so handy

o deter-
rve from
ge xlow
the area

pose you
centered
between

3.11 can be
' theorem.]
total sam-
h graders.
dents who

ies of grade
re indepen-
(grade) and

Bayes' Rule

CONTENTS

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

If you see that there are clouds, what is the probability that soon there will be rain? If you know that it is raining, by hearing it patter on the roof, what is the probability that there are clouds? Notice that p(clouds | rain) is not equal

Doing Bayesian Data Analysis: A Tutorial with R and BUGS. DOI: 10.1016/B978-0-12-381485-2.00004-3

© 2011, Elsevier Inc. All rights reserved.

to p(rain | clouds). If someone smiles at you, what is the probability that they love you? If someone loves you, what is the probability that they will smile at you? Notice that p(smile | love) is not equal to p(love | smile).

Let's consider an example for which we can determine specific numbers. Suppose I have a standard deck of playing cards, which has 52 cards altogether. There are four suits: hearts, diamonds, clubs, and spades. Within each suit, there are 13 values: ace, two, three,. . . , ten, jack, queen, and king. I shuffle the cards and draw one at random without showing it to you. I look at the card, and tell you (truthfully) that it is a queen. Given that you know it is a queen, what is the probability that it is a heart? Think about it a moment: There are four queens in the deck, and only one of them is a heart. So the probability that the card is a heart is 1/4. We can write this as a conditional probability: $p(\heartsuit \mid Q) = \frac{1}{4}$.

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

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

4.1 BAYES' RULE

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

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

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

4.1.1 Derived from Definitions of Conditional Probability

Recall from the definition of conditional probability, back in Equations 3.11 and 3.12 on p. 45, that $p(y \mid x) = p(y, x)/p(x)$. In words, the definition simply says that the probability of y given x is the probability that they happen together relative to the probability that x happens at all. We used this definition quite naturally when computing the conditional probabilities for the example, presented earlier, regarding hearts and queens in a deck of cards.

Now we just do some very simple algebraic manipulations. First, multiply both sides of $p(y \mid x) = p(y, x)/p(x)$ by $p(x)$ to get $p(y \mid x)p(x) = p(y, x)$. Second, notice that we can do the analogous manipulation starting with $p(x \mid y) = p(y, x)/p(y)$ to get $p(x \mid y)p(y) = p(y, x)$. Now we have two different expressions equal to $p(y, x)$, so we know those expressions equal each other: $p(y \mid x)p(x) = p(x \mid y)p(y)$. Divide both sides of that last expression by $p(x)$ to arrive at

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

But we are not done yet, because we can rewrite the denominator in terms of $p(x \mid y)$ also. Toward that goal, recall that $p(x) = \sum_y p(x, y)$. That was Equation 3.9, on p. 44, if you're keeping score. We also know that $p(x, y) = p(x \mid y)p(y)$. Combining those equations yields $p(x) = \sum_y p(x, y) = \sum_y p(x \mid y)p(y)$. Substitute that into the denominator of Equation 4.1 to get

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

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

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

4.1.2 Intuited from a Two-Way Discrete Table

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

The suit and value on playing cards are independent. (The idea of independent attributes was discussed in Section 3.4.3.) Let's now confirm Bayes' rule for two attributes that are not independent. Recall the case of tossing a coin three times and counting the number of heads and the number of switches between heads and tails, as tabulated back in Table 3.3 (p. 43), and repeated here for convenience:

	Number of Heads				Marginal
Number of Switches	**0**	**1**	**2**	**3**	**(Number of Switches)**
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
Marginal (Number of Heads)	1/8	3/8	3/8	1/8	

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

A valuable intuition, for understanding conditional probabilities and Bayes' rule, comes from restricting our spatial attention to a single row or column of the conjoint probability table. Suppose someone tosses a coin three times and tells us that the sequence contains one switch. Given that knowledge, we can restrict our attention to the row of the table corresponding to one switch. We know that one of the conjoint events *within that row* must have happened, but we don't know which one. The relative probabilities of events within that row have not changed, but we know that the total probability within that row must now sum to 1.0. To achieve that transformation mathematically, we simply divide the cell probabilities in the one-switch row by its original row total. This preserves the relative probabilities within the row but makes the total probability equal to 1.0. Dividing a set of values by their sum is called *normalizing* the values. When we normalize the cell probabilities in the i^{th} row, we get the conditional probabilities of the columns, given the row value. In particular, when we normalize the one-switch row, we get the conditional probabilities for number of heads: $p(0H \mid 1S) = 0/(4/8) = 0$, $p(1H \mid 1S) = (2/8)/(4/8) = 0.5$, $p(2H \mid 1S) = (2/8)/(4/8) = 0.5$, and $p(3H \mid 1S) = 0/(4/8) = 0$.

The idea of restricting attention to a single column or row of the conjoint probability table yields a way of intuiting Bayes' rule in general. The key to Bayes' rule is to notice, from the definition of conditional probability (Equations 3.11 and 3.12 on p. 45), that the conjoint probability of the i^{th} row (R_i) and the j^{th} column (C_j) can be reexpressed either as $p(R_i \mid C_j)p(C_j)$ or as $p(C_j \mid R_i)p(R_i)$. These alternative expressions of the conjoint probability $p(R_i, C_j)$ have been entered into the i, j^{th} cell of Table 4.1.

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

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

Table 4.1 Making Bayes' Rule Not Merely Special but Spatial

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

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

4.1.3 The Denominator as an Integral over Continuous Values

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

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

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

4.2 APPLIED TO MODELS AND DATA

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

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

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

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

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

It helps to think about the application of Bayes' rule to data and models in terms of a two-way table, shown in Table 4.2. The columns of Table 4.2 correspond to specific values of the model parameter, and the rows of Table 4.2 correspond to specific values of the data. Each cell of the table holds the conjoint probability of the specific combination of parameter value θ and data value D.

Tab
Moc

Data

D v

Mar

That is, $p(D, \theta)$ is
value and param
parameter values

The prior probab
which appears ir
of each possible

When we obser
restricting our at
sponding to the
by dividing the
Thus, the poster
row, normalized

We need to defir
rule have names

where the evider

The "prior," $p(\theta$
"posterior," $p(\theta \mid$
taken into accou
could be generat
is the probabilit

Table 4.2 Applying Bayes' Rule to Data and Model Parameter

		Model Parameter		
Data	⋯	**θ value**	⋯	**Marginal**
⋮		⋮		
D value	⋯	$p(D,\theta)$ $= p(D \mid \theta)p(\theta)$ $= p(\theta \mid D)p(D)$	⋯	$p(D)$
⋮		⋮		
Marginal		$p(\theta)$		

That is, $p(D,\theta)$ is the probability of getting that particular combination of data value and parameter value, across all possible combinations of data values and parameter values.

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

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

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

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

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

$$p(D) = \int d\theta \; p(D \mid \theta)p(\theta) \tag{4.5}$$

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

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

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

where the evidence is

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

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

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

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

Terminological aside: The quantity $p(D \mid M)$, which is called the *evidence* in this book, is sometimes instead called the *marginal likelihood* or *prior predictive* by other authors. The term "evidence" is common in the machine learning literature (e.g., Bishop, 2006, MacKay, 2003). Whenever I refer to the "evidence" for a model, I am referring to $p(D \mid M)$ as defined in Equation 4.7. This usage might be a little confusing in the context of model comparison when considering the equation $p(M1 \mid D) = p(D \mid M1)p(M1)/p(D)$, where $p(D \mid M1)$ plays the *role* of the likelihood, not the evidence. This apparent confusion is cleared up when abbreviated terminology is expanded to its full specificity. The factor $p(D \mid M)$ is not merely "the evidence," it is "the evidence for model M." On the other hand, the factor $p(D)$, in the context of the equation $p(M1 \mid D) =$

$p(D\,|\,M1)p(M1)/p(D)$, is not the evidence for *a* model but is the evidence for the entire *set* of models under consideration: $p(D) = \sum_i p(D\,|\,M_i)p(M_i)$. The term "likelihood" also deserves expansion. In Equation 4.6, the likelihood is more fully stated as "the likelihood of parameter value θ in model M for data D." That is, the likelihood is referring to the parameter θ. On the other hand, in the context of model comparison, the factor $p(D\,|\,M1)$, in the equation $p(M1\,|\,D) = p(D\,|\,M1)p(M1)/p(D)$, is the "likelihood of the *model* $M1$ for the data D." To reiterate, the term "evidence" is merely a word to refer to $p(D\,|\,M)$. As we will see, its value does not have much meaning by itself. Instead, $p(D\,|\,M)$ can only be interpreted in the context of other models.

4.2.1 Data Order Invariance

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

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

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

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

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

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

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

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

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

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

4.2.2 An Example with Coin Flipping

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

First, we specify our prior beliefs. We denote the bias as $\theta = p(H)$, the probability of the coin coming up heads. To keep the example straightforward, suppose that we believe there are only three possible values for the coin's bias. Either the coin is fair, with $\theta = 0.50$, or the coin is biased, with $\theta = 0.25$ or $\theta = 0.75$. We believe that the coin is probably fair, but there's some smaller chance it could be biased high or low. This prior probability is graphed in the top panel of Figure 4.1. It shows three "spikes," one over each value of θ that we think could be possible. The spike over $\theta = 0.5$ is tallest, indicating that we believe it to be most likely. Note that the heights of the spikes are probability masses, not densities, because each spike indicates the probability of its specific, discrete value of θ.

Next, we flip the coin to get some data, D, and determine the likelihood, $p(D \mid \theta)$. Suppose we flip the coin 12 times and it comes up heads 3 times. According to our model of the coin, the probability of coming up heads is θ, and the probability of coming up tails is $1 - \theta$. Moreover, the flips are independent of each other, and therefore we can multiply the probabilities of the individual flips to get the probability of the combination of flips. Consequently, the probability of a specific sequence of three heads and nine tails is $p(D \mid \theta) = \theta^3 (1 - \theta)^9$. The resulting likelihood for each value of θ is plotted in the middle panel of Figure 4.1. Notice that the likelihood is highest for

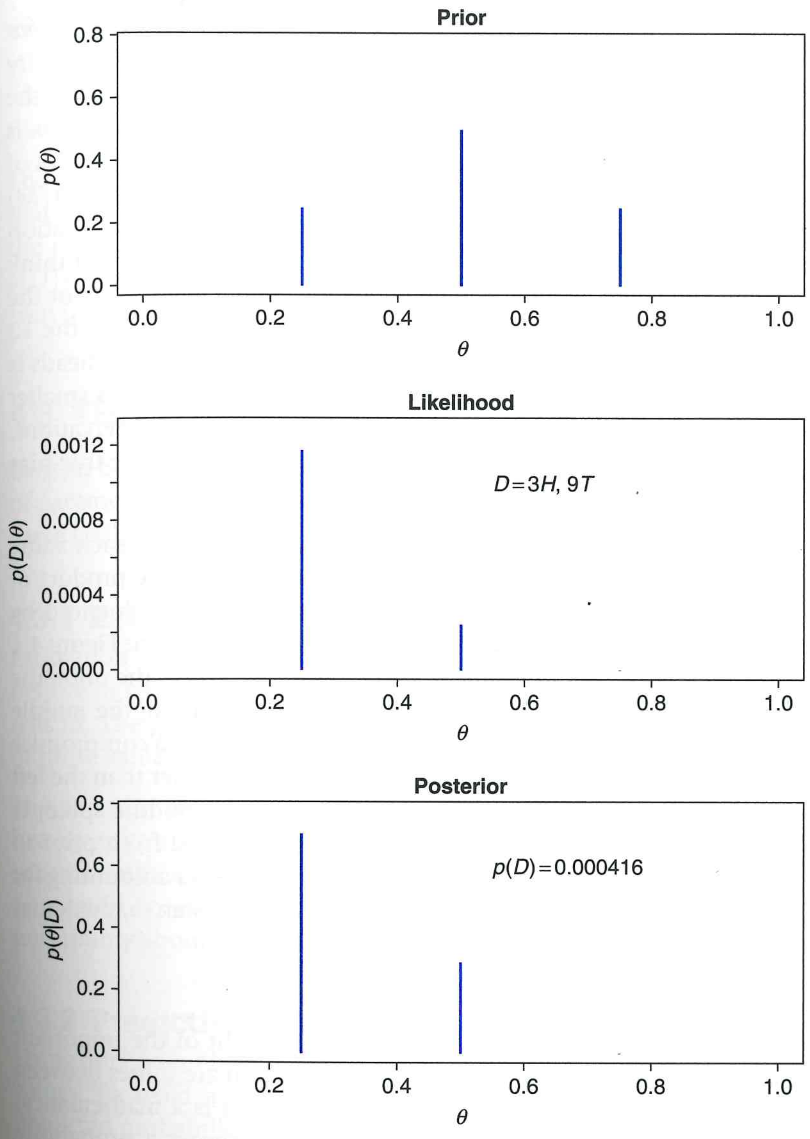


FIGURE 4.1
Bayesian updating of beliefs about the bias of a coin. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. (The R code that generated this graph is in Section 4.4.1 (`BayesUpdate.R`).)

$\theta = 0.25$ and lowest for $\theta = 0.75$. This peak at $\theta = 0.25$ makes sense, because the data have 25% heads, so they are more likely if $\theta = 0.25$ than if $\theta = 0.50$ or $\theta = 0.75$. The value of θ that maximizes the likelihood is called the *maximum likelihood estimate* of θ.

The lower panel of Figure 4.1 includes the value of $p(D \mid M)$, the evidence for the model. Recall from Equation 4.7 that the evidence is the overall probability of the data, averaging across the available parameter values weighted by the degree to which we believe in them: $p(D \mid M) = \sum_\theta p(D \mid \theta, M) p(\theta, M)$. This is the normalizer for the posterior distribution, hence it is displayed in the plot of the posterior distribution. The value is displayed as $p(D)$ instead of as $p(D \mid M)$ because there is only one model in this context, and therefore the M notation is suppressed. When you see the value of $p(D)$ in Figure 4.1, you might think that $p(D)$ is terribly small, until you remember that we are talking about the conjoint probability of several things happening together (i.e., exactly the 12 flips we observed). The probability of 1 head is θ. The probability of 2 heads is θ^2, which is smaller than θ. The probability of 3 heads is θ^3, which is smaller yet. As the set of data D gets bigger, in terms of containing more observations, $p(D)$ gets smaller, regardless of how closely the model θ matches the true bias in the coin.

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

4.2.2.1 $p(D \mid \theta)$ Is Not θ

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

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

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

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

4.3 THE THREE GOALS OF INFERENCE

Back in Section 2.2 (p. 12), I introduced three goals of inference: estimation of parameter values, prediction of data values, and model comparison. Each of these goals are now given precise mathematical expressions.

4.3.1 Estimation of Parameter Values

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

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

4.3.2 Prediction of Data Values

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

$$p(y) = \int d\theta p(y \mid \theta) p(\theta)$$

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

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

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

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

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

4.3.3 Model Comparison

You may recall from earlier discussion (p. 58) that Bayes' rule is also useful for comparing models. Equation 4.8 indicated that the posterior beliefs in the models involve the evidences of the models. Notice that in this third goal (i.e., model comparison), the evidence term appears again, just as it appeared for the goals of parameter estimation and data prediction.

One of the nice features of Bayesian model comparison is that there is an automatic accounting for model complexity when assessing the degree to which we should believe in the model. This might be best explained with an example. Recall the coin-flipping example discussed earlier, illustrated in Figure 4.1 and reproduced in the left side of Figure 4.2. In that example, we supposed that the bias θ could take on only three possible values. This restriction made the model rather simple. We could instead entertain a more complex model that allows for many more possible values of θ. One such model is illustrated in

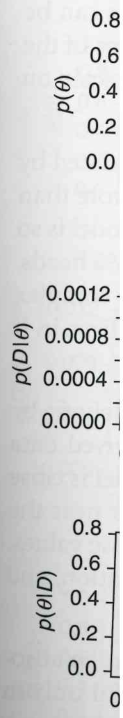


FIGURE 4.2
A simple mo
distributions
addressed b
the lower-left
the lower-righ
that generate

the right sid
only 3. The
triangular s
with lesser b

The complex
more opport
flips has 37%
that outcom
that are in bo
values in the
there are so m

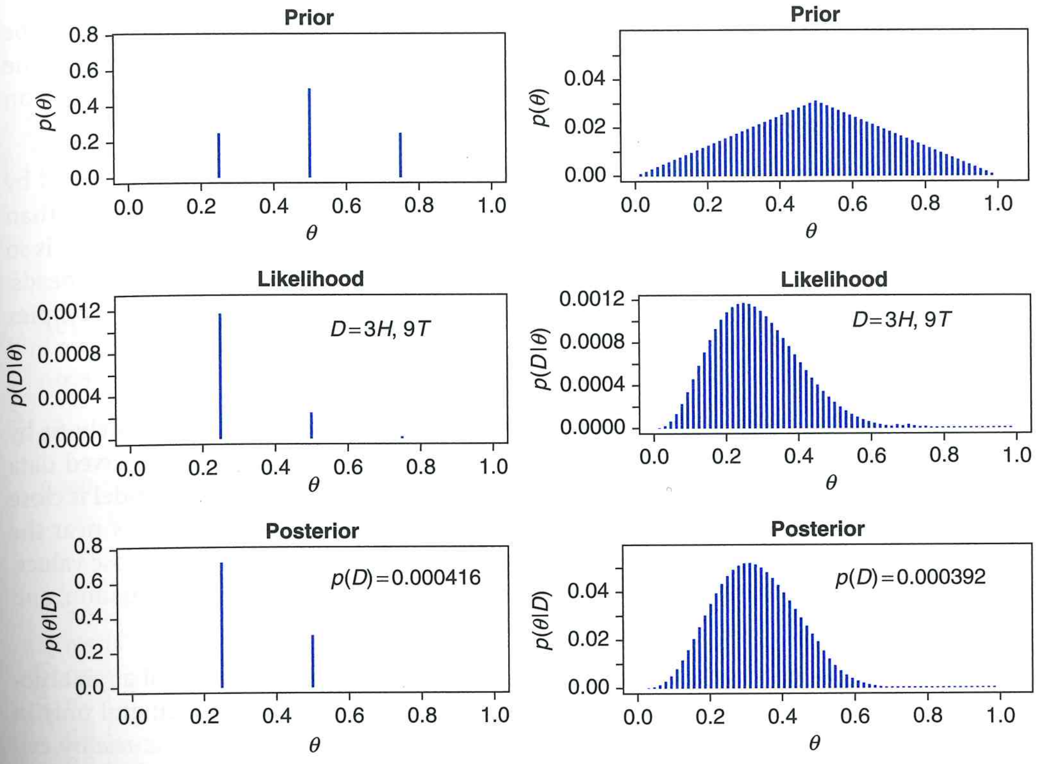


FIGURE 4.2
A simple model in the left column and a complex model in the right column. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. The same data are addressed by both models. The evidence $p(D \mid M_{\text{simple}})$ for the simple model is displayed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

the right side of Figure 4.2. This model has 63 possible values of θ instead of only 3. The shape of the prior beliefs in the complex model follows the same triangular shape as in the simple model; there is highest belief in $\theta = 0.50$, with lesser belief in more extreme values.

The complex model has many more available values for θ, and so it has much more opportunity to fit arbitrary data sets. For example, if a sequence of coin flips has 37% heads, the simple model does not have a θ value very close to that outcome, but the complex model does. On the other hand, for θ values that are in both the simple and complex models, the prior probability on those values in the simple model is much higher than in the complex model. Because there are so many possibilities in the complex model, the prior beliefs have to get spread out, very shallowly, over a larger range of possibilities. This can be seen in Figure 4.2 by inspecting the numerical scales on the vertical axes of the prior beliefs. The scale on the simple model is much larger than the scale on the complex model.

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

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

The evidence for a model, $p(D \mid M)$, is not particularly meaningful as an absolute magnitude for a single model. The evidence is most meaningful only in the context of the Bayes factor, $p(D \mid M1)/p(D \mid M2)$, which is the *relative* evidence for two models, when considering an observed data set D.[2] Regardless of which model wins, the winning model does not need to be a good model of the data. The model comparison process merely tells us about the *relative* evidence for each model. The winning model is better than the other models in the competition, but the winning model might merely be less bad than the horrible competitors. In later chapters we will explore ways to assess whether the winning model is actually a viable model of the data.

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

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

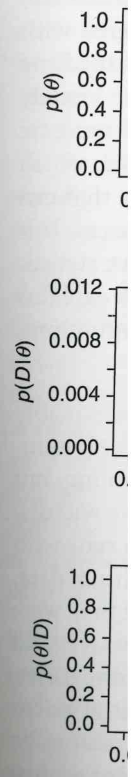


FIGURE 4.3
A simple mo
distributions
addressed b
the lower-left
the lower-rig
that generate

4.3.4 Wh

All three go
which usua
of this diffic
jugate" prio
function sin
tional form
doesn't wor
functions th
is reasonab

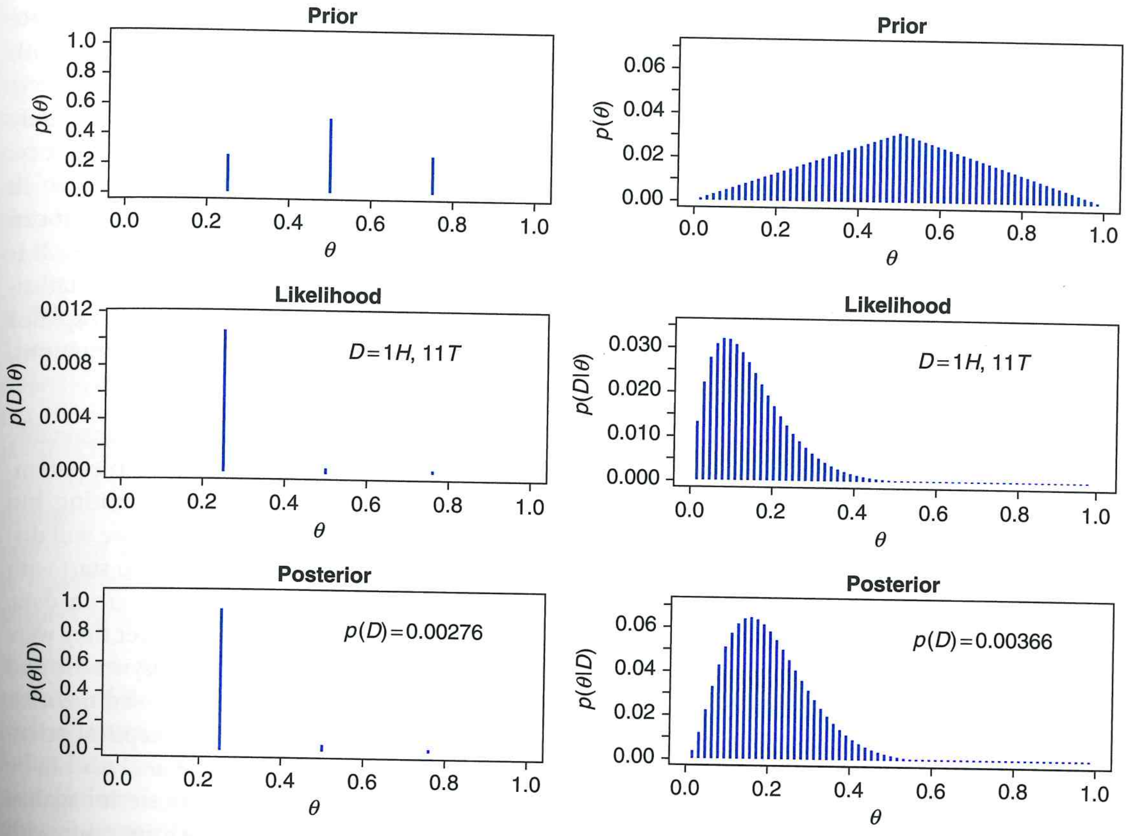


FIGURE 4.3
A simple model in the left column and a complex model in the right column. The prior and posterior distributions indicate probability masses at discrete candidate values of θ. The same data are addressed by both models. The evidence $p(D \mid M_{\text{simple}})$ for the simple model is displayed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

4.3.4 Why Bayesian Inference Can Be Difficult

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

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

4.3.5 Bayesian Reasoning in Everyday Life

4.3.5.1 *Holmesian Deduction*

Despite the difficulty of exact Bayesian inference in complex mathematical models, the essence of Bayesian reasoning is frequently used in everyday life. One example has been immortalized in the words of Sherlock Holmes to his friend Dr. Watson: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Arthur Conan Doyle, *The Sign of Four*, 1890, Chapter 6). This reasoning is actually a consequence of Bayesian belief updating, as expressed in Equation 4.4. Let me restate it this way: "How often have I said to you that when $p(D \mid \theta_i) = 0$ for all $i \neq j$, then, no matter how small the prior $p(\theta_j) > 0$ is, the posterior $p(\theta_j \mid D)$ must equal one." Somehow it sounds better the way Holmes said it. The intuition behind Holmes's deduction is clear, though: When we reduce belief in some possibilities, we necessarily increase our belief in the remaining possibilities (*if* our set of possibilities exhausts all conceivable options). Thus, according to Holmesian deduction, when the data make some options less believable, we increase belief in the other options.

4.3.5.2 Judicial Exoneration

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

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

4.4 R CODE

4.4.1 R Code for Figure 4.1

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

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

Then try

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

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

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

(BayesUpdate.R)

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

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

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

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

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

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

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

# Plot the likelihood:
plot( Theta , pDataGivenTheta , type="h" , lwd=3 , main="Likelihood" ,
      xlim=c(0,1) , xlab=bquote(theta) ,
```

```
          y
          c
text( .
          b

# Plot
plot( T
          x
          y
          c
text( .
          b
```

4.5 EXE

Exercise 4.1 **important rol** the probabili true presence have the valu is absent. Th This is our pr

Suppose that that if a perso We denote a The observed about the val as $p(D= + |$ means that 5 indicates that $+|\theta=$☺$)=0$

Suppose we s the test, and i person has th $p(\theta=$☹$|D=+$ intuitive answ people have ar hit rate of the

Hint: The foll stand the poss case of Table bottom margi our attention t

```
        ylim=c(0,1.1*max(pDataGivenTheta)) , ylab=bquote(paste("p(D|",theta,")")),
        cex.axis=1.2 , cex.lab=1.5 , cex.main=1.5 )
text( .55 , .85*max(pDataGivenTheta) , cex=2.0 ,
        bquote( "D=" * .(nHeads) * "H," * .(nTails) * "T" ) , adj=c(0,.5) )

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

4.5 EXERCISES

Exercise 4.1. **[Purpose: Application of Bayes' rule to disease diagnosis, to see the important role of prior probabilities.]** Suppose that in the general population, the probability of having a particular rare disease is 1 in a 1000. We denote the true presence or absence of the disease as the value of a parameter, θ, that can have the value $\theta = \ddot\frown$ if the disease is present, or the value $\theta = \ddot\smile$ if the disease is absent. The base rate of the disease is therefore denoted $p(\theta = \ddot\frown) = 0.001$. This is our prior belief that a person selected at random has the disease.

Suppose that there is a test for the disease that has a 99% hit rate, which means that if a person has the disease, then the test result is positive 99% of the time. We denote a positive test result as $D = +$ and a negative test result as $D = -$. The observed test result is a bit of data that we will use to modify our belief about the value of the underlying disease parameter. The hit rate is expressed as $p(D = + \mid \theta = \ddot\frown) = 0.99$. The test also has a false alarm rate of 5%. This means that 5% of the time when the disease is not present, the test falsely indicates that the disease is present. We denote the false alarm rate as $p(D = + \mid \theta = \ddot\smile) = 0.05$.

Suppose we sample a person at random from the population, administer the test, and it comes up positive. What is the posterior probability that the person has the disease? Mathematically expressed, we are asking, what is $p(\theta = \ddot\frown \mid D = +)$? Before determining the answer from Bayes' rule, generate an intuitive answer and see if your intuition matches the Bayesian answer. Most people have an intuition that the probability of having the disease is near the hit rate of the test (which in this case is 0.99).

Hint: The following table of conjoint probabilities might help you understand the possible combinations of events. (The following table is a specific case of Table 4.2, p. 57.) The prior probabilities of the disease are on the bottom marginal. When we know that the test result is positive, we restrict our attention to the row marked $D = +$.

	$\theta = \ddot{\frown}$	$\theta = \ddot{\smile}$	
$D = +$	$p(D=+,\theta=\ddot{\frown})$ $= p(D=+\mid\theta=\ddot{\frown})p(\theta=\ddot{\frown})$	$p(D=+,\theta=\ddot{\smile})$ $= p(D=+\mid\theta=\ddot{\smile})p(\theta=\ddot{\smile})$	$p(D=+)$
$D = -$	$p(D=-,\theta=\ddot{\frown})$ $= p(D=-\mid\theta=\ddot{\frown})p(\theta=\ddot{\frown})$	$p(D=-,\theta=\ddot{\smile})$ $= p(D=-\mid\theta=\ddot{\smile})p(\theta=\ddot{\smile})$	$p(D=-)$
	$p(\theta=\ddot{\frown})$	$p(\theta=\ddot{\smile})$	

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

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

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

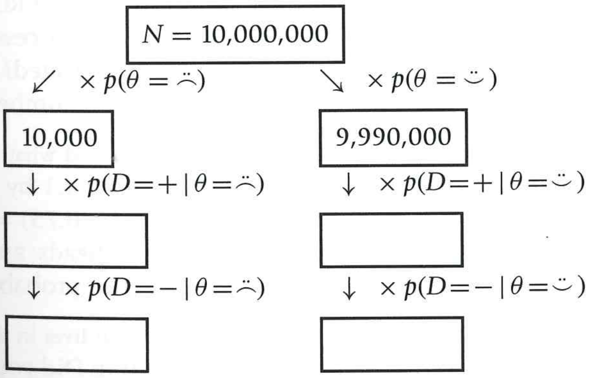
(A) Suppose that the population consists of 100,000 people. Compute how many people should fall into each cell of the table in the hint shown in Exercise 4.1. To compute the expected frequency of people in a cell, just multiply the cell probability by the size of the population. To get you started, a few of the cells of the frequency table are filled in here:

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

Notice the frequencies on the lower margin of the table. They indicate that out of 100,000 people, only 100 have the disease, whereas 99,900 do not have the disease. These marginal frequencies instantiate the prior probability that $p(\theta = \ddot\frown) = 0.001$. Notice also the cell frequencies in the column $\theta = \ddot\frown$, which indicate that of 100 people with the disease, 99 have a positive test result and 1 has a negative test result. These cell frequencies instantiate the hit rate of 0.99. Your job for this part of the exercise is to fill in the frequencies of the remaining cells of the table.

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

(C) Now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumulate more data. Krauss, Martignon, & Hoffrage (1999) called this type of representation a *Markov* representation. Suppose now we start with a population of $N = 10{,}000{,}000$ people. We expect 99.9% of them (i.e., 9,990,000) not to have the disease, and just 0.1% (i.e., 10,000) to have the disease. Now consider how many people we expect to test positive. Of the 10,000 people who have the disease, 99% (i.e., 9900), will be expected to test positive. Of the 9,990,000 people who do not have the disease, 5%, (i.e., 499,500) will be expected to test positive. Now consider retesting everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest? Use this diagram to compute your answer:



$N = 10{,}000{,}000$
$\times\, p(\theta = \ddot\frown)$
$\times\, p(\theta = \ddot\smile)$
10,000
9,990,000
$\times\, p(D{=}+ \mid \theta{=}\ddot\frown)$
$\times\, p(D{=}+ \mid \theta{=}\ddot\smile)$
$\times\, p(D{=}- \mid \theta{=}\ddot\frown)$
$\times\, p(D{=}- \mid \theta{=}\ddot\smile)$

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

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

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

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

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

Suppose that a new study is conducted and finds that the ROI is activated. If the prior probability that the task involves language processing is 0.5, what is the posterior probability, given that the ROI is activated? (Hint: Poldrack (2006) reports that it is 0.69. Your job is to derive this number.)

Exercise 4.6. **[Purpose: To make sure you really understand what is being shown in Figure 4.1.]** Derive the posterior distribution in Figure 4.1 by hand. The prior has $p(\theta=0.25) = 0.25$, $p(\theta=0.50) = 0.50$, and $p(\theta=0.75) = 0.25$. The data consist of a specific sequence of flips with three heads and nine tails, so $p(D\,|\,\theta) = \theta^3\,(1-\theta)^9$. Hint: Check that your posterior probabilities sum to 1.

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

CHAPTER 18

Metric Predicted Variable with One Nominal Predictor

CONTENTS

Familywise error rates breed rumors of incest,
Hounding for quarry in multiple t tests.
Barking at research, poor dog got run over;
Should have done Bayesian oneway ANOVA.

In this chapter we consider situations with a metric predicted variable and a nominally scaled predictor variable. These cases occur frequently in real-world research. For example, we may want to predict weight loss (a metric variable) as a function of which diet the person follows (e.g., low-carb, vegetarian, or low-fat). As another example, we may want to predict severity of psychosis (measured on a metric scale) as a function of which antipsychotic drug the person takes. Or we may want to predict income as a function of political party affiliation. This combination of predicted and predictor scale types occurs in the first row, fourth cell, of Table 14.1 (p. 385).

Doing Bayesian Data Analysis: A Tutorial with R and BUGS. DOI: 10.1016/B978-0-12-381485-2.00018-3

© 2011, Elsevier Inc. All rights reserved.

In traditional NHST analyses, these situations are addressed by "oneway analysis of variance" (ANOVA). The term *oneway* refers to the fact that a single nominal variable is being used as the predictor. The phrase *analysis of variance* refers to the fact that the overall variance across all the data is decomposed (i.e., analyzed) into two parts: variance within the levels of the nominal predictors and variance between the levels of the nominal predictors. The variance within levels of the nominal predictor is called *noise* or *error* (i.e., variability that cannot be predicted by the predictor). The complementary variance between the levels of the nominal predictor is called the *effect* of the predictor. Usually we do the research with the goal of detecting an effect, which means that we would like the magnitude of the variance between levels to be large compared to the noise within levels. The ratio of variance between to variance within is called the *F-ratio*. In the Bayesian approach, we rarely if ever refer to the F-ratio. But because the model we use is based on the model of traditional ANOVA, we will refer to our analysis as Bayesian ANOVA or sometimes BANOVA.

18.1 BAYESIAN ONEWAY ANOVA

The basic idea of oneway ANOVA was introduced in Section 14.1.6.1, p. 368. The predictor is a variable measured on a nominal scale. For example, if income is predicted as a function of political party affiliation, notice that the predictor has nominal levels such as libertarian, green, democratic, republican, and so on. We denote the predictor variable as $\overrightarrow{x}$, which is a vector with one component per nominal level. For example, suppose that the predictor is political party affiliation, with Green as level 1, Democrat as level 2, Republican as level 3, Libertarian as level 4, and Other as level 5. Then Democrat is represented as $\overrightarrow{x} = \langle 0, 1, 0, 0, 0\rangle$, and Libertarian is represented as $\overrightarrow{x} = \langle 0, 0, 0, 1, 0\rangle$. Political party affiliation is being treated here as a categorical label only, with no ordering along a liberal-conservative scale.

The formal model indicates how to derive the predicted value from the predictor. The idea is that there is a baseline quantity of the predicted variable, and each level of the predictor indicates a deflection above or below that baseline. We will denote the baseline value of the prediction as β_0. The deflection for the j^{th} level of the predictor is denoted β_j. When the predictor has value $\overrightarrow{x}_i = \langle \ldots, x_{ji}, \ldots\rangle$, then the predicted value is

$$\begin{aligned}\mu_i &= \beta_0 + \sum_j \beta_j x_{ji} \\ &= \beta_0 + \overrightarrow{\beta} \cdot \overrightarrow{x}_i \qquad (18.1)\end{aligned}$$

where the notation $\overrightarrow{\beta} \cdot \overrightarrow{x}$ denotes the *dot product* of the vectors. In Equation 18.1, the coefficient β_j indicates how much μ changes when $\overrightarrow{x}$ changes from neutral to level j. In other words, β_j indicates how much μ changes when

$\overrightarrow{x}$ changes from all $x_j = 0$ to $x_j = 1$. The baseline is constrained such that the deflections sum to zero across the levels of $\overrightarrow{x}$:

$$\sum_{j=1} \beta_j = 0 \tag{18.2}$$

The expression of the model in Equation 18.1 is not complete without the constraint in Equation 18.2. Examples were shown in Figure 14.4, p. 370, and it is worth the effort to go now to that Figure for a quick review.

The predicted value, μ_i, in Equation 18.1 is for the central tendency in the data. The data themselves are assumed to be randomly generated around that central tendency. As usual, we will assume a normal distribution, $y_i \sim N(\mu_i, \tau)$, where τ is the precision of the normal distribution. As discussed in previous chapters, if the data have outliers, a t distribution may be used instead.

18.1.1 The Hierarchical Prior

Our primary interest is in estimating the deflection parameters, β_j, for each level of $\overrightarrow{x}$. We could just put a separate prior on each parameter and estimate them separately from each other. It is typical, however, that the levels of $\overrightarrow{x}$ are not utterly unrelated to each other, and therefore data from one level may inform estimates in another level. For example, the deflections for republicans, libertarians, and greens can inform an estimate of the deflection for democrats. Thus, if the deflection for libertarians is $+1.0$, for republicans is $+0.5$, and for greens is -1.0, then the deflection for democrats should be somewhere in that general range, and not out at, say, -12.0. At the least, we might have prior beliefs that the deflections for most levels of $\overrightarrow{x}$ may be small, with only a few deflections being large, and therefore we can let the various levels mutually inform each other's estimates based on this structural assumption.

The form of the hierarchical model for oneway BANOVA is displayed in Figure 18.1 (Gelman, 2005, 2006). In the upper middle of the diagram is the normal distribution that describes the distribution of deflections, β_j, across levels of $\overrightarrow{x}$. This normal distribution has a mean at zero, reflecting the fact that the deflections are constrained to fall both above and below the baseline, because they must sum to zero. Importantly, the precision of this normal distribution, τ_β, is estimated, not preset at a constant. Thus, if many of the levels of $\overrightarrow{x}$ have a small deflection in the data, then the precision τ_β is estimated to be high, and this in turn shrinks the estimates of other β_j.

The prior for τ_β derives from the recommendation of Gelman (2006). First, the precision is converted to standard deviation: $\tau_\beta = 1/\sigma_\beta^2$. Then a folded-$t$ distribution is used as a prior on σ_β. The folded-t is just the positive side of the usual t distribution. Notice that it is defined only over nonzero values, as is required for σ_β, and it extends to positive infinity. Unlike

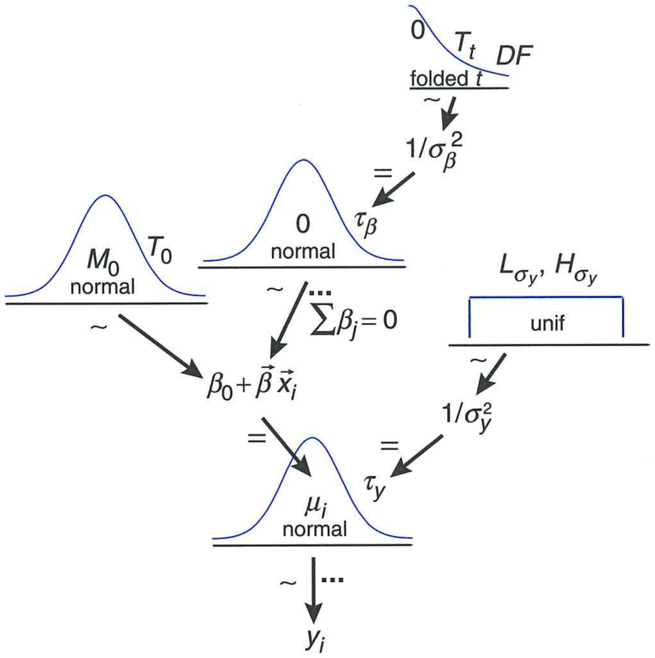


FIGURE 18.1

Hierarchical dependencies for model of oneway Bayesian ANOVA. The baseline is β_0, and the deflection away from that baseline for the j^{th} level of x is β_j. The standard deviation of the β_j's has a folded-t prior. The variance within levels of x is estimated by the precision τ_y, which is here assumed to be homogeneous across groups, although it need not be in general.

the gamma(ϵ, ϵ) distribution that is often used for precisions, however, the folded-t does not have infinite density near zero. Because noisy data can never rule out deflections of all zero, there can be unintended distortions in the estimates if the prior places extreme densities at either end of the scale (see Gelman, 2006, for details).

A folded-t prior could also be used for the noise σ_y, but we will use a uniform, again as recommended by Gelman (2006). One motivation is that a uniform may have a more intuitive form than a folded-t when expressing a prior belief. A second reason is that the infelicities of estimation that affect σ_β are not present so prominently at this level in the model, because the within-level noise is typically not near zero, and there are enough data points to overwhelm any mildly informed prior.

18.1.1.1 Homogeneity of Variance

The model described here assumes equal variances across all levels of $\vec{x}$. As a concrete example, the model assumes that the variance of income is the same for republicans as for democrats as for libertarians as for greens. This assumption of homogeneous variances is vestigial from two precursors. First, the analogous assumption is made in linear regression, and ANOVA may be construed, mathematically, as a special case of linear regression. Second, homogeneity of variance is assumed in NHST ANOVA to simplify derivation of F distributions. Neither of these precursors actually demands that we make the assumption of equal variances in BANOVA.

The model described here assumes homogeneity of variance merely for simplicity in presentation. By assuming equal variances for all levels of x, the focus could be on the estimation of the group deflection parameters, β_j. Also, by assuming equal variances, the results of BANOVA can be more directly compared to the results of NHST ANOVA, if such a comparison desired.

In principle, the BANOVA model can (and should) estimate difference variances for each level of $\vec{x}$. The model in Figure 18.1 can be expanded analogously to the model in Figure 16.11, p. 436. Instead of a single precision, τ_y, used for all levels of $\vec{x}$, a separate precision τ_j is estimated for each level of $\vec{x}$, as in the lower right of Figure 16.11. A higher-level distribution describes the spread of the τ_j across levels of $\vec{x}$. This structure provides shrinkage of the estimates of the τ_j, to the extent that the data suggest homogeneity of variance. Exercise 18.3 has you give this scheme a test drive.

18.1.2 Doing It with R and BUGS

As usual, every arrow in the hierarchical diagram of Figure 18.1 has a corresponding line in the BUGS model specification. The parameters that appear as "β_j" in Figure 18.1 are denoted by "a[j]" in the model specification.

To understand the way that the model is specified in the BUGS code, it is important to understand how the data are formatted. The $\vec{x}$ values in the program are coded as integer indices 1, 2, 3,. . . , and *not* as vectors $\langle 1, 0, 0, \ldots \rangle$, $\langle 0, 1, 0, \ldots \rangle$, $\langle 0, 0, 1, \ldots \rangle$,. . . . By coding $\vec{x}$ as integers, then nested indexing can be used instead of dot products of vectors. Thus, $\vec{\beta} \cdot \vec{x}$ becomes coded as a[x], not inprod(a[],x[]). For the i^{th} observation, the value of x is coded as x[i]. Thus, x[i]$\in \{1, 2, 3, \ldots,$NxLvl$\}$ for $i \in \{1, \ldots,$Ntotal$\}$, where NxLvl is the number of levels of $\vec{x}$ and Ntotal is the total number of observations.

Here is the BUGS model specification (ANOVAonewayBRugs.R):

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu[i] , tau )
    mu[i] <- a0 + a[x[i]]
  }
  #
  tau <- pow( sigma , -2 )
  sigma ~ dunif(0,10) # y values are assumed to be standardized
```

```
    #
    a0 ~ dnorm(0,0.001) # y values are assumed to be standardized
    #
    for ( j in 1:NxLvl ) { a[j] ~ dnorm( 0.0 , atau ) }
    atau <- 1 / pow( aSD , 2 )
    aSD <- abs( aSDunabs ) + .1
    aSDunabs ~ dt( 0 , 0.001 , 2 )
}
```

The constraint, that the deflections sum to zero, does not appear in the model specification. The BUGS code estimates the baseline and deflections without the constraint, but the MCMC estimates are recentered at zero by subsequent R code. The noncentered baseline is denoted in the BUGS model as `a0`, and the noncentered deflections are denoted `a[j]`. Those noncentered estimates are transformed to respect the sum-to-zero constraint merely by subtracting the mean of the `a[j]`'s from each `a[j]`, and adding the mean to the baseline. Thus, `b[j] = a[j] - mean(a)` and `b0 = a0 + mean(a)`.

The constants for the top-level priors are set with the assumption that the data values, y, have been standardized according to Equation 16.1, p. 425. (Of course, the x values cannot be standardized because they are nominal.) This standardization makes it easier to establish reasonable default priors for a range of applications, without having to change the prior constants when the application changes, for example, from income, on the order of 10^5 dollars, to width of hairs, on the order of 10^{-1} millimeters. Nevertheless, when there is strong prior information, it should be incorporated. Exercise 18.2 has you explore robustness of the results when you use different priors.

There is one other trick in the BUGS model specification that is not in the hierarchical diagram of Figure 18.1. One line of the BUGS model specifies that the standard deviation of the group effects, denoted `aSDunabs`, comes from a t distribution: `aSDunabs ~ dt(0,0.001,2)`. Another line takes the absolute value to "fold" the t distribution onto the nonnegative numbers: `aSD <- abs(aSDunabs) + .1`. But that line also mysteriously adds a small constant, namely 0.1. This constant keeps `aSD` from venturing extremely close to zero. The reason for keeping `aSD` away from zero is that shrinkage can become overwhelmingly strong when there are many groups with few data points per group. This becomes especially problematic in the next chapter when we consider interaction of factors.

It turns out that MCMC sampling for this model can be extremely inefficient. One important way to reduce burn-in time is to start the chain at reasonable positions. We start the overall baseline at the grand mean of the data, and start the deflections at the level means minus the grand mean. The variances are also initialized near the corresponding data variances. The full code, including initialization of chains, is presented in Section 18.4.1 (`ANOVAonewayBRugs.R`).

Because the chains can be highly autocorrelated, extensive thinning is needed, keeping a step only once out of several hundred. Running such long chains can take a long time and become boring for your computer, which would rather be searching the web for exciting software updates. In the examples presented here, we simply tolerate the modest waiting times. But there are various methods for reparameterizing the models so that the chains are sampled with less autocorrelation (e.g., Gelman, 2006; Gelman & Hill, 2007, Ch. 19).

One tempting but inappropriate approach is to impose the sum-to-zero constraint in the BUGS model specification like this:

```
a[1] <- -sum( a[2:NxLvl] )
for ( j in 2:NxLvl ) { a[j] ~ dnorm( 0.0 , atau ) }
```

Notice that the first deflection is forced to equal the negative sum of the remaining deflections; therefore the first deflection is not an estimated parameter. Only deflections indexed 2 and higher have a prior specification. This approach works fine when the prior on the deflections has no hyperprior, that is, when `atau` is a constant (Ntzoufras, 2009). But when `atau` is itself being estimated, it must be informed by all the deflections, not only by deflections 2 and higher. For example, it might be that group 1 is very different from groups 2 through NxLvl, whereas groups 2 through NxLvl are nearly equal. This situation would cause the estimate of the precision `atau` to be artificially high, because it would not be affected by the group 1. Therefore, despite the fact that this approach to model specification reduces autocorrelation dramatically, the approach is not appropriate when we are using a hyperprior to estimate the deflections.

18.1.3 A Worked Example

With all the emphasis these days on physical fitness and muscle building, it's only appropriate to consider an example about muscles. In particular, we'd like to know if geographical location influences muscle size, which might be affected by the weather or amount of daylight. Consider some data regarding muscles from five geographic locations: (1) Tillamook, Oregon; (2) Newport, Oregon; (3) Petersburg, Alaska; (4) Magadan, Russia (Pacific coast); and (5) Tvarminne, Finland. The values in the data set are the length of the anterior adductor muscle scar divided by total muscle length, in the mussel species Mytilus trossulus. These ratios of scar length to total length tend to be between 5% and 15% (McDonald, 2009; McDonald, Seed, & Koehn, 1991).

Results of the BRugs program listed in Section 18.4.1 (`ANOVAonewayBRugs.R`) are shown in Figure 18.2. The histograms in the upper row show the (marginal) posterior distributions of the β_j values for the five geographical locations. These β_j values are deflections away from the baseline β_0, which is not shown. Some things to keep in mind when interpreting the results: First, the estimates of deflection are subject to shrinkage, because the model

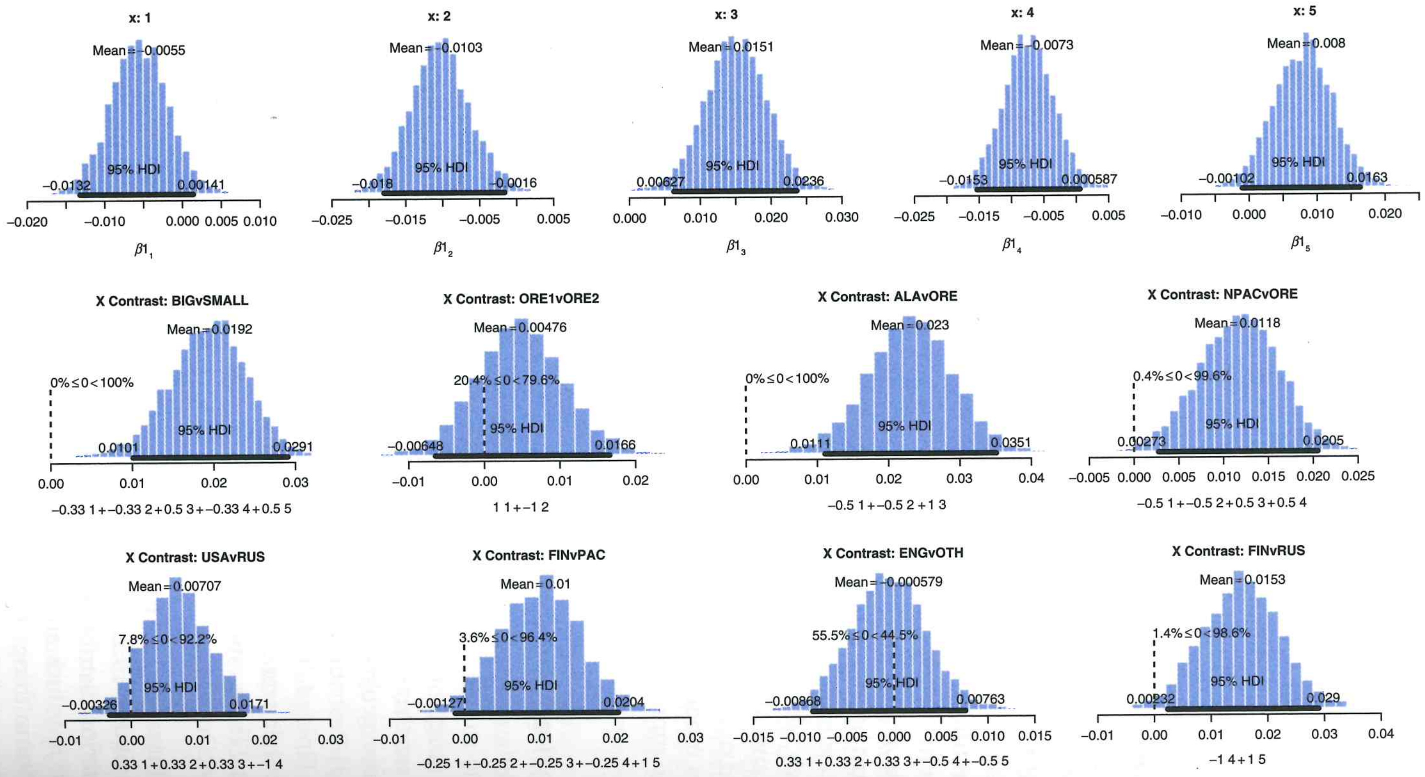


FIGURE 18.2

Upper row: Posterior estimates of β_j values for data from McDonald (2009; McDonald et al., 1991). The values indicate deflections by each group away from the overall central tendency. *Middle and bottom rows*: Various complex comparisons of β_j values. For example, the bottom row, third panel, compares the three English-speaking sites against the two non-English-speaking sites (where they might say "midiya myshtsy" or "simpukka lihas" instead of "mussel muscle").

incorporates the prior structural assumption that all the deflections come from the same overarching distribution. The mean deflections shown in Figure 18.2 are, in fact, a little smaller than the deflections of the actual sample means. Second, the model assumes that the precision is the same for all groups (i.e., there is homogeneity of variance). The posterior β_j values are the ones that are believable when also assuming homogeneous variances. If the groups actually have wildly different variances, then the estimates for β_j may be distorted. Third, the marginal distributions on the β_j cannot be used to directly infer differences between groups, because the parameters might be correlated. Indeed, the deflections tend to be negatively correlated, because increasing the estimated deflection for one group suggests decreasing the estimated deflection for another, if they are to remain symmetric around the baseline. To judge differences between groups, the differences must be computed directly.

18.1.3.1 Contrasts and Complex Comparisons

The middle and bottom rows of Figure 18.2 shows several comparisons for the mussel muscle results. A comparison amounts to a difference between an average of some groups and an average of other groups. For example, to compare the four Pacific Ocean mussels against the one non-Pacific (Baltic Sea) mussel, we multiply the deflections (β_j's) of the first four groups by 1/4 to get their average, and subtract it from the deflection (β_5) of the fifth group to get the difference. The difference is called a *contrast*, and when the comparison involves a contrast of averages, instead of a contrast of two specific groups, it is sometimes called a *complex* comparison. The contrast is fully specified by the coefficients on the groups, which can be placed into a vector of *contrast coefficients*. For example, the contrast coefficients for comparing Pacific Ocean mussels against Baltic Sea mussels are $-1/4, -1/4, -1/4, -1/4, +1$. Notice that the coefficients sum to zero. We compute the difference at every step in the MCMC chain, and examine the resulting distribution of believable differences. The distribution for this particular example is shown in the bottom row, second panel, of Figure 18.2, where it can be seen that just over 96% of the believable differences lie on one side of zero, and the 95% HDI just spans zero. From these results we may not want to declare categorically that there is a credible difference between Finland and the other sites; the decision depends on how you set your HDI and ROPE. Regardless of your decision rule, the posterior does tells us the most believable difference and the uncertainty in that difference.

Figure 18.2 shows a variety of comparisons that might be of interest. For example, the first panel of the middle row compares the two sites with the biggest muscles against the three other sites. This sort of comparison would be labeled "post hoc" by traditional analyses, because we might not have specified which sites would be biggest before collecting the data. The second panel in the middle row contrasts the two sites in Oregon. The third panel in the middle row compares the Alaska site against the average of the two Oregon sites.

We can make all the comparisons shown in Figure 18.2, and as many others as we like, without worrying about inflated false alarm rates, because the posterior distribution does not change when we consider additional comparisons. The posterior distribution is the best inference we can make based on the data we have and the prior beliefs we started with. It is possible that the random data in our sample are spuriously unrepresentative of the underlying population, but we cannot know. Fortunately, because of the incorporation of our prior knowledge about how estimates in the different locations can mutually inform each other, the estimates undergo shrinkage, which helps to mitigate the effect of rogue data. In many applications, the shrinkage yields decisions similar to those that would result from NHST "corrections" for multiple comparisons. But unlike NHST corrections, the shrinkage in the Bayesian approach is based on explicit structural prior knowledge, and is not affected by which or how many comparisons are intended. (For previous discussion of these issues, see Section 17.2, regarding decisions about multiple regression coefficients, and Section 11.4, regarding multiple comparisons of groups.)

18.1.3.2 Is There a Difference?

The contrasts and complex comparisons in Figure 18.2 were judged to be credibly nonzero if the 95% HDI excluded (a ROPE around) zero. A difference would be deemed to be practically equivalent to zero if its HDI fell entirely within a ROPE. This decision procedure is attractive because all the group β_j's are simultaneously estimated, with mutually informed shrinkage, and from priors that are also appropriately informed (which entails also being agreeable to a skeptical audience).

Some researchers prefer to pose the question "Is there a difference?" as a model comparison on two priors. One prior expresses the null hypothesis that the contrast has zero magnitude; the other prior expresses a complementary hypothesis that any magnitude contrast is possible. This approach was discussed extensively in Section 12.2.

There are two attractions to the two-prior, model-comparison approach. One attraction is that the model comparison can yield posterior odds in favor of the null, unlike NHST, which can only reject a null hypothesis but never accept it. Another attraction is that the complementary hypothesis is usually intended to be an "automatic" uninformed prior that is chosen for mathematical felicity. The hope is that an automatic prior obviates debate about how prior information should be expressed.

As was argued in Section 12.2, the two-prior approach should be applied cautiously. First, it is important to emphasize that the two-prior approach only indicates which prior is relatively less unbelievable. If either prior is theoretically untenable in the first place, then the "automatic" model comparison is automatically uninformative. Thus, the two-prior approach should only be applied to situations in which (1) it is theoretically appropriate to posit that a particular contrast really can be exactly zero, and (2) the alternative prior incorporates prior knowledge about the plausible magnitude of the difference.

As an example, consider a situation presented by Solari, Liseo, & Sun (2008, Table 3, p. 495). There were nine groups, with a metric dependent variable. The dependent variable was the acetic acid content of tomatoes, and the nine groups were different types of manuring during growth of the tomatoes. The mean of group 3 appeared to be different than other groups. To test whether group 3 was different, the authors conducted a Bayesian model comparison of two priors: The null-hypothesis prior had all nine groups with identical means. The alternative prior had group 3 with a separately estimated mean, while the other eight groups had identical means. The resulting Bayes factor (BF) strongly favored the alternative prior. Does this result suggest that the alternative prior is what we should believe? Unfortunately, no. The BF tells us that the prior with eight equal means and one different mean for group 3 is more believable than the prior with nine equal means (assuming that the priors on the two hypotheses were 50-50). But the prior with eight equal means, on groups other than group 3, is already untenable because we do not believe that the eight groups have identical means. Moreover, the estimate of difference, between group 3 and the other groups, is not what we want, because the estimate does not take into account variation among the eight other groups.

When instead we conduct a Bayesian analysis using the BANOVA model, we obtain a posterior that simultaneously estimates all the separate group deflections, with shrinkage, from a plausibly informed prior. The complex comparison of group 3 against the other eight groups is shown in Figure 18.3, where it can be seen that the magnitude of the contrast is credibly greater than zero. In this application, there is no need to pursue a BF approach to group comparisons.

It is also worth reiterating that the two-prior, model-comparison approach can arrive at a conclusion opposite that of the one-prior, estimation approach. Recall Figure 12.5, p. 308, which showed that a model comparison preferred the null hypothesis of identical groups to the alternative hypothesis of all different groups, even though an estimation of effects in the alternative hypothesis showed a credible difference among groups. The point in that case was that the null model, even though it was a poor model, was less bad than the alternative model. Follow-up model comparisons would be required to narrow down which combination of group equivalences was least implausible. Even after that, we would not necessarily want to believe that any of the groups are truly equivalent, because we know in advance that they were treated differently. Instead, we desire an estimate of the differences and the precision of the estimate. That situation involved a dichotomous dependent variable, but the analogous situation can arise for metric dependent variables.

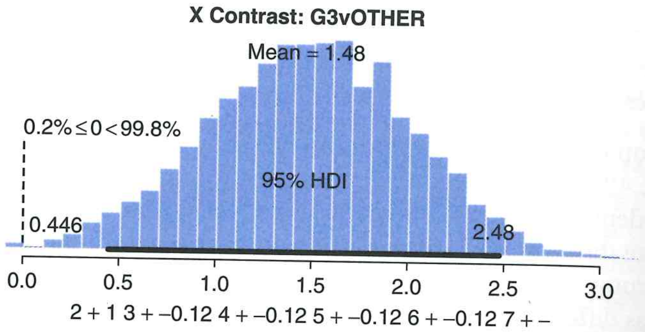


FIGURE 18.3

A comparison of group 3 versus the average of other groups, for the data in Solari et al. (2008, Table 3, p. 495). The specification of contrast coefficients on the x axis overflows the margins of the figure because there are so many groups. The contrast coefficients on the nine groups are $-1/8, -1/8, +1, -1/8, \ldots$, which, when rounded to two decimal places, appear as $-0.12, -0.12, +1, -0.12, \ldots$. (Reprinted with permission from Figure 6 of Kruschke, 2010a,b).

The two-prior, model-comparison approach can be appropriate in situations where actual equivalence is tenable and the goal is to identify which conditions are plausibly equivalent, or situations in which zero-magnitude effects are tenable and the goal is to identify which conditions have zero effect. In those situations, it behooves the researcher to pursue the model-comparison or related approaches (see, e.g., Berry & Hochberg, 1999; Gopalan & Berry, 1998; Mueller, Parmigiani, & Rice, 2007; Scott & Berger, 2006). Moreover, Bayesian model comparison is highly advisable when the two models are genuinely viable competitors that express different explanations of the data. In these situations, it is important that the priors in the two models are equivalently informed so that neither model is at a disadvantage because of an infelicity in an arbitrary, automatic prior.

18.2 MULTIPLE COMPARISONS

In 20th-century null-hypothesis significance testing (NHST), there is immense literature regarding how to compute the "true" significance (i.e., probability of false alarm) of an apparent difference between groups, when the analyst is conducting comparisons of multiple groups. The problem is that when more comparisons are conducted, there are more opportunities for a spuriously large difference to appear by accident. In other words, there are more opportunities for false alarms. Notice that this problem of inflated false alarm rates arises because NHST is based on the intentions of the analyst. If the analyst intends to make lots of comparisons between various combinations of groups, then there is greater opportunity for false alarms. If the analyst intends to make only a few comparisons between groups, then there is less opportunity for false alarms.

For example, consider again the sea mussel data. Group 4 (Pacific coast Russia) and group 5 (Finland) seem to be different, and it is meaningful to plan a comparison between them because of their geographical difference. If we run a two-group t test, we get $t = 2.53, p = 0.028$, which denotes a significant difference. On the other hand, if we run a post hoc test of all pairwise comparisons, using Tukey's "Honest Significant Difference" correction, then we find that $p = 0.093$, and the difference is *not* significant. So do Russia and Finland really differ? According to NHST, the answer depends on your intentions: If you intended to compare only those two locales, then they are significantly different, but if you intended to make all pairwise comparisons, then they are not significantly different.

Section 11.4, p. 281, discussed multiple comparisons in NHST, in the context of a dichotomous dependent variable. Here we reiterate those ideas in the context of a metric dependent variable.

Suppose that we have two groups: One group is patients treated with a placebo and a second group is patients treated with a totally ineffective drug. We measure a metric variable (e.g., body temperature). Because there is no actual difference between the treatments, the underlying distributions of body temperatures are identical for the two groups; we will suppose that they are normally distributed with equal means and equal variances. When we run an experiment, we are collecting a random sample of data from each of the groups. The random samples might show a spuriously large difference between their means, just by chance, despite the fact that on average, in the long run, the groups are identical.

To determine how often the spuriously large differences occur, we can simulate conducting the experiment over and over. For every simulated experiment, we compute the difference of means between the samples from the groups. The difference of sample means is in units of the original measurement scale (e.g., degrees Fahrenheit or degrees Celsius). To get rid of the arbitrary influence of the measurement scale, we standardize the difference of means and call the result the t statistic. Because the true difference between groups is zero, the t value typically will be near zero. Occasionally, by chance, the t value will be far above or far below zero. The lowest curve in Figure 18.4 shows the probability that the sampled t value falls above the critical t value on the abscissa. For example, the probability that the sampled t value falls above $t_{crit} = 2.23$ is $p(FA) = 0.05$; this is marked by an arrow. In NHST, the decision rule is to reject the null hypothesis if the sample t exceeds a critical value that is selected to keep false alarms to only 5%. Thus, when comparing group 1 with

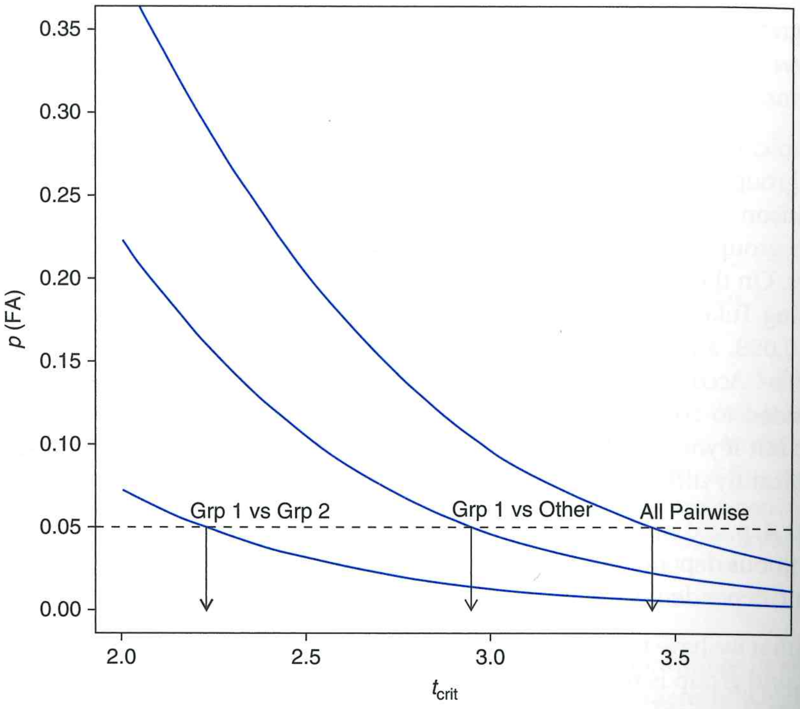


FIGURE 18.4
The probability of false alarm as a function of critical t value, with separate curves for different sets of comparisons. All groups have $N = 6$ fixed by intention. The curve labeled "Grp 1 vs Grp 2" is for a single comparison of two groups, and corresponds with the usual two-group t distribution. The curve labeled "Grp 1 vs Other" refers to four paired comparisons, of Group 1 versus each of the other four groups. The last curve is for the set of all 10 paired comparisons. (Reprinted with permission from Figure 2 of Kruschke, 2010a.)

group 2, we would reject the null hypothesis if $t > 2.23$, because that would happen only 5% of the time by chance alone.

Now consider an expanded experiment, in which there is a placebo treatment and four distinct drugs, for a total of five treatment groups. According to the null hypothesis, the five treatment groups have identical distributions of body temperatures (normally distributed with equal means and variances). However, because of random sampling in any particular experiment, some treatment samples will have higher or lower mean temperatures than other treatment samples. Suppose that before we collect any real data, we plan to compare the placebo group (group 1) with each of the four drug groups (i.e., we plan four pairwise comparisons). Each of these comparisons might yield a fairly large difference merely by chance, even when there is truly no difference in the underlying distributions. We can determine how often these chance extremes happen by running a Monte Carlo simulation. For a simulated experiment, we randomly sample six scores from each of the five groups, and compute the t values of each of the four comparisons. The simulated experiment is repeated many times. For each candidate t_{crit}, we see what proportion of simulated experiments had a comparison that exceeded that critical value. The middle curve of Figure 18.4 shows the result. Notice that at any given value of t_{crit}, there is now a much higher probability that the simulated experiment will have at least one comparison with larger t. In particular, to keep the false alarm rate down to 5%, t_{crit} must be about 2.95 instead of 2.33.

If we did not plan only four tests but instead decided to compare every group with every other group, then we would have even more opportunity for false alarms. With five groups, there are 10 pairwise comparisons. If we simulate experiments from equal distributions as before, but this time consider all 10 t values, the probability of false alarm is higher yet, as shown in the right curve of Figure 18.4. The critical value has risen even higher, to approximately 3.43.[1]

Now, suppose we actually run the experiment. We randomly assign 30 people to the five groups, six people per group. The first group gets the placebo, and the other four groups get the corresponding four drugs. *We are careful to make this a double-blind experiment: Neither the subjects nor experimenters know who is getting which treatment. Moreover, no one knows whether any other person is even in the experiment.* We collect the data. Our first question is to compare the placebo and the first drug (i.e., group 1 versus group 2). We compute the t statistic for the data from the two groups and find that $t = 2.95$. Do we decide that the two treatments had significantly different effects?

The answer, bizarrely, depends on the intentions of the person we ask. Suppose, for instance, that we handed the data from the first two groups to a research assistant, who is asked to test for a difference between groups. The assistant runs a t test and finds $t = 2.95$, declaring it to be *highly significant* because it greatly exceeds the critical value of 2.23 for a two-group t test. Suppose, on the other hand, that we handed the data from all five groups to a different research assistant, who is asked to compare the first group against each of the other four. This assistant runs a t test of group 1 versus group 2 and finds $t = 2.95$, declaring it to be *marginally significant* because it just squeezes past the critical value of 2.95 for these four planned comparisons. Suppose, on yet another hand, that we handed the data from all five groups to a different research assistant, who is told to conduct all pairwise comparisons post hoc because we have no strong hypotheses about which treatments will have beneficial or detrimental or neutral effects. This assistant runs a t test of group 1

[1] For a discussion of various correction procedures and when to use them, see Figure 5.1 of Maxwell & Delaney (2004). If you must learn NHST methods, this is an excellent resource.

versus group 2 and finds $t = 2.95$, declaring it to be *not significant* because it fails to exceed the critical value of 3.43 that is used for post hoc pairwise comparisons. Notice that regardless of which assistant analyzed the data, the t value for the two groups stayed the same because the data of the two groups stayed the same. Indeed, the data were completely uninfluenced by the intentions of the analyst. So why should the interpretation of the data be influenced by the intentions of the analyst? It shouldn't.

If you believe that the interpretation should be influenced by the intentions of the analyst, how do you determine the intentions of the analyst? Did the analyst truly plan only those particular comparisons, or did the analyst really plan others but jettison them once the data were in? Or did the analyst actually plan fewer comparisons but realize later that additional comparisons should be made to address other theoretical issues? Or did the analyst actually plan to include two other treatment groups in the study but then not actually include those groups in the analysis because of administrative errors committed during the data collection? Or what if the experiment was planned by a team of people, some of whom planned some comparisons and others of whom planned other comparisons? Conclusion: Establishing the true intentions of the analyst is not only pointless, it is also impossible.

Multiple comparisons are not a problem in a Bayesian analysis (e.g., Gelman, Hill, & Yajima, 2009). The posterior distribution is a fixed entity in high-dimensional parameter space, and making comparisons between groups is simply examining that posterior distribution from different perspectives or margins. The posterior does not change when new comparisons come to mind.

The posterior is not immune to spurious coincidences of rogue data, of course. False alarms are mitigated, however, by incorporating prior knowledge into the structure of the model. The estimates of the groups are mutually informative via estimation of higher-level structure, and shrinkage of estimates across groups attenuates false alarms. The attenuation of false alarms is governed by the data, not by unknowable intentions.

18.3 TWO-GROUP BAYESIAN ANOVA AND THE NHST *t* TEST

The idea behind an NSHT t test is simple: We have two groups, each with a mean. We compute the difference of the means and standardize that difference relative to the standard deviation of the scores within the groups. The resulting standardized difference is called the t value. We want to know whether the observed t value is significantly different from zero, so we compare the t value to a sampling distribution of t values (Gosset, 1908). The sampling distribution assumes that the intention of the researcher was to stop when there were exactly $N1$ values observed for the first group, and exactly $N2$ values observed for the second group.

The t test is a special case of NHST ANOVA when there are only two groups. More specifically, when the two groups are assumed to have equal variances in the underlying population, then the t value squared equals the F value in two-group ANOVA. (And what's an F value, you may ask? The F value is the summary statistic used in NHST ANOVA to express how much the groups differ from each other. It's the ratio of the variance between group means, relative to the variance within groups.)

In typical applications of BANOVA, the prior on the between-group variance is only mildly informed. In this case, a BANOVA on two groups imposes little shrinkage on the group estimates because there are so few groups. It is only when several groups "gang up" that they strongly inform the estimate of the variation between groups and therefore constrain the estimates of other groups. When the prior on the variance within groups is also vague, the results of a two-group BANOVA closely agree with the results of an NHST t test. Exercise 18.1 has you explore this correspondence.

18.4 R CODE

18.4.1 Bayesian Oneway ANOVA

(ANOVAonewayBRugs.R)

```
graphics.off()
rm(list=ls(all=TRUE))
fnroot = "ANOVAonewayBrugs"
library(BRugs)          # Kruschke, J. K. (2010). Doing Bayesian data analysis:
                        # A Tutorial with R and BUGS. Academic Press / Elsevier.
#------------------------------------------------------------------------------
# THE MODEL.

modelstring = "
# BUGS model specification begins here...
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu[i] , tau )
    mu[i] <- a0 + a[x[i]]
  }
  #
  tau <- pow( sigma , -2 )
  sigma ~ dunif(0,10) # y values are assumed to be standardized
  #
  a0 ~ dnorm(0,0.001) # y values are assumed to be standardized
  #
  for ( j in 1:NxLvl ) { a[j] ~ dnorm( 0.0 , atau ) }
  atau <- 1 / pow( aSD , 2 )
  aSD <- abs( aSDunabs ) + .1
  aSDunabs ~ dt( 0 , 0.001 , 2 )
}
# ... end BUGS model specification
" # close quote for modelstring
# Write model to a file, and send to BUGS:
writeLines(modelstring,con="model.txt")
modelCheck( "model.txt" )

#------------------------------------------------------------------------------
# THE DATA.

# Specify data source:
dataSource = c( "McDonaldSK1991" , "SolariLS2008" , "Random" )[1]
# Load the data:

if ( dataSource == "McDonaldSK1991" ) {
  fnroot = paste( fnroot , dataSource , sep="" )
  datarecord = read.table( "McDonaldSK1991data.txt", header=T ,
                           colClasses=c("factor","numeric") )
  y = as.numeric(datarecord$Size)
  Ntotal = length(datarecord$Size)
  x = as.numeric(datarecord$Group)
  xnames = levels(datarecord$Group)
  NxLvl = length(unique(datarecord$Group))
  contrastList = list( BIGvSMALL = c(-1/3,-1/3,1/2,-1/3,1/2) ,
                       ORE1vORE2 = c(1,-1,0,0,0) ,
                       ALAvORE = c(-1/2,-1/2,1,0,0) ,
                       NPACvORE = c(-1/2,-1/2,1/2,1/2,0) ,
                       USAvRUS = c(1/3,1/3,1/3,-1,0) ,
                       FINvPAC = c(-1/4,-1/4,-1/4,-1/4,1) ,
                       ENGvOTH = c(1/3,1/3,1/3,-1/2,-1/2) ,
                       FINvRUS = c(0,0,0,-1,1) )
}

if ( dataSource == "SolariLS2008" ) {
  fnroot = paste( fnroot , dataSource , sep="" )
  datarecord = read.table("SolariLS2008data.txt", header=T ,
                           colClasses=c("factor","numeric") )
  y = as.numeric(datarecord$Acid)
  Ntotal = length(datarecord$Acid)
  x = as.numeric(datarecord$Type)
  xnames = levels(datarecord$Type)
  NxLvl = length(unique(datarecord$Type))
  contrastList = list( G3vOTHER = c(-1/8,-1/8,1,-1/8,-1/8,-1/8,-1/8,-1/8,-1/8) )
}

if ( dataSource == "Random" ) {
  fnroot = paste( fnroot , dataSource , sep="" )
  #set.seed(47405)
  ysdtrue = 4.0
  a0true = 100
  atrue = c( 2 , -2 ) # sum to zero
  npercell = 8
  datarecord = matrix( 0, ncol=2 , nrow=length(atrue)*npercell )
  colnames(datarecord) = c("y","x")
  rowidx = 0
  for ( xidx in 1:length(atrue) ) {
    for ( subjidx in 1:npercell ) {
      rowidx = rowidx + 1
      datarecord[rowidx,"x"] = xidx
      datarecord[rowidx,"y"] = ( a0true + atrue[xidx] + rnorm(1,0,ysdtrue) )
    }
  }
  datarecord = data.frame( y=datarecord[,"y"] , x=as.factor(datarecord[,"x"]) )
  y = as.numeric(datarecord$y)
  Ntotal = length(y)
  x = as.numeric(datarecord$x)
  xnames = levels(datarecord$x)
  NxLvl = length(unique(x))
  # Construct list of all pairwise comparisons, to compare with NHST TukeyHSD:
  contrastList = NULL
  for ( g1idx in 1:(NxLvl-1) ) {
    for ( g2idx in (g1idx+1):NxLvl ) {
      cmpVec = rep(0,NxLvl)
      cmpVec[g1idx] = -1
      cmpVec[g2idx] = 1
      contrastList = c( contrastList , list( cmpVec ) )
    }
  }
}

# Specify the data in a form that is compatible with BRugs model, as a list:
ySDorig = sd(y)
yMorig = mean(y)
z = ( y - yMorig ) / ySDorig
datalist = list(
  y = z ,
  x = x ,
  Ntotal = Ntotal ,
  NxLvl = NxLvl
)
# Get the data into BRugs:
modelData( bugsData( datalist ) )

#------------------------------------------------------------------------------
# INTIALIZE THE CHAINS.

# Autocorrelation within chains is large, so use several chains to reduce
# degree of thinning. But we still have to burn-in all the chains, which takes
# more time with more chains (on serial CPUs).
nchain = 5
modelCompile( numChains = nchain )

if ( F ) {
   modelGenInits() # often won't work for diffuse prior
} else {
  #  initialization based on data
  theData = data.frame( y=datalist$y , x=factor(x,labels=xnames) )
  a0 = mean( theData$y )
  a = aggregate( theData$y , list( theData$x ) , mean )[,2] - a0
  ssw = aggregate( theData$y , list( theData$x ) ,
                  function(x){var(x)*(length(x)-1)} )[,2]
  sp = sqrt( sum( ssw ) / length( theData$y ) )
  genInitList <- function() {
    return(
        list(
            a0 = a0 ,
            a = a ,
            sigma = sp ,
            aSDunabs = sd(a)
        )
    )
  }
  for ( chainIdx in 1 : nchain ) {
    modelInits( bugsInits( genInitList ) )
  }
}

#------------------------------------------------------------------------------
# RUN THE CHAINS

# burn in
BurnInSteps = 10000
modelUpdate( BurnInSteps )
# actual samples
samplesSet( c( "a0" ,  "a" , "sigma" , "aSD" ) )
stepsPerChain = ceiling(5000/nchain)
thinStep = 750
modelUpdate( stepsPerChain , thin=thinStep )

#------------------------------------------------------------------------------
# EXAMINE THE RESULTS

source("plotChains.R")
source("plotPost.R")

checkConvergence = T
if ( checkConvergence ) {
   sumInfo = plotChains( "a0" , saveplots=T , filenameroot=fnroot )
   sumInfo = plotChains( "a" , saveplots=T , filenameroot=fnroot )
   sumInfo = plotChains( "sigma" , saveplots=T , filenameroot=fnroot )
   sumInfo = plotChains( "aSD" , saveplots=T , filenameroot=fnroot )
}

# Extract and plot the SDs:
sigmaSample = samplesSample("sigma")
aSDSample = samplesSample("aSD")
windows()
layout( matrix(1:2,nrow=2) )
par( mar=c(3,1,2.5,0) , mgp=c(2,0.7,0) )
plotPost( sigmaSample , xlab="sigma" , main="Cell SD" , breaks=30 )
plotPost( aSDSample , xlab="aSD" , main="a SD" , breaks=30 )
```

```
dev.copy2eps(file=paste(fnroot,"SD.eps",sep=""))

# Extract a values:
a0Sample = samplesSample( "a0" )
chainLength = length(a0Sample)
aSample = array( 0 , dim=c( datalist$NxLvl , chainLength ) )
for ( xidx in 1:datalist$NxLvl ) {
   aSample[xidx,] = samplesSample( paste("a[",xidx,"]",sep="") )
}

# Convert to zero-centered b values:
mSample = array( 0, dim=c( datalist$NxLvl , chainLength ) )
for ( stepIdx in 1:chainLength ) {
    mSample[,stepIdx ] = ( a0Sample[stepIdx] + aSample[,stepIdx] )
}
b0Sample = apply( mSample , 2 , mean )
bSample = mSample - matrix(rep( b0Sample ,NxLvl),nrow=NxLvl,byrow=T)
# Convert from standardized b values to original scale b values:
b0Sample = b0Sample * ySDorig + yMorig
bSample = bSample * ySDorig

# Plot b values:
windows(datalist$NxLvl*2.75,2.5)
layout( matrix( 1:datalist$NxLvl , nrow=1 ) )
par( mar=c(3,1,2.5,0) , mgp=c(2,0.7,0) )
for ( xidx in 1:datalist$NxLvl ) {
    plotPost( bSample[xidx,] , breaks=30 ,
              xlab=bquote(beta*1[.(xidx)]) ,
              main=paste("x:",xnames[xidx])  )
}
dev.copy2eps(file=paste(fnroot,"b.eps",sep=""))

# Display contrast analyses
nContrasts = length( contrastList )
if ( nContrasts > 0 ) {
   nPlotPerRow = 5
   nPlotRow = ceiling(nContrasts/nPlotPerRow)
   nPlotCol = ceiling(nContrasts/nPlotRow)
   windows(3.75*nPlotCol,2.5*nPlotRow)
   layout( matrix(1:(nPlotRow*nPlotCol),nrow=nPlotRow,ncol=nPlotCol,byrow=T) )
   par( mar=c(4,0.5,2.5,0.5) , mgp=c(2,0.7,0) )
   for ( cIdx in 1:nContrasts ) {
       contrast = matrix( contrastList[[cIdx]],nrow=1) # make it a row matrix
       incIdx = contrast!=0
       histInfo = plotPost( contrast %*% bSample , compVal=0 , breaks=30 ,
                 xlab=paste( round(contrast[incIdx],2) , xnames[incIdx] ,
                             c(rep("+",sum(incIdx)-1),"") , collapse=" " ) ,
                 cex.lab = 1.0 ,
                 main=paste( "X Contrast:", names(contrastList)[cIdx] ) )
   }
   dev.copy2eps(file=paste(fnroot,"xContrasts.eps",sep=""))
}

#==============================================================================
# Do NHST ANOVA and t tests:

theData = data.frame( y=y , x=factor(x,labels=xnames) )
aovresult = aov( y ~ x , data = theData ) # NHST ANOVA
cat("\n------------------------------------------------------------------\n\n")
print( summary( aovresult ) )
cat("\n------------------------------------------------------------------\n\n")
print( model.tables( aovresult , "means" ) , digits=4 )
windows()
boxplot( y ~ x , data = theData )
cat("\n------------------------------------------------------------------\n\n")
print( TukeyHSD( aovresult , "x" , ordered = FALSE ) )
windows()
plot( TukeyHSD( aovresult , "x" ) )
if ( T ) {
  for ( xIdx1 in 1:(NxLvl-1) ) {
    for ( xIdx2 in (xIdx1+1):NxLvl ) {
      cat("\n-------------------------------------------------------------\n\n")
      cat( "xIdx1 = " , xIdx1 , ", xIdx2 = " , xIdx2 ,
           ", M2-M1 = " , mean(y[x==xIdx2])-mean(y[x==xIdx1]) , "\n" )
      print( t.test( y[x==xIdx2] , y[x==xIdx1] , var.equal=T ) ) # t test
    }
  }
}
cat("\n------------------------------------------------------------------\n\n")

#==============================================================================
```

18.5 EXERCISES

Exercise 18.1. **[Purpose: To notice that Bayesian ANOVA with two groups tends to agree with an NHST *t* test.]** The BRugs program of Section 18.4.1 (`ANOVAonewayBRugs.R`) allows you to specify random data. It executes a Bayesian ANOVA, and at the end of the program it also conducts an NHST ANOVA and *t* tests (using R's `aov` and `t.test` functions). Run the program ten times with different random data by commenting out the `set.seed` command. Specify `ysdtrue` = `4.0`, `atrue` = `c(2,-2)` (which implies two groups because there are two deflections) and `npercell` = 8. For each run, record, by hand, (1) how much of the posterior difference between means falls on one side of zero (see the posterior histogram with the main title "X Contrast" and *x* axis labeled "$-1\ 1 + 1\ 2$"), (2) whether the 95% HDI excludes zero, and (3) the confidence interval and *p* value of the NHST *t* test. Do the *t* test and the BANOVA usually agree in their decisions about whether the group means are different?

Exercise 18.2. **[Purpose: To understand the influence of the prior in Bayesian ANOVA.]** In the model section of the BRugs program of Section 18.4.1 (`ANOVAonewayBRugs.R`), and correspondingly in the diagram of Figure 18.1, there are several constants that determine the prior. These constants include the mean value of the baseline (M_0 in the diagram), the precision on the baseline (T_0 in the diagram), the precision of the folded-*t* distribution (T_t in the diagram), and the upper value of the uniform distribution on σ_y (H_{σ_y} in the diagram). Because the data are standardized, M_0 should be set at zero, and T_0 can be modest (not terribly small). H_{σ_y} also can be set to a modest value because the data are standardized. But what about the precision of the folded-*t* distribution, T_t? This constant modulates the degree of shrinkage: A large value of T_t indicates prior knowledge that the groups do not differ much, and it imposes a high degree of shrinkage that must be overcome by the data.

Run the program on the mussel data using a small value of T_t, such as 1.0E-6, and a large value of T_t, such as 1000. Are the results very different? Discuss which prior value might be appropriate.

Exercise 18.3. **[Purpose: To understand Bayesian ANOVA with*out* assuming equal variances.]** Modify the program in Section 18.4.1 (`ANOVAonewayBRugs.R`) so that it allows a different variance for each group, with the different variances coming from a hyperdistribution that has its precision informed by the data. In other words, instead of assuming the same τ_y ($= 1/\sigma_y^2$) for all the levels of *x*, we allow each group to have its own variance. Denote the precision of the j^{th} group as τ_j, analogous to the deflection β_j. Just as the group deflections are assumed to come from a higher-level distribution, we will assume that the group SDs come from a higher-level distribution. Because SDs must be non-negative, use a gamma density for the higher-level distribution. The gamma distribution has two parameters for which you need to establish a prior. *See the right side of Figure 16.11 for guidance.* Corresponding code is offered in a hint, below. Run the program on the mussel muscle data. Are the conclusions about the group means any different than when assuming equal variances across groups?

Hint regarding the conclusion: The posteriors on the group means are only a little different in this case, because the group variances are roughly the same. But because the group variances are less constrained when they are all allowed to be different, they are less certain. Therefore, the group means are a little less certain, and thus the differences of means are a little less certain.

Programming hints: Here are some code snippets, showing the model specification and chain initialization.

(ANOVAonewayNonhomogvarBrugs.R)

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu[i] , tau[x[i]] )
    mu[i] <- a0 + a[x[i]]
  }
  a0 ~ dnorm(0,0.001)
  for ( j in 1:NxLvl ) {
     a[j] ~ dnorm( 0.0 , atau )
     tau[j] ~ dgamma( sG , rG )
  }
  sG <- pow(m,2)/pow(d,2)
  rG <- m/pow(d,2)
  m ~ dgamma(1,1)
  d ~ dgamma(1,1)
  atau <- 1 / pow( aSD , 2 )
  aSD <- abs( aSDunabs ) + .1
  aSDunabs ~ dt( 0 , 0.001 , 2 )
}
```

(ANOVAonewayNonhomogvarBrugs.R)

```
# initialization based on data
theData = data.frame( y=datalist$y , x=factor(x,labels=xnames) )
a0 = mean( theData$y )
a = aggregate( theData$y , list( theData$x ) , mean )[,2] - a0
tau = 1/(aggregate( theData$y , list( theData$x ) , sd )[,2])^2
genInitList <- function() {
  return(
      list(
          a0 = a0 ,
          a = a ,
          tau = tau ,
          m = mean( tau ) ,
          d = sd( tau ) ,
          aSDunabs = sd(a)
      )
  )
}
for ( chainIdx in 1 : nchain ) {
  modelInits( bugsInits( genInitList ) )
}
```

CHAPTER 19

Metric Predicted Variable with Multiple Nominal Predictors

CONTENTS

Sometimes I wonder just how it could be, that
Factors aligned so you'd end up with me.
All of the priors made everyone think, that
Our interaction was destined to shrink.

In this chapter we consider situations with a metric predicted variable and multiple nominal predictor variables. For example, we might want to predict income (a metric variable) on the basis of political party affiliation (a nominal variable) and ethnicity (another nominal variable). Or we may want to predict response time (a metric variable) on the basis of hand used for the response (a nominal value: dominant hand or nondominant hand) and modality of

Doing Bayesian Data Analysis: A Tutorial with R and BUGS. DOI: 10.1016/B978-0-12-381485-2.00019-5

© 2011, Elsevier Inc. All rights reserved.

stimulus (another nominal value: visual, auditory, or tactile). These situations are modeled by the cell in the first row and last column of Table 14.1, p. 385.

In traditional NHST, this situation is known as multifactor ANOVA. We use the same underlying model, but without reference to F sampling distributions; instead we use hierarchical priors that provide additional structural constraints. Multifactor ANOVA is a straightforward extension of the model presented in the previous chapter, but with a new concept of interaction between nominal variables. Just as multiple regression considered interaction of metric predictors, multifactor ANOVA considers interaction of nominal predictors.

19.1 BAYESIAN MULTIFACTOR ANOVA

Recall from the previous chapter that in oneway ANOVA, we describe the effect of each level of the predictor as a deflection away from an overall baseline, where the baseline is the central tendency across all levels of the predictor. In multifactor ANOVA, the same idea applies to two or more predictors, and the deflections resulting from each predictor are *added*. We'll use notation analogous to the previous chapter, but with extra subscripts to indicate the different predictors, just as we used in multiple regression on continuous predictors.

The mathematical notation was introduced as a case of the generalized linear model in Section 14.1.6.2, p. 370. Suppose we have two nominal predictors, cleverly denoted $\vec{x}_1$ and $\vec{x}_2$. These predictor vectors can only take on values of $\langle 1, 0, 0, \ldots \rangle$, $\langle 0, 1, 0, \ldots \rangle$, and so on, with the j^{th} component having the value 1 when the predictor has its j^{th} nominal level.

When the effects of the two predictors are additive, the predicted tendency is as follows:

$$\begin{aligned} y &= \beta_0 + \vec{\beta}_1 \vec{x}_1 + \vec{\beta}_2 \vec{x}_2 \\ &= \beta_0 + \sum_{j=1}^{J_1} \beta_{1,j} x_{1,j} + \sum_{k=1}^{J_2} \beta_{2,k} x_{2,k} \end{aligned}$$

To make the parameter values unique, we include the constraints

$$\sum_{j=1}^{J_1} \beta_{1,j} = 0 \quad \text{and} \quad \sum_{j=k}^{J_2} \beta_{2,k} = 0$$

Those equations repeat Equations 14.7 and 14.8. In words, the value β_0 establishes the overall baseline from which the predictors indicate deflections. When predictor x_1 has value $x_{1,j}$, a deflection of $\beta_{1,j}$ is added to the baseline, and when predictor x_2 has value $x_{2,k}$, a deflection of $\beta_{2,k}$ is also added

to the
the pre
deflecti

19.1.1

The eff
there is
a hot-ai
its levit
is a non

Figure 1
The abs
indicate
the sam
each pa

In the l
additive
the level
deflection
that the
the inte
predicto

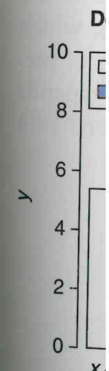


FIGURE 1
An interact
interaction
arrows. Th
the value o
$\vec{x}_2$ depen

to the baseline. The deflections may be negative. Indeed, across all levels of the predictors, the constraints demand as much negative deflection as positive deflection, so that the deflections sum to zero for each predictor.

19.1.1 Interaction of Nominal Predictors

The effect of two predictors may be nonadditive, in which case we say that there is an "interaction" of the predictors. For example, if a flame is put under a hot-air balloon, its levity will increase. And if hydrogen is added to a balloon, its levity will increase. But if hydrogen and flame are added to a balloon, there is a nonadditive interaction, such that levity is not increased.

Figure 19.1 displays a simple interaction. Both predictors have only two levels. The abscissa groups the two levels of predictor $\vec{x}_1$, and the shading of the bars indicates the two levels of predictor $\vec{x}_2$. All three panels of Figure 19.1 show the same data, but the nature of the interaction is highlighted differently in each panel.

In the left panel of Figure 19.1, the dashed parallelogram indicates the best *additive* model for the data. The dashed lines indicate the average change when the levels of the predictors change. *The vertical arrows highlight the nonadditive deflections, away from the additive average, that constitute the interaction.* Notice that the arrows sum to zero across each edge of the parallelogram. Thus, the interaction components do not change the average deflections of each predictor.

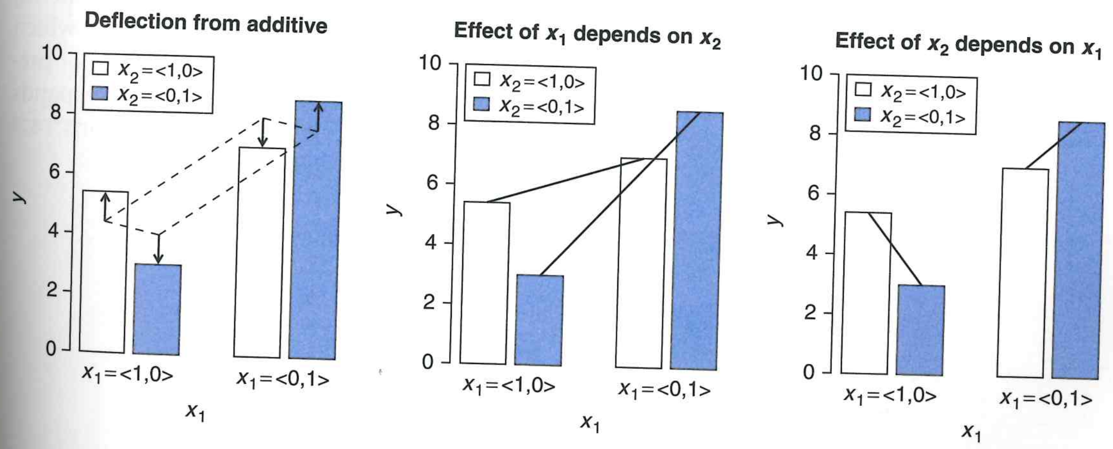


FIGURE 19.1
An interaction of nominal variables $\vec{x}_1$ and $\vec{x}_2$, parsed three ways. The left panel emphasizes that the interaction involves a nonadditive, torsion-like deflection away from the additive model, as indicated by the arrows. The middle panel shows the same data, with lines that emphasize that the effect of $\vec{x}_1$ depends on the value of $\vec{x}_2$. The right panel again shows the same data, but with lines that emphasize that the effect of $\vec{x}_2$ depends on the value of $\vec{x}_1$.

The middle and right panels of Figure 19.1 highlight different interpretations of the interaction. The middle panel shows that the effect of $\overrightarrow{x}_1$, that is, the amount that y changes when $\overrightarrow{x}_1$ changes, depends on the level of $\overrightarrow{x}_2$: When $\overrightarrow{x}_2 = \langle 1, 0\rangle$, there is only a small change in y when $\overrightarrow{x}_1$ changes, but when $\overrightarrow{x}_2 = \langle 0, 1\rangle$, there is a larger change in y when $\overrightarrow{x}_1$ changes. The right panel makes the same point but with the roles of the predictors reversed: When $\overrightarrow{x}_1 = \langle 1, 0\rangle$, the effect of $\overrightarrow{x}_2$ is to decrease y, but when $\overrightarrow{x}_1 = \langle 0, 1\rangle$, the effect of $\overrightarrow{x}_2$ is to increase y.

The average deflection from baseline due to a predictor is called the *main effect* of the predictor. The main effects of the predictors correspond to the dashed lines in the left panel of Figure 19.1. When there is nonadditive interaction between predictors, the effect of one predictor depends on the level of the other predictor. The deflection from baseline for a predictor, at a fixed level of the other predictor, is called the *simple effect* of the predictor at the level of the other predictor. When there is interaction, the simple effects do not equal the main effect.

It may be edifying to compare Figure 19.1, which shows interaction of *nominal* predictors, with Figure 17.8, p. 470, which shows interaction of *metric* predictors. The essential notion of interaction is the same in both cases: Interaction is the nonadditive portion of the prediction, and interaction means that the effect of one predictor depends on the level of the other predictor.

The mathematical formalism for nonadditive interactions was introduced in Section 14.1.6.3, p. 371, and is repeated here. The nonadditive components, indicated by the vertical arrows in Figure 19.1, are denoted $\beta_{1\times 2,j,k}$, which means the interaction of predictors 1 and 2 (denoted 1 × 2) at level j of predictor 1 and level k of predictor 2. The formal expression merely expands the additive model by including the interaction. Recall from Equations 14.9 and 14.10 that the model with interaction term can be written as

$$\begin{aligned} y &= \beta_0 + \overrightarrow{\beta}_1 \overrightarrow{x}_1 + \overrightarrow{\beta}_2 \overrightarrow{x}_2 + \overrightarrow{\beta}_{1\times 2} \overrightarrow{x}_{1\times 2} \\ &= \beta_0 + \sum_{j=1}^{J_1} \beta_{1,j} x_{1,j} + \sum_{k=1}^{J_2} \beta_{2,k} x_{2,k} + \sum_{j=1}^{J_1} \sum_{k=1}^{J_2} \beta_{1\times 2,j,k} x_{1\times 2,j,k} \end{aligned}$$

with the constraints

$$\sum_{j=1}^{J_1} \beta_{1,j} = 0 \quad \text{and} \quad \sum_{k=1}^{J_2} \beta_{2,k} = 0 \quad \text{and}$$

$$\sum_{j=1}^{J_1} \beta_{1\times 2,j,k} = 0 \;\forall k \quad \text{and} \quad \sum_{k=1}^{J_2} \beta_{1\times 2,j,k} = 0 \;\forall j$$

In those
equation
level of th
panel of
along eve

Our goal
observed
bars in Fi
near the
data at ea
actually p
and the da

19.1.2 T

The comp
look daur
Figure 18.

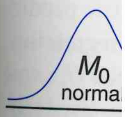


FIGURE 19.
Hierarchical c
Figure 18.1.

In those last equations, the symbol "∀" means "for all." In words, the last two equations simply mean that the interaction deflections sum to zero along every level of the two predictors. A graphic example of this was presented in the left panel of Figure 19.1, which shows that the heights of the arrows sum to zero along every edge of the parallelogram.

Our goal is to estimate the additive and interactive deflections, based on the observed data. It is important to understand that the observed data are *not* the bars in Figure 19.1; instead, the data are swarms of points at various heights near the heights of the bars. The bars represent the central tendency of the data at each combination of the predictors. Thus, what the equations above actually predict is the central tendency μ at each combination of predictors, and the data are typically modeled as being normally distributed around μ.

19.1.2 The Hierarchical Prior

The complete generative model of the data is shown in Figure 19.2. It might look daunting, but it really is merely the diagram for oneway ANOVA, in Figure 18.1, with the hyperprior replicated for each predictor and interaction.

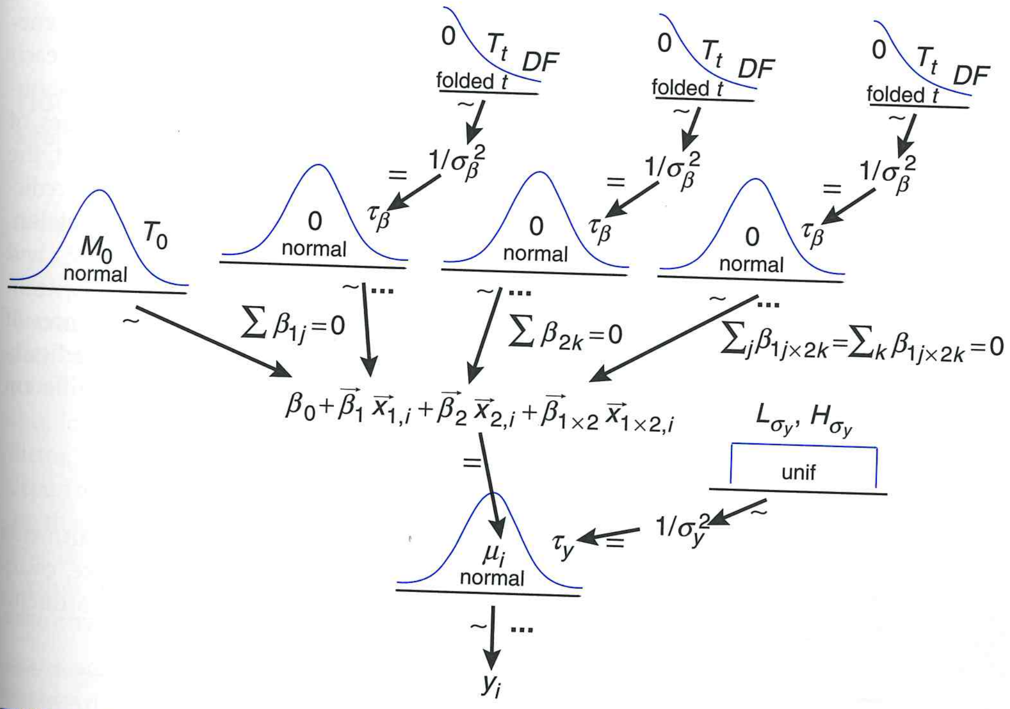


FIGURE 19.2
Hierarchical dependencies for model of two-way Bayesian ANOVA. Compare with Figure 18.1.

The lowest level of Figure 19.2 indicates that the observed data points, y_i, are distributed normally around the predicted value μ_i. Moving upward in the diagram, the arrow impinging on μ_i indicates that the predicted value is baseline plus additive deflection due to each predictor plus interactive deflection due to the combination of predictors. The upper levels of the diagram indicate prior structural assumptions about the deflections. We assume that the deflections produced by a predictor are centered at zero, and we allow the variance (i.e., precision) of the deflections to be estimated from the data. Thus, if most of the deflections are small, the estimated variance is small, and the hyperdistribution creates shrinkage in the estimates of other deflections.

A key conceptual aspect of the hyperdistributions is that they apply separately to the different predictors and interactions. In other words, there is not just one hyperdistribution that governs all deflections for all predictors and interactions. This division of generative structure reflects a prior assumption that the magnitude of the effect of one predictor might not be informative regarding the magnitude of the effect of a different predictor. But within a predictor, the magnitude of deflection produced by one level may inform the magnitude of deflection produced by other levels of that same predictor.[1]

As was assumed in the case of oneway ANOVA, we will assume homogeneity of variance: The variability of the observed data is the same within each combination of predictors. This is indicated in Figure 19.2 by the *single* parameter σ_y that is used in the likelihood function, regardless of the values of the predictors. As before, there are two reasons for this assumption. First, the assumption is a natural simplification in multiple regression on metric predictors, and ANOVA can be construed as a special case of multiple regression. Second, the assumption of equal variances is made in NHST ANOVA, and we will also make it here in BANOVA to facilitate comparing across the techniques. But there is no requirement in BANOVA to assume equal variances. If the situation suggests that different levels of the predictors produce radically different variances in the data, then the hierarchical prior can allow different variances.

19.1.3 An Example in R and BUGS

Figure 19.3 shows the mean annual salaries of faculty in four departments at three levels of seniority. The four departments are business finance, counseling and educational psychology, chemistry, and theater. These departments

[1] By analogy to multiple regression, if there are many predictors included in a model, it is reasonable in principle to include a higher-level distribution *across predictors* such that the estimated variance of one predictor informs the estimated variance of another predictor. This would be especially useful if the application includes many nominal predictors, each with many levels. Such applications are rare.

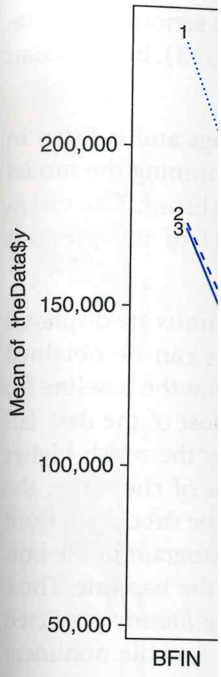


FIGURE 19.3
Mean annual salaries

are the nominal le
ity are full profess
professors are usua
doctoral studies. A
their doctoral or p
to 40 years post g
should, be treated
dictor, denoted $\vec{x}_2$
department and of
ing that the change
goal is to estimate
bership, the main
seniority.

The display of the r
binations of depart
In other words, the
not necessarily equa

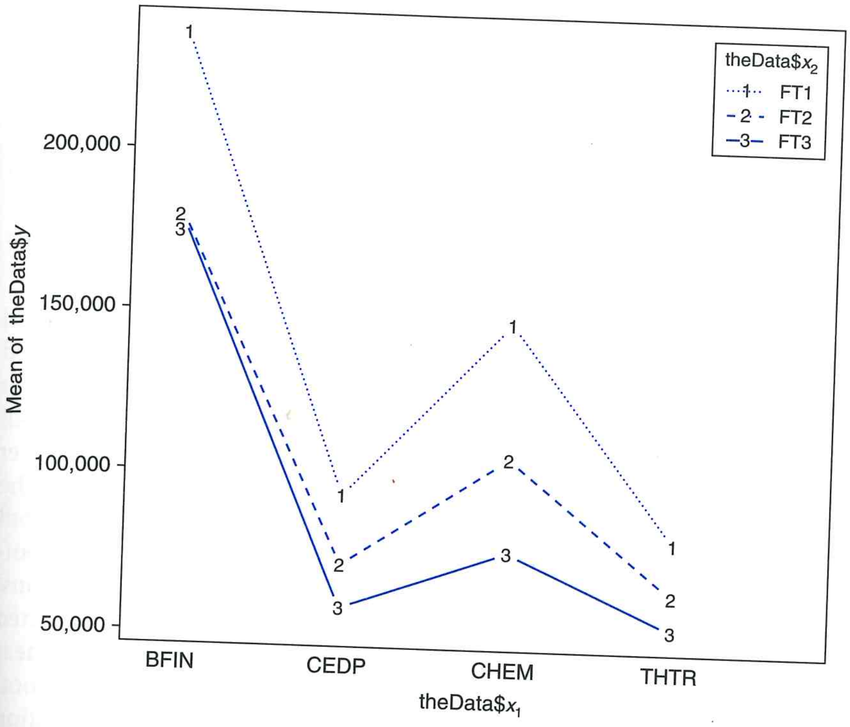


FIGURE 19.3
Mean annual salaries of faculty in four departments at three levels of seniority.

are the nominal levels of a predictor denoted $\overrightarrow{x}_1$. The three levels of seniority are full professor, associate professor, and assistant professor. Assistant professors are usually within 7 years after completing their doctoral or postdoctoral studies. Associate professors are usually within about 10 years of their doctoral or postdoctoral studies. Full professors are anywhere from 10 to 40 years post graduate school. Although seniority could, and perhaps should, be treated as an ordinal variable, we will treat it as a nominal predictor, denoted $\overrightarrow{x}_2$. A glance at the means suggests that there are effects of department and of seniority. There appears also to be an interaction, meaning that the change in salary due to seniority depends on the department. Our goal is to estimate the baseline salary, the main effect of department membership, the main effect of seniority, and the interaction of department and seniority.

The display of the means in Figure 19.3 obscures the fact that different combinations of department and seniority had different numbers of data points. In other words, the number of associate professors in business finance was not necessarily equal to the number of full professors in theater. In traditional

NHST ANOVA, this sort of "unbalanced" design can cause serious computational difficulties (e.g., Maxwell & Delaney, 2004, pp. 320–343). But Bayesian ANOVA has no problem with unbalanced designs.

The model of Figure 19.2 was implemented in R and BRugs and is listed in Section 19.3.1 (`ANOVAtwowayBRugs.R`). Several tricks for running the model in BUGS are described in that section, before the program listing. The essentials, however, are much like the oneway ANOVA model of the previous chapter.

The results are shown in Figure 19.4. (The means and HDI limits are displayed with only three significant digits, but more precise values can be obtained directly from the program.) The top-left histogram shows that the baseline for these four departments is 111,381. Notice, however, that most of the data fall below this baseline because the overall data are skewed by the much higher salaries in one department. For salaries in the department of chemistry, the fourth histogram in the top row indicates that 2164 should be subtracted from the baseline. For salaries of assistant professors, the first histogram in the bottom row indicates that 20,100 should be subtracted from the baseline. Thus, for assistant professors in the department of chemistry, the *linearly* predicted salary is 111,381 − 2164 − 20,100 = 89,117. But there is a notable nonlinear interaction component for that combination: The fourth histogram of the bottom row shows that 10,938 must be subtracted from the linear combination to get the mean estimate for that combination, namely, 78,179.

19.1.4 Interpreting the Posterior

In most applications, we are interested not only in estimation of effects for each group, but we are also interested in deciding whether two groups are credibly different. Just as we compared groups in oneway ANOVA in the previous chapter, we can compare groups in multifactor ANOVA.

The top and middle rows of Figure 19.5 show selected contrasts of levels of the main effects. We may ask whether there is a credible difference in salaries, on average, between business finance (BFIN) and counseling and educational psychology (CEDP). The top-left histogram indicates that the average difference is about $122,000, and the 95% HDI falls far from zero. We may also ask whether there is a credible difference in salaries, on average, between CEDP and theater (THTR). The top-right histogram indicates that the average difference is about $7780, but the 95% HDI spans zero, which indicates that the uncertainty in the estimated difference is fairly large relative to the estimated difference itself. The middle row of Figure 19.5 shows contrasts regarding levels of seniority: There is a credible difference between full professors (FT1) and associate professors (FT2), and between FT2 and assistant professors (FT3).

It is important to understand that the main effects of department and seniority are *average* effects, when the other factors are collapsed. For example, the

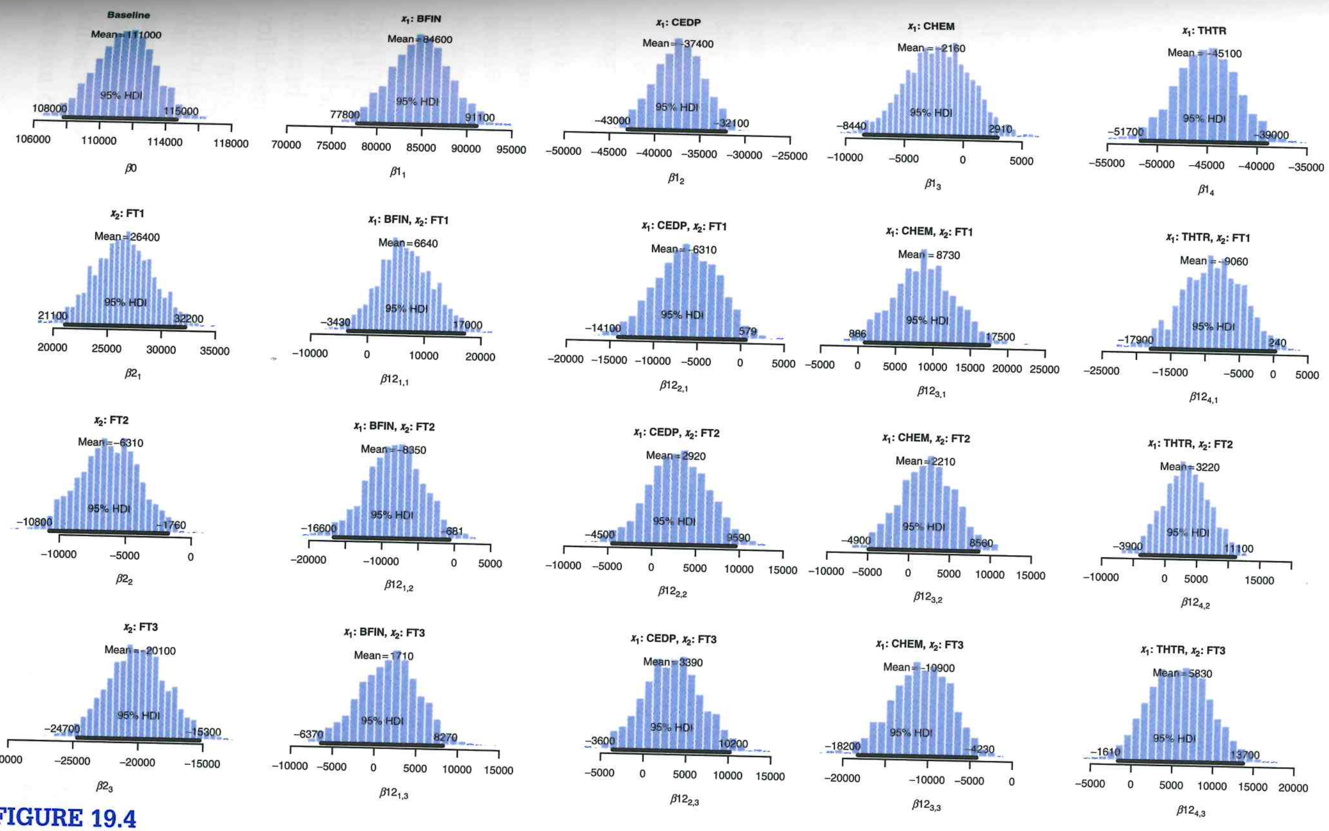


FIGURE 19.4

Posterior distribution for data in Figure 19.3. Baseline (β_0) is shown in upper left. Remainder of top row is main effect of $\vec{x}_1$ (department). Remainder of left column is main effect of $\vec{x}_2$ (seniority). Remaining cells show the interaction effects.

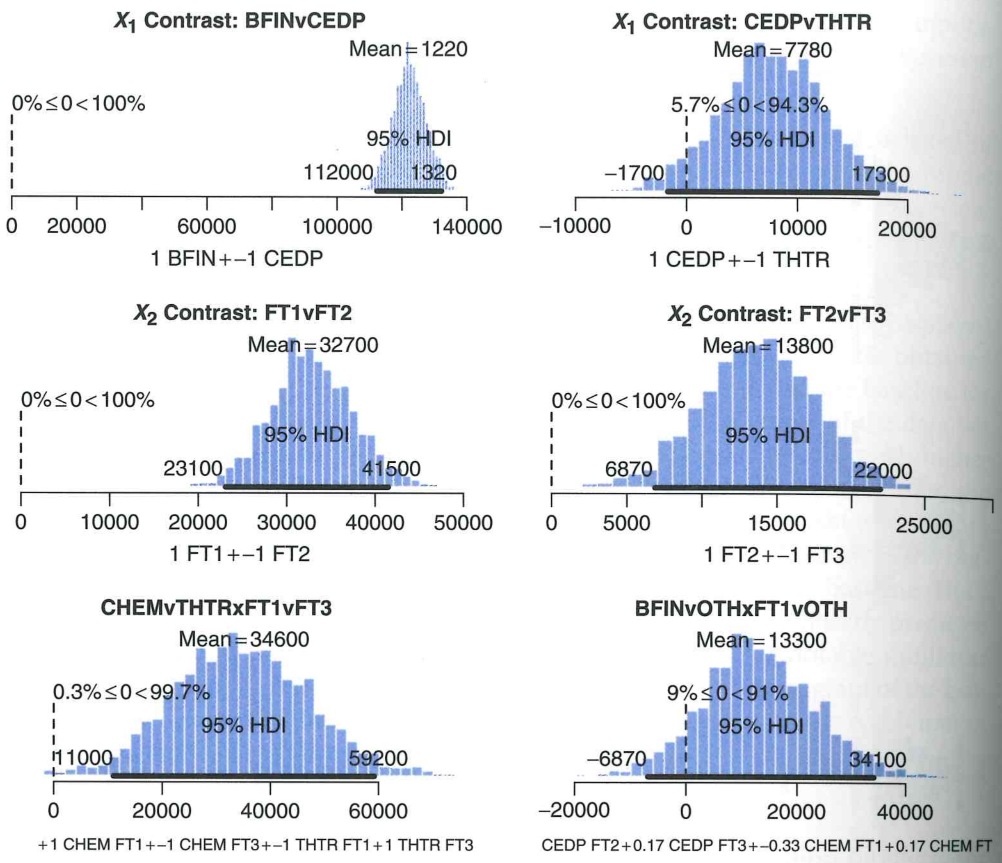


FIGURE 19.5

Selected contrasts for posterior in Figure 19.4. The values for the means and HDI limits are rounded to three leading digits. The x-axis labels of the bottom row are obscured because they exceed the boundaries of the plot.

contrast between FT2 and FT3 (middle row, right panel of Figure 19.5) is the average difference between FT2 and FT3, collapsed across all departments. But if you look at the data in Figure 19.3, you can see that the difference between FT2 and FT3 is not the same in every department: There is a fairly large difference in CHEM, but a very small difference in BFIN. The effect of changing from FT2 to FT3 depends on the department, which means that there is interaction.

Main effects must be interpreted and described cautiously when there are interactions. It would be a mistake to say that "the" difference between FT2 and FT3 is 13,800. Instead, that is the average difference across departments. The actual difference within any particular department might be quite different. Similarly, it would be a mistake to say that "the" effect of FT3 is to subtract 20,100 from the baseline, because the effect of seniority interacts with department.

19.1.4.1 Combining Metric and Nominal Predictors: ANCOVA

Consider again the salary data in Figure 19.3. You can see that the mean salary for FT1's in chemistry is much higher than in theater. This difference might be attributable solely to being in one department or the other. But the difference might also be attributable to some other factor, such as years on the job. In other words, the FT1's in chemistry might happen to have been employed for decades, whereas the FT1's in theater might happen to be relatively young. If we had the age of each employee, or, better yet, the number of years that the employee had been at the current level of seniority, then we could include that information as an additional predictor of salary. We could then assess whether department membership contributed any predictiveness beyond number of years on the job.

When a nominal predictor, such as department membership, is combined with a metric predictor, such as years on the job, the model is sometimes referred to as analysis of *co*variance, or ANCOVA. The metric predictor is the "covariate."

Programming ANCOVA in BUGS is a trivial combination of the models we've used for linear regression and ANOVA. Denote the nominal group membership for individual `i` as `xNom[i]`, and denote the metric covariate value as `xMet[i]`. Then the core of the BUGS model specification is

```
mu[i] <- a0 + a[ xNom[i] ] + bMet * xMet[i]
y[i] ~ dnorm( mu[i] , tau )
```

where `a[]` is the deflection of each group from baseline, and `bMet` is the regression coefficient on the covariate. As in standard ANOVA, the deflections `a[]` and intercept `a0` should be transformed so that the deflections sum to zero.

When initializing the chains for ANCOVA in BUGS, it can help to start at the maximum likelihood estimate (MLE). The `lm()` function in R provides the MLE. If we type

```
xNom = factor( xNom ) # makes xNom a "factor"
lmInfo = lm( y ~ xNom + xMet )
```

then `lmInfo` is a list of information about the linear model that "best" fits the data. The best fitting coefficients are stored in `lmInfo$coef`. The first components of `lmInfo$coef` are the deflections for the levels of `xNom` and the last component of `lmInfo$coef` is the slope coefficient for `xMet` (because of the ordering of the variables in the call to `lm`). The deflections are parameterized relative to the first component, however. To convert to the parameterization that we use, in which the deflections sum to zero around a baseline, we can use the following code:

```
NxNomLevels = length( levels( xNom ) ) # number of levels of xNom
# Next line adds first xNom component to other xNom components:
a = c( lmInfo$coef[1] , lmInfo$coef[1] + lmInfo$coef[2:NxNomLevels] )
a0Init = mean( a )
aInit = a - mean( a )
```

ANCOVA models can also involve a separate slope for every level of the nominal predictor. Then the core of the BUGS model specification is

```
mu[i] <- a0 + a[ xNom[i] ] + ( bMet + bMetI[ xNom[i] ] ) * xMet[i]
y[i] ~ dnorm( mu[i] , tau )
```

where bMet+bMetI[xNom[i]] is the group-specific regression coefficient on the covariate for the xNom[i]th group. The coefficient bMet is the overall slope, and deflection bMetI[xNom[i]] adjusts steeper or shallower for each group. To make the slopes identifiable, the group-specific deflections are constrained to sum to zero: $\sum_{\mathtt{xNom}} \mathtt{bMetI[xNom]} = 0$. Just as in multiple regression, a hyperprior can be put on the group-specific slopes, whereby the group-specific slopes come from a normal (or *t*) distribution, and the precision of that distribution is itself estimated.

The ANCOVA model, with a distinct intercept and slope for each group, closely resembles the model for repeated-measures simple linear regression in Section 16.3, p. 433. The model in that section had a distinct intercept and slope for each *subject*. If the subject variable in that model is considered to be a nominal predictor analogous to xNom here, then that model is essentially equivalent to one used here. The two model expressions are different, however, in how naturally they generalize to situations with more predictors. The formulation in the present section uses the general ANOVA formulation for the group-specific coefficients (i.e., deflections that sum to zero) and therefore generalizes naturally to situations with multiple nominal predictors.

Additional information about non-Bayesian ANCOVA can be found in a variety of other sources. A brief Bayesian treatment can be found in the book by Ntzoufras (2009), but beware that the formulation there uses no hyperprior on the nominal or metric coefficients, and the method used there to implement the sum-to-zero constraint cannot be used with hyperpriors, as was discussed previously on p. 497.

19.1.4.2 Interaction Contrasts

Just as we can ask whether differences among particular levels of predictors are credible, we can ask whether interactions among particular combinations of predictors are credible. Consider again the data in Figure 19.3. The difference between full professors (FT1) and assistant professors (FT3) appears to be large in the chemistry department (CHEM) but smaller in the theater department (THTR). Is the simple effect of seniority bigger in chemistry than it is in theater? In other words, is (CHEM.FT1−CHEM.FT3)−(THTR.FT1−THTR.FT3) credibly nonzero?

This sort of difference of differences is called an *interaction contrast*. In general, an interaction contrast is constructed by taking any set of contrast coefficients on $\vec{x}_1$, and any set of contrast coefficients on $\vec{x}_2$, and computing their outer product. The outer product was described in Section 8.8.1 (`BernTwoGrid.R`), p. 178. Formally, the outer product of two vectors is denoted by the symbol "⊗." To provide an example of an interaction contrast as an outer product of main-effect contrasts, we will recast the one we are presently considering, namely, (CHEM.FT1−CHEM.FT3)−(THTR.FT1−THTR.FT3), in generic notation. Notice that CHEM is level 3 of predictor 1, and hence can be written as $\vec{x}_{1,3}$. Writing the other components in the same fashion, the interaction contrast is $(\vec{x}_{1,3}.\vec{x}_{2,1} - \vec{x}_{1,3}.\vec{x}_{2,3}) - (\vec{x}_{1,4}.\vec{x}_{2,1} - \vec{x}_{1,4}.\vec{x}_{2,3})$. That can be algebraically rearranged to highlight the coefficients on the particular combinations:

$$\mathbf{(+1)}\,\vec{x}_{1,3}.\vec{x}_{2,1} + \mathbf{(-1)}\,\vec{x}_{1,3}.\vec{x}_{2,3} + \mathbf{(-1)}\,\vec{x}_{1,4}.\vec{x}_{2,1} + \mathbf{(+1)}\,\vec{x}_{1,4}.\vec{x}_{2,3}$$

Those highlighted coefficients can be obtained as the outer product of main-effect contrasts, namely, the contrast $\vec{c}_1 = \langle 0, 0, +1, -1\rangle$, which expresses CHEM minus THTR, and the contrast $\vec{c}_2 = \langle +1, 0, -1\rangle$, which expresses FT1 minus FT3:

$$\vec{c}_1 \otimes \vec{c}_2 = \begin{array}{cccccc} & \vec{x}_{1,1} & \vec{x}_{1,2} & \vec{x}_{1,3} & \vec{x}_{1,4} & \\ \langle & 0 & 0 & +1 & -1 & \rangle \end{array} \bigotimes \begin{array}{ccccc} & \vec{x}_{2,1} & \vec{x}_{2,2} & \vec{x}_{2,3} & \\ \langle & +1 & 0 & -1 & \rangle \end{array}$$

$$= \begin{array}{c|ccc|} & \vec{x}_{2,1} & \vec{x}_{2,2} & \vec{x}_{2,3} \\ \vec{x}_{1,1} & 0 & 0 & 0 \\ \vec{x}_{1,2} & 0 & 0 & 0 \\ \vec{x}_{1,3} & +1 & 0 & -1 \\ \vec{x}_{1,4} & -1 & 0 & +1 \end{array}$$

Notice that the coefficients in the matrix match the highlighted coefficients in the difference of differences that was expressed a few sentences previously. The posterior of this interaction contrast is shown in the bottom-left histogram of Figure 19.5. The mean of 34,600 indicates that the difference between FT1 and FT2 is about 34,600 greater for CHEM than for THTR. The 95% HDI clearly excludes zero, indicating that this interaction contrast is credibly nonzero.

Interaction contrasts can involve "complex" comparisons just as simply as pairwise comparisons. For example, suppose we are interested in comparing BFIN against the average of the other nonbusiness departments, specifically for a contrast between FT1 and lesser ranks. This interaction contrast is expressed as $\langle +1, -1/3, -1/3, -1/3 \rangle \otimes \langle +1, -1/2, -1/2 \rangle$. The posterior of this contrast is shown in the bottom-right histogram of Figure 19.5. (The label on the x axis exceeds the margins of the figure because there are 12 combinations of levels involved in the contrast specification.) The result suggests that there is considerable uncertainty in the larger difference, between FT1 and other ranks, in BFIN than in other departments. Therefore, we would not want to conclude that the interaction contrast is credibly nonzero. Exercise 19.2 gives you hands-on practice with specification of interaction contrasts.

19.1.5 Noncrossover Interactions, Rescaling, and Homogeneous Variances

When interpreting interactions, it can be important to consider the scale on which the data are measured. This is because an interaction means nonadditive effects when measured in the current scale. If the data are nonlinearly transformed to a different scale, then the nonadditivity can also change.

Consider an example, using utterly fictional numbers merely for illustration. Suppose the average salary of Democratic women is 10 monetary units, for Democratic men it's 12 units, for Republican women it's 15 units, and for Republican men it's 18 units. These data indicate that there is a nonadditive interaction of political party and gender, because the change in pay from women to men is 2 units for Democrats, but 3 units for Republicans. Another way of describing the interaction is to notice that the change in pay from Democrats to Republicans is 5 units for women but 6 units for men. A researcher might be tempted to interpret the interaction as indicating some extra advantage attained by Republican men, or some special disadvantage suffered by Democratic women. But such an interpretation may be inappropriate, because a mere rescaling of the data makes the interaction disappear, as will be described next.

Salary raises and comparisons are often measured by percentages and ratios, not by additive or subtractive differences. Consider the salary data in percentage terms. Among Democrats, men make 20% more than women. Among Republicans, the men again make 20% more than the women. Among women, Republicans make 50% more than Democrats. Among men, Republicans again make 50% more than Democrats. In these ratio terms, there is no interaction of gender and political party: Change from female to male predicts a 20% increase in salary regardless of party, and change from Democrat to Republican predicts a 50% increase in salary regardless of gender.

Equal ratios are transformed to equal distances by a logarithmic transformation. If we measure salary in terms of the logarithm of monetary units, then the salary of Democratic women is $\log_{10}(10) = 1.000$, the salary of Democratic men is $\log_{10}(12) = 1.079$, the salary of Republican women is $\log_{10}(15) = 1.176$, and the salary of Republican men is $\log_{10}(18) = 1.255$. With this logarithmic scaling, the increase in salary from women to men is 0.079 for both parties, and the increase from Democrat to Republican is 0.176 for both genders. In other words, when salary is measured on a logarithmic scale, there is no interaction of gender and political party.

It may seem strange to measure salary on a logarithmic scale, but there are many situations for which the scale is arbitrary. The pitch of a sound can be measured in terms of frequency (i.e., cycles per second) or in terms of perceived pitch, which is essentially the logarithm of the frequency. The magnitude of an earthquake can be measured by its energy or by its value on the Richter scale, which is the logarithm of energy. The pace of a dragster on a racetrack can be measured by the average speed during the run or by the duration from start to finish (which is the reciprocal of average speed). Thus, measurement scales are not unique and are instead determined by convention.

The general issue is illustrated in Figure 19.6. Suppose that predictor x_1 has two levels, as does predictor x_2. Suppose we have three data points at each combination of levels, yielding 12 data points altogether. The means at each combination of levels are shown in the top-left graph of Figure 19.6. You can see that there is an interaction, with the effect of x_1 being bigger when $x_2 = 2$ than when $x_2 = 1$. But this interaction goes away when the data are transformed by taking the logarithm, as shown in the lower-left graph. Each individual data point was transformed, and then the means in each cell were computed. Of course, the transformation can go the other way: Data with no interaction, as in the lower-left plot, can be made to have an interaction when they are rescaled as in the upper-left plot, via an exponential transformation.

The transformability from interaction to noninteraction is only possible for *noncrossover* interactions. This terminology is merely a description of the graph: The lines do not cross over each other (and they have the same sign slope). In this situation, the y axis can have different portions stretched or shrunken so that the lines become parallel. If, however, the lines cross, as in the middle column of Figure 19.6, then there is no way to uncross the lines merely by stretching or shrinking intervals of the y axis. The right column of Figure 19.6 shows the same data as the middle column, but it is plotted with the roles of x_1 and x_2 exchanged. When plotted this way, the lines do not cross, but they do have *opposite-sign slopes* (i.e., one slope is positive and the other slope is negative). There is no way that stretching or shrinking the y axis can change the signs of the slopes, hence the interaction cannot be removed merely by

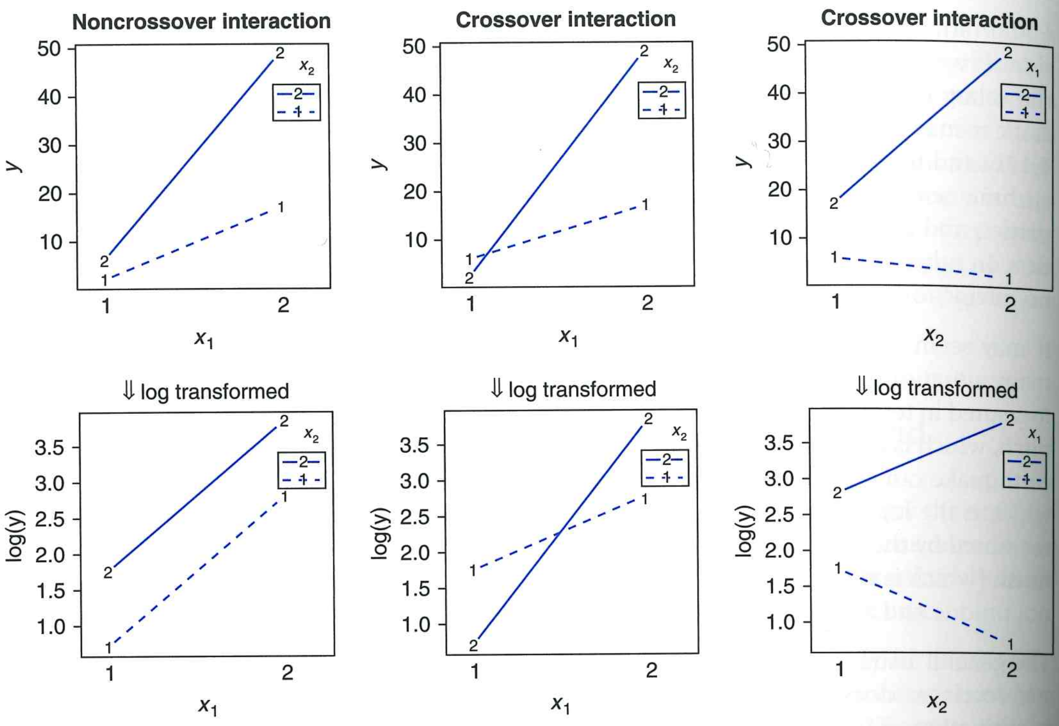


FIGURE 19.6
Top row shows means of original data; bottom row shows means of logarithmically transformed data. Left column shows a noncrossover interaction; middle and right columns show the same crossover interaction plotted against x_1 or x_2.

transforming the data. Because these data have crossing lines when plotted as in the middle column, they are said to have a crossover interaction even when they are plotted as in the right column. (Test your understanding: Is the interaction in Figure 19.1 a crossover interaction?)

It is important to note that the transformation applies to individual raw data values, not to the means of the conditions. A consequence of transforming the data, therefore, is changes in the variances of the data within each condition. For example, suppose one condition has data values of 100, 110, and 120, whereas a second condition has data values of 1100, 1110, and 1120. For both conditions, the variance is 66.7 (i.e., there is homogeneity of variance). When the data are logarithmically transformed, the variance of the first group becomes 1.05e−3, but the variance of the second group becomes two orders of magnitude smaller, namely, 1.02e−5 (i.e., there is *not* homogeneity of variance).

Therefore, when applying the hierarchical model of Figure 19.2, we must be aware that it assumes homogeneity of variance. If we transform the data, we

are chang
variances
either the
of varianc

The mod
which m
mally dis
shape of t
non-norm
function.

In summa
noncrossc
interactio
is the onl
nant scale
nonadditi
reasonabl
might be r
ever, no re
change the
Be sure tha
homogene
on whatev
"hands on

19.2 RE W

In many s
the predic
ple can pr
appear in
tone. The s
nondomin
ity and ha
combinatio
either a to
or nondom
uation is so
predictors
design. I fa
essential as

are changing the variances within the levels of the predictors. The transformed variances might or might not be fairly homogeneous. If they are not, then either the data should be transformed in such a way as to respect homogeneity of variance or the model should be changed to allow unequal variances.

The models we have been using also assume a normal likelihood function, which means that the data at any level of the predictors should be normally distributed. When the data are transformed to a different scale, the shape of their distribution also changes. If the distributions become radically non-normal, it may be misleading to use a model with a normal likelihood function. For a discussion of these issues, review Section 15.1.4, p. 399.

In summary, this section has made two main points. First, if you have a noncrossover interaction, be careful what you claim about it. A noncrossover interaction merely means nonadditivity *in the scale you are using*. If this scale is the only meaningful scale, or if the this scale is the overwhelmingly dominant scale used in that field of research, then you can cautiously interpret the nonadditive interaction with respect to that scale. But if transformed scales are reasonable, then keep in mind that nonadditivity is scale specific, and there might be no interaction in a different scale. With a crossover interaction, however, no rescaling can undo the interaction. Second, nonlinear transformations change the within-cell variances and the shapes of the within-cell distributions. Be sure that the model you are using is appropriate to the homogeneity or nonhomogeneity of variances in the data and to the shapes of the distributions, on whatever scale you are using. Exercise 19.1 has you consider these issues "hands on."

19.2 REPEATED MEASURES, A.K.A. WITHIN-SUBJECT DESIGNS

In many situations, a single "subject" contributes data to multiple levels of the predictors. For example, suppose we are interested in how quickly people can press a button in response to a stimulus onset. The stimulus could appear in the visual modality as a light, or in the auditory modality as a tone. The subject could respond with his or her dominant hand or with the nondominant hand. Thus, there are two nominal predictors, namely modality and hand. The new aspect is that a single subject contributes data to all combinations of the predictors. On many successive trials, the subject gets either a tone or light and is instructed to respond with either the dominant or nondominant hand. Because every subject is measured many times, this situation is sometimes called a *repeated measures* design. Because the levels of the predictors change within subjects, this situation is also called a *within-subject* design. I favor the latter terminology because it more explicitly connotes the essential aspect of the design, that the same subject contributes data in more than one condition. Within-subject designs are contrasted with *between-subject* designs, in which different subjects contribute data to different levels of the predictors.

When every subject contributes many data points to every combination of predictors, then the model of the situation is a straightforward extension of the models we've already considered. We merely add "subject" as another predictor in the model, with each individual subject being a level of the predictor. If there is one predictor other than subject, the model becomes

$$y = \beta_0 + \overrightarrow{\beta}_1 \overrightarrow{x}_1 + \overrightarrow{\beta}_S \overrightarrow{x}_S + \overrightarrow{\beta}_{1\times S} \overrightarrow{x}_{1\times S}$$

This is exactly the two-predictor model we have already considered, with the second predictor being subject. When there are two predictors other than subject, the model becomes

$$\begin{aligned} y = \beta_0 &+ \overrightarrow{\beta}_1 \overrightarrow{x}_1 + \overrightarrow{\beta}_2 \overrightarrow{x}_2 + \overrightarrow{\beta}_S \overrightarrow{x}_S + \overrightarrow{\beta}_{1\times 2} \overrightarrow{x}_{1\times 2} + \overrightarrow{\beta}_{1\times S} \overrightarrow{x}_{1\times S} \\ &+ \overrightarrow{\beta}_{2\times S} \overrightarrow{x}_{2\times S} + \overrightarrow{\beta}_{1\times 2\times S} \overrightarrow{x}_{1\times 2\times S} \end{aligned}$$

This model includes all the two-way interactions of the factors, plus the three-way interaction. Again, subject merely plays the role of the third predictor.

The preceding model, which includes all the high-order interactions with subject, is fine in principle but may be overkill in practice. Unless you have specific theoretical motivations to seek out and interpret high-order interactions of subject with other predictors, there is little reason to model them, and there is difficulty making sense of them even if you did model them. Instead, if you have many data points from each subject in every cell, an alternative approach is to apply a Bayesian ANOVA model to each subject's data, and then put a higher-order prior across the subject parameter estimates, so that different subjects mutually inform each other's estimates and provide a stable group-level estimate. Thus, every subject has a baseline, β_{0s}, and there is a higher-order, group-level prior on the distribution of β_{0s} across subjects. Each predictor also has subject-specific estimates, with the effect of the j^{th} level of predictor 1 denoted $\beta 1s, j$. Each of these effect parameters has a higher, group-level prior across subjects. (This was the modeling approach taken for repeated measures in simple linear regression in Section 16.3, p. 433.) Finally, the group-level effects have a hyperprior that provides shrinkage on the effects of a predictor. In other words, the shrinkage prior, on effects of a predictor, is set at the group level, not at the subject level.

There are other situations, however, in which each subject contributes only one datum to a combination of the other predictors. For example, in the case of the response-time study described earlier, perhaps we have only the median response time of the subject in each combination of hand and modality. As another example, suppose the value to be predicted is IQ, as measured by a lengthy exam, with one predictor being noisy versus quiet exam environment and the other predictor being paper versus computerized exam format. Although it is conceivable that subjects could be repeatedly tested in each condition, it would be challenging enough to get people to sit through all four combinations even once. Thus, each subject would contribute one value to each condition.

In the situation when each subject contributes only one datum per condition, the models described earlier, with all the interaction terms, are not *identifiable*, meaning that there are more parameters than data points. The simplest case of this situation is trying to estimate the mean and variance of a normal distribution from a single data point. A Bayesian analysis can still be conducted, but there will be high uncertainty in the parameter estimates, governed largely by the priors. Therefore, instead of attempting to estimate all the interactions of subjects with other predictors, we assume a simpler model in which the only influence of subjects is on the baseline:

$$y = \beta_0 + \vec{\beta}_S \vec{x}_S + \vec{\beta}_1 \vec{x}_1 + \vec{\beta}_2 \vec{x}_2 + \vec{\beta}_{1\times 2} \vec{x}_{1\times 2}$$

In other words, we assume a main effect of subject but no interaction of subject with other predictors. In this model, the subject effect (deflection) is constant across treatments, and the treatment effects (deflections) are constant across subjects. Notice that the model makes no requirement that every subject contributes a datum to every condition. Indeed, the model allows zero or more than one datum per subject per condition. As mentioned earlier, the computations in Bayesian ANOVA make no assumptions or requirements that the design is "balanced." If you do have many observations per subject in every combination of predictors, then one of the previously described models may be considered.

19.2.1 Why Use a Within-Subject Design? And Why Not?

The primary reason to use a within-subject design is that you can achieve much greater precision in the estimates of the effects than in a between-subject design. For example, suppose you are interested in measuring the effect on response time of using the dominant versus nondominant hand. Suppose there is a population of four subjects from whom you could measure data. If we could measure every subject in every condition, we would know that for the first subject, his or her response times for dominant and nondominant hands are 300 and 320 msec. For the second subject, the response times are 350 and 370. For the third subject, the response times are 400 and 420, and for the fourth subject, the response times are 450 and 470. Thus, for every subject, the difference between dominant and nondominant hands is 20 msec. Suppose we have the resources to measure only two data points in each condition. We measure response times from the dominant hands of two subjects. Should we measure response times from the nondominant hands of the *same* two subjects or the nondominant hands of two *other* subjects? If we measure from the same two subjects, then the estimated effect for each subject is 20 msec, and we have high certainty in the magnitude of the effect. If we measure from two other subjects, then the estimated effect of dominant versus nondominant hand is the average of the first two subjects versus the average of the second two subjects, and the difference is badly affected by random sampling. The between-subject design yields lower precision in the estimate of the effect. Exercise 19.3 has you examine, hands on, a case of this improvement in precision.

Because of the gain in precision, it is desirable to use within-subject designs. But there are many dangers of within-subject designs that need to be considered before they are applied in any particular situation. The key problem is that, in most situations, when you measure the subject you change the subject, and therefore subsequent measurements are not measuring the same subject. The simplest examples of this are mere fatigue or generic practice effects. In measures of response time, if you measure repeatedly from the same subject, you will find improvement over the first several trials because of the subject gaining practice with the task, but after a while, as the subject tires, there will be a decline in performance. The problem is that if you measure the dominant hand in the early trials and the nondominant hand in the later trials, then the effect of practice or fatigue will contaminate the effect of handedness. The repeated measurement process affects and contaminates the measure that is supposed to be a signature of the predictor.

Practice and fatigue effects can be overcome by randomly distributing and repeating the conditions throughout the repeated measures, *if* the practice and fatigue effects influence all conditions equally. Thus, if practice improves both the dominant and nondominant hand by 50 msec, then the difference between dominant and nondominant hands is unaffected by practice. But practice might affect the nondominant hand much more than the dominant hand. You can imagine that in complex designs with many predictors, each with many levels, it can become difficult to justify an assumption that repeated measures have comparable effects on all conditions.

Worse yet, in some situations there can be *differential carryover effects* from one condition to the next. For example, having just experienced practice in the visual modality with the nondominant hand might improve subsequent performance in the auditory modality with the nondominant hand, but it might not improve subsequent performance in the visual modality with the dominant hand. Thus, the carryover effect is different for different subsequent conditions.

When you suspect strong differential carryover effects, you may be able to explicitly manipulate the ordering of the conditions and measure the carryover effects, but this might be impossible mathematically and impractical, depending on the specifics of your situation. In this case, you must revert to a between-subject design and simply include many subjects to attenuate between-subject noise.

In general, all the models we have been using assume independence of observations. The probability of the combined data is the product of the probabilities of the individual data points. When we use repeated measures, this assumption is much less easy to justify. On the one hand, when we repeatedly flip a coin, we might be safe to assume that its underlying bias does not change much from one flip to the next. But, on the other hand, when we repeatedly test the response time of a human subject, it is less easy to justify an assumption that the underlying response time remains unaffected by the previous trial. Researchers will often make the assumption of independence merely as an approximation of convenience, hoping that by arranging conditions randomly across many repeated measures, the differential carryover effects will be minimized.

19.3 R CODE

19.3.1 Bayesian Two-Factor ANOVA

Several implementation details of the program are the same as the oneway ANOVA program of the previous chapter:

- Data are normalized so that prior constants can be more generic.
- Initialization of chains is based on the data. It is important to do this, otherwise burn-in can take forever.
- Because there is nasty autocorrelation, we use a large thinning constant and we also use multiple chains. For a reminder of the issues of burn-in and thinning, see Section 23.2, p. 623.

A new detail arises in how the uncentered parameter estimates are recentered to respect the sum-to-zero constraints. The uncentered estimates from BUGS are `a0`, `a1[]`, `a2[]`, and `a1a2[,]`. By definition of the ANOVA model, the predicted mean of cell *i*, *j* is

```
m[i,j] = a0 + a1[i] + a2[j] + a1a2[i,j]
```

We use these predicted means to construct the zero-centered parameters. First, `b0` is the mean across all the predicted means:

```
b0 = mean( m[,] )
```

Then the main effect deflections are the marginal means minus the overall mean:

```
b1[i] = mean( m[i,] ) - b0
b2[j] = mean( m[,j] ) - b0
```

It is easy (honest!) to check that those deflections do indeed sum to zero (i.e., sum(b1[]) = 0 and sum(b2[]) = 0). Finally, the interaction deflections are the residuals after the additive effect of b1 and b2 is taken into account:

```
b1b2[i,j] = m[i,j] - ( b0 + b1[i] + b2[j] )
```

Again, it is easy to check that the rows and columns of b1b2[,] all sum to zero.

In the data section of the program, one option is to load data from the article of Qian & Shen (2007). The program here uses a hierarchical structure similar to that used by Qian & Shen (2007), but their program did not recenter the parameters as is done here. It may be instructive to compare the results of the program here with the results reported by Qian & Shen (2007).

BUGS for many factors. The program that follows applies only for cases of two nominal predictors. If you have many nominal predictors, along with their two-way, three-way, and higher-order interactions, it becomes unwieldy to explicitly and separately name all the deflection parameters. Instead, it can be more elegant to use a technique of *dummy coding*, whereby we essentially revert back to using vectors for coding the values of the predictors instead of integer indices. That is, $\overrightarrow{x}_1 = \textit{level}\ 2$ is coded by the "dummy" vector $\langle 0, 1, 0, \ldots \rangle$ instead of by the integer index 2. Interactions are represented by matrices of dummy codes that have been flattened into vectors. For an example of programming this technique in BUGS, see Ntzoufras (2009, Ch. 6). Unfortunately, those examples do not incorporate the higher-level prior structure emphasized in Figure 19.2.

(ANOVAtwowayBRugs.R)

```
graphics.off()
rm(list=ls(all=TRUE))
fnroot = "ANOVAtwowayBrugs"
library(BRugs)         # Kruschke, J. K. (2010). Doing Bayesian data analysis:
                       # A Tutorial with R and BUGS. Academic Press / Elsevier.
#------------------------------------------------------------------------------
# THE MODEL.

modelstring = "
# BUGS model specification begins here...
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dnorm( mu[i] , tau )
    mu[i] <- a0 + a1[x1[i]] + a2[x2[i]] + a1a2[x1[i],x2[i]]
  }
  #
  tau <- pow( sigma , -2 )
  sigma ~ dunif(0,10) # y values are assumed to be standardized
  #
  a0 ~ dnorm(0,0.001) # y values are assumed to be standardized
  #
  for ( j1 in 1:Nx1Lvl ) { a1[j1] ~ dnorm( 0.0 , a1tau ) }
  a1tau <- 1 / pow( a1SD , 2 )
  a1SD <- abs( a1SDunabs ) + .1
  a1SDunabs ~ dt( 0 , 0.001 , 2 )
  #
  for ( j2 in 1:Nx2Lvl ) { a2[j2] ~ dnorm( 0.0 , a2tau ) }
  a2tau <- 1 / pow( a2SD , 2 )
  a2SD <- abs( a2SDunabs ) + .1
  a2SDunabs ~ dt( 0 , 0.001 , 2 )
  #
  for ( j1 in 1:Nx1Lvl ) { for ( j2 in 1:Nx2Lvl ) {
    a1a2[j1,j2] ~ dnorm( 0.0 , a1a2tau )
  } }
  a1a2tau <- 1 / pow( a1a2SD , 2 )
  a1a2SD <- abs( a1a2SDunabs ) + .1
  a1a2SDunabs ~ dt( 0 , 0.001 , 2 )
}
# ... end BUGS model specification
" # close quote for modelstring
# Write model to a file, and send to BUGS:
writeLines(modelstring,con="model.txt")
modelCheck( "model.txt" )

#------------------------------------------------------------------------------
# THE DATA.
# Specify data source:
dataSource = c( "QianS2007" , "Salary" , "Random" , "Ex19.3" )[4]

# Load the data:
if ( dataSource == "QianS2007" ) {
  fnroot = paste( fnroot , dataSource , sep="" )
  datarecord = read.table( "QianS2007SeaweedData.txt" , header=TRUE , sep="," )
  # Logistic transform the COVER value:
  # Used by Appendix 3 of QianS2007 to replicate Ramsey and Schafer (2002).
  datarecord$COVER = -log( ( 100 / datarecord$COVER ) - 1 )
  y = as.numeric(datarecord$COVER)
  x1 = as.numeric(datarecord$TREAT)
  x1names = levels(datarecord$TREAT)
  x2 = as.numeric(datarecord$BLOCK)
  x2names = levels(datarecord$BLOCK)
  Ntotal = length(y)
  Nx1Lvl = length(unique(x1))
  Nx2Lvl = length(unique(x2))
```

```
                           dim=c(NROW(contrast),NCOL(contrast),chainLength) )
      contrastLab = ""
      for ( x1idx in 1:Nx1Lvl ) {
        for ( x2idx in 1:Nx2Lvl ) {
          if ( contrast[x1idx,x2idx] != 0 ) {
            contrastLab = paste( contrastLab , "+" ,
                                 signif(contrast[x1idx,x2idx],2) ,
                                 x1names[x1idx] , x2names[x2idx] )
          }
        }
      }
      histInfo = plotPost( apply( contrastArr * b1b2Sample , 3 , sum ) ,
               compVal=0 , breaks=30 , xlab=contrastLab , cex.lab = 0.75 ,
               main=paste( names(x1x2contrastList)[cIdx] ) )
  }
  dev.copy2eps(file=paste(fnroot,"x1x2Contrasts.eps",sep=""))
}

#=============================================================================
# Do NHST ANOVA:

theData = data.frame( y=y , x1=factor(x1,labels=x1names) ,
                            x2=factor(x2,labels=x2names) )
windows()
interaction.plot( theData$x1 , theData$x2 , theData$y , type="b" )
dev.copy2eps(file=paste(fnroot,"DataPlot.eps",sep=""))
aovresult = aov( y ~ x1 * x2 , data = theData )
cat("\n-----------------------------------------------------------------\n\n")
print( summary( aovresult ) )
cat("\n-----------------------------------------------------------------\n\n")
print( model.tables( aovresult , type = "effects", se = TRUE ) , digits=3 )
cat("\n-----------------------------------------------------------------\n\n")

#=============================================================================
```

19.4 EXERCISES

Exercise 19.1. **[Purpose: To inspect an interaction for transformed data.]** Consider the data plotted in Figure 19.3, p. 521.

(A) Is the interaction a crossover interaction or not? Briefly explain your answer.

(B) Suppose we are interested in salaries thought of in terms of *percentage* (i.e., ratio) differences rather than additive differences. Therefore we take the logarithm, base 10, of the individual salaries (the R code has this option built into to the data section, where the salary data are loaded). Run the analysis on the transformed data, producing the results and contrasts analogous to those in Figures 19.4 and 19.5. Do any of the conclusions change?

(C) Examine the within-cell variances in the original and in the transformed data. (Hint: Try using the `aggregate` function on the data. As a guide, see how the function is used to initialize `a1a2`. Instead of applying the mean to the aggregated data, apply the standard deviation. The result is the within-cell standard deviations. Are they all roughly the same?) Do the original or the transformed data better respect the assumption of homogeneous variances?

Exercise 19.2. **[Purpose: To investigate a case of two-factor Bayesian ANOVA.]** In the data specification of the program in Section 19.3.1 (ANOVAtwowayBRugs), you can load data from Qian & Shen (2007), regarding how quickly seaweed regenerates when in the presence of different types of grazers. Data were collected from eight different tidal areas of the Oregon coast; this predictor ($\vec{x}_2$) is referred to as the *Block*. In each of the eight Blocks, there were six different combinations of seaweed grazers established by the researchers. This predictor ($\vec{x}_1$) had six levels: control, with no grazers allowed; only small fish allowed; small and large fish allowed; only limpets allowed; limpets and small fish allowed; and all three grazers allowed. The predicted variable was the percentage of the plot covered by regenerated seaweed, logarithmically transformed.

(A) Load the data and run the program. You will find that there are too many levels of the two predictors to fit all the posterior histograms into a single multipanel display. Therefore, modify the plotPost.R program so that it produces only the mean and HDI limits, marked by a horizontal bar with a circle at the mean (without a histogram) and perhaps without a main title. Name your program something other than plotPost.R, and use it in the plotting section at the end of the program instead of plotPost.R. Show your results. (A secondary goal of this part of the exercise is to give you experience modifying graphics in R to suit your own purposes.) Hints: There are many ways to do this, but here are some options. To suppress plotting of the histogram, just put this argument in the hist function: plot=F. To suppress a title on a plot, just use the argument main="". To adjust the font size, specify the "character expansion": cex for text, cex.lab for axis labels, and so forth. To reduce the margins around a plot, so there is more room for the plot itself, try variations of these margin specifications: par(mar=c(2,0.5,1,0.5), mgp=c(0.5,0,0)). The par command needs to be called before the plots are made.

(B) The program already includes contrasts that consider whether there is an effect of small fish, an effect of large fish, and an effect of limpets. What conclusions do you reach from the posteriors of these contrasts?

(C) Construct a contrast of the average of Blocks 3 and 4 versus the average of Blocks 1 and 2. Show your specification, the graph of the posterior on the contrast, and state your conclusion.

(D) Is the effect of limpets different in Block 6 than in Block 7? To answer this question, construct an interaction contrast using an outer product (Hint: refer to the already-coded L_effect for the contrast that specifies the effect of limpets). Is the effect of small fish different in Blocks 1 and 7 than in Blocks 3 and 5? For both questions, show the contrast vectors that you constructed and show the posterior of the contrast, and state your conclusion.

Exercise 19.3. **[Purpose: To notice that within-subject designs can be more sensitive (hence more powerful) than between-subject designs.]** Consider these data:

$\vec{x}_1$	$\vec{x}_2$	y	S
1	1	101	1
1	1	102	2
1	1	103	3
1	1	105	4
1	1	104	5
1	2	104	1
1	2	105	2
1	2	107	3
1	2	106	4
1	2	108	5
2	1	105	1
2	1	107	2
2	1	106	3
2	1	108	4
2	1	109	5
2	2	109	1
2	2	108	2
2	2	110	3
2	2	111	4
2	2	112	5

Note: The table is split into two halves so it fits the page more compactly. The continuation of the first column appears in the fifth column. The continuation of the second column appears in the sixth column, and so forth.

(A) Ignoring the last column, which indicates the subject who generated the data, conduct a Bayesian ANOVA using $\vec{x}_1$ and $\vec{x}_2$ as predictors of y. Show the code you used to load the data, and show the resulting posterior histograms of β_0, $\beta_{1,j}$, $\beta_{2,k}$, and $\beta_{1\times 2,jk}$. Also show the posterior of the contrast $\beta_{1,2} - \beta_{1,1}$ (i.e., the marginal difference between levels 1 and 2 of factor 1, also called the *main effect of factor 1*) and the posterior of the contrast $\beta_{2,2} - \beta_{2,1}$ (i.e., the marginal difference between levels 1 and 2 of factor 2, also called the *main effect of factor 2*).

(B) Now include the subject as a predictor by expanding the model to include a deflection from baseline due to subject. (Do not include any subject interaction terms.) Again show the posteriors of the β's requested in the previous part. Are the certainties on the estimates and contrasts different than in the previous part? In what way, and why?
(Hint regarding the answer: Figure 19.7 shows posterior histograms for the main effect of factor 2, when the data are considered to be between subject or within subject. Notice that the means are essentially the same in both histograms, but the uncertainties are very different!

Programming hints: The model specification without a subject factor is

```
mu[i] <- a0 + a1[x1[i]] + a2[x2[i]] + a1a2[x1[i],x2[i]]
```

but with a subject factor becomes

```
mu[i] <- a0 + a1[x1[i]] + a2[x2[i]] + a1a2[x1[i],x2[i]] + aS[S[i]]
```

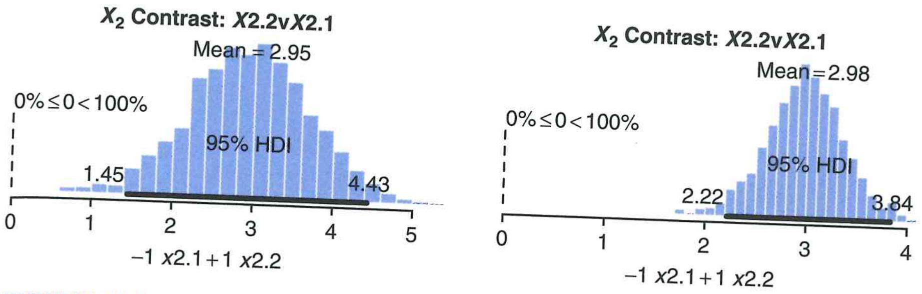


FIGURE 19.7

For Exercise 19.3. *Left panel:* Posterior for difference between levels of factor 2 when data are considered to be between subject. *Right panel:* Posterior for difference between levels of factor 2 when data are considered to be within subject. Notice that the means are (essentially) the same in both histograms, but the uncertainties are very different!

where `S[i]` is the subject number for the `i`th datum, and `aS[]` are the deflections from baseline for each subject. You must, of course, specify a prior on `aS[]` analogous to the prior on `a1[]`.

The conversion of the `a[]` values to zero-centered `b[]` values proceeds analogously to what was explained at the beginning of Section 19.3.1 (`ANOVAtwowayBRugs.R`). The code merely needs to be expanded to include the additional subject factor. Here is a guide (ANOVAtwowayBRugsWithinSubj.R):

```
# Convert the a values to zero-centered b values.
# m12Sample is predicted cell means at every step in MCMC chain:
m12Sample = array( 0, dim=c( datalist$Nx1Lvl , datalist$Nx2Lvl ,
                             datalist$NSLvl , chainLength ) )
for ( stepIdx in 1:chainLength ) {
  for ( a1idx in 1:Nx1Lvl ) {
    for ( a2idx in 1:Nx2Lvl ) {
      for ( aSidx in 1:NSLvl ) {
        m12Sample[ a1idx , a2idx , aSidx , stepIdx ] = (
           a0Sample[stepIdx]
           + a1Sample[a1idx,stepIdx]
           + a2Sample[a2idx,stepIdx]
           + a1a2Sample[a1idx,a2idx,stepIdx]
           + aSSample[aSidx,stepIdx] )
      }
    }
  }
}

# b0Sample is mean of the cell means at every step in chain:
b0Sample = apply( m12Sample , 4 , mean )
# b1Sample is deflections of factor 1 marginal means from b0Sample:
b1Sample = ( apply( m12Sample , c(1,4) , mean )
             - matrix(rep( b0Sample ,Nx1Lvl),nrow=Nx1Lvl,byrow=T) )
# b2Sample is deflections of factor 2 marginal means from b0Sample:
b2Sample = ( apply( m12Sample , c(2,4) , mean )
           - matrix(rep( b0Sample ,Nx2Lvl),nrow=Nx2Lvl,byrow=T) )
# bSSample is deflections of factor S marginal means from b0Sample:
bSSample = ( apply( m12Sample , c(3,4) , mean )
           - matrix(rep( b0Sample ,NSLvl),nrow=NSLvl,byrow=T) )
# linpredSample is linear combination of the marginal effects:
linpredSample = 0*m12Sample
for ( stepIdx in 1:chainLength ) {
  for ( a1idx in 1:Nx1Lvl ) {
    for ( a2idx in 1:Nx2Lvl ) {
      for ( aSidx in 1:NSLvl ) {
        linpredSample[a1idx,a2idx,aSidx,stepIdx ] = (
           b0Sample[stepIdx]
           + b1Sample[a1idx,stepIdx]
           + b2Sample[a2idx,stepIdx]
           + bSSample[aSidx,stepIdx] )
      }
    }
  }
}
# b1b2Sample is the interaction deflection, i.e., the difference
# between the cell means and the linear combination:
b1b2Sample = apply( m12Sample - linpredSample , c(1,2,4) , mean )

# Convert from standardized b values to original scale b values:
b0Sample = b0Sample * ySDorig + yMorig
b1Sample = b1Sample * ySDorig
b2Sample = b2Sample * ySDorig
bSSample = bSSample * ySDorig
b1b2Sample = b1b2Sample * ySDorig
```

Exercise 19.4. **[Purpose: To conduct a power analysis for Bayesian ANOVA, for within-subject versus between-subject designs.]** Conduct power analyses for the between-subject and within-subject versions of the previous exercise. Specifically, suppose the goal is for the 95% HDI of the contrast on factor 2 to have a width of 2.0 or less. Conduct a retrospective power analysis for this goal, for the within-subject version and the between-subject version. Caution: This exercise demands a lot of programming and could be time consuming, but the results drive home the point that within-subject designs can be more powerful than between-subject designs.