CHAPTER 1

What's in This Book (Read This First!)

Contents

Oh honey I'm searching for love that is true,
But driving through fog is so dang hard to do.
Please paint me a line on the road to your heart,
I'll rev up my pick up and get a clean start.[1]

1.1. REAL PEOPLE CAN READ THIS BOOK

This book explains how to actually *do* Bayesian data analysis, by real people (like you), for realistic data (like yours). The book starts at the basics, with elementary notions of probability and programming. You do not need to already know statistics and programming. The book progresses to advanced hierarchical models that are used in realistic data analysis. This book is speaking to a person such as a first-year graduate student or advanced undergraduate in the social or biological sciences: Someone who grew up in Lake Wobegon,[2] but who is not the mythical being that has the previous training of a nuclear physicist and then decided to learn about Bayesian statistics. (After the publication of the first edition, one of those mythical beings contacted me about

[1] This chapter provides a road map to the book, which hopes to have you fall in love with Bayesian analysis even if you previously had unhappy relationships with statistics. The poem plays with those ideas.

[2] A popular weekly radio show on National Public Radio, called *A Prairie Home Companion*, features fictional anecdotes about a small town named Lake Wobegon. The stories, written and orated by Garrison Keillor, always end with the phrase, "And that's the news from Lake Wobegon, where all the women are strong, all the men are good looking, and all the children are above average." So, if you grew up there, …

Doing Bayesian Data Analysis, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.00001-5

the book! So, even if you *do* have the previous training of a nuclear physicist, I hope the book speaks to you too.)

Details of prerequisites and the contents of the book are presented below. But first things first: As you may have noticed from the beginning of this chapter, the chapters commence with a stanza of elegant and insightful verse composed by a famous poet. The quatrains[3] are formed of dactylic[4] tetrameter,[5] or, colloquially speaking, "country waltz" meter. The poems regard conceptual themes of the chapter via allusion from immortal human motifs in waltz timing.

> If you do not find them to be all that funny,
> If they leave you wanting back all of your money,
> Well, honey, some waltzing's a small price to pay, for
> All the good learning you'll get if you stay.

1.1.1. Prerequisites

There is no avoiding mathematics when doing data analysis. On the other hand, this book is definitely not a mathematical statistics textbook, in that it does *not* emphasize theorem proving or formal analyses.[6] But I *do* expect that you are coming to this book with a dim knowledge of basic calculus. For example, if you understand expressions like $\int x \, dx = \frac{1}{2}x^2$, you're probably good to go. Notice the previous sentence said "understand" the statement of the integral, not "generate" the statement on your own. When mathematical derivations are helpful for understanding, they will usually be presented with a thorough succession of intermediate steps, so you can actually come away feeling secure and familiar with the trip and destination, rather than just feeling car sick after being thrown blindfolded into the back seat and driven around curves at high speed.

The beginnings of your journey will go more smoothly if you have had some basic experience programming a computer, but previous programming experience is not crucial. A computer program is just a list of commands that the computer can execute.

[3] *quatrain* [noun]: Four lines of verse. Unless it's written "*qua* train," in which case it's a philosopher comparing something to a locomotive.

[4] *dactylic* [adjective]: A metrical foot in poetry comprising one stressed and two unstressed syllables. Not to be confused with a pterodactyl, which was a flying dinosaur, and which probably sounded nothing like a dactyl unless it fell from the sky and bounced twice: THUMP-bump-bump.

[5] *tetrameter* [noun]: A line of verse containing four metrical feet. Not to be confused with a quadraped, which has four feet, but is averse to lines.

[6] The first edition mentioned at this point that "any mathematical statistician would be totally bummed at the informality, dude." The statement was meant to be funny, with the slang self-referentially instantiating informality. Even the oracle of truth, Wikipedia, says that "'Dude' is generally used informally to address someone" (http://en.wikipedia.org/wiki/Dude, retrieved February 02, 2014) and "[slang] lowers, if temporarily, 'the dignity of formal or serious speech or writing'" (http://en.wikipedia.org/wiki/Slang, retrieved February 02, 2014). But some readers were offended by such undignified writing, and therefore the joke is now only available to people who read footnotes.

For example, if you've ever typed an equal sign in an Excel spreadsheet cell, you've written a programming command. If you've ever written a list of commands in Java, C, Python, Basic or any other computer programming language, then you're set. We will be using programming languages called R, JAGS, and Stan, which are free and thoroughly explained in this book.

1.2. WHAT'S IN THIS BOOK

This book has three major parts. The first part covers foundations: The basic ideas of Bayesian reasoning, models, probabilities, and programming in R.

The second part covers all the crucial ideas of modern Bayesian data analysis while using the simplest possible type of data, namely dichotomous data such as agree/disagree, remember/forget, male/female, etc. Because the data are so simplistic, the focus can be on Bayesian techniques. In particular, the modern techniques of "Markov chain Monte Carlo" (MCMC) are explained thoroughly and intuitively. Because the data are kept simple in this part of the book, intuitions about the meaning of hierarchical models can be developed in glorious graphic detail. This second part of the book also explores methods for planning how much data will be needed to achieve a desired degree of precision in the conclusions, broadly known as "power analysis."

The third part of the book applies the Bayesian methods to realistic data. The applications are organized around the type of data being analyzed, and the type of measurements that are used to explain or predict the data. Different types of measurement scales require different types of mathematical models, but otherwise the underlying concepts are always the same. More details of coverage are provided below.

The chapters of the book are designed to be read in order, for a "grand tour" of basic applied Bayesian analysis. Especially through parts one and two, the chapters probably make the most sense if read in order. But shorter routes are possible, as described next.

1.2.1. You're busy. What's the least you can read?

Here is a minimalist excursion through the book:

- Chapter 2: The idea of Bayesian inference and model parameters. This chapter introduces important concepts; don't skip it.
- Chapter 3: The R programming language. Read the sections about installing the software, including the extensive set of programs that accompany this book. The rest can be skimmed and returned to later when needed.
- Chapter 4: Basic ideas of probability. Merely skim this chapter if you have a high probability of already knowing its content.
- Chapter 5: Bayes rule!
- Chapter 6: The simplest formal application of Bayes rule, referenced throughout the remainder of the book.

- Chapter 7: MCMC methods. This chapter explains the computing method that makes contemporary Bayesian applications possible. You don't need to study all the mathematical details, but you should be sure to get the gist of the pictures.
- Chapter 8: The JAGS programming language for implementing MCMC.
- Chapter 16: Bayesian estimation of two groups. All of the foundational concepts from the aforementioned chapters, applied to the case of comparing two groups.

1.2.2. You're *really* busy! Isn't there even less you can read?

If all you want is a conceptual foundation and the fastest possible hands-on experience, and if you have some previous knowledge of classical statistics such as a t test, then I recommend the following. First, read Chapter 2 of this book for the conceptual foundation. Then read the article by Kruschke (2013a), which describes Bayesian estimation of two groups (analogous to a traditional t test). Essentially, you've just leapfrogged to Chapter 16 of this book. For your hands-on experience, the article has accompanying software, and there is a version that has been implemented in JavaScript for use in your web browser without need to install other software. For details, see the Web site http://www.indiana.edu/~kruschke/BEST/.

1.2.3. You want to enjoy the view a little longer. But not *too* much longer

After the minimalist excursion suggested above, if you want to delve into further specific applications, you will need to read these sections:

- Chapter 9: Hierarchical models. Many realistic applications involve hierarchical, or "multilevel," structure. One of the things that makes Bayesian methods so exciting is their seamless applicability to hierarchical models.
- Chapter 15: Overview of the generalized linear model. To know what type of model might apply to your data, you need to know the canonical catalog of conventional models, many of which fall under the umbrella of the generalized linear model.
- Individual chapters from Chapters 16–24. Go to the chapter relevant to the data structure you're interested in (which you'll understand because you previously read Chapter 15).
- Chapter 13: Statistical power analysis and planning of research, from a Bayesian perspective. This chapter is not essential on a first reading, but it's important not to skip forever. After all, failing to plan is planning to fail.
- Section 25.1, which has recommendations for how to report a Bayesian analysis. If you want your research to influence other people, you've got to be able to tell them about it. (Well, I suppose there are other means of persuasion, but you'll have to learn those from other sources.)

1.2.4. If you just gotta reject a null hypothesis…

Traditional statistical methods are often focused on rejecting a null hypothesis, as opposed to estimating magnitudes and their uncertainty. For a Bayesian perspective on null hypotheses, read these chapters:

- Chapter 11: The perils of p values in traditional null-hypothesis significance testing.
- Chapter 12: Bayesian approaches to null value assessment.

1.2.5. Where's the equivalent of traditional test X in this book?

Because many readers will be coming to this book after having already been exposed to traditional statistics that emphasize null hypothesis significance testing (NHST), this book provides Bayesian approaches to many of the usual topics in NHST textbooks. Table 1.1 lists various tests covered by standard introductory statistics textbooks, along with the location of their Bayesian analogues in this book.

The array of tests mentioned in Table 1.1 are all cases of what is called the "generalized linear model." For those of you already familiar with that term, you can glance ahead to Table 15.3, p. 444, to see which chapters cover which cases. For those of you not yet familiar with that term, do not worry, because all of Chapter 15 is devoted to introducing and explaining the ideas.

A superficial conclusion from Table 1.1 might be, "Gee, the table shows that traditional statistical tests do something analogous to Bayesian analysis in every case, therefore it's pointless to bother with Bayesian analysis." Such a conclusion would be wrong. First, traditional NHST has deep problems, some of which are discussed in Chapter 11. Second, Bayesian analysis yields richer and more informative inferences than NHST, as will be shown in numerous examples throughout the book.

Table 1.1 Bayesian analogues of null hypothesis significance tests.

Traditional analysis name	Bayesian analogue
Binomial test	Chapters 6–9 and 21
t test	Chapter 16
Simple linear regression	Chapter 17
Multiple linear regression	Chapter 18
One-way ANOVA	Chapter 19
Multifactor ANOVA	Chapter 20
Logistic regression	Chapter 21
Multinomial logistic regression	Chapter 22
Ordinal regression	Chapter 23
Chi-square test (contingency table)	Chapter 24
Power analysis (sample size planning)	Chapter 13

1.3. WHAT'S NEW IN THE SECOND EDITION?

The basic progression of topics remains the same as the first edition, but all the details have been changed, from cover to cover. The book and its programs have been completely rewritten. Here are just a few highlights of the changes:

- There are all new programs in JAGS and Stan. The new programs are designed to be much easier to use than the scripts in the first edition. In particular, there are now compact high-level scripts that make it easy to run the programs on your own data. This new programming was a major undertaking by itself.
- The introductory Chapter 2, regarding the basic ideas of how Bayesian inference re-allocates credibility across possibilities, is completely rewritten and greatly expanded.
- There are completely new chapters on the programming languages R (Chapter 3), JAGS (Chapter 8), and Stan (Chapter 14). The lengthy new chapter on R includes explanations of data files and structures such as lists and data frames, along with several utility functions. (It also has a new poem that I am particularly pleased with.) The new chapter on JAGS includes explanation of the RunJAGS package which executes JAGS on parallel computer cores. The new chapter on Stan provides a novel explanation of the concepts of Hamiltonian Monte Carlo. The chapter on Stan also explains conceptual differences in program flow between Stan and JAGS.
- Chapter 5 on Bayes' rule is greatly revised, with a new emphasis on how Bayes' rule re-allocates credibility across parameter values from prior to posterior. The material on model comparison has been removed from all the early chapters and integrated into a compact presentation in Chapter 10.
- What were two separate chapters on the Metropolis algorithm and Gibbs sampling have been consolidated into a single chapter on MCMC methods (as Chapter 7).
- There is extensive new material on MCMC convergence diagnostics in Chapters 7 and 8. There are explanations of autocorrelation and effective sample size. There is also exploration of the stability of the estimates of the highest density interval (HDI) limits. New computer programs display the diagnostics, as well.
- Chapter 9 on hierarchical models includes extensive new and unique material on the crucial concept of shrinkage, along with new examples.
- All the material on model comparison, which was spread across various chapters in the first edition, in now consolidated into a single focused chapter (Chapter 10) that emphasizes its conceptualization as a case of hierarchical modeling.
- Chapter 11 on null hypothesis significance testing is extensively revised. It has new material for introducing the concept of sampling distribution. It has new illustrations of sampling distributions for various stopping rules, and for multiple tests.
- Chapter 12, regarding Bayesian approaches to null value assessment, has new material about the region of practical equivalence (ROPE), new examples of accepting the

null value by Bayes factors, and new explanation of the Bayes factor in terms of the Savage-Dickey method.

- Chapter 13, regarding statistical power and sample size, has an extensive new section on sequential testing, and recommends making the research goal be precision of estimation instead of rejecting or accepting a particular value.
- Chapter 15, which introduces the generalized linear model, is fully revised, with more complete tables showing combinations of predicted and predictor variable types.
- Chapter 16, regarding estimation of means, now includes extensive discussion of comparing two groups, along with explicit estimation of effect size.
- Chapter 17, regarding regression on a single metric predictor, now includes extensive examples of robust regression in JAGS and Stan. New examples of hierarchical regression, including quadratic trend, graphically illustrate shrinkage in estimates of individual slopes and curvatures. The use of weighted data is also illustrated.
- Chapter 18, on multiple linear regression, includes a new section on Bayesian variable selection, in which various candidate predictors are probabilistically included in the regression model.
- Chapter 19, on one-factor ANOVA-like analysis, has all new examples, including a completely worked out example analogous to analysis of covariance (ANCOVA), and a new example involving heterogeneous variances.
- Chapter 20, on multi-factor ANOVA-like analysis, has all new examples, including a completely worked out example of a split-plot design that involves a combination of a within-subjects factor and a between-subjects factor.
- Chapter 21, on logistic regression, is expanded to include examples of robust logistic regression, and examples with nominal predictors.
- There is a completely new chapter (Chapter 22) on multinomial logistic regression. This chapter fills in a case of the generalized linear model (namely, a nominal predicted variable) that was missing from the first edition.
- Chapter 23, regarding ordinal data, is greatly expanded. New examples illustrate single-group and two-group analyses, and demonstrate how interpretations differ from treating ordinal data as if they were metric.
- There is a new section (25.4) that explains how to model censored data in JAGS.
- Many exercises are new or revised.

Oh, and did I mention that the cover is different? The correspondence of the doggies to Bayes' rule is now made explicit: The folded ears of the posterior doggie are a compromise between the perky ears and floppy ears of the likelihood and prior doggies. The marginal likelihood is not usually computed in MCMC methods, so the doggie in the denominator gets sleepy with nothing much to do. I hope that what's between the covers of this book is as friendly and engaging as the doggies on the cover.

1.4. GIMME FEEDBACK (BE POLITE)

I have worked thousands of hours on this book, and I want to make it better. If you have suggestions regarding any aspect of this book, please do email me: johnkruschke@gmail.com. Let me know if you've spotted egregious errors or innocuous infelicities, typo's or thoughto's. Let me know if you have a suggestion for how to clarify something. Especially let me know if you have a good example that would make things more interesting or relevant. I'm interested in complete raw data from research that is interesting to a broad audience, and which can be used with acknowledgement but without fee. Let me know also if you have more elegant programming code than what I've cobbled together. The outside margins of these pages are intentionally made wide so that you have room to scribble your ridicule and epithets before re-phrasing them into kindly stated suggestions in your email to me. Rhyming couplets are especially appreciated.

Since the publication of the first edition I have received hundreds of email messages from readers. I have replied to many, but I am truly embarrassed by the fact that I have not been able to reply to all of them, and some have gone unacknowledged. If I don't respond to your email in a timely manner, it is likely that your message got buried under the avalanche of subsequent emails. You are welcome to send a follow-up email to try to provoke a response from me. In any case, if your email is lengthy or has attachments or asks about a problem in some complicated analysis you are trying, chances are that I'll react by saying to myself, "That's very interesting, but I'll have to think about it when I have more time and reply later." Then, no matter what time it is, later never comes. Whether I am able to reply or not, I appreciate all your messages.

1.5. THANK YOU!

I would like to thank the many readers who posted reviews and recommendations of the first edition on sites such as Amazon.com, Goodreads.com, blogs, and social networking sites. At the time of this writing, there were 46 reviews on Amazon.com, 6 at Amazon.co.uk (United Kingdom), 2 at Amazon.ca (Canada), and 1 at Amazon.cn (China). There were 4 reviews at Goodreads.com plus many ratings. Numerous people have reviewed or recommended the book on their blogs. Many people have given a "shout out" to the book on social networking sites. I am very grateful to all of you for taking the time to write reviews, and pleased that your reviews have generally been very positive! I hope that this second edition elicits continued impulse to post new reviews about the second edition, as the revisions have been a huge effort to make the book even better.

I also thank the authors of professional reviews of first edition, including Andrews (2011), Barry (2011), Colvin (2013), Ding (2011), Fellingham (2012), Goldstein (2011), Smithson (2011), and Vanpaemel and Tuerlinckx (2012). My apologies to any reviewers whom I have missed; please let me know. I think it is valuable to raise the visibility of Bayesian methods in professional circles, and I am very grateful to all these authors for taking the time and effort to compose reviews.

Several people have written computer programs that extended or improved programs related to the first edition of the book. In particular, the programs for Bayesian estimation of two groups ("BEST"; Kruschke, 2013a, cf. Chapter 16 of this book) were repackaged in R by Mike Meredith, in JavaScript by Rasmus Bååth, and in Python by Andrew Straw. For links to their work, see http://www.indiana.edu/~kruschke/BEST/. Systems for creating hierarchical diagrams, akin to Figure 9.13 (p. 252), were created by Rasmus Bååth for LibreOffice and R, and by Tinu Schneider for LATEX and TikZ. For links to their work, see http://doingbayesiandataanalysis.blogspot.com/2013/10/diagrams-for-hierarchical-models-new.html. Thank you all for your extensive efforts and contributions to making Bayesian methods accessible to a wider audience.

Many colleagues have organized workshops or courses for *Doing Bayesian Data Analysis*. A list of workshops appears at https://sites.google.com/site/doingbayesiandataanalysis/. At each of those workshops many more people were involved than I can possibly mention by name here, but they include William Jacoby and Dieter Burrell at the Interuniversity Consortium for Political and Social Research Summer Program at the University of Michigan; William Pridemore, James Russell, James Walker, and Jeff DeWitt at the Indiana University Social Science Research Commons; Hans-Joachim Knopf at the University of St. Gallen Summer School in Empirical Research Methods, Switzerland; Hans Olav Melberg at the University of Oslo, Norway; Mark Nawrot and colleagues at North Dakota State University; Ulf Ahlstrom and colleagues at the Federal Aviation Administration Human Factors Lab, New Jersey; Tim Pleskac and colleagues at Michigan State University; John Curtin at the University of Wisconsin, Madison; Michael Roberts and colleagues including Chester Fornari, Bryan Hanson, and Humberto Barreto at DePauw University; Jan Andrews, Ming-Wen An, and colleagues at Vassar College; Kelly Goedert and colleagues at Seton Hall University; Gregory Francis and Zygmunt Pizlo at Purdue University; Boicho Kokinov and colleagues at the New Bulgarian University, Sofia; Nathalie Rothert and the workshop program committees at the Association for Psychological Science; Duncan Brumby and the tutorials program committees of the Cognitive Science Society; Andrew Delamater at the Eastern Psychological Association (and thanks to James McClelland for introducing my talk there); William Merriman at the Midwestern Psychological Association; Xiangen Hu and Michael Jones at the Society for Computers in Psychology. My apologies to those whom I have inadvertently left off this list. I thank you all for your time and effort. I hope that the workshops and courses genuinely facilitate doing Bayesian data analysis by the attendees.

The book has been used by a number of instructors in their courses, and a few of them have sent me notes of their experiences. In particular, Jeffrey Witmer at Oberlin

College sent extensive comments. Adrian Brasoveanu at the University of California, Santa Cruz, and John Miyamoto at the University of Washington, also conveyed or posted information about their courses. The review by Vanpaemel and Tuerlinckx (2012) reported experiences from classroom use. Thanks to all the instructors who have boldly tried the first edition in their courses. I hope that the second edition is even more useful for the classroom and self study.

Over recent years there have been many students in my classes who have made insightful comments and suggestions. Some of these include Young Ahn, Gregory Cox, Junyi Dai, Josh de Lueew, Andrew Jahn, Arash Khodadadi, and Torrin Liddell. Thanks to Torrin Liddell also for help with checking the proofs of the book. I thank Anne Standish for researching and recommending a discussion forum for the book's blog. I am grateful to the many people who have left thoughtful comments at the blog and discussion forum. Thanks also to several careful readers who reported errors in the first edition. And thanks to Jacob Hodes for being so interested in Bayesian analysis that he would travel to a conference to talk with me about it.

The first edition of this book was 6 years in the making, and many colleagues and students provided helpful comments during that period. The most extensive comments came from Luiz Pessoa, Michael Kalish, Jerome Busemeyer, and Adam Krawitz; thank you all! Particular sections were insightfully improved by helpful comments from Michael Erickson, Robert Nosofsky, Geoff Iverson, and James L. (Jay) McClelland. Various parts of the book benefited indirectly from communications with Woojae Kim, Charles Liu, Eric-Jan Wagenmakers, and Jeffrey Rouder. Pointers to data sets were offered by Teresa Treat and Michael Trosset, among others. Very welcome supportive feedback was provided by Michael D. Lee, and also by Adele Diederich. A Bayesian-supportive working environment was provided by many colleagues including Richard Shiffrin, Jerome Busemeyer, Peter Todd, James Townsend, Robert Nosofsky, and Luiz Pessoa. Other department colleagues have been very supportive of integrating Bayesian statistics into the curriculum, including Linda Smith and Amy Holtzworth-Munroe. Various teaching assistants have provided helpful comments; in particular I especially thank Noah Silbert and Thomas Wisdom for their excellent assistance. As this book has evolved over the years, suggestions have been contributed by numerous students, including Aaron Albin, Thomas Smith, Sean Matthews, Thomas Parr, Kenji Yoshida, Bryan Bergert, and perhaps dozens of others who have contributed insightful questions or comments that helped me tune the presentation in the book.

Gratefully acknowledged are the creators of the software R (R Core Team, 2013), RStudio (RStudio, 2013), JAGS (Plummer, 2003, 2012), RunJAGS (Denwood, 2013), BUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; A. Thomas, O'Hara, Ligges, & Sturtz, 2006), and Stan (Stan Development Team, 2012). Also gratefully acknowledged are the creators of the typesetting software LaTeX (http://www.latex-project.org/) and MikTeX (http://miktex.org/), the editor TeXmaker (http://www.xm1math.net/

texmaker/), and the drawing application of LibreOffice (http://www.libreoffice.org/), in which this book was composed by the author.

To all the people who have made contributions but whom I have inadvertently forgotten to mention, I extend my apologies and genuine appreciation.

Finally, my deepest gratitude goes to Dr. Rima Hanania, who has been my constant and most esteemed companion throughout all the years of writing this book.