratings. GI ratings were least severe.

Analysis of Ipsative z-Scores: The computing formula for a standardized z-score is: $(\text{observation's score} - \text{mean score}) / \text{SD}$. For a *normative* z-score, z_N , mean and SD are based on a sample: conceptually z_N measures the magnitude of an observation's score relative to the population of scores *for all observations*. For an ipsative z-score, z_i , mean and SD are based only on the data from the observation: conceptually z_i measures the magnitude of any observation's score relative to all scores in the population of scores *for the observation*. In this manner z_i expresses data on a scale enabling direct comparison of different single-case series, and eliminates inter-series variability attributable to base-rate differences (noise) in both mean and variance when individual series are combined for use in sample-based studies (see Chapter 2).

Distributions and descriptive statistics for z_i rating data for this patient's series of 297 entries are presented in Table 9.22 (grey shading indicates z-scores less than the mean) and Table 9.23 (Mean = CV = 0, SD = 1, and SEM = 0.058, for all symptoms), respectively.

Table 9.22: Ipsative z-Score Distributions for Nine Symptoms: *Patient's Perspective*

<u>Rating</u>	<u>Percentile</u>	<u>Pain</u>	<u>Stiffness</u>	<u>Fatigue</u>	<u>Concentration</u>	<u>Memory</u>	<u>Anxiety</u>	<u>Depression</u>	<u>GI</u>	<u>Sleep</u>
-3.449	99.97								1	
-2.704	99.66		1							
-2.579	99.51				2					
-2.341	99.04			5						
-2.211	98.65								4	
-2.078	98.12		11							
-1.992	97.68				18					
-1.824	96.59					23				
-1.703	95.57	26								
-1.648	95.03			25						
-1.592	94.43								12	
-1.451	92.66		25							
-1.405	92.00				24					
-1.168	87.86						74			
-1.115	86.76					48				
-1.047	85.24	48								
-0.973	83.47								61	
-0.964	83.25							97		
-0.955	83.02			50						
-0.825	79.53		56							
-0.818	79.33				36					
-0.406	65.76					69				
-0.391	65.21	74								
-0.374	64.58							233		
-0.354	64.95								85	
-0.319	62.51							85		
-0.261	60.30			85						
-0.231	59.13				69					
-0.198	57.85		66							
-0.008	50.32						173			
0.265	39.55	65							54	
0.303	38.09					80				
0.326	37.22							43		
0.356	36.09				69					
0.428	33.43		71							
0.432	33.29			60						
0.593	27.66								44	
0.884	18.84								40	
0.922	17.83	49								
0.943	17.28				50					
0.970	16.60							37		
1.012	15.58					51				
1.055	14.57		42							
1.125	13.03			49						
1.152	12.47						32			
1.503	6.64								22	
1.530	6.30				22					
1.560	5.94								8	
1.577	5.74	27								
1.615	5.32							22		
1.681	4.64		18							
1.721	4.26				22					
1.816	3.47			23						
2.117	1.71				7					
2.121	1.70								18	

2.234	1.27	7		
2.260	1.19			8
2.308	1.05	7		
2.312	1.04			11
2.430	0.75		4	
2.527	0.58			3
2.890	0.19	1		
2.904	0.18			2
3.472	0.03			6
3.494	0.02			5
3.549	0.02			2
4.194	<0.01			1
4.461	<0.01			1
4.632	<0.01			1
5.428	<0.01			2
8.330	<0.01			1

Note: Rating is ipsative z-score. Percentile is the 1-sided percent of normally-distributed z-scores that exceed the tabled value: that is, the percent of ratings (days) which would be associated with a worse symptom rating.

Table 9.23: Ipsative z-Score Descriptive Statistics for Nine Symptoms: *Patient's Perspective*

Symptom	Median	Skewness	Kurtosis
Pain	0.265	0.20	-0.53
Stiffness	-0.198	-0.01	-0.30
Fatigue	-0.261	-0.03	-0.59
Concentration	-0.231	-0.25	-0.33
Memory	0.303	0.05	-0.56
Anxiety	-0.008	1.42	3.32
Depression	-0.319	1.17	1.26
Gastrointestinal	-0.374	4.34	24.1
Sleep	-0.354	0.28	-0.03

Comparing Table 9.19 and Table 9.22 indicates the patient *doesn't* use numbers and labels on the response scale in the same manner across scales as assumed in the investigator's perspective. Instead, the way the patient uses numbers and labels on the response scale to report the personal experience of each symptom differs as a function of the mean and variability of each symptom. The experience being rated interacts with the patient's interpretation of the numbers and labels of the scale, and thus the identical numbers and labels indicate different levels of symptom severity, depending on the symptom being rated.

Data were not normally distributed (raw and z_i data produced identical results): for all nine symptoms, $p < 0.01$ for Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests for normality. Therefore, the percentile of the standard normal distribution having worse (greater) severity ratings—provided in Table 9.22 for every z_i rating, is a poor interpretative heuristic.

Viewed in terms of *relative negative severity*, the best (least severe) symptom reported in the series was one single rating of sleep issues, with $z_i = -3.449$ (Table 9.22). In normally distributed data, a z-score as extremely negative as this is better than for 99.97% of the population of ratings on this symptom *for the patient*. The next least severe ipsative symptom ratings occurred for stiffness and concentration.

Viewed in terms of *relative positive severity*, and in terms of *absolute overall severity*, the worst (most severe) symptom reported in the series was one single rating of GI issues, with $z_i = 8.330$: in normally distributed data, for a z-score as extremely positive as this only $4.4 \times 10^{-14}\%$ of all possible days are worse on this symptom *for the patient*. In addition to eight other extreme positive ipsative GI ratings, next-worst ipsative ratings occurred for anxiety and depression. Results of all possible comparisons between pairs of ipsatively standardized attributes by UniODA are summarized in Table 9.24. x

Table 9.24: z-Score Summary of UniODA Comparisons of All Pairs of Patient Symptoms: *Patient's Perspective*

	<u>Stiffness</u>	<u>Fatigue</u>	<u>Concentration</u>	<u>Memory</u>	<u>Anxiety</u>	<u>Depression</u>	<u>Gastrointestinal</u>	<u>Sleep</u>
<u>Pain</u>	St > Pa* 18.5, 19.2	Fa > Pa 22.9, 24.2	Co > Pa* 22.9, 24.2	Me > Pa 24.6, 25.5	Pa > An 33.3, 37.4	De > Pa 100, 57.1	Ga > Pa 49.8, 66.6	SI > Pa* 23.6, 25.0
<u>Stiffness</u>	St > Fa 24.2, 24.7	St > Co 19.9, 21.4	St > Me 20.5, 22.2	St > An 29.6, 34.2	St > De 30.0, 30.1	St > Ga 47.1, 47.6	St > SI 23.6, 24.0	
<u>Fatigue</u>		Co > Fa 28.6, 29.5	Fa > Me* 20.2, 21.7	An > Fa 30.6, 31.8	Fa > De 34.3, 34.3	Fa > Ga 51.5, 51.7	Fa > SI* 28.0, 28.9	
<u>Concentration</u>			Co > Me 23.9, 25.4	Co > An 33.3, 37.1	Co > De 34.3, 34.8	Co > Ga 51.5, 51.7	Co > SI 28.0, 28.9	
<u>Memory</u>				Me > An 36.0, 39.7	De > Me 23.9, 56.8	Ga > Me 47.1, 65.4	Me > SI 25.9, 27.0	
<u>Anxiety</u>					An > De 36.4, 37.1	An > Ga 53.5, 53.6	An > SI 30.0, 31.2	
<u>Depression</u>						De > Ga* 45.8, 46.4	De > SI 26.3, 57.6	
<u>Gastrointestinal</u>							SI > Ga 52.2, 52.3	

Note: All $p < 0.05$ at the experimentwise criterion. An asterisk (*) marking an inequality indicates that the direction of the effect, and *ESS* and *ESP* obtained using raw and ipsatively standardized data were identical. An inequality indicated in color is an example of paradoxical confounding: red indicates that the opposite effect was obtained for raw data; green shows an effect found here for which no statistically significant model was obtained using raw data; and blue shows an effect with overestimated strength in raw score analysis (more than one type of paradox was noted for some effects, but only the most significant manifestation—such as finding the opposite effect—is reported here). *ESS* and *ESP* values (the left and right entries in second row of every table cell, respectively) shaded in grey are weak effects; green are moderate effects; red are strong effects; and blue are very strong effects.

All 36 pairwise comparisons met the criterion for experimentwise $p < 0.05$: twenty times more than expected by chance. Strongest effects were obtained for comparisons with GI issues, and to a lesser extent, for ratings of sleep and depression issues. Table 9.25 cross-tabulates the ordered *qualitative ESS* and *ESP* levels obtained by UniODA models using raw and z-score data.

Table 9.25: *ESS* and *ESP* of UniODA Models Developed Using Raw and Ipsatively Standardized Data

	ESS		ESP	
	Raw	z-Score	Raw	z-Score
No Effect	6	0	6	0
Weak	6	12	5	9
Moderate	5	19	6	18
Strong	3	4	3	9
Very Strong	16	1	16	0

These data were analyzed using Gen UniODA: performance index (*ESS*, *ESP*) was treated as the Gen (multi-index) variable, data (raw, ipsative) was treated as the class variable, and qualitative effect strength (dummy-coded as Likert-type item with no effect = 1, weak = 2, ..., very strong = 5) was treated as the ordered attribute. The Gen UniODA model was: if effect = very strong then predict raw data were analyzed, otherwise predict z-score data were analyzed (all $p < 0.0002$; $37.5 < ESS < 41.7$; $57.8 < ESP < 60.5$). Overall the model correctly classified almost all of the raw score-based effects (sensitivity = 98.6%), but less than half of the ipsative score-based effects (sensitivity = 41.2%).

Comparison of Table 9.21 and Table 9.24 reveals that of the 36 comparisons of symptom pairs performed for raw data, 28 (78%) were confounded by three different types of paradox when considered from the perspective of ipsative data. As indicated in Table 9.24, ten (28%) raw-score models identified the opposite effect (raw and ipsative data analyses arrived at opposite conclusions); six (17%) raw-score models missed statistically significant effects identified using z-scores; and twelve (33%) effects identified for raw scores over-estimated parallel effects found using z-scores (one moderate effect for raw scores was a weak effect for z-scores; one strong effect for raw scores was a moderate effect using z-scores; three very strong effects for raw scores were strong effects for z-scores; and seven very strong effects identified using raw scores were moderate effects for z-scores). Only six of the raw- and z-score-based models were identical.

SDH structure underlying pairwise differences found for ipsative data is much more complex than corresponding structure observed for raw scores. Analysis begins as before: ratings of stiffness exceeded all other ratings with $p < 0.05$, so at this point the SDH is mapped as shown.

LOW -----|-- HIGH
St

Co ratings exceeded remaining ratings, so the SDH is now mapped as shown.

LOW -----|---|-- HIGH
Co St

Fa ratings exceed all other ratings except for An ratings, so the SDH is now mapped as shown.

LOW -----|---|---|----|-- HIGH
Fa An Co St

At this point the mapping procedure can no longer integrate the remaining inequalities in the pairwise comparisons table: unlike results obtained using raw data for which a linear model was fit, for

ipsative data a multidimensional model is needed to map UniODA pairwise comparison findings (a similar situation was encountered using UniODA to sequentially decompose serial structure in stratigraphic samples that also were expressed as a set of pairs of inequalities identified in optimal Markov analysis⁴). Presently the primary structure identified, illustrated above, represents 21 (58%) of 36 inequalities in the pairwise comparisons table.

The search begins for a second structure that explains the 15 remaining unexplained inequalities in the pairwise-comparison table: two additional models were identified.⁵⁹ Together the three models, summarized in Table 9.26, explain the set of 36 inequalities in the pairwise comparisons table.

Table 9.26: Three Symptom Dominance Hierarchies Identified Using Ipsative z-Scores

Model	Structure of Dominance Hierarchy	Pairwise Comparisons Explained (%)	
		Model	Cumulative
1	Stiffness > Concentration > Anxiety > Fatigue	21 (58%)	21 (58%)
2	GI > Memory > Sleep > Pain > Anxiety	9 (25%)	30 (83%)
3	Anxiety > Depression > GI > Sleep	6 (16%)	36 (100%)

By the ipsative perspective not all ratings made by the patient on nine symptom scales over 297 sequential days can be explained by a single SDH. By the ipsative perspective three SDHs are needed. All three structures identified by z-scores are reminiscent of phase-shift models produced in Markov models.⁴ For example it is plausible that moving from most to least severe in the first model, increased stiffness is followed by greater concentration issues next, which in-turn increase anxiety resulting in greater fatigue.

UniODA model cutpoints for raw and z-score data analyses are given in Table 9.27 for all pairwise comparisons, and Table 9.28 lists different ipsative z-score cutpoints used in UniODA models for each raw-score cutpoint used: the precision and sensitivity of ipsative data cannot be matched by raw scores.

Table 9.27: UniODA Model Cutpoints for Raw (Above Diagonal) and z-Score (Below Diagonal) Data

	Pain	Stiffness	Fatigue	Concentration	Memory	Anxiety	Depression	GI	Sleep
Pain	[REDACTED]	4	5	4	4	2	2	1	4
Stiffness	-0.29	[REDACTED]	5	4	4	2	3	1	3
Fatigue	-0.33	-0.23	[REDACTED]	5	5	3	3	2	5
Concentration	-0.31	0.39	-0.25	[REDACTED]	3	2	3	1	3
Memory	0.28	0.37	-0.33	0.33	[REDACTED]	2	3	2	5
Anxiety	0.13	0.21	-0.13	0.17	0.15	[REDACTED]	1	0	2
Depression	-1.01	-0.26	-0.29	-0.28	-1.04	-0.16	[REDACTED]	0	3
GI	-0.38	-0.29	-0.32	-0.30	-0.39	-0.19	-0.35	[REDACTED]	2
Sleep	-0.37	-0.28	-0.31	-0.29	0.28	-0.18	-0.97	-0.36	[REDACTED]

Note: GI = gastrointestinal. Cutpoints for UniODA models based on *raw* data are indicated *above* the diagonal, and cutpoints for UniODA models based on *ipsatively standardized* data are indicated *beneath* the diagonal.

Table 9.28: Comparing UniODA Model Cutpoints for Raw and Ipsatively Standardized Data

Raw Data Cutpoints	Ipsative z-Score Cutpoints
0	-0.19, -0.35
1	-0.16, -0.29, -0.30, -0.38
2	0.21, 0.17, 0.15, 0.13, -0.18, -0.29, -0.36, -0.39, -1.01
3	0.33, -0.13, -0.28, -0.29, -0.97, -1.04
4	0.39, 0.28, -0.23, -0.29, -0.31, -0.37
5	-0.23, -0.25, -0.29, -0.31, -0.33

Chapter 10

Hierarchically Optimal Classification Tree Analysis

The explicitly optimal (maximum-accuracy) classification tree analysis (CTA) algorithm was discovered in 1996, and is known as hierarchically optimal CTA or HO-CTA.¹ In 2010 the second-generation algorithm called enumerated optimal CTA or EO-CTA was developed, that typically identifies more accurate and parsimonious models than HO-CTA (Chapter 11). In 2014 the third-generation, and current state-of-the-art algorithm known as globally-optimal CTA (GO-CTA) was discovered, that is the conceptual equivalent of quantum mechanics for classical data (Chapter 12).

Despite the development of more accurate EO and GO models, techniques used to identify HO-CTA models are important for two reasons. First, learning how to obtain an HO-CTA model improves one's understanding of the operation of CTA algorithms, thereby enhancing conceptual clarity and improving experimental design, hypothesis development, measurement practices, and interpretative skills. Second, UniODA and MegaODA software facilitates systematic manipulation and precise exploration within CTA models and in the development of alternative models. Therefore, the mechanical steps required to obtain an HO-CTA model are now illustrated.²

Obtaining an HO-CTA Model

This example investigates factors that discriminate patients who are likely to recommend an Emergency Department (ED) to others versus patients who are ambivalent about recommending the ED.² The study was set in an urban 800 bed university-based level 1 Trauma center having an annual census of 48,000 patients.³ One week post discharge, patients were mailed a survey assessing their satisfaction with care they received in the ED. The survey elicited ratings of the likelihood of recommending the ED to others, and satisfaction with aspects of administration, nurse, physician, laboratory, and staff care of family and friends. A total of 2,109 surveys having completed recommendation ratings were returned in a six-month period (17% return rate). Likelihood to recommend ("recom") was rated using a five-point Likert-type scale: scores of 3 (fair, $N = 239$) indicate *ambivalence*; and scores of 4 (good, $N = 584$) reflect *likely to recommend*. Analysis included a total of 823 patients responding with recommendation ratings of 3 or 4. For this exposition satisfaction ratings of aspects of care received from nurses were used as potential attributes: n1 = courtesy; n2 = took problem seriously; n3 = attention; n4 = informed patient about treatment; n5 = concern for privacy; n6 = technical skill. Satisfaction items were completed via five-point Likert-type scales: 1 = very poor satisfaction; 2 = poor; 3 = fair; 4 = good; 5 = very good satisfaction.

Determining the Minimum N for HO-CTA Model Endpoints

The first step in developing any CTA model is to determine *a priori* the minimum appropriate sample size for any (for every) model endpoint: issues requiring consideration in this context include statistical power and model cross-generalizability.

To ensure adequate statistical power in the absence of strong information regarding anticipated effect strength, an excellent heuristic is to use $ESS = 37.5$, a value that lies in the middle of the range that is used to define a moderate effect ($25 \leq ESS < 50$). Examination of Table 3.14 reveals that a minimum endpoint sample size of $N = 40$ for a Cohen's d value of between 0.7 and 0.8 corresponds to an ESS value

of 37.5 (*ESS* values in Table 3.14 are divided by 100). Referring to Table 3.13 reveals that statistical power for this sample size for $p < 0.05$ lies near 90%, the standard for statistical power in funded research. The exact minimum precision approach (Chapter 3) will generate a more conservative (i.e., larger) minimum denominator than the former approach for a given hypothesis, p , and level of power.

Described in Chapter 2, estimating the cross-sample generalizability of a model is accomplished by leave-one-out (jackknife), hold-out, and multisample generalizability analyses. For very large samples, using statistical power criteria to establish a minimum endpoint denominator can produce over-fit models having limited cross-generalizability when applied to independent random samples, particularly when the class variable is skewed, and/or when the replication sample is smaller than the sample that was used to create the model.

For example, imagine a CTA model based on a sample of $N = 10,000$ observations, with $N = 9,700$ *class 0* observations and $N = 300$ *class 1* observations. Also imagine a minimum endpoint denominator of $N = 40$ is specified on the basis of statistical power analysis, and that at least one endpoint in the model has $N = 40$: proportional distribution would yield $N = (300 / 9,700) \times 40 = 1.237$ *class 1* observations in this endpoint. Finally, imagine a replication study is conducted for a sample of $N = 1,000$ observations: in this case, if an identical effect emerged, then the endpoint with $N = 40$ for the larger sample would have $N = 4$ (and $N = 0.1237$ *class 1* observations)—and thus inadequate statistical power, for the smaller sample.

In such circumstances a heuristic approach involves setting the minimum endpoint denominator to be 5% or 10% of the total sample. This heuristic is motivated by K-Fold cross-validation methodology used in machine learning, in which a sample is separated into between three (3-Fold) at the low end, and ten sets of 10-Fold independent random subsamples at the high end, to obtain an estimate of the cross-generalizability prediction error of the model (see work of Dr. Shuichi Shinmura).⁴⁻⁹ Two variations of this methodology involving minimizing the maximum classification error of a multi-sample (Gen) model are presented in Examples 10.4 and 10.5 in Yarnold and Soltysik.¹⁰ However, this approach is inappropriate when representation of the class categories is highly skewed, and/or when one or more class categories are relatively rare. Thus the 5% (corresponding to a 20-Fold subsample) and 10% (10-Fold subsample) heuristics are designed to avoid overfitting on the one hand, and to preserve statistical power (limited by the small proportion of *class 1* observations) on the other hand.

In the present application, the total sample is $N = 823$ observations, and 5% of this value is 41.25 observations. Thus, upon consideration of both statistical power and cross-generalizable considerations, the minimum endpoint value in this application is rounded-up to a value of 42 observations. To enter the HO-CTA model, *the attribute with the highest ESS value must meet the criterion for experimentwise significance, and also must have 42 or more observations per endpoint*.

Growing the HO-CTA Model

To identify the initial (root) node of the HO-CTA model, UniODA is conducted for every attribute available to discriminate the class variable—here, self-rating of the likelihood to recommend the ED to others (3 or 4), for the entire sample. The attribute yielding the highest value for the *ESS* statistic, with both endpoints meeting or exceeding the minimum N criterion, is selected as the root node of the HO-CTA model if $p < 0.05$. *ESS* is the critical criterion by which the HO-CTA model is grown, and is the criterion that the HO-CTA model maximizes. *ESS* is based on the mean sensitivity (i.e., proportion of observations in a given class category that are correctly classified) of the model across all class categories. An errorless model yields a mean sensitivity of 1, and for a two-category problem, if the two class categories cannot be discriminated then a chance model yields mean sensitivity = 0.5. For a two-category problem, $ESS = [(mean\ sensitivity - 0.5) / 0.5] \times 100\%$. If the model is errorless (correctly classifies all observations) then $ESS = [(1 - 0.5) / 0.5] \times 100\% = 100$. If the model correctly classifies half the observations of each class category, or if the model correctly classifies all the observations in one category and misclassifies all the observations in the other category, $ESS = [(0.5 - 0.5) / 0.5] \times 100\% = 0$. Thus, $ESS=0$ is the level of classification accuracy expected by chance alone, and $ESS=100$ is perfect, errorless classification. UniODA analysis conducted to identify the root node was performed using the following UniODA and MegaODA software syntax:

```

OPEN recom.dat;
ATTR n1 to n6;
OUTPUT recom.out;
MISSING all (-9);
VARS recom n1 to n6;
MC ITER 10000;
CLASS recom;
GO;

```

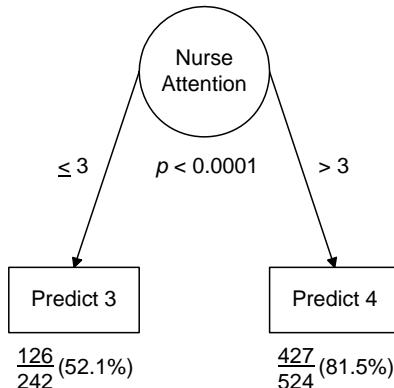
Rating of attention paid to the patient by the nurse (n3) yielded greatest *ESS* = 35.1, $p < 0.0001$. To guard against overfitting, CTA models only include attributes with p that satisfies the experimentwise criterion for statistical significance. In ODA software this is accomplished by using a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure, with *a priori* alpha splitting if appropriate for the investigation (Chapter 2). Presently the UniODA model was: if $n3 \leq 3$ then predict recom = 3; and if $n3 > 3$ then predict recom = 4. Table 10.1 presents the confusion table for this model applied to the data (note that the sample is reduced to $N = 766$ due to missing data for n3).

Table 10.1: Confusion Table for First UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	3	126	97
	4	116	427

When a recommended likelihood score of 3 was predicted 116 observations were misclassified, and when a recommended likelihood score of 4 was predicted 97 observations were misclassified. The sensitivity of this model for class category 3 is $126 / (126 + 97) = 0.565$, and the sensitivity of this model for class category 4 is $427 / (427 + 116) = 0.786$. The mean sensitivity is thus 0.676, and $ESS = [(0.676 - 0.5) \times 100\% = 35.1]$. Figure 10.1 illustrates the HO-CTA model as it exists at this point in the analysis.

Figure 10.1: HO-CTA Model after First Step of Analysis



In the second step of the analysis, an attribute that can improve classification accuracy for the left-hand endpoint is sought. This second analysis was accomplished by including one additional UniODA (or MegaODA) command before the GO command:

```
INCLUDE n3<4;
```

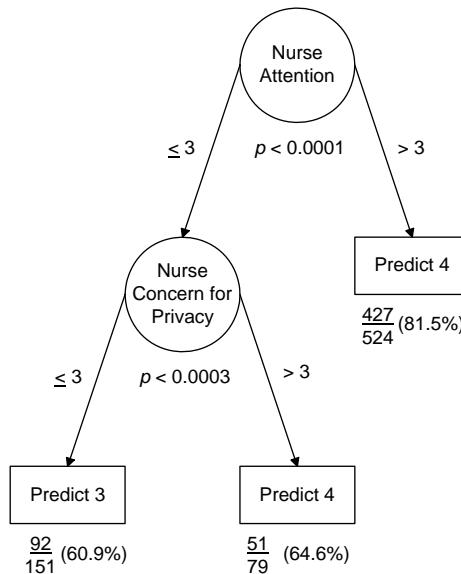
The rating of nurse concern for privacy (n5) yielded greatest *ESS* = 23.0, $p < 0.0003$. The UniODA model was: if $n5 \leq 3$ then predict that recom = 3; otherwise predict recom = 4. Table 10.2 presents the confusion table for this model applied to the data.

Table 10.2: Confusion Table for Second UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	3	92	28
	4	59	51

When the model predicted a recommended likelihood score of 3 a total of 59 observations were misclassified, and when the model predicted a recommended likelihood score of 4 a total of 28 observations were misclassified. Figure 10.2 illustrates the HO-CTA model as it exists at this point in analysis.

Figure 10.2: HO-CTA Model after Second Step of Analysis



An integrated confusion table is created to ascertain the accuracy of the model at this point in its development. In Figure 10.2 the left-most endpoint correctly predicts 92 of 151 (60.9%) observations were class 3; the middle endpoint correctly predicts 51 of 79 (64.6%) observations were class 4; and the right-most endpoint correctly predicts 427 of 524 (81.5%) observations were class 4. The integrated confusion table, for which $ESS = 31.4$, is shown in Table 10.3. Note that the sample was reduced to $N = 754$ (versus $N = 823$ with complete recommendation ratings) because of missing data for the two attributes.

Table 10.3: Integrated Confusion Table after Second UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	3	92	125
	4	59	478

In the third step of the analysis, an attribute that can improve classification accuracy for the left-most endpoint of the HO-CTA model is sought. This analysis was accomplished via the following modified UniODA (MegaODA) command:

INCLUDE n3<4 n5<4;

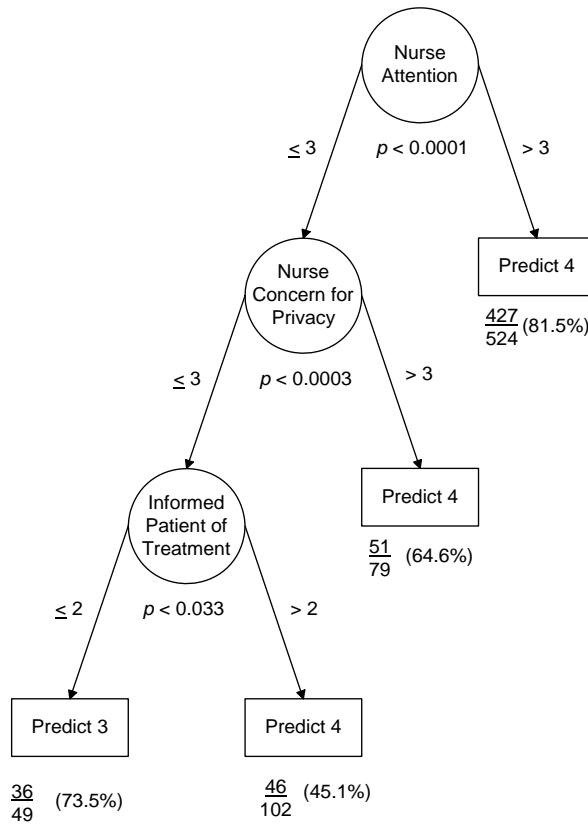
Rating of information about treatment (n_4) yielded greatest $ESS = 17.1, p < 0.033$. The UniODA model was: if $n_4 \leq 2$ then predict that recom = 3; and if $n_4 > 2$ then predict recom = 4. Table 10.4 presents the confusion table for this model applied to the data.

Table 10.4: Confusion Table for Third UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	3	36	56
	4	13	46

As seen in Table 10.4, when the model predicted a recommended likelihood score of 3 a total of 13 observations were misclassified, and when the model predicted a recommended likelihood score of 4 a total of 56 observations were misclassified. Figure 10.3 illustrates the HO-CTA model as it exists at this point in the analysis.

Figure 10.3: HO-CTA Model after Third Step of Analysis



To ascertain the accuracy of the model at this point in its development, an integrated confusion table is created. In Figure 10.3, the left-most endpoint correctly predicts 36 of 49 (73.5%) observations were class 3; the second-from-the-left endpoint correctly predicts 42 of 102 (45.1%) observations were class 4; the third-from-the-left endpoint correctly predicts 51 of 79 (64.6%) observations were class 4; and the right-most endpoint correctly predicts that 427 of 524 (81.5%) observations were class 4. The integrated confusion table, for which $ESS = 13.8$, is shown in Table 10.5. Note that the sample was reduced to $N = 750$ because of missing data for the included attributes.

Table 10.5: Integrated Confusion Table after Third UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	<u>3</u>	36	185
	<u>4</u>	13	516

Note that because the left-most endpoint has only 49 observations and the third-from-the-left endpoint has only 79 observations, no additional endpoints may be added at either branch since there are too few observations remaining to satisfy the minimum requirement of 42 observations per endpoint.

In the fourth step of the analysis, an attribute that can improve classification accuracy for the second-from-the-left endpoint of the HO-CTA model is sought. This fourth analysis was accomplished using the following modified UniODA (MegaODA) code:

```
INCLUDE n3<4 n5<4 n4>2;
```

Because none of the attributes achieved a Type I error rate that was statistically significant at the experimentwise criterion, this branch of the HO-CTA model cannot be expanded.

In the fifth step of the analysis, an attribute that can improve classification accuracy for the right-most endpoint of the HO-CTA model is sought. This fifth analysis was accomplished using the following modified UniODA (MegaODA) code:

```
INCLUDE n3>3;
```

The rating of nurse concern for privacy ($n5$) yielded greatest $ESS = 10.6$, $p < 0.042$. The UniODA model was: if $n5 \leq 3$ then predict that $recom = 3$; if $n5 > 3$ then predict $recom = 4$. Table 10.6 presents the confusion table for this model applied to the data.

Table 10.6: Confusion Table for Fifth UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	<u>3</u>	22	71
	<u>4</u>	54	359

As seen in Table 10.6, when the model predicted a recommended likelihood score of 3 a total of 54 observations were misclassified, and when the model predicted a recommended likelihood score of 4 a total of 71 observations were misclassified. Figure 10.4 shows the HO-CTA model at this point in analysis.

Controlling Experimentwise Type I Error

Because of the requirement that all Type I error estimates in the model are statistically significant at the experimentwise criterion, the model shown in Figure 10.4 is untenable. Using the sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure to control alpha inflation¹⁰, p -values associated with each node in the HO-CTA model are arranged in order of decreasing magnitude: the largest (least significant) p -value is at the top of the list, and the smallest (most significant) p -value is at the bottom of the list. Table 10.7 illustrates this for the model in Figure 10.4.

Figure 10.4: HO-CTA Model after Fifth Step of Analysis

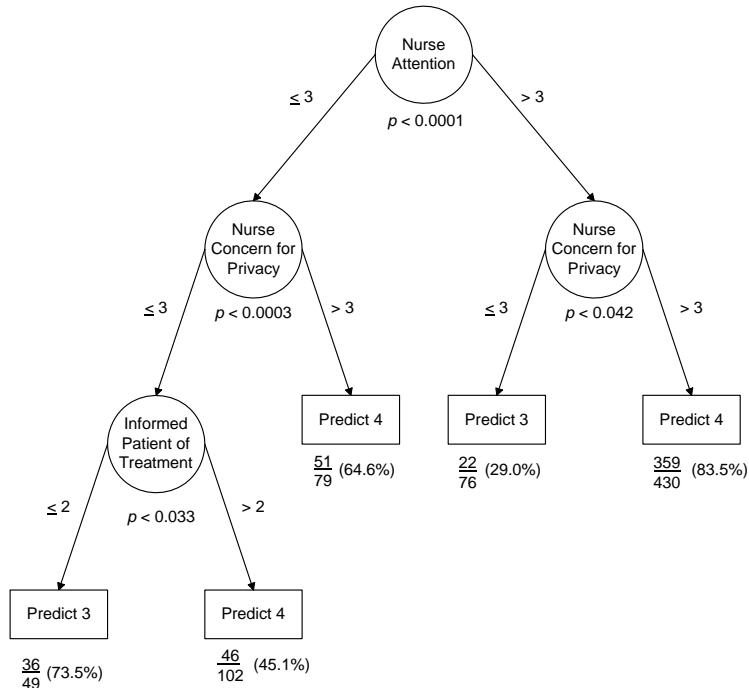


Table 10.7: Actual p -Values and Corresponding Sidak Critical p -Values

Actual p -value	Sidak Critical p -Value
0.042	0.05000
0.033	0.02533
0.0003	0.01696
0.0001	0.01275

Each actual p -value is compared with the corresponding Sidak critical p -value starting at the bottom of the ordered list. At each step of the procedure the actual and critical p -value is compared. If the actual p -value is less than or equal to the critical p -value, then the actual p -value is statistically significant at the experimentwise criterion of $p < 0.05$. However, if the actual p -value is greater than the critical p -value, then the actual p -value is not statistically significant at the experimentwise criterion of $p < 0.05$.

In the first step of the evaluation of the statistical significance of the actual p -values, because the most statistically significant actual p -value ($p < 0.0001$) is smaller than the corresponding critical p -value ($p < 0.01275$), this actual p -value is statistically significant with experimentwise $p < 0.05$.

In the second step of the evaluation of the statistical significance of the actual p -values, because the second-most statistically significant actual p -value ($p < 0.0003$) is smaller than the corresponding critical p -value ($p < 0.01696$), this actual p -value is also statistically significant with experimentwise $p < 0.05$.

In the third step of the evaluation of the statistical significance of the actual p -values, because the third-most statistically significant actual p -value ($p < 0.033$) is *larger* than the corresponding critical p -value ($p < 0.02533$), this actual p -value is *not* statistically significant with experimentwise $p < 0.05$. Thus, the HO-CTA node with this actual p -value is not statistically reliable.

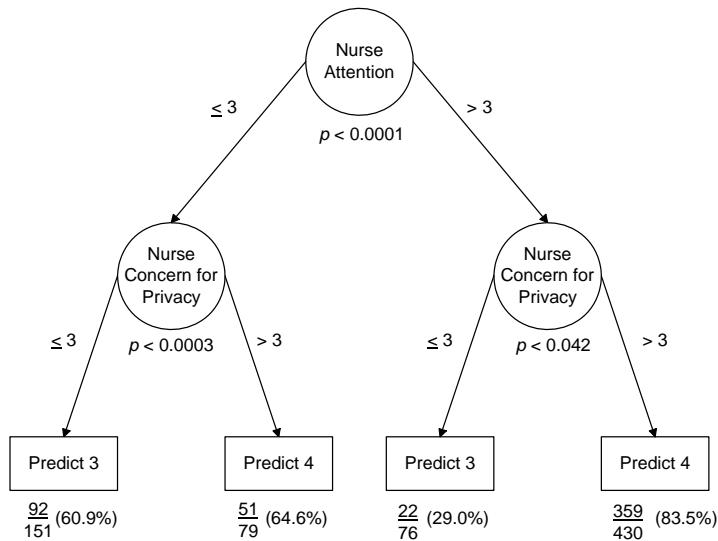
In this methodology, once a statistically unreliable p -value is identified, then the actual p -value that failed to fall at or beneath the Sidak critical p -value, and all of the less-statistically significant actual p -values higher in the ordered list, are considered statistically unreliable at the experimentwise criterion. Note that had the third p -value instead been lower than the Sidak criterion ($p < 0.02533$), then in the least

statistically fourth and final step of the evaluation of the statistical significance of the actual p -values, because the significant actual p -value ($p < 0.042$) is less than the corresponding critical p -value ($p < 0.05$), this actual p -value would have been statistically significant with experimentwise $p < 0.05$.

In the construction of HO-CTA models the standard is to eliminate the non-statistically-significant comparison that corresponds to the *deepest node* in the tree model. Presently this means that the node indicating that the nurse kept the patient aware of treatment progress is dropped from the model.

Figure 10.5 presents the final fully-grown HO-CTA model that meets the *a priori* criterion that all actual p -values are statistically significant with experimentwise $p < 0.05$ (in Table 10.7 the second actual p -value from the top of the list is dropped, and only the three remaining actual p -values are evaluated).

Figure 10.5: Corrected HO-CTA Model after Fifth Step of Analysis



To ascertain the accuracy of the model at this point in the development, an integrated confusion table is created, shown in Table 10.8 ($ESS = 31.9$). Note that the sample was reduced to $N = 736$ because of missing data on included attributes.

Table 10.8: Integrated Confusion Table after Corrected Fifth UniODA Analysis

		Predicted Recommendation	
		3	4
Actual	3	3	114
	4	113	410

In the sixth step of the analysis an attribute that can improve classification accuracy for the right-most endpoint of the HO-CTA model is sought. This sixth analysis was conducted by using the following modified UniODA (MegaODA) command:

INCLUDE n3>3 n5 >3;

Because none of the attributes achieved a Type I error rate that was statistically significant at the experimentwise criterion, this branch of the HO-CTA model cannot be expanded.

A table of critical Sidak values for up to 200 comparisons is provided as Appendix A in Yarnold and Soltysik¹⁰, and Chapter 4 of that text¹⁰ illustrates *a priori* alpha splitting, a procedure used to partition

the experimentwise Type I error rate between various analyses presented within a single project (i.e., a single manuscript) and prevent overly conservative criteria for statistical reliability.

Pruning the Fully-Grown HO-CTA Model to Ensure Maximum-Accuracy

Growing the initial HO-CTA model has been completed. However, subsequent to the initial development of this methodology, it was discovered that all such initial full-grown HO-CTA models must be *pruned* in order to explicitly maximize *ESS* and identify the final, maximum-accuracy HO-CTA model.¹¹

Figure 10.6A:
L1 Sub-Branch and Confusion Table

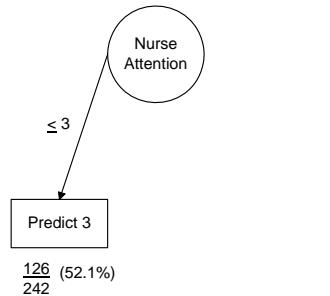
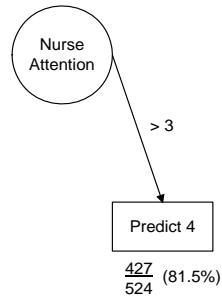


Figure 10.6C:
R1 Sub-Branch and Confusion Table



		L1 Predicted	
		3	4
Actual	3	126	0
	4	116	0

		R1 Predicted	
		3	4
Actual	3	0	97
	4	0	427

Figure 10.6B:
L2 Sub-Branch and Confusion Table

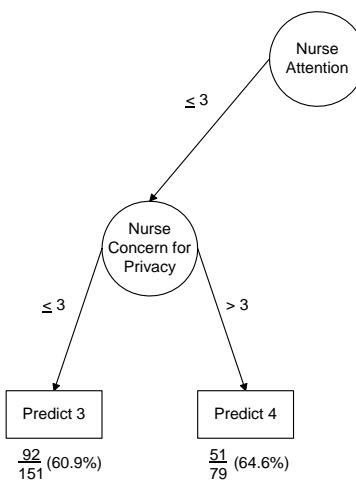
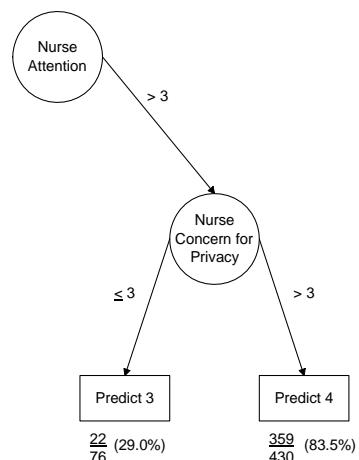


Figure 10.6D:
R2 Sub-Branch and Confusion Table



		L2 Predicted	
		3	4
Actual	3	92	28
	4	59	51

		R1 Predicted	
		3	4
Actual	3	22	71
	4	54	359

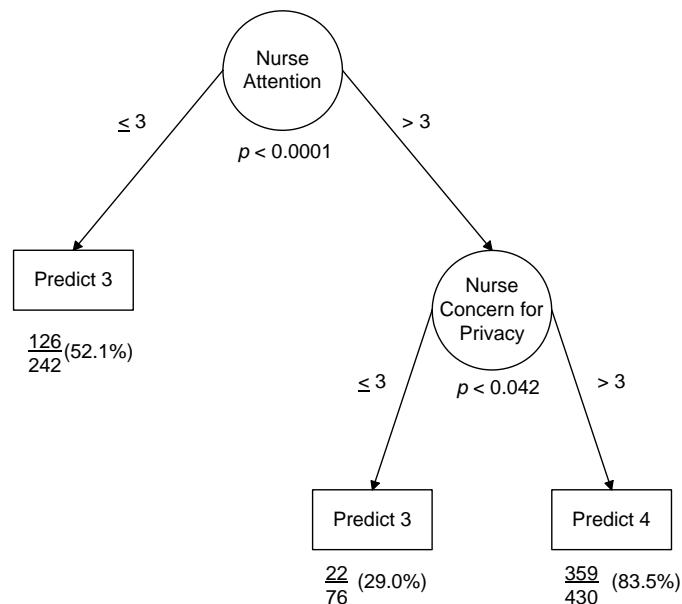
Such pruning involves deconstructing the initial HO-CTA model (Figure 10.5) into all possible nested sub-branches, and then selecting the combination of sub-branches that explicitly maximizes *ESS*. Sub-branches are constructed separately for the branches emanating from the left-hand side of the root (top) node of the model, and for branches emanating from the right-hand side of the root node. The sub-branches are indicated by a letter (L for left-hand side, R for right-hand side) and a number (the number of nodes in the sub-branch). Figures 10.6A-10.6D show the two left-hand sub-branches, and the two right-hand sub-branches, for the HO-CTA model in Figure 10.5.

For the final step of the maximum accuracy pruning procedure, Table 10.9 presents integrated confusion tables for all four possible combinations of left (L1, L2) and right (R1, R2) sub-branches, and their associated *ESS*. As seen in Table 10.9, the combination L1-R2 has the greatest *ESS* = 35.1, and thus is selected as the maximum-accuracy HO-CTA model (Figure 10.7).

Table 10.9: Classification Results for Every Combination of Left (L1-L2) and Right (R1-R2) Sub-Branch

<u>Model</u>	<u>Confusion Table</u>		<u>Model</u>	<u>Confusion Table</u>	
<i>L1-R1</i>	Predicted		<i>L1-R2</i>	Predicted	
	<u>3</u>	<u>4</u>		<u>3</u>	<u>4</u>
Actual	<u>3</u> 126	97	Actual	<u>3</u> 148	71
	<u>4</u> 116	427		<u>4</u> 170	359
	ESS=35.1			ESS=35.4	
<i>L2-R1</i>	Predicted		<i>L2-R2</i>	Predicted	
	<u>3</u>	<u>4</u>		<u>3</u>	<u>4</u>
Actual	<u>3</u> 92	125	Actual	<u>3</u> 114	99
	<u>4</u> 59	478		<u>4</u> 113	410
	ESS=31.4			ESS=31.9	

Figure 10.7: Final Pruned Maximum-Accuracy HO-CTA Model



As seen, construction of a maximum-accuracy HO-CTA model is a complex, analysis-intensive enterprise. HO-CTA models reward analytic rigor with accurate, parsimonious models that are impossible to obtain using legacy linear-based statistical methods.

Additional considerations imperative in UniODA and CTA modeling, not illustrated presently, are treatment of categorical variables, correct transformation of serial data, assessing cross-generalizability of HO-CTA models, and the use of weights. With respect to treatment of categorical variables, unlike general linear model or maximum-likelihood paradigms, in the ODA paradigm multicategorical variables involving more than two response categories are *not* transformed into a series of binary (dummy) variables normed against a reference category; instead the multicategorical attribute is simply treated as being a categorical attribute having multiple categorical options.¹²⁻¹⁴ With respect to serial measurements, ipsative standardization is essential to prevent anomalous measurement artifacts including paradoxical confounding.¹⁵⁻¹⁷ Potential cross-generalizability of maximum-accuracy models is estimated via “leave-one-out” one-sample jackknife analysis, and is assessed using hold-out validity samples, via commands offered in UniODA and MegaODA software.¹⁸ If individual observations are assigned weights, the HO-CTA model will maximize weighted classification accuracy.¹⁹

Methodology discussed above focuses on identification of the HO-CTA model yielding maximum accuracy normed against chance—greatest possible integrated *ESS*. The process of obtaining an HO-CTA model sometimes identifies non-linear models (sub-branches) that accurately classify (*ESS*) or make point predictions (effect strength for predictive value²⁰ or *ESP*) of important class categories.²¹

For example, identification of adverse drug reactions (ADRs) is critical to improved patient safety, and warfarin is known to be associated with a high rate of ADRs.²² HO-CTA was used to predict ADRs attributable to medications taken in addition to warfarin, for a sample of 2,289 hospital inpatients. Data were collected from June 2000 to November 2001 at a private teaching hospital in Chicago. Information on ADRs was collected by Pharmacy as part of the in-house ADR program which involves examination of caregiver reports, surveillance, and medical record reviews. A model that met the experimentwise Type I error rate criterion but that wasn't pruned to explicitly maximize *ESS* was identified. Relatively high ADR rates were identified for patients on warfarin who also received zolpidem tartrate, tamsulosin HCL, famotidine, nitroglycerin, and rofecoxib (Table 10.10). The model achieved moderate classification accuracy (*ESS* = 38.0), correctly classifying 1,323 of 2,246 patients (58.9%) without an ADR, and 34 of 43 patients (79.1%) experiencing an ADR. The CTA model enumerated more than one billion systems of linear inequalities, requiring ten CPU-hours to solve on a 650 MHz Pentium-4 microcomputer.

Table 10.10: Medications Identified by CTA which are Associated with Significant ADRs When Taken Concurrently with Warfarin

<u>Medication</u>	<u>Ratio</u>	<u>%</u>	<u>p≤</u>	<u>ESS</u>
Zolpidem Tartrate	12/304	3.95	0.01	14.9
Tamsulosin HCL	5/72	6.94	0.004	12.7
Famotidine	10/409	2.44	0.05	17.3
Nitroglycerin	4/106	3.77	0.022	18.2
Rofecoxib	3/66	4.55	0.016	20.4

Note: Ratio = number of ADRs / number of patients with indicated medication. % = Ratio x 100%.

Using only information about the medications administered after the patient was admitted to the hospital, the CTA model reveals that 80% of the ADRs involving warfarin occurred in approximately 40% of the patients receiving one (or more) of five additional medications.

This discussion focuses on *how to obtain* a HO-CTA model, but it does not consider *how to report* the findings of a HO-CTA model. Well-known reporting statistics, such as confusion tables, and summary indices including sensitivities, predictive values, and overall classification accuracy were discussed here, and model diagrams and normed accuracy (*ESS* and *ESP*) scores were discussed previously. Chapter 11 discusses construction of staging tables (easy-to-use scoring templates); computation of odds, odds ratios

and propensity scores; the use of pie charts to visually represent identified strata; and the attribute importance in discrimination (A/D) statistic—the optimal analogue to R^2 used in linear modeling. Chapter 12 discusses the definition of an ideal statistical model; identification of all-possible maximum ESS solutions for a given application; assessing the quality of an empirical model in light of the theoretical ideal; and computation of exact discrete confidence intervals for performance parameters for models and chance.

Forward HO-CTA

Imagine that a researcher wishes to investigate the efficacy of a given attribute (A_1) in predicting a class variable. A secondary objective is to determine whether adding other attributes (A_2, A_3 , etc.) increases ESS above and beyond what is achieved using only A_1 . This analysis may be conducted with the partial UniODA procedure described in Chapter 8, by substituting A_1 for the confounding variable.

Reverse HO-CTA

HO-CTA may be reversed for applications involving an ordered class variable and categorical attributes: whereas regression analysis is used to make point predictions for the dependent measure based on values of the independent variables, reverse HO-CTA is used to find domains on the dependent measure that are explained by independent variables. This is illustrated for a serial multi-attribute single-case design.

Self-monitoring and review tool (SMART) is an interactive, internet-based, self-monitoring and feedback system for helping individuals to identify and monitor the relationships between their personal behaviors, stressors, management strategies, and symptom levels over time.²³ SMART involves longitudinal collection and statistical analysis of self-monitoring data, with the ultimate objective of timely delivery of personalized feedback derived from the data.²⁴

This study examines longitudinal data for an individual using SMART to rate the intensity of nine fibromyalgia (FM) symptoms experienced over 297 consecutive days. Rated using 10-point Likert-type response scales, the symptoms are pain, stiffness, fatigue, concentration problems, memory problems, anxiety, depression, gastrointestinal problems, and sleep problems.²³ Symptoms are rated via 10-point (0-9) Likert-type scales: increasing values indicate worsening symptoms. For reverse HO-CTA daily symptom ratings were coded as positive if it exceeded the median value, and as negative otherwise. Ratings are treated as independent variables in multiple regression analysis (MRA), and as attributes in reverse HO-CTA. Presently the dependent (MRA) or class (reverse HO-CTA) variable is 500 mb geopotential height anomaly (HT500) measured in meters, an atmospheric pressure index independently recorded by the investigator for each day in the longitudinal record.¹⁹ Descriptive statistics for study variables are provided in Table 10.11 (analysis of ipsatively standardized data is appropriate here¹⁶ and will be reported later).

Table 10.11: Descriptive Statistics

<u>Variable</u>	<u>Mean</u>	<u>SD</u>	<u>Median</u>
HT500	5575	158	5565
Pain	4.6	1.5	5
Stiffness	5.3	1.6	5
Fatigue	6.4	1.4	6
Concentration	5.4	1.7	5
Memory	5.6	1.4	6
Anxiety	1.0	0.9	1
Depression	1.5	1.6	1
Gastrointestinal	0.4	1.0	0
Sleep	5.6	1.6	5

Multiple Regression Analysis

Among the most widely used statistical analysis methods, MRA requires little in the way of introduction.²⁵ The present data were first analyzed by MRA for expository purposes. The first analysis used raw data, HT500 as the dependent variable, and symptom ratings as the independent variables. Using all nine symptoms the model had $R^2 = 0.34$ [$F(9,287) = 16.4, p < 0.0001$], with concentration problems ($p < 0.0001$), fatigue ($p < 0.0001$), and anxiety ($p < 0.02$) explaining statistically reliable unique variance. The analysis using only these three independent variables found that the effect for anxiety was statistically unreliable (indicating the presence of paradoxical confounding), so the final model used concentration problems and fatigue (p 's < 0.0001) as independent variables: $R^2 = 0.23$; $F(2,294) = 43.1, p < 0.0001$. The regression model relating HT500 to patient symptoms was:

$$\text{HT500} = 5716.6 + 36.0 \times \text{concentration problems} - 68.8 \times \text{fatigue}.$$

The model shows that for this person, increasing HT500 is associated with *decreasing fatigue* and *increasing concentration problems*.

In the second analysis dummy-variable MRA was performed with median-based binary symptom indicators as independent variables (concentration problems was dropped to prevent multicollinearity). With all eight symptoms the model had $R^2 = 0.17$ [$F(8,288) = 7.5, p < 0.0001$], with stiffness ($p < 0.0001$) and anxiety ($p < 0.002$) explaining statistically reliable unique variance. The final model relating HT500 to symptoms using stiffness ($p < 0.0001$) and anxiety ($p < 0.04$) as independent variables [$R^2 = 0.13, F(2,294) = 22.8, p < 0.0001$] was: $\text{HT500} = 5618.2 - 110.6 \times \text{stiffness} + 49.2 \times \text{anxiety}$. The model shows that for this person, increasing HT500 is associated with *decreasing stiffness* and *increasing anxiety*.

Motivating Reverse CTA

A challenging issue for the regression results, in light of the study aim, is how to use the regression models to alert patients about forthcoming symptom-inducing or symptom-relieving weather. For example, if a forecast calls for higher HT500, what does this imply about levels of stiffness and anxiety that the patient should anticipate experiencing? The regression model may be rearranged to estimate a given symptom based on HT500, it is impossible to estimate both of the symptoms simultaneously.

By their structure it is obvious that regression models are useful for *predicting the level of HT500 from patient symptoms*.

Instead what is needed is precisely the opposite functionality, *predicting the patient symptoms based on HT500*. As demonstrated below, reverse HO-CTA provides this functionality.

Defining Attributes

In conventional HO-CTA the class variable is binary, and attributes may be categorical or ordered.²⁶ In reverse HO-CTA the class variable is ordered, and *attributes must be categorical*. The present research simulates alerts about relatively severe symptom days, defined for each day and symptom as positive if the rating is greater than the median value for the symptom (Table 10.11), and as negative otherwise.

Reverse HO-CTA begins by determining the attributes to include in analysis, using structural decomposition. In step one the strength of the relationship between HT500 and each of the nine categorical variables was evaluated for the total sample, and the model for stiffness had greatest associated ESS. The UniODA model was: if $\text{HT500} \leq 5675$ ppm predict positive, otherwise predict negative. The model yielded moderate $ESS = 40.0$: it correctly classified 60% of $N = 214$ days having HT500 below the cutpoint, and 89% of $N = 83$ days having HT500 above the cutpoint ($p < 0.0001$).

In step two depression had greatest ESS (41.4, $p < 0.0001$), and the UniODA model (if $\text{HT500} \leq 5490$ ppm then predict positive) correctly classified 60% of $N = 40$ days with HT500 below the cutpoint, and 80% of $N = 54$ days with HT500 above the cutpoint.

In the third and final step anxiety had greatest ESS (75.0, $p < 0.05$): the UniODA model (if $\text{HT500} > 5560$ ppm then predict positive) correctly classified 100% of $N = 18$ days having HT500 below the cutpoint,

and 33.3% of $N = 9$ days having HT500 above the cutpoint. Thus, stiffness, depression and anxiety were selected as the attributes to use in reverse CTA.

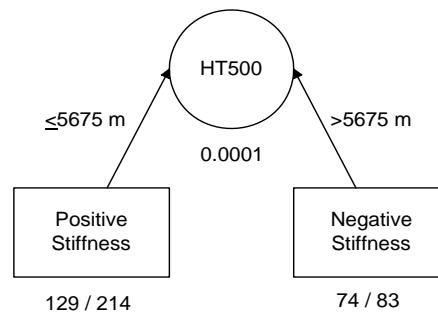
Obtaining the Model

Because no automated software is available that is capable of performing reverse HO-CTA, the analysis is conducted manually vis-à-vis UniODA or MegaODA software. Because only a small number of attributes were identified presently in structural decomposition analysis (a common finding), *enumerated* reverse HO-CTA is demonstrated. Decomposition analysis determined that the model for stiffness had strongest *ESS*, and it is arbitrarily selected as the root attribute in the first of the three HO-CTA models required to conduct the enumeration (all three attributes will be used as the root). Decomposition and reverse HO-CTA are both performed using the following UniODA and MegaODA software syntax:

```
OPEN example.dat; ATTR stiff depress anxiety;
OUTPUT example.out; MC ITER 25000;
VARS ht500 stiff depress anxiety; GO;
CLASS ht500;
```

An illustration of the root of this reverse HO-CTA model is provided in Figure 10.8. Ordinarily the attributes are shown in nodes and class variable in endpoints, but the opposite order occurs for reverse HO-CTA. Similarly to conventional HO-CTA arrows indicate paths from class variables to attributes, but in reverse HO-CTA arrows point *up* the tree. In contrast to conventional HO-CTA where one reads down the tree starting from the root, in reverse HO-CTA one reads *up the tree starting from the endpoints*. As seen, when $HT500 \leq 5675$ m, stiffness is correctly predicted to be positive on 129 of 214 (60.3%) days, and when $HT500 > 5675$ m, stiffness is correctly predicted to be negative on 74 of 83 (89.2%) days.

Figure 10.8: Root of First Reverse HO-CTA Model



The next analytic step is evaluating if an attribute should be added to the left endpoint, requiring *adding* the command:

```
exclude ht500>5675;
```

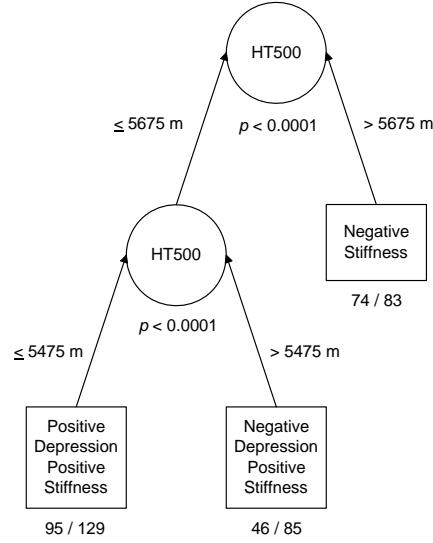
Depression had greatest *ESS* ($28.4, p < 0.0003$) and was thus added to the model as is illustrated in Figure 10.9.

No additional attributes could be added as left- ($p's > 0.14$) or right-hand ($p's > 0.25$) branches off of depression, so construction of the left-hand side of the model is complete.

The next step involves assessing whether to add an attribute to the right endpoint. This is accomplished by *modifying* the prior command:

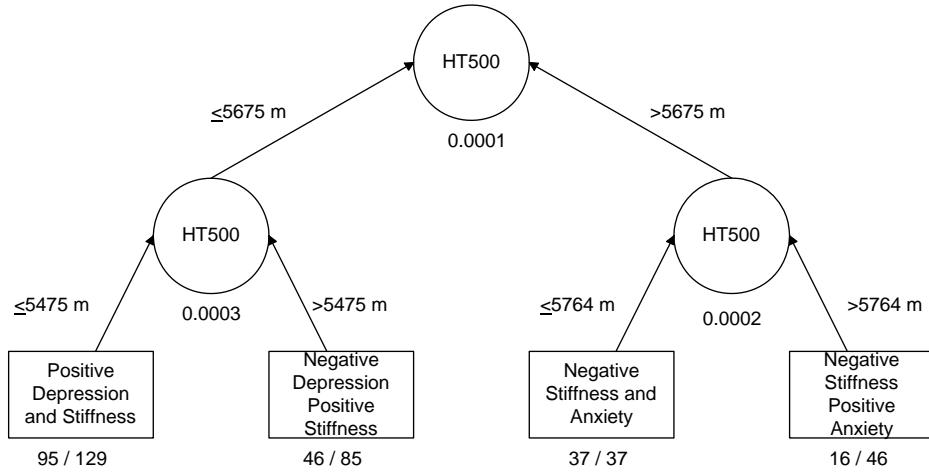
```
exclude ht500<=5675;
```

Figure 10.9: Attribute Added on Left Branch



The UniODA model for anxiety had the greatest *ESS* and therefore was added to the model. The classification for the left endpoint was perfect so no attribute could be added, and no statistically significant attributes emerged at the right-hand branch: thus, initial construction of this reverse HO-CTA model is complete (Figure 10.10).

Figure 10.10: Complete Non-Pruned Reverse HO-CTA Model Having Stiffness as the Root Variable



The next step involves pruning the full HO-CTA model in order to find the (sub)model having the greatest overall *ESS*. The reverse HO-CTA model in Figure 10.9 had the greatest overall *ESS* of 44.5 (versus 30.4 for the full model in Figure 10.10).

The foregoing procedure is now repeated twice—once using depression as the root, and again using anxiety as the root. Not illustrated here, *ESS* of the resulting pruned reverse HO-CTA models was 39.3 and 22.7, respectively.

Thus, the reverse HO-CTA model illustrated in Figure 10.9 was the strongest, most parsimonious representation of the longitudinal symptom data that was identified. Interpreting the model is simplified by examining the corresponding staging table presented as Table 10.12: *N* is the number of days in the series with HT500 falling in the indicated domain (stage); % is the percent of the *N* days in which the indicated symptom was in fact present; and Odds are given for a *bad symptom day*.

Table 10.12: Staging Table Predicting Patient Symptoms as a Function of HT100 Values

<u>Stage</u>	<u>HT500</u>	<u>Symptom</u>	<u>N</u>	<u>%</u>	<u>Odds</u>
1	>5675 m	None	83	89.2	1:8
2	>5475 m	Stiffness	85	53.4	1:1
3	≤5475 m	Stiffness, Depression	129	76.0	3:1

In the case of the present individual, the odds of a bad symptom day are 1 in 9 if the pressure is high ($HT500 > 5675\text{m}$); 1 in 2 (for stiffness) when pressure falls to an intermediate level ($HT500 > 5475\text{m}$); and 3 in 4 when the pressure falls to a low level ($HT500 \leq 5475\text{m}$). Several weather services predict HT500 days to two weeks or longer into the future. Thus, providing individuals with short-term and intermediate-range alerts regarding the odds of experiencing good- and bad-symptom days is a realistic opportunity.²⁷

HO-CTA in Applied Research

Researchers in numerous laboratories have undertaken the analysis-intensive, complex task of manually constructing HO-CTA models using UniODA, the only software that can accomplish this feat. In disciplines such as medicine, psychology, neurology, education, criminal science, pharmacology, and engineering the HO-CTA model obtained was more accurate, parsimonious, and theoretically apropos than any alternative analysis published in applications of inquiry. A thorough review of the research using HO-CTA is warranted but lies outside the objectives of this book. The following review covers a portion of our research using HO-CTA reported in literatures in medicine and allied medical disciplines prior to 2010, and is presented to demonstrate the diversity of phenomena that have been successfully explored in these areas.

HO-CTA was first used in medicine to predict in-hospital mortality attributable to *Pneumocystis carinii* pneumonia, or PCP.²⁸ Analysis was performed for 1,193 patients discharged alive ($N = 988$) or who died in-hospital ($N = 205$). Manually-derived using UniODA software the HO-CTA model selected alveolar-arterial oxygen gradient (AaPo₂—difference in partial pressure of oxygen between pulmonary system and blood: higher values indicate more severe pneumonia), body mass index (a measure of nutritional status, predictive of poor short- and long-term survival rates), and prior AIDS (a binary indicator of whether the current episode of PCP was the first clinical evidence of full-blown AIDS) as attributes. The HO-CTA model yielded a relatively weak *ESS* = 21.2, but alternative statistical methods used with these data (e.g., logistic regression analysis, regression-based recursive partitioning) yielded *ESS* values to an order of magnitude weaker. Subsequent research employing the ordinal severity staging algorithm created by this model as a prognostic index confirmed stage is a powerful ordinal risk factor for in-hospital mortality from PCP.^{29,30} Recently, with increasing use of PCP prophylaxis and multidrug antiretroviral therapy the clinical manifestations of HIV infection have changed dramatically. Because predictors of inpatient mortality for PCP may have also changed, HO-CTA was used to develop a new staging system for predicting inpatient mortality for patients with HIV-associated PCP admitted between 1995 and 1997.³¹ Chart reviews were performed for 1,660 patients hospitalized with HIV-associated PCP at 78 hospitals in seven metropolitan areas in the USA. HO-CTA identified a five-category staging system (*ESS* = 33.1) using three predictors: wasting, alveolar-arterial oxygen gradient (AaPO₂), and serum albumin level. Mortality rate increased with stage: 3.7% for Stage 1; 8.5% for Stage 2; 16.1% for Stage 3; 23.3% for Stage 4; and 49.1% for Stage 5.

During the mid-1990s, community-acquired pneumonia (CAP) began to account for an increasing proportion of the pulmonary infections in people with HIV infection: hospital mortality rates for HIV-associated CAP ranged to 28%. A staging system was thus developed for categorizing mortality risk of patients with HIV-associated CAP using information available prior to hospital admission.³² Data were obtained for a retrospective medical records review of 1,415 patients hospitalized with HIV-associated CAP from 1995 to 1997 at 86 hospitals in seven metropolitan areas. The overall inpatient mortality rate was 9.1%. Predictors of mortality in the HO-CTA model included presence of neurologic symptoms, respiratory rate of 25 breaths/minute or greater, and creatinine $> 1.2\text{ mg/dL}$. A five-category staging system yielding a mortality rate of 2.3% for stage 1, 5.8% for stage 2, 12.9% for stage 3, 22.0% for stage 4, and 40.5% for stage 5: *ESS*

= 45.5. This staging system proved to be useful for guiding clinical decisions about the intensity of patient care, and for case-mix adjustment in research addressing variation in hospital mortality rates. HO-CTA was also used to discriminate CAP versus inhalational anthrax cases.³³ Limiting effects of a bioterrorist anthrax attack necessitates rapid detection of the earliest victims, so a study was run to improve physicians' ability to rapidly detect inhalational anthrax victims. A case-control study compared chest radiograph findings from 47 patients from historical inhalational anthrax cases and 188 community-acquired pneumonia control subjects. HO-CTA was employed to derive an algorithm of chest radiograph findings and clinical characteristics that accurately discriminated inhalational anthrax versus community-acquired pneumonia. A nearly perfect HO-CTA model (*ESS* = 98.3) with three attributes (chest radiograph finding of mediastinal widening, altered mental status, and elevated hematocrit) was 100% sensitive and 98.3% specific. The most recent investigation in this line of research utilized HO-CTA to derive an algorithm for emergency department (ED) triage for rapid ordering of chest radiography for CAP, accounting for 1.5 million annual ED patient visits in the US.³⁴ An ED-based retrospective matched case-control study was conducted with 100 radiographic confirmed CAP cases and 100 radiographic confirmed influenza-like-illness control cases. A HO-CTA model was obtained that used three attributes (temperature, tachycardia, and hypoxemia on room air pulse oximetry), which was 70.8% sensitive and 79.1% specific: *ESS* = 49.9.

Encouraged by the initial and continuing success of HO-CTA in modeling heretofore poorly-understood outcomes, the use of HO-CTA began to proliferate across a variety of substantive areas within medicine. Areas of representation may be broadly categorized as representing clinical medicine, psychosocial aspects of medicine, and allied health disciplines.

Clinical Medicine

HO-CTA has been used in clinical medicine applications such as predicting adverse drug events (ADEs) via hospital administrative data in exploratory data-mining research, and pharmacoalgorithms in confirmatory theory-building research; self-selection for interventional management on the basis of clinical and psychosocial factors; gestational age at delivery after placement of an emergent cerclage; and mortality from thienopyridine-associated thrombotic thrombocytopenic purpura and from complications of HIV on the basis of nutritional information for a large observational database.

Hospital administrative data are appealing for the surveillance of ADEs because of their uniform availability, but expert-generated surveillance rules have limited accuracy. To assess whether rules based on nonlinear associations among available administrative data are more accurate, HO-CTA was applied to administrative data, and used to derive and validate surveillance rules for predicting bleeding/anticoagulation and delirium/psychosis ADEs.³⁵ Using a retrospective cohort design, a random sample of $N = 3,987$ patient admission records were drawn from all 41 Utah acute-care hospitals: reviewers identified ADEs by implicit chart review; pharmacists assigned Medical Dictionary for Regulatory Activities codes to ADE descriptions for identification of clinical groups of events; and the hospitals provided patient demographic, admission, and ICD-9-CM data. Incidence proportions were 0.8% for drug-induced bleeding/anticoagulation problems and 1.0% for drug-induced delirium/psychosis. The HO-CTA model for bleeding had strong sensitivity (86%), and fair positive predictive value (12%). The HO-CTA model for delirium had excellent sensitivity (94%), but low positive predictive value (3%). Poisoning and ADE codes designed for targeted ADEs had low sensitivities, and degraded model accuracy when forced to enter the model. These findings indicate that HO-CTA is a promising method for rapidly developing clinically meaningful surveillance rules for administrative data. The models obtained for drug-induced bleeding and anticoagulation problems may be useful for retrospective ADE screening and rate estimation.

In contrast to the preceding classical exploratory data-mining application of HO-CTA to study ADEs, Belknap began with the axiom that a prescription may be conceptualized as a health-care program implemented by a physician in the form of instructions that govern the plan of care for an individual patient.³⁶ Software design principles and debugging methods were used to create a "Patient-oriented Prescription for Analgesia" (POPA), the rate and extent of adoption of POPA by physicians was assessed, and HO-CTA was conducted to evaluate whether POPA would reduce the rate of both severe and fatal opioid-associated ADEs. The study involved a population of $N = 153,260$ hospitalized adults, $N = 50,576$ (33%) of whom received parenteral opioids. Hospital-wide the use of POPA increased to 62% of opioid prescrip-

tions, and opioid-associated severe/fatal ADEs fell from an initial peak of seven/month to zero/month during the final six months of the study.

In a study of treatment bias in observational outcomes research, HO-CTA was used to assess the role of clinical and psychosocial factors in predicting self-selection for interventional management (lower extremity bypass surgery or angioplasty) for patients with intermittent claudication.³⁷ A total of $N = 532$ patients with mild to moderate lower extremity vascular disease, and without prior peripheral revascularization procedures or symptoms of disease progression, were enrolled in a prospective outcomes study at the time of an initial referral visit for claudication to one of 16 Chicago-area vascular surgery offices or clinics. Study variables were derived from lower extremity blood flow records and patient questionnaires, and follow-up home health visits were used to ascertain frequency of lower extremity revascularization procedures within six months (13.3%). Ten patient attributes were used in the HO-CTA model: sensitivity = 67.6%; specificity = 92.9%; ESS = 57.7%. Initial ankle-brachial index (used to classify 100% of sample), leg symptom status over the previous six months (used to classify 89% of sample), self-reported community walking distance (used to classify 74% of sample) and prior willingness to undergo a lower extremity hospital procedure (used to classify 39% of sample) were the most influential attributes in the model, and are critical control variables for a valid observational study of treatment effectiveness.

HO-CTA was also used to develop a predictive model for predicting gestational age at delivery, after placement of an emergent cerclage in the second trimester, for $N = 116$ women with documented cervical change on physical examination.³⁸ HO-CTA was employed to predict delivery prior to 24 weeks, between 24 and 27 6/7 weeks, or 28 weeks or later. Delivery prior to 24 weeks was best predicted by the presence of prolapsed membranes and gestational age at cerclage placement; delivery between 24 and 27 6/7 weeks was best predicted by parity alone; and delivery of at least 28 weeks was best predicted by cervical dilation and length, presence of prolapsed membranes, and parity. When choosing a single model to predict delivery at the three different gestational age periods, the HO-CTA model predicting delivery at 28 weeks yielded the most accurate results. Findings for outcome after emergent cerclage are informative for both patients and physicians.

HO-CTA was also used in research describing clinical and laboratory findings for a sample of $N = 128$ patients with thienopyridine-associated thrombotic thrombocytopenic purpura (TTP).^{39,40} Duration of thienopyridine exposure, clinical and laboratory findings, and survival were recorded for all subjects, and ADAMTS13 activity (39 patients) and inhibitor (30 patients) were measured for a subset of the individuals. Among the patients who developed TTP more than two weeks after thienopyridine exposure, therapeutic plasma exchange (TPE) increased likelihood of survival (84% versus 38%, $p < 0.05$). In contrast, among the patients who developed TTP within two weeks of starting thienopyridines, survival was 77% with TPE and 78% without. Findings suggested that TTP drug toxicity occurs by two different mechanistic pathways, characterized primarily by time of onset before versus after two weeks of thienopyridine administration.

A retrospective observational database with detailed medical records on $N = 2,179$ HIV-positive patients who attended the Johannesburg General Hospital HIV clinic was mined to assess the effect of nutrition on health in this population.⁴¹ Times to progression or death were calculated from the patient's first clinic visit. HO-CTA showed that by using race alone, one can predict progression to AIDS in ≤ 1 year for 79.3% of nonwhite patients, and predict no progression for 59.4% of white patients. For the nonwhite patients, the next most useful predictor of progression was the use of multivitamins: multivitamin tablets (MVI), vitamin B complex tablets (VBC), or pyridoxine used in the clinic. The median progression time to AIDS was 32.0 weeks for patients without vitamins and 72.7 weeks for patients who took vitamin B, and the median survival was 144.8 weeks for patients without vitamins and 264.6 weeks for patients who took vitamin B. These findings demonstrate that HO-CTA can elucidate clinically-relevant relationships within large patient populations, such as observational databases.

Finally, HO-CTA was used to identify risk factors for venous thromboembolism (VTE) during the rehabilitation phase of spinal cord injury for a sample of $N = 243$ patients with acute spinal cord injury, $N = 51$ of whom had VTE, and $N = 8$ of whom died.⁴² The attributes included type and location of spinal cord injury, American Spinal Injury Association classification, concomitant injuries, surgical procedures, complications, preexisting illnesses, and use of antithrombotic prophylaxis. A three-attribute HO-CTA model was obtained that identified patient groups differing in likelihood of experiencing deep vein thrombosis (DVT). The group having the highest likelihood of DVT was patients with cancer over the age of 35 years, though

women without cancer between the ages of 36 and 58 years, and cancer-free men with flaccid paralysis, were also at increased risk.

Psychosocial Aspects of Medicine

HO-CTA has been employed in the investigation of psychosocial aspects of medicine, for example quality-of-care and patient satisfaction in the emergency department (ED); severity-of-illness and quality-of-life of asthma patients; age and functional status of ambulatory internal medicine patients; and literacy and hospitalization rates of general medicine outpatients.

HO-CTA was used to identify perceptions that predict patient (dis)satisfaction with care received in the ED.⁴³ Data were responses: (a) to a survey mailed to all discharged patients over a 6-month period (the Academic Hospital); or (b) to a telephone-based interview of a random sample of discharged patients over a 1-year period (the Community Hospital). The survey and interview assessed overall satisfaction, as well as satisfaction with perceived waiting times, information delivery, and expressive quality of the staff, nurses, and physicians. Data for $N = 1,176$ patients (the training sample) and $N = 1,101$ patients (holdout sample) who rated overall satisfaction as either “very good” or “very poor” (Academic Hospital), and for $N = 856$ patients (the training sample) and $N = 431$ patients (holdout sample) who rated overall satisfaction as either “excellent” or “poor” (Community Hospital), were retained for analysis. For both hospitals, HO-CTA models efficiently achieved *ESS* values near 90 (p 's < 0.0001). Findings reveal overall (dis)satisfaction with care received in the ED is nearly perfectly predictable on the basis of patient-rated expressive qualities of ED staff, and suggest that interventions that reinforce positive expressive provider behaviors may reduce the number of dissatisfied patients in half.

HO-CTA was used to associate severity-of-illness (assessed via the Asthma Severity Index, or ASI) with quality-of-life (QOL) in a prospective study of clinical and psychological correlates of adverse asthma outcomes.⁴⁴ Data were collected at study intake and then every three months thereafter for one year for $N = 13$ adults with asthma, and included a QOL scale, the ASI, spirometry, history, and a physical exam. A perfect HO-CTA model was obtained that included a query about bodily pain in the last four weeks, and a self-assessment of one's general health.

Research reporting the first HO-CTA ever published discriminated $N = 65$ geriatric (≥ 65 years of age) and $N = 85$ non-geriatric ambulatory medical patients using five functional status subscales, and five single-item measures hypothesized to be relevant to functional status (assessing physical limitations, social support, and satisfaction with health).¹ The findings revealed four strata (“patient clusters”): relatively active non-geriatric adults; relatively inactive geriatric adults; inactive, depressed, socially isolated young women; and active, happy, socially connected geriatric adults.

Finally, although higher hospitalization rates are reported among patients with low literacy, prior research failed to determine the preventability of these admissions or to consider other determinants of hospitalization such as social support. HO-CTA was used to evaluate whether low literacy is a predictor for preventability of hospitalization when considered in the context of social support, sociodemographics, risk behavior and health status for a sample of $N = 400$ patients admitted to the general medicine wards in a university-affiliated Veterans Affairs hospital.⁴⁵ Two board-certified internists independently assessed the preventability of hospitalization and determined the primary preventable cause through blinded medical chart reviews. Significant predictors of having a preventable cause of hospitalization were binge alcohol drinking, lower social support for medical care, three or fewer annual clinic visits, and 12 or more people talked to weekly.

Allied Health Disciplines

HO-CTA has been used in research in allied health disciplines, including studies designed to enhance psychological diagnostic accuracy, model psychosocial adaptation, improve long-term functional status, and predict adolescent psychiatric inpatient hospitalization.

The utility of the Behavioral Assessment System for Children (BASC) and Child Behavior Checklist (CBCL) Parent scales was assessed in terms of discriminating between: (a) students with attention deficit-hyperactivity disorder (ADHD) versus non-ADHD students, and (b) inattentive-type versus combined-type

ADHD-afflicted students.⁴⁶ For both the BASC and the CBCL, a different HO-CTA model was developed for each of the two diagnostic predictions. In distinguishing ADHD versus non-ADHD students the BASC model was more parsimonious and accurate than the CBCL model, whereas for differentiating inattentive versus combined types the CBCL model was superior. The results demonstrate the diagnostic utility of the BASC and CBCL, and describe salient behavioral dimensions associated with subtypes of ADHD. Also, a Bayesian method for estimating the efficiency of a HO-CTA model versus chance, for any given base-rate of *class 1* (and also of *class 0*) membership, is presented.

Demographic and clinical correlates of lifetime substance use disorders were studied in a cohort of $N = 325$ recently hospitalized psychiatric patients.⁴⁷ Attributes including gender (male), age (younger), education (less), time in jail, conduct disorder symptoms, and antisocial personality disorder symptoms were predictive of substance use disorders. HO-CTA was successful in predicting 74% to 86% of the alcohol, cannabis, and cocaine use disorders.

Seventy patients with chronic fatigue syndrome were randomly assigned to the control ($N = 33$) or experimental ($N = 37$) group.⁴⁸ All patients continued usual medical care, but the experimental subjects also underwent a 9-week-long, 2-hours-per-week training in mindfulness meditation and medical qigong practices. HO-CTA was employed to model change in the SF36 12-month Health Transition score: patients were categorized as "improvers" versus "non-improvers". The model achieved very strong *ESS* = 80.5, based on SF36 Role Functioning Physical score and frequency of mind/body self-healing practice.

Finally, HO-CTA was used to explore predictors of inpatient hospital admission decisions using a sample of $N = 13,245$ children in foster care over a four-year period.⁴⁹ As hypothesized, clinical variables including suicidality, psychotism and dangerousness predicted psychiatric admissions; however, family problems, and the location of hospital screening, impacted decision making in a subsample of cases. The HO-CTA model developed in Year 1 reliably and consistently predicted admission decisions across the next three years.

Conclusion

HO-CTA identified the most accurate and interpretable statistical models ever published in applications in which it was employed in the field of medicine and allied medical disciplines. Programmatic research that spanned more than a decade revealed that the HO-CTA models replicated better across time and sample than alternative models of identical pulmonary phenomena. This is encouraging, especially because HO-CTA models published prior to 2010 preceded the discovery of optimal pruning to maximize model *ESS*.

HO-CTA models were published by independent laboratories in medicine as well as in disciplines including psychology, social work, transportation science, criminal justice, education, political science and computer programming, for example. While a review of accumulated literature is clearly warranted, it lies outside of the scope of the present work. For a partial list of publications that use UniODA and MegaODA software (which is required to manually construct an HO-CTA model) is available on the Publications tab of the *Optimal Data Analysis* eJournal webpage.⁵⁰

In addition to discovery of optimal pruning, software capable of conducting *automated* HO-CTA became available in 2010.²⁶ Two major advantages of automated HO-CTA software both involve pruning.

First, when an HO-CTA model is derived *manually* the Bonferroni procedure is conducted as the model is grown. As model growth proceeds the attributes that are in close proximity to the root variable, that have an associated Type I error of $0.01 < p < 0.05$, must be forced out of the model as an increasing number of attributes load on the lower branches: this can greatly increase complexity of the modeling process. However, when conducting *automated* HO-CTA analysis recursive trimming and re-development is user-transparent: the computer simply executes the HO-CTA growth algorithm until it is completed.

Second, the automated software always conducts pruning to explicitly maximize model accuracy, an arduous task to accomplish manually for complex models.

Chapter 11

Enumerated Optimal Classification Tree Analysis

Fourteen years after development of HO-CTA, a second-generation method that is known as *enumerated* optimal classification tree analysis (EO-CTA) was developed, that can identify substantially more accurate and parsimonious models than are obtained by a single iteration of the HO-CTA algorithm.¹

Obtaining an EO-CTA Model

Manual development of HO-CTA models via UniODA or MegaODA statistical software identified models in many disciplines that were less complex, and more accurate and theoretically apropos than corresponding models developed via the general linear model and maximum-likelihood paradigms. However, manual construction of maximum-accuracy HO-CTA models is complicated and analysis-intensive. This requisite rigorous computation motivated the development of automated CTA statistical software that is capable of identifying HO-CTA models using the attribute having highest *ESS* as the root node, as well as previously inconceivable EO-CTA models evaluating all possible combinations of attributes that meet the Type I error rate criterion in the *first three nodes* of the model. Availability of automated CTA software yielded models in many fields that were less complex, and more accurate and theoretically apropos than corresponding models developed via parametric²⁻²⁴ or HO-CTA²⁵⁻²⁸ methods (while warranted, a review lies outside of the scope of this work). We begin by demonstrating how to obtain HO-CTA and EO-CTA models via automated CTA software, and then comparing the two models that were developed using the identical data.

Context of the Exposition

As described in exposition of the development of an HO-CTA model (Chapter 10), data for the exposition of an EO-CTA model came from a study investigating factors increasing the likelihood of an ambivalent Emergency Department (ED) patient recommending the ED to others. The study was set in an urban 800 bed university-based level 1 Trauma center with annual census of 48,000 patients. One week after being discharged patients were mailed a survey assessing satisfaction with the care they received in the ED. The survey elicited ratings of the likelihood of recommending the ED to others, and satisfaction with aspects of administration, nurse, physician, laboratory, and care of family and friends. A total of $N = 2,109$ surveys with complete recommendation ratings were returned in a six-month period (17% return rate). Likelihood to recommend (“recom” in the CTA syntax) was rated using a five-point Likert-type scale: scores of 3 (fair, $N = 239$) indicate *ambivalence*; and scores of 4 (good, $N = 584$) reflect *likely to recommend*. Analysis thus included a total of $N = 823$ patients responding with recommendation ratings of 3 or 4.

As in the demonstration of the development of the HO-CTA model (Chapter 10), presently only the satisfaction ratings of aspects of care received from nurses were employed as potential attributes: n1 = courtesy; n2 = took patient’s problem seriously; n3 = attention paid to patient; n4 = informed patient about treatment; n5 = concern for patient privacy; and n6 = technical skill. Satisfaction items were rated via five-point Likert-type scales: scores of 1 = very poor satisfaction, 2 = poor, 3 = fair, 4 = good, and 5 = very good satisfaction. Data file requirements for CTA software are the same as the requirements for both UniODA and MegaODA software (Chapter 3).

Determining the Minimum N for EO-CTA Model Endpoints

The first step in developing any CTA model is to determine *a priori* the minimum appropriate sample size for any (every) endpoint in the model. As is detailed in exposition of HO-CTA analysis of the present data, consideration of statistical power and generalizability considerations determined the minimum endpoint value in this application is 42 observations (Chapter 10). In order to enter the EO-CTA model, the attribute must meet the criterion for experimentwise statistical significance, and must also have an endpoint with 42 or more observations (this is controlled using the MINDENOM function, illustrated below)..

Obtaining the EO-CTA Model

The HO-CTA and EO-CTA models for this application were both generated using the following CTA⁶ syntax (Appendix C):

```
OPEN recom.dat;                                MC ITER 10000 CUTOFF .05 STOP 99.9;
OUTPUT recom.out;                               PRUNE .05;
VARS recom n1 to n6;                           ENUMERATE;
CLASS recom;                                    MINDENOM 42;
ATTR n1 to n6;                                 GO;
MISSING all (-9);
```

Note that syntax used to operate CTA software is identical to syntax used to operate UniODA and MegaODA software, except for the three commands following the Monte Carlo (MC) simulator. Here MC simulation is set to terminate when 99.9% confidence that $p < 0.05$ has been obtained (UniODA and MegaODA software have identical capability). The PRUNE command specifies use of Sidak-based experimentwise pruning at the specified Type I error rate (p -value). The ENUMERATE command specifies an EO-CTA model is sought: eliminating this command obtains only the HO-CTA model; expressing this command obtains both the HO-CTA and EO-CTA models. The MINDENOM command specifies the minimum N that is required in every endpoint of the model. Automated CTA software required a total of 4 CPU seconds to conduct the HO-CTA analysis, and an additional 48 CPU seconds to conduct the EO-CTA analysis, when run on a 3 GHz Intel Pentium D microcomputer.

The HO-CTA model identified automatically using CTA software, identical to the HO-CTA model manually identified using UniODA (MegaODA) software, is presented as Figure 10.7 in Chapter 10 (p. 240).

The EO-CTA model identified automatically using CTA software is presented as Figure 11.1, and Table 11.1 presents the confusion table for this model (note that the sample is reduced to $N = 748$ due to missing data). As seen, when the model predicted a recommended likelihood score of 3 a total of $N = 103$ observations were misclassified, and when the model predicted a recommended likelihood score of 4 a total of $N = 92$ observations were misclassified. The model correctly classified 57.2% of the class category 3 observations, and 80.7% of the class category 4 observations: ESS = 37.9.

Table 11.1: Confusion Table for EO-CTA Model

		Predicted Recommendation	
		3	4
Actual	3	123	92
	4	103	430

Developed using this EO-CTA model, Table 11.2 presents the staging table for predicting the self-rated likelihood of a patient recommending the ED to others: Stage is an ordinal index and p_{rec} is a more granular ordered index of the self-rated likelihood of patient recommendation.

Figure 11.1: EO-CTA Model

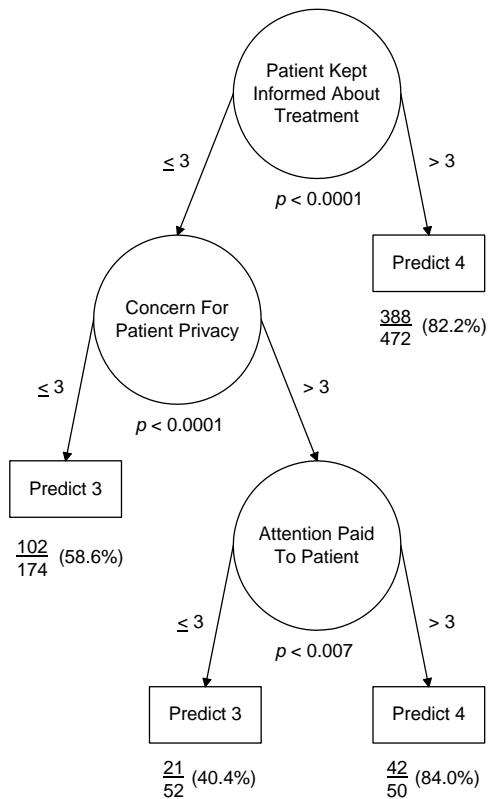


Table 11.2: Staging Table for Predicting Likelihood of Recommending ED to Others

Stage	Informed About Treatment	Concern for Patient Privacy	Attention Paid To Patient	N	p _{recom}	Odds
1	≤ 3	> 3	≤ 3	52	0.404	2:3
2	≤ 3	≤ 3	----	174	0.586	3:2
3	> 3	----	----	472	0.822	9:2
4	≤ 3	> 3	> 3	50	0.840	5:1

Note: p_{recom} = self-rated likelihood of recommending ED to others,
Odds = self-rated odds of recommending ED to others.

The attribute importance in discrimination (AID) statistic is conceptually similar to the partial R^2 statistic in regression analysis: both statistics indicate the incremental importance of every attribute in the model with respect to predicting the value of the class variable. The most important attribute is the root node—nurse informed patient about treatment: this attribute was used in predicting the class category status of every observation ($AID = 100\%$). The second-most-important attribute was concern for patient privacy, instrumental in classification of $[(174 + 52 + 50) / 748 \times 100\%] = AID = 36.9\%$ of the observations. The last, least instrumental attribute was attention paid to patient: $[(52 + 50) / 748 \times 100\%] = AID = 13.6\%$ of the observations.

The most important substantive revelation of this EO-CTA model is the importance of the nurse keeping the patient informed about treatment: for $472 / 748 = 63.1\%$ of the sample, 4 of 5 patients rating

this attribute as good or very good were likely to recommend the ED to others. For the remaining 36.9% of the sample rating this attribute as fair or worse, $(42 + 21) / 102 = 62\%$ were likely to recommend the ED to others if the nurse's concern for privacy was rated as good or very good. Actionable behaviors such as these should be emphasized in an effort to maximize positive patient recommendations of the ED.

It is informative to consider similarities and differences between the 3-attribute EO-CTA model ($ESS = 37.9$; Figure 11.1) and the less complex 2-attribute HO-CTA model ($ESS = 35.4$; Figure 10.7).

With respect to *similarities*, the EO-CTA and HO-CTA models include two of the same attributes—concern for patient privacy and attention paid to patient—each of which has the same optimal cut-point (i.e., 3) in both models. In addition, values > 3 for both attributes produce nearly identical predictive values for the deepest right-hand endpoint in both models (84.0% for the EO-CTA model, 83.5% for the HO-CTA model), even though this endpoint involves markedly different N s for the two models: 42 / 50 in the EO-CTA model, 359 / 430 in the HO-CTA model (corresponding to an endpoint N roughly 8.5 times greater in the HO-CTA model).

With respect to *differences* between the two models, although the EO-CTA and HO-CTA models include two of the same attributes, these two attributes appear in opposite order in the two models: in the EO-CTA model concern for patient privacy enters before attention paid to patient, whereas in the HO-CTA model attention paid to patient enters before concern for patient privacy. Furthermore, in the EO-CTA model this combination of concern for patient privacy and attention paid to patient is relevant only for the patients who were relatively dissatisfied with how well they were informed about their treatment; in the HO-CTA model this same two-attribute combination (albeit in opposite order of entry) constitutes the full tree model. Thus, for this particular set of attributes, the EO-CTA model qualifies the HO-CTA model by clarifying that the interaction of nurse attention and nurse concern for privacy in predicting likelihood of recommending the ED to others is most applicable to patients who are less satisfied with the degree to which the nurse kept them informed about their treatment.

HO-CTA versus EO-CTA Models

A few studies compare HO-CTA versus EO-CTA models developed for the same application using identical data (e.g., Chapter 9, Exploratory Methods).

In-Hospital Mortality from *Pneumocystis cariini* Pneumonia (PCP)

The first direct comparison of HO-CTA and EO-CTA involved competing severity-of-illness models for staging risk of in-hospital mortality from *Pneumocystis cariini* pneumonia (PCP) in the early AIDS era.²⁶ Research involved a sample of $N = 1,339$ patients hospitalized with HIV-associated PCP between 1987 and 1990—when hospital mortality rates were as high as 60%. The initial HO-CTA model (the first published in the field of medicine, before maximum- ESS pruning was discovered) used five nodes to correctly classify 34.1% of $N = 205$ patients who died, and 87.0% of $N = 988$ living patients ($N = 146$ patients were missing data on some attributes in the model). The relatively weak $ESS = 21.2$ nevertheless represented an order-of-magnitude gain in ESS versus the best prior linear model (logistic regression), and more than doubled the ESS achieved by the best prior suboptimal tree model (regression-based recursive partitioning). Using three nodes the *pruned* HO-CTA model correctly classified 74.6% dead and 59.1% living patients, yielding moderate $ESS = 33.7$. Finally, the EO-CTA model (Figure 11.2) had 69.5% sensitivity and 70.1% specificity, yielding moderate $ESS = 39.7$. Analyses were completed in 278 CPU seconds using a 3 GHz Intel Pentium D microcomputer.

Research studying a sample of $N = 1,660$ patients hospitalized with HIV-associated PCP between 1995 and 1997—the period marking early adoption of non-nucleoside reverse transcriptase and protease inhibitors as HIV therapy (the Highly Active Antiretroviral Therapy or HAART Era), is considered next.²⁶ Using four nodes (wasting, $AaPo_2$ —used twice, and Albumin) the HO-CTA model correctly classified 59.4% of $N = 128$ patients who died, and 73.7% of $N = 1,066$ patients who lived ($N = 466$ patients had missing data for model attributes), yielding moderate $ESS = 33.1$. After pruning to maximize ESS , the resulting two-attribute model had 53.8% sensitivity (correct prediction of dead patients), 84.3% specificity (correct prediction of living patients), and moderate $ESS = 45.2$. Finally, an EO-CTA model was conducted allowing

a jackknife-unstable attribute to enter the model if it met the Bonferroni criterion for statistical significance and the jackknife *ESS* exceeded training or jackknife *ESS* of alternative attributes. To facilitate direct comparison of models, the three-attribute EO-CTA model was developed using only attributes selected by the manually derived model: wasting, $AaPo_2$, and Albumin. The EO-CTA model (Figure 11.3) had 65.4% sensitivity, 88.2% specificity, and a relatively strong *ESS* = 53.7.

Figure 11.2: EO-CTA Model for Predicting PCP Inpatient Mortality Prior to 1995

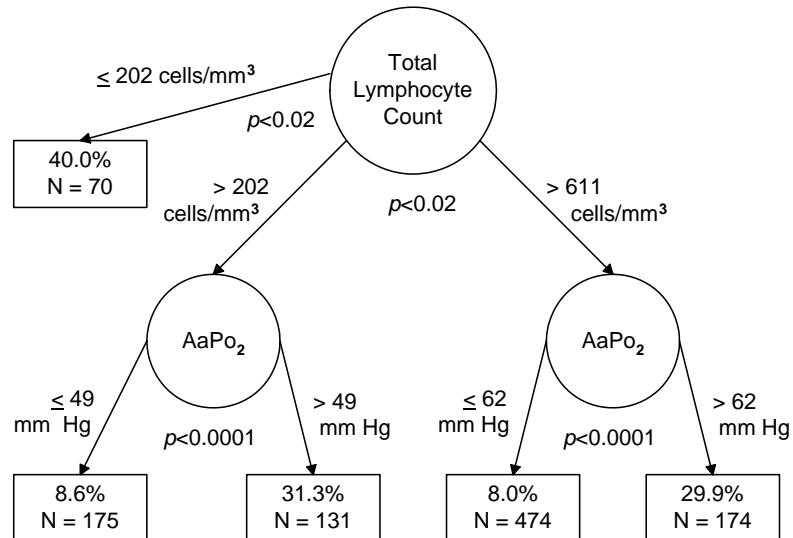
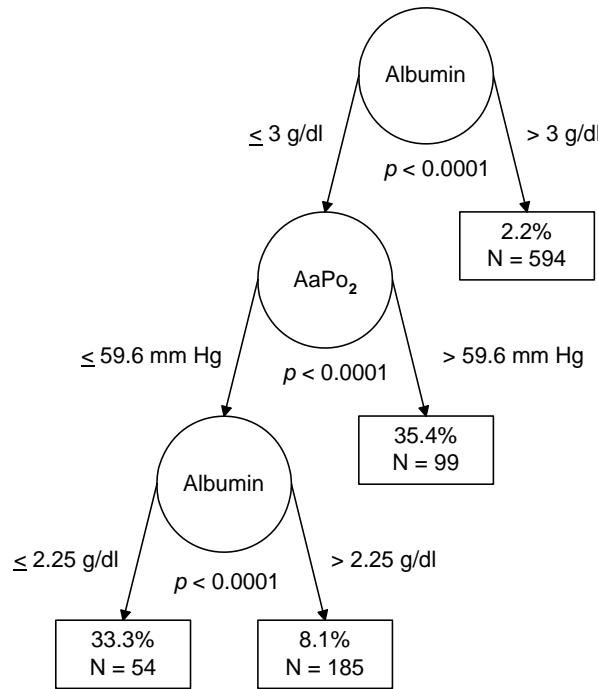


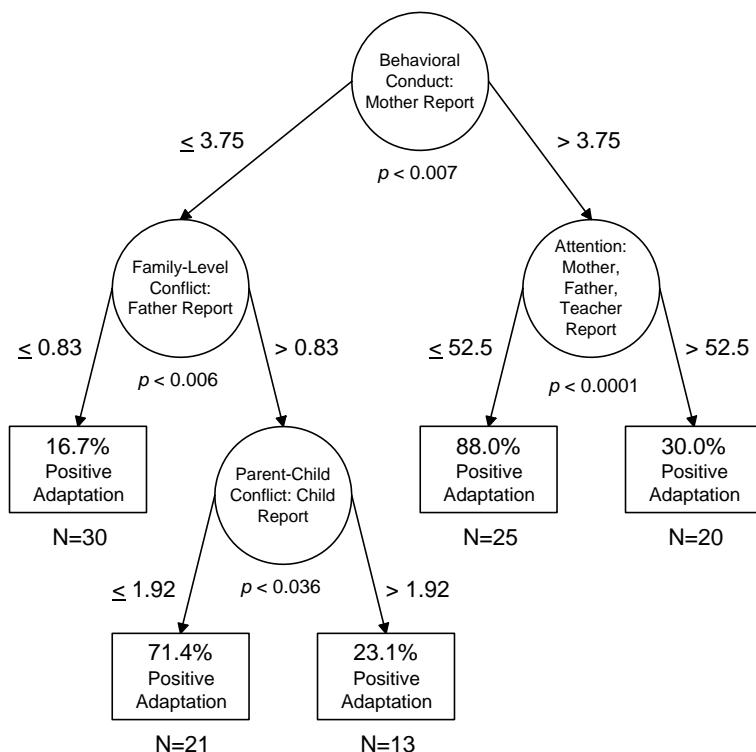
Figure 11.3: EO-CTA Model for Predicting PCP Inpatient Mortality After 1995



Psychosocial Adaptation in Early Adolescence

A prospective study of how individual- and family-level multimethod, multi-informant attributes predict psychosocial adaptation (i.e., scholastic success, social acceptance, positive self-worth) in early adolescence was conducted for a sample of $N = 68$ families of children with spina bifida and $N = 68$ comparison families of healthy children.²⁷ HO-CTA identified a five-mode model involving intrinsic motivation, estimated verbal IQ, behavioral conduct, coping style, and physical appearance to predict psychosocial adaptation in early adolescence: health status was not a factor in the model. The model achieved $ESS = 55.0$. An EO-CTA model was obtained next: attributes were only allowed to enter the model if their associated ESS was stable (did not diminish) in jackknife (leave-one-out) validity analysis. The EO-CTA model used four nodes and achieved $ESS = 57.0$ (Figure 11.4).

Figure 11.4: Enumerated CTA Model Predicting Psychosocial Adaptation in Young Adolescence



The EO-CTA model has several important similarities to the HO- CTA model. First, neither health status (spina bifida vs. able-bodied) nor socioeconomic status emerged as factors in either model. This suggests that both CTA models were able to identify factors that were more predictive of psychosocial adaptation than the group differences often identified in pediatric research. Second, the factor “behavioral conduct in the classroom” emerged in both models. This demonstrates consistency between the models and reinforces the relationship between behavioral control in the classroom and psychosocial adaptation. There were also important differences between the two models. Counter to original hypotheses, the HO-CTA model didn’t identify any family-level variables, nor did it include any variables based on mother or father report. In contrast, the EO-CTA model supported the original hypothesis by using two family-level variables in the model and including three variables based in part on mother and father report. Another difference between the two models is that in the HO-CTA model all of the factors were based on characteristics of the child and two of the factors represented more internalized child qualities (i.e., intrinsic motivation, coping style). In comparison, only half of the EO-CTA model attributes focused on child factors and these included only externalized or observable behaviors (i.e., conduct, attention).

In summary, the EO-CTA model presents a more parsimonious way of classifying this sample and supports the researchers' original hypotheses by including family-level factors and information from multiple informants (parents, teachers, child). However, it identifies a substantially different constellation of factors in the classification of psychosocial adaptation as compared to the HO-CTA model. Many theoretically important factors that emerged in the HO-CTA model that are well supported in pediatric research on psychosocial adaptation (e.g., motivation, IQ, coping style, and attractiveness) were not included in the EO-CTA model. Instead, the EO-CTA model selected a narrower constellation of factors that was highly focused on behavioral presentation and family-level conflict. These models likely represent two theoretically viable and empirically supported paths to psychosocial adaptation.

Person-Environment Fit Theory and Freshman Attrition

Person-Environment (PE) fit theory was used to explore the relationship between student involvement and freshman retention.²⁹ Incoming freshmen ($N = 382$) were followed longitudinally in a two-wave panel study, the summer before beginning college, and again during spring of their freshman year. Involvement levels, a variety of summer and spring preferences, and spring perceptions regarding specific aspects of the college environment were assessed. Twelve PE fit indicators were derived and compared with respect to their relationship with student involvement and retention. Parametric linear discriminant analysis was compared with HO-CTA to identify the most accurate classification model for use in designing potential attrition interventions. Parametric analysis yielded a weak effect ($ESS = 5.1$), but was 14% more accurate than HO-CTA in classifying returners (97% vs. 85%). HO-CTA yielded a relatively strong effect ($ESS = 68.8$), and was 962% more accurate than parametric analysis in classifying dropouts (8% vs. 84%). The HO-CTA model identified nine student clusters—five of returners and four of dropouts, revealing that the different subgroups of freshmen chose to return (and to stay) for different reasons. Findings suggest students' end-of-the-year preferences are more important influences upon retention than are anticipated preferences, college perceptions, or PE fit levels. EO-CTA was then conducted²⁵ and the model correctly classified 96% of returners, 87% of dropouts, and yielded strong $ESS = 82.4$. However, the EO-CTA model incorporated 14 attributes rather than eight as used in the HO-CTA model, and thus was more complex.

The HO-CTA and EO-CTA models share several important similarities. In particular, both models include predictors of drop-out that reflect less desire to identify oneself as a member of the university, develop socially and spiritually, connect with faculty outside class, and work in a challenging and competitive academic arena. These points of convergence demonstrate generalizability between the models and reinforce prior evidence suggesting these attributes are related to attrition. There were also important differences between the two models. In the EO-CTA model five attributes are pretest variables that reflect what students said they were hoping to find at Loyola, whereas the HO-CTA model included no pretest attributes. The EO-CTA model thus provides a more effective means for policy makers to target students prospectively who are at risk of dropping out.

At the time this research was conducted statistical power analysis focused on overall sample size (this is true of all parametric multivariable and multivariate methods), rather than on the endpoint sample size, without consideration of model geometry in this regard (Chapter 3). Presently the N of sample strata (model endpoints) that were identified by the CTA models was approximately consistent between models, but varied considerably within model. For the HO-CTA model the largest endpoint strata ($N = 176$, 50.6% of classified sample) is 29-times larger than the smallest strata ($N = 6$, 1.7% of classified sample). For the EO-CTA model the largest endpoint strata ($N = 125$, 35.7% of classified sample) is 42-times larger than the smallest strata ($N = 3$, 0.9% of classified sample). Because of complexity in the models, few attributes in either CTA model influenced the classification decisions for a substantial minority of the total sample as evaluated using the A/D statistic. Considered from a methodological perspective, highly granular solutions may yield nearly perfect classification performance, but model endpoints with small N can be criticized on the basis of inadequate statistical power, and of the potential cross-generalizability of the model: small endpoint denominators leave little room for inconsistent results before effects found in training analysis vanish in validity analysis. This research suggested that systematic investigation is needed to understand the interplay between sample size, number of attributes, tree depth, minimum endpoint denominator, granularity, and the training and validity performance of models developed using CTA models.

Considered from a policy perspective in some applications the cost of misclassification is extreme and misclassifications are to be avoided to the fullest extent possible. Considering the lifetime of effort, achievement and sacrifice that typically precedes acceptance into college, and the opportunity loss to individual, family, school and society associated with attrition, using available resources to monitor and assist the $N = 47$ students (one in every eight incoming students) who are predicted to drop seems wholly appropriate. If attributes in the EO-CTA model are actionable on the part of the student or counselor, successful targeted efforts could aid in the circumvention of loss of 33 / 38, or 87% of the students (seven of eight) who would otherwise drop, together representing 9% (one in eleven) of the freshman class.

Parsing Attributes

Years spent using automated CTA algorithms to analyze a vast number and variety of data sets, addressing directional, non-directional, and weighted hypotheses, routinely yielded new substantive discoveries in all fields of application that the ODA laboratory explored. However, occasionally a new methodological discovery emerged. For example, CTA algorithms sometimes identified models including a *parse*—a node for which an ordered attribute had three or more emanating branches. This is demonstrated here.

The study used ordered questionnaire-based scores on past- (reminiscence), present- (savor the moment), and future-focused (anticipation) savoring beliefs to discriminate $N = 117$ extreme Type A and $N = 131$ extreme Type B college undergraduates.⁴ Comparing mean scores on the three subscales using Student's *t*-test, no statistically reliable effect emerged for scores on the reminiscence [$t(244) = 1.2, p < 0.25$], savor the moment [$t(246) = 0.7, p < 0.49$], or anticipation [$t(246) = 1.2, p < 0.23$] subscales.

UniODA statistical analysis was performed to investigate the independent associations between savoring belief subscales and A/B Type. For reminiscence a statistically reliable, ecologically weak effect emerged ($p < 0.04, ESS = 16.6$), that was stable in LOO validity analysis ($p < 0.007$). The UniODA model was: if reminiscence ≤ 5.93 (53rd percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals Type As had significantly higher reminiscence scores than Type Bs. The model correctly classified 56% of the Type Bs, and 61% of the Type As. The model was correct 62% of the time a prediction of Type B was made, and 55% of the time a prediction of Type A was made.

For savor the moment a statistically marginal, ecologically weak effect emerged ($p < 0.08, ESS = 14.7$), that was stable in LOO validity analysis ($p < 0.005$). The UniODA model was: if savor the moment ≤ 6.19 (77th percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals that Type As had marginally higher savor the moment scores compared to Type Bs. The model correctly classified 84% of the Type Bs, and 31% of the Type As. The model was correct 58% of the time that a prediction of Type B was made, and 63% of the time that a prediction of Type A was made.

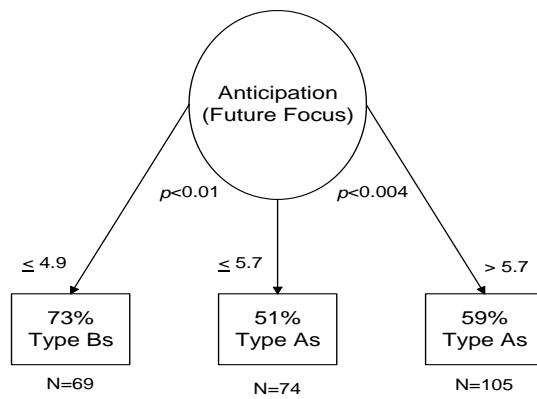
Finally, for anticipation a statistically reliable, ecologically weak effect emerged ($p < 0.003, ESS = 20.2$), that was stable in LOO validity analysis ($p < 0.002$). The UniODA model was: if anticipation ≤ 5.69 (58th percentile in the sample), then predict Type B; otherwise predict Type A. This model reveals Type As had significantly higher anticipation scores compared to Type Bs. The model correctly classified 67% of the Type Bs, and 53% of the Type As. The model was correct 62% of the time that a prediction of Type B was made, and 59% of the time that a prediction of Type A was made.

Figure 11.5 is the EO-CTA model obtained to discriminate A/B Type, treating the reminiscence, savor the moment, and anticipation subscale scores, and gender, as potential attributes: *ESS* = 24.1 (near the boundary between relatively weak versus moderate effect strength). Only the anticipation subscale emerged as a statistically significant attribute in the model, and a three-endpoint parse was identified. As seen, extreme Type B undergraduates are *more likely* (3:1 odds) than extreme Type As to score at *lowest levels* on the anticipation dimension of savoring beliefs: the cut-point 4.9 represents the 28nd percentile on this dimension for the sample. And, while A/B Types are *comparably likely* to score at *intermediate levels* on anticipation (1:1 odds), Type As are modestly *more likely* (3:2 odds) to score at *highest levels* on anticipation: the cut-point 5.7 represents the 58th percentile on this dimension for the sample.

Bivariate optimal results reveal an interesting pattern of differences between A/B Types in terms of their perceived ability to savor positive experiences retrospectively, concurrently, and prospectively. Concerning past-focused savoring, Type As reported a *greater capacity* than Type Bs to derive enjoyment by reminiscing about positive memories, contrary to the *a priori* hypothesis. Concerning present-focused

savoring, there was only a marginally significant A-B difference in the perceived capacity to savor the moment. Concerning future-focused savoring, UniODA revealed that Type As perceived higher capacity to derive enjoyment through anticipation relative to Type Bs, and EO-CTA revealed the specific thresholds of anticipation subscale scores that reliably discriminated As and Bs. In particular, significantly more Type Bs and fewer Type As scored below the 28th percentile on anticipation, and significantly more Type As and fewer Type Bs score above the 58th percentile on anticipation; whereas As and Bs were equally likely to fall between the 28th and 58th percentile on anticipation. Thus, while UniODA analysis is consistent with the *a priori* hypothesis, EO-CTA analysis provides strong evidence to support the *a priori* hypothesis. In sum, Type As, relative to Type Bs, believe they are more capable of enjoying positive memories through reminiscence and marginally more capable of enjoying positive moments; and are less likely to report a lower capacity (< 28th percentile) and more likely to report a higher capacity (> 58th percentile) to derive joy through anticipation.

Figure 11.5: EO-CTA Model Discriminating A/B Type Using Three Savoring Belief Dimensions



The difference between the results of the UniODA and EO-CTA analyses of anticipation for Type As and Type Bs highlights the potential benefit of considering nonlinear effects when testing research hypotheses. The UniODA model reflects the cut-score on anticipation that produces the highest possible accuracy in classifying As and Bs when selecting a single cut-point to discriminate these groups on the basis of anticipation. The EO-CTA model, in contrast, represents the combination of reminiscence, savoring the moment, and anticipation subscale scores that produces the highest possible accuracy in classifying As and Bs. The three-endpoint parse that emerged in the EO-CTA model reveals that the hypothesized A-B difference in the capacity to anticipate exists at the lower and upper range of the Anticipation subscale, but not in the middle range of the subscale. Whereas 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. Thus, the EO-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. Clearly, researchers would be wise to examine the possibility of nonlinear effects in testing bivariate relationships, in order to avoid missing important and informative research conclusions. CTA is the only statistical methodology available which is capable of identifying *explicitly optimal* parsed models such as the model which was obtained presently.

Modeling Moderating Effects

The most appropriate methodology for identifying inter-attribute interactions is obviously a non-linear method such as CTA that is specifically engineered to identify inter-attribute interactions. HO- and EO-CTA are the only non-linear methods that *explicitly maximize* the (weighted) classification accuracy (either *PAC* or *ESS*) that is achieved by the model for the sample. If there are no interactions, then CTA will explicitly identify the most accurate linear model. CTA requires no distributional assumptions (p are exact), may be used with any combination of attribute measurement scales (e.g., dichotomous, multicategorical, ordinal, integer, ratio), and present intuitive models in which parameters are expressed in their original units.

Enigma

Readers approach the end of a long analytic odyssey unmasking a scientific ecology that is being strangled by the universal, reflexive application of a plethora of inextricably confounded legacy statistical traditions. This statistical safari further revealed that data are instantly reanimated if they are examined in the light of a new universally-adaptable paradigm: seemingly magical transformations expose otherwise invisible phenomena; exact methods remove the specter of unsatisfied assumptions; and simple, transparent algorithms routinely obtain the most accurate and theoretically coherent statistical models ever identified in numerous empirical applications. Nevertheless, some perplexing findings and yet unanswered questions signal that not all is yet ideal in the ODA statistical oasis.

For example, for applications involving a single attribute it is possible for CTA to identify a model having three or more endpoints that yields greater *ESS* than the corresponding UniODA model with two endpoints. In other words if two groups (class categories) are compared on one attribute via CTA, it is possible that more than one “type” of one (or each) class category exists in the sample data: that not all males are identical, not all Type A undergraduates are identical, and so forth? Is UniODA the simplest (limiting minimal) case of a CTA analysis?

It isn’t unusual for ODA models to have *ESS* values that lie “near” (assessed using eyeball analysis) to rule-of-thumb criteria for qualitative effect strength levels, such as a relatively weak effect ($ESS < 25$), a moderate effect ($25 \leq ESS < 50$) and so forth. A conceptually related issue is how to evaluate if *ESS* values of competing alternative models are similar, or reliably dissimilar. Can an exact confidence interval methodology unambiguously characterize the classification performance parameters of ODA models?

For a given application the *ESS* for an EO-CTA model often exceeds (but, of course, is never less than) the *ESS* obtained by an HO-CTA model. The EO-CTA algorithm only enumerates the first three nodes of the model, suggesting the possible existence of even more accurate models. Is it possible to explicitly determine the theoretical maximum level (“upper limit”) of *ESS* obtainable for an empirical application?

In CTA analysis the minimum endpoint *N* is established on the basis of statistical power analysis and then a maximum-accuracy (*PAC* or *ESS*, depending on the research objective) model is identified that satisfies the minimum *N* constraint. If the resulting model isn’t theoretically cogent then an alternative CTA model is needed that also satisfies the minimum *N* constraint: can all of the unique models that exist be identified for a given application?

Finally, as endpoint minimum *N* decreases, and as number of model endpoints increases, model *ESS* typically increases. Perfect $ESS = 100$ is of course the ultimate objective of a classification algorithm, but cross-generalizability (reproducibility) of the model when applied to classify an independent random sample is a collateral ultimate objective.³⁰ *ESS* is a measure of predictive accuracy, and number of model endpoints is a measure of model complexity—the opposite of parsimony. Decision-makers responsible for outcomes would probably prefer a model that performs moderately in training and validity analysis versus a model that is accurate in training analysis but performs poorly in validity analysis. If more than one model exists for a given application is there an algorithmic method to select the model that represents the best combination of accuracy and parsimony? Is it possible to compare models of different outcomes vis-à-vis a normed measure of their distances from corresponding theoretically ideal models?

The short answer to all of these questions is “yes”.

Chapter 12

Globally Optimal Statistical Analysis

In 2014 a third-generation maximum-accuracy CTA methodology was discovered, called *globally-optimal* CTA (GO-CTA). Statistically motivated by *novometric theory*, GO-CTA constitutes a conceptual counterpart to quantum mechanics for classical data.^{1,2}

Defining a Theoretically Ideal Statistical Model

Few statistical classification models reported in the literature yield (nearly) perfect prediction of the class variable, no matter how many attributes are in the model: mediocre classification accuracy is the norm. Perhaps attributable to their scarcity, the character of a (nearly) perfect classification model is not widely discussed. Thus, as a first step toward improving the predictive accuracy of statistical models it is helpful to characterize the nature of an excellent statistical classification model.¹

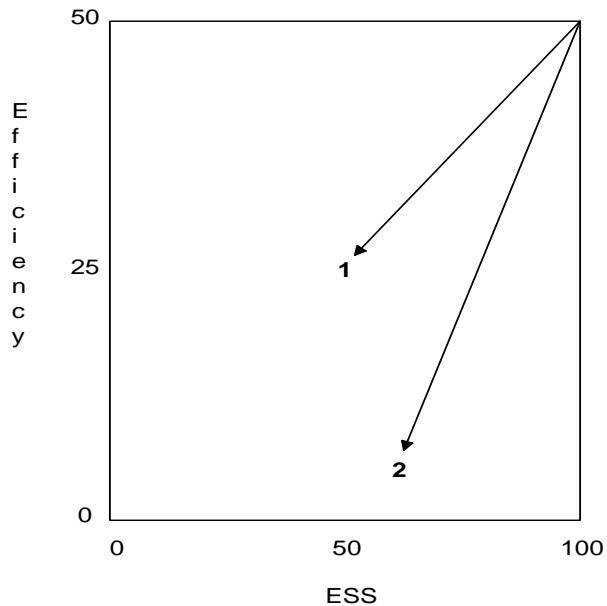
Simply stated, an excellent statistical classification model explains a class variable *accurately* and *parsimoniously*. In the ODA paradigm predictive accuracy is assessed using the *ESS* statistic. Regardless of the number of levels constituting the class variable, of attributes used in the model or their measurement metrics, or the sample size, *ESS* is a *normed* statistic: for every statistical classification problem $ESS = 0$ is the level of classification accuracy expected by chance for the application, and $ESS = 100$ represents errorless classification. And, in the ODA paradigm, parsimony is assessed using a statistic known as *efficiency*, defined as *ESS* divided by the number of distinct strata—the number of endpoints that are identified by the classification model. As a means of illustrating these concepts, consider three statistical discrimination problems each involving a class variable having two levels, for example gender: males versus females.

First, imagine a model that perfectly predicts the gender of all observations in the sample. If the model separated (parsed) the observations into two strata—one consisting entirely of females and the other entirely of males, then the classification model would achieve $ESS = 100$ and $efficiency = 50$ (a solution with two strata can only occur for a model involving a single attribute and two emanating branches). Figure 12.1 illustrates the *ESS*-by-*efficiency* space: the perfect accuracy of the two-strata model is located (illustrated) in the extreme upper right-hand corner. Consider two different less-than-perfect two-strata ODA models, indicated as 1 and 2, with different *ESS* and *efficiency* values. As seen, the distance of model 1 from the perfect solution (indicated by an arrow) is less than the distance of model 2 from the perfect solution. Model 1 is therefore a better approximation of (is closer to) the perfect solution, and is thus superior to model 2—that is a worse approximation of (is further from) the perfect solution. If these two ODA models are the only models that existed for this application then model 1 would be selected as the GO-CTA model in this application.

Next, imagine a perfect classification model separating observations into three strata, with at least one strata consisting entirely of males, another consisting entirely of females, and the third strata consisting entirely of either males or females. The perfect classification model would achieve $ESS = 100$ and $efficiency = 33.3$ (a solution with three strata can only occur for a model with one or two attributes). Figure 12.1 also illustrates the *ESS*-by-*efficiency* space, in which the perfect three-strata model is located in the extreme upper right-hand corner of the space (in Figure 12.1 change the *efficiency* values of “50” to “33.3”, and “25” to “16.6”). Consider two different less-than-perfect three-strata ODA models, indicated as 1 and 2, having different *ESS* and *efficiency* values. As seen, the distance of model 1 from the perfect

solution is less than the distance of model 2 from the perfect solution. Model 1 is a better approximation of the perfect solution, and thus is superior to model 2, which is a worse approximation of the perfect solution. If these two ODA models were the only models possible for this specific application, then model 1 would be the GO-CTA model in this application.

Figure 12.1: Two-Strata Models



Finally, imagine a model separating observations into four strata, with at least one strata consisting entirely of males, another entirely of females, and the third and fourth strata each consisting entirely of either males or females. The classification model would achieve $ESS = 100$ and $efficiency = 25$ (a solution with four strata can only occur with a model having one, two, or three attributes). Figure 12.1 also illustrates the ESS -by- $efficiency$ space, with the perfect four-strata model located in the extreme upper right-hand corner of the space (in Figure 12.1 change the efficiency values of “50” to “25”, and “25” to “12.5”). Consider two different less-than-perfect four-strata ODA models, indicated as 1 and 2, having different ESS and $efficiency$ values. As seen, the distance of model 1 from the perfect solution is less than the distance of model 2 from the perfect solution. Model 1 is thus a better approximation of the perfect solution, and therefore is superior to model 2—a worse approximation of the perfect solution. If these two ODA models were the only models possible for this specific application, then model 1 would be the GO-CTA model in this application.

Comparing Empirical and Theoretically Ideal GO-CTA Models

Comparing the quality of an empirical model to a corresponding theoretically ideal model is intuitively conceptualized as the Euclidean distance of the empirical result from the upper right-hand corner of a unit square Cartesian space defined by two orthogonal axes, with predictive accuracy (ESS) used as the abscissa, and parsimony—quantified as ESS divided by the number of *strata* (endpoints) in the model—used as the ordinate. However, this conceptual perspective is unproductive as a means of computing the distance between an empirical and a theoretically ideal model because if an interactive transformation (Chapter 2) is used to obtain a unit *efficiency* scale separately for problems of varying complexity (number of *strata*), then all of the models in the *descendant family* (discussed ahead, this is the set of all possible models that exist for the application) lie along the proper diagonal between chance (0,0) and the ideal model (1,1). The distance between empirical and theoretically ideal model in this approach is a perfect function of ESS . Thus, interactive transformation used to standardize *efficiency* into unit scale separately

by number of *strata* is necessary to obtain a unit square space for models differing in complexity, but this thereby eliminates the role of model complexity.³

Accordingly, the distance of an empirical model from a theoretically ideal statistical classification model is defined as the *number of additional equivalent effects* that are needed in order to obtain perfect classification for the sample.³ For example, imagine a 3-strata model achieved $ESS = 75$, with *efficiency* = $75 / 3 = 25$. If *one additional attribute* is identified yielding an equivalent effect of $ESS = 25$, then $ESS = 100$ and an ideal model for this sample is obtained.

The distance of an empirical classification model from a theoretically ideal statistical classification model is easily computed (*strata* is number of strata in the model): $Distance (D) = [100 / (ESS / strata)] - strata$. This definition considers both accuracy (*ESS*) and parsimony (*strata*) in computing the distance of an empirically-obtained model from the corresponding theoretically ideal model for any given sample and application. In the example above, $D = [100 / (75 / 3)] - 3 = (100 / 25) - 3 = 1$.

Consider comparing two different CTA models identified in the same application. For example imagine model 1 has $ESS = 50$ and *strata* = 3, and model 2 has $ESS = 60$ and *strata* = 4. Here $D = 3.00$ for model 1 and $D = 2.67$ for model 2: because D is smallest for model 2, model 2 is the GO-CTA model.

Finally, consider comparing GO-CTA models obtained in two independent applications (the class variable in the applications needn't be the same). For example imagine model 1 has $ESS = 35$, *strata* = 3, and $D = 5.57$, and model 2 has $ESS = 45$, *strata* = 4, and $D = 4.89$. Having the smallest D statistic, model 2 is closest to a theoretical ideal model and thus is the "best" GO-CTA model.

Novometric Theory

Novometrics means new (Latin: *novo*) measurement and connotes the algorithm used to explicitly identify a GO-CTA model.¹ For a given application the novometric algorithm first obtains the *descendant family* of all optimal models that exist between the *class* variable and the *attribute(s)*, and then identifies the model with the minimum D statistic as being the GO-CTA model. After the four axioms that underlie novometric theory are introduced, similarities and differences of axioms of quantum mechanics and novometrics are considered, and the algorithms described in second and third novometric axioms are demonstrated.

Axiom 1: Statistical Sample

As used herein, a *statistical sample* (*sample*) has N or more observations per model endpoint—providing a minimally sufficient level of statistical power to enable EO-CTA to test the confirmatory or exploratory hypothesis that the class variable is accurately predicted by the attribute(s): the null hypothesis is that the attribute(s) don't predict the class variable. The data required for each observation include *class* category, score on each *attribute* (however, depending on model geometry, not every attribute must necessarily be used to classify each observation), and *weight* (if no weight is specified then a unit weight of 1 is used so that all observations are equally-valued). Data for serially-recorded attributes are ipsatively standardized (Chapters 2 and 3).

Axiom 2: Structural Decomposition Analysis

In multiattribute applications the attribute subset producing the GO-CTA model for the *sample* is identified using *structural decomposition analysis* (SDA), a procedure for identifying attributes that successively maximize *ESS* for monotonically diminishing partitions of the *sample*.⁴ Conceptually SDA is analogous to principal components analysis (PCA): SDA explicitly maximizes predictive accuracy, whereas PCA explicitly maximizes explained variance.⁵ SDA isn't used in analyses involving a single attribute.

In step one of SDA, using EO-CTA the attribute yielding the minimum D statistic, with exact $p < 0.05$, is selected (selected attributes are omitted from latter steps). In step two all observations that were correctly classified using the EO-CTA model in step one are deleted from the *sample*, leaving only the misclassified observations. The attribute with the minimum D statistic and $p < 0.05$ in the reduced sample

is selected. This procedure is continued until all observations in either of the class categories are correctly classified, $p > 0.05$, or an insufficient number of observations remain to satisfy Axiom 1.

Axiom 3: Descendant Family

The GO-CTA model for the *sample* lies within the descendant family of models obtained by applying the *minimum denominator search algorithm* (MDSA) to an *initially-unrestricted* EO-CTA model configured to predict the class variable using only the attribute(s) selected by SDA. An initially-unrestricted CTA model is obtained when automated analysis is conducted vis-à-vis CTA statistical software using syntax that *doesn't restrict* algorithm freedom regarding admissible combinations of model characteristics, such as maximum branch depth, maximum number of attributes, or minimum endpoint sample size.⁶

EO-CTA models terminate in endpoints that constitute different unique strata, and each endpoint represents some number of observations. Represent the minimum number of observations obtained for an endpoint for an initial, unrestricted EO-CTA model in step one of the MDSA as N_1 . The second EO-CTA model in the descendant family is obtained by constraining the size of the smallest strata (the minimum denominator) to $N_1 + 1$, and CTA model i in the descendant family is found by constraining the size of the smallest strata to $N_{i-1} + 1$. MDSA terminates if a CTA model is no longer feasible.

Axiom 4: Model Reproducibility

Validity analyses conducted via hold-out, leave-one-out (one-sample jackknife) or bootstrap methods for static data, or by test-retest methods for dynamic (repeated-measures) data, provide an estimate of the cross-generalizability of D , ESS , ESP , and other classification performance indices for the GO-CTA model.⁷

Novometric Theory and Quantum Mechanics

It is important to understand conceptual consistencies and to distinguish between novometric theory and quantum mechanics (QM), a field of physics that derives from the finding that some physical phenomena change in discrete amounts (Latin: *quanta*), rather than along a continuum.⁸ In QM amount of information is measured by von Neumann entropy, a generalization of classical information theory (a mathematical model of communication) applied to quantum phenomena.⁹ A popular unit of quantum information is a binary system known as a *qubit*, and the information content of a message is measured in terms of the minimum number of binary models (n qubits) required to store the message. This is analogous to a *bit* in binary-log classical information theory, where the mean number of bits needed to store or communicate one symbol in a message is called entropy.

A fundamental premise of the ODA paradigm and novometric theory is that classical phenomena are fundamentally discrete in nature, and no assumptions are made regarding hypothetical underlying parent distributions for the class variable or attribute(s).⁴ For example, rather than modeling the ordered attribute temperature of water, an ODA model might employ either of the two critical thresholds (one for freezing, one for evaporating) that define qualitative states, or might discover another critical threshold value for predicting some other qualitative outcome (e.g., a chemical reaction). Conceptually reminiscent of information theory, novometry involves the use of binary *parses* to create sample strata, and the information content of a model relating two variables (a class variable is modeled using one or more attributes) is measured in terms of the minimum number of strata (model endpoints) required to achieve the best combination of accuracy and parsimony in the classification of the phenomenon.

In QM the *Particle in a Box* model of energy held in a confined space is conceptually analogous to the premise of sample strata in novometry. According to this model the energy of a particle in infinite space has continuous solutions, but as constraints are imposed, such as physical confinement, discrete solutions occur which represent the only possible solutions. For single confined particle systems these solutions represent discrete energy levels. In novometric theory the size of the box corresponds to the number of observations (N) in the sample (the population corresponds to the universe), and the discrete measurement levels correspond to sample strata: patient strata for classifying disease incidence; market segments for classifying consumer preference; or storm categories for classifying drilling platform struc-

tural damage, for example. The set of all possible solutions (the *descendant family*) for a given application (sample) is identified in novometry using the MDSA. Table 12.1 compares axioms of QM and novometrics, and Table 12.2 provides a summary of intuitive concepts of these theoretical models.

Table 12.1: Axioms of Quantum Mechanics and Novometrics

Quantum Mechanics	Novometry
The state of a system S is represented by a normalized element of a vector \mathbf{v} of a Hilbert space \mathbf{H} which is complete (calculus applies).	A random <i>sample</i> consists of a <i>class</i> variable, one or more <i>attributes</i> , a <i>weight</i> (unit-weighted observations are equally-valued), and a number of observations N yielding at least minimally adequate statistical power for testing the hypothesis that the attributes predict the class variable. The null hypothesis is the attributes can't predict the class variable.
The possible outcomes of measurement of an observable \mathbf{O} are the eigenvalues of \mathbf{O} .	In applications involving two or more attributes, the specific subset of attributes that yield the GO-CTA model in the <i>sample</i> is identified by a <i>structural decomposition analysis</i> (SDA), an iterative application of EO-ODA conceptually analogous to principal components analysis, but that explicitly maximizes classification accuracy rather than variance.
The value of the measurement of an observable is one of the observable eigenvalues. The probability of obtaining a particular eigenvalue is given by the modulus square of the inner product of the state vector of the system with the corresponding eigenvector.	The GO-CTA model for the <i>sample</i> lies within the <i>descendant family</i> of models obtained by applying the <i>minimum denominator search algorithm</i> (MDSA) to an initially-unrestricted enumerated EO-CTA model configured to predict the class variable using only the attribute(s) selected by SDA. The MDSA enables the discovery of all possible EO-CTA models in the <i>sample</i> that originate from an initially unrestricted model.
If we repeat the experiment after the first measurement, we will obtain again the same result with probability 1. This is the well-known collapse of the wavefunction.	Validity analyses conducted by hold-out, LOO (jackknife), or bootstrap methods for static data, or by test-retest methods for dynamic data, provide an estimate of the cross-generalizability of D , <i>ESS</i> , <i>ESP</i> and other classification performance indices for the GO-CTA model. Exact confidence intervals are used to assess overlap and reliability of model and chance performance.

Table 12.2: Intuitive Interpretation of Quantum Mechanics and Novometrics

Quantum Mechanics	Novometry
QM derives from the finding that some physical phenomena change in discrete amounts (Latin: <i>quanta</i>), and not along a continuum. In QM information is measured by von Neumann entropy, a generalization of classical information theory (a mathematical model of communication) applied to quantum phenomena. A popular unit of quantum information is a binary system called a <i>qubit</i> , and the information content of a message is measured in as the minimum number of binary models (n qubits) required to store the message. This is analogous to a <i>bit</i> in binary-log classical information theory, where the mean number of bits needed to store or communicate one symbol in a message is called <i>entropy</i> .	Novometry derives from the observation that for every S involving classical phenomena, results of all classification methodologies assign observations to discrete strata. In novometry the information (accuracy) obtained by a model is measured using the normed <i>ESS</i> statistic. The number of <i>strata</i> needed to define the model is identified using binary parses. The complexity of the model is measured using the <i>efficiency</i> statistic. The distance of an empirical model from the corresponding theoretically ideal model is measured using the D statistic.

In QM the *Particle in a Box* model of energy held in a confined space is conceptually analogous to the premise of sample strata in novometry. According to this QM model the energy of a particle in infinite space has continuous solutions, but as constraints are imposed, such as physical confinement, discrete solutions occur which represent the only possible solutions. For single confined particle systems these solutions represent discrete energy levels.

In novometric theory the size of the box corresponds to the amount of data (N) in the sample (the population corresponds to the universe), and the discrete measurement levels correspond to sample strata. The set of all possible solutions (the *descendant family*) for a given application (sample) is identified in novometry using the MDSA.

Exact Discrete Confidence Intervals

Parametric statistical classification methods use 95% confidence intervals (CIs) to assess whether model parameters and performance indices overlap chance ("error")—thus indicating the absence of statistical reliability. Although chance performance is represented by the value zero, statistical models rarely attain a result of zero when applied to random data, especially for smaller samples.^{10,11}

In novometry 95% CIs are also used to assess the overlap and statistical reliability of CTA model performance. In contrast to parametric methods however, a fundamental premise of novometric theory is that classical phenomena, including optimal CIs, are fundamentally discrete (are not continuous), and no assumptions are made concerning hypothetical underlying parent distributions. For novometric analysis exact discrete 95% CIs are developed using Fisher's randomization procedure for the EO- and GO-CTA models, and using a bootstrap methodology for chance^{4,12} (see Chapter 2).

The development of exact discrete 95% CIs in novometric analysis for model- and chance-based classification performance is illustrated using an application investigating the relationship between the temperature of an Emergency Department (ED) patient and the disease status [community-acquired pneumonia (CAP) or influenza-like illness (ILI)] of the patient.¹² MDSA identified a descendant family of two optimal solutions for the sample of $N = 200$ patients, and findings for the GO-CTA model identified in the second step of the MDSA are summarized in Table 12.3.

Table 12.3: Summary of Second Step of MDSA Procedure for Discriminating CAP and ILI Patients

Step	Strata	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
2	2	88	41.4	20.7	2.83
			26.0-56.5	13.0-28.2	5.69-1.54
			0.19-14.2	0.10-7.12	1051-12.1

Table 12.4: Model Bootstrap

Quantile	Estimate
100% Max	74.85
99%	62.60
95%	56.48
90%	53.25
75% Q3	48.99
50% Median	41.67
25% Q1	35.50
10%	29.58
5%	25.96
1%	19.65
0% Min	7.39

Table 12.5: Chance Fisher's Randomization

Quantile	Estimate
100% Max	31.13
99%	18.76
95%	14.25
90%	12.19
75% Q3	8.06
50% Median	4.32
25% Q1	2.25
10%	0.19
5%	0.19
1%	0.19
0% Min	0.19

Bootstrap methodology using 10,000 iterations of a 50% resample with replacement⁴ is used to obtain the exact discrete 95% CI for model *ESS* for the GO-CTA model. The resulting estimates of *ESS* were sorted and cumulated, yielding the results given in Table 12.4. In Table 12.3 and Table 12.4 the 5% bound of the discrete 95% CI for the model is 25.96 (rounded as 26.0), and the 95% bound is 56.48 (rounded as 56.5).

For the chance effect 10,000 iterations of Fisher's randomization procedure is performed using a 50% resample with replacement, and the results are cumulated: as seen in Table 12.5 the 5% bound of the discrete 95% CI for the model is given as 0.19 (Table 12.3), and the 95% bound is rounded as 14.2 (Table 12.3).

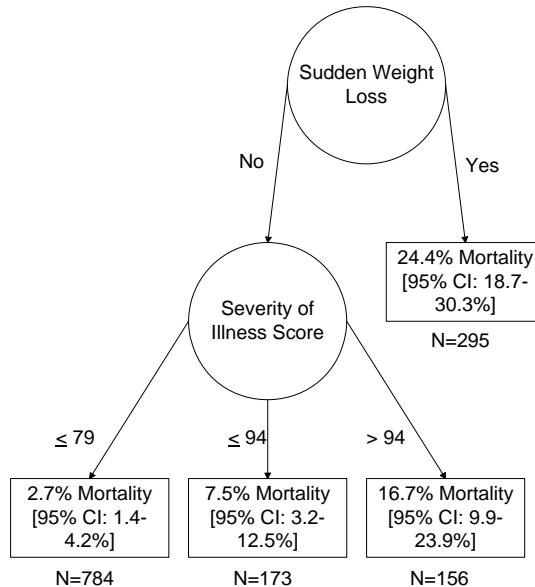
For this GO-CTA model the 95% CIs for classification performance clearly exceed the 95% CIs for chance performance in this application.

Model Endpoint Redundancy

For any optimal model in a descendant family it is possible that exact discrete 95% CIs for two or more model endpoints may overlap. When such redundancy occurs the overlapping endpoints fail to provide reliably unique estimates of outcome (e.g., *class category* = 1) variable prevalence.

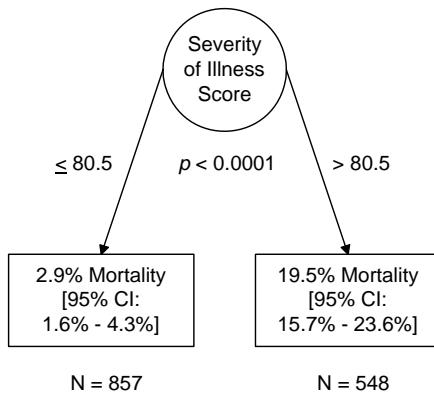
To illustrate such redundancy consider research predicting in-hospital mortality for a sample of $N = 1,660$ patients hospitalized with HIV-associated *Pneumocystis cariini* pneumonia.¹³ The next-to-final ($ESS = 46.58$, $D = 4.59$, Figure 12.2) and the final ($ESS = 46.42$, $D = 2.31$, Figure 12.3) optimal models within the descendant family are compared.¹⁴ Both models included an ordered measure of severity-of-illness, and the more complex next-to-final model included a binary indicator of whether the observation experienced sudden weight loss. As seen for the more complex model (Figure 12.2), counting from the left, the 95% CI of the first and second, second and third, and third and fourth endpoints overlapped, thereby indicating endpoint redundancy across the model domain.

Figure 12.2: Redundant Endpoints



In contrast, 95% CIs for the endpoints of the GO-CTA model (Figure 12.3) didn't overlap, and there is no evidence of redundancy: total N between the models differs due to missing data.

Figure 12.3: Non-redundant Endpoints



In applied settings involving decision-making the presence of non-redundant endpoints offers the advantage of representing strata with unambiguously different likelihood of membership in *class category* 1, so strata can thus be unequivocally ranked in terms of the outcome dimension. In theoretical settings in contrast, endpoint redundancy *must* occur for applications involving (nearly) ideal models having three or more predicted strata. In an ideal statistical model (Figure 12.1) the representation of *class category* = 1 observations in every endpoint is either exactly 100, or exactly 0. An ideal statistical model that involves E endpoints has $E - 2$ perfectly redundant endpoints. For example, an ideal two-strata model has $2 - 2 = 0$ redundant endpoints; an ideal three-strata model has $3 - 2 = 1$ perfectly redundant endpoint; and an ideal four-strata model has $4 - 2 = 2$ perfectly redundant endpoints (either one redundant endpoint from each class category, or two redundant endpoints from one of the two class categories).

Obtaining a GO-CTA Model: Binary Class Variable, One Attribute

Novometry is demonstrated for an application with a binary class variable and one ordered attribute. The data are drawn from the Surveillance, Epidemiology, and End Results (SEER) Program, which collects and publishes cancer incidence and survival data so as to assemble and report estimates of cancer incidence, survival, mortality, other measures of cancer burden, and patterns of care in the USA.¹⁵

In this example cancer incidence is parsed separately by gender (male, female) and by race (white, African American) to evaluate if these class variables can be used to identify discrete patient strata that differ in cancer incidence. Cancer incidence rate is the number of new cancers of a specific site (type) occurring in a specified population in one year, expressed as number of new cancers for every 100,000 population at risk. In SEER data the number of new cancers may include multiple primary cancers that occur in one patient; the primary site reported is the site of origin and not the metastatic site; and the population used in computation depends on the rate being calculated (e.g., for cancer sites occurring in one sex, the sex-specific population is used). SEER provides an age-adjusted rate weighted by the proportion of people in the corresponding age groups of a standard population, but raw data were used to avoid possible paradoxical confounding induced by linear-model-based adjustment.

All cancer categories in the SEER database were analyzed in the order they are provided.¹ Each analysis involved either $N = 608$ or $N = 304$ observations, as indicated in Table notes. Under proportional reduction in sample size over successive parses: for $N = 608$ one binary parse will create two strata, each having $N = 304$ observations; two binary parses will create four strata each with $N = 152$ observations; and three binary parses will create eight strata, each having $N = 76$ observations. For a two-tailed analysis with a binary class variable, an ordered attribute, and endpoints having $N = 76$ observations, moderate $ESS = 32.5$ (for $p < 0.01$) and $ESS = 28.7$ (for $p < 0.05$) are required for power of at least 90% (Chapter 3). There is adequate statistical power for CTA models involving eight strata for $N = 608$, and four strata for $N = 304$.

The MDSA algorithm was used to identify the descendant family of optimal models that underlie these data. In the first step of the analysis an unrestricted EO-CTA model was obtained using the following CTA software⁶ syntax (Appendix C):

```

OPEN seer.dat;                               MISSING all (-9);
OUTPUT seer.out;                            MC ITER 10000 CUTOFF .05 STOP 99.9;
VARS sex rate;                             PRUNE .05;
CLASS sex;                                 ENUMERATE;
ATTR rate;                                GO;

```

Table 12.6 summarizes the descendant family of three optimal models identified in the present sample, by using sex to parse cancer incidence data for *all cancer types combined*.

Table 12.6: Parsing Cancer Incidence by Sex: All Sites Combined

Cancer Site	Strata	MinD	ESS	Efficiency	D
All Sites	6	2	33.2	5.54	12.07
			25.4-41.2	4.22-6.87	17.62-8.56
			0.33-7.57	0.06-1.26	1812-73.3
	5	63	33.2	6.64	10.06
			25.3-41.2	5.05-8.23	14.76-7.14
			0.33-6.91	0.07-1.38	1510-67.4
	3	80	31.9	10.6	6.40
			22.8-40.6	7.60-13.5	10.16-4.39
			0.33-7.57	0.11-2.52	603-36.6

Note: Sex is male or female. Cancer Site is type of cancer. *Strata* is number of model endpoints. *MinD* (minimum denominator) is the smallest *N* for any strata. *Efficiency (ESS / Strata)* is a normed index of relative strength of the class variable(s) used in identifying sample strata. Results for each step of the MDSA are tabled. For each model the first line gives point estimates; the second line gives 95% CIs for discrete distributions obtained by bootstrap analysis; and the third line gives 95% CIs for chance obtained using Monte Carlo analysis involving 100,000 iterations. For this analysis *N* = 608.

For the first EO-CTA model that emerged the *ESS* point estimate (33.2), and the upper (41.2) and lower (25.4) bounds of the 95% CI for *ESS*, indicate a moderate effect. In contrast the 95% CI for chance-based *ESS* is relatively weak and lies well beneath the 95% CI for model-based *ESS*. Six patient strata were identified, the smallest of which had *N* = 2 observations: this endpoint minimum denominator is too small and violates Axiom 1, disqualifying this model on statistical grounds.

In step two of the MDSA the second EO-CTA model in the descendant family was identified by adding the following CTA software syntax (forcing the endpoint minimum denominator of the second model to be at least one observation larger than occurred in the prior first model) to the initial program:

```
MINDENOM 3;
```

A five-strata CTA model was identified (Table 12.6) yielding moderate *ESS* equivalent to that of the initial six-strata model as assessed by 95% CI overlap. The smallest strata for this model had *N* = 63.

In step three of the MDSA the third EO-CTA model in the descendant family was identified by replacing the following CTA software syntax (that forces the endpoint minimum denominator of the third model to be at least one observation larger than in the prior second model) to the initial program:

```
MINDENOM 64;
```

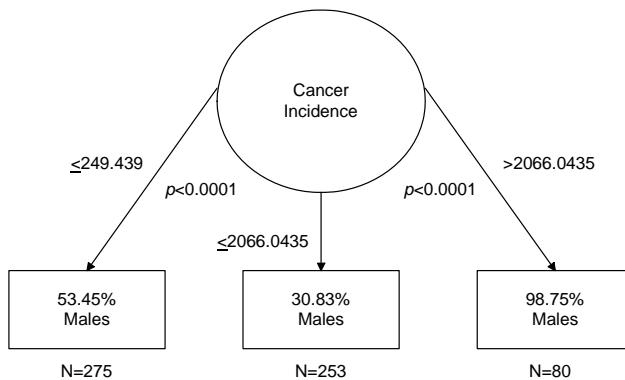
A three-strata CTA model was identified (Table 12.6) yielding moderate *ESS* equivalent to that of the initial six-strata model as assessed by 95% CI overlap. The smallest strata for this model had *N* = 63.

When the minimum denominator was set to $N = 81$, no EO-CTA model could be identified: the descendant family of optimal models for this application is thus completed, and it consists of three models.

A three-strata CTA model was identified (Table 12.6) yielding moderate *ESS* equivalent to that of the initial six-strata model as assessed by 95% CI overlap (the three-strata model had significantly greater efficiency than the six-strata model). The lower-bound of *ESS* for the three-attribute model was 22.8, falling into the qualitative category of relatively weak: the three-strata model thus represents a weak-moderate effect. Having comparable strength and greater parsimony and efficiency (as assessed via 95% CI or by point estimate) than other models in the descendant family, the *D* statistic indicates that the three-strata model is closest to the perfect prototype (Figure 12.1) and thus is selected as the GO-CTA model of the relationship between sex and all-site cancer incidence. The smallest strata for this model had $N = 63$. When the minimum denominator was set to $N = 81$ no EO-CTA model could be identified: the descendant family of optimal models for this application is thus completed, consisting of three models.

The three-strata GO-CTA model is illustrated in Figure 12.4. The strata having the lowest cancer incidence ($\leq 0.249439\%$) was approximately equally represented by males and females, and it comprised 275 / 608 or 45.2% of the total sample. In contrast, the strata having the highest cancer incidence ($> 2.0660435\%$) was dominated (98.8%) by males, and comprised 13.2% of the total sample. And, the strata having intermediate cancer incidence ($0.24944\% - 2.0660434\%$) was largely composed of females (69.2%) and comprised 41.6% of the total sample. Endpoints weren't redundant: exact discrete 95% CIs for the three endpoints from left-to-right were 46.38%-60.61%, 24.24%-37.70%, and 95.12%-100%, respectively.

Figure 12.4: Three-Strata GO-CTA Model for Sex Parsing All Cancer Sites



In contrast, when race was used to parse cancer incidence data for all cancer sites combined no statistically reliable model was found, suggesting race is unrelated to cancer incidence. Table 12.8 reveals this result to be a Simpson's Paradox in which the combined data mask actual effects (Chapter 9). Clearly, observations having different types of cancer *shouldn't* be combined in research in which race is used as a variable. If the combining of cancer types is considered (e.g., to increase N), it first must be assessed if paradoxical confounding is induced: if confounding is present then determine if it can be circumvented (Chapter 9). If confounding exists and it can't be circumvented, then cancer types can't be combined and instead must be treated as separate class or attribute categories (Chapters 2 and 3).

Table 12.7 summarizes novometric analysis using sex to parse Oral Cavity and Pharynx cancer incidence. Note that the Oral Cavity and Pharynx, Gum and Other Mouth, and Other Oral Cavity and Pharynx categories combine different types of cancer and may thus be confounded. On the basis of 95% CIs, five models had weak-moderate effects, and four models had moderate effects. For every analysis a single model was identified and all except two models had two strata, offering limited opportunity for granular parsing in applications involving multiple attributes.

Table 12.7: Parsing Cancer Incidence by Sex: Oral Cavity and Pharynx

Cancer Site	Strata	MinD	ESS	Efficiency	D
Oral Cavity and Pharynx	2	120	39.5 33.1-46.0 0-6.58	19.7 16.6-23.0 0-3.29	3.08 4.02-2.35 n/a-28.4
Lip	2	62	19.7 14.5-25.3 0-4.61	9.87 7.27-12.6 0-3.31	8.13 11.8-5.94 n/a-28.2
Tongue	2	120	37.5 30.8-44.1 0-6.58	18.8 15.4-22.1 0-3.29	3.32 4.49-2.52 n/a-28.4
Salivary Gland	3	90	20.1 10.5-29.3 0.33-7.57	6.69 3.51-9.77 0.11-2.52	11.9 25.5-7.24 906-36.7
Floor of Mouth	2	124	33.6 26.6-40.5 0-6.58	16.8 13.3-20.3 0-2.19	3.95 5.52-2.93 n/a-43.7
Gum and Other Mouth	3	103	30.6 21.6-39.3 0.33-8.22	10.2 7.20-13.1 0.11-2.74	6.80 10.9-4.63 906-33.5
Nasopharynx	2	136	30.3 22.7-37.5 0-6.58	15.1 7.55-12.5 0-2.19	4.62 11.2-6.00 n/a-43.7
Tonsil	2	162	37.5 30.0-45.1 0-7.24	18.8 15.0-22.5 0-3.62	3.32 4.67-2.44 n/a-25.6
Oropharynx	2	143	34.5 27.1-41.7 0.33-6.91	17.3 13.6-20.9 0.16-3.46	3.78 5.35-2.78 623-26.9
Hypopharynx	2	136	38.8 31.8-46.0 0-6.58	19.4 15.9-23.0 0-2.19	3.15 4.29-2.35 n/a-2.37
Other Oral Cavity and Pharynx	2	120	25.7 18.7-32.8 0-6.58	12.8 9.34-16.4 0-2.19	5.81 8.71-4.10 n/a-43.7

See Note to Table 12.6. n / a indicates undefined (infinity) due to division by zero. In Figure 12.1 models with ESS = 0 are indicated at the lower left-hand corner (coordinates of 0, 0) of the ESS-by-efficiency space.

Table 12.8 summarizes novometric analysis using race to parse Oral Cavity and Pharynx cancer incidence. The Gum and Other Mouth and the Other Oral Cavity and Pharynx categories combine different types of cancer, and thus may be confounded. For example, analysis for the combined Oral Cavity and Pharynx category found no statistically reliable model suggesting race doesn't predict cancer incidence for this site—clearly an example of paradoxical confounding masking actual effects (Table 12.8).

Table 12.8: Parsing Cancer Incidence by Race: Oral Cavity and Pharynx

Cancer Site	<i>Strata</i>	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
Lip	2	289	50.3	25.2	1.98
			42.3-58.4	21.2-29.2	2.72-1.42
			0.33-8.22	0.16-4.11	623-22.3
Tongue	5	29	35.5	7.11	9.06
			26.8-44.5	5.35-8.91	13.7-6.22
			0-7.89	0-1.58	n/a-58.3
Tongue	4	122	29.3	7.32	9.66
			20.2-38.3	5.05-9.58	15.8-6.44
			0-7.57	0-1.89	n/a-48.9
Tongue	3	135	20.7	6.91	11.5
			12.9-28.5	4.31-9.49	20.2-7.54
			0.33-6.91	0.11-2.30	906-40.5
Salivary Gland	2	93	16.1	8.06	10.4
			9.58-22.8	4.79-11.4	18.9-6.77
			0.33-5.59	0.16-2.80	623-33.7
Floor of Mouth	2	223	11.5	5.76	15.4
			2.28-20.5	1.14-10.2	85.7-7.80
			0.33-7.57	0.16-3.78	623-24.5
Gum and Other Mouth	3	113	21.4	7.13	11.0
			13.8-29.1	4.58-9.70	18.8-7.31
			0.33-6.91	0.11-2.30	906-40.5
Nasopharynx	3	88	29.0	9.65	7.36
			23.0-35.1	7.67-11.7	10.0-5.55
			0-5.26	0-1.75	n/a-54.1
Nasopharynx	2	113	26.3	13.2	5.58
			19.3-33.3	9.64-16.7	8.37-3.99
			0-6.58	0-3.29	n/a-28.4
Tonsil	5	48	27.0	5.39	13.6
			18.8-35.0	3.76-7.00	21.6-9.29
			0-7.24	0-1.45	n/a-64.0
Tonsil	3	106	17.8	5.92	13.9
			10.9-24.9	3.62-8.30	24.6-9.05
			0-5.92	0-1.97	n/a-47.8
Tonsil	2	209	14.8	7.40	11.5
			6.03-23.7	3.02-11.8	31.1-6.47
			0.33-7.57	0.16-3.78	623-24.5
Oropharynx	4	41	32.9	8.22	8.17
			23.9-41.7	5.97-10.4	12.8-5.62
			0-7.89	0-1.97	n/a-46.8
Oropharynx	3	158	26.3	8.77	8.40
			18.3-34.3	6.11-11.4	13.4-5.77
			0-7.24	0-2.41	n/a-38.5
Oropharynx	2	291	21.4	10.7	7.35
			12.4-30.6	6.21-15.3	14.1-4.54
			0.33-8.22	0.16-4.11	623-22.3

Hypopharynx	3	165	19.4 10.9-28.0 0.33-7.57	6.47 3.64-9.32 0.11-2.52	12.5 24.5-7.73 906-36.7
Other Oral Cavity and Pharynx	6	35	29.9 20.9-38.8 0.33-8.22	4.99 3.48-6.46 0.06-1.37	14.0 22.7-9.48 1661-67.0
	4	37	27.0 17.7-36.2 0-7.89	6.74 4.42-9.05 0-1.97	10.8 18.6-7.05 n/a-46.8
	3	118	26.6 18.1-35.4 0-7.57	8.88 6.04-11.8 0-2.52	8.26 13.6-5.47 n/a-36.7
	2	271	17.4 7.99-26.6 0-7.57	8.72 4.00-13.3 0-3.78	9.47 23.0-5.52 n/a-24.5

See Note to Table 12.6. The class variable race was white or African American.
No model emerged for combined Oral Cavity and Pharynx.

In this category Lip cancer produced the weakest effect in sex analyses but the strongest effect in race analyses: the race model is the first in the present example to achieve a relatively strong *ESS* point estimate (greater than overall *ESS* obtained by many published multiattribute models), and a moderate-strong effect as assessed by 95% CI. The 95% CIs for *ESS* and *D* lie outside the 95% CIs for these statistics for all other models in this category. The *D* statistic is less than two, suggesting that race ultimately may emerge as an attribute in an excellent statistical model predicting Lip cancer.

All other models identified in this category had *D* statistics exceeding three, suggesting that using these attributes in multiattribute CTA models would likely yield complex model geometries (Figure 3.5): in most cases analyzed presently (involving models having many endpoints) this ultimately means including the attribute (sex or race) in a complex model would likely yield a greater *D* statistic and worse reproducibility in validity analysis versus eliminating the attribute in a simpler model.

Substantively it may interesting that race yielded a significantly better model (the lower bound of the *ESS* 95% CI for race exceeds the upper bound of the *ESS* 95% CI for sex) for predicting Lip cancer, and that sex produced superior models than race for predicting Tongue, Floor of Mouth, Tonsil, and Hypopharynx cancer.

Finally, note that although exact $p < 0.05$, model 95% CIs for *ESS* and *D* overlap corresponding CIs for chance for the two-strata models for Floor of Mouth cancer, and also for Tonsil cancer.

Considered as a whole, sex and race both generally predict different manifestations of oral and pharynx cancer at a weak-to-moderate level of accuracy and efficiency. MDSA analyses identified many descendant families for race but none for sex. The most powerful model that emerged was for race used to parse Lip cancer incidence.

Table 12.9 summarizes novometric analysis using sex to parse Digestive System cancer incidence. As seen, no statistically reliable model emerged for 11 cancer categories, and the eight models identified were all binary (two-strata) parses that returned relatively weak effects for all models except for a weak-moderate finding for Esophagus cancer. For Anus, Anal Canal and Anorectum cancer the lower bound of the 95% CI for model *ESS*, *efficiency*, and *D* overlapped corresponding 95% CIs for chance, indicating sex is not predictive of this category of cancer.

Table 12.10 gives novometric results using race to parse digestive system cancer incidence. As seen, no statistically reliable model was identified for 13 cancer categories, and the six models that were identified had descendant families having two or three members. Note that although exact $p < 0.05$, the model 95% CIs for *ESS* and *D* overlap corresponding CIs for chance for the two-strata models for Stomach cancer, Hepatic Fracture cancer, and Splenic Flexure cancer.

Table 12.9: Parsing Cancer Incidence by Sex: Digestive System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Digestive System	2	48	13.2 8.34-18.1 0-3.95	6.58 4.17-9.06 0-1.98	13.2 22.0-9.04 n/a-48.5
Esophagus	2	113	32.6 26.1-39.2 0.33-6.25	16.3 13.0-19.6 0.16-3.12	4.13 5.69-3.10 623-30.1
Stomach	2	100	15.8 9.16-22.6 0-5.92	7.90 4.58-11.3 0-2.96	10.7 19.8-6.85 n/a-31.8
Splenic Flexure	2	72	12.5 6.58-18.6 0-5.26	6.25 3.29-9.28 0-2.63	14.0 28.4-8.78 n/a-36.0
Sigmoid Colon	2	76	11.8 5.80-17.9 0-5.26	5.90 2.90-8.97 0-2.63	14.9 32.5-9.15 n/a-36.0
Rectum and Recto-sigmoid Junction	2	73	14.8 8.93-20.9 0.33-4.93	7.40 4.46-10.4 0.16-2.46	11.5 20.4-7.62 623-38.7
Rectum	2	122	15.1 7.71-22.5 0-6.58	7.56 3.86-11.3 0-3.29	11.2 23.9-6.85 n/a-28.4
Anus, Anal Canal and Anorectum	2	83	11.5 4.99-17.9 0.33-5.59	5.76 2.50-8.96 0.16-2.88	15.4 38.0-9.16 623-32.7

See Note to Table 12.6. No model emerged for Small Intestine; Colon and Rectum; Colon excluding Rectum; Cecum; Appendix; Ascending Colon; Hepatic Fracture; Transverse Colon; Descending Colon; Large Intestine NOS; or Rectosigmoid Junction.

Table 12.10: Parsing Cancer Incidence by Race: Digestive System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Esophagus	5	39	29.6 20.6-38.4 0-7.89	5.92 4.12-7.69 0-1.58	11.9 19.3-8.00 n/a-58.3
	3	184	24.0 15.5-32.7 0.33-7.57	8.00 5.15-10.9 0.11-2.52	9.50 16.4-6.17 906-36.7
	2	221	16.8 7.93-25.8 0.33-7.57	8.39 3.96-12.9 0.16-3.78	9.92 23.3-5.75 623-24.5
Stomach	3	110	19.1 10.2-28.1 0-7.89	6.36 3.40-9.36 0-2.63	12.7 26.4-7.68 n/a-35.0

	2	274	11.2 1.75-20.7 0-7.89	5.59 0.88-10.3 0-3.94	15.9 112-7.71 n/a-23.4
Small Intestine	4	59	27.6 19.1-36.2 0-7.24	6.91 4.78-9.06 0-1.81	10.5 16.9-7.04 n/a-51.2
	3	155	22.7 14.4-31.0 0.33-6.91	7.57 4.78-10.3 0.11-2.30	10.2 17.9-6.71 905-40.5
	6	25	42.1 33.5-50.5 0-7.89	7.02 5.59-8.42 0-1.32	8.25 11.9-5.88 n/a-69.8
Appendix	4	42	39.8 31.4-48.2 0.33-7.57	9.95 7.85-12.0 0.08-1.89	6.05 8.74-2.33 1246-48.9
	2	191	33.9 25.7-42.0 0.33-7.57	16.9 12.8-21.0 0.16-3.78	3.92 5.81-2.76 623-24.5
	3	78	15.1 8.93-21.2 0-5.26	5.04 2.98-7.06 0-1.75	16.8 30.6-11.2 n/a-54.1
Hepatic Fracture	2	194	13.2 4.43-21.9 0-7.24	6.58 2.22-10.9 0-3.62	13.2 43.0-7.17 n/a-25.6
	3	193	22.0 13.7-30.5 0.33-7.57	7.35 4.57-10.2 0.11-2.52	10.6 18.9-6.80 905-36.7
	2	195	11.5 2.50-20.3 0.33-7.57	5.76 1.25-10.1 0.16-3.78	15.4 78.0-7.90 623-24.5

See Note to Table 12.8. No model emerged for combined Digestive System; Colon and Rectum; Colon excluding Rectum; Cecum; the Ascending, Transverse, Descending, or Sigmoid Colon; Large Intestine, NOS; Rectum and Rectosigmoid Junction; Rectosigmoid Junction; Rectum; or Anus, Anal Canal and Anorectum.

Table 12.11 summarizes novometric analysis using sex to parse Liver and Intrahepatic Bile Duct cancer incidence. As seen, no statistically reliable model was identified for four of the cancer categories, all five models that were identified were binary parses, and accuracy point estimates were weak except for the weak-to-moderate finding for Liver cancer. For Intrahepatic Bile Duct cancer, and for Gallbladder cancer, the lower bounds of the 95% CIs for model *ESS* overlapped the corresponding 95% CIs for chance, indicating that sex is not reliably predictive of these cancers.

Table 12.11: Parsing Cancer Incidence by Sex: Liver and Intrahepatic Bile Duct

Cancer Site	Strata	MinD	ESS	Efficiency	D
Liver and Intra-Hepatic Bile Duct	2	102	23.0 16.3-29.9 0-5.92	11.5 8.14-14.9 0-2.96	6.70 10.3-4.71 n/a-31.8

Liver	2	119	25.3 18.2-32.4 0.33-6.25	12.7 9.11-16.2 0.16-3.12	5.87 8.98-4.17 623-30.1
Intrahepatic Bile Duct	2	185	11.5 3.05-20.3 0.33-6.91	5.76 1.52-10.2 0.16-3.46	15.4 63.8-7.80 623-26.9
Gallblader	2	241	14.8 5.67-24.0 0.33-7.57	7.40 2.84-12.0 0.16-3.78	11.5 33.2-6.33 623-24.5
Peritoneum, Omentum and Mesentery	2	136	21.1 13.5-28.9 0-6.58	10.5 6.74-14.4 0-3.29	7.52 12.8-4.94 n/a-28.4

See Note to Table 12.6. No model emerged for Other Biliary; Pancreas; Retroperitoneum; or Other Digestive Organs.

Table 12.12 summarizes novometric analysis using race to parse Liver and Intrahepatic Bile Duct cancer incidence. As seen, no statistically reliable model was identified for two of the cancer categories, and three of the seven identified models were binary parses. For Other Biliary cancer the lower bounds of the 95% CIs for model ESS overlapped the corresponding 95% CIs for chance, indicating race is not reliably predictive of this cancer.

Table 12.12: Parsing Cancer Incidence by Race: Liver and Intrahepatic Bile Duct

Cancer Site	Strata	MinD	ESS	Efficiency	D
Liver and Intrahepatic Bile Duct	4	2	19.4 12.3-26.6 0.33-6.25	4.85 3.07-6.64 0.08-1.56	16.6 28.6-11.1 1246-60.1
	3	54	18.8 11.8-25.8 0.33-6.25	6.25 3.93-8.59 0.11-2.08	13.0 22.4-8.64 906-45.1
	7	2	28.0 21.0-34.9 0.33-6.25	3.99 3.00-4.99 0.05-0.89	18.1 26.3-13.0 1993-105
	4	55	25.3 17.3-33.3 0.33-6.91	6.33 4.32-8.32 0.08-1.73	11.8 19.1-8.02 1246-53.8
Intrahepatic Bile Duct	2	283	25.3 16.2-34.4 0.33-7.57	12.7 8.08-17.2 0.16-3.78	5.87 10.4-3.81 623-24.5
Other Biliary	3	3	17.4 8.54-26.2 0.33-7.57	5.81 2.85-8.72 0.11-2.52	14.2 32.1-8.47 906-36.7
	2	216	16.4 7.56-25.4 0-7.89	8.22 3.78-12.7 0-3.94	10.2 24.5-5.87 n/a-23.4

Retroperitoneum	2	124	28.3 21.2-35.4 0-6.58	14.1 10.6-17.7 0-3.29	5.09 7.43-3.65 n/a-28.4
Peritoneum, Omentum and Mesentery	2	304	37.5 28.8-46.3 0-7.89	18.8 14.4-23.2 0-3.94	3.32 4.94-2.31 n/a-23.4
Other Digestive Organs	3	119	24.0 16.9-31.1 0.33-6.25	8.00 5.63-10.4 0.11-2.08	9.50 14.8-6.62 906-45.1
	2	264	22.4 13.3-31.3 0-7.89	11.2 6.64-15.7 0-3.94	6.93 13.1-4.37 n/a-23.4

See Note to Table 12.8. No model emerged for Gallbladder or Pancreas.

Table 12.13 summarizes novometric analysis using sex to parse Respiratory System cancer incidence. As seen, all of the identified models were binary (two-strata) parses. The 95% CIs for *ESS* show that sex produced a weak effect for Nose, Nasal Cavity and Middle Ear cancer; a moderate effect for Combined Respiratory System, Lung and Bronchus, and Trachea, Mediastinum, and Other Respiratory Organ cancers; and a moderate-strong effect for Larynx cancer. For Pleura cancer, however, the lower bounds of the 95% CIs for model *ESS* and efficiency overlapped corresponding 95% CIs for chance, indicating sex is not predictive of this cancer. All models identified in this category except for Larynx cancer had *D* statistics exceeding three, suggesting their use in multiattribute CTA models would likely yield untenably complex model geometries (Figure 3.5)

Table 12.13: Parsing Cancer Incidence by Sex: Respiratory System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Respiratory System	2	99	29.3 22.9-35.8 0.33-6.25	14.6 11.4-17.9 0.16-3.12	4.85 6.77-3.59 623-30.1
Nose, Nasal Cavity and Middle Ear	2	138	17.1 9.30-25.0 0-6.58	8.56 4.65-12.5 0-3.29	9.68 19.5-6.00 n/a-28.4
Larynx	2	163	43.8 36.4-51.0 0.33-6.91	21.9 18.2-25.5 0.16-3.46	2.57 3.49-1.92 623-26.9
Lung and Bronchus	2	111	27.3 20.5-34.0 0.33-6.25	13.6 10.2-17.0 0.16-3.12	5.35 7.81-3.88 623-30.1
Pleura	2	137	13.5 5.78-21.2 0.33-6.91	6.70 2.89-10.6 0.16-3.46	12.9 32.6-7.43 623-26.9
Trachea, Media- stium, Other Respiratory Organs	2	296	29.0 19.7-38.0 0-7.89	14.5 9.86-19.0 0-3.94	4.90 8.14-3.26 n/a-23.4

See Note to Table 12.6.

Table 12.14 gives novometric results for race parsing Respiratory System cancer incidence. As seen, no statistically reliable model emerged for the Combined Cancer category, suggesting that race isn't predictive of respiratory system cancers—an example of paradoxical confounding masking effects. All models identified in this category except the GO-CTA (two-parse) and the three-parse models for Trachea, Mediastinum, and Other cancer had D statistics exceeding three.

Table 12.14: Parsing Cancer Incidence by Race: Respiratory System

Cancer Site	Strata	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
Nose, Nasal Cavity and Middle Ear	2	115	24.0	12.0	6.33
			17.0-31.0	8.51-15.5	9.75-4.45
			0.33-6.25	0.16-3.12	623-30.1
Larynx	5	29	31.6	6.32	10.8
			22.5-40.6	4.51-8.13	17.2-7.30
			0-7.89	0-1.58	n/a-58.3
Lung and Bronchus	3	170	24.0	8.00	9.50
			15.1-32.9	5.02-11.0	16.9-6.09
			0.33-7.57	0.11-2.52	906-36.7
Pleura	2	91	15.5	5.15	16.4
			8.81-22.2	2.94-7.38	31.0-10.6
			0.33-5.59	0.11-1.86	906-50.8
Trachea, Medi- stium, Other	6	48	30.3	15.1	4.62
			22.2-38.3	11.1-19.2	7.01-3.21
			0-7.24	0-3.62	n/a-25.6
Trachea, Medi- stium, Other	5	72	56.2	9.38	4.67
			48.2-64.0	8.02-10.7	6.47-3.35
			0.33-8.22	0.06-1.37	1661-67.0
Trachea, Medi- stium, Other	3	110	53.0	10.6	4.43
			44.9-60.8	8.98-12.2	6.14-3.20
			0.33-7.57	0.07-1.51	1424-61.2
Trachea, Medi- stium, Other	2	180	50.0	16.7	2.99
			41.7-58.2	13.9-19.4	4.19-2.15
			0-7.89	0-2.63	n/a-35.0

See Note to Table 12.8. No model emerged for Combined Respiratory System.

Table 12.15 gives novometric analysis using sex to parse Bones and Joints cancer incidence. For the two-strata model the lower bounds of 95% CIs for model accuracy and efficiency overlapped corresponding 95% CIs for chance. The four-strata race model had a weak-moderate effect.

Table 12.16 gives novometric results using race to parse Bones and Joints cancer incidence. The two-strata *race* model for Bones and Joints cancer is more significantly 95% CI non-overlap) more accurate (greater *ESS*), parsimonious (greater *efficiency*), and closer to being an ideal classification model (smaller *D*) than the four-strata *sex* model.

Table 12.15: Parsing Cancer Rate by Sex: Bones and Joints

Cancer Site	<i>Strata</i>	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
Bones and Joints	4	58	20.4	5.10	15.6
			12.5-28.2	3.13-7.04	27.9-10.2
	2	251	0-6.58	0-1.64	n/a-57.0
			14.8	7.40	11.5
			5.76-24.2	2.88-12.1	32.7-6.26
			0.33-7.57	0.16-3.78	623-24.5

See Note to Table 12.6.

Table 12.16: Parsing Cancer Rate by Race: Bones and Joints

Cancer Site	<i>Strata</i>	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
Bones and Joints	2	198	29.6	14.8	4.76
			21.3-38.2	10.6-19.1	7.43-3.24
			0-7.24	0-3.62	n/a-25.6

See Note to Table 12.8.

No models emerged parsing any of the Skin excluding Basal and Squamous cancer incidence categories by sex. Table 12.17 gives novometric analysis using race to parse this category: strong (*ESS* > 50) models having *D* < 2 (by 95% CI) were obtained for Skin excluding Basal and Squamous cancer, and for Melanoma of the Skin. The two-strata model for Skin excluding Basal and Squamous cancer correctly classified 93.1% of African Americans and 68.4% of whites; for Melanoma of the Skin correctly classified 95.1% of African Americans and 72.0% of whites; and for Other Non-Epithelial Skin cancer correctly classified 65.5% of African Americans and 67.4% of whites.

Table 12.17: Parsing Cancer Incidence by Race: Skin excluding Basal and Squamous

Cancer Site	<i>Strata</i>	<i>MinD</i>	<i>ESS</i>	<i>Efficiency</i>	<i>D</i>
Skin excluding Basal and Squamous	5	14	66.1	13.2	2.58
			59.2-72.9	11.8-14.6	3.48-1.85
	4	51	0.33-7.57	0.07-1.51	1424-61.2
			63.8	16.0	2.25
			56.6-71.0	14.1-17.8	3.09-1.62
Melanoma of the Skin	2	229	0-7.89	0-1.97	n/a-46.8
			61.5	30.8	1.25
			54.4-68.6	27.2-34.3	1.68-0.92
	5	25	0.33-7.57	0.16-3.79	623-24.4
			75.3	15.1	1.62
Other Non-Epithelial Skin	4	63	69.2-81.2	13.8-16.2	2.25-1.17
			0.33-8.22	0.07-1.64	1424-56.0
	2	234	71.4	17.8	1.62
			64.4-77.8	16.1-19.5	2.10-1.13
Cervix Uteri	2	112	0.33-8.22	0.08-2.06	1246-44.5
			67.1	33.6	0.98
			60.4-73.7	30.2-36.8	1.31-0.72
Prostate	2	100	0-7.89	0-3.94	n/a-23.4

Other Non-Epithelial Skin	4	83	32.9 24.0-41.9 0-7.89	8.22 6.00-10.5 0-1.97	8.17 12.7-5.52 n/a-46.8
---------------------------	---	----	-----------------------------	-----------------------------	-------------------------------

See Note to Table 12.8.

Table 12.18 summarizes novometric analysis using race to parse Female Genital System cancer incidence. No statistically reliable model was identified for the combined Female Genital System cancer category, suggesting that race is not predictive of Female Genital System cancers—yet another example of paradoxical confounding. For the two-strata Ovary cancer model, the 95% CI for model *ESS* overlapped the 95% CI for chance.

Table 12.18: Parsing Cancer Incidence by Race: Female Genital System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Cervix Uteri	3	84	40.8 28.5-52.5 0-10.5	13.6 9.50-17.5 0-3.50	4.35 7.53-2.71 n/a-25.6
			39.5 28.5-50.6 0-10.5	19.7 14.2-25.3 0-5.25	3.08 5.04-1.95 n/a-17.0
			18.4 8.92-27.7 0-7.89	9.21 4.46-13.9 0-3.94	8.86 20.4-5.19 n/a-23.4
	2	90	20.4 10.6-30.3 0.66-8.55	10.2 5.28-15.1 0.33-4.28	7.80 16.9-4.62 301-21.4
Corpus and Uterus, NOS	2	80	35.5 22.6-48.0 0-10.5	11.8 7.55-16.0 0-3.50	5.47 10.2-3.25 n/a-25.6
			17.1 7.68-26.8 0-10.5	8.56 3.84-13.4 0-5.25	9.68 24.0-5.46 n/a-17.0
			29.0 18.0-40.1 0-9.21	9.65 6.00-13.4 0-3.07	7.36 13.7-4.46 n/a-29.6

See Note to Table 12.8. No model emerged for Combined Female Genital System; Vagina; Vulva; or Other Female Genital Organs. Here, *N* = 304.
NOS = Not Otherwise Specified.

Table 12.19 gives novometric analysis using race to parse Male Genital System cancer incidence. For the two-strata Prostate cancer model, the 95% CI for model *ESS* overlapped the 95% CI for chance.

Table 12.19: Parsing Cancer Incidence by Race: Male Genital System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Male Genital System	4	29	36.8 24.7-48.8 0-10.5	9.21 6.19-12.2 0-2.62	6.86 12.2-4.20 n/a-34.2
			29.0 18.0-40.1 0-9.21	9.65 6.00-13.4 0-3.07	7.36 13.7-4.46 n/a-29.6
			29.0 18.0-40.1 0-9.21	9.65 6.00-13.4 0-3.07	7.36 13.7-4.46 n/a-29.6

	2	99	23.0 10.8-35.1 0.66-9.87	11.5 5.42-17.5 0.33-4.94	6.70 16.5-3.71 301-18.2
Prostate	4	35	32.9 20.9-44.7 0-10.5	8.22 5.24-11.2 0-2.62	8.17 15.1-4.93 n/a-34.2
	3	36	23.7 15.8-31.8 0-6.58	7.89 5.26-10.6 0-2.19	9.67 16.0-6.43 n/a-42.7
	2	71	19.1 8.01-30.1 0.66-9.87	9.54 4.01-15.1 0.33-4.94	8.48 22.9-4.62 301-18.2
Testes	2	118	46.1 34.2-57.2 0-10.5	23.0 17.1-28.6 0-5.25	2.35 3.85-1.50 n/a-17.0
Other Male Genital Organs	2	104	35.5 23.8-47.2 0-10.5	17.8 11.9-23.6 0-5.25	3.62 6.40-2.24 n/a-17.0

See Note to Table 12.8. No model emerged for Penis cancer. $N = 304$.

Table 12.20 gives novometric results using sex to parse Urinary System cancer incidence.

Table 12.20: Parsing Cancer Incidence by Sex: Urinary System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Urinary System	2	87	26.0 19.9-32.2 0.33-5.59	13.0 9.93-16.1 0.16-2.80	5.69 8.07-4.21 623-33.7
Urinary Bladder	2	88	24.3 18.2-30.7 0-5.26	12.2 9.12-15.4 0-2.63	6.20 8.96-4.49 n/a-36.0
Kidney and Renal Pelvis	2	106	21.7 14.8-28.6 0-5.92	10.9 7.41-14.3 0-2.96	7.17 11.5-4.99 n/a-31.8
Ureter	2	56	13.8 8.58-19.2 0-4.61	6.91 4.29-9.61 0-2.31	12.5 21.3-8.41 n/a-41.3
Other Urinary Organs	2	78	16.4 10.5-22.8 0-5.26	8.22 5.26-11.4 0-2.63	10.2 17.0-6.77 n/a-36.0

See Note to Table 12.6.

Table 12.21 gives novometric results for race parsing urinary system cancer incidence. For the two-strata Other Urinary Organs cancer model, the 95% CIs for model and chance ESS overlapped.

Table 12.21: Parsing Cancer Incidence by Race: Urinary System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Urinary Bladder	2	82	15.8 9.54-22.1 0-5.26	7.90 4.77-11.1 0-2.63	10.7 19.0-7.01 n/a-36.0
Ureter	4	80	32.2 24.3-40.1 0-7.24	8.06 6.07-10.0 0-1.81	8.41 12.5-6.00 n/a-51.2
	2	288	28.3 19.1-37.3 0-7.89	14.1 9.53-18.7 0-3.94	5.09 8.49-3.35 n/a-23.4
Other Urinary Organs	3	151	24.0 16.1-31.2 0.33-6.91	8.00 5.37-10.6 0.11-2.30	9.50 15.6-6.43 906-40.5
	2	250	14.5 5.09-23.6 0-7.89	7.24 2.54-11.8 0-3.94	11.8 37.4-6.47 n/a-23.4

See Note to Table 12.8. No model emerged for Combined Urinary System or Kidney and Renal Pelvis.

Table 12.22 summarizes novometric analysis using sex to parse Eye and Orbit cancer incidence.

Table 12.22: Parsing Cancer Incidence by Sex: Eye and Orbit

Cancer Site	Strata	MinD	ESS	Efficiency	D
Eye and Orbit	2	38	11.2 6.85-15.7 0-3.95	5.60 3.42-7.86 0-1.98	15.9 27.2-10.7 n/a-48.5

See Note to Table 12.6.

Table 12.23 summarizes novometric analysis using race to parse Eye and Orbit cancer incidence: a strong ($ESS > 50$) two-parse GO-CTA model having $D < 2$ (by 95% CI) was obtained.

Table 12.23: Parsing Cancer Incidence by Race: Eye and Orbit

Cancer Site	Strata	MinD	ESS	Efficiency	D
Eye and Orbit	7	27	71.1 64.3-77.6 0-7.89	10.2 9.19-11.1 0-1.13	2.80 3.89-2.01 n/a-81.5
	6	47	68.1 61.3-74.6 0.33-7.57	11.3 10.2-12.4 0.06-1.26	2.85 3.80-2.06 1661-73.4
	5	51	62.8 55.6-70.1 0.33-7.57	12.6 11.1-14.0 0.07-1.51	2.94 4.01-2.14 1424-61.2

2	202	62.5 55.8-69.0 0-7.24	31.2 27.9-34.5 0-3.62	1.21 1.58-0.90 n/a-25.6
---	-----	-----------------------------	-----------------------------	-------------------------------

See Note to Table 12.8.

Table 12.24 gives novometric analysis using sex to parse Brain and Other Nervous system cancer incidence: both models had overlapping 95% CI bounds for model and chance classification performance.

Table 12.24: Parsing Cancer Incidence by Sex: Brain and Other Nervous System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Brain and Other Nervous System	2	275	14.8 5.39-24.0 0.33-7.57	7.40 2.70-12.0 0.16-3.78	11.5 35.0-6.33 623-24.5
Brain	2	282	15.1 5.83-24.4 0-7.89	7.56 2.92-12.2 0-3.94	11.2 32.3-6.20 n/a-23.4

See Note to Table 12.6. No model emerged for Cranial Nerves, Other Nervous System.

Table 12.25 summarizes novometric analysis using race to parse Brain and Other Nervous System cancer incidence.

Table 12.25: Parsing Cancer Incidence by Race: Brain and Other Nervous System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Brain and Other Nervous System	2	163	26.6 18.6-34.6 0.33-6.91	13.3 9.32-17.3 0.16-3.46	5.52 8.73-3.78 623-26.9
Brain	2	171	27.3 19.2-35.5 0.33-6.91	13.6 9.62-17.6 0.16-3.46	5.35 8.40-3.68 623-26.9
Cranial Nerves, Other Nervous System	3	48	43.1 35.5-50.4 0.33-6.91	14.4 11.8-16.8 0.11-2.30	3.94 5.47-2.95 906-40.5
	2	119	34.5 27.8-41.3 0.33-6.25	17.3 13.9-20.7 0.15-3.12	3.78 5.19-2.83 665-30.1

See Note to Table 12.8.

Table 12.26 gives novometric results using sex to parse Endocrine System cancer incidence, and Table 12.27 gives novometric results using race to parse Endocrine System cancer incidence.

Table 12.26: Parsing Cancer Incidence by Sex: Endocrine System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Endocrine System	2	224	38.8 30.2-47.1 0-7.89	19.4 15.1-23.5 0-3.94	3.15 4.62-2.26 n/a-23.4
Thyroid	2	210	43.4 35.4-51.3 0-7.24	21.7 17.7-25.7 0-3.62	2.61 3.65-1.89 n/a-25.6
Other Endocrine including Thymus	2	146	15.1 7.36-23.2 0-6.58	7.56 3.68-11.6 0-3.29	11.2 25.2-6.62 n/a-28.4

See Note to Table 12.6.

Table 12.27 gives novometric results using race to parse endocrine system cancer incidence.

Table 12.27: Parsing Cancer Incidence by Race: Endocrine System

Cancer Site	Strata	MinD	ESS	Efficiency	D
Endocrine System	2	265	22.7 13.5-31.8 0.33-8.22	11.4 6.76-15.9 0.16-4.11	6.77 12.8-4.29 623-22.3
Thyroid	2	225	22.0 13.2-30.9 0.33-7.57	11.0 6.59-15.5 0.16-3.78	7.09 13.2-4.45 623-24.5
Other Endocrine including Thymus	4	31	33.6 26.9-40.3 0-5.92	8.39 6.73-10.1 0-1.48	7.92 10.9-5.90 n/a-63.6
	3	79	32.6 24.4-40.7 0.33-6.91	10.9 8.14-13.6 0.11-2.30	6.17 9.29-4.35 906-40.5

See Note to Table 12.8.

Table 12.28 gives novometric results using sex to parse Lymphoma incidence. Two-parse models for Lymphoma, Non-Hodgkin Lymphoma, Non-Hodgkin Lymphoma-Nodal, and Non-Hodgkin Lymphoma-Extranodal had overlapping model and chance ESS 95% CIs.

Table 12.28: Parsing Cancer Incidence by Sex: Lymphoma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Lymphoma	2	284	15.8 6.54-25.1 0-7.89	7.90 3.27-12.6 0-3.94	10.7 28.6-5.94 n/a-23.4
Hodgkin Lymphoma	2	289	36.5 27.8-45.4 0.33-7.57	18.3 13.9-22.7 0.16-3.78	3.46 5.19-2.41 623-24.5

Hodgkin-Nodal	2	292	36.8 28.2-45.4 0-7.89	18.4 14.1-22.7 0-3.94	3.43 5.09-2.41 n/a-23.4
Non-Hodgkin Lymphoma	4	73	19.1 10.7-27.4 0-7.24	4.77 2.67-6.84 0-1.81	17.0 33.5-10.6 n/a-51.2
	2	299	12.8 3.41-22.3 0.33-8.22	6.42 1.71-11.2 0.16-4.11	13.6 56.5-6.93 623-22.3
Non-Hodgkin Lymphoma- Nodal	2	107	12.2 5.01-19.2 0.33-5.59	6.08 2.51-9.62 0.15-2.80	14.4 37.8-8.40 665-33.7
Non-Hodgkin Lymphoma- Extranodal	2	240	13.8 4.59-22.8 0-7.89	6.91 2.30-11.4 0-3.94	12.5 41.5-6.77 n/a-23.4

See Note to Table 12.6. No model emerged for Hodgkin-Extranodal.

Table 12.29 summarizes novometric analysis using race to parse Lymphoma incidence.

Table 12.29: Parsing Cancer Incidence by Race: Lymphoma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Lymphoma	2	94	16.4 9.89-23.1 0-5.92	8.22 4.94-11.6 0-2.96	10.2 18.2-6.62 n/a-31.8
Hodgkin Lymphoma	4	30	31.6 22.6-40.4 0-7.89	7.90 5.64-10.1 0-1.97	8.66 13.7-5.90 n/a-46.8
	2	280	26.3 17.2-35.3 0-7.89	13.2 8.60-17.7 0-3.94	5.58 9.63-3.65 n/a-23.4
Hodgkin-Nodal	4	33	30.3 21.5-39.0 0-7.89	7.56 5.38-9.74 0-1.97	9.23 14.6-6.27 n/a-46.8
	2	261	25.3 16.0-34.2 0.33-7.57	12.7 8.02-17.1 0.16-3.78	5.87 10.5-3.85 623-24.5
Hodgkin- Extranodal	2	251	47.7 39.4-55.8 0.33-7.57	23.8 19.7-27.9 0.16-3.78	2.20 3.08-1.58 623-24.5
Non-Hodgkin Lymphoma	2	109	14.8 7.65-21.8 0.33-6.25	7.40 3.82-10.9 0.16-3.12	11.5 24.2-7.17 623-30.1
Non-Hodgkin Lymphoma- Nodal	2	87	16.8 10.2-23.3 0.33-5.59	8.39 5.13-11.6 0.16-2.80	9.92 17.5-6.62 623-33.7

Non-Hodgkin	4	53	28.3 19.5-37.1 0-7.89	7.07 4.87-9.27 0-1.97	10.1 16.5-6.79 n/a-46.8
-------------	---	----	-----------------------------	-----------------------------	-------------------------------

See Note to Table 12.8.

Table 12.30 presents novometric analysis using sex to parse Myeloma incidence: the model and chance 95% CIs for model performance overlapped.

Table 12.30: Parsing Cancer Incidence by Sex: Myeloma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Myeloma	2	111	12.8 5.63-20.0 0.33-6.25	6.42 2.82-10.0 0.16-3.12	13.6 33.5-8.00 623-30.1

See Note to Table 12.6.

Table 12.31 summarizes novometric analysis using race to parse myeloma incidence. Note that identical error effects emerged for sex and race analyses.

Table 12.31: Parsing Cancer Incidence by Race: Myeloma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Myeloma	2	76	18.4 12.4-24.6 0.33-6.25	9.21 6.21-12.3 0.16-3.12	8.86 14.1-6.13 623-30.1

See Note to Table 12.8.

Table 12.32 summarizes novometric analysis using sex to parse Leukemia incidence. Models for Acute Lymphocytic Leukemia and Chronic Myeloid Leukemia had overlapping model and chance 95% CIs.

Table 12.32: Parsing Cancer Incidence by Sex: Leukemia

Cancer Site	Strata	MinD	ESS	Efficiency	D
Leukemia	4	41	24.3 15.4-33.1 0-7.89	6.08 3.86-8.26 0-1.97	12.4 21.9-8.11 n/a-46.8
	2	84	15.1 8.67-21.4 0-5.26	7.56 4.34-10.7 0-2.63	11.2 21.0-7.35 n/a-36.0
Lymphocytic Leukemia	2	50	15.8 11.0-20.9 0-3.95	7.90 5.49-10.4 0-1.98	10.7 16.2-7.62 n/a-48.5
Acute Lymphocytic Leukemia	2	275	16.1 6.64-25.6 0.33-8.22	8.06 3.32-12.8 0.16-4.11	10.4 28.1-5.81 623-22.3

Chronic Lymphocytic Leukemia	2	72	15.1 9.21-21.1 0-5.26	7.56 4.61-10.5 0-2.63	11.2 19.7-7.52 n/a-36.0
Other Lymphocytic Leukemia	2	189	20.7 11.9-29.1 0.33-7.57	10.4 5.97-14.5 0.16-3.78	7.62 14.8-4.90 623-24.5
Myeloid and Monocytic Leukemia	2	64	12.5 6.77-18.2 0-4.61	6.25 3.38-9.08 0-2.31	14.0 27.6-9.01 n/a-41.3
Acute Myeloid Leukemia	2	49	12.2 7.20-17.3 0.33-4.28	6.08 3.60-8.66 0.16-2.14	14.4 25.8-9.55 623-44.7
Chronic Myeloid Leukemia	2	239	15.5 6.27-24.5 0.33-8.22	7.73 3.14-12.3 0.16-4.11	10.9 29.8-6.13 623-22.3

See Note to Table 12.6. No model emerged for Acute Monocytic Leukemia; Other Myeloid/Monocytic Leukemia; Other Leukemia; Other Acute Leukemia; or Aleukemic, Subleukemic and NOS = not otherwise specified.

Table 12.33 summarizes novometric analysis using race to parse leukemia incidence. Models for Leukemia and Lymphocytic Leukemia had overlapping model and chance *ESS* 95% CIs. A strong (*ESS* > 50) two-parse model with *D* < 2 (by 95% CI) was obtained for Acute Monocytic Leukemia.

Table 12.33: Parsing Cancer Incidence by Race: Leukemia

Cancer Site	Strata	MinD	ESS	Efficiency	D
Leukemia	2	119	11.5 4.00-18.8 0.33-6.25	5.76 2.00-9.41 0.16-3.12	15.4 48.0-8.63 623-30.1
Lymphocytic Leukemia	2	109	12.2 4.99-19.6 0.33-6.25	6.08 2.50-9.79 0.16-3.12	14.4 38.0-8.21 623-30.1
Acute Lymphocytic Leukemia	2	138	31.6 24.3-38.9 0-6.58	15.8 12.2-19.5 0-3.29	4.33 6.20-3.13 n/a-28.4
Other Lymphocytic Leukemia	2	228	38.8 30.2-47.2 0-7.89	19.4 15.1-23.6 0-3.94	3.15 4.62-2.24 n/a-23.4
Acute Monocytic Leukemia	4	69	60.9 53.4-68.3 0.33-8.22	15.2 13.3-17.1 0.08-2.06	2.58 3.52-1.85 1246-44.5
	2	218	57.2 50.0-64.4 0-7.89	28.6 25.0-32.2 0-3.94	1.50 2.00-1.11 n/a-23.4
Chronic Myeloid Leukemia	5	31	17.4 11.8-23.3 0.33-4.93	3.49 2.36-4.65 0.07-0.99	23.7 37.4-16.5 1424-96.0

	3	52	15.1 8.42-22.0 0-5.92	5.04 2.81-7.34 0-1.97	16.8 32.6-10.6 n/a-47.8
Other Myeloid/ Monocytic Leukemia	2	224	46.7 38.5-54.6 0-7.89	23.4 19.2-27.3 0-3.94	2.27 3.21-1.66 n/a-23.4
Other Leukemia	2	91	26.0 19.7-32.2 0.33-5.59	13.0 9.86-16.1 0.16-2.80	5.69 5.08-4.21 623-33.7
Other Acute Leukemia	3	2	43.8 36.6-50.8 0.33-6.91	14.6 12.2-16.9 0.11-2.30	3.85 5.20-2.92 906-40.5
	2	159	43.1 35.7-50.3 0.33-6.91	21.5 17.9-25.2 0.16-3.46	2.65 3.59-1.97 623-26.9
Aleukemic, Leukemic and NOS	3	102	33.6 27.1-40.0 0-5.92	11.2 9.05-13.3 0-1.97	5.93 8.05-4.52 n/a-47.8
	2	185	30.6 22.3-38.7 0.33-7.57	15.3 11.2-19.4 0.16-3.78	4.54 6.93-3.15 623-24.7

See Note to Table 12.8. No model emerged for Chronic Lymphocytic Leukemia; Myeloid and Monocytic Leukemia; or Acute Myeloid Leukemia.

Finally, novometric analysis was conducted using sex (Table 12.34) or race (Table 12.34) to parse Mesothelioma and Kaposi Sarcoma incidence.

Table 12.34: Parsing Cancer Incidence by Sex: Mesothelioma and Kaposi Sarcoma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Mesothelioma	2	158	23.7 15.8-31.7 0-7.24	11.8 7.88-15.8 0-3.62	6.47 10.7-4.33 n/a-25.6
Kaposi Sarcoma	2	192	47.4 39.8-55.0 0-7.24	23.7 19.9-27.5 0-3.62	2.22 3.03-1.64 n/a-25.6

See Note to Table 12.6.

Table 12.35: Parsing Cancer Incidence by Race: Mesothelioma and Kaposi Sarcoma

Cancer Site	Strata	MinD	ESS	Efficiency	D
Mesothelioma	2	235	27.3 18.2-36.1 0.33-7.57	13.6 9.08-18.0 0.16-3.78	5.35 9.01-3.56 623-24.5

Kaposi Sarcoma	2	253	19.4 10.1-28.5 0.33-7.57	9.71 5.06-14.2 0.16-3.78	8.30 17.8-5.04 623-24.5
----------------	---	-----	--------------------------------	--------------------------------	-------------------------------

See Note to Table 12.8.

Research involving human subjects routinely includes “subject variables” such as gender and race as potential predictive attributes: it is *intrinsically assumed* that the different categories of a variable are *homogeneous within-group*, and *explicitly hypothesized* that the categories are *heterogeneous between-group*. Obtained using novometric analysis, parsed models that identify different manifestations (i.e., sub-groups) of a specific class category (e.g., females) indicate that the intrinsic assumption of within-group homogeneity must be explicitly evaluated: different types of males (e.g., males having a high or having a low associated cancer incidence rate), and/or different types of females may be present in the sample. And, regarding *assessing* between-group heterogeneity, novometry demonstrated that although a point estimate of *Type I error* for an application may meet the criterion for statistical reliability (i.e., $p < 0.05$), the exact discrete 95% CIs for *classification performance* achieved by the model and by chance may overlap, indicating statistically unreliable heterogeneity. Furthermore, in this demonstration use of the D statistic revealed that although many effects were *statistically significant* ($p < 0.05$), few were *ecologically significant* ($ESS > 50$, $D < 2$). Presently a moderate-to-strong effect was identified for five cancer sites on the basis of sex, and for eight sites on the basis of race. Race also emerged as a strong predictor for three cancer sites. Race was also much more likely than sex to identify GO-CTA models involving one attribute and three or more strata: 29 versus 3 cancer sites, respectively. Rarely seen prior to the development of novometry, unfolding research suggests that granular parses of ordered attributes may underlie many classical phenomena. Finally, findings indicated that race is particularly susceptible to different types of paradoxical confounding. This issue may be exacerbated if more than two ethnic groups (as were studied presently) are involved. Therefore the different groups should be represented as a multicategorical variable, rather than being combined and risking paradoxical confounding (Chapter 2).

Obtaining a GO-CTA Model: Binary Class Variable, Multiple Attributes

Identification of a GO-CTA model for an application involving a binary class variable and multiple potential attributes is illustrated for an application discriminating a sample of $N = 823$ Emergency Department (ED) patients self-reporting a “fair” versus a “good” likelihood of recommending the ED to others (this class variable, called SAT in CTA software syntax presented below, was coded 0 and 1, respectively).²

The study was set in an urban 800-bed university-based level-1 Trauma center having an annual census of $N = 48,000$ patients. One week after being discharged, patients were mailed a survey assessing satisfaction with care received in the ED. The survey elicited self-ratings of the likelihood of the patient recommending the ED to others, and satisfaction with the administrative, nurse, physician, and laboratory facets of care. A total of $N = 2,109$ surveys having completed recommendation ratings were returned in a six-month period (17% return rate). Survey items were completed via five-point Likert-type scales: scores of 1 (very poor, $N = 182$) and 2 (poor, $N = 92$) indicate unlikely to recommend; scores of 3 (fair, $N = 239$) indicate ambivalence; and scores of 4 (good, $N = 584$) and 5 (very good, $N = 1,012$) indicate the patient is likely to recommend the ED to others. In this example the potential attributes were satisfaction ratings of aspects of care received from physicians: p1 = waiting time; p2 = physician courtesy; p3 = physician took patient’s problem seriously; p4 = concern for patient’s comfort; p5 = explanation of test/treatment; and p6 = explanation of illness/injury. Items were completed using five-point Likert-type scales: 1 = very poor satisfaction, 2 = poor, 3 = fair, 4 = good, and 5 = very good satisfaction.

Using exact minimum precision statistical power analysis (Figure 3.2), for a moderate effect $N = 32$ observations per class category (thus minimum endpoint denominator = $2 \times 32 = 64$ observations) are needed to obtain 90% power for a confirmatory test (generalized $p < 0.05$) of the alternative hypothesis that patients who are more satisfied with care they received are more likely to recommend the ED to others (the null hypothesis is this is not true). For a relatively strong effect $N = 12$ observations per class category, or $N = 24$ observations per endpoint, are needed.

The GO-CTA model is identified in two ways. In the first analysis MDSA is used to identify the descendant family of all optimal models underlying the data. In the second analysis SDA is used to identify the attributes that underlie the GO-CTA model in order to demonstrate the analytic efficiency gain that is realized by reducing the number of attributes evaluated by MDSA.

MDSA without SDA

The first step of MDSA analysis identifies the initial optimal model in the descendant family, for which the minimum endpoint denominator attains the smallest (most granular) value that is possible for the sample. This initial model is identified by conducting an *unrestricted* EO-CTA analysis in which CTA software⁶ (see Appendix C) syntax *doesn't* restrict the endpoint minimum denominator:

```

OPEN satis.dat;                               MISSING all (-9);
OUTPUT satis.out;                            MC ITER 10000 CUTOFF .05 STOP 99.9;
VARS sat p1 to p6;                          PRUNE .05;
CLASS sat;                                  ENUMERATE;
ATTR p1 to p6;                             GO;
```

Unrestricted analyses occasionally identify GO-CTA models, usually in applications having a small sample and few attributes. However, in more complex applications initial optimal models are often untenable due to limitations such as excessive complexity, unacceptably small (i.e., statistically underpowered) minimum endpoint sample size, or unreasonably large D statistics.

Presently all six of the attributes are ordered Likert-type scores. However, imagine one (or more) of the attributes (p1 and p4, for example) was categorical with two or more qualitative categories: in this case one additional command (CAT p1 p4;) would be used.

Table 12.36 summarizes MDSA results for the present example (the exact 95% CIs for model and chance performance are not presented). As seen, the initial optimal model identified in MDSA step 1 used all of the attributes except for p2, and was nearly ideal as indicated by its minute D statistic (all 14 models in the descendant family were nearly ideal).

Table 12.36: Summary of MDSA Findings *without SDA*

Step	Attribute(s)	ESS	Endpoints	Efficiency	D	Minimum Endpoint N	CPU Seconds
1	1, 3-6	96.63	10	9.663	0.349	2	124
2	1, 3-5	96.40	9	10.711	0.336	3	109
3	1-3, 6	95.81	8	11.976	0.350	4	106
4	1, 3-5	95.47	7	13.639	0.332	8	106
5	1, 3, 4	94.59	6	15.765	0.343	18	103
6	1, 2, 4, 6	91.40	6	15.233	0.565	20	107
7	1, 4, 5	91.22	4	22.805	0.385	38	106
8	1, 6	90.66	4	22.665	0.412	39	105
9	1, 2, 4	88.78	4	22.195	0.506	42	104
10	1, 4	85.96	3	28.653	0.490	94	102
11	6	85.88	2	42.940	0.329	123	66
12	3	82.70	2	41.350	0.418	201	37
13	2	78.92	2	39.460	0.534	215	32
14	1	78.39	2	39.195	0.551	233	19

Note that the minimum endpoint $N = 2$ for the initial optimal model identified in step 1. The next model in the descendant family is obtained in MDSA step 2 by conducting a *restricted* EO-CTA analysis by

forcing the minimum endpoint denominator to $N \geq 3$. This is accomplished by adding the following CTA software syntax to the program presented above:

MIND 3;

And, for example, the seventh optimal model in the descendant family is identified in MDSA step 7 by increasing the minimum endpoint denominator to $N \geq 20$ (Table 12.36) by modifying the syntax:

MIND 20;

This procedure is continued until the minimum endpoint denominator is increased sufficiently so that no additional models can be identified, at which point the MDSA analysis is completed.

Presently the minimum endpoint N for all of the optimal models identified in the first six MDSA steps violate Axiom 1 (adequate statistical power) of novometric theory, and all optimal models identified beginning in step 7 satisfy Axiom 1 so long as there are at least $N = 12$ observations in each class category.

Identified in MDSA step 11, the GO-CTA model in this application is a two-strata parse (a UniODA model) of p6: if explanation of illness/injury = good or very good then predict that the patient is likely to recommend; otherwise predict that the patient is unlikely to recommend the ED to others. For the model the exact discrete 95% CIs are 78.1 – 92.8 for *ESS*, 39.1 – 46.4 for *efficiency*, and 0.56 – 0.16 for *D*; and for chance, corresponding exact discrete 95% CIs are 0.07 – 6.46, 0.04 – 3.23, and 2,855 – 29.0, respectively. The GO-CTA model is presented in Figure 12.5, and Table 12.37 gives the associated confusion table.

Figure 12.5: GO-CTA Model Predicting Likelihood of Recommending the ED to Others

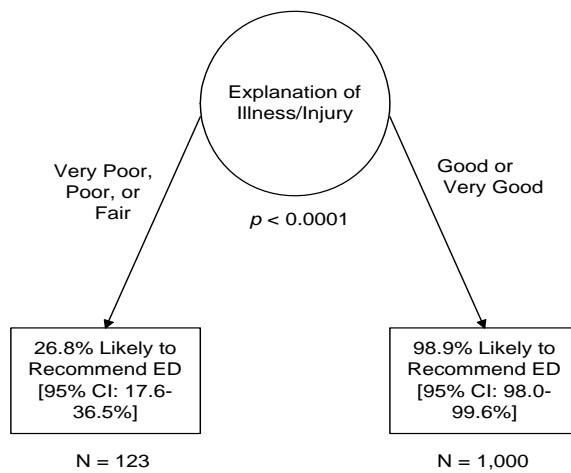


Table 12.37: Confusion Table for GO-CTA Model

		Predicted Recommendation	
		3	4
Actual	3	90	11
	4	33	989

As seen the GO-CTA model correctly classified $90 / 101 = 89.1\%$ of the ambivalent patients, and $989 / 1,000 = 98.9\%$ of the patients who are likely to recommend the ED to others. Note that because the number of actual code = 3 (unlikely to recommend) observations misclassified by the model, $N = 11$, is less than the minimum number that is required in each class category by the *a priori* statistical power analysis, the right-hand endpoint has inadequate statistical power to support further model development from this

endpoint for this sample. And, although sufficient statistical power exists for the actual code = 4 (likely to recommend) observations that were misclassified by the GO-CTA model, models that included more attributes for this model left-hand branch increased the distance of the resulting EO-CTA model from a theoretically ideal model.

The descendant family (Table 12.36) required a total of 1,226 CPU seconds to identify by running CTA software⁶ on a 3 GHz Intel Pentium D microcomputer.

MDSA with SDA

Originally known as an “iterative ODA-based decomposition procedure”, SDA was initially developed as an algorithm used to identify structure underlying sequential data in Markov, turnover, and autocorrelation tables.⁴ An optimal (maximum-accuracy) analogue of principal components analysis⁵, in novometry SDA is (technically) used to identify a family of models that (conceptually) define *the set of attributes* that most accurately predict a monotonically decreasing number of events in a contingency table. Axioms 2 and 3 of novometric theory state the GO-CTA model for a given application is identified by applying MDSA to the attribute subset that is identified by SDA. Prior to the discovery of novometric theory, SDA defined model accuracy on the basis of the *ESS* obtained by iterative UniODA analyses—that (for ordered attributes) all produced two-strata (two endpoint) parses. However novometric analysis makes it possible to identify the descendant family for each individual attribute, and the identification of optimal multicategorical parses involving three or more strata isn’t uncommon. Therefore, in novometry SDA defines model accuracy on the basis of the *D* statistic, which is normed across parsimony—that is, the number of strata in the model.

SDA consists of *S* steps. In step *S* = 1, EO-CTA and MDSA are applied to each individual attribute, and the GO-CTA model that satisfies the minimum denominator criterion, has $p < 0.05$, and achieves the minimum *D* statistic is selected as the first model (i.e., attribute).

In step *S* = 2 the observations correctly classified in *S* = 1 are deleted from the total data set, and the attribute used in the GO-CTA identified in *S* = 1 is deleted from the CTA syntax (the ATTR command). As was done for *S* = 1, for this reduced sample EO-CTA and MDSA are applied to each attribute, and the GO-CTA model that satisfies the minimum denominator criterion, has $p < 0.05$, and yields the minimum *D* statistic is selected as the second model (i.e., attribute).

SDA terminates if all of the observations are correctly classified, if all of the observations in either class category are correctly classified, if Axiom 1 is violated, or if $p > 0.05$. Upon termination the attribute set identified in prior SDA steps are retained for MDSA analysis.

Here EO-CTA was conducted separately for each attribute, and only two-strata models emerged. In the first step of the SDA, p6 yielded the strongest EO-CTA model: *ESS* = 85.55, *D* = 0.338. As discussed above, when observations correctly classified by this model are eliminated from the total sample in order to conduct the next SDA step, the number of misclassified REC = 3 (ambivalent) observations remaining in the reduced sample is too small and violates Axiom 1 concerning adequate statistical power. Thus, in this application the SDA terminates after the first step, and the GO-CTA model for the application is an EO-CTA model that is identified using MDSA analysis. However, step 1 of the SDA revealed that the GO-CTA model for p6 is a binary parse, and no additional models exist for the sample data. Thus, for MDSA conducted with SDA, step 11 in Table 12.36 summarizes the findings: the GO-CTA model was discovered in one step requiring 123 CPU seconds to complete—which is 10% of the CPU time required to identify the GO-CTA model for this application using MDLA without first using SDA to identify the crucial attribute(s).

In this example using SDA to identify the optimal attribute set (OAS) for analysis vis-à-vis MDSA resulted in an order-of-magnitude reduction in the computational resources used to identify the GO-CTA model. Applied research conducted in the ODA laboratory has shown that using SDA to identify the OAS in more complex applications (e.g., samples with $N \geq 1,000$ observations assessed on 40 or more attributes having missing data and yielding weak-to-moderate *ESS*) is often crucial: otherwise the initial unrestricted step of MDSA may be computationally intractable. The record “intractable” analysis conducted in the ODA laboratory was an unrestricted run that failed to solve after one CPU-week elapsed. In contrast, using SDA to obtain the OAS, EO-CTA identified the descendant family and GO-CTA model for this application in a few CPU minutes.

Obtaining a GO-CTA Model: Unrestricted Class Variable and Attribute(s)

The algorithm that identifies (weighted) GO-CTA models for applications with class variables measured at any precision level ranging from binary to mult categorial to real numbers was recently discovered, and laboratory software has been successfully tested for applications having ordinal as well as mult categorial class variables. Unlike the UniODA, MegaODA and CTA software systems, commercially-available special-purpose software designed to conduct automated unrestricted GO-CTA analysis isn't yet available. When unrestricted GO-CTA software becomes commercially-available it will be announced on the News tab and will be available through the Resources tab of the *Optimal Data Analysis* eJournal: www.ODAJournal.com.

Analysis Involving Missing Data

When developing a linear multiattribute statistical model in an application involving missing data, every observation missing data on any attribute included in the model is omitted from analysis. In contrast, with CTA an observation is omitted only if missing data for any of the attributes that are actually used to assign the observation to a specific model endpoint. Nevertheless, as is well-illustrated in the following example, samples having attributes with sporadically-missing data are capable of inducing analytic pandemonium.¹⁶

A retrospective Emergency Department-based matched case-control study obtained data on nine attributes (Table 12.38) for $N = 100$ radiographically-confirmed cases of community-acquired pneumonia (CAP) and $N = 100$ cases of influenza-like illness (ILI). SDA was dominated by attributes having small N and strong predictive value, thus rendering an analysis sample unable to provide sufficient statistical power to satisfy Axiom 1 of novometric theory.

Table 12.38: Attributes and Number of Missing Values used by Descendant Family Optimal Models

<u>Attribute</u>	<u>Used in Models</u>	<u>N Missing</u>
WBC	1-10	105
Temperature	1-3, 5-9, 11-14	6
Pulse Oximetry	1,2,13	12
Dyspnea	1-3, 8, 11	29
Respiration Rate	1, 7	7
Sore Throat	4	122
Fever	4, 5, 12	20
Heart Rate	5, 11, 12	6
Wheezing	6, 10	1

Note: WBC = White Blood Cell count. Fever = binary indicator of whether or not temperature exceeded 100.4 degrees Fahrenheit. Model number indicates MDSA step number (Table 12.39).

Summarized in Table 12.39, a descendant family of 14 unique CTA models emerged. Using exact minimum precision statistical power analysis (Figure 3.3), for a relatively strong effect $N = 16$ observations per class category (minimum endpoint denominator = 32 observations) are needed to achieve 90% power for an exploratory test (experimentwise $p < 0.05$, up to a five-strata model) of the alternative hypothesis that CAP and ILI patients can be discriminated: the null hypothesis is that this is not true.

Optimal models in steps 1 – 10 violate Axiom 1. Although the models identified in steps 1 and 2 are statistically untenable, it may be encouraging to note that the upper-bound of the exact 95% CIs for these models overlap theoretical ideal. Of the models satisfying Axiom 1, the two-parse model identified in step 14 had the smallest D point estimate. Compared to the final model the next-best model (step 13) used an additional endpoint (50% greater complexity) and nevertheless had a greater D statistic (36.5% further from theoretically ideal). Model 14 is thus selected as the GO-CTA model in this application.

Table 12.39: Summary of MDSA Discriminating CAP and ILI Patients

<u>Step</u>	<u>Strata</u>	<u>MinD</u>	<u>ESS</u>	<u>Efficiency</u>	<u>D</u>
1	8	2	95.0	11.9	0.42
			80.0-100	10.0-12.5	2.00-0
			0.83-20.0	0.10-2.50	992-32.0
2	6	6	89.4	14.9	0.71
			74.3-100	12.4-16.7	2.08-0
			0.00-19.2	0.00-3.20	n/a-25.5
3	5	7	81.1	16.2	1.17
			64.2-94.1	12.8-18.8	2.81-0.32
			0.00-19.2	0.00-3.84	n/a-21.0
4	4	9	72.4	18.1	1.52
			45.6-94.1	11.4-23.5	4.77-0.26
			1.89-29.7	0.47-7.42	209-9.48
5	6	11	77.6	12.9	1.73
			64.6-88.9	10.8-14.8	3.26-0.76
			0.88-17.5	0.15-2.92	661-28.2
6	4	18	69.3	17.3	1.77
			53.8-83.3	13.4-20.8	3.46-0.81
			1.65-19.2	0.41-4.80	240-16.8
7	4	19	63.9	16.0	2.26
			47.9-78.4	12.0-19.6	4.33-1.10
			1.57-18.8	0.39-4.70	252-17.3
8	4	21	57.7	14.4	2.93
			40.8-73.7	10.2-18.4	5.80-1.43
			0.28-17.8	0.07-4.46	1425-18.4
9	4	21	56.7	14.2	3.05
			35.3-75.0	8.82-18.8	7.34-1.32
			1.67-22.8	0.42-5.70	234-13.5
10	3	30	54.0	18.0	2.55
			40.5-67.5	13.5-22.5	4.41-1.44
			0.30-20.6	0.10-6.87	997-11.6
11	4	35	50.9	12.7	3.86
			35.1-66.3	8.78-16.6	7.39-2.02
			0.27-14.7	0.07-3.68	1425-23.2
12	4	42	50.7	12.7	3.89
			35.0-65.6	8.75-16.4	7.43-2.10
			1.02-14.7	0.26-3.68	381-23.2
13	3	45	43.8	14.6	3.85
			28.3-58.8	9.44-19.6	7.59-2.10
			0.75-14.3	0.25-4.77	397-18.0
14	2	88	41.5	20.8	2.82
			26.0-56.5	13.0-28.2	5.69-1.55
			0.19-14.2	0.10-7.12	1051-12.0

Table 12.40 gives descriptive information on all 14 models in the descendant family (Attributes = number of different attributes used in the model; model sensitivity is percent of correctly classified CAP and ILI patients). The GO-CTA model was the most simple (parsimonious) optimal model in the family, and it classified the largest number of observations ($N = 194$), and the largest number of both CAP ($N = 99$) and ILI ($N = 95$) patients. Note how missing data induce instability in the number of patients with CAP and ILI that are classified by successive models in the descendant family, particularly for the ILI patients, and for steps 1 through 10 of the MDSA analysis.

Table 12.40: Descriptive Information for Optimal Models in the Descendant Family

Model	<u>N</u>	Attributes	Model Sensitivity			
			CAP	<u>N</u>	ILI	<u>N</u>
1	92	5	100	72	95.0	20
2	92	4	94.4	72	95.0	20
3	92	3	86.1	72	95.0	20
4	49	3	84.8	33	87.5	16
5	133	4	82.9	76	94.7	57
6	111	3	75.0	76	94.3	35
7	123	3	72.4	76	91.5	47
8	138	3	75.9	83	81.8	55
9	95	2	66.2	74	90.5	21
10	111	2	54.0	76	100	35
11	167	3	75.6	90	75.3	77
12	178	3	72.5	91	78.2	87
13	186	2	79.0	95	64.8	91
14	194	1	65.7	99	75.8	95

How can research in this area be improved, what do these results suggest? The most informative attribute with excessive missing values presently may be white blood cell count: all of the highly-accurate (Table 12.39) optimal models 1, 2, 3, and 5 through 9 used white blood cell count as well as temperature as attributes (Table 12.38). These results motivate a focused follow-up study with an independent random sample having complete data for three measures: diagnosis (CAP or ILI), temperature, and white blood cell count. Statistical power analysis should assume a two-attribute, three-endpoint model and a relatively strong effect, and should test the *a priori* hypothesis that higher temperature and white blood cell count both predict CAP.

The Best is Yet to Come

Anyone who just completed reading this book, especially if previously unfamiliar with the ODA paradigm, has covered a vast conceptual territory. We generally recommend that all completed initial readers return to the beginning and skim the material quickly once over, to maximize conceptual understanding vis-à-vis integrating the concepts presented in the book as a whole, and by tying together any loose-ends that may have been created by the complexity of the material, especially since the specific presentation sometimes required simultaneous understanding of multiple new, not-yet-introduced procedures or concepts to be fully understood. This is the nature of a *new statistical paradigm*—of a new theoretical perspective and a new analytic methodology.

Also reflecting the nature of a new statistical paradigm, much more than what has already been learned lies ahead waiting to be discovered and developed. Universally true, contributing in exploration and development of the ODA (or any) paradigm requires active participation vis-à-vis theoretical and

empirical research, as well as teaching and mentoring. Ultimately this will require use of ODA software systems available for individual and institutional use (see the resources tab at www.ODAJournal.com).

Questions and comments concerning the material presented in this book may be addressed to the authors as is described in the About tab at the link above. New developments in maximum-accuracy theory and methods are also available at the link above: for example, peer-reviewed journal articles vis-à-vis the open source *Optimal Data Analysis* eJournal (sending ODA a link to your work involving maximum-accuracy methods will ensure it appears in the ODA web-page list of published manuscripts); publications in other journals and books vis-à-vis the Publications tab; resources such as new books or software, and seminars vis-à-vis the Resources tab; and news vis-à-vis the News tab.

We look forward to reading about your discoveries obtained using optimal statistical procedures in the journals. Thank you for your interest in our work, and our optimal wishes for great success in yours.

Paul and Rob

Appendix A

UniODA™ and MegaODA™

Command Syntax

UniODA and MegaODA software contain a flexible scripting language enabling users to precisely specify the nature of the analysis: a myriad of experimental structures may be defined by using combinations of available commands. This software can be used to analyze problems involving 500 variables, $N = 65,536$ observations (MegaODA is capable of analyzing an unlimited number of attributes and samples as large as three million observations), and 16 groups (used in the Gen feature). Following is an alphabetical list of the software syntax, and explanations of associated keywords.

ATTRIBUTE

Syntax ATTRIBUTE *variable list* ;

Alias ATTR

Remarks ATTR specifies the attribute(s) to be used in the analysis, and it is mandatory unless TABLE input is specified. A separate analysis will be run for each attribute named. The TO keyword may be used to define multiple variables in the variable list. For example, the command:

ATTR A1 to A4 ;

indicates that A1, A2, A3, and A4 will all be treated as attributes. Elaboration of the TO keyword may be found in the discussion under VARS.

CATEGORICAL

Syntax CATEGORICAL {ON | OFF} ;

 CATEGORICAL *variable list* ;

Alias CAT

Remarks CAT specifies that categorical analysis will be used, and is mandatory when the attribute to be analyzed is categorical. Using the ON keyword indicates that all attributes in the variable list are categorical. CAT with no parameters is the same as CAT ON. The TO keyword may be used in the variable list (see discussion under VARS). When using TABLE input, CAT analysis is assumed, and it is not necessary to specify this command.

CLASS

Syntax CLASS *variable list* ;

 CLASS {ROW | COL} ;

Remarks This mandatory command specifies the class variable to be used in the analysis. If more than one class variable is named, a separate analysis will be run for each. ROW and COL are used for TABLE input to indicate whether the rows or columns of the table are to be used as the class variable. Otherwise, the TO keyword may be used in the variable list (see the discussion under VARS).

DATA

Syntax DATA ;

Remarks The DATA command indicates that the entries that follow the command are data to be used in the current analysis. The END statement terminates the data block. For example, the following commands enter hypothetical data on two variables for each of two observations:

```
DATA;  
1 2  
3 4  
END;
```

Each line should correspond to a single observation. The END statement must be the only command on the line it appears in.

DEGEN

Syntax DEGEN {ON | OFF} ;

DEGEN *variable list* ;

Alias DEGENERATE

Remarks DEGEN specifies whether or not degenerate cutpoints are allowed. If DEGEN OFF is specified, the resulting ODA solution must have at least one observation assigned to each predicted class. DEGEN allows flexibility in data sets which have small or no representation in some classes. The default is OFF. DEGEN with no parameters is the same as DEGEN ON. The TO keyword may be used in the variable list (see the discussion under VARS).

DIRECTION

Syntax DIRECTION {< | LT | > | GT | OFF} *value list* ;

Aliases DIR, DIRECTIONAL

Remarks The DIRECTION command specifies the presence and nature of a directional (i.e., an a priori or one-tailed) hypothesis. The parameter < or LT indicates that the class values in the value list are ordered in the “less than” direction. The parameter > or GT indicates the class values are ordered in the “greater than” direction. The value list must contain every value of the class variable currently defined. The default is OFF.

EXCLUDE

Syntax EXCLUDE *variable* {= | <> | < | > | <= | >= | OFF} *value* (,*value2*,...) ... ;

Aliases	EX, EXCL
Remarks	This command excludes observations with the indicated <i>value of variable</i> . For example,
	EXCLUDE C=3 ;
	tells ODA to drop all observations with the value of 3 for variable C. Also, the command
	EXCLUDE A=1 G>=902 ;
	drops all observations with the values of 1 for variable A or values greater than or equal to 902 for variable G. Commas in the exclude string enable the user to exclude multiple values of a variable with a single command:
	EXCLUDE B=1,3 ;
	excludes all observations which have a value of 1 or 3 for variable B. Multiple EXCLUDE commands may be entered, up to a maximum of 100 clauses. The system will exclude observations that satisfy any of the EXCLUDE clauses. EXCLUDE is not allowed with TABLE input.

FREE

See VARS.

GO

Syntax	GO ;
Remarks	The GO command begins execution of the currently defined analysis.

GEN

Syntax	GEN {OFF} <i>variable</i> ;
	GEN TABLE <i>g</i> ;
Alias	GROUP
Remarks	The GEN command specifies the variable whose (integer) values indicate groups in a multisample (GEN) analysis. If TABLE has been specified, then GEN TABLE <i>g</i> indicates that <i>g</i> tables are present, corresponding to <i>g</i> GEN groups. The default is OFF.

HOLDOUT

Syntax	HOLDOUT <i>path\file name</i> ;
Alias	HOLD
Remarks	HOLDOUT specifies the file name to be used for holdout (validity) analysis. The variable list for the holdout file must be in the same order as that for the main input file. OFF turns the holdout indicator off.

ID

Syntax ID {OFF} *variable* ;

Remarks The ID command defines the ID variable that is to be printed in the long report. The default is OFF.

INCLUDE

Syntax INCLUDE *variable* {= | < | > | <= | >= | OFF} *value* (,*value2*,...) ... ;

Aliases IN, INCL

Remarks The INCLUDE command functions in the same way as the EXCLUDE command, except that ODA will keep only those observations with the indicated *values* for *variable*. If multiple INCLUDE statements exist, only those observations will be kept which satisfy all these INCLUDE statements. INCLUDE is not allowed with TABLE input.

LOO

Syntax LOO {ON | OFF} ;

Remarks The LOO command specifies that a leave-one-out (jackknife) analysis will be performed. LOO is not allowed in WEIGHTed CATEGORICAL problems. The default is OFF. LOO with no parameters is the same as LOO ON.

MCARLO

Syntax MCARLO {ITERATIONS *value* | SECONDS *value* | TARGET *value* | SIDAK *value* | STOP *value* | STOPUP *value* | ADJUST | OFF} ;

Alias MC

Remarks The MCARLO command controls Monte Carlo analysis for estimating Type I error, or *p*. The keywords specify a number of stopping criteria; if any criterion is met, then the analysis stops. ITERATIONS (ITER) specifies the maximum number of Monte Carlo iterations, SECONDS (SEC) specifies the maximum number of seconds before the analysis terminates. TARGET specifies a target significance level. SIDAK adjusts the target to reflect a Sidak (Bonferroni) adjustment of the TARGET level, in which *value* is an integer that indicates the number of experiments involved in the adjustment (see Chapter 3). STOP indicates the confidence level (in percent) that the estimated Type I error rate is *less* than the TARGET value, at which point the analysis stops. STOPUP indicates the confidence level (in percent) that the estimated Type I error rate is *greater* than the TARGET value, at which point the analysis stops. For example, the command

```
MCARLO ITER 1000 SEC 30 TARGET .01 STOP 99.9 STOPUP 99 ;
```

indicates that a Monte Carlo analysis will be conducted, and will stop when one of the following occurs: (1) 1000 iterations have been executed, (2) 30 seconds have elapsed, (3) a confidence level of 99.9% has been obtained for *p* < 0.01, or (4) a confidence level of 99% has been obtained for *p* > 0.01. The default Monte Carlo method is conservative in the estimation of significance, in that a Monte Carlo iteration, whose optimal value is tied with

the optimal value obtained from the original analysis, is always counted toward a higher significance level. Specifying ADJUST will adjust for this boundary by splitting these tied iterations in half. The default is OFF.

MISSING

Syntax MISSING {*variable list* | ALL} (*value*) ;

Alias MISS

Remarks The MISSING command tells ODA to treat observations with value (*value*) as missing for each variable on the list. For example, the command

```
MISSING A B C (-1) ;
```

indicates that observations with variables A, B, or C equal to -1 will be dropped if they are present in a CLASS, ATTRIBUTE, WEIGHT, GROUP, or ID variable. ALL specifies that the indicated missing value applies to all variables. The TO keyword may be used in the variable list (see the discussion under VARS).

OPEN

Syntax OPEN {*path\file name* | DATA} ;

Remarks The OPEN command specifies the data file to be processed by ODA. This file must be in ASCII format. DATA indicates that a DATA statement, with inline data following, appears in the command stream.

OUTPUT

Syntax OUTPUT *path\file name* {APPEND} ;

Remarks The OUTPUT command specifies the output file containing the results of the ODA run. The default is ODA.OUT. APPEND indicates that the report is to be appended to the end of an already existing output file.

PRIMARY

Syntax PRIMARY {MAXSENS | MEANSENS | SAMPLEREP | BALANCED | SENS *value* | DISTANCE | RANDOM | GENMEAN | GENSENS *value* | DEFAULT} ;

Alias PRI

Remarks The PRIMARY command specifies the primary criterion for choosing among multiple optimal solutions. MAXSENS (or MAXPAC) is maximum sensitivity. MEANSENS (MEANPAC) is the mean of the sensitivities of the separate classes. SAMPLEREP (SREP) selects the pattern of predicted class membership most closely resembling the sample class membership. BALANCED (BAL) selects the solution in which the sensitivity of the actual classes is most similar amongst each other. SENS (PAC) selects the solution with maximum sensitivity of class *value*. DISTANCE (DIST) selects the solution with smallest maximum (over all cutpoints) distance between the cutpoints and their boundaries. RANDOM (RAND) selects a randomly chosen solution. GENMEAN is used only when GEN is in effect: it selects the solution with maximum mean (weighted) sensitivity over all GEN groups. GENSENS

selects the solution with the maximum (weighted) sensitivity of group *value*. The default is MAXSENS when PRIORS is ON and MEANSENS otherwise.

PRIORS

Syntax PRIORS {ON | OFF} ;

Remarks The PRIORS command indicates whether or not the ODA criterion will be weighted by the reciprocal of sample class membership. The default is ON. PRIORS with no parameters is the same as PRIORS ON.

QUIT

Syntax QUIT ;

Remarks Use the QUIT command to exit from ODA immediately.

REPORT

Syntax REPORT {SHORT | LONG} ;

Alias REP

Remarks The REPORT command specifies whether the short or long report is to be generated. The LONG report additionally prints the predicted and actual class memberships for each observation (ordered analysis) or for each cell (categorical analysis). The default is SHORT.

RESET

Syntax RESET ;

Remarks Use this command to reset all parameters to their default values.

SECONDARY

Syntax SECONDARY {MAXSENS | MEANSENS | SAMPLEREP | BALANCED |

 SENS *value* | DISTANCE | RANDOM | GENMEAN | GENSENS

value | DEFAULT} ;

Alias SEC

Remarks The SECONDARY command specifies the secondary criterion for choosing among multiple optimal solutions. The default is SAMPLEREP. See the entry for PRIMARY for definitions of the above criteria.

SEED

Syntax SEED {*value* | TIME | 0} ;

Remarks The SEED command supplies the seed value for random number generation. TIME or 0 indicate that the current time will be used for the seed. If this command is not present, the time at program initiation will be used.

TABLE

Syntax TABLE *row (col)* ;

Alias FREE TABLE

Remarks The TABLE command is used for categorical analysis only, and indicates that a *row*- by-*col* table is present in the input file. If only *row* is entered, a square *row*-by-*row* table is assumed. For the sake of illustration, imagine that the following 2-by-2 table constitutes the data one wishes to analyze:

	Column 1	Column 2
Row 1	5	6
Row 2	7	8

In the ODA script for this illustration, the statement TABLE 2 would be used to indicate that the 2-by-2 table was to be input. CLASS ROW would indicate that the rows were to be considered the class variable. If the table is rectangular, the CLASS command should reflect the smaller value of *row* or *col*. For example, if

TABLE 4 3 ;

was entered, the user should then enter

CLASS COL ; .

When using TABLE input, a CATEGORICAL analysis is assumed, and it is not necessary to specify this command.

TITLE

Syntax TITLE *title* ;

Remarks The TITLE command specifies the title to be printed in the report. TITLE with no parameters erases the currently defined title.

VARS

Syntax VARS *variable list* ;

Alias FREE

Remarks The VARS command specifies a list of variable names corresponding to fields in the input data set. The TO keyword may be used to define multiple variables in the variable list. For example, the command

VARS A B C X1 TO X5 ;

specifies that the input file contains, in order, variables A, B, C, X1, X2, X3, X4, and X5, and that there is at least one blank space separating all adjacent data. Alternatively, the data points may be separated by a single comma (with no spaces).

The TO keyword may only be used to input a range of variables that have the same name except for the integer at the end of the name: the integers must be positive and ascending, increasing one unit per variable. Thus, VAR1 TO VAR10 is admissible (defining 10 variables). In contrast, VAR10 TO VAR1, VARA TO VARJ, or A TO X10, are not admissible.

The data for each observation may all exist on a single line of the data set, or may be spread on multiple adjacent lines. It is not recommended that a new observation be included on a line that contains data from the previous observation.

WEIGHT

Syntax WEIGHT {*variable* | OFF} ;

Alias RETURN

Remarks The optional WEIGHT command specifies the weight variable for the analysis. The data values for the WEIGHT variable supply the weight the corresponding observation. The default is OFF.

Running ODA Software

There are basically two methods for running ODA software (UniODA, MegaODA, CTA). The first method involves the “Programmers File Editor” (PFE), an intuitive, easy-to-use integrated editing application for text files created by Alan Phillips. With PFE one can write and edit ODA scripts and data files, execute analyses, and view outputs from multiple runs. The use of PFE is described in our initial book on the ODA paradigm¹ that is available at many academic libraries, and/or vis-à-vis inter-library-loan, world-wide. The second method is using the command line editor, in the MS-DOS command prompt window. The use of the DOS (command) prompt is described in Chapter 3 and also in your Windows™ documentation. The decision regarding which system to use is a matter of personal preference, both systems get the job done. Among the authors Paul prefers to use the DOS prompt window, and Rob prefers to use PFE.

Appendix B

MegaODA Time Trials

MegaODA™ software is capable of conducting UniODA analysis for an unlimited number of attributes and samples as large as three million observations. To minimize computational burden associated with Monte Carlo (MC) simulation used to estimate the exact Type I error rate (p), the first step in statistical analysis is identifying effects that are *not* statistically significant (ns).

Study 1: Identifying ns Effects

This study presents an experimental simulation exploring the ability of MegaODA to identify ns effects in a host of designs involving a binary class variable under challenging discrimination conditions (all data are random) for sample sizes of $N = 10^5$ and $N = 10^6$.

A data set was constructed that had three independently generated random numbers for each of $N = 10^6$ observations. The first attribute was binary, defined on the basis of a random probability value (p) generated from a uniform distribution: BINARY = 0 if $p \leq 0.05$; BINARY = 1 if $p > 0.05$. The second attribute was an ordered 5-point Likert-type scale created using another random p generated from another uniform distribution: LIKERT = 1 if $p < 0.2$; LIKERT = 2 if $0.2 \leq p < 0.4$; LIKERT = 3 if $0.4 \leq p < 0.6$; LIKERT = 4 if $0.6 \leq p < 0.8$; and LIKERT = 5 if $p > 0.8$. The third attribute, a bounded real-number, was a random probability value generated from a uniform distribution: RANDOM = p .

Observations were assigned a unique ID number (an integer between 1 and 10^6) used to partition the sample. Two analyses are reported, the first for observations having ID $\leq 100,000$, and the second for the complete sample. The trials begin using the former relatively large sample.

Relatively Large Samples

For the sample of $N = 100,000$, BINARY was evenly distributed with $N = 49,978$ Class = 0 and $N = 50,022$ Class = 1 observations. Table B.1 presents descriptive statistics for LIKERT and RANDOM data separately by BINARY—the class variable in analyses reported herein.

Table B.1: Descriptive Statistics for LIKERT and RANDOM Data by *Class*: $N = 100,000$

<u>Variable</u>	<u>Statistic</u>	<u>Class 0</u>	<u>Class 1</u>
LIKERT	Mean	2.996	3.012
	SD	1.418	1.412
	Median	3	3
	Skewness	0.001	-0.014
	Kurtosis	-1.308	-1.297
RANDOM	Mean	0.501	0.500
	SD	0.288	0.290
	Median	0.501	0.497
	Skewness	-0.003	0.007
	Kurtosis	-1.197	-1.210

Categorical Attribute: In the first pair of time trials the LIKERT attribute was treated as though it had been measured using a *categorical scale* with five response categories: this configuration is called a rectangular categorical design. The exploratory hypothesis that class can be discriminated using the five-category categorical attribute was tested using the following UniODA and MegaODA syntax:

```

OPEN total.dat;                                CAT likert;
OUTPUT total.out;                             EX id>100000;
VARS id binary likert random;                MC ITER 1000 TARGET .05 STOPUP 99.9;
CLASS binary;                                 GO;
ATTR likert;

```

MC simulation is parameterized to target generalized “per-comparison” $p < 0.05$ for a single test of a statistical hypothesis. A maximum allowance of 1000 MC experiments is indicated. STOPUP ceases simulation when the specified confidence that estimated p exceeds target p is achieved. Analysis stopped after 100 MC iterations: estimated exact $p < 0.23$, confidence for target $p > 0.10$ is $> 99.9\%$. With negligible *ESS* (0.65) and *ESP* (0.68), the model required < 1 CPU second to solve (analyses herein were run using a 3 GHz Intel Pentium D microcomputer). A leave-one-out (LOO) validity analysis conducted in a separate run finished in < 1 CPU second.

The exploratory hypothesis that class can be discriminated vis-à-vis the five-category categorical attribute, with observations *weighted* by a *continuous* variable (RANDOM), was tested using the $N = 10^5$ sample by adding the following syntax:

```

WEIGHT random;
GO;

```

Analysis stopped at 200 MC iterations: estimated exact $p < 0.12$, confidence for target $p > 0.05$ is $> 99.99\%$. Having a negligible weighted *ESS* (0.91) and *ESP* (0.95), analysis was completed in 1 CPU second. LOO analysis is not available for weighted categorical designs.

Ordinal Attribute: The second set of two time trials treated LIKERT as being an *ordinal* attribute having five discrete response categories. The exploratory hypothesis that class can be discriminated using the five-level ordinal attribute was tested using the prior MegaODA syntax, after first commenting-out the WEIGHT and CAT commands:

```

*CAT likert;
*WEIGHT random;
GO;

```

Analysis stopped in 300 MC iterations: estimated exact $p < 0.094$; confidence for target $p > 0.05$ is $> 99.9\%$. With negligible *ESS* (0.65) and *ESP* (0.68), the model was obtained in 2 CPU seconds (the model identified was identical to the model for the corresponding categorical analysis).

The exploratory hypothesis that class can be discriminated using the five-level ordinal attribute, with observations *weighted* by a *continuous* variable (RANDOM), was tested by amending the syntax:

```

WEIGHT random;
GO;

```

Analysis stopped at 1,000 MC iterations: estimated exact $p < 0.026$, confidence for target $p > 0.05$ is $> 0.1\%$. With negligible weighted *ESS* (0.91) and *ESP* (0.95), the model was obtained in 6 CPU seconds. MC simulation also reported confidence for target $p < 0.05$ is $> 99.99\%$.

Continuous Attribute: The third set of two time trials treated RANDOM as a *continuous* attribute. The exploratory hypothesis that class can be discriminated via *continuous* RANDOM scores was tested via syntax amended as shown:

```

ATTR random;
*WEIGHT random;
GO;

```

Analysis stopped in 100 MC iterations: estimated exact $p < 0.21$; confidence for target $p > 0.10$ is $> 99.9\%$. With negligible *ESS* (0.65) and *ESP* (0.79) analysis required 99 CPU seconds to complete.

The exploratory hypothesis that class can be discriminated using *continuous* LIKERT scores, with observations *weighted* by an *ordinal* variable (RANDOM), was tested by amending the code:

```

WEIGHT likert;
GO;

```

Analysis stopped in 100 MC iterations: estimated exact $p < 0.72$, confidence for target $p > 0.10$ is $> 99.99\%$. With negligible weighted *ESS* (0.66) and *ESP* (0.80), analysis was completed in 49 CPU seconds.

Multiple-Sample Analysis: The final time trial for relatively large samples used the Generalizability (Gen) procedure that identifies the UniODA model that—simultaneously and independently applied to multiple samples—*maximizes* the *minimum ESS* yielded by the model across the samples. The exploratory hypothesis that class can be discriminated using a Gen model involving a *continuous* attribute (RANDOM) independently applied to five different *samples* (LIKERT) was tested via the following syntax:

CLASS binary;	GEN likert;
ATTR random;	GO;

Analysis stopped in 100 MC iterations: estimated exact $p < 0.33$; confidence for target $p > 0.10$ is $> 99.99\%$. Having negligible *ESS* (0.49) and *ESP* (1.13) the model required 78 CPU seconds to complete.

Big Data Samples

For the $N = 10^6$ sample BINARY was evenly distributed: $N = 499,928$ *Class* = 0 and $N = 500,072$ *Class* = 1 observations. Descriptive statistics for LIKERT and RANDOM data are presented in Table B.2 separately by BINARY, the class variable in analyses presented below.

Table B.2: Descriptive Statistics for LIKERT and RANDOM Data by Class: $N = 1,000,000$

Variable	Statistic	Class 0	Class 1
LIKERT	Mean	2.997	3.003
	SD	1.414	1.415
	Median	3	3
	Skewness	0.001	-0.004
	Kurtosis	-1.300	-1.301
RANDOM	Mean	0.500	0.500
	SD	0.289	0.289
	Median	0.501	0.500
	Skewness	-0.002	0.001
	Kurtosis	-1.200	-1.202

All seven time trial experiments run for the relatively large ($N = 10^5$) sample are also run for the BIG DATA sample. UniODA and MegaODA syntax is the same as used previously, with EX commented-out:

```
*EX ID>100000;
```

Categorical attribute: Analysis for the exploratory hypothesis that class can be discriminated by the five-category categorical attribute stopped after 100 MC iterations: estimated exact $p < 0.19$, confidence for target $p > 0.05$ is $> 99.99\%$. With negligible ESS (0.20) and ESP (0.21), the model required 3 CPU seconds to solve. LOO analysis conducted in a separate run required < 1 CPU second to complete.

Analysis for the exploratory hypothesis that class can be discriminated using the five-category categorical attribute, with observations *weighted* by a *continuous* variable (RANDOM) stopped after 200 MC iterations: estimated exact $p < 0.12$, confidence for target $p > 0.05$ is $> 99.99\%$. Returning negligible weighted ESS (0.26) and ESP (0.27), the model required 9 CPU seconds to solve.

Ordinal attribute: Analysis for the exploratory hypothesis that class can be discriminated via the five-category ordinal attribute stopped after 300 MC iterations: estimated exact $p < 0.107$, confidence for target $p > 0.05$ is $> 99.99\%$. With negligible ESS (0.20) and ESP (0.21), the model required 42 CPU seconds to solve.

Analysis for the exploratory hypothesis that class can be discriminated on the basis of the five-category ordinal attribute, with observations *weighted* by a *continuous* variable (RANDOM) stopped after 1,000 MC iterations: estimated exact $p < 0.049$, confidence for target $p > 0.05$ is $> 47.69\%$, confidence for $p < 0.05$ is $> 52.31\%$. Having a negligible weighted ESS (0.26) and ESP (0.27), the model required 144 CPU seconds to solve. This is an example of the most resource intensive scenario: clarification of the estimated exact p in this application would require more MC experiments.

Continuous attribute: Analysis for the exploratory hypothesis that class can be discriminated via *continuous* RANDOM scores terminated after 14 MC iterations: estimated exact $p < 0.58$, confidence for target $p > 0.05$ is $> 99.99\%$. With negligible ESS (0.16) and ESP (0.18), analysis finished in 140 CPU seconds.

Analysis for the exploratory hypothesis that class can be discriminated by *continuous* RANDOM scores, with observations *weighted* by an *ordinal* variable (LIKERT) terminated after 28 MC iterations: estimated exact $p < 0.86$, confidence for target $p > 0.10$ is $> 99.99\%$. With negligible weighted ESS (0.17) and ESP (0.18), the model required 127 CPU seconds to solve.

Multiple-sample analysis: Analysis for the exploratory hypothesis that class can be discriminated by a single model using a *continuous* attribute (RANDOM) independently applied to five different *samples* (LIKERT) terminated after only 12 MC iterations: estimated exact $p < 0.17$, confidence for target $p > 0.05$ is $> 99.98\%$. With negligible ESS (0.09) and ESP (0.11), the model required 131 CPU seconds to complete.

Study 2: Identifying Statistically Significant Effects

In research involving multiple tests of statistical hypotheses the efficiency of Monte Carlo (MC) simulation used to estimate p is maximized using a two-step procedure. The first step involves identifying the effects that are *ns*. The second step is verifying that remaining effects have $p < 0.05$ at either the generalized or the experimentwise criterion, necessary in order to reject the null hypothesis and accept the alternative hypothesis that a statistically significant effect occurred. This study uses MC simulation to investigate the ability of MegaODA to identify $p < 0.05$ effects in a host of designs with a binary class variable and ordered attribute for mildly or weakly challenging discrimination conditions (i.e., moderate or modest distribution overlap, respectively), target p 's of 0.01 and 0.001, and sample sizes of $N = 100,000$ and $N = 1,000,000$.

The data set constructed in Study 1 was used to create two additional data sets. The first data set featured a *relatively weak* effect ($ESS \leq 25$) and a *moderate* effect ($25 < ESS \leq 50$): the latter was created by *adding* $\frac{1}{2}$ SD to both the LIKERT and RANDOM data of $class = 1$ observations. Descriptive statistics for the data are presented in Table B.3 (CV=coefficient of variation).

The second data set featured a *moderate* to *relatively strong* effect ($50 < ESS \leq 75$). It was created using the first data set, but also *subtracting* $\frac{1}{2}$ SD from the LIKERT and the RANDOM data of the $class = 0$ observations. Then the value 1 was added to all of the weights, because in ODA software weights must all be positive numbers. Descriptive statistics for the data are presented in Table B.4.

Table B.5 presents results for analysis with an *ordinal attribute* and a *continuous weight*, and Table B.6 gives results for analysis with a *continuous attribute* and an *ordinal weight*. Comparing findings *within* Table B.5 and Table B.6 using MegaODA revealed that sample size was a statistically significant discriminator of solution time at the generalized criterion, and comparison *between* Table B.5 and Table B.6

showed the use of a continuous attribute was a statistically significant discriminator of solution time at the experimentwise criterion. The CPU time used in rule-in analyses with continuous attributes can be reduced by dividing the attribute into fewer segments and utilizing weights so as to perfectly reproduce the actual score.

Table B.3: Descriptive Statistics for LIKERT and RANDOM by Class: Weak-Moderate *ESS*

Variable	Statistic	<i>N</i> =49,978		<i>N</i> =50,022		Variable	Statistic	<i>N</i> =499,928		<i>N</i> =500,072	
		Class 0	Class 1	Class 0	Class 1			Class 0	Class 1	Class 0	Class 1
LIKERT	Mean	2.996	3.719	LIKERT		Mean	2.997	3.710	RANDOM		
	SD	1.418	1.412			SD	1.414	1.415			
	Median	3	3.707			Median	3	3.707			
	CV	47.3	38.0			CV	47.2	38.1			
	Skewness	0.001	-0.014			Skewness	0.001	-0.004			
	Kurtosis	-1.308	-1.297			Kurtosis	-1.300	-1.301			
RANDOM	Mean	0.501	0.750	RANDOM		Mean	0.500	0.750			
	SD	0.288	0.290			SD	0.289	0.289			
	Median	0.501	0.747			Median	0.501	0.750			
	CV	57.6	38.7			CV	57.7	38.5			
	Skewness	-0.003	0.007			Skewness	-0.002	0.001			
	Kurtosis	-1.197	-1.210			Kurtosis	-1.199	-1.202			

Table B.4: Descriptive Statistics for LIKERT and RANDOM by Class: Moderate-Strong *ESS*

Variable	Statistic	<i>N</i> =49,978		<i>N</i> =50,022		Variable	Statistic	<i>N</i> =499,928		<i>N</i> =500,072	
		Class 0	Class 1	Class 0	Class 1			Class 0	Class 1	Class 0	Class 1
LIKERT	Mean	2.289	3.719	LIKERT		Mean	2.290	3.710	RANDOM		
	SD	1.418	1.412			SD	1.414	1.415			
	Median	2.293	3.707			Median	2.293	3.707			
	CV	62.0	38.0			CV	61.7	38.1			
	Skewness	0.001	-0.014			Skewness	0.001	-0.004			
	Kurtosis	-1.308	-1.300			Kurtosis	-1.300	-1.301			
RANDOM	Mean	1.251	1.750	RANDOM		Mean	1.250	1.750			
	SD	0.288	0.290			SD	0.289	0.289			
	Median	1.251	1.747			Median	1.251	1.750			
	CV	23.1	16.6			CV	23.1	16.5			
	Skewness	-0.003	0.007			Skewness	-0.002	0.001			
	Kurtosis	-1.200	-1.210			Kurtosis	-1.200	-1.202			

Solution speeds ranged from 5 to more than 83,000 CPU seconds running MegaODA software on a 3 GHz Intel Pentium D microcomputer. Using MegaODA it is straightforward to rapidly rule-in $p < 0.05$ for weak and moderate effects by MC simulation with large samples and with BIG DATA in designs having ordinal attributes with or without the use of weights applied to observations. Although significantly longer time was required to solve problems involving continuous attributes, even the most computer-intensive analyses investigated presently were completed in less than one day.

Table B.5: Simulation Results for *Ordinal Attribute*, Continuous Weight

Effect Strength	<u>ESS</u>	<u>ESP</u>	<u>N</u>	<u>Target p</u>	CPU Seconds	MC Iterations	<u>Weighted</u>
Weak	20.5	24.4	100,000	0.01	5	700	No
Weak	20.5	24.4	100,000	0.001	59	9,300	No
Weak	20.3	65.2	100,000	0.01	5	700	Yes
Weak	20.3	65.2	100,000	0.001	66	9,300	Yes
Weak	20.2	24.1	1,000,000	0.01	104	700	No
Weak	20.2	24.1	1,000,000	0.001	1,647	9,300	No
Weak	20.0	65.2	1,000,000	0.01	112	700	Yes
Weak	20.0	65.2	1,000,000	0.001	1,581	9,300	Yes
Moderate	40.4	42.0	100,000	0.01	4	700	No
Moderate	40.4	42.0	100,000	0.001	64	9,300	No
Moderate	40.2	70.1	100,000	0.01	4	700	Yes
Moderate	40.2	70.1	100,000	0.001	61	9,300	Yes
Moderate	40.2	41.8	1,000,000	0.01	109	700	No
Moderate	40.2	41.8	1,000,000	0.001	1,462	9,300	No
Moderate	40.0	70.0	1,000,000	0.01	94	700	Yes
Moderate	40.0	70.0	1,000,000	0.001	1,508	9,300	Yes

Table B.6: Simulation Results for *Continuous Attribute*, Ordinal Weight

Effect Strength	<u>ESS</u>	<u>ESP</u>	<u>N</u>	<u>Target p</u>	CPU Seconds	MC Iterations	<u>Weighted</u>
Moderate	25.2	56.8	100,000	0.01	831	700	No
Moderate	25.2	56.8	100,000	0.001	3,669	5,000	No
Moderate	24.8	62.3	100,000	0.01	822	700	Yes
Moderate	24.8	62.3	100,000	0.001	3,336	5,000	Yes
Moderate	25.1	33.5	1,000,000	0.01	4,708	700	No
Moderate	25.1	33.5	1,000,000	0.001	83,596	5,000	No
Moderate	25.0	62.2	1,000,000	0.01	31,311	700	Yes
Moderate	25.0	62.2	1,000,000	0.001	80,175	5,000	Yes
Strong	49.9	56.7	100,000	0.01	274	700	No
Strong	49.9	56.7	100,000	0.001	1,727	5,000	No
Strong	49.8	76.4	100,000	0.01	414	700	Yes
Strong	49.8	76.4	100,000	0.001	942	5,000	Yes
Strong	50.1	52.4	1,000,000	0.01	2,650	700	No
Strong	50.1	52.4	1,000,000	0.001	19,979	5,000	No
Strong	50.0	76.4	1,000,000	0.01	3,417	700	Yes
Strong	50.0	76.4	1,000,000	0.001	18,621	5,000	Yes

Study 3: Binary Designs

This third and final time trial of MegaODA software studies the fastest-to-analyze application, a 2×2 cross-classification table also known as a binary design. Study 2 investigated rule-in analysis of ordered attributes, but rule-in hasn't yet been considered for categorical attributes. Study 3 thus presents time

trials for rule-out and rule-in analysis conducted for a binary design. In this study the special-purpose TABLE algorithm available in UniODA and MegaODA software (designed to maximize solution speed) is used to conduct the analyses, and MC simulation is not run because exact p is computed (see Chapter 2).

Table B.7: Experimental Data for Study of Categorical Attributes: $N = 10^5$ Sample

<i>Weak Effect</i>			<i>Moderate Effect</i>		
Attribute			Attribute		
Class Variable	0	1	Class Variable	0	1
0	25,000	25,000	0	35,000	15,000
1	20,000	30,000	1	15,000	35,000

Seen in Table A2.7, the study involved four 2×2 tables featuring *weak* ($ESS = 10.0$, $ESP = 10.1$) and *moderate* ($ESS = ESP = 40.0$) effects for designs having $N = 100,000$ and $N = 1,000,000$ (each tabled value was multiplied by ten to create the $N = 10^6$ data). The following UniODA and MegaODA syntax was used to analyze these 2×2 tables.

OPEN DATA;	DATA;
OUTPUT weak100K.out;	25000 25000
CATEGORICAL ON;	20000 30000
TABLE 2;	END DATA;
CLASS ROW;	GO;

All analyses involving large or BIG DATA samples, and weak or moderate effects, were completed in fractions of a CPU second: reported use of 0 CPU seconds in ODA software output indicates < 0.5 CPU second elapsed. Chaining 100 sequential runs of the largest problem to estimate solution time revealed that the largest individual analyses required approximately 0.15 CPU seconds to complete.

Appendix C

CTA™ Command Syntax

ATTRIBUTE

Syntax ATTRIBUTE *variable list* ;

Alias ATTR

Remarks The ATTRIBUTE command lists the attribute(s) to be used in the analysis. The TO keyword may be used to define multiple attributes in the list. For example, the command

ATTR A1 to A4;

indicates that A1, A2, A3 and A4 will be treated as attributes. Further exposition of the TO keyword is found in the discussion for VARS.

CATEGORICAL

Syntax CATEGORICAL {ON | OFF} ;
CATEGORICAL *variable list* ;

Alias CAT

Remarks The CATEGORICAL command specifies that categorical analysis will be used, and is required when the attribute to be analyzed is categorical. Using the ON keyword indicates that all variables in the variable list are categorical. CAT with no parameters is the same as CAT ON. The TO keyword may be used in the variable list (see the discussion under VARS).

CLASS

Syntax CLASS *variable list* ;

Remarks The mandatory CLASS command specifies the class variable to be used in the analysis. A separate analysis will be run for each class variable named. The TO keyword may be used in the variable list (see discussion under VARS).

DIRECTION

Syntax DIRECTION {< | LT | > | GT |
OFF} *value list* ;

Aliases DIR, DIRECTIONAL

Remarks The DIRECTION command defines the presence and nature of a directional (*i.e., a priori*, one-tailed, or confirmatory) hypothesis. The parameter < or LT indicates that the class values in the value list are ordered in the “less than” direction. The parameter > or GT indicates the class values are ordered in the “greater than” direction. The value list must contain every value of the class variable currently defined. The default is OFF.

ENUMERATE

Syntax ENUMERATE {ROOT} {MINOBS *value*} ;

Remarks The ENUMERATE command with no options specifies that all combinations of attributes in the top three nodes will evaluated. ENUMERATE ROOT specifies that only the top node will have all attributes evaluated. ENUMERATE MINOBS *value* allows only solution trees with at least *value* observations in them.

EXCLUDE

Syntax EXCLUDE *variable* {= | <> | < | > |
<= | >= | OFF} *value* (,*value2*,...);

Aliases EX, EXCL

Remarks This command excludes observations having the indicated *value* of *variable*. For example,

 EXCLUDE D=4 ;

drops all observations with the value of 4 for attribute D. The command

 EXCLUDE B=2 Z>=113 ;

drops all observations with the value of 2 for attribute B or values greater than or equal to 113 for attribute Z. Commas in the exclude string enable the user to exclude multiple values of a variable using a single command:

 EXCLUDE C=2,4 ;

excludes all observations having a value of 2 or 4 for attribute C. Multiple EXCLUDE commands may be entered, up to a maximum of 100 clauses. Observations which satisfy any of the EXCLUDE clauses will be excluded.

FORCENODE

Syntax FORCENODE *node var* ;

Remarks The FORCENODE command forces CTA to insert the attribute *var* at node *node* in the solution tree. If the UniODA solution for this attribute is not significant, or this node is subsequently pruned, an error message will be printed.

GO

Syntax GO ;

Remarks The GO command begins execution of the currently defined analysis.

INCLUDE

Syntax	INCLUDE <i>variable</i> {= < > <= >= OFF} <i>value</i> (, <i>value2</i> ,...);
Aliases	IN, INCL
Remarks	The INCLUDE command functions in the same manner as the EXCLUDE command, except that only those observations with the indicated <i>value</i> for <i>variable</i> are included. If multiple INCLUDE statements exist, only those observations will be kept which satisfy all these INCLUDE statements.

LOO

Syntax	LOO { <i>pvalue</i> STABLE};
Remarks	The LOO command indicates that leave-one-out analysis will be performed for every attribute in the tree. LOO STABLE allows only attributes with LOO ESS equal to the ESS for that attribute. LOO <i>pvalue</i> allows only those attributes in the solution tree which have an ESS that yields a $p \leq pvalue$.

MCARLO

Syntax	MCARLO {ITERATIONS <i>value</i> CUTOFF <i>pvalue</i> STOP <i>confvalue</i> };
Alias	MC
Remarks	The MCARLO command controls Monte Carlo analysis for estimating Type I error, or <i>p</i> . The keywords specify stopping criteria; if any criterion is met, then the analysis stops. ITERATIONS (ITER) specifies the maximum number of Monte Carlo iterations. STOP <i>xxx</i> indicates the confidence level (in percent), which will stop processing for the current attribute, if the estimated Type I error rate (specified with the CUTOFF keyword) drops below this level. For example, the command

```
MCARLO ITER 70000  
CUTOFF .05 STOP 99.9 ;
```

indicates a Monte Carlo analysis will be conducted, and will stop when one of the following occurs: (1) 70,000 iterations have been executed, (2) a confidence level of less than 99.9% that $p < .05$ has been obtained.

MAXLEVEL

Syntax	MAXLEVEL <i>value</i> ;
Remarks	The MAXLEVEL command specifies the deepest level or depth allowed in the solution tree.

MINDENOM

Syntax	MINDENOM <i>value</i> ;
Remarks	The MINDENOM command specifies that only attributes which yield a denominator of <i>value</i> or more will be allowed in the solution tree.

MISSING

Syntax	MISSING {variable list ALL} (value) ;
Alias	MISS
Remarks	The MISSING command tells ODA to treat observations with value (value) as missing for each variable on the list. For example, the command
MISSING X Y Z (-4) ;	
indicates that observations with attributes X, Y, or Z equal to -4 will be dropped if they are present in a CLASS, ATTRIBUTE, WEIGHT, or GROUP variable. ALL specifies that the indicated missing value applies to all variables. The TO keyword may be used in the attribute list (see discussion under VARS).	

OPEN

Syntax	OPEN {path\file name DATA} ;
Remarks	The OPEN command specifies the data file to be processed by ODA. This file must be in ASCII format. DATA indicates that a DATA statement, with inline data following, appears in the command stream.

OUTPUT

Syntax	OUTPUT path\file name {APPEND} ;
Remarks	The OUTPUT command specifies the output file containing the results of the ODA run. The default is ODA.OUT. APPEND indicates that the report is to be appended to the end of an already existing output file.

PRIORS

Syntax	PRIORS {ON OFF} ;
Remarks	The PRIORS command indicates whether the ODA criterion will be weighted by the reciprocal of sample class membership. The default is ON. PRIORS with no parameters is the same as PRIORS ON.

PRUNE

Syntax	PRUNE pvalue {NOPRIORS} ;
Remarks	The PRUNE command indicates the p-value with which to optimally prune the classification tree. The NOPRIORS keyword should be used when PRIORS is turned OFF.

SKIPNODE

Syntax	SKIPNODE node ;
Remarks	The SKIPNODE command specifies that the node node will be empty of any attribute in the solution tree.

TITLE

Syntax TITLE *title* ;

Remarks The TITLE command specifies the title to be printed in the report. TITLE with no parameters erases the currently defined title.

USEFISHER

Syntax USEFISHER *value* ;

Remarks The USEFISHER command specifies that all probability calculations for categorical variable will be determined by Fisher's exact test, rather than by Monte Carlo.

VARS

Syntax VARS *variable list* ;

Remarks The VARS command specifies a list of attribute names corresponding to fields in the input data set. The TO keyword may be used to define multiple variables in the variable list. For example, the command

VARS X Y Z V1 TO V4 ;

specifies that the input file contains, in order, variables X, Y, Z, V1, V2, V3, and V4, and that there is at least one blank space separating all adjacent data. Alternatively, the data points may be separated by a single comma (with no spaces).

The TO keyword may be used to input a range of variables which have the same name except for the integer at the end of the name: the integers must be positive and ascending, increasing one unit per variable. Thus, VAR1 TO VAR10 is admissible (defining 10 variables). In contrast, VAR10 TO VAR1, VARA TO VARJ, or A TO X10, are not admissible. The data for each observation may all exist on a single line of the data set, or may be placed on multiple adjacent lines. It is not recommended that a new observation is included on a line containing data from the previous observation.

WEIGHT

Syntax WEIGHT {*variable* | OFF} ;

Alias RETURN

Remarks The optional WEIGHT command specifies the weight variable for the analysis. The data values for the WEIGHT variable supply the weight the corresponding observation. The default is OFF.

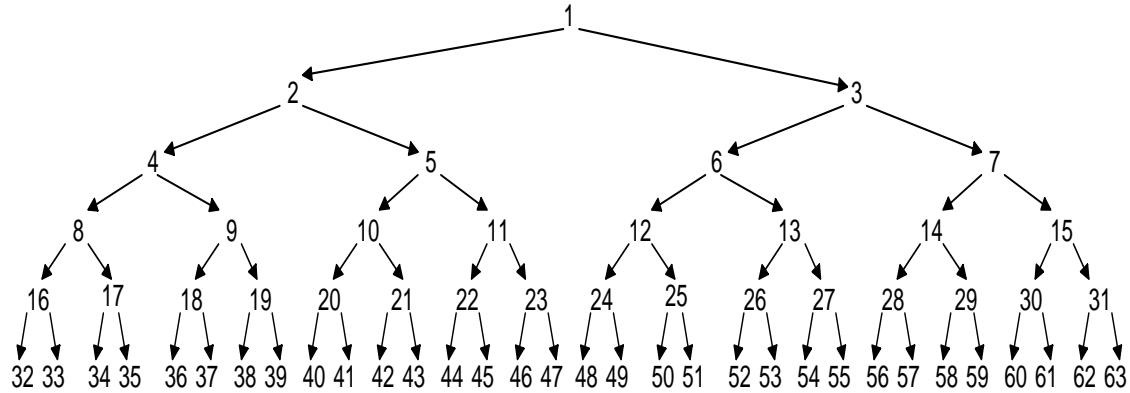
Interpreting Automated CTA Software Output

Automated CTA software reports models using an intuitive shorthand notation to describe model nodes. To facilitate clarity, Figure C.1 is a schematic illustration of the node structure underlying all CTA models.

It is a straightforward matter to determine the "identity number" of a node existing at a deeper depth than is illustrated in this five-level-deep tree: depth level 1 of the tree includes node 1; level 2 includes nodes 2 and 3; level 3 includes nodes 4 - 7; level 4 includes nodes 8 - 15; level 5 includes nodes 16 - 31; and level 6 includes nodes 32 - 63. For node X the identify number of the node emanating from the

left-hand side of X is $2X$, and from the right-hand side of X is $2X+1$. For example, from node 47, node 94 (2×47) emanates to the left, and node 95 ($94 + 1$) emanates to the right. From node 94, node 188 emanates to the left, node 189 to the right, etcetera. After the root attribute all even-numbered nodes lie on the left-hand branch, and all odd-numbered nodes on the right-hand branch of the tree.

Figure C.1: CTA Node Structure



CTA software produces output employing node identity numbers to describe the CTA model: an example of CTA software output is given in Figure C.2 (hypothetical data). Respectively, automated CTA software output lists attribute name (D2, D3 and D4 loaded in the hypothetical CTA model); node identity number; tree depth level; sample size for the analysis indicated; ESS for the attribute; whether jackknife (LOO) validity analysis was stable or unstable; jackknife ESS; p for jackknife ESS; attribute metric (ORD = ordered, CAT = categorical); and CTA model shorthand.

Figure C.2: Sample CTA Software Output (Hypothetical Data)

ATTRIBUTE	NODE	LEV	OBS	p	ESS	LOO	ESSL	LOOP	TYP	MODEL
<hr/>										
D2	1	1	704	.000	48.44%	STABLE	48.44%	.000	ORD	<=6.2-->4,165/242,68.18%*
										>6.2-->5,375/462,81.17%
D3	3	2	292	.000	41.60%	STABLE	41.60%	.000	ORD	<=4.5-->4,29/63,46.03%
										>4.5-->5,206/229,89.96%*
D4	6	3	62	.039	28.99%	STABLE	28.99%	.039	ORD	<=1.9-->4,18/30,60.00%*
										>1.9-->5,22/32,68.75%*

The root attribute (D2) is listed first in the report. For each attribute the report first indicates the cutpoint and outcome for the left-hand branch emanating from the attribute, and second for the right-hand branch. *Branches ending in model endpoints are marked by an asterisk*. The left-hand branch that emanates from D2 has a cutpoint of ≤ 6.2 units: observations having D2 scores ≤ 6.2 units are predicted to be a member of *class* = 4, and this branch terminates in an endpoint representing $N = 242$ observations, of whom $N = 165$ (68.18%) are correctly classified. The remaining $N = 242 - 165 = 77$ observations with D2 scores ≤ 6.2 units were in reality members of *class* = 5, and therefore were misclassified by this branch of the CTA model.

The right-hand branch emanating from D2 has a cutpoint of > 6.2 units: observations having D2 scores > 6.2 units are predicted to be members of *class* = 5, but this branch does not terminate in a model endpoint. Rather, the model includes attribute D3 at node 3.

The left-hand branch emanating from D3 has a cutpoint of ≤ 4.5 units: observations having D3 scores ≤ 4.5 units are predicted to be members of *class* = 4, but this branch does not terminate in a model endpoint.

The right-hand branch from D3 has a cutpoint of > 4.5 units: observations with D3 scores > 4.5 units are predicted to be members of *class* = 5. This branch terminates in a model endpoint representing a total of $N = 229$ observations, of whom $N = 206$ (89.96%) are correctly classified. The remaining $N = 229 - 206 = 23$ observations having D3 scores > 4.5 units were members of *class* = 4, and thus were misclassified by this branch of the CTA model.

Both branches emanating from D4 terminate in a model endpoint: this is always true for the last attribute listed in the output. The left-hand branch has a cutpoint of ≤ 1.9 units: observations having D4 scores ≤ 1.9 units are predicted to be members of *class* = 4; this endpoint represents $N = 30$ observations of whom $N = 18$ (60.00%) are correctly classified and $N = 30 - 18 = 12$ (40.00%) are misclassified. And, the right-hand branch has a cutpoint of > 1.9 units: observations with D4 scores > 1.9 units are predicted to be members of *class* = 5; this endpoint represents $N = 32$ observations of whom $N = 22$ (68.75%) are correctly classified and $32 - 22 = 10$ (31.25%) are misclassified.

To construct an illustration of the final CTA model, referring to Figure C.2 select nodes 1, 3 and 6: these are attributes and are depicted as circles (Chapter 10). Model branches are depicted using arrows emanating from the left-hand side of the root attribute (D2), the right-hand side of D3, and both sides of D4. Branches terminate in model endpoints depicted using rectangles. Exact p is indicated beneath each attribute, cutpoint values are indicated adjacent to arrows, and the outcome (e.g., the percent of *class* = 1 membership) and the number of observations is reported for each endpoint.

Appendix D

Troubleshooting ODA Software

Including the 2005 and 2016 ODA books, and articles published in the eJournal (www.ODAJournal.com) and in pay-for-view journals, many examples of designs addressed using optimal (maximum-accuracy) methods are available, and many provide ODA software command syntax successfully used in analysis. Following steps and suggested practices provided in this book will decrease the likelihood that errors will occur, especially for projects with a small-to-moderate number of attributes and observations.

The 2005 ODA book gives troubleshooting tips, but experience shows the most common programming error is omitting a semi-colon used to indicate the end of a command; the most common typing error is misspelling; and the most common error is failing to indicate one or more missing values in the data set. No matter what the size of the data set, omitting the value of even one single variable for even one single observation—and failing to indicate the presence of a missing value—can wreak havoc in analysis.

In ODA the old adage “garbage in, garbage out” applies—data must be correct, and intuitive, transparent models make data errors obvious: that some software written for legacy statistical methods gives opaque answers for data ODA software diagnoses as “off-kilter” begs the question of exactly what legacy software is doing when data are incorrect.

In the ODA laboratory, when a data set fails, we “return to formula”, rewriting the original complete data set into a new data set involving a small random subset of the sample of observations—including data for only two variables: the class variable and one attribute. We use a template ODA program—known to work for test problems (e.g., as provided with the 2005 ODA book) to analyze the new data set. If there is a problem, then the data are incorrect. If there is no problem we repeat this procedure using all of the data, and if there is no problem then the data for the two first variables are approved as ready for analysis. This procedure is iterated, each pass adding one new attribute. If this procedure ends prior to completion the variable inducing termination has a problem, and when this procedure completes without a problem the analysis is ready to begin.

Even if program and data are working nominally, research can become an analytic adventure if one is fortunate to be in the right place. It is not uncommon for those using ODA for the first time in a new application—presented with previously unexplored hypothesis structure and/or data geometry, to stage multiple attempts before the final analytic methodology is established. If necessary, ODA staff may be reached (see the eJournal About tab).

Appendix E

Weather Prediction Results

Table E.1: Temperature Prediction via Weighted CTA by US State, for January, February, and March of 2008, using Ipsative Mode Scores, and Published and Computed Raw Mode Scores

State	Month	Ipsative Modes	Published		Computed	
			WESS	Normative Modes	WESS	Normative Modes
Alabama	Jan	B,EE,JJ,MM,2	97.43	EAWR,NAO,PNA	71.30	2,3,9
	Feb	A,C,I,EE,PP	93.80	NAO,SCA	57.74	3,6
	Mar	DD,GG	51.55	-	-	-
Arkansas	Jan	C,R,EE,MM,XX,2	98.54	EPNP,PNA,WP	74.63	3,5,8
	Feb	CC,DD,RR,VV	88.90	EPNP,NAO	63.35	3,5,10
	Mar	II	38.63	-	-	-
Arizona	Jan	C,H,U,YY,1	93.22	NAO,POL,WP	75.80	2,6
	Feb	F,II,PP	72.65	-	-	-
California	Jan	C,BB,GG,VV,WW,YY	98.89	PNA,WP	52.83	2,6
	Feb	RR,TT	74.87	EAWR,EPNP,PNA	76.04	-
Colorado	Jan	I,V,T,SS,WW	95.62	-	-	2,6
	Feb	M,O,P,Q,BB,3	91.70	-	-	-
	Mar	J,SS,1	72.76	NAO	39.74	1,5
Connecticut	Jan	E,K,LL,2	96.43	EA,EAWR,EPNP,NAO,WP	86.62	3,4,5
	Feb	PP,2	50.44	-	-	-
Delaware	Jan	V,EE,MM,2	95.15	EAWR,EPNP,NAO,WP	84.63	3,5,7
	Feb	HH,JJ,PP,SS	73.41	NAO	42.84	3
	Mar	J	37.97	-	-	-
Florida	Jan	A,G,O,MM,PP,YY	98.95	EAWR,EPNP,PNA	89.19	2,3,6
	Feb	D,Q,CC,LL,RR	93.22	NAO	40.50	5
	Mar	K,DD,EE,GG	75.80	-	-	-

Georgia	Jan	P,EE,MM,PP,2	98.13	EAWR,EPNP,PNA	84.04	2,3,9	70.66
	Feb	A,C,H,EE,PP	92.34	NAO,SCA	57.16	3,5	73.47
Iowa	Jan	H,L,V,2	93.51	EPNP,SCA,WP	76.74	3,4,7,8	84.57
	Feb	D,DD,JJ	80.95	EAWR	49.09	3,7	44.18
	Mar	J,HH,LL,PP,1	87.26	PNA	41.15	-	-
Idaho	Jan	C,I,MM,SS,ZZ	94.56	-	-	2,3,6	81.59
	Feb	D,Q,R,BB	86.91	PNA	60.78	-	-
	Mar	D,R,Y,RR	93.86	NAO,PNA,SCA	83.99	1,5	63.82
Illinois	Jan	B,D,E,V,EE,WW,2	99.36	EPNP,PNA,WP	83.52	3,4,8	86.62
	Feb	D,DD,GG,PP	83.40	EAWR,NAO,SCA	66.04	-	-
	Mar	-	-	PNA	39.86	-	-
Indiana	Jan	D,E,K,V,EE,WW	96.61	EPNP,PNA,WP	82.70	3,5,8	82.35
	Feb	K,U,NN,RR	71.01	EAWR,NAO,POL	73.58	3	40.44
	Mar	L,II	57.04	PNA	39.39	1,10	57.22
Kansas	Jan	F,Q,GG,WW,1	96.73	EPNP,WP	59.44	1,3,6,9	69.43
	Feb	V,CC,FF,UU	80.19	EAWR,NAO	60.55	3,6,7,9	82.82
	Mar	D,H,FF	73.00	-	-	-	-
Kentucky	Jan	E,J,V,PP,2	96.20	EAWR,EPNP,NAO	79.37	3,5	73.82
	Feb	F,I,Q,U,RR	96.55	NAO	53.36	3,6	60.14
Louisiana	Jan	U,V,EE,LL,3	96.20	NAO,PNA	69.37	1,2,6	84.22
	Feb	A,C,EE,PP	79.37	NAO	53.71	3,5,6,10	79.19
	Mar	D,DD	52.95	-	-	-	-
Massachusetts	Jan	E,I,K,LL,2	97.72	EA,EAWR,EPNP,NAO,WP	90.06	3,4,5	73.70
	Mar	-	-	-	-	2	38.92
Maryland	Jan	E,G,L,V,RR,UU	98.54	EAWR,EPNP,WP	84.28	3,5,8	71.30
	Feb	Y,RR,XX	69.96	NAO,POL	55.00	3	46.41
Maine	Jan	E,O,LL,2	95.21	EPNP,WP	61.60	3,8	65.81
	Feb	Q,RR,1	76.04	-	-	7	39.63
	Mar	Q	39.10	-	-	-	-
Michigan	Jan	D,E,GG,II	97.37	EAWR,EPNP,WP	81.71	3,5,7,8	86.56
	Feb	I,DD,GG,HH	82.76	EAWR,NAO	53.65	3,7	51.43
	Mar	J,L	57.51	PNA,SCA	59.73	2	44.18
Minnesota	Jan	C,E,CC,1,2	95.73	EAWR,EPNP,PNA,WP	88.49	4,5,8	79.78
	Feb	F,Q,NN,RR	78.08	EAWR	44.71	7,10	61.19
	Mar	J,O,1	82.70	PNA,WP	56.81	2	40.68
Missouri	Jan	D,E,F,EE,GG	94.92	EPNP,PNA,WP	85.74	3,4,7,8	93.98
	Feb	EE,RR,SS,TT,VV	93.44	EAWR,EPNP,NAO,POL	77.93	3,5,7	76.52

Mississippi	Jan	I , V, EE, 2	96.20	EPNP, NAO, PNA	86.91	1, 2, 6	78.73
	Feb	A , C, EE, PP	79.54	NAO	52.78	3, 6, 10	71.95
	Mar	DD, GG	51.32	-	-	-	-
Montana	Jan	E , F, L, ZZ, 2	96.67	EPNP, PNA, SCA, WP	84.04	2, 6, 9	75.45
	Feb	A , G, Q, R	85.62	PNA	47.05	7	49.80
	Mar	CC, GG, TT, 3	80.60	PNA	45.35	1	39.22
North Carolina	Jan	E , Y, MM, XX	95.38	EAWR, EPNP, PNA	86.15	3, 5	71.60
	Feb	D , T, Y, RR, VV	89.83	NAO, SCA	54.94	3, 9	56.52
North Dakota	Jan	C , E, L, WW	96.90	EPNP, PNA, SCA, WP	91.41	1, 3, 5, 7	80.89
	Feb	D , Q, II, RR	94.21	EAWR, PNA	61.84	7	45.35
	Mar	J, GG, 1	77.91	PNA	43.83	-	-
Nebraska	Jan	A , V, DD, 1, 2	95.56	EPNP, WP	57.10	1, 3, 9	70.19
	Feb	Q , DD, RR, TT	86.44	EAWR	43.83	-	-
	Mar	D, LL	74.81	-	-	-	-
New Hampshire	Jan	E , K, JJ, LL, 2	97.49	EA, EPNP, WP	71.89	3, 5, 7	70.72
	Feb	-	-	-	-	7	39.28
	Mar	-	-	-	-	2	40.56
New Jersey	Jan	E , K, H, LL	98.48	EA, EAWR, EPNP, NAO, WP	87.38	3, 4, 5	76.74
	Feb	Y, RR, 1	70.72	-	-	3	40.68
New Mexico	Jan	G , T, RR, UU, ZZ	97.84	EA, NAO	64.64	1, 6	84.16
	Feb	F , G, RR, VV, 1	88.43	NAO	43.25	6	42.84
	Mar	G, Y, 3	73.52	-	-	-	-
Nevada	Jan	C , I, V, SS, ZZ	96.43	-	-	2, 3, 6	86.62
	Feb	RR, TT, WW	76.62	EA, PNA	60.43	-	-
	Mar	1	38.81	NAO	41.44	-	-
New York	Jan	II, MM, XX, 2	97.02	EA, EAWR, EPNP, NAO, WP	89.42	3, 4, 5	77.79
	Mar	L	38.98	-	-	-	-
Ohio	Jan	E , L, V, RR	96.67	EAWR, EPNP, WP	80.65	3, 5, 8	79.43
	Feb	D, GG, HH, PP	81.71	NAO, POL	59.03	3	39.98
	Mar	L, II	56.22	-	-	1, 10	55.93
Oklahoma	Jan	F , K, Q, DD, E, 2	96.90	EA, EPNP	59.15	8	63.35
	Feb	H, EE, RR, TT, VV	85.86	EPNP, NAO	67.15	3, 6, 7	74.17
	Mar	D, J	49.09	-	-	-	-
Oregon	Jan	C , I, EE, MM, PP	91.88	NAO, PNA, WP	83.99	2, 3, 5	81.18
	Feb	Q, R, NN, 3	86.15	PNA	61.72	1, 3, 7	63.35
	Mar	F, R, V, SS, 2	82.58	NAO, PNA, POL	69.08	-	-

Pennsylvania	Jan	E,J,HH,YY	96.96	EAWR,EPNP,NAO,WP	85.80	3,5,8	72.36
	Feb	Q,RR	58.45	NAO	43.42	3,7	56.81
	Mar	L	39.80	-	-	-	-
Rhode Island	Jan	E,K,LL,2	96.84	EA,EAWR,EPNP,NAO,WP	86.50	3,4,5	75.34
	Feb	G,K,2	73.52	-	-	-	-
	Mar	J,Q,CC,EE,XX	71.54	-	-	-	-
South Carolina	Jan	Q,R,MM,RR	96.73	EAWR,EPNP,PNA	85.91	2,3,9	70.89
	Feb	D,Q,JJ,RR	90.01	NAO,SCA	55.29	3,6	62.01
South Dakota	Jan	C,E,L,2	97.25	EPNP,SCA,WP	88.02	5,8	62.30
	Feb	D,Q,II,RR	92.69	EAWR	47.69	7	42.20
	Mar	D,J,DD,1	87.67	-	-	-	-
Tennessee	Jan	I,Q,V,EE,3	94.86	EAWR,EPNP,NAO,PNA	77.85	3,5	69.02
	Feb	D,T,U,RR,TT	87.38	NAO	53.13	3,6	56.75
Texas	Jan	C,EE,GG,NN,RR	92.17	NAO,PNA,POL	68.73	1,2,6	82.99
	Feb	A,M,JJ,RR,WW,3	94.62	NAO	51.96	3,5,10	74.34
	Mar	Y,FF,LL,PP	72.36	-	-	-	-
Utah	Jan	C,I,V,BB,SS,ZZ	96.32	-	-	1,2,6	84.34
	Feb	Q,CC,DD,NN	80.25	PNA	44.59	-	-
	Mar	1	41.61	NAO	43.13	1,5	58.62
Virginia	Jan	E,H,L,V,RR	97.37	EAWR,EPNP,PNA	85.68	3,5	72.06
	Feb	A,H,Y,RR,VV	92.87	NAO	49.50	3,5,9	56.98
Vermont	Jan	E,CC,JJ,LL,2	99.12	EA,EPNP,NAO,WP	73.41	3,5,7	71.30
	Mar	Q	42.72	-	-	-	-
Washington	Jan	L,O,CC,EE,VV	97.78	EA,NAO,PNA,WP	91.06	2,5,6	76.68
	Feb	M,R,EE,WW	88.37	PNA	67.45	1,7	58.56
	Mar	D,H,PP,TT,XX,2	92.93	PNA	57.39	-	-
Wisconsin	Jan	E,M,GG,UU,ZZ	97.84	EAWR,EPNP,PNA	79.31	3,5,8	75.04
	Feb	Q,RR,ZZ,1	74.87	EAWR	44.54	7	48.39
	Mar	L,T,CC,GG,NN	93.10	PNA,SCA	65.81	2	43.60
West Virginia	Jan	E,H,V,EE,LL	98.19	EAWR,EPNP,PNA,SCA	83.46	3,5	76.74
	Feb	D,T,U,LL,RR,TT	95.91	NAO	52.54	3	42.31
Wyoming	Jan	K,DD,MM,YY,ZZ	92.11	-	-	2,3,5	77.50
	Feb	C,G,Q,DD	84.57	-	-	-	-
	Mar	D,F,LL,SS	89.89	NAO	43.37	1	41.03

Table E.2: **Precipitation** Prediction via Weighted CTA by US State, for January, February, and March of 2008, using Ipsative Mode Scores, and Published and Computed Raw Mode Scores

State	Month	Ipsative Modes	Published		Computed	
			WEss	Normative Modes	WEss	Normative Modes
Alabama	Jan	C,O,P,MM,NN	89.01	EA,SCA	64.47	8
	Feb	A,R,T,V,II	87.03	EA	39.45	-
	Mar	I,YY	59.56	-	-	-
Arkansas	Jan	C,R,FF,MM,YY	90.01	NAO,PNA	76.27	1,3,9
	Feb	Q	39.98	-	-	-
	Mar	HH	39.28	-	-	-
Arizona	Jan	G,LL,SS	73.47	EPNP	39.63	9
	Feb	I,J,L,1	87.14	EPNP,SCA	62.83	3,5
	Mar	G,Q,T,JJ,SS	84.51	PNA	38.11	5,7,9
California	Jan	BB,LL,NN,SS,2	94.62	EA	48.92	3,6,8
	Feb	V,SS,XX	68.79	-	-	-
	Mar	C,R,U,SS	84.57	NAO	44.07	-
Colorado	Jan	D,EE	59.44	PNA	52.48	-
	Feb	NN,XX	65.75	SCA	45.59	3,7
	Mar	II,SS,3	76.21	PNA	45.47	-
Connecticut	Jan	V,BB,XX	87.67	-	-	5
	Feb	P,HH	77.26	EAWR	43.54	10
	Mar	G,H,J	51.32	POL	44.07	-
Delaware	Jan	B,RR	57.74	EAWR,NAO	51.49	2
	Feb	C,BB,EE	70.31	-	-	6
	Mar	CC,DD,EE,PP	90.77	NAO,WP	55.35	-
Florida	Jan	F,O,BB,CC,DD	92.11	EA	43.66	3
	Feb	T,EE,VV,2	94.62	-	-	-
	Mar	C,D,O,SS,TT	89.60	-	-	4,5
Georgia	Jan	O,MM,NN	73.76	EA	68.32	6,8
	Feb	C,J,T,SS,WW	91.88	-	-	3
Iowa	Jan	GG,NN	59.38	EAWR,PNA	60.61	1
	Feb	G,I,R,PP	77.97	-	-	-
	Mar	T,EE	56.52	-	-	-
Idaho	Jan	E,L,T,GG,WW,1	98.48	EPNP,PNA,SCA	75.75	1,6,8,9
	Feb	J,M,U,NN,XX	85.86	EA,POL	64.52	5,7
	Mar	I,U,HH,LL,3	89.54	EA,NAO,WP	80.77	2,4,5

Illinois	Jan	H,Q,R,MM,NN	92.99	PNA	50.32	9	46.05
	Feb	Q,U,BB,HH	81.06	-	-	7	39.51
	Mar	E,J,JJ,UU	87.90	-	-	-	-
Indiana	Jan	F,I,EE,HH,PP	91.23	NAO,PNA	72.59	9	40.56
	Feb	R,EE,LL,XX	83.46	-	-	4,7	66.45
	Mar	O,JJ,SS	74.05	-	-	-	-
Kansas	Jan	E,Y,GG,LL	84.72	-	-	3,6	55.91
	Feb	F,K,M,FF	78.08	-	-	-	-
	Mar	D,H,R,2	81.12	PNA	41.55	-	-
Kentucky	Jan	A,V,HH,PP	89.42	PNA,SCA	69.67	1,6	79.95
	Feb	Q,V,II,LL,TT	86.09	-	-	7	50.96
	Mar	G,NN,XX	75.39	-	-	2	53.83
Louisiana	Jan	H,DD,FF,WW	80.77	EA,EPNP	51.61	-	-
	Feb	C,P,T	71.24	-	-	2	40.09
	Mar	A,E,K,FF,WW	85.80	-	-	6,7	64.87
Massachusetts	Jan	-	-	-	-	2	39.45
	Feb	I,SS,WW,1	78.43	-	-	-	-
	Mar	C,G,HH	66.74	POL	50.85	-	-
Maryland	Jan	G,H,WW	69.73	-	-	-	-
	Feb	E,P,Q,YY	88.54	-	-	6	42.96
	Mar	I,HH,RR,VV	94.80	SCA,WP	53.19	-	-
Maine	Jan	HH,WW,YY,2	86.44	-	-	5	39.22
	Feb	J,NN,WW	70.25	-	-	1,5	65.40
	Mar	I,J,HH,SS,1	86.85	POL	43.78	-	-
Michigan	Jan	H,Q,T,GG,MM	86.97	PNA	50.38	1,6	53.95
	Feb	D,DD	68.26	-	-	-	-
Minnesota	Jan	P,FF,GG	77.62	-	-	-	-
	Mar	Q,YY,3	78.43	-	-	-	-
Missouri	Jan	O,Q,R,EE,SS	89.77	PNA	51.55	2,3,8	72.88
	Feb	Q,U	58.85	-	-	-	-
	Mar	L,JJ	62.77	-	-	-	-
Mississippi	Jan	U,V,MM,XX	91.12	EAWR	40.50	-	-
	Feb	J,NN	48.51	-	-	-	-
	Mar	CC,FF,2	73.12	-	-	-	-
Montana	Jan	L,V,FF,GG,VV	96.90	PNA	60.08	2,3,5,6	83.23
	Feb	M,O,B,BB	89.13	EAWR,PNA,POL	76.97	2,7	71.83
	Mar	B,H,M,Q,TT	85.62	-	-	-	-

North Carolina	Jan	MM	41.15	WP	38.98	-	-
	Feb	F,L,R,PP,YY	84.40	-	-	-	-
	Mar	G,EE,PP	73.41	-	-	-	-
North Dakota	Jan	C,D,L,HH	83.34	PNA	46.70	-	-
	Feb	L>NN,WW	61.72	-	-	-	-
	Mar	I	45.35	-	-	-	-
Nebraska	Jan	Q,EE,PP	75.86	-	-	9	39.98
	Feb	M,V,WW,XX	84.34	SCA	39.28	8,10	52.02
	Mar	FF,MM,NN	73.70	PNA	44.77	-	-
New Hampshire	Jan	Q,HH,WW,2	86.44	-	-	5	46.23
	Feb	NN,WW,2	70.89	-	-	-	-
	Mar	H,R,P,HH	85.74	POL	48.57	-	-
New Jersey	Feb	E,P,U,JJ	77.32	EAWR	40.68	-	-
	Mar	J,P,JJ,2	76.50	POL,SCA	56.52	-	-
New Mexico	Jan	O,EE,GG,LL	89.17	-	-	9	46.72
	Feb	A,O,EE,RR,WW	78.08	-	-	3,6	51.96
	Mar	Q,GG,SS	80.19	NAO,PNA	54.88	1,7	55.52
Nevada	Jan	U,LL,SS,YY	89.01	-	-	1	47.34
	Feb	V,DD,RR,SS,XX	92.69	-	-	-	-
	Mar	C,G,U,SS	72.82	EA,NAO	58.09	-	-
New York	Mar	D,H,R,HH,NN	87.90	EPNP	40.39	-	-
Ohio	Jan	U,BB,HH,MM	77.85	NAO,PNA,WP	75.39	1,6	60.08
	Feb	F,P,R,TT	95.79	EAWR	39.45	7	54.24
	Mar	I,SS	62.83	-	-	2,9	55.29
Oklahoma	Jan	D,L,EE,FF,UU	90.88	WP	40.68	6,9	68.73
	Feb	YY	41.03	-	-	1,6	61.48
	Mar	D,H,Q,II	86.85	-	-	2,5	62.77
Oregon	Jan	D,GG,LL,XX,YY	99.59	EPNP,PNA,SCA	78.61	1,6,8,9	89.66
	Feb	P,LL,3	72.36	EA,POL	74.28	6	45.18
	Mar	I,V,FF	76.91	EA,NAO	52.54	2	44.54
Pennsylvania	Jan	J,P,U,MM	69.14	-	-	-	-
	Feb	E,Q,II,TT,WW	90.24	EAWR	40.56	2,7	52.07
	Mar	J,O,SS,XX	79.84	-	-	3	39.86
Rhode Island	Jan	JJ,LL,NN,UU	83.11	-	-	-	-
	Feb	E,P,U	86.15	EAWR	42.14	-	-
	Mar	CC	39.63	EA,POL	71.60	5,9	53.42

South Carolina	Jan	T,JJ	67.45	EA,WP	74.40	6,8	54.59
	Feb	L,R,CC,PP	75.69	-	-	-	-
South Dakota	Jan	Q,FF,TT	76.10	-	-	-	-
	Feb	A,U,LL,ZZ	87.90	-	-	-	-
	Mar	A,H,GG,WW	76.10	-	-	5,10	63.30
Tennessee	Jan	E,P,V,HH,ZZ	90.65	PNA	68.44	1,2,6	80.42
	Mar	I,M	58.27	-	-	2	42.02
Texas	Jan	L,JJ	65.81	EAWR, POL, SCA	50.15	1,6,7,9	88.90
	Feb	F,V,SS,TT,ZZ	89.95	-	-	3,7	59.15
	Mar	D,J,R,XX,2	87.61	-	-	5,7,9	77.56
Utah	Jan	J,SS,XX	77.32	PNA	40.04	1	43.13
	Feb	B,F,M,DD,XX	91.93	-	-	3	49.56
	Mar	NN,SS,WW	73.47	NAO	40.33	2	39.45
Virginia	Jan	G,I	71.71	-	-	-	-
	Feb	C,Q,NN	79.31	EA	40.09	6,8	71.42
	Mar	F,K,CC,MM,PP,RR	96.08	-	-	-	-
Vermont	Jan	H,Q,V	77.15	-	-	5	49.74
	Feb	C,J,K,M,FF	87.03	-	-	-	-
	Mar	J	40.74	EPNP, WP	60.43	-	-
Washington	Jan	J,GG,NN,2	90.65	EA,EAWR	52.78	1,9	57.10
	Feb	-	-	EA, POL	54.35	5,6	61.84
	Mar	I,FF	55.58	EA	45.06	2,10	60.90
Wisconsin	Jan	A,MM,PP	76.97	PNA	47.63	1	48.39
	Feb	G,J,P,R	85.33	-	-	1,7	59.61
	Mar	Q,R,YY,1	83.69	-	-	-	-
West Virginia	Jan	HH,MM,3	81.94	EA,NAO,PNA	73.64	1,6,8	70.72
	Feb	A,C,Q,R	81.77	-	-	-	-
	Mar	D,G,L,M,JJ	94.16	SCA	38.63	2	41.26
Wyoming	Jan	T,YY,1	85.86	EA,PNA,SCA,WP	79.60	2,9	59.91
	Feb	CC,JJ,RR,WW	74.40	SCA	39.22	-	-
	Mar	D,G,BB,HH,TT	86.09	-	-	6	41.67

Chapter References

Chapter 1

- ¹Soltysik RC, Yarnold PR (1993). *ODA 1.0: Optimal Data Analysis for DOS*. Chicago: Optimal Data Analysis, Inc.
- ²Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.
- ³Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, 1, 10-22. URL: <http://optimalprediction.com/files/pdf/V1A2.pdf>
- ⁴Yarnold PR (2014). “A statistical guide for the ethically perplexed” (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, 3, 30-31. URL: <http://optimalprediction.com/files/pdf/V3A12.pdf>
- ⁵Grimm LG, Yarnold PR (Eds). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books, 1995.
- ⁶Grimm LG, Yarnold PR (Eds). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books, 2000.
- ⁷Bryant FB (2015). The Loyola experience (1993-2009): Optimal data analysis in the Department of Psychology. *Optimal Data Analysis*, 1, 4-9. URL: <http://optimalprediction.com/files/pdf/V1A1.pdf>
- ⁸Bryant FB (1994). Analyze your data optimally using ODA 1.0. *Decision Line*, 25, 16-19.
- ⁹Bryant FB (2005). How to make the best of your data [Review of Optimal Data Analysis]. *PsycCRITIQUES-Contemporary Psychology: APA Review of Books*, 50, Article 5 (7 pp). URL: file:///C:/Users/Paul/Downloads/How_to_Make_the_Best_of_Your_Data.pdf
- ¹⁰Yarnold PR, Bryant FB (1994). A measurement model for the Type A self-rating inventory. *Journal of Personality Assessment*, 62, 102-115. DOI: 10.1207/s15327752jpa6201_10
- ¹¹Russell RL, Bryant FB, Estrada AU (1996). Confirmatory P-technique analyses of therapist discourse: High- versus low-quality child therapy sessions. *Journal of Consulting and Clinical Psychology*, 64, 1366-1376. DOI: 10.1037/0022-006X.64.6.1366
- ¹²Brockway JH (1997). *Here today, gone tomorrow: Understanding freshman attrition using person-environment fit theory*. Doctoral dissertation, Loyola University Chicago (112 pp).
- ¹³Elling KA (2000). *Predicting children's emotional responsiveness during therapy sessions*. Doctoral dissertation, Loyola University Chicago (119 pp).
- ¹⁴Bivens AJ (2001). *Accurate classification of child molesters using context variation and Hierarchical Optimal Classification Tree Analysis*. Doctoral dissertation, Loyola University Chicago (110 pp).
- ¹⁵Coakley RM (2004). *Constructing a prospective model of psychosocial resilience in early adolescents with spina bifida: An application of optimal data analysis in pediatric psychology*. Doctoral dissertation, Loyola University Chicago (233 pp).
- ¹⁶Suzuki H (2005). *Prospectively tracing profiles of juvenile delinquents and non-delinquents: An optimal data analysis*. Master's thesis, Loyola University Chicago (87 pp).

¹⁷Hurley CL (2005). *Medical, demographic, and psychological predictors of morbidity and mortality in autologous bone marrow transplant patients*. Doctoral dissertation, Loyola University Chicago (192 pp).

¹⁸Wolf JL (2005). *A meta-analysis of primary preventive interventions targeting the mental health of children and adolescents: A review spanning 1992–2003*. Doctoral dissertation, Loyola University Chicago (128 pp).

¹⁹Kapunga CT (2006). *Individual, parental and peer influences associated with risky sexual behaviors among African-American adolescents*. Doctoral dissertation, Loyola University Chicago (125 pp).

²⁰Laforce M (2006). *A classification profile of high-risk sexual behavior among men who have sex with men*. Doctoral dissertation, Loyola University Chicago (94 pp).

²¹Jandasek BN (2008). *Predictors of social competence in adolescents with spina bifida*. Doctoral dissertation, Loyola University Chicago (199 pp).

²²Snowden J (2008). *Predictors of stepping-up from foster homes to residential treatment: A profile of children in the child welfare system*. Doctoral dissertation, Loyola University Chicago (136 pp).

²³Donenberg GR, Bryant FB, Emerson E, Wilson HW, Pasch KE (2003). Tracing the roots of early sexual debut among adolescents in psychiatric care. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42, 594-608. DOI: 10.1097/01.CHI.0000046833.09750.91

²⁴Coakley RM, Holmbeck GN, Bryant FB (2006). Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: An application of optimal data analysis. *Journal of Pediatric Psychology*, 31, 1084-1099. DOI: 10.1093/jpepsy/jsj032

²⁵Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Clinical Child and Adolescent Psychology*, 36, 8-18. DOI: 10.1080/15374410709336564

²⁶Smart CM, Nelson NW, Sweet JJ, Bryant FB, Berry DTR, Granacher RP, Heilbronner RL (2008). Use of MMPI-2 to predict cognitive effort: A hierarchically optimal classification tree analysis. *Journal of the International Neuropsychological Society*, 14, 842-852. DOI: 10.1017/S1355617708081034

²⁷Han SD, Suzuki H, Drake AI, Jak AJ, Houston WS, Bondi MW (2009). Clinical, cognitive, and genetic predictors of change in job status following traumatic brain injury in a military population. *Journal of Head Trauma Rehabilitation*, 24, 57-64. DOI: 10.1097/HTR.0b013e3181957055

²⁸Lyons AM, Leon SC, Zaddach C, Luboyeski EJ, Richards M (2009). Predictors of clinically significant sexual concerns in a child welfare population. *Journal of Child and Adolescent Trauma*, 2, 28-45. DOI: 10.1080/19361520802675884

²⁹Bryant FB, Yarnold PR (1995). Comparing five alternative factor-models of the Student Jenkins Activity Survey: Separating the wheat from the chaff. *Journal of Personality Assessment*, 64, 145-158. DOI: 10.1207/s15327752jpa6401_10

³⁰Bryant FB, Yarnold PR, Grimm LG (1996). Toward a measurement model of the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality*, 30, 223-247. DOI: 10.1006/jrpe.1996.0015

³¹Yarnold PR, Bryant FB, Nightingale SD, Martin GJ (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health, and Medicine*, 1, 207-221. DOI: 10.1080/13548509608400019

³²Layden BL, Minadeo N, Suhy J, Metreger T, Foley K, Borge G, Crayton J, Bryant FB, Mota de Freitas D (2004). Bi-chemical and psychiatric predictors of Li⁺ response and toxicity in Li⁺-treated bipolar patients. *Bipolar Disorders*, 6, 53-61. DOI: 10.1046/j.1399-5618.2003.00093.x

³³Hoffman LAD (2000). *Marital interaction and depression: A test of the interactional systems model of depression*. Doctoral dissertation, Loyola University Chicago (194 pp).

³⁴Collinge WC, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

³⁵Collinge W, Yarnold PR, Soltysik RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103/1947-2714.118920

³⁶Trafimow D, Marks M (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1-2. DOI: 10.1080/01973533.2015.1012991

³⁷Yarnold PR (2015). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>

³⁸Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, 13, 1015-1021. DOI: 10.1002/sim.4780131004

³⁹Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine*, 10, 601-606. DOI: 10.1007/BF02602743

⁴⁰Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine*, 17, 2405-2414. DOI: 10.1002/(SICI)1097-0258(19981030)17:203.0.CO;2-F

Chapter 2

¹Grimm LG, Yarnold PR (Eds). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books, 1995.

²Grimm LG, Yarnold PR (Eds). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books, 2000.

³Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

⁴Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, 2, 194-197. URL: <http://optimalprediction.com/files/pdf/V2A29.pdf>

⁵Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis*, 2, 202-205. URL: <http://optimalprediction.com/files/pdf/V2A31.pdf>

⁶Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, 2, 220-221. URL: <http://optimalprediction.com/files/pdf/V2A35.pdf>

⁷Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, 54, 646-653. DOI: 10.1177/0013164494054003007

⁸Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, 74, 223-238. DOI: 10.1023/A:1018922421450

⁹Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, 22, 739-752. DOI: 10.1111/j.1540-5915.1991.tb00362

¹⁰Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, 1, 144-160. URL: <http://odajournal.com/2013/09/19/62/>

¹¹Bradley JV (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

¹²Noreen EW (1989). *Computer-intensive methods for testing hypotheses: An introduction*. New York: Wiley.

¹³Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989). *Numerical recipes: The art of scientific computing*. Cambridge: University Press.

¹⁴Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo estimation of two-category two-tailed p . *Optimal Data Analysis*, 1, 43-45. URL: <http://optimalprediction.com/files/pdf/V1A8.pdf>

¹⁵Mielke PW (1984). Meteorological applications of permutation techniques based on distance functions. In P.R. Krishnaiah & P.K. Sen (Eds.), *Handbook of statistics, Volume 4: Nonparametric methods*. New York: North-Holland.

¹⁶Mielke PW (1991). The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Science Reviews*, 31, 55-71. DOI: 10.1175/1520-0493(1981)109<0120:AOMRPP>2.0.CO;2

¹⁷de Cani JS (1984). Balancing Type I risk and loss of power in ordered Bonferroni procedures. *Journal of Educational Psychology*, 6, 1035-1037. DOI: 10.1037/0022-0663.76.6.1035

¹⁸Ryan TA (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26-47. DOI: 10.1037/h0042478

¹⁹Holland BS, Copenhaver MD (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, 43, 417-423. URL: <http://www.jstor.org/stable/2531823>

²⁰Holland BS, Copenhaver MD (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, 104, 145-149. DOI: 10.1037/0033-2909.104.1.145

²¹Maxwell SE, Delaney HD (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

²²Rosenthal R (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

²³Ryan TA (1985). "Ensemble-adjusted" p values: How are they to be weighted? *Psychological Bulletin*, 97, 521-526. DOI: 10.1037/0033-2909.97.3.521

²⁴Rosenthal R, Rubin DB (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*, 6, 1028-1034. DOI: 10.1037//0022-0663.76.6.1028

²⁵Green MA (1988). Evaluating the discriminatory power of a multiple regression model. *Statistics in Medicine*, 7, 519-524. DOI: 10.1002/sim.4780070408

²⁶Hosmer DW, Lemeshow S (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theoretical Methods*, A9, 1043-1069. DOI: 10.1080/03610928008827941

²⁷Lachenbruch PA (1975). *Discriminant analysis*. New York: Hafner.

²⁸Sorum M (1972). Three probabilities of misclassification. *Technometrics*, 14, 309-316. DOI: 10.1080/00401706.1972.10488917

²⁹Azar B (1997). APA task force urges a harder look at data. *APA Monitor*, 3, 26. DOI: 10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2

³⁰Baumeister RF, Tice DM (1996). Should we abandon $p < .05$? (Editorial). *Dialogue*, 11, 11.

³¹Cohen J (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003. DOI: 10.1037/0003-066X.49.12.997

³²Goodman SN, Royall R (1988). Evidence and scientific research. *American Journal of Public Health*, 78, 1568-1574. DOI: 10.2105/AJPH.78.12.1568

- ³³Hagen RL (1997). In praise of the null hypothesis significance test. *American Psychologist*, 52, 15-24. DOI: <http://dx.doi.org/10.1037/0003-066X.52.1.15>
- ³⁴Feinstein AR (1988). Statistical significance versus clinical importance. *Quality of Life and Cardiovascular Care*, 4, 99-102.
- ³⁵Kraemer HC (1992). *Evaluating medical tests*. Newbury Park, CA: Sage.
- ³⁶Bacus JW, Gose EE (1972). Leukocyte pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*, 513-526. DOI: <http://dx.doi.org/10.1109/TSMC.1972.4309161>
- ³⁷Eisenbeis RA (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, 32, 875-900. DOI: 10.1111/j.1540-6261.1977.tb01995.x
- ³⁸Nishikawa K, Kubota Y, Ooi T (1983). Classification of proteins into groups based on amino acid composition and other characters, II: Grouping into four types. *Journal of Biochemistry*, 94, 997-1007.
- ³⁹Yarnold PR (1992). Statistical analysis for single-case designs. In: F.B. Bryant, L. Heath, E. Posavac, J. Edwards, S. Tindale, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Social psychological applications to social issues, Volume 2: Methodological issues in applied social research*. New York: Plenum (pp. 177-197).
- ⁴⁰Sokal RR, Sneath PHA (1963). *Principles of numerical taxonomy*. San Francisco: Freeman.
- ⁴¹Friedman GD (1987). *Primer of epidemiology* (3rd ed.). New York: McGraw-Hill.
- ⁴²Rosner B (1982). *Fundamentals of biostatistics*. Boston: Duxbury.
- ⁴³Yarnold PR (2013). Minimum standards for reporting UniODA findings. *Optimal Data Analysis*, 2, 63-68. URL: <http://optimalprediction.com/files/pdf/V2A11.pdf>
- ⁴⁴Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119. URL: <http://optimalprediction.com/files/pdf/V2A19.pdf>
- ⁴⁵Dawes RM (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, 26, 422-424. DOI: 10.1037/h0044612
- ⁴⁶Greenblatt RL, Mozdzierz GI, Murphry TJ, Trimakas K (1992). A comparison of non-adjusted and bootstrapped methods: Bootstrapped diagnosis might be worth the trouble. *Educational and Psychological Measurement*, 52, 181-187. DOI: 10.1177/001316449205200123
- ⁴⁷McLachlan GJ (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.
- ⁴⁸Stevens J (1992). Applied multivariate statistics for the social sciences (2nd ed.). Hillsdale, NJ: Erlbaum.
- ⁴⁹Rorer LG, Dawes RM (1982). A base-rate bootstrap. *Journal of Consulting and Clinical Psychology*, 50, 419-425. DOI: 10.1037/0022-006X.50.3.419
- ⁵⁰Widiger TA (1983). Utilities and fixed rules: Comments on Finn. *Journal of Abnormal Psychology*, 92, 495-498. DOI: 10.1037/0021-843X.92.4.495
- ⁵¹Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, 66, 660-672. DOI: 10.1037/0022-006X.66.4.660
- ⁵²Wainer H (1991). Adjusting for differential base rates: Lord's Paradox again. *Psychological Bulletin*, 109, 147-151. DOI: 10.1037/0033-2909.109.1.147
- ⁵³Meehl PE, Rosen A (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194-216. DOI: doi/10.1037/h0048070

⁵⁴Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, 56, 656-667. DOI: 10.1177/0013164496056004007

⁵⁵Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1, 67-100. URL: <http://optimalprediction.com/files/pdf/V2A33.pdf>

⁵⁶Yarnold PR, Feinglass J, Martin GJ, McCarthy WJ (1999). Comparing three pre-processing strategies for longitudinal data for individual patients: An example in functional outcomes research. *Evaluation and the Health Professions*, 22, 254-277. DOI: 10.1177/01632789922034301

⁵⁷Yarnold PR (2013). Surfing the *Index of Consumer Sentiment*: Identifying statistically significant monthly and yearly changes. *Optimal Data Analysis*, 2, 211-216. URL: <http://optimalprediction.com/files/pdf/V2A33.pdf>

⁵⁸<http://research.stlouisfed.org/fred2/series/UMCSENT/>

⁵⁹Yarnold PR (2013). The most recent, earliest, and Kth significant changes in an ordered series: Traveling backwards in time to assess when annual crude mortality rate most recently began increasing in McLean County, North Dakota. *Optimal Data Analysis*, 2, 143-147. URL: <http://optimalprediction.com/files/pdf/V2A21.pdf>

⁶⁰Yarnold PR (2013). UniODA and small samples. *Optimal Data Analysis*, 2, 71. URL: <http://optimalprediction.com/files/pdf/V2A13.pdf>

⁶¹Yarnold PR (1982). On comparing interscale difference scores within a profile. *Educational and Psychological Measurement*, 42, 1037-1045. DOI: 10.1177/001316448204200410

⁶²Yarnold PR (1988). Classical test theory methods for repeated-measures N=1 research designs. *Educational and Psychological Measurement*, 48, 913-919. DOI: 10.1177/0013164488484006

⁶³Mueser KT, Yarnold PR, Foy DW (1991). Statistical analysis for single-case designs: Evaluating outcomes of imaginal exposure treatment of chronic PTSD. *Behavior Modification*, 15, 134-155. DOI: 10.1177/01454455910152002

⁶⁴Yarnold PR. (1992). Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, S Tindale (Eds.), *Social Psychological Applications to Social Issues, Volume 2: Methodological Issues in Applied Social Research*. New York, NY: Plenum, pp. 177-197.

⁶⁵Yarnold PR, Soltysik RC (2013). Ipsative transformations are *essential* in the analysis of serial data. *Optimal Data Analysis*, 2, 94-97. URL: <http://optimalprediction.com/files/pdf/V2A17.pdf>

⁶⁶Lamiell JT (1981). Toward an ideothetic psychology of personality. *American Psychologist*, 36, 276-289. DOI: 10.1037/0003-066X.36.3.276

⁶⁷Rogers JH, Widiger TA (1989). Comparing ideothetic, ipsative, and normative indices of consistency. *Journal of Personality*, 57, 847-869. DOI: 10.1111/j.1467-6494.1989.tb00497.x

⁶⁸Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, 3, 19. URL: <http://optimalprediction.com/files/pdf/V3A7.pdf>

⁶⁹Yarnold PR (2010). Aggregated vs. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis*, 1, 46-49. URL: <http://optimalprediction.com/files/pdf/V1A8.pdf>

⁷⁰Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D. (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke*, 27, 1516-1520. Abstracted in *American College of Physicians Journal Club*, 1997, 126, 43. DOI: 10.1161/01.STR.27.9.1516

⁷¹Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: <http://optimalprediction.com/files/pdf/V2A27.pdf>

- ⁷¹Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 162,393-398. DOI: 10.1164/ajrccm.162.2.9909014
- ⁷²Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmeil JS, Sipler AM, Chan C, Goetz MB, Schwartz DN, Bennett CL (2000). A new pre-admission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine*, 161, 1081-1086. DOI: 10.1164/ajrccm.161.4.9906072
- ⁷³Efron B, Gong G (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37, 36-48. DOI: 10.1080/00031305.1983.10483087
- ⁷⁴Dunn OJ, Vardy PD (1966). Probabilities of correct classification in discriminant analysis. *Biometrics*, 22, 908-924. DOI: 10.2307/2528081
- ⁷⁵Frank RE, Massy WF, Morrison GD (1965). Bias in multiple discriminant analysis. *Journal of Marketing Research*, 2, 250-258. URL: <http://www.jstor.org/stable/3150183>
- ⁷⁶Fukunaga K, Kessell DL (1971). Estimation of classification error. *IEEE Transactions on Computers*, 20, 1521-1527. DOI: <http://dx.doi.org/10.1109/T-C.1971.223165>
- ⁷⁷Hills M (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society*, 28, 1-20. URL: <http://www.jstor.org/stable/2984268>
- ⁷⁸Lachenbruch PA (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23, 639-645. URL: <http://www.jstor.org/stable/2528418>
- ⁷⁹Lachenbruch PA (1975). *Discriminant analysis*. New York: Hafner.
- ⁸⁰Lachenbruch PA, Mickey MR (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11. DOI: 10.1080/00401706.1968.10490530
- ⁸¹Kshirsagar AM (1972). *Multivariate analysis*. New York: Dekker.
- ⁸²Geisser S (1975). The predictive sample Reuse method with applications. *Journal of the American Statistical Association*, 70, 320-328. DOI: 10.1080/01621459.1975.10479865
- ⁸³Stone M (1974). Cross-validatory choice and assessment of statistical problems. *Journal of the Royal Statistical Society*, 36, 111-147. URL: <http://www.jstor.org/stable/2984809>
- ⁸⁴Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest*, 123: 1151-1160. DOI: 10.1378/chest.123.4.1151
- ⁸⁵Yarnold PR (2013). Assessing hold-out validity of CTA models using UniODA. *Optimal Data Analysis*, 2, 31-36. URL: <http://optimalprediction.com/files/pdf/V2A5.pdf>
- ⁸⁶Simpson EH (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, B, 13, 238-241. URL: <http://www.jstor.org/stable/2984065>
- ⁸⁷Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>
- ⁸⁸Blyth CR (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67, 364-366. DOI: 10.1080/01621459.1972.10482387
- ⁸⁹Hintzman DL (1993). On variability, Simpson's paradox, and the relation between recognition and recall: Reply to Tulving and Flexser. *Psychological Review*, 100, 143-148. DOI: 10.1037/0033-295X.100.1.143

⁹⁰Preuss L, Vorkauf H (1997). The knowledge content of statistical data. *Psychometrika*, 62, 133-161. DOI: 10.1007/BF02294784

⁹¹Hintzman DL (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, 87, 398-410. DOI: 10.1037/0033-295X.87.4.398

⁹²Martin E (1981). Simpson's paradox revisited: A reply to Hintzman. *Psychological Review*, 88, 372-374.

⁹³Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete multivariate analysis*. Cambridge: University Press.

⁹⁴Flexser AJ, Tulving E (1993). Recognition-failure constraints and the average maximum. *Psychological Review*, 100, 149-153. DOI: 10.1037/0033-295X.100.1.149

⁹⁵Woolson RF (1987). *Statistical methods for the analysis of biomedical data*. New York: Wiley.

⁹⁶Dowdney L, Rogers C, Dunn G (1993). Influences upon attendance at out patient facilities—the contribution of linear-logistic modeling. *Psychological Medicine*, 23, 195-201. DOI: 10.1037/0033-295X.100.1.149

⁹⁷Hagenaars JA (1990). *Categorical longitudinal data*. Newbury Park, CA: Sage.

⁹⁸Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, 56, 430-442. DOI: 10.1177/0013164496056003005

⁹⁹McClish DK (1992). Combining and comparing area estimates across studies or strata. *Medical Decision Making*, 12, 274-279. DOI: 10.1177/0272989X9201200405

¹⁰⁰Midgette AS, Stukel TA, Littenberg B (1993). A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Medical Decision Making*, 13, 253-257. DOI: 10.1177/0272989X9301300313

Chapter 3

¹Nunnally JC (1978). *Psychometric theory* (2nd Ed.). New York: McGraw-Hill.

²Kazdin AE (1992). *Research design in clinical psychology* (2nd Ed.). Boston: Allyn & Bacon

³Thompson DA, Yarnold PR, Adams SL, Spacone AB (1996). How accurate are patient's waiting time perceptions? *Annals of Emergency Medicine*, 28, 652-656. DOI: 10.1016/S0196-0644(96)70089-6

⁴Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119. URL: <http://optimalprediction.com/files/pdf/V2A19.pdf>

⁵Yarnold PR (1987). Norms for the Glass model of the short student version of the Jenkins Activity Survey. *Social and Behavioral Science Documents*, 16, 60. MS# 2777.

⁶Yarnold PR, Lyons JS (1987). Norms for college undergraduates on the Bem Sex-Role Inventory and the Wiggins Interpersonal Behavior Circle. *Journal of Personality Assessment*, 51, 595-599. DOI: 10.1207/s15327752jpa5104_11

⁷Yarnold 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. DOI: 10.1207/s15327752jpa5203_2

⁸Collinge WC, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

⁹Collinge W, Yarnold PR, Soltysik, RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103%2F1947-2714.118920

¹⁰Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, 2, 37-42. URL: <http://optimalprediction.com/files/pdf/V2A6.pdf>

¹¹Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

¹²Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, 2, 172-176. URL: <http://optimalprediction.com/files/pdf/V2A26.pdf>

¹³Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: <http://optimalprediction.com/files/pdf/V2A27.pdf>

¹⁴Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books.

¹⁵Grimm LG, Yarnold PR (1995). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books.

¹⁶Yarnold PR (2015). An example of nonlinear UniODA. *Optimal Data Analysis*, 4, 124-128. URL: <http://optimalprediction.com/files/pdf/V4A24.pdf>

¹⁷<http://www.biostathandbook.com/kruskalwallis.html>

¹⁸Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, 3, 30-31. URL: <http://optimalprediction.com/files/pdf/V3A12.pdf>

¹⁹Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke*, 27, 1516-1520. Abstracted in *American College of Physicians Journal Club* (1997), 126, 43. DOI: 10.1161/01.STR.27.9.1516

²⁰Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, 56, 430-442. DOI: 10.1177/0013164496056003005

²¹<http://bmdshapi.com/>

²²Behavioral Measurement Database Services, BMDS, PO Box 110287, Pittsburgh, PA, USA 15232-0787; 1-412-687-6850; bmdshapi@aol.com; <http://bmdshapi.com/index.html>

²³Levenson T, Grammer LC, Yarnold PR, Patterson R (1997). Cost-effective management of malignant potentially fatal asthma. *Allergy and Asthma Proceedings*, 18, 73-78. DOI: 10.2500/108854197778605455

²⁴Yarnold PR (2013). Percent oil-based energy consumption and average percent GDP growth: a small sample UniODA analysis. *Optimal Data Analysis*, 2, 60-61. URL: <http://optimalprediction.com/files/pdf/V2A10.pdf>

²⁵Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, 22, 739-752. DOI: 10.1111/j.1540-5915.1991.tb00362

²⁶Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, 54, 646-653. DOI: 10.1177/0013164494054003007

²⁷Soltysik RC, Yarnold PR (2013). Statistical power of optimal discrimination with one attribute and two classes: One-tailed hypotheses. *Optimal Data Analysis*, 2, 26-30. URL: <http://optimalprediction.com/files/pdf/V2A4.pdf>

²⁸Odeh RE, Evans JO (1974). Algorithm AS 70: The percentage points of the normal distribution. *Applied Statistics*, 23: 96-97. DOI: 10.2307/2347061

²⁹Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

³⁰Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D. (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke*, 27, 1516-1520. Abstracted in *American College of Physicians Journal Club*, 1997, 126, 43. DOI: 10.1161/01.STR.27.9.1516

³¹<http://panko.shidler.hawaii.edu/SSR/>

³²Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software. *Optimal Data Analysis*, 2, 2-6. URL: <http://optimalprediction.com/files/pdf/V2A1.pdf>

³³Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression away from the mean. *Optimal Data Analysis*, 2, 19-25. URL: <http://optimalprediction.com/files/pdf/V2A3.pdf>

³⁴Bollen KA (1989). *Structural equations with latent variables*. New York: Wiley.

³⁵Kline, RB. (2011). *Principles and practice of structural equation modeling* (3rd ed). New York: Guilford Press.

³⁶Yarnold PR, Broftt GC (2013). Comparing knot strength with UniODA. *Optimal Data Analysis*, 2, 54-59. URL: <http://optimalprediction.com/files/pdf/V2A9.pdf>

³⁷Philpott L (2008). *The complete handbook of fishing knots, leaders, and lines*. New York: Skyhorse Publishing.

³⁸Yarnold PR, Soltysik RC (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences*, 22, 1158-1164. DOI: 10.1111/j.1540-5915.1991.tb01912.x

Chapter 4

¹Bowker AH (1948). Bowker's test for symmetry. *Journal of the American Statistical Association*, 43, 572–574. URL: <http://www.jstor.org/stable/2280710>

²Yarnold JK (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, 65, 864-886. URL: <http://www.jstor.org/stable/2284594>

³Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

⁴Agresti A (1990). *Categorical data analysis*. Hoboken, NJ, Wiley (pp. 356-357, 360-361).

⁵Yarnold PR (2015). UniODA vs. Bowker's test for symmetry: Diagnosis before vs. after treatment. *Optimal Data Analysis*, 4, 29-31. URL: <http://optimalprediction.com/files/pdf/V4A9.pdf>

⁶<http://www.econ.upf.edu/~michael/stanford/maeb5.pdf>

⁷Yarnold PR (2014). UniODA vs. Bray-Curtis dissimilarity index for count data. *Optimal Data Analysis*, 3, 115-116. URL: <http://optimalprediction.com/files/pdf/V3A28.pdf>

⁸Maxwell SE, Delaney HD (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

⁹Larichev OI, Olson DL, Moshkovich HM, Mechitov AJ (1995). Numerical vs cardinal measurements in multiattribute decision making: How exact is enough? *Organizational Behavior and Human Decision Processes*, 64, 9-21. DOI: 10.1006/obhd.1995.1085

¹⁰Smithson M (1987). *Fuzzy set analysis for behavioral and social sciences*. New York: Springer-Verlag.

¹¹Rubin PA (1992). *Mathematical programming and alternative classification models*. Invited address presented at the TIMS/ORSA Joint National Meetings, Orlando.

¹²Pearson K (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157-175. DOI: 10.1080/14786440009463897

¹³Cochran WG (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, 10, 417-451. DOI: 10.2307/3001616

¹⁴Mosteller F (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63, 1-28. DOI: 10.1080/01621459.1968.11009219

¹⁵Parshall CG, Kromrey JD (1996). Tests of independence in contingency tables with small samples: A comparison of statistical power. *Educational and Psychological Measurement*, 56, 26-44. DOI: 10.1177/0013164496056001002

¹⁶Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.

¹⁷Bishop YMM, Feinberg SE, Holland PW (1985). *Discrete multivariate analysis*. Cambridge: University Press.

¹⁸Goodman LA (1968). The analysis of cross-classified data: Independence, quasi-independence, and interaction in contingency tables with or without missing cells. *Journal of the American Statistical Association*, 63, 1091-1131. DOI: 10.1080/01621459.1968.10480916

¹⁹Hagenaars JA (1990). *Categorical longitudinal data*. Newbury Park, CA: Sage.

²⁰Yarnold PR (2010). UniODA vs. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis*, 1, 62-65. URL: <http://optimalprediction.com/files/pdf/V1A12.pdf>

²¹Yarnold PR (2014). UniODA vs. chi-square: Audience effect on smile production in infants. *Optimal Data Analysis*, 3, 3-5. URL: <http://optimalprediction.com/files/pdf/V3A1.pdf>

²²Jones SS, Collins K, Hong HW (1991). An audience effect on smile production in 10-month-old infants. *Psychological Science*, 2, 45-49. URL: <http://www.jstor.org/stable/40062583>

²³Yarnold PR, Broftt GC (2013). ODA range test vs. one-way analysis of variance: Comparing strength of alternative line connections. *Optimal Data Analysis*, 2, 198-201. URL: <http://optimalprediction.com/files/pdf/V2A30.pdf>

²⁴Yarnold PR (2013). ODA range test vs. one-way analysis of variance: Patient race and lab results. *Optimal Data Analysis*, 2, 206-210. URL: <http://optimalprediction.com/files/pdf/V2A32.pdf>

²⁵Cox MK, Key CH (1993). Post hoc pairwise comparisons for the chi-square test of homogeneity of proportions. *Educational and Psychological Measurement*, 53, 951-962. DOI: 10.1177/0013164493053004008

²⁶Seaman MA, Hill CC (1996). Pairwise comparisons for proportions: a note on Cox and Key. *Educational and Psychological Measurement*, 56, 452-459. DOI: 10.1177/0013164496056003007

²⁷Snyder DK, Wills RM, Grady-Fletcher A (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: a four-year follow up study. *Journal of Consulting and Clinical Psychology*, 59, 138-141. DOI: 10.1037/0022-006X.59.1.138

²⁸Hyde JS, Plant EA (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist*, 50, 159-161. DOI: 10.1037/0003-066X.50.3.159

²⁹Cochran WG (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-266. DOI: <http://www.jstor.org/stable/2332378>

³⁰Yarnold PR (2015). UniODA vs. Cochran's Q test: Comparing success of alternatives. *Optimal Data Analysis*, 4, 116-117. URL: <http://optimalprediction.com/files/pdf/V4A20.pdf>

³¹<https://www.medcalc.org/manual/cochranq.php>

³²<http://webword.com/moving/cochransq.html>

³³Yarnold PR (2015). UniODA vs. Cochran's Q test: Evaluating success rate in web usability testing. *Optimal Data Analysis*, 4, 118-119. URL: <http://optimalprediction.com/files/pdf/V4A21.pdf>

³⁴<http://psych.unl.edu/psycrs/handcomp/hccochran.PDF>

³⁵Yarnold PR (2015). UniODA vs. Cochran's Q test: Pet store reptile display behavior by holiday. *Optimal Data Analysis*, 4, 120-121. URL: <http://optimalprediction.com/files/pdf/V4A22.pdf>

³⁶Cohen (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46. DOI: 10.1177/001316446002000104

³⁷Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine*, 9, 1103-1115. DOI: 10.1002/sim.4780090917

³⁸Feingold M (1992). The equivalence of Cohen's kappa and Pearson's chi-square statistics in the 2 x 2 table. *Educational and Psychological Measurement*, 52, 57-61. DOI: 10.1177/001316449205200105

³⁹Davies M, Fleiss JL (1982). Measuring agreement for multinomial data. *Biometrics*, 38, 1047-1051. DOI: URL: <http://www.jstor.org/stable/2529886>

⁴⁰Fleiss JL (1986). *The design and analysis of clinical experiments*. New York: Wiley.

⁴¹Soeken KL, Prescott PA (1986). Issues in the use of Kappa to estimate reliability. *Medical Care*, 24, 733-741. URL: <http://www.jstor.org/stable/3765100>

⁴²Carmelli D, Dame A, Swan G, Rosenman R (1991). Long-term changes in Type A behavior: A 27-year follow-up of the Western Collaborative Group Study. *Journal of Behavioral Medicine*, 14, 593-606. DOI: 10.1007/BF00867173

⁴³Yarnold PR (2014). UniODA vs. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, 3, 14-16. URL: <http://optimalprediction.com/files/pdf/V3A5.pdf>

⁴⁴Kwoh CK, O'Connor GT, Regan-Smith MG, Olmstead EM, Brown LA, Burnett JB, Hochman RF, King K, Morgan GJ (1992). Concordance between clinician and patient assessment of physical and mental health status. *Journal of Rheumatology*, 19, 1031-1037.

⁴⁵Yarnold PR (2014). UniODA vs. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient's physical and mental health functioning. *Optimal Data Analysis*, 3, 12-13. URL: <http://optimalprediction.com/files/pdf/V3A4.pdf>

⁴⁶Everett JE (1990). Discrimination measure using contingency tables. *Multivariate Behavioral Research*, 25, 371-386. DOI: 10.1207/s15327906mbr2503_8

⁴⁷Wright RE (2005). Logistic Regression. In: LG Grimm, PR Yarnold (eds.), *Reading and understanding multivariate statistics*. Washington, DC: APA Books (pp. 217-244).

⁴⁸Rosengren A, Welin L, Tsipogianni A, Wilhelmsen L (1989). Impact of cardiovascular risk factors on coronary heart disease and mortality among middle aged diabetic men: A general population study. *British Medical Journal*, 299, 1127-1131. DOI: 10.1136/bmj.299.6708.1127

- ⁴⁹Yarnold PR (2014). UniODA vs. logistic regression analysis: Serum cholesterol and coronary heart disease and mortality among middle aged diabetic men. *Optimal Data Analysis*, 3, 17-18. URL: <http://optimalprediction.com/files/pdf/V3A6.pdf>
- ⁵⁰Beck JR, Pauker SG (1983). The Markov process in medical prognosis. *Medical Decision Making*, 3, 419-458. DOI: 10.1177/0272989X8300300403
- ⁵¹Billingsley P (1961). *Statistical inference for Markov processes*. Chicago: University of Chicago Press.
- ⁵²Coombs CH, Dawes RM, Tversky A (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.
- ⁵³Disney RL (1971). Probability and stochastic processes. In: H.B. Maynard (Ed.), *Industrial engineering handbook*. New York: McGraw-Hill (ps. 10.32-10.51).
- ⁵⁴Hanneman RA (1988). *Computer-assisted theory building: Modeling dynamic social systems*. Newbury Park, CA: Sage.
- ⁵⁵Kemeny JG, Snell JL (1976). *Finite Markov chains*. New York: Springer.
- ⁵⁶Kruskal JB (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25, 201-237. DOI: 10.1137/1025045
- ⁵⁷Parzen E (1962). *Stochastic processes*. San Francisco: Holden-Day.
- ⁵⁸Rau JG (1970). *Optimization and probability in systems engineering*. New York: Van Nostrand.
- ⁵⁹Raush HL (1965). Interaction sequences. *Journal of Personality and Social Psychology*, 2, 487-499. DOI: 10.1037/h0022478
- ⁶⁰Driese SG, Dott RH (1984). Model for sandstone-carbonate "cyclothem" based on Upper Member of Morgan Formation (Middle Pennsylvanian) of northern Utah and Colorado. *The American Association of Petroleum Geologists Bulletin*, 68, 574-597.
- ⁶¹McNemar Q (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157. DOI: 10.1007/BF02295996
- ⁶²Fagerland MW, Lydersen S, Laake P (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13: 91. DOI: 10.1186/1471-2288-13-91
- ⁶³http://en.wikipedia.org/wiki/McNemar%27s_test
- ⁶⁴Yarnold PR (2015). UniODA vs. McNemar's test for correlated proportions: Diagnosis of disease before vs. after treatment. *Optimal Data Analysis*, 4, 24-26. URL: <http://optimalprediction.com/files/pdf/V4A7.pdf>
- ⁶⁵Haberman SJ (1979). *Analysis of qualitative data, Volume 2: New developments*. New York: Academic Press.
- ⁶⁶Bakeman R, Quera V (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, 118, 272-284. DOI: 10.1037/0033-2909.118.2.272
- ⁶⁷Gilbert N (1993). *Analyzing tabular data: Log linear and logistic models for social researchers*. London: University of London College Press.
- ⁶⁸Loo R (1996). Construct validity and classification stability of the revised Learning Style Inventory (LSI-1985). *Educational and Psychological Measurement*, 56, 529-536. DOI: 10.1177/0013164496056003015
- ⁶⁹Kolb DA (1984). *Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.
- ⁷⁰Foa U (1971). Interpersonal and economic resources. *Science*, 171, 345-351. DOI: 10.1126/science.171.3969.345

⁷¹Goodman LA (1962). Statistical methods for analyzing processes of change. *American Journal of Sociology*, 68, 57-78. URL: <http://www.jstor.org/stable/2774180>

Chapter 5

¹Kendall MG, Babington SB (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10, 275–287. URL: <http://www.jstor.org/stable/2235668>

²Legendre P (2005). Species associations: The Kendall Coefficient of Concordance revisited. *Journal of Agricultural, Biological and Environmental Statistics*, 10, 226–245. DOI: 10.1198/108571105X46642

³<http://www.real-statistics.com/reliability/kendalls-w/>

⁴Yarnold JK (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, 65, 864-886. URL: <http://www.jstor.org/stable/2284594>

⁵Yarnold PR (2014). UniODA vs. Kendall's Coefficient of Concordance (W): Multiple rankings of multiple movies. *Optimal Data Analysis*, 3, 121-123. URL: <http://optimalprediction.com/files/pdf/V3A30.pdf>

⁶Kruskal W, Wallis WA (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583–621. DOI: 10.1080/01621459.1952.10483441

⁷Conover WJ (1999). *Practical nonparametric statistics* (3rd Ed.). Hoboken, NJ, Wiley.

⁸<http://www.biostathandbook.com/kruskalwallis.html>

⁹Yarnold PR (2015). UniODA vs. Kruskal-Wallace test: Farming method and corn yield. *Optimal Data Analysis*, 4, 113-115. URL: <http://optimalprediction.com/files/pdf/V4A19.pdf>

¹⁰<http://www.biostathandbook.com/kruskalwallis.html>

¹¹Yarnold PR (2015). UniODA vs. Kruskal-Wallace test: Gender and dominance of free-ranging domestic dogs in the outskirts of Rome. *Optimal Data Analysis*, 4, 122-123. URL: <http://optimalprediction.com/files/pdf/V4A23.pdf>

¹²Mann HB, Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60. URL: <http://www.jstor.org/stable/2236101>

¹³Newcombe RG (2005). Confidence intervals for an effect size measure based on the Mann-Whitney statistic, Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25, 543-557. DOI: 10.1002/sim.2323

¹⁴<http://www.saburchill.com/IBbiology/stats/002.html>

¹⁵Yarnold PR (2014). UniODA vs. Mann-Whitney U test: Sunlight and petal width. *Optimal Data Analysis*, 4, 3-5. URL: <http://optimalprediction.com/files/pdf/V4A1.pdf>

¹⁶<http://users.sussex.ac.uk/~grahamh/RM1web/MannWhitneyHandout%202011.pdf>

¹⁷Yarnold, P.R. (2014). UniODA vs. Mann-Whitney U test: Comparative effectiveness of laxatives. *Optimal Data Analysis*, 4, 6-8. URL: <http://optimalprediction.com/files/pdf/V4A2.pdf>

¹⁸Yarnold PR, Broffet GC (2013). ODA range test vs. one-way analysis of variance: Comparing strength of alternative line connections. *Optimal Data Analysis*, 2, 198-201. URL: <http://optimalprediction.com/files/pdf/V2A30.pdf>

¹⁹Yarnold PR, Broffet GC (2013). Comparing knot strength using UniODA. *Optimal Data Analysis*, 2, 54-59. URL: <http://optimalprediction.com/files/pdf/V2A9.pdf>

²⁰Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, DC: APA Books.

- ²¹Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, 16, 1451-1463. DOI: 10.1002/(SICI)1097-0258(19970715)16:13<1451::AID-SIM571>3.0.CO;2-F
- ²²Yarnold PR (2013). ODA range test vs. one-way analysis of variance: Patient race and lab results. *Optimal Data Analysis*, 2, 206-210. URL: <http://optimalprediction.com/files/pdf/V2A32.pdf>
- ²³Olsson U (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460. DOI: 10.1007/BF02296207 .
- ²⁴Tallis GM (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18, 342-353. URL: <http://www.jstor.org/stable/2527476>
- ²⁵Uebersax JS (2006). The tetrachoric and polychoric correlation coefficients. *Statistical Methods for Rater Agreement* web site: <http://john-uebersax.com/stat/tetra.htm>.
- ²⁶Yarnold PR (2014). UniODA vs. polychoric correlation: Number of lambs born over two years. *Optimal Data Analysis*, 3, 113-114. URL: <http://optimalprediction.com/files/pdf/V3A27.pdf>
- ²⁷Allen MJ, Yen WM (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- ²⁸Brown FG (1983). *Principles of educational and psychological testing* (3rd Ed.). New York: Holt.
- ²⁹Carmines EG, Zeller RA (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- ³⁰Cromack TR (1989). Measurement considerations in clinical research. In: C.B. Royeen (Ed.), *Clinical research handbook: An analysis for the service professions*. Thorofare, NJ: Slack (ps. 47-69).
- ³¹Ebel RL (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- ³²Ghiselli EE (1964). *Theory of psychological measurement*. New York: McGraw-Hill.
- ³³Gulliksen H (1950). *Theory of mental tests*. New York: Wiley.
- ³⁴Lord FM, Novick MR (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- ³⁵Magnusson D (1967). *Test theory*. Reading, MA: Addison-Wesley.
- ³⁶Nunnally JC (1978). *Psychometric theory* (2nd Ed.). New York: McGraw-Hill.
- ³⁷Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.
- ³⁸Yarnold PR (1984). The reliability of a profile. *Educational and Psychological Measurement*, 44, 49-59. DOI: 10.1177/0013164484441005
- ³⁹Cronbach LJ, Rajaratnam N, Gleser GC (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 15, 137-163. DOI: 10.1111/j.2044-8317.1963.tb00206.x
- ⁴⁰Kuder GF, Richardson MW (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160. DOI: 10.1007/BF02288391
- ⁴¹Yarnold PR (1988). Classical test theory methods for repeated-measures $N=1$ research designs. *Educational and Psychological Measurement*, 48, 913-919. DOI: 10.1177/001316448484006
- ⁴²Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete multivariate analysis*. Cambridge: University Press.

⁴³Fleiss JL (1986). *The design and analysis of clinical experiments*. New York: Wiley.

⁴⁴Woolson RF (1987). *Statistical methods for the analysis of biomedical data*. New York: Wiley.

⁴⁵Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.

⁴⁶Mueser KT, Sayers SL, Schooler NR, Mance RM, Haas GL (1993). A multisite investigation of the reliability of the *Scale for the Assessment of Negative Symptoms*. *The American Journal of Psychiatry*, 151, 1453-1462. DOI: 10.1176/ajp.151.10.1453

⁴⁷Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Gleister R, Tanabe P (2004). Assessment of inter-observer reliability of two five-level triage and acuity scales: A randomized controlled trial. *Canadian Journal of Emergency Medicine*, 6, 240-245. DOI: 10.1017/S1481803500009192

⁴⁸Yarnold PR (2014). How to assess inter-observer reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, 3, 42-49. URL: <http://optimalprediction.com/files/pdf/V3A15.pdf>

⁴⁹Royeen CB (1989). *Clinical research handbook: An analysis for the service professions*. Thorofare, NJ: SLACK.

⁵⁰Zegers FE (1991). Coefficients for interrater agreement. *Applied Psychological Measurement*, 15, 321-333. DOI: 10.1177/014662169101500401

⁵¹Yarnold PR (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, 3, 50-54. URL: <http://optimalprediction.com/files/pdf/V3A16.pdf>

⁵²Mirhaghi A, Heydari A, Mazlom R, Hasanzadeh F (2015). Reliability of the Emergency Severity Index: Meta-Analysis. *Sultan Qaboos University Medical Journal*, 15, e71-77.

⁵³Yarnold PR (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis*, 4, 21-23. URL: <http://optimalprediction.com/files/pdf/V4A6.pdf>

⁵⁴Brown W (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322. DOI: 10.1111/j.2044-8295.1910.tb00207.x

⁵⁵Spearman C (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295. DOI: 10.1111/j.2044-8295.1910.tb00206.x

⁵⁶Guttman L (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282. DOI: 10.1007/BF02288892

⁵⁷Lyerly R (1958). The Kuder-Richardson formula 21 as a split-half coefficient, and some remarks on its basic assumption. *Psychometrika*, 23, 267-270. DOI: 10.1007/BF02289239

⁵⁸Rulon PJ (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Education Review*, 9, 99-103.

⁵⁹Yarnold PR (1994). Comparing the split-half reliability of androgyny and sex-typing measures. *Australian Journal of Psychology*, 46, 164-169. DOI: 10.1080/00049539408259491

⁶⁰Guyatt G, Walter S, Norman G (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, 40, 171-178. DOI: 10.1016/0021-9681(87)90069-5

⁶¹Shea JA, Norcini JJ, Baranowski RA, Langdon LO, Poop RL (1992). A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. *Evaluation in the Health Professions*, 15, 325-340. DOI: 10.1177/016327879201500305

⁶²Kahneman D, Slovic P, Tversky A (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge: University Press.

⁶³Saal FE, Downey RG, Lahey MA(1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428. DOI: 10.1037/0033-2909.88.2.413

⁶⁴Royeen CB (1989). *Clinical research handbook: An analysis for the service professions*. Thorofare, NJ: SLACK.

⁶⁵Bryant FB, Yarnold PR (1990). The impact of Type A behavior on subjective life quality: Bad for the heart, good for the soul? *Journal of Social Behavior and Personality, 5*, 369-404.

⁶⁶Raaflaub KA, Talbert RJA (2009). *Geography and ethnography: Perceptions of the world in pre-modern societies*. New York: Wiley.

⁶⁷nces.ed.gov/pubs2009/2009081.pdf

⁶⁸Colizza V, Vespignani A, Hardy EF (2007). *Impact of air travel on global spread of infectious diseases*. Indiana University.

⁶⁹Charlez PA (1997). *Rock mechanics: petroleum applications*. Paris: Editions Technip.

⁷⁰Colborn T, Schultz K, Herrick L, Kwiatkowski C (2014). An exploratory study of air quality near natural gas operations. *Human and Ecological Risk Assessment: An International Journal, 20*, 86-105. DOI: 10.1080/10807039.2012.749447

⁷¹Bamberger M, Oswald RE (2012). Impacts of gas drilling on human and animal health. *New Solutions: A Journal of Environmental and Occupational Health Policy, 22*, 51-77. DOI: 10.2190/NS.22.1.e

⁷²*Chemicals Used in Hydrolic Fracturing* (April 18, 2011). Committee on Energy and Commerce, US House of Representatives.

⁷³Colborn T, Kwiatkowski C, Schultz K, Bach-ran M (2011). Natural gas operations from public health perspective. *Human and Ecological Risk Assessment: An International Journal, 17*, 1039-1056. DOI: 10.1080/10807039.2011.605662

⁷⁴<http://chevrontoxicco.com/about/environmental-impacts/produced-water>

⁷⁵<http://www.netl.doe.gov/technologies/pwmis/intropw/>

⁷⁶<http://rt.com/op-edge/fracking-radioactive-uranium-danger-ecology-057/>

⁷⁷Clark CE, Veil JA (2009). *Produced Water Volumes and Management Practices in the United States*, ANL/EVS/R-09/1, Environmental Science Division, Argonne National Laboratory.

⁷⁸Jacquet J (2009). *Energy boomtowns and natural gas: Implications for Marcellus Shale local governments and rural communities*. NERC RD Rural Development Paper No. 43, Pennsylvania State University.

⁷⁹*Injuries, illnesses, and fatalities in the coal mining industry* (2010). US Bureau of Labor Statistics.

⁸⁰<http://ndhealth.gov/vital/stats.htm>

⁸¹Yarnold PR (2013). Statistically significant increases in crude mortality rate of North Dakota counties occurring after massive environmental usage of toxic chemicals and biocides began there in1998: An optimal static statistical map. *Optimal Data Analysis, 2*, 98-105. URL: <http://optimalprediction.com/files/pdf/V2A18.pdf>

⁸²<http://data.worldbank.org/indicator/SP.DYN.CDRT.IN>

⁸³<http://www.npwrc.usgs.gov/resource/habitat/climate/wind.htm>

⁸⁴http://www.aoa.gov/AoARoot/AoA_Programs/HPW/Behavioral/docs2/North%20Dakota.pdf

⁸⁵Yarnold PR (2013). The most recent, earliest, and Kth significant changes in an ordered series: Traveling backwards in time to assess when annual crude mortality rate most recently began increasing in McLean County, North Dakota. *Optimal Data Analysis*, 2, 143-147. URL: <http://optimalprediction.com/files/pdf/V2A21.pdf>

⁸⁶Hanley JA, McNeil BJ (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36. DOI: 10.1148/radiology.143.1.7063747

⁸⁷<http://www.medicalbiostatistics.com/roccurve.pdf>

⁸⁸Yarnold PR (2014). UniODA vs. ROC analysis: Computing the “optimal” cut-point. *Optimal Data Analysis*, 3, 117-120. URL: <http://optimalprediction.com/files/pdf/V3A29.pdf>

⁸⁹Triantaphyllou E, Shu B, Sanchez N, Ray T (1998). Multi-criteria decision making: An operations research approach. *Encyclopedia of Electrical and Electronics Engineering*, 15, 175-186.

⁹⁰Appleton DR (1995). Pitfalls in the interpretation of studies: III. *Journal of the Royal Society of Medicine*, 88, 241-243. DOI: 10.1177/014107689508800501

⁹¹Yarnold PR (2014). UniODA vs. t-Test: Comparing two migraine treatments. *Optimal Data Analysis*, 3, 6-8. URL: <http://optimalprediction.com/files/pdf/V3A2.pdf>

⁹²Martin GJ, Magid NM, Myers G, Barnett PS, Schaad JW, Weiss JS, Lesch M, Singer DH (1987). Heart rate variability and sudden cardiac death secondary to coronary artery disease during ambulatory electrocardiographic monitoring. *American Journal of Cardiology*, 60, 86-89. 10.1016/0002-9149(87)90990-8

⁹³Collinge W, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

⁹⁴Collinge W, Yarnold PR, Soltysik RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103%2F1947-2714.118920

⁹⁵Raatikka VP, Rytkonen M, Nayha S, Hassi J (2007). Prevalence of cold-related complaints, symptoms and injuries in the general population: The INRISK 2002 cold substudy. *International Journal of Biometeorology*, 51, 441-448. DOI: 10.1007/s00484-006-0076-1

⁹⁶Yunus MB, Holt GS, Masi AT, Aldag JC (1988). Fibromyalgia syndrome among the elderly: Comparison with younger patients. *Journal of the American Geriatric Society*, 36, 987-995. DOI: 10.1111/j.1532-5415.1988.tb04364.x

⁹⁷Hagglund KJ, Deuser WE, Buckelew SP, Hewett J, Kay DR (1994). Weather, beliefs about weather, and disease severity among patients with fibromyalgia. *Arthritis Care Research*, 7, 130-135. DOI: 10.1002/art.1790070306

⁹⁸Bennett RM, Jones J, Turk DC, Russell IJ, Matallana L (2007). An internet survey of 2,596 people with fibromyalgia. *BMC Musculoskeletal Disorders*, 8, 27. DOI: 10.1186/1471-2474-8-27

⁹⁹Strusberg I, Mendelberg RC, Serra HA, Strusberg AM (2002). Influence of weather conditions on rheumatic pain. *Journal of Rheumatology*, 29, 335-338.

¹⁰⁰Miranda LC, Parente M, Silva C, Clemente-Coelho P, Santos H, Cortes S, Medeiros D, Ribeiro JS, Barcelos F, Sousa M, Miguel C, Figueiredo R, Mediavilla M, Simoes E, Silva M, Patto JV, Madeira H, Ferreira J, Micaelo M, Leitao R, Las V, Faustino A, Teixeira A (2007). Perceived pain and weather changes in rheumatic patients. *Acta Reumatologica Portuguesa*, 32, 351-361.

¹⁰¹Martinez JE, Cruz CG, Aranda C, Boulos FC, Lagoa LA (2003). Disease perceptions of Brazilian fibromyalgia patients: do they resemble perceptions from other countries? *International Journal of Rehabilitation Research*, 26, 223-227. DOI: 10.1097/00004356-200309000-00010

- ¹⁰²Guedj D, Weinberger A (1990). Effect of weather conditions on rheumatic patients. *Annals of the Rheumatic Diseases*, 49, 158-159. DOI: 10.1136/ard.49.3.15
- ¹⁰³Hawley DJ, Wolfe F, Lue FA, Moldofsky H (2001). Seasonal symptom severity in patients with rheumatic diseases: a study of 1,424 patients. *Journal of Rheumatology*, 28, 1900-1909.
- ¹⁰⁴Fors EA, Sexton H (2002). Weather and the pain in fibromyalgia: are they related? *Annals of the Rheumatic Diseases*, 61, 247-250. DOI: 10.1136/ard.61.3.247
- ¹⁰⁵Loza E, Abasolo L, Jover JA, Carmona L (2008). Burden of disease across chronic diseases: A health survey that measured prevalence, function, and quality of life. *The Journal of Rheumatology*, 35, 159-165.
- ¹⁰⁶http://climate.ncsu.edu/images/climate/enso/geo_heights.php
- ¹⁰⁷Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, 2, 37-42. URL: <http://optimalprediction.com/files/pdf/V2A6.pdf>
- ¹⁰⁸Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>
- ¹⁰⁹Bryant FB (2000). Assessing the validity of measurement. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding more multivariate statistics*. Washington DC: APA Books (pp. 99-146).
- ¹¹⁰Cook TD, Campbell DT (1978). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- ¹¹¹Cronbach LJ, Meehl PE (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302. DOI: 10.1037/h0040957
- ¹¹²Yarnold 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. DOI: 10.1207/s15327752jpa5203_2
- ¹¹³Jenkins CD, Zyzanski SJ, Ryan TJ, Flessas A, Tannenbaum SI (1977). Social insecurity and coronary-prone Type A responses as identifiers of severe atherosclerosis. *Journal of Consulting and Clinical Psychology*, 45, 1060-1067. DOI: 10.1037/0022-006X.45.6.1060
- ¹¹⁴Yarnold PR, Soltysik RC (2014). Emergency Severity Index (Version 3) score predicts hospital admission. *Optimal Data Analysis*, 3, 20-22. URL: <http://optimalprediction.com/files/pdf/V3A8.pdf>
- ¹¹⁵Campbell DT, Fiske DW (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105. URL: <http://psycnet.apa.org/doi/10.1037/h0046016>
- ¹¹⁶Friedman M, Rosenman RH (1974). *Type A behavior and your heart*. New York: Knopf.
- ¹¹⁷Glass DC (1977). *Behavior patterns, stress, and coronary disease*. Hillsdale, NJ: Erlbaum.
- ¹¹⁸Bryant FB, Yarnold PR, Morgan L (1991). Type A behavior and reminiscence in college undergraduates. *Journal of Research in Personality*, 25, 418-433. DOI: 10.1016/0092-6566(91)90031-K
- ¹¹⁹Smith JL, Bryant FB (2013). Are we having fun yet? Savoring, Type A behavior, and vacation enjoyment. *International Journal of Well-Being*, 3, 1-19. DOI: 10.1145/10.5502/ijw.v3i1.1
- ¹²⁰Bryant FB, Yarnold PR (2014). Type A behavior and savoring among college undergraduates: Enjoy achievements now—not later. *Optimal Data Analysis*, 3, 25-27. URL: <http://optimalprediction.com/files/pdf/V3A10.pdf>
- ¹²¹Yarnold PR, Mueser KT (1988). Student version of the Jenkins Activity Survey. In: M Hersen & AS Bellack (Eds.), *Dictionary of Behavioral Assessment Techniques*. Beverly Hills, CA: Pergamon, pp. 454-455.

¹²²Yarnold PR, Bryant FB (1988). Seven transliterations of the short version of the student Jenkins Activity Survey. *Social and Behavioral Science Documents*, 18, 18-19. MS# 2854.

¹²³Yarnold PR, Bryant FB, Grimm LG (1987). Comparing the short and long versions of the student Jenkins Activity Survey. *Journal of Behavioral Medicine*, 10, 75-90.

¹²⁴Bryant FB, Yarnold PR (1989). A measurement model for the short form of the student Jenkins Activity Survey. *Journal of Personality Assessment*, 53, 188-191. DOI: 10.1207/s15327752jpa5301_21

¹²⁵Yarnold PR, Bryant FB, Litsas F (1989). Type A behavior and psychological androgyny among Greek college students. *European Journal of Personality*, 3, 249-268. DOI: 10.1002/per.2410030403

¹²⁶Bryant FB, Yarnold PR (1995). Comparing five alternative factor-models of the Student Jenkins Activity Survey: Separating the wheat from the chaff. *Journal of Personality Assessment*, 64, 145-158. DOI: 10.1207/s15327752jpa6401_10

¹²⁷Yarnold PR, Mueser KT, Grau BW, Grimm LG (1986). The reliability of the student version of the Jenkins Activity Survey. *Journal of Behavioral Medicine*, 9, 401-414. DOI: 10.1007/BF00845123

¹²⁸Yarnold PR, Mueser KT (1989). Meta-analysis of the reliability of Type A behavior measures. *British Journal of Medical Psychology*, 62, 43-50. DOI: 10.1111/j.2044-8341.1989.tb02809.x

¹²⁹Yarnold PR (1987). Norms for the Glass model of the SJAS. *Social and Behavioral Sciences Documents*, 16, 60-65.

¹³⁰Bryant FB (2003). Savoring Beliefs Inventory (SBI): A scale for measuring beliefs about savoring. *Journal of Mental Health*, 12, 175-196. DOI: 10.1080/0963823031000103489

¹³¹Bryant FB, Veroff J (2007). *Savoring: A new model of positive experience*. Mahwah, NJ: Erlbaum.

¹³²Byant 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. URL: <http://optimalprediction.com/files/pdf/V3A14.pdf>

¹³³Bryant FB, Yarnold PR (2014). Type A Behavior and savoring among college undergraduates: Enjoy achievements now—not later. *Optimal Data Analysis*, 3, 25-27. URL: <http://optimalprediction.com/files/pdf/V3A10.pdf>

Chapter 6

¹LG Grimm, PR Yarnold (1995). *Reading and understanding multivariate statistics*. Washington DC: APA Books.

²LG Grimm, PR Yarnold (2000). *Reading and understanding more multivariate statistics*. Washington DC: APA Books.

³Licht MH (1995). Multiple regression and correlation. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 19-64).

⁴<http://people.duke.edu/~rnau/testing.htm>

⁵Kendall M (1975). *Multivariate Analyses*. New York: Hafner (Chapter 7).

⁶Bradley JV (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

⁷Toh RS, Hu MY (2008). Averaging to minimize or eliminate regression toward the mean to measure pure experimental effects. *Psychological Reports*, 102, 665-677. DOI: 10.2466/pr0.102.3.665-677

⁸Yarnold PR, Bryant FB, Solysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression away from the mean. *Optimal Data Analysis*, 2, 19-25. URL: <http://optimalprediction.com/files/pdf/V2A3.pdf>

⁹Yarnold PR, Solysik RC (1991). Refining two-group multivariable models using univariate optimal discriminant analysis. *Decision Sciences*, 22, 1158-1164. DOI: 10.1111/j.1540-5915.1991.tb01912.x

- ¹⁰Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, 54, 73-85. DOI: 10.1177/0013164494054001007
- ¹¹Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington DC: APA Books.
- ¹²Nanna MJ, Sawilowsky SS (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, 3, 55-67. DOI: 10.1037/1082-989X.3.1.55
- ¹³Bryant FB, Veroff J (2007). *Savoring: a new model of positive experience*. Mahwah, NY: Erlbaum.
- ¹⁴Scheier MF, Carver CS (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247. DOI: 10.1037/0278-6133.4.3.219
- ¹⁵Bryant FB (2003). Savoring Beliefs Inventory (SBI): A scale for measuring beliefs about savoring. *Journal of Mental Health*, 12, 175-196. DOI: 10.1080/0963823031000103489
- ¹⁶Rosenberg M (1965). *Society and the adolescent self-image*. Princeton NJ, Princeton University Press.
- ¹⁷Bryant FB, Yarnold PR, Grimm LG (1996). Toward a measurement model of the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality*, 30, 223-247. DOI: <http://dx.doi.org/10.1006/jrpe.1996.0015>
- ¹⁸Yarnold PR (2013). Maximum-accuracy multiple regression analysis: Influence of registration on overall satisfaction ratings of Emergency Room patients. *Optimal Data Analysis*, 2, 72-74. URL: <http://optimalprediction.com/files/pdf/V2A14.pdf>
- ¹⁹Yarnold PR (2013). Assessing technician, nurse, and doctor ratings as predictors of overall satisfaction ratings of Emergency Room patients: A maximum-accuracy multiple regression analysis. *Optimal Data Analysis*, 2, 76-85. URL: <http://optimalprediction.com/files/pdf/V2A15.pdf>
- ²⁰Yarnold PR (2013). Creating a data set with SAS™ and maximizing ESS of a multiple regression analysis model for a Likert-type dependent variable using UniODA™ and MegaODA™ software. *Optimal Data Analysis*, 2, 191-193. URL: <http://optimalprediction.com/files/pdf/V2A28.pdf>
- ²¹Yarnold PR (2015). Maximizing ESS of regression models in applications with dependent measures with domains exceeding ten values. *Optimal Data Analysis*, 4, 12-13. URL: <http://optimalprediction.com/files/pdf/V4A4.pdf>
- ²²Fisher RA (1950). *Statistical methods for research workers* (11th Ed., revised). London, UK: Oliver and Boyd. Ltd.
- ²³Pedhazur EJ (1982). *Multiple regression in behavioral research* (2nd Ed.). New York, NY: Holt, Rinehart and Winston.
- ²⁴Kleinbaum DG, Kupper LL, Muller KE (1988). *Applied regression analysis and other multivariable methods* (2nd Ed.). Boston, MA: PWS-Kent.
- ²⁵Silva APD, Stam A (1995). Discriminant analysis. In LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 277-318).
- ²⁶Stevens J (1992). *Applied multivariate statistics for the social sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.
- ²⁷Green PE (1978). *Analyzing multivariate data*. Hillsdale, IL: Dryden.
- ²⁸Finn MA, Stalans LJ (1996). Police referrals to shelter and mental health treatment: Examining their decisions in domestic assault cases. *Crime and Delinquency*, 41, 467-480. DOI: 10.1177/0011128795041004006
- ²⁹Stalans LS, Finn MA (1995). How novice and experienced officers interpret wife assaults: Normative and efficiency frames. *Law and Society Review*, 29, 301-335. DOI: 10.2307/3054013

³⁰Weinfurt KP, Bryant FB, Yarnold PR (1994). The factor structure of the Affect Intensity Measure: In search of a measurement model. *Journal of Research in Personality*, 28, 314-331. DOI: 10.1006/jrpe.1994.1023

³¹Yarnold PR, Bryant FB (1994). A measurement model for the Type A Self-Rating Inventory. *Journal of Personality Assessment*, 62, 102-115. DOI: 10.1207/s15327752jpa6201_10

³²Yarnold PR, Martin GJ, Soltysik RC, Nightingale SD (1993). Androgyny predicts empathy for trainees in medicine. *Perceptual and Motor Skills*, 77, 576-578. DOI: 10.2466/pms.1993.77.2.576

Chapter 7

¹LG Grimm, PR Yarnold (1995). *Reading and understanding multivariate statistics*. Washington DC: APA Books.

²LG Grimm, PR Yarnold (2000). *Reading and understanding more multivariate statistics*. Washington DC: APA Books.

³Stevens J (1992). *Applied multivariate analysis for the social sciences* (2nd Ed.). Hillsdale NJ: Erlbaum.

⁴Aldrich JH, Cnudde, C (1975). Probing the bounds of conventional wisdom: Comparison of regression, probit, and discriminant analysis. *American Journal of Political Science*, 19, 571-608. URL: <http://www.jstor.org/stable/2110547>

⁵Yarnold BM (1990). Federal court outcomes in asylum-related appeals 1980-1987: A highly politicized process. *Policy Sciences*, 23, 291-306. DOI: 10.1007/BF00141323

⁶Yarnold BM (1990). *Refugees without refuge: Formation and failed implementation of U.S. political asylum policy in the 1980s*. Lanham, MD: University Press of America.

⁷Yarnold BM (1990). The Refugee Act of 1980 and de-politicization of refugee/asylum admissions: Failed policy implementation. *American Politics Quarterly*, 18, 527-536. DOI: 10.1177/1532673X9001800408

⁸Yarnold BM (1991). *International fugitives: A new role for the International Court of Justice*. New York, NY: Praeger.

⁹Yarnold BM (1992). *Politics and the courts: Toward a general theory of public law*. New York, NY: Praeger.

¹⁰Yarnold BM (1993). *Abortion politics in the federal courts: Right versus right*. New York, NY: Paragon.

¹¹Hagle T, Mitchell G (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36, 762-784. URL: <http://www.jstor.org/stable/2111590>

¹²Yarnold BM, Yarnold PR (2010). Maximizing the accuracy of probit models via UniODA. *Optimal Data Analysis*, 1, 41-42. URL: <http://optimalprediction.com/files/pdf/V1A6.pdf>

¹³Wright RE (1995). Logistic regression. In: LG Grimm, PR Yarnold, *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 217-244).

¹⁴Friedman GD (1987). *Primer of epidemiology* (3rd Ed.). New York, NY: McGraw-Hill.

¹⁵Gilbert N (1993). *Analyzing tabular data: Log-linear and logistic models for social researchers*. London, UK: UCL Press.

¹⁶Agresti A (1990). *Categorical data analysis*. New York, NY: Wiley.

¹⁷Aldrich JH, Nelson FD (1984). *Linear probability, logit, and probit models*. Beverly Hills, CA: Sage.

¹⁸Hosmer DW, Lemeshow S (1989). *Applied logistic regression*. New York, NY: Wiley.

¹⁹Yarnold PR, Soltysik RC (1991). Refining two-group multivariable models using univariate optimal discriminant analysis. *Decision Sciences*, 22, 1158-1164. DOI: 10.1111/j.1540-5915.1991.tb01912.x

²⁰Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, 54, 73-85. DOI: 10.1177/0013164494054001007

²¹Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington DC: APA Books.

²²Bryant FB (2010). How to save the binary class variable and predicted probability of group membership from logistic regression analysis to an ASCII space-delimited file in *SPSS 17 For Windows. Optimal Data Analysis*, 1, 161-162. URL: <http://optimalprediction.com/files/pdf/V1A17.pdf>

²³Yarnold PR, Bryant FB (2013). Analysis involving categorical attributes having many response categories. *Optimal Data Analysis*, 2, 69-70. URL: <http://optimalprediction.com/files/pdf/V2A12.pdf>

²⁴Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete multivariate analysis*. Cambridge: University Press

²⁵Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.

²⁶Hagenaars JA (1990). *Categorical longitudinal data*. Newbury Park, CA: Sage.

²⁷Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: <http://optimalprediction.com/files/pdf/V2A27.pdf>

²⁸Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmiel J, Sipler A, Chan C, Goetz MB, Schwartz D, Bennett CL (2000). A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine*, 161, 1081-1086.

²⁹Yarnold JK (1970). The minimum expectation of χ^2 goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Society*, 65, 864-886. URL: <http://www.jstor.org/stable/2284594>

³⁰Kleinbaum DG, Kupper LL, Muller KE (1988). *Applied regression analysis and other multivariable methods* (2nd Ed.). Boston, MA: PWS-Kent.

³¹Green PE (1978). *Analyzing multivariate data*. Hillsdale, IL: Dryden.

³²Pedhazur EJ (1982). *Multiple regression in behavioral research* (2nd Ed.). New York, NY: CBS College Publishing.

³³Tabachnick BG, Fidell LS (1983). *Using multivariate statistics*. New York, NY: Harper and Row.

³⁴Kendall M (1975). *Multivariate Analyses*. New York: Hafner (Chapter 7).

³⁵Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, 1, 144-160. URL: <http://odajournal.com/2013/09/19/62/>

Chapter 8

¹Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, 56, 430-442. DOI: 10.1177/0013164496056003005

²Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, 56, 656-667. DOI: 10.1177/0013164496056004007

³Gehrlein WV (1986). General mathematical programming formulations for the statistical classification problem. *Operations Research Letters*, 5, 299-304. DOI: 10.1016/0167-6377(86)90068-4

⁴Joachimsthaler EA, Stam A (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioral Research*, 25, 427-454. DOI: 10.1207/s15327906mbr2504_2

⁵Koehler GJ, Erenguc SS (1990). Minimizing misclassifications in linear discriminant analysis. *Decision Sciences*, 21, 63-74. DOI: 10.1111/j.1540-5915.1990.tb00317.x

⁶Rubin PA (1990). Heuristic solution procedures for a mixed-integer programming discriminant model. *Mangerial and Decision Economics*, 11, 255-266. DOI: 10.1002/mde.4090110407

⁷Stam A, Joachimsthaler EA (1990). A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research*, 46, 113-122. DOI: 10.1016/0377-2217(90)90304-T

⁸Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, 13, 1015-1021. DOI: 10.1002/sim.4780131004

⁹Soltysik RC, Yarnold PR (2010). Two-group MultiODA: a mixed-integer programming solution with bounded M . *Optimal Data Analysis*, 1, 30-37. URL: <http://optimalprediction.com/files/pdf/V1A4.pdf>

¹⁰Karmarkar N (1984). A new polynomial time algorithm for linear programming. *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, 302-311. URL: <http://dl.acm.org/citation.cfm?id=808695>

¹¹Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

¹²Curry RH, Yarnold PR, Bryant FB, Martin GJ, Hughes RL (1988). A path analysis of medical school and residency performance: implications for housestaff selection. *Evaluation in the Health Professions*, 11, 113-129. DOI: 10.1177/016327878801100108

¹³Bajgier SM, Hill AV (1982). An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences*, 13, 604-612. DOI: 10.1111/j.1540-5915.1982.tb01185.x

¹⁴Cornell R, Luginbuhl RC, Yeo C (1989). *SAS/OR user's guide, version 6*. Durham, NC: SAS Institute.

¹⁵LG Grimm, PR Yarnold (1995). *Reading and understanding multivariate statistics*. Washington DC: APA Books.

¹⁶Asparoukhov OK, Stam A (1997). Mathematical programming formulations for two-group classification with binary variables. *Annals of Operations Research*, 74, 89-112. DOI: 10.1023/A:1018995010063

¹⁷Rubin PA (1997). Solving mixed-integer classification problems by decomposition. *Annals of Operations Research*, 74, 51-64. DOI: 10.1023/A:1018990909155

¹⁸Silva APD, Stam A (1997). A mixed-integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. *Annals of Operations Research*, 74, 129-157. DOI: 10.1023/A:1018962102794

¹⁹Pfetsch ME (2008). Branch-and-cut for the maximum feasible subsystem problem. *SIAM Journal on Optimization*, 19, 21-38. URL: <http://dx.doi.org/10.1137/050645828>

²⁰Bremner D, Chen D (2009). A branch and cut algorithm for the halfspace depth problem. arXiv:0910.1923v1. URL: <http://arxiv.org/abs/0910.1923>

²¹Soltysik RC, Yarnold PR (1994). The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research*, 21, 735-745. DOI: 10.1016/0305-0548(94)90003-5

²²Rubin PA (1999). Adapting the Warmack-Gonzalez algorithm to handle discrete data. *European Journal of Operational Research*, 113, 632-642. DOI: 10.1016/S0377-2217(97)00448-7

²³Loucopoulos C, Pavur R (1997). Computational characteristics of a new mathematical programming model for the three-group discriminant problem. *Computers & Operations Research*, 24, 179-191. DOI: 10.1016/S0305-0548(96)00046-9

²⁴Adem J, Gochet W (2006). Mathematical programming based heuristics for improving LP-generated classifiers for the multiclass supervised classification problem. *European Journal of Operational Research*, 168, 181-199. DOI: 10.1016/j.ejor.2004.04.031

²⁵Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine*, 10, 601-606. DOI: 10.1007/BF02602743

²⁶Green DM, Swets JA (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

²⁷Kraemer HC (1992). *Evaluating medical tests*. Newbury Park, CA: Sage.

²⁸Loke WH (1989). Diagnostic evaluations using signal detection analysis. *Indian Journal of Psychological Medicine*, 12, 87-91.

²⁹Swets JA (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47, 522-532. DOI: 10.1037/0003-066X.47.4.522

³⁰Rubin PA (1992). *Mathematical programming and alternative classification models*. Invited address presented at the TIMS/ORSA Joint National Meetings, Orlando, FL.

³¹Smithson M (1987). *Fuzzy set analysis for behavioral and social sciences*. New York, NY: Springer-Verlag.

³²Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, 3, 30-31. URL: <http://optimalprediction.com/files/pdf/V3A12.pdf>

³³Page CV (1977). Heuristics for signature table analysis as a pattern recognition technique. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7, 77-86. DOI: 10.1109/TSMC.1977.4309658

³⁴Reynolds HT (1977). *The analysis of cross-classifications*. New York, NY: Free Press.

³⁵Yarnold PR, Soltysik RC, Curry RH, Martin GJ (1989). *In quest of the best: Resident selection based on application information and mixed integer programming*. Paper presented at the Annual Meeting of the Society of Behavioral Medicine, San Francisco, CA.

³⁶Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine*, 17, 2405-2414. DOI: 10.1002/(SICI)1097-0258(19981030)17:20<2405::AID-SIM928>3.0.CO;2-F

Chapter 9

¹Grimm LG, PR Yarnold (1995). *Reading and understanding multivariate statistics*. Washington DC: APA Books.

²Grimm LG, PR Yarnold (2000). *Reading and understanding more multivariate statistics*. Washington DC: APA Books.

³Bryant FB, Yarnold PR (1995). Principal components analysis and exploratory and confirmatory factor analysis. In: LG Grimm, PR Yarnold (Ed's.), *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 99-136).

⁴Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

⁵Yarnold PR (2015). Evaluating non-confounded association of an attribute and a class variable using partial UniODA. *Optimal Data Analysis*, 4, 32-35. URL: <http://optimalprediction.com/files/pdf/V4A10.pdf>

⁶Yarnold PR, Michelson EA, Thompson DA, Adams SL (1998). Predicting patient satisfaction: A study of two emergency departments. *Journal of Behavioral Medicine*, 21, 545-563. DOI: 10.1023/A:1018796628917

⁷Thompson DA, Yarnold PR (1995). Relating patient satisfaction to waiting time perceptions and expectations: The disconfirmation paradigm. *Academic Emergency Medicine*, 2, 1057-1062. DOI: 10.1111/j.1553-2712.1995.tb03150.x

⁸Thompson DA, Yarnold PR, Adams SL, Spacone AB (1996). How accurate are patient's waiting time perceptions? *Annals of Emergency Medicine*, 28, 652-656. DOI: 10.1016/S0196-0644(96)70089-6

⁹Thompson DA, Yarnold PR, Williams DR, Adams SL (1996). The effects of actual waiting time, perceived waiting time, information delivery and expressive quality on patient satisfaction in the emergency department. *Annals of Emergency Medicine*, 28, 657-665. DOI: 10.1016/S0196-0644(96)70090-2

¹⁰Yarnold PR, Soltysik RC (2010). Unconstrained covariate adjustment in CTA. *Optimal Data Analysis*, 1, 38-40. URL: <http://optimalprediction.com/files/pdf/V1A5.pdf>

¹¹Curtis JR, Yarnold PR, Schwartz DN, Wein-stein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 162, 393-398. DOI: 10.1164/ajrccm.162.2.9909014

¹²Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, 1, 144-160. URL: <http://optimalprediction.com/files/pdf/V1A16.pdf>

¹³Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law & Human Behavior*, 28, 53-271. DOI: 10.1023/B:LAHU.0000029138.92866.af

¹⁴Yarnold PR (2015). Optimal statistical analysis involving a confounding variable. *Optimal Data Analysis*, 4, 87-103. URL: <http://optimalprediction.com/files/pdf/V4A16.pdf>

¹⁵Levinson W, Lesser CS, Epstein RM (2015). Developing physician communication skills for patient-centered care. *Health Affairs*, 29, 1310-1318. DOI: 10.1377/hlthaff.2009.0450

¹⁶Podolsky A, Stern DT, Peccoralo L (2015). The courteous consult: A CONSULT card and training to improve resident consults. *Journal of Graduate Medical Education*, 7, 113-117. DOI: 10.4300/JGME-D-14-00207.1

¹⁷Sayah A, Rogers L, Devarajan K, Kingsley-Rocker L, Lobon LF (2014). Minimizing ED waiting times and improving patient flow and experience of care. *Emergency Medicine International*, 1984, article ID 981472. DOI: 10.1155/2014/981472

¹⁸Conrad P, Barker KK (2010). The social construction of illness: Key insights and policy implications. *Journal of Health and Social Behavior*, 51, S67-S69. DOI: 10.1177/0022146510383495

¹⁹Fitzcharles MA, Yunus MB (2012). The clinical concept of fibromyalgia as a changing paradigm in the past 20 years. *Pain Research and Treatment*, 2012, article ID 184835. DOI: 10.1155/2012/184835

²⁰Grahame R (2001). Time to take hypermobility seriously (in adults and children). *Rheumatology*, 40, 485-487. DOI: 10.1093/rheumatology/40.5.485

²¹Welch SJ (2013). Twenty years of patient satisfaction research applied to the Emergency Department: A qualitative review. *American Journal of Medical Quality*, 25, 64-72. DOI: 10.1177/1062860609352536

²²Cheng SH, Yang MC, Chiang TL (2003). Patient satisfaction with and recommendation of a hospital: Effects of interpersonal and technical aspects of hospital care. *International Journal for Quality in Health Care*, 15, 345-355. DOI: 10.1093/intqhc/mzg045

²³Yarnold PR (1984). Note on the multidisciplinary scope of psychological androgyny theory. *Psychological Reports*, 55, 936-938. DOI: 10.2466/pr0.1984.54.3.936

²⁴Yarnold PR (1993). A brief measure of psychological androgyny for use in predicting physicians' decision making. *Academic Medicine*, 68, 312. DOI: 10.1097/00001888-199304000-00027

²⁵Yarnold PR (1994). Comparing the split-half reliability of androgyny and sex-typing measures. *Australian Journal of Psychology*, 46, 164-169. DOI: 10.1080/00049539408259491

²⁶Yarnold PR, Bryant FB, Litsas F (1989). Type A behavior and psychological androgyny among Greek college students. *European Journal of Personality*, 3, 249-268. DOI: 10.1002/per.2410030403

²⁷Yarnold PR, Nightingale SD, Curry RH, Martin GJ (1991). Psychological androgyny and preference in loss-framed gambles of medical students: Possible implications for resource utilization. *Medical Decision Making*, 11, 176-179. DOI: 10.1177/0272989X9101100306

²⁸Yarnold PR, Nightingale SD, Curry RH, Martin GJ (1990). Psychological androgyny and preference for intubation in a hypothetical case of end-stage lung disease. *Medical Decision Making*, 10, 215-222. DOI: 10.1177/0272989X9001000309

²⁹Yarnold PR, Martin GJ, Soltysik RC, Nightingale SD (1993). Androgyny predicts empathy for trainees in medicine. *Perceptual and Motor Skills*, 77, 576-578. DOI: 10.2466/pms.1993.77.2.576

³⁰Yarnold PR, Bryant FB, Nightingale SD, Martin GJ (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health, and Medicine*, 1, 207-221. DOI: 10.1080/13548509608400019

³¹Nightingale SD, Yarnold PR, Greenberg MS (1991). Sympathy, empathy, and physician resource utilization. *Journal of General Internal Medicine*, 6, 420-423. DOI: 10.1007/BF02598163

³²Yarnold PR (1990). Androgyny and sex-typing as continuous independent factors, and a glimpse of the future. *Multivariate Behavioral Research*, 25, 407-419. DOI: 10.1207/s15327906mbr2503_10

³³Yarnold PR (2015). Optimal statistical analysis involving multiple confounding variables. *Optimal Data Analysis*, 4, 107-112. URL: <http://optimalprediction.com/files/pdf/V4A18.pdf>

³⁴Robins JM, Hernan MA, Brumback B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560. URL: <http://www.jstor.org/stable/3703997?origin=JSTOR-pdf>

³⁵Yarnold PR (2015). GO-CTA vs. marginal structural model: Observed data from a point-treatment study, stratified by known confounder. *Optimal Data Analysis*, 4, 104-106. URL: <http://optimalprediction.com/files/pdf/V4A17.pdf>

³⁶Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>

³⁷Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, 56, 430-442. DOI: 10.1177/0013164496056003005

³⁸Bryant 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. URL: <http://optimalprediction.com/files/pdf/V1A19.pdf>

³⁹Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1, 67-100. URL: <http://optimalprediction.com/files/pdf/V2A33.pdf>

⁴⁰<http://data.giss.nasa.gov/gistemp/>

⁴¹Barnston AG, Livezey RE (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115, 1083-1126. DOI: 10.1175/1520-0493(1987)115%3C1083:CSAPOL%3E2.0.CO;2

⁴²<http://www.cpc.noaa.gov/data/teledoc/telepatcalc.shtml>

- ⁴³<http://www.cdc.noaa.gov/cgi-bin/Timeseries/timeseries1.pl>
- ⁴⁴Yarnold PR (1992). Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, and S Tindale (Eds.), *Social psychological applications to social issues, volume 2: Methodological issues in applied social research*. New York, NY: Plenum (pp. 177-197).
- ⁴⁵Cade BS, Richards JD (2005). *User manual for blossom statistical software*. Fort Collins, CO: US Geological Survey.
- ⁴⁶Kwok R, Cunningham GF, Pang SS (2004). Fram Strait sea ice outflow. *Journal of Geophysical Research*, 109, 1009-1029. DOI: 10.1029/2003JC001785
- ⁴⁷Holland MM (2003). The north Atlantic oscillation–Arctic oscillation in the CCSM2 and its influence on Arctic climate variability. *Journal of Climate*, 16, 2767–2781. 10.1175/1520-0442(2003)016%3C2767:TNAOI%3E2.0.CO;2
- ⁴⁸Hilmer M, Jung T (2000). Evidence for a recent change in the link between north Atlantic oscillation and Arctic sea ice export. *Geophysical Research Letters*, 27, 989–992. DOI: 10.1029/1999GL010944
- ⁴⁹Willoughby HE, Rappaport EN, Marks FD (2007). Hurricane forecasting: the state of the art. *Natural Hazards Review*, 8, 45-49. DOI: 10.1061/(ASCE)1527-6988(2007)8:3(45)
- ⁵⁰Charlton AJ, Polvani LM (2007). A new look at stratospheric sudden warming events: part I. Climatology and modeling benchmarks. *Journal of Climate*, 20, 449-469. DOI: 10.1175/JCLI3996.1
- ⁵¹Charlton AJ, Polvani LM, Perlitz J, Sassi F, Manzini E (2007). A new look at stratospheric sudden warming events: part II. Evaluation of numerical model simulations. *Journal of Climate*, 20, 470-488. DOI: 10.1175/JCLI3994.1
- ⁵²Troccoli A (2008). Management of weather and climate risk in the energy industry. Proceedings of the NATO advanced research workshop on weather/climate risk management for the energy sector, Santa Maria di Leuca, Italy 6-10 October 2008. Series: *NATO science for peace and security series C: environmental security*. New York, NY: Springer.
- ⁵³Babkina AM (2004). *El Niño: overview and bibliography*. Haup-Pauge, NY: Nova Science Publishers.
- ⁵⁴Kalnay E (2002). *Atmospheric modeling, data assimilation and predictability*. Cambridge, UK: Cambridge Univ Press.
- ⁵⁵Yarnold PR (2013). Comparing attributes measured with “identical” Likert-type scales in single-case designs with UniODA. *Optimal Data Analysis*, 2, 148-153. URL: <http://optimalprediction.com/files/pdf/V2A22.pdf>
- ⁵⁶Collinge WC, Solysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>
- ⁵⁷Collinge W, Yarnold PR, Solysik, RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103%2F1947-2714.118920
- ⁵⁸Yarnold PR, Feinglass J, Martin GJ, McCarthy WJ (1999). Comparing three pre-processing strategies for longitudinal data for individual patients: An example in functional outcomes research. *Evaluation and the Health Professions*, 22, 254-277. DOI: 10.1177/01632789922034301
- ⁵⁹Yarnold 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. URL: <http://optimalprediction.com/files/pdf/V2A25.pdf>

Chapter 10

- ¹Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, 56, 656-667. DOI: 10.1177/0013164496056004007

²Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, 4, 36-53. URL: <http://optimalprediction.com/files/pdf/V4A11.pdf>

³Yarnold PR (2014). Increasing the likelihood of an ambivalent patient recommending the Emergency Department to others, *Optimal Data Analysis*, 3, 89-91. URL: <http://optimalprediction.com/files/pdf/V3A20.pdf>

⁴Bengio Y, Grandvalet Y (2005). Bias in estimating the variance of K-Fold cross-validation. In: P Duchesne, B RE Millard (Ed's.), *Statistical modeling and analysis for complex data problems*. New York, NY: Springer (pp. 75-95).

⁵Rodriguez JD, Perez A, Lozano JA (2009). Sensitivity analysis of K-Fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 569-575. DOI: 10.1109/TPAMI.2009.187

⁶Witten IH, Frank E, Hall MA (2011). *Data Mining: Practical Machine Learning Tools and Technique* (3rd Ed.). San Francisco, CA: Morgan Kaufmann.

⁷<http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf>

⁸<http://www.autonlab.org/tutorials/overfit10.pdf>

⁹http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf

¹⁰Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC: APA Books.

¹¹Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, 1, 23-29. URL: <http://optimalprediction.com/files/pdf/V1A3.pdf>

¹²Yarnold PR (2010). Aggregated vs. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis*, 1, 46-49. URL: <http://optimalprediction.com/files/pdf/V1A8.pdf>

¹³Yarnold PR, Bryant FB (2013). Analysis involving categorical attributes having many categories. *Optimal Data Analysis*, 2, 69-70. URL: <http://optimalprediction.com/files/pdf/V2A12.pdf>

¹⁴Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, 2, 172-176. URL: <http://optimalprediction.com/files/pdf/V2A12.pdf>

¹⁵Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: <http://optimalprediction.com/files/pdf/V2A27.pdf>

¹⁶Yarnold PR, Soltysik RC (2013). Ipsative transformations are *essential* in the analysis of serial data. *Optimal Data Analysis*, 2, 94-97. URL: <http://optimalprediction.com/files/pdf/V2A17.pdf>

¹⁷Yarnold PR (2013). Comparing attributes measured with “identical” Likert-type scales in single-case designs with UniODA. *Optimal Data Analysis*, 2, 148-153. URL: <http://optimalprediction.com/files/pdf/V2A22.pdf>

¹⁸Yarnold PR (2013). Assessing hold-out validity of CTA models using UniODA. *Optimal Data Analysis*, 2, 31-36. URL: <http://optimalprediction.com/files/pdf/V2A5.pdf>

¹⁹Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1, 67-100. URL: <http://optimalprediction.com/files/pdf/V1A13.pdf>

²⁰Yarnold PR (2013). Standards for reporting UniODA findings expanded to include *ESP* and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119. URL: <http://optimalprediction.com/files/pdf/V2A19.pdf>

²¹Stalans LJ, Seng M (2006). Identifying subgroups at high risk of dropping out of domestic batterer treatment: The buffering effects of a high school education. *International Journal of Offender Therapy and Comparative Criminology*, 10, 1-19. DOI: 10.1177/0306624X06290204

²²Soltysik RC, Yarnold PR (2014). Hierarchically optimal classification tree analysis of adverse drug reactions secondary to warfarin therapy. *Optimal Data Analysis*, 3, 23-24. URL: <http://optimalprediction.com/files/pdf/V3A9.pdf>

²³Collinge W, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

²⁴Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, 2, 43-48. URL: <http://optimalprediction.com/files/pdf/V2A6.pdf>

²⁵Licht MH (1995). Multiple regression and correlation. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington, DC: APA Books (pp. 19-64).

²⁶Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, 1: 144-160. URL: <http://optimalprediction.com/files/pdf/V1A16.pdf>

²⁷Yarnold PR, Soltysik RC (2013). Reverse CTA: An optimal analog to analysis of variance. *Optimal Data Analysis*, 2, 43-47. URL: <http://optimalprediction.com/files/pdf/V2A7.pdf>

²⁸Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, 16, 1451-1463. DOI: 10.1002/(SICI)1097-0258(19970715)16:13<1451::AID-SIM571>3.0.CO;2-F

²⁹Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 162, 393-398. DOI: 10.1164/ajrccm.162.2.9909014

³⁰Kim B, Lyons TM, Parada JP, Uphold CR, Yarnold PR, Hounshell JB, Sipler AM, Goetz MB, DeHovitz JA, Weinstein RA, Campo RE, Bennett CL (2001). HIV-related *Pneumocystis carinii* pneumonia in older patients hospitalized in the early HAART era. *Journal of General Internal Medicine*, 16, 583-589. DOI: 10.1046/j.1525-1497.2001.016009583.x

³¹Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmeil JS, Sipler AM, Chan C, Goetz MB, Schwartz DN, Bennett CL (2000). A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine*, 161, 1081-1086. DOI: 10.1164/ajrccm.161.4.9906072

³²Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest*, 123, 1151-1160. DOI: 10.1378/chest.123.4.1151

³³Kyriacou DN, Yarnold PR, Stein AC, Schmitt BP, Soltysik RC, Nelson RR, Frerichs RR, Noskin GA, Belknap SB, Bennett CL (2007). Discriminating inhalational anthrax from community-acquired pneumonia using chest radiograph findings and a clinical algorithm. *Chest*, 131, 489-495. DOI: 10.1378/chest.06-1687

³⁴Kyriacou DM, Yarnold PR, Soltysik RC, Wunderink RG, Schmitt BP, Parada JP, Adams JG (2008). Derivation of a triage algorithm for chest radiography of community-acquired pneumonia in the emergency department. *Academic Emergency Medicine*, 15, 40-44. DOI: 10.1111/j.1553-2712.2007.00011.x

³⁵Nebeker JR, Yarnold PR, Soltysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Wu X (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care*, 45, S81-S88. DOI: 10.1097/MLR.0b013e3180616c2c

³⁶Belknap SM, Moore H, Lanzotti SA, Yarnold PR, Getz M, Deitrick DL, Peterson A, Akeson J, Maurer T, Soltysik RC, Storm GA, Brooks I (2008). Application of software design principles and debugging methods to an analgesia

prescription reduces risk of severe injury from medical use of opioids. *Clinical Pharmacology and Therapeutics*, 84, 385-392. DOI: DOI: 10.1038/clpt.2008.24

³⁷Feinglass J, Yarnold PR, Martin GJ, McCarthy WJ (1998). A classification tree analysis of selection for discretionary treatment. *Medical Care*, 36, 740-747. URL: <http://www.jstor.org/stable/3767410>

³⁸Grobman WA, Terkildsen MF, Soltysik RC, Yarnold PR (2008). Predicting outcome after emergent cerclage using classification tree analysis. *American Journal of Perinatology*, 25, 443-448. DOI: 10.1055/s-0028-1083843

³⁹Zakaria A, Bandarenko N, Pandey DK, Auerbach A, Raisch DW, Kim B, Kwaan HC, McKoy JM, Schmitt BP, Davidson CJ, Yarnold PR, Gorelick PB, Bennett CL (2004). Clopidogrel-associated TTP: An update of pharmacovigilance efforts conducted by independent researchers, pharmaceutical suppliers, and the Food and Drug Administration. *Stroke*, 35, 533-538. DOI: 10.1161/01.STR.0000109253.66918.5E

⁴⁰Bennett CL, Kim B, Zakaria A, Bandarenko N, Pandey DK, Buffie CG, McKoy JM, Tvar AD, Cursio JFR, Yarnold PR, Kwaan HC, De Masi D, Sarode R, Raife TJ, Kiss JE, Raisch DW, Davidson C, Sadler JE, Ortell TL, Zheng XL, Kato S, Matsumoto M, Uemura M, Fujimura Y (2007). Two mechanistic pathways for thienopyridine-associated thrombotic thrombocytopenic purpura: A report from the Surveillance, Epidemiology, and Risk Factors for Thrombotic Thrombocytopenic Purpura (SERF-TTP) research group and the Research on Adverse Drug events And Reports (RADAR) project. *Journal of American College of Cardiology*, 50, 1138-1143. DOI: 10.1016/j.jacc.2007.04.093

⁴¹Kanter AS, Spencer DC, Steinberg MH, Soltysik RC, Yarnold PR, Graham NM (1999). Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. *Journal of Acquired Immune Deficiency Syndrome*, 21, 252-253. URL: http://journals.lww.com/jaids/Citation/1999/07010/Supplemental_Vitamin_B_and_Progression_to_AIDS_and.11.aspx

⁴²Green D, Hartwig D, Chen D, Soltysik RC, Yarnold PR (2003). Spinal cord injury risk assessment for thromboembolism (SPIRATE study). *American Journal of Physical and Medical Rehabilitation*, 82, 950-956. URL: http://journals.lww.com/ajpmr/Abstract/2003/12000/Spinal_Cord_Injury_Risk_Assessment_for.7.aspx

⁴³Yarnold PR, Michelson EA, Thompson DA, Adams SL (1998). Predicting patient satisfaction: A study of two emergency departments. *Journal of Behavioral Medicine*, 21, 545-563. DOI: 10.1023/A:1018796628917

⁴⁴Kucera CM, Greenberger PA, Yarnold PR, Choy AC, Levenson T (1999). An attempted prospective testing of an asthma severity index and a quality of life survey for 1 year in ambulatory patients with asthma. *Allergy and Asthma Proceedings*, 20, 29-38. DOI: 10.2500/108854199778681521

⁴⁵Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine*, 21, 140-145. DOI: 10.1111/j.1525-1497.2005.00300.x

⁴⁶Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, 66, 660-672. DOI: 10.1037/0022-006X.66.4.660

⁴⁷Mueser KT, Yarnold PR, Rosenberg SD, Drake RE, Swett C, Miles KM, Hill D (2000). Substance use disorder in hospitalized severely mentally ill psychiatric patients: Prevalence, correlates, and sub-groups. *Schizophrenia Bulletin*, 26, 179-193. URL: <http://psycnet.apa.org/journals/szb/26/1/179/>

⁴⁸Collinge W, Yarnold PR, Raskin E (1998). Use of mind/body self-healing practice predicts positive health transition in chronic fatigue syndrome: a controlled study. *Subtle Energies & Energy Medicine*, 9, 171-190. URL: <http://journals.sfu.ca/seemj/index.php/seemj/article/view/256>

⁴⁹Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Clinical Child and Adolescent Psychology*, 36, 8-18. DOI: 10.1080/15374410709336564

⁵⁰<http://ODAJournal.com>

Chapter 11

¹Yarnold PR, Bryant FB (2015). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis*, 4, 54-61. URL: <http://optimalprediction.com/files/pdf/V4A12.pdf>

²Arozullah AM, Gordon HS, Yarnold PR, Soltysik RC, Ferreira MR, Wolf MS, Molokie R, Bhoopalam N, Bennett CL (2008). Predictors of prostate cancer stage at presentation. *Journal of General Internal Medicine*, 23, 376.

³Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine*, 21, 140-145. DOI:10.1111/j.1525-1497.2005.00300.x

⁴Bryant 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. URL: <http://optimalprediction.com/files/pdf/V3A14.pdf>

⁵Bryant FB, Yarnold PR (2014). Type A behavior, pessimism and optimism among college undergraduates. *Optimal Data Analysis*, 3, 32-35. URL: <http://optimalprediction.com/files/pdf/V3A13.pdf>

⁶Collinge WC, Kahn J, Walton T, Kozak L, Bauer-Wu S, Fletcher K, Yarnold PR, Soltysik RC (2013). Touch, caring, and cancer: Randomized controlled trial of a multimedia caregiver education program. *Supportive Care in Cancer*, 21, 1405-1414. DOI: 10.1007/s00520-012-1682-6

⁷Collinge WC, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

⁸Collinge WC, Yarnold PR, Soltysik RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103/1947-2714.118920

⁹Grobman WA, Terkildsen MF, Soltysik RC, Yarnold PR (2008). Predicting outcome after emergent cerclage using classification tree analysis. *American Journal of Perinatology*, 25, 443-448. DOI: 10.1055/s-0028-1083843

¹⁰Kyriacou DN, Yarnold PR, Stein AC, Schmitt BP, Soltysik RC, Nelson RR, Frerichs RR, Noskin GA, Belknap SB, Bennett CL (2007). Discriminating inhalational anthrax from community-acquired pneumonia using chest radiograph findings and a clinical algorithm. *Chest*, 131, 489-495.

¹¹Kyriacou DM, Yarnold PR, Soltysik RC, Wunderink RG, Schmitt BP, Parada JP, Adams JG (2008). Derivation of a triage algorithm for chest radiography of community-acquired pneumonia in the emergency department. *Academic Emergency Medicine*, 15, 40-44. DOI: 10.1111/j.1553-2712.2007.00011.x

¹²Lyons AM, Leon SC, Zaddach C, Luboyeski EJ, Richards M (2009). Predictors of clinically significant sexual concerns in a child welfare population. *Journal of Child and Adolescent Trauma*, 2, 28-45. DOI: 10.1080/19361520802675884

¹³Nebeker JR, Yarnold PR, Soltysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Xu W (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care*, 45, S81-S88. URL: <http://www.effectivehealthcare.ahrq.gov/repFiles/MedCare/s81.pdf>

¹⁴Sieracki JH, Fuller AK, Leon SC, Jhe Bai G, Bryant FB (2015). The role of race, socioeconomic status, and System of Care services in placement decision-making. *Children and Youth Services Review*, DOI: 10.1016/j.childyouth.2014.12.013

¹⁵Smith JH, Bryant FB, Njus D, Posavac EJ (2010). Here today, gone tomorrow: Understanding freshman attrition using Person-Environment Fit Theory. *Optimal Data Analysis*, 1, 101-124. URL: <http://optimalprediction.com/files/pdf/V1A14.pdf>

¹⁶Snowden J, Leon S, Sieracki J (2008). Predictors of children in foster care being adopted: A classification tree analysis. *Children and Youth Services Review*, 30, 1318-1327. DOI: 10.1016/j.childyouth.2008.03.014

¹⁷Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Child and Adolescent Psychology*, 36, 8-18. DOI: 10.1080/15374410709336564

¹⁸Soltysik RC, Yarnold PR (2014). Hierarchically optimal classification tree analysis of adverse drug reactions secondary to warfarin therapy. *Optimal Data Analysis*, 3, 23-24. URL: <http://optimalprediction.com/files/pdf/V3A9.pdf>

¹⁹Stalans LJ, Hacker R, Talbot ME (2010). Comparing nonviolent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behavior*, 37, 613-628. DOI: 10.1177/0093854810363794

²⁰Stalans LJ, Seng M (2006). Identifying subgroups at high risk of dropping out of domestic batterer treatment: The buffering effects of a high school education. *International Journal of Offender Therapy and Comparative Criminology*, 10, 1-19. DOI: 10.1177/0306624X06290204

²¹Stoner AM, Leon SC, Fuller AK (2013). Predictors of reduction in symptoms of depression for children and adolescents in foster care. *Journal of Child and Family Studies*, 22, DOI 10.1007/s10826-013-9889-9

²²Suzuki H, Bryant FB, Edwards JD (2010). Tracing prospective profiles of juvenile delinquency: An optimal classification tree analysis. *Optimal Data Analysis*, 1, 125-143. URL: <http://optimalprediction.com/files/pdf/V1A15.pdf>

²³Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, 3, 102-106. URL: <http://optimalprediction.com/files/pdf/V3A24.pdf>

²⁴Yarnold PR, Soltysik RC, Collinge WC (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, 2, 37-42. URL: <http://optimalprediction.com/files/pdf/V2A6.pdf>

²⁵Yarnold PR, Bryant FB, & Smith JH. (2013). Manual vs. automated CTA: Predicting freshman attrition. *Optimal Data Analysis*, 2, 48-53. URL: <http://optimalprediction.com/files/pdf/V2A8.pdf>

²⁶Yarnold PR, Soltysik RC (2010). Manual vs. automated CTA: Optimal preadmission staging for inpatient mortality from *Pneumocystis cariini* pneumonia. *Optimal Data Analysis*, 1, 50-54. URL: <http://optimalprediction.com/files/pdf/V1A9.pdf>

²⁷Coakley RM, Holmbeck GN, Bryant FB, Yarnold PR (2010). Manual vs. automated CTA: Predicting adolescent psychosocial adaptation. *Optimal Data Analysis*, 1, 55-58. URL: <http://optimalprediction.com/files/pdf/V1A10.pdf>

²⁸Yarnold PR (2015). Optimal statistical analysis involving a confounding variable. *Optimal Data Analysis*, 4, 87-103. URL: <http://optimalprediction.com/files/pdf/V4A16.pdf>

²⁹Smith JH, Bryant FB, Njus D, Posavac EJ (2010). Here today, gone tomorrow: Understanding freshman attrition using person-environment fit theory. *Optimal Data Analysis*, 1, 101-124. URL: <http://optimalprediction.com/files/pdf/V1A14.pdf>

³⁰Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>

Chapter 12

¹Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, 3, 55-77. URL: <http://optimalprediction.com/files/pdf/V3A17.pdf>

²Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, 3, 78-84. URL: <http://optimalprediction.com/files/pdf/V3A18.pdf>

³Yarnold PR (2015). Distance from a theoretically ideal statistical classification model defined as the number of additional equivalent effects needed to obtain perfect classification for the sample. *Optimal Data Analysis*, 4, 81-86. URL: <http://optimalprediction.com/files/pdf/V4A15.pdf>

⁴Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC: APA Books.

⁵Bryant FB, Yarnold PR (1995). Principal components, and exploratory and confirmatory factor analysis. In: L.G Grimm, PR Yarnold (Eds.), *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books (pp. 99-136).

⁶Soltysik RC, Yarnold PR (2010). Introduction to automated CTA software. *Optimal Data Analysis*, 1, 144-160. URL: <http://optimalprediction.com/files/pdf/V4A12.pdf>

⁷Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119. URL: <http://optimalprediction.com/files/pdf/V2A19.pdf>

⁸Shanker R (1994). *Principles of quantum mechanics, 2nd edition*. New York, NY: Springer.

⁹Wilde MM (2013). *Quantum information theory*. Cambridge, UK: Cambridge University Press.

¹⁰Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, DC: APA Books.

¹¹Grimm LG, Yarnold PR (2000). *Reading and understanding more multivariate statistics*. Washington, DC: APA Books.

¹²Yarnold PR, Soltysik RC (2014). Discrete 95% confidence intervals for ODA model- and chance-based classifications. *Optimal Data Analysis*, 3, 110-112. URL: <http://optimalprediction.com/files/pdf/V3A26.pdf>

¹³Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmeil JS, Sipler AM, Chan C, Goetz MB, Schwartz DN, Bennett CL (2000). A new pre-admission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine*, 161, 1081-1086. DOI: 10.1164/ajrccm.161.4.9906072

¹⁴Yarnold (2014). Illustrating how 95% confidence intervals indicate model redundancy. *Optimal Data Analysis*, 3, 96-97. URL: <http://optimalprediction.com/files/pdf/V3A22.pdf>

¹⁵Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, based on the November 2011 submission.

¹⁶Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, 3, 102-106. URL: <http://optimalprediction.com/files/pdf/V3A24.pdf>

Appendix A

¹Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC: APA Books.

Alphabetical References

- Adem J, Gochety W (2006). Mathematical programming based heuristics for improving LP-generated classifiers for the multiclass supervised classification problem. *European Journal of Operational Research*, 168, 181-199. DOI: 10.1016/j.ejor.2004.04.031
- Agresti A (1990). *Categorical data analysis*. Hoboken, NJ, Wiley (pp. 356-357, 360-361).
- Aldrich JH, Cnudde, C (1975). Probing the bounds of conventional wisdom: Comparison of regression, probit, and discriminant analysis. *American Journal of Political Science*, 19, 571-608. URL: <http://www.jstor.org/stable/2110547>
- Aldrich JH, Nelson FD (1984). *Linear probability, logit, and probit models*. Beverly Hills, CA: Sage.
- Allen MJ, Yen WM (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Appleton DR (1995). Pitfalls in the interpretation of studies: III. *Journal of the Royal Society of Medicine*, 88, 241-243. DOI: 10.1177/014107689508800501
- Arozullah AM, Gordon HS, Yarnold PR, Soltysik RC, Ferreira MR, Wolf MS, Molokie R, Bhoopalam N, Bennett CL (2008). Predictors of prostate cancer stage at presentation. *Journal of General Internal Medicine*, 23, 376.
- Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine*, 21, 140-145. DOI: 10.1111/j.1525-1497.2005.00300.x
- Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest*, 123: 1151-1160. DOI: 10.1378/chest.123.4.1151
- Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmeil JS, Sipler AM, Chan C, Goetz MB, Schwartz DN, Bennett CL (2000). A new pre-admission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine*, 161, 1081-1086. DOI: 10.1164/ajrccm.161.4.9906072
- Asparoukhov OK, Stam A (1997). Mathematical programming formulations for two-group classification with binary variables. *Annals of Operations Research*, 74, 89-112. DOI: 10.1023/A:1018995010063
- Azar B (1997). APA task force urges a harder look at data. *APA Monitor*, 3, 26. DOI: 10.1641/0006-3568(2001)051[1051:SSTAR]2.0.CO;2
- Babkina AM (2004). *El Niño: overview and bibliography*. Haup-Pauge, NY: Nova Science Publishers.
- Bacus JW, Gose EE (1972). Leukocyte pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*, 513-526. DOI: <http://dx.doi.org/10.1109/TSMC.1972.4309161>
- Bajgier SM, Hill AV (1982). An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences*, 13, 604-612. DOI: 10.1111/j.1540-5915.1982.tb01185.x
- Bakeman R, Quera V (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, 118, 272-284. DOI: 10.1037/0033-2909.118.2.272
- Bamberger M, Oswald RE (2012). Impacts of gas drilling on human and animal health. *New Solutions: A Journal of Environmental and Occupational Health Policy*, 22, 51-77. DOI: 10.2190/NS.22.1.e

Barnston AG, Livezey RE (1987). Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115, 1083-1126. DOI: 10.1175/1520-0493(1987)115%3C1083:CSAPOL%3E2.0.CO;2

Baumeister RF, Tice DM (1996). Should we abandon $p < .05$? (Editorial). *Dialogue*, 11, 11.

Beck JR, Pauker SG (1983). The Markov process in medical prognosis. *Medical Decision Making*, 3, 419-458. DOI: 10.1177/0272989X8300300403

Behavioral Measurement Database Services, BMDS, PO Box 110287, Pittsburgh, PA, USA 15232-0787; 1-412-687-6850; bmdshapi@aol.com; <http://bmdshapi.com/index.html>

Belknap SM, Moore H, Lanzotti SA, Yarnold PR, Getz M, Deitrick DL, Peterson A, Akeson J, Maurer T, Soltysik RC, Storm GA, Brooks I (2008). Application of software design principles and debugging methods to an analgesia prescription reduces risk of severe injury from medical use of opioids. *Clinical Pharmacology and Therapeutics*, 84, 385-392. DOI: DOI: 10.1038/clpt.2008.24

Bengio Y, Grandvalet Y (2005). Bias in estimating the variance of K-Fold cross-validation. In: P Duchesne, B RE Millard (Ed's.), *Statistical modeling and analysis for complex data problems*. New York, NY: Springer (pp. 75-95).

Bennett CL, Kim B, Zakarija A, Bandarenko N, Pandey DK, Buffie CG, McKoy JM, Tvar AD, Cursio JFR, Yarnold PR, Kwaan HC, De Masi D, Sarode R, Raife TJ, Kiss JE, Raisch DW, Davidson C, Sadler JE, Ortel TL, Zheng XL, Kato S, Matsumoto M, Uemura M, Fujimura Y (2007). Two mechanistic pathways for thienopyridine-associated thrombotic thrombocytopenic purpura: A report from the Surveillance, Epidemiology, and Risk Factors for Thrombotic Thrombocytopenic Purpura (SERF-TTP) research group and the Research on Adverse Drug events And Reports (RADAR) project. *Journal of American College of Cardiology*, 50, 1138-1143. DOI: 10.1016/j.jacc.2007.04.093

Bennett RM, Jones J, Turk DC, Russell IJ, Matallana L (2007). An internet survey of 2,596 people with fibromyalgia. *BMC Musculoskeletal Disorders*, 8, 27. DOI: 10.1186/1471-2474-8-27

Billingsley P (1961). *Statistical inference for Markov processes*. Chicago: University of Chicago Press.

Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete multivariate analysis*. Cambridge: University Press.

Bivens AJ (2001). *Accurate classification of child molesters using context variation and Hierarchical Optimal Classification Tree Analysis*. Doctoral dissertation, Loyola University Chicago (110 pp).

Blyth CR (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67, 364-366. DOI: 10.1080/01621459.1972.10482387

Bollen KA (1989). *Structural equations with latent variables*. New York: Wiley.

Bowker AH (1948). Bowker's test for symmetry. *Journal of the American Statistical Association*, 43, 572-574. URL: <http://www.jstor.org/stable/2280710>

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

Bremner D, Chen D (2009). A branch and cut algorithm for the halfspace depth problem. arXiv:0910.1923v1. URL: <http://arxiv.org/abs/0910.1923>

Brockway JH (1997). *Here today, gone tomorrow: Understanding freshman attrition using person-environment fit theory*. Doctoral dissertation, Loyola University Chicago (112 pp).

Brown FG (1983). *Principles of educational and psychological testing* (3rd Ed.). New York: Holt.

Brown W (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322. DOI: 10.1111/j.2044-8295.1910.tb00207.x

- Bryant FB (1994). Analyze your data optimally using ODA 1.0. *Decision Line*, 25, 16-19.
- Bryant FB (2000). Assessing the validity of measurement. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding more multivariate statistics*. Washington DC: APA Books (pp. 99-146).
- Bryant FB (2003). Savoring Beliefs Inventory (SBI): A scale for measuring beliefs about savoring. *Journal of Mental Health*, 12, 175-196. DOI: 10.1080/0963823031000103489
- Bryant FB (2015). The Loyola experience (1993-2009): Optimal data analysis in the Department of Psychology. *Optimal Data Analysis*, 1, 4-9. URL: <http://optimalprediction.com/files/pdf/V1A1.pdf>
- Bryant FB (2005). How to make the best of your data [Review of Optimal Data Analysis]. *PsycCRITIQUES-Contemporary Psychology: APA Review of Books*, 50, Article 5 (7 pp). URL: file:///C:/Users/Paul/Downloads/How_to_Make_the_Best_of_Your_Data.pdf
- Bryant FB (2010). How to save the binary class variable and predicted probability of group membership from logistic regression analysis to an ASCII space-delimited file in *SPSS 17 For Windows*. *Optimal Data Analysis*, 1, 161-162. URL: <http://optimalprediction.com/files/pdf/V1A17.pdf>
- Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software. *Optimal Data Analysis*, 2, 2-6. URL: <http://optimalprediction.com/files/pdf/V2A1.pdf>
- 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. URL: <http://optimalprediction.com/files/pdf/V1A19.pdf>
- Bryant FB, Veroff J (2007). *Savoring: a new model of positive experience*. Mahwah, NY: Erlbaum.
- Bryant FB, Yarnold PR (1989). A measurement model for the short form of the student Jenkins Activity Survey. *Journal of Personality Assessment*, 53, 188-191. DOI: 10.1207/s15327752jpa5301_21
- Bryant FB, Yarnold PR (1990). The impact of Type A behavior on subjective life quality: Bad for the heart, good for the soul? *Journal of Social Behavior and Personality*, 5, 369-404.
- Bryant FB, Yarnold PR (1995). Comparing five alternative factor-models of the Student Jenkins Activity Survey: Separating the wheat from the chaff. *Journal of Personality Assessment*, 64, 145-158. DOI: 10.1207/s15327752jpa6401_10
- Bryant FB, Yarnold PR (1995). Principal components, and exploratory and confirmatory factor analysis. In: L.G Grimm, PR Yarnold (Eds.), *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books (pp. 99-136).
- Bryant FB, Yarnold PR (2014). Type A Behavior and savoring among college undergraduates: Enjoy achievements now—not later. *Optimal Data Analysis*, 3, 25-27. URL: <http://optimalprediction.com/files/pdf/V3A10.pdf>
-
- Bryant FB, Yarnold PR (2014). Type A behavior, pessimism and optimism among college undergraduates. *Optimal Data Analysis*, 3, 32-35. URL: <http://optimalprediction.com/files/pdf/V3A13.pdf>
- 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. URL: <http://optimalprediction.com/files/pdf/V3A14.pdf>
- Bryant FB, Yarnold PR, Grimm LG (1996). Toward a measurement model of the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality*, 30, 223-247. DOI: 10.1006/jrpe.1996.0015
- Bryant FB, Yarnold PR, Morgan L (1991). Type A behavior and reminiscence in college undergraduates. *Journal of Research in Personality*, 25, 418-433. DOI: 10.1016/0092-6566(91)90031-K
- Cade BS, Richards JD (2005). *User manual for blossom statistical software*. Fort Collins, CO: US Geological Survey.

- Campbell DT, Fiske DW (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105. URL: <http://psycnet.apa.org/doi/10.1037/h0046016>
- Carmelli D, Dame A, Swan G, Rosenman R (1991). Long-term changes in Type A behavior: A 27-year follow-up of the Western Collaborative Group Study. *Journal of Behavioral Medicine*, 14, 593-606. DOI: 10.1007/BF00867173
- Carmines EG, Zeller RA (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research*, 74, 223-238. DOI: 10.1023/A:1018922421450
- Charlez PA (1997). *Rock mechanics: petroleum applications*. Paris: Editions Technip.
- Charlton AJ, Polvani LM (2007). A new look at stratospheric sudden warming events: part I. Climatology and modeling benchmarks. *Journal of Climate*, 20, 449-469. DOI: 10.1175/JCLI3996.1
- Charlton AJ, Polvani LM, Perlitz J, Sassi F, Manzini E (2007). A new look at stratospheric sudden warming events: part II. Evaluation of numerical model simulations. *Journal of Climate*, 20, 470-488. DOI: 10.1175/JCLI3994.1
- Chemicals Used in Hydrolic Fracturing* (April 18, 2011). Committee on Energy and Commerce, US House of Representatives.
- Cheng SH, Yang MC, Chiang TL (2003). Patient satisfaction with and recommendation of a hospital: Effects of interpersonal and technical aspects of hospital care. *International Journal for Quality in Health Care*, 15, 345-355. DOI: 10.1093/intqhc/mzg045
- Clark CE, Veil JA (2009). *Produced Water Volumes and Management Practices in the United States*, ANL/EVS/R-09/1, Environmental Science Division, Argonne National Laboratory.
- Coakley RM (2004). *Constructing a prospective model of psychosocial resilience in early adolescents with spina bifida: An application of optimal data analysis in pediatric psychology*. Doctoral dissertation, Loyola University Chicago (233 pp).
- Coakley RM, Holmbeck GN, Bryant FB (2006). Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: An application of optimal data analysis. *Journal of Pediatric Psychology*, 31, 1084-1099. DOI: 10.1093/jpepsy/jsj032
- Coakley RM, Holmbeck GN, Bryant FB, Yarnold PR (2010). Manual vs. automated CTA: Predicting adolescent psychosocial adaptation. *Optimal Data Analysis*, 1, 55-58. URL: <http://optimalprediction.com/files/pdf/V1A10.pdf>
- Cochran WG (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-266. DOI: <http://www.jstor.org/stable/2332378>
- Cochran WG (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*, 10, 417-451. DOI: 10.2307/3001616
- Cohen J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46. DOI: 10.1177/001316446002000104
- Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen J (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003. DOI: 10.1037/0003-066X.49.12.997
- Colborn T, Kwiatkowski C, Schultz K, Bach-Ran M (2011). Natural gas operations from public health perspective. *Human and Ecological Risk Assessment: An International Journal*, 17, 1039-1056. DOI: 10.1080/10807039.2011.605662
- Colborn T, Schultz K, Herrick L, Kwiatkowski C (2014). An exploratory study of air quality near natural gas operations. *Human and Ecological Risk Assessment: An International Journal*, 20, 86-105. DOI: 10.1080/10807039.2012.749447

Colizza V, Vespignani A, Hardy EF (2007). *Impact of air travel on global spread of infectious diseases*. Indiana University.

Collinge WC, Kahn J, Walton T, Kozak L, Bauer-Wu S, Fletcher K, Yarnold PR, Soltysik RC (2013). Touch, caring, and cancer: Randomized controlled trial of a multimedia caregiver education program. *Supportive Care in Cancer*, 21, 1405-1414. DOI: 10.1007/s00520-012-1682-6

Collinge W, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: initial design and alpha test. *Optimal Data Analysis*, 1, 163-175. URL: <http://optimalprediction.com/files/pdf/V1A18.pdf>

Collinge W, Yarnold PR, Raskin E (1998). Use of mind/body self-healing practice predicts positive health transition in chronic fatigue syndrome: a controlled study. *Subtle Energies & Energy Medicine*, 9, 171-190. URL: <http://journals.sfu.ca/seemj/index.php/seemj/article/view/256>

Collinge W, Yarnold PR, Soltysik, RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, 5, 546-553. DOI: 10.4103%2F1947-2714.118920

Conrad P, Barker KK (2010). The social construction of illness: Key insights and policy implications. *Journal of Health and Social Behavior*, 51, S67-S69. DOI: 10.1177/0022146510383495

Conover WJ (1999). *Practical nonparametric statistics* (3rd Ed.). Hoboken, NJ, Wiley.

Cook TD, Campbell DT (1978). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Coombs CH, Dawes RM, Tversky A (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Cornell R, Luginbuhl RC, Yeo C (1989). *SAS/OR user's guide, version 6*. Durham, NC: SAS Institute.

Cox MK, Key CH (1993). Post hoc pairwise comparisons for the chi-square test of homogeneity of proportions. *Educational and Psychological Measurement*, 53, 951-962. DOI: 10.1177/0013164493053004008

Cromack TR (1989). Measurement considerations in clinical research. In: C.B. Royeen (Ed.), *Clinical research handbook: An analysis for the service professions*. Thorofare, NJ: Slack (ps. 47-69).

Cronbach LJ, Meehl PE (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302. DOI: 10.1037/h0040957

Cronbach LJ, Rajaratnam N, Gleser GC (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 15, 137-163. DOI: 10.1111/j.2044-8317.1963.tb00206.x

Curry RH, Yarnold PR, Bryant FB, Martin GJ, Hughes RL (1988). A path analysis of medical school and residency performance: implications for housestaff selection. *Evaluation in the Health Professions*, 11, 113-129. DOI: 10.1177/016327878801100108

Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 162, 393-398. DOI: 10.1164/ajrccm.162.2.9909014

Davies M, Fleiss JL (1982). Measuring agreement for multinomial data. *Biometrics*, 38, 1047-1051. DOI: URL: <http://www.jstor.org/stable/2529886>

Dawes RM (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, 26, 422-424. DOI: 10.1037/h0044612

- de Cani JS (1984). Balancing Type I risk and loss of power in ordered Bonferroni procedures. *Journal of Educational Psychology*, 6, 1035-1037. DOI: 10.1037/0022-0663.76.6.1035
- Disney RL (1971). Probability and stochastic processes. In: H.B. Maynard (Ed.), *Industrial engineering handbook*. New York: McGraw-Hill (ps. 10.32-10.51).
- Donenberg GR, Bryant FB, Emerson E, Wilson HW, Pasch KE (2003). Tracing the roots of early sexual debut among adolescents in psychiatric care. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42, 594-608. DOI: 10.1097/01.CHI.0000046833.09750.91
- Dowdney L, Rogers C, Dunn G (1993). Influences upon attendance at out-patient facilities—the contribution of linear-logistic modeling. *Psychological Medicine*, 23, 195-201. DOI: 10.1037/0033-295X.100.1.149
- Dries SG, Dott RH (1984). Model for sandstone-carbonate "cyclothem" based on Upper Member of Morgan Formation (Middle Pennsylvanian) of northern Utah and Colorado. *The American Association of Petroleum Geologists Bulletin*, 68, 574-597.
- Dunn OJ, Vardy PD (1966). Probabilities of correct classification in discriminant analysis. *Biometrics*, 22, 908-924. DOI: 10.2307/2528081
- Ebel RL (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Efron B, Gong G (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37, 36-48. DOI: 10.1080/00031305.1983.10483087
- Eisenbeis RA (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, 32, 875-900. DOI: 10.1111/j.1540-6261.1977.tb01995.x
- Elling KA (2000). *Predicting children's emotional responsiveness during therapy sessions*. Doctoral dissertation, Loyola University Chicago (119 pp).
- Everett JE (1990). Discrimination measure using contingency tables. *Multivariate Behavioral Research*, 25, 371-386. DOI: 10.1207/s15327906mbr2503_8
- Fagerland MW, Lydersen S, Laake P (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13: 91. DOI: 10.1186/1471-2288-13-91
- Feinglass J, Yarnold PR, Martin GJ, McCarthy WJ (1998). A classification tree analysis of selection for discretionary treatment. *Medical Care*, 36, 740-747. URL: <http://www.jstor.org/stable/3767410>
- Feingold M (1992). The equivalence of Cohen's kappa and Pearson's chi-square statistics in the 2 x 2 table. *Educational and Psychological Measurement*, 52, 57-61. DOI: 10.1177/001316449205200105
- Feinstein AR (1988). Statistical significance versus clinical importance. *Quality of Life and Cardiovascular Care*, 4, 99-102.
- Finn MA, Stalans LJ (1996). Police referrals to shelter and mental health treatment: Examining their decisions in domestic assault cases. *Crime and Delinquency*, 41, 467-480. DOI: 10.1177/001128795041004006
- Fisher RA (1950). *Statistical methods for research workers* (11th Ed., revised). London, UK: Oliver and Boyd. Ltd.
- Fitzcharles MA, Yunus MB (2012). The clinical concept of fibromyalgia as a changing paradigm in the past 20 years. *Pain Research and Treatment*, 2012, article ID 184835. DOI: 10.1155/2012/184835
- Fleiss JL (1986). *The design and analysis of clinical experiments*. New York: Wiley.
- Flexser AJ, Tulving E (1993). Recognition-failure constraints and the average maximum. *Psychological Review*, 100, 149-153. DOI: 10.1037/0033-295X.100.1.149

- Foa U (1971). Interpersonal and economic resources. *Science*, 171, 345-351. DOI: 10.1126/science.171.3969.345
- Fors EA, Sexton H (2002). Weather and the pain in fibromyalgia: are they related? *Annals of the Rheumatic Diseases*, 61, 247-250. DOI: 10.1136/ard.61.3.247
- Frank RE, Massy WF, Morrison GD (1965). Bias in multiple discriminant analysis. *Journal of Marketing Research*, 2, 250-258. URL: <http://www.jstor.org/stable/3150183>
- Friedman GD (1987). *Primer of epidemiology* (3rd ed.). New York: McGraw-Hill.
- Friedman M, Rosenman RH (1974). *Type A behavior and your heart*. New York: Knopf.
- Fukunaga K, Kessell DL (1971). Estimation of classification error. *IEEE Transactions on Computers*, 20, 1521-1527. DOI: <http://dx.doi.org/10.1109/T-C.1971.223165>
- Gehrlein WV (1986). General mathematical programming formulations for the statistical classification problem. *Operations Research Letters*, 5, 299-304. DOI: 10.1016/0167-6377(86)90068-4
- Geisser S (1975). The predictive sample Reuse method with applications. *Journal of the American Statistical Association*, 70, 320-328. DOI: 10.1080/01621459.1975.10479865
- Ghiselli EE (1964). *Theory of psychological measurement*. New York: McGraw-Hill.
- Gilbert N (1993). *Analyzing tabular data: Log linear and logistic models for social researchers*. London: University of London College Press.
- Glass DC (1977). *Behavior patterns, stress, and coronary disease*. Hillsdale, NJ: Erlbaum.
- Goodman LA (1962). Statistical methods for analyzing processes of change. *American Journal of Sociology*, 68, 57-78. URL: <http://www.jstor.org/stable/2774180>
- Goodman LA (1968). The analysis of cross-classified data: Independence, quasi-independence, and interaction in contingency tables with or without missing cells. *Journal of the American Statistical Association*, 63, 1091-1131. DOI: 10.1080/01621459.1968.10480916
- Goodman SN, Royall R (1988). Evidence and scientific research. *American Journal of Public Health*, 78, 1568-1574. DOI: 10.2105/AJPH.78.12.1568
- Grahame R (2001). Time to take hypermobility seriously (in adults and children). *Rheumatology*, 40, 485-487. DOI: 10.1093/rheumatology/40.5.485
- Green D, Hartwig D, Chen D, Solysik RC, Yarnold PR (2003). Spinal cord injury risk assessment for thromboembolism (SPIRATE study). *American Journal of Physical and Medical Rehabilitation*, 82, 950-956. URL: http://journals.lww.com/ajpmr/Abstract/2003/12000/Spinal_Cord_Injury_Risk_Assessment_for_7.aspx
- Green DM, Swets JA (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Green MA (1988). Evaluating the discriminatory power of a multiple regression model. *Statistics in Medicine*, 7, 519-524. DOI: 10.1002/sim.4780070408
- Green PE (1978). *Analyzing multivariate data*. Hillsdale, IL: Dryden.
- Greenblatt RL, Mozdzierz GJ, Murphry TJ, Trimakas K (1992). A comparison of non-adjusted and bootstrapped methods: Bootstrapped diagnosis might be worth the trouble. *Educational and Psychological Measurement*, 52, 181-187. DOI: 10.1177/001316449205200123
- Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, DC: APA Books.
- Grimm LG, Yarnold PR (2000). *Reading and understanding more multivariate statistics*. Washington, DC: APA Books.

- Grobman WA, Terkildsen MF, Soltysik RC, Yarnold PR (2008). Predicting outcome after emergent cerclage using classification tree analysis. *American Journal of Perinatology*, 25, 443-448. DOI: 10.1055/s-0028-1083843
- Guedj D, Weinberger A (1990). Effect of weather conditions on rheumatic patients. *Annals of the Rheumatic Diseases*, 49, 158-159. DOI: 10.1136/ard.49.3.15
- Gulliksen H (1950). *Theory of mental tests*. New York: Wiley.
- Guttman L (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282. DOI: 10.1007/BF02288892
- Guyatt G, Walter S, Norman G (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, 40, 171-178. DOI: 10.1016/0021-9681(87)90069-5
- Haberman SJ (1979). *Analysis of qualitative data, Volume 2: New developments*. New York: Academic Press.
- Hagen RL (1997). In praise of the null hypothesis significance test. *American Psychologist*, 52, 15-24.
DOI: <http://dx.doi.org/10.1037/0003-066X.52.1.15>
- Hagenaars JA (1990). *Categorical longitudinal data*. Newbury Park, CA: Sage.
- Hagglund KJ, Deuser WE, Buckelew SP, Hewett J, Kay DR (1994). Weather, beliefs about weather, and disease severity among patients with fibromyalgia. *Arthritis Care Research*, 7, 130-135. DOI: 10.1002/art.1790070306
- Hagle T, Mitchell G (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36, 762-784. URL: <http://www.jstor.org/stable/2111590>
- Han SD, Suzuki H, Drake AI, Jak AJ, Houston WS, Bondi MW (2009). Clinical, cognitive, and genetic predictors of change in job status following traumatic brain injury in a military population. *Journal of Head Trauma Rehabilitation*, 24, 57-64. DOI: 10.1097/HTR.0b013e3181957055
- Hanley JA, McNeil BJ (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36. DOI: 10.1148/radiology.143.1.7063747
- Hanneman RA (1988). *Computer-assisted theory building: Modeling dynamic social systems*. Newbury Park, CA: Sage.
- Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D. (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke*, 27, 1516-1520. Abstracted in *American College of Physicians Journal Club*, 1997, 126, 43. DOI: 10.1161/01.STR.27.9.1516
- Hawley DJ, Wolfe F, Lue FA, Moldofsky H (2001). Seasonal symptom severity in patients with rheumatic diseases: a study of 1,424 patients. *Journal of Rheumatology*, 28, 1900-1909.
- Hills M (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society*, 28, 1-20. URL: <http://www.jstor.org/stable/2984268>
- Hilmer M, Jung T (2000). Evidence for a recent change in the link between north Atlantic oscillation and Arctic sea ice export. *Geophysical Research Letters*, 27, 989-992. DOI: 10.1029/1999GL010944
- Hintzman DL (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, 87, 398-410. DOI: 10.1037/0033-295X.87.4.398
- Hintzman DL (1993). On variability, Simpson's paradox, and the relation between recognition and recall: Reply to Tulving and Flexser. *Psychological Review*, 100, 143-148. DOI: 10.1037/0033-295X.100.1.143
- Hoffman LAD (2000). *Marital interaction and depression: A test of the interactional systems model of depression*. Doctoral dissertation, Loyola University Chicago (194 pp).

Holland BS, Copenhaver MD (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, 43, 417-423.
URL: <http://www.jstor.org/stable/2531823>

Holland BS, Copenhaver MD (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, 104, 145-149. DOI: 10.1037/0033-2909.104.1.145

Holland MM (2003). The north Atlantic oscillation–Arctic oscillation in the CCSM2 and its influence on Arctic climate variability. *Journal of Climate*, 16, 2767–2781. 10.1175/1520-0442(2003)016%3C2767:TNAOI%3E2.0.CO;2

Hosmer DW, Lemeshow S (1989). *Applied logistic regression*. New York, NY: Wiley.

Hosmer DW, Lemeshow S (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theoretical Methods*, A9, 1043-1069. DOI: 10.1080/03610928008827941

<http://bmdshapi.com/>

<http://chevronotoxico.com/about/environmental-impacts/produced-water>

http://climate.ncsu.edu/images/climate/enso/geo_heights.php

<http://data.giss.nasa.gov/gistemp/>

<http://data.worldbank.org/indicator/SP.DYN.CDRT.IN>

http://en.wikipedia.org/wiki/McNemar%27s_test

<http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf>

<http://ndhealth.gov/vital/stats.htm>

<http://ODAJournal.com>

<http://panko.shidler.hawaii.edu/SSR/>

<http://people.duke.edu/~rnau/testing.htm>

<http://psych.unl.edu/psychrs/handcomp/hccochran.PDF>

http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf

<http://research.stlouisfed.org/fred2/series/UMCSENT/>

<http://rt.com/op-edge/fracking-radioactive-uranium-danger-ecology-057/>

<http://users.sussex.ac.uk/~grahamh/RM1web/MannWhitneyHandout%202011.pdf>

<http://webword.com/moving/cochransq.html>

http://www.aoa.gov/AoARoot/AoA_Programs/HPW/Behavioral/docs2/North%20Dakota.pdf

<http://www.autonlab.org/tutorials/overfit10.pdf>

<http://www.biostathandbook.com/kruskalwallis.html>

<http://www.biostathandbook.com/kruskalwallis.html>

<http://www.cdc.noaa.gov/cgi-bin/Timeseries/timeseries1.pl>

- <http://www.cpc.noaa.gov/data/teledoc/telepatcalc.shtml>
<http://www.econ.upf.edu/~michael/stanford/maeb5.pdf>
<http://www.medicalbiostatistics.com/roccurve.pdf>
<http://www.netl.doe.gov/technologies/pwmis/intropw/>
<http://www.npwrc.usgs.gov/resource/habitat/climate/wind.htm>
<http://www.real-statistics.com/reliability/kendalls-w/>
<http://www.saburchill.com/IBbiology/stats/002.html>
<https://www.medcalc.org/manual/cochranq.php>

Hurley CL (2005). *Medical, demographic, and psychological predictors of morbidity and mortality in autologous bone marrow transplant patients*. Doctoral dissertation, Loyola University Chicago (192 pp).

Hyde JS, Plant EA (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist*, 50, 159-161. DOI: 10.1037/0003-066X.50.3.159

Injuries, illnesses, and fatalities in the coal mining industry (2010). US Bureau of Labor Statistics.

Jacquet J (2009). *Energy boomtowns and natural gas: Implications for Marcellus Shale local governments and rural communities*. NERCRD Rural Development Paper No. 43, Pennsylvania State University.

Jandasek BN (2008). *Predictors of social competence in adolescents with spina bifida*. Doctoral dissertation, Loyola University Chicago (199 pp).

Jenkins CD, Zyzanski SJ, Ryan TJ, Flessas A, Tannenbaum SI (1977). Social insecurity and coronary-prone Type A responses as identifiers of severe atherosclerosis. *Journal of Consulting and Clinical Psychology*, 45, 1060-1067. DOI: 10.1037/0022-006X.45.6.1060

Joachimsthaler EA, Stam A (1990). Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioral Research*, 25, 427-454. DOI: 10.1207/s15327906mbr2504_2

Jones SS, Collins K, Hong HW (1991). An audience effect on smile production in 10-month-old infants. *Psychological Science*, 2, 45-49. URL: <http://www.jstor.org/stable/40062583>

Kahneman D, Slovic P, Tversky A (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge: University Press.

Kalnay E (2002). *Atmospheric modeling, data assimilation and predictability*. Cambridge, UK: Cambridge University Press.

Kanter AS, Spencer DC, Steinberg MH, Solysik RC, Yarnold PR, Graham NM (1999). Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. *Journal of Acquired Immune Deficiency Syndrome*, 21, 252-253. URL: http://journals.lww.com/jaids/Citation/1999/07010/Supplemental_Vitamin_B_and_Progression_to_AIDS_and.11.aspx

Kapunga CT (2006). *Individual, parental and peer influences associated with risky sexual behaviors among African-American adolescents*. Doctoral dissertation, Loyola University Chicago (125 pp).

Karmarkar N (1984). A new polynomial time algorithm for linear programming. *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, 302-311. URL: <http://dl.acm.org/citation.cfm?id=808695>

Kazdin AE (1992). *Research design in clinical psychology* (2nd Ed.). Boston: Allyn & Bacon

- Kemeny JG, Snell JL (1976). *Finite Markov chains*. New York: Springer.
- Kendall M (1975). *Multivariate Analyses*. New York: Hafner (Chapter 7).
- Kendall MG, Babington SB (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10, 275–287.
URL: <http://www.jstor.org/stable/2235668>
- Kim B, Lyons TM, Parada JP, Uphold CR, Yarnold PR, Hounshell JB, Sipler AM, Goetz MB, DeHovitz JA, Weinstein RA, Campo RE, Bennett CL (2001). HIV-related *Pneumocystis carinii* pneumonia in older patients hospitalized in the early HAART era. *Journal of General Internal Medicine*, 16, 583-589. DOI: 10.1046/j.1525-1497.2001.016009583.x
- Kleinbaum DG, Kupper LL, Muller KE (1988). *Applied regression analysis and other multivariable methods* (2nd Ed.). Boston, MA: PWS-Kent.
- Kline, RB. (2011). *Principles and practice of structural equation modeling* (3rd ed). New York: Guilford Press.
- Koehler GJ, Erenguc SS (1990). Minimizing misclassifications in linear discriminant analysis. *Decision Sciences*, 21, 63-74. DOI: 10.1111/j.1540-5915.1990.tb00317.x
- Kolb DA (1984). *Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice Hall.
- Kraemer HC (1992). *Evaluating medical tests*. Newbury Park, CA: Sage.
- Kruskal JB (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25, 201-237. DOI: 10.1137/1025045
- Kruskal W, Wallis WA (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583–621. DOI: 10.1080/01621459.1952.10483441
- Kshirsagar AM (1972). *Multivariate analysis*. New York: Dekker.
- Kucera CM, Greenberger PA, Yarnold PR, Choy AC, Levenson T (1999). An attempted prospective testing of an asthma severity index and a quality of life survey for 1 year in ambulatory patients with asthma. *Allergy and Asthma Proceedings*, 20, 29-38. DOI: 10.2500/108854199778681521
- Kuder GF, Richardson MW (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160. DOI: 10.1007/BF02288391
- Kwoh CK, O'Connor GT, Regan-Smith MG, Olmstead EM, Brown LA, Burnett JB, Hochman RF, King K, Morgan GJ (1992). Concordance between clinician and patient assessment of physical and mental health status. *Journal of Rheumatology*, 19, 1031-1037.
- Kwok R, Cunningham GF, Pang SS (2004). Fram Strait sea ice outflow. *Journal of Geophysical Research*, 109, 1009-1029. DOI: 10.1029/2003JC001785
- Kyriacou DM, Yarnold PR, Solysik RC, Wunderink RG, Schmitt BP, Parada JP, Adams JG (2008). Derivation of a triage algorithm for chest radiography of community-acquired pneumonia in the emergency department. *Academic Emergency Medicine*, 15, 40-44. DOI: 10.1111/j.1553-2712.2007.00011.x
- Kyriacou DN, Yarnold PR, Stein AC, Schmitt BP, Solysik RC, Nelson RR, Frerichs RR, Noskin GA, Belknap SB, Bennett CL (2007). Discriminating inhalational anthrax from community-acquired pneumonia using chest radiograph findings and a clinical algorithm. *Chest*, 131, 489-495. DOI: 10.1378/chest.06-1687
- Lachenbruch PA (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23, 639-645. URL: <http://www.jstor.org/stable/2528418>
- Lachenbruch PA (1975). *Discriminant analysis*. New York: Hafner.

Lachenbruch PA, Mickey MR (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11. DOI: 10.1080/00401706.1968.10490530

Laforce M (2006). *A classification profile of high-risk sexual behavior among men who have sex with men*. Doctoral dissertation, Loyola University Chicago (94 pp).

Lamiell JT (1981). Toward an ideothetic psychology of personality. *American Psychologist*, 36, 276-289. DOI: 10.1037/0003-066X.36.3.276

Larichev OI, Olson DL, Moshkovich HM, Mechitov AJ (1995). Numerical vs cardinal measurements in multiattribute decision making: How exact is enough? *Organizational Behavior and Human Decision Processes*, 64, 9-21. DOI: 10.1006/obhd.1995.1085

Layden BL, Minadeo N, Suhy J, Metreger T, Foley K, Borge G, Crayton J, Bryant FB, Mota de Freitas D (2004). Bi-chemical and psychiatric predictors of Li⁺ response and toxicity in Li⁺-treated bipolar patients. *Bipolar Disorders*, 6, 53-61. DOI: 10.1046/j.1399-5618.2003.00093.x

Legendre P (2005). Species associations: The Kendall Coefficient of Concordance revisited. *Journal of Agricultural, Biological and Environmental Statistics*, 10, 226-245. DOI: 10.1198/108571105X46642

Levenson T, Grammer LC, Yarnold PR, Patterson R (1997). Cost-effective management of malignant potentially fatal asthma. *Allergy and Asthma Proceedings*, 18, 73-78. DOI: 10.2500/108854197778605455

Levinson W, Lesser CS, Epstein RM (2015). Developing physician communication skills for patient-centered care. *Health Affairs*, 29, 1310-1318. DOI: 10.1377/hlthaff.2009.0450

Licht MH (1995). Multiple regression and correlation. In: LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 19-64).

Loke WH (1989). Diagnostic evaluations using signal detection analysis. *Indian Journal of Psychological Medicine*, 12, 87-91.

Loo R (1996). Construct validity and classification stability of the revised Learning Style Inventory (LSI-1985). *Educational and Psychological Measurement*, 56, 529-536. DOI: 10.1177/0013164496056003015

Lord FM, Novick MR (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loucopoulos C, Pavur R (1997). Computational characteristics of a new mathematical programming model for the three-group discriminant problem. *Computers & Operations Research*, 24, 179-191. DOI: 10.1016/S0305-0548(96)00046-9

Loza E, Abasolo L, Jover JA, Carmona L (2008). Burden of disease across chronic diseases: A health survey that measured prevalence, function, and quality of life. *The Journal of Rheumatology*, 35, 159-165.

Lyerly R (1958). The Kuder-Richardson formula 21 as a split-half coefficient, and some remarks on its basic assumption. *Psychometrika*, 23, 267-270. DOI: 10.1007/BF02289239

Lyons AM, Leon SC, Zaddach C, Luboyeski EJ, Richards M (2009). Predictors of clinically significant sexual concerns in a child welfare population. *Journal of Child and Adolescent Trauma*, 2, 28-45. DOI: 10.1080/19361520802675884

Magnusson D (1967). *Test theory*. Reading, MA: Addison-Wesley.

Mann HB, Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60. URL: <http://www.jstor.org/stable/2236101>

Martin E (1981). Simpson's paradox revisited: A reply to Hintzman. *Psychological Review*, 88, 372-374.

Martin GJ, Magid NM, Myers G, Barnett PS, Schaad JW, Weiss JS, Lesch M, Singer DH (1987). Heart rate variability and sudden cardiac death secondary to coronary artery disease during ambulatory electrocardiographic monitoring. *American Journal of Cardiology*, 60, 86-89. DOI: 10.1016/0002-9149(87)90990-8

Martinez JE, Cruz CG, Aranda C, Boulos FC, Lagoa LA (2003). Disease perceptions of Brazilian fibromyalgia patients: do they resemble perceptions from other countries? *International Journal of Rehabilitation Research*, 26, 223-227. DOI: 10.1097/00004356-200309000-00010

Maxwell SE, Delaney HD (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

McClish DK (1992). Combining and comparing area estimates across studies or strata. *Medical Decision Making*, 12, 274-279. DOI: 10.1177/0272989X9201200405

McLachlan GJ (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

McNemar Q (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157. DOI: 10.1007/BF02295996

Meehl PE, Rosen A (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194-216. DOI: 10.1037/h0048070

Midgette AS, Stukel TA, Littenberg B (1993). A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Medical Decision Making*, 13, 253-257. DOI: 10.1177/0272989X9301300313

Mielke PW (1984). Meteorological applications of permutation techniques based on distance functions. In P.R. Krishnaiah & P.K. Sen (Eds.), *Handbook of statistics, Volume 4: Nonparametric methods*. New York: North-Holland.

Mielke PW (1991). The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Science Reviews*, 31, 55-71. DOI: 10.1175/1520-0493(1981)109<0120:AOMRPP>2.0.CO;2

Miranda LC, Parente M, Silva C, Clemente-Coelho P, Santos H, Cortes S, Medeiros D, Ribeiro JS, Barcelos F, Sousa M, Miguel C, Figueiredo R, Mediavilla M, Simoes E, Silva M, Patto JV, Madeira H, Ferreira J, Micaelo M, Leitao R, Las V, Faustino A, Teixeira A (2007). Perceived pain and weather changes in rheumatic patients. *Acta Reumatologica Portuguesa*, 32, 351-361.

Mirhaghi A, Heydari A, Mazlom R, Hasanzadeh F (2015). Reliability of the Emergency Severity Index: Meta-Analysis. *Sultan Qaboos University Medical Journal*, 15, e71-77.

Mosteller F (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63, 1-28. DOI: 10.1080/01621459.1968.11009219

Mueser KT, Sayers SL, Schooler NR, Mance RM, Haas GL (1993). A multisite investigation of the reliability of the Scale for the Assessment of Negative Symptoms. *The American Journal of Psychiatry*, 151, 1453-1462. DOI: 10.1176/ajp.151.10.1453

Mueser KT, Yarnold PR, Foy DW (1991). Statistical analysis for single-case designs: Evaluating outcomes of imaginal exposure treatment of chronic PTSD. *Behavior Modification*, 15, 134-155. DOI: 10.1177/01454455910152002

Mueser KT, Yarnold PR, Rosenberg SD, Drake RE, Swett C, Miles KM, Hill D (2000). Substance use disorder in hospitalized severely mentally ill psychiatric patients: Prevalence, correlates, and sub-groups. *Schizophrenia Bulletin*, 26, 179-193. URL: <http://psycnet.apa.org/journals/szb/26/1/179/>

Nanna MJ, Sawilowsky SS (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, 3, 55-67. DOI: 10.1037/1082-989X.3.1.55

nces.ed.gov/pubs2009/2009081.pdf

Nebeker JR, Yarnold PR, Solysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Wu X (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care*, 45, S81-S88. DOI: 10.1097/MLR.0b013e3180616c2c

- Newcombe RG (2005). Confidence intervals for an effect size measure based on the Mann-Whitney statistic, Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25, 543-557. DOI: 10.1002/sim.2323
- Nightingale SD, Yarnold PR, Greenberg MS (1991). Sympathy, empathy, and physician resource utilization. *Journal of General Internal Medicine*, 6, 420-423. DOI: 10.1007/BF02598163
- Nishikawa K, Kubota Y, Ooi T (1983). Classification of proteins into groups based on amino acid composition and other characters, II: Grouping into four types. *Journal of Biochemistry*, 94, 997-1007.
- Noreen EW (1989). *Computer-intensive methods for testing hypotheses: An introduction*. New York: Wiley.
- Nunnally JC (1978). *Psychometric theory* (2nd Ed.). New York: McGraw-Hill.
- Odeh RE, Evans JO (1974). Algorithm AS 70: The percentage points of the normal distribution. *Applied Statistics*, 23: 96-97. DOI: 10.2307/2347061
- Olsson U (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460. DOI: 10.1007/BF02296207
- Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*, 66, 660-672. DOI: 10.1037/0022-006X.66.4.660
- Page CV (1977). Heuristics for signature table analysis as a pattern recognition technique. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7, 77-86. DOI: 10.1109/TSMC.1977.4309658
- Parshall CG, Kromrey JD (1996). Tests of independence in contingency tables with small samples: A comparison of statistical power. *Educational and Psychological Measurement*, 56, 26-44. DOI: 10.1177/0013164496056001002
- Parzen E (1962). *Stochastic processes*. San Francisco: Holden-Day.
- Pearson K (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157-175. DOI: 10.1080/14786440009463897
- Pedhazur EJ (1982). *Multiple regression in behavioral research* (2nd Ed.). New York, NY: Holt, Rinehart and Winston.
- Pfetsch ME (2008). Branch-and-cut for the maximum feasible subsystem problem. *SIAM Journal on Optimization*, 19, 21-38. URL: <http://dx.doi.org/10.1137/050645828>
- Philpott L (2008). *The complete handbook of fishing knots, leaders, and lines*. New York: Skyhorse Publishing.
- Podolsky A, Stern DT, Peccoraro L (2015). The courteous consult: A CONSULT card and training to improve resident consults. *Journal of Graduate Medical Education*, 7, 113-117. DOI: 10.4300/JGME-D-14-00207.1
- Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine*, 9, 1103-1115. DOI: 10.1002/sim.4780090917
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989). *Numerical recipes: The art of scientific computing*. Cambridge: University Press.
- Preuss L, Vorkauf H (1997). The knowledge content of statistical data. *Psychometrika*, 62, 133-161. DOI: 10.1007/BF02294784
- Raaflaub KA, Talbert RJA (2009). *Geography and ethnography: Perceptions of the world in pre-modern societies*. New York: Wiley.

- Raatikka VP, Rytkenen M, Nayha S, Hassi J (2007). Prevalence of cold-related complaints, symptoms and injuries in the general population: The INRISK 2002 cold substudy. *International Journal of Biometeorology*, 51, 441-448. DOI: 10.1007/s00484-006-0076-1
- Rau JG (1970). *Optimization and probability in systems engineering*. New York: Van Nostrand.
- Raush HL (1965). Interaction sequences. *Journal of Personality and Social Psychology*, 2, 487-499. DOI: 10.1037/h0022478
- Reynolds HT (1977). *The analysis of cross-classifications*. New York: Free Press.
- Robins JM, Hernan MA, Brumback B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560. URL: <http://www.jstor.org/stable/3703997?origin=JSTOR-pdf>
- Rodriguez JD, Perez A, Lozano JA (2009). Sensitivity analysis of K-Fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 569-575. DOI: 10.1109/TPAMI.2009.187
- Rogers JH, Widiger TA (1989). Comparing ideothetic, ipsative, and normative indices of consistency. *Journal of Personality*, 57, 847-869. DOI: 10.1111/j.1467-6494.1989.tb00497.x
- Rorer LG, Dawes RM (1982). A base-rate bootstrap. *Journal of Consulting and Clinical Psychology*, 50, 419-425. DOI: 10.1037/0022-006X.50.3.419
- Rosenberg M (1965). *Society and the adolescent self-image*. Princeton NJ, Princeton University Press.
- Rosengren A, Welin L, Tsipogianni A, Wilhelmsen L (1989). Impact of cardiovascular risk factors on coronary heart disease and mortality among middle aged diabetic men: A general population study. *British Medical Journal*, 299, 1127-1131. DOI: 10.1136/bmj.299.6708.1127
- Rosenthal R (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- Rosenthal R, Rubin DB (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*, 6, 1028-1034. DOI: 10.1037//0022-0663.76.6.1028
- Rosner B (1982). *Fundamentals of biostatistics*. Boston: Duxbury.
- Royeen CB (1989). *Clinical research handbook: An analysis for the service professions*. Thorofare, NJ: SLACK.
- Rubin PA (1990). Heuristic solution procedures for a mixed-integer programming discriminant model. *Mangerial and Decision Economics*, 11, 255-266. DOI: 10.1002/mde.4090110407
- Rubin PA (1992). *Mathematical programming and alternative classification models*. Invited address presented at the TIMS/ORSA Joint National Meetings, Orlando, FL.
- Rubin PA (1997). Solving mixed-integer classification problems by decomposition. *Annals of Operations Research*, 74, 51-64. DOI: 10.1023/A:1018990909155
- Rubin PA (1999). Adapting the Warmack-Gonzalez algorithm to handle discrete data. *European Journal of Operational Research*, 113, 632-642. DOI: 10.1016/S0377-2217(97)00448-7
- Rulon PJ (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Education Review*, 9, 99-103.
- Russell RL, Bryant FB, Estrada AU (1996). Confirmatory P-technique analyses of therapist discourse: High- versus low-quality child therapy sessions. *Journal of Consulting and Clinical Psychology*, 64, 1366-1376. DOI: 10.1037/0022-006X.64.6.1366
- Ryan TA (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26-47. DOI: 10.1037/h0042478

- Ryan TA (1985). "Ensemble-adjusted" *p* values: How are they to be weighted? *Psychological Bulletin*, 97, 521-526. DOI: 10.1037/0033-2909.97.3.521
- Saal FE, Downey RG, Lahey MA (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413-428. DOI: 10.1037/0033-2909.88.2.413
- Sayah A, Rogers L, Devarajan K, Kingsley-Rocker L, Lobon LF (2014). Minimizing ED waiting times and improving patient flow and experience of care. *Emergency Medicine International*, 1984, article ID 981472. DOI: 10.1155/2014/981472
- Scheier MF, Carver CS (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247. DOI: 10.1037/0278-6133.4.3.219
- Seaman MA, Hill CC (1996). Pairwise comparisons for proportions: a note on Cox and Key. *Educational and Psychological Measurement*, 56, 452-459. DOI: 10.1177/0013164496056003007
- Shanker R (1994). *Principles of quantum mechanics*, 2nd edition. New York, NY: Springer.
- Shea JA, Norcini JJ, Baranowski RA, Langdon LO, Poop RL (1992). A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. *Evaluation in the Health Professions*, 15, 325-340. DOI: 10.1177/016327879201500305
- Silva APD, Stam A (1995). Discriminant analysis. In LG Grimm, PR Yarnold (Eds.), *Reading and understanding multivariate statistics*. Washington DC: APA Books (pp. 277-318).
- Silva APD, Stam A (1997). A mixed-integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. *Annals of Operations Research*, 74, 129-157. DOI: 10.1023/A:1018962102794
- Simpson EH (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, B*, 13, 238-241. URL: <http://www.jstor.org/stable/2984065>
- Sieracki JH, Fuller AK, Leon SC, Jhe Bai G, Bryant FB (2015). The role of race, socioeconomic status, and System of Care services in placement decision-making. *Children and Youth Services Review*, DOI: 10.1016/j.childyouth.2014.12.013
- Smart CM, Nelson NW, Sweet JJ, Bryant FB, Berry DTR, Granacher RP, Heilbronner RL (2008). Use of MMPI-2 to predict cognitive effort: A hierarchically optimal classification tree analysis. *Journal of the International Neuropsychological Society*, 14, 842-852. DOI: 10.1017/S1355617708081034
- Smith JL, Bryant FB (2013). Are we having fun yet? Savoring, Type A behavior, and vacation enjoyment. *International Journal of Well-Being*, 3, 1-19. DOI: 10.1145/10.5502/ijw.v3i1.1
- Smith JH, Bryant FB, Njus D, Posavac EJ (2010). Here today, gone tomorrow: Understanding freshman attrition using Person-Environment Fit Theory. *Optimal Data Analysis*, 1, 101-124. URL: <http://optimalprediction.com/files/pdf/V1A14.pdf>
- Smith JH, Bryant FB, Njus D, Posavac EJ (2010). Here today, gone tomorrow: Understanding freshman attrition using person-environment fit theory. *Optimal Data Analysis*, 1, 101-124. URL: <http://optimalprediction.com/files/pdf/V1A14.pdf>
- Smithson M (1987). *Fuzzy set analysis for behavioral and social sciences*. New York: Springer-Verlag.
- Snowden J (2008). *Predictors of stepping-up from foster homes to residential treatment: A profile of children in the child welfare system*. Doctoral dissertation, Loyola University Chicago (136 pp).
- Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Clinical Child and Adolescent Psychology*, 36, 8-18. DOI: 10.1080/15374410709336564

- Snowden J, Leon S, Sieracki J (2008). Predictors of children in foster care being adopted: A classification tree analysis. *Children and Youth Services Review*, 30, 1318-1327. DOI: 10.1016/j.childyouth.2008.03.014
- Snyder DK, Wills RM, Grady-Fletcher A (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: a four-year follow up study. *Journal of Consulting and Clinical Psychology*, 59, 138-141. DOI: 10.1037/0022-006X.59.1.138
- Soeken KL, Prescott PA (1986). Issues in the use of Kappa to estimate reliability. *Medical Care*, 24, 733-741. URL: <http://www.jstor.org/stable/3765100>
- Sokal RR, Sneath PHA (1963). *Principles of numerical taxonomy*. San Francisco: Freeman.
- Soltysik RC, Yarnold PR (1993). *ODA 1.0: Optimal Data Analysis for DOS*. Chicago: Optimal Data Analysis, Inc.
- Soltysik RC, Yarnold PR (1994). The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research*, 21, 735-745. DOI: 10.1016/0305-0548(94)90003-5
- Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: One-tailed hypotheses. *Educational and Psychological Measurement*, 54, 646-653. DOI: 10.1177/0013164494054003007
- Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis*, 1, 144-160. URL: <http://odajournal.com/2013/09/19/62/>
- Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, 1, 67-100. URL: <http://optimalprediction.com/files/pdf/V2A33.pdf>
- Soltysik RC, Yarnold PR (2010). Two-group MultiODA: a mixed-integer programming solution with bounded M . *Optimal Data Analysis*, 1, 30-37. URL: <http://optimalprediction.com/files/pdf/V1A4.pdf>
- Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, 2, 194-197. URL: <http://optimalprediction.com/files/pdf/V2A29.pdf>
- Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis*, 2, 202-205. URL: <http://optimalprediction.com/files/pdf/V2A31.pdf>
- Soltysik RC, Yarnold PR (2013). Statistical power of optimal discrimination with one attribute and two classes: One-tailed hypotheses. *Optimal Data Analysis*, 2, 26-30. URL: <http://optimalprediction.com/files/pdf/V2A4.pdf>
- Soltysik RC, Yarnold PR (2014). Hierarchically optimal classification tree analysis of adverse drug reactions secondary to warfarin therapy. *Optimal Data Analysis*, 3, 23-24. URL: <http://optimalprediction.com/files/pdf/V3A9.pdf>
- Sorum M (1972). Three probabilities of misclassification. *Technometrics*, 14, 309-316. DOI: 10.1080/00401706.1972.10488917
- Spearman C (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295. DOI: 10.1111/j.2044-8295.1910.tb00206.x
- Stalans LS, Finn MA (1995). How novice and experienced officers interpret wife assaults: Normative and efficiency frames. *Law and Society Review*, 29, 301-335. DOI: 10.2307/3054013
- Stalans LJ, Hacker R, Talbot ME (2010). Comparing nonviolent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behavior*, 37, 613-628. DOI: 10.1177/0093854810363794
- Stalans LJ, Seng M (2006). Identifying subgroups at high risk of dropping out of domestic batterer treatment: The buffering effects of a high school education. *International Journal of Offender Therapy and Comparative Criminology*, 10, 1-19. DOI: 10.1177/0306624X06290204

Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law & Human Behavior*, 28, 53-271. DOI: 10.1023/B:LAHU.0000029138.92866.af

Stam A, Joachimsthaler EA (1990). A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research*, 46, 113-122. DOI: 10.1016/0377-2217(90)90304-T

Stevens J (1992). *Applied multivariate statistics for the social sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Stone M (1974). Cross-validatory choice and assessment of statistical problems. *Journal of the Royal Statistical Society*, 36, 111-147. URL: <http://www.jstor.org/stable/2984809>

Stoner AM, Leon SC, Fuller AK (2013). Predictors of reduction in symptoms of depression for children and adolescents in foster care. *Journal of Child and Family Studies*, 22, DOI 10.1007/s10826-013-9889-9

Strusberg I, Mendelberg RC, Serra HA, Strusberg AM (2002). Influence of weather conditions on rheumatic pain. *Journal of Rheumatology*, 29, 335-338.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, based on the November 2011 submission.

Suzuki H (2005). *Prospectively tracing profiles of juvenile delinquents and non-delinquents: An optimal data analysis*. Master's thesis, Loyola University Chicago (87 pp).

Suzuki H, Bryant FB, Edwards JD (2010). Tracing prospective profiles of juvenile delinquency: An optimal classification tree analysis. *Optimal Data Analysis*, 1, 125-143. URL: <http://optimalprediction.com/files/pdf/V1A15.pdf>

Swets JA (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47, 522-532. DOI: 10.1037/0003-066X.47.4.522

Tabachnick BG, Fidell LS (1983). *Using multivariate statistics*. New York, NY: Harper and Row.

Tallis GM (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18, 342-353. URL: <http://www.jstor.org/stable/2527476>

Thompson DA, Yarnold PR (1995). Relating patient satisfaction to waiting time perceptions and expectations: The disconfirmation paradigm. *Academic Emergency Medicine*, 2, 1057-1062. DOI: 10.1111/j.1553-2712.1995.tb03150.x

Thompson DA, Yarnold PR, Adams SL, Spacone AB (1996). How accurate are patient's waiting time perceptions? *Annals of Emergency Medicine*, 28, 652-656. DOI: 10.1016/S0196-0644(96)70089-6

Thompson DA, Yarnold PR, Williams DR, Adams SL (1996). The effects of actual waiting time, perceived waiting time, information delivery and expressive quality on patient satisfaction in the emergency department. *Annals of Emergency Medicine*, 28, 657-665. DOI: 10.1016/S0196-0644(96)70090-2

Toh RS, Hu MY (2008). Averaging to minimize or eliminate regression toward the mean to measure pure experimental effects. *Psychological Reports*, 102, 665-677. DOI: 10.2466/pr0.102.3.665-677

Trafimow D, Marks M (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1-2. DOI: 10.1080/01973533.2015.1012991

Triantaphyllou E, Shu B, Sanchez N, Ray T (1998). Multi-criteria decision making: An operations research approach. *Encyclopedia of Electrical and Electronics Engineering*, 15, 175-186.

Troccoli A (2008). Management of weather and climate risk in the energy industry. Proceedings of the NATO advanced research workshop on weather/climate risk management for the energy sector, Santa Maria di Leuca, Italy

6-10 October 2008. Series: *NATO science for peace and security series C: environmental security*. New York, NY: Springer.

Uebersax JS (2006). The tetrachoric and polychoric correlation coefficients. *Statistical Methods for Rater Agreement* web site: <http://john-uebersax.com/stat/tetra.htm>.

Wainer H (1991). Adjusting for differential base rates: Lord's Paradox again. *Psychological Bulletin*, 109, 147-151. DOI: 10.1037/0033-2909.109.1.147

Weinfurt KP, Bryant FB, Yarnold PR (1994). The factor structure of the Affect Intensity Measure: In search of a measurement model. *Journal of Research in Personality*, 28, 314-331. DOI: 10.1006/jrpe.1994.1023

Welch SJ (2013). Twenty years of patient satisfaction research applied to the Emergency Department: A qualitative review. *American Journal of Medical Quality*, 25, 64-72. DOI: 10.1177/1062860609352536

Widiger TA (1983). Utilities and fixed rules: Comments on Finn. *Journal of Abnormal Psychology*, 92, 495-498. DOI: 10.1037/0021-843X.92.4.495

Wilde MM (2013). *Quantum information theory*. Cambridge, UK: Cambridge University Press.

Willoughby HE, Rappaport EN, Marks FD (2007). Hurricane forecasting: the state of the art. *Natural Hazards Review*, 8, 45-49. DOI: 10.1061/(ASCE)1527-6988(2007)8:3(45)

Witten IH, Frank E, Hall MA (2011). *Data Mining: Practical Machine Learning Tools and Technique* (3rd Ed.). San Francisco, CA: Morgan Kaufmann.

Wolf JL (2005). *A meta-analysis of primary preventive interventions targeting the mental health of children and adolescents: A review spanning 1992–2003*. Doctoral dissertation, Loyola University Chicago (128 pp).

Woolson RF (1987). *Statistical methods for the analysis of biomedical data*. New York: Wiley.

Worster A, Gilboy N, Fernandes CM, Eitel D, Eva K, Gleister R, Tanabe P (2004). Assessment of inter-observer reliability of two five-level triage and acuity scales: A randomized controlled trial. *Canadian Journal of Emergency Medicine*, 6, 240-245. DOI: 10.1017/S1481803500009192

Wright RE (2005). Logistic Regression. In: LG Grimm, PR Yarnold (eds.), *Reading and understanding multivariate statistics*. Washington, DC: APA Books (pp. 217-244).

Yarnold BM (1990). *Refugees without refuge: Formation and failed implementation of U.S. political asylum policy in the 1980s*. Lanham, MD: University Press of America.

Yarnold BM (1990). Federal court outcomes in asylum-related appeals 1980-1987: A highly politicized process. *Policy Sciences*, 23, 291-306. DOI: 10.1007/BF00141323

Yarnold BM (1990). The Refugee Act of 1980 and de-politicization of refugee/asylum admissions: Failed policy implementation. *American Politics Quarterly*, 18, 527-536. DOI: 10.1177/1532673X9001800408

Yarnold BM (1991). *International fugitives: A new role for the International Court of Justice*. New York, NY: Praeger.

Yarnold BM (1992). *Politics and the courts: Toward a general theory of public law*. New York, NY: Praeger.

Yarnold BM (1993). *Abortion politics in the federal courts: Right versus right*. New York, NY: Paragon.

Yarnold BM, Yarnold PR (2010). Maximizing the accuracy of probit models via UniODA. *Optimal Data Analysis*, 1, 41-42. URL: <http://optimalprediction.com/files/pdf/V1A6.pdf>

Yarnold JK (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, 65, 864-886. URL: <http://www.jstor.org/stable/2284594>

Yarnold PR (1982). On comparing interscale difference scores within a profile. *Educational and Psychological Measurement*, 42, 1037-1045. DOI: 10.1177/001316448204200410

Yarnold PR (1984). Note on the multidisciplinary scope of psychological androgyny theory. *Psychological Reports*, 55, 936-938. DOI: 10.2466/pr0.1984.54.3.936

Yarnold PR (1984). The reliability of a profile. *Educational and Psychological Measurement*, 44, 49-59. DOI: 10.1177/0013164484441005

Yarnold PR (1987). Norms for the Glass model of the short student version of the Jenkins Activity Survey. *Social and Behavioral Science Documents*, 16, 60. MS# 2777.

Yarnold PR (1988). Classical test theory methods for repeated-measures N=1 research designs. *Educational and Psychological Measurement*, 48, 913-919. DOI: 10.1177/00131644848484006

Yarnold PR (1990). Androgyny and sex-typing as continuous independent factors, and a glimpse of the future. *Multivariate Behavioral Research*, 25, 407-419. DOI: 10.1207/s15327906mbr2503_10

Yarnold PR (1992). Statistical analysis for single-case designs. In: F.B. Bryant, L. Heath, E. Posavac, J. Edwards, S. Tindale, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Social psychological applications to social issues, Volume 2: Methodological issues in applied social research*. New York: Plenum (pp. 177-197).

Yarnold PR (1993). A brief measure of psychological androgyny for use in predicting physicians' decision making. *Academic Medicine*, 68, 312. DOI: 10.1097/00001888-199304000-00027

Yarnold PR (1994). Comparing the split-half reliability of androgyny and sex-typing measures. *Australian Journal of Psychology*, 46, 164-169. DOI: 10.1080/00049539408259491

Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, 56, 430-442. DOI: 10.1177/0013164496056003005

Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, 56, 656-667. DOI: 10.1177/0013164496056004007

Yarnold PR (2010). Aggregated vs. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis*, 1, 46-49. URL: <http://optimalprediction.com/files/pdf/V1A8.pdf>

Yarnold PR (2010). UniODA vs. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis*, 1, 62-65. URL: <http://optimalprediction.com/files/pdf/V1A12.pdf>

Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, 2, 172-176. URL: <http://optimalprediction.com/files/pdf/V2A26.pdf>

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. URL: <http://optimalprediction.com/files/pdf/V2A25.pdf>

Yarnold PR (2013). Assessing hold-out validity of CTA models using UniODA. *Optimal Data Analysis*, 2, 31-36. URL: <http://optimalprediction.com/files/pdf/V2A5.pdf>

Yarnold PR (2013). Assessing technician, nurse, and doctor ratings as predictors of overall satisfaction ratings of Emergency Room patients: A maximum-accuracy multiple regression analysis. *Optimal Data Analysis*, 2, 76-85. URL: <http://optimalprediction.com/files/pdf/V2A15.pdf>

Yarnold PR (2013). Comparing attributes measured with “identical” Likert-type scales in single-case designs with UniODA. *Optimal Data Analysis*, 2, 148-153. URL: <http://optimalprediction.com/files/pdf/V2A22.pdf>

Yarnold PR (2013). Creating a data set with SAS™ and maximizing ESS of a multiple regression analysis model for a Likert-type dependent variable using UniODA™ and MegaODA™ software. *Optimal Data Analysis*, 2, 191-193. URL: <http://optimalprediction.com/files/pdf/V2A28.pdf>

Yarnold PR (2013). Maximum-accuracy multiple regression analysis: Influence of registration on overall satisfaction ratings of Emergency Room patients. *Optimal Data Analysis*, 2, 72-74. URL: <http://optimalprediction.com/files/pdf/V2A14.pdf>

Yarnold PR (2013). Minimum standards for reporting UniODA findings. *Optimal Data Analysis*, 2, 63-68. URL: <http://optimalprediction.com/files/pdf/V2A11.pdf>

Yarnold PR (2013). ODA range test vs. one-way analysis of variance: Patient race and lab results. *Optimal Data Analysis*, 2, 206-210. URL: <http://optimalprediction.com/files/pdf/V2A32.pdf>

Yarnold PR (2013). Percent oil-based energy consumption and average percent GDP growth: A small sample UniODA analysis. *Optimal Data Analysis*, 2, 60-61. URL: <http://optimalprediction.com/files/pdf/V2A10.pdf>

Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, 2, 106-119. URL: <http://optimalprediction.com/files/pdf/V2A19.pdf>

Yarnold PR (2013). Statistically significant increases in crude mortality rate of North Dakota counties occurring after massive environmental usage of toxic chemicals and biocides began there in 1998: An optimal static statistical map. *Optimal Data Analysis*, 2, 98-105. URL: <http://optimalprediction.com/files/pdf/V2A18.pdf>

Yarnold PR (2013). Surfing the *Index of Consumer Sentiment*: Identifying statistically significant monthly and yearly changes. *Optimal Data Analysis*, 2, 211-216. URL: <http://optimalprediction.com/files/pdf/V2A33.pdf>

Yarnold PR (2013). The most recent, earliest, and Kth significant changes in an ordered series: Traveling backwards in time to assess when annual crude mortality rate most recently began increasing in McLean County, North Dakota. *Optimal Data Analysis*, 2, 143-147. URL: <http://optimalprediction.com/files/pdf/V2A21.pdf>

Yarnold PR (2013). UniODA and small samples. *Optimal Data Analysis*, 2, 71. URL: <http://optimalprediction.com/files/pdf/V2A13.pdf>

Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: <http://optimalprediction.com/files/pdf/V2A27.pdf>

Yarnold PR (2014). “A statistical guide for the ethically perplexed” (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, 3, 30-31. URL: <http://optimalprediction.com/files/pdf/V3A12.pdf>

Yarnold PR (2014). “Breaking-up” an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, 3, 19. URL: <http://optimalprediction.com/files/pdf/V3A7.pdf>

Yarnold PR (2014). How to assess inter-observer reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, 3, 42-49. URL: <http://optimalprediction.com/files/pdf/V3A15.pdf>

Yarnold PR (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, 3, 50-54. URL: <http://optimalprediction.com/files/pdf/V3A16.pdf>

Yarnold (2014). Illustrating how 95% confidence intervals indicate model redundancy. *Optimal Data Analysis*, 3, 96-97. URL: <http://optimalprediction.com/files/pdf/V3A22.pdf>

Yarnold PR (2014). Increasing the likelihood of an ambivalent patient recommending the Emergency Department to others, *Optimal Data Analysis*, 3, 89-91. URL: <http://optimalprediction.com/files/pdf/V3A20.pdf>

Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis*, 3, 107-109. URL: <http://optimalprediction.com/files/pdf/V3A25.pdf>

Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, 3, 102-106. URL: <http://optimalprediction.com/files/pdf/V3A24.pdf>

Yarnold PR (2014). UniODA vs. Bray-Curtis dissimilarity index for count data. *Optimal Data Analysis*, 3, 115-116. URL: <http://optimalprediction.com/files/pdf/V3A28.pdf>

Yarnold PR (2014). UniODA vs. chi-square: Audience effect on smile production in infants. *Optimal Data Analysis*, 3, 3-5. URL: <http://optimalprediction.com/files/pdf/V3A1.pdf>

Yarnold PR (2014). UniODA vs. chi-square: Discriminating inhibited and uninhibited infant profiles. *Optimal Data Analysis*, 3, 9-11. URL: <http://optimalprediction.com/files/pdf/V3A3.pdf>

Yarnold PR (2014). UniODA vs. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, 3, 14-16. URL: <http://optimalprediction.com/files/pdf/V3A5.pdf>

Yarnold PR (2014). UniODA vs. Kendall's Coefficient of Concordance (W): Multiple rankings of multiple movies. *Optimal Data Analysis*, 3, 121-123. URL: <http://optimalprediction.com/files/pdf/V3A30.pdf>

Yarnold PR (2014). UniODA vs. logistic regression analysis: Serum cholesterol and coronary heart disease and mortality among middle aged diabetic men. *Optimal Data Analysis*, 3, 17-18. URL: <http://optimalprediction.com/files/pdf/V3A6.pdf>

Yarnold, PR (2014). UniODA vs. Mann-Whitney U test: Comparative effectiveness of laxatives. *Optimal Data Analysis*, 4, 6-8. URL: <http://optimalprediction.com/files/pdf/V4A2.pdf>

Yarnold PR (2014). UniODA vs. Mann-Whitney U test: Sunlight and petal width. *Optimal Data Analysis*, 4, 3-5. URL: <http://optimalprediction.com/files/pdf/V4A1.pdf>

Yarnold PR (2014). UniODA vs. polychoric correlation: Number of lambs born over two years. *Optimal Data Analysis*, 3, 113-114. URL: <http://optimalprediction.com/files/pdf/V3A27.pdf>

Yarnold PR (2014). UniODA vs. ROC analysis: Computing the “optimal” cut-point. *Optimal Data Analysis*, 3, 117-120. URL: <http://optimalprediction.com/files/pdf/V3A29.pdf>

Yarnold PR (2014). UniODA vs. t-Test: Comparing two migraine treatments. *Optimal Data Analysis*, 3, 6-8. URL: <http://optimalprediction.com/files/pdf/V3A2.pdf>

Yarnold PR (2014). UniODA vs. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient’s physical and mental health functioning. *Optimal Data Analysis*, 3, 12-13. URL: <http://optimalprediction.com/files/pdf/V3A4.pdf>

Yarnold PR (2015). An example of nonlinear UniODA. *Optimal Data Analysis*, 4, 124-128. URL: <http://optimalprediction.com/files/pdf/V4A24.pdf>

Yarnold PR (2015). Distance from a theoretically ideal statistical classification model defined as the number of additional equivalent effects needed to obtain perfect classification for the sample. *Optimal Data Analysis*, 4, 81-86. URL: <http://optimalprediction.com/files/pdf/V4A15.pdf>

- Yarnold PR (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis*, 4, 21-23. URL: <http://optimalprediction.com/files/pdf/V4A6.pdf>
- Yarnold PR (2015). Evaluating non-confounded association of an attribute and a class variable using partial UniODA. *Optimal Data Analysis*, 4, 32-35. URL: <http://optimalprediction.com/files/pdf/V4A10.pdf>
- Yarnold PR (2015). GO-CTA vs. marginal structural model: Observed data from a point-treatment study, stratified by known confounder. *Optimal Data Analysis*, 4, 104-106. URL: <http://optimalprediction.com/files/pdf/V4A17.pdf>
- Yarnold PR (2015). Maximizing ESS of regression models in applications with dependent measures with domains exceeding ten values. *Optimal Data Analysis*, 4, 12-13. URL: <http://optimalprediction.com/files/pdf/V4A4.pdf>
- Yarnold PR (2015). Optimal statistical analysis involving a confounding variable. *Optimal Data Analysis*, 4, 87-103. URL: <http://optimalprediction.com/files/pdf/V4A16.pdf>
- Yarnold PR (2015). Optimal statistical analysis involving multiple confounding variables. *Optimal Data Analysis*, 4, 107-112. URL: <http://optimalprediction.com/files/pdf/V4A18.pdf>
- Yarnold PR (2015). UniODA vs. Bowker's test for symmetry: Diagnosis before vs. after treatment. *Optimal Data Analysis*, 4, 29-31. URL: <http://optimalprediction.com/files/pdf/V4A9.pdf>
- Yarnold PR (2015). UniODA vs. Cochran's Q test: Comparing success of alternatives. *Optimal Data Analysis*, 4, 116-117. URL: <http://optimalprediction.com/files/pdf/V4A20.pdf>
- Yarnold PR (2015). UniODA vs. Cochran's Q test: Evaluating success rate in web usability testing. *Optimal Data Analysis*, 4, 118-119. URL: <http://optimalprediction.com/files/pdf/V4A21.pdf>
- Yarnold PR (2015). UniODA vs. Cochran's Q test: Pet store reptile display behavior by holiday. *Optimal Data Analysis*, 4, 120-121. URL: <http://optimalprediction.com/files/pdf/V4A22.pdf>
- Yarnold PR (2015). UniODA vs. Kruskal-Wallace test: Farming method and corn yield. *Optimal Data Analysis*, 4, 113-115. URL: <http://optimalprediction.com/files/pdf/V4A19.pdf>
- Yarnold PR (2015). UniODA vs. Kruskal-Wallace test: Gender and dominance of free-ranging domestic dogs in the outskirts of Rome. *Optimal Data Analysis*, 4, 122-123. URL: <http://optimalprediction.com/files/pdf/V4A23.pdf>
- Yarnold PR (2015). UniODA vs. McNemar's test for correlated proportions: Diagnosis of disease before vs. after treatment. *Optimal Data Analysis*, 4, 24-26. URL: <http://optimalprediction.com/files/pdf/V4A7.pdf>
- Yarnold PR, Brofft GC (2013). Comparing knot strength with UniODA. *Optimal Data Analysis*, 2, 54-59. URL: <http://optimalprediction.com/files/pdf/V2A9.pdf>
- Yarnold PR, Brofft GC (2013). ODA range test vs. one-way analysis of variance: Comparing strength of alternative line connections. *Optimal Data Analysis*, 2, 198-201. URL: <http://optimalprediction.com/files/pdf/V2A30.pdf>
- 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. DOI: 10.1207/s15327752jpa5203_2
- Yarnold PR, Bryant FB (1988). Seven transliterations of the short version of the student Jenkins Activity Survey. *Social and Behavioral Science Documents*, 18, 18-19. MS# 2854.
- Yarnold PR, Bryant FB (1994). A measurement model for the Type A self-rating inventory. *Journal of Personality Assessment*, 62, 102-115. DOI: 10.1207/s15327752jpa6201_10
- Yarnold PR, Bryant FB (2013). Analysis involving categorical attributes having many response categories. *Optimal Data Analysis*, 2, 69-70. URL: <http://optimalprediction.com/files/pdf/V2A12.pdf>

- Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, 4, 36-53. URL: <http://optimalprediction.com/files/pdf/V4A11.pdf>
- Yarnold PR, Bryant FB (2015). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis*, 4, 54-61. URL: <http://optimalprediction.com/files/pdf/V4A12.pdf>
- Yarnold PR, Bryant FB, Grimm LG (1987). Comparing the short and long versions of the student Jenkins Activity Survey. *Journal of Behavioral Medicine*, 10, 75-90.
- Yarnold PR, Bryant FB, Litsas F (1989). Type A behavior and psychological androgyny among Greek college students. *European Journal of Personality*, 3, 249-268. DOI: 10.1002/per.2410030403
- Yarnold PR, Bryant FB, Nightingale SD, Martin GJ (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health, and Medicine*, 1, 207-221. DOI: 10.1080/13548509608400019
- Yarnold PR, Bryant FB, & Smith JH. (2013). Manual vs. automated CTA: Predicting freshman attrition. *Optimal Data Analysis*, 2, 48-53. URL: <http://optimalprediction.com/files/pdf/V2A8.pdf>
- Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression away from the mean. *Optimal Data Analysis*, 2, 19-25. URL: <http://optimalprediction.com/files/pdf/V2A3.pdf>
- Yarnold PR, Feinglass J, Martin GJ, McCarthy WJ (1999). Comparing three pre-processing strategies for longitudinal data for individual patients: An example in functional outcomes research. *Evaluation and the Health Professions*, 22, 254-277. DOI: 10.1177/01632789922034301
- Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, 54, 73-85. DOI: 10.1177/0013164494054001007
- Yarnold PR, Lyons JS (1987). Norms for college undergraduates on the Bem Sex-Role Inventory and the Wiggins Interpersonal Behavior Circle. *Journal of Personality Assessment*, 51, 595-599. DOI: 10.1207/s15327752jpa5104_11
- Yarnold PR, Martin GJ, Soltysik RC, Nightingale SD (1993). Androgyny predicts empathy for trainees in medicine. *Perceptual and Motor Skills*, 77, 576-578. DOI: 10.2466/pms.1993.77.2.576
- Yarnold PR, Michelson EA, Thompson DA, Adams SL (1998). Predicting patient satisfaction: A study of two emergency departments. *Journal of Behavioral Medicine*, 21, 545-563. DOI: 10.1023/A:1018796628917
- Yarnold PR, Mueser KT (1988). Student version of the Jenkins Activity Survey. In: M Hersen & AS Bellack (Eds.), *Dictionary of Behavioral Assessment Techniques*. Beverly Hills, CA: Pergamon, pp. 454-455.
- Yarnold PR, Mueser KT (1989). Meta-analysis of the reliability of Type A behavior measures. *British Journal of Medical Psychology*, 62, 43-50. DOI: 10.1111/j.2044-8341.1989.tb02809.x
- Yarnold PR, Mueser KT, Grau BW, Grimm LG (1986). The reliability of the student version of the Jenkins Activity Survey. *Journal of Behavioral Medicine*, 9, 401-414. DOI: 10.1007/BF00845123
- Yarnold PR, Nightingale SD, Curry RH, Martin GJ (1990). Psychological androgyny and preference for intubation in a hypothetical case of end-stage lung disease. *Medical Decision Making*, 10, 215-222. DOI: 10.1177/0272989X9001000309
- Yarnold PR, Nightingale SD, Curry RH, Martin GJ (1991). Psychological androgyny and preference in loss-framed gambles of medical students: Possible implications for resource utilization. *Medical Decision Making*, 11, 176-179. DOI: 10.1177/0272989X9101100306
- Yarnold PR, Soltysik RC (1991). Refining two-group multivariable models using univariate optimal discriminant analysis. *Decision Sciences*, 22, 1158-1164. DOI: 10.1111/j.1540-5915.1991.tb01912.x
- Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, 22, 739-752. DOI: 10.1111/j.1540-5915.1991.tb00362

Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington DC: APA Books.

Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, 1, 10-22. URL: <http://optimalprediction.com/files/pdf/V1A2.pdf>

Yarnold PR, Soltysik RC (2010). Manual vs. automated CTA: Optimal preadmission staging for inpatient mortality from *Pneumocystis cariini* pneumonia. *Optimal Data Analysis*, 1, 50-54. URL: <http://optimalprediction.com/files/pdf/V1A9.pdf>

Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, 1, 23-29. URL: <http://optimalprediction.com/files/pdf/V1A3.pdf>

Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo estimation of two-category two-tailed *p*. *Optimal Data Analysis*, 1, 43-45. URL: <http://optimalprediction.com/files/pdf/V1A8.pdf>

Yarnold PR, Soltysik RC (2010). Unconstrained covariate adjustment in CTA. *Optimal Data Analysis*, 1, 38-40. URL: <http://optimalprediction.com/files/pdf/V1A5.pdf>

Yarnold PR, Soltysik RC (2010). UniODA vs. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis*, 1, 62-66. URL: <http://optimalprediction.com/files/pdf/V1A12.pdf>

Yarnold PR, Soltysik RC (2013). Ipsiative transformations are *essential* in the analysis of serial data. *Optimal Data Analysis*, 2, 94-97. URL: <http://optimalprediction.com/files/pdf/V2A17.pdf>

Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, 2, 220-221. URL: <http://optimalprediction.com/files/pdf/V2A35.pdf>

Yarnold PR, Soltysik RC (2013). Reverse CTA: An optimal analog to analysis of variance. *Optimal Data Analysis*, 2, 43-47. URL: <http://optimalprediction.com/files/pdf/V2A7.pdf>

Yarnold PR, Soltysik RC (2014). Discrete 95% confidence intervals for ODA model- and chance-based classifications. *Optimal Data Analysis*, 3, 110-112. URL: <http://optimalprediction.com/files/pdf/V3A26.pdf>

Yarnold PR, Soltysik RC (2014). Emergency Severity Index (Version 3) score predicts hospital admission. *Optimal Data Analysis*, 3, 20-22. URL: <http://optimalprediction.com/files/pdf/V3A8.pdf>

Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, 3, 55-77. URL: <http://optimalprediction.com/files/pdf/V3A17.pdf>

Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, 3, 78-84. URL: <http://optimalprediction.com/files/pdf/V3A18.pdf>

Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, 16, 1451-1463. DOI: 10.1002/(SICI)1097-0258(19970715)16:13<1451::AID-SIM571>3.0.CO;2-F

Yarnold PR, Soltysik RC, Collinge W (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, 2, 37-42. URL: <http://optimalprediction.com/files/pdf/V2A6.pdf>

Yarnold PR, Soltysik RC, Curry RH, Martin GJ (1989). *In quest of the best: Resident selection based on application information and mixed integer programming*. Paper presented at the Annual Meeting of the Society of Behavioral Medicine, San Francisco, CA.

Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine*, 17, 2405-2414. DOI: 10.1002/(SICI)1097-0258(19981030)17:203.0.CO;2-F

Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, 13, 1015-1021. DOI: 10.1002/sim.4780131004

Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine*, 10, 601-606. DOI: 10.1007/BF02602743

Yunus MB, Holt GS, Masi AT, Aldag JC (1988). Fibromyalgia syndrome among the elderly: Comparison with younger patients. *Journal of the American Geriatric Society*, 36, 987-995. DOI: 10.1111/j.1532-5415.1988.tb04364.x

Zakarija A, Bandarenko N, Pandey DK, Auerbach A, Raisch DW, Kim B, Kwaan HC, McKoy JM, Schmitt BP, Davidson CJ, Yarnold PR, Gorelick PB, Bennett CL (2004). Clopidogrel-associated TTP: An update of pharmacovigilance efforts conducted by independent researchers, pharmaceutical suppliers, and the Food and Drug Administration. *Stroke*, 35, 533-538. DOI: 10.1161/01.STR.0000109253.66918.5E

Zegers FE (1991). Coefficients for interrater agreement. *Applied Psychological Measurement*, 15, 321-333. DOI: 10.1177/014662169101500401

Index

- Aggregated Confusion Table – 43, 154
- ASCII data – see Data Set Design
- Attribute – 9
 - Categorical scales – 72, 160
 - Don't Parse – 28
 - Measurement scales – 40, 42
 - Ordered scales – 99
 - Parse – 258, 270
 - Rectangular categorical design – 160
- Attribute Importance in Discrimination (AID) Index – 20, 253
- Autocorrelation – 24
- Backward-stepping analysis – 126
- Base rate – 16
- Bias, reliable – 113
 - Algorithm – 17
 - Applied research – 3
 - Attribute Importance in Discrimination (AID) Index – 20
- Boolean ODA – 180
- Class Variable – 9, 41
- Classification accuracy – 1, 15
- Classification tree analysis (CTA) – 18, 190
 - Controlling Experimentwise P Value – 236
 - Determining endpoint minimum N – 231, 252 (see also Power analysis)
 - Enumerated optimal – 251
 - Forward CTA – 242
 - Globally optimal – 261
 - Growing the model – 232
 - Hierarchically optimal – 231
 - Modeling moderating effects -- 259
 - Pruning to maximize ESS – 239
 - Reverse CTA – 242
- Classifying observations using ODA/CTA models – 18
- Compositional dissimilarity – 74
- Confirmatory study – 45
- Confounding
 - and Correlation – 199
 - by combining Groups – 198
 - by combining Time periods – 205
 - by Covariates – 183
 - and CTA – 190

Exploratory ODA methods – 187
in Single-case studies -- 219
Partial UniODA – 183, 189
Reliability assessment – 119
Simpson’s paradox – 39, 133
Symbolic representation – 201
Unconstrained covariates – 185

Confusion table (matrix) – 15
Ecological significance – 15
Enumerated – 141
Globally optimal – 261
Hierarchical – 136
Parses – 139
Versus linear models – 8

D (distance) statistic – 263

Data pre-processing – 22

Data set design – 60

Data transformations – see Transformations

Degenerate model – 115

Descriptive statistics – 67
Exporting source data file – 62
Missing data – 61
Numeric variables – 61
Quality assurance – 64
Variable labels – 61

Efficiency analysis – 17

Effect Strength for Predictive Value (ESP) – 16

Effect Strength for Sensitivity (ESS) – 16
Aggregated Confusion Table – 43, 48
Rule-of-thumb – 17

Exact ODA – 181

Examples
Allied health disciplines – 249
Asylum-related appeals to the Federal courts – 159
Atmospheric pressure and physical symptoms – 46, 131
City and type of insurance – 166
Clinical medicine – 247
Clinical trial of two migraine treatments – 129
Clinician and patient agreement – 85
Community Acquired Pneumonia and in-hospital mortality – 33; vs. Influenza-like illness – 293
Comparative effectiveness of laxatives – 104
Comparing success of alternatives -- 81
Congressional voting on Pinckney Gag Rule – 76
Corn yield by farming method – 100
Dominance rankings of free-ranging dogs – 100
Drug effect on disease – 90, 91

Duration of membrane rupture and Cesarean delivery – 128
Ecological categories and sampling sites – 75
Emergency Department triage coding and hospital admission – 135
Exchange of material and psychological resources – 93
Fracking and crude mortality rate – 122
Gender and city – 163; demographic characteristics – 169; insurance – 165
Gender and income – 44
Gender, race, and cancer incidence – 268
Geriatric vs. non-geriatric medical patients – 19
Heart rate variability and sudden cardiac death – 130
Identifying adverse drug reactions – 241
In-hospital mortality and PCP pneumonia – 185, 246
Index of Consumer Sentiment – 22
Inter-method agreement of ESI and CTAS triage scores – 117;
Inter-rater agreement, Canadian Triage and Acuity Scale (CTAS) triage scores – 116; Emergency Severity Index (ESI) triage scores – 114; facial expression – 111; movie ratings – 99;
Intubating pneumonia patients – 30
“Junk science” in the courtroom – 204
Knot breaking strength – 68
Likelihood a discharged patient recommends Emergency Department – 183, 187, 231, 251, 289
Long-term diagnosis stability – 84; inter-method agreement – 84, 85
Looking forward to receiving a good grade – 146
Modeling vacation enjoyment – 144
Mortality status and city – 161; gender – 161; insurance – 162
Number of faculty by rank, gender, and year – 86; lambs born to ewes over two years – 110
Optimistic benefit-finding in the face of adversity – 145
Outcomes of marital therapy – 80
Patient satisfaction with Emergency Department doctors – 151; nurses – 150; technicians – 149
Person-environment fit theory and college freshman attrition – 257
Petal width and sunshine – 102
Plaintiff gender and age – 80
Point-treatment study stratified by measured confounder – 197
Political affiliation of parents and children – 77
Predicting export of Arctic sea ice – 217
Predicting temperature and precipitation anomalies – 205, 216, 320
Protein vs. DNA polymorphisms – 50
Psychosocial adaptation in early adolescence – 256
Psychosocial aspects of medicine – 249
Race and pneumonia laboratory test outcomes – 107
Rat brain blood samples – 25
Region of residence across time – 72
Reliability of androgyny 4-fold typology scores – 120
Reptile display by store and holiday – 83
Rock type order in carbonite units – 89
Serum cholesterol and coronary heart disease – 87
Smile production in infants – 79

Strength of fishing line connections – 105; gender differences – 81
Success rate in Web usability testing – 82
Temporal stability of affective experience – 121; learning styles – 92; negative symptom ratings – 219, 222
Type A behavior and coronary artery disease – 134; savoring behaviors – 140, and beliefs – 137
Voting intentions across time – 97

Fisher's exact test – 14

Forward-stepping analysis – 23, 126

GEN (multisample generalizability analysis) – 37, 38, 87

Generalizability (see also GEN) – 32

- Hold-out – 33
- Leave-one-out (LOO), jackknife – 32

Hold-out validation – 33

Improving future models, role of residuals – 66

Integer coefficient models – 182

Iterative structural decomposition – 4, 73, 93

Legacy statistical methods, general

- Accuracy – 7
- Assumptions – 7, 17
- Insensitivity (Robust) – 7
- Simpson's paradox – 8, 18

Legacy statistical methods, specific

- ANOVA – 105, 107, 155
- Bowker's test for symmetry – 72
- Bray-Curtis Dissimilarity Index – 74
- Chi-Square – 75, 78
- Cochran's Q test --- 81
- Cohen's kappa – 84, 85, 113
- Eyeball analysis – 92, 94, 96
- Generalizability theory – 117
- Kendall's Coefficient of Concordance – 99
- Kendall's tau – 132
- Kruskal-Wallace test – 50, 100
- Linear discriminant function – 156
- Logistic regression analysis – 87, 159, 185
- Log-Linear model – 73, 76, 86, 89, 97, 158
- Mann-Whitney *U* test – 102
- Marginal structural model – 197
- Markov processes – 88, 97
- McNemar's test for correlated proportions – 72, 90
- Polychoric correlation – 110
- Principal components analysis – 207, 209, 210
- Probit model – 158
- Regression analysis – 142, 144-157, 206, 243 (see also Regression analysis)
- ROC analysis – 128
- t* test – 129, 138

Test-retest correlation – 84
Turnover table – 92
Leave-one-out (LOO) analysis – see Generalizability
Linear versus non-linear model – 17, 169
Longitudinal analysis – 121 (see also Single-case design)
Map making – 121
Marginal structural model – 197
Maximum accuracy and ODA – 1
Measurement
 Don't parse attributes – 28
 Instrumentation – 52
 Precision dimension – 14
 Predictive accuracy – 40
 Scales – 40, 42
Mixed-integer programming – 173, 210
Moderation and CTA – 190
Multiattribute ODA models – 71
MultiODA – 173
 Binary attributes – 178
 Optimal attribute subset selection – 177
 Special-purpose models – 180
 Warmack search algorithm – 179
 Weighted classification – 176
Multisample generalizability analysis – see GEN
Multitrait–multimethod matrix – 137
N-of-1 Design – 24
Novometric theory
 and Quantum Mechanics – 264
 Axioms – 263
 Exact discrete confidence intervals for model and chance – 266
 MDSA and SDA algorithms – 290, 292
 Missing data – 293
 Model endpoint redundancy – 267
 Multiple attribute models – 289
 Single attribute models – 268
ODA Model
 Categorical attribute – 11
 Direction – 9
 Invariance over monotonic transformations – 105
 Optimal threshold – 9
 Ordered attribute – 9
ODA paradigm
 Commercial applications – 6
 Genesis of name – 1
 Improving science – 6
 Learning – 2

Publishing – 2
Teaching – 2
ODA software (CTA, MegaODA, UniODA)
 CTA command syntax – 312
 MegaODA time trials – 305
 Programmers File Editor – 64
 Troubleshooting -- 319
 UniODA and MegaODA command syntax – 297
 Using DOS prompt – 64
Optimal discriminant analysis – 4
Optimal range test – 107, 163, 165
Optimal solution – 1
Optimal values – 11
Overall PAC – 11
P Value
 Alpha inflation – 14
 Computation – 11
 Ensemble adjusted –14
 Experimentwise criterion – 14
 Generalized criterion – 14
 Multiple comparisons – 14
 Planned comparisons – 14
 Sidak (Bonferroni-Like) correction – 14
 Simulation studies – 12
PAC (and overall PAC) – 11, 15
Pairwise comparisons
 All possible comparisons – 101
 Optimal range test – 107, 163, 165
Paradoxical confounding – see Confounding
Partial UniODA – 183, 189
Pie chart – 21
Power analysis – 53
 Exact minimum precision approach – 56
 Model geometry -- 59
 Parametric approach – 54
Predictive value – 15
 And base rate – 16
 Efficiency analysis – 17
Pre-processing data – 22
Prior odds – 43
Propensity score and staging table – 19
Raw scores (see also Transformations) – 34
Regression analysis (see also Legacy statistical methods, specific) – 243
 Assumptions – 142
 Confounding – 206
 Regression away from mean – 144

Regression toward mean – 143

Reliability analysis – 111

- Inter-method (Parallel Forms) – 85, 117
- Inter-rater – 111, 114
- Paradoxical confounding – 119
- Reliable bias – 113
- Split-half – 119
- Temporal (test-retest) – 84, 120

Repeated measures analysis -- 121

Repeated measures analysis – see Longitudinal analysis

Reporting findings – 66

Reproducibility (see also Generalizability, Validity) – 32

Research funding – 5

Residuals – 66

Sample size – see Power Analysis

Satisfaction

- Customer – 6
- Care received in the Emergency Department – 148, 187

Sensitivity – 15

Simpson’s paradox – see Confounding

Single-case analysis – 24, 133

Spilt-half validation – 32

Split-half validation – 33

Staging Tables – 19, 136, 171, 187

- Propensity scores – 19

State transition table – 89, 97

Statistical reliability – see *P* value

Statistical significance – see *P* value

Structural decomposition – see Iterative structural decomposition

Suboptimal – 1

Symptom dominance hierarchy – 223

Tau ODA – 181

Template ODA – 182

Temporal analysis – see longitudinal analysis, single-case analysis

Theoretically ideal statistical model – 261

- Comparing empirical and theoretically ideal models – 262

Time series analysis – see longitudinal analysis, single-case analysis

Transformations – 21

- Interactive transformation – 28, 222
- Ipsative standardization – 25, 221
- Normative Standardization – 25, 36
- Raw scores -- 34
- Weighting observations – 22

UniODA

- Algorithm, ordered data – 9
- Algorithm, categorical data – 11

Nonlinear – 50
Optimal value, distribution – 11
Unit coefficient ODA models – 182
Validity analysis – 134
 Construct validity – 134
 Convergent and discriminant validity – 137
Warmack search algorithm – 179
Weighted solution – 1
Weighting observations – 22, 43

About the Authors

Paul R. Yarnold, Ph.D.

After receiving his Ph.D. in academic social psychology at the University of Illinois at Chicago in 1984, Dr. Yarnold joined the faculty of Northwestern University Medical School (appointments in General Internal Medicine, Allergy/Immunology, and Emergency Medicine) and the University of Illinois at Chicago (Adjunct faculty, Psychology). In medical school Paul worked primarily in the design and statistical analysis aspects of active research ongoing in areas of General Internal Medicine, Cardiology, Surgery, Psychiatry, Emergency Medicine, Infectious Disease, Pharmacy, Physical Medicine and Rehabilitation, Geriatrics, Comparative/Translational, Oncology, Dermatology, OB/GYN, Allergy/Immunology, and Informatics until 2010, when as Research Professor of Medicine, and Adjunct Professor of Psychology, he left full-time academia to become an entrepreneurial scientist. Paul is an elected Fellow of the Society of Behavioral Medicine, and an elected Fellow of Divisions 5 (Measurement, Evaluation, and Statistics) and 38 (Health Psychology) of the American Psychological Association. He co-discovered the Optimal Data Analysis paradigm and novometric theory; founded the eJournal *Optimal Data Analysis*; published close to 400 articles on a myriad of topics largely in the areas of statistics, medicine and psychology; authored two best-selling statistics books; obtained more than \$16M in grant awards; and earned several patents. In April, 2016, Paul had a total of 13,387 citations and an H index of 52 (Google Scholar), and an impact factor of 1,122.50 (Research Gate).

Robert C. Solysik, M.S.

Robert is a consultant with Optimal Data Analysis, LLC, and co-editor of the journal *Optimal Data Analysis*. He is a former Senior Research Associate at Northwestern University, and scientific programmer for Feinberg School of Medicine and for the Veterans Affairs Chicago Health Care System. He has served on numerous NIH-sponsored studies including the Fibromyalgia Wellness Project and other interactive intervention systems. With Yarnold he co-discovered the ODA paradigm of maximum accuracy statistics, and created the software systems including UniODA™, and CTA™. He is co-author with Yarnold of *Optimal Data Analysis: A Guidebook for Windows* (APA Press, 2004).

"The ways in which these techniques can be applied are too numerous to count. It is only a matter of time until these tools will replace conventional methods for use in everything from clinical trials, to causal inference, to predictive modeling, to application in everyday practice. Only the limits of the human imagination pose a constraint for the application of the maximum-accuracy paradigm." --- *Ariel Linden, DrPH, Division of General Internal Medicine, Medical School, University of Michigan, Ann Arbor*

"This book is destined to become a classic in the field of statistics. The unified analytic paradigm of optimally accurate methods that the authors present is the most extraordinary quantitative breakthrough I've seen during my 45 years in the field of statistics. This modern-day analytic paradigm represents a quantum leap beyond the inferential statistical methods of the 19th and 20th century. Had Pearson and Fisher had access to this maximum accuracy distribution-free analytic method, they would have embraced it for its power and precision, its optimal accuracy, its elegant simplicity, and its virtually unlimited applicability. Contemporary statisticians ought to do the same, for this is truly the wave of the future." — *Fred B. Bryant, PhD, Department of Psychology, Loyola University, Chicago*

Procedures to identify mathematical models that explicitly yield optimal (maximum accuracy) solutions for samples were widely studied in the past century, with literatures emerging in fields such as symbolic logic, discrete mathematics, operations research, mathematical programming, set theory, decision-making, systems engineering, algorithms, automated manufacturing, computer science, machine intelligence, finance, transportation science, management science, and numerical taxonomy. Broad-spectrum consensus among disparate experts indicates that predictive accuracy is an objective function worthy of being optimized.

In the Optimal (or "optimizing") Data Analysis (ODA) statistical paradigm, an optimization algorithm is first utilized to identify the model that explicitly maximizes predictive accuracy for the sample, and then the resulting optimal performance is evaluated in the context of an application-specific exact statistical architecture. Discovered in 1990, the first and most basic ODA model was a distribution-free machine learning algorithm used to make maximum accuracy classifications of observations into one of two categories (pass or fail) on the basis of their score on an ordered attribute (test score). When the first book on ODA was written in 2004 a cornucopia of indisputable evidence had already amassed demonstrating that statistical models identified by ODA were more flexible, transparent, intuitive, accurate, parsimonious, and generalizable than competing models instead identified using an unintegrated menagerie of legacy statistical methods. Understanding of ODA methodology skyrocketed over the next decade, and 2014 produced the development of novometric theory – the conceptual analogue of quantum mechanics for the statistical analysis of classical data. This point was selected to pause to write *Maximizing Predictive Accuracy*, as a means of organizing and making sense of all that has so-far been learned about ODA, through November of 2015.

For researchers exploring ODA for the first time, the most appreciated aspect will likely be the intellectually transparent, intuitive presentation involving minimal use of a few simple equations: the optimal, maximum-accuracy paradigm is clear in its derivation, application, computation, interpretation, and dissemination. For researchers that use ODA in their work, the most appreciated reward is unquestionably the unmatched flexibility, simplicity, and accuracy of the statistical models – and their generalizability across time and sample. ODA accommodates all metrics, requires no distributional assumptions, allows for analytic weighting of individual observations, explicitly maximizes predictive accuracy (overall, or normed against chance), and supports multiple methods of assessing validity. Finally, amply illustrated in *Maximizing Predictive Accuracy*, conducting maximum-accuracy statistical analysis is made astonishingly straightforward, simple, and fast using commercially-available special-purpose software.

Optimal Data Analysis LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646