Statistical Rethinking

A Bayesian Course
with Examples
in R and Stan

Richard McElreath

This version compiled November 9, 2015

Contents

Preface xi
Audience xi
Teaching strategy xii
How to use this book xii
Installing the `rethinking` R package xvi
Acknowledgments xvi

Chapter 1. The Golem of Prague 1
1.1. Statistical golems 1
1.2. Statistical rethinking 4
1.3. Three tools for golem engineering 10
1.4. Summary 16

Chapter 2. Small Worlds and Large Worlds 19
2.1. The garden of forking data 20
2.2. Building a model 28
2.3. Components of the model 32
2.4. Making the model go 37
2.5. Summary 45
2.6. Practice 45

Chapter 3. Sampling the Imaginary 49
3.1. Sampling from a grid-approximate posterior 52
3.2. Sampling to summarize 53
3.3. Sampling to simulate prediction 61
3.4. Summary 68
3.5. Practice 69

Chapter 4. Linear Models 71
4.1. Why normal distributions are normal 72
4.2. A language for describing models 77
4.3. A Gaussian model of height 78
4.4. Adding a predictor 92
4.5. Polynomial regression 110
4.6. Summary 115
4.7. Practice 115

Chapter 5. Multivariate Linear Models 119
5.1. Spurious association 121
5.2. Masked relationship 135
5.3. When adding variables hurts 141

5.4. Categorical variables 152
5.5. Ordinary least squares and `lm` 159
5.6. Summary 162
5.7. Practice 162

Chapter 6. Overfitting, Regularization, and Information Criteria 165
6.1. The problem with parameters 167
6.2. Information theory and model performance 174
6.3. Regularization 186
6.4. Information criteria 188
6.5. Using information criteria 195
6.6. Summary 205
6.7. Practice 205

Chapter 7. Interactions 209
7.1. Building an interaction 211
7.2. Symmetry of the linear interaction 223
7.3. Continuous interactions 225
7.4. Interactions in design formulas 235
7.5. Summary 236
7.6. Practice 236

Chapter 8. Markov Chain Monte Carlo 241
8.1. Good King Markov and His island kingdom 242
8.2. Markov chain Monte Carlo 245
8.3. Easy HMC: `map2stan` 247
8.4. Care and feeding of your Markov chain 255
8.5. Summary 263
8.6. Practice 263

Chapter 9. Big Entropy and the Generalized Linear Model 267
9.1. Maximum entropy 268
9.2. Generalized linear models 280
9.3. Maximum entropy priors 288
9.4. Summary 289

Chapter 10. Counting and Classification 291
10.1. Binomial regression 292
10.2. Poisson regression 311
10.3. Other count regressions 322
10.4. Summary 328
10.5. Practice 329

Chapter 11. Monsters and Mixtures 331
11.1. Ordered categorical outcomes 331
11.2. Zero-inflated outcomes 342
11.3. Over-dispersed outcomes 346
11.4. Summary 351
11.5. Practice 352

Chapter 12. Multilevel Models 355
12.1. Example: Multilevel tadpoles 357
12.2. Varying effects and the underfitting/overfitting trade-off 364

12.3. More than one type of cluster 370
12.4. Multilevel posterior predictions 376
12.5. Summary 384
12.6. Practice 384

Chapter 13. Adventures in Covariance 387
13.1. Varying slopes by construction 389
13.2. Example: Admission decisions and gender 398
13.3. Example: Cross-classified chimpanzees with varying slopes 403
13.4. Continuous categories and the Gaussian process 410
13.5. Summary 419
13.6. Practice 419

Chapter 14. Missing Data and Other Opportunities 423
14.1. Measurement error 424
14.2. Missing data 431
14.3. Summary 439
14.4. Practice 439

Chapter 15. Horoscopes 441

Endnotes 445

Bibliography 457

Citation index 465

Topic index 467

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 21st 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 21st century really requires programming, of at least a minor sort. The code is 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 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.

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:

$$p_1 = \log(0.01^{200})$$
$$p_2 = 200 \times \log(0.01)$$

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 21st 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.

First, 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 itself. 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. Do not use them.

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 )
```

Finally, 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` and its dependencies with this code:

R code 0.5

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

Note that `rethinking` is not on the CRAN package archive, at least not yet. There's no real benefit to having a package on CRAN. 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 16th 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, with the Sun always shining on some portion of it. Its ruler was also Holy Roman Emperor, and his seat of power was Prague. The Emperor in the late 16th 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 known in 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 16th 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.

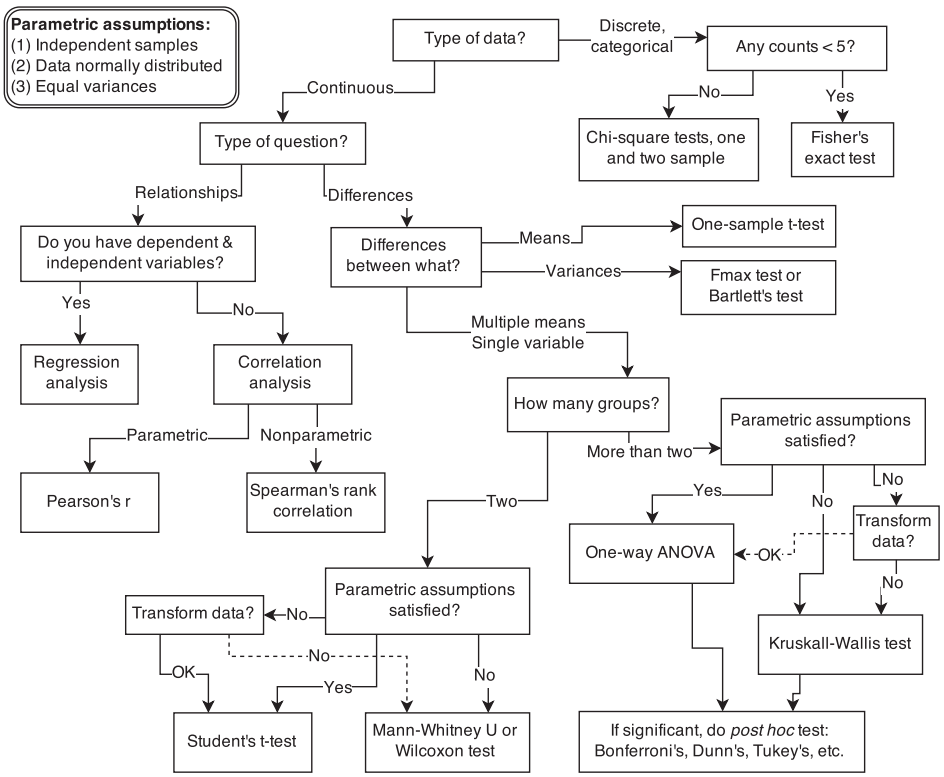

FIGURE 1.1. Example decision tree, or flowchart, for selecting an appropriate statistical procedure. Beginning at the top, the user answers a series of questions about measurement and intent, arriving eventually at the name of a procedure. Many such decision trees are possible.

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.

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 innovative 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 6).

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. In such settings, pre-manufactured golems may do nothing useful at all. Worse, they might wreck Prague. And if we keep adding new types of tools, soon there will be far too many to keep track of.

Instead, 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 framework 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, which just makes 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 statistical 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, as Popper recognized and argued.[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 very often work, then learning that it doesn't work via falsification shouldn't change your mind. But it might help you do better science, because it will open your eyes to the 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, many evolutionary biologists became interested in the proposal that the vast majority of evolution—changes in gene frequency—are caused not by natural selection, by 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.

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 thousands of possible detailed processes that can be described as "neutral," depending upon choices about, for example, 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]

In order to challenge these process models with evidence, they have to be made into statistical models. This usually means deriving the expected frequency distribution of some quantity—a "statistic"—in the model. For example, a very common statistic in this context is the frequency distribution (histogram) of the frequency of different genetic variants (alleles).

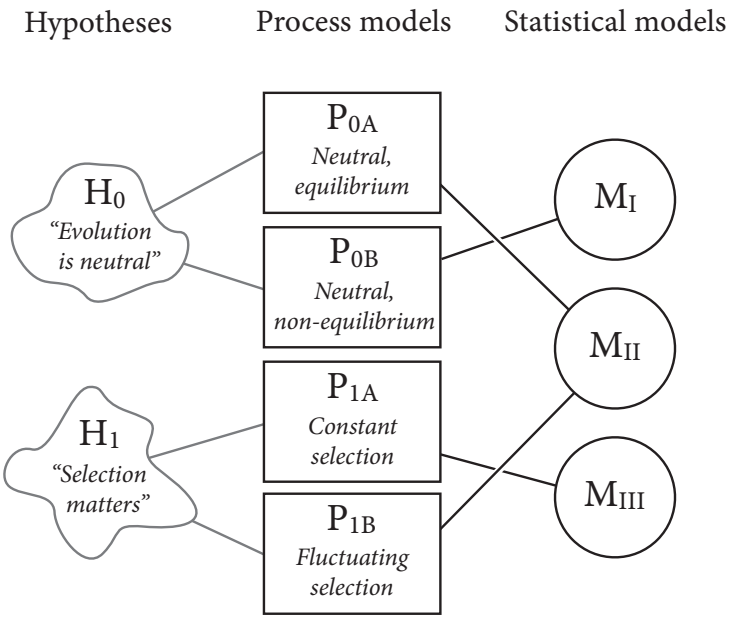

FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of "neutral" models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses 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.

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 at all clear what we are 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.

And while you might think that more routine statistical models, like linear regressions (Chapter 4), don't carry such risk, think again. A typical "null" in these contexts is just that there is zero *average* difference between groups. But there are usually many different ways for this average to be close to or consistent with zero, just as there are many different ways to get a power law. This recognition lies behind many common practices in statistical inference, such as consideration of unobserved variables and sampling bias.

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. In other words, explicitly compare predictions of more than one model, and you can save yourself from some ordinary kinds of folly.

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 9. 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. Not only can we not identify it, but we don't have to.

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 will 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, if any, 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 evidence satisfying all parties has yet (by 2015) 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 sub-atomic 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, among other things.[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 himself 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. Three 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 measure, forecast, and argue. Doing research benefits from the ability to produce and manipulate statistical 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. 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 the engineering training, you're always at someone else's mercy.

It can be hard to get a good education in statistical model building and criticism, though. Applied statistical modeling in the early 21st century is marked by the heavy use of several engineering tools that are almost always absent from introductory, and even many advanced, statistics courses. These tools aren't really new, but they are newly popular. And many of the recent advances in statistical inference depend upon computational innovations that feel more like computer science than classical statistics, so it's not clear who is responsible for teaching them, if anyone.

There are many tools worth learning. In this book I've chosen to focus on three broad ones that are in demand in both the social and biological sciences. These tools are:

(1) Bayesian data analysis
(2) Multilevel models
(3) Model comparison using information criteria

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 rest of this section, I provide brief introductions to the three tools.

1.3.1. Bayesian data analysis. For the classical Greeks and Romans, wisdom and chance were enemies. Minerva (Athena), symbolized by the owl, was the personification of wisdom. Fortuna (Tyche), symbolized by the wheel of fortune, was the personification of luck, both good and bad. Minerva was mindful and measuring, while Fortuna was fickle and unreliable. Only a fool would rely on Fortuna, while all wise folk appealed to Minerva.[21]

The rise of probability theory changed that. Statistical inference compels us instead to rely on Fortuna as a servant of Minerva, to use chance and uncertainty to discover reliable knowledge. All flavors of statistical inference have this motivation. But Bayesian data analysis embraces it most fully, by using the language of chance to describe the plausibility of different possibilities.

There are many ways to use the term "Bayesian." But mainly it denotes a particular interpretation of probability. In modest terms, Bayesian inference is no more than counting the numbers of ways things can happen, according to our assumptions. Things that can happen more ways are more plausible. And since probability theory is just a calculus for counting, this means that we can 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. Once you accept this gambit, the rest follows logically. Once we have defined our assumptions, Bayesian inference forces a purely logical way of processing that information to produce inference.

Chapter 2 explains this in depth. For now, it will help to have another probability concept 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 countable events and their frequencies in very large samples.[22] 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 20th century, put it:

> [...] 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 [...][23]

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.[24]

But these different attitudes towards probability do enforce different trade-offs. It will help to encounter a 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

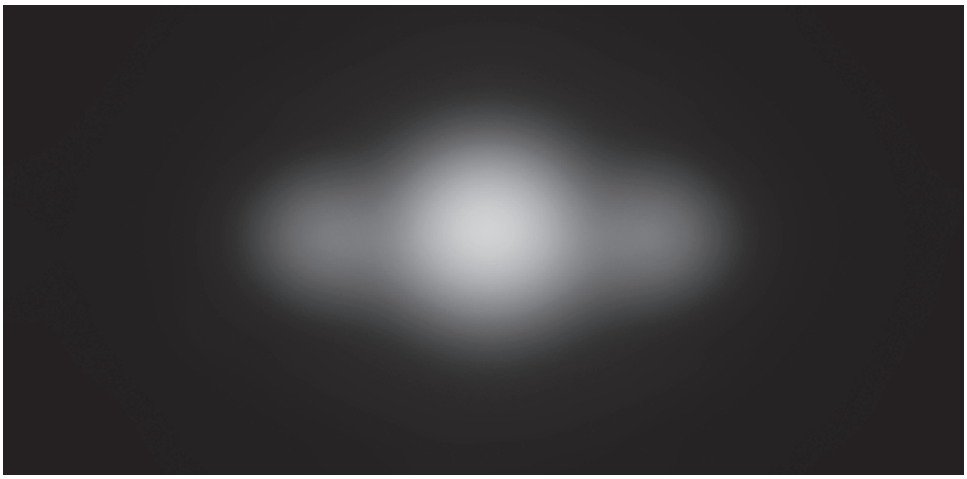

FIGURE 1.3. Saturn, much like Galileo must have seen it. The true shape is uncertain, but not because of any sampling variation. Probability theory can still help.

identify probability with frequency. As a result, the field of image reconstruction and processing is dominated by Bayesian algorithms.[25]

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 of Bayesian analysis 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**.[26] But this book neither describes nor advocates it.

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 9.

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 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.

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 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.[27] 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, if ever.

None of this ensures that Bayesian models will be more correct than non-Bayesian models. 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: A little history. Bayesian statistical inference is much older than the typical tools of introductory statistics, most of which were developed in the early 20th century. Versions of the Bayesian approach were applied to scientific work in the late 1700s and repeatedly in the 19th 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:

> [...] the theory of inverse probability is founded upon an error, and must be wholly rejected.[28]

Bayesian data analysis became increasingly accepted within statistics during the second half of the 20th 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.[29] 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. 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.

There are four typical and complementary reasons to use multilevel models:

(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.* Frequently, scholars pre-average some data to construct variables for a regression analysis. This can be dangerous, because averaging removes variation. It therefore manufactures false confidence. Multilevel models allow us to 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. This is why 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 the better multivariate tools. The same will eventually be true of multilevel models. But scholarly culture and curriculum still have a little catching up to do.

Rethinking: Multilevel election forecasting. One of the older applications of multilevel modeling is to forecast the outcomes of democratic elections. In the early 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 12 and 13. Tukey developed and used such models for NBC through 1978.[30] Contemporary election prediction and poll aggregation remains an active topic for multilevel modeling.[31]

1.3.3. Model comparison and information criteria. Beginning seriously in the 1960s and 1970s, statisticians began to develop a peculiar family of metrics for comparing structurally different models: **INFORMATION CRITERIA**. All of these criteria aim to let us compare models based upon future predictive accuracy. But they do so in more and less general ways and by using different approximations for different model types. So the number of unique information criteria in the statistical literature has grown quite large. Still, they all share this common enterprise.

The most famous information criterion is AIC, the Akaike (ah-kah-ee-kay) information criterion. AIC and related metrics—we'll discuss DIC and WAIC as well—explicitly build a model of the prediction task and use that model to estimate performance of each model you might wish to compare. Because the prediction is modeled, it depends upon assumptions. So information criteria do not in fact achieve the impossible, by seeing the future. They are still golems.

AIC and its kin are known as "information" criteria, because they develop their measure of model accuracy from **INFORMATION THEORY**. Information theory has a scope far beyond comparing statistical models. But it will be necessary to understand a little bit of general information theory, in order to really comprehend information criteria. So in Chapter 6, you'll also find a conceptual crash course in information theory.

What AIC and its kin actually do for a researcher is help with two common difficulties in model comparison.

(1) The most important statistical phenomenon that you may have never heard of is **OVERFITTING**.[32] Overfitting is the subject of Chapter 6. For now, you can understand overfitting with this mantra: *fitting is easy; prediction is hard*. 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. So if we wish to make good predictions, we

cannot judge our models simply on how well they fit our data. Information criteria provide estimates of predictive accuracy, rather than merely fit. So they compare models where it matters.

(2) A major benefit of using AIC and its kin is that they allow for comparison of multiple non-null models to the same data. Frequently, several plausible models of a phenomenon are known to us. The neutral evolution debate (page 6) is one example. In some empirical contexts, like social networks and evolutionary phylogenies, there are no reasonable or uniquely "null" models. This was also true of the neutral evolution example. In such cases, it's not only a good idea to explicitly compare models. It's also mandatory. Information criteria aren't the only way to conduct the comparison. But they are an accessible and widely used way.

Multilevel modeling and Bayesian data analysis have been worked on for decades and centuries, respectively. Information criteria are comparatively very young. 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[33]—and a great deal is known about them, both from analysis and experience.

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.[34] In short, if ancient north-east 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.

1.4. 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, multilevel models, and information theoretic model comparison.

The remainder of the book is organized into four interdependent 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 four chapters, 4 through 7, build multiple linear regression as a Bayesian tool. These chapters also move rather slowly, largely because of the emphasis on plotting results, including interaction effects. Problems of model complexity—overfitting—also feature prominently. So you'll also get an introduction to information theory in Chapter 6.

(3) The third part of the book, Chapters 8 through 11, presents generalized linear models of several types. Chapter 8 is something of a divider, as it introduces Markov

chain Monte Carlo, used to fit the non-linear models in Chapters 10 through 14. Chapter 9 introduces maximum entropy as an explicit procedure to help us design these models. Then Chapters 10 and 11 detail the models themselves.

(4) The last part, Chapters 12 through 14, gets around to multilevel models, both linear and generalized linear, as well as specialized types that address measurement error, missing data, and spatial correlation modeled through Gaussian processes. This material is fairly advanced, but it proceeds in the same mechanistic way as earlier material.

The final chapter, Chapter 15, 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. The solutions to these problems are available online.

12 Multilevel Models

In the year 1985, Clive Wearing lost his mind, but not his music.[154] Wearing was a musicologist and accomplished musician, but the same virus that causes cold sores, *Herpes simplex*, snuck into his brain and ate his hippocampus. The result was chronic anterograde amnesia—he cannot form new long-term memories. He remembers how to play the piano, though he cannot remember that he played it 5 minutes ago. Wearing now lives moment to moment, unaware of anything more than a few minutes into the past. Every cup of coffee is the first he has ever had.

Many statistical models also have anterograde amnesia. As the models move from one cluster—individual, group, location—in the data to another, estimating parameters for each cluster, they forget everything about the previous clusters. They behave this way, because the assumptions force them to. Any of the models from previous chapters that used dummy variables (page 152) to handle categories are programmed for amnesia. These models implicitly assume that nothing learned about any one category informs estimates for the other categories—the parameters are independent of one another and learn from completely separate portions of the data. This would be like forgetting you had ever been in a café, each time you go to a new café. Cafés do differ, but they are also alike.

Anterograde amnesia is bad for learning about the world. We want models that instead use all of the information in savvy ways. This does not mean treating all clusters as if they were the same. Instead it means learning simultaneously about each cluster while learning about the population of clusters. Doing both estimation tasks at the same time allows us to transfer information across clusters, and that transfer improves accuracy. That is the value of remembering.

Consider cafés again. Suppose we program a robot to visit two cafés, order coffee, and estimate the waiting times at each. The robot begins with a vague prior for the waiting times, say with a mean of 5 minutes and a standard deviation of 1. After ordering a cup of coffee at the first café, the robot observes a waiting time of 4 minutes. It updates its prior, using Bayes' theorem of course, with this information. This gives it a posterior distribution for the waiting time at the first café.

Now the robot moves on to a second café. When this robot arrives at the next café, what is its prior? It could just use the posterior distribution from the first café as its prior for the second café. But that implicitly assumes that the two cafés have the same average waiting time. Cafés are all pretty much the same, but they aren't identical. Likewise, it doesn't make much sense to ignore the observation from the first café. That would be anterograde amnesia.

So how can the coffee robot do better? It needs to represent the population of cafés and learn about that population. The distribution of waiting times in the population becomes the prior for each café. But unlike priors in previous chapters, this prior is actually learned

from the data. This means the robot tracks a parameter for each café as well as at least two parameters to describe the population of cafés: an average and a standard deviation. As the robot observes waiting times, it updates everything: the estimates for each café as well as the estimates for the population. If the population seems highly variable, then the prior is flat and uninformative and, as a consequence, the observations at any one café do very little to the estimate at another. If instead the population seems to contain little variation, then the prior is narrow and highly informative. An observation at any one café will have a big impact on estimates at any other café.

In this chapter, you'll see the formal version of this argument and how it leads us to **MULTILEVEL MODELS**. These models remember features of each cluster in the data as they learn about all of the clusters. Depending upon the variation among clusters, which is learned from the data as well, the model pools information across clusters. This pooling tends to improve estimates about each cluster. This improved estimation leads to several, more pragmatic sounding, benefits of the multilevel approach. I mentioned them in Chapter 1. They are worth repeating.

(1) *Improved estimates for repeat sampling*. When more than one observation arises from the same individual, location, or time, then traditional, single-level models either maximally underfit or overfit the data.

(2) *Improved estimates for imbalance in sampling*. When some individuals, locations, or times are sampled more than others, multilevel models automatically cope with differing uncertainty across these clusters. This prevents over-sampled clusters from unfairly dominating inference.

(3) *Estimates of 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) *Avoid averaging, retain variation*. Frequently, scholars pre-average some data to construct variables. This can be dangerous, because averaging removes variation, and there are also typically several different ways to perform the averaging. Averaging therefore both manufactures false confidence and introduces arbitrary data transformations. Multilevel models allow us to preserve the uncertainty and avoid data transformations.

All of these benefits flow out of the same strategy and model structure. You learn one basic design and you get all of this for free.

When it comes to regression, multilevel regression deserves to be the default approach. There are certainly contexts in which it would be better to use an old-fashioned single-level model. But the contexts in which multilevel models are superior are much more numerous. It is better to begin to build a multilevel analysis, and then realize it's unnecessary, than to overlook it. And once you grasp the basic multilevel strategy, it becomes much easier to incorporate related tricks such as allowing for measurement error in the data and even modeling missing data itself (Chapter 14).

There are costs of the multilevel approach. The first is that we have to make some new assumptions. We have to define the distributions from which the characteristics of the clusters arise. Luckily, conservative maximum entropy distributions do an excellent job in this context. Second, there are new estimation challenges that come with the full multilevel approach. These challenges lead us headfirst into MCMC estimation. Third, multilevel models can be hard to understand, because they make predictions at different levels of the data. In

many cases, we are interested in only one or a few of those levels, and as a consequence, model comparison using metrics like DIC and WAIC becomes more subtle. The basic logic remains unchanged, but now we have to make more decisions about which parameters in the model we wish to focus on.

This chapter has the following progression. First, we'll work through an extended example of building and fitting a multilevel model for clustered data. Then we'll simulate clustered data, to demonstrate the improved accuracy the approach delivers. This improved accuracy arises from the same underfitting and overfitting trade-off you met in Chapter 6. Then we'll finish by looking at contexts in which there is more than one type of clustering in the data. All of this work lays a foundation for more advanced multilevel examples in the next two chapters.

Rethinking: A model by any other name. Multilevel models go by many different names, and some statisticians use the same names for different specialized variants, while others use them all interchangeably. The most common synonyms for "multilevel" are **HIERARCHICAL** and **MIXED EFFECTS**. The type of parameters that appear in multilevel models are most commonly known as **RANDOM EFFECTS**, which itself can mean very different things to different analysts and in different contexts.[155] And even the innocent term "level" can mean different things to different people. There's really no cure for this swamp of vocabulary aside from demanding a mathematical or algorithmic definition of the model. Otherwise, there will always be ambiguity.

12.1. Example: Multilevel tadpoles

The heartwarming focus of this example are experiments exploring Reed frog (*Hyperolius spinigularis*) tadpole mortality.[156] The natural history background to these data is very interesting. Take a look at the full paper, if amphibian life history dynamics interests you. But even if it doesn't, load the data and acquaint yourself with the variables:

R code 12.1

```
library(rethinking)
data(reedfrogs)
d <- reedfrogs
str(d)
```

```
'data.frame': 48 obs. of  5 variables:
 $ density : int  10 10 10 10 10 10 10 10 10 10 ...
 $ pred    : Factor w/ 2 levels "no","pred": 1 1 1 1 1 1 1 1 2 2 ...
 $ size    : Factor w/ 2 levels "big","small": 1 1 1 1 2 2 2 2 1 1 ...
 $ surv    : int  9 10 7 10 9 9 10 9 4 9 ...
 $ propsurv: num  0.9 1 0.7 1 0.9 0.9 1 0.9 0.4 0.9 ...
```

For now, we'll only be interested in number surviving, `surv`, out of an initial count, `density`. In the practice at the end of the chapter, you'll consider the other variables, which are experimental manipulations.

There is a lot of variation in these data. Some of the variation comes from experimental treatment. But a lot of it comes from other sources. Think of each row as a "tank," an experimental environment that contains tadpoles. There are lots of things peculiar to each tank that go unmeasured, and these unmeasured factors create variation in survival across tanks, even when all the predictor variables have the same value. These tanks are an example

of a *cluster* variable. Multiple observations, the tadpoles in this case, are made within each cluster.

So we have repeat measures and heterogeneity across clusters. If we ignore the clusters, assigning the same intercept to each of them, then we risk ignoring important variation in baseline survival. This variation could mask association with other variables. If we instead estimate a unique intercept for each cluster, using a dummy variable for each tank, we instead practice anterograde amnesia. After all, tanks are different but each tank does help us estimate survival in the other tanks. So it doesn't make sense to forget entirely, moving from one tank to another.

A multilevel model, in which we simultaneously estimate both an intercept for each tank and the variation among tanks, is what we want. This will be a **VARYING INTERCEPTS** model. Varying intercepts are the simplest kind of **VARYING EFFECTS**.[157] For each cluster in the data, we use a unique intercept parameter. This is no different than the categorical variable examples from previous chapters, except now we also adaptively learn the prior that is common to all of these intercepts. This adaptive learning is the absence of amnesia discussed at the start of the chapter. When what we learn about each cluster informs all the other clusters, we learn the prior simultaneous to learning the intercepts.

Here is a model for predicting tadpole mortality in each tank, using the regularizing priors of earlier chapters:

$$\begin{aligned} s_i &\sim \text{Binomial}(n_i, p_i) && \text{[likelihood]} \\ \text{logit}(p_i) &= \alpha_{\text{TANK}[i]} && \text{[unique log-odds for each tank } i\text{]} \\ \alpha_{\text{TANK}} &\sim \text{Normal}(0, 5) && \text{[weakly regularizing prior]} \end{aligned}$$

And you can fit this to the data in the standard way, using `map` or `map2stan`. We'll use `map2stan` from here onwards, because the next model will not work in `map`.

R code 12.2

```
library(rethinking)
data(reedfrogs)
d <- reedfrogs

# make the tank cluster variable
d$tank <- 1:nrow(d)

# fit
m12.1 <- map2stan(
    alist(
        surv ~ dbinom( density , p ) ,
        logit(p) <- a_tank[tank] ,
        a_tank[tank] ~ dnorm( 0 , 5 )
    ),
    data=d )
```

If you inspect the estimates, `precis(m12.1,depth=2)`, you'll see 48 different intercept offsets, one for each tank. To get each tank's expected survival probability, just take one of the `a_tank` values and then use the logistic transform. So far there is nothing new here.

Now let's fit the multilevel model, which adaptively pools information across tanks. All that is required to enable adaptive pooling is to make the prior for the `a_tank` parameters a function of its own parameters. Here is the multilevel model, in mathematical form, with

the changes from the previous model highlighted in blue:

$$
\begin{aligned}
s_i &\sim \text{Binomial}(n_i, p_i) && \text{[likelihood]} \\
\text{logit}(p_i) &= \alpha_{\text{TANK}[i]} && \text{[log-odds for tank on row } i\text{]} \\
\alpha_{\text{TANK}} &\sim \text{Normal}(\alpha, \sigma) && \text{[varying intercepts prior]} \\
\alpha &\sim \text{Normal}(0, 1) && \text{[prior for average tank]} \\
\sigma &\sim \text{HalfCauchy}(0, 1) && \text{[prior for standard deviation of tanks]}
\end{aligned}
$$

Notice that the prior for the α_{TANK} intercepts is now a function of two parameters, α and σ. This is where the "multi" in multilevel arises.[158] The Gaussian distribution with mean α and standard deviation σ is the prior for each tank's intercept. But that prior itself has priors for α and σ. So there are two *levels* in the model, each resembling a simpler model. In the top level, the outcome is s, the parameters are α_{TANK}, and the prior is $\alpha_{\text{TANK}} \sim \text{Normal}(\alpha, \sigma)$. In the second level, the "outcome" variable is the vector of intercept parameters, α_{TANK}. The parameters are α and σ, and their priors are $\alpha \sim \text{Normal}(0, 1)$ and $\sigma \sim \text{HalfCauchy}(0, 1)$. For more explanation of the σ prior, see the Overthinking box on the next page.

These two parameters, α and σ, are often referred to as **HYPERPARAMETERS**. They are parameters for parameters. And their priors are often called **HYPERPRIORS**. In principle, there is no limit to how many "hyper" levels you can install in a model. For example, different populations of tanks could be embedded within different regions of habitat. But in practice there are limits, both because of computation and our ability to understand the model.

Rethinking: Why Gaussian tanks? In the multilevel tadpole model, the population of tanks is assumed to be Gaussian. Why? The least satisfying answer is "convention." The Gaussian assumption is extremely common. A more satisfying answer is "pragmatism." The Gaussian assumption is easy to work with, and it generalizes easily to more than one dimension. This generalization will be important for handling varying slopes in the next chapter. But my preferred answer is instead "entropy." If all we are willing to say about a distribution is the mean and variance, then the Gaussian is the most conservative assumption (Chapter 9). There is no rule requiring the Gaussian distribution of varying effects, though. So if you have a good reason to use another distribution, then do so. The practice problems at the end of the chapter provide an example.

Fitting the model to data estimates both levels simultaneously, in the same way that our robot at the start of the chapter learned both about each café and the variation among cafés. But you cannot fit this model with `map`. Why? Because the likelihood must now average over the level 2 parameters α and σ. But `map` just hill climbs, using static values for all of the parameters. It can't see the levels. For more explanation, see the Overthinking box further down. You can however fit this model with `map2stan`:

R code 12.3

```
m12.2 <- map2stan(
    alist(
        surv ~ dbinom( density , p ) ,
        logit(p) <- a_tank[tank] ,
        a_tank[tank] ~ dnorm( a , sigma ) ,
        a ~ dnorm(0,1) ,
        sigma ~ dcauchy(0,1)
    ), data=d , iter=4000 , chains=4 )
```

This model fit provides estimates for 50 parameters: one overall sample intercept α, the variance among tanks σ, and then 48 per-tank intercepts. Let's check WAIC though to see the effective number of parameters. We'll compare the earlier model, `m12.1`, with the new multilevel model:

R code 12.4

```
compare( m12.1 , m12.2 )
```

```
        WAIC pWAIC dWAIC weight    SE  dSE
m12.2 1010.2  38.0   0.0      1 37.94   NA
m12.1 1023.3  49.4  13.1      0 43.01 6.54
```

There are two facts to note here. First, the multilevel model has only 38 effective parameters. There are 12 fewer effective parameters than actual parameters, because the prior assigned to each intercept shrinks them all towards the mean α. In this case, the prior is reasonably strong. Check the mean of `sigma` with `precis` or `coef` and you'll see it's around 1.6. This is a **REGULARIZING PRIOR**, like you've used in previous chapters, but now the amount of regularization has been learned from the data itself.[159] Second, notice that the multilevel model `m12.2` has fewer effective parameters than the ordinary fixed model `m12.1`. This is despite the fact that the ordinary model has fewer actual parameters, only 48 instead of 50. The extra two parameters in the multilevel model allowed it to learn a more aggressive regularizing prior, to adaptively regularize. This resulted in a less flexible posterior and therefore fewer effective parameters.

Overthinking: MAP fails, MCMC wins. Why doesn't MAP estimation, using for example `map`, work with multilevel models? When a prior is itself a function of parameters, there are two levels of uncertainty. This means that the probability of the data, conditional on the parameters, must average over each level. Ordinary MAP estimation cannot handle the averaging in the likelihood, because in general it's not possible to derive an analytical solution. That means there is no unified function for calculating the log-posterior. So your computer cannot directly find its minimum (the maximum of the posterior).

Some other computational approach is needed. It is possible to extend the mode-finding optimization strategy to these models, but we don't want to be stuck with optimization in general. One reason is that the posterior of these models is routinely non-Gaussian. More generally, as models become more complex, a phenomenon known as *concentration of measure* guarantees that the posterior mode will be far from the posterior median. So we really need to give up optimization as a strategy. One robust solution is MCMC.

To appreciate the impact of this adaptive regularization, let's plot and compare the posterior medians from models `m12.1` and `m12.2`. The code that follows is long, only because it decorates the plot with informative labels. The basic code is just the first part, which extracts samples and computes medians.

R code 12.5

```
# extract Stan samples
post <- extract.samples(m12.2)

# compute median intercept for each tank
# also transform to probability with logistic
d$propsurv.est <- logistic( apply( post$a_tank , 2 , median ) )
```

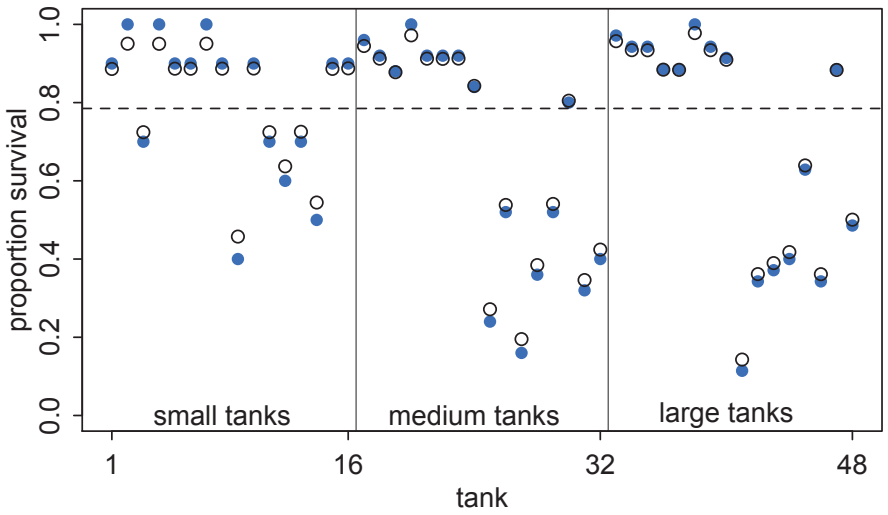

FIGURE 12.1. Empirical proportions of survivors in each tadpole tank, shown by the filled blue points, plotted with the 48 per-tank estimates from the multilevel model, shown by the black circles. The dashed line locates the overall average proportion of survivors across all tanks. The vertical lines divide tanks with different initial densities of tadpoles: small tanks (10 tadpoles), medium tanks (25), and large tanks (35). In every tank, the posterior median from the multilevel model is closer to the dashed line than the empirical proportion is. This reflects the pooling of information across tanks, to help with inference about each tank.

```
# display raw proportions surviving in each tank
plot( d$propsurv , ylim=c(0,1) , pch=16 , xaxt="n" ,
    xlab="tank" , ylab="proportion survival" , col=rangi2 )
axis( 1 , at=c(1,16,32,48) , labels=c(1,16,32,48) )

# overlay posterior medians
points( d$propsurv.est )

# mark posterior median probability across tanks
abline( h=logistic(median(post$a)) , lty=2 )

# draw vertical dividers between tank densities
abline( v=16.5 , lwd=0.5 )
abline( v=32.5 , lwd=0.5 )
text( 8 , 0 , "small tanks" )
text( 16+8 , 0 , "medium tanks" )
text( 32+8 , 0 , "large tanks" )
```

You can see the result in FIGURE 12.1. The horizontal axis is tank index, from 1 to 48. The vertical is proportion of survivors in a tank. The filled blue points show the raw proportions, computed from the observed counts. These values are already present in the data frame, in the `propsurv` column. The black circles are instead the varying intercept medians. The horizontal dashed line at about 0.8 is the estimated median survival proportion in the population of tanks, α. It is not the same as the empirical mean survival. The vertical gray lines divide tanks with different initial counts of tadpoles—10 (left), 25 (middle), and 35 (right).

First, notice that in every case, the multilevel estimate is closer to the dashed line than the raw empirical estimate is. It's as if the entire distribution of black circles has been shrunk towards the dashed line at the center of the data, leaving the blue points behind on the outside. This phenomenon is sometimes called **SHRINKAGE**, and it results from regularization (as in Chapter 6). Second, notice that the estimates for the smaller tanks have shrunk farther from the blue points. As you move from left to right in the figure, the initial densities of tadpoles increase from 10 to 25 to 35, as indicated by the vertical dividers. In the smallest tanks, it is easy to see differences between the open estimates and empirical blue points. But in the largest tanks, there is little difference between the blue points and open circles. Varying intercepts for the smaller tanks, with smaller sample sizes, shrink more. Third, note that the farther a blue point is from the dashed line, the greater the distance between it and the corresponding multilevel estimate. Shrinkage is stronger, the further a tank's empirical proportion is from the global average α.

All three of these phenomena arise from a common cause: pooling information across clusters (tanks) to improve estimates. What **POOLING** means here is that each tank provides information that can be used to improve the estimates for all of the other tanks. Each tank helps in this way, because we made an assumption about how the varying log-odds in each tank related to all of the others. We assumed a distribution, the normal distribution in this case. Once we have a distributional assumption, we can use Bayes' theorem to optimally (in the small world only) share information among the clusters.

What does the inferred population distribution of survival look like? We can visualize it by sampling from the posterior distribution, as usual. First we'll plot 100 Gaussian distributions, one for each of the first 100 samples from the posterior distribution of both α and σ. Then we'll sample 8000 new log-odds of survival for individual tanks. The result will be a posterior distribution of variation in survival in the population of tanks. Before we do the sampling though, remember that "sampling" from a posterior distribution is not a simulation of empirical sampling. It's just a convenient way to characterize and work with the uncertainty in the distribution. Now the sampling:

R code 12.6

```
# show first 100 populations in the posterior
plot( NULL , xlim=c(-3,4) , ylim=c(0,0.35) ,
    xlab="log-odds survive" , ylab="Density" )
for ( i in 1:100 )
    curve( dnorm(x,post$a[i],post$sigma[i]) , add=TRUE ,
    col=col.alpha("black",0.2) )

# sample 8000 imaginary tanks from the posterior distribution
sim_tanks <- rnorm( 8000 , post$a , post$sigma )

# transform to probability and visualize
```

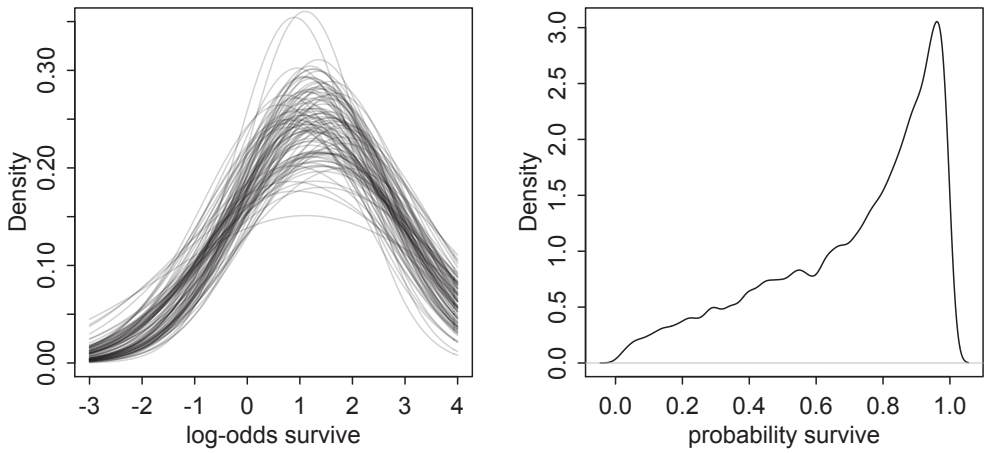

FIGURE 12.2. The inferred population of survival across tanks. Left: 100 Gaussian distributions of the log-odds of survival, sampled from the posterior of `m12.2`. Right: Survival probabilities for 8000 new simulated tanks, averaging over the posterior distribution on the left.

```
dens( logistic(sim_tanks) , xlab="probability survive" )
```

The results are displayed in FIGURE 12.2. Notice that there is uncertainty about both the location, α, and scale, σ, of the population distribution of log-odds of survival. All of this uncertainty is propagated into the simulated probabilities of survival.

Rethinking: Varying intercepts as over-dispersion. In the previous chapter (page 346), the beta-binomial and gamma-Poisson models were presented as ways for coping with **OVER-DISPERSION** of count data. Varying intercepts accomplish the same thing, allowing count outcomes to be over-dispersed. They accomplish this, because when each observed count gets its own unique intercept, but these intercepts are pooled through a common distribution, the predictions expect over-dispersion just like a beta-binomial or gamma-Poisson model would. Compared to a beta-binomial or gamma-Poisson model, a binomial or Poisson model with a varying intercept on every observed outcome will often be easier to estimate and easier to extend. There will be an example of this approach, later in this chapter.

Overthinking: Priors for variance components. The examples in this book use weakly regularizing half-Cauchy priors for variance components, the σ parameters that estimate the variation across clusters in the data. These Cauchy priors work very well in routine multilevel modeling. But there are two common contexts in which they can be problematic. First, sometimes there isn't much information in the data with which to estimate the variance. For example, if you only have 5 clusters, then that's something like trying to estimate a variance with 5 data points. Second, in non-linear models with logit and log links, floor and ceiling effects sometimes render extreme values of the variance equally plausible as more realistic values. In such cases, the trace plot for the variance parameters may swing around over very large values. It can do this, because the Cauchy prior has a very thick and long tail, extending into very large values. Such large values are typically *a priori* impossible. Often, the chain

will still sample validly, but it might be highly inefficient, exhibiting small `n_eff` values and possibly many divergent iterations.

To improve such a model, instead of using half-Cauchy priors for the variance components, you can use exponential priors. For example:

$$\begin{aligned} s_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{TANK}[i]} \\ \alpha_{\text{TANK}} &\sim \text{Normal}(\alpha, \sigma) \\ \alpha &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Exponential}(1) \end{aligned}$$

The exponential prior—`dexp(1)` in R code—has a much thinner tail than the Cauchy does. This induces more conservatism in estimates and can help your Markov chain converge correctly. The exponential is also the maximum entropy prior for the standard deviation, provided all we want to say *a priori* is the expected value. That is to say that the only information contained in an exponential prior is the mean value and the positive constraint.

Again, using the exponential instead of the Cauchy isn't usually necessary. But there are cases, especially with non-linear models with ceiling or floor effects, in which the variance components can be only weakly identified. In those cases, you are going to have to add more strongly regularizing priors in order to make any inference at all. And of course, it is typically useful to try different priors to ensure that inference either is insensitive to them or rather to measure how inference is altered.

12.2. Varying effects and the underfitting/overfitting trade-off

Varying intercepts are just regularized estimates, but adaptively regularized by estimating how diverse the clusters are while estimating the features of each cluster. This fact is not easy to grasp, so if it still seems mysterious, this section aims to further relate the properties of multilevel estimates to the foundational underfitting/overfitting dilemma from Chapter 6.

A major benefit of using varying effects estimates, instead of the empirical raw estimates, is that they provide more accurate estimates of the individual cluster (tank) intercepts.[160] On average, the varying effects actually provide a better estimate of the individual tank (cluster) means. The reason that the varying intercepts provide better estimates is that they do a better job of trading off underfitting and overfitting.

To understand this in the context of the reed frog example, suppose that instead of experimental tanks we had natural ponds, so that we might be concerned with making predictions for the same clusters in the future. We'll approach the problem of predicting future survival in these ponds, from three perspectives:

(1) Complete pooling. This means we assume that the population of ponds is invariant, the same as estimating a common intercept for all ponds.
(2) No pooling. This means we assume that each pond tells us nothing about any other pond. This is the model with amnesia.
(3) Partial pooling. This means using an adaptive regularizing prior, as in the previous section.

First, suppose you ignore the varying intercepts and just use the overall mean across all ponds, α, to make your predictions for each pond. A lot of data contributes to your estimate of α, and so it can be quite precise. However, your estimate of α is unlikely to exactly match the mean of any particular pond. As a result, the total sample mean underfits the data. This is the **COMPLETE POOLING** approach, pooling the data from all ponds to produce a single

estimate that is applied to every pond. This sort of model is equivalent to assuming that the variation among ponds is zero—all ponds are identical.

Second, suppose you use the survival proportions for each pond to make predictions. This means using a separate intercept for each pond. The blue points in FIGURE 12.1 are this same kind of estimate. In each particular pond, quite little data contributes to each estimate, and so these estimates are rather imprecise. This is particularly true of the smaller ponds, where less data goes into producing the estimates. As a consequence, the error of these estimates is high, and they are rather overfit to the data. Standard errors for each intercept can be very large, and in extreme cases, even infinite. These are sometimes called the **NO POOLING** estimates. No information is shared across ponds. It's like assuming that the variation among ponds is infinite, so nothing you learn from one pond helps you predict another.

Third, when you estimate varying intercepts, you use **PARTIAL POOLING** of information to produce estimates for each cluster that are less underfit than the grand mean and less overfit than the no-pooling estimates. As a consequence, they tend to be better estimates of the true per-cluster (per-pond) means. This will be especially true when ponds have few tadpoles in them, because then the no pooling estimates will be especially overfit. When a lot of data goes into each pond, then there will be less difference between the varying effect estimates and the no pooling estimates.

To demonstrate this fact, we'll simulate some tadpole data. That way, we'll know the true per-pond survival probabilities. Then we can compare the no-pooling estimates to the partial pooling estimates, by computing how close each gets to the true values they are trying to estimate. The rest of this section shows how to do such a simulation.

Learning to simulate and validate models and model fitting in this way is extremely valuable. Once you start using more complex models, you will want to ensure that your code is working and that you understand the model. You can help in this project by simulating data from the model, with specified parameter values, and then making sure that your method of estimation can recover the parameters within tolerable ranges of precision. Even just simulating data from a model structure has a huge impact on understanding.

12.2.1. The model. The first step is to define the model we'll be using. I'll use the same basic multilevel binomial model as before, but now with "ponds" instead of "tanks":

$$\begin{aligned} s_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{POND}[i]} \\ \alpha_{\text{POND}} &\sim \text{Normal}(\alpha, \sigma) \\ \alpha &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{HalfCauchy}(0, 1) \end{aligned}$$

So to simulate data from this process, we need to assign values to:

- α, the average log-odds of survival in the entire population of ponds
- σ, the standard deviation of the distribution of log-odds of survival among ponds
- α_{POND}, a vector of individual pond intercepts, one for each pond

We'll also need to assign sample sizes, n_i, to each pond. But once we've made all of those choices, we can easily simulate counts of surviving tadpoles, straight from the top-level binomial process, using `rbinom`. We'll do it all one step at a time.

Note that the priors are part of the model when we estimate, but not when we simulate. Why? Because priors are epistemology, not ontology. They represent the initial state of information of our robot, not a statement about how nature chooses parameter values.

12.2.2. Assign values to the parameters. I'm going to assign specific values representative of the actual tadpole data, to make the upcoming plot that demonstrates the increased accuracy of the varying effects estimates. But you can come back to this step later and change them to whatever you want.

Here's the code to initialize the values of α, σ, the number of ponds, and the sample size n_i in each pond.

R code 12.7

```
a <- 1.4
sigma <- 1.5
nponds <- 60
ni <- as.integer( rep( c(5,10,25,35) , each=15 ) )
```

I've chosen 60 ponds, with 15 each of initial tadpole density 5, 10, 25, and 35. I've chosen these densities to illustrate how the error in prediction varies with sample size. The use of `as.integer` in the last line arises from a subtle issue with how Stan, and therefore `map2stan`, works. See the Overthinking box at the bottom of the page for an explanation.

The values $\alpha = 1.4$ and $\sigma = 1.5$ define a Gaussian distribution of individual pond log-odds of survival. So now we need to simulate all 60 of these intercept values from the implied Gaussian distribution with mean α and standard deviation σ:

R code 12.8

```
a_pond <- rnorm( nponds , mean=a , sd=sigma )
```

Go ahead and inspect the contents of `a_pond`. It should contain 60 log-odds values, one for each simulated pond.

Finally, let's bundle some of this information in a data frame, just to keep it organized.

R code 12.9

```
dsim <- data.frame( pond=1:nponds , ni=ni , true_a=a_pond )
```

Go ahead and inspect the contents of `dsim`, the simulated data. The first column is the pond index, 1 through 60. The second column is the initial tadpole count in each pond. The third column is the true log-odds survival for each pond.

Overthinking: Data types and Stan models. There are two basic types of numerical data in R, integers and real values. A number like "3" could be either. Inside your computer, integers and real ("`numeric`") values are represented differently. For example, here is the same vector of values generated as both:

R code 12.10

```
class(1:3)
class(c(1,2,3))
```

```
[1] "integer"
[1] "numeric"
```

Usually, you don't have to manage these types, because R manages them for you. But when you pass values to Stan, or another external program, often the internal representation does matter. In particular, Stan and `map2stan` sometimes require explicit integers. For example, in a binomial model,

the "size" variable that specifies the number of trials must be of integer type. Stan may provide a mysterious warning message about a function not being found, when the size variable is instead of "real" type, or what R calls `numeric`. Using `as.integer` before passing the data to Stan or `map2stan` will resolve the issue.

12.2.3. Simulate survivors. Now we're ready to simulate the binomial survival process. Each pond i has n_i potential survivors, and nature flips each tadpole's coin, so to speak, with probability of survival p_i. This probability p_i is implied by the model definition, and is equal to:

$$p_i = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}$$

The model uses a logit link, and so the probability is defined by the logistic function.

Putting the logistic into the random binomial function, we can generate a simulated survivor count for each pond:

R code 12.11

```
dsim$si <- rbinom( nponds , prob=logistic(dsim$true_a) , size=dsim$ni )
```

As usual with R, if you give it a list of values, it returns a new list of the same length. In the above, each paired α_i (`dsim$true_a`) and n_i (`dsim$ni`) is used to generate a random survivor count with the appropriate probability of survival and maximum count. These counts are stored in a new column in `dsim`.

12.2.4. Compute the no-pooling estimates. We're ready to start analyzing the simulated data now. The easiest task is to just compute the no-pooling estimates. We can accomplish this straight from the empirical data, just by calculating the proportion of survivors in each pond. I'll keep these estimates on the probability scale, instead of translating them to the log-odds scale, because we'll want to compare the quality of the estimates on the probability scale later.

R code 12.12

```
dsim$p_nopool <- dsim$si / dsim$ni
```

Now there's another column in `dsim`, containing the empirical proportions of survivors in each pond. These are the same no-pooling estimates you'd get by fitting a model with a dummy variable for each pond and flat priors that induce no regularization.

12.2.5. Compute the partial-pooling estimates. Now to fit the model to the simulated data, using `map2stan`. I'll use a single long chain in this example, but keep in mind that you need to use multiple chains to check convergence to the right posterior distribution. In this case, it's safe. But don't get cocky.

R code 12.13

```
m12.3 <- map2stan(
    alist(
        si ~ dbinom( ni , p ),
        logit(p) <- a_pond[pond],
        a_pond[pond] ~ dnorm( a , sigma ),
        a ~ dnorm(0,1),
        sigma ~ dcauchy(0,1)
    ),
```

```
    data=dsim , iter=1e4 , warmup=1000 )
```

We've fit the basic varying intercept model above. You can take a look at the estimates for α and σ with the usual `precis` approach:

R code 12.14
```
precis(m12.3,depth=2)
```

```
            Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a_pond[1]   1.45   0.95      -0.11       2.89  9000    1
a_pond[2]   1.47   0.95      -0.02       2.96  9000    1
...
a_pond[59]  1.81   0.47       1.02       2.52  7314    1
a_pond[60]  2.03   0.50       1.24       2.82  9000    1
a           1.13   0.23       0.78       1.50  5848    1
sigma       1.59   0.22       1.25       1.93  2705    1
```

I've abbreviated the output, since there are 60 intercept parameters, one for each pond.

Now that we've found these estimates, let's compute the predicted survival proportions and add those proportions to our growing simulation data frame. To indicate that it contains the partial pooling estimates, I'll call the column `p.partpool`.

R code 12.15
```
estimated.a_pond <- as.numeric( coef(m12.3)[1:60] )
dsim$p_partpool <- logistic( estimated.a_pond )
```

If we want to compare to the true per-pond survival probabilities used to generate the data, then we'll also need to compute those, using the `true_a` column:

R code 12.16
```
dsim$p_true <- logistic( dsim$true_a )
```

The last thing we need to do, before we can plot the results and realize the point of this lesson, is to compute the absolute error between the estimates and the true varying effects. This is easy enough, using the existing columns:

R code 12.17
```
nopool_error <- abs( dsim$p_nopool - dsim$p_true )
partpool_error <- abs( dsim$p_partpool - dsim$p_true )
```

Now we're ready to plot. This is enough to get the basic display:

R code 12.18
```
plot( 1:60 , nopool_error , xlab="pond" , ylab="absolute error" ,
    col=rangi2 , pch=16 )
points( 1:60 , partpool_error )
```

I've decorated this plot with some additional information, displayed in Figure 12.3. Your own plot will look different, because of simulation variance. The pattern displayed in the figure is the central tendency. To see how to quickly re-run the model on newly simulated data, without re-compiling, see the Overthinking box at the end of this section.

The filled blue points in Figure 12.3 display the no-pooling estimates. The black circles show the varying effect estimates. The horizontal axis is the pond index, from 1 through

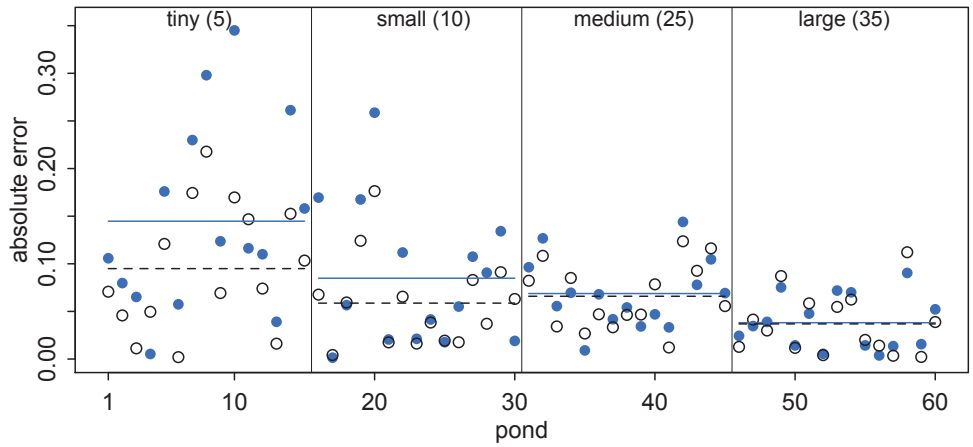

FIGURE 12.3. Error of no-pooling and partial pooling estimates, for the simulated tadpole ponds. The horizontal axis displays pond number. The vertical axis measures the absolute error in the predicted proportion of survivors, compared to the true value used in the simulation. The higher the point, the worse the estimate. No-pooling shown in blue. Partial pooling shown in black. The blue and dashed black lines show the average error for each kind of estimate, across each initial density of tadpoles (pond size). Smaller ponds produce more error, but the partial pooling estimates are better on average, especially in smaller ponds.

60. The vertical axis is the distance between the mean estimated probability of survival and the actual probability of survival. So points close to the bottom had low error, while those near the top had a large error, more than 20% off in some cases. The vertical lines divide the groups of ponds with different initial densities of tadpoles. And finally, the horizontal blue and black line segments show the average error of the no-pooling and partial pooling estimates, respectively, for each group of ponds with the same initial size.

The first thing to notice about this plot is that both kinds of estimates are much more accurate for larger ponds, on the right side. This arises because more data means better estimates, usually. In the small ponds, sample size is small, and neither kind of estimate can work magic. Therefore, prediction suffers on the left side of the plot. Second, note that the blue line is always above the black dashed line. This indicates that the no-pool estimates, shown by the blue points, have higher average error in each group of ponds. So even though both kinds of estimates get worse as sample size decreases, the varying effect estimates have the advantage, on average. Third, the distance between the blue line and the black dashed line grows as ponds get smaller. So while both kinds of estimates suffer from reduced sample size, the partial pooling estimates suffer less.

Okay, so what are we to make of all of this? Remember, back in FIGURE 12.1 (page 361), the smaller tanks demonstrated more shrinkage towards the mean. Here, the ponds with the smallest sample size show the greatest improvement over the naive no-pooling estimates. This is no coincidence. Shrinkage towards the mean results from trying to negotiate the underfitting and overfitting risks of the grand mean on one end and the individual means

of each pond on the other. The smaller tanks/ponds contain less information, and so their varying estimates are influenced more by the pooled information from the other ponds. In other words, small ponds are prone to overfitting, and so they receive a bigger dose of the underfit grand mean. Likewise, the larger ponds shrink much less, because they contain more information and are prone to less overfitting. Therefore they need less correcting. When individual ponds are very large, pooling in this way does hardly anything to improve estimates, because the estimates don't have far to go. But in that case, they also don't do any harm, and the information pooled from them can substantially help prediction in smaller ponds.

The partially pooled estimates are better on average. They adjust individual cluster (pond) estimates to negotiate the trade-off between underfitting and overfitting. This is a form of regularization, just like in Chapter 6, but now with an amount of regularization that is learned from the data itself.

But there are some cases in which the no-pooling estimates are better. These exceptions often result from ponds with extreme probabilities of survival. The partial pooling estimates shrink such extreme ponds towards the mean, because few ponds exhibit such extreme behavior. But sometimes outliers really are outliers.

Overthinking: Repeating the pond simulation. This model samples pretty quickly. Compiling the model takes up most of the execution time. Luckily the compilation only has to be done once. Then you can pass new data to the compiled model and get new estimates. Once you've compiled `m12.3` once, you can use this code to re-simulate ponds and sample from the new posterior, without waiting for the model to compile again:

R code 12.19

```
a <- 1.4
sigma <- 1.5
nponds <- 60
ni <- as.integer( rep( c(5,10,25,35) , each=15 ) )
a_pond <- rnorm( nponds , mean=a , sd=sigma )
dsim <- data.frame( pond=1:nponds , ni=ni , true_a=a_pond )
dsim$si <- rbinom( nponds,prob=logistic( dsim$true_a ),size=dsim$ni )
dsim$p_nopool <- dsim$si / dsim$ni
newdat <- list(si=dsim$si,ni=dsim$ni,pond=1:nponds)
m12.3new <- map2stan( m12.3 , data=newdat , iter=1e4 , warmup=1000 )
```

The `map2stan` function reuses the compiled model in `m12.3`, passes it the new data, and returns the new samples in `m12.3new`. This is a useful trick, in case you want to perform a simulation study of a particular model structure. And if you ever want to extract the actual compiled Stan model, it is held in `m12.3@stanfit`, and you can always view its code with `stancode(m12.3)` and the input data (which is augmented a bit) with `m12.3@data`.

12.3. More than one type of cluster

We can use and often should use more than one type of cluster in the same model. For example, the observations in `data(chimpanzees)`, which you met back in Chapter 10, are lever pulls. Each pull is within a cluster of pulls belonging to an individual chimpanzee. But each pull is also within an experimental block, which represents a collection of observations that happened on the same day. So each observed pull belongs to both an `actor` (1 to 7) and a `block` (1 to 6). There may be unique intercepts for each actor as well as for each block.

So in this section we'll reconsider the chimpanzees data, using both types of clusters simultaneously. This will allow us to use partial pooling on both categorical variables, `actor` and `block`, at the same time. We'll also get estimates of the variation among actors and among blocks.

Rethinking: Cross-classification and hierarchy. The kind of data structure in `data(chimpanzees)` is usually called a **cross-classified** multilevel model. It is cross-classified, because actors are not nested within unique blocks. If each chimpanzee had instead done all of his or her pulls on a single day, within a single block, then the data structure would instead be *hierarchical*. However, the model specification would typically be the same. So the model structure and code you'll see below will apply both to cross-classified designs and hierarchical designs. Other software sometimes forces you to treat these differently, on account of using a conditioning engine substantially less capable than MCMC. There are other types of "hierarchical" multilevel models, types that make adaptive priors for adaptive priors. It's turtles all the way down, recall (page 13). You'll see an example in the next chapter. But for the most part, people (or their software) nearly always use the same kind of model in both cases.

12.3.1. Multilevel chimpanzees. Let's proceed by taking the full chimpanzees model from Chapter 10 (`m10.4`, page 299) and first adding varying intercepts on actor. To add varying intercepts to this model, we just replace the fixed regularizing prior with an adaptive prior. But this time, I'll put the mean α up in the linear model, rather than down in the prior. Why? Because it will pave the way to adding more varying effects later. You'll see why, once we've pushed forward a little.

Here is the multilevel chimpanzees model in mathematical form, with the varying intercept components highlighted in blue:

$$\begin{aligned} L_i &\sim \text{Binomial}(1, p_i) \\ \text{logit}(p_i) &= \alpha + \alpha_{\text{ACTOR}[i]} + (\beta_P + \beta_{PC} C_i) P_i \\ \alpha_{\text{ACTOR}} &\sim \text{Normal}(0, \sigma_{\text{ACTOR}}) \\ \alpha &\sim \text{Normal}(0, 10) \\ \beta_P &\sim \text{Normal}(0, 10) \\ \beta_{PC} &\sim \text{Normal}(0, 10) \\ \sigma_{\text{ACTOR}} &\sim \text{HalfCauchy}(0, 1) \end{aligned}$$

Notice that α is inside the linear model, not inside the Gaussian prior for α_{ACTOR}. This is mathematically equivalent to what you did with the tadpoles earlier in the chapter. You can always take the mean out of a Gaussian distribution and treat the distribution as a constant plus a Gaussian distribution centered on zero.

This might seem a little weird at first, so it might help train your intuition by experimenting in R. These two lines of code sample values from two identical Gaussian distributions, with mean 10 and standard deviation 1:

R code 12.20

```
y1 <- rnorm( 1e4 , 10 , 1 )
y2 <- 10 + rnorm( 1e4 , 0 , 1 )
```

Inspect the distributions of values in `y1` and `y2`. You'll see they are the same. This feature of the Gaussian distribution arises from the independence of the mean and standard deviation. Most distributions do not have this property. But we'll exploit it here. And sometimes a given combination of model and data is more efficiently fit using one form or the other. I'll say more about this in the next chapter.

Here's the corresponding `map2stan` code for the model with varying intercepts on `actor`, but not yet on `block`. Note that the linear model contains α, the varying intercepts mean. The adaptive prior for the intercepts themselves has a mean of zero.

R code 12.21

```
library(rethinking)
data(chimpanzees)
d <- chimpanzees
d$recipient <- NULL     # get rid of NAs

m12.4 <- map2stan(
    alist(
        pulled_left ~ dbinom( 1 , p ) ,
        logit(p) <- a + a_actor[actor] + (bp + bpC*condition)*prosoc_left ,
        a_actor[actor] ~ dnorm( 0 , sigma_actor ),
        a ~ dnorm(0,10),
        bp ~ dnorm(0,10),
        bpC ~ dnorm(0,10),
        sigma_actor ~ dcauchy(0,1)
    ) ,
    data=d , warmup=1000 , iter=5000 , chains=4 , cores=3 )
```

Inspect the trace plot, `plot(m12.4)`, and the posterior distribution for `sigma_actor`. Make sure the effective numbers of samples and `Rhat` values look alright. If you need to review these MCMC diagnostics, glance back at Chapter 8.

Now that the mean of the population of actors, α (`a`), is in the linear model, it's important to notice now that the `a_actor` parameters are *deviations* from `a`. So for any given row *i*, the total intercept is $\alpha + \alpha_{\text{ACTOR}[i]}$. The part that varies across actors is just the deviation from the grand mean α. To compute the total intercept for each `actor`, you need to add samples of `a` to samples of `a_actor`:

R code 12.22

```
post <- extract.samples(m12.4)
total_a_actor <- sapply( 1:7 , function(actor) post$a + post$a_actor[,actor] )
round( apply(total_a_actor,2,mean) , 2 )
```

```
[1] -0.71  4.59 -1.02 -1.02 -0.71  0.23  1.76
```

12.3.2. Two types of cluster. To add the second cluster type, `block`, we merely replicate the structure for the `actor` cluster. This means the linear model gets yet another varying intercept, $\alpha_{\text{BLOCK}[i]}$, and the model gets another adaptive prior and yet another standard deviation parameter. Here is the mathematical form of the model, with the new pieces of the machine

highlighted in blue:

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} + (\beta_P + \beta_{PC} C_i) P_i \\
\alpha_{\text{ACTOR}} &\sim \text{Normal}(0, \sigma_{\text{ACTOR}}) \\
\alpha_{\text{BLOCK}} &\sim \text{Normal}(0, \sigma_{\text{BLOCK}}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_P &\sim \text{Normal}(0, 10) \\
\beta_{PC} &\sim \text{Normal}(0, 10) \\
\sigma_{\text{ACTOR}} &\sim \text{HalfCauchy}(0, 1) \\
\sigma_{\text{BLOCK}} &\sim \text{HalfCauchy}(0, 1)
\end{aligned}
$$

Each cluster variable needs its own standard deviation parameter that adapts the amount of pooling across units, be they actors or blocks. These are σ_{ACTOR} and σ_{BLOCK}, respectively. Finally, note that there is only one global mean parameter α, and both of the varying intercept parameters are centered at zero. We can't identify a separate mean for each varying intercept type, because both intercepts are added to the same linear prediction. So it is conventional to define varying intercepts with a mean of zero, so there's no risk of accidentally creating hard-to-identify parameters. There's a practice problem at the end of the chapter that leads you to explore what happens when you forget and instead include two grand mean α parameters.

Now to fit the model that uses both `actor` and `block`:

R code 12.23

```
# prep data
d$block_id <- d$block  # name 'block' is reserved by Stan

m12.5 <- map2stan(
    alist(
        pulled_left ~ dbinom( 1 , p ),
        logit(p) <- a + a_actor[actor] + a_block[block_id] +
                    (bp + bpc*condition)*prosoc_left,
        a_actor[actor] ~ dnorm( 0 , sigma_actor ),
        a_block[block_id] ~ dnorm( 0 , sigma_block ),
        c(a,bp,bpc) ~ dnorm(0,10),
        sigma_actor ~ dcauchy(0,1),
        sigma_block ~ dcauchy(0,1)
    ) ,
    data=d, warmup=1000 , iter=6000 , chains=4 , cores=3 )
```

If all goes well, you'll end up with 20,000 samples from 4 independent chains. As always, be sure to inspect the trace plots and the diagnostics. As soon as you start trusting the machine, the machine will betray your trust. In this case, you might see for the first time a warning about *divergent iterations*:

```
Warning message:
In map2stan(alist(pulled_left ~ dbinom(1, p), logit(p) <- a + a_actor[actor] +  :
  There were 3 divergent iterations during sampling.
Check the chains (trace plots, n_eff, Rhat) carefully to ensure they are valid.
```

We'll have a lot more to say about these in the next chapter. For now, they are safe to ignore. Just do as stated and inspect `n_eff` and `Rhat`.

This is easily the most complicated model we've fit in the book so far. So let's look at the estimates and take note of a few important features:

R code 12.24

```
precis(m12.5,depth=2) # depth=2 displays varying effects
plot(precis(m12.5,depth=2)) # also plot
```

```
              Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a_actor[1]  -1.17   0.93      -2.57       0.29  2333    1
a_actor[2]   4.14   1.59       1.92       6.33  4543    1
a_actor[3]  -1.48   0.94      -2.91      -0.02  2310    1
a_actor[4]  -1.48   0.94      -2.92      -0.01  2360    1
a_actor[5]  -1.17   0.94      -2.63       0.27  2354    1
a_actor[6]  -0.22   0.93      -1.65       1.25  2359    1
a_actor[7]   1.32   0.96      -0.15       2.84  2496    1
a_block[1]  -0.18   0.22      -0.53       0.11  3848    1
a_block[2]   0.04   0.18      -0.23       0.34  8595    1
a_block[3]   0.05   0.19      -0.23       0.35  7237    1
a_block[4]   0.00   0.18      -0.30       0.28  9532    1
a_block[5]  -0.04   0.18      -0.34       0.25  8327    1
a_block[6]   0.11   0.20      -0.17       0.43  5420    1
a            0.46   0.92      -0.98       1.88  2263    1
bp           0.82   0.26       0.41       1.25  6890    1
bpc         -0.13   0.30      -0.62       0.33  8360    1
sigma_actor  2.25   0.90       1.02       3.33  4892    1
sigma_block  0.22   0.18       0.01       0.43  2079    1
```

The `precis` plot is shown in the left-hand part of FIGURE 12.4.

First, notice that the number of effective samples, `n_eff`, varies quite a lot across parameters. This is common in complex models. Why? There are many reasons for this. But in this sort of model the most common reason is that some parameter spends a lot of time near a boundary. Here, that parameter is `sigma_block`. It spends a lot of time near its minimum of zero. As a consequence, you may also see a warning about "divergent iterations." You can wait until the next chapter to explore what these mean and what to do about them. For now, you can trust the `Rhat` values above, but later you'll see how to make sampling more efficient for models like these.

Second, compare `sigma_actor` to `sigma_block` and notice that the estimated variation among actors is a lot larger than the estimated variation among blocks. This is easy to appreciate, if we plot the marginal posterior distributions of these two parameters:

R code 12.25

```
post <- extract.samples(m12.5)
dens( post$sigma_block , xlab="sigma" , xlim=c(0,4) )
dens( post$sigma_actor , col=rangi2 , lwd=2 , add=TRUE )
text( 2 , 0.85 , "actor" , col=rangi2 )
text( 0.75 , 2 , "block" )
```

And this plot appears on the right in FIGURE 12.4. While there's uncertainty about the variation among actors, this model is confident that actors vary more than blocks. You can easily

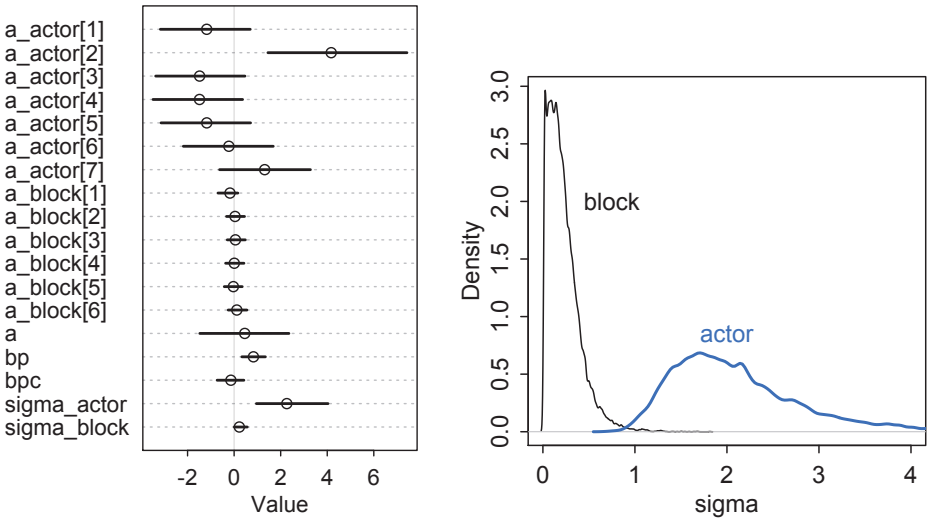

FIGURE 12.4. Left: Posterior means and 89% highest density intervals for `m12.5`. The greater variation across actors than blocks can be seen immediately in the `a_actor` and `a_block` distributions. Right: Posterior distributions of the standard deviations of varying intercepts by actor (blue) and experimental block (black).

see this variation in the varying intercept estimates: the `a_actor` distributions are much more scattered than are the `a_block` distributions.

As a consequence, adding `block` to this model hasn't added a lot of overfitting risk. Let's compare the model with only varying intercepts on `actor` to the model with both kinds of varying intercepts:

```
compare(m12.4,m12.5)
```

R code 12.26

```
       WAIC pWAIC dWAIC weight    SE  dSE
m12.4 531.5   8.1   0.0   0.65 19.50   NA
m12.5 532.7  10.5   1.2   0.35 19.74 1.94
```

Look at the `pWAIC` column, which reports the "effective number of parameters." While `m12.5` has 7 more parameters than `m12.4` does, it has only about 2.5 more effective parameters. Why? Because the posterior distribution for `sigma_block` ended up close to zero. This means each of the 6 `a_block` parameters is strongly shrunk towards zero—they are relatively inflexible. In contrast, the `a_actor` parameters are shrunk towards zero much less, because the estimated variation across actors is much larger, resulting in less shrinkage. But as a consequence, each of the `a_actor` parameters contributes much more to the `pWAIC` value.

You might also notice that the difference in WAIC between these models is small, only 1.2. This is especially small compared the standard deviation of the difference, 1.94. These two models imply nearly identical predictions, and so their expected out-of-sample accuracy is nearly identical. The block parameters have been shrunk so much towards zero that they do very little work in the model.

If you are feeling the urge to "select" `m12.4` as the best model, pause for a moment. There is nothing to gain here by selecting either model. The comparison of the two models tells a richer story—whether we include block or not hardly matters, and the `a_block` and `sigma_block` estimates tell us why. Furthermore, the standard error of the difference in WAIC between the models is twice as large as the difference itself. By retaining and reporting both models, we and our readers learn more about the experiment.

12.3.3. Even more clusters. Adding more types of clusters proceeds the same way. At some point the model may become too complex to reliably fit to data. But Hamiltonian Monte Carlo is very capable with varying effects. It can easily handle tens of thousands of varying effect parameters. Sampling will be slow in such cases, but it will work.

So don't be shy—if you have a good theoretical reason to include a cluster variable, then you also have good theoretical reason to partially pool its parameters. As you've seen, the overfitting risk induced by including varying intercepts can be quite small, when there is little variation among the clusters. The multilevel model adaptively regularizes, helping us discover the relevance of different kinds of clusters within the data. In this way, you can think of the `sigma` parameter for each cluster as a crude measure of the cluster's relevance for explaining variation in the outcome.

12.4. Multilevel posterior predictions

Way back in Chapter 3 (page 64), I commented on the importance of **MODEL CHECKING**. Software does not always work as expected, and one robust way to discover mistakes is to compare the sample to the posterior predictions of a fit model. The same procedure, producing implied predictions from a fit model, is very helpful for understanding what the model means. Every model is a merger of sense and nonsense. When we understand a model, we can find its sense and control its nonsense. But as models get more complex, it is very difficult to impossible to understand them just by inspecting tables of posterior means and intervals. Exploring implied posterior predictions helps much more.

Another role for constructing implied predictions is in computing **INFORMATION CRITERIA**, like DIC and WAIC. These criteria provide simple estimates of out-of-sample model accuracy, the KL divergence. In practical terms, information criteria provide a rough measure of a model's flexibility and therefore overfitting risk. This was the big conceptual mission of Chapter 6.

All of this advice applies to multilevel models as well. We still often need model checks, counterfactual predictions for understanding, and information criteria. The introduction of varying effects does introduce nuance, however.

First, we should no longer expect the model to exactly retrodict the sample, because adaptive regularization has as its goal to trade off poorer fit in sample for better inference and hopefully better fit out of sample. That is what shrinkage does for us. Of course, we should never be trying to really retrodict the sample. But now you have to expect that even a perfectly good model fit will differ from the raw data in a systematic way that reflects shrinkage.

Second, "prediction" in the context of a multilevel model requires additional choices. If we wish to validate a model against the specific clusters used to fit the model, that is one thing. But if we instead wish to compute predictions for new clusters, other than the ones observed in the sample, that is quite another. We'll consider each of these in turn, continuing to use the chimpanzees model from the previous section.

12.4.1. Posterior prediction for same clusters. When working with the same clusters as you used to fit a model, varying intercepts are just parameters. The only trick is to ensure that you use the right intercept for each case in the data. If you use `link` and `sim` to do your work for you, this is handled automatically. But otherwise, there are no tricks.

For example, in `data(chimpanzees)`, there are 7 unique actors. These are the clusters. The varying intercepts model, `m12.4`, estimated an intercept for each, in addition to two parameters to describe the mean and standard deviation of the population of actors. We'll construct posterior predictions (retrodictions), using both the automated `link` approach and doing it from scratch, so there is no confusion.

Before computing predictions, note that we should no longer expect the posterior predictive distribution to match the raw data, even when the model worked correctly. Why? The whole point of partial pooling is to shrink estimates towards the grand mean. So the estimates should not necessarily match up with the raw data, once you use pooling.

The code needed to compute posterior predictions is just like the code from Chapter 10. Here it is again, computing and plotting posterior predictions for actor number 2:

R code 12.27

```
chimp <- 2
d.pred <- list(
    prosoc_left = c(0,1,0,1),   # right/left/right/left
    condition = c(0,0,1,1),     # control/control/partner/partner
    actor = rep(chimp,4)
)
link.m12.4 <- link( m12.4 , data=d.pred )
pred.p <- apply( link.m12.4 , 2 , mean )
pred.p.PI <- apply( link.m12.4 , 2 , PI )
```

And the plotting code is exactly the same as before (page 297).

To construct the same calculations without using `link`, we just have to remember the model. The only difficulty is that when we work with the samples from the posterior, the varying intercepts will be a matrix of samples. Let's take a look:

R code 12.28

```
post <- extract.samples(m12.4)
str(post)
```

```
List of 5
 $ a_actor    : num [1:8000, 1:7] -1.842 -0.225 -1.811 -0.759 -1.882 ...
 $ a          : num [1:8000(1d)] 1.291 -0.632 0.285 -0.109 1.229 ...
 $ bp         : num [1:8000(1d)] 1 1.064 1.087 0.254 0.908 ...
 $ bpC        : num [1:8000(1d)] -0.272 -0.539 -0.295 0.375 -0.218 ...
 $ sigma_actor: num [1:8000(1d)] 2.13 2.49 2.32 1.51 4.12 ...
```

The `a_actor` matrix has samples on the rows and actors on the columns. So to plot, for example, the density for actor 5:

R code 12.29

```
dens( post$a_actor[,5] )
```

The `[,5]` means "all samples for actor 5."

To construct posterior predictions, we build our own link function. I'll use the `with` function here, so we don't have to keep typing `post$` before every parameter name:

R code 12.30

```
p.link <- function( prosoc_left , condition , actor ) {
    logodds <- with( post ,
        a + a_actor[,actor] + (bp + bpC * condition) * prosoc_left
    )
    return( logistic(logodds) )
}
```

The linear model is identical to the one used to define the model, but with a single comma added inside the brackets after `a_actor`. Now to compute predictions:

R code 12.31

```
prosoc_left <- c(0,1,0,1)
condition <- c(0,0,1,1)
pred.raw <- sapply( 1:4 , function(i) p.link(prosoc_left[i],condition[i],2) )
pred.p <- apply( pred.raw , 2 , mean )
pred.p.PI <- apply( pred.raw , 2 , PI )
```

At some point, you will have to work with a model that `link` will mangle. At that time, you can return to this section and peer hard at the code above and still make progress. No matter what the model is, if it is a Bayesian model, then it is *generative*. This means that predictions are made by pushing samples up through the model to get distributions of predictions. Then you summarize the distributions to summarize the predictions.

12.4.2. Posterior prediction for new clusters. Often, the particular clusters in the sample are not of any enduring interest. In the chimpanzees data, for example, these particular 7 chimpanzees are just seven individuals. We'd like to make inferences about the whole species, not just those seven individuals. So the individual actor intercepts aren't of interest, but the distribution of them definitely is.

One way to grasp the task of construction posterior predictions for new clusters is to imagine leaving out one of the clusters when you fit the model to the data. For example, suppose we leave out actor number 7 when we fit the chimpanzees model. Now how can we assess the model's accuracy for predicting actor number 7's behavior? We can't use any of the `a_actor` parameter estimates, because those apply to other individuals. But we can make good use of the `a` and `sigma_actor` parameters, because those describe the population of actors.

First, let's see how to construct posterior predictions for a now, previously unobserved *average* actor. By "average," I mean an individual chimpanzee with an intercept exactly at `a` (α), the population mean. This simultaneously implies a varying intercept of zero. Since there is uncertainty about the population mean, there is still uncertainty about this average individual's intercept. But as you'll see, the uncertainty is much smaller than it really should be, if we wish to honestly represent the problem of what to expect from a new individual.

The first step is to make a new data list to compute predictions over. You've done this in previous chapters. Here is our new list, representing the four different treatments:

R code 12.32

```
d.pred <- list(
    prosoc_left = c(0,1,0,1),   # right/left/right/left
    condition = c(0,0,1,1),     # control/control/partner/partner
    actor = rep(2,4) )          # placeholder
```

Next, we're going to make a matrix of zeros, to replace the varying intercept samples. It's easiest to just keep the same dimension as the original matrix. In this case that means using 1000 samples for each of 7 actors. But all of the samples will be set to zero:

R code 12.33

```
# replace varying intercept samples with zeros
# 1000 samples by 7 actors
a_actor_zeros <- matrix(0,1000,7)
```

That's the only new trick. Now we just pass this new matrix to `link` using the optional `replace` argument. Make sure the new matrix is named the same as the varying intercept matrix, `a_actor`. Otherwise it won't replace anything that appears in the model.

R code 12.34

```
# fire up link
# note use of replace list
link.m12.4 <- link( m12.4 , n=1000 , data=d.pred ,
    replace=list(a_actor=a_actor_zeros) )

# summarize and plot
pred.p.mean <- apply( link.m12.4 , 2 , mean )
pred.p.PI <- apply( link.m12.4 , 2 , PI , prob=0.8 )
plot( 0 , 0 , type="n" , xlab="prosoc_left/condition" ,
    ylab="proportion pulled left" , ylim=c(0,1) , xaxt="n" ,
    xlim=c(1,4) )
axis( 1 , at=1:4 , labels=c("0/0","1/0","0/1","1/1") )
lines( 1:4 , pred.p.mean )
shade( pred.p.PI , 1:4 )
```

The result is displayed in FIGURE 12.5, on the left. The gray region shows the 80% interval for an actor with an average intercept. This kind of calculation makes it easy to see the impact of `prosoc_left`, as well as uncertainty about where the average is, but it doesn't show the variation among actors.

To show the variation among actors, we'll need to use `sigma_actor` in the calculation. We can again smuggle this into `link` by using the `replace` argument. This time however, we'll simulate a matrix of new varying intercepts from a Gaussian distribution defined by the adaptive prior in the model itself:

$$\alpha_{\text{ACTOR}} \sim \text{Normal}(0, \sigma_{\text{ACTOR}})$$

This implies that once we have samples for σ_{ACTOR}, we can simulate new actor intercepts from this distribution. Here's the code to do just that, using `rnorm`:

R code 12.35

```
# replace varying intercept samples with simulations
post <- extract.samples(m12.4)
a_actor_sims <- rnorm(7000,0,post$sigma_actor)
a_actor_sims <- matrix(a_actor_sims,1000,7)
```

Now pass the simulated intercepts into `link`. Note the `replace` list, which inserts the simulations into the posterior.

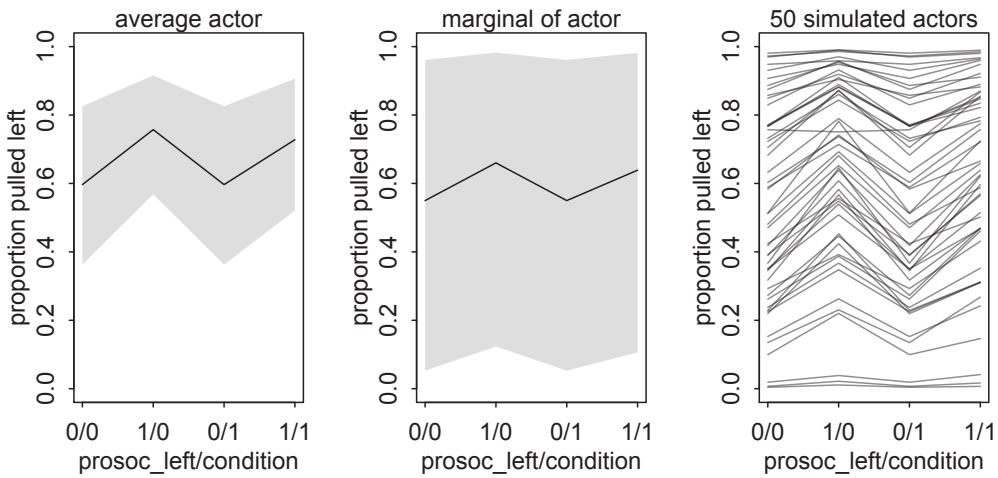

FIGURE 12.5. Posterior predictive distributions for the chimpanzees varying intercept model, `m12.4`. The solid lines are posterior means and the shaded regions are 80% percentile intervals. Left: Setting the varying intercept `a_actor` to zero produces predictions for an *average* actor. These predictions ignore uncertainty arising from variation among actors. Middle: Simulating varying intercepts using the posterior standard deviation among actors, `sigma_actor`, produces predictions that account for variation among actors. Right: 50 simulated actors with unique intercepts sampled from the posterior. Each simulation maintains the same parameter values across all four treatments.

R code 12.36

```
link.m12.4 <- link( m12.4 , n=1000 , data=d.pred ,
    replace=list(a_actor=a_actor_sims) )
```

Summarizing and plotting is exactly as before, and the result is displayed in the middle of FIGURE 12.5. These posterior predictions are *marginal* of actor, which means that they average over the uncertainty among actors. In contrast, the predictions on the left just set the actor to the average, ignoring variation among actors.

At this point, students usually ask, "So which one should I use?" The answer is, "It depends." Both are useful, depending upon the question. The predictions for an average actor help to visualize the impact of treatment. The predictions that are marginal of actor illustrate how variable different chimpanzees are, according to the model. You probably want to compute both for yourself, when trying to understand a model. But which you include in a report will depend upon context.

In this case, we can do better by making a plot that displays both the treatment effect and the variation among actors. We can do this by forgetting about intervals and instead simulating a series of new actors in each of the four treatments. By drawing a line for each actor across all four treatments, we'll be able to visualize both the zig-zag impact of `prosoc_left` as well as the variation among individuals.

What we'll do now is write a new function that simulates a new actor from the estimated population of actors and then computes probabilities of pulling the left lever for each of the four treatments. These simulations will not average over uncertainty in the posterior. We'll get that uncertainty into the plot by using multiple simulations, each with a different sample from the posterior. Here's the function:

R code 12.37

```
post <- extract.samples(m12.4)
sim.actor <- function(i) {
    sim_a_actor <- rnorm( 1 , 0 , post$sigma_actor[i] )
    P <- c(0,1,0,1)
    C <- c(0,0,1,1)
    p <- logistic(
        post$a[i] +
        sim_a_actor +
        (post$bp[i] + post$bpC[i]*C)*P
    )
    return(p)
}
```

This function takes a single argument, `i`, which is just the index of a sample from the posterior distribution. It then draws a random intercept for the actor, using `rnorm` and a particular value of `sigma_actor`. Then it computes probabilities `p` for each of the four treatments, using the same linear model, but with different predictor values inside the `P` and `C` vectors. Because these vectors are of length 4, the code spits out 4 values for `p`.

Now to use this function to plot 50 simulations:

R code 12.38

```
# empty plot
plot( 0 , 0 , type="n" , xlab="prosoc_left/condition" ,
    ylab="proportion pulled left" , ylim=c(0,1) , xaxt="n" , xlim=c(1,4) )
axis( 1 , at=1:4 , labels=c("0/0","1/0","0/1","1/1") )

# plot 50 simulated actors
for ( i in 1:50 ) lines( 1:4 , sim.actor(i) , col=col.alpha("black",0.5) )
```

The result is shown in the right-hand plot of FIGURE 12.5. Each trend is a simulated actor, across all four treatments on the horizontal axis. It is much easier in this plot to see both the zig-zag impact of treatment and the variation among actors that is induced by the posterior distribution of `sigma_actor`.

Also note the interaction of treatment and the variation among actors. Because this is a binomial model, in principle all parameters interact, due to ceiling and floor effects. For actors with very large intercepts, near the top of the plot, treatment has very little effect. These actors have strong handedness preferences. But actors with intercepts nearer the mean are influenced by treatment.

12.4.3. Focus and multilevel prediction. All of this is confusing at first. There is no uniquely correct way to always construct the predictions, and the calculations themselves probably seem a little magical. In time, it makes a lot more sense. The fact is that multilevel models contain parameters with different **FOCUS**. Focus here means which level of the model the

parameter makes direct predictions for. It helps to organize the issue into three common cases.

First, when retrodicting the sample, the parameters that describe the population of clusters, such as α and σ_{ACTOR} in `m12.4`, do not influence prediction directly. Recall that these population parameters are often called **HYPERPARAMETERS**, as they are parameters for parameters. These hyperparameters had their effects during estimation, by shrinking the varying effect parameters towards a common mean. The prediction focus here is on the top level of parameters, not the deeper hyperparameters.

Second, the same is true when forecasting a new observation for a cluster that was present in the sample. For example, if we want to predict what chimpanzee number 2 will do in the next experiment, we should probably bet she'll pull the left lever, because her varying intercept was very large. The focus is again on the top level.

The third case is different. When instead we wish to forecast for some new cluster that was not present in the sample, such as a new individual or school or year or location, then we need the hyper-parameters. The hyper-parameters tell us how to forecast a new cluster, by generating a distribution of new per-cluster intercepts. This is what we did in the previous section, simulating new chimpanzees.

This is also the right thing to do whenever varying effects are used to model **OVER-DISPERSION** (page 363). In that case, we need to simulate intercepts in order to account for the over-dispersion. Here's a quick example, using the Oceanic societies example from Chapter 10, but now adding a varying intercept to each society. Here's the mathematical form of the model, with the varying intercept pieces highlighted in blue:

$$\begin{aligned} T_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \alpha + \alpha_{\text{SOCIETY}[i]} + \beta_P \log P_i \\ \alpha &\sim \text{Normal}(0, 10) \\ \beta_P &\sim \text{Normal}(0, 1) \\ \alpha_{\text{SOCIETY}} &\sim \text{Normal}(0, \sigma_{\text{SOCIETY}}) \\ \sigma_{\text{SOCIETY}} &\sim \text{HalfCauchy}(0, 1) \end{aligned}$$

T is `total_tools`, P is `population`, and i indexes each society. The above is just a varying intercept model, but with a varying intercept for every observation. As a result, σ_{SOCIETY} ends up being an estimate of the over-dispersion among societies. Another way to think of this is that the varying intercepts α_{SOCIETY} are residuals for each society. By also estimating the distribution of these residuals, we get an estimate of the excess variation, relative to the Poisson expectation.

And here is the code to fit the over-dispersed Poisson model:

R code 12.39

```
# prep data
library(rethinking)
data(Kline)
d <- Kline
d$logpop <- log(d$population)
d$society <- 1:10

# fit model
m12.6 <- map2stan(
```

```
    alist(
        total_tools ~ dpois(mu),
        log(mu) <- a + a_society[society] + bp*logpop,
        a ~ dnorm(0,10),
        bp ~ dnorm(0,1),
        a_society[society] ~ dnorm(0,sigma_society),
        sigma_society ~ dcauchy(0,1)
    ),
    data=d ,
    iter=4000 , chains=3 )
```

This model samples very efficiently, despite using 13 parameters to describe 10 observations. Remember: Varying effect parameters are adaptively regularized. So they are not completely flexible and induce much less overfitting risk. In this case, WAIC should tell you that the effective number of parameters is about 5, not 13. If you have hundreds or thousands of observations in the data, this approach still works fine. You just end up with hundreds or thousands of varying intercept estimates. You won't care about the estimates themselves. But you will care about the hyperparameters that describe the population of varying intercepts.

Now to generate posterior predictions that visualize the over-dispersion. You can display posterior predictions (retrodictions) by using `postcheck(m12.6)`. But those predictions just use the varying intercepts, `a_society`, directly. They do not use the hyper-parameters. To instead see the general trend that the model expects, we'll need to simulate counterfactual societies, using the hyper-parameters α and σ_{SOCIETY}. This is the same procedure that we used for new chimpanzee actors earlier.

R code 12.40

```
post <- extract.samples(m12.6)
d.pred <- list(
    logpop = seq(from=6,to=14,length.out=30),
    society = rep(1,30)
)
a_society_sims <- rnorm(20000,0,post$sigma_society)
a_society_sims <- matrix(a_society_sims,2000,10)
link.m12.6 <- link( m12.6 , n=2000 , data=d.pred ,
    replace=list(a_society=a_society_sims) )
```

And this code will display the raw data and the new prediction envelope:

R code 12.41

```
# plot raw data
plot( d$logpop , d$total_tools , col=rangi2 , pch=16 ,
    xlab="log population" , ylab="total tools" )

# plot posterior median
mu.median <- apply( link.m12.6 , 2 , median )
lines( d.pred$logpop , mu.median )

# plot 97%, 89%, and 67% intervals (all prime numbers)
mu.PI <- apply( link.m12.6 , 2 , PI , prob=0.97 )
shade( mu.PI , d.pred$logpop )
```

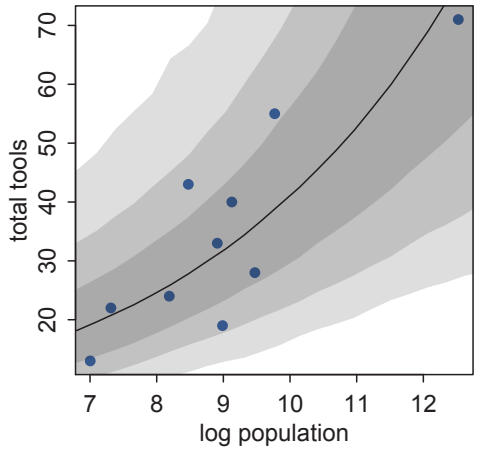

FIGURE 12.6. Posterior predictions for the over-dispersed Poisson island model, `m12.6`. The shaded regions are, inside to out: 67%, 89%, and 97% intervals of the expected mean. Marginalizing over the varying intercepts results in a much wider prediction region than we'd expect under a pure Poisson process.

```
mu.PI <- apply( link.m12.6 , 2 , PI , prob=0.89 )
shade( mu.PI , d.pred$logpop )
mu.PI <- apply( link.m12.6 , 2 , PI , prob=0.67 )
shade( mu.PI , d.pred$logpop )
```

The result is displayed in FIGURE 12.6. The envelope of predictions is a lot wider here than it was back in Chapter 10. This is a consequence of the varying intercepts, combined with the fact that there is much more variation in the data than a pure-Poisson model anticipates.

12.5. Summary

This chapter has been an introduction to the motivation, implementation, and interpretation of basic multilevel models. It focused on varying intercepts, which achieve better estimates of baseline differences among clusters in the data. They achieve better estimates, because they simultaneously model the population of clusters and use inferences about the population to pool information among parameters. From another perspective, varying intercepts are adaptively regularized parameters, relying upon a prior that is itself learned from the data. All of this is a foundation for the next chapter, which extends these concepts to additional types of parameters and models.

12.6. Practice

Easy.

12E1. Which of the following priors will produce more *shrinkage* in the estimates? (a) $\alpha_{\text{TANK}} \sim \text{Normal}(0, 1)$; (b) $\alpha_{\text{TANK}} \sim \text{Normal}(0, 2)$.

12E2. Make the following model into a multilevel model.

$$\begin{aligned} y_i &\sim \text{Binomial}(1, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{GROUP}[i]} + \beta x_i \\ \alpha_{\text{GROUP}} &\sim \text{Normal}(0, 10) \\ \beta &\sim \text{Normal}(0, 1) \end{aligned}$$

12E3. Make the following model into a multilevel model.

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha_{\text{GROUP}[i]} + \beta x_i \\ \alpha_{\text{GROUP}} &\sim \text{Normal}(0, 10) \\ \beta &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{HalfCauchy}(0, 2) \end{aligned}$$

12E4. Write an example mathematical model formula for a Poisson regression with varying intercepts.

12E5. Write an example mathematical model formula for a Poisson regression with two different kinds of varying intercepts, a cross-classified model.

Medium.

12M1. Revisit the Reed frog survival data, `data(reedfrogs)`, and add the `predation` and `size` treatment variables to the varying intercepts model. Consider models with either main effect alone, both main effects, as well as a model including both and their interaction. Instead of focusing on inferences about these two predictor variables, focus on the inferred variation across tanks. Explain why it changes as it does across models.

12M2. Compare the models you fit just above, using WAIC. Can you reconcile the differences in WAIC with the posterior distributions of the models?

12M3. Re-estimate the basic Reed frog varying intercept model, but now using a Cauchy distribution in place of the Gaussian distribution for the varying intercepts. That is, fit this model:

$$\begin{aligned} s_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{TANK}[i]} \\ \alpha_{\text{TANK}} &\sim \text{Cauchy}(\alpha, \sigma) \\ \alpha &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{HalfCauchy}(0, 1) \end{aligned}$$

Compare the posterior means of the intercepts, α_{TANK}, to the posterior means produced in the chapter, using the customary Gaussian prior. Can you explain the pattern of differences?

12M4. Fit the following cross-classified multilevel model to the `chimpanzees` data:

$$\begin{aligned} L_i &\sim \text{Binomial}(1, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} + (\beta_P + \beta_{PC} C_i) P_i \\ \alpha_{\text{ACTOR}} &\sim \text{Normal}(\alpha, \sigma_{\text{ACTOR}}) \\ \alpha_{\text{BLOCK}} &\sim \text{Normal}(\gamma, \sigma_{\text{BLOCK}}) \\ \alpha, \gamma, \beta_P, \beta_{PC} &\sim \text{Normal}(0, 10) \\ \sigma_{\text{ACTOR}}, \sigma_{\text{BLOCK}} &\sim \text{HalfCauchy}(0, 1) \end{aligned}$$

Each of the parameters in those comma-separated lists gets the same independent prior. Compare the posterior distribution to that produced by the similar cross-classified model from the chapter. Also compare the number of effective samples. Can you explain the differences?

Hard.

12H1. In 1980, a typical Bengali woman could have 5 or more children in her lifetime. By the year 200, a typical Bengali woman had only 2 or 3. You're going to look at a historical set of data, when contraception was widely available but many families chose not to use it. These data reside in `data(bangladesh)` and come from the 1988 Bangladesh Fertility Survey. Each row is one of 1934 women. There are six variables, but you can focus on three of them for this practice problem:

(1) `district`: ID number of administrative district each woman resided in
(2) `use.contraception`: An indicator (0/1) of whether the woman was using contraception
(3) `urban`: An indicator (0/1) of whether the woman lived in a city, as opposed to living in a rural area

The first thing to do is ensure that the cluster variable, `district`, is a contiguous set of integers. Recall that these values will be index values inside the model. If there are gaps, you'll have parameters for which there is no data to inform them. Worse, the model probably won't run. Look at the unique values of the `district` variable:

R code 12.42
```
sort(unique(d$district))
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 55 56 57 58 59 60 61
```

District 54 is absent. So `district` isn't yet a good index variable, because it's not contiguous. This is easy to fix. Just make a new variable that is contiguous. This is enough to do it:

R code 12.43
```
d$district_id <- as.integer(as.factor(d$district))
sort(unique(d$district_id))
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
[51] 51 52 53 54 55 56 57 58 59 60
```

Now there are 60 values, contiguous integers 1 to 60.

Now, focus on predicting `use.contraception`, clustered by `district_id`. Do not include `urban` just yet. Fit both (1) a traditional fixed-effects model that uses dummy variables for district and (2) a multilevel model with varying intercepts for district. Plot the predicted proportions of women in each district using contraception, for both the fixed-effects model and the varying-effects model. That is, make a plot in which district ID is on the horizontal axis and expected proportion using contraception is on the vertical. Make one plot for each model, or layer them on the same plot, as you prefer. How do the models disagree? Can you explain the pattern of disagreement? In particular, can you explain the most extreme cases of disagreement, both why they happen where they do and why the models reach different inferences?

12H2. Return to the Trolley data, `data(Trolley)`, from Chapter 11. Define and fit a varying intercepts model for these data. Cluster intercepts on individual participants, as indicated by the unique values in the `id` variable. Include `action`, `intention`, and `contact` as ordinary terms. Compare the varying intercepts model and a model that ignores individuals, using both WAIC and posterior predictions. What is the impact of individual variation in these data?

12H3. The Trolley data are also clustered by `story`, which indicates a unique narrative for each vignette. Define and fit a cross-classified varying intercepts model with both `id` and `story`. Use the same ordinary terms as in the previous problem. Compare this model to the previous models. What do you infer about the impact of different stories on responses?

Endnotes

Chapter 1

1. I draw this metaphor from Collins and Pinch (1998), *The Golem: What You Should Know about Science*. It is very similar to E. T. Jaynes' 2003 metaphor of statistical models as robots, although with a less precise and more monstrous implication. [1]

2. There are probably no algorithms nor machines that never break, bend, or malfunction. A common citation for this observation is Wittgenstein (1953), *Philosophical Investigations*, section 193. Malfunction will interest us, later in the book, when we consider more complex models and the procedures needed to fit them to data. [2]

3. See Mulkay and Gilbert (1981). I sometimes teach a PhD core course that includes some philosophy of science, and PhD students are nearly all shocked by how little their casual philosophy resembles that of Popper or any other philosopher of science. The first half of Ian Hacking's *Representing and Intervening* (1983) is probably the quickest way into the history of the philosophy of science. It's getting out of date, but remains readable and broad minded. [4]

4. Maybe best to begin with Popper's last book, *The Myth of the Framework* (1996). I also recommend interested readers to go straight to a modern translation of Popper's earlier *Logic of Scientific Discovery*. Chapters 6, 8, 9 and 10 in particular demonstrate that Popper appreciated the difficulties with describing science as an exercise in falsification. Other later writings, many collected in *Objective knowledge: An evolutionary approach*, show that Popper viewed the generation of scientific knowledge as an evolutionary process that admits many different methods. [4]

5. Meehl (1967) observed that this leads to a methodological paradox, as improvements in measurement make it easier to reject the null. But since the research hypothesis has not made any specific quantitative prediction, more accurate measurement doesn't lead to stronger corroboration. See also Andrew Gelman's comments in a September 5, 2014 blog post: http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/. [5]

6. George E. P. Box is famous for this dictum. As far as I can tell, his first published use of it was as a section heading in a 1979 paper (Box, 1979). Population biologists like myself are more familiar with a philosophically similar essay about modeling in general by Richard Levins, "The Strategy of Model Building in Population Biology" (Levins, 1966). [5]

7. Ohta and Gillespie (1996). [5]

8. Hubbell (2001). The theory has been productive in that it has forced greater clarity of modeling and understanding of relations between theory and data. But the theory has had its difficulties. See Clark (2012). For a more general skeptical attitude towards "neutrality," see Proulx and Adler (2010). [5]

9. For direct application of Kimura's model to cultural variation, see for example Hahn and Bentley (2003). All of the same epistemic problems reemerge here, but in a context with much less precision of theory. Hahn and Bentley have since adopted a more nuanced view of the issue. See their comment to Lansing and Cox (2011), as well as the similar comment by Feldman. [5]

10. Gillespie (1977). [5]

11. Lansing and Cox (2011). See objections by Hahn, Bentley, and Feldman in the peer commentary to the article. [7]

12. See Cho (2011) for a December 2011 summary focusing on debates about measurement. [8]

13. For an autopsy of the experiment, see http://profmattstrassler.com/articles-and-posts/particle-physics-basics/neutrinos/neutrinos-faster-than-light/opera-what-went-wrong/. [9]

14. See Mulkay and Gilbert (1981) for many examples of "Popperism" from practicing scientists, including famous ones. [9]

15. For an accessible history of some measurement issues in the development of physics and biology, including early experiments on relativity and abiogenesis, I recommend Collins and Pinch (1998). Some scientists have read this book as an attack on science. However, as the authors clarify in the second edition, this was not their intention. Science makes myths, like all cultures do. That doesn't necessarily imply that science does not work. See also Daston and Galison (2007), which tours concepts of objective measurement, spanning several centuries. [9]

16. The first chapter of Sober (2008) contains a similar discussion of *modus tollens*. Note that the statistical philosophy of Sober's book is quite different from that of the book you are holding. In particular, Sober is weakly anti-Bayesian. This is important, because it emphasizes that rejecting *modus tollens* as a model of statistical inference has nothing to do with any debates about Bayesian versus non-Bayesian tools. [9]

17. Popper himself had to deal with this kind of theory, because the rise of quantum mechanics in his lifetime presented rather serious challenges to the notion that measurement was unproblematic. See Chapter 9 in his *Logic of Scientific Discovery*, for example. [9]

18. See the Afterword to the 2nd edition of Collins and Pinch (1998) for examples of textbooks getting it wrong by presenting tidy fables about the definitiveness of evidence. [10]

19. A great deal has been written about the sociology of science and the interface of science and public interest. Interested novices might begin with Kitcher (2011), *Science in a Democratic Society*, which has a very broad topical scope and so can serve as an introduction to many dilemmas. [10]

20. Yes, even procedures that claim to be free of assumptions do have assumptions and are a kind of model. All systems of formal representation, including numbers, do not directly reference reality. For example, there is more than one way to construct "real" numbers in mathematics, and there are important consequences in some applications. In application, all formal systems are like models. See http://plato.stanford.edu/entries/philosophy-mathematics/ for a short overview of some different stances that can be sustained towards reasoning in mathematical systems. [10]

21. Saint Augustine, in *City of God*, famously admonished against trusting in luck, as personified by Fortuna: "How, therefore, is she good, who without discernment comes to both the good and to the bad?" See also the introduction to Gigerenzer et al. (1990). Rao (1997) presents a page from an old book of random numbers, commenting upon how seemingly useless such a thing would have been in previous eras. [10]

22. Most scholars trace frequentism to British logician John Venn (1834–1923), as for example presented in his 1876 book. Speaking of the proportion of male births in all births, Venn said, "probability *is* nothing but that proportion" (page 84). Venn taught Fisher some of his maths, so this may be where Fisher acquired his opposition to Bayesian probability. [11]

23. Fisher (1956). See also Fisher (1955), the first major section of which discusses the same point. Some people would dispute that Fisher was a "frequentist," because he championed his own likelihood methods over the methods of Neyman and Pearson. But Fisher definitely rejected the broader Bayesian approach to probability theory. See Endnote 28. [11]

24. This last sentence is a rephrasing from Lindley (1971): "A statistician faced with some data often embeds it in a family of possible data that is just as much a product of his fantasy as is a prior distribution." Dennis V. Lindley (1923–2013) was a prominent defender of Bayesian data analysis when it had very few defenders. [11]

25. It's hard to find an accessible introduction to image analysis, because it's a very computational subject. At the intermediate level, see Marin and Robert (2007), Chapter 8. You can hum over their mathematics and still

acquaint yourself with the different goals and procedures. See also Jaynes (1984) for spirited comments on the history of Bayesian image analysis and his pessimistic assessment of non-Bayesian approaches. There are better non-Bayesian approaches since. [12]

26. Binmore (2009) describes the history within economics and related fields and provides a critique that I am sympathetic to. [12]

27. See Gigerenzer et al. (2004). [13]

28. Fisher (1925), page 9. See Gelman and Robert (2013) for reflection on intemperate anti-Bayesian attitudes from the middle of last century. [13]

29. See McGrayne (2011) for a non-technical history of Bayesian data analysis. See also Fienberg (2006), which describes (among many other things) applied use of Bayesian multilevel models in election prediction, beginning in the early 1960s. [13]

30. See Fienberg (2006), page 24. [15]

31. See Wang et al. (2015) for a vivid example. [15]

32. I borrow this phrasing from Silver (2012). Silver's book is a well-written, non-technical survey of modeling and prediction in a range of domains. [15]

33. See Theobald (2010) for a fascinating example in which multiple non-null phylogenetic models are contrasted. [16]

34. See Sankararaman et al. (2012) for a thorough explanation, including why current evidence suggests that there really was interbreeding. [16]

Chapter 2

35. Morison (1942). Globe illustration modified from public domain illustration at the Wikipedia entry for Martin Behaim. In addition to underestimating the circumference, Colombo also overestimated the size of Asia and the distance between mainland China and Japan. [19]

36. This distinction and vocabulary derive from Savage (1962). [19]

37. See Robert (2007) for thorough coverage of the decision-theoretic optimality of Bayesian inference. [19]

38. See Simon (1969) and chapters in Gigerenzer et al. (2000). [20]

39. See Cox (1946). Jaynes (2003) and Van Horn (2003) explain the Cox theorem and its role in inference. [24]

40. See Gelman and Robert (2013) for examples. [24]

41. I first encountered this globe tossing strategy in Gelman and Nolan (2002). Since I've been using it in classrooms, several people have told me that they have seen it in other places, but I've been unable to find a primeval citation, if there is one. [28]

42. There is actually a set of theorems, the *No Free Lunch* theorems. These theorems—and others which are similar but named and derived separately—effectively state that there is no optimal way to pick priors (for Bayesians) or select estimators or procedures (for non-Bayesians). See Wolpert and Macready (1997) for example. [31]

43. This is a subtle point that will be expanded in other places. On the topic of accuracy of assumptions versus information processing, see e.g. Appendix A of Jaynes (1985): The Gaussian, or normal, error distribution needn't be physically correct in order to be the most useful assumption. [32]

44. Kronecker (1823–1891), an important number theorist, was quoted as stating "God made the integers, all else is the work of man" (*Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk*). There appears to be no consensus among mathematicians about which parts of mathematics are discovered rather than invented. But all admit that applied mathematical models are "work of man." [32]

152. Hilbe (2011) is an entire book devoted to gamma-Poisson regression. [350]

153. Jung et al. (2014). [352]

Chapter 12

154. Wearing's wife Deborah has written a book about their life after the illness (Wearing, 2005). His story has also appeared in a number of documentaries. A quick internet search will turn up a number of news articles, as well. [355]

155. See section 6, page 20, of Gelman (2005) for an entertaining list of wildly different definitions of "random effect." [357]

156. Vonesh and Bolker (2005). [357]

157. I adopt the terminology of Gelman (2005), who argues that the common term *random effects* hardly aids with understanding, for most people. Indeed, it seems to encourage misunderstanding, partly because the terms *fixed* and *random* mean different things to different statisticians. See pages 20–21 of Gelman's paper. I fully realize, however, that by trying to spread Gelman's alternative jargon, I am essentially spitting into a very strong wind. [358]

158. It's also common for the "multi" to refer to multiple linear models. This is especially true in the literature on "hierarchical linear models." Regardless, we're talking about the same kind of robot here. [359]

159. Note that there is still uncertainty about the regularization. So this model isn't exactly the same as just assuming a regularizing prior with a constant standard deviation 1.6. Instead the intercepts for each tank average over the uncertainty in σ (and α). [360]

160. This fact has been understood much longer than multilevel models have been practical to use. See Stein (1955) for an influential non-Bayesian paper. [364]

Chapter 13

161. Lewandowski et al. (2009). The "LKJ" part of the name comes from the first letters of the last names of the authors, who themselves called the approach the "onion method." For use in Bayesian models, see the explanation in the latest version of the Stan reference manual. [394]

162. See Gelfand et al. (1995), as well as Roberts and Sahu (1997). See also Papaspiliopoulos et al. (2007) for a more recent overview. See Betancourt and Girolami (2013) for a discussion focusing of Hamiltonian Monte Carlo. [408]

163. See Neal (1998) for a highly cited overview, with notes on implementation. [410]

164. See MacKay and Neal (1994); Neal (1996). [419]

Chapter 14

165. Joseph Bertrand, 1889, *Calcul des probabilités*. [423]

166. See Carroll et al. (2012) for an overview of both Bayesian and non-Bayesian approaches to measurement error. The topic of measurement error is often very specific to different disciplines and contexts, because the nature of error can be very specific. [425]

167. See Molenberghs et al. (2014) for an overview of contemporary approaches, Bayesian and otherwise. [431]

168. See Rubin (1976); Rubin and Little (2002) for background and additional terminology. Section 4 of Rubin's 1976 article is valuable for the clear definitions of causes of missing data. [432]

Chapter 15

169. See Speed (1986) for extended comments like this, aimed at statisticians. You can find a copy of this essay online with a quick internet search. [441]

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, 30:9–14.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society, Series D*, 43(4):495–504.

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

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

Betancourt, M. J. and Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. arXiv:1312.0906.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admission: Data from Berkeley. *Science*, 187(4175):398–404.

Binmore, K. (2009). *Rational Decisions*. Princeton University Press.

Bolker, B. (2008). *Ecological Models and Data in R*. Princeton University Press.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. and Wilkinson, G., editors, *Robustness in Statistics*. Academic Press, New York.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, 143:383–430.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Pub. Co., Reading, Mass.

Breiman, L. (1968). *Probability*. Addison-Wesley Pub. Co.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X., editors (2011). *Handbook of Markov Chain Monte Carlo*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, 2nd edition.

Campbell, D. T. (1985). Toward an epistemologically-relevant sociology of science. *Science, Technology, & Human Values*, 10(1):38–48.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2012). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press, 2nd edition.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Caticha, A. and Griffin, A. (2007). Updating probabilities. In Mohammad-Djafari, A., editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 872 of *AIP Conf. Proc.*

Cho, A. (2011). Superluminal neutrinos: Where does the time go? *Science*, 334(6060):1200–1201.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Clark, J. S. (2012). The coherence problem with the unified neutral theory of biodiversity. *Trends in Ecology and Evolution*, 27:198–2002.

Collins, H. M. and Pinch, T. (1998). *The Golem: What You Should Know about Science*. Cambridge University Press, 2nd edition.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–10.

Cushman, F., Young, L., and Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12):1082–1089.

Daston, L. J. and Galison, P. (2007). *Objectivity*. MIT Press, Cambridge, MA.

Elias, P. (1958). Two famous papers. *IRE Transactions: on Information Theory*, 4:99.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904.

Ferguson, C. J. and Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7(6):555–561.

Feynman, R. (1967). *The character of physical law*. MIT Press.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1–40.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society B*, 17(1):69–78.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner, New York, NY.

Fontani, M., Costa, M., and Orna, M. V. (2014). *The Lost Elements: The Periodic Table's Shadow Side*. Oxford University Press, Oxford.

Forer, B. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44:118–123.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345:1502–1505.

Frank, S. (2007). *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. Princeton University Press, Princeton, NJ.

Frank, S. A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, 22:1563–1585.

Frank, S. A. (2011). Measurement scale in maximum entropy models of species abundance. *Journal of Evolutionary Biology*, 24:485–496.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, 38(2):306–347.

Galton, F. (1989). Kinship and correlation. *Statistical Science*, 4(2):81–86.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika*, (82):479–488.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–534.

Gelman, A., Carlin, J. C., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013a). *Bayesian Data Analysis*. Chapman & Hall/CRC, 3rd edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Hwang, J., and Vehtari, A. (2013b). Understanding predictive information criteria for Bayesian models.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was

posited ahead of time. Technical report, Department of Statistics, Columbia University.
Gelman, A. and Loken, E. (2014). Ethics and statistics: The AAA tranche of subprime science. *CHANCE*, 27(1):51–56.
Gelman, A. and Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks*. Oxford University Press.
Gelman, A. and Robert, C. P. (2013). "Not only defended but also applied": The perceived absurdity of Bayesian inference. *The American Statistician*, 67(1):1–5.
Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.
Gelman, A. and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, 25:165–173.
Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331.
Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.
Gigerenzer, G. and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102:684–704.
Gigerenzer, G., Krauss, S., and Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In Kaplan, D., editor, *The Sage handbook of quantitative methodology for the social sciences*, pages 391–408. Sage Publications, Inc., Thousand Oaks.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Kruger, L. (1990). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press.
Gigerenzer, G., Todd, P., and The ABC Research Group (2000). *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford.
Gilad, Y. and Mizrahi-Man, O. (2015). A reanalysis of mouse encode comparative gene expression data. *F1000Research*, 4(121).
Gillespie, J. H. (1977). Sampling theory for alleles in a random environment. *Nature*, 266:443–445.
Grafen, A. and Hails, R. (2002). *Modern Statistics for the Life Sciences*. Oxford University Press, Oxford.
Griffin, A. (2008). *Maximum Entropy: The Universal Method for Inference*. PhD thesis, University of Albany, State University of New York, Department of Physics.
Grosberg, A. (1998). Entropy of a knot: Simple arguments about difficult problem. In Stasiak, A., Katrich, V., and Kauffman, L. H., editors, *Ideal Knots*, pages 129–142. World Scientific.
Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge MA.
Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press, Cambridge.
Hahn, M. W. and Bentley, R. A. (2003). Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society B*, 270:S120–S123.
Harte, J. (2011). *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford.
Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.
Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis & Prevention*, 36:495–500.
Henrion, M. and Fischoff, B. (1986). Assessing uncertainty in physcial constants. *American Journal of Physics*, 54:791–798.
Herndon, T., Ash, M., and Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2):257–279.
Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, Cambridge, 2nd edition.

Hinde, K. and Milligan, L. M. (2011). Primate milk synthesis: Proximate mechanisms and ultimate perspectives. *Evolutionary Anthropology*, 20:9–23.

Hoffman and Gelman (2011). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo.

Horton, R. (2015). What is medicine's 5 sigma? *The Lancet*, 385(April 11):1380.

Howell, N. (2000). *Demography of the Dobe !Kung*. Aldine de Gruyter, New York.

Howell, N. (2010). *Life Histories of the Dobe !Kung: Food, Fatness, and Well-being over the Life-span*. Origins of Human Behavior and Culture. University of California Press.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.

Hull, D. L. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. University of Chicago Press, Chicago, IL.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correction. *Perspectives on Psychological Science*, 7(6):645–654.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In Harper, W. L. and Hooker, C. A., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, page 175.

Jaynes, E. T. (1984). The intutive inadequancy of classical statistics. *Epistemologia*, 7:43–74.

Jaynes, E. T. (1985). Highly informative priors. *Bayesian Statistics*, 2:329–360.

Jaynes, E. T. (1986). Monkeys, kangaroos and *N*. In Justice, J. H., editor, *Maximum-Entropy and Bayesian Methods in Applied Statistics*, page 26. Cambridge University Press, Cambridge.

Jaynes, E. T. (1988). The relation of Bayesian and maximum entropy methods. In Erickson, G. J. and Smith, C. R., editors, *Maximum Entropy and Bayesian Methods in Science and Engineering*, volume 1, pages 25–29. Kluwer Academic Publishers.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jung, K., Shavitt, S., Viswanathan, M., and Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences USA*, 111(24):8782–8787.

Kadane, J. B. (2011). *Principles of Uncertainty*. Chapman & Hall/CRC.

Kitcher, P. (2000). Reviving the sociology of science. *Philosophy of Science*, 67:S33–S44.

Kitcher, P. (2011). *Science in a Democratic Society*. Prometheus Books, Amherst, New York.

Kline, M. A. and Boyd, R. (2010). Population size predicts technological complexity in Oceania. *Proc. R. Soc. B*, 277:2559–2564.

Kruscke, J. K. (2011). *Doing Bayesian Data Analysis*. Academic Press, Burlington, MA.

Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons, NY.

Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician*, 41(4):340.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

Lansing, J. S. and Cox, M. P. (2011). The domain of the replicators: Selection, neutrality, and cultural evolution (with commentary). *Current Anthropology*, 52:105–125.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1):19–49.

Lee, R. B. and DeVore, I., editors (1976). *Kalahari Hunter-Gatherers: Studies of the !Kung San and Their Neighbors*. Harvard University Press, Cambridge.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100:1989–2001.

Lightsey, J. D., Rommel, S. A., Costidis, A. M., and Pitchford, T. D. (2006). Methods used during gross necropsy to determine watercraft-related mortality in the Florida manatee (*Trichechus manatus*

latirostris). *Journal of Zoo and Wildlife Medicine*, 37(3):262–275.

Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., Gingeras, T. R., Ecker, J. R., and Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U.S.A.*, 111(48):17224–17229.

Lindley, D. V. (1971). Estimation of many parameters. In Godambe, V. P. and Sprott, D. A., editors, *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS Book*. CRC Press.

MacKay, D. J. C. and Neal, R. M. (1994). Automatic relevance determination for neural networks. Technical report, Cambridge University.

Mangel, M. and Samaniego, F. (1984). Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association*, 79:259–267.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42:109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition.

McElreath, R. and Smaldino, P. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS One*, 10(8):e0136088. doi:10.1371/journal.pone.0136088.

McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

McHenry, H. M. and Coffing, K. (2000). *Australopithecus* to *Homo*: Transformations in body and mind. *Annual Review of Anthropology*, 29:125–146.

Meehl, P. E. (1956). Wanted—a good cookbook. *The American Psychologist*, 11:263–272.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34:103–115.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66:195–244.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. CRC Press.

Morison, S. E. (1942). *Admiral of the Ocean Sea: A Life of Christopher Columbus*. Little, Brown and Company, Boston.

Mulkay, M. and Gilbert, G. N. (1981). Putting philosophy to work: Karl Popper's influence on scientific practice. *Philosophy of the Social Sciences*, 11:389–407.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag, New York.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. In Bernardo, J. M., editor, *Bayesian Statistics*, volume 6, pages 475–501. Oxford University Press.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384.

Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, 17:354–74.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, 14(9):1105–1107.

Nunn, N. and Puga, D. (2011). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*.

Nuzzo, R. (2014). Statistical errors. *Nature*, 506:150–152.

Ohta, T. and Gillespie, J. H. (1996). Development of neutral and nearly neutral theories. *Theoretical Population Biology*, 49:128–142.

O'Rourke, K. and Detsky, A. S. (1989). Meta-analysis in medical research: Strong encouragement for higher quality in individual research efforts. *Journal of Clinical Epidemiology*, 42(10):1021–1024.

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, (22):59–73.

Pearl, J. (2014). Understanding Simpson's paradox. *The American Statistician*, 68:8–13.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7:887–902.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, New York.

Popper, K. (1996). *The Myth of the Framework: In Defence of Science and Rationality*. Routledge.

Proulx, S. R. and Adler, F. R. (2010). The standard of neutrality: still flapping in the breeze? *Journal of Evolutionary Biology*, 23:1339–1350.

Rao, C. R. (1997). *Statistics and Truth: Putting Chance To Work*. World Scientific Publishing.

Reilly, C. and Zeringue, A. (2004). Improved predictions of lynx trappings using a biologial model. In Gelman, A. and Meng, X., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 297–308. John Wiley and Sons.

Reinhart, C. and Rogoff, K. (2010). Growth in a time of debt. *American Economic Review*, 100(2):573–578.

Rice, K. (2010). A decision-theoretic formulation of Fisher's approach to testing. *The American Statistician*, 64(4):345–349.

Riley, S. J., DeGloria, S. D., and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, 5:23–27.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 2. CRC Press.

Robert, C. P. (2007). *The Bayesian Choice: from decision-theoretic foundations to computational implementation*. Springer Texts in Statistics. Springer, 2nd edition.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, (59):291–317.

Rommel, S. A., Costidis, A. M., Pitchford, T. D., Lightsey, J. D., Snyder, R. H., and Haubold, E. M. (2007). Forensic methods for characterizing watercraft from watercraft-induced wounds on the Florida manatee (*Trichechus manatus latirostris*). *Marine Mammal Science*, 23(1):110–132.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society A*, 147(5):656–666.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rubin, D. B. and Little, R. J. A. (2002). *Statistical analysis with missing data*. Wiley, New York, 2nd edition.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genetics*, 8(10):e1002947.

Savage, L. J. (1962). *The Foundations of Statistical Inference*. Methuen.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sedlemeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2):309–316.

Senn, S. (2003). A conversation with John Nelder. *Statistical Science*, 18:118–131.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions: on Information Theory*, 2:3.

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J., and Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. *Nature*, 437:1357–1359.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin Press, New York.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359–1366.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2013). Life after p-hacking. SSRN Scholarly Paper ID 2205186, Social Science Research Network, Rochester, NY.

Simon, H. (1969). *The Sciences of the Artificial*. MIT Press, Cambridge, Mass.

Simpson, D. P., Martins, T. G., Riebler, A., Fuglstad, G.-A., Rue, H., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv:1403.4630v3*.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241.

Sober, E. (2008). *Evidence and Evolution: The logic behind the science*. Cambridge University Press, Cambridge.

Speed, T. (1986). Questions, answers and statistics. In *International Conference on Teaching Statistics 2*. International Association for Statistical Education.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivatiate normal distribution. In *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, volume 1, pages 197–206, Berkeley. University of California Press.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39(1):44–47.

Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature*, 465:219–222.

Van Horn, K. S. (2003). Constructing a logic of plausible inference: A guide to Cox's theorem guide to Cox's theorem. *International Journal of Approximate Reasoning*, 34:3–24.

Vehtari, A. and Gelman, A. (2014). WAIC and cross-validation in Stan. Technical report, Aalto University.

Venn, J. (1876). *The Logic of Chance*. Macmillan and co, New York, 2nd edition.

Vonesh, J. R. and Bolker, B. M. (2005). Compensatory larval responses shift trade-offs associated with predator-induced hatching plasticity. *Ecology*, 86:1580–1591.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326.

Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Technical report, Statistical Research Group, Columbia University.

Wald, A. (1950). *Statistical Decision Functions*. J. Wiley, New York.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Wearing, D. (2005). *Forever Today: A True Story of Lost Memory and Never-Ending Love*. Doubleday.

Welsh, Jr., H. H. and Lind, A. (1995). Habitat correlates of the Del Norte salamander, Plethodon elongatus (Caudata: Plethodontidae) in northwestern California. *Journal of Herpetology*, 29:198–210.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31:949–952.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society, Series C*, 31(2):144–148.

Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, 31:131–144.

Wittgenstein, L. (1953). *Philosophical Investigations*.

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, page 67.

Citation index

Akaike (1973), 451, 457
Akaike (1974), 451, 457
Akaike (1978), 450, 451, 457
Akaike (1981), 451, 457
Baker (1994), 453, 457
Berger and Berry (1988), 448, 457
Berger (1985), 451, 457
Betancourt and Girolami (2013), 454, 457
Bickel et al. (1975), 453, 457
Binmore (2009), 447, 457
Bolker (2008), 453, 457
Box and Tiao (1973), 448, 457
Box (1979), 445, 457
Box (1980), 448, 457
Breiman (1968), 448, 457
Brooks et al. (2011), 452, 457
Burnham and Anderson (2002), 450, 451, 457
Campbell (1985), 455, 457
Carroll et al. (2012), 454, 457
Casella and George (1992), 452, 457
Caticha and Griffin (2007), 452, 453, 457
Cho (2011), 446, 457
Claeskens and Hjort (2008), 451, 457
Clark (2012), 445, 458
Collins and Pinch (1998), 445, 446, 458
Cox (1946), 447, 458
Cushman et al. (2006), 453, 458
Daston and Galison (2007), 446, 458
Elias (1958), 450, 458
Fanelli (2012), 455, 458
Ferguson and Heene (2012), 455, 458
Feynman (1967), 448, 458
Fienberg (2006), 447, 451, 458
Fisher (1925), 447, 448, 458
Fisher (1955), 446, 458
Fisher (1956), 446, 448, 458
Fontani et al. (2014), 455, 458
Forer (1949), 455, 458
Franco et al. (2014), 455, 458
Frank (2007), 453, 458
Frank (2009), 448, 450, 458
Frank (2011), 450, 458
Fullerton (2009), 453, 458
Galton (1989), 449, 458
Gelfand et al. (1995), 454, 458
Gelman and Hill (2007), 450, 458
Gelman and Loken (2013), 455, 458
Gelman and Loken (2014), 455, 459
Gelman and Nolan (2002), 447, 459
Gelman and Robert (2013), 447, 459
Gelman and Rubin (1992), 452, 459
Gelman and Rubin (1995), 451, 459
Gelman and Stern (2006), 450, 459
Gelman et al. (2013a), 448, 451, 458
Gelman et al. (2013b), 451, 458
Gelman (2005), 454, 458
Gelman (2006), 452, 458
Geman and Geman (1984), 452, 459
Gigerenzer and Hoffrage (1995), 448, 459
Gigerenzer et al. (1990), 446, 452, 459
Gigerenzer et al. (2000), 447, 459
Gigerenzer et al. (2004), 447, 459
Gilad and Mizrahi-Man (2015), 455, 459
Gillespie (1977), 445, 459
Grafen and Hails (2002), 450, 452, 459
Griffin (2008), 452, 453, 459
Grosberg (1998), 452, 459
Grünwald (2007), 450, 451, 459
Hacking (1983), 445, 459
Hahn and Bentley (2003), 445, 459
Harte (2011), 450, 459
Hastie et al. (2009), 450, 459
Hastings (1970), 452, 459
Hauer (2004), 448, 459
Henrion and Fischoff (1986), 448, 459
Herndon et al. (2014), 455, 459
Hilbe (2011), 454, 459
Hinde and Milligan (2011), 450, 459
Hoffman and Gelman (2011), 452, 460
Horton (2015), 455, 460
Howell (2000), 449, 460
Howell (2010), 449, 460
Hubbell (2001), 445, 460
Hull (1988), 455, 460

Ioannidis (2005), 448, 451, 455, 460
Ioannidis (2012), 455, 460
Jaynes (1976), 450, 460
Jaynes (1984), 447, 460
Jaynes (1985), 447, 448, 460
Jaynes (1986), 449, 460
Jaynes (1988), 452, 453, 460
Jaynes (2003), 445, 447, 449–453, 460
Jung et al. (2014), 454, 460
Kadane (2011), 448, 460
Kitcher (2000), 455, 460
Kitcher (2011), 446, 460
Kline and Boyd (2010), 453, 460
Kruscke (2011), 452, 460
Kullback and Leibler (1951), 450, 460
Kullback (1959), 450, 460
Kullback (1987), 450, 460
Lambert (1992), 453, 460
Lansing and Cox (2011), 445, 446, 460
Laudan (1981), 455, 460
Lee and DeVore (1976), 449, 460
Levins (1966), 445, 460
Lewandowski et al. (2009), 454, 460
Lightsey et al. (2006), 451, 460
Lin et al. (2014), 455, 461
Lindley (1971), 446, 461
Lunn et al. (2013), 451, 461
MacKay and Neal (1994), 454, 461
Mangel and Samaniego (1984), 451, 461
Marin and Robert (2007), 446, 461
McCullagh and Nelder (1989), 453, 461
McCullagh (1980), 453, 461
McElreath and Smaldino (2015), 455, 461
McGrayne (2011), 447, 461
McHenry and Coffing (2000), 450, 461
Meehl (1956), 455, 461
Meehl (1967), 445, 461
Meehl (1990), 449, 461
Metropolis and Ulam (1949), 452, 461
Metropolis et al. (1953), 452, 461
Molenberghs et al. (2014), 454, 461
Morison (1942), 447, 461
Mulkay and Gilbert (1981), 445, 446, 461
Neal (1996), 454, 461
Neal (1998), 454, 461
Nelder and Wedderburn (1972), 453, 461
Nettle (1998), 452, 461
Nieuwenhuis et al. (2011), 450, 461
Nunn and Puga (2011), 451, 461
Nuzzo (2014), 453, 461
O'Rourke and Detsky (1989), 455, 462
Ohta and Gillespie (1996), 445, 462
Papaspiliopoulos et al. (2007), 454, 462
Pearl (2014), 449, 453, 462
Polson and Scott (2012), 452, 462
Popper (1963), 455, 462
Popper (1996), 445, 455, 462
Proulx and Adler (2010), 445, 462
Rao (1997), 446, 452, 462
Reilly and Zeringue (2004), 449, 462
Reinhart and Rogoff (2010), 455, 462
Rice (2010), 448, 462
Riley et al. (1999), 451, 462
Robert and Casella (2011), 452, 462
Robert (2007), 447, 448, 451, 462
Roberts and Sahu (1997), 454, 462
Rommel et al. (2007), 451, 462
Rosenbaum (1984), 450, 462
Rosenthal (1979), 455, 462
Rubin and Little (2002), 454, 462
Rubin (1976), 454, 462
Rubin (2005), 449, 462
Sankararaman et al. (2012), 447, 462
Savage (1962), 447, 462
Schwarz (1978), 451, 462
Sedlemeier and Gigerenzer (1989), 455, 462
Senn (2003), 453, 462
Shannon (1948), 450, 462
Shannon (1956), 450, 463
Silk et al. (2005), 453, 463
Silver (2012), 447, 463
Simmons et al. (2011), 451, 453, 455, 463
Simmons et al. (2013), 455, 463
Simon (1969), 447, 463
Simpson et al. (2014), 452, 463
Simpson (1951), 449, 453, 463
Sober (2008), 446, 463
Speed (1986), 454, 455, 463
Stein (1955), 454, 463
Stigler (1981), 450, 463
Stone (1977), 451, 463
Theobald (2010), 447, 463
Van Horn (2003), 447, 463
Vehtari and Gelman (2014), 451, 463
Venn (1876), 446, 463
Vonesh and Bolker (2005), 454, 463
Wald (1939), 451, 463
Wald (1943), 451, 463
Wald (1950), 451, 463
Wang et al. (2015), 447, 463
Watanabe (2010), 451, 463
Wearing (2005), 454, 463
Welsh and Lind (1995), 453, 463
Williams (1975), 453, 463
Williams (1980), 452, 453, 464
Williams (1982), 453, 464
Wittgenstein (1953), 445, 464
Wolpert and Macready (1997), 447, 464

Topic index

absolute effect, 296
aggregated binomial regression, 292
AIC, 189
Akaike weight, 198, 199
Akaike information criterion, 189
autocorrelation, of samples, 255
automatic relevance determination, 419
`axis`, 114

barplot, 203
Bayes factor, 192
Bayes' theorem, 36
Bayesian imputation, 424
Bayesian information criterion, 167, 192
Bayesian updating, 29
Bayesianism, 12
Bertrand's box paradox, 423
beta-binomial, 328, 347
bias-variance trade-off, 174
bias-variance tradeoff, *see also* overfitting
binomial distribution, 275
binomial regression, 291
Buridan's ass, 223
burn-in, 256

categorical, 323
categorical variable, 152
categorical variables, 120
Cauchy distribution, 249, 260
centering, 98, 99, 225, 230
Cholesky decomposition, 409
`coeftab`, 201
Colombo, Cristoforo, 19
Columbus, Christopher, 19
`compare`, 198
complete pooling, 364
complete-case, 432
concentration of measure, 360
conditional independence, 279
conditioning, 209
confidence interval, 54
consistency, model, 190
consistent, 190
continuous mixture, 346
correlation matrix, prior for, 393
correlation, among parameters, 99
credible interval, 54
cross entropy, 180
cross-classified, 371
cross-classified multilevel model, 403
cross-validation, 187
cumulative link, 331, 332
Curse of Tippecanoe, 205

data compression, 172
data dredging, 205
`data(Howell1)`, 87, 96, 153
`data(milk)`, 135, 155, 196
`data(WaffleDivorce)`, 121
`dbetabinom`, 348
design formula, 159
deviance, 182
Deviance Information Criterion, 190
deviance, computed example, 182
DIC, 190
divergence, *see also* Kullback-Leibler divergence, 179, 272, 274
divergent iterations, 373, 405
`dmvnormNC`, 405, 409
dotchart, 203
dummy data, 62
dummy variable, 153

effective samples, 321
ensemble, 203
`ensemble`, 204
entropy, cross, 180
event history analysis, 327
exchangeable, 81
exponential distribution, 282
exponential distribution, as prior, 364
exponential family, 7, 75, 282
`extract.samples`, 90

factor analysis, 149
focus, 381

folk theorem of statistical computing, 262
frequentist, 11

Galton, Francis, 92
gamma distribution, 283
gamma-Poisson, 350
Gaussian process regression, 410
Gaussian processes, 388
generalized linear model, 280, 281
  and information criteria, 288
generalized linear models, 268
geometric distribution, 328
Gibbs sampling, 245
`GPL2`, 413
grid approximation, 39

Hamiltonian Monte Carlo, 246
hierarchical model, 357
highest posterior density interval, 56
Histomancy, 282
hyperparameters, 359, 382
hyperpriors, 359

imputation, 432
index variable, 158
information criteria, 15, 166, 189
information criteria, multilevel models, 376
information entropy, 28, 178, 268
information theory, 15, 76, 166, 177
  Kullback-Leibler divergence, 179
interaction, 120, 210
interaction, continuous, 225
inverse-logit, 285

large world, 19
likelihood, 27, 32
likelihood, average, 37
likelihood, marginal, 37
linear model, 92
linear model, generalized, 280
linear regression, 71
`link`, 104, 107
link function, 281, 284
LKJcorr probability density, 393
log link, 286
log pointwise predictive density, 191
`log_sum_exp`, 193
logistic, 285
logistic regression, 292
logit link, 284
loss function, 59
lppd, 191

`map`, 42
`map2stan`, 249
Markov chain Monte Carlo, 44, 241
maxent, 179
maximum a posteriori, 42, 59
maximum entropy, 7, 76, 179, 268
maximum entropy classifier, 323
maximum entropy distribution, 271
maximum entropy, binomial distribution, 279
maximum entropy, Gaussian, 274
maximum entropy, Wallis derivation, 271
maximum likelihood estimate, 44
MCAR, 432
MCMC, 241
`mcreplicate`, 185
measurement error, 424
Metropolis algorithm, 245
Minimum Description Length, 172
missing data, 424
misspecified, 392
mixed effects model, 357
mixture model, 342
mixture, continuous, 351
model averaging, 196, 203
model checking, 64, 376
model comparison, 196
model selection, 195
modus tollens, 7
multicollinearity, 141
multilevel model, 14, 356
multilevel model, cross-classified, 403
multilevel model, non-centered parameterization, 408
multinomial distribution, 323
multinomial-Poisson transformation, 325
multinomial-Poisson transformation, derivation, 327
multivariate regression, 119

`n_eff`, 255
negative-binomial, 328, 350
no pooling, 365
non-centered parameterization, 403, 405, 408
non-identifiability, 149
number of samples, effective, 255

Ockham's razor, 165
OLS, 159
omitted variable bias, 150
`optim`, 228
ordered categorical, 331
ordinary least squares, 159
over-dispersed, 328
over-dispersion, 346, 363, 382
overfitting, 3, 15, 20, 141, 166–168

`pairs`, 147
parameter, 27
parameters, 34
partial pooling, 365

patristic distance, 418
penalized likelihood, 35
percentile intervals, 56
phylogenetic regression, 418
point estimate, 58
Poisson distribution, 283
Poisson regression, 291
polynomial regression, 110
pooling, 362
post-treatment bias, 141, 150
posterior distribution, 36
posterior predictive distribution, 65
posterior probability, 27
power analysis, 61
prequential, 195
principle components, 149
principle of indifference, 26
prior, 34
prior probability, 27
proportional change in odds, 296

quadratic approximation, 41

random effects, 357
randomization, 28
regularizing, 35
regularizing prior, 166, 186, 360
relative effect, 296
residuals, 282
ridge regression, 188
`rlkjcorr`, 394
`rugged`, dataset, 211

sampling distribution, 11, 63
sensitivity analysis, 287
shrinkage, 362, 401, 428
`sim`, 108, 110
`sim.train.test`, 184, 185
Simpson's paradox, 119, 309
simulation, 61
small world, 19
Stan, 241
standard error, 44
standardize, 111
Stanislaw Ulam, 242
stargazing, 167
start values, 88
stochastic, 78
subjective Bayesian, 35
survival analysis, 327

trace plot, 253
triptych, 233

underfitting, 166, *see also* overfitting, 172

variance-covariance, 89
varying effects, 358, 387
varying intercepts, 358, 362
varying intercepts, example, 398
varying slopes, 388, 389, 392
varying slopes, example, 399

WAIC, 190, 191
weakly informative, 35
Widely Applicable Information Criterion, 190

zero-augmented, 331
zero-inflated, 331, 342