

CONTENTS

Preface	xi
About the Author	xv
Chapter 1 Why Data Cleaning Is Important: Debunking the Myth of Robustness	1
Origins of Data Cleaning	2
Are Things Really That Bad?	5
Why Care About Testing Assumptions and Cleaning Data?	8
How Can This State of Affairs Be True?	8
The Best Practices Orientation of This Book	10
Data Cleaning Is a Simple Process; However . . .	11
One Path to Solving the Problem	12
For Further Enrichment	13
SECTION I: BEST PRACTICES AS YOU PREPARE FOR DATA COLLECTION	17
Chapter 2 Power and Planning for Data Collection: Debunking the Myth of Adequate Power	19
Power and Best Practices in Statistical Analysis of Data	20
How Null-Hypothesis Statistical Testing Relates to Power	22
What Do Statistical Tests Tell Us?	23
How Does Power Relate to Error Rates?	26
Low Power and Type I Error Rates in a Literature	28
How to Calculate Power	29
The Effect of Power on the Replicability of Study Results	31
Can Data Cleaning Fix These Sampling Problems?	33
Conclusions	34
For Further Enrichment	35
Appendix	36

Chapter 3 Being True to the Target Population: Debunking the Myth of Representativeness	43
Sampling Theory and Generalizability	45
Aggregation or Omission Errors	46
Including Irrelevant Groups	49
Nonresponse and Generalizability	52
Consent Procedures and Sampling Bias	54
Generalizability of Internet Surveys	56
Restriction of Range	58
Extreme Groups Analysis	62
Conclusion	65
For Further Enrichment	65
Chapter 4 Using Large Data Sets With Probability Sampling Frameworks: Debunking the Myth of Equality	71
What Types of Studies Use Complex Sampling?	72
Why Does Complex Sampling Matter?	72
Best Practices in Accounting for Complex Sampling	74
Does It Really Make a Difference in the Results?	76
So What Does All This Mean?	80
For Further Enrichment	81
SECTION II: BEST PRACTICES IN DATA CLEANING AND SCREENING	85
Chapter 5 Screening Your Data for Potential Problems: Debunking the Myth of Perfect Data	87
The Language of Describing Distributions	90
Testing Whether Your Data Are Normally Distributed	93
Conclusions	100
For Further Enrichment	101
Appendix	101
Chapter 6 Dealing With Missing or Incomplete Data: Debunking the Myth of Emptiness	105
What Is Missing or Incomplete Data?	106
Categories of Missingness	109
What Do We Do With Missing Data?	110
The Effects of Listwise Deletion	117
The Detrimental Effects of Mean Substitution	118

The Effects of Strong and Weak Imputation of Values	122
Multiple Imputation: A Modern Method of Missing Data Estimation	125
Missingness Can Be an Interesting Variable in and of Itself	128
Summing Up: What Are Best Practices?	130
For Further Enrichment	131
Appendixes	132

Chapter 7 Extreme and Influential Data Points: Debunking the Myth of Equality 139

What Are Extreme Scores?	140
How Extreme Values Affect Statistical Analyses	141
What Causes Extreme Scores?	142
Extreme Scores as a Potential Focus of Inquiry	149
Identification of Extreme Scores	152
Why Remove Extreme Scores?	153
Effect of Extreme Scores on Inferential Statistics	156
Effect of Extreme Scores on Correlations and Regression	156
Effect of Extreme Scores on t-Tests and ANOVAs	161
To Remove or Not to Remove?	165
For Further Enrichment	165

Chapter 8 Improving the Normality of Variables Through Box-Cox Transformation: Debunking the Myth of Distributional Irrelevance 169

Why Do We Need Data Transformations?	171
When a Variable Violates the Assumption of Normality	171
Traditional Data Transformations for Improving Normality	172
Application and Efficacy of Box-Cox Transformations	176
Reversing Transformations	181
Conclusion	184
For Further Enrichment	185
Appendix	185

Chapter 9 Does Reliability Matter? Debunking the Myth of Perfect Measurement 191

What Is a Reasonable Level of Reliability?	192
Reliability and Simple Correlation or Regression	193
Reliability and Partial Correlations	195
Reliability and Multiple Regression	197
Reliability and Interactions in Multiple Regression	198

Protecting Against Overcorrecting During Disattenuation	199
Other Solutions to the Issue of Measurement Error	200
What If We Had Error-Free Measurement?	200
An Example From My Research	202
Does Reliability Influence Other Analyses?	205
The Argument That Poor Reliability Is Not That Important	206
Conclusions and Best Practices	207
For Further Enrichment	208

SECTION III: ADVANCED TOPICS IN DATA CLEANING 211

Chapter 10 Random Responding, Motivated Misresponding, and Response Sets: Debunking the Myth of the Motivated Participant 213

What Is a Response Set?	213
Common Types of Response Sets	214
Is Random Responding Truly Random?	216
Detecting Random Responding in Your Research	217
Does Random Responding Cause Serious Problems With Research?	219
Example of the Effects of Random Responding	219
Are Random Responders Truly Random Responders?	224
Summary	224
Best Practices Regarding Random Responding	225
Magnitude of the Problem	226
For Further Enrichment	226

Chapter 11 Why Dichotomizing Continuous Variables Is Rarely a Good Practice: Debunking the Myth of Categorization 231

What Is Dichotomization and Why Does It Exist?	233
How Widespread Is This Practice?	234
Why Do Researchers Use Dichotomization?	236
Are Analyses With Dichotomous Variables Easier to Interpret?	236
Are Analyses With Dichotomous Variables Easier to Compute?	237
Are Dichotomous Variables More Reliable?	238
Other Drawbacks of Dichotomization	246
For Further Enrichment	250

Chapter 12	The Special Challenge of Cleaning Repeated Measures Data: Lots of Pits in Which to Fall	253
	Treat All Time Points Equally	253
	What to Do With Extreme Scores?	257
	Missing Data	258
	Summary	258
Chapter 13	Now That the Myths Are Debunked . . . : Visions of Rational Quantitative Methodology for the 21st Century	261
	Name Index	265
	Subject Index	269