Bayesian Applications in Environmental and Ecological Studies with R and Stan

Modern ecological and environmental sciences are dominated by observational data. As a result, traditional statistical training often leaves scientists ill-prepared for the data analysis tasks they encounter in their work. Bayesian methods provide a more robust and flexible tool for data analysis, as they enable information from different sources to be brought into the modelling process. This book provides a Bayesian framework for model formulation, parameter estimation, and model evaluation in the context of analyzing environmental and ecological data.

Features:

- An accessible overview of Bayesian methods in environmental and ecological studies
- Emphasizes the hypothetical deductive process, particularly model formulation
- Necessary background material on Bayesian inference and Monte Carlo simulation
- Detailed case studies, covering water quality monitoring and assessment, ecosystem response to urbanization, fisheries ecology, and more
- Advanced chapter on Bayesian applications, including Bayesian networks and a change point model
- Complete code for all examples, along with the data used in the book, are available via GitHub

The book is primarily aimed at graduate students and researchers in the environmental and ecological sciences, as well as environmental management professionals. This is a group of people representing diverse subject matter fields, which could benefit from the potential power and flexibility of Bayesian methods.

Chapman & Hall/CRC

Applied Environmental Series

Recently Published Titles

Statistical Geoinformatics for Human Environment Interface
Wayne L. Myers, Ganapati P. Patil

Introduction to Hierarchical Bayesian Modeling for Ecological Data
Eric Parent, Etienne Rivot

Handbook of Spatial Point-Pattern Analysis in Ecology
Thorsten Wiegand, Kirk A. Moloney

Introduction to Ecological Sampling
Bryan F.J. Manly, Jorge A. Navarro Alberto

Future Sustainable Ecosystems: Complexity, Risk, and Uncertainty
Nathaniel K Newlands

Environmental and Ecological Statistics with R
Second Edition
Song S. Qian

Statistical Methods for Field and Laboratory Studies in Behavioral Ecology
Scott Pardo, Michael Pardo

Biometry for Forestry and Environmental Data with Examples in R
Lauri Mehtätalo, Juha Lappi

Bayesian Applications in Environmental and Ecological Studies with R and Stan
Song S. Qian, Mark R. DuFour, Ibrahim Alameddine

For more information about this series, please visit: https://www.crcpress.com/Chapman--HallCRC-Applied-Environmental-Statistics/book-series/ CRCAPPENVSTA

Bayesian Applications in Environmental and Ecological Studies with R and Stan

Song S. Qian
Mark R. DuFour
Ibrahim Alameddine

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2023
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

ISBN: 9781138497399 (hbk)
ISBN: 9781032290072 (pbk)
ISBN: 9781351018784 (ebk)

DOI: 10.1201/9781351018784

Typeset in CMR10
by Cenveo Publisher Services

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

To the Reckhow Lab

Contents

Foreword

I arrived at Duke University in 1980, eager to begin a research program focusing on statistical analysis and uncertainty analysis based on lake water quality data. I had learned about Bayesian analysis in my PhD program at Harvard, but I had yet to apply Bayesian techniques in my research. As a water quality modeler focused on statistical methods, I began to realize that Bayes' Theorem provided a logical framework for learning with new information. So, why not utilize the expert knowledge of water quality scientists to augment the information in water quality data? This expert judgment could become an informative prior probability to be updated with a likelihood function based on data. Ironically, at this time in the mid-1980s, Bayesian analysis was dominated by noninformative priors. This emphasis was gradually changing. So, in the mid-1980s I began research work with informative priors. An article in *Science* magazine [Malakoff, 1999] describes the emergence of applied Bayesian statistics, beginning with a brief description of my experiences attempting to apply Bayesian analysis to a water quality modeling problem in the mid-1980s.

Over the next 30+ years, I was fortunate to work with exceptional Ph.D. students who became my colleagues and friends. Beginning with Bill Warren-Hicks in the late 1980s, I established a significant intellectual collaboration with Robert Wolpert, a Bayesian statistician who successfully endeavored to understand the water quality modeling issues as he worked with us on Bayesian methods. That work with Bill and Robert was followed by Craig Stow's dissertation, which was the first of several addressing water quality in the Florida Everglades region. At the same time, Robert and I worked with Laura Steinberg, my Ph.D. advisee from Civil and Environmental Engineering, on a Bayesian model for PCB transport in the Hudson River. During this time, I continued to learn from my Ph.D. students and my Duke colleagues. Yet it was so much more than a rich learning experience. These students became my friends, and they became friends and colleagues with each other. Song Qian was with our group in the mid-1990s, and he helped me understand nonparametric Bayes. Song, Laura Steinberg, Robert Wolpert, and Michael Lavine increased my depth of understanding of Bayesian methods during those years. Beginning with this work, Song became the most adept Bayesian statistical modeler in my research lab. From 1996 to 2004, on a part-time basis, I became Director of the University of North Carolina Water Resources Research Institute. During that time, Craig Stow assumed some of my Duke teaching responsibilities, as well as advising my grad students. In this position, Craig developed into an excellent scholar and mentor, producing several outstanding Bayesian water quality modeling papers. In the late 1990s, I became interested in Bayesian networks. Mark Borsuk was a member of our grad student group at that time, and in his dissertation research, Mark demonstrated the merits of Bayesian networks applied to a water quality modeling problem. Mark is a terrific communicator and scholar, as confirmed by his academic position at Duke. Thanks to Bill, Craig, Laura, Song, and Mark, our program at Duke became the academic center for Bayesian water quality modeling. At that same time, our lab group included several Ph.D. students who contributed to the sharing of knowledge and camaraderie through their own research; these students/colleagues were Conrad Lamon, Barbara Van Harn Adams, Pauline Vaas, and Tom Stockton. Prior to their Ph.D. research, Pauline and Tom were professional master's students with me in the 1980s. Tom and I were distance running companions, competing in local 5/10 Ks and training/running in the Boston Marathon! Also during that time, we were fortunate to have George Arhonditsis as a post-doc in my lab. George's intellect and engaging personality were great assets to our lab. Ph.D. student Melissa Kenney provided a similar energy and leadership in my lab, particularly with my professional master's students. During that time, Song and I worked with my Ph.D. student Roxalana Kashuba along with USGS scientists on Bayesian Methods to assess the effect of urbanization on aquatic ecosystems using the USGS NAWQA data. Also during that time, Farnaz Nojavan Asghari added intellectually and internationally to our lab; aside from US researchers in the Reckhow lab, students/colleagues in our group were from China, Lebanon, Iran, Greece, and South Korea. Around that time, Drew Gronewold and Ibrahim Alameddine joined our lab group. Drew and Ibrahim continued to expand the application of Bayesian statistics in water quality modeling. They were not only great scholars; they were great mentors, as evidenced by each assuming academic positions following completion of the Ph.D. As I moved toward academic retirement in 2010, my group included two Ph.D. students from South Korea, Boknam Lee and Yoonkyung Cha. Boknam did some terrific work on Bayes nets focused on pollutant export from the hundreds of high-intensity hog operations in coastal North Carolina. Yoonkyung collaborated with Craig Stow at NOAA, modeling the *spirogyra* and *cladafora* that accumulated on the shores and beaches of Saginaw Bay. This work brought back distant memories; in the mid-1960s as a high school student, I would go with my classmates to Lake Erie beaches in Canada where we would have to wade through shoreline accumulations of *cladafora* to get to clear swimming water! I am delighted and humbled that Song and Ibrahim have dedicated their book to our lab group. As with many of life's endeavors, I ventured on an academic career that was fraught with uncertainty, least of which was the uncertainty in the water quality modeling that was the focus of my career! Without doubt, the most gratifying outcome of my academic career is to have worked with the wonderful scholars and friends who were part of the “Reckhow lab.” This honor that Song, Mark, and Ibrahim have bestowed on our research group is something that I will cherish forever.

– Kenneth H. Reckhow

In my final year of graduate school at Duke University, we had a new student from China enter the program. We came from very different backgrounds and had very different life experiences, but we had something in common – we were both excited. He was excited to be starting and I was excited to be finishing. Another thing we had in common was a shared recognition of the potential to contribute to environmental decision-making by immersing ourselves in learning more about quantitative approaches. Although we overlapped in graduate school, what we learned there was very different. During my tenure, Bayesian statistics was fairly controversial and, for most applied problems, not very practical. By the time Song enrolled, fast, cheap computing and the coevolution of modern software resulted in an explosion of new possibilities and applications. I learned about Bayesian statistics; Song learned how to really make it work. Those complementary experiences served us well as we became friends, colleagues, and frequent collaborators (and PhDs). Song has become an ace programmer who dives deeply into statistical theory for use in environmental applications and typically offers-up novel approaches to problems I’ve been pondering. Our collaborations often work something like this – I approach him with data from a particular system (often a lake) and ask: “what if we did this?” He’ll think it over and reply: “here’s a better approach.” And it usually is. Song’s (or rather Dr. Qian’s) updated book captures his interest in both statistics and environmental science and reflects what is now 25+ years of experience. Along with our respected colleagues Drs. DuFour and Alameddine, he presents a range of applied problems, a bit of statistical theory for context, and code to reveal the stories lurking in the data. I’ve learned a lot from working with Song over the years and I know readers will learn a lot from this book.

– Craig A. Stow

Preface

We wrote this book to summarize our journey of learning and using statistics in our careers as environmental scientists. Statistics is a tool for inductive reasoning, using hypothetical deduction as the main inference method. The hypothetical deductive nature of statistical inference often dictates how statistics is taught and learned: we learn the deductive part of statistics and largely ignore the hypothetical nature of the deductive process. The statistical curriculum in a typical environmental sciences, biology, or ecology graduate program remains strongly influenced by biostatistics – a sub-field of statistics that focuses on statistical models developed to assess experimental data. Because proper experimental design can ensure that the resulting data meet the intended models, we learn which statistical model is suitable for each type of experiment.

The modern ecological and environmental sciences are, however, dominated by observational data. As a result, our traditional statistical training often leaves us ill-prepared for the data analysis tasks in our professional work. The three of us worked for a few years after college before going back to graduate school. A common motive for pursuing a graduate degree was the need to be better educated in statistics. In different ways, we were attracted by the potential of Bayesian statistics, although we initially had no idea what it was. Through our graduate studies and, more importantly, our professional work, we gradually learned the art of applying statistics and honed our abilities to use Bayesian statistics. We decided to write this book to summarize what we learned and re-learned in our combined approximately 45 years of professional life. In the process of writing the book, we revisited many of our published papers and reviewed them with critical eyes. We want to present the process of statistical application in our field, with an emphasis on how to propose and justify the model – the hypothetical part of the statistical inference. The deductive part is quite straightforward under a Bayesian framework – through Bayes' theorem to derive the posterior distribution, combining the prior (summarizing what we already know) and the likelihood (the support of the proposed model from data). By organizing our work based on the type of response variable data, we hope to provide our colleagues with a collection of case studies that can help them avoid some of the detours we encountered.

Chapter 1 is an overview of the book, including a summary of our understanding of applied statistics, an overview of the general principles in application, and a summary of some of the examples used in the book. Chapter 2 documents some of the most commonly used methods for generating random numbers from a probability distribution. Chapter 3 introduces the general process of the Bayesian inference, including our understanding of where and how to start when proposing a proper informative prior. These three chapters form the methodological basis of the rest of the book. The methods discussed in Chapter 2 are largely incorporated in most modern computer software for Monte Carlo simulation. They are unlikely to be directly used by a practitioner. A summary of our experience in writing Stan code is included in this chapter. Chapter 4 summarizes a number of examples with response variables approximated by the normal distribution. Examples in Chapter 4 are mostly related to environmental monitoring and assessment. Chapter 5 describes some commonly used models for modeling count data. The focus of Chapter 5 is on imperfect detection – a common feature in many ecological data sets. Chapter 6 is a collection of examples to illustrate the Bayesian hierarchical model (BHM). We see BHM as a tool for properly analyzing "big data" – data collected from multiple sources. We argue that properly modeling the hierarchical structure of the data is the key to resolving Simpson's paradox. Chapter 7 includes two relatively independent topics – the Bayesian network model and the Bayesian change point/threshold model. The book ends with a number of concluding remarks in Chapter 8, where we revisit the general themes of statistical inference with references to some of the examples covered in the book, as well as a discussion of the connection between classical hypothesis testing and Bayesian statistics. Our approach to the subject is application oriented, which may be justifiably criticized by both classical and Bayesian statisticians.

A casual reader can start the book with Chapter 1, Chapter 3, and Sections 2.5–2.6, then go to a chapter (Chapters 4–7) of choice (e.g., matches the application in hand), and finish with Chapter 8.

The R and Stan code printed in the book are not complete by themselves. They should be used as a reference. Complete code for all examples, along with the data used in the book, are available at the book's GitHub repository (`github.com/songsqian/beesrstan`). The GitHub repository is regularly updated.

Song S. Qian	Mark R. DuFour	Ibrahim Alameddine
Sylvania, Ohio	Sandusky, Ohio	Beirut
USA	USA	Lebanon

Acknowledgments

This book is possible due largely to the support and mentoring we had from Professor Kenneth H. Reckhow. Ken fostered a warm and collegiate environment in his lab. Ken's conviction and enthusiasm for adopting Bayesian statistics as a means to better understand modern environmental problems was and continues to be irresistible. Ken is serious about two things in life: Bayesian statistics and his support for the Blue Devils. His pioneering work in the use of Bayesian statistics in environmental modeling and management was seminal and remains inspiring. His lab attracted researchers from different backgrounds and varied prior experiences. His dedication to simplifying complex problems, ensuring that the forest is not missed for the trees, and his curiosity to explore topics at the fringe of his expertise have helped us become productive researchers and educators, each with a unique trajectory and a different focus area. In our years with Ken, we were given the freedom and encouraged to explore. Our dissertation research often established novel linkages between surface water quality modeling and research questions raised by policy makers, city planners, ecologists, toxicologists, and engineers. Ken always found a way to build on the individual strength of each of his students and was always open to exploring a new avenue as long as we could justify the approach to him. The process of developing a research topic, rather than being handed one, and the focus on being able to convincingly defend the adopted rationale made us all better researchers. The only mandatory task that Ken ever assigned to his students was to sit for Robert L. Winkler's class in Bayesian statistics and decision theory. While we all struggled through the class and, in the process, developed a habit of working with our fellow lab mates to work through hard problems, we all came out of that class with a better appreciation of Ken's commitment to Bayesian statistics as a means for environmental modeling and decision making. Over the years, the Reckhow lab has become a big family that now spans the globe with colleagues and collaborators who continue to work together. Many of us can trace many of the most important events that shaped our development journey to the time we spent at Ken's lab. For the first author (SSQ), the most memorable quote from Ken was "if a regression model has an R^2 value of more than 0.9, you should check to see if there is anything wrong with it." Reading Efron and Morris [1977] (a group activity organized by Laura J. Steinberg) in Spring 1991 is another important event in his tenure as a graduate student. Although the content of the paper was beyond his comprehension at the time, the paper was the initial inspiration for his pursuit of the Bayesian hierarchical modeling approach. The paper led to numerous discussions with Ken, and later with Craig A. Stow in the years after graduate school (sometimes over a shot of bourbon), and these discussions inspired many exciting ideas. For the third author (IA), Ken's summer book club that was dedicated to exploring the inner workings of Bayesian networks and SSQ's unwavering diligence to adopt the hierarchical Bayesian modeling framework as a means to tease out the signal from the noise, when data are collected at different spatial scales, are two landmark moments. Over the years, we had chances to collaborate with many of Ken's students, including E. Conrad Lamon, Mark E. Borsuk, Yoonkyung Cha, Andrew D. Gronewold, Farnaz Nojavan Asghari, and George B. Arhonditsis.

Ken's lab was open and inclusive. Each of us had committee members from statistics and other departments. Professors Michael L. Lavine and Robert L. Wolpert were popular committee members among Ken's students. We owe them gratitude for their tutelage of the fine points of applied statistics.

For the second author (MRD), his journey in contributing to this book began with an opportunity to join Professor Christine M. Mayer's lab and explore the use of Bayesian statistical applications in fisheries. The appeal of Bayesian methods was great while the struggle was real. As luck would have it, SSQ soon joined the department and immediately became a mentor and friend. Ken's emphasis on thought-provoking discussion, freedom to intellectually explore, development of coherent research topics, and culminating in statically rigorous applications continues to generate a learning and working environment that allows each of our students to grow as a researcher and person, and future students will continue to benefit from this influence.

Outside the extended "Reckhow family," we owe gratitude to many colleagues. Curt J. Richardson helped to clarify many questions we had regarding wetland nutrient retention in the Everglades examples. Boping Han and Dengsheng Lu discussed many potential applications of Bayesian statistics in limnology and remote sensing. Thomas F. Cuffney, Jonathan G. Kennen, Mary C. Freeman, Jason May, and Gerald McMahon were instrumental in shaping up the EUSE example. Michael J. Messner and Jonathon Koplos introduced us to the world of drinking water safety assessment. The idea of imperfect detection and the snake fungal disease example were given to us by Jennifer A. Moore. Christine M. Mayer, who mentored MRD, gave us fisheries insight to make the discussions of walleye examples less fishy.

We are indebted to Freya E. Rowland, Patrick M. Kočovský, Jason C. Doll, Jean Adams, and Robin White who read an early version of the book. They corrected many errors and provided helpful comments and suggestions. Our editors Rob Calver and Vaishali Singh oversaw the entire process, the book proposal, the writing, and the manuscript review. Without them, the book would have been impossible.

All errors remaining in the book are ours. A live errata is on the book's GitHub repository (`https://github.com/songsqian/BeesRStan`).

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Chapter 1

Overview

This is a book for practitioners. Our goal is to introduce the procedure of using Bayesian statistics in environmental and ecological studies. As a practical application-oriented book, we use examples from the literature, most of them our work, to illustrate our approach to an applied statistics problem. In this chapter, we outline the inductive nature of statistical inference, the three basic questions in statistical inference, our guiding principles, and examples we use in multiple chapters. Our guiding principles are further summarized in Chapter 8, with an emphasis on the iterative nature of using Bayesian statistics in a problem (Figure 1.5).

1.1 Two Modes of Reasoning: Deduction vs. Induction

Bayesian statistics is named after the British Presbyterian minister Thomas Bayes (1702–1761), whose posthumously published paper entitled "An essay towards solving a problem in the doctrine of chances" appeared in the *Philosophical Transactions of the Royal Society of London* in 1763. In the paper, Bayes presented a solution to a problem of inverse probability regarding the quantification of a binomial distribution probability. Some speculated that the motivation of the paper was to respond to David Hume's essay in which Hume questioned the validity of inductive reasoning. Inference from cause to effect is an example of deductive reasoning, while the inverse problem, inference about the cause based on observed effects, is inductive reasoning. Although the differences are philosophical, the validity of induction brings into question all empirical claims, including science and religion (e.g., miracles as evidence supporting the existence of God). At the time, the concept of probability was only explored in the context of gambling. Although methods for calculating the probability of an effect (e.g., the chance of having a hand with four aces in a poker game) from a cause (e.g., dealt from a fair deck of cards) was well understood, the inverse problem (e.g., what is the chance that the deck is loaded if a poker player receives four aces in four consecutive hands) was hardly obvious at the time. The statistical question answered in Bayes' essay is related to a binomial problem. A deduction in this problem is to calculate the probability of observing x positives (or successes) in n trials when the probability of success is known, a question long answered by Bernoulli. Bayes' paper is about how to calculate the probability of success after observing x successes in n trials. This problem is actually very hard. The solution Bayes provided requires an initial guess of what the likely probability (p) would be. For example, we may limit the likely value to be one of the five values: $p_1 = 0.1, p_2 = 0.2, p_3 = 0.3, p_4 = 0.4$, and $p_5 = 0.5$. Then we can use Bayes' theorem to calculate the probability of each of the five possible values to be the true probability of success provided we also know, a priori, how likely each of the five probabilities are to be true.

Suppose we don't know how likely each of the five values are to be true. One way to express this ignorance is to assume that the five values are equally likely. In other words, we set $\Pr(p = p_i) = 0.2$ for $i = 1, \cdots, 5$. For a binomial process with probability of success p_i, the number of successes X is a random variable and its distribution is described by the binomial distribution. The probability of observing x successes in n trial is $\Pr(Y = y, n|p_i) = \binom{n}{y} p_i^y (1-p_i)^{n-y}$. Using Bayes' theorem, we can update the probability of being the true probability of success (p_i) after observing $data = \{y, n\}$:

$$\Pr(p = p_i|data) = \frac{\Pr(p_i)\Pr(data|p_i)}{\sum_{j=1}^{5}\Pr(p_i)\Pr(data|p = p_j)}. \tag{1.1}$$

If we suppose that we conducted a study and obtained $x = 3$ successes in $n = 10$ trials, then equation (1.1) can be tabulated (Table 1.1).

TABLE 1.1: A discrete binomial distribution problem – tabulated for easy calculation and accuracy check.

	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	sum
$\Pr(p_i)$	0.2	0.2	0.2	0.2	0.2	
$\Pr(data\|p_i)$	0.0574	0.2013	0.2668	0.2150	0.1172	
$\Pr(p_i) \times \Pr(data\|p_i)$	0.0115	0.0403	0.0534	0.0430	0.0234	0.1715
$\Pr(p_i\|data)$	0.0669	0.2347	0.3111	0.2506	0.1366	

This tabulated method was initially used by Pierre-Simon Laplace, who was responsible for the general form of Bayes' theorem we use now. Using the general form of equation (1.1), we can tabulate the calculation, so that the tedious arithmetic is organized and easy to check for accuracy. Although we now can easily automate the calculation process using an R function, Table (1.1) is conceptually appealing.

```
#### R code ####
y <- 3
n <- 10
pri <- (1:5)/10
```

```
prior <- rep(0.2,5)
likelihood <- dbinom(y,n, pri)
post <- likelihood*prior/sum(likelihood*prior)

## in a function ##
bayes_binom <- function(y, n, p_i=pri, prior=NULL){
  k <- length(p_i)
  if (is.null(prior)) prior <- rep(1/k, k)
  likelihood <- dbinom(y, n, p_i)
  return(likelihood * prior / (sum(likelihood * prior)))
}
### End ###
```

The posterior probabilities $\Pr(p_i|y)$ show that, upon observing $y = 3$ successes in $n = 10$ trials, we believe that the probability of success is most likely to be $p = 0.3$, with a "posterior" probability of 0.3111.

Let's suppose that we carried out another study with a result of $y_2 = 2$ successes in $n_2 = 12$ trials. Because of the previous study, we have a better understanding the relative likelihood of the five possible values. Our interpretation of the new data should be based on the most recent information. That is, we should use the estimated posterior probabilities as the prior probabilities:

```
#### R code ###
> bayes_binom(y=2, n=12, prior=post)
[1] 0.10108542 0.43676278 0.34264060 0.10505992 0.01445127
>
### End ###
```

The information contained in the two batches of data should be the same with or without our two successive applications of Bayes' theorem. That is, our posterior probabilities should be the same if we use the flat prior (0.2) and the combined data of $\{y = 5, n = 22\}$:

```
#### R code ###
> bayes_binom(y=5, n=22)
[1] 0.10108542 0.43676278 0.34264060 0.10505992 0.01445127
>
### End ###
```

The above example illustrates the difference between deductive and inductive reasoning. Using deductive reasoning, we start from what we know to predict the outcome. As long as what we know is correct, the prediction will be correct. If we know the probability of success is $p = 0.3$, we can easily calculate the likelihood of observing $x = 3$ successes in $n = 10$ trials (`dbinom(x=3, n=10, p=0.3)`). Induction is the inverse process of figuring out the likely value of the probability of success when observing the data. In the binomial example, we start the process by providing an initial guess (the prior) and Bayes' theorem updates the prior with data. The updating process can be iterative.

The updating process is straightforward, but often tedious. In the above example, we limited the estimation accuracy of p_i to one decimal point (increment of 0.1). If we reduce the incremental increase to 0.05, we have nine potential values.

```
#### R code ####
pi <- seq(0.1,0.5,0.05)
post <- bayes_binom(y=5, n=22, p_i=pi)

print(cbind(pi, post))
print(pi_gvn_y <- pi[post==max(post)])
### End ###
```

Using this method we can achieve an arbitrary level of accuracy; by decreasing the increments thus increasing the number of potential values of p_i. This approach was used in the 1990s and is often known as Bayesian Monte Carlo (BMC), where thousands of potential values are used. When the number of potential values increases, equation (1.1) can be expressed in a continuous form:

$$\pi(p|y) = \frac{\pi(p)L(y|p)}{\int_p \pi(p)L(y|p)dp} \tag{1.2}$$

where $\pi(\cdot)$ represents a probability density function. For example, $\pi(p)$ is now the prior probability density function of the binomial distribution parameter p. $L(\cdot)$ is the likelihood function. Because the observed data are discrete integers (e.g., $\{y = 5, n = 22\}$), the likelihood is a function proportional to the probability of observing y. In this case, $L(y|p) \propto p^y(1-p)^{n-y}$. Now all we need is a probability density function of p (the prior probability) to describe our uncertainty on the value.

In practice, we often use a beta distribution to describe the uncertainty in a probability variable. The beta distribution has two shape parameters, α and β: $p \sim beta(\alpha, \beta)$ with a probability density function $\pi(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$, where $\Gamma(\cdot)$ is the gamma function. The mean of the distribution is $\alpha/(\alpha + \beta)$, the mode is $(\alpha - 1)/(\alpha + \beta - 2)$ (for $\alpha > 1$ and $\beta > 1$), and the variance is $\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$. By using a properly selected combination of the two parameters, we can use a beta distribution to approximate our uncertainty about p. For example, if we are entirely ignorant about the probability, we can assume that any given value of p is equally likely as any other value, or $\pi(p) = 1$ (the flat prior), which is $beta(1, 1)$. Sometimes based on the available information, we may believe that the probability is more likely to be either 0 or 1 than any other value. A beta distribution with $\alpha = \beta < 1$ can be used (Figure 1.1).

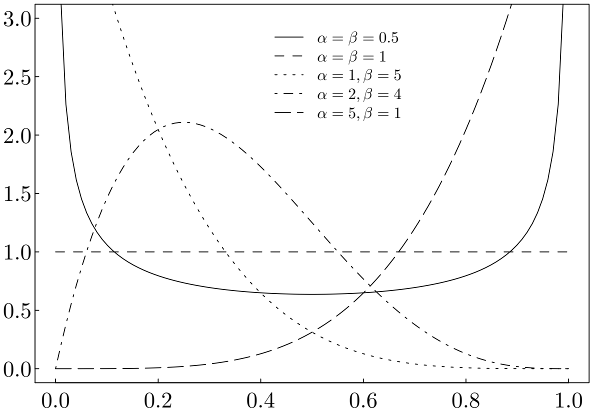

FIGURE 1.1: The beta distribution – commonly used to represent uncertainty in a variable bounded by 0 and 1.

Combining the prior $beta(\alpha,\ \beta)$ and the likelihood function, we derive the posterior distribution:

$$\begin{aligned}\pi(p|y) &= \frac{p^{\alpha-1}(1-p)^{\beta-1}p^y(1-p)^{n-y}}{\int_{p=1}^1 p^{\alpha-1}(1-p)^{\beta-1}p^y(1-p)^{n-y}dp}\\ &= \frac{p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{\int_p p^{y+\alpha-1}(1-p)^{n-y+\beta-1}dp}.\end{aligned}$$

Recognizing that the numerator is proportional to the density function of the beta distribution with parameters $y+\alpha$ and $n-y+\beta$ and the denominator is a constant, we can deduce that the posterior distribution is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. A quick shortcut for solving the integral in the denominator is to multiply both the denominator and the numerator by the beta distribution constant $\frac{\Gamma(y+\alpha+n-y+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}$ such that the denominator is an integral of a beta distribution density, which is 1. Using the data $\{y = 5, n = 22\}$ and a flat prior $beta(1, 1)$, the posterior of p is $beta(6, 18)$. This distribution has a mean of 6/(6+18) = 0.25, a mode of (6-1)/(24-2) = 0.2273, and the 90% credible interval of the posterior distribution is (0.1202, 0.4039) (Figure 1.2).

This binomial example illustrates a general framework of the Bayesian inductive inference. In any applied problem, the first step is always the identification of the response variable. In the binomial example, the response is the number of successes in a given number of trials. A probability distribution model is then applied to the response variable to summarize the data-generating process. A probability distribution model is defined through one or more unknown parameters. The unknown parameter(s) is the target of the inference once observations of the response variable are available.

Characterizing the data generating process is what Fisher called the problem of model formulation, the first step of a statistical modeling problem [Fisher, 1922]. The second step is Fisher's problem of parameter estimation – quantifying the unknown parameter(s) using observed data. The Bayesian approach is to derive the posterior distribution of the parameter, which is proportional to the product of the prior density function and the likelihood function. The difficulties of using the Bayesian approach are (1) finding the appropriate prior distribution and (2) computation (potentially high-dimensional integration).

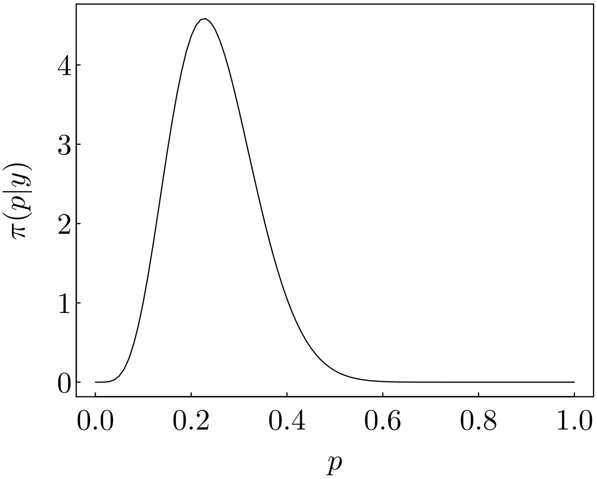

FIGURE 1.2: The posterior distribution of the binomial parameter is a beta distribution.

Many methods can be found in the literature for deriving an appropriate prior distribution. Frequently, researchers choose a class of priors that represents little or no information. The flat prior $beta(1, 1)$ is an example of a "non-informative" prior. In our work, we often find that these non-informative priors are worrisome, especially for parameters without a theoretical upper or lower bound. In the rest of the book we will discuss our understanding of the prior and how to derive a relevant prior in an application whenever appropriate.

1.1.1 Testing for Snake Fungal Disease

We use an example to explain the use of the Bayesian framework and show the difference between deduction and induction. When we have the model parameters, we often want to know the likelihood of an outcome (deduction). Conversely, when we have the outcome, we may want to infer the model parameter values (induction). This example is typical of many applied problems.

Snake fungal disease (SFD), caused by the fungus *Ophidiomyces ophiodiicola*, is an emergent pathogen known to affect at least 30 snake species from 6 families in eastern North America and Europe [Allender et al., 2015, Lorch et al., 2016]. SFD was detected in eastern massasaugas (*Sistrurus catenatus*), a small, federally threatened rattlesnake species in Michigan, USA in 2013 [Tetzlaff et al., 2015]. The estimated SFD prevalence ranges from 3–17% in three Michigan populations [Hileman et al., 2017].

A commonly used method for detecting SFD is quantitative polymerase chain reaction (qPCR) to identify the fungal DNA using a skin swab. The method often leads to a false negative because swabbing can miss the disease agent. Hileman et al. [2017] show that a single swab of an eastern massasauga with clinical signs of SFD (skin lesions) can often result in a false negative; a positive result (detecting fungal DNA on an individual snake) does not always indicate that the individual has SFD (false positive). In other words, the test is imperfect and we are uncertain of the snake's true state regardless of the test result. Such uncertainty is quantified using probabilities, specifically, conditional probabilities.

For the purpose of discussion, we used a sample of 20 snakes, 5 of which tested positive for SFD. As the effectiveness of using qPCR for testing SFD is still under study, we use optimistic hypothetical rates of false positive (7%) and false negative (5%). With these facts, how we carry out further analysis depends on the objective of the study. If we are interested in the 20 individual snakes, the question is likely whether the five positive snakes are truly infected and whether the 15 negative snakes are truly free of the fungal disease. If the objective is to learn about the status of the disease in the snake population, the question is likely, "what is the prevalence of the disease in the population?"

To simplify the presentation, we use "+" to represent a positive test result and "−" to represent a negative test result. Likewise, we use *pr* and *ab* to denote the presence and absence of the fungus, respectively. That is, the unknown true state of the world is represented by *pr* or *ab* and the observed test results are + or −. Once we observe a + or − from a snake, we want to know the likelihood that the snake is infected or not. Using a probability symbol, we want to learn $\Pr(pr|+)$ (and $\Pr(ab|-)$).

Deduction: The following calculation shows that the disease prevalence is necessary in the deductive process. To summarize, we know the following quantities:

$$\Pr(pr) = 0.03, \Pr(+|ab) = 0.07, \text{ and } \Pr(-|pr) = 0.05$$

and observed a positive test result (+, data). If the objective is specifically the disease status of the 20 individual snakes, we have a deductive problem. There are two possible true states of the world: each snake is either infected (*pr*) or not infected (*ab*). From the known quantities and the data, we want to learn $\Pr(pr|+)$, the probability that *pr* is the true state of the world.

Before a formal statistical analysis, let us consider a hypothetical population of 10,000 snakes. Suppose that we can test all 10,000 snakes. The known prevalence of the disease (3%) tells us that 300 snakes are infected. These 300 infected snakes (and a 5% false negative rate) will result in 15 false negatives and 285 true positives. The 9,700 healthy snakes (and a 7% false positive rate) will result in 679 false positives and 9,021 true negatives. After testing all 10,000 snakes, there will be a total of 964 (679+285) positives and 9,036 (9,021+15) negatives. Among the 964 positives, 285 are truly infected. As a result, the likelihood of a positive snake being truly infected is 285/964 = 0.2956. This calculation is a deductive reasoning process. That is, we start from the known causes of a positive result to infer the probability of a positive snake being truly infected.

Bayes' theorem can be interpreted by this deductive calculation:

$$\Pr(pr|+) = \frac{\Pr(pr)\Pr(+|pr)}{\Pr(pr)\Pr(+|pr) + \Pr(ab)\Pr(+|ab)}. \tag{1.3}$$

When we know the prevalence $\Pr(pr)$ (3%), we can see that Bayes' theorem represents the same deductive calculation as illustrated in the previous paragraph. Before a snake is tested, we don't know whether the snake is infected or not. The prevalence of SFD allows us to make a probabilistic statement about the snake's status. Once the snake is tested, the snake population is now divided into two virtual populations: those would-be positive snakes and those would-be negative snakes. Bayes' theorem updates the prevalence in these two virtual populations. In the would-be positive population, the prevalence is now 0.2956. In other words, we can update the probability statement with regard to the positive snake.

Induction: If the prevalence $\Pr(pr)$ is unknown, and the goal of testing snakes is to estimate the prevalence, we then have an induction problem. That is, we infer the cause (the prevalence) from the effects (observing five positives in 20 snakes). Bayes' theorem of equation (1.3) cannot be used directly because the population prevalence ($\Pr(pr)$) is unknown; therefore, testing just one snake is no longer a viable option. Let us now return to the original data of 5 positives from a sample of 20 snakes.

To simplify the discussion, we will first assume that the test is perfect with $\Pr(+|ab) = 0$ and $\Pr(-|pr) = 0$, which brings us back to the binomial problem. As before, we can simplify the problem by limiting the prevalence value to one decimal point, that is, $\Pr(pr) = \theta = 0.0, 0.1, 0.2, \cdots, 0.9, 1.0$. More generally, $\theta = \theta_1, \cdots, \theta_k$. Alternatively, we can use a beta distribution (e.g., $beta(1, 1)$) to represent the prior distribution of the prevalence. The posterior distribution of the prevalence is also a beta distribution ($beta(1 + 5, 1 + 20 - 5)$).

The problem is considerably more complicated because the test is imperfect. With an imperfect test, the probability of observing a positive result is no longer the prevalence. Using the Bayesian notation, we use θ to represent the unknown parameter of interest. In this case, it is the prevalence $\theta = \Pr(pr)$. Also, we use $f_p = \Pr(+|ab)$ to represent the false positive probability and $f_n = \Pr(-|pr)$ to represent the false negative probability. The data we obtain from testing a number of snakes are a binomial variable. The probability of observing a positive result is $p_+ = \theta(1-f_n) + (1-\theta)f_p$. The probability of observing $y = 5$ positive in $n = 20$ trials is:

$$L(y=5, n=20|p_+) = \binom{20}{5} p_+^y (1-p_+)^{n-y}.$$

Using Bayes' theorem (equation (1.2), the posterior distribution of θ is:

$$\begin{aligned}\pi(\theta|y) &= \frac{\theta^{1-1}(1-\theta)^{1-1}\times p_+^y(1-p_+)^{(n-y)}}{\int_\theta \theta^{1-1}(1-\theta)^{1-1}\times p_+^y(1-p_+)^{n-y}d\theta} \\ &= \frac{(\theta(1-f_n)+(1-\theta)f_p)^y(1-\theta(1-f_n)-(1-\theta)f_p)^{n-y}}{\int_\theta(\theta(1-f_n)+(1-\theta)f_p)^y(1-\theta(1-f_n)-(1-\theta)f_p)^{n-y}d\theta}.\end{aligned} \tag{1.4}$$

The integral in the denominator is no longer simple. However, we can derive numerical solutions using the characteristics of a probability density function, namely, the density function integrates to unity: $\int_0^1 \pi(\theta|y)d\theta = 1$, or the area under the density curve is 1. We can approximate the area under the curve by summing the numerator values on the right-hand side of equation (1.4) evaluated over an evenly spaced grid of θ between 0 and 1 and multiplying the sum by the width of the grid. Dividing these same numerator values by the approximate area under the curve, we have the estimated posterior density function, both numerically and graphically (Figure 1.3).

```
post_impft <- function(y=5, n=20, fp=0.07, fn=0.05, k=100){
    theta <- seq(0, 1,, k)
    fpst <- theta*(1-fn) + (1-theta)*fp
    post <- y*log(fpst) + (n-y)*log(1-fpst)
    return(exp(post)/(theta[2]*sum(exp(post))))
}

plot(seq(0, 1, , 100), (post2 <- post_impft()), type="l",
     xlab="$\\theta$", ylab="$\\pi(\\theta|y)$")
### End ###
```

With these two examples, we show the difference between deduction and induction. The most important difference is the target of the inference. When we know the prevalence, we want to know the likelihood of infection of an individual snake (deduction). Consequently, Bayes' theorem calculated $\Pr(pr|+) = 0.2956$ is specific to individual snakes. The flow of information goes from the cause and data to the quantity of interest. When the prevalence is unknown, our goal becomes the estimation of the prevalence based on the test result (induction). The flow of information goes from data back to the cause (Figure 1.4). While the deductive calculation is definite, the inductive process requires additional information about the quantity of interest.

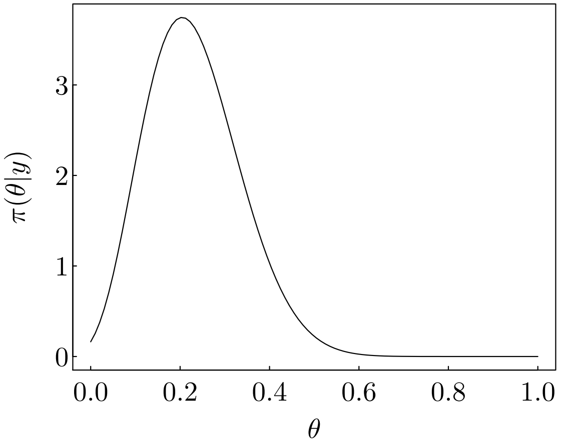

FIGURE 1.3: The posterior distribution of the unknown prevalence of snake fungal disease is graphically displayed.

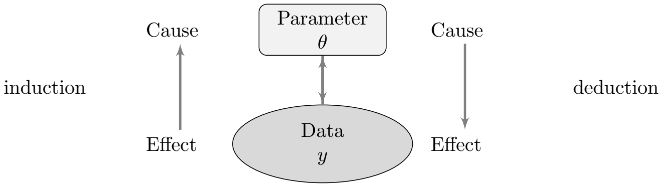

FIGURE 1.4: Deduction reasoning goes from cause to effect, whereas induction goes from effect to cause.

1.2 Bayesian versus Classical Statistics

Neyman and Pearson started their 1933 paper that developed the classical statistical hypothesis testing theory (the Neyman-Pearson lemma) with a review of existing methods for testing statistical hypothesis. They cited Bayes as the first to develop a test to learn about a causal relationship – what are the probabilities that the observed data (e.g., $y = 5, n = 20$) are caused by several likely events (e.g., p_j)? They quickly declared that their work was about a different kind of hypothesis testing – a procedure for discovering a "characteristic" of the data, upon which one can determine whether to reject the hypothesis of interest. The procedure ensures that the hypothesis will be rejected only infrequently when it is correct, and rejected when it is wrong with a high probability. This is the classical null hypothesis testing, where a null hypothesis (e.g., $p = 0.3$) is set against an alternative hypothesis ($p \neq 0.3$). A test statistic, for example, number of successes (y) in n trials, a binomial random variable with $p = 0.3$ under the null hypothesis, is compared to a set of criteria. In this case, if $2 \leq y \leq 10$, we do not reject the null hypothesis and otherwise we do. This procedure ensures that when the null hypothesis is true (i.e., $p = 0.3$) the probability of rejecting the null is approximately $\alpha = 0.05$. Neyman and Pearson [1933] showed that when the null is not true, this procedure will reject the null hypothesis with the highest probability among any other kind of tests. In the introduction section of their paper, Neyman and Pearson stated that the objective of such procedure is not to determine how likely it is taht the null hypothesis is true. Rather, the procedure provides a rational "rule of behavior:"

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x < x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.

This long-run frequency interpretation of statistics is the foundation of all classical statistics. For example, confidence intervals are used to measure the uncertainty in almost all estimated parameters. The interpretation of a confidence interval is not about the likely range of the parameter, rather a rule of behavior. That is, the parameter of interest is a fixed number and the confidence interval is random. The probabilistic statement of being 95% confident is about the confidence interval, not about the parameter of interest – if we repeat the same estimation process and construct the confidence interval each time, 95% of the intervals will include the true parameter value. But for the specific confidence interval at hand, there is no hope to determine whether the true parameter value is inside or outside of the confidence interval. This is a rule of behavior, in that, if we use the confidence interval as a guide to make decisions about the parameter of interest, we will be correct 95% of the time. This rule of behavior may be fine in some situations, but is often confusing when used in scientific research.

In scientific research, we are interested in the underlying causal relationship, that is, we want to know the value of the SFD prevalence θ. Any specific value of the prevalence (e.g., $\theta = 0.3$) is unlikely to be exactly true. As a result, a long-run frequency approach (e.g., testing whether $\theta = 0.3$) is simply unsatisfactory. The Bayesian approach is focused on the understanding of the causal relationship. As a result, the Bayesian approach is more focused on the problem of estimation. Although the core component of the Bayesian framework, the likelihood function, is the same used in the classical statistics, the Bayesian view of the role of data in our quest for the truth is fundamentally different from the long-run frequency perspective. However, scientists are often naturally Bayesian, in that, we rarely interpret statistical results in terms of long-run frequency. Because graduate-level statistics instruction in scientific disciplines is nearly inevitably based on classical statistics, we are accustomed to the standardized classical statistical models, from t-test, to linear regression and generalized linear models, and sometimes multilevel models (most likely referred to as the mixed effect models). When Bayesian statistics is introduced, we often use non-informative priors on familiar models, which leads to posterior distributions of parameters of interest similar to the classical statistical estimates (most likely using the maximum likelihood estimator). As a result, some question the need for Bayesian statistics, especially with the added computational burden.

Because of the shared likelihood function in both classical and Bayesian statistics, they also share some basic principles as we discussed earlier. We would start a statistical inference problem with the problem of model formulation and likely arrive at the same probabilistic model, regardless of our statistical denomination. Suppose that the Bayesian statistician has no prior knowledge for the parameter of interest; both the classical statistician and the Bayesian statistician summarize information in the data using the same likelihood function. The difference between the two lies in how they present the information represented in the likelihood function. The classical statistician would present the parameter value that maximizes the likelihood function as the best estimate of the unknown parameter (i.e., developing the maximum likelihood estimator). Finding the value is a mathematical optimization process, mainly through derivative operation. The Bayesian statistician would normalize the likelihood function and present the unknown parameter as a probability distribution. The Bayesian computation is mostly solving integrals. For example, in the snake fungal disease example, the likelihood function of the problem in equation (1.4) is

$$L = (\theta(1-f_n) + (1-\theta)f_p)^y(1-\theta(1-f_n) - (1-\theta)f_p)^{n-y}.$$

The maximum likelihood estimator (MLE) of θ is often derived by setting the first derivative of the likelihood function to 0. Because the log transformation is a monotonic transformation, we can find the MLE using the log-likelihood, which often simplifies the operation. In this case,

$$\frac{d\log(L)}{d\theta} = \frac{y(1-f_n-f_p)}{\theta(1-f_n-f_p)+f_p} - \frac{(n-y)(1-f_n-f_p)}{1-\theta(1-f_n-f_p)-f_p}.$$

Setting the right-hand side to be 0, the MLE of θ is

$$\theta_{MLE} = \frac{y - nf_p}{n(1 - f_n - f_p)}.$$

The posterior distribution function, on the other hand, requires the integration of the likelihood function. In many cases, the integral may not have an analytic solution. As a result, numerical approximation is necessary.

Classical statistics has a tradition of emphasizing easy access, for example, the early effort of tabulating commonly used probability distributions. As a result, efficient and fast analytic/numeric algorithms for a large number of standard problems, including linear, nonlinear, and the generalized linear models, were developed and implemented long before the advent of powerful personal computers, making classical statistics methods readily available for scientific applications. The same did not happen in Bayesian statistics.

However, routine applications of classical statistics are limited to a number of standard models. These methods limit the choices of models in an applied problem. As a result, practitioners are advised to design their experiment with the intended statistical methods in mind. In ecology, the institutionalized approach of randomized experimental design for causal inference and the concerns of "pseudo-replication" are examples of this influence. The lack of computer algorithms for standard models in Bayesian statistics, however, has required us to formulate a model, which is perhaps why we have perceived it as more difficult. A purposely formulated model is often more realistic, hence more relevant.

Because of the commonality of the classical and Bayesian statistics, especially the inductive nature of both, applications of statistics should not be limited to one or the other. A more effective data analysis approach often combines elements of both. In this book, we often use classical algorithms to explore potential models for a problem. If a standard classical model fits the data, we can either directly use the output for inference (e.g., interpreting the 95% confidence interval as the range of the middle 95% of the respective posterior distribution, or the credible interval) or use Monte Carlo simulation based on classical sampling distribution theory to approximate the posterior distribution. In most applications, standard models in classical statistics are often inadequate. Nevertheless, using these standard models as a starting point can often lead to better model specification for subsequent Bayesian analysis.

1.3 Guiding Principles

Throughout the book, we describe the Bayesian analysis as part of the inference process of a specific environmental and ecological problem. As such, when we use the term "Bayesian analysis," we imply that it is part of the larger scientific problem, of which we aim to understand the underlying environmental/ecological processes and develop models to describe them. This emphasis takes the Bayesian analysis beyond the simple application of Bayes' theorem to derive the posterior distribution. In this process, we need to develop a procedure of inquiry of the most appropriate model in a study, and this should be guided by the three basic problems of statistics: specification, estimation, and distribution [Fisher, 1922]. Both classical and Bayesian statistics must address these three problems; however, classical biostatistic studies typically address these problems in a linear fashion. In contrast, many environmental and ecological studies require feedback among these three basic problems, which we refer to moving forward as model formulation, parameter estimation, and model evaluation (Figure 1.5). At each step, we have feedback on the model itself, which we argue is particularly suited for a Bayesian approach.

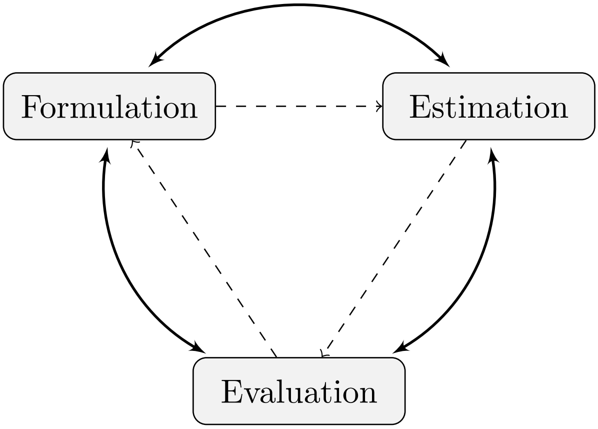

FIGURE 1.5: Bayesian analysis is an iterative process among the three problems of a statistical analysis (solid arrows), while the classical statistics inference is largely a linear process (dashed arrows).

To find the appropriate model for a study requires a process of proposing and evaluating alternative models. Comparing multiple alternative models was recognized long ago as an effective method to guard against personal and professional bias. In Chamberlin's words, proposing only a single model can “menace the integrity of the intellectual processes” because the single model quickly instills “parental affections” in the mind of the researcher [Chamberlin, 1890]. Chamberlin advocated the method of “multiple working hypotheses.” The disadvantage of this approach is the burden of proposing alternative models, especially when the first model that springs into our mind is always the preferred model. In Chamberlin's words: “we cannot put into words more than a single line of thought at the same time; and even in that the order of expression must be conformed to the idiosyncrasies of the language, and the rate must be relatively slow.” We emphasize that proposing the right model cannot be isolated from addressing the questions of parameter estimation and model evaluation. As a result, a Bayesian analysis is characterized by the iterative interactions among the three questions, rather than the linear process of moving from model formulation to parameter estimation to model evaluation. Throughout the book, we use examples to emphasize the steps of finding the most appropriate model.

1.4 Examples

In this book, we emphasize practical applications of Bayesian statistics with examples from several large studies. In this section, we introduce some of the examples that appear in multiple chapters. We used materials from these studies in multiple chapters because of the complicated nature of these problems. In this section, we summarize the scientific background of these examples.

1.4.1 Gill-net Monitoring Data

Three of the focal studies in this book involve fisheries-related research, including from studies of populations of Lake Erie walleye (*Sander vitreus*) [DuFour et al., 2019, 2021] and Hudson River Atlantic sturgeon (*Acipenser oxyrinchus).* A common feature of these studies is the use of gill nets as sampling gear to collect target fish in study areas. Gill nets are passive sampling gear, as they are placed in a designated location underwater and capture fish that swim into them. Panels of the net are attached on the top to floaters and on the bottom to weights so that the net fishes vertically, and can be set on the bottom or suspended in the water column. A gill net is set in a straight line and is characterized by its mesh size. A gill net is typically set (or soaked) in water for a predetermined period of time. The net is then retrieved and the number of fish caught in the net is counted, identified to the species level, and fish length and weight are measured. The number of fish caught in the net divided by the time the net was deployed is called the catch per unit effort (CPUE). In many fisheries studies, CPUE is used as an index of the population abundance. Fisheries management agencies carry out standard surveys to routinely monitor fish population. Standardized surveys use consistent sampling protocols to allow comparison of survey results through time. As a result, changes in CPUE over time are assumed to indicate changes in fish population.

In many fisheries studies, CPUE is usually treated as a continuous response variable in analysis, implying that the population is proportional to CPUE.

Because our interests are typically in the population, the process linking the population to CPUE (the data-generating process) can invalidate this implicit assumption, due to circumstances of data collection, spatial and temporal scales represented by the data, uneven fish distribution, and environmental influences. Specifically, we are interested in quantifying the proportionality of CPUE and the underlying population, and how the proportion constant changes spatially and under different environmental conditions. To accomplish the goal, the Lake Erie walleye project used a paired hydroacoustic survey to better quantify the number of fish in the vicinity of each gill net. We used the hydroacoustic estimated number of walleyes as a surrogate of the population of fish available to the passive gill net. In doing so, DuFour et al. [2019] showed that the gill net's catchability (measured as the ratio of CPUE and the underlying population size) varies among three distinct regions of Lake Erie. By sharing information across regional boundaries to improve region-specific estimates of survey gear efficiency (Chapter 6), we can improve the overall accuracy of the estimated population trends. In addition, the Lake Erie walleye example also demonstrates the advantage of using the Bayesian approach for quantifying uncertainty and its propagation in estimating intermediate variables (Chapter 2). We use the Atlantic sturgeon example to explicitly account for the processes of resulting in a zero count in the model to better estimate the relative abundance indices (Chapter 5).

1.4.2 Effects of Urbanization on Stream Ecosystems

The topical study on the Effects of Urbanization on Stream Ecosystems (EUSE) was part of U.S. Geological Survey's (USGS) National Water Quality Assessment Program (NAWQA). The study was designed to mimic a typical ecological experiment to study the effect of a dose-response effect of a treatment, which is the level of urbanization. As the levels of urbanization cannot be manipulated and applied to selected watersheds, the EUSE study selected nine metropolitan areas (or regions) across the continental U.S. These metropolitan regions (Atlanta, Georgia (ATL); Boston, Massachusetts (BOS); Birmingham, Alabama (BIR); Denver, Colorado (DEN); Dallas-Fort Worth, Texas (DFW); Milwaukee-Green Bay, Wisconsin (MGB); Portland, Oregon (POR); Raleigh, North Carolina (RAL); and Salt Lake City, Utah (SLC)) represent different environmental settings.

Within each region, 30 watersheds were identified based on a multimetric national urban intensity index (NUII) to represent gradients of urbanization within relatively homogeneous environmental settings [McMahon and Cuffney, 2000, Cuffney and Falcone, 2008]. These watersheds are similar in all other aspects except their levels of urbanization and data from these watersheds were collected using the same sampling protocol. The sampling design was guided by the principle of a randomized experiment for causal inference, even though the treatment (levels of urbanization) cannot be randomly assigned.

The intended analysis method was regression: modeling of various watershed-level indicators calculated based on observed biological data as functions of the watershed urbanization level. The biological data are counts of individuals of various species representing aquatic biota (fish, invertebrates, and algae) in samples collected using the same sampling protocol throughout the EUSE studies. For example, benthic macroinvertebrate communities are widely used to represent stream ecological conditions. Species in these communities are relatively long-lived (compared to algae), to integrate the temporal changes in water quality, and are of limited mobility (compared to fish), to reflect the impact of activities in the immediate upstream watershed.

Ecologists often use univariate metrics, instead of the counts of each species, to describe a community [Barbour and Paul, 2010, Barbour et al., 1999, e.g.]. For macroinvertebrates and algae, the total number of individuals counted is typically limited to a manageable value (e.g., 300 or 500 individuals). As a result, many of the metrics are based on the relative abundances of individual species. These univariate metrics are often used as response variables in regression analysis, with predictors represented by various water chemistry variables and physical habitat variables.

Many authors have examined metrics representing the rate and form of biological responses. Through regression analysis, researchers with the EUSE project identified watershed characteristics most strongly associated with biological responses and compared responses among urban areas [Coles et al., 2004, Cuffney et al., 2005, Brown et al., 2009, Cuffney et al., 2010]. As mentioned above, the biological responses used in these studies were univariate biological metrics calculated based on the observed counts of individual species of the target ecological community. For example, the metric Cuffney et al. [2005] used for measuring the level of pollution in an aquatic environment is the average tolerance level of benthic macroinvertebrates, a richness-weighted average of individual taxa tolerance values (RichTOL). Later, RichTOL was used in Qian et al. [2010] to illustrate the multilevel modeling approach to quantify the region-specific effects of the main treatment (watershed-level urbanization), as well as to explore the regional-level factors (e.g., annual average temperature and precipitation) that influence the urbanization effect (the regression slope). The multilevel modeling approach was further detailed in Qian [2016].

The basic form of the macroinvertebrate community data is counts of various species (taxa) from stream segments near the outlets of respective watersheds. These taxa counts are commonly used to derive univariate indicators or metrics to represent different features of the community. We use univariate indicators (in Chapter 6) to discuss the multilevel model, especially the multivariate normal priors for a Bayesian hierarchical model. We also directly use the observed count data to discuss two types of models: (1) the zero-inflated Poisson model and how such models can be used in a hierarchical model and (2) the multinomial regression model. Through these applications, we hope to show the multi-faceted nature of the study.

Although the EUSE studies were designed to mimic a randomized experiment, the data collected from these studies are largely observational. For example, there were differences in land use patterns and weather conditions among the nine metropolitan regions, which make the comparisons across the regions tentative. In all metropolitan areas except RAL and SLC, the distribution of levels of urbanization in sampling watersheds was skewed toward the lower end of the urban gradient because the number of streams available for study dropped precipitously at higher levels of urbanization (Figure 1.6). DEN, DFW, and MGB have high levels of antecedent agricultural land use compared with the other six metropolitan areas and are known to respond to urban development differently [Qian et al., 2010]. SLC differs from the other metropolitan areas by having development that has progressed from the valley floor up the Wasatch Mountains to the east of Salt Lake City. The upper limit of development is determined by the water supply infrastructure resulting in a sudden jump in urban intensity (i.e., ca. 0 to 20% developed land) as the water supply boundary is crossed. As a result, averages of % developed land in RAL and SLC are higher than the averages in the other four low antecedent agriculture regions [Qian and Cuffney, 2014].

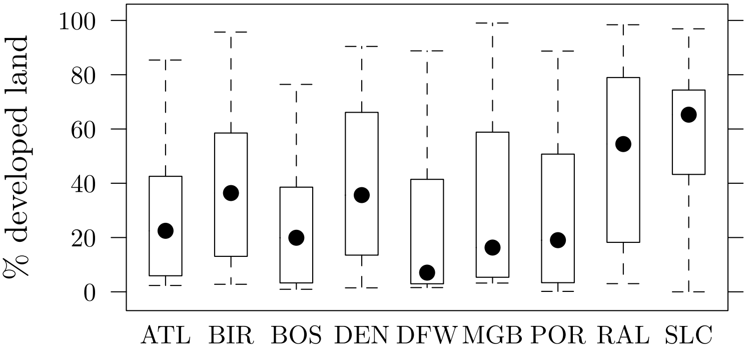

FIGURE 1.6: Watershed level % developed land distribution in the nine metropolitan regions.

1.4.3 Everglades Studies

Data from various aspects of the Everglades research are used in this book. These research activities were prompted, in large part, by a series of legal actions focused on the protection of the Everglades ecosystem in south Florida. Qian and Lavine [2003] summarize these legal actions and the subsequent research activities supporting the establishment of an environmental standard for phosphorus in the Everglades. In this section, we summarize some of these studies.

Legal actions related to the Everglades environmental protection started in the late 1980s. The current Everglades is largely represented by the Everglades National Park in the southern tip of the Florida peninsula. Over one hundred years ago, the Everglades was nearly one million hectares, covering almost the entire area south of Lake Okeechobee [Davis, 1943]. In the 1940s, settlers drained a small portion of the land for agriculture. Through the U.S. federal project Central and Southern Project for Flood Control and Other Purposes of 1948, the Everglades was systematically drained through the establishment of a system of canals, pumping stations, water storage areas, and levees [Light and Dineen, 1994]. As a result, large agricultural tracts were established within the Everglades, south of Lake Okeechobee, leading to the increased input of nutrient enriched agriculture runoff to the remaining Everglades [Snyder and Davidson, 1994]. The Everglades is a historically phosphorus-limited freshwater wetland ecosystem [Steward and Ornes, 1975b,a, Swift and Nicholas, 1987, Flora et al., 1988, Richardson et al., 1997]. The intense agriculture activity in the area resulted in an increased supply of phosphorus to the wetlands causing changes to its ecosystem, including major alterations in the water chemistry and elevated phosphorus concentrations in soils, extensive shifts in algal species, and altered community structure in areas with high and moderate phosphorus enrichment.

In a legal action brought by the federal government in 1988, the U.S. Department of Justice accused the state of Florida (represented by the South Florida Water Management District and the then Florida Department of Environmental Regulation) of violations of state water quality standards, particularly phosphorus, in the Loxahatchee National Wildlife Refuge and the Everglades National Park. In the settlement reached in 1991, the state of Florida recognized the severe harm to the Everglades National Park to the south and the Loxahatchee Wildlife Refuge to the east of the Everglades Agriculture Area. The 1992 consent decree of the settlement commits all parties to achieve the water quality and quantity needed to preserve and restore the unique flora and fauna of the Park and Refuge and to require agricultural growers to use best management practices to control and cleanse discharges from the Everglades Agricultural Area.

The 1991 settlement agreement was superseded by the 1994 Everglades Forever Act (EFA), requiring compliance with all water quality standards in the entire Everglades by December 31, 2006. The EFA authorized the Everglades Construction Project, including schedules for construction and operation of six storm water treatment areas to remove phosphorus from the EAA runoff. The EFA created a research program to understand phosphorus impacts on the Everglades and to develop additional treatment technologies. Finally, the EFA required the Florida Department of Environmental Protection (FDEP) to establish a numeric criterion for phosphorus.

To accomplish the task of setting environmental criterion for phosphorus, several studies focused on how the Everglades ecosystem responded to the elevated phosphorus input through both observational and experimental studies. A prominent mesocosm experimental study was carried out by the Duke University Wetland Center (DUWC) in the interior of a large water conservation area (the Water Conservation Area – 2A, or WCA2A), where the remaining wetland ecosystem is known not to be impacted by the agricultural runoff. In this experiment, six experimental flumes (10m × 2m channels) were constructed. At the upstream end, an automatic feeding machine pumped a continuous stream of water into the flume. Five of the channels were dosed with different levels of phosphorus, the sixth was a control. The purpose of the channels was to create phosphorus concentration gradients. Total phosphorus (TP) concentrations were measured biweekly at each meter mark along each channel for six years (1992–1998). The mean TP concentrations in those channels ranged from 10 to 75 μg/L. After the system stabilized, biological samples were collected regularly from 1995 to 1998. These biological samples were used to derive attributes representing biological responses at several trophic levels. Finally, an analysis was done to see what level of TP resulted in significant ecological change.

Data from DUWC's mesocosm study (known as the dosing study) were used to estimate the Everglades ecosystem's TP thresholds. The concept of an ecological threshold is based on the ecological concept of ecosystem resilience and alternative stable states. An ecosystem and its functions are resilient to disturbance when such disturbance is within certain limits. Because of the resilience, indicators of the ecosystem are often stable even when the disturbance increases. In ecological studies, we can observe the behavior of several similar ecosystems with different levels of disturbances (observational study), or create small replicas of the target ecosystems and apply different levels of disturbance (experimental study). The varying levels of disturbance (in this case, the disturbance is TP) is often known as the disturbance gradient. A threshold is a point along a disturbance gradient across which an ecosystem (or certain aspects of the ecosystem) changes abruptly from one stable state to another. The six experimental flumes created a TP gradient both within and among the flumes. In the DUWC's Bayesian analysis, the sequence of the observed ecological response variable is ordered along the TP gradient. In Chapter 7 we use data from the dosing study as part of the change point model examples. In Chapter 6, the separately estimated ecological thresholds for several ecological indicators are combined to derive an ecosystem-level threshold.

The task of establishing treatment wetlands for removing phosphorus requires an understanding of the effectiveness of a constructed wetland in removing phosphorus. Observation data from WCA2A were used in several studies to study the effectiveness of WCA2A as a treatment wetland. Because only a portion of WCA2A was affected by the agricultural runoff, historical soil TP in the wetland was of interest. Knowing the historical (or background)

TP levels, we can estimate the size of WCA2A that is impacted by agricultural runoff [Qian, 1997, e.g.,]. Using the changes in algal species (diatom) composition along the TP gradient in present-day Everglades, we illustrate the multinomial missing data problem in Chapter 5. Qian and Richardson [1997] used the ecological threshold concept to propose a separate model for estimating a wetland's phosphorus retention capacity. We used the data from Qian and Richardson [1997] as an example of the Gibbs sampler in Chapter 2.

1.4.4 Compliance Assessment under the U.S. Clean Water Act

The U.S. Clean Water Act (CWA, 33 U.S.C. §1251 et seq. (1972)) requires that states in the U.S. periodically submit a list of impaired waters, that is, waters that are "too polluted or otherwise degraded to meet water quality standards." This requirement is part of Section 303(d) of CWA, and the process of compiling the list is known as 303(d) listing. Once a water is listed, CWA further requires that the state develop total maximum daily load (TMDL) programs for mitigating the impairment. With a TMDL program implemented, monitoring is usually the basis for assessing compliance and determining if any management modifications are needed. The U.S. Environmental Protection Agency is responsible for developing rules and regulations to implement the law including setting specific standards for compliance assessment. Enforcement actions will take effect once a noncompliance is identified. Consequently, compliance assessment is often the important first step of environmental management.

As inference based on monitoring data is associated with errors because of, for example, measurement uncertainty, sampling error, seasonal and other periodicity caused auto-correlation, statistical analysis is inevitably an important consideration in compliance assessment based on monitoring data. In the U.S., most states developed procedures of standard compliance assessment based on statistical null hypothesis testing, mostly comparing an upper percentile (e.g., 80–90%) to the established environmental standard [Keller and Cavallaro, 2008]. Qian and Miltner [2018] discussed legal and management background of the environmental standard compliance assessment in the U.S., including the legal definition of a water quality standard as the mean concentration of a pollutant, the varying practices among states, and the "magnitude, duration, and frequency (MDF)" components of a water quality standard. They concluded that the practice of using hypothesis testing, common in nearly all states, was designed to address statistical uncertainty (sampling and measurement error). The MDF components of a water quality standard define (1) the harmful level of a pollutant (magnitude) determined by a toxicity study for toxic pollutants and by reference conditions for non-toxic pollutants such as nutrients, (2) the assessment period (duration) for which the mean concentration is estimated (e.g., annual mean concentration), and

(3) a recurring probability (frequency) which defines how to address the estimation uncertainty in the estimated annual mean [Qian, 2015]. The frequency component is the most confusing aspect of the MDF components. Qian [2015] suggested that the frequency component defines an upper quantile of the sampling distribution of the estimated mean. In subsequent chapters, we use the topic to discuss the advantage of a Bayesian estimation-oriented approach in compliance assessment.

1.5 Summary

Although we emphasize the two modes of inference in Bayesian and classical statistics, both follow the hypothetical deduction approach as described by Fisher [Fisher, 1922]. The common starting point of statistical inference is a probability distribution assumption on the response variable. The assumed probability distribution includes parameter(s) that are potentially function(s) of predictor variable(s). This step establishes the hypothesis in the hypothetical deduction process. In classical statistics, model parameters are estimated using the maximum likelihood method, while the estimation uncertainty is represented by the sampling distributions of these parameters. In Bayesian statistics, model parameters are estimated via Bayes' theorem producing posterior distributions which simultaneously characterize estimation uncertainty. The oft-neglected third step of the hypothetical deduction process is model evaluation. In classical statistics, this step is based on various hypothesis testing procedures. In classical statistics, a model parameter is a fixed (but unknown) constant and the estimate is a random variable. Model evaluation relies on the sampling distribution of the estimated parameter – a probability distribution of all potential estimates based on random sampling of the same population. As such, the classical statistical inference relies on not just the estimated parameter, but all unrealized potential estimates. When using Bayesian statistics, a probability is not necessarily representing a long-run frequency. As a result, we can use a probability distribution to represent our uncertainty about a parameter. The prior and posterior distributions represent what we know and do not know before and after observing data, respectively. Instead of using hypothesis testing, we use predictive distribution in Bayesian statistics for model evaluation.

We emphasize the iterative nature of model development – model formulation, parameter estimation, and model evaluation are inter-dependent in an applied problem.

This is a book for practitioners. Consequently, we want to emphasize applications, especially the thought-process behind the examples we used throughout the book. Because there is only one equation in Bayesian statistics, the process of applying Bayesian statistics is inevitably centered around two questions. First, "What is the statistical model describing the response variable?"

A model is parameterized with one or more parameters. Once these parameters are identified, the second question is, “Do we have any prior knowledge about these parameters?” If we do, how can our knowledge be quantified in terms of a probability distribution of these parameters? Once we have the answers to these two questions, the remaining tasks of a Bayesian application are mostly related to computation – calculating the posterior distribution and making inference (e.g., through posterior predictive inference).