

TABLE OF CONTENTS

MOTIVATION	IX
WHO IS THE READER OF THIS BOOK?	XI
ACKNOWLEDGMENTS	XIII
NOTATIONS	XV
1. USING STATISTICS IN BUSINESS	1
1.1. When are statistics useful?	1
1.2. Statistical inference in business	6
1.3. Primary uses of statistics in online A/B testing	8
1.4. The necessity for counterfactual reasoning	11
1.5. Establishing causality	15
1.6. Ruling out alternative explanations for the data	18
1.7. Statistical methods and efficient use of data	20
1.8. Caveats in using statistical methods in business	22
2. ESTIMATING UNCERTAINTY. STATISTICAL SIGNIFICANCE, P-VALUES, OTHER ESTIMATES.	25
2.1. Substantive hypotheses	25
2.2. Statistical hypotheses	27
2.3. Standard deviation and Z-Scores	31
2.4. p -value and type I errors	38
2.5. p -value: utility and interpretation	41
2.6. Confidence intervals	44
2.7. Misinterpretations of p -values and confidence intervals	50
2.8. p -values and confidence intervals in decision-making	53
2.9. Maximum likelihood estimate	55

2.10. Severity	57
3. STATISTICAL ASSUMPTIONS. ASSESSING MODEL ADEQUACY	59
3.1. The importance of statistical assumptions	59
3.2. Probabilistic assumptions of statistical models	61
3.3. Probabilistic assumptions in different practical cases	63
3.4. Assumptions imposed by the design of the experiment	65
3.5. Assessing statistical adequacy through statistical tests	66
3.6. Assessing statistical adequacy through A/A tests	67
4. STATISTICAL POWER AND SAMPLE SIZE CALCULATIONS	71
4.1. Errors of the second kind (type II errors)	71
4.2. Statistical power	74
4.3. The role of statistical power in A/B testing	79
4.4. Minimum effect of interest vs. minimum detectable effect	81
4.5. Underpowered and overpowered tests	83
4.6. Sample size calculations	85
4.7. False positive and false negative rates	93
4.8. Misunderstandings about statistical power	95
5. TYPES OF STATISTICAL HYPOTHESES	97
5.1. One-sided tests and confidence intervals	98
5.2. Misconceptions about one-sided tests	101
5.3. Strong superiority tests	103
5.4. Non-inferiority tests	106
5.5. p -value notation	109
6. TESTS WITH MORE THAN ONE VARIANT	111
6.1. Type I error in an A/B/n test	111
6.2. p -value and CI corrections for testing multiple variants	114
6.3. Sample size calculations for A/B/n tests	128
6.4. Dynamically dropping or adding variants	122

6.5. Do factorial designs make sense in online A/B testing?	123
6.6. Testing the perfect shade of blue	128
7. SEGMENTATION AND MULTIPLE PERFORMANCE INDICATORS ...	131
7.1. The Šidák correction	131
7.2. Analyses of segments of the sample of an online experiment ...	132
7.3. Designs with more than one primary parameter of interest ...	135
7.4. Designs with secondary parameters of interest	138
8. WORKING WITH CONTINUOUS DATA	139
8.1. Standard deviation, p-values, and confidence intervals	139
8.2. Statistical power and sample size calculations	141
8.3. Is the Normal distribution assumption adequate?	143
8.4. A workaround for incomplete ARPU data	146
9. PERCENTAGE CHANGE	149
9.1. Percentage change (lift) vs. absolute change	149
9.2. Confidence intervals for percentage change	153
9.3. <i>p</i> -values for percentage change	154
9.4. Sample size calculations for percentage change	155
10. SEQUENTIAL TESTING: CONTINUOUS MONITORING OF DATA ...	157
10.1. The issue of repeated significance tests on accumulating data ...	158
10.2. The sequential probability ratio test	161
10.3. Fixed analysis time group sequential trials	164
10.4. Alpha-spending functions and efficacy boundaries	167
10.5. Beta-spending and futility boundaries	171
10.6. Expected sample size and efficiency of sequential A/B tests ...	173
10.7. Estimation following a group sequential A/B test	181
11. OPTIMAL SIGNIFICANCE THRESHOLDS AND SAMPLE SIZES	187
11.1. Defining “success” for a business experiment	188
11.2. Costs, benefits, risks, and rewards in A/B testing	191

11.3. Test parameters and their relationship to costs and benefits	195
11.4. Distribution of expected effect sizes	199
11.5. Calculating risk/reward ratios and key points	203
11.6. Testing with 50% confidence threshold?	209
11.7. Inherent cost of A/B testing	211
11.8. Limitations of Risk/Reward calculations	214
12. EXTERNAL VALIDITY a.k.a. GENERALIZABILITY OF A/B TEST RESULTS	217
12.1. What is external validity (generalizability)?	217
12.2. Threats to the external validity of online experiments	219
12.3. Improving the generalizability of A/B test results	223
12.4. Representative samples and sequential tests	224
12.5. Running multiple concurrent A/B tests	226
13. MISCELLANEOUS TOPICS	233
13.1. Equal or unequal allocation between test groups?	233
13.2. Holdout groups	236
13.3. Time to event analysis. Hazard ratio	238
13.4. Meta-analyses of A/B test results	242
13.5. Adaptive Designs	245
13.6. A word on multi-armed bandits	247
13.7. Bayesian methods	249
14. COMMUNICATING STATISTICAL RESULTS	253
14.1. Changing the perception of data variability	254
14.2. Translating business questions into statistical models	256
14.3. Presenting statistical results	259
EPILOGUE	269
REFERENCES	271
INDEX	281

MOTIVATION

The most straightforward way to explain why this book exists is via a brief description of my journey from a statistical know-nothing to an author of a book on statistics.

Some years back, I set out to learn more about the application of state-of-the-art scientific methods to the business world of data-driven decision-making. Starting with analyses of observational data, I quickly shifted my focus to online controlled experiments. These are commonly referred to as A/B tests, or split tests.

At the time, I had no formal training in statistics, and only college-level understanding of mathematics, so, to be honest, I didn't even know where to start! Available books on A/B testing barely had anything to say about statistics, so I started reading online blog posts and educational resources from universities, such as online lectures and courses, as well as the odd scientific paper.

In doing so, I had to face an entirely **new jargon** full of counterintuitive terms, such as 'statistical significance', which has little to do with significance, 'statistical power' which has nothing to do with power in the casual sense, 'confidence interval', which has nothing to do with any kind of confidence, and so on. And, above all else, I had to familiarize myself with a notation full of small and capital Greek letters (α , β , γ , δ , θ , μ etc.), which would sometimes mean different things in different contexts, while different letters would also denote the same concept.

To make things worse, there were ample examples of **vague or conflicting information**. There were dozens of definitions for what a p-value is and how it should be interpreted. Almost nobody seemed to care to define what a family of hypotheses is supposed to be when discussing the Family-Wise Error Rate. One source would claim one-tailed tests are preferable, while others would swear by two-tailed tests, and scare you with the heavens coming down on you if you were so reckless as to consider a one-tailed test.

Practitioners and academics alike were battling over which approach is best overall, or squabbling over the merits of particular applications - frequentist inference vs. decision-theoretic vs. Bayesian approaches. To make matters

even more confusing, there seemed to be noticeable schisms within each school of thought.

Most confusing of all, statistics as such turned out to be very context-dependent - it meant different things in different scientific and business fields. Practitioners in those fields had, over time, developed somewhat separate branches of statistics. Therefore, statistics would mean something different for you depending on whether you come from physics, medicine, social studies, econometrics, environmental studies, or industrial quality control.

It was simply a nightmare attempting to navigate this fractured jungle of jargon, conflicting stances, and math-heavy explanations. Yet, I persevered! And through painstaking reading, practice, implementing/coding methods, and countless simulation runs, I was able to garner a good enough understanding of the matter to begin writing methodological white papers and in-depth articles, to start delivering lectures and courses on statistics in A/B testing, and to become a developer of statistical tools.

From my current position, I see both the immense **value** of statistical methods applied to business risk management, estimation and prediction problems, and the immense **harm** done by improper applications or misguided understanding of those same methods. Thus, in-depth explanations of the practical application of statistical methods, as well as common errors and how to avoid them, are key elements of this work.

Furthermore, in 2019 the difficulties that I went through are about as severe as they were a few years before, despite the valiant efforts of some in the statistics and A/B testing communities. Addressing common mistakes, misconceptions, and misapplications of statistical methods is, therefore, a central part of this work.

My aim with this book is to **carve a clear path through the statistical jungle**, and thus save the reader weeks, months or even years of wandering around in circles, falling into gorges, and crossing rivers, metaphorically speaking! While the book does use the established jargon, each term is explained with painstaking detail and accuracy using the simplest language possible. Math and formulas are kept to a sanitary minimum in order to facilitate reading, while also satisfying the needs of technically-inclined readers, who will also find the detailed references supporting each chapter to be particularly useful. Since this is a book on statistical methods, and not on decision theory per se, the text also sticks to the frequentist error-statistical approach, and only briefly touches on current decision-theoretic and Bayesian methods.

WHO IS THE READER OF THIS BOOK?

This book aims to introduce the complex topic of statistical estimation and inference to readers with somewhere between little and no mathematical and statistical background. The text makes few assumptions, and builds each topic from the ground up, explaining the rationale behind each concept, and following it with a multitude of practical examples from the world of online A/B testing. It contains detailed explanations, so that one can understand the statistical methods deeply enough in order to correctly put them into practice, but steers clear of some of the difficult parts of set theory, calculus, etc., which are typical of many other books on statistics.

A background in conversion rate optimization, operating an online business or mobile app, design of user experiences, or similar, would be helpful for an easier reading, as the primary audience for this work is conversion rate optimization professionals who design, execute, and analyze A/B tests in an online environment. However, due to the similarities with other fields where controlled experimentation is possible and valuable, the book can be a useful guide to A/B testing in areas other than website and mobile application development. Product managers and growth experts should benefit from it regardless of the particular product or service they are focused on.

The overall framing of the presentation is always mindful of the topic of business objectives achieved through statistical methods. This will be useful for those readers who have some experience of statistics in other disciplines, and who are now looking to understand the use of statistical tools in facilitating decision-making through online A/B testing.

However, I must emphasize that the unique research presented in this book, in terms of ideas, models, and simulation results, will be valuable to all readers, regardless of background.