*This book is again dedicated to Jack Gill, who is still there when I need him.*

# *Contents*

# *List of Figures*

# *List of Tables*

# *Preface to the Third Edition*

## General Comments

Welcome to the third edition of BMSBSA. When the first edition appeared in 2002 Bayesian methods were still considered a bit exotic in the social sciences. Many distrusted the use of prior distributions and some mysterious simulation process that involved non-iid sampling and uncertainty about convergence. The world is completely different now, and Bayesian modeling has become standard and MCMC is well-understood and trusted. Of course it helps that Moore's Law (doubling of computing power every two years, presumably until we reach the 7 nanometer threshold) continues without pause making our computers notably faster and allowing longer sampling procedures and parallel process without an agonizingly long wait. In this context the third edition spends less time justifying procedures and more time providing implementation details on these procedures. This is also an opportunity to expand the set of examples.

## Changes from the Second Edition

As expected there are a number of additions in the new edition. First, and most laboriously, the number of exercises has been doubled such that there are now twenty in each chapter. The former exercises are now the odd-numbered exercises with the answer key being fully publicly distributed (not just to instructors). The new exercises emphasize recent developments in Bayesian inference and Bayesian computing as a way to include more modern material. All of the chapters have been refreshed, although some more than others. The basic material has not changed in over a century, so there is no need to dramatically alter basic material. Conversely, Bayesian stochastic simulation (MCMC) has undergone dramatic developments in the last decade, including having become routine in applied settings. New MCMC material includes Hamiltonian Monte Carlo and expanded model implementation. Second, there are two new chapters. A chapter on Bayesian decision theory is long overdue, and this is now Chapter 8. It includes discussion of both Bayesian and frequentist decision theory since this is where the two paradigms are most intertwined. Included topics are: loss

functions, risk, decision rules, and regression-model applications. The chapter finishes with two important topics not featured in previous editions: James-Stein estimation and empirical Bayes. While empirical Bayes was briefly covered in the second edition, its connection with James-Stein estimation was ignored. This section now covers the important topics in this area and provides Bayesian context. Also new is a chapter on practical implementation of MCMC methods (Chapter 11). This covers mechanical issues with the `BUGS` language, including calling the software from `R`. The goal is to provide a detailed introduction to the essential software for running Bayesian hierarchical regression models. Relatedly the chapter on hierarchical models is greatly expanded. This is an area of great applied interest right now and provides a strong motivation for the Bayesian paradigm. Finally, on a more practical side, there is a wealth of new examples and applications. These are chosen from a variety of social science disciplines and are intended to illustrate the key principles of the relevant chapter. In addition, the BaM package in `R` that accompanies this manuscript has been greatly expanded with new datasets and new code. This includes new procedures for calling `BUGS` packages from `R`.

---

## Course Plans

The recommended course plans remain essentially the same as outlined in the preface to the second edition. The one critical difference is adding Chapter 11 (Implementing Bayesian Models with Markov Chain Monte Carlo) to a basic course or comprehensive course. The longer length of the text means that not all chapters are practical in a one-semester course. For a standard introductory Bayesian social science graduate course, the most succinct set of chapters are:

- ▷ Chapter 1: Background and Introduction
- ▷ Chapter 2: Specifying Bayesian Models
- ▷ Chapter 3: The Normal and Students'-$t$ Models
- ▷ Chapter 4: The Bayesian Prior
- ▷ Chapter 5: The Bayesian Linear Model
- ▷ Chapter 10: Basics of Markov Chain Monte Carlo
- ▷ Chapter 11: Implementing Bayesian Models with Markov Chain Monte Carlo
- ▷ Chapter 12: Bayesian Hierarchical Models
- ▷ Chapter 14: Utilitarian Markov Chain Monte Carlo.

This assumes some knowledge of basic Monte Carlo methods of the students. Chapter 11 and Chapter 14 could also be assigned reading rather than part of lectures since the focus is on very practical concerns.

## Support

As done in the last two editions of this text, there is a dedicated website provided to support readers: `http://stats.wustl.edu/BMSBSA3`. This site has software, errata, comments, and the answer key for odd-numbered exercises. All of the code is also provided in the associated `R` package, `BaM`, which has been substantially updated to include new code and data. Where possible `BUGS` code is included in this package. Note that in many cases the code relies on multiple `R` packages, as well as stand-alone software such as `JAGS` and `WinBUGS`, so changes over time may introduce incompatibilities that need to be worked out. In many cases this amounts to downloading the most recent version of some software. Relevant updates will be posted at the dedicated webpage when they come to my attention.

## Acknowledgments

Since this edition has been seven years in the making, there are many people to thank. First, I appreciate the support from the Department of Political Science, the Division of Biostatistics, and the Department of Surgery (Public Health Sciences) at Washington University for being supportive home environments. This includes support from the Transdisciplinary Research on Energetics and Cancer (TREC) grant and Graham Colditz and Sarah Gehlert. I also thank the Summer School in Social Science Data Analysis at The University of Essex leadership, Thomas Plümper and Vera Troeger, since some of the datasets and software were developed in the process of teaching there. In particular, new ideas for the chapter on hierarchical models were created in the welcoming community that is Essex. The manuscript was finished while on sabbatical Spring 2014 at the University of Minnesota Division of Biostatistics (with additional financial support from the Departments of Political Science, Psychology, Sociology, and Statistics). I thank Brad Carlin, Sudipto Banerjee, John Freeman, and Niels Waller for making this happen and providing a pleasant (but really cold!) environment to wrap up this project. Finally, I acknowledge the support of NSF grants: DMS-0631632 and SES-0631588, which are my last in cooperation with George Casella. Numerous people have sent useful comments, notably Gary McDonald, Kentaro Fukumoto, Ben Begozzi, Patrick Brandt, Yuta Kamahara, Bruce Desmarais, Skyler Cranmer, Ryan Bakker, and Gary in particular was relentless in his comments (in a good way!). In the Spring of 2011 twelve graduate students at Washington University participated in a weekly reading group around the manuscript giving useful and insightful comments on methods, computing, and even the writing. I thank Peter Casey, Adriana Crespo-Tenorio, Constanza Figueroa Schibber, Morgan Hazelton, Rachael Hinkley, Chia-yi Lee, Michael Nelson, Christopher Pope, Elizabeth Rose, and Alicia Uribe. All of these stu-

dents have finished their studies and gone on to successful careers. This is the first edition of this book after George Casella left us. He made a big impact on the first two issues as my neighbor, close friend, colleague, and coauthor. As Christian Robert states, George simply made everyone's work better for those fortunate to be around him. This work is certainly better because of George's influence.

# *Preface to the Second Edition*

## Starters

Wow, over five years have elapsed since the first edition appeared. Bayesian methods in the social sciences have grown and changed dramatically. This is a positive and encouraging development. When I was writing the first version I would often get questions from social science colleagues about why I would write on a seemingly obscure branch of statistics. This is clearly no longer an issue and Bayesian approaches appear to have a prominent role in social science methodology. I hope that the first edition contributed to this development.

Bayesian methods continue to become more important and central to statistical analysis, broadly speaking. Seemingly, no issue of the *Journal of the American Statistical Association* arrives without at least one Bayesian application or theoretical development. While this upward trend started in the 1990s after we discovered Markov chain Monte Carlo hiding in statistical physics, the trend accelerates in the 21st century. A nice foretelling is found in the 1999 *Science* article by David Malakoff, "Bayesian Boom," complete with anecdotes about popular uses in biology and computing as well as quotes from John Geweke. Back in 1995, the Bayesian luminary Bruno de Finetti predicted that by the year 2020 we would see a paradigm shift to Bayesian thinking (quoted in Smith [1995]). I believe we are fully on track to meet this schedule.

Bayesian computing is broader and more varied than it was at the writing of the first edition. In addition to `BUGS` and `WinBUGS`, we now routinely use `MCMCpack`, `JAGS`, `openbugs`, `bayesm`, and even the new `SAS` MCMC procedure. The diagnostic routines in `R`, `BOA`, and `CODA` continue to be useful and are more stable than they were. Of the course the lingua franca of `R` is critical, and many researchers use `C` or `C++` for efficiency. Issues of statistical computing remain an important component of the book. It is also necessary to download and use the `R` packages `CODA` and `BOA` for MCMC diagnostics.

Bayesian approaches are also increasingly popular in related fields not directly addressed in this text. There is now an interesting literature in archaeology that is enjoyable to read (Reese 1994, Freeman 1976, Laxton *et al.* 1994), and the best starting point is the seminal paper by Litton and Buck (1995) that sets the agenda for Bayesian archaeometrics. Researchers in this area have also become frustrated with the pathetic state of the null hypothesis significance test in the social and behavioral sciences (Cowgill 1977). One area where

Bayesian modeling is particularly useful is in archaeological forensics, where researchers make adult-age estimates of early humans (Lucy *et al.* 1996, Aykroyd *et al.* 1999).

---

## Changes from the First Edition

A reader of the first edition will notice many changes in this revision. Hopefully these constitute improvements (they certainly constituted a lot of work). First, the coverage of Markov chain Monte Carlo is greatly expanded. The reason for this is obvious, but bears mentioning. Modern applied Bayesian work is integrally tied to stochastic simulation and there are now several high-quality software alternatives for implementation. Unfortunately these solutions can be complex and the theoretical issues are often demanding. Coupling this with easy-to-use software, such as `WinBUGS` and `MCMCpack`, means that there are users who are unaware of the dangers inherent in MCMC work. I get a fair number of journal and book press manuscripts to review supporting this point. There is now a dedicated chapter on MCMC theory covering issues like ergodicity, convergence, and mixing. The last chapter is an extension of sections from the first edition that now covers in greater detail tools like: simulated annealing (including its many variants), reversible jump MCMC, and coupling from the past. Markov chain Monte Carlo research is an incredibly dynamic and fast growing literature and the need to get some of these ideas before a social science audience was strong. The reader will also note a substantial increase on MCMC examples and practical guidance. The objective is to provide detailed advice on day-to-day issues of implementation. Markov chain Monte Carlo is now discussed in detail in the first chapter, giving it the prominent position that it deserves. It is my belief that Gibbs sampling is as fundamental to estimation as maximum likelihood, but we (collectively) just do not realize it yet. Recall that there was about 40 years between Fisher's important papers and the publication of Birnbaum's Likelihood Principle. This second edition now provides a separate chapter on Bayesian linear models. Regression remains the favorite tool of quantitative social scientists, and it makes sense to focus on the associated Bayesian issues in a full chapter. Most of the questions I get by email and at conferences are about priors, reflecting sensitivity about how priors may affect final inferences. Hence, the chapter on forms of prior distributions is longer and more detailed. I have found that some forms are particularly well-suited to the type of work that social and behavioral researchers do. One of the strengths of Bayesian methods is the ease with which hierarchical models can be specified to recognize different levels and sources in the data. So there is now an expanded chapter on this topic alone, and while Chapter **??** focuses exclusively on hierarchical model specifications, these models appear throughout the text reflecting their importance in Bayesian statistics.

Additional topics have crept into this edition, and these are covered at varied levels from a basic introduction to detailed discussions. Some of these topics are older and well-

known, such as Bayesian time-series, empirical Bayes, Bayesian decision theory, additional prior specifications, model checking with posterior data prediction, the deviance information criterion (DIC), methods for computing highest posterior density (HPD) intervals, convergence theory, metropolis-coupling, tempering, reversible jump MCMC, perfect sampling, software packages related to `BUGS`, and additional models based on normal and Student's-t assumptions.

Some new features are more structural. There is now a dedicated `R` package to accompany this book, `BaM` (for "**Ba**yesian **M**ethods"). This package includes data and code for the examples as well as a set of functions for practical purposes like calculated HPD intervals. These materials and more associated with the book are available at the dedicated Washington University website: `http://stats.wustl.edu/BMSBSA`. The second edition includes three appendices covering basic maximum likelihood theory, distributions, and `BUGS` software. These were moved to separate sections to make referencing easier and to preserve the flow of theoretical discussions. References are now contained in a single bibliography at the end for similar reasons. Some changes are more subtle. I've changed all instances of "noninformative" to "uninformative" since the first term does not really describe prior distributions. Markov chain Monte Carlo techniques are infused throughout, befitting their central role in Bayesian work. Experience has been that social science graduate students remain fairly tepid about empirical examples that focus on rats, lizards, beetles, and nuclear pumps. Furthermore, as of this writing there is no other comprehensive Bayesian text in the social sciences, outside of economics (except the out-of-print text by Phillips [1973]).

---

## Road Map

To begin, the prerequisites remain the same. Readers will need to have a basic working knowledge of linear algebra and calculus to follow many of the sections. My math text, *Essential Mathematics for Political and Social Research* (2006), provides an overview of such material. Chapter 1 gives a brief review of the probability basics required here, but it is certainly helpful to have studied this material before. Finally, one cannot understand Bayesian modeling without knowledge of maximum likelihood theory. I recognize graduate programs differ in their emphasis on this core material, so Appendix A covers these essential ideas.

The second edition is constructed in a somewhat different fashion than the first. The most obvious difference is that the chapter on generalized linear models has been recast as an appendix, as mentioned. Now the introductory material flows directly into the construction of basic Bayesian statistical models and the procession of core ideas is not interrupted by a non-Bayesian discussion of standard models. Nonetheless, this material is important to have close at hand and hopefully the appendix approach is convenient. Another notable

change is the "promotion" of linear models to their own chapter. This material is important enough to stand on its own despite the overlap with Bayesian normal and Student's-t models. Other organization changes are found in the computational section where considerable extra material has been added, both in terms of theory and practice. Markov chain Monte Carlo set the Bayesians free, and remains an extremely active research field. Keeping up with this literature is a time-consuming, but enjoyable, avocation.

There are a number of ways that a graduate course could be structured around this text. For a basic-level introductory course that emphasizes theoretical ideas, the first seven chapters provide a detailed overview without considering many computational challenges. Some of the latter chapters are directed squarely at sophisticated social scientists who have not yet explored some of the subtle theory of Markov chains. Among the possible structures, consider the following curricula.

**Basic Introductory Course**

- ▷ Chapter 1: **Background and Introduction**
- ▷ Chapter **??**: **Specifying Bayesian Models**
- ▷ Chapter **??**: **The Normal and Student's-t Models**
- ▷ Chapter **??**: **The Bayesian Linear Model**
- ▷ Chapter **??**: **Bayesian Hierarchical Models**

**Thorough Course without an Emphasis on Computing**

- ▷ Chapter 1: **Background and Introduction**
- ▷ Chapter **??**: **Specifying Bayesian Models**
- ▷ Chapter **??**: **The Normal and Student's-t Models**
- ▷ Chapter **??**: **The Bayesian Linear Model**
- ▷ Chapter **??**: **The Bayesian Prior**
- ▷ Chapter **??**: **Assessing Model Quality**
- ▷ Chapter **??**: **Bayesian Hypothesis Testing and the Bayes Factor**
- ▷ Chapter **??**: **Bayesian Hierarchical Models**

**A Component of a Statistical Computing Course**

- ▷ Chapter **??**: **Specifying Bayesian Models**
- ▷ Chapter **??**: **Monte Carlo and Related Iterative Methods**
- ▷ Chapter **??**: **Basics of Markov Chain Monte Carlo**
- ▷ Chapter **??**: **Some Markov Chain Monte Carlo Theory**

- ▷ Chapter **??**: **Utilitarian Markov Chain Monte Carlo**
- ▷ Chapter **??**: **Markov Chain Monte Carlo Extensions**

**A Component of an Estimation Course**

- ▷ **??**: **Generalized Linear Model Review**
- ▷ Chapter 1: **Background and Introduction**
- ▷ Chapter **??**: **Specifying Bayesian Models**
- ▷ Chapter **??**: **The Bayesian Linear Model**
- ▷ Chapter **??**: **Bayesian Hypothesis Testing and the Bayes Factor**

Of course I am eager to learn about how instructors use these chapters independent of any advice here.

---

## Acknowledgments

So many people have commented on this edition, the previous edition, related papers, associated conference presentations, and classes taught from the book that I am unlikely to remember them all. Apologies to anyone left out from this list. This edition received three detailed, formal reviews from Patrick Brandt, Andrew Gelman, and Andrew Martin. Their comments were invaluable and dramatically improved the quality of this work.

A substantial part of the writing of this book was done during the 2006-2007 academic year while I was Visiting Professor at Harvard University's Department of Government and Fellow at the Institute for Quantitative Social Science. I thank Gary King for inviting me into that dynamic and rewarding intellectual environment. The Fall semester of that year I taught a graduate seminar titled *Bayesian Hierarchical Modeling*, which enabled me to produce and distribute chapter material on a weekly basis. Participants in the seminar provided excellent critiques of the principles and exposition. These students and guests included: Justin Grimmer, Jonathan Harris, Jennifer Katkin, Elena Llaudet, Clayton Nall, Emre Ozaltin, Lindsay Page, Omar Wasow, Lefteris Anastasopoulos, Shelden Bond, Janet Lewis, Serban Tanasa, and Lynda Zhang. The Teaching Fellows for this seminar were Skyler Cranmer and Andrew Thomas, who were instrumental in improving various technical sections. I also thank Jens Hainmueller, Dominik Hangartner, and Holger Lutz Kern for productive discussions of statistical computing during the Spring semester of that year.

Since 2000 I have taught a course based on this book at the Inter-University Consortium for Political and Social Research (ICPSR) Summer Program at the University of Michigan. Many of the highly motivated students in this program had constructive comments on the

material. I also benefited immensely from interactions with colleagues and administrators in the program, including Bob Andersen, David Armstrong, Ken Bollen, Dieter Burrell, John Fox, Charles Franklin, Hank Heitowit, Bill Jacoby, Jim Johnson, Dean Lacy, Scott Long, Jim Lynch, Tim McDaniel, Sandy Schneider, Bob Stine, and Lee Walker. The three teaching assistants over this period were incredibly helpful in developing homework and computer lab assignments: Ryan Bakker, Leslie Johns, and Yu-Sung Su. Overall, ICPSR is unmatched as a high-intensity teaching and learning Summer experience.

I have tried to record those making comments on the first edition or the manuscript version of the second edition. In addition to those already mentioned, useful critiques came from: Attic Access, Larry Bartels, Neal Beck, Jack Buckley, Sid Chib, Boyd Collier, Skyler J. Cranmer, Chris Dawes, Daniel J. Denis, Persi Diaconis, Alexis Dinno, Hanni Doss, George Duncan, James Fowler, Justin Gross, Josue Guzman, Michael Herrmann, Jim Hobert, Kosuke Imai, Brad Jones, Lucas Leemann, Jane Li, Rod Little, John Londregan, Enrico Luparini, Jonathan Nagler, Shunsuke Narita, Keith T. Poole, Kevin Quinn, Rafal Raciborski, Michael Smithson, John Sprague, Rick Waterman, Bruce Western, and Chris Zorn. I also want to give a special thank you to my friend and coauthor George Casella who has provided irreplaceable brotherly guidance over the last eight years.

The research on Dirichlet process priors (Chapter **??**) was supported by National Science Foundation grants DMS-0631632 and SES-0631588. My work on elicited priors (Chapter **??**) was helped by research with Lee Walker that appeared in the *Journal of Politics*. The discussion of dynamic tempered transitions (Chapter **??**) draws from an article written with George Casella that appeared in *Political Analysis*. Comments from the editors, Bill Jacoby and Bob Erikson, made these works better and more focused. The education policy example (Chapter **??**) using a Bayesian linear model is drawn from an article I wrote with my former graduate student Kevin Wagner. The discussion of convergence in Chapter **??** benefited from my recent article on this topic in *Political Analysis*. Finally, while there is no direct overlap, my understanding of detailed statistical computing principles benefited from the book project with Micah Altman and Michael McDonald.

---

## References

Aykroyd, Robert G., Lucy, David, Pollard, Mark A., and Roberts, Charlotte A. (1999). Nasty, Brutish, But Not Necessarily Short: A Reconsideration of the Statistical Methods Used to Calculate Age At Death From Adult Human Skeletal and Dental Age Indicators. *American Antiquity* **64**, 55-70.

Cowgill, G. L. (1977). The Trouble With Significance Tests and What We Can Do About It. *Philosophical Transactions of the Royal Society* **327**, 331-338.

Freeman, P. R. (1976). A Bayesian Approach to the Megalithic Yard. *Journal of the Royal Statistical Association, Series A* **139**, 279-295.

Gill, Jeff. (2006). *Essential Mathematics for Political and Social Research.* Cambridge, England: Cambridge University Press.

Laxton, R. R., Cavanaugh, W. G., Litton, C. D., Buck, C. E., and Blair, R. (1994). The Bayesian Approach to Archaeological Data Analysis: An Application of Change-Point Analysis to Prehistoric Domes. *Archeologia e Calcolatori* **5**, 53-68.

Litton, C. D. and Buck, C. E. (1995). The Bayesian Approach to the Interpretation of Archaeological Data. *Archaeometry* **37**, 1-24.

Lucy, D., Aykroyd, R. G., Pollard, A. M., and Solheim, T. (1996). A Bayesian Approach to Adult Human Age Estimation from Dental Observations by Johanson's Age Changes. *Journal of Forensic Sciences* **41**, 189-194.

Malakoff, David. (1999). Bayes Offers a 'New' Way to Make Sense of Numbers. *Science* **286**, 1460-1464.

Phillips, L. D. (1973). *Bayesian Statistics for Social Scientists.* Thomas Nelson and Sons, London.

Reese, R. (1994). Are Bayesian Statistics Useful to Archaeological Reasoning? *Antiquity* **68**, 848-850.

Smith, A. F. M. (1995). A Conversation with Dennis Lindley. *Statistical Science* **10**, 305-319.

# *Preface to the First Edition*

## Contextual Comments

This book is intended to fill a void. There is a reasonably wide gap between the background of the median empirically trained social or behavioral scientist and the full weight of Bayesian statistical inference. This is unfortunate because, as we will see in the forthcoming chapters, there is much about the Bayesian paradigm that suits the type of data and data analysis performed in the social and behavioral sciences. Consequently, the goal herein is to bridge this gap by connecting standard maximum likelihood inference to Bayesian methods by emphasizing linkages between the standard or classical approaches and full probability modeling via Bayesian methods.

This is far from being an exclusively theoretical book. I strongly agree that "theoretical satisfaction and practical implementation are the twin ideals of coherent statistics" (Lindley 1980), and substantial attention is paid to the mechanics of putting the ideas into practice. Hopefully the extensive attention to calculation and computation basics will enable the interested readers to immediately try these procedures on their own data. Coverage of various numerical techniques from detailed posterior calculations to computational-numerical integration is extensive because these are often the topics that separate theory and realistic practice.

The treatment of theoretical topics in this work is best described as "gentle but rigorous": more mathematical derivation details than related books, but with more explanation as well. This is not an attempt to create some sort of "Bayes-Lite" or "Bayes for Dummies" (to paraphrase the popular self-help works). Instead, the objective is to provide a Bayesian methods book tailored to the interests of the social and behavioral sciences. It therefore features data that these scholars care about, focuses more on the tools that they are likely to require, and speaks in a language that is more compatible with typical prerequisites in associated departments.

There is also a substantial effort to put the development of Bayesian methods in a historical context. To many, the principles of Bayesian inference appear to be something that "came out of left field," and it is important to show that not only are the fundamentals of Bayesian statistics older than the current dominant paradigms, but that their history and development are actually closely intertwined.

## Outline of the Book

This book is laid out as follows. Chapter 1 gives a high-level, brief introduction to the basic philosophy of Bayesian inference. I provide some motivations to justify the time and effort required to learn a new way of thinking about statistics through Bayesian inference. Chapter 2 (*now Appendix A*) provides the necessary technical background for going on: basic likelihood theory, the generalized linear model, and numerical estimation algorithms.

Chapter 1 describes the core idea behind Bayesian thinking: updating prior knowledge with new data to give the *posterior distribution*. Examples are used to illustrate this process and some historical notes are included. The normal model and its relatives are no less important in Bayesian statistics than in non-Bayesian statistics, and Chapter **??** outlines the key basic normal models along with extensions.

Specifying prior distributions is a key component of the Bayesian inference process and Chapter **??** goes through the typology of priors. The Bayesian paradigm has a cleaner and more introspective approach to assessing the quality of fit and robustness of researcher-specified models, and Chapter **??** outlines procedures so that one can test the performance of various models. Chapter **??** is a bit more formal about this process; it outlines a number of ways to explicitly test models against each other and to make decisions about unknown parameters.

The most modern component of this book begins with Chapter **??**, which is an introduction to Monte Carlo and related methods. These topics include the many varieties of numerical integration and importance sampling, and culminating with the EM algorithm. While none of these tools are exclusively Bayesian in nature, Bayesians generally make more use of them than others. Chapter **??** formally introduces Markov chain Monte Carlo (MCMC). These are the tools that revolutionized Bayesian statistics and led to the current renaissance. This chapter includes both theoretical background on Markov chains as well as practical algorithmic details. Chapter **??** discusses hierarchical models that give the Bayesian researcher great flexibility in specifying models through a general framework. These models often lead to the requirement of MCMC techniques and the examples in this chapter are illustrated with practical computing advice. Finally, Chapter 11 (*in the first edition*) discusses necessary details about the mechanics of running and testing MCMC inferences.

The structure of each chapter is reasonably uniform. The basic ideas are enumerated early in the chapter and several of the chapters include an advanced topics section to further explore ideas that are unlikely to be of interest to every reader or to the first-time reader. All chapters have exercises that are intended to give practice developing the central ideas of each topic, including computer-based assignments.

There are unfortunately several topics that I have not had the space to cover here. Foremost is Bayesian decision theory. Many social and behavioral scientists do not operate

in a data-analytic environment where the explicit cost of making a wrong decision can be quantified and incorporated into the model. This may be changing and there are a number of areas that are currently oriented toward identifying loss and risk, such as applied public policy. In the meantime, readers who are focused accordingly are directed to the books by Berger (1985), Winkler (1972), Robert (2001), and the foundational work of Wald (1950). The second major topic that is mentioned only in passing is the growing area of empirical Bayes. The best introduction is the previously noted text of Carlin and Louis (2001). See also the extensive empirical Bayes reference list in Section **??**. I would very much have liked to cover the early, but exciting developments in perfect sampling (coupling from the past). See the original work by Propp and Wilson (1996).

Bayesian game theory is an important topic that has been omitted. Some of the better known citations are Raiffa (1982), Blackwell and Girshick (1954), Savage (1954), and Bayarri and DeGroot (1991). The Bayesian analysis of survival data as a distinct subspecialty is somewhat understudied. The recent book by Ibrahim, Chen, and Sinha (2001) goes a long way toward changing that. Chapter **??** provides the essentials for understanding Markov chains in general. The study of Markov chains extends well beyond basic MCMC and the mathematical references that I often find myself reaching for are Meyn and Tweedie (1993), Norris (1997), and Nummelin (1984). The Bayesian hierarchical models covered in Chapter **??** naturally and easily extend into meta-analysis, a subject well-covered in the social sciences by Cooper and Hedges (1994), Hunter and Schmidt (1990), and Lipsey and Wilson (2001).

---

## Background and Prerequisites

This is not a book for a first-semester course in social and behavioral statistics. Instead, it is intended to extend the training of graduate students and researchers who have already experienced a one-year (roughly) sequence in social statistics. Therefore good prerequisites include intermediate-level, regression-oriented texts such as Fox (1997), Gujarati (1995), Hanushek and Jackson (1977), Harrell (2001), Neter *et al.* (1996), and Montgomery *et al.* (2001). Essentially it is assumed that the reader is familiar with the basics of the linear model, simple inference, multivariate specifications, and some nonlinear specifications.

A rudimentary understanding of matrix algebra is required, but this does not need to go beyond the level of Chapter 1 in Greene (2000), or any basic undergraduate text. The essential manipulations that we will use are matrix multiplication, inversion, transposition, and segmentation. The calculus operations done here are more conceptual than mechanical; that is, it is more important to understand the meaning of differentiation and integration operations rather than to be an expert on the technicalities. A knowledge at the level of Kleppner and Ramsey's (1985) self-teaching primer is sufficient to follow the calculations.

The core philosophical approach taken with regard to model specification comes from the generalized linear model construct of Nelder and Wedderburn (1972), elaborated in McCullagh and Nelder (1989). This is an integrated theoretical framework that unifies disparate model specifications by re-expressing models based on making the appropriate choice of model configuration based on the structure of the outcome variable and the nature of the dispersion. This fundamental way of thinking is independent of whether the model is Bayesian (see Dey, Ghosh, and Mallick 2000) or classical (see Fahrmeir and Tutz 2001).

---

## Software

The concepts and procedures in this book would be of little practical value without a means of directly applying them. Consequently, there is an emphasis here on demonstrating ideas with statistical software. *All* code in `R` and `BUGS` and *all* data are posted at the dedicated webpage:

`http://web.clas.ufl.edu/~jgill/BMSBSA`.

A great deal of the material in this book focuses on developing examples using the `R` and `BUGS` statistical packages. Not only are these extremely high-quality analytical tools, they are also widely distributed free of charge.

It is hard to overstate the value of the `R` statistical environment. `R` is the Open Source implementation of the `S` statistical language (from AT&T-Bell Labs), which has become the *de facto* standard computing language in academic statistics because of its power, flexibility, and sense of community. `R` was initially written by Robert Gentleman and Ross Ihak at the University of Auckland, but is now supported by a growing group of dedicated scholars. An important aspect of `R` is the user community itself, and the user-written packages have been shown to be an effective way for scholars to share and improve new methods.

The homesite for `R` (see the details in Chapter 2, *now Appendix A*), contains documentation on installation and learning the language. In addition, because `R` is "non-unlike" `S`, any published book on S-Plus will be useful. The standard text for statistical modeling in `S` is the work of Venables and Ripley (1999). The forthcoming book by Fox (2002) is a particularly helpful and well-written introduction to doing applied work in `S`. In addition, an increasing number of applied methodology books that feature the `S` language have appeared, and I try to keep up with these on a webpage:

`http://web.clas.ufl.edu/~jgill/s-language.help.html`.

Any applied Bayesian today that wants to feel good about the state of the world with regard to software need only look at Press' 1980 summary of available Bayesian analysis programs. This is a disparate, even tortured, list of mainframe-based programs that generally only implement one or two procedures each and require such pleasantries as "Raw data on

Paper tape." In contrast, the `BUGS` package makes Bayesian analysis using MCMC pleasant and engaging by taking the odious mechanical aspects away from the user, allowing one to focus on the more interesting aspects of model specification and testing. This unbelievable gift to the Bayesian statistical community was developed at the MRC Biostatistics Unit in Cambridge:

`http://www.mrc-bsu.cam.ac.uk/bugs/`.

---

## Acknowledgments

I am indebted to many people for criticisms, suggestions, formal reviews, and influential conversations. These include Alan Agresti, Micah Altman, Attic Access, Sammy Barkin, Neal Beck, Jim Booth, Brad Carlin, George Casella, Lauren Cowles, John Fox, Wayne Francis, Charles Franklin, Malay Ghosh, Hank Heitowit, Jim Hobert, Bill Jacoby, Renee Johnson, Gary King, Andrew Martin, Michael Martinez, Mike McDonald, Ken Meier, Elias Moreno, Brad Palmquist, Kevin Quinn, Christian Robert, Stephen Sen, Jason Wittenberg, Sam Wu, Chris Zorn, and anonymous reviewers. I am especially grateful to George Casella for his continued advice and wisdom; I've learned as much sitting around George's kitchen table or trying to keep up with him running on the beach as I have in a number of past seminars. Andrew Martin also stands out for having written detailed and indispensable reviews (formally and informally) of various drafts and for passing along insightful critiques from his students.

A small part of the material in Chapter **??** was presented as a conference paper at the Fifth International Conference on Social Science Methodology in Cologne, Germany, October 2000. Also, some pieces of Chapter **??** were taken from another conference paper given at the Midwestern Political Science Association Annual Meeting, Chicago, April 2001.

This book is a whole lot better due to input from students at the University of Florida and the University of Michigan Summer Program, including: Javier Aprricio-Castello, Jorge Aragon, Jessica Archer, Sam Austin, Ryan Bakker, David Conklin, Jason Gainous, Dukhong Kim, Eduardo Leoni, Carmela Lutmar, Abdel-hameed Hamdy Nawar, Per Simonnson, Jay Tate, Tim McGraw, Nathaniel Seavy, Lee Walker, and Natasha Zharinova. For research assistance, I would also like to thank Ryan Bakker, Kelly Billingsley, Simon Robertshaw, and Nick Theobald.

I would like to also thank my editor, Bob Stern, at Chapman & Hall for making this process much more pleasant than it would have been without his continual help. This volume was appreciatively produced and delivered "camera-ready" with LaTeX using the $\mathcal{AMS}$ packages, pstricks, and other cool typesetting macros and tools from the TeX world. I cannot imagine the bleakness of a life restricted to the low-technology world of word processors.

## References


Bayarri, M. J. and DeGroot, M. H. (1991). What Bayesians Expect of Each Other. *Journal of the American Statistical Association* **86**, 924-932.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second Edition. New York: Springer-Verlag.

Blackwell, D. and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions.* New York: Wiley.

Carlin, B. P. and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis.* Second Edition. New York: Chapman & Hall.

Cooper, H. and Hedges, L. (eds.) (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective.* New York: Marcel Dekker.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Second Edition. New York: Springer.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods.* Thousand Oaks, CA: Sage.

Fox, J. (2002). *An R and S-Plus Companion to Applied Regression.* Thousand Oaks, CA: Sage.

Greene, W. (2000). *Econometric Analysis.* Fourth Edition. Upper Saddle River, NJ: Prentice Hall.

Gujarati, D. N. (1995). *Basic Econometrics.* New York: McGraw-Hill.

Hanushek, E. A. and Jackson, J. E. (1977). *Statistical Methods for Social Scientists.* San Diego: Academic Press.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer-Verlag.

Hunter, J. E. and Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Thousand Oaks, CA: Sage.

Ibrahim, J. G., Chen, M-H., and Sinha, D. (2001). *Bayesian Survival Analysis.* New York: Springer-Verlag.

Kleppner, D. and Ramsey, N. (1985). *Quick Calculus: A Self-Teaching Guide.* New York: Wiley Self Teaching Guides.

Lindley, D. V. (1980). Jeffreys's Contribution to Modern Statistical Thought. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* Arnold Zellner (ed.). Amsterdam: North Holland.

Lipsey, M. W. and Wilson, D. B. (2001). *Practical Meta-Analysis.* Thousand Oaks, CA: Sage.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Second Edition. New York: Chapman & Hall.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* New York: Springer-Verlag.

Montgomery, D. C. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis.* Third Edition. New York: John Wiley & Sons.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* **135**, 370-85.

Neter, J., Kutner, M. H., Nachtsheim, C., and Wasserman, W. (1996). *Applied Linear Regression Models.* Chicago: Irwin.

Norris, J. R. (1997). *Markov Chains.* Cambridge: Cambridge University Press.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators.* Cambridge: Cambridge University Press.

Press, S. J. (1980). Bayesian Computer Programs. In *Studies in Bayesian Econometrics and Statistics*, S. E. Fienberg and A. Zellner, eds., Amsterdam: North Holland, pp. 429-442..

Propp, J. G. and Wilson, D. B. (1996). Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. *Random Structures and Algorithms* **9**, 223-252.

Raiffa, H. (1982). *The Art and Science of Negotiation.* Cambridge: Cambridge University Press.

Robert, C. P. (2001). *The Bayesian Choice: A Decision Theoretic Motivation.* Second Edition. New York: Springer-Verlag.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, Third Edition. New York: Springer-Verlag.

Wald, A. (1950). *Statistical Decision Functions.* New York: Wiley.

Winkler, R. L. (1972). *Introduction to Bayesian Inference and Decision.* New York: Holt, Rinehart, and Winston.

# *Chapter 1*

# *Background and Introduction*

## 1.1 Introduction

Vitriolic arguments about the merits of Bayesian versus classical approaches seem to have faded into a quaint past of which current researchers in the social sciences are, for the most part, blissfully unaware. In fact, it almost seems odd well into the 21st century that deep philosophical conflicts dominated the last century on this issue. What happened? Bayesian methods always had a natural underlying advantage because all unknown quantities are treated probabilistically, and this is the way that statisticians and applied statisticians really prefer to think. However, without the computational mechanisms that entered into the field we were stuck with models that couldn't be estimated, prior distributions (distributions that describe what we know before the data analysis) that incorporated uncomfortable assumptions, and an adherence to some bankrupt testing notions. Not surprisingly, what changed all this was a dramatic increase in computational power and major advances in the algorithms used on these machines. We now live in a world where there are very few model limitations, other than perhaps our imaginations. We therefore live in world now where researchers are for the most part comfortable specifying Bayesian and classical models as it suits their purposes.

It is no secret that Bayesian methods require a knowledge of classical methods as well as some additional material. Most of this additional material is either applied calculus or statistical computing. That is where this book comes in. The material here is intended to provide an introduction to Bayesian methods all the way from basic concepts through advanced computational material. Some readers will therefore be primarily interested in different sections. Also it means that this book is not strictly a textbook, a polemic, nor a research monograph. It is intended to be all three.

Bayesian applications in medicine, the natural sciences, engineering, and the social sciences have been increasing at a dramatic rate since the middle of the early 1990s. Interestingly, the Mars Rovers are programmed to think Bayesianly while they traverse that planet. Currently seismologists perform Bayesian updates of aftershocks based on the mainshock and previous patterns of aftershocks in the region. Bayesian networks are built in computational biology, and the forefront of quantitative research in genomics is now firmly Bayesian.

So why has there been a noticeable increase in interest in Bayesian statistics? There are actually several visible reasons. *First*, and perhaps most critically, society has radically increased its demand for statistical analysis of all kinds. A combined increase in clinical trials, statistical genetics, survey research, general political studies, economic analysis, government policy work, Internet data distribution, and marketing research have led to golden times for applied statisticians. *Second*, many introspective scholars who seriously evaluate available paradigms find that alternatives to Bayesian approaches are fraught with logical inconsistencies and shortcomings. *Third*, until recent breakthroughs in statistical computing, it was easy to specify realistic Bayesian statistical models that could not provide analytically tractable summary results.

There is therefore ample motivation to understand the basics of Bayesian statistical methodology, and not just because it is increasingly important in mainstream analysis of data. The Bayesian paradigm rests on a superior set of underlying assumptions and includes procedures that allow researchers to include reliable information in addition to the sample, to talk about findings in intuitive ways, and to set up future research in a coherent manner. At the core of the data-analytic enterprise, these are key criteria to producing useful statistical results.

Statistical analysis is the process of "data reduction" with the goal of separating out underlying systematic effects from the noise inherent in all sets of observations. Obviously there is a lot more to it than that, but the essence of what we do is using models to distill findings out of murky data. There are actually three general steps in this process: collection, analysis, and assessment. For most people, data collection is not difficult in that we live in an age where data are omnipresent. More commonly, researchers possess an abundance of data and seek meaningful patterns lurking among the various dead-ends and distractions. Armed with a substantive theory, many are asking: what should I do now? Furthermore, these same people are often frustrated when receiving multiple, possibly conflicting, answers to that question.

Suppose that there exists a model-building and data analysis process with the following desirable characteristics:

- ▷ overt and clear model assumptions,
- ▷ a principled way to make probability statements about the real quantities of theoretical interest,
- ▷ an ability to update these statements (i.e., learn) as new information is received,
- ▷ systematic incorporation of previous knowledge on the subject,
- ▷ missing information handled seamlessly as part of the estimation process,
- ▷ recognition that population quantities can be changing over time rather than forever fixed,
- ▷ the means to model all data types including hierarchical forms,
- ▷ straightforward assessment of both model quality and sensitivity to assumptions.

As the title of this book suggests, the argument presented here is that the practice of Bayesian statistics possesses all of these qualities. Press (1989) adds the following practical advantages to this list:

- ▷ it often results in shorter confidence/credible intervals,
- ▷ it often gives smaller model variance,
- ▷ predictions are usually better,
- ▷ "proper" prior distributions (Chapter **??**) give models with good frequentist properties,
- ▷ reasonably "objective" assumptions are available,
- ▷ hypotheses can be tested without pre-determination of testing quality measures.

This text will argue much beyond these points and assert that the type of data social and behavioral scientists routinely encounter makes the Bayesian approach ideal in ways that traditional statistical analysis cannot match. These natural advantages include avoiding the assumption of infinite amounts of forthcoming data, recognition that fixed-point assumptions about human behavior are dubious, and a direct way to include existing expertise in a scientific field.

What reasons are there for *not* worrying about Bayesian approaches and sticking with the, perhaps more comfortable, traditional mindset? There are several reasons why a reader may not want to worry about the principles in this text for use in their research, including:

- ▷ their population parameters of interest are truly fixed and unchanging under all realistic circumstances,
- ▷ they do *not* have any prior information to add to the model specification process,
- ▷ it is necessary for them to provide statistical results as if data were from a *controlled experiment*,
- ▷ they care more about "significance" than effect size,
- ▷ computers are slow or relatively unavailable for them,
- ▷ they prefer very automated, "cookbook" type procedures.

So why do so-called classical approaches dominate Bayesian usage in the social and behavioral sciences? There are several reasons for this phenomenon. *First*, key figures in the development of modern statistics had strong prejudices against aspects of Bayesian inference for narrow and subjective reasons. *Second*, the cost of admission is higher in the form of additional mathematical formalism. *Third*, until recently realistic model specifications sometimes led to unobtainable Bayesian solutions. *Finally*, there has been a lack of methodological introspection in a number of disciplines. The primary mission of this text is to make the second and third reasons less of a barrier through accessible explanation, detailed examples, and specific guidance on calculation and computing.

It is important to understand that the Bayesian way does not mean throwing away one's

comfortable tools, and it is not itself just another tool. Instead it is a way of *organizing* one's toolbox and is also a way of doing statistical work that has sharply different philosophical underpinnings. So adopting Bayesian methods means keeping the usual set of methods, such as linear regression, ANOVA, generalized linear models, tabular analysis, and so on. In fact, many researchers applying statistics in the social sciences are not actually frequentists since they cannot assume an *infinite* stream of iid (independent and identically distributed) data coming from a controlled experimental setup. Instead, most of these analysts can be described as "likelihoodists," since they obtain one sample of observational data that is contextual and will not be repeated, then perform standard likelihood-based (Fisherian) inference to get coefficient estimates.

Aside from underlying philosophical differences, many readers will be comforted in finding that Bayesian and non-Bayesian analyses often agree. There are two important instances where this is *always* true. *First*, when the included prior information is very uninformative (there are several ways of providing this), summary statements from Bayesian inference will match likelihood point estimates. Therefore a great many researchers are Bayesians who do not know it yet. *Second*, when the data size is very large, the form of the prior information used does not matter and there is agreement again. Other circumstances also exist in which Bayesian and non-Bayesian statistical inferences lead to the same results, but these are less general than the two mentioned. In addition to these two important observations, all hierarchical models are overtly Bayesian since they define distributional assumptions at levels. These are popular models due to their flexibility with regard to the prevalence of different levels of observed aggregation in the same dataset. We will investigate Bayesian hierarchical models in Chapter **??**.

We will now proceed to a detailed justification for the use of modern Bayesian methods.

---

## 1.2 General Motivation and Justification

With Bayesian analysis, assertions about unknown model parameters are not expressed in the conventional way as single point estimates along with associated reliability assessed through the standard null hypothesis significance test. Instead the emphasis is on making probabilistic statements using distributions. Since Bayesians make no fundamental distinction between unobserved data and unknown parameters, the world is divided into: immediately available quantities, and those that need to be described probabilistically. Before observing some data, these descriptions are called *prior distributions*, and after observing the data these descriptions are called *posterior distributions*. The quality of the modeling process is the degree to which a posterior distribution is more informed than a prior distribution for some unknown quantity of interest. Common descriptions of posterior distributions include standard quantile levels, the probability of occupying some affected region of the

sample space, the predictive quantities from the posterior, and Bayesian forms of confidence intervals called credible intervals.

It is important to note here that the pseudo-frequentist null hypothesis significance test (NHST) is not just sub-optimal, it is *wrong*. This is the dominant hypothesis testing paradigm as practiced in the social sciences. Serious problems include: a logical inconsistency coming from probabilistic modus tollens, confusion over the order of the conditional probability, chasing significance but ignoring effect size, adherence to completely arbitrary significance thresholds, and confusion about the probability of rejection. There is a general consensus amongst those that have paid attention to this issue that the social sciences have been seriously harmed by the NHST since it has led to fixations with counting stars on tables rather than looking for effect sizes and general statistical reliability. See the recent discussions in Gill (1999) and Denis (2005) in particular for detailed descriptions of these problems and how they damage statistical inferences in the social sciences. Serious criticism of the NHST began shortly after its creation in the early 20th century by textbook writers who blended Fisherian likelihoodism with Neyman and Pearson frequentism in an effort to offend neither warring and evangelical camp. An early critic of this unholy union was Rozeboom (1960) who noticed its "strangle-hold" on social science inference. In 1962 Arthur Melton wrote a parting editorial in the *Journal of Experimental Psychology* revealing that he had held authors to a 0.01 p-value standard: "In editing the *Journal* there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were significant at the .05 level, whether by one- or two-tailed test!" This had the effect of accelerating the criticism and led to many analytical and soul-searching articles discussing the negative consequences of this procedure in the social sciences, including: Barnett (1973), Berger (2003), Berger, Boukai, and Wang (1997), Berger and Sellke (1987), Berkhardt and Schoenfeld (2003), Bernardo (1984), Brandstätter (1999), Carver (1978, 1993), Cohen (1962, 1977, 1988, 1992, 1994), Dar, Serlin and Omer (1994), Falk and Greenbaum (1995), Gigerenzer (1987, 1998a, 1998b, 2004), Gigerenzer and Murray (1987), Gliner, Leech and Morgan (2002), Greenwald (1975), Greenwald, *et al.* (1996), Goodman (1993, 1999), Haller and Krauss (2002), Howson and Urbach (1993), Hunter (1997), Hunter and Schmidt (1990), Kirk (1996), Krueger (2001), Lindsay (1995), Loftus (1991, 1993), Macdonald (1997), McCloskey and Ziliak (1996), Meehl (1978, 1990, 1997), Moran and Soloman (2004), Morrison and Henkel (1969, 1970), Nickerson (2000), Oakes (1986), Pollard (1993), Pollard and Richardson (1987), Robinson and Levin (1997), Rosnow and Rosenthal (1989), Schmidt (1996), Schmidt and Hunter (1977), Schervish (1996), Sedlmeier and Gigerenzer (1989), Thompson (1996, 1997, 2002a, 2002b, 2004), Wilkinson (1977), Ziliak and McCloskey (2007). And this is only a small sample of the vast literature describing the NHST as bankrupt. Conveniently some of the more influential articles listed above are reprinted in Harlow *et al.* (1997). We will return to this point in Chapter **??** (starting on page **??**) in the discussion of Bayesian hypothesis testing and model comparison.

This focus on distributional inference leads to two key assumptions for Bayesian work.

*First*, a specific parametric form is assumed to describe the distribution of the data given parameter values. Practically, this is used to construct a *likelihood function* (**??**) to incorporate the contribution of the full sample of data. Note that this is an inherently parametric setup, and although nonparametric Bayesian modeling is a large and growing field, it exists beyond the scope of the basic setup. *Second*, since unknown parameters are treated as having distributional qualities rather than being fixed, an assumed prior distribution on the parameters of interest unconditional on the data is given. This reflects either uncertainty about a truly fixed parameter or recognition that the quantity of interest actually behaves in some stochastic fashion.

With those assumptions in hand, the essentials of Bayesian thinking can be stated in three general steps:

**1.** Specify a probability model that includes some prior knowledge about the parameters for unknown parameter values.

**2.** Update knowledge about the unknown parameters by conditioning this probability model on observed data.

**3.** Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

Notice that this process does not include an unrealistic and artificial step of making a contrived decision based on some arbitrary quality threshold. The value of a given Bayesian model is instead found in the description of the distribution of some parameter of interest in probabilistic terms. Also, there is nothing about the process contained in the three steps above that cannot be repeated as new data are observed. It is often convenient to use the conventional significance thresholds that come from Fisher, but Bayesians typically do not ascribe any major importance to being barely on one side or the other. That is, Bayesian inference often prescribes something like a 0.05 threshold, but it is rare to see work where a 0.06 finding is not taken seriously as a likely effect.

Another core principle of the Bayesian paradigm is the idea that the data are fixed once observed. Typically (but not always) these data values are assumed to be *exchangeable*; the model results are not changed by reordering the data values. This property is more general than, and implied by, the standard assumption that the data are *independent and identically distributed* (iid): independent draws from the same distribution, and also implies a common mean and variance for the data values (Leonard and Hsu 1999, p.1). Exchangeability allows us to say that the data generation process is conditional on the unknown model parameters in the same way for every data value (de Finetti 1974, Draper *et al.* 1993, Lindley and Novick 1981). Essentially this is a less restrictive version of the standard iid assumption. Details about the exchangeability assumption are given in Chapter **??**. We now turn to a discussion of probability basics as a precursor to Bayesian mechanics.

## 1.3 Why Are We Uncertain about Uncertainty?

The fundamental principles of probability are well known, but worth repeating here. Actually, it is relatively easy to *intuitively* define the properties of a probability function: *(1)* its range is bounded by zero and one for all the values in the applicable domain, *(2)* it sums or integrates to one over this domain, and *(3)* the sum or integral of the functional outcome (probability) of disjoint events is equal to the functional outcome of the union of these events. These are the Kolmogorov axioms (1933), and are given in greater detail in Gill (2006), Chapter 7. The real problem lies in describing the actual meaning of probability statements. This difficulty is, in fact, at the heart of traditional disagreements between Bayesians and non-Bayesians.

The frequentist statistical interpretation of probability is that it is a limiting relative frequency: the long-run behavior of a nondeterministic outcome or just an observed proportion in a population. This idea can be traced back to Laplace (1814), who defined probability as the number of successful events out of trials observed. Thus if we could simply repeat the experiment or observe the phenomenon enough times, it would become apparent what the future probability of reoccurrence will be. This is an enormously useful way to think about probability but the drawback is that frequently it is not possible to obtain a large number of outcomes from exactly the same event-generating system (Kendall 1949, Placket 1966).

A competing view of probability is called "subjective" and is often associated with the phrase "degree of belief." Early proponents included Keynes (1921) and Jeffreys (1961), who observed that two people could look at the same situation and assign different probability statements about future occurrences. This perspective is that probability is *personally* defined by the conditions under which a person would make a bet or assume a risk in pursuit of some reward. Subjective probability is closely linked with the idea of decision-making as a field of study (see, for instance, Bernardo and Smith [1994, Chapter 2]) and the principle of selecting choices that maximize personal utility (Berger 1985).

These two characterizations are necessarily simplifications of the perspectives and de Finetti (1974, 1975) provides a much deeper and more detailed categorization, which we will return to in Chapter **??**. To de Finetti, the ultimate arbiter of subjective probability assignment is the conditions under which individuals will wager their own money. In other words, a person will not violate a personal probability assessment if it has financial consequences. Good (1950) makes this idea more axiomatic by observing that people have personal probability assessments about many things around them rather than just one, and in order for these disparate comparative statements to form a *body of beliefs* they need to be free of contradictions. For example, if a person thinks that $A$ is more likely to occur than $B$, and $B$ is more likely to occur than $C$, then this person cannot coherently believe that $C$ is more likely than $A$ (transitivity). Furthermore, Good adds the explicitly Bayesian idea that people are constantly updating these personal probabilities as new information is observed,

although there is evidence that people have subadditive notions of probability when making calculations (the probability of some event plus the probability of its complement do not add to certainty).

The position underlying nearly all Bayesian work is the subjective probability characterization, although there have been many attempts to "objectify" Bayesian analysis (see Chapter **??**). Prior information is formalized in the Bayesian framework and this prior information can be subjective in the sense that the researcher's experience, intuition, and theoretical ideas are included. It is also common to base the prior information on previous studies, experiments, or just personal observations and this process is necessarily subject to a limited (although possibly large) number of observations rather than the infinite number assumed under the frequentist view. We will return to the theme of subjectivity contained in prior information in Chapter **??** and elsewhere, but the principal point is that *all* statistical models include subjective decisions, and therefore we should *ceteris paribus* prefer one that is the most explicit about assumptions. This is exactly the sense that the Bayesian prior provides readers with a specific, formalized statement of currently assumed knowledge in probabilistic terms.

### 1.3.1 Required Probability Principles

There are some simple but important principles and notational conventions that must be understood before proceeding. We will not worry too much about measure theory until Chapter **??**, and the concerned reader is directed to the first chapter of any mathematical statistics text or the standard reference works of Billingsley (1995), Chung (1974), and Feller (1990, Volumes 1 and 2). Abstract events are indicated by capital Latin letters: $A$, $B$, $C$, etc. A probability function corresponding to some event $A$ is always indicated by $p(A)$. The complement of the event $A$ is denoted $A^c$, and it is a consequence of Kolmogorov's axioms listed above that $p(A^c) = 1 - p(A)$. The union of two events is indicated by $A \cup B$ and the intersection by $A \cap B$. For any two events: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$. If two events are independent, then $p(A \cap B) = p(A)p(B)$, but not necessarily the converse (the product relationship does not imply independence).

Central to Bayesian thinking is the idea of conditionality. If an event $B$ is material to another event $A$ in the sense that the occurrence or non-occurrence of $B$ affects the probability of $A$ occurring, then we say that $A$ is conditional on $B$. It is a basic tenet of Bayesian statistics that we update our probabilities as new relevant information is observed. This is done with the definition of conditional probability given by: $p(A|B) = p(A \cap B)/p(B)$, which is read as "the probability of $A$ given $B$ is equal to the probability of $A$ and $B$ divided by the probability of $B$."

In general the quantities of interest here are random variables rather than the simple discrete events above. A random variable $X$ is defined as a measurable function from a probability space to a state space. This can be defined very technically (Shao 2005, p.7), but for our purposes it is enough to understand that the random variable connects

possible occurrences of some data value with a probability structure that reflects the relative "frequency" of these occurrences. The function is thus defined over a specific state space of all possible realizations of the random variable, called support. Random variables can be discrete or continuous. For background details see Casella and Berger (2002), Shao (2005), or the essay by Doob (1996). The expression $p(X \cap Y)$ is usually denoted as $p(X, Y)$, and is referred to as the *joint distribution* of random variables $X$ and $Y$. Marginal distributions are then simply $p(X)$ and $p(Y)$. Restating the principle above in this context, for two independent random variables the joint distribution is just the product of the marginals, $p(X, Y) = p(X)p(Y)$. Typically we will need to integrate expressions like $p(X, Y)$ to get marginal distributions of interest. Sometimes this is done analytically, but more commonly we will rely on computational techniques.

We will make extensive use of expected value calculations here. Recall that if a random variable $X$ is distributed $p(X)$, the expected value of some function of the random variable, $h(X)$, is

$$E[h(X)] = \begin{cases} \sum_{i=1}^{k} h(x_i)p(x_i) & k\text{-category discrete case} \\ \int_{\mathcal{X}} h(X)p(X)dx & \text{continuous case.} \end{cases} \tag{1.1}$$

Commonly $h(X) = X$, and we are simply concerned with the expectation of the random variable itself. In the discrete case this is a very intuitive idea as the expected value can be thought of as a probability-weighted average over possible events. For the continuous case there are generally limits on the integral that are dictated by the support of the random variable, and sometimes these are just given by $[-\infty, \infty]$ with the idea that the PDF (probability density function) indicates zero and non-zero regions of density. Also $p(X)$ is typically a conditional statement: $p(X|\theta)$. For the $k \times 1$ vector $\mathbf{X}$ of discrete random variables, the expected value is: $E[\mathbf{X}] = \sum \mathbf{X}p(\mathbf{X})$. With the expected value of a function of the continuous random vector, it is common to use the *Riemann-Stieltjes integral* form (found in any basic calculus text): $E[f(\mathbf{X})] = \int_{\mathcal{X}} f(\mathbf{X})dF(\mathbf{X})$, where $F(\mathbf{X})$ denotes the joint distribution of the random variable vector $\mathbf{X}$. The principles now let us look at Bayes' Law in detail.

## 1.4 Bayes' Law

The Bayesian statistical approach is based on updating information using what is called Bayes' Law (and synonymously Bayes' Theorem) from his famous 1763 essay. The Reverend Thomas Bayes was an amateur mathematician whose major contribution (the others remain rather obscure and do not address the same topic) was an essay found and published two years after his death by his friend Richard Price. The enduring association of an important branch of statistics with his name actually is somewhat of an exaggeration of

the generalizeability of this work (Stigler 1982). Bayes was the first to explicitly develop this famous law, but it was Laplace (1774, 1781) who (apparently independently) provided a more detailed analysis that is perhaps more relevant to the practice of Bayesian statistics today. See Stigler (1986) for an interesting historical discussion and Sheynin (1977) for a detailed technical analysis. Like Bayes, Laplace assumed a uniform distribution for the unknown parameter, but he worried much less than Bayes about the consequences of this assumption. Uniform prior distributions are simply "flat" distributions that assign equal probability for every possible outcome.

Suppose there are two events of interest $A$ and $B$, which are not independent. We know from basic axioms of probability that the conditional probability of $A$ given that $B$ has occurred is given by:

$$p(A|B) = \frac{p(A,B)}{p(B)}, \tag{1.2}$$

where $p(A|B)$ is read as "the probability of $A$ given that $B$ has occurred, $p(A,B)$ is the "the probability that both $A$ and $B$ occur" (i.e., the joint probability) and $p(B)$ is just the unconditional probability that $B$ occurs. Expression (1.2) gives the probability of $A$ after some event $B$ occurs. If $A$ and $B$ are independent here then $p(A,B) = p(A)p(B)$ and (1.2) becomes uninteresting.

We can also define a different conditional probability in which $A$ occurs first:

$$p(B|A) = \frac{p(B,A)}{p(A)}. \tag{1.3}$$

Since the probability that $A$ and $B$ occur is the same as the probability that $B$ and $A$ occur ($p(A,B) = p(B,A)$), then we can rearrange (1.2) and (1.3) together in the following way:

$$p(A,B) = p(A|B)p(B)$$

$$p(B,A) = p(B|A)p(A)$$

$$p(A|B)p(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(A)}{p(B)}p(B|A). \tag{1.4}$$

The last line is the famous Bayes' Law. This is really a device for "inverting" conditional probabilities. Notice that we could just as easily produce $p(B|A)$ in the last line above by moving the unconditional probabilities to the left-hand side in the last equality.

We can also use Bayes' Law with the use of odds, which is a common way to talk about uncertainty related to probability. The odds of an event is the ratio of the probability of an event happening to the probability of the event not happening. So for the event $A$, the odds of this event is simply:

$$Odds = \frac{p(A)}{1-p(A)} = \frac{p(A)}{p(\neg A)}, \tag{1.5}$$

which is this ratio expressed in two different ways. Note the use of "$\neg A$" for "not $A$," which is better notation when the complement of $A$ isn't specifically defined and we care only that event $A$ did not happen. Since this statement is not conditional on any other quantity, we can call it a "prior odds." If we make it conditional on $B$, then it is called a "posterior odds," which is produced by multiplying the prior odds by the reverse conditional with regard to $B$:

$$\frac{p(A|B)}{p(\neg A|B)} = \frac{p(A)}{p(\neg A)} \frac{p(B|A)}{p(B|\neg A)}. \tag{1.6}$$

The last ratio, $p(B|A)/p(B|\neg A)$ is the "likelihood ratio" for $B$ under the two conditions for $A$. This is actually the ratio of two expressions of Bayes' Law in the sense of (1.4), which we can see with the introduction of the ratio $p(B)/p(B)$:

$$\frac{p(A|B)}{p(\neg A|B)} = \frac{p(A)/p(B)}{p(\neg A)/p(B)} \frac{p(B|A)}{p(B|\neg A)}. \tag{1.7}$$

This ratio turns out to be very useful in Bayesian model consideration since it implies a test between the two states of nature, $A$ and $\neg A$, given the observation of some pertinent information $B$.

■ **Example 1.1: Testing with Bayes' Law.** How is this useful? As an example, hypothetically assume that 2% of the population of the United States are members of some extremist **M**ilitia group ($p(M) = 0.02$), a fact that some members might attempt to hide and therefore not readily admit to an interviewer. A survey is 95% accurate on positive **C**lassification, $p(C|M) = 0.95$, ("sensitivity") and the unconditional probability of classification (i.e., regardless of actual militia status) is given by $p(C) = 0.05$. To illustrate how $p(C)$ is really the normalizing constant obtained by accumulating over all possible events, we will stipulate the additional knowledge that the survey is 97% accurate on negative classification, $p(C^c|M^c) = 0.97$ ("specificity"). The unconditional probability of classifying a respondent as a militia member results from accumulation of the probability across the sample space of survey events using the Total Probability Law: $p(C) = p(C \cap M) + p(C \cap M^c) = p(C|M)p(M) + [1 - p(C^c|M^c)]p(M^c) = (0.95)(0.02) + (0.03)(0.98) \cong 0.05$.

Using Bayes' Law, we can now derive the probability that someone positively classified by the survey as being a militia member really *is* a militia member:

$$p(M|C) = \frac{p(M)}{p(C)} p(C|M) = \frac{0.02}{0.05}(0.95) = 0.38. \tag{1.8}$$

The startling result is that although the probability of correctly classifying an individual as a militia member given they really are a militia member is 0.95, the probability that an individual really is a militia member given that they are positively classified is only 0.38.

The highlighted difference here between the order of conditional probability is often substantively important in a policy or business context. Consider the problem of

designing a home pregnancy test. Given that there exists a fundamental business trade-off between the reliability of the test and the cost to consumers, no commercially viable product will have perfect or near-perfect test results. In designing the chemistry and packaging of the test, designers will necessarily have to compromise between the probability of **PR**egnancy given positive **T**est results, $p(\mathbf{PR}|\mathbf{T})$, and the probability of positive test results given pregnancy, $p(\mathbf{T}|\mathbf{PR})$. Which one is more important? Clearly, it is better to maximize $p(\mathbf{T}|\mathbf{PR})$ at the expense of $p(\mathbf{PR}|\mathbf{T})$, as long as the reduction in the latter is reasonable: it is preferable to give a higher number of false positives, sending women to consult their physician to take a more sensitive test, than to fail to notify many pregnant women. This reduces the possibility that a woman who does not realize that she is pregnant might continue unhealthy practices such as smoking, drinking, or maintaining a poor diet. Similarly, from the perspective of general public health, it is better to have preliminary tests for deadly contagious diseases designed to be similarly conservative with respect to false positives.

### 1.4.1 Bayes' Law for Multiple Events

It would be extremely limiting if Bayes' Law only applied to two alternative events. Fortunately the extension to multiple events is quite easy. Suppose we observe some data $\mathbf{D}$ and are interested in the relative probabilities of three events $A$, $B$, and $C$ conditional on these data. These might be rival hypotheses about some social phenomenon for which the data are possibly revealing. Thinking just about event $A$, although any of the three could be selected, we know from Bayes' Law that:

$$p(A|\mathbf{D}) = \frac{p(\mathbf{D}|A)p(A)}{p(\mathbf{D})}. \tag{1.9}$$

We also know from the Total Probability Law and the definition of conditional probability that:

$$\begin{aligned} p(\mathbf{D}) &= p(A \cap \mathbf{D}) + p(B \cap \mathbf{D}) + p(C \cap \mathbf{D}) \\ &= p(\mathbf{D}|A)p(A) + p(\mathbf{D}|B)p(B) + p(\mathbf{D}|C)p(C). \end{aligned} \tag{1.10}$$

This means that if we substitute the last line into the expression for Bayes' Law, we get:

$$p(A|\mathbf{D}) = \frac{p(\mathbf{D}|A)p(A)}{p(\mathbf{D}|A)p(A) + p(\mathbf{D}|B)p(B) + p(\mathbf{D}|C)p(C)}, \tag{1.11}$$

which demonstrates that the conditional distribution for any of the rival hypotheses can be produced as long as there exist unconditional distributions for the three rival hypotheses, $p(A)$, $p(B)$, and $p(C)$, and three statements about the probability of the data given these three hypotheses, $p(\mathbf{D}|A)$, $p(\mathbf{D}|B)$, $p(\mathbf{D}|C)$. The first three probability statements are called prior distributions because they are unconditional from the data and therefore presumably determined before observing the data. The second three probability statements are merely

PDF (probability density function) or PMF (probability mass function) statements in the conventional sense. All this means that a posterior distribution, $p(A|\mathbf{D})$, can be determined through Bayes' Law to look at the weight of evidence for any one of several rival hypotheses or claims.

There is a more efficient method for making statements like (1.11) when the number of outcomes increases. Rather than label the three hypotheses as we have done above, let us instead use $\theta$ as an unknown parameter whereby different regions of its support define alternative hypotheses. So statements may take the form of "Hypothesis A: $\theta < 0$," or any other desired statement. To keep track of the extra outcome, denote the three hypotheses as $\theta_i,\ i = 1, 2, 3$. Now (1.11) is given more generally for $i = 1, 2, 3$ as:

$$p(\theta_i|\mathbf{D}) = \frac{p(\mathbf{D}|\theta_i)p(\theta_i)}{\sum_{j=1}^{3} p(\mathbf{D}|\theta_j)p(\theta_j)} \tag{1.12}$$

for the posterior distribution of $\theta_i$. This is much more useful and much more in line with standard Bayesian models in the social and behavioral sciences because it allows us to compactly state Bayes' Law for any number of discrete outcomes/hypotheses, say $k$ for instance:

$$p(\theta_i|\mathbf{D}) = \frac{p(\theta_i)p(\mathbf{D}|\theta_i)}{\sum_{j=1}^{k} p(\theta_j)p(\mathbf{D}|\theta_j)}. \tag{1.13}$$

Consider also that the denominator of this expression averages over the $\theta$ variables and therefore just produces the *marginal distribution of the sample data*, which we could overtly label as $p(\mathbf{D})$. Doing this provides a form that very clearly looks like the most basic form of Bayes' Law: $p(\theta_i|\mathbf{D}) = p(\theta_i)p(\mathbf{D}|\theta_i)/p(\mathbf{D})$. We can contrast this with the standard likelihood approach in the social sciences (King 1989, p.22), which overtly ignores information available through a prior and has no use for the denominator above: $L(\hat{\theta}|y) \propto p(y|\hat{\theta})$, in King's notation using proportionality since the objective is simply to find the mode and curvature around this mode, thus making constants unimportant. Furthermore, in the continuous case, where the support of $\theta$ is over some portion of the real line, and possibly all of it, the summation in (1.13) is replaced with an integral. The continuous case is covered in the next chapter.

■ **Example 1.2: Monty Hall.** The well-known Monty Hall problem (Selvin 1975) can be analyzed using Bayes' Law. Suppose that you are on the classic game show *Let's Make a Deal* with its personable host Monty Hall, and you are to choose one of three doors, A, B, or C. Behind two of the doors are goats and behind the third door is a new car, and each door is equally likely to award the car. Thus, the probabilities of selecting the car for each door at the beginning of the game are simply:

$$p(A) = \frac{1}{3}, \qquad p(B) = \frac{1}{3}, \qquad p(C) = \frac{1}{3}.$$

After you have picked a door, say A, before showing you what is behind that door Monty opens another door, say B, revealing a goat. At this point, Monty gives you

the opportunity to switch doors from A to C if you want to. What should you do? The psychology of this approach is to suggest the idea to contestants that they must have picked the correct door and Monty is now trying to induce a change. A naïve interpretation is that you should be indifferent to switching due to a perceived probability of 0.5 of getting the car with either door since there are two doors left. To see that this is false, recall that Monty is not a benign player in this game. He is deliberately trying to deny you the car. Therefore consider his probability of opening door B. Once you have picked door A, success is clearly conditional on what door of the three possibilities actually provides the car since Monty has this knowledge and the contestant does not. After the first door selection, we can define the three conditional probabilities as follows:

| | |
|---|---|
| The probability that Monty opens door B, given the car is behind A: | $p(B_{\text{Monty}}\|A) = \frac{1}{2}$ |
| The probability that Monty opens door B, given the car is behind B: | $p(B_{\text{Monty}}\|B) = 0$ |
| The probability that Monty opens door B, given the car is behind C: | $p(B_{\text{Monty}}\|C) = 1.$ |

Using the definition of conditional probability, we can derive the following three joint probabilities:

$$p(B_{\text{Monty}}, A) = p(B_{\text{Monty}}|A)p(A) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

$$p(B_{\text{Monty}}, B) = p(B_{\text{Monty}}|B)p(B) = 0 \times \frac{1}{3} = 0$$

$$p(B_{\text{Monty}}, C) = p(B_{\text{Monty}}|C)p(C) = 1 \times \frac{1}{3} = \frac{1}{3}.$$

Because there are only three possible events that cover the complete sample space, and these events are non-overlapping (mutually exclusive), they form a partition of the sample space. Therefore the sum of these three events is the unconditional probability of Monty opening door B, which we obtain with the Total Probability Law:

$$\begin{aligned} p(B_{\text{Monty}}) &= p(B_{\text{Monty}}, A) + p(B_{\text{Monty}}, B) + p(B_{\text{Monty}}, C) \\ &= \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

Now we can apply Bayes' Law to obtain the two probabilities of interest:

$$p(A|B_{\text{Monty}}) = \frac{p(A)}{p(B_{\text{Monty}})} p(B_{\text{Monty}}|A) = \frac{\frac{1}{3}}{\frac{1}{2}} \times \frac{1}{2} = \frac{1}{3}$$

$$p(C|B_{\text{Monty}}) = \frac{p(C)}{p(B_{\text{Monty}})} p(B_{\text{Monty}}|C) = \frac{\frac{1}{3}}{\frac{1}{2}} \times 1 = \frac{2}{3}.$$

Therefore you are twice as likely to win the car if you switch to door C! This example demonstrates that Bayes' Law is a fundamental component of probability calculations, and the principle will be shown to be the basis for an inferential system of statistical analysis. For a nice generalization to $N$ doors, see McDonald (1999).

---

## 1.5 Conditional Inference with Bayes' Law

To make the discussion more concrete and pertinent, consider a simple problem in sociology and crime studies. One quantity of interest to policy-makers is the recidivism rate of prisoners released after serving their sentence. The quantity of interest is the probability of committing an additional crime and returning to prison. Notice that this is a very elusive phenomenon. Not only are there regional, demographic, and individualistic differences, but the aggregate probability is also constantly in flux, given entries and exits from the population as well as exogenous factors (such as the changing condition of the economy). Typically, we would observe a change in law or policy at the state or federal level, and calculate a point estimate from observed recidivism that follows.

Perhaps we should not assume that there is some fixed value of the recidivism probability, $\mathcal{A}$, and that it should be estimated with a single point, say $\bar{\mathcal{A}}$. Instead, consider this unknown quantity in probabilistic terms as the random variable $A$, which means conceptualizing a distribution for the probability of recidivism. Looking at data from previous periods, we might have some reasonable guess about the distribution of this probability parameter, $p(A)$, which is of course the prior distribution since it is not conditional on the information at hand, $B$.

In all parametric statistical inference, a model is proposed and tested in which an event has some probability of occurring given a specific value of the parameter. This is the case for both Bayesian and traditional approaches, and is just a recognition that the researcher must specify a data generation model. Let us call this quantity $p(B|A)$, indicating that for posited values of recidivism, we would expect to see a particular pattern of events. For instance, if recidivism suddenly became much higher in a particular state, then there might be pressure on the legislature to toughen sentencing and parole laws. This is a probability model and we do not need to have a specific value of $A$ to specify a parametric form (i.e., PMF or PDF). Of course what we are really interested in is $p(A|B)$, the (posterior) distribution of $A$ after

having observed an event, which we obtain using Bayes' Law: $p(A|B) = \frac{p(A)}{p(B)}p(B|A)$. From a public policy perspective, this is equivalent to asking how do recidivism rates change for given statutes.

We are still missing one component of the right-hand-side of Bayes' Law here, the *unconditional* probability of generating the legal or policy event, $p(B)$. This is interpretable as the denominator of (1.13), but to a Bayesian this is an unimportant *probability* statement since $B$ has already been observed and therefore has probability one of occurring. Recall that for Bayesians, observed quantities are fixed and unobserved quantities are assigned probability statements. So there is no point in treating $B$ probabilistically if the actual facts are sitting on our desk right now. This does not mean that everything is known about all possible events, missing events, or events occurring in the future. It just means that everything is known about *this* event. So the only purpose for $p(B)$ in this context is to make sure that $p(A|B)$ sums or integrates to one.

This last discussion suggests simply treating $p(B)$ as a normalizing constant since it does not change the *relative* probabilities for $A$. Maybe this is a big conceptual leap, but if we could recover unconditional $p(B)$ later, it is convenient to just use it then to make the conditional statement, $p(A|B)$, a properly scaled probability statement. So if $p(A|B)$ summed or integrated to five instead of one, we would simply divide everywhere by five and lose nothing but the agony of carrying $p(B)$ through the calculations. If we temporarily ignore $p(B)$, then:

$$p(A|B) \propto p(A)p(B|A), \tag{1.14}$$

where "$\propto$" means "proportional to" (i.e., the *relative* probabilities are preserved). So the final estimated probability of recidivism (in our example problem) given some observed behavior, is proportional to prior notions about the distribution of the probability times the parametric model assumed to be generating the new observed event. The conditional probability of interest on the left-hand side of (1.14) is a balance between things we have already seen or believe about recidivism, $p(A)$, and the contribution from the new observation, $p(B|A)$. It is important to remember that there are occasions where the data are more influential than the prior and vice-versa. This is comforting since if the data are poor in size or information we want to rely more on prior knowledge, prior research, researcher or practitioner information and so on. Conversely, if the data are plentiful and highly informed, then we should not care much about the form of the prior information. Remarkably, the Bayesian updating process in (1.14) has this trade-off automatically built-in to the process.

As described, this is an ideal paradigm for inference in the social and behavioral sciences, since it is consentaneously desirable to build models that test theories with newly observed events or data, but also based on previous research and knowledge. We never start a data analysis project with absolutely no *a priori* notions whatsoever about the state of nature (or at least we should not!). This story actually gets better. As the number of events increases, $p(B|A)$ becomes progressively more influential in determining $p(A|B)$. That is, the greater the number of our new observations, the less important are our previous convictions: $p(A)$.

Also, if either of the two distributions, $p(A)$ and $p(B|A)$, are widely dispersed relative to the other, then this distribution will have less of an impact on the final probability statement. We will see this principle detailed-out in Chapter **??**. The natural weighting of these two distributions suitably reflects relative levels of uncertainty in the two quantities.

### 1.5.1 Statistical Models with Bayes' Law

The statistical role of the quantities in (1.14) has not yet been identified since we have been talking abstractly about "events" rather than conventional data. The goal of inference is to make claims about unknown quantities using data currently in hand. Suppose that we designate a generic Greek character to denote an unobserved parameter that is the objective of our analysis. As is typical in these endeavors, we will use $\theta$ for this purpose. What we usually have available to us is generically (and perhaps a little vaguely) labeled $\mathbf{D}$ for data. Therefore, the objective is to obtain a probabilistic statement about $\theta$ given $\mathbf{D}$: $p(\theta|\mathbf{D})$.

Inferences in this book, and in the majority of Bayesian and non-Bayesian statistics, are made by first specifying a parametric model for the data generating process. This defines what the data should be expected to look like given a specific probabilistic function conditional on unknown variable values. These are the common probability density functions (continuous data) and probability mass functions (discrete data) that we already know, such as normal, binomial, chi-square, etc., denoted by $p(\mathbf{D}|\theta)$.

Now we can relate these two conditional probabilities using (1.14):

$$\pi(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta), \tag{1.15}$$

where $p(\theta)$ is a formalized statement of the prior knowledge about $\theta$ before observing the data. If we know little, then this prior distribution should be a vague probabilistic statement and if we know a lot then this should be a very narrow and specific claim. The right-hand side of (1.15) implies that the *post*-data inference for $\theta$ is a compromise between prior information and the information provided by the new data, and the left-hand side of (1.15) is the posterior distribution of $\theta$ since it provides the updated distribution for $\theta$ after conditioning on the data.

Bayesians describe $\pi(\theta|\mathbf{D})$ to readers via distributional summaries such as means, modes, quantiles, probabilities over regions, traditional-level probability intervals, and graphical displays. Once the posterior distribution has been calculated via (1.15), everything about it is known and it is entirely up to the researcher to highlight features of interest. Often it is convenient to report the posterior mean and variance in papers and reports since this is what non-Bayesians do by default. We can calculate the posterior mean using an expected value calculation, confining ourselves here to the continuous case:

$$E[\theta|\mathbf{D}] = \int_{-\infty}^{\infty} \theta\pi(\theta|\mathbf{D})d\theta \tag{1.16}$$

and the posterior variance via a similar process:

$$\begin{aligned}
\text{Var}[\theta|\mathbf{D}] &= E\left[(\theta - E[\theta|\mathbf{D}])^2|\mathbf{D}\right] \\
&= \int_{-\infty}^{\infty} (\theta - E[\theta|\mathbf{D}])^2 \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \left(\theta^2 - 2\theta E[\theta|\mathbf{D}]) + E[\theta|\mathbf{D}]^2\right) \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \theta^2 \pi(\theta|\mathbf{D}) d\theta - 2E[\theta|\mathbf{D}] \int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta + \left(\int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta\right)^2 \\
&= E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2 \qquad (1.17)
\end{aligned}$$

given some minor regularity conditions about switching the order of integration (see Casella and Berger 2002, Chapter 1). An obvious summary of this posterior would then be the vector $(E[\theta|\mathbf{D}], \sqrt{\text{Var}[\theta|\mathbf{D}]})$, although practicing Bayesians tend to prefer reporting more information.

Researchers sometimes summarize the Bayesian posterior distribution in a deliberately traditional, non-Bayesian way in an effort to communicate with some readers. The posterior mode corresponds to the maximum likelihood point estimate and is calculated by:

$$M(\theta) = \underset{\theta}{\text{argmax}} \pi(\theta|\mathbf{D}), \qquad (1.18)$$

where argmax function specifies the value of $\theta$ that maximizes $\pi(\theta|\mathbf{D})$. Note that the denominator of Bayes' Law is unnecessary here since the function has the same mode with or with including it. The accompanying measure of curvature (e.g., Fisher Information, defined in Appendix **??**) can be calculated with standard analytical tools or more conveniently from MCMC output with methods introduced in Chapter **??**. The posterior median is a slightly less popular choice for a Bayesian point estimate, even though its calculation from MCMC output is trivial from just sorting empirical draws and determining the mid-point.

■ **Example 1.3: Summarizing a Posterior Distribution from Exponential Data.** Suppose we had generic data, $\mathbf{D}$, distributed $p(\mathbf{D}|\theta) = \theta e^{-\theta \mathbf{D}}$, which can be either a single scalar or a vector for our purposes. Thus $\mathbf{D}$ is exponentially distributed with the support $[0{:}\infty)$; see Appendix **??** for details on this probability density function. We also need to specify a prior distribution for $\theta$: $p(\theta) = 1$, where $\theta \in [0{:}\infty)$. Obviously this prior distribution does not constitute a "proper" distribution in the Kolmogorov sense since it does not integrate to one (infinity, in fact). We should not let this bother us since this effect is canceled out due to its presence in both the numerator and denominator of Bayes' Law (a principle revisited in Chapter **??** in greater detail).

This type of prior is often used to represent high levels of prior uncertainty, although it is not completely *uninformative*. Using (1.15) now, we get:

$$\pi(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta) = (1)\theta e^{-\theta\mathbf{D}} = \theta e^{-\theta\mathbf{D}}. \tag{1.19}$$

This posterior distribution has mean:

$$E[\theta|\mathbf{D}] = \int_0^\infty (\theta)\left(\theta e^{-\theta\mathbf{D}}\right) d\theta = \frac{2}{\mathbf{D}^3}, \tag{1.20}$$

which is found easily using two iterations of integration-by-parts. Also, the expectation of $\theta^2|\mathbf{D}$ is:

$$E[\theta^2|\mathbf{D}] = \int_0^\infty (\theta^2)\left(\theta e^{-\theta\mathbf{D}}\right) d\theta = \frac{6}{\mathbf{D}^4}, \tag{1.21}$$

which is found using three iterations of integration-by-parts now. So the posterior variance is:

$$\text{Var}[\theta|\mathbf{D}] = E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2 = 6\mathbf{D}^{-4} - 4\mathbf{D}^{-6}. \tag{1.22}$$

The notation would be slightly different if $\mathbf{D}$ were a vector.

Using these quantities we can perform an intuitive Bayesian hypothesis test, such as asking what is the posterior probability that $\theta$ is positive $p(\theta|\mathbf{D}) > 0$. In the context of a regression coefficient, this would be the probability that increases in the corresponding $X$ explanatory variable have a positive affect on the $Y$ outcome variable. Testing will be discussed in detail in Chapter **??**. We can also use these derived quantities to create a Bayesian version of a confidence interval, the credible interval for some chosen $\alpha$ level:

$$\left[E[\theta|\mathbf{D}] - \sqrt{\text{Var}[\theta|\mathbf{D}]}f_{\alpha/2} : E[\theta|\mathbf{D}] + \sqrt{\text{Var}[\theta|\mathbf{D}]}f_{1-\alpha/2}\right], \tag{1.23}$$

where $f_{\alpha/2}$ and $f_{1-\alpha/2}$ are lower and upper tail values for some assumed or empirically observed distribution for $\theta$ (Chapter **??**).

The purpose of this brief discussion is to highlight the fact that conditional probability underlies the ability to update previous knowledge about the distribution of some unknown quantity. This is precisely in line with the iterative scientific method, which postulates theory improvement through repeated specification and testing with data. The Bayesian approach combines a formal structure of rules with the mathematical convenience of probability theory to develop a process that "learns" from the data. The result is a powerful and elegant tool for scientific progress in many disciplines.

## 1.6 Science and Inference

This is a book about the scientific process of discovery in the social and behavioral sciences. Data analysis is best practiced as a theory-driven exploration of collected observations with the goal of uncovering important and unknown effects. This is true regardless of academic discipline. Yet some fields of study are considered more rigorously analytical in this pursuit than others.

The process described herein is that of *inference*: making probabilistic assertions about unknown quantities. It is important to remember that "in the case of uncertain inference, however, the very uncertainty of uncertain predictions renders question of their proof or disproof almost meaningless" (Wilkinson 1977). Thus, confusion sometimes arises in the interpretation of the inferential process as a scientific, investigative endeavor.

### 1.6.1 The Scientific Process in Our Social Sciences

Are the social and behavioral sciences truly "scientific"? This is a question asked about fields such as sociology, political science, economics, anthropology, and others. It is not a question about whether serious, rigorous, and important work has been done in these endeavors; it is a question about the research process and whether it conforms to the empirico-deductive model that is historically associated with the natural sciences. From a simplistic view, this is an issue of the conformance of research in the social and behavioral sciences to the so-called scientific method. Briefly summarized, the scientific method is characterized by the following steps:

- ▷ Observe or consider some phenomenon.
- ▷ Develop a theory about the cause(s) of this phenomenon and articulate it in a specific hypothesis.
- ▷ Test this hypothesis by developing a model to fit experimentally generated or collected observational data.
- ▷ Assess the quality of the fit to the model and modify the theory if necessary, repeating the process.

This is sometimes phrased in terms of "prediction" instead of theory development, but we will use the more general term. If the scientific method as a process were the defining criterion for determining what is scientific and what is not, then it would be easy to classify a large proportion of the research activities in the social and behavioral sciences as scientific. However useful this typology is in teaching children about empirical investigation, it is a poor standard for judging academic work.

Many authors have posited more serviceable definitions. Braithwaite (1953, p.1) notes:

> The function of a science, in this sense of the word, is to establish general laws covering the

> behavior of the empirical events or objects with which the science in question is concerned, and thereby to enable us to connect together our knowledge of the separately known events, and to make reliable predictions of events as yet unknown.

The core of this description is the centrality of empirical observation and subsequent accumulation of knowledge. Actually, "science" is the Latin word for knowledge. Legendary psychologist B. F. Skinner (1953, p.11) once observed that "science is unique in showing a cumulative process." It is clear from the volume and preservation of published research that social and behavioral scientists *are* actively engaged in empirical research and knowledge accumulation (although the quality and permanence of this foundational knowledge might be judged to differ widely by field). So what is it about these academic pursuits that makes them only suspiciously scientific to some? The three defining characteristics about the *process* of scientific investigation are empiricism, objectivity, and control (Singleton and Straight 2004). This is where there is lingering and sometimes legitimate criticism of the social and behavioral sciences as being "unscientific."

The social and behavioral sciences are partially empirical (data-oriented) and partially normative (value-oriented), the latter because societies develop norms about human behavior, and these norms permeate academic thought prior to the research process. For instance, researchers investigating the onset and development of AIDS initially missed the effects of interrelated social factors such as changes in behavioral risk factors, personal denial, and reluctance to seek early medical care on the progress of the disease as a sociological phenomenon (Kaplan *et al.* 1987). This is partially because academic investigators as well as health professionals made normative assumptions about individual responses to sociological effects. Specifically, researchers investigating human behavior, whether political, economic, sociological, psychological, or otherwise, cannot completely divorce their prior attitudes about some phenomenon of interest the way a physicist or chemist can approach the study of the properties of thorium: atomic number 90, atomic symbol Th, atomic weight 232.0381, electron configuration $[Rn]7^s26d^2$. This criticism is distinct from the question of objectivity; it is a statement that students of human behavior are themselves human.

We are also to some extent driven by the quality and applicability of our tools. Many fields have radically progressed after the introduction of new analytical devices. Therefore, some researchers may have a temporary advantage over others, and may be able to answer more complex questions: "It comes as no particular surprise to discover that a scientist formulates problems in a way which requires for their solution just those techniques in which he himself is especially skilled" (Kaplan 1964). The objective of this book is to "level the pitch" by making an especially useful tool more accessible to those who have thus far been accordingly disadvantaged.

### 1.6.2 Bayesian Statistics as a Scientific Approach to Social and Behavioral Data Analysis

The standard frequentist interpretation of probability and inference assumes an infinite series of trials, replications, or experiments using the same research design. The "objectivist" paradigm is typically explained and justified through examples like multiple tosses of a coin, repeated measurements of some physical quantity, or samples from some ongoing process like a factory output. This perspective, which comes directly from Neyman and Pearson (1928a, 1928b, 1933a, 1933b, 1936a, 1936b), and was formalized by Von Mises (1957) among others, is combined with an added Fisherian fixation with p-values in typical inference in the social and behavioral sciences (Gill 1999). Efron (1986), perhaps overly kindly, calls this a "rather uneasy alliance."

Very few, if any, social scientists would be willing to seriously argue that human behavior fits this objectivist long-run probability model. Ideas like "personal utility," "legislative ideal points," "cultural influence," "mental states," "personality types," and "principal-agent goal discrepancy" do not exist as parametrically uniform phenomena in some physically tangible manner. In direct contrast, the Bayesian or "subjective" conceptualization of probability is the degree of belief that the individual researcher is willing to personally assign and defend. This is the idea that an individual *personally* assigns a probability measure to some event as an expression of uncertainty about some event that may only be relevant to one observational situation or experiment.

The central idea behind subjective probability is the assignment of a prior probability based on what information one currently possesses and under what circumstances one would be willing to place an even wager. Naturally, this probability is updated as new events occur, therefore incorporating serial events in a systematic manner. The core disagreement between the frequentist notion of objective probability and the Bayesian idea of subjective probability is that frequentists see probability measure as a property of the outside world and Bayesians view probability as a personal internalization of observed uncertainty. The key defense of the latter view is the inarguable point that all statistical models are subjective: decisions about variable specifications, significance thresholds, functional forms, and error distributions are completely nonobjective.[1] In fact, there are instances when Bayesian subjectivism is more "objective" than frequentist objectivism with regard to the impact of irrelevant information and arbitrary decision rules (e.g., Edwards, Lindman, and Savage 1963, p.239).

[1] As a brief example, consider common discussions of reported analyses in social science journals and books that talk about reported model parameters being "of the wrong sign." What does this statement mean? The author is asserting that the statistical model has produced a regression coefficient that is positive when it was *a priori* expected to be negative or vice versa. What is this statement in effect? It is a prior statement about knowledge that existed before the model was constructed. Obviously this is a form of the Bayesian prior without being specifically articulated as such.

Given the existence of subjectivity in all scientific data analysis endeavors,[2] one should prefer the inferential paradigm that gives the most *overt* presentation of model assumptions. This is clearly the Bayesian subjective approach since both prior information and posterior uncertainty are given with specific, clearly stated model assumptions. Conversely, frequentist models are rarely presented with caveats such as "Caution: the scientific conclusions presented here depend on repeated trials that were never performed," or "Warning: prior assumptions made in this model are not discussed or clarified." If there is a single fundamental scientific tenet that underlies the practice and reporting of empirical evidence, it is the idea that all important model characteristics should be provided to the reader. It is clear then which of the two approaches is more "scientific" by this criterion. While this discussion specifically contrasts Bayesian and frequentist approaches, likelihood inference is equally subjective in every way, and as already explained, ignores available information.

These ideas of what sort of inferences social scientists make are certainly not new or novel. There is a rich literature to support the notion that the Bayesian approach is more in conformance with widely accepted scientific norms and practices. Poirer (1988, p.130) stridently makes this point in the case of prior specifications:

> I believe that subjective prior beliefs should play a *formal* role so that it is easier to investigate their impact on the results of the analysis. Bayesians must live with such honesty whereas those who introduce such beliefs informally need not.

The core of this argument is the idea that if the prior contains information that pertains to the estimation problem, then we are foolish to ignore it simply because it does not neatly fit into some familiar statistical process. For instance, Theil and Goldberger (1961) suggested "mixed" estimation some time ago, which is a way to incorporate prior knowledge about coefficients in a standard linear regression model by mixing earlier estimates into the estimation process and under very general assumptions is found to be simultaneously best linear unbiased with respect to both sample and prior information (see also Theil [1963]). This notion of combining information from multiple sources is not particularly controversial among statisticians, as observed by Samaniego and Reneau (1994, p.957):

> If a prior distribution contains "useful" information about an unknown parameter, then the Bayes estimator with respect to that prior will outperform the best frequentist rule. Otherwise, it will not.

A more fundamental advantage to Bayesian statistics is that both prior and posterior parameter estimates are assumed to have a distribution and therefore give a more realistic picture of uncertainty that is also more useful in applied work:

> With conventional statistics, the only uncertainty admitted to the analysis is

[2] See Press and Tanur (2001) for a fascinating account of the role of researcher-introduced subjectivity in a number of specific famous scientific breakthroughs, including discoveries by Galileo, Newton, Darwin, Freud, and Einstein.

> sampling uncertainty. The Bayesian approach offers guidance for dealing with the myriad sources of uncertainty faced by applied researchers in real analyses.

Western (1999, p.20). Lindley (1986, p.7) expresses a more biting statement of preference:

> Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly and was honest enough to admit he might have been wrong.

This book rests on the perspective, sampled above, that the Bayesian approach is not only useful for social and behavioral scientists, but it also provides a more compatible methodology for analyzing data in the manner and form in which it arrives in these disciplines. As we describe in subsequent chapters, Bayesian statistics establishes a rigorous analytical platform with clear assumptions, straightforward interpretations, and sophisticated extensions. For more extended discussions of the advantages of Bayesian analysis over alternatives, see Berger (1986b), Dawid (1982), Efron (1986), Good (1976), Jaynes (1976), and Zellner (1985). We now look at how the Bayesian paradigm emerged over the last 250 years.

---

## 1.7 Introducing Markov Chain Monte Carlo Techniques

In this section we briefly discuss Bayesian computation and give a preview of later chapters. The core message is that these algorithms are relatively simple to understand in the abstract.

Markov chain Monte Carlo (MCMC) set the Bayesians free. Prior to 1990, it was relatively easy to specify an interesting and realistic model with actual data whereby standard results were unobtainable. Specifically, faced with a high dimension posterior resulting from a regression-style model, it was often very difficult or even impossible to perform multiple integration across the parameter space to produce a regression table of marginal summaries. The purpose of MCMC techniques is to replace this difficult analytical integration process with iterative work by the computer. When calculations similar to (1.16) are multidimensional, there is a need to summarize each marginal distribution to provide useful results to readers in a table or other format for journal submission. The basic principle behind MCMC techniques is that if an iterative chain of computer-generated values can be set up carefully enough, and run long enough, then *empirical* estimates of integral quantities of interest can be obtained from summarizing the observed output. If each visited multidimensional location is recorded as a row vector in a matrix, then the marginalization for some parameter of interest is obtained simply by summarizing the individual dimension down the corresponding column. So we replace an analytical problem with a sampling problem, where the sampling process has the computer perform the difficult and repetitive processes. This is an enormously important idea to Bayesians and to others since it frees researchers

from having to make artificial simplifications to their model specifications just to obtain describable results.

These Markov chains are successive quantities that depend probabilistically only on the value of their immediate predecessor: the *Markovian property*. In general, it is possible to set up a chain to estimate multidimensional probability structures (i.e., desired probability distributions), by starting a Markov chain in the appropriate sample space and letting it run until it settles into the target distribution. Then when it runs for some time confined to this particular distribution, we can collect summary statistics such as means, variances, and quantiles from the simulated values. This idea has revolutionized Bayesian statistics by allowing the empirical estimation of probability distributions that could not be analytically calculated.

### 1.7.1 Simple Gibbs Sampling

As a means of continuing the discussion about conditional probability and covering some basic principles of the `R` language, this section introduces an important, and frequently used Markov chain Monte Carlo tool, the Gibbs sampler. The idea behind a Gibbs sampler is to get a marginal distribution for each variable by iteratively conditioning on interim values of the others in a continuing cycle until samples from this process empirically approximate the desired marginal distribution. Standard regression tables that appear in journals are simply marginal descriptions. There will be much more on this topic in Chapter **??** and elsewhere, but here we will implement a simple but instructive example.

As outlined by Example 2 in Casella and George (1992), suppose that we have two conditional distributions, where they are conditional on each other such that the parameter of one is the variable of interest in the other:

$$f(x|y) \propto y \exp[-yx], \quad f(y|x) \propto x \exp[-xy], \quad 0 < x, y < B < \infty. \tag{1.24}$$

These conditional distributions are both exponential probability density functions (see Appendix **??** for details). The upper bound, $B$, is important since without it there is no finite joint density and the Gibbs sampler will not work. It is possible, but not particularly pleasant, to perform the correct integration steps to obtain the desired marginal distributions: $f(x)$ and $f(y)$. Instead we will let the Gibbs sampler do the work computationally rather than us do it analytically.

The Gibbs sampler is defined by first identifying conditional distributions for each parameter in the model. These are conditional in the sense that they have dependencies on other parameters, and of course the data, which emerge from the model specification. The "transition kernel" for the Markov chain is created by iteratively cycling through these distributions, drawing values that are conditional on the latest draws of the dependencies. It is proven that this allows us to run a Markov chain that eventually settles into the desired limiting distribution that characterizes the marginals. In other language, it is an iterative process that cycles through conditional distributions until it reaches a stable status whereby

future samples characterize the desired distributions. The important theorem here assures us that when we reach this stable distribution, the autocorrelated sequence of values can be treated as an iid sample from the marginal distributions of interest. The amazing part is that this is accomplished simply by ignoring the time index, i.e., putting the values in a "bag" and just "shaking it up" to lose track of the order of occurrence. Gibbs sampling is actually even more general than this. Chib (1995) showed how Gibbs sampling can be used to compute the marginal distribution of the sample data, i.e., the denominator of (1.13), by using the individual parameter draws. This quantity is especially useful in Bayesian hypothesis testing and model comparison, as we shall see in Chapter **??**. The second half of this text applies this tool and similar methods of estimation.

For two parameters, $x$ and $y$, this process involves a starting point, $[x_0, y_0]$, and the cycles are defined by drawing random values from the conditionals according to:

$$\begin{array}{ll}
x_1 \sim f(x|y_0), & y_1 \sim f(y|x_1) \\
x_2 \sim f(x|y_1), & y_2 \sim f(y|x_2) \\
x_3 \sim f(x|y_2), & y_3 \sim f(y|x_3) \\
\vdots & \vdots \\
\vdots & \vdots \\
x_m \sim f(x|y_{m-1}), & y_m \sim f(y|x_m).
\end{array}$$

If we are successful, then after some reasonable period the values $x_j$, $y_j$ are safely assumed to be empirical samples from the correct marginal distribution. There are many theoretical and practical concerns that we are ignoring here, and the immediate objective here is to give a rough overview.

The following steps indicate how the Gibbs sampler is set up and run:

- ▷ Set the initial values: $B = 10$, and $m = 50,000$. $B$ is the parameter that ensures that the joint distribution is finite, and $m$ is the desired number of generated values for $x$ and $y$.
- ▷ Create $x$ and $y$ vectors of length $m$ where the first value of each is a starting point uniformly distributed over the support of $x$ and $y$, and all other vector values are filled in with unacceptable entries greater than $B$.
- ▷ Run the chain for $m = 50,000 - 1$ iterations beginning at the starting points. At each iteration, fill-in and save only sampled exponential values that are less than $B$, repeating this sampling procedure until an acceptable value is drawn to replace the unacceptable $B + 1$ in that position.
- ▷ Throw away some early part of the chain where it has not yet converged.
- ▷ Describe the marginal distributions of $x$ and $y$ with the remaining empirical values.

This leads to the following `R` code, which can be retyped verbatim, obtained from the book's webpage, or the book's `R` package `BaM` to replicate this example:

```
B <- 10; m <- 50000
gibbs.expo <- function(B,m) {
    x <- c(runif(1,0,B),rep((B+1),length=(m-1)))
    y <- c(runif(1,0,B),rep((B+1),length=(m-1)))
    for (i in 2:m)  {
        while(x[i] > B) x[i] <- rexp(1,y[i-1])
        while(y[i] > B) y[i] <- rexp(1,x[i])
    }
    return(cbind(x,y))
}

gibbs.expo(B=5, m=500)
```

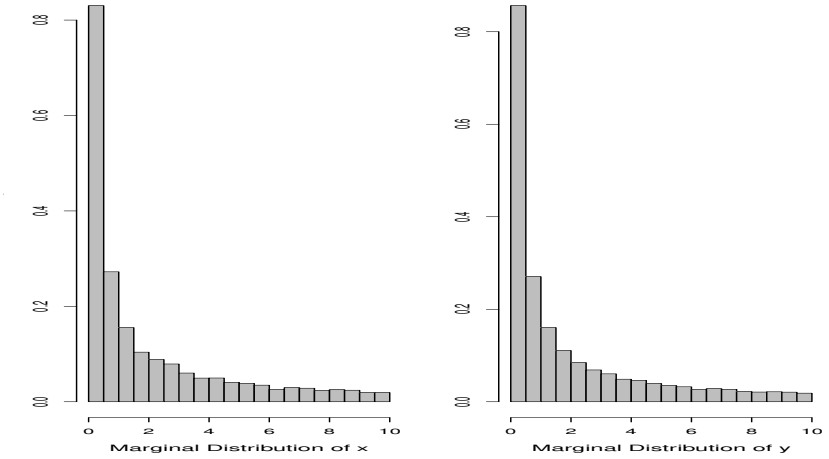


**FIGURE 1.1**: Gibbs Sampling, Marginal Exponentials

These samples are summarized by histograms of the empirical results for $x$ and $y$ in Figure 1.1, where $m = 50,000$ samples are drawn and the first $40,000$ are discarded (these are called "burn-in" values). It is clear from the figure that the marginal distributions are exponentially distributed. We can recover parameters by using the empirical draws to calculate sample statistics. This part of the MCMC process is actually quite trivial once we are convinced that there has been convergence of the Markov chain. In later chapters we will see this process in a more realistic, and therefore detailed, setting. This example is intended to give an indication of activities to come and to reinforce the linkage between Bayesian inference and modern statistical computing.

### 1.7.2 Simple Metropolis Sampling

Another Markov chain Monte Carlo tool with wide use is the Metropolis algorithm from statistical physics (Metropolis *et al.* 1953). The Metropolis algorithm is more flexible than

Gibbs sampling because it works with the joint distribution rather than a full listing of conditional distributions for the parameters in the model. As a result, many variations have been developed to satisfy particular sampling challenges posed by complicated models and ill-behaved target functions. Later chapters will cover these extensions in detail.

The essential idea behind the Metropolis algorithm is that, while we cannot easily generate values from the joint (posterior) distribution of interest, we can often find a "similar" distribution that is easy to sample from. Obviously we need to make sure that this alternative distribution is defined over the same support as the target distribution and that it does not radically favor areas of low density of this target. Once a candidate point in multivariate space has been produced by this candidate-generating distribution we will accept or reject it based upon characteristics of the target distribution. The algorithm is characterized by the following steps.

**1.** The candidate-generating distribution proposes that we move to some other point by drawing a point from *its* generating mechanism.

**2.** If this point produces a step on the target distribution that is of *higher* density, then we will always go there.

**3.** If this point produces a step on the target distribution that is of *lower* density, then we will go there probabilistically proportional to how much lower the step is in density.

Thus it is easy to see that the Markov chain "wanders around" the target density describing it as it goes and favoring higher density regions. The nice part is that the Markov chain will also explore other lower density regions as well, but with lower probability as we would want. Analogously, consider locking a house cat in large room with features that are attractive to cats (the high density regions of the posterior), and features that are unattractive to cats (the low density regions of the posterior). Anyone who has spent time with house cats can see at least some Markovian feature to their nature, as well as an innate curiosity. As our feline Markov chain wanders the room in a memory-less state, we record the coordinates of their travel. Over time we will find that the cat spends more time in the attractive areas, but still occasionally investigates the unattractive areas. If this attractiveness is proportional to density we want to describe, then the cat eventually produces description of the posterior distribution.

We can more precisely describe this algorithm. Suppose we have a two-dimensional target distribution, $p(x, y)$, which can be a posterior distribution from a Bayesian model, or any other form that is hard to marginalize, i.e., produce individual distributions $p(x)$ and $p(y)$. A single Metropolis step is produced by:

**1.** Sample $(x', y')$ from the candidate-generating distribution, $q(x', y')$.

**2.** Sample a value $u$ from $u[0:1]$.

**3.** If

$$a(x', y'|x, y) = \frac{p(x', y')}{p(x, y)} > u$$

then accept $(x', y')$ as the new destination.

**4.** Otherwise keep $(x, y)$ as the new destination.

The result is a chain of values, $[(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots]$. Note that unlike the Gibbs sampler, the Metropolis algorithm does not necessarily have to move to a new position at each iteration, and the decision to stay put is considered a Markovian step to the current position (time is consumed by the step).

There are a few technical details that we will worry about in much more detail later beginning in Chapter **??**. Often the candidate-generating distribution produces values conditional on the current position, $q(x', y'|x, y)$, but this is not strictly necessary. The basic version described here requires that the candidate-generating distribution be symmetrical in its arguments, $q(x', y'|x, y) = q(x, y|x', y')$. Also, the choice of candidate-generating distribution can be complicated by the need to match irregularities in the target distribution. Finally, it is important in real applications to run the Markov chain for some initial period to let it settle into the distribution of interest before recording values.

Consider a problem similar to that above, but where we have a joint distribution for the parameters and not the desired marginals (or conditionals as used in the Gibbs sampler). The bivariate exponential distribution for $x, y \in [0{:}\infty]$ is given by the function:

$$p(x, y) = \exp[-(\lambda_1 + \lambda)x - (\lambda_2 + \lambda)y - \lambda \max(x, y)], \tag{1.25}$$

with non-negative parameters $\lambda_1$, $\lambda_2$, and $\lambda$. This model is common in reliability analysis (Marshall and Olkin 1967), where the interpretation is that the first two parameters are the event intensities for systems 1 and 2, and the non-subscripted parameter is the shared intensity between systems. In this literature events are usually machine failures, but for our purposes they can be death, graduation, cabinet dissolutions, divorce, cessation of war, and so on. In this example we have the parameters:

$$\lambda_1 = 0.5, \quad \lambda_2 = 0.5, \quad \lambda = 0.01, \quad B = \max(x) = \max(y) = 8,$$

which produces the bivariate distribution shown in the first panel of Figure 1.2. The maximum in the function makes it a little harder to analytically calculate marginal distributions with integration, so we might want to apply MCMC to save trouble. This is exactly analogous to the process where complicated Bayesian model specifications sometimes make it difficult to describe marginal posteriors for parameters of interest.

To implement the Metropolis algorithm we need a candidate-generating distribution from which to draw potential destinations for the Markov chain. Typically researchers look for some convenient distribution from the commonly used form since software such as `R` makes drawing values trivial. Here we will exploit the stipulated bounds on the problem and note that the bivariate exponential is enclosed in a big box with length and width equal to $B = 8$ and maximum height equal to one from the form of (1.25). The process is further covered in Chapter **??**, but note here that it is easy to draw points inside this box from scaled uniforms. Nicely, we do not have to rescale the distribution of $q(x', y')$ because the values are drawn from this distribution but inserted into $p()$. It is important to note, without getting too far ahead of ourselves, that a better fitting candidate-generating distribution could be found and that drawing from uniform boxes is not particularly efficient.

To begin we define our function in R according to:

```
biv.exp <- function(x,y,L1,L2,L)
           exp( -(L1+L)*x - (L2+L)*y -L*max(x,y) )
```

So it will return density values for given $(x, y)$ pairs and specific parameters. The candidate-generating function is:

```
cand.gen <- function(max.x,max.y)
            c(runif(1,0,max.x),runif(1,0,max.y))
```

where we could have stipulated the $B$ value but left the function slightly more general. Markov chains require starting positions and we arbitrarily select $(x = 0.5, y = 0.5$ here. The algorithm is now given to be the following R code, which (again) can be retyped verbatim to replicate the example:

```
m <-5000; x<-0.5; y<-0.5; L1<-0.5; L2<-0.5; L<-0.01; B<-8
for (i in 1:m)  {
    cand.val <- cand.gen(B,B)
    a <- biv.exp(cand.val[1],cand.val[2],L1,L2,L)
            / biv.exp(x[i],y[i],L1,L2,L)
    if (a > runif(1)) {
        x <- c(x,cand.val[1])
        y <- c(y,cand.val[2])
    }
    else  {
        x <- c(x,x[i])
        y <- c(y,y[i])
    }
}
```

The resulting values are shown by the histograms in the second and third panels of Figure 1.2, where the algorithm has been run for $m = 5,000$ iterations but the first $3,000$ are discarded. We could also simply summarize the resulting marginals for $x$ and $y$ empirically with means, quantiles, or other simple statistics. The Metropolis algorithm shown here will be expanded and generalized in Chapter **??** by loosening restrictions on the candidate-generating distribution and allowing for hybrid processes that accommodate difficult features in the target distribution. The two MCMC algorithms described here form the basis for all practical work needed to estimate complex Bayesian models in the social sciences.

## 1.8 Historical Comments

Statistics is a relatively new field of scientific endeavor. In fact, for much of its history it was subsumed to various natural sciences as a combination of foster-child and household

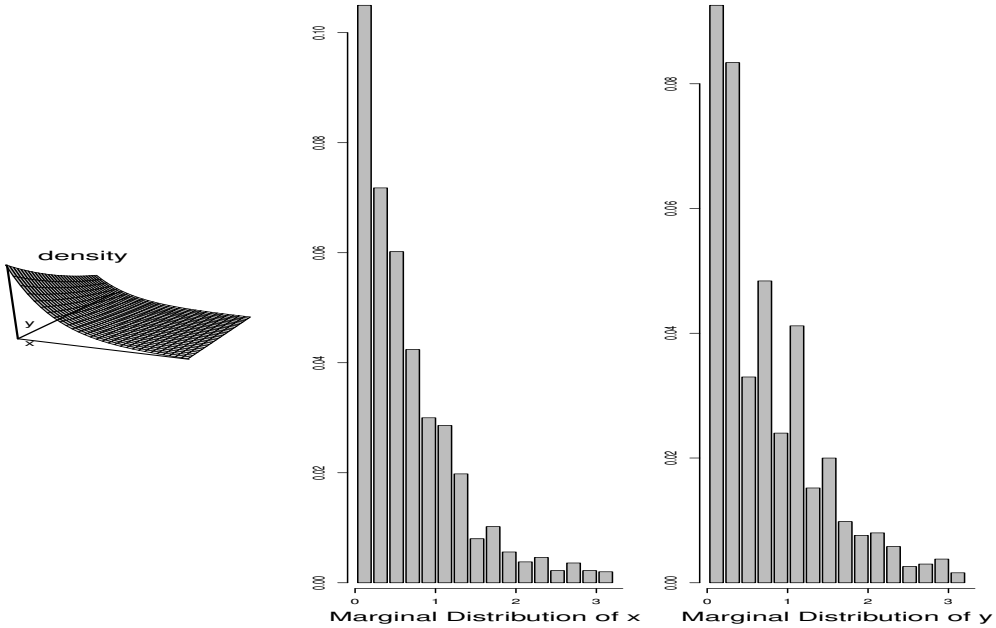


**FIGURE 1.2**: Metropolis Sampling, Bivariate Exponential

maid: unwanted by its natural parents (mathematics and philosophy), yet necessary to clean things up. Beginning with the work of Laplace (1774, 1781, 1811), Gauss (1809, 1823, 1855), Legendre (1805), and de Morgan (1837, 1838, 1847), statistics began to emerge as a discipline worthy of study on its own merits. The first renaissance occurred around the turn of the last century due to the monumental efforts of Galton (1869, 1875, 1886, 1892), Fisher (1922, 1925a, 1925b, 1934), Neyman and (Egon) Pearson (1928a, 1928b, 1933a, 1933b, 1936a, 1936b), Gossett (as Student, 1908a, 1908b), Edgeworth (1892a, 1892b, 1893a, 1893b), (Karl) Pearson (1892, 1900, 1907, 1920), and Venn (1866). Left out of the twin intellectual developments of frequentist inference from Neyman and Pearson and likelihood inference from Fisher (see Chapter **??**, Section **??** for details), was the Bayesian paradigm. Sir Thomas Bayes' famous (and only) essay was published in 1763, two years after his death (Bayes chose to perish before publishing), suggesting to some that he was ambivalent about the approach of applying a uniform prior to a binomial probability parameter. This ingenious work unintendedly precipitated a philosophy about how researcher-specified theories are fit to empirical observations. Interestingly, it was not until the early 1950s that Bayesian statistics became a self-aware branch (Fienberg 2006).

Fisher in particular was hostile to the Bayesian approach and was often highly critical, though not always with substantiated claims: Bayesianism "which like an impenetrable jungle arrests progress towards precision of statistical concepts" (1922, p.311). Fisher also worked to discredit Bayesianism and inverse probability (Bayesianism with an assumed uniform prior) by pressuring peers and even misquoting other scholars (Zabell 1989). Yet Fisher (1935) develops *fiducial inference*, which is an attempt to apply inverse probability

without uniform priors, but this approach fails; Efron (1998, p.105) calls this "Fisher's biggest blunder." In fact, Lindley (1958) later proved that fiducial inference is consistent *only* when it is made equivalent to Bayesian inference with a uniform prior. The Neyman-Pearson paradigm was equally unkind to the development of Bayesian statistics, albeit on a less vindictive level. If one is willing to subscribe to the idea of an infinite sequence of samples, then the Bayesian prior is unimportant since the data will overwhelm this prior. Although there are scenarios where this is a very reasonable supposition, generally these are far more difficult to come by in the social and behavioral sciences.

Although Bayesianism had suffered "a nearly lethal blow" from Fisher and Neyman by the 1930s (Zabell 1989), it was far from dead. Scholars such as Jeffreys (1961), Good (1950), Savage (1954, 1962), de Finetti (1972, 1974, 1975), and Lindley (1961, 1965) reactivated interest in Bayesian methods in the middle of the last century in response to observed deficiencies in classical techniques. Lindley and Novick (1978, 1981) published important applied work in education psychology that carefully studied exchangeability and utility from a Bayesian perspective, and Novick *et al.* (1976) developed an early Bayesian software program for estimating simple models: CADA, Computer Assisted Data Analysis. Unfortunately many of the specifications developed by these modern Bayesians, while superior in theoretical foundation, led to mathematical forms that were intractable.[3] Fortunately, this problem has been largely resolved in recent years by a revolution in statistical computing techniques, and this has led to a second renaissance for the Bayesian paradigm (Berger 2001).

Markov chain Monte Carlo (MCMC) techniques solve a lingering problem in Bayesian analysis, and thus earn a special place in this work. Often Bayesian model specifications considered either interesting or realistic produced inference problems that were analytically intractable because they led to high-dimension integral calculations that were impossible to solve analytically. Previous numerical techniques for performing these integrations were often difficult and highly specialized tasks (e.g., Shaw 1988, Stewart and Davis 1986, van Dijk and Kloek 1982, Tierney and Kadane 1986). Beginning with the foundational work of Metropolis *et al.* (1953), Hastings (1970), Peskun (1973), Geman and Geman (1984), and the critical synthesizing essay of Gelfand and Smith (1990), there is now a voluminous literature on Markov chain Monte Carlo. In fact, modern Bayesian statistical practice is intimately and intrinsically tied to stochastic simulation techniques and as a result, these tools are an integral part of this book. We introduce these tools in this chapter in Section 1.7 and in much greater detail in Chapter **??**.

Currently the most popular method for generating samples from posterior distributions using Markov chains is the `WinBUGS` program and its Unix-based precursor `BUGS` and the more recent functional equivalent `JAGS`. The name `BUGS` is a pseudo-acronym for *Bayesian inference Using Gibbs Sampling*, referring to the most frequently used method for producing Markov chains. In what constitutes a notable and noble contribution to the Bayesian

[3] This led one observer (Evans 1994) to compare Bayesians to "unmarried marriage guidance counselors."

statistical world, the Community Statistical Research Project at the MRC Biostatistics Unit and the Imperial College School of Medicine at St. Mary's, London provide this high-quality software to users free of charge, and it can be downloaded from their web page: `http://www.mrc-bsu.cam.ac.uk/software/bugs/`. These authors have even made available extensive documentation at the same site by Spiegelhalter *et al.* (1996a, 1996b, 2000, 2012). Alternative ways to use `WinBUGS` with `R` as the interface are: `BRugs`, `rbugs`, and `R2WinBUGS`. There are also facilities for calling `WinBUGS` from `SAS`, `stata`, and `excel`. The `JAGS` program (Just Another Gibbs Sampler) is an engine for the `BUGS` language that has nearly the same structure as `WinBUGS`, with only a few syntactical differences. Authored by Martyn Plummer, it is extremely well-developed software that runs on non-windows platforms and is command-line driven rather than point-and-click. It can be downloaded at `http://www-ice.iarc.fr/~martyn/software/jags/`. There are also facilities for calling `JAGS` from `R`: `R2jags`, `Rjags`, and `runjags`. Most of the `BUGS` code in this text are run with `JAGS` from the command window. Other high-quality `R` packages using or providing MCMC computing include: `BMS`, `dclone`, `eco`, `glmdm`, `HI`, `lmm`, `MasterBayes`, `mcmc`, `MCMCglmm`, `MCMCpack`, `MNP`, `pscl`, `spBayes`, `tgp`, and `zic`. Of these `MCMCpack` is the most general, whereas most of the others are MCMC implementations to solve a specific problem. Given the rapid pace of `R` package development, this list is growing rapidly.

---

## 1.9 Exercises

1.1 Restate the three general steps of Bayesian inference from page 6 in your own words.

1.2 Given $k$ possible disjoint (non-overlapping) events labeled: $E_1, \ldots, E_k$ where $k$ could even be infinity, denote $p(E_i)$ as the mapping from events $E_i$ to [0:1] space. Write the Kolmogorov axioms of probability in technical detail.

1.3 Rewrite Bayes' Law when the two events are independent. How do you interpret this?

1.4 Equation (1.11) on page 12 showed that $p(A|\mathbf{D}) = p(\mathbf{D}|A)p(A)/(p(\mathbf{D}|A)p(A) + p(\mathbf{D}|B)p(B) + p(\mathbf{D}|C)p(C))$. Rewrite this expression for $p(A|\mathbf{D})$ when there are arbitrary $k \in \mathcal{I}^+$ events including $A$.

1.5 Suppose $f(\theta|\mathbf{X})$ is the posterior distribution of $\theta$ given the data $\mathbf{X}$. Describe the shape of this distribution when the mode, $\underset{\theta}{\text{argmax}} f(\theta|\mathbf{X})$, is equal to the mean, $\int_\theta \theta f(\theta|\mathbf{X}) d\theta$.

1.6 The Rényi countable additivity axiom is defined by: (1) for any events $E_1$ and $E_2$, $p(E_1|E_2) \geq 0$ (and reversed), $p(E_i|E_i) = 1$, (2) for disjoint sets $E_1, \ldots$ and

another arbitrary event $D$, $p(\cup_{i=1}^{\infty} E_i|D) = \sum_{i=1}^{\infty} p(E_i|D)$, and (3) for every subset of events, $E_i, E_j, E_k$, with $E_j \subseteq E_k$ and $p(E_j|E_k) > 0$, we get $p(E_i|E_j) = p(E_i, E_j|E_k)/p(E_j|E_k)$. Show that the Kolmogorov axioms are a special case.

1.7 Using `R` run the Gibbs sampling function given on page 26. What effect do you see in varying the $B$ parameter? What is the effect of producing 200 sampled values instead of 50,000?

1.8 Some authors have objected to the uniform prior, $p(\theta) = 1, \theta \in [0{:}1]$ to describe unknown probabilities in a binomial model and suggested instead the Haldane prior: $p(\theta) \propto [\theta^{-1}(1-\theta)]^{-1}$ (Haldane [1938], Novick and Hall [1965], Villegas [1977], ). Plot this prior and the uniform prior over [0:1] in the same graph.

1.9 Rerun the Metropolis algorithm on page 30 in `R` but replacing the uniform generation of candidate values in `cand.gen` with a normal truncated to fit in the appropriate range. What differences do you observe?

1.10 The Gibbs sampler described in Section 1.7.1 from Casella and George (1992) was originally done as follows: (1) set initial values for $B = 5$, $k = 15$, $m = 5,000$, and the set of accepted values $(x, y)$ as an empty object, (2) run $m$ chains of length $k + 1$ where the first value is the uniformly distributed starting point $[0 : B]$) and the rest are sampled conditional exponential values that are less than $B$, (3) save only the last value from the $x$ and $y$ series, $x_{16}$ and $y_{16}$ to the stored Markov chain until 5,000 of each are obtained. Implement this alternative algorithm in `R` and compare it to the output shown in Figure 1.1 on page 27.

1.11 If $p(\mathbf{D}|\theta) = 0.5$, and $p(\mathbf{D}) = 1$, calculate the value of $p(\theta|\mathbf{D})$ for priors $p(\theta)$, $[0.001, 0.01, 0.1, 0.9]$.

1.12 Buck, Cavanaugh, and Litton (1996) demonstrate the use of Bayesian statistics for radiocarbon dating of Early Bronze Age archaeological samples (seeds and bones) from St. Veit-Klinglberg, Austria. These ten age data points are produced by the Oxford accelerator dating facilities:

| Context # | $\mu_i$ | $\sigma_i$ |
|---|---|---|
| 758 | 3275 | 75 |
| 814 | 3270 | 80 |
| 1235 | 3400 | 75 |
| 493 | 3190 | 75 |
| 925 | 3420 | 65 |
| 923 | 3435 | 60 |
| 1168 | 3160 | 70 |
| 358 | 3340 | 80 |
| 813 | 3270 | 75 |
| 1210 | 3200 | 70 |

Given the model $X_i|\mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, (1) calculate the probability that sample 358 originates between 3300 to 3400 years ago, (2) generate 10,000 samples from the distribution for sample 493 and sample 923 and plot a histogram of these in the same figure (side-by-side), (3) give the proportion of values that overlap, and (4) how do you interpret this overlap probabilistically with regard to the age of the samples?

1.13 Sometimes Bayesian results are given as *posterior odds ratios*, which for two possible alternative hypotheses is expressed as:

$$\text{odds}(\theta_1, \theta_2) = \frac{p(\theta_1|\mathbf{D})}{p(\theta_2|\mathbf{D})}.$$

If the prior probabilities for $\theta_1$ and $\theta_2$ are identical, how can this be re-expressed using Bayes' Law?

1.14 Using the posterior distribution in (1.19) on page 19, produce the posterior mean for $\theta$ in (1.20) and the posterior variance for $\theta$ in (1.21).

1.15 Suppose we had data, $\mathbf{D}$, distributed $p(\mathbf{D}|\theta) = \theta e^{-\theta \mathbf{D}}$ as in Section 1.5.1 starting on page 18, but now $p(\theta) = 1/\theta$, for $\theta \in (0{:}\infty)$. Calculate the posterior mean.

1.16 Modify the Gibbs sampler in Section 1.7.1 starting on page 25 to sample from two mutually conditional gamma distributions instead of exponential distributions. The exponential distribution is a simplified form of the rate parameter gamma distribution where the first (shape) parameter is 1 (Appendix B, page **??**). Set the two relevant shape parameters to values of your choosing $\alpha > 1$. Produce a graphs of the marginal draws.

1.17 Since the posterior distribution is a compromise between prior information and the information provided by the new data, then it is interesting to compare relative strengths. Perform an experiment where you flip a coin 10 times, recording the data as zeros and ones. Produce the posterior expected value (mean) for two priors on $p$ (the probability of a heads): a uniform distribution between zero and one, and