Contents

Preface – How to Read This Book  page xv
Acknowledgments  xvii

PART I INTRODUCTORY TOPICS FOR EVERYONE  1

1 Introduction and Motivation  3
Online Controlled Experiments Terminology  5
Why Experiment? Correlations, Causality, and Trustworthiness  8
Necessary Ingredients for Running Useful Controlled Experiments  10
Tenets  11
Improvements over Time  14
Examples of Interesting Online Controlled Experiments  16
Strategy, Tactics, and Their Relationship to Experiments  20
Additional Reading  24

2 Running and Analyzing Experiments: An End-to-End Example  26
Setting up the Example  26
Hypothesis Testing: Establishing Statistical Significance  29
Designing the Experiment  32
Running the Experiment and Getting Data  34
Interpreting the Results  34
From Results to Decisions  36

3 Twyman’s Law and Experimentation Trustworthiness  39
Misinterpretation of the Statistical Results  40
Confidence Intervals  43
Threats to Internal Validity  43
Threats to External Validity  48
### Table of Contents

<table>
<thead>
<tr>
<th>Segment Differences</th>
<th>52</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simpson’s Paradox</td>
<td>55</td>
</tr>
<tr>
<td>Encourage Healthy Skepticism</td>
<td>57</td>
</tr>
</tbody>
</table>

#### 4 Experimentation Platform and Culture

- Experimentation Maturity Models | 58 |
- Infrastructure and Tools | 66 |

#### PART II SELECTED TOPICS FOR EVERYONE

- 5 Speed Matters: An End-to-End Case Study | 79 |
  - Key Assumption: Local Linear Approximation | 83 |
  - How to Measure Website Performance | 84 |
  - The Slowdown Experiment Design | 86 |
  - Impact of Different Page Elements Differs | 87 |
  - Extreme Results | 89 |

#### 6 Organizational Metrics

- Metrics Taxonomy | 90 |
- Formulating Metrics: Principles and Techniques | 94 |
- Evaluating Metrics | 96 |
- Evolving Metrics | 97 |
- Additional Resources | 98 |
- SIDEBAR: Guardrail Metrics | 98 |
- SIDEBAR: Gameability | 100 |

#### 7 Metrics for Experimentation and the Overall Evaluation Criterion

- From Business Metrics to Metrics Appropriate for Experimentation | 102 |
- Combining Key Metrics into an OEC | 104 |
- Example: OEC for E-mail at Amazon | 106 |
- Example: OEC for Bing’s Search Engine | 108 |
- Goodhart’s Law, Campbell’s Law, and the Lucas Critique | 109 |

#### 8 Institutional Memory and Meta-Analysis

- What Is Institutional Memory? | 111 |
- Why Is Institutional Memory Useful? | 112 |

#### 9 Ethics in Controlled Experiments

- Background | 116 |
- Data Collection | 121 |
- Culture and Processes | 122 |
- SIDEBAR: User Identifiers | 123 |
# Table of Contents

## Part III: Complementary and Alternative Techniques to Controlled Experiments

### 10 Complementary Techniques
- The Space of Complementary Techniques
- Logs-based Analysis
- Human Evaluation
- User Experience Research (UER)
- Focus Groups
- Surveys
- External Data
- Putting It All Together

### 11 Observational Causal Studies
- When Controlled Experiments Are Not Possible
- Designs for Observational Causal Studies
- Pitfalls
- SIDEBAR: Refuted Observational Causal Studies

## Part IV: Advanced Topics for Building an Experimentation Platform

### 12 Client-Side Experiments
- Differences between Server and Client Side
- Implications for Experiments
- Conclusions

### 13 Instrumentation
- Client-Side vs. Server-Side Instrumentation
- Processing Logs from Multiple Sources
- Culture of Instrumentation

### 14 Choosing a Randomization Unit
- Randomization Unit and Analysis Unit
- User-level Randomization

### 15 Ramping Experiment Exposure: Trading Off Speed, Quality, and Risk
- What Is Ramping?
- SQR Ramping Framework
- Four Ramp Phases
- Post Final Ramp
Table of Contents

16 Scaling Experiment Analyses
   Data Processing 177
   Data Computation 178
   Results Summary and Visualization 180

PART V ADVANCED TOPICS FOR ANALYZING EXPERIMENTS 183

17 The Statistics behind Online Controlled Experiments 185
   Two-Sample t-Test 185
   p-Value and Confidence Interval 186
   Normality Assumption 187
   Type I/II Errors and Power 189
   Bias 191
   Multiple Testing 191
   Fisher’s Meta-analysis 192

18 Variance Estimation and Improved Sensitivity: Pitfalls and Solutions 193
   Common Pitfalls 193
   Improving Sensitivity 196
   Variance of Other Statistics 198

19 The A/A Test 200
   Why A/A Tests? 200
   How to Run A/A Tests 205
   When the A/A Test Fails 207

20 Triggering for Improved Sensitivity 209
   Examples of Triggering 209
   A Numerical Example (Kohavi, Longbotham et al. 2009) 212
   Optimal and Conservative Triggering 213
   Overall Treatment Effect 214
   Trustworthy Triggering 215
   Common Pitfalls 216
   Open Questions 217

21 Sample Ratio Mismatch and Other Trust-Related Guardrail Metrics 219
   Sample Ratio Mismatch 219
   Debugging SRMs 222
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td><strong>Leakage and Interference between Variants</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examples</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>Some Practical Solutions</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td>Detecting and Monitoring Interference</td>
<td>230</td>
</tr>
<tr>
<td>23</td>
<td><strong>Measuring Long-Term Treatment Effects</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What Are Long-Term Effects?</td>
<td>235</td>
</tr>
<tr>
<td></td>
<td>Reasons the Treatment Effect May Differ between Short-Term and Long-Term</td>
<td>236</td>
</tr>
<tr>
<td></td>
<td>Why Measure Long-Term Effects?</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>Long-Running Experiments</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>Alternative Methods for Long-Running Experiments</td>
<td>241</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>Index</td>
<td>266</td>
</tr>
</tbody>
</table>