

Statistics for Social and Behavioral Sciences

Advisors:

S.E. Fienberg W.J. van der Linden

Statistics for Social and Behavioral Sciences

Brennan: Generalizability Theory

DeBoeck/Wilson: Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach

Devlin/Fienberg/Resnick/Roeder: Intelligence, Genes, and Success: Scientists Respond to The Bell Curve

Dorans/Pommerich/Holland: Linking and Aligning Scores and Scales

Finkelstein/Levin: Statistics for Lawyers, 2nd ed.

Gastwirth: Statistical Science in the Courtroom

Handcock/Morris: Relative Distribution Methods in the Social Sciences

Johnson/Albert: Ordinal Data Modeling

Kolen/Brennan: Test Equating, Scaling, and Linking: Methods and Practices, 2nd ed.

Longford: Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician

Lynch: Introduction to Applied Bayesian Statistics and Estimation for Social Scientists

Morton/Rolph: Public Policy and Statistics: Case Studies from RAND

van der Linden: Linear Models for Optimal Test Design

von Davier/Carstensen: Multivariate and Mixture Distribution Rasch Models

von Davier/Holland/Thayer: The Kernel Method of Test Equating

Zeisel/Kaye: *Prove It with Figures*: Empirical Methods in Law and Litigation

Scott M. Lynch

Introduction to Applied Bayesian Statistics and Estimation for Social Scientists

With 89 Figures

 Springer

Scott M. Lynch
Department of Sociology and
Office of Population Research
Princeton University
Princeton, NJ 08544
slynch@princeton.edu

Series Editors:

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213–3890
USA

Wim J. van der Linden
Department of Measurement
and Data Analysis
Faculty of Behavioral Sciences
University of Twente
7500 AE Enschede
The Netherlands

Library of Congress Control Number: 2007929729

ISBN 978-0-387-71264-2

e-ISBN 978-0-387-71265-9

SAS[®] and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.

STATA[®] and STATA[®] logo are registered trademarks of StataCorp LP.

Printed on acid-free paper.

© 2007 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

For my Barbara

Preface

This book was written slowly over the course of the last five years. During that time, a number of advances have been made in Bayesian statistics and Markov chain Monte Carlo (MCMC) methods, but, in my opinion, the market still lacks a *truly* introductory book written explicitly for social scientists that thoroughly describes the actual process of Bayesian analysis using these methods. To be sure, a variety of introductory books are available that cover the basics of the Bayesian approach to statistics (e.g., Gill 2002 and Gelman et al. 1995) and several that cover the foundation of MCMC methods (e.g., beginning with Gilks et al. 1996). Yet, a highly applied book showing how to use MCMC methods to complete a Bayesian analysis involving typical social science models applied to typical social science data is still sorely lacking. The goal of this book is to fill this niche.

The Bayesian approach to statistics has a long history in the discipline of statistics, but prior to the 1990s, it held a marginal, almost cult-like status in the discipline and was almost unheard of in social science methodology. The primary reasons for the marginal status of the Bayesian approach include (1) philosophical opposition to the use of “prior distributions” in particular and the subjective approach to probability in general, and (2) the lack of computing power for completing realistic Bayesian analyses. In the 1990s, several events occurred simultaneously to overcome these concerns. First, the explosion in computing power nullified the second limitation of conducting Bayesian analyses, especially with the development of sampling based methods (e.g., MCMC methods) for estimating parameters of Bayesian models. Second, the growth in availability of longitudinal (panel) data and the rise in the use of hierarchical modeling made the Bayesian approach more appealing, because Bayesian statistics offers a natural approach to constructing hierarchical models. Third, there has been a growing recognition both that the enterprise of statistics is a subjective process in general and that the use of prior distributions need not influence results substantially. Additionally, in many problems, the use of a prior distribution turns out to be advantageous.

The publication of Gelfand and Smith's 1990 paper describing the use of MCMC simulation methods for summarizing Bayesian posterior distributions was the watershed event that launched MCMC methods into popularity in statistics. Following relatively closely on the heels of this article, Gelman et al.'s (1995) book, *Bayesian Data Analysis*, and Gilks et al.'s (1996) book, *Markov Chain Monte Carlo in Practice*, placed the Bayesian approach in general, and the application of MCMC methods to Bayesian statistical models, squarely in the mainstream of statistics. I consider these books to be classics in the field and rely heavily on them throughout this book.

Since the mid-1990s, there has been an explosion in advances in Bayesian statistics and especially MCMC methodology. Many improvements in the recent past have been in terms of (1) monitoring and improving the performance of MCMC algorithms and (2) the development of more refined and complex Bayesian models and MCMC algorithms tailored to specific problems. These advances have largely escaped mainstream social science.

In my view, these advances have gone largely unnoticed in social science, because purported introductory books on Bayesian statistics and MCMC methods are not truly introductory for this audience. First, the mathematics in introductory books is often too advanced for a mainstream social science audience, which begs the question: "introductory *for whom?*" Many social scientists do not have the probability theory and mathematical statistics background to follow many of these books beyond the first chapter. This is not to say that the material is impossible to follow, only that more detail may be needed to make the text and examples more readable for a mainstream social science audience.

Second, many examples in introductory-level Bayesian books are at best foreign and at worst irrelevant to social scientists. The probability distributions that are used in many examples are not typical probability distributions used by social scientists (e.g., Cauchy), and the data sets that are used in examples are often atypical of social science data. Specifically, many books use small data sets with a limited number of covariates, and many of the models are not typical of the regression-based approaches used in social science research. This fact may not seem problematic until, for example, one is faced with a research question requiring a multivariate regression model for 10,000 observations measured on 5 outcomes with 10 or more covariates. Nonetheless, research questions involving large-scale data sets are not uncommon in social science research, and methods shown that handle a sample of size 100 measured on one or two outcomes with a couple of covariates simply may not be directly transferrable to a larger data set context. In such cases, the analyst without a solid understanding of the linkage between the model and the estimation routine may be unable to complete the analysis. Thus, some discussion tailored to the practicalities of *real* social science data and computing is warranted.

Third, there seems to be a disjunction between introductory books on Bayesian theory and introductory books on applied Bayesian statistics. One

of the greatest frustrations for me, while I was learning the basics of Bayesian statistics and MCMC estimation methods, was (and is) the lack of a book that links the theoretical aspects of Bayesian statistics and model development with the application of modern estimation methods. Some examples in extant books may be substantively interesting, but they are often incomplete in the sense that discussion is truncated after model development without adequate guidance regarding how to estimate parameters. Often, suggestions are made concerning how to go about implementing only certain aspects of an estimation routine, but for a person with no experience doing this, these suggestions are not enough.

In an attempt to remedy these issues, this book takes a step back from the most recent advances in Bayesian statistics and MCMC methods and tries to bridge the gap between Bayesian theory and modern Bayesian estimation methods, as well as to bridge the gap between Bayesian statistics books written as “introductory” texts for statisticians and the needs of a mainstream social science audience. To accomplish this goal, this book presents very little that is new. Indeed, most of the material in this book is now “old-hat” in statistics, and many references are a decade old (In fact, a second edition of Gelman et al.’s 1995 book is now available). However, the trade-off for not presenting much new material is that this book explains the process of Bayesian statistics and modern parameter estimation via MCMC simulation methods in great depth. Throughout the book, I painstakingly show the modeling process from model development, through development of an MCMC algorithm to estimate its parameters, through model evaluation, and through summarization and inference.

Although many introductory books begin with the assumption that the reader has a solid grasp of probability theory and mathematical statistics, I do not make that assumption. Instead, this book begins with an exposition of the probability theory needed to gain a solid understanding of the statistical analysis of data. In the early chapters, I use contrived examples applied to (sometimes) contrived data so that the forest is not lost for the trees: The goal is to provide an understanding of the issue at hand rather than to get lost in the idiosyncratic features of real data. In the latter chapters, I show a Bayesian approach (or approaches) to estimating some of the most common models in social science research, including the linear regression model, generalized linear models (specifically, dichotomous and ordinal probit models), hierarchical models, and multivariate models.

A consequence of this choice of models is that the parameter estimates obtained via the Bayesian approach are often very consistent with those that could be obtained via a classical approach. This may make a reader ask, “then what’s the point?” First, there are many cases in which a Bayesian approach and a classical approach will not coincide, but from my perspective, an introductory text should establish a foundation that can be built upon, rather than beginning in unfamiliar territory. Second, there are additional benefits to taking a Bayesian approach beyond the simple estimation of model

parameters. Specifically, the Bayesian approach allows for greater flexibility in evaluating model fit, comparing models, producing samples of parameters that are not directly estimated within a model, handling missing data, “tweaking” a model in ways that cannot be done using canned routines in existing software (e.g., freeing or imposing constraints), and making predictions/forecasts that capture greater uncertainty than classical methods. I discuss each of these benefits in the examples throughout the latter chapters.

Throughout the book I thoroughly flesh out each example, beginning with the development of the model and continuing through to developing an MCMC algorithm (generally in R) to estimate it, estimating it using the algorithm, and presenting and summarizing the results. These programs should be straightforward, albeit perhaps tedious, to replicate, but some programming is inherently required to conduct Bayesian analyses. However, once such programming skills are learned, they are incredibly freeing to the researcher and thus well worth the investment to acquire them. Ultimately, the point is that the examples are thoroughly detailed; nothing is left to the imagination or to guesswork, including the mathematical contortions of simplifying posterior distributions to make them recognizable as known distributions.

A key feature of Bayesian statistics, and a point of contention for opponents, is the use of a prior distribution. Indeed, one of the most complex things about Bayesian statistics is the development of a model that includes a prior and yields a “proper” posterior distribution. In this book, I do not concentrate much effort on developing priors. Often, I use uniform priors on most parameters in a model, or I use “reference” priors. Both types of priors generally have the effect of producing results roughly comparable with those obtained via maximum likelihood estimation (although not in interpretation!). My goal is not to minimize the importance of choosing appropriate priors, but instead it is not to overcomplicate an introductory exposition of Bayesian statistics and model estimation. The fact is that most Bayesian analyses explicitly attempt to minimize the effect of the prior. Most published applications to date have involved using uniform, reference, or otherwise “noninformative” priors in an effort to avoid the “subjectivity” criticism that historically has been levied against Bayesians by classical statisticians. Thus, in most Bayesian social science research, the prior has faded in its importance in differentiating the classical and Bayesian paradigms. This is not to say that prior distributions are unimportant—for some problems they may be very important or useful—but it is to say that it is not necessary to dwell on them.

The book consists of a total of 11 chapters plus two appendices covering (1) calculus and matrix algebra and (2) the basic concepts of the Central Limit Theorem. The book is suited for a highly applied one-semester graduate level social science course. Each chapter, including the appendix but excluding the introduction, contains a handful of exercises at the end that test the understanding of the material in the chapter at both theoretical and applied levels. In the exercises, I have traded quantity for quality: There are relatively few exercises, but each one was chosen to address the essential material in

the chapter. The first half of the book (Chapters 1-6) is primarily theoretical and provides a generic introduction to the theory and methods of Bayesian statistics. These methods are then applied to common social science models and data in the latter half of the book (Chapters 7-11). Chapters 2-4 can each be covered in a week of classes, and much of this material, especially in Chapters 2 and 3, should be review material for most students. Chapters 5 and 6 will most likely each require more than a week to cover, as they form the nuts and bolts of MCMC methods and evaluation. Subsequent chapters should each take 1-2 weeks of class time. The models themselves should be familiar, but the estimation of them via MCMC methods will not be and may be difficult for students without some programming and applied data analysis experience. The programming language used throughout the book is R, a freely available and common package used in applied statistics, but I introduce the program WinBugs in the chapter on hierarchical modeling. Overall, R and WinBugs are syntactically similar, and so the introduction of WinBugs is not problematic. From my perspective, the main benefit of WinBugs is that some derivations of conditional distributions that would need to be done in order to write an R program are handled automatically by WinBugs. This feature is especially useful in hierarchical models. All programs used in this book, as well as most data, and hints and/or solutions to the exercises can be found on my Princeton University website at: www.princeton.edu/~slynch.

Acknowledgements

I have a number of people to thank for their help during the writing of this book. First, I want to thank German Rodriguez and Bruce Western (both at Princeton) for sharing their advice, guidance, and statistical knowledge with me as I worked through several sections of the book. Second, I thank my friend and colleague J. Scott Brown for reading through virtually all chapters and providing much-needed feedback over the course of the last several years. Along these same lines, I thank Chris Wildeman and Steven Shafer for reading through a number of chapters and suggesting ways to improve examples and the general presentation of material. Third, I thank my statistics thesis advisor, Valen Johnson, and my mentor and friend, Ken Bollen, for all that they have taught me about statistics. (They cannot be held responsible for the fact that I may not have learned well, however). For their constant help and tolerance, I thank Wayne Appleton and Bob Jackson, the senior computer folks at Princeton University and Duke University, without whose support this book could not have been possible. For their general support and friendship over a period including, but not limited to, the writing of this book, I thank Linda George, Angie O’Rand, Phil Morgan, Tom Espenshade, Debby Gold, Mark Hayward, Eileen Crimmins, Ken Land, Dan Beirute, Tom Rice, and John Moore. I also thank my son, Tyler, and my wife, Barbara, for listening to me ramble incessantly about statistics and acting as a sounding

board during the writing of the book. Certainly not least, I thank Bill McCabe for helping to identify an egregious error on page 364. Finally, I want to thank my editor at Springer, John Kimmel, for his patience and advice, and I acknowledge support from NICHD grant R03HD050374-01 for much of the work in Chapter 10 on multivariate models.

Despite having all of these sources of guidance and support, all the errors in the book remain my own.

Princeton University

Scott M. Lynch
April 2007

Contents

Preface	vii
Contents	xiii
List of Figures	xix
List of Tables	xxvii
1 Introduction	1
1.1 Outline	3
1.2 A note on programming	5
1.3 Symbols used throughout the book	6
2 Probability Theory and Classical Statistics	9
2.1 Rules of probability	9
2.2 Probability distributions in general	12
2.2.1 Important quantities in distributions	17
2.2.2 Multivariate distributions	19
2.2.3 Marginal and conditional distributions	23
2.3 Some important distributions in social science	25
2.3.1 The binomial distribution	25
2.3.2 The multinomial distribution	27
2.3.3 The Poisson distribution	28
2.3.4 The normal distribution	29
2.3.5 The multivariate normal distribution	30
2.3.6 t and multivariate t distributions	33
2.4 Classical statistics in social science	33
2.5 Maximum likelihood estimation	35
2.5.1 Constructing a likelihood function	36
2.5.2 Maximizing a likelihood function	38
2.5.3 Obtaining standard errors	39

2.5.4	A normal likelihood example	41
2.6	Conclusions	44
2.7	Exercises	44
2.7.1	Probability exercises	44
2.7.2	Classical inference exercises	45
3	Basics of Bayesian Statistics	47
3.1	Bayes' Theorem for point probabilities	47
3.2	Bayes' Theorem applied to probability distributions	50
3.2.1	Proportionality	51
3.3	Bayes' Theorem with distributions: A voting example	53
3.3.1	Specification of a prior: The beta distribution	54
3.3.2	An alternative model for the polling data: A gamma prior/ Poisson likelihood approach	60
3.4	A normal prior–normal likelihood example with σ^2 known	62
3.4.1	Extending the normal distribution example	65
3.5	Some useful prior distributions	68
3.5.1	The Dirichlet distribution	69
3.5.2	The inverse gamma distribution	69
3.5.3	Wishart and inverse Wishart distributions	70
3.6	Criticism against Bayesian statistics	70
3.7	Conclusions	73
3.8	Exercises	74
4	Modern Model Estimation Part 1: Gibbs Sampling	77
4.1	What Bayesians want and why	77
4.2	The logic of sampling from posterior densities	78
4.3	Two basic sampling methods	80
4.3.1	The inversion method of sampling	81
4.3.2	The rejection method of sampling	84
4.4	Introduction to MCMC sampling	88
4.4.1	Generic Gibbs sampling	88
4.4.2	Gibbs sampling example using the inversion method	89
4.4.3	Example repeated using rejection sampling	93
4.4.4	Gibbs sampling from a real bivariate density	96
4.4.5	Reversing the process: Sampling the parameters <i>given</i> the data	100
4.5	Conclusions	103
4.6	Exercises	105
5	Modern Model Estimation Part 2: Metroplis–Hastings Sampling	107
5.1	A generic MH algorithm	108
5.1.1	Relationship between Gibbs and MH sampling	113

- 5.2 Example: MH sampling when conditional densities are difficult to derive 115
- 5.3 Example: MH sampling for a conditional density with an unknown form 118
- 5.4 Extending the bivariate normal example: The full multiparameter model. 121
 - 5.4.1 The conditionals for μ_x and μ_y 122
 - 5.4.2 The conditionals for σ_x^2 , σ_y^2 , and ρ 123
 - 5.4.3 The complete MH algorithm 124
 - 5.4.4 A matrix approach to the bivariate normal distribution problem 126
- 5.5 Conclusions 128
- 5.6 Exercises 129

- 6 Evaluating Markov Chain Monte Carlo Algorithms and Model Fit 131**
 - 6.1 Why evaluate MCMC algorithm performance? 132
 - 6.2 Some common problems and solutions. 132
 - 6.3 Recognizing poor performance 135
 - 6.3.1 Trace plots 135
 - 6.3.2 Acceptance rates of MH algorithms 141
 - 6.3.3 Autocorrelation of parameters 146
 - 6.3.4 “ \hat{R} ” and other calculations 147
 - 6.4 Evaluating model fit 153
 - 6.4.1 Residual analysis 154
 - 6.4.2 Posterior predictive distributions 155
 - 6.5 Formal comparison and combining models 159
 - 6.5.1 Bayes factors 159
 - 6.5.2 Bayesian model averaging 161
 - 6.6 Conclusions 163
 - 6.7 Exercises 163

- 7 The Linear Regression Model 165**
 - 7.1 Development of the linear regression model 165
 - 7.2 Sampling from the posterior distribution for the model parameters 168
 - 7.2.1 Sampling with an MH algorithm 168
 - 7.2.2 Sampling the model parameters using Gibbs sampling . . 169
 - 7.3 Example: Are people in the South “nicer” than others? 174
 - 7.3.1 Results and comparison of the algorithms 175
 - 7.3.2 Model evaluation 178
 - 7.4 Incorporating missing data 182
 - 7.4.1 Types of missingness 182
 - 7.4.2 A generic Bayesian approach when data are MAR: The “niceness” example revisited 186

7.5	Conclusions	191
7.6	Exercises	192
8	Generalized Linear Models	193
8.1	The dichotomous probit model	195
8.1.1	Model development and parameter interpretation	195
8.1.2	Sampling from the posterior distribution for the model parameters	198
8.1.3	Simulating from truncated normal distributions	200
8.1.4	Dichotomous probit model example: Black–white differences in mortality	206
8.2	The ordinal probit model	217
8.2.1	Model development and parameter interpretation	218
8.2.2	Sampling from the posterior distribution for the parameters	220
8.2.3	Ordinal probit model example: Black–white differences in health	223
8.3	Conclusions	228
8.4	Exercises	229
9	Introduction to Hierarchical Models	231
9.1	Hierarchical models in general	232
9.1.1	The voting example redux	233
9.2	Hierarchical linear regression models	240
9.2.1	Random effects: The random intercept model	241
9.2.2	Random effects: The random coefficient model	251
9.2.3	Growth models	256
9.3	A note on fixed versus random effects models and other terminology	264
9.4	Conclusions	268
9.5	Exercises	269
10	Introduction to Multivariate Regression Models	271
10.1	Multivariate linear regression	271
10.1.1	Model development	271
10.1.2	Implementing the algorithm	275
10.2	Multivariate probit models	277
10.2.1	Model development	278
10.2.2	Step 2: Simulating draws from truncated multivariate normal distributions	283
10.2.3	Step 3: Simulation of thresholds in the multivariate probit model	289
10.2.4	Step 5: Simulating the error covariance matrix	295
10.2.5	Implementing the algorithm	297
10.3	A multivariate probit model for generating distributions	303

10.3.1	Model specification and simulation	307
10.3.2	Life table generation and other posterior inferences	310
10.4	Conclusions	315
10.5	Exercises	317
11	Conclusion	319
A	Background Mathematics	323
A.1	Summary of calculus	323
A.1.1	Limits	323
A.1.2	Differential calculus	324
A.1.3	Integral calculus	326
A.1.4	Finding a general rule for a derivative	329
A.2	Summary of matrix algebra	330
A.2.1	Matrix notation	330
A.2.2	Matrix operations	331
A.3	Exercises	335
A.3.1	Calculus exercises	335
A.3.2	Matrix algebra exercises	335
B	The Central Limit Theorem, Confidence Intervals, and Hypothesis Tests	337
B.1	A simulation study	337
B.2	Classical inference	338
B.2.1	Hypothesis testing	339
B.2.2	Confidence intervals	342
B.2.3	Some final notes	344
	References	345
	Index	353

List of Figures

2.1	Sample Venn diagram: Outer box is sample space; and circles are events A and B	11
2.2	Two uniform distributions.	15
2.3	Histogram of the importance of being able to express unpopular views in a free society (1 = Not very important...6 = One of the most important things).	16
2.4	Sample probability density function: A linear density.	17
2.5	Sample probability density function: A bivariate plane density.	20
2.6	Three-dimensional bar chart for GSS data with “best” planar density superimposed.	22
2.7	Representation of bivariate cumulative distribution function: Area under bivariate plane density from 0 to 1 in both dimensions.	23
2.8	Some binomial distributions (with parameter $n = 10$).	27
2.9	Some Poisson distributions.	29
2.10	Some normal distributions.	31
2.11	Two bivariate normal distributions.	32
2.12	The $t(0, 1, 1)$, $t(0, 1, 10)$, and $t(0, 1, 120)$ distributions (with an $N(0, 1)$ distribution superimposed).	34
2.13	Binomial (top) and Bernoulli (bottom) likelihood functions for the OH presidential poll data.	37
2.14	Finding the MLE: Likelihood and log-likelihood functions for the OH presidential poll data.	39
3.1	Three beta distributions with mean $\alpha/(\alpha + \beta) = .5$	56
3.2	Prior, likelihood, and posterior for polling data example: The likelihood function has been normalized as a density for the parameter K	58

3.3 Posterior for polling data example: A vertical line at $K = .5$ is included to show the area needed to be computed to estimate the probability that Kerry would win Ohio. 59

3.4 Some examples of the gamma distribution. 61

4.1 Convergence of sample means on the true beta distribution mean across samples sizes: Vertical line shows sample size of 5,000; dashed horizontal lines show approximate confidence band of sample estimates for samples of size $n = 5,000$; and solid horizontal line shows the true mean. 81

4.2 Example of the inversion method: Left-hand figures show the sequence of draws from the $U(0, 1)$ density (upper left) and the sequence of draws from the density $f(x) = (1/40)(2x + 3)$ density (lower left); and the right-hand figures show these draws in histogram format, with true density functions superimposed. 83

4.3 The three-step process of rejection sampling. 86

4.4 Sample of 1,000 draws from density using rejection sampling with theoretical density superimposed. 87

4.5 Results of Gibbs sampler using the inversion method for sampling from conditional densities. 92

4.6 Results of Gibbs sampler using the inversion method for sampling from conditional densities: Two-dimensional view after 5, 25, 100, and 2,000 iterations. 93

4.7 Results of Gibbs sampler using rejection sampling to sample from conditional densities. 97

4.8 Results of Gibbs sampler using rejection sampling to sample from conditional densities: Two-dimensional view after 5, 25, 100, and 2,000 iterations. 98

4.9 Results of Gibbs sampler for standard bivariate normal distribution with correlation $r = .5$: Two-dimensional view after 10, 50, 200, and 2,000 iterations. 100

4.10 Results of Gibbs sampler for standard bivariate normal distribution: Upper left and right graphs show marginal distributions for x and y (last 1,500 iterations); lower left graph shows contour plot of true density; and lower right graph shows contour plot of true density with Gibbs samples superimposed. 101

4.11 Samples from posterior densities for a mean and variance parameter for NHIS years of schooling data under two Gibbs sampling approaches: The solid lines are the results for the marginal-for- σ^2 -but conditional-for- μ approach; and the dashed lines are the results for the full conditionals approach. . . 104

5.1	Example of symmetric proposals centered over the previous and candidate values of the parameters.	110
5.2	Example of asymmetric proposals centered at the mode over the previous and candidate values of the parameters.	111
5.3	Example of asymmetry in proposals due to a boundary constraint on the parameter space.	113
5.4	Trace plot and histogram of m parameter from linear density model for the 2000 GSS free speech data.	117
5.5	Trace plot and histogram of ρ parameter from bivariate normal density model for the 2000 GSS free speech and political participation data.	120
5.6	Trace plot and histogram of ρ from bivariate normal model for the 2000 GSS free speech and political participation data.	126
6.1	Trace plot for first 1,000 iterations of an MH algorithm sampling parameters from planar density for GSS free speech and political participation data.	136
6.2	Trace plot for first 4,000 iterations of an MH algorithm sampling parameters from planar density for GSS free speech and political participation data.	137
6.3	Trace plot for all 50,000 iterations of an MH algorithm sampling parameters from planar density for GSS free speech and political participation data with means superimposed.	138
6.4	Trace plot of cumulative and batch means for MH algorithm sampling parameters from planar density for GSS free speech and political participation data.	139
6.5	Trace plot of cumulative standard deviation of m_2 from MH algorithm sampling parameters from planar density for GSS free speech and political participation data.	140
6.6	Posterior density for m_2 parameter from planar density MH algorithm and two proposal densities: $U(-.0003, .0003)$ and $U(-.003, .003)$	142
6.7	Trace plot of the first 1,000 iterations of an MH algorithm for the planar density with a (relatively) broad $U(-.003, .003)$ proposal density.	143
6.8	Posterior density for m_2 parameter from planar density MH algorithm and two proposal densities: $U(-.003, .003)$ and $N(0, .00085)$	144
6.9	Two-dimensional trace plot for initial run of MH algorithm sampling parameters from planar density for GSS free speech and political participation data: Two possible bivariate normal proposal densities are superimposed.	145

6.10	Trace plot for 5000 iterations of an MH algorithm sampling parameters from planar density for GSS free speech and political participation data: Bivariate normal proposal density with correlation -0.9	146
6.11	Autocorrelation plots of parameter m_2 from MH algorithms for the planar density: Upper figure is for the MH algorithm with independent uniform proposals for m_1 and m_2 ; and lower figure is for the MH algorithm with a bivariate normal proposal with correlation -0.9	148
6.12	Histogram of marginal posterior density for m_2 parameter: Dashed reference line is the posterior mean.	149
6.13	Trace plot of first 600 sampled values of ρ from MH algorithms with three different starting values (GSS political participation and free speech data).	151
6.14	Two-dimensional trace plot of sampled values of σ_x^2 and σ_y^2 from MH algorithms with three different starting values (GSS political participation and free speech data).	152
6.15	Trace plot of the scale reduction factor \hat{R} for ρ , σ_x^2 , and σ_y^2 across the first 500 iterations of the MH algorithms.	153
6.16	Posterior predictive distributions for the ratio of the mean to the median in the bivariate normal distribution and planar distribution models: Vertical reference line is the observed value in the original data.	158
6.17	Correlations between observed cell counts and posterior predictive distribution cell counts for the bivariate normal and planar distribution models.	160
7.1	Scale reduction factors by iteration for all regression parameters.	176
7.2	Trace plot of error variance parameter in three different MCMC algorithms.	177
7.3	Posterior predictive distributions: (1) The distribution of all replicated samples and the distribution of the original data; (2) the distribution of the ratio of the sample mean of y to the median, with the observed ratio superimposed; and (3) the distribution of the ranges of predicted values with the observed range superimposed.	180
7.4	Posterior predictive distributions (solid lines) and observed values (dashed vertical lines) for persons 1,452 and 1,997.	181
7.5	Distribution of age, education, and income for entire sample and for 31 potential outliers identified from posterior predictive simulation (reference lines at means).	183
7.6	Examples of types of missingness: No missing, missing are not OAR, and missing are not MAR.	185

8.1	Depiction of simulation from a truncated normal distribution. . .	201
8.2	Depiction of rejection sampling from the tail region of a truncated normal distribution.	203
8.3	Trace plot and autocorrelation function (ACF) plot for intercept parameter in dichotomous probit model example (upper plots show samples thinned to every 10 th sample; lower plot shows the ACF after thinning to every 24 th post-burn-in sample).	209
8.4	Sampled latent health scores for four sample members from the probit model example.	211
8.5	Model-predicted traits and actual latent traits for four observations in probit model (solid line = model-predicted latent traits from sample values of β ; dashed line = latent traits simulated from the Gibbs sampler).	212
8.6	Latent residuals for four observations in probit model (reference line at 0).	213
8.7	Distribution of black and age-by-black parameters (Cases 2 and 3 = both values are positive or negative, respectively; Cases 1 and 4 = values are of opposite signs. Case 4 is the typically-seen/expected pattern).	215
8.8	Distribution of black-white crossover ages for ages > 0 and < 200 (various summary measures superimposed as reference lines).	216
8.9	Depiction of latent distribution for Y^* and Y with thresholds superimposed.	220
8.10	Trace plot for threshold parameters in ordinal probit model. . .	224
8.11	Distributions of latent health scores for five persons with different observed values of y	225
8.12	Distributions of R^2 and pseudo- R^2 from OLS and probit models, respectively.	228
9.1	Two-dimensional trace plot of $\alpha_{(0)}$ and $\alpha_{(1)}$ parameters (dashed lines at posterior means for each parameter).	249
9.2	Scatterplots of four persons' random intercepts and slopes from growth curve model of health (posterior means superimposed as horizontal and vertical dashed lines).	261
9.3	Predicted trajectories and observed health for four persons: The solid lines are the predicted trajectories based on the posterior means of the random intercepts and slopes from Figure 9.2; and the dashed lines are the predicted trajectories based on the individuals' covariate profiles and posterior means of the parameters in Table 9.4	265

10.1 Depiction of a bivariate outcome space for continuous latent variables Z_1 and Z_2 : The observed variables Y_1 and Y_2 are formed by the imposition of the vectors of thresholds in each dimension; and the darkened area is the probability that an individual's response falls in the $[2, 3]$ cell, conditional on the distribution for Z 282

10.2 Comparison of truncated bivariate normal simulation using naive simulation and conditional/decomposition simulation: Upper plots show the true contours (solid lines) of the bivariate normal distribution with sampled values superimposed (dots); and lower plots show histograms of the values sampled from the two dimensions under the two alternative sampling approaches. 287

10.3 Truncated bivariate normal distribution: Marginal distribution for x_1 when truncation of x_2 is ignored. 288

10.4 Comparison of truncated bivariate normal distribution simulation using naive simulation and two iterations of Gibbs sampling: Upper plots show the true contours (solid lines) of the bivariate normal distribution with sampled values superimposed (dots); and lower plots show histograms of the values sampled from the two dimensions under the two alternative approaches to sampling. 290

10.5 Gibbs sampling versus Cowles' algorithm for sampling threshold parameters. 293

10.6 Posterior predictive distributions for the probability a 30-year-old white male from the South self-identifies as a conservative Republican across time. 304

10.7 Representation of state space for a three state model. 305

10.8 Two-dimensional outcome for capturing a three-state state space. 306

10.9 Trace plots of life table quantities computed from Gibbs samples of bivariate probit model parameters. 315

10.10 Histograms of life table quantities computed from Gibbs samples of bivariate probit model parameters. 316

A.1 Generic depiction of a curve and a tangent line at an arbitrary point. 324

A.2 Finding successive approximations to the area under a curve using rectangles. 328

B.1 Distributions of sample means for four sample sizes ($n = 1, 2, 30,$ and 100). 339

B.2 Empirical versus theoretical standard deviations of sample means under the CLT. 340

B.3 Sampling distributions of variances in the simulation. 341

B.4	z -based confidence intervals for the mean: Top plot shows intervals for samples of size $n = 2$; second plot shows intervals for samples of size $n = 30$; third plot shows intervals for samples of size $n = 100$	343
-----	--	-----

List of Tables

1.1	Some Symbols Used Throughout the Text	7
2.1	Cross-tabulation of importance of expressing unpopular views with importance of political participation.	21
3.1	CNN/USAToday/Gallup 2004 presidential election polls.	57
4.1	Cell counts and marginals for a hypothetical bivariate dichotomous distribution.	94
6.1	Posterior predictive tests for bivariate normal and planar distribution models.	159
7.1	Descriptive statistics for variables in OLS regression example (2002 and 2004 GSS data, $n = 2,696$).	175
7.2	Results of linear regression of measures of “niceness” on three measures of region.	178
7.3	Results of regression of empathy on region with and without missing data: Missing data assumed to be MAR.	190
8.1	Link functions and corresponding generalized linear models (individual subscripts omitted).	194
8.2	Descriptive statistics for NHANES/NHEFS data used in dichotomous probit model example (baseline $n = 3,201$).	207
8.3	Gibbs sampling results for dichotomous probit model predicting mortality.	214
8.4	Maximum likelihood and Gibbs sampling results for ordinal probit model example.	226
8.5	Various summary measures for the black-white health crossover age.	229

9.1	Results of hierarchical model for voting example under different gamma hyperprior specifications.	240
9.2	Results of hierarchical model for two-wave panel of income and Internet use data.	247
9.3	Results of “growth” model for two-wave panel of income and Internet use data.	257
9.4	Results of growth curve model of health across time.	262
10.1	Results of multivariate regression of measures of “niceness” on three measures of region.	277
10.2	Political orientation and party affiliation.	279
10.3	Multivariate probit regression model of political orientation and party affiliation on covariates (GSS data, $n = 37,028$). . . .	300
10.4	Results of multivariate probit regression of health and mortality on covariates.	310
B.1	Percentage of confidence intervals capturing the true mean in simulation.	343

Introduction

The fundamental goal of statistics is to summarize large amounts of data with a few numbers that provide us with some sort of insight into the process that generated the data we observed. For example, if we were interested in learning about the income of individuals in American society, and we asked 1,000 individuals “What is your income?,” we would probably not be interested in reporting the income of all 1,000 persons. Instead, we would more likely be interested in a few numbers that summarized this information—like the mean, median, and variance of income in the sample—and we would want to be able to use these sample summaries to say something about income in the population. In a nutshell, “statistics” is the process of constructing these sample summaries and using them to infer something about the population, and it is the inverse of probabilistic reasoning. Whereas determining probabilities or frequencies of events—like particular incomes—is a *deductive* process of computing probabilities given certain parameters of probability distributions (like the mean and variance of a normal distribution), statistical reasoning is an *inductive* process of “guessing” best choices for parameters, given the data that have been observed, and making some statement about how close our “guess” is to the real population parameters of interest. Bayesian statistics and classical statistics involving maximum likelihood estimation constitute two different approaches to obtaining “guesses” for parameters and for making inferences about them. This book provides a detailed introduction to the Bayesian approach to statistics and compares and contrasts it with the classical approach under a variety of statistical models commonly used in social science research.

Regardless of the approach one takes to statistics, the process of statistics involves (1) formulating a research question, (2) collecting data, (3) developing a probability model for the data, (4) estimating the model, and (5) summarizing the results in an appropriate fashion in order to answer the research question—a process often called “statistical inference.” This book generally assumes that a research question has been formulated and that a random sample of data has already been obtained. Therefore, this book focuses on

model development, estimation, and summarization/inference. Under a classical approach to statistics, model estimation is often performed using canned procedures within statistical software packages like SAS[®], STATA[®], and SPSS[®]. Under the Bayesian approach, on the other hand, model estimation is often performed using software/programs that the researcher has developed using more general programming languages like R, C, or C++. Therefore, a substantial portion of this book is devoted to explaining the mechanics of model estimation in a Bayesian context. Although I often use the term “estimation” throughout the book, the modern Bayesian approach to statistics typically involves simulation of model parameters from their “posterior distributions,” and so “model estimation” is actually a misnomer.

In brief, the modern Bayesian approach to model development, estimation, and inference involves the following steps:

1. Specification of a “likelihood function” (or “sampling density”) for the data, given the model parameters.
2. Specification of a “prior distribution” for the model parameters.
3. Derivation of the “posterior distribution” for the model parameters, given the likelihood function and prior distribution.
4. Simulation of parameters to obtain a sample from the “posterior distribution” of the parameters.
5. Summarization of these parameter samples using basic descriptive statistical calculations.

Although this process and its associated terminology may seem foreign at the moment, the goal of this book is to thoroughly describe and illustrate these steps. The first step—as well as the associated parameter estimation method of maximum likelihood—is perhaps well understood by most quantitative researchers in the social sciences. The subsequent steps, on the other hand, are not, especially Step 4. Yet advances in Step 4 have led to the recent explosion in the use of Bayesian methods. Specifically, the development of Markov chain Monte Carlo (MCMC) sampling methods, coupled with exponential growth in computing capabilities, has made the use of Bayesian statistics more feasible because of their relative simplicity compared with traditional numerical methods. When approximation methods of estimation were more common, such methods generally relied on normality assumptions and asymptotic arguments for which Bayesians often criticize classical statistics. With the advent of MCMC sampling methods, however, more complicated and realistic applications can be undertaken, and there is no inherent reliance on asymptotic arguments and assumptions. This has allowed the benefits of taking a Bayesian approach over a classical approach to be realized.

1.1 Outline

For this book, I assume only a familiarity with (1) classical social science statistics and (2) matrix algebra and basic calculus. For those without such a background, or for whom basic concepts from these subjects are not fresh in memory, there are two appendices at the end of the book. Appendix A covers the basic ideas of calculus and matrix algebra needed to understand the concepts of, and notation for, mathematical statistics. Appendix B briefly reviews the Central Limit Theorem and its importance for classical hypothesis testing using a simulation study.

The first several chapters of this book lay a foundation for understanding the Bayesian paradigm of statistics and some basic modern methods of estimating Bayesian models. Chapter 2 provides a review of (or introduction to) probability theory and probability distributions (see DeGroot 1986, for an excellent background in probability theory; see Billingsley 1995 and Chung and AitSahlia 2003, for more advanced discussion, including coverage of Measure theory). Within this chapter, I develop several simple probability distributions that are used in subsequent chapters as examples before jumping into more complex real-world models. I also discuss a number of real univariate and multivariate distributions that are commonly used in social science research.

Chapter 2 also reviews the classical approach to statistical inference from the development of a likelihood function through the steps of estimating the parameters involved in it. Classical statistics is actually a combination of at least two different historical strains in statistics: one involving Fisherian maximum likelihood estimation and the other involving Fisherian and Neyman and Pearsonian hypothesis testing and confidence interval construction (DeGroot 1986; Edwards 1992; see Hubbard and Bayarri 2003 for a discussion of the confusion regarding the two approaches). The approach commonly followed today is a hybrid of these traditions, and I lump them both under the term “classical statistics.” This chapter spells out the usual approach to deriving parameter estimates and conducting hypothesis tests under this paradigm.

Chapter 3 develops Bayes’ Theorem and discusses the Bayesian paradigm of statistics in depth. Specifically, I spend considerable time discussing the concept of prior distributions, the classical statistical critique of their use, and the Bayesian responses. I begin the chapter with examples that use a point-estimate approach to applying Bayes’ Theorem. Next, I turn to more realistic examples involving probability distributions rather than points estimates. For these examples, I use real distributions (binomial, poisson, and normal for sampling distributions and beta, gamma, and inverse gamma for prior distributions). Finally, in this chapter, I discuss several additional probability distributions that are not commonly used in social science research but are commonly used as prior distributions by Bayesians.

Chapter 4 introduces the rationale for MCMC methods, namely that sampling quantities from distributions can help us produce summaries of them that allow us to answer our research questions. The chapter then describes

some basic methods of sampling from arbitrary distributions and then develops the Gibbs sampler as a fundamental method for sampling from high-dimensional distributions that are common in social science research.

Chapter 5 introduces an alternative MCMC sampling method that can be used when Gibbs sampling cannot be easily employed: the Metropolis-Hastings algorithm. In both Chapters 4 and 5, I apply these sampling methods to distributions and problems that were used in Chapters 2 and 3 in order to exemplify the complete process of performing a Bayesian analysis *up to, but not including*, assessing MCMC algorithm performance and evaluating model fit.

Chapter 6 completes the exemplification of a Bayesian analysis by showing (1) how to monitor and assess MCMC algorithm performance and (2) how to evaluate model fit and compare models. The first part of the chapter is almost entirely devoted to technical issues concerning MCMC implementation. A researcher must know that his/her estimation method is performing acceptably, and s/he must know how to use the output to produce appropriate estimates. These issues are generally nonissues for most classical statistical analyses, because generic software exists for most applications. However, they are important issues for Bayesian analyses, which typically involve software that is developed by the researcher him/herself. A benefit to this additional step in the process of analysis—evaluating algorithm performance—is that it requires a much more intimate relationship with the data and model assumptions than a classical analysis, which may have the potential to lull researchers into a false sense of security about the validity of parameter estimates and model assumptions.

The second part of the chapter is largely substantive. All researchers, classical or Bayesian, need to determine whether their models fit the data at hand and whether one model is better than another. I attempt to demonstrate that the Bayesian paradigm offers considerably more information and flexibility than a classical approach in making these determinations. Although I cannot and do not cover all the possibilities, in this part of the chapter, I introduce a number of approaches to consider.

The focus of the remaining chapters (7-10) is substantive and applied. These chapters are geared to developing and demonstrating MCMC algorithms for specific models that are common in social science research. Chapter 7 shows a Bayesian approach to the linear regression model. Chapter 8 shows a Bayesian approach to generalized linear models, specifically the dichotomous and ordinal probit models. Chapter 9 introduces a Bayesian approach to hierarchical models. Finally, Chapter 10 introduces a Bayesian approach to multivariate models. The algorithms developed in these chapters, although fairly generic, should not be considered endpoints for use by researchers. Instead, they should be considered as starting points for the development of algorithms tailored to user-specific problems.

In contrast to the use of sometimes contrived examples in the first six chapters, almost all examples in the latter chapters concern real probability

distributions, real research questions, and real data. To that end, some additional beneficial aspects of Bayesian analysis are introduced, including the ability to obtain posterior distributions for parameters that are not directly estimated as part of a model, and the ease with which missing data can be handled.

1.2 A note on programming

Throughout the text, I present R programs for virtually all MCMC algorithms in order to demystify the linkage between model development and estimation. R is a freely available, downloadable programming package and is extremely well suited to Bayesian analyses (www.r-project.org). However, R is only one possible programming language in which MCMC algorithms can be written. Another package I use in the chapter on hierarchical modeling is WinBugs. WinBugs is a freely available, downloadable software package that performs Gibbs sampling with relative ease (www.mrc-bsu.cam.ac.uk/bugs). I strongly suggest learning how to use WinBugs if you expect to routinely conduct Bayesian analyses. The syntax of WinBugs is very similar to R, and so the learning curve is not steep once R is familiar. The key advantage to WinBugs over R is that WinBugs derives conditional distributions for Gibbs sampling for you; the user simply has to specify the model. In R, on the other hand, the conditional distributions must be derived mathematically by the user and then programmed. The key advantage of R over WinBugs, however, is that R—as a generic programming language—affords the user greater flexibility in reading data from files, modeling data, and writing output to files. For learning how to program in R, I recommend downloading the various documentation available when you download the software. I also recommend Venables and Ripley’s books for S and S-Plus[®] programming (1999, 2000). The S and S-Plus languages are virtually identical to R, but they are not freely available.

I even more strongly recommend learning a generic programming language like C or C++. Although I show R programs throughout the text, I have used UNIX-based C extensively in my own work, because programs tend to run *much* faster in UNIX-based C than in any other language. First, UNIX systems are generally faster than other systems. Second, C++ is the language in which many software packages are written. Thus, writing a program in a software package’s language when that language itself rests on a foundation in C/C++ makes any algorithm in that language inherently slower than it would be if it were written directly in C/C++.

C and C++ are not difficult languages to learn. In fact, if you can program in R, you can program in C, because the syntax for many commands is close to identical. Furthermore, if you can program in SAS or STATA, you can learn C very easily. The key differences between database programming languages like SAS and generic programming languages like C are in terms