STATISTICS DONE WRONG

THE WOEFULLY COMPLETE GUIDE

ALEX REINHART

no starch press

STATISTICS DONE WRONG

STATISTICS DONE WRONG

The Woefully Complete Guide

by Alex Reinhart

no starch press

San Francisco

19 18 17 16 15   1 2 3 4 5 6 7 8 9

ISBN-10: 1-59327-620-6
ISBN-13: 978-1-59327-620-1

Publisher: William Pollock
Production Editor: Alison Law
Cover Illustration: Josh Ellingson
Developmental Editors: Greg Poulos and Leslie Shen
Technical Reviewer: Howard Seltman
Copyeditor: Kim Wimpsett
Compositor: Alison Law
Proofreader: Emelie Burnette

For information on distribution, translations, or bulk sales,
please contact No Starch Press, Inc. directly:

No Starch Press, Inc.
245 8th Street, San Francisco, CA 94103
phone: 415.863.9900; info@nostarch.com
www.nostarch.com

Library of Congress Cataloging-in-Publication Data

```
Reinhart, Alex, 1991-
  Statistics done wrong : the woefully complete guide / by Alex Reinhart.
       pages cm
  Includes index.
  Summary: "Discusses how to avoid the most common statistical errors
  in modern research, and perform more accurate statistical analyses"
  - Provided by publisher.
  ISBN 978-1-59327-620-1 - ISBN 1-59327-620-6
 1.  Statistics-Methodology. 2.  Missing observations (Statistics)  I. Title.
  QA276.R396 2015
  519.5-dc23
                                                          2015002128
```

The first principle is that you must not fool yourself,
and you are the easiest person to fool.

—Richard P. Feynman

To consult the statistician after an experiment is finished is
often merely to ask him to conduct a post mortem examination.
He can perhaps say what the experiment died of.

—R.A. Fisher

About the Author

Alex Reinhart is a statistics instructor and PhD student at Carnegie Mellon University. He received his BS in physics at the University of Texas at Austin and does research on locating radioactive devices using physics and statistics.

BRIEF CONTENTS

Preface xv
Introduction 1
Chapter 1: An Introduction to Statistical Significance 7
Chapter 2: Statistical Power and Underpowered Statistics 15
Chapter 3: Pseudoreplication: Choose Your Data Wisely 31
Chapter 4: The p Value and the Base Rate Fallacy 39
Chapter 5: Bad Judges of Significance 55
Chapter 6: Double-Dipping in the Data 63
Chapter 7: Continuity Errors 73
Chapter 8: Model Abuse 79
Chapter 9: Researcher Freedom: Good Vibrations? 89
Chapter 10: Everybody Makes Mistakes 97
Chapter 11: Hiding the Data 105
Chapter 12: What Can Be Done? 119
Notes 131
Index 147

CONTENTS IN DETAIL

PREFACE **xv**

Acknowledgments xvii

INTRODUCTION **1**

1
AN INTRODUCTION TO STATISTICAL SIGNIFICANCE **7**

The Power of p Values 8
Psychic Statistics 10
Neyman-Pearson Testing 11
Have Confidence in Intervals 12

2
STATISTICAL POWER AND UNDERPOWERED STATISTICS **15**

The Power Curve 15
The Perils of Being Underpowered 18
Wherefore Poor Power? 20
Wrong Turns on Red 21
Confidence Intervals and Empowerment 22
Truth Inflation 23
Little Extremes 26

3
PSEUDOREPLICATION: CHOOSE YOUR DATA WISELY **31**

Pseudoreplication in Action 32
Accounting for Pseudoreplication 33
Batch Biology 34
Synchronized Pseudoreplication 35

4
THE P VALUE AND THE BASE RATE FALLACY **39**

The Base Rate Fallacy ... 40
A Quick Quiz ... 41
The Base Rate Fallacy in Medical Testing ... 42
How to Lie with Smoking Statistics ... 43
Taking Up Arms Against the Base Rate Fallacy ... 45
If At First You Don't Succeed, Try, Try Again ... 47
Red Herrings in Brain Imaging ... 51
Controlling the False Discovery Rate ... 52

5
BAD JUDGES OF SIGNIFICANCE **55**

Insignificant Differences in Significance ... 55
Ogling for Significance ... 59

6
DOUBLE-DIPPING IN THE DATA **63**

Circular Analysis ... 64
Regression to the Mean ... 67
Stopping Rules ... 68

7
CONTINUITY ERRORS **73**

Needless Dichotomization ... 74
Statistical Brownout ... 75
Confounded Confounding ... 76

8
MODEL ABUSE **79**

Fitting Data to Watermelons ... 80
Correlation and Causation ... 84
Simpson's Paradox ... 85

9
RESEARCHER FREEDOM: GOOD VIBRATIONS? **89**

A Little Freedom Is a Dangerous Thing ... 91
Avoiding Bias ... 93

10
EVERYBODY MAKES MISTAKES 97

Irreproducible Genetics 98
Making Reproducibility Easy 100
Experiment, Rinse, Repeat 102

11
HIDING THE DATA 105

Captive Data 106
Obstacles to Sharing 107
Data Decay 108
Just Leave Out the Details 110
Known Unknowns 110
Outcome Reporting Bias 111
Science in a Filing Cabinet 113
Unpublished Clinical Trials 114
Spotting Reporting Bias 115
Forced Disclosure 116

12
WHAT CAN BE DONE? 119

Statistical Education 121
Scientific Publishing 124
Your Job 126

NOTES 131

INDEX 147

PREFACE

A few years ago I was an undergraduate physics major at the University of Texas at Austin. I was in a seminar course, trying to choose a topic for the 25-minute presentation all students were required to give. "Something about conspiracy theories," I told Dr. Brent Iverson, but he wasn't satisfied with that answer. It was too broad, he said, and an engaging presentation needs to be focused and detailed. I studied the sheet of suggested topics in front of me. "How about scientific fraud and abuse?" he asked, and I agreed.

In retrospect, I'm not sure how scientific fraud and abuse is a narrower subject than conspiracy theories, but it didn't matter. After several slightly obsessive hours of research, I realized that scientific fraud isn't terribly interesting—at least, not compared to all the errors scientists commit *unintentionally*.

Woefully underqualified to discuss statistics, I nonetheless dug up several dozen research papers reporting on the numerous statistical errors routinely committed by scientists, read and outlined them, and devised a presentation that satisfied Dr. Iverson. I decided that as a future scientist (and now a self-designated statistical pundit), I should take a course in statistics.

Two years and two statistics courses later, I enrolled as a graduate student in statistics at Carnegie Mellon University. I still take obsessive pleasure in finding ways to do statistics wrong.

Statistics Done Wrong is a guide to the more egregious statistical fallacies regularly committed in the name of science. Because many scientists receive no formal statistical training—and because I do not want to limit my audience to the statistically initiated—this book assumes no formal statistical training. Some readers may easily skip through the first chapter, but I suggest at least skimming it to become familiar with my explanatory style.

My goal is not just to teach you the names of common errors and provide examples to laugh at. As much as is possible without detailed mathematics, I've explained *why* the statistical errors are errors, and I've included surveys showing how common most of these errors are. This makes for harder reading, but I think the depth is worth it. A firm understanding of basic statistics is essential for everyone in science.

For those who perform statistical analyses for their day jobs, there are "Tips" at the end of most chapters to explain what statistical techniques you might use to avoid common pitfalls. But this is not a textbook, so I will not teach you how to use these techniques in any technical detail. I hope only to make you aware of the most common problems so you are able to pick the statistical technique best suited to your question.

In case I pique your curiosity about a topic, a comprehensive bibliography is included, and every statistical misconception is accompanied by references. I omitted a great deal of mathematics in this guide in favor of conceptual understanding, but if you prefer a more rigorous treatment, I encourage you to read the original papers.

I must caution you before you read this book. Whenever we understand something that few others do, it is tempting to find every opportunity to prove it. Should *Statistics Done Wrong* miraculously become a *New York Times* best seller, I expect to see what Paul Graham calls "middlebrow dismissals" in response to any science news in the popular press. Rather than taking the time to understand the interesting parts of scientific research, armchair statisticians snipe at news articles, using the vague description of the study regurgitated from some overenthusiastic university press release to criticize the statistical design of the research.*

This already happens on most websites that discuss science news, and it would annoy me endlessly to see this book used to justify it. The first comments on a news article are always complaints about how "they didn't control for this variable" and "the sample size is too small," and 9 times out of 10, the commenter never read the scientific paper to notice that their complaint was addressed in the third paragraph.

This is stupid. A little knowledge of statistics is not an excuse to reject all of modern science. A research paper's statistical methods can be judged only in detail and in context with the rest of its methods: study design, measurement techniques, cost constraints, and goals. Use your statistical knowledge to better understand the strengths, limitations, and potential biases of research, not to shoot down any paper that seems to misuse a *p* value or contradict your personal beliefs. Also, remember that a conclusion supported by poor statistics can still be correct—statistical and logical errors do not make a conclusion wrong, but merely unsupported.

In short, please practice statistics responsibly. I hope you'll join me in a quest to improve the science we all rely on.

Acknowledgments

Thanks to James Scott, whose statistics courses started my statistical career and gave me the background necessary to write this book; to Raye Allen, who made James's homework assignments much more fun; to Matthew Watson and Moriel Schottlender, who gave invaluable feedback and suggestions on my drafts; to my parents, who gave suggestions and feedback; to Dr. Brent Iverson, whose seminar first motivated me to learn about statistical abuse; and to all the scientists and statisticians who have broken the rules and given me a reason to write.

My friends at Carnegie Mellon contributed many ideas and answered many questions, always patiently listening as I tried to explain some new statistical error. My professors, particularly Jing Lei, Valérie Ventura, and Howard Seltman, prepared me with the necessary knowledge. As technical reviewer, Howard

*Incidentally, I think this is why conspiracy theories are so popular. Once you believe you know something nobody else does (the government is out to get us!), you take every opportunity to show off that knowledge, and you end up reacting to all news with reasons why it was falsified by the government. Please don't do the same with statistical errors.

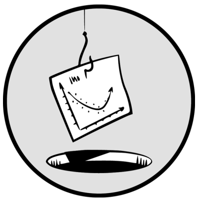
caught several embarrassing errors; if any remain, they're my responsibility, though I will claim they're merely in keeping with the title of the book.

My editors at No Starch dramatically improved the manuscript. Greg Poulos carefully read the early chapters and wasn't satisfied until he understood every concept. Leslie Shen polished my polemic in the final chapters, and the entire team made the process surprisingly easy.

I also owe thanks to the many people who emailed me suggestions and comments when the guide became available online. In no particular order, I thank Axel Boldt, Eric Franzosa, Robert O'Shea, Uri Bram, Dean Rowan, Jesse Weinstein, Peter Hozák, Chris Thorp, David Lovell, Harvey Chapman, Nathaniel Graham, Shaun Gallagher, Sara Alspaugh, Jordan Marsh, Nathan Gouwens, Arjen Noordzij, Kevin Pinto, Elizabeth Page-Gould, and David Merfield. Without their comments, my explanations would no doubt be less complete.

Perhaps you can join this list. I've tried my best, but this guide will inevitably contain errors and omissions. If you spot an error, have a question, or know a common fallacy I've missed, email me at *alex@refsmmat.com*. Any errata or updates will be published at *http://www.statisticsdonewrong.com/*.

INTRODUCTION



In the final chapter of his famous book *How to Lie with Statistics*, Darrell Huff tells us that "anything smacking of the medical profession" or backed by scientific laboratories and universities is worthy of our trust—not unconditional trust but certainly more trust than we'd afford the media or politicians. (After all, Huff's book is filled with the misleading statistical trickery used in politics and the media.) But few people complain about statistics done by trained scientists. Scientists seek understanding, not ammunition to use against political opponents.

Statistical data analysis is fundamental to science. Open a random page in your favorite medical journal and you'll be deluged with statistics: t tests, p values, proportional hazards models, propensity scores, logistic regressions, least-squares fits, and confidence intervals. Statisticians have provided scientists with tools of enormous power to find order and meaning in the most complex of datasets, and scientists have embraced them with glee.

They have not, however, embraced statistics *education*, and many undergraduate programs in the sciences require no statistical training whatsoever.

Since the 1980s, researchers have described numerous statistical fallacies and misconceptions in the popular peer-reviewed scientific literature and have found that many scientific papers—perhaps more than half—fall prey to these errors. Inadequate statistical power renders many studies incapable of finding what they're looking for, multiple comparisons and misinterpreted *p* values cause numerous false positives, flexible data analysis makes it easy to find a correlation where none exists, and inappropriate model choices bias important results. Most errors go undetected by peer reviewers and editors, who often have no specific statistical training, because few journals employ statisticians to review submissions and few papers give sufficient statistical detail to be accurately evaluated.

The problem isn't fraud but poor statistical education—poor enough that some scientists conclude that most published research findings are probably false.[1] Review articles and editorials appear regularly in leading journals, demanding higher statistical standards and tougher review, but few scientists hear their pleas, and journal-mandated standards are often ignored. Because statistical advice is scattered between frequently misleading textbooks, review articles in assorted journals, and statistical research papers difficult for scientists to understand, most scientists have no easy way to improve their statistical practice.

The methodological complexity of modern research means that scientists without extensive statistical training may not be able to understand most published research in their fields. In medicine, for example, a doctor who took one standard introductory statistics course would have sufficient knowledge to fully understand only about a fifth of research articles published in the *New England Journal of Medicine*.[2] Most doctors have even less training—many medical residents learn statistics informally through journal clubs or short courses, rather than through required courses.[3] The content that *is* taught to medical students is often poorly understood, with residents averaging less than 50% correct on tests of statistical methods commonly used in medicine.[4] Even medical school faculty with research training score less than 75% correct.

The situation is so bad that even the authors of surveys of statistical knowledge lack the necessary statistical knowledge to formulate survey questions—the numbers I just quoted are misleading because the survey of medical residents included a multiple-choice question asking residents to define a *p* value and gave four incorrect definitions as the only options.[5] We can give the authors some leeway because many introductory statistics textbooks also poorly or incorrectly define this basic concept.

When the designers of scientific studies don't employ statistics with sufficient care, they can sink years of work and thousands of dollars into research that cannot possibly answer the questions it is meant to answer. As psychologist Paul Meehl complained,

> Meanwhile our eager-beaver researcher, undismayed by logic-of-science considerations and relying blissfully on the "exactitude" of modern statistical hypothesis-testing, has produced a long publication list and been promoted to a full professorship. In terms of his contribution to the enduring body of psychological knowledge, he has done hardly anything. His true position is that of a potent-but-sterile intellectual rake, who leaves in his merry path a long train of ravished maidens but no viable scientific offspring.[6]

Perhaps it is unfair to accuse most scientists of intellectual infertility, since most scientific fields rest on more than a few misinterpreted *p* values. But these errors have massive impacts on the real world. Medical clinical trials direct our health care and determine the safety of powerful new prescription drugs, criminologists evaluate different strategies to mitigate crime, epidemiologists try to slow down new diseases, and marketers and business managers try to find the best way to sell their products—it all comes down to statistics. Statistics done wrong.

Anyone who's ever complained about doctors not making up their minds about what is good or bad for you understands the scope of the problem. We now have a dismissive attitude toward news articles claiming some food or diet or exercise might harm us—we just wait for the inevitable second study some months later, giving exactly the opposite result. As one prominent epidemiologist noted, "We are fast becoming a nuisance to society. People don't take us seriously anymore, and when they do take us seriously, we may unintentionally do more harm than good."[7] Our instincts are right. In many fields, initial results tend to be contradicted by later results. It seems the pressure to publish exciting results early and often has surpassed the responsibility to publish carefully checked results supported by a surplus of evidence.

Let's not judge so quickly, though. Some statistical errors result from a simple lack of funding or resources. Consider the mid-1970s movement to allow American drivers to turn right at red lights, saving gas and time; the evidence suggesting this would cause no more crashes than before was statistically flawed, as you will soon see, and the change cost many lives. The only factor holding back traffic safety researchers was a lack of data. Had they the money to collect more data and perform more studies—and the time to collate results from independent researchers in many different states—the truth would have been obvious.

While Hanlon's razor directs us to "never attribute to malice that which is adequately explained by incompetence," there are some published results of the "lies, damned lies, and statistics" sort. The pharmaceutical industry seems particularly tempted to bias evidence by neglecting to publish studies that show their drugs do not work;* subsequent reviewers of the literature may be pleased to find that 12 studies indicate a drug works, without knowing that 8 other unpublished studies suggest it does not. Of course, it's likely that such results would not be published by peer-reviewed journals even if they were submitted—a strong bias against unexciting results means that studies saying "it didn't work" never appear and other researchers never see them. Missing data and publication bias plague science, skewing our perceptions of important issues.

Even properly done statistics can't be trusted. The plethora of available statistical techniques and analyses grants researchers an enormous amount of freedom when analyzing their data, and it is trivially easy to "torture the data until it confesses." Just try several different analyses offered by your statistical software until one of them turns up an interesting result, and then pretend this is the analysis you intended to do all along. Without psychic powers, it's almost impossible to tell when a published result was obtained through data torture.

In "softer" fields, where theories are less quantitative, experiments are difficult to design, and methods are less standardized, this additional freedom causes noticeable biases.[8]

*Readers interested in the pharmaceutical industry's statistical misadventures may enjoy Ben Goldacre's *Bad Pharma* (Faber & Faber, 2012), which caused a statistically significant increase in my blood pressure while I read it.

Researchers in the United States must produce and publish interesting results to advance their careers; with intense competition for a small number of available academic jobs, scientists cannot afford to spend months or years collecting and analyzing data only to produce a statistically insignificant result. Even without malicious intent, these scientists tend to produce exaggerated results that more strongly favor their hypotheses than the data should permit.

In the coming pages, I hope to introduce you to these common errors and many others. Many of the errors are prevalent in vast swaths of the published literature, casting doubt on the findings of thousands of papers.

In recent years there have been many advocates for statistical reform, and naturally there is disagreement among them on the best method to address these problems. Some insist that *p* values, which I will show are frequently misleading and confusing, should be abandoned altogether; others advocate a "new statistics" based on confidence intervals. Still others suggest a switch to new Bayesian methods that give more-interpretable results, while others believe statistics as it's currently taught is just fine but used poorly. All of these positions have merits, and I am not going to pick one to advocate in this book. My focus is on statistics as it is currently used by practicing scientists.