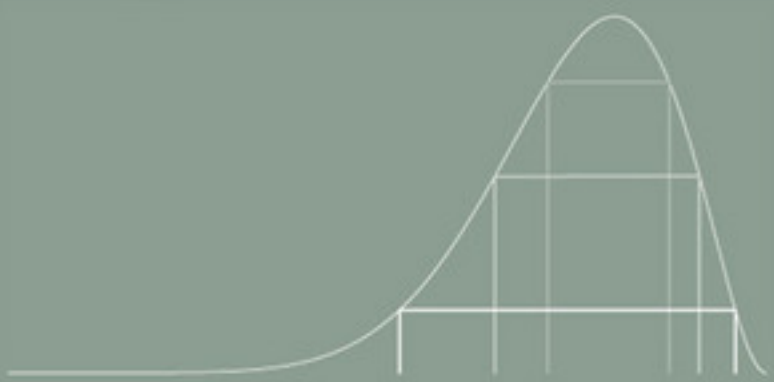
Springer Texts in Statistics

Peter D. Hoff

A First Course in Bayesian Statistical Methods



Springer

Springer Texts in Statistics

Series Editors:
G. Casella
S. Fienberg
I. Olkin

For other titles published in this series, go to
http://www.springer.com/series/417

Peter D. Hoff

A First Course in Bayesian Statistical Methods

Peter D. Hoff
Department of Statistics
University of Washington
Seattle WA 98195-4322
USA
hoff@stat.washington.edu

ISSN 1431-875X
ISBN 978-0-387-92299-7 e-ISBN 978-0-387-92407-6
DOI 10.1007/978-0-387-92407-6
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929120

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book originated from a set of lecture notes for a one-quarter graduate-level course taught at the University of Washington. The purpose of the course is to familiarize the students with the basic concepts of Bayesian theory and to quickly get them performing their own data analyses using Bayesian computational tools. The audience for this course includes non-statistics graduate students who did well in their department's graduate-level introductory statistics courses and who also have an interest in statistics. Additionally, first- and second-year statistics graduate students have found this course to be a useful introduction to statistical modeling. Like the course, this book is intended to be a self-contained and compact introduction to the main concepts of Bayesian theory and practice. By the end of the text, readers should have the ability to understand and implement the basic tools of Bayesian statistical methods for their own data analysis purposes. The text is not intended as a comprehensive handbook for advanced statistical researchers, although it is hoped that this latter category of readers could use this book as a quick introduction to Bayesian methods and as a preparation for more comprehensive and detailed studies.

Computing

Monte Carlo summaries of posterior distributions play an important role in the way data analyses are presented in this text. My experience has been that once a student understands the basic idea of posterior sampling, their data analyses quickly become more creative and meaningful, using relevant posterior predictive distributions and interesting functions of parameters. The open-source R statistical computing environment provides sufficient functionality to make Monte Carlo estimation very easy for a large number of statistical models, and example R-code is provided throughout the text. Much of the example code can be run "as is" in R, and essentially all of it can be run after downloading the relevant datasets from the companion website for this book.

Acknowledgments

The presentation of material in this book, and my teaching style in general, have been heavily influenced by the diverse set of students taking CSSS-STAT 564 at the University of Washington. My thanks to them for improving my teaching. I also thank Chris Hoffman, Vladimir Minin, Xiaoyue Niu and Marc Suchard for their extensive comments, suggestions and corrections for this book, and to Adrian Raftery for bibliographic suggestions. Finally, I thank my wife Jen for her patience and support.

Seattle, WA *Peter Hoff*
March 2009

Contents

1

Introduction and examples

1.1 Introduction

We often use probabilities informally to express our information and beliefs about unknown quantities. However, the use of probabilities to express information can be made formal: In a precise mathematical sense, it can be shown that probabilities can numerically represent a set of rational beliefs, that there is a relationship between probability and information, and that Bayes' rule provides a rational method for updating beliefs in light of new information. The process of inductive learning via Bayes' rule is referred to as *Bayesian inference*.

More generally, *Bayesian methods* are data analysis tools that are derived from the principles of Bayesian inference. In addition to their formal interpretation as a means of induction, Bayesian methods provide:

- parameter estimates with good statistical properties;
- parsimonious descriptions of observed data;
- predictions for missing data and forecasts of future data;
- a computational framework for model estimation, selection and validation.

Thus the uses of Bayesian methods go beyond the formal task of induction for which the methods are derived. Throughout this book we will explore the broad uses of Bayesian methods for a variety of inferential and statistical tasks. We begin in this chapter with an introduction to the basic ingredients of Bayesian learning, followed by some examples of the different ways in which Bayesian methods are used in practice.

Bayesian learning

Statistical induction is the process of learning about the general characteristics of a population from a subset of members of that population. Numerical values of population characteristics are typically expressed in terms of a parameter θ, and numerical descriptions of the subset make up a dataset y. Before a dataset

P.D. Hoff, *A First Course in Bayesian Statistical Methods*,
Springer Texts in Statistics, DOI 10.1007/978-0-387-92407-6_1,

is obtained, the numerical values of both the population characteristics and the dataset are uncertain. After a dataset y is obtained, the information it contains can be used to decrease our uncertainty about the population characteristics. Quantifying this change in uncertainty is the purpose of Bayesian inference.

The *sample space* $\mathcal{Y}$ is the set of all possible datasets, from which a single dataset y will result. The *parameter space* Θ is the set of possible parameter values, from which we hope to identify the value that best represents the true population characteristics. The idealized form of Bayesian learning begins with a numerical formulation of joint beliefs about y and θ, expressed in terms of probability distributions over $\mathcal{Y}$ and Θ.

1. For each numerical value $\theta \in \Theta$, our *prior distribution* $p(\theta)$ describes our belief that θ represents the true population characteristics.
2. For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, our *sampling model* $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true.

Once we obtain the data y, the last step is to update our beliefs about θ:

3. For each numerical value of $\theta \in \Theta$, our *posterior distribution* $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y.

The posterior distribution is obtained from the prior distribution and sampling model via *Bayes' rule*:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}}\,.$$

It is important to note that Bayes' rule does not tell us what our beliefs should be, it tells us how they should change after seeing new information.

1.2 Why Bayes?

Mathematical results of Cox (1946, 1961) and Savage (1954, 1972) prove that if $p(\theta)$ and $p(y|\theta)$ represent a rational person's beliefs, then Bayes' rule is an optimal method of updating this person's beliefs about θ given new information y. These results give a strong theoretical justification for the use of Bayes' rule as a method of quantitative learning. However, in practical data analysis situations it can be hard to precisely mathematically formulate what our prior beliefs are, and so $p(\theta)$ is often chosen in a somewhat ad hoc manner or for reasons of computational convenience. What then is the justification of Bayesian data analysis?

A famous quote about sampling models is that "all models are wrong, but some are useful" (Box and Draper, 1987, pg. 424). Similarly, $p(\theta)$ might be viewed as "wrong" if it does not accurately represent our prior beliefs. However, this does not mean that $p(\theta|y)$ is not useful. If $p(\theta)$ approximates our beliefs, then the fact that $p(\theta|y)$ is optimal under $p(\theta)$ means that it will also

generally serve as a good approximation to what our posterior beliefs should be. In other situations it may not be *our* beliefs that are of interest. Rather, we may want to use Bayes' rule to explore how the data would update the beliefs of a variety of people with differing prior opinions. Of particular interest might be the posterior beliefs of someone with weak prior information. This has motivated the use of "diffuse" prior distributions, which assign probability more or less evenly over large regions of the parameter space.

Finally, in many complicated statistical problems there are no obvious non-Bayesian methods of estimation or inference. In these situations, Bayes' rule can be used to generate estimation procedures, and the performance of these procedures can be evaluated using non-Bayesian criteria. In many cases it has been shown that Bayesian or approximately Bayesian procedures work very well, even for non-Bayesian purposes.

The next two examples are intended to show how Bayesian inference, using prior distributions that may only roughly represent our or someone else's prior beliefs, can be broadly useful for statistical inference. Most of the mathematical details of the calculations are left for later chapters.

1.2.1 Estimating the probability of a rare event

Suppose we are interested in the prevalence of an infectious disease in a small city. The higher the prevalence, the more public health precautions we would recommend be put into place. A small random sample of 20 individuals from the city will be checked for infection.

Parameter and sample spaces

Interest is in θ, the fraction of infected individuals in the city. Roughly speaking, the parameter space includes all numbers between zero and one. The data y records the total number of people in the sample who are infected. The parameter and sample spaces are then as follows:

$$\Theta = [0, 1] \qquad \mathcal{Y} = \{0, 1, \ldots, 20\}\,.$$

Sampling model

Before the sample is obtained the number of infected individuals in the sample is unknown. We let the variable Y denote this to-be-determined value. If the value of θ were known, a reasonable sampling model for Y would be a binomial$(20, \theta)$ probability distribution:

$$Y|\theta \sim \text{binomial}(20, \theta)\,.$$

The first panel of Figure 1.1 plots the binomial$(20, \theta)$ distribution for θ equal to 0.05, 0.10 and 0.20. If, for example, the true infection rate is 0.05, then the probability that there will be zero infected individuals in the sample $(Y = 0)$ is 36%. If the true rate is 0.10 or 0.20, then the probabilities that $Y = 0$ are 12% and 1%, respectively.

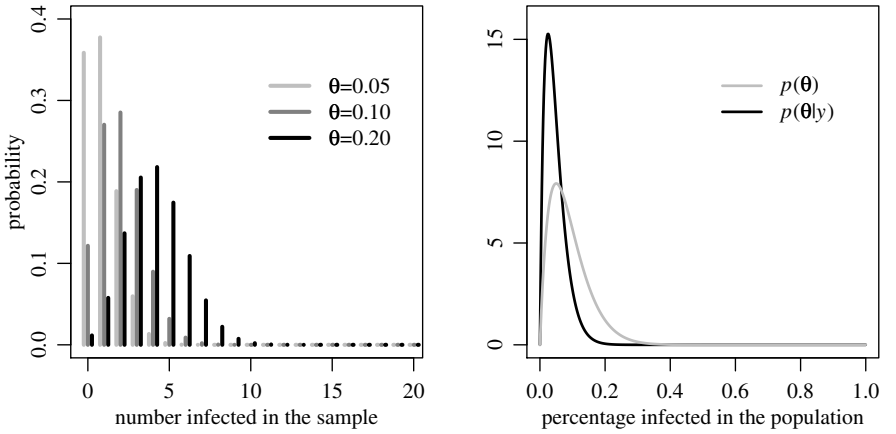


Fig. 1.1. Sampling model, prior and posterior distributions for the infection rate example. The plot on the left-hand side gives binomial$(20, \theta)$ distributions for three values of θ. The right-hand side gives prior (gray) and posterior (black) densities of θ.

Prior distribution

Other studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalence of 0.10. This prior information suggests that we use a prior distribution $p(\theta)$ that assigns a substantial amount of probability to the interval (0.05, 0.20), and that the expected value of θ under $p(\theta)$ is close to 0.10. However, there are infinitely many probability distributions that satisfy these conditions, and it is not clear that we can discriminate among them with our limited amount of prior information. We will therefore use a prior distribution $p(\theta)$ that has the characteristics described above, but whose particular mathematical form is chosen for reasons of computational convenience. Specifically, we will encode the prior information using a member of the family of beta distributions. A beta distribution has two parameters which we denote as a and b. If θ has a beta(a, b) distribution, then the expectation of θ is $a/(a + b)$ and the most probable value of θ is $(a - 1)/(a - 1 + b - 1)$. For our problem where θ is the infection rate, we will represent our prior information about θ with a beta(2,20) probability distribution. Symbolically, we write

$$\theta \sim \text{beta}(2, 20).$$

This distribution is shown in the gray line in the second panel of Figure 1.1. The expected value of θ for this prior distribution is 0.09. The curve of the prior distribution is highest at $\theta = 0.05$ and about two-thirds of the area under the curve occurs between 0.05 and 0.20. The prior probability that the infection rate is below 0.10 is 64%.

$$\begin{aligned} \mathrm{E}[\theta] &= 0.09 \\ \mathrm{mode}[\theta] &= 0.05 \\ \Pr(\theta < 0.10) &= 0.64 \\ \Pr(0.05 < \theta < 0.20) &= 0.66\,. \end{aligned}$$

Posterior distribution

As we will see in Chapter 3, if $Y|\theta \sim \text{binomial}(n, \theta)$ and $\theta \sim \text{beta}(a, b)$, then if we observe a numeric value y of Y, the posterior distribution is a $\text{beta}(a+y, b+n-y)$ distribution. Suppose that for our study a value of $Y = 0$ is observed, i.e. none of the sample individuals are infected. The posterior distribution of θ is then a $\text{beta}(2, 40)$ distribution.

$$\theta|\{Y = 0\} \sim \text{beta}(2, 40)$$

The density of this distribution is given by the black line in the second panel of Figure 1.1. This density is further to the left than the prior distribution, and more peaked as well. It is to the left of $p(\theta)$ because the observation that $Y = 0$ provides evidence of a low value of θ. It is more peaked than $p(\theta)$ because it combines information from the data and the prior distribution, and thus contains more information than in $p(\theta)$ alone. The peak of this curve is at 0.025 and the posterior expectation of θ is 0.048. The posterior probability that $\theta < 0.10$ is 93%.

$$\begin{aligned} \mathrm{E}[\theta|Y = 0] &= 0.048 \\ \mathrm{mode}[\theta|Y = 0] &= 0.025 \\ \Pr(\theta < 0.10|Y = 0) &= 0.93. \end{aligned}$$

The posterior distribution $p(\theta|Y = 0)$ provides us with a model for learning about the city-wide infection rate θ. From a theoretical perspective, a rational individual whose prior beliefs about θ were represented by a beta(2,20) distribution now has beliefs that are represented by a beta(2,40) distribution. As a practical matter, if we accept the beta(2,20) distribution as a reasonable measure of prior information, then we accept the beta(2,40) distribution as a reasonable measure of posterior information.

Sensitivity analysis

Suppose we are to discuss the results of the survey with a group of city health officials. A discussion of the implications of our study among a diverse group of people might benefit from a description of the posterior beliefs corresponding to a variety of prior distributions. Suppose we were to consider beliefs represented by $\text{beta}(a, b)$ distributions for values of (a, b) other than (2,20). As mentioned above, if $\theta \sim \text{beta}(a, b)$, then given $Y = y$ the posterior distribution of θ is $\text{beta}(a + y, b + n - y)$. The posterior expectation is

$$\begin{aligned}\mathrm{E}[\theta|Y=y] &= \frac{a+y}{a+b+n} \\ &= \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} \\ &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0,\end{aligned}$$

where $\theta_0 = a/(a+b)$ is the prior expectation of θ and $w = a+b$. From this formula we see that the posterior expectation is a weighted average of the sample mean $\bar{y}$ and the prior expectation θ_0. In terms of estimating θ, θ_0 represents our prior guess at the true value of θ and w represents our confidence in this guess, expressed on the same scale as the sample size. If

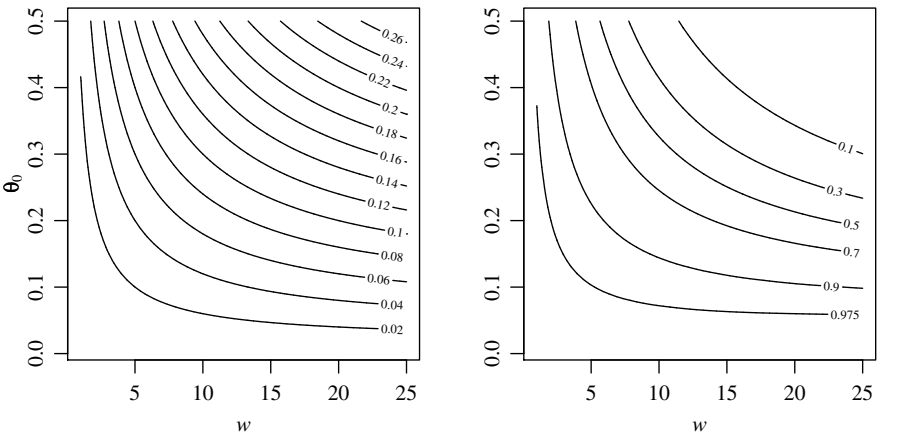


Fig. 1.2. Posterior quantities under different beta prior distributions. The left- and right-hand panels give contours of $\mathrm{E}[\theta|Y=0]$ and $\Pr(\theta < 0.10|Y=0)$, respectively, for a range of prior expectations and levels of confidence.

someone provides us with a prior guess θ_0 and a degree of confidence w, then we can approximate their prior beliefs about θ with a beta distribution having parameters $a = w\theta_0$ and $b = w(1-\theta_0)$. Their approximate posterior beliefs are then represented with a beta$(w\theta_0 + y, w(1-\theta_0) + n - y)$ distribution. We can compute such a posterior distribution for a wide range of θ_0 and w values to perform a *sensitivity analysis*, an exploration of how posterior information is affected by differences in prior opinion. Figure 1.2 explores the effects of θ_0 and w on the posterior distribution via contour plots of two posterior quantities. The first plot gives contours of the posterior expectation $\mathrm{E}[\theta|Y=0]$, and the second gives the posterior probabilities $\Pr(\theta < 0.10|Y=0)$. This latter plot may be of use if, for instance, the city officials would like to recommend a vaccine to the general public unless they were reasonably sure that the current infection rate was less than 0.10. The plot indicates, for example, that