BAYESIAN DATA ANALYSIS FOR THE BEHAVIORAL AND NEURAL SCIENCES

Non-Calculus Fundamentals

This textbook bypasses the need for advanced mathematics by providing in-text computer code, allowing students to explore Bayesian data analysis without the calculus background normally considered a prerequisite for this material. Now, students can use the best methods without needing advanced mathematical techniques. This approach goes beyond 'frequentist' concepts of *p*-values and null hypothesis testing, using the full power of modern probability theory to solve real-world problems. The book offers a fully self-contained course, and demonstrates analysis techniques using over 100 worked examples crafted specifically for students in the behavioral and neural sciences. The book presents two general algorithms that help students solve the measurement and model selection (also called 'hypothesis testing') problems most frequently encountered in real-world applications.

Todd E. Hudson is a professor of rehabilitation medicine at New York University's Grossman School of Medicine, holding cross-appointments in neurology, and also in the Department of Biomedical Engineering at the New York University Tandon School of Engineering. Dr. Hudson has taught statistics, perception and sensory processes, experimental design, and/or advanced topics in neurobiology and behavior at several major universities, including Brandeis University and Columbia University. He co-founded, and serves as Chief Scientific Advisor to, Tactile Navigation Tools, LLC, which develops navigation aids for the visually impaired.

Bayesian Data Analysis for the Behavioral and Neural Sciences

NON-CALCULUS FUNDAMENTALS

TODD E. HUDSON

New York University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108835565
DOI: 10.1017/9781108890984

First published 2021

Printed in the United Kingdom by TJ Books Limited, Padstow, Cornwall 2021

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Hudson, Todd Erik, author.
Title: Bayesian data analysis for the behavioral and neural sciences : non-calculus fundamentals / Todd E. Hudson.
Description: Cambridge ; New York, NY : Cambridge University Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020052398 | ISBN 9781108835565 (hardback) | ISBN 9781108812900 (paperback) | ISBN 9781108890984 (ebook)
Subjects: LCSH: Social sciences – Statistical methods. | Social sciences – Data processing. | Psychology – Statistical methods. | Psychology – Data processing. | Neurosciences – Statistical methods. | Neurosciences – Data processing. | Bayesian statistical decision theory.
Classification: LCC HA29 .H813 2021 | DDC 519.5/42–dc23
LC record available at https://lccn.loc.gov/2020052398

ISBN 978-1-108-83556-5 Hardback
ISBN 978-1-108-81290-0 Paperback

Additional resources for this publication at www.cambridge.org/hudsonbayesian

To my wife, Elizabeth, and my sons,
Gabriel William and Julian Leonard,
who showed me there was even more to life

Cambridge University Press
978-1-108-83556-5 — Bayesian Data Analysis for the Behavioral and Neural Sciences
Todd E. Hudson
Frontmatter
More Information

Contents

Preface

A scientist should live two lifetimes: one studying the empirical results and experimental methods of a scientific or medical specialty area, and the other spent studying the art and science of extracting information from data. In this text I will assume that you have made a start on the first, and have learned something of neuroscience, education, economics, psychology, and/or medical research. Here, we will take the first steps toward completing your training, and begin a study of modern data analysis.

Why is data analysis so important? The reason for the extreme importance of data analysis is twofold: First, a working knowledge of the logically valid techniques for extracting information from data will naturally provide insights into experimental design, and allow you to create experiments that are well suited to answering the scientific questions in which you are interested. Second, it gives you the ability to take the data you have obtained, and efficiently extract *all* of the information contained therein. Neither of these is truly possible without a solid understanding of data analysis, and by extension probability.

This text is meant to be either a first or a second exposure to data analysis. First contact occasionally occurs prior to the university level, and occasionally waits until graduate school. As a first exposure (that is, prior to any other course that would have need of the concept of probability), this text is meant to provide a self-contained treatment of the basics of data analysis ***without the need for calculus***.[1] This, in fact, was my motivation for writing a textbook at all. There are certainly other excellent, probably more

[1] Here is a quick test to know if you have the mathematics background for this text:

Choose the lowest row of the table with an equation you can confidently solve:	
$x + x$	(a) $2x$ (b) $x/2$ (c) x^2 (d) x^{-2}
$\sum_{i=1}^{5} 2i$	(a) 2 (b) 30 (c) 5 (d) 10
$e^0 + \ln(1)$	(a) 0+1 (b) $1 - 1$ (c) 1+0 (d) $-1 + 1$
$\int d\theta \cos(\theta) + \int d\vartheta \sec^2(\vartheta)$	(a) $\sin\theta + \cos\vartheta$ (b) $\cos - \sin\vartheta$ (c) $\tan(\theta) - \sin(\vartheta)$ (d) $\sin(\theta) + \tan(\vartheta)$

The correct answers progress from (a) to (d) by row from top to bottom. If you can solve all rows' equations, you are ready for the current text, including the optional calculus-based material. If you solved the first three, you are fully prepared for the current text. If you solved only one equation from the first or second row, but were familiar with the elements of the equation on the third row (even if you could not immediately solve it), you should do fine with this text after a little refamiliarization with the concept of a logarithm (a refresher is provided in **Appendix B**). If you were unable to solve either of the first two rows' equations, you will need to take an algebra and/or precalculus course before attempting this text.

complete and certainly more rigorously mathematical textbooks on the topics of probability and Bayesian data analysis already available. They are simply not appropriate as a first exposure because they require advanced calculus *as their starting point*.

If this is *not* your first course in data analysis, it is likely that first contact was provided by a 'frequentist statistics' course (most typically a single-term course in the second university year, although such courses are now more common at the preuniversity level). These courses focus on memorizing a series of 'frequentist statistical tests,' with little or no attempt at explaining their origins, along with a cookbook-style recipe for application; for example, *if your data consist of two sets of continuous observations, you test for a difference in their means using thus-and-so statistical test; if the result is greater than a particular value, the meaning of the frequentist test is....* Because the cookbook approach avoids any attempt at explaining the origins of tests (the most commonly used explanations require advanced mathematics), there is always a great deal of confusion on the part of students. This has, inevitably, led to large-scale misuse of frequentist statistical tests – this is true even in a professional setting, because the majority of frequentist tests are computed via software package, where the user of that software need not be aware of the subtleties that went into producing the output, or whether the match between analysis and data was truly appropriate. This has led to no end of problems and confusion, as you might imagine. Confusion is exacerbated by the fact that frequentist statistics courses are based on a definition of probability that is incompatible with modern probability theory, and that ultimately leads to logical inconsistency.[2]

To avoid the problems of a frequentist statistics course and still provide a text that can be used by undergraduate students, I have done three things: First, I use the modern formulation of the concept of probability, which allows a unified and logical treatment of both probability theory and data analysis. This allows also for straightforward explanation of the origins of each of the resulting concepts and analyses. Second, I avoid the requirement of advanced calculus-based methods. Third, in lieu of advanced mathematics[3] I make use of computer-based approximation techniques. Approximate techniques are suitable both for the student who has not yet taken calculus, as well as the professional performing calculations for which no analytical solution exists; approximate techniques are therefore useful regardless of your level of mathematics training. Although many of the same mathematical topics are covered in several other excellent texts on probability and data analysis, applications here are geared toward students in the behavioral, neural, and biomedical sciences (i.e., psychology, education, neuroscience, biomedical engineering, economics, biology, and medicine).

For students who have taken a frequentist statistics course, it is my hope that the information presented here will provide a useful counterpoint to what you have been

2 One such inconsistency, the 'optional stopping problem,' is demonstrated in Chapter 2.

3 Although I have taken pains to eliminate calculus-based techniques in this text (which constitutes a self-contained course in the basic elements of Bayesian data analysis), you should still be prepared for what is to come: there will be equations. Familiarity with the use of equations is fundamental for any real understanding of data analysis, even in a non-calculus introduction. Indeed, since data analysis and probability theory are topics in mathematics, there will in fact be a fair number of equations. If you are not yet comfortable with them, just keep in mind that there are no equations in this text that will require anything of you beyond basic arithmetic (i.e., addition, subtraction, multiplication, division, and by extension exponents and logarithms); a refresher of exponents and logarithms is provided in **Appendix B**, and an in-depth demonstration of how to solve marginalization equations in terms of sums and multiplications can be found in **Appendix C**.

previously taught, and give you the skills necessary to perform the analyses most suited to extracting information from your experimental data. With this in mind, I have attempted to provide explicit comparisons between modern methods and some of those taught in an old-fashioned frequentist statistics course, at least to the extent possible without becoming bogged down in digression.

What, you should ask, makes this text different and useful? Aside from presenting techniques of Bayesian data analysis in a non-calculus context, I have also based my presentation of the course material on *applying* the concepts to realistic problems of data analysis. This is very different from the commonly seen 'cookbook approach,' where the student's goal is to memorize a flowchart in which the data and experiment are put into categories (e.g., 'a multifactorial design with repeated measurements'), and this categorization dictates the test to be performed: the focus is therefore on memorizing the mapping from experiment and categorization to frequentist test.[4] Here, we will focus on the *process* through which one must pass in understanding a dataset, rather than in presenting a set of ready-made equations to memorize and use in predefined circumstances. To this end, the culmination of this text is a pair of data-analytic algorithms that are meant to *guide our thinking* when analyzing data. The steps of these algorithms will guide us as we construct each analysis, whether or not the data conform to a prepackaged set of descriptors (e.g., the deliberately mysterious 'repeated-measures multifactorial design' mentioned above). The initial chapters will focus on building these algorithms up from Bayes' theorem, and the final chapters demonstrate their use in a set of worked problems.

Two final points must be made before we begin: First, data analysis is always an exercise in logic. This logic is fundamental to the portion of the scientific enterprise that uses data of any variety, because it provides the set of rules that tell us what can be reasonably concluded from the data we have collected. Probability theory encodes this logic, and gives the mathematics that allows us to optimally extract information from data; this is in part a purely formal exercise, and in part based on our information about the data and the scientific hypotheses under investigation. Second (and finally), I must emphasize that the analysis of data is *not* entirely a mechanical exercise (i.e., 'plug-and-chug,' 'turn-the-crank'): it will require some thought. In consequence, you will sometimes struggle to construct the best analysis, and you will sometimes make mistakes. To understand your data and analyze it fully, you must be willing to try, make mistakes, revise, and

[4] This 'cookbook approach' has led to the proliferation of frequentist statistical software packages that are meant to be used by those without the requisite training to understand their inner workings (rather than simply to relieve a trained expert of the need to perform seemingly endless tedious calculations). In one recent example that appeared in my e-mail, the package was advertised with the headline: *You don't need to be a statistician to perform statistical analysis on your data*, which I assume is meant to imply that you don't need to understand statistical data analysis to generate results from the software. This view is supported by their further claims regarding the product:
- *Ease of use!*
- *Stat Wizard runs the right tests*
- *Interprets the results for you*
- *Creates charts to visualize results*
- *Accurate results without worry*
Apparently, one need only feed in numbers, and the software will do all the work ... including **interpret the results for you!** The software is meant to generate reports that look as if they were produced by a knowledgable expert who had thought through and tailored an analysis to the problem at hand, rather than automatic outputs that could have been generated by a monkey with a keyboard. Such statistical packages are the second ingredient in a recipe for disaster. That disaster is currently playing itself out in the scientific journals, and is a major contributor to the 'crisis of reproducibility' in the behavioral and neural sciences.

try again. There is an increasingly persistent misperception, even among professionals, that a 'good' scientist should never make mistakes – a belief that leads to a constellation of negative consequences, the most insidious being a fear of simply *getting started and trying something*. We will learn in the following pages, however, that crafting just the right analysis for the scientific question we are interested in answering will generally require trial-and-error. Trial-and-error is fundamental to the actual practice of the professional scientist; it is how we learn what works in the real world, on real datasets, and allows us to hone our skills and further the scientific enterprise. The good scientist must learn to understand the logic that connects data to theory and theory to experiment, and you will find a good deal of trial and error along the way. I encourage you to keep this in mind as you begin your journey into the sometimes challenging, but ever rewarding study of data analysis.

Acknowledgments

I am grateful to my colleagues for their interest in my work, especially Leonard Matin, Michael Landy, Kateri McRae, Peter Sokol-Hessner, John Martone, Janet Rucker, and JR Rizzo, who encouraged me during the writing process and motivated me to simplify and focus my presentation of the material.