# Statistical Rethinking

Second edition published 2020
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

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



First edition published by CRC Press 2015

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



*Library of Congress Cataloging-in-Publication Data*

Library of Congress Control Number:2019957006

ISBN: 978-0-367-13991-9 (hbk)
ISBN: 978-0-429-02960-8 (ebk)

# Contents

# Preface to the Second Edition

It came as a complete surprise to me that I wrote a statistics book. It is even more surprising how popular the book has become. But I had set out to write the statistics book that I wish I could have had in graduate school. No one should have to learn this stuff the way I did. I am glad there is an audience to benefit from the book.

It consumed five years to write it. There was an initial set of course notes, melted down and hammered into a first 200-page manuscript. I discarded that first manuscript. But it taught me the outline of the book I really wanted to write. Then, several years of teaching with the manuscript further refined it.

Really, I could have continued refining it every year. Going to press carries the penalty of freezing a dynamic process of both learning how to teach the material and keeping up with changes in the material. As time goes on, I see more elements of the book that I wish I had done differently. I've also received a lot of feedback on the book, and that feedback has given me ideas for improving it.

So in the second edition, I put those ideas into action. The major changes are:

**The R package has some new tools**. The `map` tool from the first edition is still here, but now it is named `quap`. This renaming is to avoid misunderstanding. We just used it to get a quadratic approximation to the posterior. So now it is named as such. A bigger change is that `map2stan` has been replaced by `ulam`. The new `ulam` is very similar to `map2stan`, and in many cases can be used identically. But it is also much more flexible, mainly because it does not make any assumptions about GLM structure and allows explicit variable types. All the `map2stan` code is still in the package and will continue to work. But now `ulam` allows for much more, especially in later chapters. Both of these tools allow sampling from the prior distribution, using `extract.prior`, as well as the posterior. This helps with the next change.

**Much more prior predictive simulation**. A prior predictive simulation means simulating predictions from a model, using only the prior distribution instead of the posterior distribution. This is very useful for understanding the implications of a prior. There was only a vestigial amount of this in the first edition. Now many modeling examples have some prior predictive simulation. I think this is one of the most useful additions to the second edition, since it helps so much with understanding not only priors but also the model itself.

**More emphasis on the distinction between prediction and inference**. Chapter 5, the chapter on multiple regression, has been split into two chapters. The first chapter focuses on helpful aspects of regression; the second focuses on ways that it can mislead. This allows as well a more direct discussion of causal inference. This means that DAGs—directed acyclic graphs—make an appearance. The chapter on overfitting, Chapter 7 now, is also more direct in cautioning about the predictive nature of information criteria and cross-validation. Cross-validation and importance sampling approximations of it are now discussed explicitly.

**New model types**. Chapter 4 now presents simple splines. Chapter 7 introduces one kind or robust regression. Chapter 12 explains how to use ordered categorical predictor variables. Chapter 13 presents a very simple type of social network model, the social relations model. Chapter 14 has an example of a phylogenetic regression, with a somewhat critical and heterodox presentation. And there is an entirely new chapter, Chapter 16, that focuses on models that are not easily conceived of as GLMMs, including ordinary differential equation models.

**Some new data examples**. There are some new data examples, including the Japanese cherry blossoms time series on the cover and a larger primate evolution data set with 300 species and a matching phylogeny.

**More presentation of raw Stan models**. There are many more places now where raw Stan model code is explained. I hope this makes a transition to working directly in Stan easier. But most of the time, working directly in Stan is still optional.

**Kindness and persistence**. As in the first edition, I have tried to make the material as kind as possible. None of this stuff is easy, and the journey into understanding is long and haunted. It is important that readers expect that confusion is normal. This is also the reason that I have not changed the basic modeling strategy in the book.

First, I force the reader to explicitly specify every assumption of the model. Some readers of the first edition lobbied me to use simplified formula tools like `brms` or `rstanarm`. Those are fantastic packages, and graduating to use them after this book is recommended. But I don't see how a person can come to understand the model when using those tools. The priors being hidden isn't the most limiting part. Instead, since linear model formulas like `y ~ (1|x) + z` don't show the parameters, nor even all of the terms, it is not easy to see how the mathematical model relates to the code. It is ultimately kinder to be a bit cruel and require more work. So the formula lists remain. You'll thank me later.

Second, half the book goes by before MCMC appears. Some readers of the first edition wanted me to start instead with MCMC. I do not do this because Bayes is not about MCMC. We seek the posterior distribution, but there are many legitimate approximations of it. MCMC is just one set of strategies. Using quadratic approximation in the first half also allows a clearer tie to non-Bayesian algorithms. And since finding the quadratic approximation is fast, it means readers don't have to struggle with too many things at once.

**Thanks**. Many readers and colleagues contributed comments that improved upon the first edition. There are too many to name individually. Several anonymous reviewers provided many pages of constructive criticism. Bret Beheim and Aki Vehtari commented on multiple chapters. My colleagues at the Max Planck Institute for Evolutionary Anthropology in Leipzig made the largest contributions, by working through draft chapters and being relentlessly honest.

Richard McElreath
Leipzig, 14 December 2019

# Preface

Masons, when they start upon a building,
Are careful to test out the scaffolding;

Make sure that planks won't slip at busy points,
Secure all ladders, tighten bolted joints.

And yet all this comes down when the job's done
Showing off walls of sure and solid stone.

So if, my dear, there sometimes seem to be
Old bridges breaking between you and me

Never fear. We may let the scaffolds fall
Confident that we have built our wall.

("Scaffolding" by Seamus Heaney, 1939–2013)

This book means to help you raise your knowledge of and confidence in statistical modeling. It is meant as a scaffold, one that will allow you to construct the wall that you need, even though you will discard it afterwards. As a result, this book teaches the material in often inconvenient fashion, forcing you to perform step-by-step calculations that are usually automated. The reason for all the algorithmic fuss is to ensure that you understand enough of the details to make reasonable choices and interpretations in your own modeling work. So although you will move on to use more automation, it's important to take things slow at first. Put up your wall, and then let the scaffolding fall.

## Audience

The principle audience is researchers in the natural and social sciences, whether new PhD students or seasoned professionals, who have had a basic course on regression but nevertheless remain uneasy about statistical modeling. This audience accepts that there is something vaguely wrong about typical statistical practice in the early twenty-first century, dominated as it is by *p*-values and a confusing menagerie of testing procedures. They see alternative methods in journals and books. But these people are not sure where to go to learn about these methods.

As a consequence, this book doesn't really argue against *p*-values and the like. The problem in my opinion isn't so much *p*-values as the set of odd rituals that have evolved around them, in the wilds of the sciences, as well as the exclusion of so many other useful tools. So the book assumes the reader is ready to try doing statistical inference without *p*-values. This isn't the ideal situation. It would be better to have material that helps you spot common mistakes and misunderstandings of *p*-values and tests in general, as all of us have to understand such things, even if we don't use them. So I've tried to sneak in a little material of that kind, but unfortunately cannot devote much space to it. The book would be too long, and it would disrupt the teaching flow of the material.

It's important to realize, however, that the disregard paid to *p*-values is not a uniquely Bayesian attitude. Indeed, significance testing can be—and has been—formulated as a Bayesian procedure as well. So the choice to avoid significance testing is stimulated instead by epistemological concerns, some of which are briefly discussed in the first chapter.

## Teaching strategy

The book uses much more computer code than formal mathematics. Even excellent mathematicians can have trouble understanding an approach, until they see a working algorithm. This is because implementation in code form removes all ambiguities. So material of this sort is easier to learn, if you also learn how to implement it.

In addition to any pedagogical value of presenting code, so much of statistics is now computational that a purely mathematical approach is anyways insufficient. As you'll see in later parts of this book, the same mathematical statistical model can sometimes be implemented in different ways, and the differences matter. So when you move beyond this book to more advanced or specialized statistical modeling, the computational emphasis here will help you recognize and cope with all manner of practical troubles.

Every section of the book is really just the tip of an iceberg. I've made no attempt to be exhaustive. Rather I've tried to explain something well. In this attempt, I've woven a lot of concepts and material into data analysis examples. So instead of having traditional units on, for example, centering predictor variables, I've developed those concepts in the context of a narrative about data analysis. This is certainly not a style that works for all readers. But it has worked for a lot of my students. I suspect it fails dramatically for those who are being forced to learn this information. For the internally motivated, it reflects how we really learn these skills in the context of our research.

## How to use this book

This book is not a reference, but a course. It doesn't try to support random access. Rather, it expects sequential access. This has immense pedagogical advantages, but it has the disadvantage of violating how most scientists actually read books.

This book has a lot of code in it, integrated fully into the main text. The reason for this is that doing model-based statistics in the twenty-first century requires simple programming. The code is really not optional. Everyplace, I have erred on the side of including too much code, rather than too little. In my experience teaching scientific programming, novices learn more quickly when they have working code to modify, rather than needing to write an algorithm from scratch. My generation was probably the last to have to learn some programming to use a computer, and so coding has gotten harder and harder to teach as time goes on. My students are very computer literate, but they sometimes have no idea what computer code looks like.

**What the book assumes**. This book does not try to teach the reader to program, in the most basic sense. It assumes that you have made a basic effort to learn how to install and process data in R. In most cases, a short introduction to R programming will be enough. I know many people have found Emmanuel Paradis' *R for Beginners* helpful. You can find it and many other beginner guides here:

http://cran.r-project.org/other-docs.html

To make use of this book, you should know already that `y<-7` stores the value 7 in the symbol `y`. You should know that symbols which end in parentheses are functions. You should recognize a loop and understand that commands can be embedded inside other commands (recursion). Knowing that R *vectorizes* a lot of code, instead of using loops, is important. But you don't have to yet be confident with R programming.

Inevitably you will come across elements of the code in this book that you haven't seen before. I have made an effort to explain any particularly important or unusual programming tricks in my own code. In fact, this book spends a lot of time explaining code. I do this because students really need it. Unless they can connect each command to the recipe and the goal, when things go wrong, they won't know whether it is because of a minor or major error. The same issue arises when I teach mathematical evolutionary theory—students and colleagues often suffer from rusty algebra skills, so when they can't get the right answer, they often don't know whether it's because of some small mathematical misstep or instead some problem in strategy. The protracted explanations of code in this book aim to build a level of understanding that allows the reader to diagnose and fix problems.

**Why R**. This book uses R for the same reason that it uses English: Lots of people know it already. R is convenient for doing computational statistics. But many other languages are equally fine. I recommend Python (especially PyMC) and Julia as well. The first edition ended up with code translations for various languages and styles. Hopefully, the second edition will as well.

**Using the code**. Code examples in the book are marked by a shaded box, and output from example code is often printed just beneath a shaded box, but marked by a fixed-width typeface. For example:

R code 0.1

```
print( "All models are wrong, but some are useful." )
```

```
[1] "All models are wrong, but some are useful."
```

Next to each snippet of code, you'll find a number that you can search for in the accompanying code snippet file, available from the book's website. The intention is that the reader follow along, executing the code in the shaded boxes and comparing their own output to that printed in the book. I really want you to execute the code, because just as one cannot learn martial arts by watching Bruce Lee movies, you can't learn to program statistical models by only reading a book. You have to get in there and throw some punches and, likewise, take some hits.

If you ever get confused, remember that you can execute each line independently and inspect the intermediate calculations. That's how you learn as well as solve problems. For example, here's a confusing way to multiply the numbers 10 and 20:

R code 0.2

```
x <- 1:2
x <- x*10
x <- log(x)
x <- sum(x)
x <- exp(x)
x
```

```
200
```

If you don't understand any particular step, you can always print out the contents of the symbol `x` immediately after that step. For the code examples, this is how you come to understand them. For your own code, this is how you find the source of any problems and then fix them.

**Optional sections**. Reflecting realism in how books like this are actually read, there are two kinds of optional sections: (1) Rethinking and (2) Overthinking. The Rethinking sections look like this:

**Rethinking: Think again**. The point of these Rethinking boxes is to provide broader context for the material. They allude to connections to other approaches, provide historical background, or call out common misunderstandings. These boxes are meant to be optional, but they round out the material and invite deeper thought.

The Overthinking sections look like this:

---

**Overthinking: Getting your hands dirty**. These sections, set in smaller type, provide more detailed explanations of code or mathematics. This material isn't essential for understanding the main text. But it does

have a lot of value, especially on a second reading. For example, sometimes it matters how you perform a calculation. Mathematics tells that these two expressions are equivalent:

$$p1 = \log(0.01^{200})$$

$$p2 = 200 \times \log(0.01)$$

But when you use R to compute them, they yield different answers:

R code 0.3

```
( log( 0.01^200 ) )
( 200 * log(0.01) )
```

```
[1] -Inf
[1] -921.034
```

The second line is the right answer. This problem arises because of rounding error, when the computer rounds very small decimal values to zero. This loses *precision* and can introduce substantial errors in inference. As a result, we nearly always do statistical calculations using the logarithm of a probability, rather than the probability itself.

---

You can ignore most of these Overthinking sections on a first read.

**The command line is the best tool**. Programming at the level needed to perform twentyfirst century statistical inference is not that complicated, but it is unfamiliar at first. Why not just teach the reader how to do all of this with a point-and-click program? There are big advantages to doing statistics with text commands, rather than pointing and clicking on menus.

Everyone knows that the command line is more powerful. But it also saves you time and fulfills ethical obligations. With a command script, each analysis documents itself, so that years from now you can come back to your analysis and replicate it exactly. You can re-use your old files and send them to colleagues. Pointing and clicking, however, leaves no trail of breadcrumbs. A file with your R commands inside it does. Once you get in the habit of planning, running, and preserving your statistical analyses in this way, it pays for itself many times over. With point-and-click, you pay down the road, rather than only up front. It is also a basic ethical requirement of science that our analyses be fully documented and repeatable. The integrity of peer review and the cumulative progress of research depend upon it. A command line statistical program makes this documentation natural. A point-and-click interface does not. Be ethical.

So we don't use the command line because we are hardcore or elitist (although we might be). We use the command line because it is better. It is harder at first. Unlike the point-and-click interface, you do have to learn a basic set of commands to get started with a command line interface. However, the ethical and cost saving advantages are worth the inconvenience.

**How you should work**. But I would be cruel, if I just told the reader to use a command-line tool, without also explaining something about how to do it. You do have to relearn some habits, but it isn't a major change. For readers who have only used menu-driven statistics software before, there will be some significant readjustment. But after a few days, it will seem natural to you. For readers who have used command-driven statistics software like Stata and SAS, there is still some readjustment ahead. I'll explain the overall approach first. Then I'll say why even Stata and SAS users are in for a change.

The sane approach to scripting statistical analyses is to work back and forth between two applications: (1) a *plain text editor* of your choice and (2) the R program running in a terminal. There are several applications that integrate the text editor with the R console. The most popular of these is RStudio. It has a lot of options, but really

it is just an interface that includes both a script editor and an R terminal.

A plain text editor is a program that creates and edits simple formatting-free text files. Common examples include Notepad (in Windows) and TextEdit (in Mac OS X) and Emacs (in most *NIX distributions, including Mac OS X). There is also a wide selection of fancy text editors specialized for programmers. You might investigate, for example, RStudio and the Atom text editor, both of which are free. Note that MSWord files are not plain text.

You will use a plain text editor to keep a running log of the commands you feed into the R application for processing. You absolutely do not want to just type out commands directly into R itself. Instead, you want to either copy and paste lines of code from your plain text editor into R, or instead read entire script files directly into R. You might enter commands directly into R as you explore data or debug or merely play. But your serious work should be implemented through the plain text editor, for the reasons explained in the previous section.

You can add comments to your R scripts to help you plan the code and remember later what the code is doing. To make a comment, just begin a line with the # symbol. To help clarify the approach, below I provide a very short complete script for running a linear regression on one of R's built-in sets of data. Even if you don't know what the code does yet, hopefully you will see it as a basic model of clarity of formatting and use of comments.

R code 0.4

```
# Load the data:
# car braking distances in feet paired with speeds in km/h

# see ?cars for details
data(cars)

# fit a linear regression of distance on speed
m <- lm( dist ~ speed , data=cars )

# estimated coefficients from the model
coef(m)

# plot residuals against speed
plot( resid(m) ~ speed , data=cars )
```

Even those who are familiar with scripting Stata or SAS will be in for some readjustment. Programs like Stata and SAS have a different paradigm for how information is processed. In those applications, procedural commands like `PROC GLM` are issued in imitation of menu commands. These procedures produce a mass of default output that the user then sifts through. R does not behave this way. Instead, R forces the user to decide which bits of information she wants. One fits a statistical model in R and then must issue later commands to ask questions about it. This more interrogative paradigm will become familiar through the examples in the text. But be aware that you are going to take a more active role in deciding what questions to ask about your models.

## Installing the `rethinking` R package

The code examples require that you have installed the `rethinking` R package. This package contains the data examples and many of the modeling tools that the text uses. The `rethinking` package itself relies upon another package, `rstan`, for fitting the more advanced models in the second half of the book.

You should install `rstan` first. Navigate your internet browser to `mc-stan.org` and follow the instructions for your platform. You will need to install both a C++ compiler (also called the "tool chain") and the `rstan` package. Instructions for doing both are at `mc-stan.org`. Then from within R, you can install `rethinking`

with this code:

R code 0.5

```
install.packages(c("coda", "mvtnorm", "devtools", "dagitty"))
library(devtools)
devtools::install_github("rmcelreath/rethinking")
```

Note that `rethinking` is not on the CRAN package archive, at least not yet. You'll always be able to perform a simple internet search and figure out the current installation instructions for the most recent version of the `rethinking` package. If you encounter any bugs while using the package, you can check `github.com/rmcelreath/rethinking` to see if a solution is already posted. If not, you can leave a bug report and be notified when a solution becomes available. In addition, all of the source code for the package is found there, in case you aspire to do some tinkering of your own. Feel free to "fork" the package and bend it to your will.

## Acknowledgments

Many people have contributed advice, ideas, and complaints to this book. Most important among them have been the graduate students who have taken statistics courses from me over the last decade, as well as the colleagues who have come to me for advice. These people taught me how to teach them this material, and in some cases I learned the material only because they needed it. A large number of individuals donated their time to comment on sections of the book or accompanying computer code. These include: Rasmus Bååth, Ryan Baldini, Bret Beheim, Maciek Chudek, John Durand, Andrew Gelman, Ben Goodrich, Mark Grote, Dave Harris, Chris Howerton, James Holland Jones, Jeremy Koster, Andrew Marshall, Sarah Mathew, Karthik Panchanathan, Pete Richerson, Alan Rogers, Cody Ross, Noam Ross, Aviva Rossi, Kari Schroeder, Paul Smaldino, Rob Trangucci, Shravan Vasishth, Annika Wallin, and a score of anonymous reviewers. Bret Beheim and Dave Harris were brave enough to provide extensive comments on an early draft. Caitlin DeRango and Kotrina Kajokaite invested their time in improving several chapters and problem sets. Mary Brooke McEachern provided crucial opinions on content and presentation, as well as calm support and tolerance. A number of anonymous reviewers provided detailed feedback on individual chapters. None of these people agree with all of the choices I have made, and all mistakes and deficiencies remain my responsibility. But especially when we haven't agreed, their opinions have made the book stronger.

The book is dedicated to Dr. Parry M. R. Clarke (1977–2012), who asked me to write it. Parry's inquisition of statistical and mathematical and computational methods helped everyone around him. He made us better.

# 1 The Golem of Prague

In the sixteenth century, the House of Habsburg controlled much of Central Europe, the Netherlands, and Spain, as well as Spain's colonies in the Americas. The House was maybe the first true world power. The Sun shone always on some portion of it. Its ruler was also Holy Roman Emperor, and his seat of power was Prague. The Emperor in the late sixteenth century, Rudolph II, loved intellectual life. He invested in the arts, the sciences (including astrology and alchemy), and mathematics, making Prague into a world center of learning and scholarship. It is appropriate then that in this learned atmosphere arose an early robot, the Golem of Prague.

A golem (goh-lem) is a clay robot from Jewish folklore, constructed from dust and fire and water. It is brought to life by inscribing *emet*, Hebrew for "truth," on its brow. Animated by truth, but lacking free will, a golem always does exactly what it is told. This is lucky, because the golem is incredibly powerful, able to withstand and accomplish more than its creators could. However, its obedience also brings danger, as careless instructions or unexpected events can turn a golem against its makers. Its abundance of power is matched by its lack of wisdom.

In some versions of the golem legend, Rabbi Judah Loew ben Bezalel sought a way to defend the Jews of Prague. As in many parts of sixteenth century Central Europe, the Jews of Prague were persecuted. Using secret techniques from the *Kabbalah*, Rabbi Judah was able to build a golem, animate it with "truth," and order it to defend the Jewish people of Prague. Not everyone agreed with Judah's action, fearing unintended consequences of toying with the power of life. Ultimately Judah was forced to destroy the golem, as its combination of extraordinary power with clumsiness eventually led to innocent deaths. Wiping away one letter from the inscription *emet* to spell instead *met*, "death," Rabbi Judah decommissioned the robot.

## 1.1. Statistical golems

Scientists also make golems.[1] Our golems rarely have physical form, but they too are often made of clay, living in silicon as computer code. These golems are scientific models. But these golems have real effects on the world, through the predictions they make and the intuitions they challenge or inspire. A concern with "truth" enlivens these models, but just like a golem or a modern robot, scientific models are neither true nor false, neither prophets nor charlatans. Rather they are constructs engineered for some purpose. These constructs are incredibly powerful, dutifully conducting their programmed calculations.

Sometimes their unyielding logic reveals implications previously hidden to their designers. These implications can be priceless discoveries. Or they may produce silly and dangerous behavior. Rather than idealized angels of reason, scientific models are powerful clay robots without intent of their own, bumbling along according to the myopic instructions they embody. Like with Rabbi Judah's golem, the golems of science are wisely regarded with both awe and apprehension. We absolutely have to use them, but doing so always entails some risk.

There are many kinds of statistical models. Whenever someone deploys even a simple statistical procedure, like a classical *t*-test, she is deploying a small golem that will obediently carry out an exact calculation, performing it the same way (nearly[2]) every time, without complaint. Nearly every branch of science relies upon the senses of statistical golems. In many cases, it is no longer possible to even measure phenomena of interest, without making use of a model. To measure the strength of natural selection or the speed of a neutrino or the number of species in the Amazon, we must use models. The golem is a prosthesis, doing the measuring for us, performing impressive calculations, finding patterns where none are obvious.

However, there is no wisdom in the golem. It doesn't discern when the context is inappropriate for its answers.

It just knows its own procedure, nothing else. It just does as it's told. And so it remains a triumph of statistical science that there are now so many diverse golems, each useful in a particular context. Viewed this way, statistics is neither mathematics nor a science, but rather a branch of engineering. And like engineering, a common set of design principles and constraints produces a great diversity of specialized applications.

This diversity of applications helps to explain why introductory statistics courses are so often confusing to the initiates. Instead of a single method for building, refining, and critiquing statistical models, students are offered a zoo of pre-constructed golems known as "tests." Each test has a particular purpose. Decision trees, like the one in FIGURE 1.1, are common. By answering a series of sequential questions, users choose the "correct" procedure for their research circumstances.

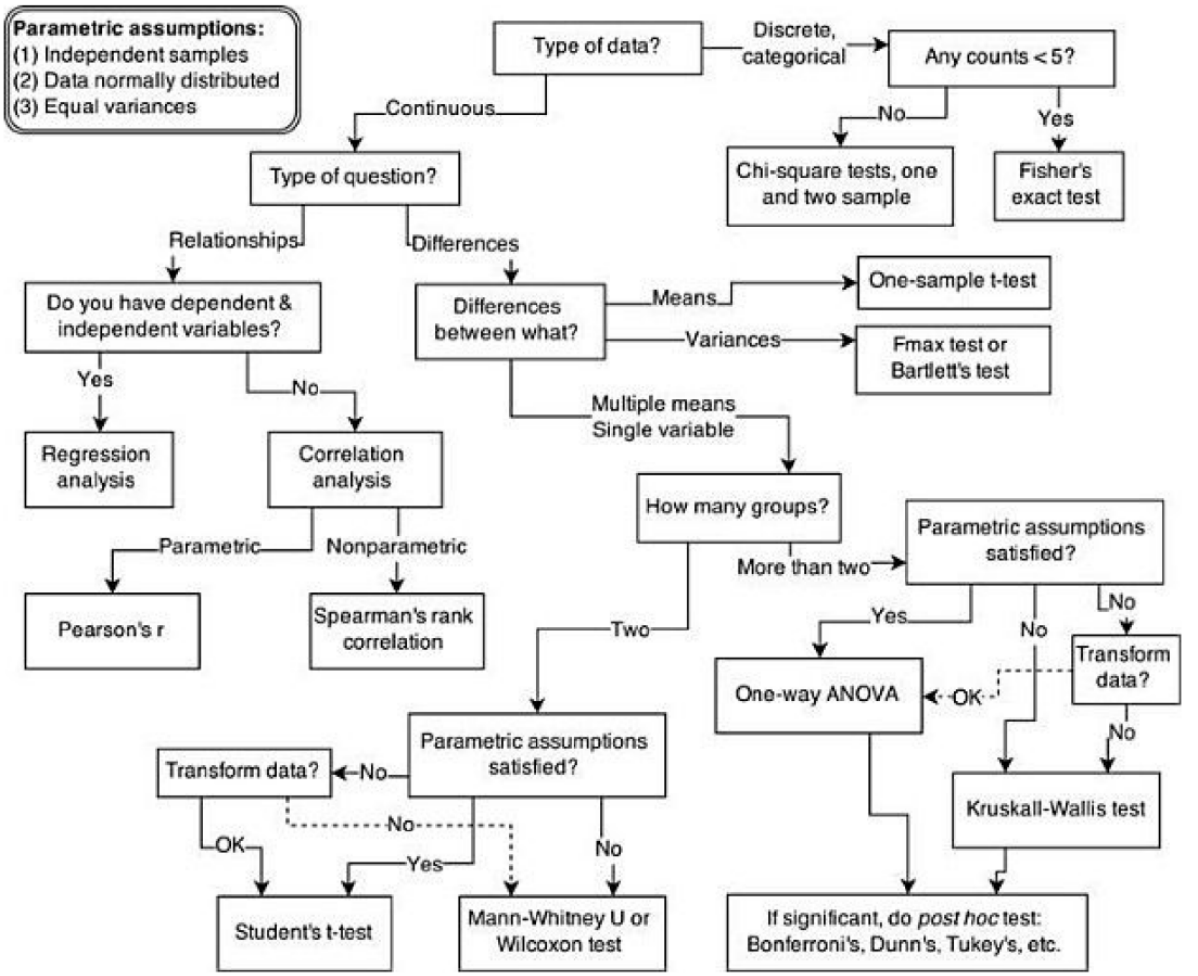



FIGURE 1.1. Example decision tree, or flowchart, for selecting an appropriate statistical procedure. Beginning at the top, the user answers a series of questions about measurement and intent, arriving eventually at the name of a procedure. Many such decision trees are possible.

Unfortunately, while experienced statisticians grasp the unity of these procedures, students and researchers rarely do. Advanced courses in statistics do emphasize engineering principles, but most scientists never get that far. Teaching statistics this way is somewhat like teaching engineering backwards, starting with bridge building and ending with basic physics. So students and many scientists tend to use charts like FIGURE 1.1 without much thought to their underlying structure, without much awareness of the models that each procedure embodies, and without any framework to help them make the inevitable compromises required by real research. It's not their fault.

For some, the toolbox of pre-manufactured golems is all they will ever need. Provided they stay within well-tested contexts, using only a few different procedures in appropriate tasks, a lot of good science can be completed. This is similar to how plumbers can do a lot of useful work without knowing much about fluid dynamics. Serious trouble begins when scholars move on to conducting innovative research, pushing the boundaries of their specialties. It's as if we got our hydraulic engineers by promoting plumbers.

Why aren't the tests enough for research? The classical procedures of introductory statistics tend to be inflexible and fragile. By inflexible, I mean that they have very limited ways to adapt to unique research contexts. By fragile, I mean that they fail in unpredictable ways when applied to new contexts. This matters, because at the boundaries of most sciences, it is hardly ever clear which procedure is appropriate. None of the traditional golems has been evaluated in novel research settings, and so it can be hard to choose one and then to understand how it behaves. A good example is *Fisher's exact test*, which applies (exactly) to an extremely narrow empirical context, but is regularly used whenever cell counts are small. I have personally read hundreds of uses of Fisher's exact test in scientific journals, but aside from Fisher's original use of it, I have never seen it used appropriately. Even a procedure like ordinary linear regression, which is quite flexible in many ways, being able to encode a large diversity of interesting hypotheses, is sometimes fragile. For example, if there is substantial measurement error on prediction variables, then the procedure can fail in spectacular ways. But more importantly, it is nearly always possible to do better than ordinary linear regression, largely because of a phenomenon known as **OVERFITTING** (Chapter 7).

The point isn't that statistical tools are specialized. Of course they are. The point is that classical tools are not diverse enough to handle many common research questions. Every active area of science contends with unique difficulties of measurement and interpretation, converses with idiosyncratic theories in a dialect barely understood by other scientists from other tribes. Statistical experts outside the discipline can help, but they are limited by lack of fluency in the empirical and theoretical concerns of the discipline.

Furthermore, no statistical tool does anything on its own to address the basic problem of inferring causes from evidence. Statistical golems do not understand cause and effect. They only understand association. Without our guidance and skepticism, pre-manufactured golems may do nothing useful at all. Worse, they might wreck Prague.

What researchers need is some unified theory of golem engineering, a set of principles for designing, building, and refining special-purpose statistical procedures. Every major branch of statistical philosophy possesses such a unified theory. But the theory is never taught in introductory—and often not even in advanced—courses. So there are benefits in rethinking statistical inference as a set of strategies, instead of a set of pre-made tools.

## 1.2. Statistical rethinking

A lot can go wrong with statistical inference, and this is one reason that beginners are so anxious about it. When the goal is to choose a pre-made test from a flowchart, then the anxiety can mount as one worries about choosing the "correct" test. Statisticians, for their part, can derive pleasure from scolding scientists, making the psychological battle worse.

But anxiety can be cultivated into wisdom. That is the reason that this book insists on working with the computational nuts and bolts of each golem. If you don't understand how the golem processes information, then you can't interpret the golem's output. This requires knowing the model in greater detail than is customary, and it requires doing the computations the hard way, at least until you are wise enough to use the push-button solutions.

There are conceptual obstacles as well, obstacles with how scholars define statistical objectives and interpret statistical results. Understanding any individual golem is not enough, in these cases. Instead, we need some statistical epistemology, an appreciation of how statistical models relate to hypotheses and the natural mechanisms of interest. What are we supposed to be doing with these little computational machines, anyway?

The greatest obstacle that I encounter among students and colleagues is the tacit belief that the proper objective of statistical inference is to test null hypotheses.[3] This is the proper objective, the thinking goes, because Karl Popper argued that science advances by falsifying hypotheses. Karl Popper (1902–1994) is possibly the most influential philosopher of science, at least among scientists. He did persuasively argue that science works better

by developing hypotheses that are, in principle, falsifiable. Seeking out evidence that might embarrass our ideas is a normative standard, and one that most scholars—whether they describe themselves as scientists or not—subscribe to. So maybe statistical procedures should falsify hypotheses, if we wish to be good statistical scientists.

But the above is a kind of folk Popperism, an informal philosophy of science common among scientists but not among philosophers of science. Science is not described by the falsification standard, and Popper recognized that.[4] In fact, deductive falsification is impossible in nearly every scientific context. In this section, I review two reasons for this impossibility.

(1) Hypotheses are not models. The relations among hypotheses and different kinds of models are complex. Many models correspond to the same hypothesis, and many hypotheses correspond to a single model. This makes strict falsification impossible.

(2) Measurement matters. Even when we think the data falsify a model, another observer will debate our methods and measures. They don't trust the data. Sometimes they are right.

For both of these reasons, deductive falsification never works. The scientific method cannot be reduced to a statistical procedure, and so our statistical methods should not pretend. Statistical evidence is part of the hot mess that is science, with all of its combat and egotism and mutual coercion. If you believe, as I do, that science does often work, then learning that it doesn't work via falsification shouldn't change your mind. But it might help you do better science. It might open your eyes to many legitimately useful functions of statistical golems.

**Rethinking: Is NHST falsificationist?** Null hypothesis significance testing, NHST, is often identified with the falsificationist, or Popperian, philosophy of science. However, usually NHST is used to falsify a null hypothesis, not the actual research hypothesis. So the falsification is being done to something other than the explanatory model. This seems the reverse from Karl Popper's philosophy.[5]

**1.2.1. Hypotheses are not models.** When we attempt to falsify a hypothesis, we must work with a model of some kind. Even when the attempt is not explicitly statistical, there is always a tacit model of measurement, of evidence, that operationalizes the hypothesis. All models are false,[6] so what does it mean to falsify a model? One consequence of the requirement to work with models is that it's no longer possible to deduce that a hypothesis is false, just because we reject a model derived from it.

Let's explore this consequence in the context of an example from population biology (FIGURE 1.2). Beginning in the 1960s, evolutionary biologists became interested in the proposal that the majority of evolutionary changes in gene frequency are caused not by natural selection, but rather by mutation and drift. No one really doubted that natural selection is responsible for functional design. This was a debate about genetic sequences. So began several productive decades of scholarly combat over "neutral" models of molecular evolution.[7] This combat is most strongly associated with Motoo Kimura (1924–1994), who was perhaps the strongest advocate of neutral models. But many other population geneticists participated. As time has passed, related disciplines such as community ecology[8] and anthropology[9] have experienced (or are currently experiencing) their own versions of the neutrality debate.

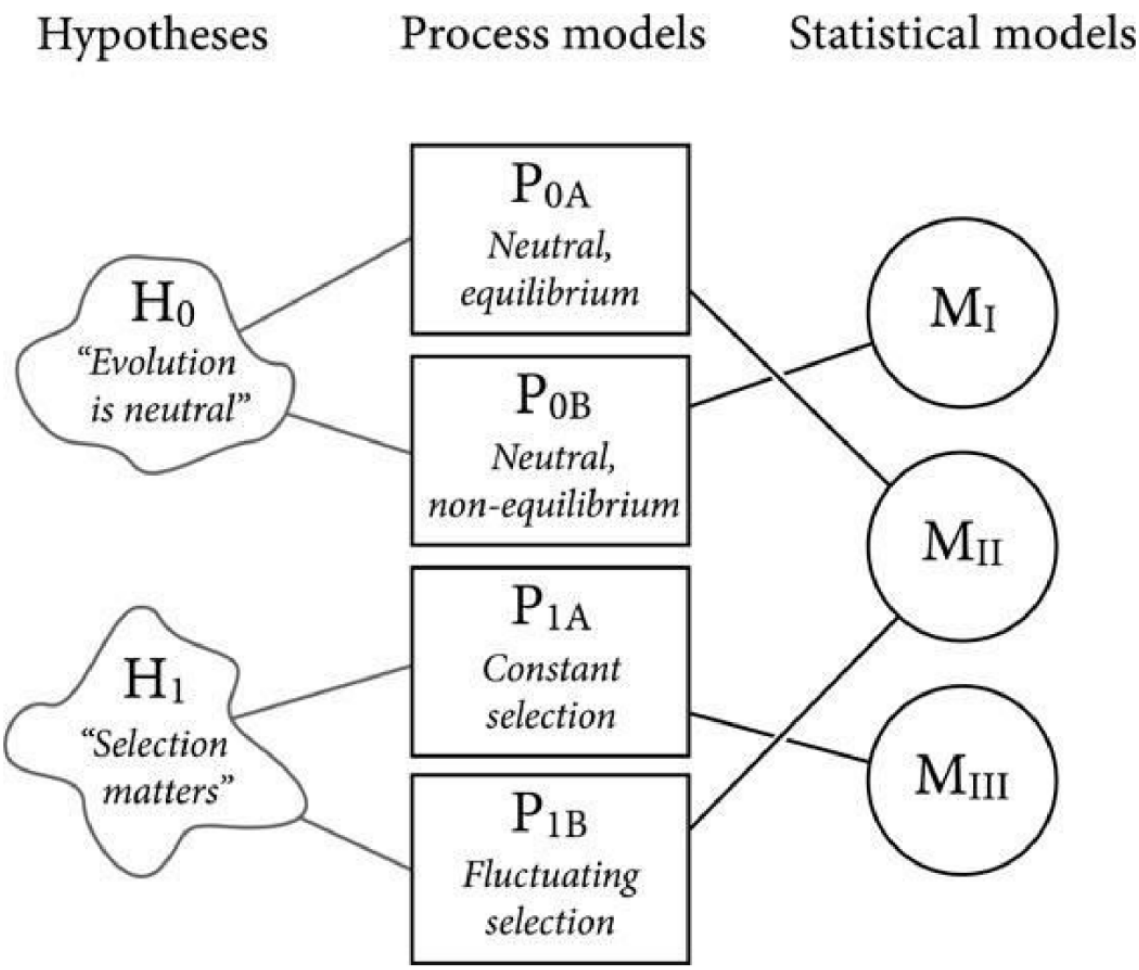



FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of "neutral" models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses rarely address process models directly. Instead, they rely upon statistical models (M), all of which reflect only some aspects of the process models. As a result, relations are multiple in both directions: Hypotheses do not imply unique models, and models do not imply unique hypotheses. This fact greatly complicates statistical inference.

Let's use the schematic in FIGURE 1.2 to explore connections between motivating hypotheses and different models, in the context of the neutral evolution debate. On the left, there are two stereotyped, informal hypotheses: Either evolution is "neutral" ($H_0$) or natural selection matters somehow ($H_1$). These hypotheses have vague boundaries, because they begin as verbal conjectures, not precise models. There are hundreds of possible detailed processes that can be described as "neutral," depending upon choices about population structure, number of sites, number of alleles at each site, mutation rates, and recombination.

Once we have made these choices, we have the middle column in FIGURE 1.2, detailed **PROCESS MODELS** of evolution. $P_{0A}$ and $P_{0B}$ differ in that one assumes the population size and structure have been constant long enough for the distribution of alleles to reach a steady state. The other imagines instead that population size fluctuates through time, which can be true even when there is no selective difference among alleles. The "selection matters" hypothesis $H_1$ likewise corresponds to many different process models. I've shown two big players: a model in which selection always favors certain alleles and another in which selection fluctuates through time, favoring different alleles.[10]

An important feature of these process models is that they express causal structure. Different process models

formalize different cause and effect relationships. Whether analyzed mathematically or through simulation, the direction of time in a model means that some things cause other things, but not the reverse. You can use such models to perform experiments and probe their causal implications. Sometimes these probes reveal, before we even turn to statistical inference, that the model cannot explain a phenomenon of interest.

In order to challenge process models with data, they have to be made into statistical models. Unfortunately, statistical models do not embody specific causal relationships. A statistical model expresses associations among variables. As a result, many different process models may be consistent with any single statistical model.

How do we get a statistical model from a causal model? One way is to derive the expected frequency distribution of some quantity—a "statistic"—from the causal model. For example, a common statistic in this context is the frequency distribution (histogram) of the frequency of different genetic variants (alleles). Some alleles are rare, appearing in only a few individuals. Others are very common, appearing in very many individuals in the population. A famous result in population genetics is that a model like $P_{0A}$ produces a *power law* distribution of allele frequencies. And so this fact yields a statistical model, $M_{II}$, that predicts a power law in the data. In contrast the constant selection process model $P_{1A}$ predicts something quite different, $M_{III}$.

Unfortunately, other selection models ($P_{1B}$) imply the same statistical model, $M_{II}$, as the neutral model. They also produce power laws. So we've reached the uncomfortable lesson:

(1) Any given statistical model (M) may correspond to more than one process model (P).
(2) Any given hypothesis (H) may correspond to more than one process model (P).
(3) Any given statistical model (M) may correspond to more than one hypothesis (H).

Now look what happens when we compare the statistical models to data. The classical approach is to take the "neutral" model as a null hypothesis. If the data are not sufficiently similar to the expectation under the null, then we say that we "reject" the null hypothesis. Suppose we follow the history of this subject and take $P_{0A}$ as our null hypothesis. This implies data corresponding to $M_{II}$. But since the same statistical model corresponds to a selection model $P_{1B}$, it's not clear what to make of either rejecting or accepting the null. The null model is not unique to any process model nor hypothesis. If we reject the null, we can't really conclude that selection matters, because there are other neutral models that predict different distributions of alleles. And if we fail to reject the null, we can't really conclude that evolution is neutral, because some selection models expect the same frequency distribution.

This is a huge bother. Once we have the diagram in FIGURE 1.2, it's easy to see the problem. But few of us are so lucky. While population genetics has recognized this issue, scholars in other disciplines continue to test frequency distributions against power law expectations, arguing even that there is only one neutral model.[11] Even if there were only one neutral model, there are so many non-neutral models that mimic the predictions of neutrality, that neither rejecting nor failing to reject the null model carries much inferential power.

So what can be done? Well, if you have multiple process models, a lot can be done. If it turns out that all of the process models of interest make very similar predictions, then you know to search for a different description of the evidence, a description under which the processes look different. For example, while $P_{0A}$ and $P_{1B}$ make very similar power law predictions for the frequency distribution of alleles, they make very dissimilar predictions for the distribution of changes in allele frequency over time. Explicitly compare predictions of more than one model, and you can save yourself from some ordinary kinds of folly.

Statistical models can be confused in other ways as well, such as the confusion caused by unobserved variables and sampling bias. Process models allow us to design statistical models with these problems in mind. The statistical model alone is not enough.

**Rethinking: Entropy and model identification**. One reason that statistical models routinely correspond to many different detailed process models is because they rely upon distributions like the normal, binomial, Poisson, and others. These distributions are members of a family, the **EXPONENTIAL FAMILY**. Nature loves the members of this family. Nature loves them because nature loves entropy, and all of the exponential family

distributions are **MAXIMUM ENTROPY** distributions. Taking the natural personification out of that explanation will wait until Chapter 10. The practical implication is that one can no more infer evolutionary process from a power law than one can infer developmental process from the fact that height is normally distributed. This fact should make us humble about what typical regression models—the meat of this book—can teach us about mechanistic process. On the other hand, the maximum entropy nature of these distributions means we can use them to do useful statistical work, even when we can't identify the underlying process.

**1.2.2. Measurement matters.** The logic of falsification is very simple. We have a hypothesis H, and we show that it entails some observation D. Then we look for D. If we don't find it, we must conclude that H is false. Logicians call this kind of reasoning *modus tollens*, which is Latin shorthand for "the method of destruction." In contrast, finding D tells us nothing certain about H, because other hypotheses might also predict D.

A compelling scientific fable that employs *modus tollens* concerns the color of swans. Before discovering Australia, all swans that any European had ever seen had white feathers. This led to the belief that all swans are white. Let's call this a formal hypothesis:

$H_0$: All swans are white.

When Europeans reached Australia, however, they encountered swans with black feathers. This evidence seemed to instantly prove $H_0$ to be false. Indeed, not all swans are white. Some are certainly black, according to all observers. The key insight here is that, before voyaging to Australia, no number of observations of white swans could prove $H_0$ to be true. However it required only one observation of a black swan to prove it false.

This is a seductive story. If we can believe that important scientific hypotheses can be stated in this form, then we have a powerful method for improving the accuracy of our theories: look for evidence that disconfirms our hypotheses. Whenever we find a black swan, $H_0$ must be false. Progress!

Seeking disconfirming evidence is important, but it cannot be as powerful as the swan story makes it appear. In addition to the correspondence problems among hypotheses and models, discussed in the previous section, most of the problems scientists confront are not so logically discrete. Instead, we most often face two simultaneous problems that make the swan fable misrepresentative. First, observations are prone to error, especially at the boundaries of scientific knowledge. Second, most hypotheses are quantitative, concerning degrees of existence, rather than discrete, concerning total presence or absence. Let's briefly consider each of these problems.

*1.2.2.1. Observation error.* All observers agree under most conditions that a swan is either black or white. There are few intermediate shades, and most observers' eyes work similarly enough that there will be little disagreement about which swans are white and which are black. But this kind of example is hardly commonplace in science, at least in mature fields. Instead, we routinely confront contexts in which we are not sure if we have detected a disconfirming result. At the edges of scientific knowledge, the ability to measure a hypothetical phenomenon is often in question as much as the phenomenon itself. Here are two examples.

In 2005, a team of ornithologists from Cornell claimed to have evidence of an individual Ivory-billed Woodpecker (*Campephilus principalis*), a species thought extinct. The hypothesis implied here is:

$H_0$: The Ivory-billed Woodpecker is extinct.

It would only take one observation to falsify this hypothesis. However, many doubted the evidence. Despite extensive search efforts and a $50,000 cash reward for information leading to a live specimen, no satisfying evidence has yet (by 2020) emerged. Even if good physical evidence does eventually arise, this episode should serve as a counterpoint to the swan story. Finding disconfirming cases is complicated by the difficulties of observation. Black swans are not always really black swans, and sometimes white swans are really black swans. There are mistaken confirmations (false positives) and mistaken disconfirmations (false negatives). Against this background of measurement difficulties, scientists who already believe that the Ivory-billed Woodpecker is

extinct will always be suspicious of a claimed falsification. Those who believe it is still alive will tend to count the vaguest evidence as falsification.

Another example, this one from physics, focuses on the detection of faster-than-light (FTL) neutrinos.[12] In September 2011, a large and respected team of physicists announced detection of neutrinos—small, neutral subatomic particles able to pass easily and harmlessly through most matter—that arrived from Switzerland to Italy in slightly faster-than-lightspeed time. According to Einstein, neutrinos cannot travel faster than the speed of light. So this seems to be a falsification of special relativity. If so, it would turn physics on its head.

The dominant reaction from the physics community was not "Einstein was wrong!" but instead "How did the team mess up the measurement?" The team that made the measurement had the same reaction, and asked others to check their calculations and attempt to replicate the result.

What could go wrong in the measurement? You might think measuring speed is a simple matter of dividing distance by time. It is, at the scale and energy you live at. But with a fundamental particle like a neutrino, if you measure when it starts its journey, you stop the journey. The particle is consumed by the measurement. So more subtle approaches are needed. The detected difference from light-speed, furthermore, is quite small, and so even the latency of the time it takes a signal to travel from a detector to a control room can be orders of magnitude larger. And since the "measurement" in this case is really an estimate from a statistical model, all of the assumptions of the model are now suspect. By 2013, the physics community was unanimous that the FTL neutrino result was measurement error. They found the technical error, which involved a poorly attached cable.[13] Furthermore, neutrinos clocked from supernova events are consistent with Einstein, and those distances are much larger and so would reveal differences in speed much better.

In both the woodpecker and neutrino dramas, the key dilemma is whether the falsification is real or spurious. Measurement is complicated in both cases, but in quite different ways, rendering both true-detection and false-detection plausible. Popper was aware of this limitation inherent in measurement, and it may be one reason that Popper himself saw science as being broader than falsification. But the probabilistic nature of evidence rarely appears when practicing scientists discuss the philosophy and practice of falsification.[14] My reading of the history of science is that these sorts of measurement problems are the norm, not the exception.[15]

*1.2.2.2. Continuous hypotheses.* Another problem for the swan story is that most interesting scientific hypotheses are not of the kind "all swans are white" but rather of the kind:

$$H_0\text{: 80\% of swans are white.}$$

Or maybe:

$$H_0\text{: Black swans are rare.}$$

Now what are we to conclude, after observing a black swan? The null hypothesis doesn't say black swans do not exist, but rather that they have some frequency. The task here is not to disprove or prove a hypothesis of this kind, but rather to estimate and explain the distribution of swan coloration as accurately as we can. Even when there is no measurement error of any kind, this problem will prevent us from applying the *modus tollens* swan story to our science.[16]

You might object that the hypothesis above is just not a good scientific hypothesis, because it isn't easy to disprove. But if that's the case, then most of the important questions about the world are not good scientific hypotheses. In that case, we should conclude that the definition of a "good hypothesis" isn't doing us much good. Now, nearly everyone agrees that it is a good practice to design experiments and observations that can differentiate competing hypotheses. But in many cases, the comparison must be probabilistic, a matter of degree, not kind.[17]

**1.2.3. Falsification is consensual.** The scientific community does come to regard some hypotheses as false. The

caloric theory of heat and the geocentric model of the universe are no longer taught in science courses, unless it's to teach how they were falsified. And evidence often—but not always—has something to do with such falsification.

But falsification is always *consensual*, not *logical*. In light of the real problems of measurement error and the continuous nature of natural phenomena, scientific communities argue towards consensus about the meaning of evidence. These arguments can be messy. After the fact, some textbooks misrepresent the history so it appears like logical falsification.[18] Such historical revisionism may hurt everyone. It may hurt scientists, by rendering it impossible for their own work to live up to the legends that precede them. It may make science an easy target, by promoting an easily attacked model of scientific epistemology. And it may hurt the public, by exaggerating the definitiveness of scientific knowledge.[19]

## 1.3. Tools for golem engineering

So if attempting to mimic falsification is not a generally useful approach to statistical methods, what are we to do? We are to model. Models can be made into testing procedures—all statistical tests are also models[20]—but they can also be used to design, forecast, and argue. Doing research benefits from the ability to produce and manipulate models, both because scientific problems are more general than "testing" and because the pre-made golems you maybe met in introductory statistics courses are ill-fit to many research contexts. You may not even know which statistical model to use, unless you have a generative model in addition.

If you want to reduce your chances of wrecking Prague, then some golem engineering know-how is needed. Make no mistake: You will wreck Prague eventually. But if you are a good golem engineer, at least you'll notice the destruction. And since you'll know a lot about how your golem works, you stand a good chance to figure out what went wrong. Then your next golem won't be as bad. Without engineering training, you're always at someone's mercy.

We want to use our models for several distinct purposes: designing inquiry, extracting information from data, and making predictions. In this book I've chosen to focus on tools to help with each purpose. These tools are:

(1) Bayesian data analysis
(2) Model comparison
(3) Multilevel models
(4) Graphical causal models

These tools are deeply related to one another, so it makes sense to teach them together. Understanding of these tools comes, as always, only with implementation—you can't comprehend golem engineering until you do it. And so this book focuses mostly on code, how to do things. But in the remainder of this chapter, I provide introductions to these tools.

**1.3.1. Bayesian data analysis.** Supposing you have some data, how should you use it to learn about the world? There is no uniquely correct answer to this question. Lots of approaches, both formal and heuristic, can be effective. But one of the most effective and general answers is to use Bayesian data analysis. Bayesian data analysis takes a question in the form of a model and uses logic to produce an answer in the form of probability distributions.

In modest terms, Bayesian data analysis is no more than counting the numbers of ways the data could happen, according to our assumptions. Things that can happen more ways are more plausible. Probability theory is relevant because probability is just a calculus for counting. This allows us to use probability theory as a general way to represent plausibility, whether in reference to countable events in the world or rather theoretical constructs like parameters. The rest follows logically. Once we have defined the statistical model, Bayesian data analysis forces a purely logical way of processing the data to produce inference.

Chapter 2 explains this in depth. For now, it will help to have another approach to compare. Bayesian probability is a very general approach to probability, and it includes as a special case another important approach,

the **FREQUENTIST** approach. The frequentist approach requires that all probabilities be defined by connection to the frequencies of events in very large samples.[21] This leads to frequentist uncertainty being premised on imaginary resampling of data—if we were to repeat the measurement many many times, we would end up collecting a list of values that will have some pattern to it. It means also that parameters and models cannot have probability distributions, only measurements can. The distribution of these measurements is called a **SAMPLING DISTRIBUTION**. This resampling is never done, and in general it doesn't even make sense—it is absurd to consider repeat sampling of the diversification of song birds in the Andes. As Sir Ronald Fisher, one of the most important frequentist statisticians of the twentieth century, put it:[22]

> […] the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination [… ]

But in many contexts, like controlled greenhouse experiments, it's a useful device for describing uncertainty. Whatever the context, it's just part of the model, an assumption about what the data would look like under resampling. It's just as fantastical as the Bayesian gambit of using probability to describe all types of uncertainty, whether empirical or epistemological.[23]

But these different attitudes towards probability do enforce different trade-offs. Consider this simple example where the difference between Bayesian and frequentist probability matters. In the year 1610, Galileo turned a primitive telescope to the night sky and became the first human to see Saturn's rings. Well, he probably saw a blob, with some smaller blobs attached to it (FIGURE 1.3). Since the telescope was primitive, it couldn't really focus the image very well. Saturn always appeared blurred. This is a statistical problem, of a sort. There's uncertainty about the planet's shape, but notice that none of the uncertainty is a result of variation in repeat measurements. We could look through the telescope a thousand times, and it will always give the same blurred image (for any given position of the Earth and Saturn). So the sampling distribution of any measurement is constant, because the measurement is deterministic—there's nothing "random" about it. Frequentist statistical inference has a lot of trouble getting started here. In contrast, Bayesian inference proceeds as usual, because the deterministic "noise" can still be modeled using probability, as long as we don't identify probability with frequency. As a result, the field of image reconstruction and processing is dominated by Bayesian algorithms.[24]

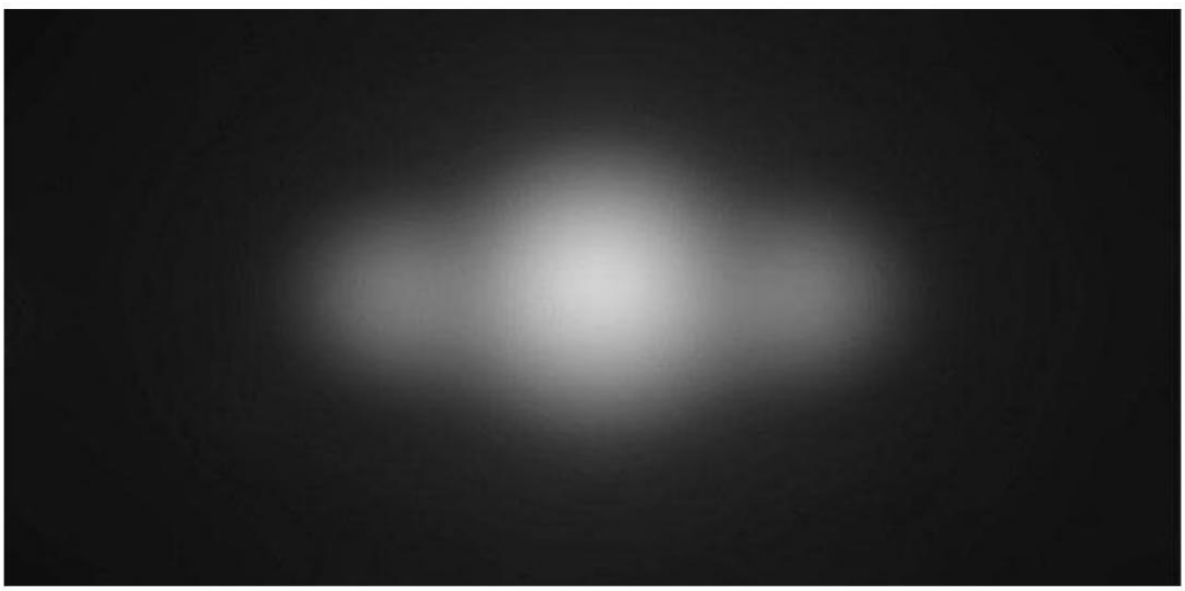


FIGURE 1.3. Saturn, much like Galileo must have seen it. The true shape is uncertain, but not because of any sampling variation. Probability theory can still help.

In more routine statistical procedures, like linear regression, this difference in probability concepts has less of an effect. However, it is important to realize that even when a Bayesian procedure and frequentist procedure give exactly the same answer, our Bayesian golems aren't justifying their inferences with imagined repeat sampling. More generally, Bayesian golems treat "randomness" as a property of information, not of the world. Nothing in the real world—excepting controversial interpretations of quantum physics—is actually random. Presumably, if we had more information, we could exactly predict everything. We just use randomness to describe our uncertainty in the face of incomplete knowledge. From the perspective of our golem, the coin toss is "random," but it's really the golem that is random, not the coin.

Note that the preceding description doesn't invoke anyone's "beliefs" or subjective opinions. Bayesian data analysis is just a logical procedure for processing information. There is a tradition of using this procedure as a normative description of rational belief, a tradition called **BAYESIANISM**.[25] But this book neither describes nor advocates it. In fact, I'll argue that no statistical approach, Bayesian or otherwise, is by itself sufficient.

Before moving on to describe the next two tools, it's worth emphasizing an advantage of Bayesian data analysis, at least when scholars are learning statistical modeling. This entire book could be rewritten to remove any mention of "Bayesian." In places, it would become easier. In others, it would become much harder. But having taught applied statistics both ways, I have found that the Bayesian framework presents a distinct pedagogical advantage: many people find it more intuitive. Perhaps the best evidence for this is that very many scientists interpret non-Bayesian results in Bayesian terms, for example interpreting ordinary *p*-values as Bayesian posterior probabilities and non-Bayesian confidence intervals as Bayesian ones (you'll learn posterior probability and confidence intervals in Chapters 2 and 3). Even statistics instructors make these mistakes.[26] Statisticians appear doomed to republish the same warnings about misinterpretation of *p*-values forever. In this sense then, Bayesian models lead to more intuitive interpretations, the ones scientists tend to project onto statistical results. The opposite pattern of mistake—interpreting a posterior probability as a *p*-value—seems to happen only rarely.

None of this ensures that Bayesian analyses will be more correct than non-Bayesian analyses. It just means that the scientist's intuitions will less commonly be at odds with the actual logic of the framework. This simplifies some of the aspects of teaching statistical modeling.

**Rethinking: Probability is not unitary**. It will make some readers uncomfortable to suggest that there is more than one way to define "probability." Aren't mathematical concepts uniquely correct? They are not. Once you adopt some set of premises, or axioms, everything does follow logically in mathematical systems. But the axioms are open to debate and interpretation. So not only is there "Bayesian" and "frequentist" probability, but there are different versions of Bayesian probability even, relying upon different arguments to justify the approach. In more advanced Bayesian texts, you'll come across names like Bruno de Finetti, Richard T. Cox, and Leonard "Jimmie" Savage. Each of these figures is associated with a somewhat different conception of Bayesian probability. There are others. This book mainly follows the "logical" Cox (or Laplace-Jeffreys-Cox-Jaynes) interpretation. This interpretation is presented beginning in the next chapter, but unfolds fully only in Chapter 10.

How can different interpretations of probability theory thrive? By themselves, mathematical entities don't necessarily "mean" anything, in the sense of real world implication. What does it mean to take the square root of a negative number? What does it mean to take a limit as something approaches infinity? These are essential and routine concepts, but their meanings depend upon context and analyst, upon beliefs about how well abstraction represents reality. Mathematics doesn't access the real world directly. So answering such questions remains a contentious and entertaining project, in all branches of applied mathematics. So while everyone subscribes to the same axioms of probability, not everyone agrees in all contexts about how to interpret probability.

**Rethinking: A little history**. Bayesian statistical inference is much older than the typical tools of introductory statistics, most of which were developed in the early twentieth century. Versions of the Bayesian approach were applied to scientific work in the late 1700s and repeatedly in the nineteenth century. But after World War I, anti-

Bayesian statisticians, like Sir Ronald Fisher, succeeded in marginalizing the approach. All Fisher said about Bayesian analysis (then called *inverse probability)* in his influential 1925 handbook was:[27]

> […] the theory of inverse probability is founded upon an error, and must be wholly rejected.

Bayesian data analysis became increasingly accepted within statistics during the second half of the twentieth century, because it proved not to be founded upon an error. All philosophy aside, it worked. Beginning in the 1990s, new computational approaches led to a rapid rise in application of Bayesian methods.[28] Bayesian methods remain computationally expensive, however. And so as data sets have increased in scale—millions of rows is common in genomic analysis, for example—alternatives to or approximations to Bayesian inference remain important, and probably always will.

**1.3.2. Model comparison and prediction.** Bayesian data analysis provides a way for models to learn from data. But when there is more than one plausible model—and in most mature fields there should be—how should we choose among them? One answer is to prefer models that make good predictions. This answer creates a lot of new questions, since knowing which model will make the best predictions seems to require knowing the future. We'll look at two related tools, neither of which knows the future: **CROSS-VALIDATION** and **INFORMATION CRITERIA**. These tools aim to compare models based upon expected predictive accuracy.

Comparing models by predictive accuracy can be useful in itself. And it will be even more useful because it leads to the discovery of an amazing fact: Complex models often make worse predictions than simpler models. The primary paradox of prediction is **OVERFITTING**.[29] Future data will not be exactly like past data, and so any model that is unaware of this fact tends to make worse predictions than it could. And more complex models tend towards more overfitting than simple ones—the smarter the golem, the dumber its predictions. So if we wish to make good predictions, we cannot judge our models simply on how well they fit our data. *Fitting is easy; prediction is hard.*

Cross-validation and information criteria help us in three ways. First, they provide useful expectations of predictive accuracy, rather than merely fit to sample. So they compare models where it matters. Second, they give us an estimate of the tendency of a model to overfit. This will help us to understand how models and data interact, which in turn helps us to design better models. We'll take this point up again in the next section. Third, cross-validation and information criteria help us to spot highly influential observations.

Bayesian data analysis has been worked on for centuries. Information criteria are comparatively very young and the field is evolving quickly. Many statisticians have never used information criteria in an applied problem, and there is no consensus about which metrics are best and how best to use them. Still, information criteria are already in frequent use in the sciences, appearing in prominent publications and featuring in prominent debates.[30] Their power is often exaggerated, and we will be careful to note what they cannot do as well as what they can.

**Rethinking: The Neanderthal in you**. Even simple models need alternatives. In 2010, a draft genome of a Neanderthal demonstrated more DNA sequences in common with non-African contemporary humans than with African ones. This finding is consistent with interbreeding between Neanderthals and modern humans, as the latter dispersed from Africa. However, just finding DNA in common between modern Europeans and Neanderthals is not enough to demonstrate interbreeding. It is also consistent with ancient structure in the African continent.[31] In short, if ancient northeast Africans had unique DNA sequences, then both Neanderthals and modern Europeans could possess these sequences from a common ancestor, rather than from direct interbreeding. So even in the seemingly simple case of estimating whether Neanderthals and modern humans share unique DNA, there is more than one process-based explanation. Model comparison is necessary.

**1.3.3. Multilevel models.** In an apocryphal telling of Hindu cosmology, it is said that the Earth rests on the back of a great elephant, who in turn stands on the back of a massive turtle. When asked upon what the turtle stands, a

guru is said to reply, "it's turtles all the way down."

Statistical models don't contain turtles, but they do contain parameters. And parameters support inference. Upon what do parameters themselves stand? Sometimes, in some of the most powerful models, it's parameters all the way down. What this means is that any particular parameter can be usefully regarded as a placeholder for a missing model. Given some model of how the parameter gets its value, it is simple enough to embed the new model inside the old one. This results in a model with multiple levels of uncertainty, each feeding into the next—a **MULTILEVEL MODEL**.

Multilevel models—also known as hierarchical, random effects, varying effects, or mixed effects models—are becoming *de rigueur* in the biological and social sciences. Fields as diverse as educational testing and bacterial phylogenetics now depend upon routine multilevel models to process data. Like Bayesian data analysis, multilevel modeling is not particularly new. But it has only been available on desktop computers for a few decades. And since such models have a natural Bayesian representation, they have grown hand-in-hand with Bayesian data analysis.

One reason to be interested in multilevel models is because they help us deal with overfitting. Cross-validation and information criteria measure overfitting risk and help us to recognize it. Multilevel models actually do something about it. What they do is exploit an amazing trick known as **PARTIAL POOLING** that pools information across units in the data in order to produce better estimates for all units. The details will wait until Chapter 13.

Partial pooling is the key technology, and the contexts in which it is appropriate are diverse. Here are four commonplace examples.

(1) *To adjust estimates for repeat sampling*. When more than one observation arises from the same individual, location, or time, then traditional, single-level models may mislead us.

(2) *To adjust estimates for imbalance in sampling*. When some individuals, locations, or times are sampled more than others, we may also be misled by single-level models.

(3) *To study variation*. If our research questions include variation among individuals or other groups within the data, then multilevel models are a big help, because they model variation explicitly.

(4) *To avoid averaging*. Pre-averaging data to construct variables can be dangerous. Averaging removes variation, manufacturing false confidence. Multilevel models preserve the uncertainty in the original, pre-averaged values, while still using the average to make predictions.

All four apply to contexts in which the researcher recognizes clusters or groups of measurements that may differ from one another. These clusters or groups may be individuals such as different students, locations such as different cities, or times such as different years. Since each cluster may well have a different average tendency or respond differently to any treatment, clustered data often benefit from being modeled by a golem that expects such variation.

But the scope of multilevel modeling is much greater than these examples. Diverse model types turn out to be multilevel: models for missing data (imputation), measurement error, factor analysis, some time series models, types of spatial and network regression, and phylogenetic regressions all are special applications of the multilevel strategy. And some commonplace procedures, like the paired *t*-test, are really multilevel models in disguise. Grasping the concept of multilevel modeling may lead to a perspective shift. Suddenly single-level models end up looking like mere components of multilevel models. The multilevel strategy provides an engineering principle to help us to introduce these components into a particular analysis, exactly where we think we need them.

I want to convince the reader of something that appears unreasonable: *multilevel regression deserves to be the default form of regression*. Papers that do not use multilevel models should have to justify not using a multilevel approach. Certainly some data and contexts do not need the multilevel treatment. But most contemporary studies in the social and natural sciences, whether experimental or not, would benefit from it. Perhaps the most important

reason is that even well-controlled treatments interact with unmeasured aspects of the individuals, groups, or populations studied. This leads to variation in treatment effects, in which individuals or groups vary in how they respond to the same circumstance. Multilevel models attempt to quantify the extent of this variation, as well as identify which units in the data responded in which ways.

These benefits don't come for free, however. Fitting and interpreting multilevel models can be considerably harder than fitting and interpreting a traditional regression model. In practice, many researchers simply trust their black-box software and interpret multilevel regression exactly like single-level regression. In time, this will change. There was a time in applied statistics when even ordinary multiple regression was considered cutting edge, something for only experts to fiddle with. Instead, scientists used many simple procedures, like t-tests. Now, almost everyone uses multivariate tools. The same will eventually be true of multilevel models. Scholarly culture and curriculum still have some catching up to do.

**Rethinking: Multilevel election forecasting**. One of the older applications of multilevel modeling is to forecast the outcomes of elections. In the 1960s, John Tukey (1915–2000) began working for the National Broadcasting Company (NBC) in the United States, developing real-time election prediction models that could exploit diverse types of data: polls, past elections, partial results, and complete results from related districts. The models used a multilevel framework similar to the models presented in Chapters 13 and 14. Tukey developed and used such models for NBC through 1978.[32] Contemporary election prediction and poll aggregation remains an active topic for multilevel modeling.[33]

**1.3.4. Graphical causal models.** When the wind blows, branches sway. If you are human, you immediately interpret this statement as causal: The wind makes the branches move. But all we see is a statistical association. From the data alone, it could also be that the branches swaying makes the wind. That conclusion seems foolish, because you know trees do not sway their own branches. A statistical model is an amazing association engine. It makes it possible to detect associations between causes and their effects. But a statistical model is never sufficient for inferring cause, because the statistical model makes no distinction between the wind causing the branches to sway and the branches causing the wind to blow. Facts outside the data are needed to decide which explanation is correct.

Cross-validation and information criteria try to guess predictive accuracy. When I introduced them above, I described overfitting as the primary paradox in prediction. Now we turn to a secondary paradox in prediction: *Models that are causally incorrect can make better predictions than those that are causally correct*. As a result, focusing on prediction can systematically mislead us. And while you may have heard that randomized controlled experiments allow causal inference, randomized experiments entail the same risks. No one is safe.

I will call this the **IDENTIFICATION** problem and carefully distinguish it from the problem of raw prediction. Consider two different meanings of "prediction." The simplest applies when we are external observers simply trying to guess what will happen next. In that case, tools like cross-validation are very useful. But these tools will happily recommend models that contain confounding variables and suggest incorrect causal relationships. Why? Confounded relationships are real associations, and they can improve prediction. After all, if you look outside and see branches swaying, it really does predict wind. Successful prediction does not require correct causal identification. In fact, as you'll see later in the book, predictions may actually improve when we use a model that is causally misleading.

But what happens when we intervene in the world? Then we must consider a second meaning of "prediction." Suppose we recruit many people to climb into the trees and sway the branches. Will it make wind? Not much. Often the point of statistical modeling is to produce understanding that leads to generalization and application. In that case, we need more than just good predictions, in the absence of intervention. We also need an accurate causal understanding. But comparing models on the basis of predictive accuracy—or $p$-values or anything else—will not necessarily produce it.

So what can be done? What is needed is a causal model that can be used to design one or more statistical models for the purpose of causal identification. As I mentioned in the neutral molecular evolution example earlier

in this chapter, a complete scientific model contains more information than a statistical model derived from it. And this additional information contains causal implications. These implications make it possible to test alternative causal models. The implications and tests depend upon the details. Newton's laws of motion for example precisely predict the consequences of specific interventions. And these precise predictions tell us that the laws are only approximately right.

Unfortunately, much scientific work lacks such precise models. Instead we must work with vaguer hypotheses and try to estimate vague causal effects. Economics for example has no good quantitative model for predicting the effect of changing the minimum wage. But the very good news is that even when you don't have a precise causal model, but only a heuristic one indicating which variables causally influence others, you can still do useful causal inference. Economics might, for example, be able to estimate the causal effect of changing the minimum wage, even without a good scientific model of the economy.

Formal methods for distinguishing causal inference from association date from the first half of the twentieth century, but they have more recently been extended to the study of measurement, experimental design, and the ability to generalize (or *transport)* results across samples.[34] We'll meet these methods through the use of a **GRAPHICAL CAUSAL MODEL**. The simplest graphical causal model is a **DIRECTED ACYCLIC GRAPH**, usually called a DAG. DAGs are heuristic—they are not detailed statistical models. But they allow us to deduce which statistical models can provide valid causal inferences, assuming the DAG is true.

But where does a DAG itself come from? The terrible truth about statistical inference is that its validity relies upon information outside the data. We require a causal model with which to design both the collection of data and the structure of our statistical models. But the construction of causal models is not a purely statistical endeavor, and statistical analysis can never verify all of our assumptions. There will never be a golem that accepts naked data and returns a reliable model of the causal relations among the variables. We're just going to have to keep doing science.

**Rethinking: Causal salad**. Causal inference requires a causal model that is separate from the statistical model. The data are not enough. Every philosophy agrees upon that much. Responses, however, are diverse. The most conservative response is to declare "causation" to be unprovable mental candy, like debating the nature of the afterlife.[35] Slightly less conservative is to insist that cause can only be inferred under strict conditions of randomization and experimental control. This would be very limiting. Many scientific questions can never be studied experimentally—human evolution, for example. Many others could in principle be studied experimentally, but it would be unethical to do so. And many experiments are really just attempts at control—patients do not always take their medication.

But the approach which dominates in many parts of biology and the social sciences is instead **CAUSAL SALAD**.[36] Causal salad means tossing various "control" variables into a statistical model, observing changes in estimates, and then telling a story about causation. Causal salad seems founded on the notion that only omitted variables can mislead us about causation. But *included* variables can just as easily confound us. When tossing a causal salad, a model that makes good predictions may still mislead about causation. If we use the model to plan an intervention, it will get everything wrong. There will be examples in later chapters.

## 1.4. Summary

This first chapter has argued for a rethinking of popular statistical and scientific philosophy. Instead of choosing among various black-box tools for testing null hypotheses, we should learn to build and analyze multiple non-null models of natural phenomena. To support this goal, the chapter introduced Bayesian inference, model comparison, multilevel models, and graphical causal models. The remainder of the book is organized into four parts.

(1) Chapters 2 and 3 are foundational. They introduce Bayesian inference and the basic tools for performing Bayesian calculations. They move quite slowly and emphasize a purely logical interpretation of probability

theory.

(2) The next five chapters, 4 through 8, build multiple linear regression as a Bayesian tool. This tool supports causal inference, but only when we analyze separate causal models that help us determine which variables to include. For this reason, you'll learn basic causal reasoning supported by causal graphs. These chapters emphasize plotting results instead of attempting to interpret estimates of individual parameters. Problems of model complexity—overfitting—also feature prominently. So you'll also get an introduction to information theory and predictive model comparison in Chapter 7.

(3) The third part of the book, Chapters 9 through 12, presents generalized linear models of several types. Chapter 9 introduces Markov chain Monte Carlo, used to fit the models in later chapters. Chapter 10 introduces maximum entropy as an explicit procedure to help us design and interpret these models. Then Chapters 11 and 12 detail the models themselves.

(4) The last part, Chapters 13 through 16, gets around to multilevel models, as well as specialized models that address measurement error, missing data, and spatial covariation. This material is fairly advanced, but it proceeds in the same mechanistic way as earlier material. Chapter 16 departs from the rest of the book in deploying models which are not of the generalized linear type but are rather scientific models expressed directly as statistical models.

The final chapter, Chapter 17, returns to some of the issues raised in this first one.

At the end of each chapter, there are practice problems ranging from easy to hard. These problems help you test your comprehension. The harder ones expand on the material, introducing new examples and obstacles. Some of the hard problems are quite hard. Don't worry, if you get stuck from time to time. Working in groups is a good way to get unstuck, just like in real research.

# 2 Small Worlds and Large Worlds

When Cristoforo Colombo (Christopher Columbus) infamously sailed west in the year 1492, he believed that the Earth was spherical. In this, he was like most educated people of his day. He was unlike most people, though, in that he also believed the planet was much smaller than it actually is—only 30,000 km around its middle instead of the actual 40,000 km (FIGURE 2.1).[37] This was one of the most consequential mistakes in European history. If Colombo had believed instead that the Earth was 40,000 km around, he would have correctly reasoned that his fleet could not carry enough food and potable water to complete a journey all the way westward to Asia. But at 30,000 km around, Asia would lie a bit west of the coast of California. It was possible to carry enough supplies to make it that far. Emboldened in part by his unconventional estimate, Colombo set sail, eventually landing in the Bahamas.

Colombo made a prediction based upon his view that the world was small. But since he lived in a large world, aspects of the prediction were wrong. In his case, the error was lucky. His small world model was wrong in an unanticipated way: There was a lot of land in the way. If he had been wrong in the expected way, with nothing but ocean between Europe and Asia, he and his entire expedition would have run out of supplies long before reaching the East Indies.

Colombo's small and large worlds provide a contrast between model and reality. All statistical modeling has these two frames: the *small world* of the model itself and the *large world* we hope to deploy the model in.[38] Navigating between these two worlds remains a central challenge of statistical modeling. The challenge is greater when we forget the distinction.

The **SMALL WORLD** is the self-contained logical world of the model. Within the small world, all possibilities are nominated. There are no pure surprises, like the existence of a huge continent between Europe and Asia. Within the small world of the model, it is important to be able to verify the model's logic, making sure that it performs as expected under favorable assumptions. Bayesian models have some advantages in this regard, as they have reasonable claims to optimality: No alternative model could make better use of the information in the data and support better decisions, assuming the small world is an accurate description of the real world.[39]

The **LARGE WORLD** is the broader context in which one deploys a model. In the large world, there may be events that were not imagined in the small world. Moreover, the model is always an incomplete representation of the large world, and so will make mistakes, even if all kinds of events have been properly nominated. The logical consistency of a model in the small world is no guarantee that it will be optimal in the large world. But it is certainly a warm comfort.

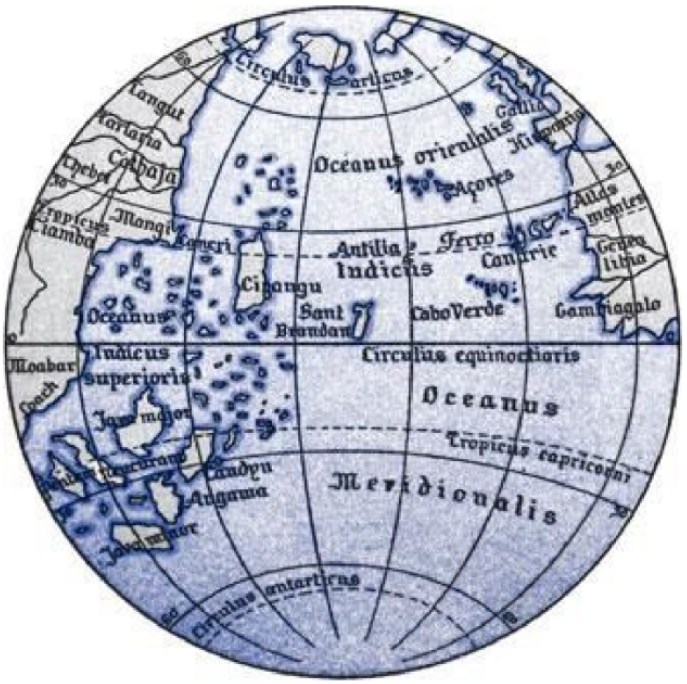



FIGURE 2.1. Illustration of Martin Behaim's 1492 globe, showing the small world that Colombo anticipated. Europe lies on the right-hand side. Asia lies on the left. The big island labeled "Cipangu" is Japan.

In this chapter, you will begin to build Bayesian models. The way that Bayesian models learn from evidence is arguably optimal in the small world. When their assumptions approximate reality, they also perform well in the large world. But large world performance has to be demonstrated rather than logically deduced. Passing back and forth between these two worlds allows both formal methods, like Bayesian inference, and informal methods, like peer review, to play an indispensable role.

This chapter focuses on the small world. It explains probability theory in its essential form: counting the ways things can happen. Bayesian inference arises automatically from this perspective. Then the chapter presents the stylized components of a Bayesian statistical model, a model for learning from data. Then it shows you how to animate the model, to produce estimates.

All this work provides a foundation for the next chapter, in which you'll learn to summarize Bayesian estimates, as well as begin to consider large world obligations.

**Rethinking: Fast and frugal in the large world**. The natural world is complex, as trying to do science serves to remind us. Yet everything from the humble tick to the industrious squirrel to the idle sloth manages to frequently make adaptive decisions. But it's a good bet that most animals are not Bayesian, if only because being Bayesian is expensive and depends upon having a good model. Instead, animals use various heuristics that are fit to their environments, past or present. These heuristics take adaptive shortcuts and so may outperform a rigorous Bayesian analysis, once costs of information gathering and processing (and overfitting, Chapter 7) are taken into account.[40] Once you already know which information to ignore or attend to, being fully Bayesian is a waste. It's neither necessary nor sufficient for making good decisions, as real animals demonstrate. But for human animals, Bayesian analysis provides a general way to discover relevant information and process it logically. Just don't think that it is the only way.

## 2.1. The garden of forking data

Our goal in this section will be to build Bayesian inference up from humble beginnings, so there is no

superstition about it. Bayesian inference is really just counting and comparing of possibilities. Consider by analogy Jorge Luis Borges' short story "The Garden of Forking Paths." The story is about a man who encounters a book filled with contradictions. In most books, characters arrive at plot points and must decide among alternative paths. A protagonist may arrive at a man's home. She might kill the man, or rather take a cup of tea. Only one of these paths is taken—murder or tea. But the book within Borges' story explores all paths, with each decision branching outward into an expanding garden of forking paths.

This is the same device that Bayesian inference offers. In order to make good inference about what actually happened, it helps to consider everything that could have happened. A Bayesian analysis is a garden of forking data, in which alternative sequences of events are cultivated. As we learn about what did happen, some of these alternative sequences are pruned. In the end, what remains is only what is logically consistent with our knowledge.

This approach provides a quantitative ranking of hypotheses, a ranking that is maximally conservative, given the assumptions and data that go into it. The approach cannot guarantee a correct answer, on large world terms. But it can guarantee the best possible answer, on small world terms, that could be derived from the information fed into it.

Consider the following toy example.

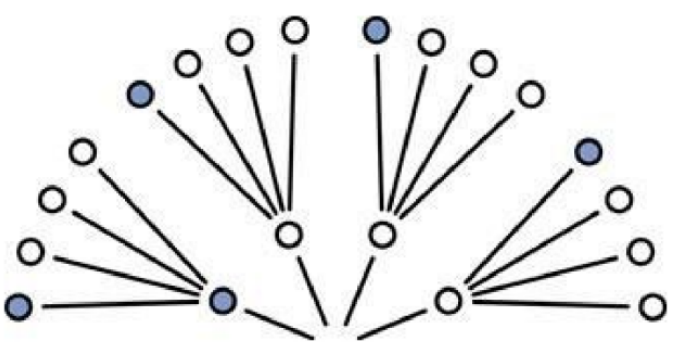
**2.1.1. Counting possibilities.** Suppose there's a bag, and it contains four marbles. These marbles come in two colors: blue and white. We know there are four marbles in the bag, but we don't know how many are of each color. We do know that there are five possibilities: (1) [○○○○], (2) [●○○○], (3) [●●○○], (4) [●●●○], (5) [●●●●]. These are the only possibilities consistent with what we know about the contents of the bag. Call these five possibilities the *conjectures*.

Our goal is to figure out which of these conjectures is most plausible, given some evidence about the contents of the bag. We do have some evidence: A sequence of three marbles is pulled from the bag, one at a time, replacing the marble each time and shaking the bag before drawing another marble. The sequence that emerges is: ●○●, in that order. These are the data.

So now let's plant the garden and see how to use the data to infer what's in the bag. Let's begin by considering just the single conjecture, [●○○○], that the bag contains one blue and three white marbles. On the first draw from the bag, one of four things could happen, corresponding to one of four marbles in the bag. So we can visualize the possibilities branching outward:

Notice that even though the three white marbles look the same from a data perspective—we just record the color of the marbles, after all—they are really different events. This is important, because it means that there are three more ways to see ○ than to see ●.

Now consider the garden as we get another draw from the bag. It expands the garden out one layer:



Now there are 16 possible paths through the garden, one for each pair of draws. On the second draw from the bag, each of the paths above again forks into four possible paths. Why? Because we believe that our shaking of the bag gives each marble a fair chance at being drawn, regardless of which marble was drawn previously. The third layer is built in the same way, and the full garden is shown in FIGURE 2.2. There are $4^3 = 64$ possible paths in total.

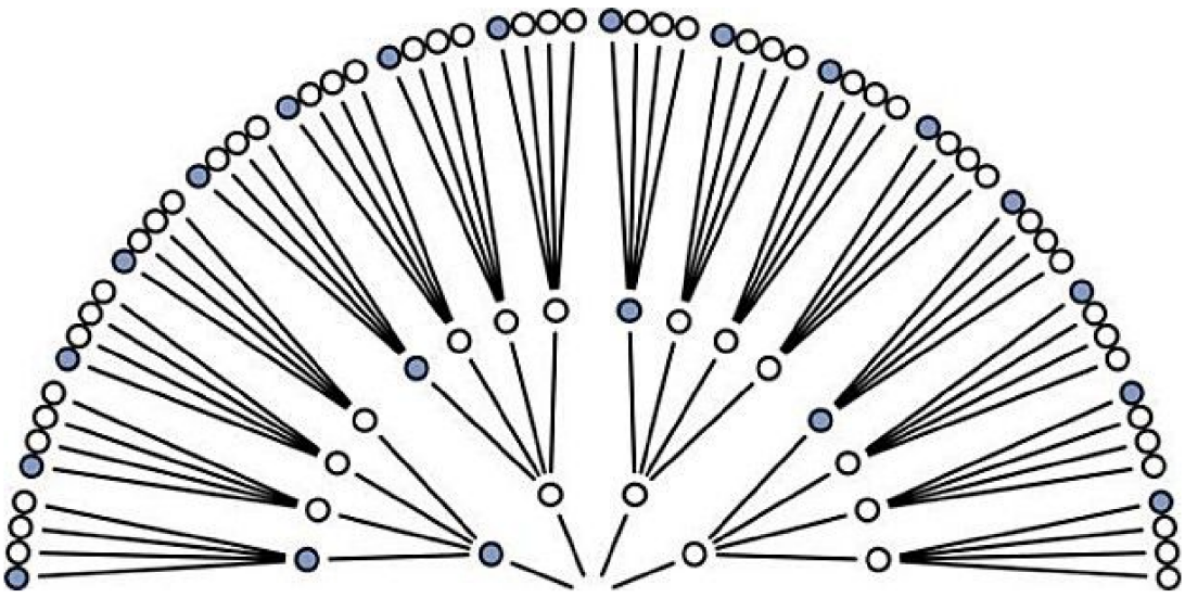


FIGURE 2.2. The 64 possible paths generated by assuming the bag contains one blue and three white marbles.

As we consider each draw from the bag, some of these paths are logically eliminated. The first draw tuned out to be ●, recall, so the three white paths at the bottom of the garden are eliminated right away. If you imagine the real data tracing out a path through the garden, it must have passed through the one blue path near the origin. The second draw from the bag produces ○, so three of the paths forking out of the first blue marble remain. As the data trace out a path, we know it must have passed through one of those three white paths (after the first blue path), but we don't know which one, because we recorded only the color of each marble. Finally, the third draw is ●. Each of the remaining three paths in the middle layer sustain one blue path, leaving a total of three ways for the sequence ●○● to appear, assuming the bag contains [●○○○]. FIGURE 2.3 shows the garden again, now with logically eliminated paths grayed out. We can't be sure which of those three paths the actual data took. But as long as we're considering only the possibility that the bag contains one blue and three white marbles, we can be sure that the data took one of those three paths. Those are the only paths consistent with both our knowledge of the bag's contents (four marbles, white or blue) and the data (●○●).

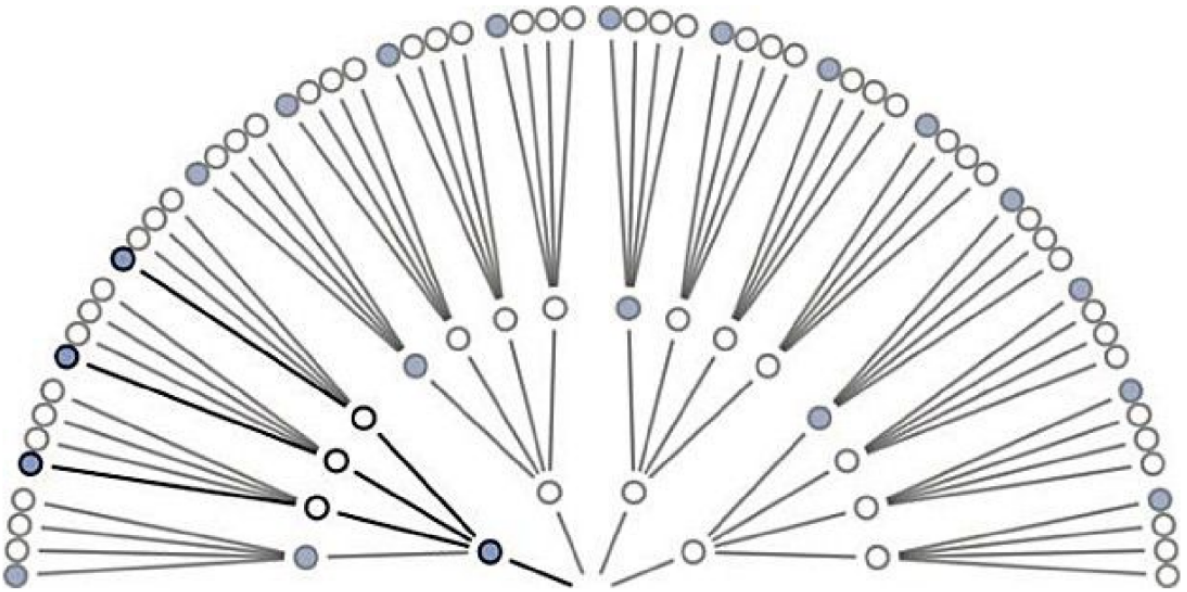


FIGURE 2.3. After eliminating paths inconsistent with the observed sequence, only 3 of the 64 paths remain.

This demonstrates that there are three (out of 64) ways for a bag containing [●○○○] to produce the data ●○●. We have no way to decide among these three ways. The inferential power comes from comparing this count to the numbers of ways each of the other conjectures of the bag's contents could produce the same data. For example, consider the conjecture [○○○○]. There are zero ways for this conjecture to produce the observed data, because even one ● is logically incompatible with it. The conjecture [●●●●] is likewise logically incompatible with the data. So we can eliminate these two conjectures, because neither provides even a single path that is consistent with the data.

FIGURE 2.4 displays the full garden now, for the remaining three conjectures: [●○○○], [●●○○], and [●●●○]. The upper-left wedge displays the same garden as FIGURE 2.3. The upper-right shows the analogous garden for the conjecture that the bag contains three blue marbles and one white marble. And the bottom wedge shows the garden for two blue and two white marbles. Now we count up all of the ways each conjecture could produce the observed data. For one blue and three white, there are three ways, as we counted already. For two blue and two white, there are eight paths forking through the garden that are logically consistent with the observed sequence. For three blue and one white, there are nine paths that survive.

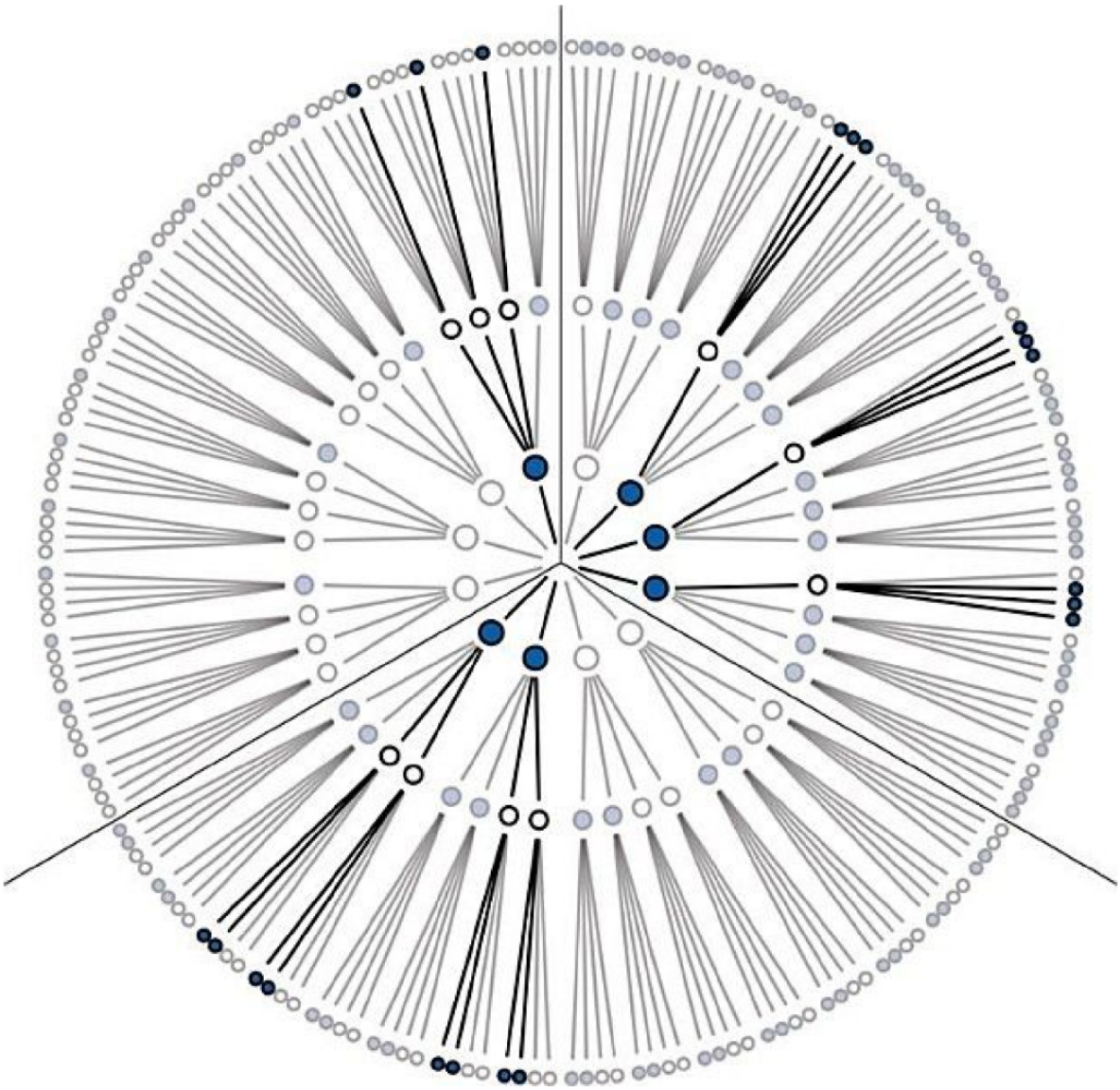


FIGURE 2.4. The garden of forking data, showing for each possible composition of the bag the forking paths that are logically compatible with the data.

To summarize, we've considered five different conjectures about the contents of the bag, ranging from zero blue marbles to four blue marbles. For each of these conjectures, we've counted up how many sequences, paths through the garden of forking data, could potentially produce the observed data, ●○●:

| Conjecture | Ways to produce ●○● |
|---|---|
| [○○○○] | $0 \times 4 \times 0 = 0$ |
| [●○○○] | $1 \times 3 \times 1 = 3$ |
| [●●○○] | $2 \times 2 \times 2 = 8$ |
| [●●●○] | $3 \times 1 \times 3 = 9$ |
| [●●●●] | $4 \times 0 \times 4 = 0$ |

Notice that the number of ways to produce the data, for each conjecture, can be computed by first counting the number of paths in each "ring" of the garden and then by multiplying these counts together. This is just a computational device. It tells us the same thing as FIGURE 2.4, but without having to draw the garden. The fact that numbers are multiplied during calculation doesn't change the fact that this is still just counting of logically

possible paths. This point will come up again, when you meet a formal representation of Bayesian inference.

So what good are these counts? By comparing these counts, we have part of a solution for a way to rate the relative plausibility of each conjectured bag composition. But it's only a part of a solution, because in order to compare these counts we first have to decide how many ways each conjecture could itself be realized. We might argue that when we have no reason to assume otherwise, we can just consider each conjecture equally plausible and compare the counts directly. But often we do have reason to assume otherwise.

**Rethinking: Justification**. My justification for using paths through the garden as measures of relative plausibility is humble: If we wish to reason about plausibility and remain consistent with ordinary logic—statements about *true* and *false*—then we should obey this procedure.[41] There are other justifications that lead to the same mathematical procedure. Regardless of how you choose to philosophically justify it, notice that it actually works. Justifications and philosophy motivate procedures, but it is the results that matter. The many successful real world applications of Bayesian inference may be all the justification you need. Twentieth century opponents of Bayesian data analysis argued that Bayesian inference was easy to justify, but hard to apply.[42] That is luckily no longer true. Indeed, the opposite is often true—scientists are switching to Bayesian approaches because it lets them use the models they want. Just be careful not to assume that because Bayesian inference is justified that no other approach can also be justified. Golems come in many types, and some of all types are useful.

**2.1.2. Combining other information.** We may have additional information about the relative plausibility of each conjecture. This information could arise from knowledge of how the contents of the bag were generated. It could also arise from previous data. Whatever the source, it would help to have a way to combine different sources of information to update the plausibilities. Luckily there is a natural solution: Just multiply the counts.

To grasp this solution, suppose we're willing to say each conjecture is equally plausible at the start. So we just compare the counts of ways in which each conjecture is compatible with the observed data. This comparison suggests that [●●●○] is slightly more plausible than [●●○○], and both are about three times more plausible than [●○○○]. Since these are our initial counts, and we are going to update them next, let's label them *prior*.

Now suppose we draw another marble from the bag to get another observation: ●. Now you have two choices. You could start all over again, making a garden with four layers to trace out the paths compatible with the data sequence ●○●●. Or you could take the previous counts—the prior counts—over conjectures (0, 3, 8, 9, 0) and just update them in light of the new observation. It turns out that these two methods are mathematically identical, as long as the new observation is logically independent of the previous observations.

Here's how to do it. First we count the numbers of ways each conjecture could produce the new observation, ●. Then we multiply each of these new counts by the prior numbers of ways for each conjecture. In table form:

| Conjecture | Ways to produce ○ | Prior counts | New count |
|---|---|---|---|
| [○○○○] | 0 | 0 | $0 \times 0 = 0$ |
| [●○○○] | 1 | 3 | $3 \times 1 = 3$ |
| [●●○○] | 2 | 8 | $8 \times 2 = 16$ |
| [●●●○] | 3 | 9 | $9 \times 3 = 27$ |
| [●●●●] | 4 | 0 | $0 \times 4 = 0$ |

The new counts in the right-hand column above summarize all the evidence for each conjecture. As new data arrive, and provided those data are independent of previous observations, then the number of logically possible ways for a conjecture to produce all the data up to that point can be computed just by multiplying the new count by the old count.

This updating approach amounts to nothing more than asserting that (1) when we have previous information suggesting there are $W_{\text{prior}}$ ways for a conjecture to produce a previous observation $D_{\text{prior}}$ and (2) we acquire new observations $D_{\text{new}}$ that the same conjecture can produce in $W_{\text{new}}$ ways, then (3) the number of ways the conjecture can account for both $D_{\text{prior}}$ as well as $D_{\text{new}}$ is just the product $W_{\text{prior}} \times W_{\text{new}}$. For example, in the table

above the conjecture [●●○○] has Wprior=8 ways to produce Dprior=●○●. It also has Wnew=2 ways to produce the new observation Dnew=●. So there are 8×2=16 ways for the conjecture to produce both $D_{\text{prior}}$ and $D_{\text{new}}$. Why multiply? Multiplication is just a shortcut to enumerating and counting up all of the paths through the garden that could produce all the observations.

In this example, the prior data and new data are of the same type: marbles drawn from the bag. But in general, the prior data and new data can be of different types. Suppose for example that someone from the marble factory tells you that blue marbles are rare. So for every bag containing [●●●○], they made two bags containing [●●○○] and three bags containing [●○○○]. They also ensured that every bag contained at least one blue and one white marble. We can update our counts again:

| Conjecture | Prior count | Factory count | New count |
|---|---|---|---|
| [○○○○] | 0 | 0 | $0 \times 0 = 0$ |
| [●○○○] | 3 | 3 | $3 \times 3 = 9$ |
| [●●○○] | 16 | 2 | $16 \times 2 = 32$ |
| [●●●○] | 27 | 1 | $27 \times 1 = 27$ |
| [●●●●] | 0 | 0 | $0 \times 0 = 0$ |

Now the conjecture [●●○○] is most plausible, but barely better than [●●●○]. Is there a threshold difference in these counts at which we can safely decide that one of the conjectures is the correct one? You'll spend the next chapter exploring that question.

**Rethinking: Original ignorance**. Which assumption should we use, when there is no previous information about the conjectures? The most common solution is to assign an equal number of ways that each conjecture could be correct, before seeing any data. This is sometimes known as the **PRINCIPLE OF INDIFFERENCE**: When there is no reason to say that one conjecture is more plausible than another, weigh all of the conjectures equally. This book does not use nor endorse "ignorance" priors. As we'll see in later chapters, the structure of the model and the scientific context always provide information that allows us to do better than ignorance.

For the sort of problems we examine in this book, the principle of indifference results in inferences very comparable to mainstream non-Bayesian approaches, most of which contain implicit equal weighting of possibilities. For example a typical non-Bayesian confidence interval weighs equally all of the possible values a parameter could take, regardless of how implausible some of them are. In addition, many non-Bayesian procedures have moved away from equal weighting, through the use of penalized likelihood and other methods. We'll discuss this in Chapter 7.

**2.1.3. From counts to probability.** It is helpful to think of this strategy as adhering to a principle of honest ignorance: *When we don't know what caused the data, potential causes that may produce the data in more ways are more plausible*. This leads us to count paths through the garden of forking data. We're counting the implications of assumptions.

It's hard to use these counts though, so we almost always standardize them in a way that transforms them into probabilities. Why is it hard to work with the counts? First, since relative value is all that matters, the size of the counts 3, 8, and 9 contain no information of value. They could just as easily be 30, 80, and 90. The meaning would be the same. It's just the relative values that matter. Second, as the amount of data grows, the counts will very quickly grow very large and become difficult to manipulate. By the time we have 10 data points, there are already more than one million possible sequences. We'll want to analyze data sets with thousands of observations, so explicitly counting these things isn't practical.

Luckily, there's a mathematical way to compress all of this. Specifically, we define the updated plausibility of each possible composition of the bag, after seeing the data, as:

$$\text{plausibility of [●○○○] after seeing ●○● } \propto \text{ ways [●○○○] can produce ●○● } \times \text{ prior plausibility [●○○○]}$$

That little $\propto$ means *proportional to*. We want to compare the plausibility of each possible bag composition. So

it'll be helpful to define $p$ as the proportion of marbles that are blue. For [●○○○], p=1/4=0.25. Also let Dnew=●○●. And now we can write:

$$\text{plausibility of p after Dnew} \propto \text{ways p can produce Dnew} \times \text{prior plausibility of p}$$

The above just means that for any value $p$ can take, we judge the plausibility of that value $p$ as proportional to the number of ways it can get through the garden of forking data. This expression just summarizes the calculations you did in the tables of the previous section.

Finally, we construct probabilities by standardizing the plausibility so that the sum of the plausibilities for all possible conjectures will be one. All you need to do in order to standardize is to add up all of the products, one for each value $p$ can take, and then divide each product by the sum of products:

$$\text{plausibility of p after Dnew} = \frac{\text{ways p can produce Dnew} \times \text{prior plausibility}}{\text{sum of products}}$$

A worked example is needed for this to really make sense. So consider again the table from before, now updated using our definitions of $p$ and "plausibility":

| Possible composition | $p$ | Ways to produce data | Plausibility |
|---|---|---|---|
| [○○○○] | 0 | 0 | 0 |
| [●○○○] | 0.25 | 3 | 0.15 |
| [●●○○] | 0.5 | 8 | 0.40 |
| [●●●○] | 0.75 | 9 | 0.45 |
| [●●●●] | 1 | 0 | 0 |

You can quickly compute these plausibilities in R:

R code 2.1

```
ways <- c( 0 , 3 , 8 , 9 , 0 )
ways/sum(ways)
```

```
[1] 0.00 0.15 0.40 0.45 0.00
```

The values in ways are the products mentioned before. And `sum(ways)` is the denominator "sum of products" in the expression near the top of the page.

These plausibilities are also *probabilities—they* are non-negative (zero or positive) real numbers that sum to one. And all of the mathematical things you can do with probabilities you can also do with these values. Specifically, each piece of the calculation has a direct partner in applied probability theory. These partners have stereotyped names, so it's worth learning them, as you'll see them again and again.

- A conjectured proportion of blue marbles, $p$, is usually called a **PARAMETER** value. It's just a way of indexing possible explanations of the data.
- The relative number of ways that a value $p$ can produce the data is usually called a **LIKELIHOOD**. It is derived by enumerating all the possible data sequences that could have happened and then eliminating those sequences inconsistent with the data.
- The prior plausibility of any specific $p$ is usually called the **PRIOR PROBABILITY**.
- The new, updated plausibility of any specific $p$ is usually called the **POSTERIOR PROBABILITY**.

In the next major section, you'll meet the more formal notation for these objects and see how they compose a simple statistical model.

**Rethinking: Randomization**. When you shuffle a deck of cards or assign subjects to treatments by flipping a

coin, it is common to say that the resulting deck and treatment assignments are *randomized*. What does it mean to randomize something? It just means that we have processed the thing so that we know almost nothing about its arrangement. Shuffling a deck of cards changes our state of knowledge, so that we no longer have any specific information about the ordering of cards. However, the bonus that arises from this is that, if we really have shuffled enough to erase any prior knowledge of the ordering, then the order the cards end up in is very likely to be one of the many orderings with high **INFORMATION ENTROPY**. The concept of information entropy will be increasingly important as we progress, and will be unpacked in Chapters 7 and 10.

## 2.2. Building a model

By working with probabilities instead of raw counts, Bayesian inference is made much easier, but it looks much harder. So in this section, we follow up on the garden of forking data by presenting the conventional form of a Bayesian statistical model. The toy example we'll use here has the anatomy of a typical statistical analysis, so it's the style that you'll grow accustomed to. But every piece of it can be mapped onto the garden of forking data. The logic is the same.

Suppose you have a globe representing our planet, the Earth. This version of the world is small enough to hold in your hands. You are curious how much of the surface is covered in water. You adopt the following strategy: You will toss the globe up in the air. When you catch it, you will record whether or not the surface under your right index finger is water or land. Then you toss the globe up in the air again and repeat the procedure.[43] This strategy generates a sequence of samples from the globe. The first nine samples might look like:

W L W W W L W L W

where W indicates water and L indicates land. So in this example you observe six W (water) observations and three L (land) observations. Call this sequence of observations the *data*.

To get the logic moving, we need to make assumptions, and these assumptions constitute the model. Designing a simple Bayesian model benefits from a design loop with three steps.

(1) Data story: Motivate the model by narrating how the data might arise.
(2) Update: Educate your model by feeding it the data.
(3) Evaluate: All statistical models require supervision, leading to model revision.

The next sections walk through these steps, in the context of the globe tossing evidence.

**2.2.1. A data story.** Bayesian data analysis usually means producing a story for how the data came to be. This story may be *descriptive*, specifying associations that can be used to predict outcomes, given observations. Or it may be *causal*, a theory of how some events produce other events. Typically, any story you intend to be causal may also be descriptive. But many descriptive stories are hard to interpret causally. But all data stories are complete, in the sense that they are sufficient for specifying an algorithm for simulating new data. In the next chapter, you'll see examples of doing just that, as simulating new data is useful not only for model criticism but also for model construction.

You can motivate your data story by trying to explain how each piece of data is born. This usually means describing aspects of the underlying reality as well as the sampling process. The data story in this case is simply a restatement of the sampling process:

(1) The true proportion of water covering the globe is $p$.
(2) A single toss of the globe has a probability $p$ of producing a water (W) observation. It has a probability $1 - p$ of producing a land (L) observation.
(3) Each toss of the globe is independent of the others.

The data story is then translated into a formal probability model. This probability model is easy to build, because the construction process can be usefully broken down into a series of component decisions. Before meeting these components, however, it'll be useful to visualize how a Bayesian model behaves. After you've become acquainted with how such a model learns from data, we'll pop the machine open and investigate its engineering.

**Rethinking: The value of storytelling**. The data story has value, even if you quickly abandon it and never use it to build a model or simulate new observations. Indeed, it is important to eventually discard the story, because many different stories correspond to the same model. As a result, showing that a model does a good job does not in turn uniquely support our data story. Still, the story has value because in trying to outline the story, often one realizes that additional questions must be answered. Most data stories are much more specific than are the verbal hypotheses that inspire data collection. Hypotheses can be vague, such as "it's more likely to rain on warm days." When you are forced to consider sampling and measurement and make a precise statement of how temperature predicts rain, many stories and resulting models will be consistent with the same vague hypothesis. Resolving that ambiguity often leads to important realizations and model revisions, before any model is fit to data.

**2.2.2. Bayesian updating.** Our problem is one of using the evidence—the sequence of globe tosses—to decide among different possible proportions of water on the globe. These proportions are like the conjectured marbles inside the bag, from earlier in the chapter. Each possible proportion may be more or less plausible, given the evidence. A Bayesian model begins with one set of plausibilities assigned to each of these possibilities. These are the prior plausibilities. Then it updates them in light of the data, to produce the posterior plausibilities. This updating process is a kind of learning, called **BAYESIAN UPDATING**. The details of this updating—how it is mechanically achieved—can wait until later in the chapter. For now, let's look only at how such a machine behaves.

For the sake of the example only, let's program our Bayesian machine to initially assign the same plausibility to every proportion of water, every value of $p$. We'll do better than this later. Now look at the top-left plot in FIGURE 2.5. The dashed horizontal line represents this initial plausibility of each possible value of $p$. After seeing the first toss, which is a "W," the model updates the plausibilities to the solid line. The plausibility of p=0 has now fallen to exactly zero—the equivalent of "impossible." Why? Because we observed at least one speck of water on the globe, so now we know there is *some* water. The model executes this logic automatically. You don't have it instruct it to account for this consequence. Probability theory takes care of it for you, because it is essentially counting paths through the garden of forking data, as in the previous section.

Likewise, the plausibility of $p > 0.5$ has increased. This is because there is not yet any evidence that there is land on the globe, so the initial plausibilities are modified to be consistent with this. Note however that the relative plausibilities are what matter, and there isn't yet much evidence. So the differences in plausibility are not yet very large. In this way, the amount of evidence seen so far is embodied in the plausibilities of each value of $p$.

In the remaining plots in FIGURE 2.5, the additional samples from the globe are introduced to the model, one at a time. Each dashed curve is just the solid curve from the previous plot, moving left to right and top to bottom. Every time a "W" is seen, the peak of the plausibility curve moves to the right, towards larger values of $p$. Every time an "L" is seen, it moves the other direction. The maximum height of the curve increases with each sample, meaning that fewer values of $p$ amass more plausibility as the amount of evidence increases. As each new observation is added, the curve is updated consistent with all previous observations.

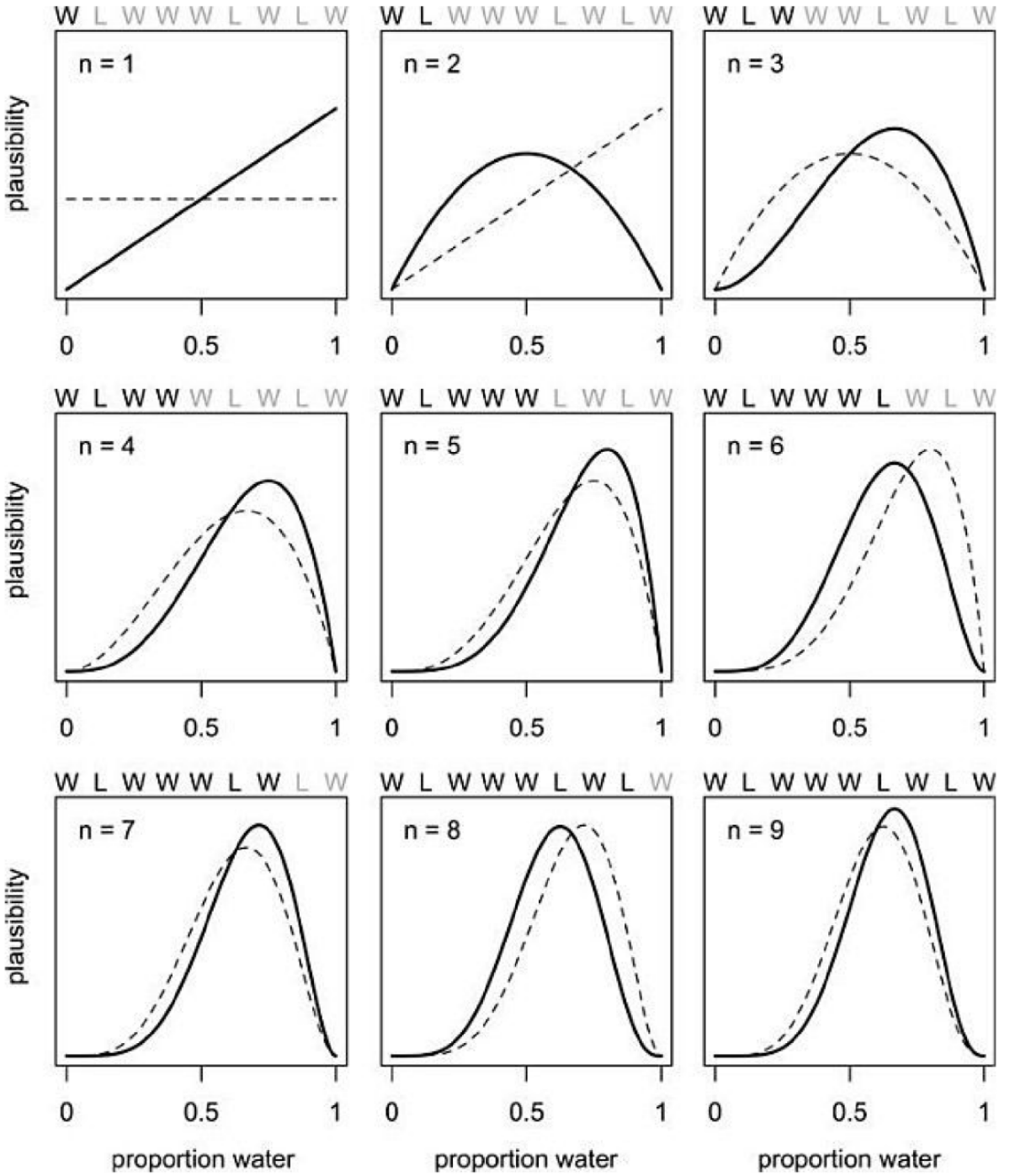



FIGURE 2.5. How a Bayesian model learns. Each toss of the globe produces an observation of water (W) or land (L). The model's estimate of the proportion of water on the globe is a plausibility for every possible value. The lines and curves in this figure are these collections of plausibilities. In each plot, previous plausibilities (dashed curve) are updated in light of the latest observation to produce a new set of plausibilities (solid curve).

Notice that every updated set of plausibilities becomes the initial plausibilities for the next observation. Every conclusion is the starting point for future inference. However, this updating process works backwards, as well as forwards. Given the final set of plausibilities in the bottom-right plot of FIGURE 2.5, and knowing the final observation (W), it is possible to mathematically divide out the observation, to infer the previous plausibility

curve. So the data could be presented to your model in any order, or all at once even. In most cases, you will present the data all at once, for the sake of convenience. But it's important to realize that this merely represents abbreviation of an iterated learning process.

**Rethinking: Sample size and reliable inference**. It is common to hear that there is a minimum number of observations for a useful statistical estimate. For example, there is a widespread superstition that 30 observations are needed before one can use a Gaussian distribution. Why? In non-Bayesian statistical inference, procedures are often justified by the method's behavior at very large sample sizes, so-called *asymptotic* behavior. As a result, performance at small samples sizes is questionable.

In contrast, Bayesian estimates are valid for any sample size. This does not mean that more data isn't helpful—it certainly is. Rather, the estimates have a clear and valid interpretation, no matter the sample size. But the price for this power is dependency upon the initial plausibilities, the prior. If the prior is a bad one, then the resulting inference will be misleading. There's no free lunch,[44] when it comes to learning about the world. A Bayesian golem must choose an initial plausibility, and a non-Bayesian golem must choose an estimator. Both golems pay for lunch with their assumptions.

**2.2.3. Evaluate.** The Bayesian model learns in a way that is demonstrably optimal, provided that it accurately describes the real, large world. This is to say that your Bayesian machine guarantees perfect inference within the small world. No other way of using the available information, beginning with the same state of information, could do better.

Don't get too excited about this logical virtue, however. The calculations may malfunction, so results always have to be checked. And if there are important differences between the model and reality, then there is no logical guarantee of large world performance. And even if the two worlds did match, any particular sample of data could still be misleading. So it's worth keeping in mind at least two cautious principles.

First, the model's certainty is no guarantee that the model is a good one. As the amount of data increases, the globe tossing model will grow increasingly sure of the proportion of water. This means that the curves in FIGURE 2.5 will become increasingly narrow and tall, restricting plausible values within a very narrow range. But models of all sorts—Bayesian or not—can be very confident about an inference, even when the model is seriously misleading. This is because the inferences are conditional on the model. What your model is telling you is that, given a commitment to this particular model, it can be very sure that the plausible values are in a narrow range. Under a different model, things might look differently. There will be examples in later chapters.

Second, it is important to supervise and critique your model's work. Consider again the fact that the updating in the previous section works in any order of data arrival. We could shuffle the order of the observations, as long as six W's and three L's remain, and still end up with the same final plausibility curve. That is only true, however, because the model assumes that order is irrelevant to inference. When something is irrelevant to the machine, it won't affect the inference directly. But it may affect it indirectly, because the data will depend upon order. So it is important to check the model's inferences in light of aspects of the data it does not know about. Such checks are an inherently creative enterprise, left to the analyst and the scientific community. Golems are very bad at it.

In Chapter 3, you'll see some examples of such checks. For now, note that the goal is not to test the truth value of the model's assumptions. We know the model's assumptions are never exactly right, in the sense of matching the true data generating process. Therefore there's no point in checking if the model is true. Failure to conclude that a model is false must be a failure of our imagination, not a success of the model. Moreover, models do not need to be exactly true in order to produce highly precise and useful inferences. All manner of small world assumptions about error distributions and the like can be violated in the large world, but a model may still produce a perfectly useful estimate. This is because models are essentially information processing machines, and there are some surprising aspects of information that cannot be easily captured by framing the problem in terms of the truth of assumptions.[45]

Instead, the objective is to check the model's adequacy for some purpose. This usually means asking and answering additional questions, beyond those that originally constructed the model. Both the questions and

answers will depend upon the scientific context. So it's hard to provide general advice. There will be many examples, throughout the book, and of course the scientific literature is replete with evaluations of the suitability of models for different jobs—prediction, comprehension, measurement, and persuasion.

**Rethinking: Deflationary statistics**. It may be that Bayesian inference is the best general purpose method of inference known. However, Bayesian inference is much less powerful than we'd like it to be. There is no approach to inference that provides universal guarantees. No branch of applied mathematics has unfettered access to reality, because math is not discovered, like the proton. Instead it is invented, like the shovel.[46]

## 2.3. Components of the model

Now that you've seen how the Bayesian model behaves, it's time to open up the machine and learn how it works. Consider three different things that we counted in the previous sections.

(1) The number of ways each conjecture could produce an observation
(2) The accumulated number of ways each conjecture could produce the entire data
(3) The initial plausibility of each conjectured cause of the data

Each of these things has a direct analog in conventional probability theory. And so the usual way we build a statistical model involves choosing distributions and devices for each that represent the relative numbers of ways things can happen.

In this section, you'll meet these components in some detail and see how each relates to the counting you did earlier in the chapter. The job in front of us is really nothing more than naming all of the variables and defining each. We'll take these tasks in turn.

**2.3.1. Variables.** Variables are just symbols that can take on different values. In a scientific context, variables include things we wish to infer, such as proportions and rates, as well as things we might observe, the data. In the globe tossing model, there are three variables.

The first variable is our target of inference, $p$, the proportion of water on the globe. This variable cannot be observed. Unobserved variables are usually called **PARAMETERS**. But while $p$ itself is unobserved, we can infer it from the other variables.

The other variables are the observed variables, the counts of water and land. Call the count of water $W$ and the count of land $L$. The sum of these two variables is the number of globe tosses: N=W+L.

**2.3.2. Definitions.** Once we have the variables listed, we then have to define each of them. In defining each, we build a model that relates the variables to one another. Remember, the goal is to count all the ways the data could arise, given the assumptions. This means, as in the globe tossing model, that for each possible value of the unobserved variables, such as $p$, we need to define the relative number of ways—the probability—that the values of each observed variable could arise. And then for each unobserved variable, we need to define the prior plausibility of each value it could take. I appreciate that this is all a bit abstract. So here are the specifics, for the globe.

*2.3.2.1. Observed variables.* For the count of water $W$ and land $L$, we define how plausible any combination of $W$ and $L$ would be, for a specific value of $p$. This is very much like the marble counting we did earlier in the chapter. Each specific value of $p$ corresponds to a specific plausibility of the data, as in FIGURE 2.5.

So that we don't have to literally count, we can use a mathematical function that tells us the right plausibility. In conventional statistics, a distribution function assigned to an observed variable is usually called a **LIKELIHOOD**. That term has special meaning in non-Bayesian statistics, however.[47] We will be able to do things with our distributions that non-Bayesian models forbid. So I will sometimes avoid the term *likelihood* and just

talk about distributions of variables. But when someone says, "likelihood," they will usually mean a distribution function assigned to an observed variable.

In the case of the globe tossing model, the function we need can be derived directly from the data story. Begin by nominating all of the possible events. There are two: *water* (W) and *land* (L). There are no other events. The globe never gets stuck to the ceiling, for example. When we observe a sample of W's and L's of length $N$ (nine in the actual sample), we need to say how likely that exact sample is, out of the universe of potential samples of the same length. That might sound challenging, but it's the kind of thing you get good at very quickly, once you start practicing.

In this case, once we add our assumptions that (1) every toss is independent of the other tosses and (2) the probability of W is the same on every toss, probability theory provides a unique answer, known as the *binomial distribution*. This is the common "coin tossing" distribution. And so the probability of observing $W$ waters and $L$ lands, with a probability $p$ of water on each toss, is:

$$\Pr(W,L|p)=(W+L)!W!L!pW(1-p)L$$

Read the above as:

> *The counts of "water" W and "land' L are distributed binomially, with probability p of "water" on each toss.*

And the binomial distribution formula is built into R, so you can easily compute the likelihood of the data—six W's in nine tosses—under any value of $p$ with:

R code 2.2

```
dbinom( 6 , size=9 , prob=0.5 )
```

```
[1] 0.1640625
```

That number is the relative number of ways to get six water, holding $p$ at 0.5 and N=W+L at nine. So it does the job of counting relative number of paths through the garden. Change the 0.5 to any other value, to see how the value changes.

Much later in the book, in Chapter 10, we'll see that the binomial distribution is rather special, because it represents the **MAXIMUM ENTROPY** way to count binary events. "Maximum entropy" might sound like a bad thing. Isn't entropy disorder? Doesn't "maximum entropy" mean the death of the universe? Actually it means that the distribution contains no additional information other than: There are two events, and the probabilities of each in each trial are $p$ and 1 - $p$. Chapter 10 explains this in more detail, and the details can certainly wait.

---

**Overthinking: Names and probability distributions**. The "`d`" in dbinom stands for *density*. Functions named in this way almost always have corresponding partners that begin with "`r`" for random samples and that begin with "`p`" for cumulative probabilities. See for example the help `?dbinom`.

---

**Rethinking: A central role for likelihood**. A great deal of ink has been spilled focusing on how Bayesian and non-Bayesian data analyses differ. Focusing on differences is useful, but sometimes it distracts us from fundamental similarities. Notably, the most influential assumptions in both Bayesian and many non-Bayesian models are the distributions assigned to data, the likelihood functions. The likelihoods influence inference for every piece of data, and as sample size increases, the likelihood matters more and more. This helps to explain why Bayesian and non-Bayesian inferences are often so similar. If we had to explain Bayesian inference using

only one aspect of it, we should describe likelihood, not priors.

*2.3.2.2. Unobserved variables.* The distributions we assign to the observed variables typically have their own variables. In the binomial above, there is $p$, the probability of sampling water. Since $p$ is not observed, we usually call it a **PARAMETER**. Even though we cannot observe $p$, we still have to define it.

In future chapters, there will be more parameters in your models. In statistical modeling, many of the most common questions we ask about data are answered directly by parameters:

- What is the average difference between treatment groups?
- How strong is the association between a treatment and an outcome?
- Does the effect of the treatment depend upon a covariate?
- How much variation is there among groups?

You'll see how these questions become extra parameters inside the distribution function we assign to the data.

For every parameter you intend your Bayesian machine to consider, you must provide a distribution of prior plausibility, its **PRIOR**. A Bayesian machine must have an initial plausibility assignment for each possible value of the parameter, and these initial assignments do useful work. When you have a previous estimate to provide to the machine, that can become the prior, as in the steps in FIGURE 2.5. Back in FIGURE 2.5, the machine did its learning one piece of data at a time. As a result, each estimate becomes the prior for the next step. But this doesn't resolve the problem of providing a prior, because at the dawn of time, when N=0, the machine still had an initial state of information for the parameter $p$: a flat line specifying equal plausibility for every possible value.

So where do priors come from? They are both engineering assumptions, chosen to help the machine learn, and scientific assumptions, chosen to reflect what we know about a phenomenon. The flat prior in FIGURE 2.5 is very common, but it is hardly ever the best prior. Later chapters will focus on prior choice a lot more.

There is a school of Bayesian inference that emphasizes choosing priors based upon the personal beliefs of the analyst.[48] While this **SUBJECTIVE BAYESIAN** approach thrives in some statistics and philosophy and economics programs, it is rare in the sciences. Within Bayesian data analysis in the natural and social sciences, the prior is considered to be just part of the model. As such it should be chosen, evaluated, and revised just like all of the other components of the model. In practice, the subjectivist and the non-subjectivist will often analyze data in nearly the same way.

None of this should be understood to mean that any statistical analysis is not inherently subjective, because of course it is—lots of little subjective decisions are involved in all parts of science. It's just that priors and Bayesian data analysis are no more inherently subjective than are likelihoods and the repeat sampling assumptions required for significance testing.[49] Anyone who has visited a statistics help desk at a university has probably experienced this subjectivity—statisticians do not in general exactly agree on how to analyze anything but the simplest of problems. The fact that statistical inference uses mathematics does not imply that there is only one reasonable or useful way to conduct an analysis. Engineering uses math as well, but there are many ways to build a bridge.

Beyond all of the above, there's no law mandating we use only one prior. If you don't have a strong argument for any particular prior, then try different ones. Because the prior is an assumption, it should be interrogated like other assumptions: by altering it and checking how sensitive inference is to the assumption. No one is required to swear an oath to the assumptions of a model, and no set of assumptions deserves our obedience.

---

**Overthinking: Prior as probability distribution**. You could write the prior in the example here as:

$$\Pr(p)=11-0=1.$$

The prior is a probability distribution for the parameter. In general, for a uniform prior from $a$ to $b$, the probability of any point in the interval is $1/(b-a)$. If you're bothered by the fact that the probability of every value of $p$ is 1, remember that every probability distribution must sum (integrate) to 1. The expression $1/(b-a)$

ensures that the area under the flat line from $a$ to $b$ is equal to 1. There will be more to say about this in Chapter 4.

**Rethinking: Datum or parameter?** It is typical to conceive of data and parameters as completely different kinds of entities. Data are measured and known; parameters are unknown and must be estimated from data. Usefully, in the Bayesian framework the distinction between a datum and a parameter is not so fundamental. Sometimes we observe a variable, but sometimes we do not. In that case, the same distribution function applies, even though we didn't observe the variable. As a result, the same assumption can look like a "likelihood" or a "prior," depending upon context, without any change to the model. Much later in the book (Chapter 15), you'll see how to exploit this deep identity between certainty (data) and uncertainty (parameters) to incorporate measurement error and missing data into your modeling.

**Rethinking: Prior, prior pants on fire**. Historically, some opponents of Bayesian inference objected to the arbitrariness of priors. It's true that priors are very flexible, being able to encode many different states of information. If the prior can be anything, isn't it possible to get any answer you want? Indeed it is. Regardless, after a couple hundred years of Bayesian calculation, it hasn't turned out that people use priors to lie. If your goal is to lie with statistics, you'd be a fool to do it with priors, because such a lie would be easily uncovered. Better to use the more opaque machinery of the likelihood. Or better yet—don't actually take this advice!—massage the data, drop some "outliers," and otherwise engage in motivated data transformation.

It is true though that choice of the likelihood is much more conventionalized than choice of prior. But conventional choices are often poor ones, smuggling in influences that can be hard to discover. In this regard, both Bayesian and non-Bayesian models are equally harried, because both traditions depend heavily upon likelihood functions and conventionalized model forms. And the fact that the non-Bayesian procedure doesn't have to make an assumption about the prior is of little comfort. This is because non-Bayesian procedures need to make choices that Bayesian ones do not, such as choice of estimator or likelihood penalty. Often, such choices can be shown to be equivalent to some Bayesian choice of prior or rather choice of loss function. (You'll meet loss functions later in Chapter 3.)

**2.3.3. A model is born.** With all the above work, we can now summarize our model. The observed variables $W$ and $L$ are given relative counts through the binomial distribution. So we can write, as a shortcut:

$$W \sim \text{Binomial}(N,p)$$

where N=W+L. The above is just a convention for communicating the assumption that the relative counts of ways to realize $W$ in $N$ trials with probability $p$ on each trial comes from the binomial distribution. And the unobserved parameter $p$ similarly gets:

$$p \sim \text{Uniform}(0,1)$$

This means that $p$ has a uniform—flat—prior over its entire possible range, from zero to one. As I mentioned earlier, this is obviously not the best we could do, since we know the Earth has more water than land, even if we do not know the exact proportion yet.

Next, let's see how to use these assumptions to generate inference.

## 2.4. Making the model go

Once you have named all the variables and chosen definitions for each, a Bayesian model can update all of the prior distributions to their purely logical consequences: the **POSTERIOR DISTRIBUTION**. For every unique combination of data, likelihood, parameters, and prior, there is a unique posterior distribution. This distribution

contains the relative plausibility of different parameter values, conditional on the data and model. The posterior distribution takes the form of the probability of the parameters, conditional on the data. In this case, it would be $\Pr(p|W, L)$, the probability of each possible value of p, conditional on the specific $W$ and $L$ that we observed.

**2.4.1. Bayes' theorem.** The mathematical definition of the posterior distribution arises from **BAYES' THEOREM**. This is the theorem that gives Bayesian data analysis its name. But the theorem itself is a trivial implication of probability theory. Here's a quick derivation of it, in the context of the globe tossing example. Really this will just be a re-expression of the garden of forking data derivation from earlier in the chapter. What makes it look different is that it will use the rules of probability theory to coax out the updating rule. But it is still just counting.

The joint probability of the data $W$ and $L$ and any particular value of $p$ is:

$$\Pr(W,L,p)=\Pr(W,L|p)\Pr(p)$$

This just says that the probability of $W$, $L$ and $p$ is the product of $\Pr(W, L|p)$ and the prior probability $\Pr(p)$. This is like saying that the probability of rain and cold on the same day is equal to the probability of rain, when it's cold, times the probability that it's cold. This much is just definition. But it's just as true that:

$$\Pr(W,L,p)=\Pr(p|W,L)\Pr(W,L)$$

All I've done is reverse which probability is conditional, on the right-hand side. It is still a true definition. It's like saying that the probability of rain and cold on the same day is equal to the probability that it's cold, when it's raining, times the probability of rain. Compare this statement to the one in the previous paragraph.

Now since both right-hand sides above are equal to the same thing, $\Pr(W, L, p)$, they are also equal to one another:

$$\Pr(W,L|p)\Pr(p)=\Pr(p|W,L)\Pr(W,L)$$

So we can now solve for the thing that we want, $\Pr(p|W,L)$:

$$\Pr(p|W,L)=\Pr(W,L|p)\Pr(p)\Pr(W,L)$$

And this is Bayes' theorem. It says that the probability of any particular value of $p$, considering the data, is equal to the product of the relative plausibility of the data, conditional on $p$, and the prior plausibility of $p$, divided by this thing $\Pr(W, L)$, which I'll call the *average probability of the data*. In word form:

$$\text{Posterior}=\frac{\text{Probability of the data}\times\text{Prior}}{\text{Average probability of the data}}$$

The average probability of the data, $\Pr(W, L)$, can be confusing. It is commonly called the "evidence" or the "average likelihood," neither of which is a transparent name. The probability $\Pr(W, L)$ is literally the *average* probability of the data. Averaged over what? Averaged over the prior. It's job is just to standardize the posterior, to ensure it sums (integrates) to one. In mathematical form:

$$\Pr(W,L)=\mathrm{E}(\Pr(W,L|p))=\int\Pr(W,L|p)\Pr(p)dp$$

The operator E means to take an *expectation*. Such averages are commonly called *marginals* in mathematical statistics, and so you may also see this same probability called a *marginal likelihood*. And the integral above just defines the proper way to compute the average over a continuous distribution of values, like the infinite possible values of $p$.

The key lesson is that the posterior is proportional to the product of the prior and the probability of the data. Why? Because for each specific value of $p$, the number of paths through the garden of forking data is the product of the prior number of paths and the new number of paths. Multiplication is just compressed counting. The

average probability on the bottom just standardizes the counts so they sum to one. So while Bayes' theorem looks complicated, because the relationship with counting paths is obscured, it just expresses the counting that logic demands.

FIGURE 2.6 illustrates the multiplicative interaction of a prior and a probability of data. On each row, a prior on the left is multiplied by the probability of data in the middle to produce a posterior on the right. The probability of data in each case is the same. The priors however vary. As a result, the posterior distributions vary.

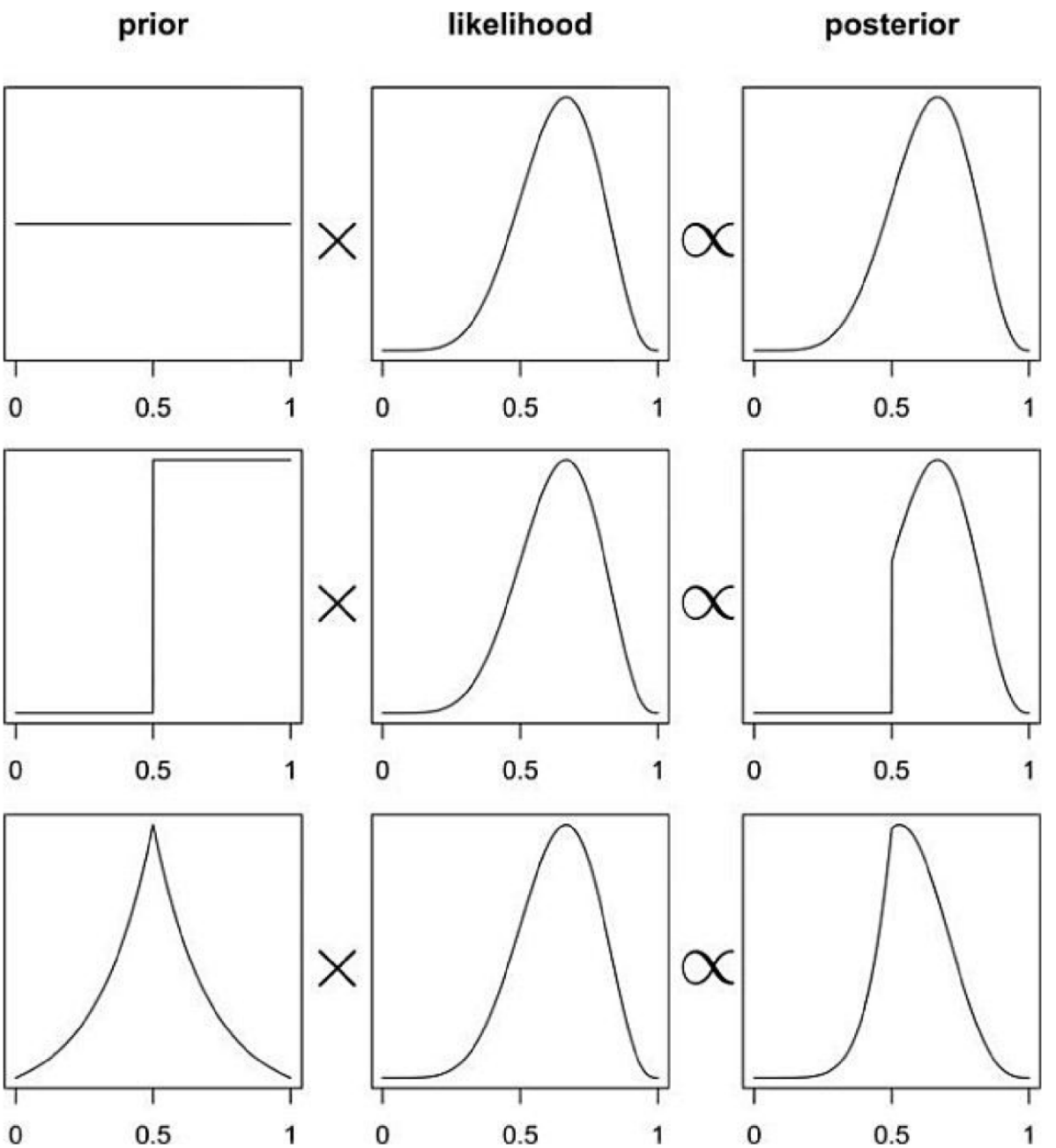



FIGURE 2.6. The posterior distribution as a product of the prior distribution and likelihood. Top: A flat prior constructs a posterior that is simply proportional to the likelihood. Middle: A step prior, assigning zero probability to all values less than 0.5, results in a truncated posterior. Bottom: A peaked prior that shifts and skews the posterior, relative to the likelihood.

**Rethinking: Bayesian data analysis isn't about Bayes' theorem**. A common notion about Bayesian data analysis, and Bayesian inference more generally, is that it is distinguished by the use of Bayes' theorem. This is a mistake. Inference under any probability concept will eventually make use of Bayes' theorem. Common introductory examples of "Bayesian" analysis using HIV and DNA testing are not uniquely Bayesian. Since all of the elements of the calculation are frequencies of observations, a non-Bayesian analysis would do exactly the same thing. Instead, Bayesian approaches get to use Bayes' theorem more generally, to quantify uncertainty about theoretical entities that cannot be observed, like parameters and models. Powerful inferences can be produced under both Bayesian and non-Bayesian probability concepts, but different justifications and sacrifices are necessary.

**2.4.2. Motors.** Recall that your Bayesian model is a machine, a figurative golem. It has built-in definitions for the likelihood, the parameters, and the prior. And then at its heart lies a motor that processes data, producing a posterior distribution. The action of this motor can be thought of as *conditioning* the prior on the data. As explained in the previous section, this conditioning is governed by the rules of probability theory, which defines a uniquely logical posterior for set of assumptions and observations.

However, knowing the mathematical rule is often of little help, because many of the interesting models in contemporary science cannot be conditioned formally, no matter your skill in mathematics. And while some broadly useful models like linear regression can be conditioned formally, this is only possible if you constrain your choice of prior to special forms that are easy to do mathematics with. We'd like to avoid forced modeling choices of this kind, instead favoring conditioning engines that can accommodate whichever prior is most useful for inference.

What this means is that various numerical techniques are needed to approximate the mathematics that follows from the definition of Bayes' theorem. In this book, you'll meet three different conditioning engines, numerical techniques for computing posterior distributions:

(1) Grid approximation
(2) Quadratic approximation
(3) Markov chain Monte Carlo (MCMC)

There are many other engines, and new ones are being invented all the time. But the three you'll get to know here are common and widely useful. In addition, as you learn them, you'll also learn principles that will help you understand other techniques.

**Rethinking: How you fit the model is part of the model**. Earlier in this chapter, I implicitly defined the model as a composite of a prior and a likelihood. That definition is typical. But in practical terms, we should also consider how the model is fit to data as part of the model. In very simple problems, like the globe tossing example that consumes this chapter, calculation of the posterior density is trivial and foolproof. In even moderately complex problems, however, the details of fitting the model to data force us to recognize that our numerical technique influences our inferences. This is because different mistakes and compromises arise under different techniques. The same model fit to the same data using different techniques may produce different answers. When something goes wrong, every piece of the machine may be suspect. And so our golems carry with them their updating engines, as much slaves to their engineering as they are to the priors and likelihoods we program into them.

**2.4.3. Grid approximation.** One of the simplest conditioning techniques is grid approximation. While most parameters are *continuous*, capable of taking on an infinite number of values, it turns out that we can achieve an excellent approximation of the continuous posterior distribution by considering only a finite grid of parameter values. At any particular value of a parameter, $p'$, it's a simple matter to compute the posterior probability: just

multiply the prior probability of $p$' by the likelihood at $p$'. Repeating this procedure for each value in the grid generates an approximate picture of the exact posterior distribution. This procedure is called **GRID APPROXIMATION**. In this section, you'll see how to perform a grid approximation, using simple bits of R code.

Grid approximation will mainly be useful as a pedagogical tool, as learning it forces the user to really understand the nature of Bayesian updating. But in most of your real modeling, grid approximation isn't practical. The reason is that it scales very poorly, as the number of parameters increases. So in later chapters, grid approximation will fade away, to be replaced by other, more efficient techniques. Still, the conceptual value of this exercise will carry forward, as you graduate to other techniques.

In the context of the globe tossing problem, grid approximation works extremely well. So let's build a grid approximation for the model we've constructed so far. Here is the recipe:

(1) Define the grid. This means you decide how many points to use in estimating the posterior, and then you make a list of the parameter values on the grid.
(2) Compute the value of the prior at each parameter value on the grid.
(3) Compute the likelihood at each parameter value.
(4) Compute the unstandardized posterior at each parameter value, by multiplying the prior by the likelihood.
(5) Finally, standardize the posterior, by dividing each value by the sum of all values.

In the globe tossing context, here's the code to complete all five of these steps:

R code 2.3

```
# define grid
p_grid <- seq( from=0 , to=1 , length.out=20 )

# define prior
prior <- rep( 1 , 20 )

# compute likelihood at each value in grid
likelihood <- dbinom( 6 , size=9 , prob=p_grid )

# compute product of likelihood and prior
unstd.posterior <- likelihood * prior

# standardize the posterior, so it sums to 1
posterior <- unstd.posterior / sum(unstd.posterior)
```

The above code makes a grid of only 20 points. To display the posterior distribution now:

R code 2.4

```
plot( p_grid , posterior , type="b" ,
   xlab="probability of water" , ylab="posterior probability" )
mtext( "20 points" )
```

You'll get the right-hand plot in FIGURE 2.7. Try sparser grids (5 points) and denser grids (100 or 1000 points). The correct density for your grid is determined by how accurate you want your approximation to be. More points means more precision. In this simple example, you can go crazy and use 100,000 points, but there won't be much

change in inference after the first 100.

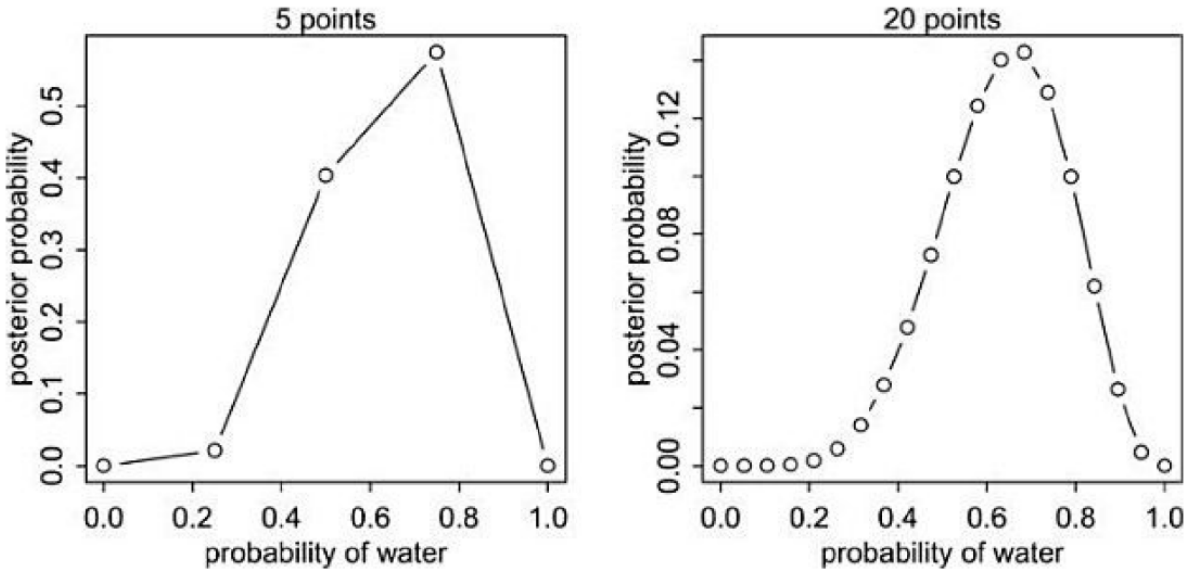



FIGURE 2.7. Computing posterior distribution by grid approximation. In each plot, the posterior distribution for the globe toss data and model is approximated with a finite number of evenly spaced points. With only 5 points (left), the approximation is terrible. But with 20 points (right), the approximation is already quite good. Compare to the analytically solved, exact posterior distribution in FIGURE 2.5 (page 30).

Now to replicate the different priors in FIGURE 2.5, try these lines of code—one at a time—for the prior grid:

R code 2.5

```
prior <- ifelse( p_grid < 0.5 , < , 1)
prior <- exp( -5*abs( p_grid - 0.5 ) )
```

The rest of the code remains the same.

---

**Overthinking: Vectorization**. One of R's useful features is that it makes working with lists of numbers almost as easy as working with single values. So even though both lines of code above say nothing about how dense your grid is, whatever length you chose for the vector `p_grid` will determine the length of the vector `prior`. In R jargon, the calculations above are *vectorized*, because they work on lists of values, *vectors*. In a vectorized calculation, the calculation is performed on each element of the input vector—`p_grid` in this case—and the resulting output therefore has the same length. In other computing environments, the same calculation would require a *loop*. R can also use loops, but vectorized calculations are typically faster. They can however be much harder to read, when you are starting out with R. Be patient, and you'll soon grow accustomed to vectorized calculations.

---

**2.4.4. Quadratic approximation.** We'll stick with the grid approximation to the globe tossing posterior, for the rest of this chapter and the next. But before long you'll have to resort to another approximation, one that makes stronger assumptions. The reason is that the number of unique values to consider in the grid grows rapidly as the number of parameters in your model increases. For the single-parameter globe tossing model, it's no problem to compute a grid of 100 or 1000 values. But for two parameters approximated by 100 values each, that's already

1002=10,000 values to compute. For 10 parameters, the grid becomes many billions of values. These days, it's routine to have models with hundreds or thousands of parameters. The grid approximation strategy scales very poorly with model complexity, so it won't get us very far.

A useful approach is **QUADRATIC APPROXIMATION**. Under quite general conditions, the region near the peak of the posterior distribution will be nearly Gaussian—or "normal"—in shape. This means the posterior distribution can be usefully approximated by a Gaussian distribution. A Gaussian distribution is convenient, because it can be completely described by only two numbers: the location of its center (mean) and its spread (variance).

A Gaussian approximation is called "quadratic approximation" because the logarithm of a Gaussian distribution forms a parabola. And a parabola is a quadratic function. So this approximation essentially represents any log-posterior with a parabola.

We'll use quadratic approximation for much of the first half of this book. For many of the most common procedures in applied statistics—linear regression, for example—the approximation works very well. Often, it is even exactly correct, not actually an approximation at all. Computationally, quadratic approximation is very inexpensive, at least compared to grid approximation and MCMC (discussed next). The procedure, which R will happily conduct at your command, contains two steps.

(1) Find the posterior mode. This is usually accomplished by some optimization algorithm, a procedure that virtually "climbs" the posterior distribution, as if it were a mountain. The golem doesn't know where the peak is, but it does know the slope under its feet. There are many well-developed optimization procedures, most of them more clever than simple hill climbing. But all of them try to find peaks.

(2) Once you find the peak of the posterior, you must estimate the curvature near the peak. This curvature is sufficient to compute a quadratic approximation of the entire posterior distribution. In some cases, these calculations can be done analytically, but usually your computer uses some numerical technique instead.

To compute the quadratic approximation for the globe tossing data, we'll use a tool in the `rethinking` package: `quap`. We're going to be using `quap` a lot in the first half of this book. It's a flexible model fitting tool that will allow us to specify a large number of different "regression" models. So it'll be worth trying it out right now. You'll get a more thorough understanding of it later.

To compute the quadratic approximation to the globe tossing data:

R code 2.6

```
library(rethinking)
globe.qa <- quap(
  alist(
    W ~ dbinom( W+L ,p) ,  # binomial likelihood
    p ~ dunif(0,1)  # uniform prior
  ),
  data=list(W=6,L=3) )
# display summary of quadratic approximation
precis( globe.qa )
```

To use quap, you provide a *formula*, a list of *data*. The formula defines the probability of the data and the prior. I'll say much more about these formulas in Chapter 4. Now let's see the output:

```
Mean StdDev 5.5% 94.5%
p 0.67 0.16 0.42 0.92
```

The function `precis` presents a brief summary of the quadratic approximation. In this case, it shows the

posterior mean value of p=0.67, which it calls the "Mean." The curvature is labeled "StdDev" This stands for *standard deviation*. This value is the standard deviation of the posterior distribution, while the mean value is its peak. Finally, the last two values in the `precis` output show the 89% percentile interval, which you'll learn more about in the next chapter. You can read this kind of approximation like: *Assuming the posterior is Gaussian, it is maximized at 0.67, and its standard deviation is 0.16*.

Since we already know the posterior, let's compare to see how good the approximation is. I'll use the analytical approach here, which uses dbeta. I won't explain this calculation, but it ensures that we have exactly the right answer. You can find an explanation and derivation of it in just about any mathematical textbook on Bayesian inference.

R code 2.7

```
# analytical calculation
W <- 6
L <- 3
curve( dbeta( x , W+1 , L+1 ) , from=0 , to=1 )
# quadratic approximation
curve( dnorm( x , 0.67 , 0.16 ) , lty=2 , add=TRUE )
```

You can see this plot (with a little extra formatting) on the left in FIGURE 2.8. The blue curve is the analytical posterior and the black curve is the quadratic approximation. The black curve does alright on its left side, but looks pretty bad on its right side. It even assigns positive probability to p=1, which we know is impossible, since we saw at least one land sample.

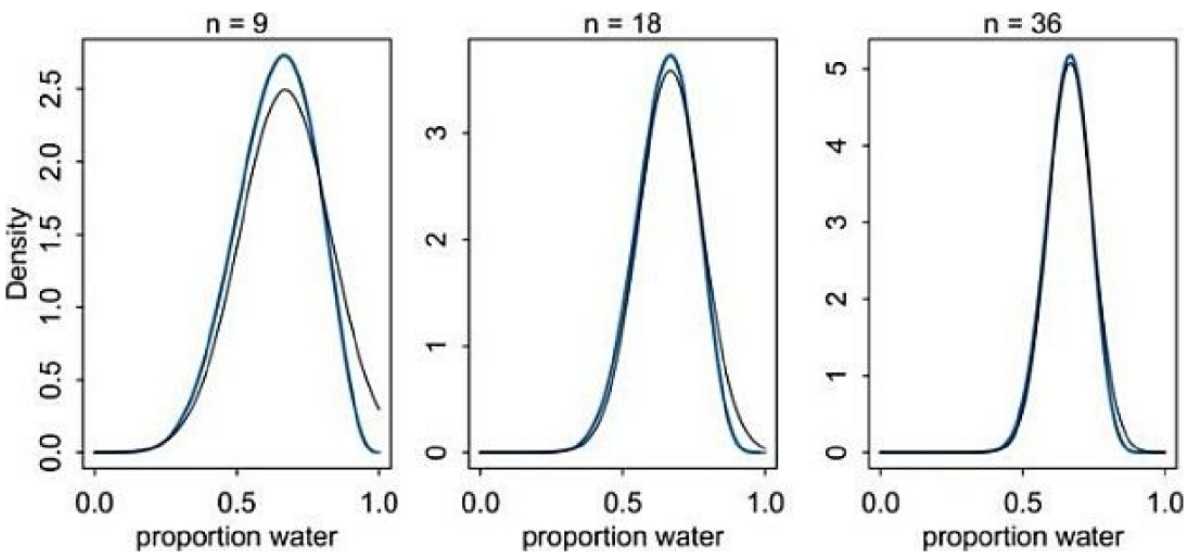



Figure 2.8 Accuracy of the quadratic approximation. In each plot, the exact posterior distribution is plotted in blue, and the quadratic approximation is plotted as the black curve. Left: The globe tossing data with n=9 tosses and w=6 waters. Middle: Double the amount of data, with the same fraction of water, n=18 and w=12. Right: Four times as much data, n=36 and w=24.

As the amount of data increases, however, the quadratic approximation gets better. In the middle of FIGURE 2.8, the sample size is doubled to n=18 tosses, but with the same fraction of water, so that the mode of the posterior is in the same place. The quadratic approximation looks better now, although still not great. At quadruple the data,

on the right side of the figure, the two curves are nearly the same now.

This phenomenon, where the quadratic approximation improves with the amount of data, is very common. It's one of the reasons that so many classical statistical procedures are nervous about small samples: Those procedures use quadratic (or other) approximations that are only known to be safe with infinite data. Often, these approximations are useful with less than infinite data, obviously. But the rate of improvement as sample size increases varies greatly depending upon the details. In some models, the quadratic approximation can remain terrible even with thousands of samples.

Using the quadratic approximation in a Bayesian context brings with it all the same concerns. But you can always lean on some algorithm other than quadratic approximation, if you have doubts. Indeed, grid approximation works very well with small samples, because in such cases the model must be simple and the computations will be quite fast. You can also use MCMC, which is introduced next.

**Rethinking: Maximum likelihood estimation**. The quadratic approximation, either with a uniform prior or with a lot of data, is often equivalent to a **MAXIMUM LIKELIHOOD ESTIMATE** (MLE) and its **STANDARD ERROR**. The MLE is a very common non-Bayesian parameter estimate. This correspondence between a Bayesian approximation and a common non-Bayesian estimator is both a blessing and a curse. It is a blessing, because it allows us to re-interpret a wide range of published non-Bayesian model fits in Bayesian terms. It is a curse, because maximum likelihood estimates have some curious drawbacks, and the quadratic approximation can share them. We'll explore these drawbacks in later chapters, and they are one of the reasons we'll turn to Markov chain Monte Carlo for the second half of the book.

**Overthinking: The Hessians are coming**. Sometimes it helps to know more about how the quadratic approximation is computed. In particular, the approximation sometimes fails. When it does, chances are you'll get a confusing error message that says something about the "Hessian." Students of world history may know that the Hessians were German mercenaries hired by the British in the eighteenth century to do various things, including fight against the American revolutionary George Washington. These mercenaries are named after a region of what is now central Germany, Hesse.

The Hessian that concerns us here has little to do with mercenaries. It is named after mathematician Ludwig Otto Hesse (1811–1874). A *Hessian* is a square matrix of second derivatives. It is used for many purposes in mathematics, but in the quadratic approximation it is second derivatives of the log of posterior probability with respect to the parameters. It turns out that these derivatives are sufficient to describe a Gaussian distribution, because the logarithm of a Gaussian distribution is just a parabola. Parabolas have no derivatives beyond the second, so once we know the center of the parabola (the posterior mode) and its second derivative, we know everything about it. And indeed the second derivative (with respect to the outcome) of the logarithm of a Gaussian distribution is proportional to its inverse squared standard deviation (its "precision": page 76). So knowing the standard deviation tells us everything about its shape.

The standard deviation is typically computed from the Hessian, so computing the Hessian is nearly always a necessary step. But sometimes the computation goes wrong, and your golem will choke while trying to compute the Hessian. In those cases, you have several options. Not all hope is lost. But for now it's enough to recognize the term and associate it with an attempt to find the standard deviation for a quadratic approximation.

**2.4.5. Markov chain Monte Carlo.** There are lots of important model types, like multilevel (mixed-effects) models, for which neither grid approximation nor quadratic approximation is always satisfactory. Such models may have hundreds or thousands or tens-of-thousands of parameters. Grid approximation routinely fails here, because it just takes too long—the Sun will go dark before your computer finishes the grid. Special forms of quadratic approximation might work, if everything is just right. But commonly, something is not just right. Furthermore, multilevel models do not always allow us to write down a single, unified function for the posterior distribution. This means that the function to maximize (when finding the MAP) is not known, but must be computed in pieces.

As a result, various counterintuitive model fitting techniques have arisen. The most popular of these is **MARKOV CHAIN MONTE CARLO** (MCMC), which is a family of conditioning engines capable of handling highly complex models. It is fair to say that MCMC is largely responsible for the insurgence of Bayesian data analysis that began in the 1990s. While MCMC is older than the 1990s, affordable computer power is not, so we must also thank the engineers. Much later in the book (Chapter 9), you'll meet simple and precise examples of MCMC model fitting, aimed at helping you understand the technique.

The conceptual challenge with MCMC lies in its highly non-obvious strategy. Instead of attempting to compute or approximate the posterior distribution directly, MCMC techniques merely draw samples from the posterior. You end up with a collection of parameter values, and the frequencies of these values correspond to the posterior plausibilities. You can then build a picture of the posterior from the histogram of these samples.

We nearly always work directly with these samples, rather than first constructing some mathematical estimate from them. And the samples are in many ways more convenient than having the posterior, because they are easier to think with. And so that's where we turn in the next chapter, to thinking with samples.

---

**Overthinking: Monte Carlo globe tossing**. If you are eager to see MCMC in action, a working Markov chain for the globe tossing model does not require much code. The following R code is sufficient for a MCMC estimate of the posterior:

R code 2.8

```
n_samples <- 1000
p <- rep( NA , n_samples )
p[1] <- 0.5
W <- 6
L <- 3
for ( i in 2:n_samples ) {
  p_new <- rnorm( 1 , p[i-1] , 0.1 )
  if ( p_new < 0 ) p_new <- abs( p_new )
  if ( p_new > 1 ) p_new <- 2 - p_new
  q0 <- dbinom( W , W+L , p[i-1] )
  q1 <- dbinom( W , W+L , p_new )
  p[i] <- ifelse( runif(1) < q1/q0 , p_new , p[i-1] )
}
```

The values in p are samples from the posterior distribution. To compare to the analytical posterior:

R code 2.9

```
dens( p , xlim=c(0,1) ) 2.9
curve( dbeta( x , W+1 , L+1 ) , lty=2 , add=TRUE )
```

It's weird. But it works. I'll explain this algorithm, the **METROPOLIS ALGORITHM**, in Chapter 9.

---

## 2.5. Summary

This chapter introduced the conceptual mechanics of Bayesian data analysis. The target of inference in Bayesian inference is a posterior probability distribution. Posterior probabilities state the relative numbers of ways each conjectured cause of the data could have produced the data. These relative numbers indicate plausibilities of the different conjectures. These plausibilities are updated in light of observations through Bayesian updating.

More mechanically, a Bayesian model is a composite of variables and distributional definitions for these variables. The probability of the data, often called the likelihood, provides the plausibility of an observation (data), given a fixed value for the parameters. The prior provides the plausibility of each possible value of the parameters, before accounting for the data. The rules of probability tell us that the logical way to compute the plausibilities, after accounting for the data, is to use Bayes' theorem. This results in the posterior distribution.

In practice, Bayesian models are fit to data using numerical techniques, like grid approximation, quadratic approximation, and Markov chain Monte Carlo. Each method imposes different trade-offs.

## 2.6. Practice

Problems are labeled Easy (E), Medium (M), and Hard (H).

**2E1.** Which of the expressions below correspond to the statement: *the probability of rain on Monday?*

(1) Pr(rain)
(2) Pr(rain|Monday)
(3) Pr(Monday|rain)
(4) Pr(rain, Monday)/ Pr(Monday)

**2E2.** Which of the following statements corresponds to the expression: Pr(Monday|rain)?

(1) The probability of rain on Monday.
(2) The probability of rain, given that it is Monday.
(3) The probability that it is Monday, given that it is raining.
(4) The probability that it is Monday and that it is raining.

**2E3.** Which of the expressions below correspond to the statement: *the probability that it is Monday, given that it is raining?*

(1) Pr(Monday|rain)
(2) Pr(rain|Monday)
(3) Pr(rain|Monday) Pr(Monday)
(4) Pr(rain|Monday) Pr(Monday)/ Pr(rain)
(5) Pr(Monday|rain) Pr(rain)/ Pr(Monday)

**2E4.** The Bayesian statistician Bruno de Finetti (1906–1985) began his 1973 book on probability theory with the declaration: "PROBABILITY DOES NOT EXIST." The capitals appeared in the original, so I imagine de Finetti wanted us to shout this statement. What he meant is that probability is a device for describing uncertainty from the perspective of an observer with limited knowledge; it has no objective reality. Discuss the globe tossing example from the chapter, in light of this statement. What does it mean to say "the probability of water is 0.7"?

**2M1.** Recall the globe tossing model from the chapter. Compute and plot the grid approximate posterior distribution for each of the following sets of observations. In each case, assume a uniform prior for $p$.

(1) W, W, W
(2) W, W, W, L
(3) L, W, W, L, W, W, W

**2M2.** Now assume a prior for $p$ that is equal to zero when $p < 0.5$ and is a positive constant when $p \geq 0.5$. Again

compute and plot the grid approximate posterior distribution for each of the sets of observations in the problem just above.

**2M3.** Suppose there are two globes, one for Earth and one for Mars. The Earth globe is 70% covered in water. The Mars globe is 100% land. Further suppose that one of these globes—you don't know which—was tossed in the air and produced a "land" observation. Assume that each globe was equally likely to be tossed. Show that the posterior probability that the globe was the Earth, conditional on seeing "land" (Pr(Earth|land)), is 0.23.

**2M4.** Suppose you have a deck with only three cards. Each card has two sides, and each side is either black or white. One card has two black sides. The second card has one black and one white side. The third card has two white sides. Now suppose all three cards are placed in a bag and shuffled. Someone reaches into the bag and pulls out a card and places it flat on a table. A black side is shown facing up, but you don't know the color of the side facing down. Show that the probability that the other side is also black is 2/3. Use the counting method (Section 2 of the chapter) to approach this problem. This means counting up the ways that each card could produce the observed data (a black side facing up on the table).

**2M5.** Now suppose there are four cards: B/B, B/W, W/W, and another B/B. Again suppose a card is drawn from the bag and a black side appears face up. Again calculate the probability that the other side is black.

**2M6.** Imagine that black ink is heavy, and so cards with black sides are heavier than cards with white sides. As a result, it's less likely that a card with black sides is pulled from the bag. So again assume there are three cards: B/B, B/W, and W/W. After experimenting a number of times, you conclude that for every way to pull the B/B card from the bag, there are 2 ways to pull the B/W card and 3 ways to pull the W/W card. Again suppose that a card is pulled and a black side appears face up. Show that the probability the other side is black is now 0.5. Use the counting method, as before.

**2M7.** Assume again the original card problem, with a single card showing a black side face up. Before looking at the other side, we draw another card from the bag and lay it face up on the table. The face that is shown on the new card is white. Show that the probability that the first card, the one showing a black side, has black on its other side is now 0.75. Use the counting method, if you can. Hint: Treat this like the sequence of globe tosses, counting all the ways to see each observation, for each possible first card.

**2H1.** Suppose there are two species of panda bear. Both are equally common in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They differ however in their family sizes. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research.

Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

**2H2.** Recall all the facts from the problem above. Now compute the probability that the panda we have is from species A, assuming we have observed only the first birth and that it was twins.

**2H3.** Continuing on from the previous problem, suppose the same panda mother has a second birth and that it is not twins, but a singleton infant. Compute the posterior probability that this panda is species A.

**2H4.** A common boast of Bayesian statisticians is that Bayesian inference makes it easy to use all of the data, even if the data are of different types.

So suppose now that a veterinarian comes along who has a new genetic test that she claims can identify the species of our mother panda. But the test, like all tests, is imperfect. This is the information you have about the test:

- The probability it correctly identifies a species A panda is 0.8.
- The probability it correctly identifies a species B panda is 0.65.

The vet administers the test to your panda and tells you that the test is positive for species A. First ignore your previous information from the births and compute the posterior probability that your panda is species A. Then redo your calculation, now using the birth data as well.

# 3 Sampling the Imaginary

Lots of books on Bayesian statistics introduce posterior inference by using a medical testing scenario. To repeat the structure of common examples, suppose there is a blood test that correctly detects vampirism 95% of the time. In more precise and mathematical notation, Pr(positive test result|vampire) = 0.95. It's a very accurate test, nearly always catching real vampires. It also make mistakes, though, in the form of false positives. One percent of the time, it incorrectly diagnoses normal people as vampires, Pr(positive test result| mortal) = 0.01. The final bit of information we are told is that vampires are rather rare, being only 0.1% of the population, implying Pr(vampire) = 0.001. Suppose now that someone tests positive for vampirism. What's the probability that he or she is a bloodsucking immortal?

The correct approach is just to use Bayes' theorem to invert the probability, to compute Pr(vampire | positive). The calculation can be presented as:

$$\Pr(\text{vampire}|\text{positive})=\Pr(\text{positive}|\text{vampire})\Pr(\text{vampire})\Pr(\text{positive})$$

where Pr(positive) is the average probability of a positive test result, that is,

$$\Pr(\text{positive})=\Pr(\text{positive}|\text{vampire})\Pr(\text{vampire})+\Pr(\text{positive}|\text{moral})(1-\Pr(\text{vampire}))$$

Performing the calculation in R:

R code 3.1

```
Pr_Positive_Vampire <- 0.95
Pr_Positive_Mortal <- 0.01
Pr_Vampire <- 0.001
Pr_Positive <- Pr_Positive_Vampire * Pr_Vampire +
   Pr_Positive_Mortal * ( 1 - Pr_Vampire )
( Pr_Vampire_Positive <- Pr_Positive_Vampire*Pr_Vampire / Pr_Positive )
```

```
[1] 0.08683729
```

That corresponds to an 8.7% chance that the suspect is actually a vampire.

Most people find this result counterintuitive. And it's a very important result, because it mimics the structure of many realistic testing contexts, such as HIV and DNA testing, criminal profiling, and even statistical significance testing (see the Rethinking box at the end of this section). Whenever the condition of interest is very rare, having a test that finds all the true cases is still no guarantee that a positive result carries much information at all. The reason is that most positive results are false positives, even when all the true positives are detected correctly.

But I don't like these examples, for two reasons. First, there's nothing uniquely "Bayesian" about them. Remember: Bayesian inference is distinguished by a broad view of probability, not by the use of Bayes' theorem. Since all of the probabilities I provided above reference frequencies of events, rather than theoretical parameters, all major statistical philosophies would agree to use Bayes' theorem in this case. Second, and more important to our work in this chapter, these examples make Bayesian inference seem much harder than it has to be. Few people find it easy to remember which number goes where, probably because they never grasp the logic of the procedure.

It's just a formula that descends from the sky. If you are confused, it is only because you are trying to understand.

There is a way to present the same problem that does make it more intuitive, however. Suppose that instead of reporting probabilities, as before, I tell you the following:

(1) In a population of 100,000 people, 100 of them are vampires.
(2) Of the 100 who are vampires, 95 of them will test positive for vampirism.
(3) Of the 99,900 mortals, 999 of them will test positive for vampirism.

Now tell me, if we test all 100,000 people, what proportion of those who test positive for vampirism actually are vampires? Many people, although certainly not all people, find this presentation a lot easier.[50] Now we can just count up the number of people who test positive: 95 + 999 = 1094. Out of these 1094 positive tests, 95 of them are real vampires, so that implies:

$$\Pr(\text{vampire}|\text{positive}) = \frac{95}{1094} \approx 0.087$$

It's exactly the same answer as before, but without a seemingly arbitrary rule.

The second presentation of the problem, using counts rather than probabilities, is often called the *frequency format* or *natural frequencies*. Why a frequency format helps people intuit the correct approach remains contentious. Some people think that human psychology naturally works better when it receives information in the form a person in a natural environment would receive it. In the real world, we encounter counts only. No one has ever seen a probability, the thinking goes. But everyone sees counts ("frequencies") in their daily lives.

Regardless of the explanation for this phenomenon, we can exploit it. And in this chapter we exploit it by taking the probability distributions from the previous chapter and sampling from them to produce counts. The posterior distribution is a probability distribution. And like all probability distributions, we can imagine drawing *samples* from it. The sampled events in this case are parameter values. Most parameters have no exact empirical realization. The Bayesian formalism treats parameter distributions as relative plausibility, not as any physical random process. In any event, randomness is always a property of information, never of the real world. But inside the computer, parameters are just as empirical as the outcome of a coin flip or a die toss or an agricultural experiment. The posterior defines the expected frequency that different parameter values will appear, once we start plucking parameters out of it.

**Rethinking: The natural frequency phenomenon is not unique**. Changing the representation of a problem often makes it easier to address or inspires new ideas that were not available in an old representation.[51] In physics, switching between Newtonian and Lagrangian mechanics can make problems much easier. In evolutionary biology, switching between inclusive fitness and multilevel selection sheds new light on old models. And in statistics, switching between Bayesian and non-Bayesian representations often teaches us new things about both approaches.

This chapter teaches you basic skills for working with samples from the posterior distribution. It will seem a little silly to work with samples at this point, because the posterior distribution for the globe tossing model is very simple. It's so simple that it's no problem to work directly with the grid approximation or even the exact mathematical form.[52] But there are two reasons to adopt the sampling approach early on, before it's really necessary.

First, many scientists are uncomfortable with integral calculus, even though they have strong and valid intuitions about how to summarize data. Working with samples transforms a problem in calculus into a problem in data summary, into a frequency format problem. An integral in a typical Bayesian context is just the total probability in some interval. That can be a challenging calculus problem. But once you have samples from the probability distribution, it's just a matter of counting values in the interval. An empirical attack on the posterior allows the scientist to ask and answer more questions about the model, without relying upon a captive

mathematician. For this reason, it is easier and more intuitive to work with samples from the posterior, than to work with probabilities and integrals directly.

Second, some of the most capable methods of computing the posterior produce nothing but samples. Many of these methods are variants of Markov chain Monte Carlo techniques (MCMC, Chapter 9). So if you learn early on how to conceptualize and process samples from the posterior, when you inevitably must fit a model to data using MCMC, you will already know how to make sense of the output. Beginning with Chapter 9 of this book, you will use MCMC to open up the types and complexity of models you can practically fit to data. MCMC is no longer a technique only for experts, but rather part of the standard toolkit of quantitative science. So it's worth planning ahead.

So in this chapter we'll begin to use samples to summarize and simulate model output. The skills you learn here will apply to every problem in the remainder of the book, even though the details of the models and how the samples are produced will vary.

**Rethinking: Why statistics can't save bad science**. The vampirism example at the start of this chapter has the same logical structure as many different *signal detection* problems: (1) There is some binary state that is hidden from us; (2) we observe an imperfect cue of the hidden state; (3) we (should) use Bayes' theorem to logically deduce the impact of the cue on our uncertainty.

Scientific inference is sometimes framed in similar terms: (1) An hypothesis is either true or false; (2) we get a statistical cue of the hypothesis' falsity; (3) we (should) use Bayes' theorem to logically deduce the impact of the cue on the status of the hypothesis. It's the third step that is hardly ever done. I'm not really a fan of this framing. But let's consider a toy example, so you can see the implications. Suppose the probability of a positive finding, when an hypothesis is true, is Pr(sig|true) = 0.95. That's the *power* of the test. Suppose that the probability of a positive finding, when an hypothesis is false, is Pr(sig|false) = 0.05. That's the false-positive rate, like the 5% of conventional significance testing. Finally, we have to state the *base rate* at which hypotheses are true. Suppose for example that 1 in every 100 hypotheses turns out to be true. Then Pr(true) = 0.01. No one knows this value, but the history of science suggests it's small. See Chapter 17 for more discussion. Now compute the posterior:

$$\Pr(\text{true}|\text{pos})=\frac{\Pr(\text{pos}|\text{true})\Pr(\text{true})}{\Pr(\text{pos})}=\frac{\Pr(\text{pos}|\text{true})\Pr(\text{true})}{\Pr(\text{pos}|\text{true})\Pr(\text{true})+\Pr(\text{pos}|\text{false})\Pr(\text{false})}$$

Plug in the appropriate values, and the answer is approximately Pr(true|pos) = 0.16. So a positive finding corresponds to a 16% chance that the hypothesis is true. This is the same low base-rate phenomenon that applies in medical (and vampire) testing. You can shrink the false-positive rate to 1% and get this posterior probability up to 0.5, only as good as a coin flip. The most important thing to do is to improve the base rate, Pr(true), and that requires thinking, not testing.[53]