Doing Bayesian Data Analysis: A Tutorial with R and BUGS

John K. Kruschke

Draft of May 11, 2010. Please do not circulate this preliminary draft. If you report Bayesian analyses based on this book, please do cite it!

Dedicated to my mother, Marilyn A. Kruschke,
and to the memory of my father, Earl R. Kruschke,
who both brilliantly exemplified and taught sound reasoning.
And, in honor of my father,
who dedicated his first book to his children,
I also dedicate this book to mine:
Claire A. Kruschke and Loren D. Kruschke

Contents

Chapter 1

This Book's Organization: Read Me First!

Contents

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

1.1 Real people can read this book

This book explains how to actually *do* Bayesian data analysis, by real people (like you), for realistic data (like yours). The book starts at the basics, with notions of probability and programming, then progresses to advanced hierarchical models that are used in realistic data analysis. In other words, you do not need to already know statistics and programming. This book is speaking to a first-year graduate student or advanced undergraduate in the social or biological sciences: Someone who grew up in Lake Wobegon[1], but who is not the mythical being that has the previous training of a nuclear physicist and then decided to learn about Bayesian statistics.

This book provides broad coverage and ease of access. Section 1.3 describes the contents in a bit more detail, but here are some highlights. This book covers Bayesian analogues

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

of all the traditional statistical tests that are presented in introductory statistics textbooks, including *t*-tests, analysis of variance (ANOVA), regression, chi-square tests, and so on. This book also covers crucial issues for *designing* research, such as statistical power and methods for determining the sample size needed to achieve a desired research goal. And you don't need to already know statistics to read this book, which starts at the beginning, including introductory chapters about concepts of probability, and an entire chapter devoted to Bayes' rule. The important concept of hierarchical modeling is introduced with unique simple examples, and the crucial methods of Markov chain Monte Carlo sampling are explained at length, starting with simple examples that, again, are unique to this book. Computer programs are thoroughly explained throughout the book, and listed in their entirety, so you can use and adapt them to your own needs.

But wait, there's more. As you may have noticed from the beginning of this chapter, the chapters commence with a stanza of elegant and insightful verse composed by a famous poet. The quatrains[2] are formed of dactylic[3] tetrameter[4], or, colloquially speaking, "country waltz" meter. The poems regard conceptual themes of the chapter via allusion from immortal human motifs often expressed by country western song lyrics, all in waltz timing.
If you do not find them to be all that funny,
if they leave you wanting back all of your money,
well honey some waltzing's a small price to pay,
for all the good learning you'll get if you stay.

1.2 Prerequisites

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

The beginnings of your journey will go more smoothly if you have had some basic experience programming a computer, but previous programming experience is not crucial. A computer program is just a list of commands that the computer can execute. For example, if you've ever typed an "=" in an Excel spreadsheet cell, you've written a programming command. If you've ever written a list of commands in Basic or Pascal or Java or any other language, then you're set. We will be using a language called R, which is *free*. More on R later.

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

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

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

- Section 2.3 introduces R.
- Chapter 4 explains Bayes' rule.
- Chapter 7 explains Markov chain Monte Carlo methods.
- Section 7.4 introduces BUGS.
- Chapter 9 explains hierarchical models.
- Chapter 13 explains varieties of power analysis.
- Chapter 14 overviews the generalized linear model and various types of data analyses that can be conducted.
- Section 23.1 summarizes how to report a Bayesian data analysis.

Figure 1.1: Essential sections of the book.

1.3 The organization of this book

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

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

The third main part of the book applies the Bayesian methods to realistic data. The applications are organized around the type of data being analyzed, and the type of measurements that are used to explain or predict the data. For example, suppose you are trying to predict college grade point average (GPA) from high school Scholastic Aptitude Test (SAT) score. In this case the data to be predicted, the GPAs, are values on a *metric* scale, and the predictor, the SAT scores, are also values on a *metric* scale. Suppose, on the other hand, that you are trying to predict college GPA from gender. In this case the predictor is a *dichotomous* value, namely, male vs. female. Different types of measurement scales require different types of mathematical models, but otherwise the underlying concepts are always the same. Table 14.1 (p. 312) shows various combinations of measurement scales and their corresponding models that are explored in detail in the third part of this book.

1.3.1 What are the essential chapters?

The foundations established in the first part of the book, and the Bayesian ideas of the second part, are important to understand. The applications to particular types of data, in the third part, can be more selectively perused as needed. Within those parts, however, there

are some chapters that are essential:

- Chapter 4 explains Bayes' rule.
- Chapter 7 explains Markov chain Monte Carlo methods.
- Chapter 9 explains hierarchical models.
- Chapter 14 overviews the generalized linear model and various types of data analyses that can be conducted.

As an emphasis of the book is *doing* Bayesian data analysis, it is also essential to learn the programming languages R and BUGS:

- Section 2.3 introduces R.
- Section 7.4 introduces BUGS.

Finally, the ultimate purpose of data analysis is to convince other people that their beliefs should be altered by the data. The results need to be communicated to a skeptical audience, and therefore additional essential reading is

- Section 23.1 summarizes how to report a Bayesian data analysis.

Another important topic is the planning of research, as opposed to the analysis of data after they have been collected. Bayesian techniques are especially nicely suited for estimating the probability that specified research goals can be achieved as a function of the sample size for the research. Therefore, although it might not be essential on a first reading, it is essential eventually to read

- Chapter 13 regarding power analysis.

Figure 1.1 puts these recommendations in a convenient reference box, re-arranged into the order of presentation in the book.

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

Because many readers will be coming to this book after having already been exposed to traditional 20th-century statistics that emphasize null hypothesis significance testing (NHST), this book will provide Bayesian approaches to the usual topics in NHST textbooks. Table 1.1 lists various tests covered by standard introductory statistics textbooks, along with their Bayesian analogues. If you have been previously contaminated by NHST, but want to know how to do an analogous Bayesian analysis, Table 1.1 may be useful.

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

Table 1.1: Bayesian analogues of 20th century null hypothesis significance tests.

Traditional Analysis Name	Bayesian Analogue
t-test for a single mean	Chapter 15
t-test for two independent groups	Chapter 18 (Section 18.3)
Simple linear regression	Chapter 16
Multiple linear regression	Chapter 17
Oneway ANOVA	Chapter 18
Multi-factor ANOVA	Chapter 19
Logistic regression	Chapter 20
Ordinal regression	Chapter 21
Binomial test	Chapters 5–9, 20
Chi-square test (contingency table)	Chapter 22
Power analysis (sample size planning)	Chapter 13

1.4 Gimme feedback (be polite)

I have worked thousands of hours on this book, and I want to make it better. If you have suggestions regarding any aspect of this book, please do e-mail me: `JohnKruschke@gmail.com`. Let me know if you've spotted egregious errors or innocuous infelicities, typo's or thoughto's. Let me know if you have a suggestion for how to clarify something. Especially let me know if you have a good example that would make things more interesting or relevant. I'm also interested in complete raw data from research that is interesting to a broad audience, and which can be used with acknowledgement but without fee. Let me know also if you have more elegant programming code than what I've cobbled together. The outside margins of these pages are intentionally made wide so that you have room to scribble your ridicule and epithets before re-phrasing them into kindly stated suggestions in your e-mail to me. Rhyming couplets are especially appreciated. If I don't respond to your e-mail in a timely manner, it is only because I can't keep up with the deluge of fan mail, not because I don't appreciate your input. Thank you in advance!

1.5 Acknowledgments

This book has been six years in the making, and many colleagues and students have provided helpful comments. The most extensive comments have come from Drs. Luiz Pessoa, Mike Kalish, Jerry Busemeyer, and Adam Krawitz; thank you all! Particular sections were insightfully improved by helpful comments from Drs. Michael Erickson, Robert Nosofsky, and Geoff Iverson. Various parts of the book benefitted indirectly from communications with Drs. Woojae Kim, Charles Liu, Eric-Jan Wagenmakers and Jeffrey Rouder. Leads to data sets were offered by Drs. Teresa Treat and Michael Trosset, among others. Very welcome supportive feedback was provided by Dr. Michael Lee, and also by Dr. Adele Diederich. A Bayesian-supportive working environment was provided by many colleagues including Drs. Richard Shiffrin, Jerome Busemeyer, Peter Todd, James Townsend, Robert Nosofsky, and Luiz Pessoa. Other department colleagues have been very supportive of integrating Bayesian statistics into the curriculum, including Drs. Linda Smith and Amy Holtzworth-Munroe. Various teaching assistants have provided helpful comments; in particular I especially thank Drs. Noah Silbert and Thomas Wisdom for their excellent as-

sistance. As this book has evolved over the years, suggestions have been contributed by numerous students, including Aaron Albin, Thomas Smith, Sean Matthews, Thomas Parr, Kenji Yoshida, Bryan Bergert, and perhaps dozens of others who have contributed insightful questions or comments that helped me tune the presentation in the book. To all the people who have made suggestions but whom I have inadvertently forgotten to mention by name, I extend my apologies and appreciation.

Part I

The Basics: Parameters, Probability, Bayes' Rule, and R

Chapter 2

Introduction: Models we believe in

Contents

I just want someone who I can believe in,
Someone at home who will not leave me grievin'.
Show me a sign that you'll always be true,
and I'll be your model of faith and virtue.

Inferential statistical methods help us decide what to believe in. With inferential statistics, we don't just introspect to find the truth. Instead, we rely on data from observations. Based on the data, what should we believe in? Should we believe that the tossed coin is fair if it comes up heads in 7 of 10 flips? Should we believe that we have cancer when the test comes back positive? Should we believe that she loves me when the daisy has 17 petals? Our beliefs can be modified when we have data, and this book is about techniques for making inferences *from* data *to* uncertain beliefs.

There might be some beliefs that cannot be decided by data, but such beliefs are dogmas that lie (double entendre intended) beyond the reach of evidence. If you are wondering about a belief that has no specific implications for concrete facts in the observable world, then inferential statistics won't help.

Why do we need hefty tomes full of mathematics to help us make decisions based on data? After all, we make lots of decisions every day without math. If we're driving, we look at the signal light and effortlessly decide whether it's red or green. We don't (consciously) go through a laborious process of mathematical statistics and finally conclude that it is probably the case that the light is red. There are two attributes of this situation that make the decision easy. First, the data about the light are numerous: An unobstructed view of the light results in a whole lot of photons striking our eyes. Second, there are only a few possible beliefs about the light, that make very distinct predictions about the photons: If the light is red, the photons are rather different than if the light is green. Consequently, the decision is easy because there is little variance in the data and little uncertainty across possible beliefs.

The math is most helpful when there is lots of variance in the data and lots of uncertainty in our beliefs. Data from scientific experiments, especially those involving humans or animals, are unmitigated heaps of variability. Theories in science tend to be rife with parameters of uncertain magnitude, and competing theories are numerous. In these situations, the mathematics of statistical inference provide precise numerical bounds on our uncertainty. The math allows us to determine accurately what the data imply for different possible beliefs. The math can tell us exactly how likely or unlikely each possibility is, even when there is an infinite spectrum of possibilities. It is this power, of precisely defining our uncertainty, that makes inferential statistics such a useful tool, worth the effort of learning.

2.1 Models of observations and models of beliefs

Suppose we flip a coin to decide which team kicks off. The teams agree to this decision procedure because they believe that the coin is fair. But how do we determine whether the coin really is fair? Even if we could study the exact minting process of the coin, and x-ray every nuance of the coin's interior, we would still need to test whether the coin really is fair when it's actually flipped. Ultimately, all we can do is flip the coin a few times and watch its behavior. From these observations we can modify our beliefs about the fairness of the coin.

Suppose we have a coin from our friend the numistmatist[1]. We notice that on the obverse is embossed the head of Tanit (of ancient Carthage), and on the reverse side is embossed a horse. The coin is gold and shows the date 350BCE. Do you believe that the coin is fair? Maybe you do, but maybe you're not very certain.[2] Let's flip it a few times. Suppose we flip it ten times and we obtain this sequence: HTTTTTHTTT. That's 2 heads and 8 tails. Now what do you think? Do you have a suspicion that maybe the coin is biased to come up tails more often than heads?

We've seen that the coin comes up horses a lot. Whoa! Let's dismount and have a heart-to-heart 'round the campfire. In that simple coin-flipping scenario we have made two sets of assumptions. First, we have assumed that the coin has some inherent fairness or bias, that we can't directly observe. All we can actually observe is an inherently probabilistic effect of that bias, namely, whether the coin comes up heads or tails on any given flip. We've made lots of assumptions about exactly how the observable head or tail relates to the unobservable

[1] *Numistmatist* [noun]: A person who studies or collects coins.

[2] A tale about coins marked BCE is a well-known joke because any coin actually minted BCE could not have been marked BCE at the time it was minted. But even a coin marked with a bogus date might be a fair flipper.

bias of the coin. For instance, we've assumed that the bias stays the same, flip after flip. We've assumed that the coin can't remember what it came up last flip, so that its flip this time is uncorrupted by its previous landings. All these assumptions are about the process that converts the unobservable bias into a probabilistic observable event. This collection of assumptions about the coin flipping process is our model of the head-tail observations.

The second set of assumptions is about our beliefs regarding the bias of the coin. We assume that we believe most strongly in the coin being fair, but we also allow for the possibility that the coin could be biased. Thus, we have a set of assumptions about how likely it is for the coin to be fair or to be biased to different amounts. This collection of assumptions is our model of our beliefs.

When we want to get specific about our model assumptions, then we have to use mathematical descriptions. A "formal" model uses mathematical formulas to precisely describe something. In this book, we'll almost always be using formal models, and so *whenever the term "model" comes up, you can assume it means a mathematical description*. In the context of statistical models, the models are typically models of probabilities. Some models describe the probabilities of observable events; e.g., we can have a formula that describes the probability that a coin will come up heads. Other models describe the extent to which we believe in various underlying possibilities; e.g., we can have a formula that describes how much we believe in each possible bias of the coin.

2.1.1 Models have parameters

Consider a model of the probability that it will rain at a particular location. This model is a formula that generates a numerical probability as its output. The probability of rain depends on many things, but in particular it might depend on elevation above sea level. Thus, the probability of rain, which is the output of the model, depends on the location's elevation, which is a value that is input to the model. The exact relationship between input and output could be modulated by another value that governs exactly how much the input affects the output. This modulating value is called a *parameter*. The model formula specifies that the input does affect the output, but the parameters govern exactly how much.

As another example, consider the probability that a coin comes up heads. We could model the probability of heads as a function of the lopsidedness of the coin. To measure lopsidedness, first consider slicing the coin like a bagel, exactly halfway between the head and tail faces. The lopsidedness is defined as the mass of the tail side minus the mass of the head side, measured in milligrams. Therefore, when lopsidedness is positive, the tail side is heavier, and heads are more likely to come up. Then we use a formula to convert from lopsidedness to probability of coming up heads. One such formula is graphed in the left panel of Figure 2.1. The S-shaped curve indicates that when the lopsidedness is zero, i.e., $x = 0$, then the coin is fair, i.e., the probability of the outcome being a head is 50%, which is written mathematically as $p(\text{datum} = H|x) = 0.50$. The S-shaped curve also shows that when the coin is positively lopsided, i.e., $x > 0$, then the coin is biased to come up heads more often, but when the coin is negatively lopsided, i.e. $x < 0$, then the coin is biased to come up heads less often.

The model, which gets us from a coin's lopsidedness to its probability of coming up heads, could have another variable that modulates the exact degree of bias stemming from a particular lopsidedness. For example, large-diameter coins might not be as affected by lopsidedness as small-diameter coins. The variable that modulates the exact probability bias is called a parameter, because its value is specified by the theorist rather than being

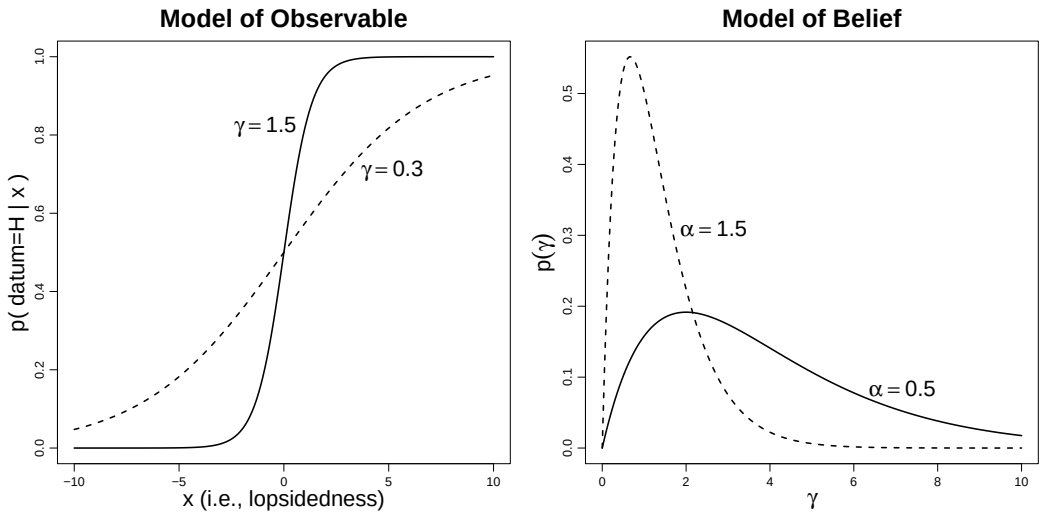


Figure 2.1: The left panel shows a model of the probability of an observable datum, i.e., a coin coming up heads, as a function of a measured input, x, the lopsidedness of a coin. The exact form of the S-shaped curve is governed by a parameter, γ. The graph shows the function for two different values of the parameter γ. The right panel shows a model of how much we believe in candidate values of γ. The exact form of the hill-shaped beliefs depends on the value of α, which is called a "hyperparameter."

measured from the world. Figure 2.1 shows the S-shaped curve for two different values of a parameter called γ (Greek letter gamma). When γ is small, then a little lopsidedness doesn't bias the coin much. When γ is large, then a little lopsidedness produces a big bias in the the coin.

We theorists might not know in advance exactly what value of the parameter to believe in, so we entertain a spectrum of possible values, each with a different degree of belief. The right panel of Figure 2.1 shows possible distributions of beliefs over the values of the parameter γ. This is a model of our beliefs. The exact shape of the hill-shaped belief distribution is governed by another parameter, this one called α (Greek letter alpha). The α parameter is in our model of beliefs about the parameter γ, and therefore the α parameter is called a *hyperparameter*. A hyperparameter is merely a parameter upon which other parameters depend. Notice in the figure that when α is large, we believe more in smallish values of γ, and when α is small, we believe in a broader spectrum of larger values of γ.

In summary, we can have a mathematical model of the probability that certain observable events happen. This mathematical model has parameters. The values of the parameters determine the exact probabilities generated by the model. Our beliefs regard the possible values of the parameters. We may believe strongly in some parameter values but less strongly in other values. The form of our beliefs about various parameter values can itself be expressed as a mathematical model, with its own (hyper-)parameters.

(Tangential aside for the philosophically inclined: The difference between a measured data value, such as lopsidedness $x = 1.2$, and a hypothetical parameter value, such as $\gamma = 1.2$, is not so clear cut as has been assumed above. On the one hand, whenever we make an observation, we are converting something in the world into a numerical scale value. This measurement process is itself a form of parameter estimation, because when we measure

the lopsidedness of a coin, we are estimating the value of the lopsidedness parameter that is most consistent with our mechanical interaction with the coin. On the other hand, when we estimate the value of a parameter in a model, we are essentially measuring the world in terms of that parameter scale.)

2.1.2 Prior and posterior beliefs

We could believe that the coin is fair, that is, that the probability of coming up heads is 50%. We could instead have other beliefs about the coin, especially if it's dated 350BCE, which no coin would be labeled if it were really minted BCE (because the people alive in 350BCE didn't yet know they were BCE). Perhaps, therefore, we also think it's possible for the coin to be biased to come up heads 20% of the time, or 80% of the time. Before observing the coin flips, we might believe that each of these three dispositions is equally likely, that is, we believe that there is a one-in-three chance that the bias is 20%, a one-in-three chance that the bias is 50%, and a one-in-three chance that the bias is 80%.

After flipping the coin and observing 2 heads in 10 flips, we will want to modify our beliefs. It makes sense that we should now believe more strongly that the bias is 20%, because we observed 20% heads in the sample. This book is about determining *exactly* how much more strongly we should believe that the bias is 20%.

Before observing the flips of the coin, we had certain beliefs about the possible biases of the coin. This is called a *prior* belief because it's our belief before taking into account some particular set of observations. After observing the flips of the coin, we had modified beliefs. These are called a *posterior* belief because they are computed after taking into account a particular set of observations. Bayesian inference gets us from prior to posterior beliefs.

There is an infelicity in the terms "prior" and "posterior," however. The terms connote the passage of time, as if the prior beliefs were held temporally before the posterior beliefs. But that is a misconception. There is no temporal ordering in the prior and posterior beliefs! Rather, the prior is simply the belief we hold by *excluding* a particular set of data, and the posterior is the belief we hold by *including* the set of data. Despite this misleading temporal connotation, the terms "prior" and "posterior" are firmly entrenched in the literature, so we'll use them too.

2.2 Three goals for inference from data

When we make observations of the world, we typically have one of three goals in mind. Each of these goals can be illustrated with the coin-flipping scenario.

2.2.1 Estimation of parameter values

One goal we may have is deciding to what extent we should believe in each of the possible parameter values. In the case of the coin, we used the observed data to determine the extent to which we should believe that the bias is 20%, 50%, or 80%. What we are determining is how much we believe in each of the available parameter values.

Because the flip of the coin is a random process, we cannot be certain of the underlying true probability of getting heads. So our posterior beliefs are an estimate. The posterior beliefs typically increase the magnitude of belief in some parameter values, while lessening the degree of belief in other parameter values. So this process of shifting our beliefs in various parameter values is called "estimation of parameter values."

2.2.2 Prediction of data values

Another goal we may have is predicting other data values, given our current beliefs about the world. For example, given that we have just observed the ball leaving the pitcher's hand and we now believe it's a curve ball, where do we predict the ball will be when it gets near the plate? Or, given that we only saw how the ball crossed over the plate, and from that we believe it was a curve ball, then what do we predict was the pitcher's grip on the ball as it was released?

Notice that "prediction" is another of those words that connotes temporal order but isn't always used that way in statistics. Prediction simply means inferring the values of some missing data based on some other included data, regardless of the actual temporal relationship of the included and missing data.

An ability to make specific predictions is one of the primary uses of mathematical models. Models help us predict the effectiveness of a flu vaccine when distributed to the general public. Models help us predict the paths of hurricanes. And models can help us predict whether the next coin flip will be heads or tails.

In Bayesian inference, to predict data values, we typically take a weighted average of our beliefs. We let each belief make its individual prediction, and then we weigh each of those predictions according to how strongly we believe in them. For example, if we believe strongly that the coin has a bias of 20% heads, and we only weakly believe in biases of 50% or 80%, then our prediction will be a mixture of the three beliefs weighted strongly toward 20%; perhaps, therefore, we predict a 30% chance of a head on the next flip.

2.2.3 Model comparison

A third goal of statistical inference is model selection, a.k.a. model comparison. If we have two different models of how something might happen, then an observation of what really does happen can influence which model we believe in most. What Bayesian inference tells us is how to shift our magnitude of belief across the available models.

As a somewhat contrived example, suppose we have two different models of the coin. One model assumes what we've described before, that the coin could have biases of 20%, 50%, or 80% heads. The second model assumes that the coin is either a perfectly fair coin or else it's a trick coin with two heads or two tails. This model allows the coin to biases of 0%, 50%, or 100% heads. Notice that the second model assumes different available parameter values than the first model.

After observing 10 flips that had 2 tails, which model do we believe in more? Let's think about the second model. Because our observations were not purely heads or purely tails, we know that the posterior beliefs for that model must load all belief on 50% heads, because we did not observe all heads or all tails. This model then is stuck asserting that the 10 observed flips with just 2 tails were generated by a fair coin, which is not very likely. The first model, on the other hand, has the belief of 20% heads available to it, which can generate the observed data with high likelihood. Therefore we should believe the first model more strongly than the second. The mathematics of Bayesian inference can tell *exactly* how much more to believe the first model than the second.

One of the nice qualities of Bayesian model comparison is that it intrinsically adjusts for model complexity. More complex models will fit data better than simple models, merely because the complex models have more flexibility. Unfortunately, more complex models will also fit random noise better than simpler models. We are interested in the model that

best fits the real trends in the data, not just the model that best fits the noise. As we will see in later chapters, Bayesian methods naturally take into account the complexity of model.

2.3 The R programming language

In this book you will learn how to actually *do* Bayesian statistics. For any but the simplest models, that means using a computer. Because the computer results are so central to doing real Bayesian statistics, examples of using the R computer programming language will be integrated into the simplest "toy" problems, so that R will not be an extra hurdle later.

The R language is great at doing Bayesian statistics for a number of reasons. First, it's free! You can get it via the web and easily install it on your computer. Second, it's already a popular language for doing Bayesian statistics, so there are lots of resources available. Third, it is a powerful and easy, general-purpose computing language, so you can use it for many other applications too.

2.3.1 Getting and installing R

It's easy to get and install R, but there are a lot of optional details in the process, and the hardest part of installation is figuring out which little details do *not* apply to you!

Basic installation is easy. Go to `http://cran.r-project.org/`. At the top of that webpage is a section headed "Download and Install R" followed by three links: Linux, MacOS, and Windows. These three links refer to the type of operating system used on your computer. Although R can be installed on any of those three types of operating system, it turns out that another package will be using extensively, called BUGS, only works on Windows. Macintosh users report that if they first install the freeware WINE (WINE Is Not an Emulator) from `http://www.winehq.org/`, and then install R and BUGS from within WINE, everything works seamlessly. The same is supposed to be true of Unix/Linux users. *Therefore, from here on, I will assume that you are using Windows or WINE.*

After clicking the Windows link, you will see a page with two links: base and contrib. Click "base". This opens a page with many links, most of which you will not need. Find the link with a label that ends with ".exe" and is described as "Setup program...". When you click the .exe link, just save the file, and after it is saved, run it. This should install R on your computer. There may be a few details that you have to navigate on your own, but remember that centuries ago lots of people crossed the oceans in tiny wooden boats without any electronics, so you can navigate the small perils of R installation.[3]

2.3.2 Invoking R and using the command line

Invoke R by double clicking the R icon in Windows. A user interface should open. In particular, one of the windows will be a command line interface. This window is constantly attentive to your every whim (well, every whim you can express in R). All you have to do is type in your wish and R will execute it as a command. For example, if you type in `show(2+2)`, followed by pressing the Enter key, R will reply with `4`. In fact, if you just type in `2+2`, without the `show` function, R will still reply with `4`.

A *program* (a.k.a. *script*) is just a list of commands that R executes. For example, you could first type in `x=2` and then, as a second command, type in `x+x`, to which R will reply `4`.

[3] Of course, lots of people failed to cross the ocean, but that's different.

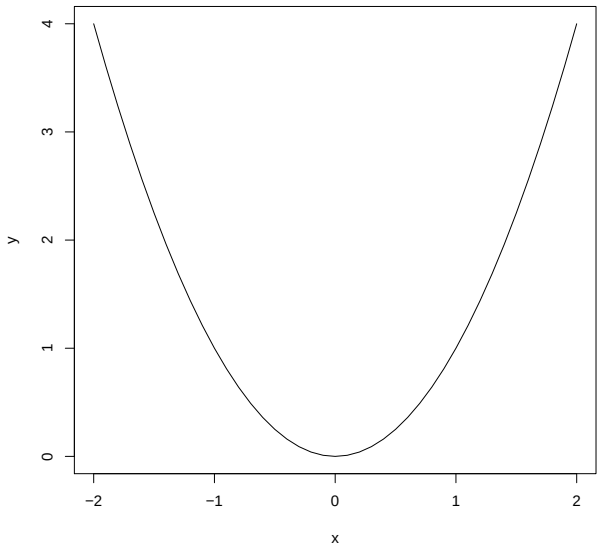


Figure 2.2: A simple graph drawn by R. The R code that generated this graph is on p. 17.

This is because R assumes that when you type in `x+x`, you are really asking for the value of the sum of the value of `x` with the value of `x`, not an algebraic reformulation such as `2*x` that some systems assume. [4]

2.3.3 A simple example of R in action

As a simple example of what R can do, let's plot a quadratic function: $y = x^2$. What looks like a smooth curve on a graph is actually a set of points connected by straight lines, but the lines are so small that the graph looks like a smooth curve. So what we have to do is tell R where all those densely packed points go.

Every point is specified by its x and y coordinates, so we have to provide R with a list of x values and a list of corresponding y values. Let's arbitrarily select x values from −2 to +2, separated by intervals of 0.1. We have R set up the list of x values by using the built-in *sequence* function: `x = seq( from = -2 , to = 2 , by = 0.1 )`. Inside R, the variable `x` now refers to a list of 31 values: −2.0, −1.9, −1.8, ..., +2.0. This sort of ordered list of numerical values is called a "vector" in R. In this textbook, programming commands are typeset in a distinctive font, `like this`, to distinguish them from English prose and to help demarcate the scope of the programming command when it is embedded in an English sentence.

Next we tell R to create the corresponding *y* values. We type in `y = x^2`. R interprets "`^`" to mean raising values to a power. Inside R, the variable `y` now refers to a vector of 31 values: 4.0, 3.61, 3.24, ..., 4.0.

All that remains is telling R to make a plot of the *x* and *y* points, connected by lines. Conveniently, R has a built-in function called `plot`, which we call by entering `plot( x , y , type="l" )`. The segment of code, `type="l"`, tells R to plot lines only and not points.

[4] You might ask, if R were a Bayesian reasoner, when you typed in `2+2`, wouldn't it reply something like, "well, I believe most strongly in 4, but the answer might be a little higher or a little lower"? A Bayesian reasoner would reply that way only if there were uncertainty introduced somewhere along the way. If the values to be added were uncertain, or if summation itself were an uncertain process, then the sum would be uncertain too. R assumes that numerals and arithmetic have no uncertainty.

If we left that part of the command out, then R would plot only points by default, not the connecting lines. The resulting plot is shown in Figure 2.2, and the complete R code that generated the graph is shown below:

(SimpleGraph.R)

```
x = seq( from = -2 , to = 2 , by = 0.1 )   # Specify vector of x values.
y = x^2                                    # Specify corresponding y values.
plot( x , y , type = "l" )                 # Make a graph of the x,y points.
dev.copy2eps( file = "SimpleGraph.eps" )   # Save the plot to an EPS file.
```

The code listing above has a few display features that will be standard throughout this book. First, the listing begins with the script's filename. This filename is relevant when you want to find the script on the website for this book. The filename is also in the index of this book. Second, the listing has line numbers in the margin. These line numbers are not part of the script, they are only part of the printed display. The line numbers are especially useful for the long programs encountered later in the book.

The last line of the program uses the function `dev.copy2eps` to save the graph to a file using a format called encapsulated postscript (eps). The command might not work on all installations of R, because some systems have not installed the eps driver. If the command doesn't work on your system, don't worry! You can save the graph in other formats, such as pdf. But how? Hmmm... Type `help('dev.copy2eps')`. Although the resulting help text can be somewhat cryptic, it reveals alternative commands such as `dev.copy2pdf`, which save the graph in pdf format.

2.3.4 Getting help in R

The `plot` function has many optional details that you can specify, such as the axis limits and labels, font sizes, etc. You can learn more about those details by getting help from R. Type the command `help('plot')` and you can read all about it. In particular, the information directs you to another command, `par`, that controls all the plot parameters. To learn about it, type `help('par')`. In general, it actually *is* helpful to use the `help` command. To get a list of all sorts of online documentation, much of it written in readable prose instead of telegraphic lists, type `help.start()`. Another useful way to find help with R is through web search. In your favorite web searcher, type in the R terms you want help with. When searching with the term "R" it helps to enclose it in square brackets, like this: [R].

A highly recommended resource is a summary of basic R commands that can be found on a compact list available at this URL:
http://cran.r-project.org/doc/contrib/Short-refcard.pdf
Other versions of reference cards can be found by searching the web with the phrase "R reference card".

Much of the time you'll learn about features of R on an as-needed basis, and usually that means you'll look for examples of the sort of thing you want to do and then imitate the example. (Or, the example might at least provoke you into realizing that there is a better way to do it than the method in the example!) Therefore, most of the examples in this book have their full R code included. Hopefully it will help you to study those examples as needed.

If you are already familiar with the programming languages Matlab or Python, you can find thesauruses of synonymous commands in R at this web site: `http://mathesaurus.sourceforge.net/`

2.3.5 Programming in R

Instead of typing all the commands one at a time at the command line, you can type them into a text document and then have R execute the document. The document is called a program.

Some important points for newbie programmers:

- Be sure you save your program in a file where you can find it again, with a filename that is easy to recognize weeks later.
- Be sure to save the program every time you make a small *working* change.
- If you are about to make a big change, save the current working version and start the modified version with a new filename. This way, when your modified version doesn't work, you still have the old working version to fall back on.
- Put lots of explanatory comments in your code, so that you can understand what the heck you were doing when you come back to the program months later. To include a comment in a program, simply type a "`#`" character, and everything after that character, on the same line, will be ignored by the R interpreter. You can see examples of comments in the code listings at the ends of the chapters.

2.3.5.1 Editing programs in R

Programs such as the one on p. 17 can be typed in any text-based word processor, but it can help enormously to use an editor that is "R friendly." The basic editor built into R is okay for small programs, but larger programs become unwieldy. One useful and free editor is Tinn-R, available at `http://www.sciviews.org/Tinn-R/`. It has many useful features such displaying comments in different colors and fonts than the command lines, highlighting matching parentheses, etc.

If you are using Windows Vista or Windows7, and if Tinn-R will not communicate with R, try the following fix: Browse to the folder Program Files > R > R-2.*. Right-click the "etc" folder and open its Properties. On the security tab, change the permissions so that the folder contents can be written to by you. This permits the Rconsole.site file to be overwritten by Tinn-R. Now go to Tinn-R, and from its menu click R > Configure > Permanent (Rprofile.site).

2.3.5.2 Variable names in R

You should name variables meaningfully, so that the programming commands are easy for a reader to understand. If you name your variables cryptically, you will curse your poor judgment when you return to the program weeks later and you have no idea what your program does.

You can use fairly long, descriptive names. If the names get too long, however, then the program becomes unwieldy to type and read. For example, suppose you want to name the crucial final output of a program. You could name it `tempfoo`, but that's not very meaningful, and might even lead you to think that the variable is unimportant. Instead, you could name it `theCrucialFinalOutputThatWillChangeTheWorldForever`, but that would be burdensome to type and read as it gets re-used in the program. So, you might best name it something like `finalOutput`, which is meaningful but not too long.

Computer programmers typically use a naming convention called *camelBack notation*. This is a way of connecting several words into a contiguous variable name without using spaces between words. For example, suppose you want to name a variable "final output". You are not allowed to name a variable with a space in it because computer compilers interpret spaces as separators of variables. One way to avoid using spaces is to connect the words with explicit connectors such as an underscore or a dot, like this: `final_output` or `final.output`. Many programmers do use those naming conventions. But the underscore notation can be difficult to read sometimes, and the dot notation is interpreted by some programming languages (other than R) as referring to subcomponents of structured variables, which confuses people who are familiar with that meaning of a dot. Therefore, the spaces are simply dropped, with successive words capitalized: `finalOutput`. The initial word is typically not capitalized, but some people have different uses for initial-capitalized variable names. R is case sensitive: the variable `myVar` is different than the variable `myvar`!

I will try to use camelBack notation in all the programs in this book. I may occasionally lapse from bactrian beauty, instead slithering into snakeback notation (`finaloutput`) or gooseneck notation (`final_output`) or ant notation (`final.output`). If you see these lower forms, quietly shun them, knowing that when you create your own programs, you will use the more highly evolved dromedary design.

2.3.5.3 Running a program

Running a program is easy, but exactly how to do it depends on how you are interacting with R.

If you are working in R's command console, first make sure that R has its working directory specified as the folder in which the program resides. Do this by selecting menu items *File* then *Change dir...*, and browsing to the appropriate folder in the pop-up dialogue box. Then you can run the program by "`source`"-ing it. Do this by selecting menu items *File* then *Source R code...*, and browsing to the program in the pop-up dialogue box. You can also type the `source("yourProgramName.R")` command directly at the command line.

You will more often be working interactively with a program that is open in an editing window. To open an R editing window, select menu items *File* then *New script* or *Open script*. Once you have a program open in an editing window, you can run the program, or merely a few lines within it, by selecting menu items *Edit* then *Run line or selection* or *Run all*. If you run the program from an editing window, every command is echoed in the command window. If you run the program by `source`ing it in the command window, then the program is executed without displaying the lines of code.

If you are working on a program in the editor Tinn-R, you will see menu buttons on the top tool bar that are equivalent to the R commands reviewed above. There is a button for setting the current working directory, there is another button for sourceing the program, and yet another button for running only the lines selected in the program being edited.

2.4 Exercises

Exercise 2.1. **[Purpose: To think about what beliefs can be altered by inference from data.]** Suppose I believe that exactly 47 angels can dance on my head. (These angels cannot be seen or felt in any way.) Is there any evidence you could provide that would change my belief?

Suppose I believe that exactly 47 anglers[5] can dance on the floor of the bait shop. Is there any evidence you could provide that would change my belief?

Exercise 2.2. **[Purpose: To get you actively manipulating mathematical models of probabilities. Notice, however, that these models have no parameters.]** Suppose we have a four-sided die from a board game. (On a tetrahedral die, each face is an equilateral triangle. When you roll the die, it lands with one face down and the other three visible as the faces of a three-sided pyramid. To read the value of the roll, you pick up the die and see what landed face down.) One side has 1-dot, the second side has 2-dots, the third side has 3-dots, and the fourth side has 4-dots. Denote the value of the bottom face as x. Consider the following three mathematical descriptions of the probabilities of x. Model A: $p(x) = 1/4$. Model B: $p(x) = x/10$. Model C: $p(x) = 12/(25x)$. For each model, determine the value of $p(x)$ for each value of x. Describe in words what kind of bias (or lack of bias) is expressed by each model.

Exercise 2.3. **[Purpose: To get you actively thinking about how data cause beliefs to shift.]** Suppose we have the tetrahedral die introduced in the previous exercise, along with the three candidate models of the die's probabilities. Suppose that initially we are not sure what to believe about the die. On the one hand, the die might be fair, with each face landing with the same probability. On the other hand, the die might be biased, with the faces that have more dots landing down more often (because the dots are created by embedding heavy jewels in the die, so that the sides with more dots are more likely to land on the bottom). On yet another hand, the die might be biased such that more dots on a face make it less likely to land down (because maybe the dots are bouncy rubber or protrude from the surface). So, initially, our beliefs about the three models can be described as $p(A) = p(B) = p(C) = 1/3$. Now we roll the die 100 times and find these results: $\#1's = 25$, $\#2's = 25$, $\#3's = 25$, $\#4's = 25$. Do these data change our beliefs about the models? Which model now seems most likely? Suppose when we rolled the die 100 times we found these results: $\#1's = 48$, $\#2's = 24$, $\#3's = 16$, $\#4's = 12$. Now which model seems most likely?

Exercise 2.4. **[Purpose: Actually doing Bayesian statistics, eventually, and the next exercises, immediately.]** Install R on your computer. (And if that's not exercise, I don't know what is.)

Exercise 2.5. **[Purpose: Being able to record and communicate the results of your analyses.]** Run the code listed on p. 17. The last line of the code saves the graph to a file in a format called "encapsulated PostScript" (abbreviated as eps), which your favorite word processor might be able to import. If your favorite word processor does not import eps files, then read the R documentation and find some other format that your word processor likes better; try `help('dev.copy2eps')`. You may find that you can just copy and paste the displayed graph directly into your document, but it can be useful to save the graph as a stand-alone file for future reference. Include the code listing and the resulting graph in a document that you compose using a word processor of your choice.

Exercise 2.6. **[Purpose: Getting experience with the details of the command syntax within R.]** Adapt the code listed in p. 17 so that it plots a cubic function ($y = x^3$) over the interval $x \in [-3, +3]$. Save the graph in a file format of your choice. Include a listing of your code, commented, and the resulting graph.

[5] *Angler* [noun]: A person who fishes with a hook and line.

Chapter 3

What is this stuff called probability?

Contents

Oh darlin' you change from one day to the next,
I'm feelin' deranged and just plain ol' perplexed.
I've learned to put up with your raves and your rants,
The mean I can handle but not variance.

Inferential statistical techniques provide precision to our uncertainty about possibilities.

Uncertainty is measured in terms of *probability*, and so we have to establish the basic properties of probability before we can make inferences about it. This chapter introduces the basic ideas of probability. If this chapter seems too abbreviated for you, an excellent beginner's introduction to the topics of this chapter has been written by Albert and Rossman (2001, pp. 227–320).

3.1 The set of all possible events

Suppose I have a coin that I am going to flip. How likely is it to come up a head? How likely is it to come up a tail? How likely is it to come up a torso? Notice that when we contemplate the likelihood of each outcome, we have a space of all possible outcomes in mind. Torso is not one of the possible outcomes. Notice also that a single flip of a coin can result in only one outcome; it can't be both heads and tails in a single flip. The outcomes are mutually exclusive.

Whenever we ask about how likely an event is, we always ask with a set of possible events in mind. This set exhausts all possible outcomes, and the outcomes are all mutually exclusive. This set is called the *sample space*.

In the previous chapter, we talked about the probability of a flipped coin coming up heads. The probability of coming up heads can be denoted with parameter label θ (Greek letter theta); for example, a coin is fair when $\theta = .5$ (spoken "theta equals point five"). We also have talked about the degree of belief that the coin is fair. The degree of belief about a parameter can be denoted $p(\theta)$. If the coin is minted by the federal government, we might have a strong belief that the coin is fair, e.g., $p(\theta = .5) = .99$, spoken "the probability (or degree of belief) that theta equals .5 is 99 percent".

Both "probability" (of head or tail) and "degree of belief" (in fairness) refer to sample spaces. The sample space for flips of a coin consists of two possible events: head and tail. The sample space for coin bias consists of a continuum of possible values: $\theta = 0$, $\theta = .01$, $\theta = .02$, $\theta = .03$, and all values in between, up to $\theta = 1$. When we flip a given coin, we are sampling from the space of head or tail. When we grab a coin at random from a sack of coins, we are sampling from the space of possible biases.

3.1.1 Coin flips: Why you should care

The fairness of a coin might be hugely consequential for high stakes games, but it isn't often in life that we flip coins. So why bother studying the statistics of coin flips?

Because coin flips are a surrogate for myriad other real-life events that we care about. For a given type of heart surgery, what is the probability that patients survive more than 1 year? For a given type of drug, what is the probability of headache as a side effect? For a particular training method, what is the probability of at least 10% improvement? For a survey question, what is the probability that people will agree or disagree? In a two candidate election, what is the probability that a person will vote for each candidate?

Whenever we are discussing coin flips, keep in mind that we could be talking about some domain in which you are actually interested! The coins are merely a generic representative of a universe of analogous applications.

3.2 Probability: Outside or inside the head

Sometimes we talk about probabilities of events that are "out there" in the world. The face of a flipped coin is such an event: We can observe the outcome of a flip, and the probability of coming up heads can be estimated by observing a bunch of flips.

But sometimes we talk about probabilities of events that are not so clearly "out there," and instead are just possible beliefs "inside the head." Our belief about the fairness of a coin is an example of an event inside the head. (The coin may have an intrinsic bias, but right now I'm referring to our *belief* about the bias.) Our beliefs refer to a space of mutually exclusive and exhaustive possibilities, but it might be strange to say that we randomly sample from our beliefs, like we randomly sample from a sack of coins. Nevertheless, the mathematical properties, of probabilities outside the head and beliefs inside the head, are the same in their essentials, as we will see.

3.2.1 Outside the head: Long-run relative frequency

For events outside the head, it's intuitive to think of probability as being the long-run relative frequency of each possible event. For example, if I say that for a fair coin the probability of heads is .5, what I mean is that if we flipped the coin many times, about 50% of the flips would come up heads. In the long run, after flipping the coin many, many times, the relative frequency of heads would be very nearly .5.

We can determine the long-run relative frequency two different ways. One way is to approximate it by actually sampling from the space many times and tallying the number of times each event happens. A second way is by deriving it mathematically. These two methods are now explored in turn.

3.2.1.1 Simulating a long-run relative frequency

Suppose we want to know the long-run relative frequency of getting heads from a fair coin. It might seem blatantly obvious that we should get about 50% heads in any long sequence of flips. But let's pretend that it's not so obvious: All we know is that there's some underlying process that generates an "H" or a "T" when we sample from it. The process has a parameter called θ, whose value is $\theta = .5$. If that's all we know, then we can approximate the long-run probability of getting an "H" by simply repeatedly sampling from the process. We sample from the process N times, tally the number of times an "H" appeared, and estimate the probability of H by the relative frequency: est.$\theta = \hat{\theta} = \#H/N$.

It gets tedious and time-consuming to manually sample a process, such as flipping a coin. Instead, we can let the computer do the repeated sampling much faster (and hopefully the computer feels less tedium than we do). Figure 3.1 shows the results of a computer simulating many flips of a fair coin. The R programming language has pseudo-random number generators built into it, which we will use often.[1] On the first flip, the computer randomly generates a head or a tail. It then computes the proportion of heads obtained so far. If the first flip was a head, then the proportion of heads is $1/1 = 1.0$. If the first flip was a tail, then the proportion of heads is $0/1 = 0.0$. Then the computer randomly generates a second head or tail, and computes the proportion of heads obtained so far. If the sequence so far is HH, then the proportion of heads is $2/2 = 1.0$. If the sequence so far

[1] Pseudo-random number generators are not actually random; they are in fact deterministic. But the properties of the sequences they generate mimic the properties of random processes.

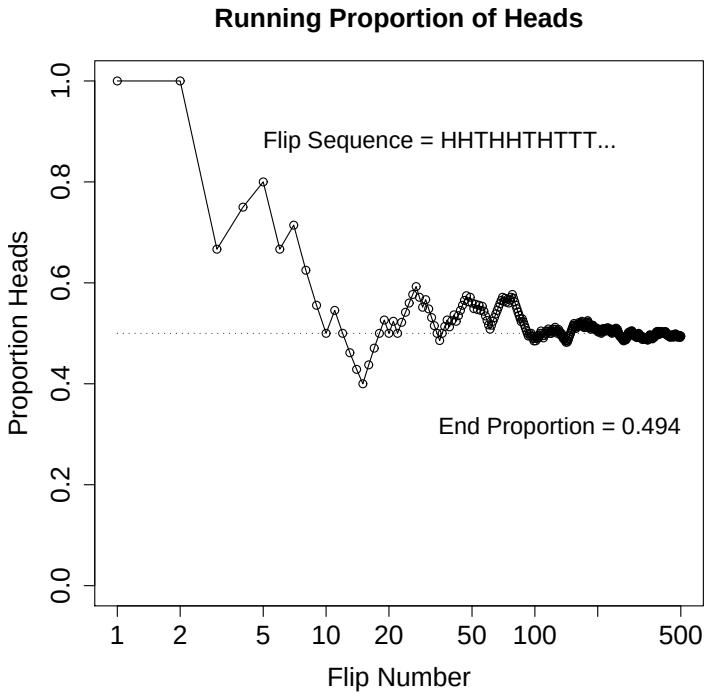


Figure 3.1: Running proportion of heads when flipping a coin. The x-axis is plotted on a logarithmic scale so that you can see the details of the first few flips but also the long-run trend after many flips. (The R code that generated this graph is in Section 3.5.1 (`RunningProportion.R`). When you run the code, your graph will look different than this one because you will generate a different random sequence of flips.)

is HT or TH, then the proportion of heads is $1/2 = 0.5$. If the sequence so far is TT, then the proportion of heads is $0/2 = 0.0$. Then the computer generates a third head or tail, and computes the proportion of heads so far, and so on for many flips. Figure 3.1 shows the running proportion of heads as the sequence continues.

Notice in Figure 3.1 that at the end of the long sequence, the proportion of heads is *near* 0.5 but not necessarily exactly equal to 0.5. This discrepancy reminds us that even this long run is still just a finite random sample, and there is no guarantee that the relative frequency of an event will match the true underlying probability of the event. That's why we say we are *approximating* the probability by the long-run relative frequency.

3.2.1.2 Deriving a long-run relative frequency

Sometimes, when the situation is simple enough mathematically, we can derive the exact long-run relative frequency. The case of the fair coin is one such simple situation.

The sample space of the coin consists of 2 possible outcomes, head and tail. By the assumption of fairness, we know that each outcome is equally likely. Therefore the long-run relative frequency of heads should be exactly one out of two, i.e., 1/2, and the long-run relative frequency of tails should also be exactly 1/2.

This technique is easily extended to other simple situations. Consider, for example, a standard six-sided die. It has six possible outcomes, namely 1 dot, 2 dots, ..., 6 dots. If we assume that the die is fair, then the long-run relative frequency of each outcome should be exactly 1/6.

Suppose that we put different dots on the faces of the six-side die. In particular, suppose that we put 1 dot on one face, 2 dots on two faces, and 3 dots on the remaining three faces. We still assume that each of the six faces is equally likely. Then the long-run relative frequency of 1 dot is exactly 1/6, and the long-run relative frequency of 2 dots is exactly 2/6, and the long-run relative frequency of 3 dots is exactly 3/6.

3.2.2 Inside the head: Subjective belief

How strongly do you believe that a coin minted by the US government is fair? If you believe that the coin could be slightly different than exactly fair, then how strongly do you believe that $\theta = .51$? Or $\theta = .49$? If instead you are considering a coin that is ancient, asymmetric, and lopsided, do you believe that it inherently has $\theta = .50$? How about a coin purchased at a magic shop? We are not talking here about the true, inherent probability that the coin will come up heads. We are talking about our degree of belief in each possible probability.

To specify our subjective beliefs, we have to specify how likely we think each possible outcome is. It can be hard to pin down mushy intuitive beliefs. In the next section we explore one way to "calibrate" subjective beliefs, and in the subsequent section we discuss ways to mathematically describe degrees of belief.

3.2.2.1 Calibrating a subjective belief by preferences

Consider a simple question that might affect travelers: How strongly do you believe that there will be a snowstorm that closes the interstate highways near Indianapolis next New Year's Day? Your job in answering that question is to provide a number between 0 and 1 that accurately reflects your belief probability. One way to come up with such a number is to calibrate your beliefs relative to other events with clear probabilities.

As a comparison event, consider a marbles-in-sack experiment. In a sack we put 10 marbles, 5 red and 5 white. We shake the sack and then draw a marble at random. The probability of getting a red marble is, of course, $5/10 = 1/2 = .5$. We will use this sack of marbles as a comparison for considering snow in Indianapolis on New Year's Day.

Consider the following two gambles that you can choose from:

- Gamble A: You get \$100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
- Gamble B: You get \$100 if you draw a red marble from a sack of marbles with 5 red and 5 white marbles.

Which gamble would you prefer? If you prefer Gamble B, that means you think there is less than a 50-50 chance of a traffic-stopping snowstorm in Indy. So at least you now know that your subjective belief about the probability of traffic-stopping snowstorm is less than .5.

We can narrow down the degree of belief by considering other comparison gambles. Consider these two gambles:

- Gamble A: You get \$100 if there is a traffic stopping snowstorm in Indianapolis next New Year's Day.
- Gamble C: You get \$100 if you draw a red marble from a sack of marbles with 1 red and 9 white marbles.

Which gamble would you prefer? If you now prefer Gamble A, that means you think there is more than a 10% chance of traffic-stopping snowstorm in Indy on New Year's Day. Taken together, the two comparison gambles have told you that your subjective probability lies somewhere between .1 and .5. We could continue to consider preferences against other candidate gambles to calibrate your subjective belief more accurately.

3.2.2.2 Describing a subjective belief mathematically

When there are several possible outcomes in a sample space, it might be too much effort to try to calibrate your subjective belief about every possible outcome. Instead, you can use a mathematical function to summarize your beliefs.

For example, you might believe that the average American woman is 5'4" tall, but be open to the possibility that the average might be somewhat above or below that value. It is too tedious and maybe impossible to specify your degree of belief that the average height is 4'1", or 4'2", or 4'3", and so on up through 6'1', 6'2", and 6'3" etc. So you might instead describe your degree of belief by a bell-shaped curve that is highest at 5'4" and drops off symmetrically above and below that most-likely height. You can change the width and center of the curve until it seems to best capture your subjective belief. Later in the book we will talk about exact mathematical formulas for functions like these, but the point now is merely to understand the idea that mathematical functions can define curves that can be used to describe degrees of belief.

3.2.3 Probabilities assign numbers to possibilities

In general, a probability, whether it's outside the head or inside the head, is just a way of assigning numbers to a set of mutually exclusive possibilities. The numbers, called "probabilities," merely need to satisfy three properties (Kolmogorov, 1956):

1. A probability value must be non-negative (i.e., zero or positive).

2. The sum of the probabilities across all events in the entire sample space must be 1.0 (i.e., one of the events in the space must happen, otherwise the space does not exhaust all possibilities).

3. For any two mutually exclusive events, the probability that one *or* the other occurs is the *sum* of their individual probabilities. For example, the probability that a fair six-sided die comes up 3-dots *or* 4-dots is $1/6 + 1/6 = 2/6$. As another example, if you believe there is a 60% chance of 0 to <3 inches of snow, and a 20% chance of 3 to <6 inches of snow, then you should believe that there is a 80% (=60%+20%) chance of 0 to <6 inches of snow.

Any assignment of numbers to events that respects those three properties will also have all the properties of probabilities that we will discuss below. So whether a probability is thought of as a long-run relative frequency of outcomes in the world, or as a magnitude of a subjective belief, it behaves the same way mathematically.

3.3 Probability distributions

A probability *distribution* is simply a list of all possible outcomes and their corresponding probabilities. For a coin, the probability distribution is trivial: We list two outcomes (head and tail) and their two corresponding probabilities (θ and $1-\theta$). For other sets of outcomes, however, the distribution can be more complex. For example, consider the total number of calories consumed by a person in a day. There is some probability that the number of calories consumed in a day will be 2000.0, some probability that the number will be less, say 1898.3, some probability that the number will be more, say 2447.9, and so forth. When

the outcomes are continuous, like calories, then the notion of probability takes on some subtleties, as we will see.

3.3.1 Discrete distributions: Probability mass

When the sample space consists of discrete outcomes, then we can talk about the probability of each distinct outcome. For example, the sample space of a flipped coin has two discrete outcomes, and we talk about the probability of head or tail. The sample space of a six-sided die has six discrete outcomes, and we talk about the probability of 1 dot, 2 dots, and so forth.

For continuous outcome spaces, we can *discretize* the space into a finite set of mutually exclusive and exhaustive "bins." For example, although calories consumed in a day is a continuous scale, we can divide up the scale into a finite number of intervals, such as <1500, 1500-2000, 2000-2500, 2500-3000, and >3000. Then we can talk about the probability of any one of those five intervals occurring: The probability of 2000-2500 is perhaps highest, with the probabilities of the other intervals dropping off from that high. Of course, the sum of the probabilities across the five intervals must sum to 1.

The probability of a discrete outcome is sometimes referred to as a probability *mass* to distinguish it from the probability density of an infinitesimal outcome, which will be defined next.

3.3.2 Continuous distributions: Rendezvous with density[†]

If you think carefully about a continuous outcome space, you realize that it becomes problematic to talk about the probability of a specific value on the continuum, as opposed to an interval on the continuum. For example, the probability that I eat exactly 2319.58372019... calories today is essentially nil, and that is true for *any* exact value you care to think of. We can, however, talk about the probability of intervals: The probability that I eat between 2000 and 2500 calories today is, say, .43. The problem with using intervals, however, is that their widths and edges are arbitrary, and very wide intervals are not very precise. So what we will do is make the intervals infinitesimally narrow, and instead of talking about the infinitesimal probability of that infinitesimal interval, we will talk about the ratio of the probability to the interval width. That ratio is called the probability density. Examples and further explanation follow.

Consider a spinner of the kind often found with board games. It has an arrow mounted on a hub in the center, and a flick of the finger makes the arrow spin around. Friction causes the arrow to eventually stop, pointing in some random direction. Along the perimeter of points reached by the arrowhead, there is a numerical scale that reads continuously from 0 to 1, wrapping around the circumference of the circle so that the 1 touches the 0, as shown in the top-left of Figure 3.2. We assume that the spinner is fair, so that any value from 0 to 1 is equally likely to be pointed at.

Let's divide the spinner into 2 equal sectors, one from 0 to .5 and the other from .5 to 1. When we spin the spinner, what is the probability that the outcome is in the first sector? Obviously the answer is 1/2. Suppose instead we divide the spinner into N equal sectors. What is the probability that that the spinner stops in any one of the N sectors? Obviously,

[†] "There is a mysterious cycle in human events. To some generations much is given. Of other generations much is expected. This generation of Americans has a rendezvous with destiny." Franklin Delano Roosevelt, 1936.

$1/N$. Notice that as we divide the spinner into more and more sectors, the probability of stopping in any one of them gets smaller and smaller. So if we want to know the probability of getting within a narrow range of a specific value, the probability is arbitrarily tiny.

But what if we instead consider the ratio of probability to sector width? The probability of stopping in a sector is $1/N$, and the width of a sector is $1/N$, so the ratio is $(1/N)/(1/N) = 1$. That ratio is called the probability *density*. It is called probability density by analogy with material density, which is defined as mass divided by volume. More loosely speaking, density is defined as the amount of stuff divided by the space it takes up. Applying that to the spinner, probability density in a sector is the amount of probability in that sector divided by the size of the sector. Again by analogy to material density, the amount of probability in the sector is called the probability *mass*. To reiterate, the probability density in an interval is the probability mass of that interval divided by the interval width.

Our goal, however, is finding the probability density *at a point*. To do that, we consider the probability density in the limit as the interval width around the point becomes infinitesimal. In the case of the fair spinner, this is easy: As N gets infinitely large, $1/N$ gets infinitesimally small. The density around a point in an interval is always $(1/N)/(1/N) = 1$, even as N grows infinitely large. In conclusion, for this case of a fair spinner ranging from 0 to 1, the probability density at every value is exactly 1.

Probability densities are not always exactly 1. For example, suppose that we re-label the spinner scale such that it starts at zero, but instead of uniformly going up to 1.0 in a 360° sweep, it goes only up to 0.5, as shown in the top-middle of Figure 3.2. Consider the sector from 0 to 0.1. This sector covers $1/5^{th}$ of the spinner, not $1/10^{th}$. Therefore, the probability mass within the sector is $1/5 = .2$, the sector width is 0.1, and the probability density within the sector is $.2/.1 = 2$. In general for this spinner, for a sector of width w, the probability mass is $2w$, and so the probability density in the sector is $2w/w = 2$. The probability density at a point is just the probability density of an infinitesimal interval around the point. Thus, for this case of a fair spinner ranging from 0 to 0.5, the probability density at every value is exactly 2. Notice that all we have changed from the previous example is the scale that goes around the circumference of the spinner. When the scale went from 0 to 1, the density was uniformly 1.0. When the scale went from 0 to 0.5, the density was uniformly 2.0.

Probability densities can be greater than 1, whereas probability masses cannot be greater than 1. Consider an analogy to a sponge. A 1 gram sponge has a mass of 1 gram regardless of how expanded or compressed it is. But the 1 gram sponge, when extremely tightly compressed into a tiny volume, can have a density as high as metal.

Probability densities are not always uniform. Continuing with the spinner example, suppose that the scale on the circumference of the spinner is logarithmic base-10, starting at 1 and wrapping around to 100 (like the x-axis of Figure 3.1, but going only to 100), as shown in the top-right of Figure 3.2. For this scale, the value of 10 appears 180° around the circle from 1, and the value 100 meets the value 1. The sector from 1 to 10 spans half the circle, and so the probability mass of the interval is .5, and the (average) probability density of this interval is $.5/(10-1) = .05556$. The sector from 10 to 100 spans the other half of the circle, and its (average) probability density is $.5/(100-10) = .00556$, ten times less than the first half. As we consider smaller intervals, you can see that the density differences will persist.

The lowly sponge can again educate our intuition. The 1 gram sponge can be squeezed at one end while the other end remains unsqueezed. The overall mass of the sponge remains 1 gram, but the density in the compressed end is much higher than the density in the uncompressed end.

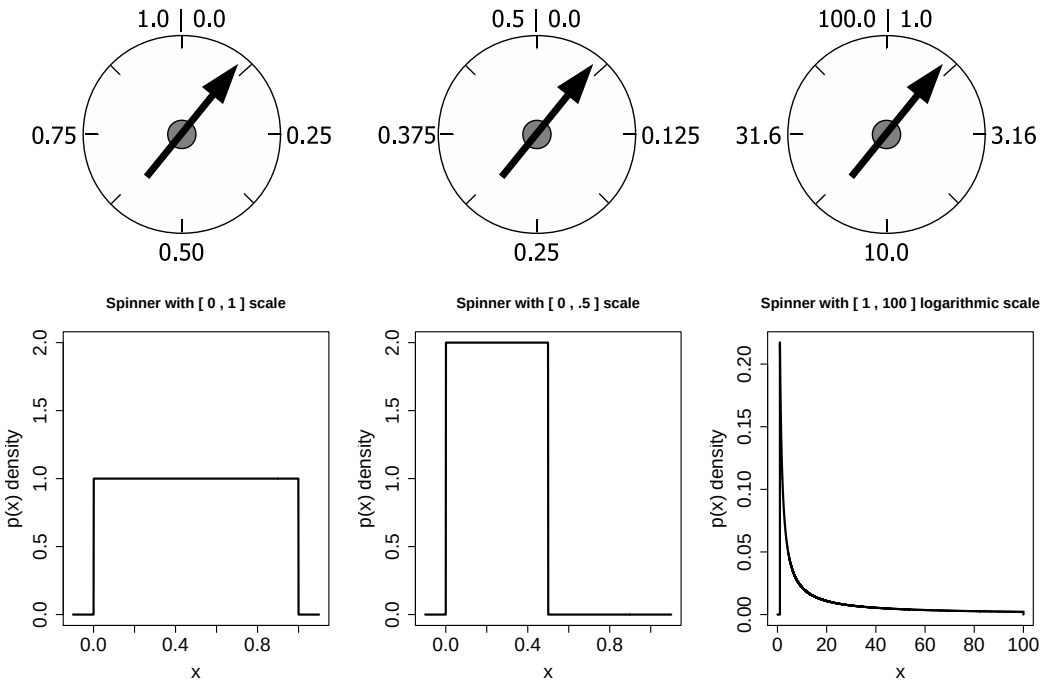


Figure 3.2: Spinners with different scales (above) and graphs of their probability density functions (below).

3.3.2.1 Properties of probability density functions

Consider again the basic spinner with a uniform scale from 0 to 1. What is the probability that the spinner lands somewhere between 0 and 1? The answer is, of course, 1, because the spinner must, by definition, have an outcome in the range 0 to 1. When we partition the spinner into N intervals, the probability of landing in any one interval is $1/N$. Notice, therefore, that the sum of the probabilities of the intervals is 1: $1/N+...+1/N = N\times 1/N = 1$.

In general, for any continuous value that is split up into intervals, the sum of the probabilities of the individual intervals must be 1, because some particular value must happen, by definition. We can write that as an equation, but we need to define some notation first. Let the continuous variable be denoted x. The width of an interval on x is denoted Δx (the symbol "Δ" is Greek letter capital delta). Let i be an index for the intervals, and let $[x_i, x_i+\Delta x]$ denote the interval between x_i and $x_i+\Delta x$. The probability mass of the i^{th} interval is denoted $p([x_i, x_i+\Delta x])$. Then —and this is the point— the sum of those probability masses is 1:

$$\sum_i p([x_i, x_i+\Delta x]) = 1.0. \tag{3.1}$$

Recall now the definition of probability density: It is the ratio of probability mass over interval width. We can re-write Equation 3.1 in terms of the density of each interval, by dividing and multiplying by Δx, as follows:

$$\sum_i \Delta x \, \frac{p([x_i, x_i+\Delta x])}{\Delta x} = 1.0 \tag{3.2}$$

In the limit, as the interval width becomes infinitesimal, we denote the width of the interval around x as dx instead of Δx, and we denote the probability *density* in the infinitesimal

interval around x simply as $p(x)$ (not to be confused with $p([x_i, x_i + \Delta x])$, which was the probability mass in an interval). Then the summation in Equation 3.2 becomes an integral:

$$\underbrace{\sum_i}_{\int} \underbrace{\Delta x}_{dx} \underbrace{\frac{p([x_i, x_i+\Delta x])}{\Delta x}}_{p(x)} = 1.0 \quad \text{that is,} \quad \int dx\, p(x) = 1.0 \tag{3.3}$$

In this book, integrals will be written with the dx term next to the integral sign, as in Equation 3.3, instead of at the far right end of the expression. Although this placement is not the most conventional notation, it is neither wrong nor unique to this book. The placement of dx next to the integral sign makes it easy to see what variable is being integrated over, without have to put subscripts on the integral sign. This usage is especially helpful later when we encounter integrals of functions that involve multiple variables.

The lower half of Figure 3.2 plots the probability density functions of the three spinner examples. The first example with the spinner had a scale that went from 0 to 1. As you may recall, we figured out that the density is 1.0 for all values of x on that scale. That density is plotted in the lower-left panel of Figure 3.2. Does Equation 3.3 work for this case? That is, does the probability density integrate to 1.0? Consider the density graph: It is a rectangle with height 1 and width 1, so its area is 1. The area under the density graph *is* the integral of the density function, so clearly Equation 3.3 is verified.

The second example with the spinner had a scale that went from 0 to .5, and we determined that the density was 2.0 for all values of x. This density is plotted in the lower-middle panel of Figure 3.2. Does Equation 3.3 work for this case? Does the probability density integrate to 1.0? The density function is a rectangle, with height 2.0 and width 0.5, so its area is 1.0.

The third example with the spinner involved a logarithmic scale that went from 1 to 100. We determined that its density on the low end of the scale was greater than its density on the high end of the scale. Although I won't derive it here, it turns out that the density at the point x is $1/(2\ln(10)x)$. (The mathematically inclined can find the method for deriving this result in Section 23.4.) This density is plotted in the lower-right panel of Figure 3.2. Notice that over the interval from 1 to 10, a typical density is around .05, which we determined earlier. For the interval from 10 to 100, a typical density is around .005, as we found before. This density function also integrates to 1.0, as it must.

To reiterate, in Equation 3.3, $p(x)$ is the probability density in the infinitesimal interval around x. Typically we let context tell us whether we are referring to a probability mass or a probability density, and use the same notation, $p(x)$, for both. For example, if x is the value of the face of a six-sided die, then $p(x)$ is a probability mass. If x is the exact point-value of number of calories consumed, then $p(x)$ is a probability density. There can be "slippage" in the usage, however. For example, if x refers to calories consumed, but the scale is discretized into intervals, then $p(x)$ is really referring to the probability mass of the interval in which x falls. In the end, you'll have to be careful and tolerant of ambiguity.

3.3.2.2 The normal probability density function

Any function that has only non-negative values and integrates to 1 (i.e., satisfies Equation 3.3) can be construed as a probability density function. Perhaps the most famous probability density function is the *normal* distribution, also known as the Gaussian distribution. A graph of the normal curve is a bell-shape; an example is shown in Figure 3.3.

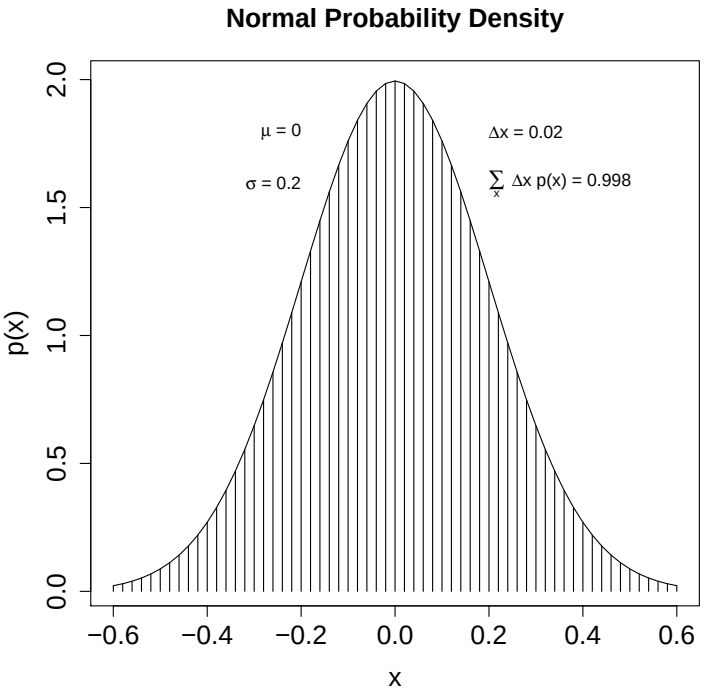


Figure 3.3: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. (The R code that generated this graph is in Section 3.5.2 (`IntegralOfDensity.R`).)

The mathematical formula for the normal probability density has two parameters: μ (Greek mu) is called the *mean* of the distribution and σ (Greek sigma) is called the *standard deviation*. The exact definitions of these terms will be provided in the next section, but for now all you need to know is that the value of μ governs where the middle of the bell shape falls on the x-axis, and the value of σ governs how wide the bell is. The exact mathematical formula for the normal probability density is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right). \tag{3.4}$$

Figure 3.3 shows an example of the normal distribution. The figure panel indicates the mean and standard deviation of the particular normal distribution that is displayed. Notice that the peak probability density can be greater than 1.0 when the standard deviation, σ, is small. In other words, when the standard deviation is small, a lot of probability mass has to get squeezed into a small interval, and consequently the probability density in that interval is high.

Figure 3.3 also illustrates that the area under the normal curve is, in fact, 1. The x axis is divided into a dense comb of small intervals, with width denoted Δx. The integral of the normal density is approximated by summing the masses of all the tiny intervals. As can be seen in the text within the graph, the sum of the interval areas is very nearly 1.0. Only rounding error, and the fact that the extreme tails of the distribution are not included in the sum, prevent the sum from being exactly 1.

The normal probability density function can be used to describe the relative frequencies of x values generated by a random process. The normal probability density function can also be used to describe our degree of belief in different x values. Let's apply these two notions to the case of guessing the height of a person selected at random. Under the first notion, the idea is that the height of the person is influenced by myriad random factors, and so the person's height will tend to be around μ, but could be somewhat larger or smaller, with a

spread indicated by σ. In this case, $p(x)$ refers to the probability density for observing a value x. Under the second notion, the idea is that we believe most strongly that the person's height is μ, but we also entertain the possibility that the person's height could be larger or smaller, with a spread indicated by σ. In this case, $p(x)$ refers to the density of our belief in value x. The mathematics of $p(x)$ are the same for either meaning, whether we think of $p(x)$ as referring to (1) the relative frequency that a random process will generate various values, or (2) degrees of belief in alternative possibilities.

3.3.3 Mean and variance of a distribution

When we have a numerical (not just categorical) value x that is generated with probability $p(x)$, we can wonder what would be its average value in the long run, if we repeatedly sampled values of x. For example, if we have a fair six-sided die, then each of its six values should come up 1/6th of the time in the long run, and so the long-run average value of the die is $(1/6)1 + (1/6)2 + (1/6)3 + (1/6)4 + (1/6)5 + (1/6)6 = 3.5$. As another example, if we play a slot machine for which we win \$100 with probability .001, we win \$5 with probability .14, and otherwise we lose \$1, then in the long run our payoff is $(.001)(\$100) + (.14)(\$5) + (.859)(-\$1) = -\$0.059$; i.e., in the long run we lose about 6 cents per pull of the bandit's arm. Notice what we did in those calculations: We weighted each possible outcome by the probability that it happens. This procedure defines the *mean* of a probability distribution, which is also called the *expected value*, and which is denoted $E[x]$:

$$E[x] = \sum_x p(x)\, x. \tag{3.5}$$

Equation 3.5 applies when the values of x are discrete, and so $p(x)$ denotes a probability mass. When the values of x are continuous, then $p(x)$ denotes a probability density and the sum becomes an integral over infinitesimal intervals:

$$E[x] = \int dx\, p(x)\, x. \tag{3.6}$$

The conceptual meaning is the same: The long-run average of the values.

The mean value of a distribution typically lies near the distribution's middle, intuitively speaking. For example, the mean of a normal distribution turns out to be the value of its parameter μ, i.e., $E[x] = \mu$. A specific case appears in Figure 3.3, where it can be seen that the bulk of the distribution is centered over $x = \mu$; see the text in the figure for the exact value of μ.

Here's an example of computing the mean of a continuous distribution, using Equation 3.6. Consider the probability density function $p(x) = 6x(1-x)$ defined over the interval $x \in [0, 1]$. That really is a probability density function: It's an upside down parabola starting at $x = 0$, peaking over $x = 0.5$, and dropping down to baseline again at $x = 1$. Because it is a symmetric distribution, intuition tells us that the mean should be at its midpoint, i.e., $x = 0.5$. Let's check that it really is:

$$\begin{aligned} E[x] &= \int_0^1 dx\, p(x)\, x \\ &= \int_0^1 dx\; 6x(1-x)\, x \end{aligned}$$

$$
\begin{aligned}
&= 6 \int_0^1 dx \ \left(x^2 - x^3\right) \\
&= 6 \left[\frac{1}{3}x^3 - \frac{1}{4}x^4\right]_0^1 \\
&= 6 \left[\left(\frac{1}{3}1^3 - \frac{1}{4}1^4\right) - \left(\frac{1}{3}0^3 - \frac{1}{4}0^4\right)\right] \\
&= 0.5
\end{aligned} \tag{3.7}
$$

The *variance* of a probability distribution is a number that represents the dispersion of the distribution away from its mean. There are many conceivable definitions of how far the values of x are dispersed from their mean, but the definition used for the specific term "variance" is based on the squared difference between x and the mean. The definition of variance is simply the mean squared deviation (MSD) of the x values from their mean:

$$\text{var}_x = \int dx \, p(x) \ (x - E[x])^2 \, . \tag{3.8}$$

Notice that Equation 3.8 is just like the formula for the mean (Equation 3.6) except that instead of integrating x weighted by x's probability, we're integrating $(x - E[x])^2$ weighted by x's probability. In other words, the variance is just the average value of $(x - E[x])^2$. For a discrete distribution, the integral in Equation 3.8 becomes a sum, analogous to the relationship between Equations 3.6 and 3.5. The square root of the variance, sometimes referred to as root mean squared deviation (RMSD), is called the *standard deviation* of the distribution.

The variance of the normal distribution turns out to be the value of its parameter σ squared; i.e., for the normal, $\text{var}_x = \sigma^2$. In other words, the standard deviation of the normal distribution is the value of the parameter σ. In a normal distribution, about 34% of the distribution lies between μ and $\mu + \sigma$ (see Exercise 3.4). Take a look at Figure 3.3 and visually identify where μ and $\mu + \sigma$ lie on the x axis (the values of μ and σ are indicated in the text within the figure) to get a visual impression of how far one standard deviation lies from the mean. Be careful, however, not to overgeneralize to distributions with other shapes: Non-normal distributions can have very different areas between their mean and first standard deviation.

3.3.3.1 Mean as minimized variance

An alternative conceptual emphasis starts with the definition of variance and derives a definition of mean, instead of starting with the mean and working to a definition of variance. Under this alternative conception, the goal is to define a value for the *central tendency* of a probability distribution. A value represents the central tendency of the distribution if the value is close to the highly probable values of the distribution. Therefore, we define the central tendency of a distribution as the value M that minimizes the long-run expected distance between it and all the other values of x. But how should we define "distance" between values? One way to define distance is as squared difference: The distance between x and M is $(x - M)^2$. One virtue of this definition is that the distance from x to M is the same as the distance from M to x, because $(x - M)^2 = (M - x)^2$. But the primary virtue of this definition is that it makes a lot of subsequent algebra tractable (which will not be rehearsed here). The central tendency is, therefore, the value M that minimizes the expected value of

$(x - M)^2$. Thus, we want the value M that minimizes $\int dx\, p(x)\,(x - M)^2$. Does that look familiar? It's essentially the formula for the variance of the distribution, in Equation 3.8, but here thought of as a function of M. Here's the punch line: The value of M that minimizes $\int dx\, p(x)\,(x-M)^2$ is, it turns out, $E[x]$. In other words, the mean of the distribution is the value that minimizes the expected squared deviation. In this way, the mean is a central tendency of the distribution.

As an aside, if the distance between M and x is defined instead as $|x-M|$, then the value that minimizes the expected distance is called the *median* of the distribution. An analogous statement applies to the *modes* of a distribution, with distance defined as zero for any exact match, and one for any mismatch.

3.3.4 Variance as uncertainty in beliefs

If $p(\theta)$ represents degrees of belief in values of θ, instead of the probability of sampling θ, then the mean of $p(\theta)$ can be thought of as the value of θ that represents our typical or central belief. The variance of θ, which measures how spread out the distribution is, can be thought of as our uncertainty about possible beliefs. If the variance is small, then we believe strongly in values of θ near the mean. If the variance is large, then we are not very certain about what value of θ to believe in. This notion of variance (or standard deviation) as representing uncertainty will re-appear often.

3.3.5 Highest density interval (HDI)

Another way of summarizing a belief distribution, which we will use often, is the highest density interval, abbreviated HDI.[2] The HDI indicates which points of a distribution we believe in most strongly, and which cover most of the distribution. Thus, the HDI summarizes the distribution by specifying an interval that spans most of the distribution, say 95% of it, such that every point inside the interval has higher believability than any point outside the interval.

If you think of the probability distribution as the profile of an island rising out the water, then the 95% HDI marks the waterline on the beach such that 95% of the island's mass is within the waterline. All the points within the waterline (above water) have higher believability than any point outside the waterline (below water). Figure 3.4 shows examples of HDIs; take a look now.

The formal definition of an HDI is just a mathematical expression of the waterline idea. What we want in the HDI is all those values of x for which we have a belief density at least as big as some value W (which is the depth of water), such that the integral over all those x values is 95% (or 99%, or whatever). Formally, the values of x in the 95% HDI are those such that $p(x) > W$ where W satisfies $\int_{x \text{ s.t. } p(x)>W} dx\ p(x) = .95$.

The width of the HDI is another way of measuring uncertainty of beliefs. If the HDI is wide, then beliefs are uncertain. If the HDI is narrow, then beliefs are fairly certain. Sometimes the goal of research is to obtain data the achieve a reasonably high degree of

[2] Some authors refer to the HDI as the HDR, which stands for highest density *region*, because a region can refer to multiple dimensions, but an interval refers to a single dimension. Because we will almost always consider the HDI of one parameter at a time, I will use "HDI" in an effort to reduce confusion. Some authors refer to the HDI as the HPD, to stand for highest probability density, but which I prefer not to use because it takes more space to write "HPD interval" than "HDI". Some authors refer to the HDI as the HPD, to stand for highest *posterior* density, but which I prefer not to use because *prior* distributions also have HDIs.

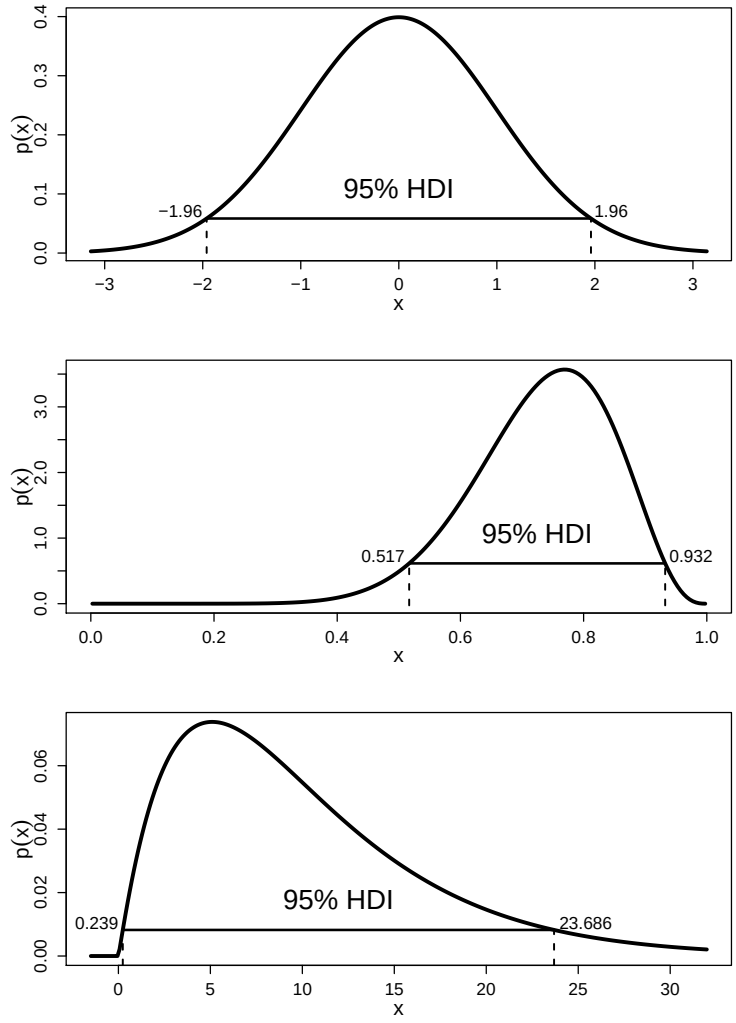


Figure 3.4: Examples of 95% highest density intervals (HDIs). For each example, all the x values inside the interval have higher density than any x value outside the interval, and the total mass of the points inside the interval is 95%. The ends of the HDIs are marked with their x values. The top panel shows a normal distribution with mean zero and standard deviation one. The middle panel shows a beta distribution (defined in Chapter 5), and the lower panel shows a gamma distribution (defined in Chapter 9). Don't fret over the definitions of the density functions; the point here is to get a visual intuition for HDIs in differently shaped distributions.

certainty about a particular value x. The desired degree of certainty can be measured as the width of the 95% HDI. For example, if x is a measure of how much a drug decreases blood pressure, the researcher may want to have an estimate with a 95% HDI no larger than 5 units on the blood pressure scale. As another example, if x is a measure of a population's preference for candidate A over candidate B, the researcher may want to have an estimate with a 95% HDI no larger than 10 percentage points.

3.4 Two-way distributions

There are many situations in which we are interested in the conjunction of two outcomes. What is the probability of being dealt a card that is both a queen *and* a heart? What is the probability of meeting a person with both red hair *and* green eyes? When playing a board game involving a die and a spinner, we have degrees of belief about both the die *and* the spinner being fair.

As a specific example for developing these ideas, imagine tossing a coin three times in a row. The sequence of flips might be TTT, or TTH, etc. Table 3.1 shows all possible sequences of three outcomes in its left column. As you can tell by counting the rows of

the table, there are 8 possible sequences. Because the coin is assumed to be fair, each row is equally likely, and so each row has probability 1/8, which is indicated in the far right column.

Table 3.1: Sample space for tossing a coin three times.

Outcome	# Heads	# Switches	Probability
TTT	0	0	1/8
TTH	1	1	1/8
THT	1	2	1/8
THH	2	1	1/8
HTT	1	1	1/8
HTH	2	2	1/8
HHT	2	1	1/8
HHH	3	0	1/8

For each sequence of three tosses, we can count the number of heads in the sequence, and the number of times that the outcome switched between heads or tails. Table 3.1 also lists the number of heads and the number of switches in each sequence. We can list the combinations of head counts and switches in a *two-way* table, as shown in Table 3.2. This two-way table shows the probability of getting a particular combination of number of heads *and* number of switches. The probability of two things happening together is called their *conjoint* probability.

Table 3.2: Conjoint probabilities for tossing a coin three times, compiled from Table 3.1.

	# Heads			
# Switches	**0**	**1**	**2**	**3**
0	1/8	0	0	1/8
1	0	2/8	2/8	0
2	0	1/8	1/8	0

One of the interesting characteristics revealed by Table 3.2 is that not all combinations of events are equally likely, and some combinations don't happen at all. For example, the probability of getting a sequence with 1 head and 1 switch is 2/8, i.e., this combination happens 25% of the time in the long run. On the other hand, it never happens that there is a sequence with 1 head and 0 switches. We will be using this table of conjoint probabilities to develop other concepts.

3.4.1 Marginal probability

When we flip the coin three times, we can count the number of heads and the number of switches, but we might be interested only in one or the other type of outcome. We can ask, what's the probability of getting 2 heads in 3 flips, without worrying about the number of switches involved. One way to determine the probabilities of these different classes of outcomes is to sum across the conjoint probabilities we've already compiled. For example, if we want to determine the probability of getting 0 heads, regardless of the number of switches in the sequence, we simply sum the probability of getting 0 heads with 0 switches and the probability of getting 0 heads with 1 switch and the probability of getting 0 heads

with 2 switches. In other words, we simply sum across rows in Table 3.2. We can do this for every number of heads, and write the sums in the lower margin of the table, as shown in Table 3.3. These summed probabilities are called the *marginal* probabilities. We can also compute the probability of each number of switches, regardless of the number of heads, by summing across the number of heads. Table 3.3 also shows these probabilities in the righthand margin.

Table 3.3: Marginal probabilities when tossing a coin three times. (This table extends Table 3.2.)

	# Heads				**Marginal**
# Switches	**0**	**1**	**2**	**3**	**(# Switches)**
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
Marginal (# Heads):	1/8	3/8	3/8	1/8	

Let's now establish some general notation for the type of example we've been considering. We suppose that we have a sample space in which every outcome has two attributes, x and y. In the previous example, the outcomes in the sample space were sequences of three flips, and the two attributes were number of heads (x) and number of switches (y). The conjoint probability of a particular combination of x and y values is denoted $p(x, y)$. For example, the conjoint probability of 2 heads and 1 switch is $p(x = 2, y = 1) = 2/8$. Notice that conjoint probabilities are symmetric: $p(x, y) = p(y, x)$.

To compute the probability distribution for x by itself, we sum $p(x, y)$ across all values of y:

$$p(x) = \sum_y p(x, y) \tag{3.9}$$

When the x and y variables are continuous, then $p(x, y)$ is a probability density, and the summation becomes an integral:

$$p(x) = \int dy\, p(x, y) \tag{3.10}$$

where the resulting marginal distribution, $p(x)$, is also a density. This summation process is called *marginalizing over* y or *integrating out* the variable y. Of course, we can also determine the probability of y by marginalizing over x.

These notions of conjoint and marginal probabilities also apply to beliefs. Consider, for example, two coins: a nickel and a dime. Suppose that we believe that they might be fair, or that they are trick coins with heads on both sides or with tails on both sides. We believe most strongly that they are both fair, but that there is a small chance that they are trick coins. Moreover, we believe that if one is a trick coin, then the other is a trick coin too. These beliefs can be captured by a joint probability table, as shown in Table 3.4.

Table 3.4 indicates that our belief that both coins are fair is 60%, and that we believe there is a 10% chance that both coins are two-tailed, and that we believe it is impossible for one coin to be fair while the other is a trick coin. Table 3.4 also shows the marginal distributions of our beliefs. The point of this example is merely to illustrate that we can talk about conjoint and marginal structure of belief distributions just as we do about data distributions.

Table 3.4: Beliefs regarding two coins.

	Nickel			Marginal
Dime	**Two Tails**	**Fair**	**Two Heads**	**(Dime)**
Two Tails	.1	0	.1	.2
Fair	0	.6	0	.6
Two Heads	.1	0	.1	.2
Marginal (Nickel):	.2	.6	.2	

3.4.2 Conditional probability

We often want to know the probability of one event, given that we know another event is true. For example, what is the probability that it will rain in the next 24 hours given that there is a thunder storm 400 miles due west of you? What is the probability that you will pass the statistics class given that you scored 88/100 on the first assignment? What is the probability that a sequence of three coin flips has 1 switch given that it has 1 head?

Let's think about that last question in detail. Refer back to Table 3.3. We want to know the probability that a sequence of three flips has 1 switch, given that it has 1 head. This means that we are only considering sequences that we know to have 1 head, which means only the column in Table 3.3 labeled "1" under "# Heads." The question is, of the probability within that column, how much of it occurs for 1 switch? We can see that the conjoint probability of 1 switch and 1 head is 2/8, and the total probability of 1-head column is $0+2/8+1/8 = 3/8$. Therefore the probability of getting 1 switch, given that the sequence has 1 head, is $(2/8)/(3/8) = 2/3$.

Notice the conditional probability of getting 1 switch, given that there is 1 head, is different from the marginal probability of getting 1 switch. The marginal probability indicates the probability of getting 1 switch on average across all numbers of heads, whereas the conditional probability restricts consideration to a particular number of heads.

Conditional probabilities have their own notation. The probability of a value of y given a value of x is denoted $p(y|x)$. For the previous example, with number of heads denoted x and number of switches denoted y, we write $p(y=1|x=1) = 2/3$.

Now that we have some general notation, we can generalize our computations from the example. Recall that to compute the conditional probability $p(y = 1|x = 1)$, we divided the conjoint probability $p(y = 1, x = 1)$ by the sum of conjoint probabilities for the given value, $\sum_y p(y, x = 1)$. Notice also that the sum of the conjoint probabilities is the marginal probability. So, in general, the conditional probability is

$$p(y|x) = \frac{p(y,x)}{\sum_y p(y,x)} = \frac{p(y,x)}{p(x)}. \tag{3.11}$$

(Notice that the equality of the denominators in Equation 3.11 was already discussed in Equation 3.9.) When y is continuous, the sum becomes an integral over the conjoint density:

$$p(y|x) = \frac{p(y,x)}{\int_y dy\, p(y,x)} = \frac{p(y,x)}{p(x)}. \tag{3.12}$$

(Notice that the equality of the denominators in Equation 3.12 was already discussed in Equation 3.10.)

Of course, we can conditionalize on the other variable, instead. That is, we can consider $p(x|y)$ instead of $p(y|x)$. It is important to recognize that, in general, $p(x|y) \neq p(y|x)$. For

example, the probability that the ground is wet, given that it's raining, is different than the probability that it's raining, given that the ground is wet. The next chapter provides an extended discussion of the relationship between $p(x|y)$ and $p(y|x)$. The relationship is called Bayes' rule.

It is also important to recognize that there is no temporal order in conditional probabilities. When we say "the probability of x given y" we do *not* mean that y has already happened and x has yet to happen. All we mean is that we are restricting our calculations of probability to a particular subset of possible events. A better gloss of $p(x|y)$ is, "among all events with value y, this proportion of them also have value x." So, for example, we can talk about the probability that it rained last night given that there are clouds this morning. This is simply referring to the proportion of all cloudy mornings like this one that had rain the night before.

Finally, as I have repeatedly emphasized, the notions of conditional probability apply to belief distributions, just as they apply to data distributions. Refer back to Table 3.4, regarding beliefs about the fairness of a nickel and a dime. Consider the probability that the dime is fair given that the nickel is fair. Using Equation 3.11, we find that the conditional probability is 1. This simply means that of all our beliefs for which the nickel is fair, 100% of them have the dime being fair.

3.4.3 Independence of attributes

Suppose I have a six-sided die and a spinner. Suppose they are fair. I flick the spinner and it points to 0.123. Given this result on the spinner, what is the probability that the rolled die will come up 3? In answering this question, you probably thought, "the spinner has no influence on the die, so the probability of the die coming up 3 is 1/6 regardless of what the spinner is pointing at." If that's what you thought, you were assuming that the spinner and the die are *independent*.

In general, when the value of y has no influence on the value of x, we know that $p(x|y) = p(x)$, for all values of x and y. Let's think a moment about what that implies. We know from the definition of conditional probability, in Equations 3.11 or 3.12, that $p(x|y) = p(x,y)/p(y)$. Combining those equations implies that $p(x) = p(x,y)/p(y)$ for all values of x and y. After multiplying both sides by $p(y)$, we get the implication that $p(x,y) = p(x)p(y)$ for all values of x and y. The implication goes the other way, too: When $p(x,y) = p(x)p(y)$ for all values of x and y, then $p(x|y) = p(x)$ for all values of x and y. Therefore either of these conditions is our mathematical definition of independence of attributes.

Consider the example back in Table 3.3 (page 37), regarding sequences of three flips of a coin. Are the attributes of number of heads and number of switches independent? You can quickly see that the answer is no. Consider, for example, the top left cell, which contains the conjoint probability of 0 heads and 0 switches (namely, 1/8). Does it equal the product of the marginal probability of 0 heads and the marginal probability of 0 switches (namely, $1/8 \times 2/8$)? No, clearly not.

As a second case, consider the example in Table 3.4 (page 38), regarding beliefs about two coins. The beliefs in that scenario were explicitly that the coins were *not* independent: If one coin was fair, then so was the other one, but if one coin was a trick coin, so was the other one. The lack of independence can be verified mathematically in Table 3.4. Consider the top left cell: Is our conjoint belief probability that both coins are two-tailed (namely, .1) equal to the product of our marginal belief probabilities that the nickel is two tailed and the

dime is two-tailed (namely, $.2 \times .2$)?

As a simple example of two attributes that *are* independent, consider the suit and value of cards in a standard deck. There are four suits, and thirteen values of each suit, making 52 cards altogether. Consider a randomly dealt card. What is the probability that it is a heart? (Answer: 13/52 = 1/4.) Suppose I look at the card without letting you see it, and I tell you that it is a Queen. Now what is the probability that it is a heart? (Answer: 1/4.) Telling you the card's value does not change the probabilities of the suits, so value and suit are independent. We can verify this in terms of cross multiplying marginal probabilities, too: Each combination of value and suit has a 1/52 chance of being dealt (in a fairly shuffled deck). Notice that 1/52 is exactly the marginal probability of any one suit (1/4) times the marginal probability of any one value (1/13).

Among other contexts, independence will come up again when we are constructing mathematical descriptions of our beliefs about more than one attribute. We will create a mathematical description of our beliefs about one attribute, and another mathematical description of our beliefs about the other attribute. Then, to describe what we believe about combinations of attributes, we will often assume independence, and simply multiply the separate beliefs to specify the conjoint beliefs.

3.5 R code

3.5.1 R code for Figure 3.1

(RunningProportion.R)

```
# Goal: Toss a coin N times and compute the running proportion of heads.
N = 500        # Specify the total number of flips, denoted N.
# Generate a random sample of N flips for a fair coin (heads=1, tails=0);
# the function "sample" is part of R:
#set.seed(47405) # Uncomment to set the "seed" for the random number generator.
flipsequence = sample( x=c(0,1) , prob=c(.5,.5) , size=N , replace=TRUE )
# Compute the running proportion of heads:
r = cumsum( flipsequence ) # The function "cumsum" is built in to R.
n = 1:N                    # n is a vector.
runprop = r / n            # component by component division.
# Graph the running proportion:
# To learn about the parameters of the plot function,
# type help('par') at the R command prompt.
# Note that "c" is a function in R.
plot( n , runprop , type="o" , log="x" ,
          xlim=c(1,N) , ylim=c(0.0,1.0) , cex.axis=1.5 ,
          xlab="Flip Number" , ylab="Proportion Heads" , cex.lab=1.5 ,
          main="Running Proportion of Heads" , cex.main=1.5 )
# Plot a dotted horizontal line at y=.5, just as a reference line:
lines( c(1,N) , c(.5,.5) , lty=3 )
# Display the beginning of the flip sequence. These string and character
# manipulations may seem mysterious, but you can de-mystify by unpacking
# the commands starting with the innermost parentheses or brackets and
# moving to the outermost.
flipletters = paste( c("T","H")[ flipsequence[ 1:10 ] + 1 ] , collapse="" )
displaystring = paste( "Flip Sequence = " , flipletters , "..." , sep="" )
text( 5 , .9 , displaystring , adj=c(0,1) , cex=1.3 )
# Display the relative frequency at the end of the sequence.
text( N , .3 , paste("End Proportion =",runprop[N]) , adj=c(1,0) , cex=1.3 )
```

```
# Save the plot to an EPS file.
dev.copy2eps( file = "RunningProportion.eps" )
```

3.5.2 R code for Figure 3.3

(IntegralOfDensity.R)

```
# Graph of normal probability density function, with comb of intervals.
meanval = 0.0             # Specify mean of distribution.
sdval = 0.2               # Specify standard deviation of distribution.
xlow  = meanval - 3*sdval # Specify low end of x-axis.
xhigh = meanval + 3*sdval # Specify high end of x-axis.
dx = 0.02                 # Specify interval width on x-axis
# Specify comb points along the x axis:
x = seq( from = xlow , to = xhigh , by = dx )
# Compute y values, i.e., probability density at each value of x:
y = ( 1/(sdval*sqrt(2*pi)) ) * exp( -.5 * ((x-meanval)/sdval)^2 )
# Plot the function. "plot" draws the intervals. "lines" draws the bell curve.
plot( x , y , type="h" , lwd=1 , cex.axis=1.5
        , xlab="x" , ylab="p(x)" , cex.lab=1.5
        , main="Normal Probability Density" , cex.main=1.5 )
lines( x , y )
# Approximate the integral as the sum of width * height for each interval.
area = sum( dx * y )
# Display info in the graph.
text( -sdval , .9*max(y) , bquote( paste(mu ," = " ,.(meanval)) )
      , adj=c(1,.5) )
text( -sdval , .8*max(y) , bquote( paste(sigma ," = " ,.(sdval)) )
      , adj=c(1,.5) )
text( sdval , .9*max(y) , bquote( paste(Delta , "x = " ,.(dx)) )
      , adj=c(0,.5) )
text( sdval , .8*max(y) ,
      bquote(
        paste( sum(,x,) , " " , Delta , "x p(x) = " , .(signif(area,3)) )
      ) , adj=c(0,.5) )
# Save the plot to an EPS file.
dev.copy2eps( file = "IntegralOfDensity.eps" )
```

3.6 Exercises

Exercise 3.1. **[Purpose: To give you some experience with random number generation in R.]** Modify the coin flipping program in Section 3.5.1 (`RunningProportion.R`) to simulate a biased coin that has $p(H) = .8$. Change the height of the reference line in the plot to match $p(H)$. Comment your code. Hint: Read the help for `sample`.

Exercise 3.2. **[Purpose: To have you work through an example of the logic presented in Section 3.2.1.2.]** Determine the exact probability of drawing a 10 from a shuffled pinochle deck. (In a pinochle deck, there are 48 cards. There are six values: 9, 10, Jack, Queen, King, Ace. There are two copies of each value in each of the standard four suits: hearts, diamonds, clubs, spades.)

(A) What is the probability of getting a 10?

(B) What is the probability of getting a *10 or Jack*?

Exercise 3.3. **[Purpose: To give you hands-on experience with a simple probability density function, in R and in calculus, and to re-emphasize that density functions can have values larger than 1.]** Consider

a spinner with a [0,1] scale on its circumference. Suppose that the spinner is slanted or magnetized or bent in some way such that it is biased, and its probability density function is $p(x) = 6x(1-x)$ over the interval $x \in [0,1]$.

(A) Adapt the code from Section 3.5.2 (`IntegralOfDensity.R`) to plot this density function and approximate its integral. Comment your code. Be careful to consider values of x only in the interval $[0,1]$. Hint: You can omit the first couple lines regarding `meanval` and `sdval`, because those parameter values pertain only to the normal distribution. Then set `xlow=0` and `xhigh=1`.

(B) Derive the exact integral using calculus. Hint: See the example, Equation 3.7.

(C) Does this function satisfy Equation 3.3?

(D) From inspecting the graph, what is the maximal value of $p(x)$?

Exercise 3.4. **[Purpose: To have you use a normal curve to describe beliefs. It's also handy to know the area under the normal curve between μ and σ.]**

(A) Adapt the code from Section 3.5.2 (`IntegralOfDensity.R`) to determine (approximately) the probability mass under the normal curve from $x = \mu - \sigma$ to $x = \mu + \sigma$. Comment your code. Hint: Just change `xlow` and `xhigh` appropriately, and change the `text` location so that the `area` still appears within the plot.

(B) Now use the normal curve to describe the following belief. Suppose you believe that women's heights follow a bell-shaped distribution, centered at 162cm with about two-thirds of all women having heights between 147cm and 177cm.

Exercise 3.5. **[Purpose: Recognize and work with the fact that Equation 3.11 can be solved for the conjoint probability, which will be crucial for developing Bayes' theorem.]** School children were surveyed regarding their favorite foods. Of the total sample, 20% were 1st graders, 20% were 6th graders, and 60% were 11th graders. For each grade, the following table shows the proportion of respondents that chose each of three foods as their favorite:

	Ice Cream	**Fruit**	**French Fries**
1st Graders	.3	.6	.1
6th Graders	.6	.3	.1
11th Graders	.3	.1	.6

From that information, construct a table of conjoint probabilities of grade and favorite food. Also, say whether grade and favorite food are independent or not, and how you ascertained the answer. Hint: You are given p(grade) and p(food|grade). You need to determine p(grade,food).

Chapter 4

Bayes' Rule

Contents

I'll love you forever in every respect
(I'll marginalize all your glaring defects)
But if you could change some to be more like me
I'd love you today unconditionally.

If you see that there are clouds, what is the probability that soon there will be rain? If you know that it is raining, by hearing it patter on the roof, what is the probability that there are clouds? Notice that p(clouds|rain) is not equal to p(rain|clouds). If someone smiles at you, what is the probability that they love you? If someone loves you, what is the probability that they will smile at you? Notice that p(smile|love) is not equal to p(love|smile).

Let's consider an example for which we can determine specific numbers. Suppose I have a standard deck of playing cards, which has 52 cards altogether. There are four suits: hearts, diamonds, clubs and spades. Within each suit, there are thirteen values: ace, two,

three, ..., ten, jack, queen, king. I shuffle the cards and draw one at random without showing it to you. I look at the card, and tell you (truthfully) that it is a queen. Given that you know it is a queen, what is the probability that it is a heart? Think about it a moment: There are 4 queens in the deck, and only one of them is a heart. So the probability that the card is a heart is 1/4. We can write this as a conditional probability: $p(\heartsuit|Q) = \frac{1}{4}$.

Now I put the card back into the deck and reshuffle. I draw another card from the deck, and this time I tell you that it is a heart. Given that you know it is a heart, what is the probability that it is a queen? Think about it a moment: There are 13 hearts in the deck, and only one of them is a queen. So the probability that the card is a queen is 1/13. We can write this as a conditional probability: $p(Q|\heartsuit) = \frac{1}{13}$.

Notice that $p(\heartsuit|Q)$ does not equal $p(Q|\heartsuit)$. Despite the inequality, the reversed conditional probabilities must be related somehow, right? Answer: Yes! What Bayes' rule tells us is the relationship between the two conditional probabilities.

4.1 Bayes' rule

Thomas Bayes (1702-1761) was a reputable mathematician and Presbyterian minister in England. His famous theorem was published posthumously in 1764. The simple rule that relates conditional probabilities has vast ramifications for statistical inference, and therefore as long as his name is attached to the rule, we'll continue to see his name in textbooks.

A crucial application of Bayes' rule is to determine the probability of a model when given a set of data. What the model itself provides is the probability of the data, given specific parameter values and the model structure. We use Bayes' rule to get from the probability of the data, given the model, to the probability of the model, given the data. This process will all be explained during the course of this chapter, and, indeed, during the rest of this book.

There is another branch of statistics, called null hypothesis significance testing (NHST), which relies on the probability of data given the model and does *not* use Bayes' rule. Chapter 11 describes NHST and its perils. This approach is often identified with another towering figure from England who lived about 200 years later than Bayes, named Ronald Fisher (1890-1962). His name, or at least the first letter of his last name, is immortalized in the most common statistic used in NHST, the F-ratio.[1] It is curious and re-assuring that the overwhelmingly dominant approach of the 20th century, i.e. NHST, is giving way in the 21st century to a Bayesian approach that had its genesis in the 18th century.

4.1.1 Derived from definitions of conditional probability

Recall from the definition of conditional probability, back in Equations 3.11 and 3.12 on p. 38, that $p(y|x) = p(y, x)/p(x)$. In words, the definition simply says that the probability of y given x is the probability that they happen together relative to the probability that x happens at all. We used this definition quite naturally when computing the conditional probabilities for the example, above, regarding hearts and queens in a deck of cards.

Now we just do some very simple algebraic manipulations. First, multiply both sides of $p(y|x) = p(y, x)/p(x)$ by $p(x)$ to get $p(y|x)p(x) = p(y, x)$. Second, notice that we can do the analogous manipulation starting with $p(x|y) = p(y, x)/p(y)$ to get $p(x|y)p(y) = p(y, x)$. Now

[1] But Fisher did not advocate the type of NHST ritual that contemporary social science performs; see Gigerenzer, Krauss, and Vitouch (2004).

we have two different expressions equal to $p(y, x)$, so we know those expressions equal each other: $p(y|x)p(x) = p(x|y)p(y)$. Divide both sides of that last expression by $p(x)$ to arrive at

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{4.1}$$

But we are not quite done yet, because we can re-write the denominator in terms of $p(x|y)$ also. Toward that goal, recall that $p(x) = \sum_y p(x, y)$. That was Equation 3.9, p. 37, if you're keeping score. We also know that $p(x, y) = p(x|y)p(y)$. Combining those equations yields $p(x) = \sum_y p(x, y) = \sum_y p(x|y)p(y)$. Substitute that into the denominator of Equation 4.1 to get

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}. \tag{4.2}$$

In Equation 4.2, the y in the numerator is a specific fixed value, whereas the y in the denominator is a variable that takes on all possible values of y over the summation. Equations 4.1 or 4.2 are called *Bayes' rule*. This simple relationship lies at the core of Bayesian inference.

4.1.2 Intuited from a two-way discrete table

It's easy to derive Bayes' rule (we just did!), but let's now get an intuition for what it means and how it works. First, let's confirm that it works for the simple case of the queen of hearts. Earlier we figured out that $p(Q|\heartsuit) = \frac{1}{13}$ and $p(\heartsuit|Q) = \frac{1}{4}$. Do those conditional probabilities satisfy Bayes' rule? Let's find out: $p(\heartsuit|Q)p(Q)/p(\heartsuit) = \frac{1}{4}\frac{4}{52}/\frac{13}{52} = \frac{1}{13} = p(Q|\heartsuit)$. It works!

The suit and value on playing cards are independent. (The idea of independent attributes was discussed in Section 3.4.3.) Let's now confirm Bayes' rule for two attributes that are not independent. Recall the case of tossing a coin three times and counting the number of heads and the number of switches between heads and tails, as tabulated back in Table 3.3 (p. 37), and repeated here for convenience:

	# Heads				**Marginal**
# Switches	**0**	**1**	**2**	**3**	**(# Switches)**
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
Marginal (# Heads):	1/8	3/8	3/8	1/8	

Consider the probability of getting one switch given that there is one head, i.e., $p(1S|1H)$, versus the probability of getting one head given that that there is one switch, i.e., $p(1H|1S)$. From the table, we can determine that $p(1S|1H) = p(1S, 1H)/p(1H) = (2/8)/(3/8) = 2/3$, and $p(1H|1S) = p(1H, 1S)/p(1S) = (2/8)/(4/8) = 1/2$. Notice that $p(1S|1H)$ does not equal $p(1H|1S)$. Then we can verify Bayes' rule: $p(1H|1S)p(1S)/p(1H) = (1/2)(4/8)/(3/8) = 2/3 = p(1S|1H)$. It works! In going through that arithmetic, essentially what we did was go through the motions of deriving Bayes' rule, using specific values instead of variables.

A valuable intuition, for understanding conditional probabilities and Bayes' rule, comes from restricting our spatial attention to a single row or column of the conjoint probability table. Suppose someone tosses a coin three times, and tells us that the sequence contains 1 switch. Given that knowledge, we can restrict our attention the row of the table corresponding to 1 switch. We know that one of the conjoint events *within that row* must

have happened, but we don't know which one. The relative probabilities of events within that row have not changed, but we know that the total probability within that row must now sum to 1.0. To achieve that transformation mathematically, we simply divide the cell probabilities in the 1-switch row by its original row total. This preserves the relative probabilities within the row, but makes the total probability equal to 1.0. Dividing a set of values by their sum is called "normalizing" the values. When we normalize the cell probabilities in the i^{th} row, we get the conditional probabilities of the columns, given the row value. In particular, when we normalize the 1-switch row, we get the conditional probabilities for number of heads: $p(0H|1S) = 0/(4/8) = 0$, $p(1H|1S) = (2/8)/(4/8) = .5$, $p(2H|1S) = (2/8)/(4/8) = .5$, and $p(3H|1S) = 0/(4/8) = 0$.

The idea of restricting attention to a single column or row of the conjoint probability table yields a way of intuiting Bayes' rule in general. The key to Bayes' rule is to notice, from the definition of conditional probability (Equations 3.11 and 3.12 on p. 38), that the conjoint probability of the i^{th} row (R_i) and the j^{th} column (C_j) can be re-expressed either as $p(R_i|C_j)p(C_j)$ or as $p(C_j|R_i)p(R_i)$. These alternative expressions of the conjoint probability $p(R_i, C_j)$ have been entered into the i, j^{th} cell of Table 4.1.

Table 4.1: A table for making Bayes' rule not merely special but spatial.

	Column			
Row	...	**j**	...	**Marginal**
⋮		⋮		
i	...	$p(R_i, C_j)$ $= p(R_i\|C_j)p(C_j)$ $= p(C_j\|R_i)p(R_i)$	...	$p(R_i)$
⋮		⋮		
Marginal:		$p(C_j)$		

Suppose we know that event R_i has happened, but we don't know the column value. In this case, the remaining possibilities are the cells in row R_i, and therefore we can restrict our attention to only the i^{th} row of Table 4.1. Because we know that R_i is true, our universe of remaining possibilities has collapsed to that row, and therefore we know that the sum of the probabilities in the row must be 1, instead of $p(R_i)$. This promotion of $p(R_i)$ to 1.0 is mathematically like dividing everything in the i^{th} row by $p(R_i)$. As mentioned before, this operation is called normalizing the probabilities in the i^{th} row so they sum to 1.0. When we normalize, the equation in the i, j^{th} cell becomes $p(R_i, C_j)/p(R_i) = p(R_i|C_j)p(C_j)/p(R_i) = p(C_j|R_i)$. This is Bayes' rule.

In summary, the key idea is that conditionalizing on a known row value is like restricting attention to only the row for which that known value is true, and then normalizing the probabilities in that row by dividing by the row's total probability. This act of spatial attention, when expressed in algebra, yields Bayes' rule.

Of course, the same relationship applies to columns instead of rows. It is arbitrary which attribute to place down the rows and which attribute to place across the columns. Thus, the analogous spatial relationship applies to columns: If we know the column value, then we restrict attention to that column, and normalize the cell probabilities to yield Bayes' rule again.

4.1.3 The denominator as an integral over continuous values

Up to this point, Bayes' rule has been presented only in the context of discrete-valued variables. It also applies to continuous variables, but probability masses become probability densities and sums become integrals. For continuous variables, Bayes' rule (Equation 4.2) becomes

$$p(y|x) = \frac{p(x|y)p(y)}{\int dy\, p(x|y)p(y)}. \tag{4.3}$$

In Equation 4.3, the y in the numerator is a specific fixed value, whereas the y in the denominator is a variable that takes on all possible values of y over the integral. It is this continuous-variable version of Bayes' rule that we will deal with most often.

4.2 Applied to models and data

One of the key applications that makes Bayes' rule so useful is when the row and column variables are data values and model parameter values, respectively. A model specifies the probability of particular data values given the model's structure and particular parameter values. For example, our usual model of coin flips says that $p(\text{datum} = H|\theta) = \theta$ and $p(\text{datum} = T|\theta) = 1 - \theta$. More generally, a model specifies

$$p(\text{ data values } |\text{ parameters values and model structure }) .$$

We use Bayes' rule to convert that to what we really want to know, which is how strongly we should believe in the model, given the data:

$$p(\text{ parameters values and model structure } |\text{ data values }) .$$

When we have observed some data, we use Bayes' rule to determine our beliefs across competing parameter values in a model, and to determine our beliefs across competing models.

It helps to think about the application of Bayes' rule to data and models in terms of a two-way table, shown in Table 4.2. The columns of Table 4.2 correspond to specific values of the model parameter, and the rows of Table 4.2 correspond to specific values of the data. Each cell of the table holds the conjoint probability of the specific combination of parameter value θ and data value D. That is, $p(D, \theta)$ is the probability of getting that particular combination of data value and parameter value, across all possible combinations of data values and parameter values.

Table 4.2: Applying Bayes' rule to data and model parameter.

	Model Parameter			
Data		θ value		**Marginal**
		$\vdots$		
D value	$\cdots$	$p(D,\theta)$ $= p(D\|\theta)p(\theta)$ $= p(\theta\|D)p(D)$	$\cdots$	$p(D)$
		$\vdots$		
Marginal:		$p(\theta)$		

The prior probability of the parameter values is the marginal distribution, $p(\theta)$, which appears in the lower margin of Table 4.2. This is simply the probability of each possible value of θ, collapsed across all possible values of data.

When we observe a particular data value, D, so we know it is true, we are restricting our attention to one specific row of Table 4.2, namely, the row corresponding to the observed value, D. The posterior distribution on θ is obtained by dividing the conjoint probabilities in that row by the row marginal, $p(D)$. Thus, the posterior probability of θ is just the conjoint probabilities in that row, normalized by $p(D)$ to sum to 1.

We need to define some notation and terms at this point. The factors of Bayes' rule have names as indicated below:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(D|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}} / \underbrace{p(D)}_{\text{evidence}} \tag{4.4}$$

where the evidence is (from the denominator of Equation 4.3)

$$p(D) = \int d\theta \, p(D|\theta)p(\theta). \tag{4.5}$$

The "prior," $p(\theta)$, is the strength of our belief in θ without the data D. The "posterior," $p(\theta|D)$, is the strength of our belief in θ when the data D have been taken into account. The "likelihood," $p(D|\theta)$, is the probability that the data could be generated by the model with parameter values θ. The "evidence," $p(D)$, is the probability of the data according to the model, determined by summing across all possible parameter values weighted by the strength of belief in those parameter values.

We talk about parameter values θ only in the context of a particular model; it's the model that gives meaning to the parameter. In some applications it can help to make the model explicit in Bayes' rule. Let's call the model M. Then, because all the probabilities are defined given that model, we can re-write Equation 4.4 as

$$\underbrace{p(\theta|D,M)}_{\text{posterior}} = \underbrace{p(D|\theta,M)}_{\text{likelihood}} \underbrace{p(\theta|M)}_{\text{prior}} / \underbrace{p(D|M)}_{\text{evidence}} \tag{4.6}$$

where the evidence is

$$p(D|M) = \int d\theta \, p(D|\theta,M)p(\theta|M). \tag{4.7}$$

It's especially handy to have the model explicitly annotated as in Equation 4.6 when you have more than one model in mind and you're using the data to help determine the strength of belief in each model. Suppose we have two models, creatively named $M1$ and $M2$. Then, by Bayes' rule, $p(M1|D) = p(D|M1)p(M1)/p(D)$ and $p(M2|D) = p(D|M2)p(M2)/p(D)$, where $p(D) = \sum_i p(D|M_i)p(M_i)$. Taking the ratio of those equations, we get

$$\frac{p(M1|D)}{p(M2|D)} = \underbrace{\frac{p(D|M1)}{p(D|M2)}}_{\text{Bayes factor}} \frac{p(M1)}{p(M2)}. \tag{4.8}$$

Equation 4.8 says that the ratio of the posterior beliefs is the ratio of the evidences (as defined in Equation 4.7) times the ratio of the prior beliefs. The ratio of the evidences is called the *Bayes factor*. Examples of all these abstract terms will be provided soon.

Terminological aside: The quantity $p(D|M)$, which is called the "evidence" in this book, is sometimes instead called the "marginal likelihood" or "prior predictive" by other authors. The term "evidence" is common in the machine learning literature (e.g., Bishop, 2006; MacKay, 2003). Whenever I refer to the "evidence" for a model, I am referring to $p(D|M)$ as defined in Equation 4.7. This usage might be a little confusing in the context of model comparison when considering the equation $p(M1|D) = p(D|M1)p(M1)/p(D)$, where $p(D|M1)$ plays the *role* of the likelihood, not the evidence. This apparent confusion is cleared up when abbreviated terminology is expanded to its full specificity. The factor $p(D|M)$ is not merely "the evidence", it is "the evidence for model M". On the other hand, the factor $p(D)$, in the context of the equation $p(M1|D) = p(D|M1)p(M1)/p(D)$, is not the evidence for *a* model, but is the evidence for the entire *set* of models under consideration: $p(D) = \sum_i p(D|M_i)p(M_i)$. The term "likelihood" also deserves expansion. In Equation 4.6, the likelihood is more fully stated as "the likelihood of parameter value θ in model M for data D". That is, the likelihood is referring to the parameter θ. On the other hand, in the context of model comparison, the factor $p(D|M1)$, in the equation $p(M1|D) = p(D|M1)p(M1)/p(D)$, is the "likelihood of the *model* $M1$ for the data D". To re-iterate, the term "evidence" is merely a word to refer to $p(D|M)$. As we will see below, its value does not have much meaning by itself. Instead, $p(D|M)$ can only be interpreted in the context of other models.

4.2.1 Data order invariance

One more nuance about Bayesian updating of beliefs. Bayes' rule in Equation 4.4 gets us from a prior belief, $p(\theta)$, to a posterior belief, $p(\theta|D)$, when we take into account some data. Now suppose we observe some *more* data, which we'll denote D'. We can then update our beliefs again, from $p(\theta|D)$ to $p(\theta|D', D)$. Here's the question: Does our final belief depend on whether we update with D first and D' second, or update with D' first and D second?

The answer is: It depends! In particular, it depends on the model function that defines the likelihood, $p(D|\theta)$. In many models, $p(D|\theta)$ does not depend in any way on other data. That is, the conjoint probability $p(D, D'|\theta)$ equals $p(D|\theta)p(D'|\theta)$. The data probabilities are independent, according to this type of model. Moreover, in many models the probability function does not change in time or depend on how many data values have been generated. The probability function is stationary. Under these conditions, when $p(D|\theta)$ and $p(D'|\theta)$ are *independent and identically distributed* (commonly referred to as "i.i.d."), then the order of updating has no effect of the final posterior.

This invariance to ordering of the data makes sense intuitively: If the likelihood function has no dependence on time or data ordering, then the posterior shouldn't have any dependence on time or data ordering either! But it's easy to prove mathematically, too. First, we'll unpack $p(\theta|D', D)$ by applying Bayes' rule on D':

$$p(\theta|D', D) = \frac{p(D'|\theta, D)\, p(\theta|D)}{\int d\theta\, p(D'|\theta, D)\, p(\theta|D)}$$

Now, notice that $p(D'|\theta, D) = p(D'|\theta)$, because the model asserts that the probability of a data value depends only on the value of θ and not on anything else, such as other data. Therefore the equation above can be re-written as

$$p(\theta|D', D) = \frac{p(D'|\theta)\, p(\theta|D)}{\int d\theta\, p(D'|\theta)\, p(\theta|D)}$$

Now we use Bayes' rule again, this time for $p(\theta|D)$, which converts the equation into

$$p(\theta|D', D) = \frac{p(D'|\theta)\, p(D|\theta)\, p(\theta)/p(D)}{\int d\theta\, p(D'|\theta)\, p(D|\theta)\, p(\theta)/p(D)}$$

Notice that $p(D)$ in that equation is a constant, and cancels out. This last equation, above, involves the product of $p(D'|\theta)$ and $p(D|\theta)$. Because multiplication can be done in either order (i.e., it is "commutative" in technical terminology), we arrive at the same formula if we start with the data in the opposite order: $p(\theta|D, D')$.

In all of the examples in this book, the likelihood functions generate i.i.d. data. One way of thinking about this assumption is as follows: We assume that every datum is equally representative of the underlying process, regardless of when the datum was observed. Older observations are just as valid and representative as more recent observations, and the underlying process that generates the data has not changed during the course of making the observations.

4.2.2 An example with coin flipping

With all the emphasis on coin flipping, by now you must be imagining flipping coins over pasture fences as you try to fall asleep. Nevertheless, imagine flipping coins once again, and try not to fall asleep. We will start with some prior beliefs about the possible bias of the coin, then flip the coin a few times, and then update our beliefs based on the observed flips.

First, we specify our prior beliefs. We denote the bias as $\theta = p(H)$, the probability of the coin coming up heads. To keep the example straightforward, suppose that we believe there are only three possible values for the coin's bias: Either the coin is fair, with $\theta = .50$, or the coin is biased with $\theta = .25$ or $\theta = .75$. We believe that the coin is probably fair, but there's some smaller chance it could be biased high or low. This prior probability is graphed in the top panel of Figure 4.1. It shows three "spikes," one over each value of θ that we think could be possible. The spike over $\theta = .5$ is tallest, indicating that we believe it to be most likely. Note that the heights of the spikes are probability masses, not densities, because each spike indicates the probability of its specific, discrete value of θ.

Next, we flip the coin to get some data, D, and determine the likelihood, $p(D|\theta)$. Suppose we flip the coin 12 times, and it comes up heads 3 times. According to our model of the coin, the probability of coming up heads is θ, and the probability of coming up tails is $1 - \theta$. Moreover, the flips are independent of each other, and therefore we can multiply the probabilities of the individual flips to get the probability of the combination of flips. Consequently, the probability of a specific sequence of 3 heads and 9 tails is $p(D|\theta) = \theta^3(1-\theta)^9$. The resulting likelihood for each value of θ is plotted in the middle panel of Figure 4.1. Notice that the likelihood is highest for $\theta = .25$ and lowest for $\theta = .75$. This peak at $\theta = .25$ makes sense, because the data have 25% heads, and so they are more likely if $\theta = .25$ than if $\theta = .50$ or $\theta = .75$. The value of θ that maximizes the likelihood is called the *maximal likelihood estimate* of θ.

The lower panel of Figure 4.1 includes the value of $p(D|M)$, the evidence for the model. Recall from Equation 4.7 that the evidence is the overall probability of the data, averaging across the available parameter values weighted by the degree to which we believe in them: $p(D|M) = \sum_\theta p(D|\theta, M)p(\theta, M)$. This is the normalizer for the posterior distribution, hence it is displayed in the plot of the posterior distribution. The value is displayed as $p(D)$ instead of as $p(D|M)$ because there is only one model in this context, and therefore the M notation is suppressed. When you see the value of $p(D)$ in Figure 4.1, you might think that $p(D)$

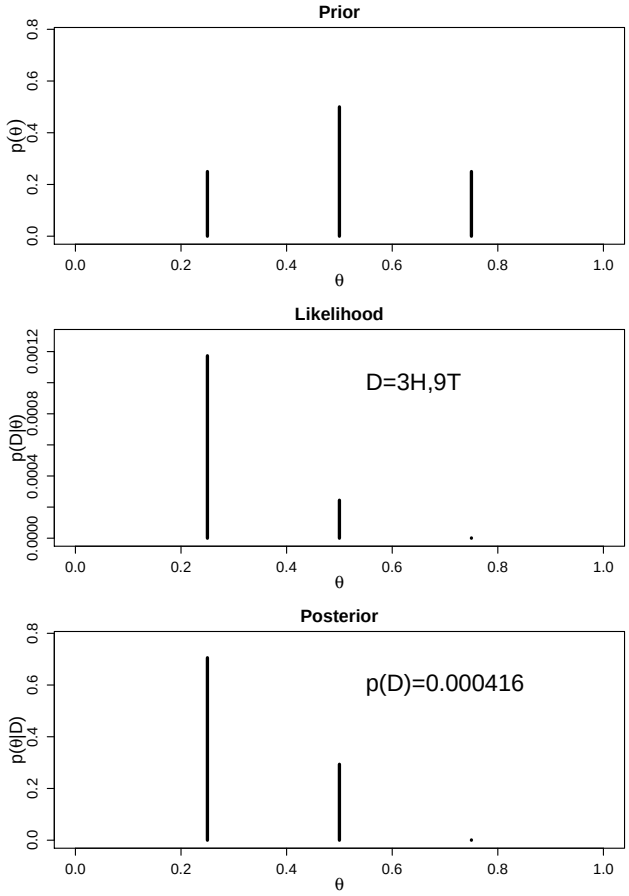


Figure 4.1: Bayesian updating of beliefs about the bias of a coin. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. (The R code that generated this graph is in Section 4.4.1 (`BayesUpdate.R`).)

is terribly small, until you remember that we are talking about the conjoint probability of several things happening together, i.e., exactly the 12 flips we observed. The probability of 1 head is θ. The probability of 2 heads is θ^2, which is smaller than θ. The probability of 3 heads is θ^3, which is smaller yet. As the set of data D gets bigger, in terms of containing more observations, $p(D)$ gets smaller, regardless of how closely the model θ matches the true bias in the coin.

The bottom panel of Figure 4.1 displays the posterior beliefs for each value of θ. According to Bayes' rule, the posterior is proportional to the product of the prior and the likelihood. So the shape of the posterior is influenced by both the prior and the likelihood. You can see this dual influence in Figure 4.1 by inspecting the relative heights of the left and middle spikes. In the prior, the middle spike is much taller than the left spike. In the likelihood, the middle spike is much shorter than the left spike. In the posterior, there is a compromise between the prior and the likelihood: The middle spike is shorter than the left spike, but not so short as in the likelihood because it (i.e., the middle spike) is buoyed up by the prior. Notice how our beliefs have changed from prior to posterior. Initially we believed most strongly in a fair coin. After accounting for the data, we believed most strongly in a biased coin. The Bayesian mathematics let us compute exactly how much our beliefs changed.

4.2.2.1 $p(D|\theta)$ is not θ

In the examples involving coin flips, it is easy to lose sight of the important fact that $p(D|\theta)$ is different from θ, even though they both are values between 0 and 1 for our current examples. The likelihood $p(D|\theta)$ is a mathematical function of θ. The value of the likelihood function is always a probability (a probability mass if θ has a finite number of values, and a probability density otherwise). The value of the parameter, however, could be on any scale, depending on the meaning of the parameter. In our examples so far, the meaning of the parameter is itself a probability, so it is easy to confuse the parameter value with the likelihood value. Adding to the confusability is the fact that, in our examples so far, the function that maps θ to $p(D=H|\theta)$ has been the identity function:

$$p(D=H|\theta) = \theta \tag{4.9}$$

and, of course, $p(D=T|\theta) = 1.0 - p(D=H|\theta) = 1.0 - \theta$. It is easy to confuse $p(D|\theta)$ with θ in our examples because the function that relates them is the identity. Later in the book, we will see many examples for which the likelihood function is not the identity function.

The point of this subsection has been to remind you that θ is a parameter that has a scale and meaning in the context of a model. The value $p(D|\theta)$, on the other hand, is a probability, and is a function of the parameter θ. Thus, $p(D|\theta)$ and θ are quite distinct entities, despite the fact that in simple models of coin flipping, $p(D=H|\theta) = \theta$.

4.3 The three goals of inference

Back in Section 2.2 (p. 13), I introduced three goals of inference: Estimation of parameter values, prediction of data values, and model comparison. Each of these goals will now be given precise mathematical expressions.

4.3.1 Estimation of parameter values

Estimation of parameter values means determining the extent to which we believe in each possible parameter value. This is precisely what Equation 4.6 tells us. The posterior distribution over the parameter values θ *is* our estimate of those values.

The posterior distribution can be narrow, with most of the probability piled heavily over a small range of θ. In this case, we are fairly certain about the possible values of θ. On the other hand, the posterior probability distribution could be wide, spread over a large range of θ. In this case, we have high uncertainty about the possible values of θ.

4.3.2 Prediction of data values

Using our current beliefs, we may want to predict the probability of future data values. To avoid notational conflicts later, I'll denote a data value as y. The predicted probability of data value y is determined by averaging the predicted data probabilities across all possible parameter values, weighted by the belief in the parameter values:

$$p(y) = \int d\theta \, p(y|\theta)p(\theta)$$

Notice that this is exactly the evidence, discussed after Equation 4.4, except that the evidence refers to a specific observed value of y, whereas here we are computing the probability of any possible value of y.

As an example, consider the prior beliefs in the top panel of Figure 4.1. For those beliefs, the predicted probability of getting a head is

$$\begin{aligned} p(y=H) &= \sum_{\theta} p(y=H|\theta)p(\theta) \\ &= p(y=H|\theta=0.25)p(\theta=0.25) \\ &\quad + p(y=H|\theta=0.50)p(\theta=0.50) \\ &\quad + p(y=H|\theta=0.75)p(\theta=0.75) \\ &= 0.25 \times 0.25 + 0.50 \times 0.50 + 0.75 \times 0.25 \\ &= 0.5 \end{aligned}$$

and the probability of getting a tail is computed analogously to be $p(y=T) = 0.5$. Notice that the predictions are probabilities of each possible data value, given the current model beliefs.

If we want to predict a particular point value for the next datum, instead of a distribution across all possible data values, it is typical to use the mean, i.e., expected value, of the predicted data distribution. Thus, the predicted value is $\overline{y} = \int dy\, y\, p(y)$. This integral only makes sense if y is on a continuum. If y is nominal, like the result of a coin flip, then the most probable value can be used as "the" predicted value. The decision to use the mean of the predicted values as our single best prediction, instead of, say, the mode or median, relies implicitly on the costs of being wrong and the benefits of being correct. These costs and benefits, called the "utilities", are considered in more advanced treatments of Bayesian decision theory. For our purposes, we will default to the mean, purely for convenience.

4.3.3 Model comparison

You may recall from earlier discussion (page 48) that Bayes' rule is also useful for comparing models. Equation 4.8 indicated that the posterior beliefs in the models involve the evidences of the models. Notice that in this third goal, i.e., model comparison, the evidence term appears again, just as it appeared for the goals of parameter estimation and data prediction.

One of the nice features of Bayesian model comparison is that there is an automatic accounting for model complexity when assessing the degree to which we should believe in the model. This might be best explained with an example. Recall the coin-flipping example discussed earlier, illustrated in Figure 4.1, and reproduced in the left side of Figure 4.2. In that example, we supposed that the bias θ could take on only three possible values. This restriction made the model rather simple. We could instead entertain a more complex model that allows for many more possible values of θ. One such model is illustrated in the right side of Figure 4.2. This model has 63 possible values of θ instead of only 3. The shape of the prior beliefs in the complex model follows the same triangular shape as in the simple model; there is highest belief in $\theta = .50$, with lesser belief in more extreme values.

The complex model has many more available values for θ, and so it has much more opportunity to fit arbitrary data sets. For example, if a sequence of coin flips has 37% heads, the simple model does not have a θ value very close to that outcome, but the complex model does. On the other hand, for θ values that are in both the simple and complex models, the prior probability on those values in the simple model is much higher than in the complex model. Because there are so many possibilities in the complex model, the prior beliefs have to get spread out, very shallowly, over a larger range of possibilities. This can be seen in

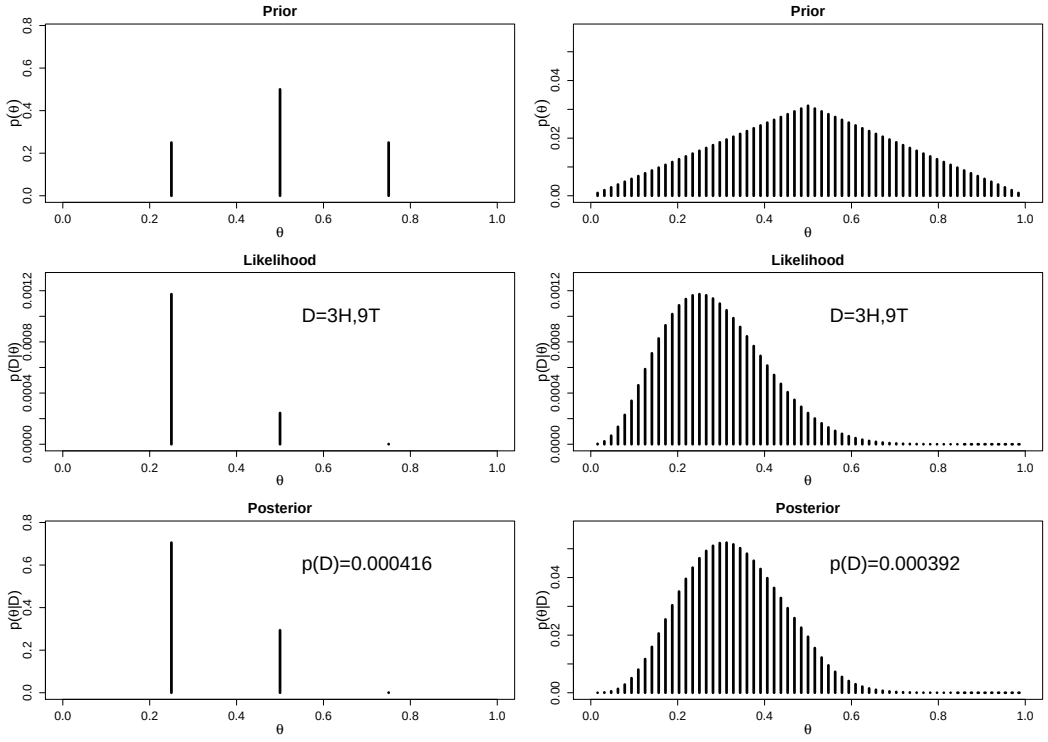


Figure 4.2: A simple in the left column and a complex model in the right column. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. The same data are addressed by both models. The evidence $p(D|M_{\text{simple}})$ for the simple model is displayed as p(D) in the lower left panel, and the evidence $p(D|M_{\text{complex}})$ for the complex model is displayed as p(D) in the lower right panel. In this case the data are such that the simple model is favored. The R code that generated these graphs is in Section 4.4.1 (`BayesUpdate.R`).

Figure 4.2 by inspecting the numerical scales on the vertical axes of the prior beliefs. The scale on the simple model is much larger than the scale on the complex model.

Therefore, if the actual data we observe happens to be well accommodated by a θ value in the simple model, we will believe in the simple model more than the complex model, because the prior on that θ value in the simple model is so high. Figure 4.2 shows a case in which this happens. The data have 25% heads, and so the evidence in the simple model is larger than the evidence in the complex model. The complex model has its prior spread too thin for us to believe in it as much as we believe in the simple model.

The complex model can be the winner if the data are not adequately fit by the simple model. For example, consider a case in which the observed data have just 1 head and 11 tails. None of the θ values in the simple model is very close to this outcome. But the complex model does have some θ values near the observed proportion, even though there is not a strong belief in those values. Figure 4.3 shows that the simple model has less evidence in this situation, and we have stronger belief in the complex model.

The evidence for a model, $p(D|M)$, is not particularly meaningful as an absolute magnitude for a single model. The evidence is most meaningful only in the context of the Bayes' factor, $p(D|M1)/p(D|M2)$, which is the *relative* evidence for two models, when considering

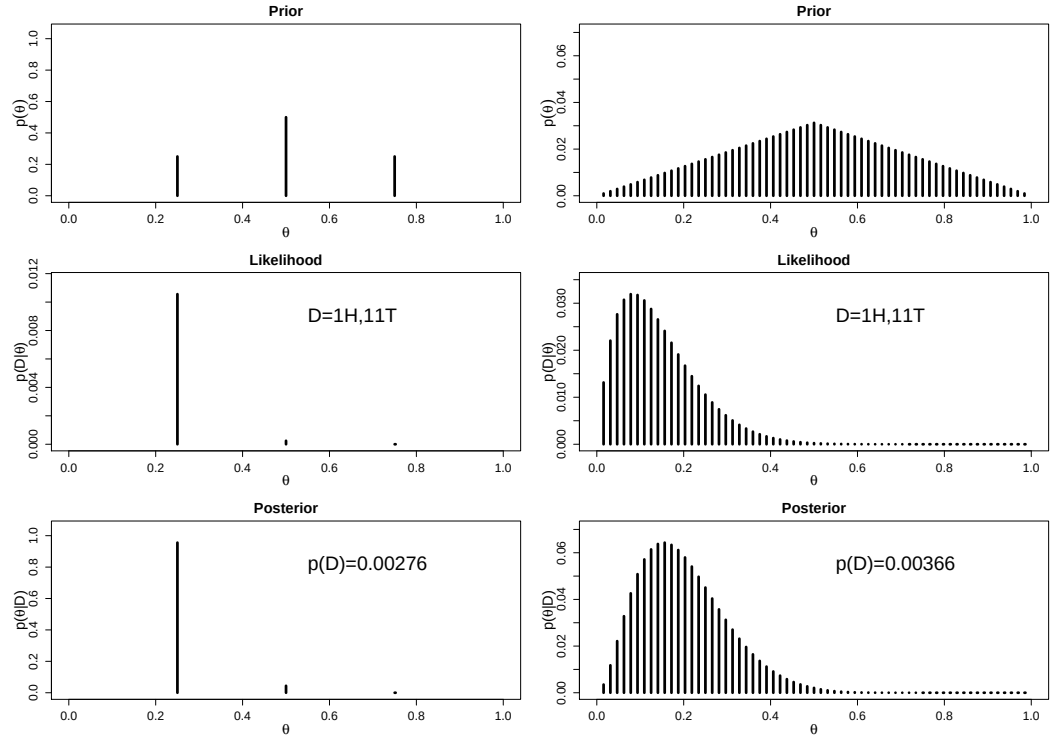


Figure 4.3: A simple in the left column and a complex model in the right column. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. The same data are addressed by both models. The evidence $p(D|M_{\text{simple}})$ for the simple model is displayed as p(D) in the lower left panel, and the evidence $p(D|M_{\text{complex}})$ for the complex model is displayed as p(D) in the lower right panel. In this case the data are such that the complex model is favored. The R code that generated these graphs is in Section 4.4.1 (`BayesUpdate.R`).

an observed data set D.[2] Regardless of which model wins, the winning model does not need to be a good model of the data. The model comparison process merely tells us about the *relative* evidence for each model. The winning model is better than the other models in the competition, but the winning model might merely be less bad than the horrible competitors. In later chapters we will explore ways to assess whether the winning model is actually a viable model of the data.

We will see in Chapter 10 that Bayesian model comparison is "really" just a case of Bayesian parameter estimation, in which a parameter that indexes the models is estimated. The individual model parameters depend on the indexical parameter, and thus the scheme involves a hierarchy of dependencies. Hierarchial models are introduced in Chapter 9. The fact that model comparison is a case of parameter estimation is mentioned here only to fend off any mistaken impression that parameter estimation and model comparison are fundamentally different.

[2]The Bayes' factor, $p(D|M1)/p(D|M2)$, is quite different than considering evidences of a single model for different candidate data sets. Specifically, $p(D1|M)/p(D2|M)$ is *not* a Bayes' factor and is not further discussed.

4.3.4 Why Bayesian inference can be difficult

All three goals involve the denominator of Bayes' formula, i.e., the evidence, which usually means computing a difficult integral. There are a few ways out of this difficulty. The traditional way is to use likelihood functions with "conjugate" prior functions. A prior function that is conjugate to the likelihood function simply makes the posterior function come out with the same functional form as the prior. That is, the math works out nicely. If this method doesn't work, an alternative is to approximate the actual functions with other functions that are easier to work with, and then show that the approximation is reasonably good under typical conditions. But this method is still pure, analytical mathematics. Yet another method is to numerically approximate the integral. When the parameter space is small, then it can be covered with a comb or grid of points and the integral can be computed by exhaustively summing across that grid. But when the parameter space gets even moderately large, there are too many grid points, and therefore other methods must be used. A large class of random sampling methods have been developed, which can be referred to as Markov chain Monte Carlo (MCMC) methods, that can numerically approximate probability distributions even for large spaces. It is the development of these MCMC methods that has allowed Bayesian statistical methods to gain practical use. The next major part of this book explains these various methods in some detail. For applications to complex situations, we will ultimately focus on MCMC methods.

Another potential difficulty of Bayesian inference is determining a reasonable prior. What distribution of beliefs should we start with, over all possible parameter values or over competing models? This question may seem daunting, but in practice it is typically addressed in straightforward manner. As will be discussed more in Chapter 11, it is actually advantageous and rational to start with an explicit prior. Prior beliefs *should* influence rational inference from data, because the role of new data is to modify our beliefs from whatever they were without the new data. Prior beliefs are *not* capricious and idiosyncratic and unknowable, but instead are based on publicly agreed facts and theories. Prior beliefs used in data analysis must be admissible by a skeptical scientific audience. When scientists disagree about prior beliefs, the analysis can be conducted with both priors, to assess the robustness of the posterior against changes in the prior. Or, the priors can be mixed together into a joint prior, with the posterior thereby incorporating the uncertainty in the prior. In summary, for most applications, specification of the prior turns out to be technically *un*problematic, although it is conceptually very important to understand the consequences of one's assumptions about the prior. Thus, the main reason that Bayesian analysis can be difficult is the computation of the evidence, and that computation is tractable in many situations via MCMC methods.

4.3.5 Bayesian reasoning in everyday life

4.3.5.1 Holmesian deduction

Despite the difficulty of exact Bayesian inference in complex mathematical models, the essence of Bayesian reasoning is frequently used in everyday life. One example has been immortalized in the words of Sherlock Holmes to his friend Dr. Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Arthur Conan Doyle, The Sign of Four, 1890, Ch. 6). This reasoning is actually a consequence of Bayesian belief updating, as expressed in Equation 4.4. Let me re-state it this way: "How often have I said to you that when $p(D|\theta_i) = 0$ for

all $i \neq j$, then, no matter how small the prior $p(\theta_j) > 0$ is, the posterior $p(\theta_j|D)$ must equal one." Somehow it sounds better the way Holmes said it. The intuition behind Holmes' deduction is clear, though: When we reduce belief in some possibilities, we necessarily increase our belief in the remaining possibilities (*if* our set of possibilities exhausts all conceivable options). Thus, according to Holmesian deduction, when the data make some options less believable, we increase belief in the other options.

4.3.5.2 Judicial exoneration

The reverse of Holmes' logic is also commonplace. For example, when an object d'art is found fallen from its shelf, our prior beliefs may indict the house cat, but when the visiting toddler is seen dancing next to the shelf, then the cat is exonerated. This downgrading of a hypothesis is sometimes called "explaining away" of a possibility by verifying a different one. This sort of exoneration also follows from Bayesian belief updating: When $p(D|\theta_j)$ is higher, then, even if $p(D|\theta_i)$ is unchanged for all $i \neq j$, $p(\theta_i|D)$ is lower. This logic of exoneration is based on competition of mutually exclusive possibilities: If the culprit is suspect A, then suspect B is exonerated.

Holmesian deduction and judicial exoneration are both expressions of the essence of Bayesian reasoning: We have a space of beliefs that are mutually exclusive and exhaust all possibilities. Therefore, if the data cause us to decrease belief in some possibilities, we must increase belief in other possibilities (as said Holmes), or, if the data cause us to increase belief in some possibilities, we must decrease belief in other possibilities (as in exoneration). What Bayes' rule tells us is exactly how much to shift our beliefs across the available possibilities.

4.4 R code

4.4.1 R code for Figure 4.1

There are several new commands used in this program. When you encounter a puzzling command in an R program, it usually helps to try the R `help` command. For example, when perusing this code, you'll come across the command `matrix`. To find out about the syntax and usage of this command, do this: At the R command line, type `help("matrix")` and you'll get some clues about how it works. Then experiment with the command at the interactive command line until you're confident about what its various arguments do. For example, try typing at the command line:

```
matrix( 1:6 , nrow=2 , ncol=3 , byrow=TRUE )
```

Then try

```
matrix( 1:6 , nrow=2 , ncol=3 , byrow=FALSE )
```

The listing below includes line numbers in the margins, to facilitate tracking the code across page splits, and to facilitate referring to specific lines of the code when you have enthusiastic conversations about it at parties.

Mac users: If you are running R under MacOS instead of in a Windows emulator such as WINE, you will need to change all the `windows()` commands to `quartz()`. Later in the book, when we use BUGS, there is no Mac equivalent and you must run the programs under WINE.

(BayesUpdate.R)

```
# Theta is the vector of candidate values for the parameter theta.
# nThetaVals is the number of candidate theta values.
# To produce the examples in the book, set nThetaVals to either 3 or 63.
nThetaVals = 3
# Now make the vector of theta values:
Theta = seq( from = 1/(nThetaVals+1) , to = nThetaVals/(nThetaVals+1) ,
             by = 1/(nThetaVals+1) )

# pTheta is the vector of prior probabilities on the theta values.
pTheta = pmin( Theta , 1-Theta ) # Makes a triangular belief distribution.
pTheta = pTheta / sum( pTheta )  # Makes sure that beliefs sum to 1.

# Specify the data. To produce the examples in the book, use either
# Data = c(1,1,1,0,0,0,0,0,0,0,0,0) or Data = c(1,0,0,0,0,0,0,0,0,0,0,0).
Data = c(1,1,1,0,0,0,0,0,0,0,0,0)
nHeads = sum( Data == 1 )
nTails = sum( Data == 0 )

# Compute the likelihood of the data for each value of theta:
pDataGivenTheta = Theta^nHeads * (1-Theta)^nTails

# Compute the posterior:
pData = sum( pDataGivenTheta * pTheta )
pThetaGivenData = pDataGivenTheta * pTheta / pData   # This is Bayes' rule!

# Plot the results.
windows(7,10) # create window of specified width,height inches.
layout( matrix( c( 1,2,3 ) ,nrow=3 ,ncol=1 ,byrow=FALSE ) ) # 3x1 panels
par(mar=c(3,3,1,0))          # number of margin lines: bottom,left,top,right
par(mgp=c(2,1,0))            # which margin lines to use for labels
par(mai=c(0.5,0.5,0.3,0.1)) # margin size in inches: bottom,left,top,right

# Plot the prior:
plot( Theta , pTheta , type="h" , lwd=3 , main="Prior" ,
      xlim=c(0,1) , xlab=bquote(theta) ,
      ylim=c(0,1.1*max(pThetaGivenData)) , ylab=bquote(p(theta)) ,
      cex.axis=1.2 , cex.lab=1.5 , cex.main=1.5 )

# Plot the likelihood:
plot( Theta , pDataGivenTheta , type="h" , lwd=3 , main="Likelihood" ,
      xlim=c(0,1) , xlab=bquote(theta) ,
      ylim=c(0,1.1*max(pDataGivenTheta)) , ylab=bquote(paste("p(D|",theta,")")),
      cex.axis=1.2 , cex.lab=1.5 , cex.main=1.5 )
text( .55 , .85*max(pDataGivenTheta) , cex=2.0 ,
      bquote( "D=" * .(nHeads) * "H," * .(nTails) * "T" ) , adj=c(0,.5) )

# Plot the posterior:
plot( Theta , pThetaGivenData , type="h" , lwd=3 , main="Posterior" ,
      xlim=c(0,1) , xlab=bquote(theta) ,
      ylim=c(0,1.1*max(pThetaGivenData)) , ylab=bquote(paste("p(",theta,"|D)")),
      cex.axis=1.2 , cex.lab=1.5 , cex.main=1.5 )
text( .55 , .85*max(pThetaGivenData) , cex=2.0 ,
      bquote( "p(D)=" * .(signif(pData,3)) ) , adj=c(0,.5) )
```

4.5 Exercises

Exercise 4.1. **[Purpose: Application of Bayes' rule to disease diagnosis, to see the important role of prior probabilities.]** Suppose that in the general population, the probability of having a particular rare disease is one in a thousand. We denote the true presence or absence of the disease as the value of a parameter, θ, that can have the value $\theta = \text{ä}$ if disease is present, or the value $\theta = \text{ü}$ if the disease is absent. The base rate of the disease is therefore denoted $p(\theta = \text{ä}) = .001$. This is our prior belief that a person selected at random has the disease.

Suppose that there is a test for the disease that has a 99% hit rate, which means that if a person has the disease, then the test result is positive 99% of the time. We denote a positive test result as $D = +$, and a negative test result as $D = -$. The observed test result is a bit of data that we will use to modify our belief about the value of the underlying disease parameter. The hit rate is expressed as $p(D = + \,|\, \theta = \text{ä}) = .99$. The test also has a false alarm rate of 5%. This means that 5% of the time when the disease is not present, the test falsely indicates that the disease is present. We denote the false alarm rate as $p(D = + \,|\, \theta = \text{ü}) = .05$.

Suppose we sample a person at random from the population, administer the test, and it comes up positive. What is the posterior probability that the person has the disease? Mathematically expressed, we are asking, what is $p(\theta = \text{ä} \,|\, D = +)$? Before determining the answer from Bayes' rule, generate an intuitive answer and see if your intuition matches the Bayesian answer. Most people have an intuition that the probability of having the disease is near the hit rate of the test (which in this case is .99).

Hint: The following table of conjoint probabilities might help you understand the possible combinations of events. (The table below is a specific case of Table 4.2, p. 47.) The prior probabilities of the disease are on the bottom marginal. When we know that the test result is positive, we restrict our attention the row marked $D = +$.

	$\theta = \text{ä}$	$\theta = \text{ü}$	
$D = +$	$p(D=+, \theta=\text{ä})$ $= p(D=+ \,\|\, \theta=\text{ä})\, p(\theta=\text{ä})$	$p(D=+, \theta=\text{ü})$ $= p(D=+ \,\|\, \theta=\text{ü})\, p(\theta=\text{ü})$	$p(D=+)$
$D = -$	$p(D=-, \theta=\text{ä})$ $= p(D=- \,\|\, \theta=\text{ä})\, p(\theta=\text{ä})$	$p(D=-, \theta=\text{ü})$ $= p(D=- \,\|\, \theta=\text{ü})\, p(\theta=\text{ü})$	$p(D=-)$
	$p(\theta = \text{ä})$	$p(\theta = \text{ü})$	

Caveat regarding interpreting the results: Remember that here we have assumed that the person was selected at random from the population; there were no other symptoms that motivated getting the test.

Exercise 4.2. **[Purpose: Iterative application of Bayes' rule, and seeing how posterior probabilities change with inclusion of more data.]** Continuing from the previous exercise, suppose that the same randomly selected person as in the previous exercise gets re-tested after the first test comes back positive, and on the re-test the result is negative. Now what is the probability that the person has the disease? *Hint: For the prior probability of the re-test, use the posterior computed from the previous exercise.* Also notice that $p(D = - \,|\, \theta = \text{ä}) = 1 - p(D = + \,|\, \theta = \text{ä})$ and $p(D = - \,|\, \theta = \text{ü}) = 1 - p(D = + \,|\, \theta = \text{ü})$.

Exercise 4.3. **[Purpose: Getting an intuition for the previous results by using "natural frequency" and "Markov" representations]**

(A) Suppose that the population consists of 100,000 people. Compute how many people should fall into each cell of the table in the Hint of Exercise 4.1. To compute the expected

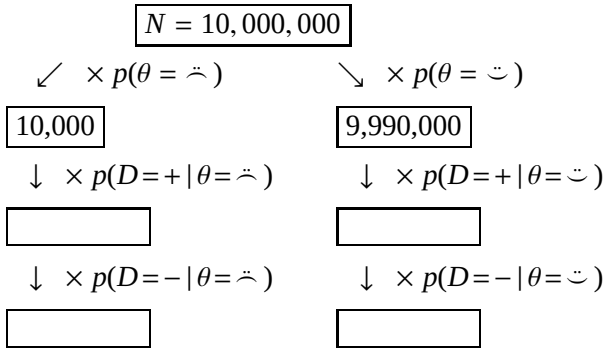
frequency of people in a cell, just multiply the cell probability by the size of the population. To get you started, a few of the cells of the frequency table are filled in here:

	$\theta = \ddot{\frown}$	$\theta = \ddot{\smile}$	
$D = +$	$\text{freq}(D=+, \theta=\ddot{\frown})$ $= p(D=+, \theta=\ddot{\frown})\, N$ $= p(D=+ \mid \theta=\ddot{\frown})\, p(\theta=\ddot{\frown})\, N$ $= 99$	$\text{freq}(D=+, \theta=\ddot{\smile})$ $= p(D=+, \theta=\ddot{\smile})\, N$ $= p(D=+ \mid \theta=\ddot{\smile})\, p(\theta=\ddot{\smile})\, N$ $=$	$\text{freq}(D=+)$ $= p(D=+)\, N$ $=$
$D = -$	$\text{freq}(D=-, \theta=\ddot{\frown})$ $= p(D=-, \theta=\ddot{\frown})\, N$ $= p(D=- \mid \theta=\ddot{\frown})\, p(\theta=\ddot{\frown})\, N$ $= 1$	$\text{freq}(D=-, \theta=\ddot{\smile})$ $= p(D=-, \theta=\ddot{\smile})\, N$ $= p(D=- \mid \theta=\ddot{\smile})\, p(\theta=\ddot{\smile})\, N$ $=$	$\text{freq}(D=-)$ $= p(D=-)\, N$ $=$
	$\text{freq}(\theta = \ddot{\frown})$ $= p(\theta = \ddot{\frown})\, N$ $= 100$	$\text{freq}(\theta = \ddot{\smile})$ $= p(\theta = \ddot{\smile})\, N$ $= 99,900$	N $= 100,000$

Notice the frequencies on the lower margin of the table. They indicate that out of 100,000 people, only 100 have the disease, while 99,900 do not have the disease. These marginal frequencies instantiate the prior probability that $p(\theta = \ddot{\frown}) = .001$. Notice also the cell frequencies in the column $\theta = \ddot{\frown}$, which indicate that of 100 people with the disease, 99 have a positive test result and 1 has a negative test result. These cell frequencies instantiate the hit rate of .99. Your job for this part of the exercise is to fill in the frequencies of the remaining cells of the table.

(B) Take a good look at the frequencies in the table you just computed for the previous part. These are the so-called "natural frequencies" of the events, as opposed to the somewhat unintuitive expression in terms of conditional probabilities (Gigerenzer & Hoffrage, 1995). From the cell frequencies alone, determine the proportion of people who have the disease, given that their test result is positive. Before computing the exact answer arithmetically, first give a rough intuitive answer merely by looking at the relative frequencies in the row $D = +$. Does your intuitive answer match the intuitive answer you provided back in Exercise 4.1? Probably not. Your intuitive answer here is probably much closer to the correct answer. Now compute the exact answer arithmetically. It should match the result from applying Bayes' rule in Exercise 4.1.

(C) Now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumulate more data. This type of representation is called a "Markov" representation by Krauss, Martignon, and Hoffrage (1999). Suppose now we start with a population of $N = 10,000,000$ people. We expect 99.9% of them (i.e., 9,990,000) not to have the disease, and just 0.1% (i.e., 10,000) to have the disease. Now consider how many people we expect to test positive. Of the 10,000 people who have the disease, 99%, i.e. 9,900, will be expected to test positive. Of the 9,990,000 people who do not have the disease, 5%, i.e. 499,500, will be expected to test positive. Now consider re-testing everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest? Use this diagram to compute your answer:



When computing the frequencies for the empty boxes above, be careful to use the proper conditional probabilities!

(D) Use the diagram in the previous part to answer this: What proportion of people, who test positive at first and then negative on retest, actually have the disease? In other words, of the total number of people at the bottom of the diagram in the previous part (those are the people who tested positive then negative), what proportion of them are in the left branch of the tree? *How does the result compare with your answer to Exercise 4.2?*

Exercise 4.4. **[Purpose: To see a hands-on example of data-order invariance.]** Consider again the disease and diagnostic test of the previous two exercises. Suppose that a person selected at random from the population gets the test and it comes back negative. Compute the probability that the person has the disease. The person then gets re-tested, and on the the second test the result is positive. Compute the probability that the person has the disease. *How does the result compare with your answer to Exercise 4.2?*

Exercise 4.5. **[Purpose: An application of Bayes' rule to neuroscience: Inferring cognitive function from brain activation.]** Cognitive neuroscientists investigate which areas of the brain are active during particular mental tasks. In many situations, researchers will observe that a certain region of the brain is active, and infer that a particular cognitive function is therefore being carried out. Poldrack (2006) cautioned that such inferences are not necessarily very firm, and need to be made with Bayes' rule in mind. Poldrack (2006) reported the following frequency table of previous studies that involved any language-related task (specifically phonological and semantic processing), and whether or not a particular region of interest (ROI) in the brain was activated:

	Language Study	*Not Language Study*
Activated	166	199
Not Activated	703	2154

Suppose that a new study is conducted and finds that the ROI is activated. If the prior probability that the task involves language processing is .5, what is the posterior probability, given that the ROI is activated? (Hint: Poldrack (2006) reports that it is 0.69. You job is to derive this number.)

Exercise 4.6. **[Purpose: To make sure you really understand what is being shown in Figure 4.1.]** Derive the posterior distribution in Figure 4.1 by hand. The prior has $p(\theta = .25) = .25$, $p(\theta = .50) = .50$, and $p(\theta = .75) = .25$. The data consist of a specific sequence of flips with 3 heads and 9 tails, so $p(D|\theta) = \theta^3 (1 - \theta)^9$. Hint: Check that your posterior probabilities sum to one.

Exercise 4.7. **[Purpose: For you to see, hands on, that $p(D)$ lives in the denominator of Bayes' rule.]** Compute $p(D)$ in Figure 4.1 by hand. Hint: Did you notice that you already computed $p(D)$ in the previous exercise?

Part II

All the Fundamentals Applied to Inferring a Binomial Proportion

Chapter 5

Inferring a Binomial Proportion via Exact Mathematical Analysis

Contents

I built up my courage to ask her to dance
By drinking too much before taking the chance.
I fell on my butt when she said see ya later;
Less priors might make my posterior beta.

I built up my courage to ask her to dance
I drank too much before taking the chance
I fell on my butt; she said see ya later
A beta prior makes posterior beta

This Part of the book addresses a simple scenario: Estimating the underlying probability that a coin comes up heads. The methods don't require that we are referring to a coin, of course. All we require in this scenario is that the space of possibilities for each datum has just two possible values that are mutually exclusive. These two values have no ordinal or metric relationship with each other, they are just nominal values. Because there are two nominal values, I refer to this sort of data as "binomial", or sometimes as "dichotomous".

We also assume that each datum is independent of the others and that the underlying probability is stationary through time. Coin flipping is the standard example of this situation: There are two possible outcomes (head or tail), the flips are (we assume) independent of each other, and the probability of getting a head is stationary through time (again, by assumption). Other examples include the proportion of free throws hit by a player in basketball, the proportion of babies born that are girls, the proportion of heart surgery patients

who survive more than a year after surgery, the proportion of people who agree with a statement on a survey, the proportion of widgets on an assembly line that are faulty, and so on. While we talk about heads and tails for coins, keep in mind that the methods could be applied to many other interesting real-world situations.

In a Bayesian analysis, we begin with some prior beliefs over possible probabilities of the coin coming up heads. Then we observe some data that consist of a set of results from flipping the coin. Then we infer the posterior distribution of our beliefs using Bayes' rule. Bayes' rule requires us to specify the likelihood function, and that is the topic of the next section.

5.1 The likelihood function: Bernoulli distribution

When we flip a coin, the result can be a head or a tail. We will denote the result by y, with $y = 1$ for head and $y = 0$ for tail. Giving the head or tail a numerical value (i.e., 1 or 0) is helpful for mathematically expressing the probabilities. But do not be lulled into thinking that somehow a head is "greater than" a tail because $1 > 0$, or that the "distance" between a head and a tail is 1 because $|1 - 0| = 1$. We will use $y = 1$ for head and $y = 0$ for tail only for convenience, but we must remember that the data are truly nominal (i.e., categorical) values without any metric or ordinal properties.

As discussed previously in Section 4.2.2.1, p. 52, the probability of the coin coming up heads is a function of an underlying parameter: $p(y=1|\theta) = f(\theta)$. We assume a particularly simple function, namely the identity: $p(y=1|\theta) = \theta$. Consequently, the probability of tails is the complement, i.e., $p(y = 0|\theta) = 1 - \theta$. These two equations can be combined into a single expression as follows:

$$p(y|\theta) = \theta^y \, (1 - \theta)^{(1-y)} \tag{5.1}$$

for y in the set $\{1, 0\}$ and θ in the interval $[0, 1]$. Notice that when $y = 1$, the righthand side of Equation 5.1 reduces to θ, and when $y = 0$, the righthand side of Equation 5.1 reduces to $1 - \theta$.

The formula in Equation 5.1 expresses the *Bernoulli distribution*. The Bernoulli distribution is a probability distribution over the two discrete values of y, for any fixed value of θ. In particular, the sum of the probabilities is 1, as must be true of a probability distribution: $\sum_y p(y|\theta) = p(y=1|\theta) + p(y=0|\theta) = \theta + (1-\theta) = 1$.

Another perspective on Equation 5.1 is to think of the data value y as fixed by an observation, and the value of θ as variable. Equation 5.1 then specifies the probability of the fixed y value if θ has some particular value. Different values of θ yield different probabilities of the datum y. When thought of in this way, Equation 5.1 is the *likelihood function* of θ.

Notice that the likelihood function is a function of a continuous value θ, whereas the Bernoulli distribution is a discrete distribution over the two values of y. The likelihood function, though it specifies a probability at each value of θ, is *not* a probability distribution. In particular, it does not integrate to 1. For example, suppose that $y = 1$. Then $\int_0^1 d\theta \, \theta^y \, (1 - \theta)^{(1-y)} = \int_0^1 d\theta \, \theta = \frac{1}{2} \neq 1$.

In Bayesian inference, the function $p(y|\theta)$ is usually thought of with the data, y, known and fixed, and the parameter, θ, uncertain and variable. Therefore, $p(y|\theta)$ is usually called the likelihood function for θ, and Equation 5.1 is called the *Bernoulli likelihood function*. Don't forget, however, that the very same function is also the probability of the datum, y.

When we flip the coin N times, we have a set of data, $D = \{y_1, \ldots, y_N\}$, where each y_i is 0 or 1. By assumption, each flip is independent of the others. (Recall the definition of independence from Section 3.4.3, p. 39.) Therefore, the probability of getting the set of N flips $D = \{y_1, \ldots, y_N\}$ is the product of the individual outcome probabilities:

$$\begin{aligned} p(\{y_1, \ldots, y_N\}|\theta) &= \prod_i p(y_i|\theta) \\ &= \prod_i \theta^{y_i} (1-\theta)^{(1-y_i)} \end{aligned} \tag{5.2}$$

If the number of heads in the set of flips is denoted $z = \sum_i^N y_i$, then Equation 5.2 can be written as

$$p(z, N|\theta) = \theta^z (1-\theta)^{(N-z)} \tag{5.3}$$

I will often lapse terminologically sloppy and refer to Equation 5.3 as the Bernoulli likelihood function for a set of flips, but please remember that the Bernoulli distribution is really Equation 5.1 and refers to a single flip.[1]

5.2 A description of beliefs: The beta distribution

In this chapter, we use purely mathematical analysis, with no numerical approximation, to derive the mathematical form of the posterior distribution of beliefs. To do this, we need a mathematical description of our prior beliefs. That is, we need a mathematical formula that describes the prior belief probability for each value of the bias θ in the interval $[0, 1]$.

In principle, we could use any probability density function supported on the interval $[0, 1]$. When we intend to apply Bayes' rule (Equation 4.4), however, there are two desiderata for mathematical tractability. First, it would be convenient if the product of $p(y|\theta)$ and $p(\theta)$, which is in the numerator of Bayes' rule, results in a function of the same form as $p(\theta)$. When this is the case, the prior and posterior beliefs are described using the same form of function. This quality allows us to include subsequent additional data and derive another posterior distribution, again of the same form as the prior. Therefore, no matter how much data we include, we always get a posterior of the same functional form. Second, we desire the denominator of Bayes' rule, namely $\int d\theta\, p(y|\theta)p(\theta)$, to be solvable analytically. This quality also depends on how the form of the function $p(\theta)$ relates to the form of the function $p(y|\theta)$. When the forms of $p(y|\theta)$ and $p(\theta)$ combine so that the posterior distribution has the same form as the prior distribution, then $p(\theta)$ is called a *conjugate prior* for $p(y|\theta)$. Notice that the prior is conjugate only with respect to a particular likelihood function.

In the present situation we are seeking a functional form for a prior density over θ that is conjugate to the Bernoulli likelihood function in Equation 5.1. If you think about it a minute, you'll notice that if the prior is of the form $\theta^a(1-\theta)^b$, then when you multiply the Bernoulli likelihood with the prior, you'll again get a function of the same form, namely

[1] Some readers might be familiar with the binomial distribution, $p(z|N, \theta) = \binom{N}{z}\theta^z (1-\theta)^{(N-z)}$, and wonder why it is not used here. The reason is that here we are considering each flip of the coin to be a distinct event, whereby each observation has just two possible values, $y \in \{0, 1\}$. The probability of the *set* of events is then the product of the individual event probabilities, as in Equation 5.2. If we instead considered a single "event" to be the flipping of N coins, then an observation of a *single* event could have $N + 1$ possible values, $z \in \{0, 1, ..., N\}$, and the probability of those values would be given by the binomial distribution. The binomial distribution is explained in Section 11.1.1, p. 216.

$\theta^{(y+a)}(1-\theta)^{(1-y+b)}$. So, to express the prior beliefs over θ, we seek a probability density function involving $\theta^a(1-\theta)^b$.

A probability density of that form is called a *beta distribution*. Formally, a beta distribution has two parameters, called a and b, and the density itself is defined as

$$\begin{aligned} p(\theta|a,b) &= \text{beta}(\theta;a,b) \\ &= \theta^{(a-1)}\,(1-\theta)^{(b-1)}/B(a,b) \end{aligned} \tag{5.4}$$

where $B(a,b)$ is simply a normalizing constant that ensures that the area under the beta density integrates to 1.0, as all probability density functions must. In other words, the normalizer for the beta distribution is $B(a,b) = \int_0^1 d\theta\,\theta^{(a-1)}\,(1-\theta)^{(b-1)}$.

Remember that the beta distribution is only defined for values of θ in the interval $[0,1]$. The values of a and b must be positive; zero and negative values don't work. Notice that in the definition of the beta distribution (Equation 5.4), the value of θ is raised to the power $a-1$, not the power a, and the value of $(1-\theta)$ is raised to the power $b-1$, not the power b. Be careful to distinguish the beta *function*, $B(a,b)$, from the beta *distribution*, $\text{beta}(\theta;a,b)$. The beta function is not a function of θ because θ has been "integrated out." In the programming language R, $\text{beta}(\theta;a,b)$ is `dbeta(`θ`,a,b)`, and $B(a,b)$ is `beta(a,b)`.[2]

Examples of the beta distribution are shown in Figure 5.1. Each panel of Figure 5.1 shows the beta distribution for particular values of a and b, as indicated inside each panel. Notice that as a gets bigger, the bulk of the distribution moves rightward over higher values of θ, but as b gets bigger, the bulk of the distribution moves leftward over lower values of θ. Notice that as a and b get bigger together, the beta distribution gets narrower.

5.2.1 Specifying a beta prior

We would like to specify a beta distribution that describes our prior beliefs. For this goal it is useful to know the mean and variance (recall Equations 3.6, p. 32, and 3.8, p. 33) of the beta distribution, so we can get a sense of which a and b values correspond to reasonable descriptions of our prior beliefs about θ. It turns out that the mean of the $\text{beta}(\theta;a,b)$ distribution is $\overline{\theta} = a/(a+b)$. Thus, when $a = b$, the mean is .5, and the bigger a is relative to b, the bigger the mean is. The standard deviation of the beta distribution is $\sqrt{\overline{\theta}(1-\overline{\theta})/(a+b+1)}$. Notice that the standard deviation gets smaller when $a+b$ gets larger.

You can think of a and b in the prior as if they were previously observed data, in which there were a heads and b tails in a total of $a+b$ flips. For example, if we have no prior knowledge other than the knowledge that the coin has a head side and a tail side, that's tantamount to having previously observed one head and one tail, which corresponds to $a = 1$ and $b = 1$. You can see in Figure 5.1 that when $a = 1$ and $b = 1$ the beta distribution is uniform: All values of θ are equally probable. As another example, if we think that the coin is probably fair but we're not very sure, then we can imagine that the previously observed data had, say, $a = 4$ heads and $b = 4$ tails. You can see in Figure 5.1 that when $a = 4$ and $b = 4$ the beta distribution is peaked at $\theta = .5$, but higher or lower values of θ are moderately probable too.

[2] Whereas it is true that $B(a,b) = \int_0^1 d\theta\,\theta^{(a-1)}\,(1-\theta)^{(b-1)}$, the beta function can also be expressed as $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, where Γ is the *Gamma function*: $\Gamma(a) = \int_0^\infty dt\,t^{(a-1)}\exp(-t)$. The Gamma function is a generalization of the factorial function, because, for integer valued a, $\Gamma(a) = (a-1)!$. In R, $\Gamma(a)$ is `gamma(a)`. Many sources define the beta function this way, in terms of the Gamma function.

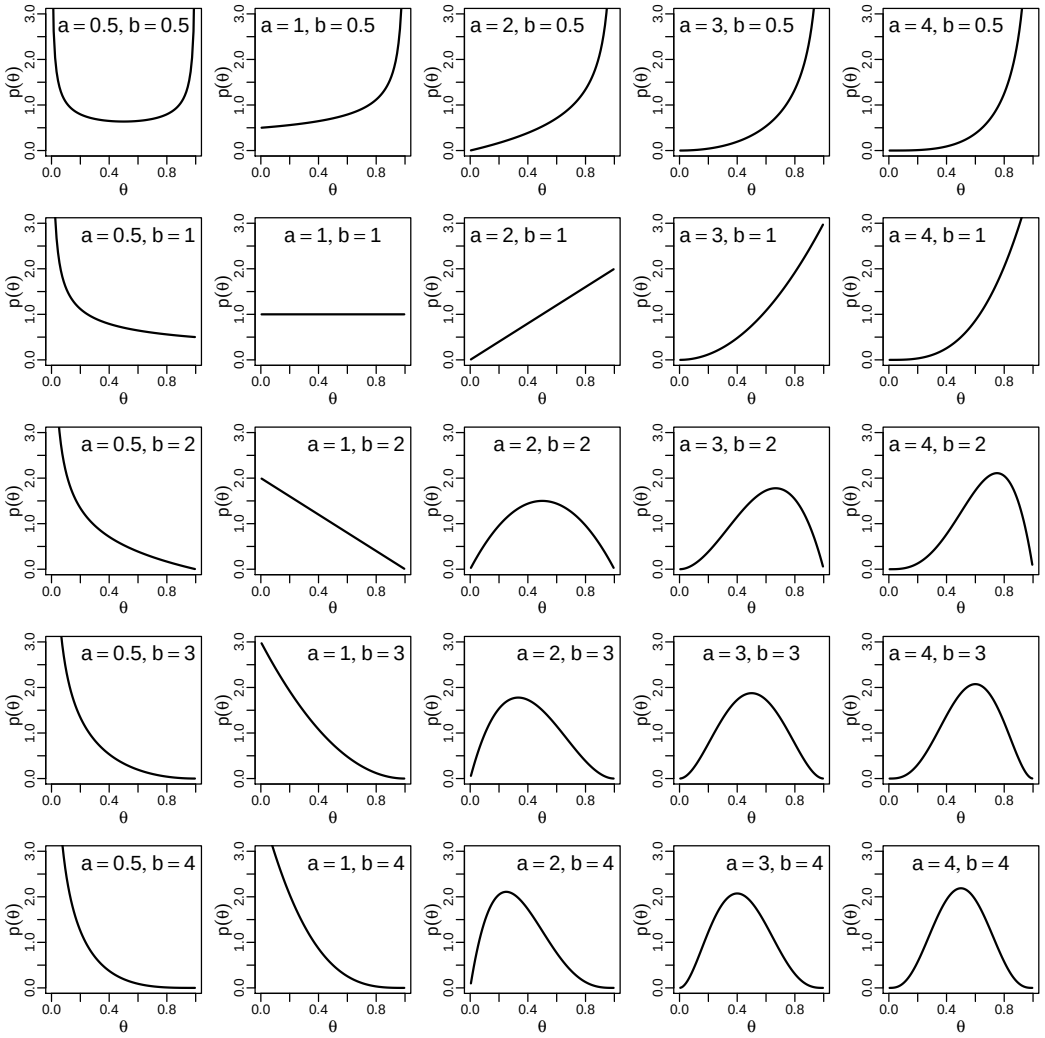


Figure 5.1: Examples of beta distributions.

Instead of thinking in terms of a heads and b tails in the prior data, it's easier to think in terms of the mean proportion of heads in the prior data and its sample size. The mean proportion of heads is $m = a/(a + b)$ and the sample size is $n = a + b$. Solving those two equations for a and b yields:

$$a = mn \quad \text{and} \quad b = (1 - m)n \tag{5.5}$$

where m is our guess for the prior mean value of the proportion θ, and n is our guess for the number of observations girding our prior belief. The value we choose for the prior n can be thought of this way: It is the number of new flips of the coin that we would need to make us teeter between the new data and the prior belief about m. If we would only need a few new flips to sway our beliefs, then our prior beliefs should be represented by a small n. If we would need a large number of new flips to sway us away from our prior beliefs about m, then our prior beliefs are worth a very large n. For example, suppose that I think the coin is fair, so $m = .5$, but I'm not highly confident about it, so maybe I imagine I've seen only $n = 8$ previous flips. Then $a = mn = 4$ and $b = (1 - m)n = 4$, which, as we saw before, is a beta distribution peaked at $\theta = .5$ and with higher or lower values less probable.

Another way of establishing the shape parameters is by starting with the mean and standard deviation of the desired beta distribution. You must be careful with this approach, because the standard deviation must make sense in the context of a beta density. In particular, the standard deviation should typically be less than 0.289, which is the standard deviation of a uniform density. For a beta density with mean m and standard deviation s, the shape parameters are:

$$a = m\left(\frac{m(1-m)}{s^2} - 1\right) \quad \text{and} \quad b = (1-m)\left(\frac{m(1-m)}{s^2} - 1\right) \tag{5.6}$$

For example, if $m = .5$ and $s = 0.28867$, Equation 5.6 implies that $a = 1$ and $b = 1$. As another example, if $m = .5$ and $s = 0.1$, then $a = 12$ and $b = 12$; i.e., a beta(θ,12,12) density has a standard deviation of 0.1.

In most applications, we will deal with beta distributions for which $a \geq 1$ and $b \geq 1$, i.e., $n \geq 2$, which reflects prior knowledge that the coin has a head side and a tail side. There are some situations, however, in which it may be convenient to use beta distributions in which $a < 1$ and/or $b < 1$. For example, we might believe that the coin is a trick coin that nearly always comes up heads or nearly always comes up tails, but we don't know which. In this case, the bimodal beta(θ; .5, .5) prior might be a useful description of our prior belief, as shown in the top left panel of Figure 5.1. Exercise 5.4 has you explore this a bit more.

5.2.2 The posterior beta

Now that we have determined a convenient prior for the Bernoulli likelihood function, let's figure out exactly what the posterior distribution is when we apply Bayes' rule (Equation 4.4, p. 48). Suppose we have a set of data comprising N flips with z heads. Substituting the Bernoulli likelihood (Equation 5.3) and the beta prior distribution (Equation 5.4) into Bayes' rule yields

$$\begin{aligned} p(\theta|z,N) &= p(z,N|\theta)p(\theta)/p(z,N) \\ &= \theta^z\ (1-\theta)^{(N-z)}\ \ \theta^{(a-1)}\,(1-\theta)^{(b-1)} \,/\, [B(a,b)p(z,N)] \\ &= \theta^{((z+a)-1)}\ (1-\theta)^{((N-z+b)-1)} \Big/ \underbrace{[B(a,b)\,p(z,N)]}_{B(z+a,N-z+b)} \quad . \end{aligned} \tag{5.7}$$

In that sequence of equations, you probably followed the collection of powers of θ and of $(1-\theta)$, but you may have balked at the transition, underbraced in the denominator, from $B(a,b)p(z,N)$ to $B(z+a,N-z+b)$. This transition was not made via some elaborate analysis of integrals. Instead, the transition was made by simply thinking about what the normalizing factor for the numerator must be. The numerator is $\theta^{((z+a)-1)}\ (1-\theta)^{((N-z+b)-1)}$, which is the numerator of a beta($\theta; z+a, N-z+b$) distribution. For the function in Equation 5.7 to be a probability distribution, as it must be, the denominator must be the normalizing factor for the corresponding beta distribution.

In other words, Equation 5.7 says this: If the prior distribution is beta($\theta; a, b$), and the data have z heads in N flips, then the posterior distribution is beta($\theta; z+a, N-z+b$). The simplicity of that updating rule is one of the beauties of the mathematical approach to Bayesian inference.

It is also revealing to think about the relationship between the prior and posterior means. The prior mean of θ is $a/(a+b)$. The posterior mean is $(z+a)/[(z+a)+(N-z+b)] =$

$(z + a)/(N + a + b)$. The posterior mean can be algebraically re-arranged into a weighted average of the prior mean, $a/(a + b)$, and the data proportion, z/N:

$$\underbrace{\frac{z+a}{N+a+b}}_{\text{posterior}} = \underbrace{\frac{z}{N}}_{\text{data}} \underbrace{\frac{N}{N+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior}} \underbrace{\frac{a+b}{N+a+b}}_{\text{weight}} . \tag{5.8}$$

Equation 5.8 indicates that the posterior mean is always somewhere between the prior mean and the proportion in the data. The mixing weight on the prior mean has N in its denominator, and so it decreases as N increases. The mixing weight on the data proportion increases as N increases. So the more data we have, the less is the influence of the prior, and the posterior mean gets closer to the proportion in the data. In particular, when $N = a + b$, the mixing weights are .5, which indicates that the prior mean and the data proportion have equal influence in the posterior. This result echoes what was said earlier (Equation 5.5) regarding how to set a and b to represent our prior beliefs: The choice of prior n (which equals $a + b - 2$) should represent the size of the new data set that would sway us away from our prior toward the data proportion.

5.3 Three inferential goals

5.3.1 Estimating the binomial proportion

The posterior distribution over θ tells us exactly how much we believe in each possible value of θ. When the posterior is a beta distribution, we can make a graph of the distribution and see in glorious detail what our new beliefs look like. We can extract numerical details of the distribution by using handy functions in R.

Figure 5.2 shows examples of posterior beta distributions. Each column of graphs show a prior beta distribution, a likelihood graph, and the resulting posterior distribution. Both columns use the same data, and therefore have the same likelihood graphs. The columns have different priors, however, hence different posteriors. The prior in the left column is uniform, which represents a prior of tremendous uncertainty wherein any bias is equally believable. The prior in the right column loads most belief over a bias of $\theta = 0.5$, indicating a moderately strong prior belief that the coin is fair. As indicated in the graphs of the likelihood, the coin is flipped 14 times and comes up heads 11 times. The posterior beta distributions are graphed in the bottom row. You can see in the left column that when the prior is uniform, then the posterior exactly mirrors the likelihood. When the prior is loaded over $\theta = 0.5$, however, the posterior is only slightly shifted away from the prior toward the proportion of heads in the data. This small shift is a graphic depiction of the relationship expressed in Equation 5.8.

The posterior distribution indicates which values of θ are relatively more credible than others. One way of summarizing where the bulk of the posterior resides is with the highest density interval (HDI), which was introduced in Section 3.3.5. The 95% HDI is an interval that spans 95% of the distribution, such that every point inside the interval has higher believability than any point outside the interval. Figure 5.2 shows the 95% HDI in the two posteriors. You can see that the 95% HDI is fairly wide when the prior is uncertain, but narrower when the prior is more certain. Thus, in this case, the posterior inherits the relative uncertainty of the prior, and the width of the 95% HDI is one measure of uncertainty.

The 95% HDI is also one way for declaring which values of the parameter are deemed "credible". Suppose we want to decide whether or not a value of interest for θ is credible,

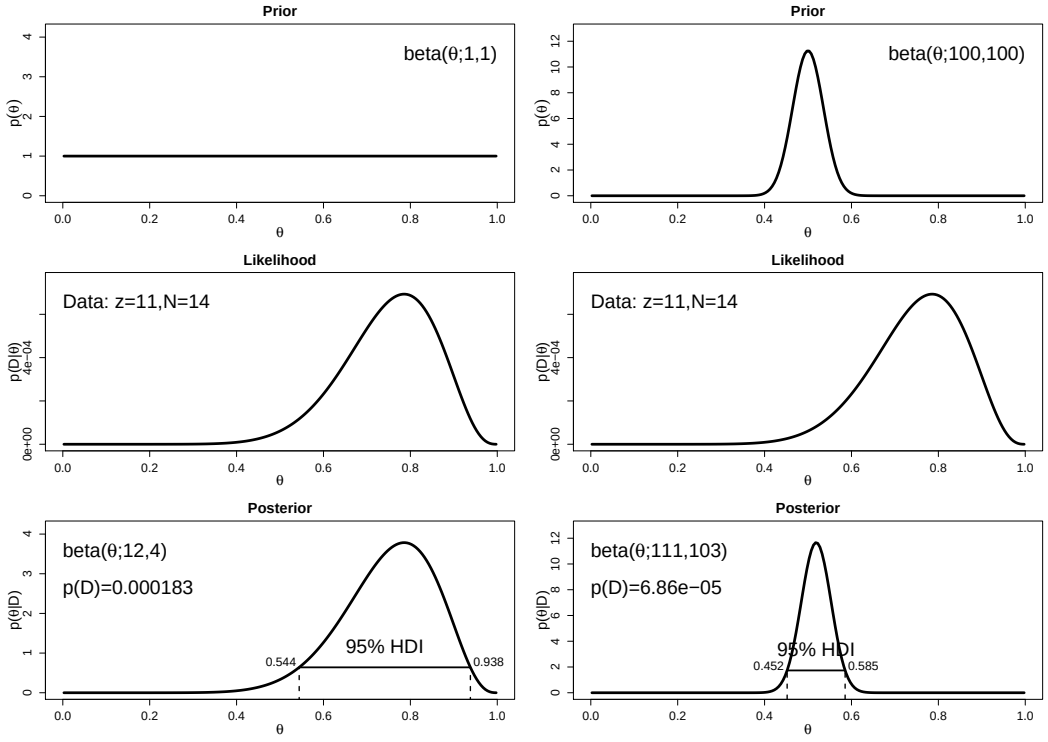


Figure 5.2: Two different beta priors, updated using the same data. (The R code that generated these graphs is in Section 5.5.1 (`BernBeta.R`).)

given the data. Consider, for example, $\theta = 0.5$, which indicates that the coin is fair. We first establish a *region of practical equivalence* (ROPE) around the value of interest, which means a small interval such that any value within the ROPE is equivalent to the value of interest for all practical purposes. Suppose that we declare the ROPE for $\theta = 0.5$ to be 0.48 to 0.52. We want to know if any values within the ROPE are reasonably credible, given the data. How should we define "reasonably credible"? One way is by saying that any points within the 95% HDI are reasonably credible. Hence, we use the following heuristic decision rule: A value of interest, such as $\theta = 0.5$, is declared to be incredible if no point in its ROPE falls within the 95% HDI of the posterior.

It is important to distinguish the two roles for the HDI just mentioned. One role for the HDI is acting as a summary of the distribution. A different role for the HDI is using it for deciding whether a value of interest is or is not credible. This latter process, of converting a rich posterior distribution to a discrete yes/no decision about credibility, involves many extra assumptions that have nothing to do with the HDI. The HDI can be a useful summary apart from whether or not it is used to decide the credibility of a point. These issues will be explored at length in Chapter 12.

5.3.2 Predicting data

As introduced back in Section 4.3.2 (p. 52), the predicted probability of a datum value y is determined by averaging that value's probability across all possible parameter values, weighted by the belief in the parameter values: $p(y) = \int d\theta\, p(y|\theta)p(\theta)$. The belief in the parameter values, $p(\theta)$, is the current posterior belief, including the data observed so far,

which we can indicate explicitly as $p(\theta|z, N)$.

In the present application, the predicted probability of heads is particularly simple, because $p(y=1|\theta) = \theta$, and therefore

$$\begin{aligned} p(y=1) &= \int d\theta\, p(y=1|\theta)\, p(\theta|z, N) \\ &= \int d\theta\, \theta\, p(\theta|z, N) \\ &= \overline{\theta}|z, N \\ &= (z + a)/(N + a + b) \end{aligned} \tag{5.9}$$

In other words, the predicted probability of heads is just the mean of the posterior distribution over θ. Recall from Equation 5.8 that the posterior mean is a weighted mixture of the prior mean and the data proportion. So the predicted probability of getting a head on the next flip is somewhere between the prior mean and the proportion of heads in the flips observed so far.

Let's make that concrete by considering a particular prior and sequence of flips. Suppose that we start with a uniform prior, i.e., beta$(\theta; 1, 1)$. We flip the coin once, and get a head. The posterior is then beta$(\theta; 2, 1)$, which has a mean of 2/3. Thus, after the first flip comes up heads, the predicted probability of heads on the next flip is 2/3. Suppose we flip the coin a second time, and again get a head. The posterior is then beta$(\theta; 3, 1)$, and the predicted probability of heads on the next flip is 3/4. Notice that even though we have flipped the coin twice and observed heads both times, we do not predict that there is 100% chance of coming up heads on the next flip, because the uncertainty of the prior beliefs is mixed with the observed data.

Consider a variation of that example in which we start with a prior of beta$(\theta; 50, 50)$, which expresses a fairly strong prior belief that the coin is fair (about 95% of the probability mass lies between $\theta = .40$ and $\theta = .60$). Suppose we flip the coin twice and get heads both times. The posterior is beta$(\theta; 52, 50)$, and hence the predicted probability of getting a head on the next flip is $52/102 \approx 51\%$ (which is much different than what we predicted when starting with a uniform prior). Because the prior $a + b$ was so large, it will take a large N to overpower the prior belief.

5.3.3 Model comparison

You may recall from the previous chapter (particularly Equation 4.8 on page 48) that Bayes' rule can be used to compare models. To do this, we compute the evidence, $p(D|M)$, for each model. The evidence is the weighted total probability of the newly observed data across all possible parameter values, where the total is computed by weighting each parameter value by its prior probability. That is, $p(D|M) = \int d\theta\, p(D|\theta, M)\, p(\theta|M)$.

In the present scenario, the data D are expressed by the values z and N. When using a Bernoulli likelihood and a beta prior, then the evidence $p(D|M)$ is $p(z, N)$ and it is especially easy to compute. In Equation 5.7, the denominator (i.e., the part with the underbrace) showed that

$$B(a, b)\, p(z, N) = B(z + a, N - z + b).$$

Solving for $p(z, N)$ reveals that

$$p(z, N) = B(z+a, N-z+b) \,/\, B(a, b). \tag{5.10}$$

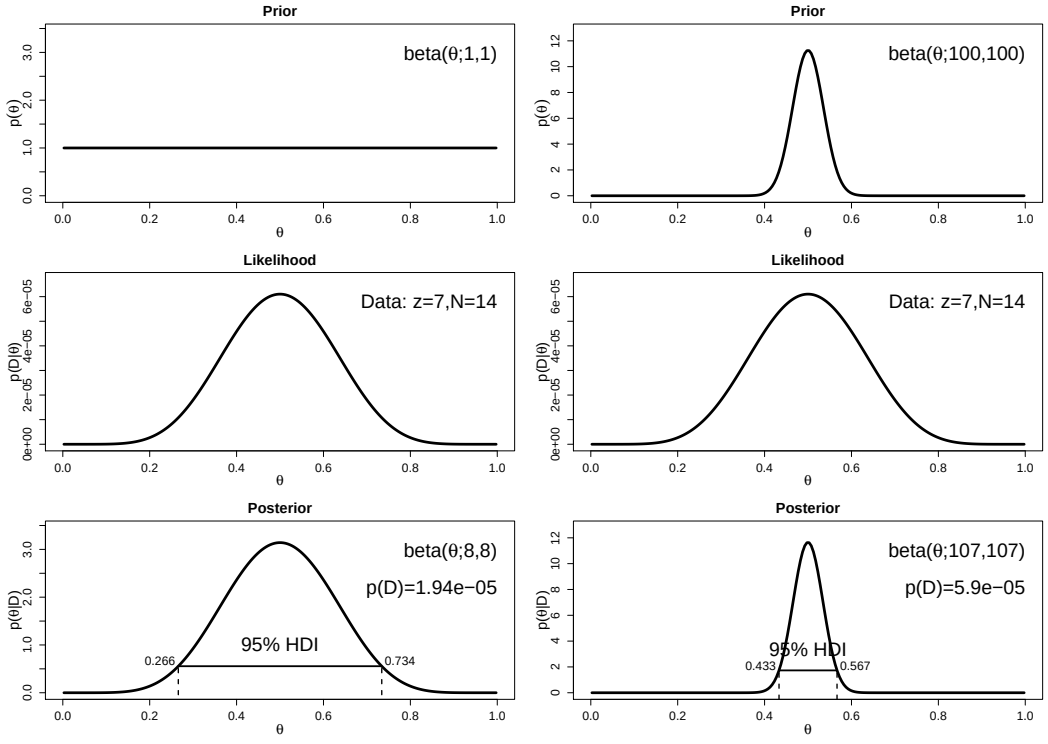


Figure 5.3: Two different beta priors, updated using the same data ($z = 7, N = 14$). The evidences are denoted by $p(D)$ (with M suppressed) in the lower panels of each column, and they favor the prior that is peaked over $\theta = 0.5$. Contrast with Figure 5.2, for which the evidences favored the uniform prior.

Thus, we can determine $p(z, N)$ using well-established beta functions, and we do not need to do any difficult integral calculus.

The lower panels of Figure 5.2 show the values of $p(z, N)$ for two different priors, for a fixed data set $z = 11, N = 14$. One prior is uniform, while the other prior is strongly peaked over $\theta = .50$. The data have a proportion of 1's that is not very close to .5, and therefore the prior that is peaked over 0.5 does not capture the data very well. The peaked prior has very low belief in values of θ near the data proportion, which means that $p(z, N)$ for the peaked-prior model is relatively small. The uniform prior, on the other hand, has relatively more belief in values of θ near the data proportion, and so its $p(D|M)$ is relatively high.

Consider a data set in which half the flips were heads, e.g., $z = 7$ and $N = 14$. Then which prior would produce the larger $p(z, N)$? Figure 5.3 shows the answer: The prior peaked over $\theta = 0.5$ is now preferred.

When we are evaluating the veracity of a model, the prior distribution for its parameters must be considered along with the likelihood function. In this chapter we are using a Bernoulli likelihood function and a beta prior. We can not say whether the Bernoulli likelihood is a "good" model of the coin flips without also specifying the values of θ that we believe in. Some values of θ may match the data well, but other values may not. If the values of θ that we believe in are not the values that match the data, then the model is not very good.

Because the prior distributions are part of the model, we can think of different prior distributions as constituting different models. We already took this approach, back in Fig-

ures 4.2 (page 54) and 4.3 (page 55). The simple model had a prior with non-zero belief on just a few values of θ, while the complex model had a prior with non-zero belief on many values of θ. We compared the models by considering the relative magnitudes of $p(D|M_{\text{simple}})$ and $p(D|M_{\text{complex}})$. We can do the same thing for the two priors in Figure 5.2. The peaked prior constitutes one model, and the uniform prior constitutes another model. We evaluate their relative veracity by comparing their values of $p(D|M_{\text{peaked}})$ and $p(D|M_{\text{uniform}})$.

We prefer the model with the higher value of $p(D|M)$, but the preference is not absolute. A tiny advantage in $p(D|M)$ should not be translated into a strong preference for one model over another. After all, the data themselves are just a random sample from the world, and they could have been somewhat different. It is only when the relative magnitudes of $p(D|M)$ are very different that we can feel confident in a preference of one model over another.

Moreover, we need to take into account our prior beliefs in the models, $p(M1)$ and $p(M2)$, as indicated in Equation 4.8 (page 48). Typically we will go into a model comparison with equal prior probabilities on the models, and so the ratio of the posterior probabilities is just the ratio of evidences (a.k.a. the Bayes factor). But if the prior probabilities of the models are not equal, then they must be factored into the posteriors. In particular, if we have a strong prior belief in a model, it takes more evidence for the other model to overcome our prior belief. For further applications of model comparison, see Exercises 5.6 and 5.7

5.3.3.1 Is the best model a good model?

Suppose we have two models, we collect some data, and find that the evidence for one model is much larger than the evidence for the other model. If our prior beliefs were equal, then the posterior beliefs strongly favor the winning model.

But is the winning model actually a good model of the data? The model comparison process has merely told us the models' *relative* believabilities, not their *absolute* believabilities. The winning model might merely be a less bad model than the horrible losing model.

Exercise 5.8 explains one way to assess whether the winning model can actually account for the data. The method used there is called a *posterior predictive check*.

5.4 Summary: How to do Bayesian inference

In this chapter, we've covered a lot of important conceptual points that are crucial for understanding what you are doing when you do Bayesian inference. But I don't want the volume of concepts to overwhelm the underlying simplicity of what you actually do. Here are the steps:

1. For the methods of this chapter to apply to your situation, the data must have two nominal values, like heads and tails. The two values must come up randomly, independently across observations, and with a single and fixed (i.e., stationary through time) probability. Denote the underlying probability of "heads" by the value θ.

2. Establish a description of your prior beliefs regarding values of θ by using a beta$(\theta; a, b)$ distribution. Decide what you think is the most probable value for θ; call it m. Decide how strongly you believe in m by considering how many new data

points (e.g., flips of the coin) it would take to sway you away from your prior belief. Call that number n. (In real research, the prior is established by considering previously published results and the audience to whom the analysis is addressed.) Then convert the m, n values to a, b values for the beta distribution by using Equation 5.5. You should check that the resulting beta distribution really captures your beliefs by looking at its graph! This is easily done by using the program in Section 5.5.1 (`BernBeta.R`). If you cannot express your prior beliefs as a beta distribution, then the methods of this chapter do not apply.

3. Observe some data. Enter them as a vector of 1's and 0's into the program of Section 5.5.1 (`BernBeta.R`). For N data points, e.g., flips of a coin, the vector should have N elements. "Heads" are coded by the value 1, and "tails" are coded by the value 0. The total number of 1's (i.e., heads) is denoted z.

4. Determine the posterior distribution of beliefs regarding values of θ. When the prior is a beta distribution, the posterior is a beta distribution too. The program of Section 5.5.1 (`BernBeta.R`) displays it graphically, along with a credible interval.

5. Make inferences from the posterior, depending on your goal. If your goal is to estimate θ, use the posterior distribution, perhaps as summarized by the 95% HDI. If your goal is to predict new data, then your predicted probability of "heads" on the next flip is the mean of the posterior, which is $(z + a)/(N + a + b)$. If your goal is to compare models (i.e., priors), then use $p(D)$ to decide which model's prior better accounts for the data. Use a posterior predictive check to get a sense of whether the better model actually mimics the data well.

5.5 R code

5.5.1 R code for Figure 5.2

This program defines a *function* in R instead of a script. A function takes input values, called "arguments," and does something with them. In general, a function in R in defined by code of the form:

```
function_name = function( arguments ) { commands }
```

The commands inside the braces can extend over many lines of code. When the function is called, it takes the values of the arguments in parentheses and uses them in the commands in the braces. You invoke the function by commanding R thus:

```
function_name( argument_values )
```

As a simple example, consider this definition:

```
asqplusb = function( a , b ) { a^2 + b }
```

We can then type

```
asqplusb( a=2 , b=1 )
```

and R returns `5`. We can get the same result be typing

```
asqplusb( 2 , 1 )
```

because unlabeled arguments are assumed to be provided in the order they were listed in the function definition. If you want to use some other ordering of arguments when calling the function, that is okay as long as the arguments are explicitly labeled. So, for example, we can get the same result by typing

```
asqplusb( b=1 , a=2 )
```

By default, a function in R returns the last value it computed in the list of commands. If you want to be sure that the function returns the intended value, you should include an explicit `return` command at the end of your function; for example:

```
asqplusb = function( a , b ) {
   c = a^2 + b
   return( c )
}
```

A very useful feature of function definitions in R is that arguments can be given default values. For example, in the following function definition, the argument `b` is given a default value of 1: `asqplusb = function( a , b=1 ) { a^2 + b }`
The function can then be called *without specifying a value for* `b`, and the value `b=1` will be assumed by default. For example, typing `asqplusb(a=2)` would return 5. A default value can be overridden by explicitly specifying its value in the function call; e.g., `asqplusb(a=2,b=2)` would return 6.

To use a function, R must know that the function exists. To point out a function to R, and make it ready to call, do the following. First, for this function `BernBeta.R`, save the full text of the function in a file, named "BernBeta.R," in a folder of your choice. Second, you must `source` that file, which essentially just runs the code. Running the function file simply defines the function for R, it does not call the function with arguments. To `source` the file, you can type the command `source("<pathtofile>/BernBeta.R")`, or, in the R command window (not an editing or graph window), click the File menu item, then the Source menu item, and then browse to the file and select it.

See the comments at the top of the code listing for details of how to use this function. When you examine the code, you may notice that the Bayesian computations constitute just a few lines; the bulk of the code is devoted to all the details of displaying the distributions! On line 57, a large window is opened. Line 58 specifies the layout of the subplots. The `layout` command takes a matrix as an argument; the integers in the matrix indicate which subpanels in the layout should be used for which plot. Thus, the first subplot is put into the matrix cells that have a 1, the second subplot is put into the matrix cells that have a 2, and so forth. Line 59 adjusts how the axis information is displayed. The `par` command has arguments such as `mar` and `mgp` which adjust the margins of the plot; type `help(par)` in R for details, and try varying the numbers in the `mar` and `mgp` vectors to explore their effects on the plot. Mathematical text in the plots uses the `expression` and `bquote` functions, which interpret their arguments as specifications for mathematical characters. For help with how to plot math characters in R, at the command line type `demo(plotmath)` and `help(plotmath)`.

The function that computes the HDI uses some advanced techniques, and its explanation is deferred to Section 23.3.3 (`HDIofICDF.R`). Nevertheless, the program that computes the HDI, called `HDIofICDF.R`, must be available for use by the program listed here, called `BernBeta.R`. Therefore, be sure that the two programs are in the same folder.

(BernBeta.R)

```
1  BernBeta = function( priorShape , dataVec , credMass=0.95 , saveGraph=F ) {
2  # Bayesian updating for Bernoulli likelihood and beta prior.
3  # Input arguments:
4  #   priorShape
5  #     vector of parameter values for the prior beta distribution.
6  #   dataVec
7  #     vector of 1's and 0's.
8  #   credMass
```

```
#     the probability mass of the equal tailed credible interval.
# Output:
#   postShape
#     vector of parameter values for the posterior beta distribution.
# Graphics:
#   Creates a three-panel graph of prior, likelihood, and posterior
#   with highest posterior density interval.
# Example of use:
# > postShape = BernBeta( priorShape=c(1,1) , dataVec=c(1,0,0,1,1) )
# You will need to "source" this function before using it, so R knows
# that the function exists and how it is defined.

# Check for errors in input arguments:
if ( length(priorShape) != 2 ) {
   stop("priorShape must have two components.") }
if ( any( priorShape <= 0 ) ) {
   stop("priorShape components must be positive.") }
if ( any( dataVec != 1 & dataVec != 0 ) ) {
   stop("dataVec must be a vector of 1s and 0s.") }
if ( credMass <= 0 | credMass >= 1.0 ) {
   stop("credMass must be between 0 and 1.") }

# Rename the prior shape parameters, for convenience:
a = priorShape[1]
b = priorShape[2]
# Create summary values of the data:
z = sum( dataVec == 1 ) # number of 1's in dataVec
N = length( dataVec )   # number of flips in dataVec
# Compute the posterior shape parameters:
postShape = c( a+z , b+N-z )
# Compute the evidence, p(D):
pData = beta( z+a , N-z+b ) / beta( a , b )
# Determine the limits of the highest density interval.
# This uses a home-grown function called HDIofICDF.
source( "HDIofICDF.R" )
hpdLim = HDIofICDF( qbeta , shape1=postShape[1] , shape2=postShape[2] )

# Now plot everything:
# Construct grid of theta values, used for graphing.
binwidth = 0.005 # Arbitrary small value for comb on Theta.
Theta = seq( from = binwidth/2 , to = 1-(binwidth/2) , by = binwidth )
# Compute the prior at each value of theta.
pTheta = dbeta( Theta , a , b )
# Compute the likelihood of the data at each value of theta.
pDataGivenTheta = Theta^z * (1-Theta)^(N-z)
# Compute the posterior at each value of theta.
pThetaGivenData = dbeta( Theta , a+z , b+N-z )
# Open a window with three panels.
windows(7,10)
layout( matrix( c( 1,2,3 ) ,nrow=3 ,ncol=1 ,byrow=FALSE ) ) # 3x1 panels
par( mar=c(3,3,1,0) , mgp=c(2,1,0) , mai=c(0.5,0.5,0.3,0.1) ) # margin specs
maxY = max( c(pTheta,pThetaGivenData) ) # max y for plotting
# Plot the prior.
plot( Theta , pTheta , type="l" , lwd=3 ,
          xlim=c(0,1) , ylim=c(0,maxY) , cex.axis=1.2 ,
          xlab=bquote(theta) , ylab=bquote(p(theta)) , cex.lab=1.5 ,
          main="Prior" , cex.main=1.5 )
if ( a > b ) { textx = 0 ; textadj = c(0,1) }
else { textx = 1 ; textadj = c(1,1) }
```

```
text( textx , 1.0*max(pThetaGivenData) ,
          bquote( "beta(" * theta * "|" * .(a) * "," * .(b) * ")"  ) ,
          cex=2.0 ,adj=textadj )
# Plot the likelihood: p(data|theta)
plot( Theta , pDataGivenTheta , type="l" , lwd=3 ,
          xlim=c(0,1) , cex.axis=1.2 , xlab=bquote(theta) ,
          ylim=c(0,1.1*max(pDataGivenTheta)) ,
          ylab=bquote( "p(D|" * theta * ")" ) ,
          cex.lab=1.5 , main="Likelihood" , cex.main=1.5 )
if ( z > .5*N ) { textx = 0 ; textadj = c(0,1) }
else { textx = 1 ; textadj = c(1,1) }
text( textx , 1.0*max(pDataGivenTheta) , cex=2.0 ,
          bquote( "Data: z=" * .(z) * ",N=" * .(N) ) ,adj=textadj )
# Plot the posterior.
plot( Theta , pThetaGivenData  ,type="l" , lwd=3 ,
          xlim=c(0,1) , ylim=c(0,maxY) , cex.axis=1.2 ,
          xlab=bquote(theta) , ylab=bquote( "p(" * theta * "|D)" ) ,
          cex.lab=1.5 , main="Posterior" , cex.main=1.5 )
if ( a+z > b+N-z ) { textx = 0 ; textadj = c(0,1) }
else { textx = 1 ; textadj = c(1,1) }
text( textx , 1.00*max(pThetaGivenData) , cex=2.0 ,
          bquote( "beta(" * theta * "|" * .(a+z) * "," * .(b+N-z) * ")"  ) ,
          adj=textadj )
text( textx , 0.75*max(pThetaGivenData) , cex=2.0 ,
          bquote( "p(D)=" * .(signif(pData,3)) ) , adj=textadj )
# Mark the HDI in the posterior.
hpdHt = mean( c( dbeta(hpdLim[1],a+z,b+N-z) , dbeta(hpdLim[2],a+z,b+N-z) ) )
lines( c(hpdLim[1],hpdLim[1]) , c(-0.5,hpdHt) , type="l" , lty=2 , lwd=1.5 )
lines( c(hpdLim[2],hpdLim[2]) , c(-0.5,hpdHt) , type="l" , lty=2 , lwd=1.5 )
lines( hpdLim , c(hpdHt,hpdHt) , type="l" , lwd=2 )
text( mean(hpdLim) , hpdHt , bquote( .(100*credMass) * "% HDI" ) ,
      adj=c(0.5,-1.0) , cex=2.0 )
text( hpdLim[1] , hpdHt , bquote(.(round(hpdLim[1],3))) ,
      adj=c(1.1,-0.1) , cex=1.2 )
text( hpdLim[2] , hpdHt , bquote(.(round(hpdLim[2],3))) ,
      adj=c(-0.1,-0.1) , cex=1.2 )
# Construct file name for saved graph, and save the graph.
if ( saveGraph ) {
  filename = paste( "BernBeta_",a,"_",b,"_",z,"_",N,".eps" ,sep="")
  dev.copy2eps( file = filename )
}
return( postShape )
} # end of function
```

5.6 Exercises

Exercise 5.1. **[Purpose: For you to see the influence of the prior in each successive flip, and for you to see another demonstration that the posterior is invariant under re-orderings of the data.]** For this exercise, use the R function of Section 5.5.1 (`BernBeta.R`). (Read the comments at the top of the code for an example of how to use it, and don't forget to `source` the function before calling it.) Notice that the function returns the posterior beta values each time it is called, so you can use the returned values as the prior values for the next function call.

(A) Start with a prior distribution that expresses some uncertainty that a coin is fair: beta(θ, 4, 4). Flip the coin once; suppose we get a head. What is the posterior distribution?

(B) Use the posterior from the previous flip as the prior for the next flip. Suppose we flip

again and get a head. Now what is the new posterior? (Hint: If you type `post = BernBeta( c(4,4) , c(1) )` for the first part, then you can type `post = BernBeta( post , c(1) )` for the next part.)

(C) Using that posterior as the prior for the next flip, flip a third time and get T. Now what is the new posterior? (Hint: Type `post = BernBeta( post , c(0) )`.)

(D) Do the same three updates but in the order T, H, H instead of H, H, T. Is the final posterior distribution the same for both orderings of the flip results?

Exercise 5.2. **[Purpose: Connecting HDIs to the real world, with iterative data collection.]** Suppose an election is approaching, and you are interested in knowing whether the general population prefers candidate A or candidate B. There is a just-published poll in the newspaper, which states that of 100 randomly sampled people, 58 preferred candidate A and the remainder preferred candidate B.

(A) Suppose that before the newspaper poll, your prior belief was a uniform distribution. What is the 95% HDI on your beliefs after learning of the newspaper poll results?

(B) Based in the newspaper poll, is it credible to believe that the population is equally divided in its preferences among candidates?

(C) You want to conduct a follow-up poll to narrow down your estimate of the population's preference. In your follow-up poll, you randomly sample 100 people and find that 57 prefer candidate A and the remainder prefer candidate B. Assuming that peoples' opinions have not changed between polls, what is the 95% HDI on the posterior?

(D) Based on your follow-up poll, is it credible to believe that the population is equally divided in its preferences among candidates?

Exercise 5.3. **[Purpose: Apply the Bayesian method to real data analysis. These data are representative of real data (Kruschke, 2009).]** Suppose you train people in a simple learning experiment, as follows. When people see the two words, "radio" and "ocean," on the computer screen, they should press the F key on the computer keyboard. They see several repetitions and learn the response well. Then you introduce another correspondence for them to learn: Whenever the words "radio" and "mountain" appear, they should press the J key on the computer keyboard. You keep training them until they know both correspondences well. Now you probe what they've learned by asking them about two novel test items. For the first test, you show them the word "radio" by itself and instruct them to make the best response (F or J) based on what they learned before. For the second test, you show them the two words "ocean" and "mountain" and ask them to make the best response. You do this procedure with 50 people. Your data show that for "radio" by itself, 40 people chose F and 10 chose J. For the word combination "ocean" and "mountain," 15 chose F and 35 chose J. Are people biased toward F or toward J for either of the two probe types? To answer this question, assume a uniform prior, and use a 95% HDI to decide which biases can be declared to be credible.

Exercise 5.4. **[Purpose: To explore an unusual prior and learn about the beta distribution in the process.]** Suppose we have a coin that we know comes from a magic-trick store, and therefore we believe that the coin is strongly biased either usually to come up heads or usually to come up tails, but we don't know which. Express this belief as a beta prior. (Hint: See Figure 5.1, upper left panel.) Now we flip the coin 5 times and it comes up heads in 4 of the 5 flips. What is the posterior distribution? (Use the R function of Section 5.5.1 (`BernBeta.R`) to see graphs of the prior and posterior.)

Exercise 5.5. **[Purpose: To get hands on experience with the goal of predicting the next datum, and to**

see how the prior influences that prediction.]

(A) Suppose you have a coin that you know is minted by the federal government and has not been tampered with. Therefore you have a strong prior belief that the coin is fair. You flip the coin 10 times and get 9 heads. What is your predicted probability of heads for the 11th flip? Explain your answer carefully; justify your choice of prior.

(B) Now you have a different coin, this one made of some strange material and marked (in fine print) "Patent Pending, International Magic, Inc." You flip the coin 10 times and get 9 heads. What is your predicted probability of heads for the 11th flip? Explain your answer carefully; justify your choice of prior. Hint: Use the prior from Exercise 5.4.

Exercise 5.6. **[Purpose: To get hands-on experience with the goal of model comparison.]** Suppose we have a coin, but we're not sure whether it's a fair coin or a trick coin. We flip it 20 times and get 15 heads. Is it more likely to be fair or trick? To answer this question, consider the value of the Bayes factor, i.e., the ratio of the evidences of the two models. When answering this question, justify your choice of priors to express the two hypotheses. Use the R function of Section 5.5.1 (`BernBeta.R`) to graph the priors and check that they reflect your beliefs; the R function will also determine the evidences from Equation 5.10.

Exercise 5.7. **[Purpose: To see how very small data sets can give strong leverage in model comparison when the model predictions are very different.]** Suppose we have a coin that we strongly believe is a trick coin, so it almost always comes up heads or it almost always comes up tails; we just don't know if the coin is the head-biased type or the tail-biased type. Thus, one model is a beta prior heavily biased toward tails, beta(θ; 1, 100), and the other model is a beta prior heavily biased toward heads, beta(θ; 100, 1). We flip the coin once and it comes up head. Based on that single flip, what is the value of the Bayes factor, i.e., the ratio of the evidences of the two models? Use the R function of Section 5.5.1 (`BernBeta.R`) to determine the evidences from Equation 5.10.

Exercise 5.8. **[Purpose: Hands-on learning about the method of posterior predictive checking.]** Following the scenario of the previous exercise, suppose we flip the coin a total of $N = 12$ times and it comes up heads in $z = 8$ of those flips. Suppose we let a beta(θ; 100, 1) distribution describe the head-biased trick coin, and we let a beta(θ; 1, 100) distribution describe the tail-biased trick coin.

(A) What are the evidences for the two models, and, what is the value of the Bayes factor?

Now for the new part, a *posterior predictive check*: Is the winning model actually a good model of the data? In other words, one model can be whoppingly better than the other, but that does not necessarily mean that the winning model is a good model; it might mean merely that the winning model is less bad than the losing model. One way to examine the veracity of the winning model is to simulate data sampled from the winning model and see if the simulated data "look like" the actual data. To simulate data generated by the winning model, we do the following: First, we will randomly generate a value of θ from the posterior distribution of the winning model. Second, using that value of θ, we will generate a sample of coin flips. Third, we will count the number of heads in the sample, as a summary of the sample. Finally, we determine whether the number of heads in a typical *simulated* sample is close to the number of heads in our *actual* sample. The program below carries out these steps. Study it, run it, and answer the questions that follow.

The program uses a `for` loop to repeat an action. For example, if you tell R: `for ( i in 1:5 ) { show(i) }`, it replies with: `1 2 3 4 5`. What `for` really does is execute the

commands within the braces for every element in the specified vector. For example, if you tell R: `for ( i in c(7,-.2) ) { show(i) }`, it replies with: `7 -.2`.

(BetaPosteriorPredictions.R)

```
# Specify known values of prior and actual data.
priorA = 100
priorB = 1
actualDataZ  = 8
actualDataN  = 12
# Compute posterior parameter values.
postA = priorA + actualDataZ
postB = priorB + actualDataN - actualDataZ
# Number of flips in a simulated sample should match the actual sample size:
simSampleSize = actualDataN
# Designate an arbitrarily large number of simulated samples.
nSimSamples = 10000
# Set aside a vector in which to store the simulation results.
simSampleZrecord = vector( length=nSimSamples )
# Now generate samples from the posterior.
for ( sampleIdx in 1:nSimSamples ) {
        # Generate a theta value for the new sample from the posterior.
        sampleTheta = rbeta( 1 , postA , postB )
        # Generate a sample, using sampleTheta.
        sampleData = sample( x=c(0,1) , prob=c( 1-sampleTheta , sampleTheta ) ,
                             size=simSampleSize , replace=TRUE )
        # Store the number of heads in sampleData.
        simSampleZrecord[ sampleIdx ] = sum( sampleData )
}
# Make a histogram of the number of heads in the samples.
hist( simSampleZrecord )
```

(B) How many samples (each of size N) were simulated?

(C) Was the same value of θ used for every simulated sample, or were different values of θ used in different samples? *Why?*

(D) Based on the simulation results, does the winning model seem to be a good model, and why or why not?

Chapter 6

Inferring a Binomial Proportion via Grid Approximation

Contents

I'm kinda coarse while the lady's refined,
I kinda stumble while she holds the line. But
both of us side-step and guess what what the answer is;
Both might feel better with (psycho-)analysis.

The previous chapter considered how to make inferences about a binomial proportion when the prior could be specified as a beta distribution. Using the beta distribution was very convenient because it made the integrals work out easily by direct formal analysis. But what if no beta distribution adequately expresses our prior beliefs? For example, our beliefs could be tri-modal: The coin might be heavily biased toward tails, or be approximately fair, or be heavily biased toward heads. No beta distribution has three "humps" like that.

In this chapter we explore one technique for numerically approximating the posterior distribution by defining the prior distribution over a fine grid of θ values. In this situation, we do not need a mathematical function of the prior over theta; we can specify any prior probability values we desire at each of the theta values. Moreover, we do not need to do any analytical (i.e., formulas only) integration. The denominator of Bayes' rule becomes a sum over many discrete θ values instead of an integral.

6.1 Bayes' rule for discrete values of θ

As in the previous chapter, the parameter θ denotes the value of a binomial proportion, such as the underlying propensity for a coin to come up heads. Previously we assumed that θ was continuous over the interval $[0, 1]$. We assumed that θ could have any value in that continuous domain. The prior probability on θ was, therefore, a probability *density* at each value of θ, such as a beta distribution.

Instead, we could assume that there are only a finite number of θ values in which we have any non-zero belief. For example, we might believe that θ can only have the values .25, .50, or .75. We already saw an example like this back in Figure 4.1 (page 51). When there are a finite number of θ values, then our prior distribution expresses the probability *mass* at each value of θ. In this situation, Bayes' rule is expressed as

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{\sum_{\theta} p(D|\theta)\, p(\theta)} \tag{6.1}$$

where the sum in the denominator is over the finite number of discrete values of θ that we are considering, and $p(\theta)$ denotes the probability *mass* at θ.

There are two niceties of dealing with the discrete version of Bayes' rule in Equation 6.1. One attraction is that some prior beliefs are easier to express with discrete values than with continuous density functions. Another felicity is that some mathematical functions, that are difficult to integrate analytically, can be approximated by evaluating the function on a fine grid of discrete values.

6.2 Discretizing a continuous prior density

If we could approximate a continuous prior density with a grid of discrete prior masses, then we could use the discrete form of Bayes' rule (in Equation 6.1) instead of the continuous form, which requires mathematically evaluating an integral. Fortunately, in some situations we can, in fact, make such an approximation. Figure 6.1 illustrates how a continuous prior density can be partitioned into a set of narrow rectangles that approximate the continuous prior. This process of discretizing the prior is straight forward: Divide the domain into a large number of narrow intervals. Draw a rectangle over each narrow interval, with height equal to the value of the density at the middle of the narrow interval. Approximate the area under the continuous density in each narrow interval by the area in the corresponding rectangle. This much of the process is illustrated in the top panels of Figure 6.1. The approximation gets better and better as the rectangles get narrower and narrower.

To discretize the density, we consider only the discrete θ values at the middles of each interval, and set the probability mass at that value to be the area of the corresponding rectangle. To make sure that the resulting discrete probabilities sum to exactly 1.0, we set each discrete probability to the corresponding interval area, and then divide by the sum of those probabilities. This discrete representation is shown in the lower panels of Figure 6.1. Notice that the scale on the y-axis has changed in going from upper to lower panels. In the upper panels, $p(\theta)$ refers to probability density at continuous values. In the lower panels, $p(\theta)$ refers to probability mass at discrete values.

When the prior density is discretized into a grid of masses, we can apply the discrete version of Bayes' rule (Equation 6.1). It is only an approximation to the true integral form, but if the grid is dense enough, the approximation can be very accurate.

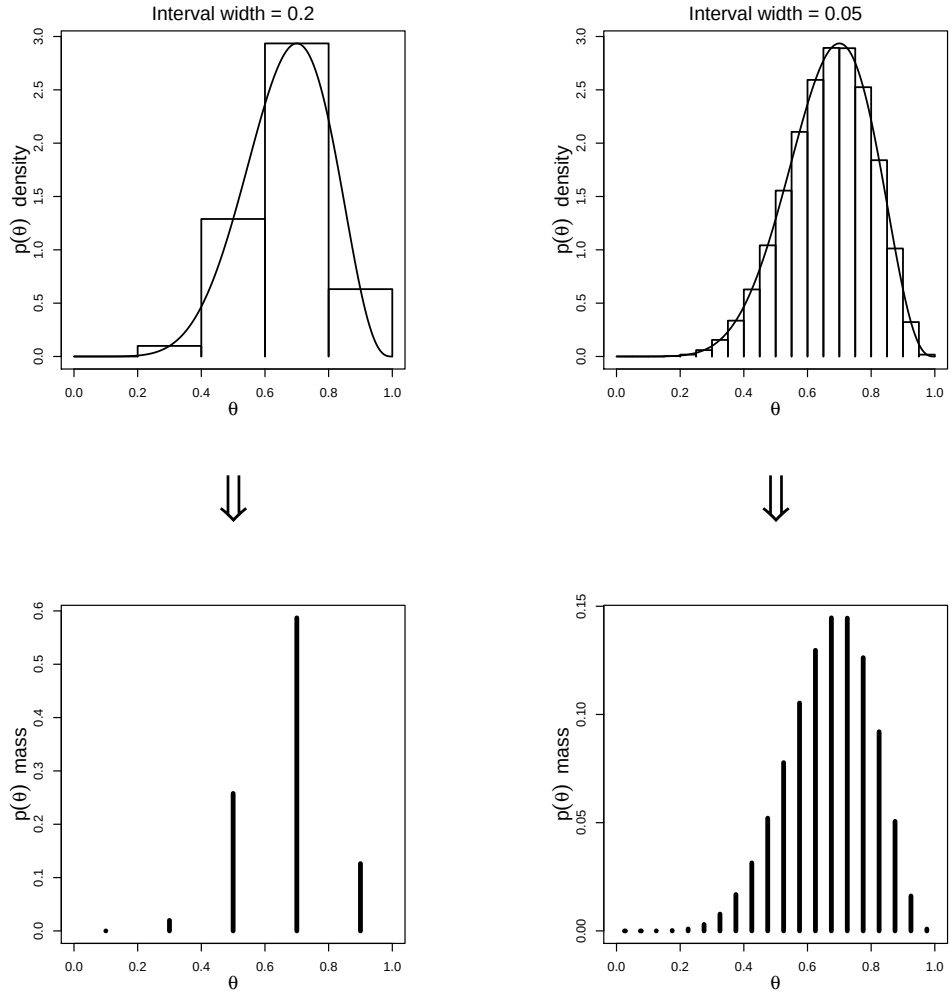


Figure 6.1: Approximation of a continuous density by a grid of discrete masses. Upper panels show a continuous density and its partitioning into rectangles. Lower panels plot the area of each rectangle; i.e., the masses. Left side shows approximation with a coarse grid; right side shows approximation with a finer grid.

6.2.1 Examples using discretized priors

Figure 6.2 shows a uniform prior discretized. Notice that because the prior is only defined at discrete points, the likelihood and posterior are only defined at those same points. The prior distribution is not represented by some mathematical function such as beta(θ, 1, 1); it is merely a list of probability masses at each discrete value of θ. Likewise, the shape of the posterior distribution is not stored as a mathematical function; it too is merely a list of probability masses at each discrete value of θ. The computations for the posterior did not involve any finesse with functions that described shapes of distributions; instead, the computations were just brute-force application of the discrete version of Bayes' rule (Equation 6.1).

The left and right sides of Figure 6.2 show the results for a coarse discretization and a finer discretization (respectively). Compare the approximation in Figure 6.2 with the exact beta-distribution updating in the left side of Figure 5.2, page 72. You can see that

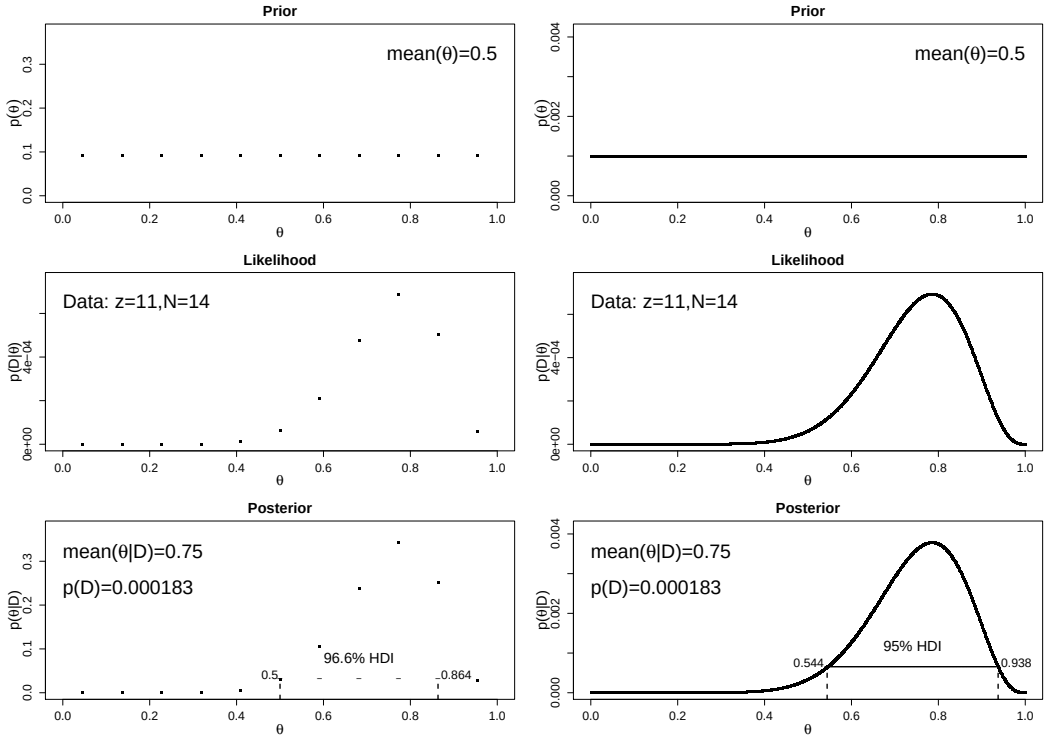


Figure 6.2: Grid approximations to Bayesian updating. The left side uses θ values at intervals of 1/11. The right side uses θ values at intervals of 1/1001. Compare this with the left side of Figure 5.2, page 72. (R code for this graph is in Section 6.7.1 (`BernGrid.R`).)

the fine-grid approximation is very accurate. It turns out in this case that even the coarse discretization does a good job at estimating $p(D)$; i.e., the denominator of Bayes' rule. The HDI, on the other hand, does differ noticeably between the two approximations. This is because the credible interval can only be specified to as fine a resolution as the sub-intervals in the discretization.

Figure 6.3 shows an arbitrarily-shaped prior, discretized. Here we see the advantaged gained by discretized priors, because no beta function could possibly imitate this prior accurately. Nevertheless, application of Bayes' rule yields the posterior distribution shown in the figure, which is as accurate as the discretized prior allows. If we desired, we could represent the prior beliefs on a finer comb of values over the parameter θ.

The joy of grid approximation is freedom from the siren song of beta priors. Eloquent as they are, beta priors can only express a limited range of beliefs. No beta function approximates the prior used in Figure 6.3, for example. You might say that these are cases for which beta ain't better. Even if you could find some complex mathematical function of θ that expressed the contours of your beliefs, such a function would probably be difficult to integrate in the denominator of Bayes' rule, and so you still couldn't determine an exact mathematical solution. Instead, you can express those beliefs approximately by specifying your degree of belief for each value of θ on a fine grid.

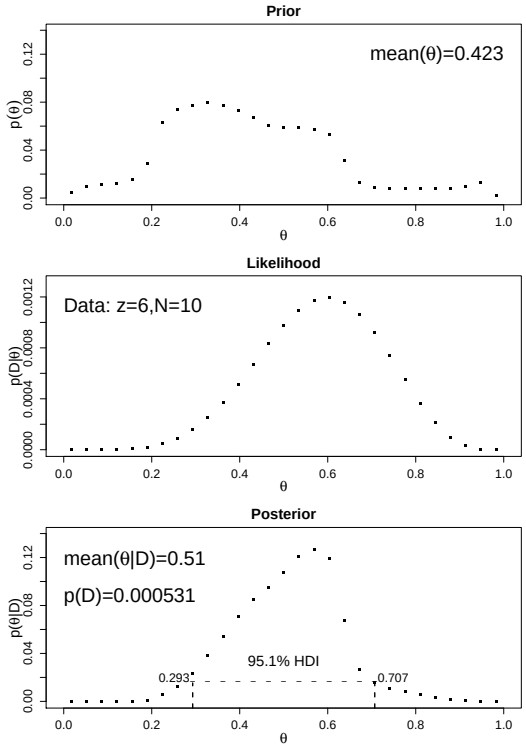


Figure 6.3: Example of an arbitrarily-shaped discretized prior. No beta function can accurately imitate this prior. (I call this the "Little Prince" prior, because it follows the profile of one of the beliefs of the character in the book by de Saint-Exupery (1943). For those of you who know what this is a picture of, you'll note the irony that despite the data showing a majority of heads, the posterior is most peaked at the elephant's tail.)

6.3 Estimation

The full list of posterior probability masses provides a complete estimate of the parameter values. Those masses can be summarized however is convenient and meaningful. Figures 6.2 and 6.3 provided two summary descriptors of the posterior, namely the mean value of θ and the 95% HDI (highest density interval).

The mean of θ is just the sum of the available parameter values weighted by the probability that they occur. Formally, that is expressed by

$$\overline{\theta}|D = \sum_{\theta} \theta \, p(\theta|D) \tag{6.2}$$

where the sum is over discrete values of θ is its grid points, and $p(\theta|D)$ is the probability mass at each grid point. The mean value is computed explicitly as this sum by the program in Section 6.7.1 (`BernGrid.R`), and the value is displayed in the plots it produces, as in Figures 6.2 and 6.3.

Recall that the HDI is defined such that the probability of any point inside the HDI is greater than the probability of any point outside the HDI, and the total probability of points in the 95% HDI is 95%. Because we are dealing with discrete masses, the sum of the masses in an interval usually will not be exactly 95%, and therefore we define the 95% HDI so it has total mass as small as possible but greater than or equal to 95%. That is why the left side of Figure 6.2 and Figure 6.3 show HDI masses slightly larger than 95%.

Figure 6.4 illustrates the 95% HDI for a bimodal posterior. The HDI is split into two separate segments. This split makes good sense: We want the HDI to represent those values of the parameter that are believable. For a bimodal posterior we should have a region of believability in two pieces. One of the attractions of grid approximation is that multimodal HDIs are easily determined.

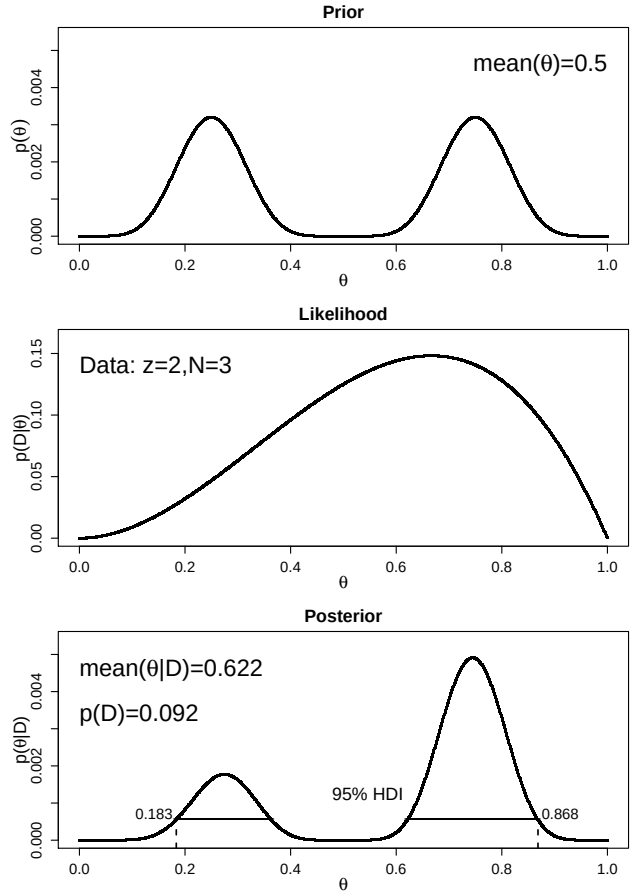


Figure 6.4: The 95% HDI of the posterior is *split* across more two distinct subintervals. Only the extreme left and right ends of the HDIs are marked in the plot, but the unmarked internal divisions *are* also endpoints of the split HDI region.

6.4 Prediction of subsequent data

A second typical goal of Bayesian inference is predicting subsequent data after incorporating an observed set of data. As has been our habit, let's denote the observed set of data as D, and the posterior distribution over parameter θ as $p(\theta|D)$. Our predicted probability for the next value of y is just the probability of that value happening for each value of θ, weighted by the posterior believability of each θ:

$$\begin{aligned} p(y|D) &= \int d\theta \, p(y|\theta) \, p(\theta|D) \\ &\approx \sum_{\theta} p(y|\theta) \, p(\theta|D) \end{aligned} \tag{6.3}$$

where $p(\theta|D)$ in the first line is a probability density, and $p(\theta|D)$ in the second line is a probability mass at discrete values of θ. In particular, for $y = 1$, Equation 6.3 becomes

$$\begin{aligned} p(y=1|D) &\approx \sum_{\theta} p(y=1|\theta) \, p(\theta|D) \\ &= \sum_{\theta} \theta \, p(\theta|D) \end{aligned} \tag{6.4}$$

which is just the average value of θ in the posterior distribution of θ. This fact is not news; we've seen this idea before in Equation 5.9 (p. 73). What's new is that we are not relying on the posterior having the form of a beta distribution. For an example, see Exercise 6.6.

6.5 Model comparison

A third typical goal of Bayesian inference is model comparison. Suppose we have two models, denoted $M1$ and $M2$, with prior beliefs $p(M1)$ and $p(M2)$. We want to determine the posterior beliefs, $p(M1|D)$ and $p(M2|D)$. Recall from Equation 4.8, p. 48, that

$$\frac{p(M1|D)}{p(M2|D)} = \frac{p(D|M1)}{p(D|M2)} \frac{p(M1)}{p(M2)}. \tag{6.5}$$

where (as in Equation 4.7)

$$p(D|M) = \int d\theta \, p(D|\theta, M) \, p(\theta|M).$$

is the "evidence" for model M. The integral for the evidence becomes a sum when we have a discrete grid approximation:

$$p(D|M) \approx \sum_{\theta} p(D|\theta, M) \, p(\theta|M) \tag{6.6}$$

where the sum is over discrete values of θ and $p(\theta|M)$ is a probability mass at each value of θ.

In our current application, we are assuming a Bernoulli likelihood function for all models. In other words, we can drop the "M" from $p(D|\theta, M)$ in Equation 6.6. The only differentiation among models is the specification of the prior beliefs over θ. For example, one model might posit that the coin is head-biased, while the other model posits that the coin is tail-biased. Therefore, combining Equations 6.5 and 6.6 yields

$$\frac{p(M1|D)}{p(M2|D)} = \frac{\sum_\theta p(D|\theta) \, p(\theta|M1)}{\sum_\theta p(D|\theta) \, p(\theta|M2)} \frac{p(M1)}{p(M2)}. \tag{6.7}$$

The expression in Equation 6.7 is useful when the priors are not beta distributions. (If the priors are beta distributions, then exact mathematical analysis yields the results described in Section 5.3.3, p. 73.) Conveniently, the R function provided in Section 6.7.1 (`BernGrid.R`) displays $p(D|M)$ in its graphical output, computed by Equation 6.6. For examples, see Exercises 6.7 and 6.8.

6.6 Summary

This chapter showed that Bayesian inference, regarding a continuous binomial proportion θ, can be achieved by approximating the continuous θ with a dense grid of discrete values. A disadvantage of this approach is that the approximation is only as good as the density of the grid. But there are several advantages of the approach. One advantage is that we have great freedom in the type of prior distributions we can specify; we are not restricted to beta distributions. Another advantage is that we can use the discrete approximation to find approximate HDI regions. We can also do posterior prediction and model comparison with arbitrary priors. The program of Section 6.7.1 (`BernGrid.R`) is provided to help you conduct these analyses.

6.7 R code

6.7.1 R code for Figure 6.2 etc.

For this function you need to set up a comb (i.e., grid) of θ values and a prior on those values. The extensive comments at the beginning of the function provide an example for how to do this. See also Exercise 6.1 for a caveat regarding how *not* to do this.

The Bayesian computations form only a few lines of this function (namely, lines 30–37). The bulk of the program is devoted to all the details of plotting the information! Section 23.3.1 (`HDIofGrid.R`) explains how the HDI is approximated.

(BernGrid.R)

```
1  BernGrid = function( Theta , pTheta , Data ,
2                       credib=.95 , nToPlot=length(Theta) ) {
3  # Bayesian updating for Bernoulli likelihood and prior specified on a grid.
4  # Input arguments:
5  #  Theta is a vector of theta values, all between 0 and 1.
6  #  pTheta is a vector of corresponding probability _masses_.
7  #  Data is a vector of 1's and 0's, where 1 corresponds to a and 0 to b.
8  #  credib is the probability mass of the credible interval, default is 0.95.
9  #  nToPlot is the number of grid points to plot; defaults to all of them.
10 # Output:
11 #  pThetaGivenData is a vector of posterior probability masses over Theta.
12 #  Also creates a three-panel graph of prior, likelihood, and posterior
13 #  probability masses with credible interval.
14 # Example of use:
15 #  # Create vector of theta values.
16 #  > binwidth = 1/1000
17 #  > thetagrid = seq( from=binwidth/2 , to=1-binwidth/2 , by=binwidth )
18 #  # Specify probability mass at each theta value.
19 #  > relprob = pmin(thetagrid,1-thetagrid) # relative prob at each theta
20 #  > prior = relprob / sum(relprob) # probability mass at each theta
21 #  # Specify the data vector.
22 #  > datavec = c( rep(1,3) , rep(0,1) ) # 3 heads, 1 tail
23 #  # Call the function.
24 #  > posterior = BernGrid( Theta=thetagrid , pTheta=prior , Data=datavec )
25 # Hints:
26 #  You will need to "source" this function before calling it.
27 #  You may want to define a tall narrow window before using it; e.g.,
28 #  > windows(7,10)
29
30 # Create summary values of Data
31 z = sum( Data==1 ) # number of 1's in Data
32 N = length( Data ) # number of flips in Data
33 # Compute the likelihood of the Data for each value of Theta.
34 pDataGivenTheta = Theta^z * (1-Theta)^(N-z)
35 # Compute the evidence and the posterior.
36 pData = sum( pDataGivenTheta * pTheta )
37 pThetaGivenData = pDataGivenTheta * pTheta / pData
38
39 # Plot the results.
40 layout( matrix( c( 1,2,3 ) ,nrow=3 ,ncol=1 ,byrow=FALSE ) ) # 3x1 panels
41 par( mar=c(3,3,1,0) , mgp=c(2,1,0) , mai=c(0.5,0.5,0.3,0.1) ) # margin settings
42 dotsize = 4 # how big to make the plotted dots
43 # If the comb has a zillion teeth, it's too many to plot, so plot only a
44 # thinned out subset of the teeth.
45 nteeth = length(Theta)
46 if ( nteeth > nToPlot ) {
```

```
  thinIdx = seq( 1, nteeth , round( nteeth / nToPlot ) )
  if ( length(thinIdx) < length(Theta) ) {
    thinIdx = c( thinIdx , nteeth ) # makes sure last tooth is included
  }
} else { thinIdx = 1:nteeth }
# Plot the prior.
meanTheta = sum( Theta * pTheta ) # mean of prior, for plotting
plot( Theta[thinIdx] , pTheta[thinIdx] , type="p" , pch="." , cex=dotsize ,
      xlim=c(0,1) , ylim=c(0,1.1*max(pThetaGivenData)) , cex.axis=1.2 ,
      xlab=bquote(theta) , ylab=bquote(p(theta)) , cex.lab=1.5 ,
      main="Prior" , cex.main=1.5 )
if ( meanTheta > .5 ) {
   textx = 0 ; textadj = c(0,1)
} else {
  textx = 1 ; textadj = c(1,1)
}
text( textx , 1.0*max(pThetaGivenData) ,
      bquote( "mean(" * theta * ")=" * .(signif(meanTheta,3)) ) ,
      cex=2.0 , adj=textadj )
# Plot the likelihood: p(Data|Theta)
plot(Theta[thinIdx] ,pDataGivenTheta[thinIdx] ,type="p" ,pch="." ,cex=dotsize
        ,xlim=c(0,1) ,cex.axis=1.2 ,xlab=bquote(theta)
        ,ylim=c(0,1.1*max(pDataGivenTheta))
        ,ylab=bquote( "p(D|" * theta * ")" )
        ,cex.lab=1.5 ,main="Likelihood" ,cex.main=1.5 )
if ( z > .5*N ) { textx = 0 ; textadj = c(0,1) }
else { textx = 1 ; textadj = c(1,1) }
text( textx ,1.0*max(pDataGivenTheta) ,cex=2.0
        ,bquote( "Data: z=" * .(z) * ",N=" * .(N) ) ,adj=textadj )
# Plot the posterior.
meanThetaGivenData = sum( Theta * pThetaGivenData )
plot(Theta[thinIdx] ,pThetaGivenData[thinIdx] ,type="p" ,pch="." ,cex=dotsize
        ,xlim=c(0,1) ,ylim=c(0,1.1*max(pThetaGivenData)) ,cex.axis=1.2
        ,xlab=bquote(theta) ,ylab=bquote( "p(" * theta * "|D)" )
        ,cex.lab=1.5 ,main="Posterior" ,cex.main=1.5 )
if ( meanThetaGivenData > .5 ) { textx = 0 ; textadj = c(0,1) }
else { textx = 1 ; textadj = c(1,1) }
text(textx ,1.00*max(pThetaGivenData) ,cex=2.0
        ,bquote( "mean(" * theta * "|D)=" * .(signif(meanThetaGivenData,3)) )
        ,adj=textadj )
text(textx ,0.75*max(pThetaGivenData) ,cex=2.0
        ,bquote( "p(D)=" * .(signif(pData,3)) ) ,adj=textadj )
# Mark the highest density interval. HDI points are not thinned in the plot.
source("HDIofGrid.R")
HDIinfo = HDIofGrid( pThetaGivenData  )
points( Theta[ HDIinfo$indices ] ,
       rep( HDIinfo$height , length( HDIinfo$indices ) ) , pch="-" , cex=1.0 )
text( mean( Theta[ HDIinfo$indices ] ) , HDIinfo$height ,
         bquote( .(100*signif(HDIinfo$mass,3)) * "% HDI" ) ,
         adj=c(0.5,-1.5) , cex=1.5 )
# Mark the left and right ends of the waterline. This does not mark
# internal divisions of an HDI waterline for multi-modal distributions.
lowLim = Theta[ min( HDIinfo$indices ) ]
highLim = Theta[ max( HDIinfo$indices ) ]
lines( c(lowLim,lowLim) , c(-0.5,HDIinfo$height) , type="l" , lty=2 , lwd=1.5)
lines( c(highLim,highLim) , c(-0.5,HDIinfo$height) , type="l" , lty=2 , lwd=1.5)
text( lowLim , HDIinfo$height , bquote(.(round(lowLim,3))) ,
      adj=c(1.1,-0.1) , cex=1.2 )
text( highLim , HDIinfo$height , bquote(.(round(highLim,3))) ,
```

```
106       adj=c(-0.1,-0.1) , cex=1.2 )
107
108 return( pThetaGivenData )
109 } # end of function
```

6.8 Exercises

Exercise 6.1. **[Purpose: Understand the discretization used for the priors in the R functions of Section 6.7.1 (BernGrid.R) and throughout this chapter.]** Consider this R code for discretizing a beta(θ, 8, 4) distribution:

```
nIntervals = 10
width = 1 / nIntervals
Theta = seq( from = width/2 , to = 1-width/2 , by = width )
approxMass = dbeta( Theta , 8 , 4 ) * width
pTheta = approxMass / sum( approxMass )
```

(A) What is the value of `sum(approxMass)`? Why is it not exactly 1?

(B) Suppose we use instead the following code to define the grid of points:

```
Theta = seq( from = 0 , to = 1 , by = width )
```

Why is this not appropriate? (Hint: Consider exactly what intervals are represented by the first and last values in `Theta`. Do those first and last intervals have the same widths as the other intervals, and if they do, do they fall entirely within the domain of the beta distribution?)

Exercise 6.2. **[Purpose: Practice specifying a non-beta prior.]** Suppose we have a coin that has a head on one side and a tail on the other. We think it might be fair, or it might be a trick coin that is heavily biased toward heads or tails. We want to express this prior belief with a single prior over θ. Therefore the prior needs to have three peaks: One near zero, one around .5, and near 1.0. But these peaks are not just isolated spikes, because we have uncertainty about the actual value of θ.

(A) Express your prior belief as a list of probability masses over a fairly dense grid of θ values. Remember to set a gradual decline around the three peaks. Briefly justify your choice. Hint: You can specify the peaks however you want, but one simple way is something like

```
pTheta = c( 50:1 , rep(1,50) , 1:50 , 50:1 , ...
pTheta = pTheta / sum( pTheta )
width = 1 / length(pTheta)
Theta = seq( from = width/2 , to = 1-width/2 , by = width )
```

(B) Suppose you flip the coin 20 times and get 15 heads. Use the R function of Section 6.7.1 (`BernGrid.R`) to display the posterior beliefs. Include the R code you used to specify the prior values.

Exercise 6.3. **[Purpose: Use the function of Section 6.7.1 (BernGrid.R) for sequential updating; i.e., use output of one function call as the prior for the next function call. Observe that data ordering does not matter]**

(A) Using the same prior that you used for the previous exercise, suppose you flip the coin just 4 times and get 3 heads. Use the R function of Section 6.7.1 (`BernGrid.R`) to display the posterior.

(B) Suppose we flip the coin an additional 16 times and get 12 heads. Now what is the posterior distribution? To answer this question, use the posterior distribution that is output

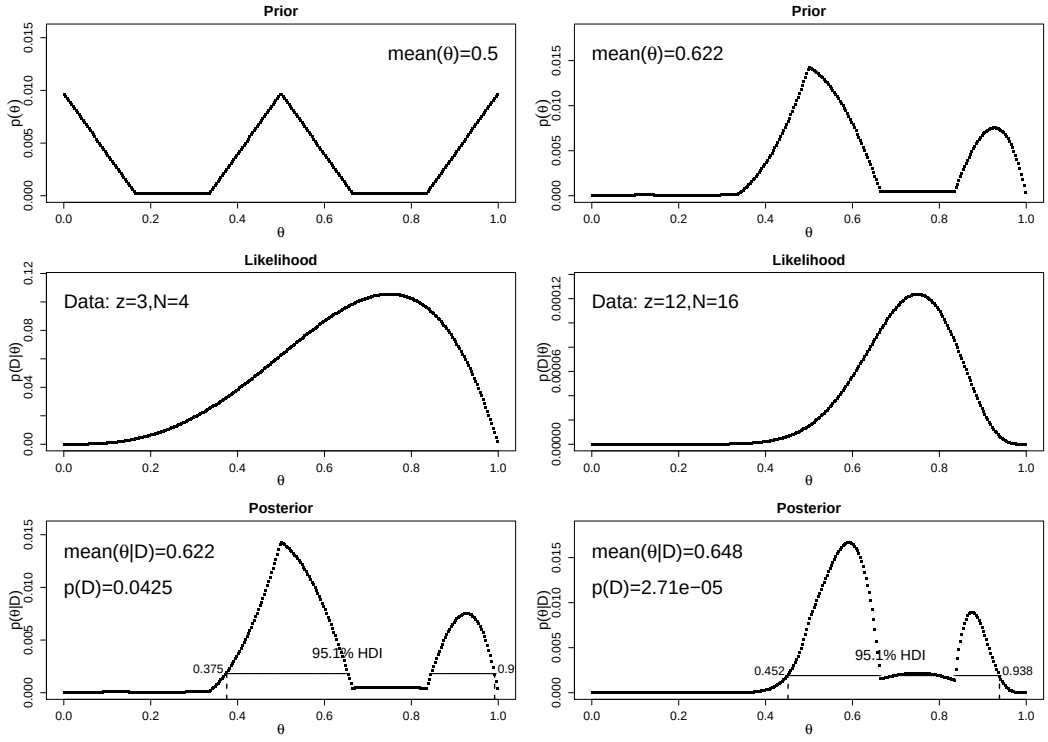


Figure 6.5: For Exercise 6.3. The posterior from the first four flips, in the left column, is used as the prior for the next sixteen flips, in the right column. Your original prior (top left panel) may look different from the original prior used here.

by the function in the previous part as the prior for this part. Show the R commands you used to call the function. Hint: The final posterior should match the posterior of Exercise 6.2, except that the graph of the prior should look like the posterior from the previous part. Figure 6.5 shows an example.

Exercise 6.4. **[Purpose: Connecting HDIs to the real world, with iterative data collection.]** Suppose an election is approaching, and you are interested in knowing whether the general population prefers candidate A or candidate B. There is a just-published poll in the newspaper, which states that of 100 randomly sampled people, 58 preferred candidate A and the remainder preferred candidate B.

(A) Suppose that before the newspaper poll, your prior belief was a uniform distribution. What is the 95% HDI on your beliefs after learning of the newspaper poll results? Use the function of Section 6.7.1 (`BernGrid.R`) to determine your answer.

(B) Based in the newspaper poll, is it credible to believe that the population is equally divided in its preferences among candidates?

(C) You want to conduct a follow-up poll to narrow down your estimate of the population's preference. In your follow-up poll, you randomly sample 100 people and find that 57 prefer candidate A and the remainder prefer candidate B. Assuming that peoples' opinions have not changed between polls, what is the 95% HDI on the posterior?

(D) Based on your follow-up poll, is it credible to believe that the population is equally divided in its preferences among candidates? Hint: Compare your answer here to your answer for Exercise 5.2.

Exercise 6.5. **[Purpose: HDIs in the (almost) real world.]** Suppose that the newly-hired quality control manager at the Acme Widget factory is trying to convince the CEO that the proportion of defective widgets coming off the assembly line is less than 10%. There are no previous data available regarding the defect rate at the factory. The manager randomly samples 500 widgets and she finds that 28 of them are defective. What do you conclude about the defect rate? Justify your choice of prior. Include graphs to explain/support your conclusion.

Exercise 6.6. **[Purpose: Using grid approximation for prediction of subsequent data.]** Suppose we believe that a coin is biased to come up heads, and we describe our prior belief as quadratically increasing: $p(\theta) \propto \theta^2$. Suppose we flip the coin 4 times and observe 2 heads and 2 tails. Based on the posterior distribution, what is the predicted probability that the next flip will yield a head? To answer this question, use the function of Section 6.7.1 (`BernGrid.R`). Define `thetagrid` as in the example in the comments at the beginning of the function. Then define `relprob = thetagrid ^ 2`, and normalize it to specify the prior. The function returns a vector of discrete posterior masses, which you might call `posterior`. Apply Equation 6.4 by computing `sum( thetagrid * posterior )`. Bonus hint: The answer is also displayed in the output graphics.

Exercise 6.7. **[Purpose: Using grid approximation to compare models.]** Suppose we have competing beliefs about the bias of a coin: One person believes the coin is head-biased, and the second person believes the coin is tail-biased. To make this specific, suppose the head-biased prior is $p(\theta|M1) \propto \theta^2$, and the tail-biased prior is $p(\theta|M2) \propto (1-\theta)^2$. Suppose that we are equally willing to entertain the two models, so $p(M1) = p(M2) = 0.5$. We flip the coin $N = 8$ times and observe $z = 6$ heads. What is the ratio of posterior beliefs? To answer this question, read the coding suggestion in Exercise 6.6 and look at $p(D)$ in the graphical output.

Exercise 6.8. **[Purpose: Model comparison in the (almost) real world.]** A pharmaceutical company claims that its new drug increases the probability that couples who take the drug will conceive a boy. They have published no studies regarding this claim, so there is no public knowledge regarding the efficacy of the drug. Suppose you conduct a study in which 50 couples, sampled at random from the general population, take the drug during a period of time while trying to conceive a baby. Suppose that eventually all couples conceive; there are 30 boys and 20 girls (no multiple births).

(A) You want to estimate the probability of conceiving a boy for couples who take the drug. What is an appropriate prior belief distribution? It cannot be the general population probability, because that is a highly peaked distribution near 0.5 that refers to non-drugged couples. Instead, the prior needs to reflect our pre-experiment uncertainty in the effect of the drug. Discuss your choice of prior with this in mind.

(B) Using your prior from the previous part, show a graph of the posterior and decide whether it is credible that couples who take the drug have a 50% chance of conceiving a boy.

(C) Suppose that the drug manufacturers make a strong claim that their drug sets the probability of conceiving a boy to very nearly 60%, with high certainty. Suppose you represent that claim by a beta(60,40) prior. Compare that claim against the skeptic who says there is no effect of the drug, and the probability of conceiving a boy is represented by a beta(50,50) prior. What is the value of $p(D)$ for each prior? What is the posterior belief in each claim? Hint: Careful when computing the posterior belief in each model,

because you need to take into account the prior belief in each model. Is the prior belief in the manufacturer's claim as strong as the prior belief in the skeptical claim?

References

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician*, *46*, 261–283.

Agresti, A., & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods & Applications*, *14*(3), 297–330.

Albert, J. H., & Rossman, A. J. (2001). *Workshop statistics: Discovery with data, a Bayesian approach*. Emeryville, CA: Key College Publishing.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis, 2nd edition*. New York: Springer.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*(2), 159–165.

Berger, R. L., Boos, D. D., & Guess, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics*, *44*(1), 103–115.

Berry, D. A. (1996). *Statistics: A Bayesian perspective*. Belmont, CA: Duxbury Press / Wadsworth.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1-2), 215–227.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bliss, C. I. (1934). The method of probits. *Science*, *79*(2037), 38–39.

Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed.). Hoboken, NJ: Wiley.

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*, 63–82.

Braumoeller, B. F. (2004). Hypothesis testing and multiplicative interaction terms. *International Organization*, *58*(04), 807–820.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*, 1–27.

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, B*, *57*(3), 473–484.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall / CRC.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Casella, G., & Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, *101*(473), 157–167.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304.

Chen, M.-H., He, X., Shao, Q.-M., & Xu, H. (2003). A Monte Carlo gap test in computing HPD regions. *Development of Modern Statistics and Related Topics: In Celebration of Professor Yaoting Zhang's 70th Birthday*, 38–52.

Chen, M. H., & Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, *8*, 69–92.

Clyde, M., & George, E. I. (2004). Model uncertainty. *Statistical Science*, 81–94.

Congdon, P. (2005). *Bayesian models for categorical data*. West Sussex, England: Wiley.

Damgaard, L. H. (2007). Technical note: How to use WinBUGS to draw inferences in animal models. *Journal of Animal Science*, *85*, 1363–1368.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–77.

de Saint-Exupery, A. (1943). *The little prince*. San Diego: Harcourt.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, *124*, 121–144.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A*, *170*, 95–113.

DeGroot, M. H. (2004). *Optimal statistical decisions*. New York: Wiley Interscience.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Feldman, H. A. (1988). Families of lines: random effects in linear regression analysis. *Journal of Applied Physiology*, *64*(4), 1721–1732.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, *40*, 575–586.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, *56*, 501–514.

Gelman, A. (2005). Analysis of variance — why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Florida: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilievel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2009). *Why we (usually) don't have to worry about multiple comparisons.* Available from `http://www.stat.columbia.edu/~gelman/research/unpublished/multiple2.pdf`

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(3), 363–373.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, *95*(452).

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*(1), 169–177.

Gill, J. (2002). *Bayesian methods for the social and behavioral sciences*. Boca Raton, Florida: CRC Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Gopalan, R., & Berry, D. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association*, 1130–1139.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, *167*(5), 523–529.

Guber, D. L. (1999). Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education*, *7*(2). Available from http://www.amstat.org/publications/JSE/secure/v7n2/datasets.guber.cfm

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*(1), 1–32.

Han, C., & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, *96*(455), 1122–1132.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A handbook of small data sets*. London: Chaprman & Hall.

Hobbs, B. P., & Carlin, B. P. (2008, January). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80.

Hoffman, P. J., Earle, T. C., & Slovic, P. (1981). Multidimensional functional learning (MFL) and some new conceptions of feedback. *Organizational Behavior and Human Performance*, *27*(1), 75–102.

Holcomb, J., & Spalsbury, A. (2005). Teaching students to use summary statistics and graphics to clean and analyze data. *Journal of Statistics Education*, *13*(3). Available from http://www.amstat.org/publications/jse/v13n3/datasets.holcomb.html

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American statistician*, *50*(2), 120–126.

Joseph, L., Wolfson, D. B., & du Berger, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*, *44*, 143–154.

Joseph, L., Wolfson, D. B., & du Berger, R. (1995b). Some comments on Bayesian sample size determination. *The Statistician*, *44*, 167–171.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keith, T. (2005). *Multiple regression and beyond*. Columbus, OH: Allyn & Bacon.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability*. New York: Chelsea.

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian inference: The general case. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 165–180). New York: Springer.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*, 201–223.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210-226.

Kruschke, J. K. (2009). Highlighting: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 153–185). Elsevier / Academic Press.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, **(**), **-**.

Learner, E. E. (1978). *Specification searches*. New York: Wiley.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.

Lindley, D. V. (1997). The choice of sample size. *The Statistician*, *46*, 129–138.

Lindley, D. V., & Phillips, L. D. (1976). Inference for a Bernoulli processs (a Bayesian view). *The American Statistician*, *30*(3), 112–119.

Lindquist, M. A., & Gelman, A. (2009). Correlations and multiple comparisons in functional imaging – a statistical perspective. *Perspectives in Psychological Science*, *4*(3), 310–313.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

MacKay, D. J. C. (2003). *Information theory, inference & learning algorithms*. Cambridge, UK: Cambridge University Press.

Marin, J.-M., & Robert, C. P. (2007). *Bayesian core: A practical approach to computational Bayesian statistics*. New York: Springer.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models, 2nd ed.* Boca Raton, FL: Chapman and Hall/CRC.

McDonald, J. H. (2009). *Handbook of biological statistics (2nd ed.)*. Baltimore, Maryland: Sparky House Publishing.

McDonald, J. H., Seed, R., & Koehn, R. K. (1991). Allozymes and morphometric characters of three species of Mytilus in the Northern and Southern Hemispheres. *Marine Biology*, *111*(3), 323–333.

McIntyre, L. (1994). Using cigarette data for an introduction to multiple regression. *Journal of Statistics Education*, *2*(1). Available from http://www.amstat.org/publications/jse/v2n1/datasets.mcintyre.html

Meng, C. Y. K., & Dempster, A. P. (1987). A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics*, *43*(2), 301–311.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calulations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1091.

Meyer, R., & Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal*, *3*(2), 198–215.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617–640.

Moore, T. L. (2006). Paradoxes in film ratings. *Journal of Statistics Education*, *14*(1). Available from www.amstat.org/publications/jse/v14n1/datasets.moore.html

Mueller, P., Parmigiani, G., & Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In J. M. Bernardo et al. (Eds.), *Bayesian statistics 8*. Oxford, UK: Oxford University Press.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Nietzsche, F. (1967). *The will to power*. New York: Random House. (Translated by W. Kaufmann and R. J. Hollingdale)

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.

Oswald, C. J. P., Yee, B. K., Rawlins, J. N. P., Bannerman, D. B., Good, M., & Honey, R. C. (2001). Involvement of the entorhinal cortex in a process of attentional modulation: Evidence from a novel variant of an IDS/EDS procedure. *Behavioral neuroscience*, *115*(4), 841–849.

Pham-Gia, T., & Turkkan, N. (1992). Sample size determination in Bayesian analysis. *The Statistician*, *41*, 389–392.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 375–383.

Qian, S. S., & Shen, Z. (2007). Ecological applications of multilevel analysis of variance. *Ecology*, *88*(10), 2489–2495.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York: Springer.

Rosa, L., Rosa, E., Sarner, L., & Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association*, *279*(13), 1005–1010.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Roy, A., Ghosal, S., & Rosenberger, W. F. (2009). Convergence properties of sequential Bayesian D-optimal designs. *Journal of Statistical Planning and Inference*, *139*, 425–440.

Sadiku, M. N. O., & Tofighi, M. R. (1999). A tutorial on simulation of queueing models. *International Journal of Electrical Engineering Education*, *36*, 102–120.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference*, *136*(7), 2144–2162.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, *28*(1), 9–12.

Solari, F., Liseo, B., & Sun, D. (2008). Some remarks on Bayesian inference for one-way ANOVA models. *Annals of the Institute of Statistical Mathematics*, *60*, 483–498.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A*, *157*, 357–416.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Thomas, A. (2004). *BRugs user manual (the R interface to BUGS).* Available from `http://mathstat.helsinki.fi/openbugs/data/Docu/BRugs%20Manual.html`

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006, March). Making BUGS open. *R News*, *6*(1), 12–17.

Tsionas, E. G. (2002). Bayesian inference in the noncentral Student-t model. *Journal of Computational and Graphical Statistics*, *11*(1), 208–221.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Walker, L. J., Gustafson, P., & Frimer, J. A. (2007). The application of Bayesian analysis to issues in developmental research. *International Journal of Behavioral Development*, *31*(4), 366.

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, *17*, 193–208.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, *46*, 185–191.

Werner, M., Stabenau, J. R., & Pollin, W. (1970). Thematic apperception test method for the differentiation of families of schizophrenics, delinquents, and "normals". *Journal of Abnormal Psychology*, *75*(2), 139–145.

Western, B., & Jackman, S. (1994). Bayesian inference for comparative research. *The American Political Science Review*, *88*(2), 412–423.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design, 3rd ed.* New York: McGraw-Hill.

The Index on the following pages is incomplete. It includes merely a few items to test the indexing facility. The index will be expanded at a later date.

Index