Data Analysis with R

R is a programming language and statistics platform that has gained enormous popularity in recent years. Frequently the tool of choice for academics, R has spread deep into the private sector and can be found in the production pipelines of some of the most advanced and successful enterprises. The power and domain-specificity of R allows the user to express complex analytics easily, quickly, and succinctly. With over 7,000 user-contributed packages, you’ll easily find support for the latest and greatest algorithms and techniques.

Starting with the basics of R and statistical reasoning, we’ll dive into advanced predictive analytics and apply those techniques to real-world data (with real-world problems). Packed with engaging examples and exercises, this book begins with a review of R and its syntax. From there, we dive straight into the fundamentals of applied statistics. Gradually, we build on this knowledge until we are performing sophisticated and powerful analytics. Finally, we solve difficulties relating to performing data analysis in practice. Here, we find solutions to working with “messy data”, large data, communicating results, and facilitating reproducibility.

Who this book is written for

Whether you are learning data analysis for the first time, or you want to deepen the understanding you already have, this book will prove to be an invaluable resource. If you have some prior programming experience and a mathematical background, then this book is ideal for you.

What you will learn from this book

- Navigate the R environment
- Describe and visualize the behavior of data and relationships between data
- Gain a thorough understanding of statistical reasoning and sampling
- Employ hypothesis tests to draw inferences from your data
- Learn Bayesian methods for estimating parameters
- Perform regression to predict continuous variables
- Apply powerful classification methods to predict categorical data
- Handle missing data gracefully using multiple imputation
- Identify and manage problematic data points
- Employ parallelization and Rcpp to scale your analyses for larger data
- Put best practices into effect to make your job easier and facilitate reproducibility
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 7 'Bayesian Methods'
- A synopsis of the book’s content
- More information on Data Analysis with R
Tony Fischetti is a data scientist at College Factual, where he gets to use R everyday to build personalized rankings and recommender systems. He graduated in cognitive science from Rensselaer Polytechnic Institute, and his thesis was strongly focused on using statistics to study visual short-term memory.

Tony enjoys writing and contributing to open source software, blogging at http://www.onthelambda.com, writing about himself in third person, and sharing his knowledge using simple, approachable language and engaging examples.

The more traditionally exciting of his daily activities include listening to records, playing the guitar and bass (poorly), weight training, and helping others.
I'm going to shoot it to you straight: there are a lot of books about data analysis and the R programming language. I'll take it on faith that you already know why it's extremely helpful and fruitful to learn R and data analysis (if not, why are you reading this preface?!) but allow me to make a case for choosing this book to guide you in your journey.

For one, this subject didn't come naturally to me. There are those with an innate talent for grasping the intricacies of statistics the first time it is taught to them; I don't think I'm one of these people. I kept at it because I love science and research and knew that data analysis was necessary, not because it immediately made sense to me. Today, I love the subject in and of itself, rather than instrumentally, but this only came after months of heartache. Eventually, as I consumed resource after resource, the pieces of the puzzle started to come together. After this, I started tutoring all of my friends in the subject—and have seen them trip over the same obstacles that I had to learn to climb. I think that coming from this background gives me a unique perspective on the plight of the statistics student and allows me to reach them in a way that others may not be able to. By the way, don't let the fact that statistics used to baffle me scare you; I have it on fairly good authority that I know what I'm talking about today.

Secondly, this book was born of the frustration that most statistics texts tend to be written in the driest manner possible. In contrast, I adopt a light-hearted buoyant approach—but without becoming agonizingly flippant.

Third, this book includes a lot of material that I wished were covered in more of the resources I used when I was learning about data analysis in R. For example, the entire last unit specifically covers topics that present enormous challenges to R analysts when they first go out to apply their knowledge to imperfect real-world data.
Lastly, I thought long and hard about how to lay out this book and which order of topics was optimal. And when I say long and hard I mean I wrote a library and designed algorithms to do this. The order in which I present the topics in this book was very carefully considered to (a) build on top of each other, (b) follow a reasonable level of difficulty progression allowing for periodic chapters of relatively simpler material (psychologists call this intermittent reinforcement), (c) group highly related topics together, and (d) minimize the number of topics that require knowledge of yet unlearned topics (this is, unfortunately, common in statistics). If you're interested, I detail this procedure in a blog post that you can read at http://bit.ly/teach-stats.

The point is that the book you're holding is a very special one—one that I poured my soul into. Nevertheless, data analysis can be a notoriously difficult subject, and there may be times where nothing seems to make sense. During these times, remember that many others (including myself) have felt stuck, too. Persevere… the reward is great. And remember, if a blockhead like me can do it, you can, too. Go you!

What this book covers

Chapter 1, Refresher, reviews the aspects of R that subsequent chapters will assume knowledge of. Here, we learn the basics of R syntax, learn R's major data structures, write functions, load data and install packages.

Chapter 2, The Shape of Data, discusses univariate data. We learn about different data types, how to describe univariate data, and how to visualize the shape of these data.

Chapter 3, Describing Relationships, goes on to the subject of multivariate data. In particular, we learn about the three main classes of bivariate relationships and learn how to describe them.

Chapter 4, Probability, kicks off a new unit by laying foundation. We learn about basic probability theory, Bayes' theorem, and probability distributions.

Chapter 5, Using Data to Reason About the World, discusses sampling and estimation theory. Through examples, we learn of the central limit theorem, point estimation and confidence intervals.

Chapter 6, Testing Hypotheses, introduces the subject of Null Hypothesis Significance Testing (NHST). We learn many popular hypothesis tests and their non-parametric alternatives. Most importantly, we gain a thorough understanding of the misconceptions and gotchas of NHST.
Chapter 7, *Bayesian Methods*, introduces an alternative to NHST based on a more intuitive view of probability. We learn the advantages and drawbacks of this approach, too.

Chapter 8, *Predicting Continuous Variables*, thoroughly discusses linear regression. Before the chapter’s conclusion, we learn all about the technique, when to use it, and what traps to look out for.

Chapter 9, *Predicting Categorical Variables*, introduces four of the most popular classification techniques. By using all four on the same examples, we gain an appreciation for what makes each technique shine.

Chapter 10, *Sources of Data*, is all about how to use different data sources in R. In particular, we learn how to interface with databases, and request and load JSON and XML via an engaging example.

Chapter 11, *Dealing with Messy Data*, introduces some of the snags of working with less than perfect data in practice. The bulk of this chapter is dedicated to missing data, imputation, and identifying and testing for messy data.

Chapter 12, *Dealing with Large Data*, discusses some of the techniques that can be used to cope with data sets that are larger than can be handled swiftly without a little planning. The key components of this chapter are on parallelization and Rcpp.

Chapter 13, *Reproducibility and Best Practices*, closes with the extremely important (but often ignored) topic of how to use R like a professional. This includes learning about tooling, organization, and reproducibility.
Bayesian Methods

Suppose I claim that I have a pair of magic rainbow socks. I allege that whenever I wear these special socks, I gain the ability to predict the outcome of coin tosses, using fair coins, better than chance would dictate. Putting my claim to the test, you toss a coin 30 times, and I correctly predict the outcome 20 times. Using a directional hypothesis with the binomial test, the null hypothesis would be rejected at alpha-level 0.05. Would you invest in my special socks?

Why not? If it's because you require a larger burden of proof on absurd claims, I don't blame you. As a grandparent of Bayesian analysis Pierre-Simon Laplace (who independently discovered the theorem that bears Thomas Bayes' name) once said: *The weight of evidence for an extraordinary claim must be proportioned to its strangeness.* Our prior belief — my absurd hypothesis — is so small that it would take much stronger evidence to convince the skeptical investor, let alone the scientific community.

Unfortunately, if you'd like to easily incorporate your prior beliefs into NHST, you're out of luck. Or suppose you need to assess the probability of the null hypothesis; you're out of luck there, too; NHST assumes the null hypothesis and can't make claims about the probability that a particular hypothesis is true. In cases like these (and in general), you may want to use Bayesian methods instead of frequentist methods. This chapter will tell you how. Join me!
The big idea behind Bayesian analysis

If you recall from Chapter 4, *Probability*, the Bayesian interpretation of probability views probability as our degree of belief in a claim or hypothesis, and Bayesian inference tells us how to update that belief in the light of new evidence. In that chapter, we used Bayesian inference to determine the probability that employees of Daisy Girl, Inc. were using an illegal drug. We saw how the incorporation of prior beliefs saved two employees from being falsely accused and helped another employee get the help she needed even though her drug screen was falsely negative.

In a general sense, Bayesian methods tell us how to dole out credibility to different hypotheses, given prior belief in those hypotheses and new evidence. In the drug example, the hypothesis suite was discrete: *drug user or not drug user*. More commonly, though, when we perform Bayesian analysis, our hypothesis concerns a continuous parameter, or many parameters. Our posterior (or updated beliefs) was also discrete in the drug example, but Bayesian analysis usually yields a continuous posterior called a *posterior distribution*.

We are going to use Bayesian analysis to put my magical rainbow socks claim to the test. Our parameter of interest is the proportion of coin tosses that I can correctly predict wearing the socks; we'll call this parameter $\theta$, or *theta*. Our goal is to determine what the most likely values of theta are and whether they constitute proof of my claim.

Refer back to the section on Bayes' theorem in Chapter 4, *Probability*. Recall that the posterior was the prior times the likelihood divided by a normalizing constant. This normalizing constant is often difficult to compute. Luckily, since it doesn't change the shape of the posterior distribution, and we are comparing relative likelihoods and probability densities, Bayesian methods often ignore this constant. So, all we need is a probability density function to describe our prior belief and a likelihood function that describes the likelihood that we would get the evidence we received given different parameter values.
The likelihood function is a *binomial function*, as it describes the behavior of Bernoulli trials; the binomial likelihood function for this evidence is shown in *Figure 7.1*:

![Figure 7.1: The likelihood function of theta for 20 out of 30 successful Bernoulli trials.](image)

For different values of theta, there are varying relative likelihoods. Note that the value of theta that corresponds to the maximum of the likelihood function is 0.667, which is the proportion of successful Bernoulli trials. This means that in the absence of any other information, the most likely proportion of coin flips that my magic socks allow me to predict is 67%. This is called the **Maximum Likelihood Estimate (MLE)**.

So, we have the likelihood function; now we just need to choose a prior. We will be crafting a representation of our prior beliefs using a type of distribution called a beta distribution, for reasons that we'll see very soon.

Since our posterior is a blend of the prior and likelihood function, it is common for analysts to use a prior that doesn't much influence the results and allows the likelihood function to *speak for itself*. To this end, one may choose to use a *non-informative prior* that assigns equal credibility to all values of theta. This type of non-informative prior is called a *flat* or *uniform prior*.

The beta distribution has two hyper-parameters, $\alpha$ (or $\text{alpha}$) and $\beta$ (or $\text{beta}$). A beta distribution with hyper-parameters $\alpha = \beta = 1$ describes such a *flat prior*. We will call this prior #1.
These are usually referred to as the beta distribution's parameters. We call them hyper-parameters here to distinguish them from our parameter of interest, theta.

![Beta distribution with parameters α = 1, β = 1](image)

Figure 7.2: A flat prior on the value of theta. This beta distribution, with alpha and beta = 1, confers an equal level of credibility to all possible values of theta, our parameter of interest.

This prior isn't really indicative of our beliefs, is it? Do we really assign as much probability to my socks giving me perfect coin-flip prediction powers as we do to the hypothesis that I'm full of baloney?

The prior that a skeptic might choose in this situation is one that looks more like the one depicted in Figure 7.3, a beta distribution with hyper-parameters \( \alpha = \beta = 50 \). This, rather appropriately, assigns far more credibility to values of theta that are concordant with a universe without magical rainbow socks. As good scientists, though, we have to be open-minded to new possibilities, so this doesn't rule out the possibility that the socks give me special powers—the probability is low, but not zero, for extreme values of theta. We will call this prior #2.
Before we perform the Bayesian update, I need to explain why I chose to use the beta distribution to describe my priors.

The Bayesian update—getting to the posterior—is performed by multiplying the prior with the likelihood. In the vast majority of applications of Bayesian analysis, we don't know what that posterior looks like, so we have to sample from it many times to get a sense of its shape. We will be doing this later in this chapter.

For cases like this, though, where the likelihood is a binomial function, using a beta distribution for our prior guarantees that our posterior will also be in the beta distribution family. This is because the beta distribution is a conjugate prior with respect to a binomial likelihood function. There are many other cases of distributions being self-conjugate with respect to certain likelihood functions, but it doesn't often happen in practice that we find ourselves in a position to use them as easily as we can for this problem. The beta distribution also has the nice property that it is naturally confined from 0 to 1, just like the proportion of coin flips I can correctly predict.
Bayesian Methods

The fact that we know how to compute the posterior from the prior and likelihood by just changing the beta distribution's hyper-parameters makes things really easy in this case. The hyper-parameters of the posterior distribution are:

$$\text{new } \alpha = \text{old } \alpha + \text{number of successes}$$

and

$$\text{new } \beta = \text{old } \beta + \text{number of failures}$$

That means the posterior distribution using prior #1 will have hyper-parameters $\alpha = 1+20$ and $\beta = 1+10$. This is shown in Figure 7.4.

![Figure 7.4: The result of the Bayesian update of the evidence and prior #1. The interval depicts the 95% credible interval (the densest 95% of the area under the posterior distribution). This interval overlaps slightly with theta = 0.5.](image)

A common way of summarizing the posterior distribution is with a credible interval. The credible interval on the plot in Figure 7.4 is the 95% credible interval and contains 95% of the densest area under the curve of the posterior distribution.

Do not confuse this with a confidence interval. Though it may look like it, this credible interval is very different than a confidence interval. Since the posterior directly contains information about the probability of our parameter of interest at different values, it is admissible to claim that there is a 95% chance that the correct parameter value is in the credible interval. We could make no such claim with confidence intervals. Please do not mix up the two meanings, or people will laugh you out of town.
Observe that the 95% most likely values for theta contain the theta value 0.5, if only barely. Due to this, one may wish to say that the evidence does not rule out the possibility that I'm full of baloney regarding my magical rainbow socks, but the evidence was suggestive.

To be clear, the end result of our Bayesian analysis is the posterior distribution depicting the credibility of different values of our parameter. The decision to interpret this as sufficient or insufficient evidence for my outlandish claim is a decision that is separate from the Bayesian analysis proper. In contrast to NHST, the information we glean from Bayesian methods—the entire posterior distribution—is much richer. Another thing that makes Bayesian methods great is that you can make intuitive claims about the probability of hypotheses and parameter values in a way that frequentist NHST does not allow you to do.

What does that posterior using prior #2 look like? It's a beta distribution with \( \text{alpha} = 50 + 20 \) and \( \text{beta} = 50 + 10 \):

```r
> curve(dbeta(x, 70, 60), # plot a beta distribution
+       xlab="θ", # name x-axis
+       ylab="posterior belief", # name y-axis
+       type="l", # make smooth line
+       yaxt='n') # remove y axis labels
> abline(v=.5, lty=2) # make line at theta = 0.5
```

![Figure 7.5: Posterior distribution of theta using prior #2](image)

---

[147]
Choosing a prior

Notice that the posterior distribution looks a little different depending on what prior you use. The most common criticism lodged against Bayesian methods is that the choice of prior adds an unsavory subjective element to analysis. To a certain extent, they’re right about the added subjective element, but their allegation that it is unsavory is way off the mark.

To see why, check out Figure 7.6, which shows both posterior distributions (from priors #1 and #2) in the same plot. Notice how priors #1 and #2—a very different priors—given the evidence, produce posteriors that look more similar to each other than the priors did.

Figure 7.6: The posterior distributions from prior #1 and #2
Now direct your attention to *Figure 7.7*, which shows the posterior of both priors if the evidence included 80 out of 120 correct trials.

![Figure 7.7: The posterior distributions from prior #1 and #2 with more evidence](image)

Note that the evidence still contains 67% correct trials, but there is now more evidence. The posterior distributions are now far more similar. Notice that now both of the posteriors' credible intervals do not contain $\theta = 0.5$; with 80 out of 120 trials correctly predicted, even the most obstinate skeptic has to concede that something is going on (though they will probably disagree that the power comes from the socks!).

Take notice also of the fact that the credible intervals, in both posteriors, are now substantially narrowing, illustrating more confidence in our estimate.
Finally, imagine the case where I correctly predicted 67% of the trials, but out of 450 total trials. The posteriors derived from this evidence are shown in Figure 7.8:

![Figure 7.8: The posterior distributions from prior #1 and #2 with even more evidence](image)

The posterior distributions are looking very similar—indeed, they are becoming identical. Given enough trials—given enough evidence—these posterior distributions will be exactly the same. When there is enough evidence available such that the posterior is dominated by it compared to the prior, it is called overwhelming the prior.

As long as the prior is reasonable (that is, it doesn't assign a probability of 0 to theoretically plausible parameter values), given enough evidence, everybody's posterior belief will look very similar.

There is nothing unsavory or misleading about an analysis that uses a subjective prior; the analyst just has to disclose what her prior is. You can't just pick a prior willy-nilly; it has to be justifiable to your audience. In most situations, a prior may be informed by prior evidence like scientific studies and can be something that most people can agree on. A more skeptical audience may disagree with the chosen prior, in which case the analysis can be re-run using their prior, just like we did in the magic socks example. It is sometimes okay for people to have different prior beliefs, and it is okay for some people to require a little more evidence in order to be convinced of something.
The belief that frequentist hypothesis testing is more objective, and therefore more correct, is mistaken insofar as it causes all parties to have a hold on the same potentially bad assumptions. The assumptions in Bayesian analysis, on the other hand, are stated clearly from the start, made public, and are auditable.

To recap, there are three situations you can come across. In all of these, it makes sense to use Bayesian methods, if that's your thing:

- You have a lot of evidence, and it makes no real difference which prior any reasonable person uses, because the evidence will overwhelm it.
- You have very little evidence, but have to make an important decision given the evidence. In this case, you'd be foolish to not use all available information to inform your decisions.
- You have a medium amount of evidence, and different posteriors illustrate the updated beliefs from a diverse array of prior beliefs. You may require more evidence to convince the extremely skeptical, but the majority of interested parties will be come to the same conclusions.

Who cares about coin flips?

Who cares about coin flips? Well, virtually no one. However, (a) **coin flips are a great simple application to get the hang of Bayesian analysis**; (b) **the kinds of problems that a beta prior and a binomial likelihood function solve go way beyond assessing the fairness of coin flips**. We are now going to apply the same technique to a real life problem that I actually came across in my work.

For my job, I had to create a career recommendation system that asked the user a few questions about their preferences and spat out some careers they may be interested in. After a few hours, I had a working prototype. In order to justify putting more resources into improving the project, I had to prove that I was on to something and that my current recommendations performed better than chance.

In order to test this, we got 40 people together, asked them the questions, and presented them with two sets of recommendations. One was the true set of recommendations that I came up with, and one was a control set—the recommendations of a person who answered the questions randomly. If my set of recommendations performed better than chance would dictate, then I had a good thing going, and could justify spending more time on the project.

Simply performing better than chance is no great feat on its own—I also wanted really good estimates of how much better than chance my initial recommendations were.
For this problem, I broke out my Bayesian toolbox! The parameter of interest is the proportion of the time my recommendations performed better than chance. If .05 and lower were very unlikely values of the parameter, as far as the posterior depicted, then I could conclude that I was on to something.

Even though I had strong suspicions that my recommendations were good, I used a uniform beta prior to preemptively thwart criticisms that my prior biased the conclusions. As for the likelihood function, it is the same function family we used for the coin flips (just with different parameters).

It turns out that 36 out of the 40 people preferred my recommendations to the random ones (three liked them both the same, and one weirdo liked the random ones better). The posterior distribution, therefore, was a beta distribution with parameters 37 and 5.

```r
> curve(dbeta(x, 37, 5), xlab="θ",
+       ylab="posterior belief",
+       type="l", yaxt='n')
```

![Figure 7.9: The posterior distribution of the effectiveness of my recommendations using a uniform prior](image)

Again, the end result of the Bayesian analysis proper is the posterior distribution that illustrates credible values of the parameter. The decision to set an arbitrary threshold for concluding that my recommendations were effective or not is a separate matter.
Let's say that, before the fact, we stated that if .05 or lower were not among the 95% most credible values, we would conclude that my recommendations were effective. How do we know what the credible interval bounds are?

Even though it is relatively straightforward to determine the bounds of the credible interval analytically, doing so ourselves computationally will help us understand how the posterior distribution is summarized in the examples given later in this chapter.

To find the bounds, we will sample from a beta distribution with hyper-parameters 37 and 5 thousands of times and find the quantiles at .025 and .975.

```r
> samp <- rbeta(10000, 37, 5)
> quantile(samp, c(.025, .975))
2.5%  97.5%
0.7674591 0.9597010
```

Neat! With the previous plot already up, we can add lines to the plot indicating this 95% credible interval, like so:

```r
# horizontal line
> lines(c(.767, .96), c(0.1, 0.1)
> # tiny vertical left boundary
> lines(c(.767, .769), c(0.15, 0.05))
> # tiny vertical right boundary
> lines(c(.96, .96), c(0.15, 0.05))
```

If you plot this yourself, you'll see that even the lower bound is far from the decision boundary—it looks like my work was worth it after all!

The technique of sampling from a distribution many many times to obtain numerical results is known as Monte Carlo simulation.

---

**Enter MCMC – stage left**

As mentioned earlier, we started with the coin flip examples because of the ease of determining the posterior distribution analytically—primarily because of the beta distribution's self-conjugacy with respect to the binomial likelihood function.

It turns out that most real-world Bayesian analyses require a more complicated solution. In particular, the hyper-parameters that define the posterior distribution are rarely known. What can be determined is the probability density in the posterior distribution for each parameter value. The easiest way to get a sense of the shape of the posterior is to sample from it many thousands of times. More specifically, we sample from all possible parameter values and record the probability density at that point.
Bayesian Methods

How do we do this? Well, in the case of just one parameter value, it's often computationally tractable to just randomly sample willy-nilly from the space of all possible parameter values. For cases where we are using Bayesian analysis to determine the credible values for two parameters, things get a little more hairy.

The posterior distribution for more than one parameter value is a called a joint distribution; in the case of two parameters, it is, more specifically, a bivariate distribution. One such bivariate distribution can be seen in Figure 7.10:

![Figure 7.10: A bivariate normal distribution](image)

To picture what it is like to sample a bivariate posterior, imagine placing a bell jar on top of a piece of graph paper (be careful to make sure Ester Greenwood isn't under there!). We don't know the shape of the bell jar but we can, for each intersection of the lines in the graph paper, find the height of the bell jar over that exact point. Clearly, the smaller the grid on the graph paper, the higher resolution our estimate of the posterior distribution is.
Note that in the univariate case, we were sampling from $n$ points, in the bivariate case, we are sampling from $n^2$ points ($n$ points for each axis). For models with more than two parameters, it is simply intractable to use this random sampling method. Luckily, there's a better option than just randomly sampling the parameter space: Markov Chain Monte Carlo (MCMC).

I think the easiest way to get a sense of what MCMC is, is by likening it to the game *hot and cold*. In this game—which you may have played as a child—an object is hidden and a searcher is blindfolded and tasked with finding this object. As the searcher wanders around, the other player tells the searcher whether she is *hot* or *cold*; *hot* if she is near the object, *cold* when she is far from the object. The other player also indicates whether the movement of the searcher is getting her closer to the object (getting *warmer*) or further from the object (getting *cooler*).

In this analogy, warm regions are areas were the probability density of the posterior distribution is high, and cool regions are the areas were the density is low. Put in this way, random sampling is like the searcher teleporting to random places in the space where the other player hid the object and just recording how hot or cold it is at that point. The guided behavior of the player we described before is far more efficient at exploring the areas of interest in the space.

At any one point, the blindfolded searcher has no memory of where she has been before. Her next position only depends on the point she is at currently (and the feedback of the other player). A memory-less transition process whereby the next position depends only upon the current position, and not on any previous positions, is called a Markov chain.

The technique for determining the shape of high-dimensional posterior distributions is therefore called Markov chain Monte Carlo, because it uses Markov chains to intelligently sample many times from the posterior distribution (Monte Carlo simulation).

The development of software to perform MCMC on commodity hardware is, for the most part, responsible for a Bayesian renaissance in recent decades. Problems that were, not too long ago, completely intractable are now possible to be performed on even relatively low-powered computers.
There is far more to know about MCMC than we have the space to discuss here. Luckily, we will be using software that abstracts some of these deeper topics away from us. Nevertheless, if you decide to use Bayesian methods in your own analyses (and I hope you do!), I’d strongly recommend consulting resources that can afford to discuss MCMC at a deeper level. There are many such resources, available for free, on the web.

Before we move on to examples using this method, it is important that we bring up this one last point: Mathematically, an infinitely long MCMC chain will give us a perfect picture of the posterior distribution. Unfortunately, we don’t have all the time in the world (universe [?]), and we have to settle for a finite number of MCMC samples. The longer our chains, the more accurate the description of the posterior. As the chains get longer and longer, each new sample provides a smaller and smaller amount of new information (economists call this diminishing marginal returns). There is a point in the MCMC sampling where the description of the posterior becomes sufficiently stable, and for all practical purposes, further sampling is unnecessary. It is at this point that we say the chain converged. Unfortunately, there is no perfect guarantee that our chain has achieved convergence. Of all the criticisms of using Bayesian methods, this is the most legitimate—but only slightly.

There are really effective heuristics for determining whether a running chain has converged, and we will be using a function that will automatically stop sampling the posterior once it has achieved convergence. Further, convergence can be all but perfectly verified by visual inspection, as we’ll see soon.

For the simple models in this chapter, none of this will be a problem, anyway.

**Using JAGS and runjags**

Although it’s a bit silly to break out MCMC for the single-parameter career recommendation analysis that we discussed earlier, applying this method to this simple example will aid in its usage for more complicated models.

In order to get started, you need to install a software program called JAGS, which stands for *Just Another Gibbs Sampler* (a Gibbs sampler is a type of MCMC sampler). This program is independent of R, but we will be using R packages to communicate with it. After installing JAGS, you will need to install the R packages `rjags`, `runjags`, and `modeest`. As a reminder, you can install all three with this command:

```
> install.packages(c("rjags", "runjags", "modeest"))
```
To make sure everything is installed properly, load the runjags package, and run the function testjags(). My output looks something like this:

```r
> library(runjags)
> testjags()
You are using R version 3.2.1 (2015-06-18) on a unix machine,
with the RStudio GUI
The rjags package is installed
JAGS version 3.4.0 found successfully using the command
'/usr/local/bin/jags'
```

The first step is to create the model that describes our problem. This model is written in an R-like syntax and stored in a string (character vector) that will get sent to JAGS to interpret. For this problem, we will store the model in a string variable called `our.model`, and the model looks like this:

```r
our.model <- "
model {
  # likelihood function
  numSuccesses ~ dbinom(successProb, numTrials)

  # prior
  successProb ~ dbeta(1, 1)

  # parameter of interest
  theta <- numSuccesses / numTrials
}
"
```

Note that the JAGS syntax allows for R-style comments, which I included for clarity.

In the first few lines of the model, we are specifying the likelihood function. As we know, the likelihood function can be described with a binomial distribution. The line:

```r
numSuccesses ~ dbinom(successProb, numTrials)
```

says the variable `numSuccesses` is distributed according to the binomial function with hyper-parameters given by variable `successProb` and `numTrials`. 
Bayesian Methods

In the next relevant line, we are specifying our choice of the prior distribution. In keeping with our previous choice, this line reads, roughly: the successProb variable (referred to in the previous relevant line) is distributed in accordance with the beta distribution with hyper-parameters 1 and 1.

In the last line, we are specifying that the parameter we are really interested in is the proportion of successes (number of successes divided by the number of trials). We are calling that theta. Notice that we used the deterministic assignment operator (<-) instead of the distributed according to operator (~) to assign theta.

The next step is to define the successProb and numTrials variables for shipping to JAGS. We do this by stuffing these variables in an R list. We do this as follows:

```r
our.data <- list(
  numTrials = 40,
  successProb = 36/40
)
```

Great! We are all set to run the MCMC.

```r
> results <- autorun.jags(our.model,
+                         data=our.data,
+                         n.chains = 3,
+                         monitor = c('theta'))
```

The function that runs the MCMC sampler and automatically stops at convergence is autorun.jags. The first argument is the string specifying the JAGS model. Next, we tell the function where to find the data that JAGS will need. After this, we specify that we want to run 3 independent MCMC chains; this will help guarantee convergence and, if we run them in parallel, drastically cut down on the time we have to wait for our sampling to be done. (To see some of the other options available, as always, you can run ?autorun.jags.) Lastly, we specify that we are interested in the variable 'theta'.

After this is done, we can directly plot the results variable where the results of the MCMC are stored. The output of this command is shown in Figure 7.11.

```r
> plot(results,
+       plot.type=c("histogram", "trace"),
+       layout=c(2,1))
```
The first of these plots is called a *trace plot*. It shows the sampled values of theta as the chain got longer. The fact that all three chains are overlapping around the same set of values is, at least in this case, a strong guarantee that all three chains have converged. The bottom plot is a bar plot that depicts the relative credibility of different values of theta. It is shown here as a bar plot, and not a smooth curve, because the binomial likelihood function is discrete. If we want a continuous representation of the posterior distribution, we can extract the sample values from the results and plot it as a density plot with a sufficiently large bandwidth:

```r
> # mcmc samples are stored in mcmc attribute
> # of results variable
> results.matrix <- as.matrix(results$mcmc)
> ```
Bayesian Methods

```r
> # extract the samples for 'theta'
> # the only column, in this case
> theta.samples <- results.matrix[, 'theta']
>
> plot(density(theta.samples, adjust=5))
```

And we can add the bounds of the 95% credible interval to the plot as before:

```r
> quantile(theta.samples, c(.025, .975))
2.5% 97.5%
0.800 0.975
> lines(c(.8, .975), c(0.1, 0.1))
> lines(c(.8, .8), c(0.15, 0.05))
> lines(c(.975, .975), c(0.15, 0.05))
```

![Figure 7.12: Density plot of the posterior distribution. Note that the x-axis starts here at 0.6](image)

Rest assured that there is only a disagreement between the two credible intervals' bounds in this example because the MCMC could only sample discrete values from the posterior since the likelihood function is discrete. This will not occur in the other examples in this chapter. Regardless, the two methods seem to be in agreement about the shape of the posterior distribution and the credible values of theta. It is all but certain that my recommendations are better than chance. Go me!
Fitting distributions the Bayesian way

In this next example, we are going to be fitting a normal distribution to the precipitation dataset that we worked with in the previous chapter. We will wrap up with Bayesian analogue to the one sample t-test.

The results we want from this analysis are credible values of the true population mean of the precipitation data. Refer back to the previous chapter to recall that the sample mean was 34.89. In addition, we will also be determining credible values of the standard deviation of the precipitation data. Since we are interested in the credible values of two parameters, our posterior distribution is a joint distribution.

Our model will look a little differently now:

```
the.model <- "
model {
  mu ~ dunif(0, 60)        # prior
  stddev ~ dunif(0, 30)    # prior
  tau <- pow(stddev, -2)

  for(i in 1:theLength){
    samp[i] ~ dnorm(mu, tau)   # likelihood function
  }
}
"
```

This time, we have to set two priors, one for the mean of the Gaussian curve that describes the precipitation data (\( \mu \)), and one for the standard deviation (\( \text{stddev} \)). We also have to create a variable called \( \tau \) that describes the precision (inverse of the variance) of the curve, because \texttt{dnorm} in JAGS takes the mean and the precision as hyper-parameters (and not the mean and standard deviation, like R). We specify that our prior for the \( \mu \) parameter is uniformly distributed from 0 inches of rain to 60 inches of rain—far above any reasonable value for the population precipitation mean. We also specify that our prior for the standard deviation is a flat one from 0 to 30. If this were part of any meaningful analysis and not just a pedagogical example, our priors would be informed in part by precipitation data from other regions like the US or my precipitation data from previous years. JAGS comes chock full of different families of distributions for expressing different priors.

Next, we specify that the variable \( \text{samp} \) (which will hold the precipitation data) is distributed normally with unknown parameters \( \mu \) and \( \tau \).
Bayesian Methods

Then, we construct an R list to hold the variables to send to JAGS:

```r
the.data <- list(
    samp = precip,
    theLength = length(precip)
)
```

Cool, let’s run it! On my computer, this takes 5 seconds.

```r
> results <- autorun.jags(the.model,
+                         data=the.data,
+                         n.chains = 3,
+                         # now we care about two parameters
+                         monitor = c('mu', 'stddev'))
```

Let’s plot the results directly like before, while being careful to plot both the trace plot and histogram from both parameters by increasing the layout argument in the call to the plot function.

```r
> plot(results,
+      plot.type=c("histogram", "trace"),
+      layout=c(2,2))
```

Figure 7.13: Output plots from the MCMC result of fitting a normal curve to the built-in precipitation data set
Figure 7.14 shows the distribution of credible values of the \( \mu \) parameter without reference to the \( \text{stddev} \) parameter. This is called a marginal distribution.

![Figure 7.14: Marginal distribution of posterior for parameter 'mu'. Dashed line shows hypothetical population mean within 95% credible interval.](image)

Remember when, in the last chapter, we wanted to determine whether the US' mean precipitation was significantly discrepant from the (hypothetical) known population mean precipitation of the rest of the world of 38 inches. If we take any value outside the 95% credible interval to indicate significance, then, just like when we used the NHST t-test, we have to reject the hypothesis that there is significantly more or less rain in the US than in the rest of the world.

Before we move on to the next example, you may be interested in credible values for both the mean and the standard deviation at the same time. A great type of plot for depicting this information is a contour plot, which illustrates the shape of a three-dimensional surface by showing a series of lines for which there is equal height. In Figure 7.15, each line shows the edges of a slice of the posterior distribution that all have equal probability density.

```r
> results.matrix <- as.matrix(results$mcmc)
> 
> library(MASS)
> # we need to make a kernel density
```
Bayesian Methods

> # estimate of the 3-d surface
> z <- kde2d(results.matrix[, 'mu'],
>            results.matrix[, 'stddev'],
>            n=50)
> plot(results.matrix)
> contour(z, drawlabels=FALSE,
>         nlevels=11, col=rainbow(11),
>         lwd=3, add=TRUE)

Figure 7.15: Contour plot of the joint posterior distribution. The purple contour corresponds to the region with the highest probability density

The purple contours (the inner-most contours) show the region of the posterior with the highest probability density. These correspond to the most likely values of our two parameters. As you can see, the most likely values of the parameters for the normal distribution that best describes our present knowledge of US precipitation are a mean of a little less than 35 and a standard deviation of a little less than 14. We can corroborate the results of our visual inspection by directly printing the `results` variable:

> print(results)

JAGS model summary statistics from 30000 samples (chains = 3; adapt+burnin = 5000):
which also shows other summary statistics from our MCMC samples and some information about the MCMC process.

The Bayesian independent samples t-test

For our last example in the chapter, we will be performing a sort-of Bayesian analogue to the two-sample t-test using the same data and problem from the corresponding example in the previous chapter—testing whether the means of the gas mileage for automatic and manual cars are significantly different.

As before, let's specify the model using non-informative flat priors:

```r
the.model <- "
model {
  # each group will have a separate mu
  # and standard deviation
  for(j in 1:2){
    mu[j] ~ dunif(0, 60) # prior
    stddev[j] ~ dunif(0, 20) # prior
    tau[j] <- pow(stddev[j], -2)
  }
  for(i in 1:theLength){
    # likelihood function
    y[i] ~ dnorm(mu[x[i]], tau[x[i]])
  }
}
"
```

Notice that the construct that describes the likelihood function is a little different now; we have to use nested subscripts for the `mu` and `tau` parameters to tell JAGS that we are dealing with two different versions of `mu` and `stddev`. 

---

[165]
Next, the data:

```r
the.data <- list(
  y = mtcars$mpg,
  # 'x' needs to start at 1 so
  # 1 is now automatic and 2 is manual
  x = ifelse(mtcars$am==1, 1, 2),
  theLength = nrow(mtcars)
)
```

Finally, let's roll!

```r
> results <- autorun.jags(the.model,
+                         data=the.data,
+                         n.chains = 3,
+                         monitor = c('mu', 'stddev'))
```

Let's extract the samples for both 'mu's and make a vector that holds the differences in the mu samples between each of the two groups.

```r
> results.matrix <- as.matrix(results$mcmc)
> difference.in.means <- (results.matrix[,1] – results.matrix[,2])
```

*Figure 7.16* shows a plot of the credible differences in means. The likely differences in means are far above a difference of zero. We are all but certain that the means of the gas mileage for automatic and manual cars are significantly different.

![Credible values for the difference in means](image)

*Figure 7.16: Credible values for the difference in means of the gas mileage between automatic and manual cars. The dashed line is at a difference of zero*
Notice that the decision to mimic the independent samples t-test made us focus on one particular part of the Bayesian analysis and didn't allow us to appreciate some of the other very valuable information the analysis yielded. For example, in addition to having a distribution illustrating credible differences in means, we have the posterior distribution for the credible values of both the means and standard deviations of both samples. The ability to make a decision on whether the samples’ means are significantly different is nice—the ability to look at the posterior distribution of the parameters is better.

**Exercises**

Practise the following exercises to reinforce the concepts learned in this chapter:

- Write a function that will take a vector holding MCMC samples for a parameter and plot a density curve depicting the posterior distribution and the 95% credible interval. Be careful of different scales on the y-axis.

- Fitting a normal curve to an empirical distribution is conceptually easy, but not very robust. For distribution fitting that is more robust to outliers, it’s common to use a t-distribution instead of the normal distribution, since the t has heavier tails. View the distribution of the shape attribute of the built-in rock dataset. Does this look normally distributed? Find the parameters of a normal curve that is a fit to the data. In JAGS, dt, the t-distribution density function, takes three parameters: the mean, the precision, and the degrees of freedom that controls the heaviness of the tails. Find the parameters after fitting a t-distribution to the data. Are the means similar? Which estimate of the mean do you think is more representative of central tendency?

- In Theseus’ paradox, a wooden ship belonging to Theseus has decaying boards, which are removed and replaced with new lumber. Eventually, all the boards in the original ship have been replaced, so that the ship is made up of completely new matter. Is it still Theseus’ ship? If not, at what point did it become a different ship?

What would Aristotle say about this? Appeal to the doctrine of the *Four Causes*. Would Aristotle’s stance still hold up if—as in Thomas Hobbes’ version of the paradox—the original decaying boards were saved and used to make a complete replica of Theseus’ original ship?
Summary

Although most introductory data analysis texts don't even broach the topic of Bayesian methods, you, dear reader, are versed enough in this matter to start applying these techniques to real problems.

We discovered that Bayesian methods could—at least for the models in this chapter—not only allow us to answer the same kinds of questions we might use the binomial, one sample t-test, and the independent samples t-test for, but provide a much richer and more intuitive depiction of our uncertainty in our estimates.

If these approaches interest you, I urge you to learn more about how to extend these to supersede other NHST tests. I also urge you to learn more about the mathematics behind MCMC.

As with the last chapter, we covered much ground here. If you made it through, congratulations!

This concludes the unit on confirmatory data analysis and inferential statistics. In the next unit, we will be concerned less with estimating parameters, and more interested in prediction. Last one there is a rotten egg!
Where to buy this book

You can buy Data Analysis with R from the Packt Publishing website.

Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.

Click here for ordering and shipping details.