Contents

Biographies xv
Foreword xvii
Preface to the Second Edition xix
Acknowledgments xxi

CHAPTER 1 Introduction and How to Use This Book 1
Introduction 1
The Organization of This Book 1
How to Use This Book 2
What Test Should I Use? 4
What Sample Size Do I Need? 5
You Don't Have to Do the Computations by Hand 6
Key Points 7
Chapter Review Questions 7
Answers to Chapter Review Questions 7
References 8

CHAPTER 2 Quantifying User Research 9
What is User Research? 9
Data from User Research 9
Usability Testing 10
Sample Sizes 10
Representativeness and Randomness 10
Three Types of Studies for User Research 12
Data Collection 12
Completion Rates 13
Usability Problems (UI Problems) 13
Task Time 14
Errors 15
Satisfaction Ratings 15
Combined Scores 15
A/B Testing 15
Clicks, Page Views, and Conversion Rates 15
Survey Data 16
Rating Scales 16
Net Promoter Scores 16
Comments and Open-Ended Data 17
Requirements Gathering 17

Key Points 17
References 18

CHAPTER 3 How Precise Are Our Estimates? Confidence Intervals 19
Introduction 19
Confidence Interval = Twice the Margin of Error 19
Confidence Intervals Provide Precision and Location 19
Three Components of a Confidence Interval 20
Confidence Interval for a Completion Rate 20
Confidence Interval History 21
Wald Interval: Terribly Inaccurate for Small Samples 21
Exact Confidence Interval 22
Adjusted-Wald: Add Two Successes and Two Failures 22
Best Point Estimates for a Completion Rate 24
Confidence Interval for a Problem Occurrence 25
Confidence Interval for Rating Scales and Other Continuous Data 26
Confidence Interval for Task-Time Data 29
Mean or Median Task Time? 30
Geometric Mean 30
Confidence Interval for Large Sample Task Times 33
Confidence Interval Around a Median 33
Key Points 35
Chapter Review Questions 36
Answers to Chapter Review Questions 37
References 38

CHAPTER 4 Did We Meet or Exceed Our Goal? 39
Introduction 39
One-Tailed and Two-Tailed Tests 42
Comparing a Completion Rate to a Benchmark 43
Small Sample Test 43
Large Sample Test 47
Comparing a Satisfaction Score to a Benchmark 48
Do at Least 75% Agree? Converting Continuous Ratings to Discrete 50
Comparing a Task Time to a Benchmark 52
Key Points 55
Chapter Review Questions 55
Answers to Chapter Review Questions 56
References 59

CHAPTER 5 Is There a Statistical Difference Between Designs? 61
Introduction 61
Comparing Two Means (Rating Scales and Task Times) 61
Within-Subjects Comparison (Paired *t*-Test) 61

Comparing Task Times 65
Between-Subjects Comparison (Two-Sample *t*-Test) 67
Assumptions of the *t*-tests 71
Comparing Completion Rates, Conversion Rates, and A/B Testing 73
Between-Subjects 73
Within-Subjects 83
Key Points 93
Chapter Review Questions 95
Answers to Chapter Review Questions 97
References 102

CHAPTER 6 What Sample Sizes Do We Need? Part 1: Summative Studies 103
Introduction 103
Why Do We Care? 103
The Type of Usability Study Matters 103
Basic Principles of Summative Sample Size Estimation 104
Estimating Values 106
Comparing Values 112
What can I Do to Control Variability? 118
Sample Size Estimation for Binomial Confidence Intervals 119
Binomial Sample Size Estimation for Large Samples 119
Binomial Sample Size Estimation for Small Samples 120
Sample Size for Comparison with a Benchmark Proportion 123
Sample Size Estimation for Chi-Squared Tests (Independent Proportions) 125
Sample Size Estimation for McNemar Exact Tests (Matched Proportions) 128
Key Points 132
Chapter Review Questions 133
Answers to Chapter Review Questions 135
References 140

CHAPTER 7 What Sample Sizes Do We Need? Part 2: Formative Studies 143
Introduction 143
Using a Probabilistic Model of Problem Discovery to Estimate Sample Sizes for Formative User Research 143
The Famous Equation $P(x \geq 1) = 1 - (1 - p)^n$ 143
Deriving a Sample Size Estimation Equation from $1 - (1 - p)^n$ 145
Using the Tables to Plan Sample Sizes for Formative User Research 146
Assumptions of the Binomial Probability Model 148
Additional Applications of the Model 149
Estimating the Composite Value of p for Multiple Problems or Other Events 149
Adjusting Small-Sample Composite Estimates of p 149

Estimating the Number of Problems Available for Discovery and the Number of Undiscovered Problems 155
What affects the Value of *p*? 157
What is a Reasonable Problem Discovery Goal? 157
Reconciling the "Magic Number Five" with "Eight is not Enough" 160
Some History—The 1980s 160
Some More History—The 1990s 161
The Derivation of the "Magic Number 5" 162
Eight is Not Enough—A Reconciliation 164
More About the Binomial Probability Formula and its Small-Sample Adjustment 167
The Origin of the Binomial Probability Formula 167
How Does the Deflation Adjustment Work? 169
Other Statistical Models for Problem Discovery 171
Criticisms of the Binomial Model for Problem Discovery 171
Expanded Binomial Models 172
Capture-Recapture Models 173
Why Not Use One of These Other Models When Planning Formative User Research? 173
Key Points 177
Chapter Review Questions 179
Answers to Chapter Review Questions 179
References 180

CHAPTER 8 Standardized Usability Questionnaires 185
Introduction 185
What is a Standardized Questionnaire? 185
Advantages of Standardized Usability Questionnaires 185
What Standardized Usability Questionnaires Are Available? 186
Assessing the Quality of Standardized Questionnaires: Reliability, Validity, and Sensitivity 187
Other Item Characteristics 187
Post-Study Questionnaires 189
QUIS (Questionnaire for User Interaction Satisfaction) 189
SUMI (Software Usability Measurement Inventory) 190
PSSUQ (Post-Study System Usability Questionnaire) 192
SUS (System Usability Scale) 198
UMUX (Usability Metric for User Experience) 211
UMUX-LITE 212
Experimental Comparison of Post-study Usability Questionnaires 214
Post-Task Questionnaires 216
ASQ (After-Scenario Questionnaire) 217

SEQ (Single Ease Question) 218
SMEQ (Subjective Mental Effort Question) 219
ER (Expectation Ratings) 220
UME (Usability Magnitude Estimation) 221
Experimental Comparisons of Post-Task Questionnaires 223
Questionnaires for Assessing Perceived Usability of Websites 225
WAMMI (Website Analysis and Measurement Inventory) 226
SUPR-Q (Standardized User Experience Percentile Rank Questionnaire) 227
Other Questionnaires for Assessing Websites 229
Other Questionnaires of Interest 231
CSUQ (Computer System Usability Questionnaire) 231
USE (Usefulness, Satisfaction, and Ease-of-Use) 232
HQ (Hedonic Quality) 232
EMO (Emotional Metric Outcomes) 233
ACSI (American Customer Satisfaction Index) 235
NPS (Net Promoter Score) 235
CxPi (Forrester Customer Experience Index) 236
TAM (Technology Acceptance Model) 237
Key Points 238
Chapter Review Questions 239
Answers to Chapter Review Questions 240
References 242

CHAPTER 9 Six Enduring Controversies in Measurement and Statistics 249
Introduction 249
Is it OK to Average Data from Multipoint Scales? 250
On One Hand 250
On the Other Hand 251
Our Recommendation 253
Do You Need to Test at Least 30 Users? 254
On One Hand 254
On the Other Hand 255
Our Recommendation 256
Should You Always Conduct a Two-Tailed Test? 257
On One Hand 257
On the Other Hand 258
Our Recommendation 258
Can You Reject the Null Hypothesis When $p > 0.05$? 259
On One Hand 259
On the Other Hand 259
Our Recommendation 261

Can You Combine Usability Metrics into Single Scores? 263
On One Hand 263
On the Other Hand 264
Our Recommendation 265
What if You Need to Run more than One Test? 265
On One Hand 265
On the Other Hand 268
Our Recommendation 268
Key Points 271
Chapter Review Questions 272
Answers to Chapter Review Questions 274
References 275

CHAPTER 10 An Introduction to Correlation, Regression, and ANOVA 277
Introduction 277
Correlation 278
How to Compute a Correlation 279
Statistical Significance of *r* 281
Confidence Intervals for *r* 282
Interpreting the Magnitude of *r* 283
Sample Size Estimation for *r* 284
Coefficient of Determination (R^2) 284
Correlation with Binary Data 285
Computing the Phi Correlation 285
Regression 288
Estimating Slopes and Intercepts 289
Confidence Intervals for Slopes and Predicted Values 289
Sample Size Estimation for Linear Regression 293
Analysis of Variance 295
Comparing More Than Two Means 295
Assessing Interactions 303
Confidence Intervals and Sample Size Estimation for ANOVA 307
Key Points 307
Chapter Review Questions 308
Answers to Chapter Review Questions 308
References 318
Appendix: Derivation of Sample Size Formulas for Regression 319
Based on Confidence Interval for Regression Slope 319
Based on Confidence Interval for Regression Intercept 319

CHAPTER 11 Wrapping Up **321**
Introduction 321
Getting More Information 321
Good Luck! 324
Key Points 324
References 324

APPENDIX A Crash Course in Fundamental Statistical Concepts **327**
Introduction 327
Types of Data 327
Populations and Samples 328
Sampling 328
Measuring Central Tendency 328
Mean 328
Median 329
Geometric Mean 329
Standard Deviation and Variance 330
The Normal Distribution 330
z-Scores 332
Area Under the Normal Curve 332
Applying the Normal Curve to User Research Data 334
Central Limit Theorem 335
Standard Error of the Mean 336
Margin of Error 337
t-Distribution 337
Significance Testing and *p*-Values 338
How Much Do Sample Means Fluctuate? 338
The Logic of Hypothesis Testing 340
Errors in Statistics 341
Key Points 342

Subject Index **345**