CHAPTER 12

Bayesian Approaches to Testing a Point ("Null") Hypothesis

Contents

Is he a loser or is he real great?
Words without actions make bad estimates.
I need big outcomes when going gets rough, 'cause
Better than nothing just ain't good enough.[1]

Suppose that you have collected some data, and now you want to answer the question, Is there a non-zero effect or not? Is the coin fair or not? Is there better-than-chance accuracy or not? Is there a difference between groups or not? In the previous chapter, I argued that answering this type of question via null hypothesis significance testing (NHST) has deep problems. This chapter describes Bayesian approaches to the question.

In the context of coin flipping, the question we are asking is whether the probability of heads has some particular value. For example, if we are asking whether the coin is fair, we are asking whether a head probability of 0.5 is credible. There are two different ways of formalizing this question in a Bayesian framework. One way to pose the question is to ask whether the value of interest ($\theta = 0.5$) falls among the most credible values

[1] This chapter explains two Bayesian approaches to evaluating null values, one approach based on parameter estimation and the other approach based on comparing nothing with something. The poem suggests the need for estimating magnitudes in real life, as opposed to merely saying something is better than nothing.

Doing Bayesian Data Analysis, 2nd Edition
http://dx.doi.org/10.1016/B978-0-12-405888-0.0.00012-X

in the posterior. A different way to pose the question sets up a dichotomy between, on the one hand, a prior distribution that allows *only* the value of interest, and, on the other hand, a prior distribution that allows a broad range of all possible values. The posterior believability of the two priors is assessed via Bayesian model comparison. This chapter will explore the two approaches in some detail. The conclusions drawn by the two approaches do not yield the same information, so it is important to choose the approach that is most appropriate to the situation and the question you want to ask of your data.

12.1. THE ESTIMATION APPROACH

Throughout this book, we have used Bayesian inference to derive a posterior distribution over a parameter of interest, such as the bias θ of a coin. We can then use the posterior distribution to discern the credible values of the parameter. If the null value is far from the credible values, then we reject the null value as not credible. But if all the credible values are virtually equivalent to the null value, then we can accept the null value. This intuitive decision procedure is given formal expression in this section.

12.1.1. Region of practical equivalence

A *region of practical equivalence* (ROPE) indicates a small range of parameter values that are considered to be practically equivalent to the null value for purposes of the particular application. For example, if we wonder whether a coin is fair, for purposes of determining which team will kick off at a football game, then we want to know if the underlying bias in the coin is reasonably close to 0.50, and we don't really care if the true bias is 0.473 or 0.528, because those values are practically equivalent to 0.50 for our application. Thus, the ROPE on the bias might go from 0.45 to 0.55. As another example, if we are assessing the efficacy of a drug versus a placebo, we might only consider using the drug if it improves the probability of cure by at least 5 percentage points. Thus, the ROPE on the difference of cure probabilities could have limits of ± 0.05. There will be more discussion later of how to set the ROPE limits.

Once a ROPE is set, we make a decision to reject a null value according the following rule:

> *A parameter value is declared to be* not *credible, or rejected, if its entire ROPE lies outside the 95% highest density interval (HDI) of the posterior distribution of that parameter.*

For example, suppose that we want to know whether a coin is fair, and we establish a ROPE that goes from 0.45 to 0.55. We flip the coin 500 times and observe 325 heads. If the prior is uniform, the posterior has a 95% HDI from 0.608 to 0.691, which falls completely outside the ROPE. Therefore we declare that the null value of 0.5 is rejected for practical purposes. *Notice that when the HDI excludes the ROPE, we are not rejecting all values within the ROPE; we are rejecting only the null value.*

Because the ROPE and HDI can overlap in different ways, there are different decisions that can be made. In particular, we can decide to "accept" a null value:

> *A parameter value is declared to be accepted for practical purposes if that value's ROPE completely contains the 95% HDI of the posterior of that parameter.*

With this decision rule, a null value of a parameter can be accepted only when there is sufficient precision in the estimate of the parameter. For example, suppose that we want to know whether a coin is fair, and we establish a ROPE that goes from 0.45 to 0.55. We flip the coin 1,000 times and observe 490 heads. If the prior is uniform, the posterior has a 95% HDI from 0.459 to 0.521, which falls completely within the ROPE. Therefore we declare that the null value of 0.5 is confirmed for practical purposes, because all of the most credible values are practically equivalent to the null value.

In principle, though rarely in practice, a situation could arise in which a highly precise HDI does not include the null value but still falls entirely within a wide ROPE. According to the decision rule, we would "accept" the null value despite it having low credibility according to the posterior distribution. This strange situation highlights the difference between the posterior distribution and the decision rule. *The decision rule for accepting the null value says merely that the most credible values are practically equivalent to the null value according to the chosen ROPE, not necessarily that the null value has high credibility.* If this situation actually arises, it could be a sign that the ROPE is too large and has been ill-conceived or is outdated because the available data are so much more precise than the ROPE.

When the HDI and ROPE overlap, with the ROPE not completely containing the HDI, then neither of the above decision rules is satisfied, and we withhold a decision. This means merely that the current data are insufficient to yield a clear decision one way or the other, according to the stated decision criteria. The posterior distribution provides complete information about the credible parameter values, regardless of the subsequent decision procedure. There are other types of decisions that could be declared, if the HDI and ROPE overlap in different ways. We will not be pursuing those other types of decisions here, but they can be useful in some applied situations. For further discussion of the ROPE, under somewhat different appellations of "range of equivalence" and "indifference zone," see, for example, Carlin and Louis (2009), Freedman, Lowe, and Macaskill (1984), Hobbs and Carlin (2008), and Spiegelhalter, Freedman, and Parmar (1994).

Aside from the intuitive appeal of using a ROPE to declare practical equivalence, there are sound logical reasons from the broader perspective of scientific method. Serlin and Lapsley (1985, 1993) pointed out that using a ROPE to *affirm* a predicted value is essential for scientific progress, and is a solution to Meehl's paradox (e.g., Meehl, 1967, 1978, 1997). Meehl started with the premise that all theories must be wrong, in the sense that they must oversimplify some aspect of any realistic scenario. The magnitude of discrepancy between theory and reality might be small, but there must be

some discrepancy. Therefore, as measurement precision improves (e.g., with collection of more data), the probability of detecting the discrepancy and disproving the theory must increase. This is how it should be: More precise data should be a more challenging test of the theory. But the logic of null hypothesis testing paradoxically yields the opposite result. In null hypothesis testing, all that it takes to "confirm" a (non-null) theory is to reject the null value. Because the null hypothesis is certainly wrong, at least slightly, increased precision implies increased probability of rejecting the null, which means increased probability of "confirming" the theory. What is needed instead is a way to affirm substantive theories, not a way to disconfirm straw-man null hypotheses. Serlin and Lapsley (1985, 1993) showed that by using a ROPE around the predicted value of a theory, the theory can be affirmed. Crucially, as the precision of data increases, and the width of the ROPE decreases, then the theory is tested more stringently.[2]

How is the size of the ROPE determined? In some domains such as medicine, expert clinicians can be interviewed, and their opinions can be translated into a reasonable consensus regarding how big of an effect is useful or important for the application. Serlin and Lapsley (1993, p. 211) said, "Admittedly, specifying [ROPE limits] is difficult. … The width of the [ROPE] depends on the state of the art of the theory and of the best measuring device available. It depends on the state of the art of the theory …[because] a historical look at one's research program or an examination of a competing research program will help determine how accurately one's theory should predict in order that it be competitive with other theories." In other words, the limits of the ROPE depend on the practical purpose of the ROPE. If the purpose is to assess the equivalence of drug-treatment outcomes, then the ROPE limits depend on the real-world costs and benefits of the treatment and the ability to measure the outcome. If the purpose is to affirm a scientific theory, then the ROPE limits depend on what other theories need to be distinguished.

The ROPE limits, by definition, cannot be uniquely "correct," but instead are established by practical aims, bearing in mind that wider ROPEs yield more decisions to accept the ROPEd value and fewer decision to reject the ROPEd value. In many situations, the exact limit of the ROPE can be left indeterminate or tacit, so that the audience of the analysis can use whatever ROPE is appropriate at the time, as competing theories and measuring devices evolve. When the HDI is far from the ROPEd value, the exact ROPE is inconsequential because the ROPEd value would be rejected for any reasonable ROPE. When the HDI is very narrow and overlaps the target value, the HDI might again fall within any reasonable ROPE, again rendering the exact ROPE inconsequential. When, however, the HDI is only moderately narrow and near the target value, the analysis can report how much of the posterior falls

[2] Serlin and Lapsley (1985, 1993) called a ROPE a "good-enough belt" and used frequentist methods, but the logic of their argument still applies.

within a ROPE as a function of different ROPE widths. An example is shown at this book's blog at http://doingbayesiandataanalysis.blogspot.com/2013/08/how-much-of-bayesian-posterior.html.

It is important to be clear that any discrete decision about rejecting or accepting a null value does *not* exhaustively capture our knowledge about the parameter value. Our knowledge about the parameter value is described by the full posterior distribution. When making a binary decision, we have merely compressed all that rich detail into a single bit of information. The broader goal of Bayesian analysis is conveying an informative summary of the posterior, and where the value of interest falls within that posterior. Reporting the limits of an HDI region is more informative than reporting the declaration of a reject/accept decision. By reporting the HDI and other summary information about the posterior, different readers can apply different ROPEs to decide for themselves whether a parameter is practically equivalent to a null value. The decision procedure is separate from the Bayesian inference. The Bayesian part of the analysis is deriving the posterior distribution. The decision procedure uses the posterior distribution, but does not itself use Bayes' rule.

In applications when the Bayesian posterior is approximated with an MCMC sample, it is important to remember the instability of the HDI limits. Recall the discussion accompanying Figure 7.13, p. 185, which indicated that the standard deviation of a 95% HDI limit for a normal distribution, across repeated runs with an MCMC sample that has an effective sample size (ESS) of 10,000, is roughly 5% of the standard deviation of parameter posterior. Thus, if the MCMC HDI limit is very near the ROPE limit, be cautious in your interpretation because the HDI limit has instability due to MCMC randomness. Analytically derived HDI limits do not suffer this problem, of course.

It may be tempting to try to adopt the use of the ROPE in NHST because an NHST confidence interval (CI) has some properties analogous to the Bayesian posterior HDI. The analogous decision rule in NHST would be to accept a null hypothesis if a 95% CI falls completely inside the ROPE. This approach goes by the name of *equivalence testing* in NHST (e.g., Rogers, Howard, & Vessey, 1993; Westlake, 1976, 1981). While the spirit of the approach is laudable, it has two main problems. One problem is technical: CIs can be difficult to determine. For example, with the goal of equivalence testing for two proportions, Dunnett and Gent (1977) devoted an entire article to various methods that approximate the CI for that particular case. With modern Bayesian methods, on the other hand, HDIs can be computed seamlessly for arbitrary complex models; indeed the case of two proportions was addressed in Section 7.4.5, p. 176, and the more complex case of hundreds of grouped proportions was addressed in Section 9.5.1, p. 253. The second problem with frequentist equivalence testing is foundational: A CI does not actually indicate the most credible parameter values. In a Bayesian approach, the 95% HDI actually includes the 95% of parameter values that are most credible. Therefore, when the 95% HDI falls within the ROPE, we can conclude that 95% of the credible

parameter values are practically equivalent to the null value. But a 95% CI from NHST says nothing directly about the credibility of parameter values. Crucially, even if a 95% CI falls within the ROPE, a change of stopping or testing intention will change the CI and the CI may no longer fall within the ROPE. For example, if the two groups being compared are intended to be compared to other groups, then the 95% CI is much wider and may no longer fall inside the ROPE. This dependency of equivalence testing on the set of tests to be conducted is pointed out by Rogers et al. (1993, p. 562). The Bayesian HDI, on the other hand, is not affected by the intended tests.

12.1.2. Some examples

The left side of Figure 9.14 (p. 256) shows the difference of batting abilities between pitchers and catchers. The 95% HDI goes from −0.132 to −0.0994. If we use a ROPE from −0.05 to +0.05 (somewhat arbitrarily), the HDI falls far outside the ROPE, and we reject the null, even taking into account the MCMC instability of the HDI limits.

The right side of Figure 9.14 (p. 256) shows the difference of batting abilities between catchers and first basemen. The 95% HDI goes from −0.0289 to 0.0 (essentially). With any non-zero ROPE, and taking into account the MCMC instability of the HDI limits, we would not want to declare that a difference of zero is rejected, instead saying that the difference is only marginally non-zero. The posterior gives the full information, indicating that there is a suggestion of difference, but the difference is small relative to the uncertainty of its estimate.

The right side of Figure 9.15 (p. 257) shows the difference of batting abilities between two individual players who provided lots of data. The 95% HDI goes from −0.0405 to +0.0368. This falls completely inside a ROPE of −0.05 to +0.05 (even taking MCMC instability into account). Thus, we could declare that a difference of zero is accepted for practical purposes. This example also illustrates that it can take a lot of data to achieve a narrow HDI; in this case both batters had approximately 600 opportunities at bat.

The `plotPost` function, in the utilities that accompany this book, has options for displaying a null value and ROPE. Details are provided in Section 8.2.5.1 (p. 205), and an example graph is shown in Figure 8.4 (p. 205). In particular, the null value is specified with the `compVal` argument (which stands for comparison value), and the ROPE is specified as a two-element vector with the `ROPE` argument. The resulting graph displays the percentage of the posterior on either side of the comparison value, and the percentage of the posterior within the ROPE and on either side of the ROPE limits.

12.1.2.1 Differences of correlated parameters

It is important to understand that the marginal distributions of two parameters do not reveal whether or not the two parameter values are different. Figure 12.1, in its left quartet, shows a case in which the posterior distribution for two parameter values

has a strong positive correlation. Two of the panels show the marginal distributions of the single parameters. Those two marginal distributions suggest that there is a lot of overlap between the two parameters values. Does this overlap imply that we should not believe that they are very different? No! The histogram of the differences shows that the true difference between parameters is credibly greater than zero, with a difference of zero falling well outside the 95% HDI, even taking into account MCMC sampling instability and a small ROPE. The upper left panel shows why: The credible values of the two parameters are highly correlated, such that when one parameter value is large, the other parameter value is also large. Because of this high correlation, the points in the joint distribution fall almost all on one side of the line of equality.

Figure 12.1 shows, in its right quartet, a complementary case. Here, the marginal distributions of the single parameters are exactly the same as before: Compare the histograms of the marginal distributions in the two quartets. Despite the fact that the marginal distributions are the same as before, the bottom right panel reveals that the difference of parameter values now straddles zero, with a difference of zero firmly in the midst of the HDI. The plot of the joint distribution shows why: Credible values of the two parameter are negatively correlated, such that when one parameter value is large, the other parameter value is small. The negative correlation causes the joint distribution to straddle the line of equality.

In summary, the marginal distributions of two parameters do not indicate the relationship between the parameter values. The joint distribution of the two parameters

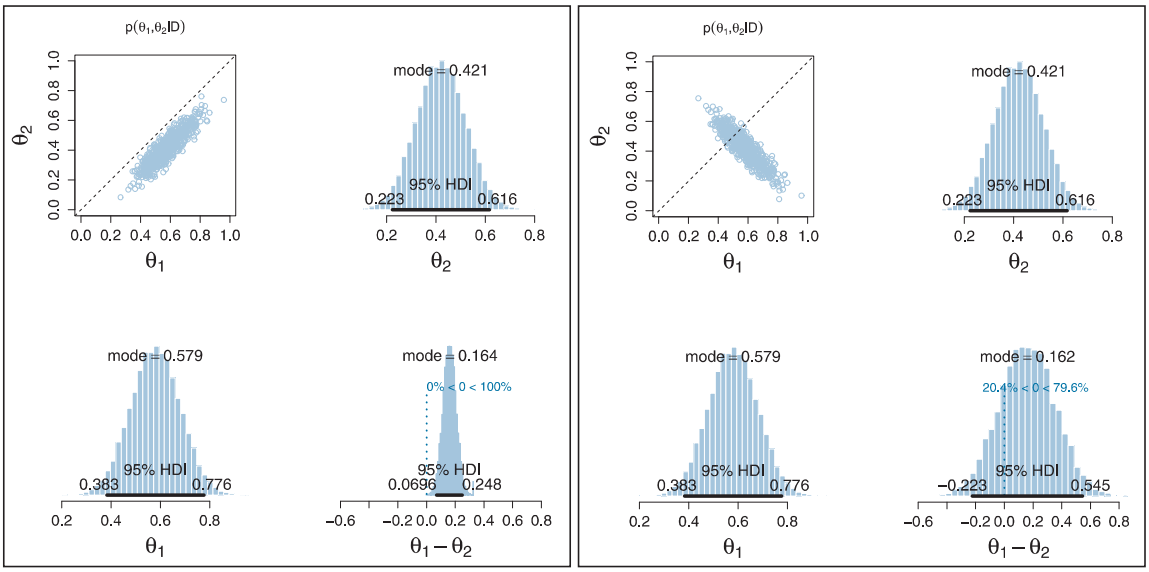

Figure 12.1 When there is a positive correlation between parameters, as shown in the left quartet, the distribution of differences is narrower than when there is a negative correlation, as shown in the right quartet.

might have positive or negative correlation (or even a non-linear dependency), and therefore the difference of the parameter values should be explicitly examined.

12.1.2.2 Why HDI and not equal-tailed interval?

I have advocated using the HDI as the summary credible interval for the posterior distribution, also used in the decision rule along with a ROPE. The reason for using the HDI is that it is very intuitively meaningful: All the values inside the HDI have higher probability density (i.e., credibility) than any value outside the HDI. The HDI therefore includes the most credible values.

Some other authors and software use an *equal-tailed interval* (ETI) instead of an HDI. A 95% ETI has 2.5% of the distribution on either side of its limits. It indicates the 2.5th percentile and the 97.5th percentile. One reason for using an ETI is that it is easy to compute.

In symmetric distributions, the ETI and HDI are the same, but not in skewed distributions. Figure 12.2 shows an example of a skewed distribution with its 95% HDI and 95% ETI marked. (It is a gamma distribution, so its HDI and ETI are easily computed to high accuracy.) Notice on the right there is a region, marked by an arrow, that is outside the HDI but inside the ETI. On the left there is another region marked by an arrow, that is inside the HDI but outside the ETI. The ETI has the strange property that parameter values in the region marked by the right arrow are *included* in the ETI, even though they have *lower credibility* than parameter values in the region marked by the left arrow that are *excluded* from the ETI. This property seems undesirable as a summary of the credible values in a distribution.

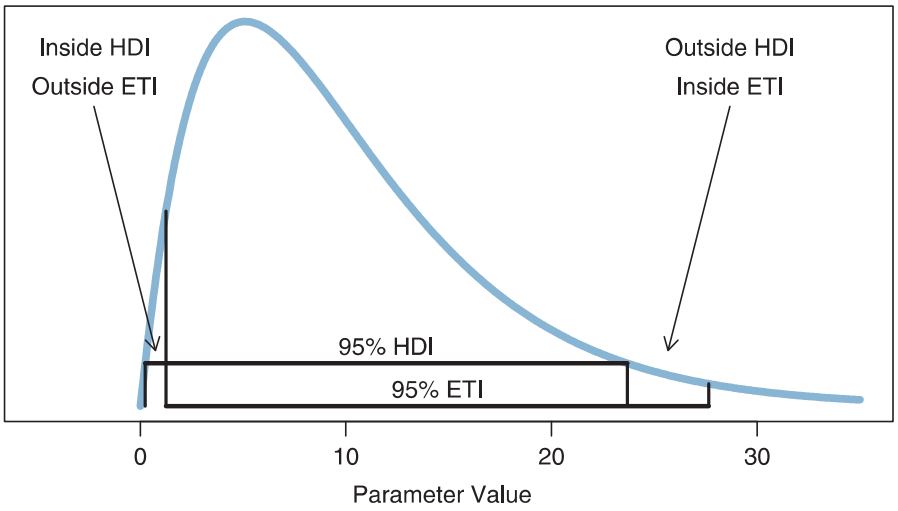

Figure 12.2 A skewed distribution has different 95% highest density interval (HDI) than 95% equal-tailed interval (ETI).

The strange property of the ETI also leads to weirdness when using it as a decision tool. If a null value and ROPE were in the region marked by the right arrow, it would be rejected by the HDI, but not by the ETI. Which decision makes more sense? I think the decision by HDI makes more sense, because it is saying that the values outside its limits have low credibility. But the decision by ETI says that values in this region are *not* rejected, even though they have low credibility. The complementary conflict happens in the region marked by the left arrow. If a null value and ROPE overlap that region, the decision by HDI would be *not* to reject, but the decision by ETI *would* be to reject. Again, I think the decision by HDI makes more sense, because these values have high credibility, even though they are in the extreme tail of the distribution.

Proponents of using the ETI point out that the ETI limits are invariant under nonlinear transformations of the parameter. The ETI limits of the transformed parameter are just the transformed limits of the original scale. This is not the case for HDI limits (in general). This property is handy when the parameters are arbitrarily scaled in abstract model derivations, or in some applied models for which parameters might be nonlinearly transformed for different purposes. But in most applications, the parameters are meaningfully defined on the canonical scale of the data, and the HDI has meaning relative to that scale. Nevertheless, it is important to recognize that if the scale of the parameter is nonlinearly transformed, the HDI limits will change relative to the percentiles of the distribution.

12.2. THE MODEL-COMPARISON APPROACH

Recall that the motivating issue for this chapter is the question, Is the null value of a parameter credible? The previous section answered the question in terms of parameter estimation. In that approach, we started with a possibly informed prior distribution and examined the posterior distribution.

In this section we take a different approach. Some researchers prefer instead to pose the question in terms of model comparison. In this framing of the question, the focus is not on estimating the magnitude of the parameter. Instead, the focus is on deciding which of two hypothetical prior distributions is least incredible. One prior expresses the hypothesis that the parameter value is exactly the null value. The alternative prior expresses the hypothesis that the parameter could be any value, according to some form of broad distribution. In some formalizations, the alternative prior is a default uninformed distribution, chosen according to mathematical desiderata. This lack of being informed is often taken as a desirable aspect of the approach, not a defect, because the method might obviate disputes about prior knowledge. We will see, however, that the model-comparison method can be extremely sensitive to the choice of "uninformed" prior for the alternative hypothesis. The model comparison is not necessarily meaningful unless both hypotheses are viable in the first place.

Recall the model-comparison framework of Figure 10.1, p. 267, where the right panel shows a special case of model comparison in which the only difference between models is their prior distribution. It is this special case that is used to express the comparison of null and alternative hypotheses. The null hypothesis is expressed as a "spike" prior on the parameter of interest, such that only the null value has non-zero prior credibility. The alternative hypothesis is expressed as a broad prior, allowing a wide range of non-null values of the parameter. We previously saw an example similar to this in Section 10.5, p. 289, when we compared the must-be-fair model with the anything's-possible model. In that case, the must-be-fair model was almost a spike-shaped prior for the null hypothesis, and the anything's-possible model was a form of alternative hypothesis.

12.2.1. Is a coin fair or not?

For the null hypothesis, the prior distribution is a "spike" at the null value. The prior probability is zero for all values of θ other than the null value. The probability of the data for the null hypothesis is

$$p(z, N|M_{\text{null}}) = \theta_{\text{null}}^{z}(1-\theta_{\text{null}})^{(N-z)} \tag{12.1}$$

where M_{null} denotes the null model, that is, the null hypothesis. For the alternative hypothesis, we assume a broad beta distribution. Recall from Footnote 5 on p. 132 that for a single coin with a beta prior, the marginal likelihood is

$$p(z, N|M_{\text{alt}}) = B(z + a_{\text{alt}}, N - z + b_{\text{alt}})/B(a_{\text{alt}}, b_{\text{alt}}) \tag{12.2}$$

This equation was expressed as functions in R immediately after Equation 10.6 on p. 270. Combining Equations 12.2 and 12.1 we get the Bayes' factor for the alternative hypothesis relative to the null hypothesis:

$$\frac{p(z, N|M_{\text{alt}})}{p(z, N|M_{\text{null}})} = \frac{B(z + a_{\text{alt}}, N - z + b_{\text{alt}})/B(a_{\text{alt}}, b_{\text{alt}})}{\theta_{\text{null}}^{z}(1-\theta_{\text{null}})^{(N-z)}} \tag{12.3}$$

For a default alternative prior, the beta distribution is supposed to be uninformed, according to particular mathematical criteria. Intuition might suggest that a uniform distribution suits this requirement, that is, beta$(\theta|1, 1)$. Instead, some argue that the most appropriate uninformed beta distribution is beta$(\theta|\epsilon, \epsilon)$, where ϵ is a small number approaching zero (e.g., Lee & Webb, 2005; Zhu & Lu, 2004). This is called the *Haldane prior* (as was mentioned in Section 10.6). A Haldane prior is illustrated in the top-left plot of Figure 12.3, using $\epsilon = 0.01$.

Let's compute the value of the Bayes' factor in Equation 12.3, for the data used repeatedly in the previous chapter on NHST, namely $z = 7$ and $N = 24$. Here are the results for various values of a_{alt} and b_{alt}:

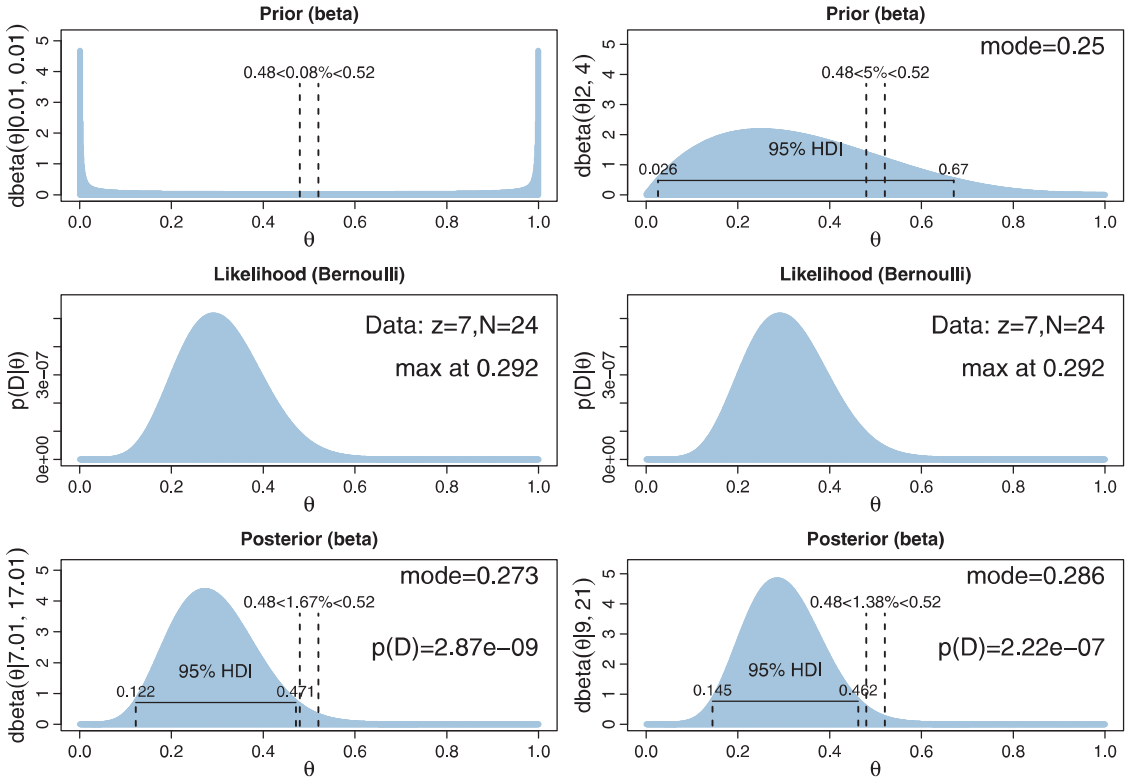

Figure 12.3 Left column: Haldane prior. Right column: Mildly informed prior. Vertical dashed lines mark a ROPE from 0.48 to 0.52. Annotation above the dashed lines indicates the percentage of the distribution within the ROPE.

$$\frac{p(z, N|M_{\text{alt}})}{p(z, N|M_{\text{null}})} = \begin{cases} 3.7227 & \text{for } a_{\text{alt}} = 2, b_{\text{alt}} = 4 \\ 1.9390 & \text{for } a_{\text{alt}} = b_{\text{alt}} = 1.000 \\ 0.4211 & \text{for } a_{\text{alt}} = b_{\text{alt}} = 0.100 \\ 0.0481 & \text{for } a_{\text{alt}} = b_{\text{alt}} = 0.010 \\ 0.0049 & \text{for } a_{\text{alt}} = b_{\text{alt}} = 0.001 \end{cases} \tag{12.4}$$

The first case, $a_{\text{alt}} = 2, b_{\text{alt}} = 4$, will be discussed later. For now, notice that when the alternative prior is uniform, with $a_{\text{alt}} = b_{\text{alt}} = 1.000$, the Bayes' factor shows a (small) preference for the alternative hypothesis, but when the alternative prior approximates the Haldane, the Bayes' factor shows a strong preference for the null hypothesis. As the alternative prior gets closer to the Haldane limit, the Bayes' factor changes by orders of magnitude. Thus, as we have seen before (e.g. Section 10.6, p. 292), the Bayes' factor is very sensitive to the choice of prior distribution.

You can see why this happens by considering Figure 12.3, which shows two priors in its two columns. The Haldane prior, in the left column, puts extremely small prior credibility on the parameter values that are most consistent with the data. Because

the marginal likelihood used in the Bayes' factor is the product of the prior and the likelihood, the marginal likelihood from the Haldane prior will be relatively small. The lower left plot of Figure 12.3 indicates that the marginal likelihood is $p(D) = 2.87 \times 10^{-9}$, which is small compared to the probability of the data from the null hypothesis, $p(D) = \theta_{\text{null}}^{z}(1-\theta_{\text{null}})^{(N-z)} = 5.96 \times 10^{-8}$. The right column of Figure 12.3 uses a mildly informed prior (discussed below) that puts modestly high probability on the parameter values that are most consistent with the data. Because of this, the marginal likelihood is relatively high, at $p(D) = 2.22 \times 10^{-7}$.

If we consider the posterior distribution instead of the Bayes' factor, we see that the posterior distribution on θ within the alternative model is only slightly affected by the prior. With $z = 7$ and $N = 24$, for the uniform $a_{\text{alt}} = b_{\text{alt}} = 1.00$, the 95% HDI is $[0.1407, 0.4828]$. For the approximate Haldane $a_{\text{alt}} = b_{\text{alt}} = 0.01$, the 95% HDI is $[0.1222, 0.4710]$, as shown in the lower-left plot of Figure 12.3. And for the mildly informed prior $a_{\text{alt}} = 2, b_{\text{alt}} = 4$, the 95% HDI is $[0.1449, 0.4624]$, as shown in the lower-right plot of Figure 12.3. (These HDI limits were accurately determined by function optimization using `HDIofICDF` in `DBDA2E-utilities.R`, not by MCMC.) The lower and upper limits vary by only about 2 percentage points. In all cases, the 95% HDI excludes the null value, although a wide ROPE might overlap the HDI. Thus, the explicit estimation of the bias parameter robustly indicates that the null value should be rejected, but perhaps only marginally. This contrasts with the Bayes' factor, model-comparison approach, which rejected the null or accepted the null depending on the alternative prior.

Of the Bayes' factors in Equation 12.4, which is most appropriate? If your analysis is driven by the urge for a default, uninformed alternative prior, then the prior that best approximates the Haldane is most appropriate. Following from that, we should strongly prefer the null hypothesis to the Haldane alternative. While this is mathematically correct, it is meaningless for an applied setting because the Haldane alternative represents nothing remotely resembling a credible alternative hypothesis. The Haldane prior sets prior probabilities of virtually zero at all values of θ except $\theta = 0$ and $\theta = 1$. There are very few applied settings where such a U-shaped prior represents a genuinely meaningful theory.

I recommended back in Section 10.6.1, p. 294, that the priors of the models should be equivalently informed. In the present application, the null-hypothesis prior is, by definition, fixed. But the alternative prior should be whatever best represents a meaningful and credible hypothesis, not a meaningless default. Suppose, for example, that we have some mild prior information that the coin is tail-biased. We express this as fictional prior data containing 1 head in 4 flips. With a uniform "proto-prior," that implies the alternative prior should be $\text{beta}(\theta|1+1, 3+1) = \text{beta}(\theta|2, 4)$, as shown in the upper-right plot of Figure 12.3. The Bayes' factor for this meaningful alternative prior is given as the first case in Equation 12.4, where is can be seen that the null is rejected. This agrees with the conclusion from explicit estimation of the parameter value in the

posterior distribution. Exercise 12.1 has you generate these examples for yourself. More extensive discussion, and an example from extrasensory perception, can be found in Kruschke (2011a).

12.2.1.1 Bayes' factor can accept null with poor precision

Figure 12.4 shows two examples in which the Bayes' factor favors the null hypothesis. In these cases, the data have a proportion of 50% heads, which is exactly consistent with the null value $\theta = 0.5$. The left column of Figure 12.4 uses a Haldane prior (with $\epsilon = 0.01$), and the data comprise only a single head in two flips. The Bayes' factor is 51.0 in favor of the null hypothesis! But should we really believe therefore that $\theta = 0.5$? No, I don't think so, because the posterior distribution on θ has a 95% HDI from 0.026 to 0.974.

The right column of Figure 12.4 uses a uniform prior. The data show 7 heads in 14 flips. The resulting Bayes' factor is 3.14 in favor of the null. But should we really believe therefore that $\theta = 0.5$? No, I don't think so, because the posterior distribution on θ has a 95% HDI from 0.266 to 0.734.

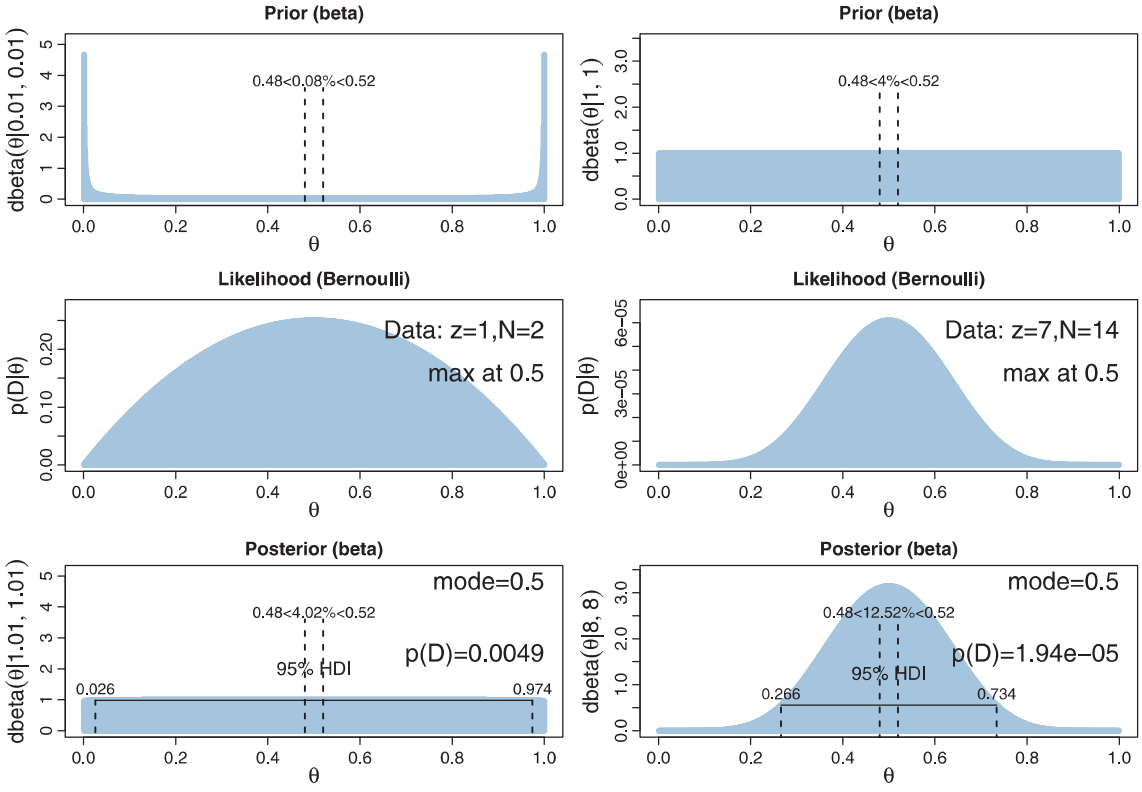

Figure 12.4 Bayes' factor (model comparison) approach can accept the null even with low precision on estimate. Left column: Haldane prior. Bayes' factor is 51.0 in favor of null, but 95% HDI extends from 0.026 to 0.974 (!). Right column: Uniform prior. Bayes' factor is 3.14 in favor of null, but the 95% HDI extends from 0.266 to 0.734 (!).

Thus, the Bayes' factor can favor the null hypothesis when the data are consistent with it, even for small quantities of data. For another example, involving continuous instead of dichotomous data, see Appendix D of Kruschke (2013a). The problem is that there is very little precision in the estimate of θ for small quantities of data. It seems inappropriate, even contradictory, to accept a particular value of θ when there is so much uncertainty about it.

When using the estimation approach instead of the model-comparison approach, accepting the null value demands that the HDI falls entirely inside a ROPE, which typically demands high precision. Narrower ROPEs require higher precision to accept the null value. For example, if we use the narrow ROPE from $\theta = 0.48$ to $\theta = 0.52$ as in Figure 12.4, we would need $N = 2400$ and $z = 1200$ for the 95% HDI to fall inside the ROPE. There is no way around this inconvenient statistical reality: high precision demands a large sample size (and a measurement device with minimal possible noise). But when we are trying to accept a specific value of θ, is seems logically appropriate that we should have a reasonably precise estimate indicating that specific value.

12.2.2. Are different groups equal or not?

In many research applications, data are collected from subjects in different conditions, groups, or categories. We saw an example with baseball batting abilities in which players were grouped by fielding position (e.g., Figure 9.14, p. 256), but there are many other situations. Experiments often have different subjects in different treatment conditions. Observational studies often measure subjects from different classifications, such as gender, location, etc. Researchers often want to ask the question, Are the groups different or not?

As a concrete example, suppose we conduct an experiment about the effect of background music on the ability to remember. As a simple test of memory, each person tries to memorize the same list of 20 words (such as "chair," "shark," "radio," etc.). They see each word for a specific time, and then, after a brief retention interval, recall as many words as they can. For simplicity, we assume that all words are equally memorable, and a person's ability to recall words is modeled as a Bernoulli distribution, with probability θ_{ij} for the ith person in the jth condition. The individual recall propensities θ_{ij} depends on condition-level parameters, ω_j and κ_j, that describe the overarching recall propensity in each condition, because $\theta_{ij} \sim \text{dbeta}(\omega_j(\kappa_j-2)+1\,,\ (1-\omega_j)(\kappa_j-2)+1\,)$.

The only difference between the conditions is the type of music being played during learning and recall. For the four groups, the music comes from, respectively, the death-metal band "Das Kruschke"[3], Mozart, Bach, and Beethoven. For the four conditions, the mean number of words recalled is 8.0, 10.0, 10.2, and 10.4.

[3] To find information regarding the death metal band Das Kruschke, search www.metal-archives.com. Appropriate to its genre, the band was short-lived. The author has no relation to the band, other than, presumably, some unknown common ancestor many generations in the past. The author was, however,

The most straight-forward way to find out whether the different types of music produce different memory abilities is to estimate the condition-level parameters and then examine the posterior differences of the parameter estimates. The histograms in the upper part of Figure 12.5 show the distributions of differences between the ω_j parameters. It can be seen that ω_1 is quite different than ω_3 and ω_4, and possibly ω_2. A difference of zero falls well outside the 95% HDI intervals, even taking into account MCMC instability and a small ROPE. From this we would conclude that Das Kruschke produces poorer memory than the classical composers.

A model-comparison approach addresses the issue a different way. It compares the full model, which has distinct ω_j parameters for the four conditions, against a restricted model, which has a shared ω_0 parameter to describe all the conditions simultaneously. The two models have equal (50/50) prior probabilities. The bottom panel of Figure 12.5 shows the results of a model comparison. The single ω_0 model is preferred over the distinct ω_j model, by about 85% to 15%. In other words, from the model comparison we might conclude that there is *no* difference in memory between the groups.

Which analysis should we believe? Is condition 1 different from some other conditions, as the parameter estimate implies, or are all the conditions the same, as the model comparison seems to imply? Consider carefully what the model comparison actually says: Given the choice between one shared mode and four different group modes, the one-mode model is less improbable. But that does not mean that the one-mode model is the best possible model. In fact, if a different model comparison is conducted, that compares the one-mode model against a different model that has one mode for group 1 and a second mode that is shared for groups 2 through 4, then the comparison favors the two-mode model. Exercise 12.2 has you carry out this alternative comparison.

In principle, we could consider all possible models formed by partitioning the four groups. For four groups, there are 15 distinct partitions. We could, in principle, put a prior probability on each of the 15 models, and then do a comparison of the 15 models (Gopalan & Berry, 1998). From the posterior probabilities of the models, we could ascertain which partition was most credible, and decide whether it is more credible than other nearly-as-credible partitions. (Other approaches have been described by D. A. Berry & Hochberg, 1999; Mueller, Parmigiani, & Rice, 2007; Scott & Berger, 2006). Suppose that we conducted such a large-scale model comparison, and found that the most credible model partitioned groups 2-4 together, separate from group 1. Does this mean that we should truly believe that there is zero difference between groups 2, 3, and 4? Not necessarily. If the group treatments are different, such as the four types of music in the present scenario, then there is almost certainly at least some small difference between their outcomes. (In fact, the simulated data do come from groups

---

in a garage band as a teenager. That band did not think it was playing death metal, although the music may have sounded that way to the critters fleeing the area.

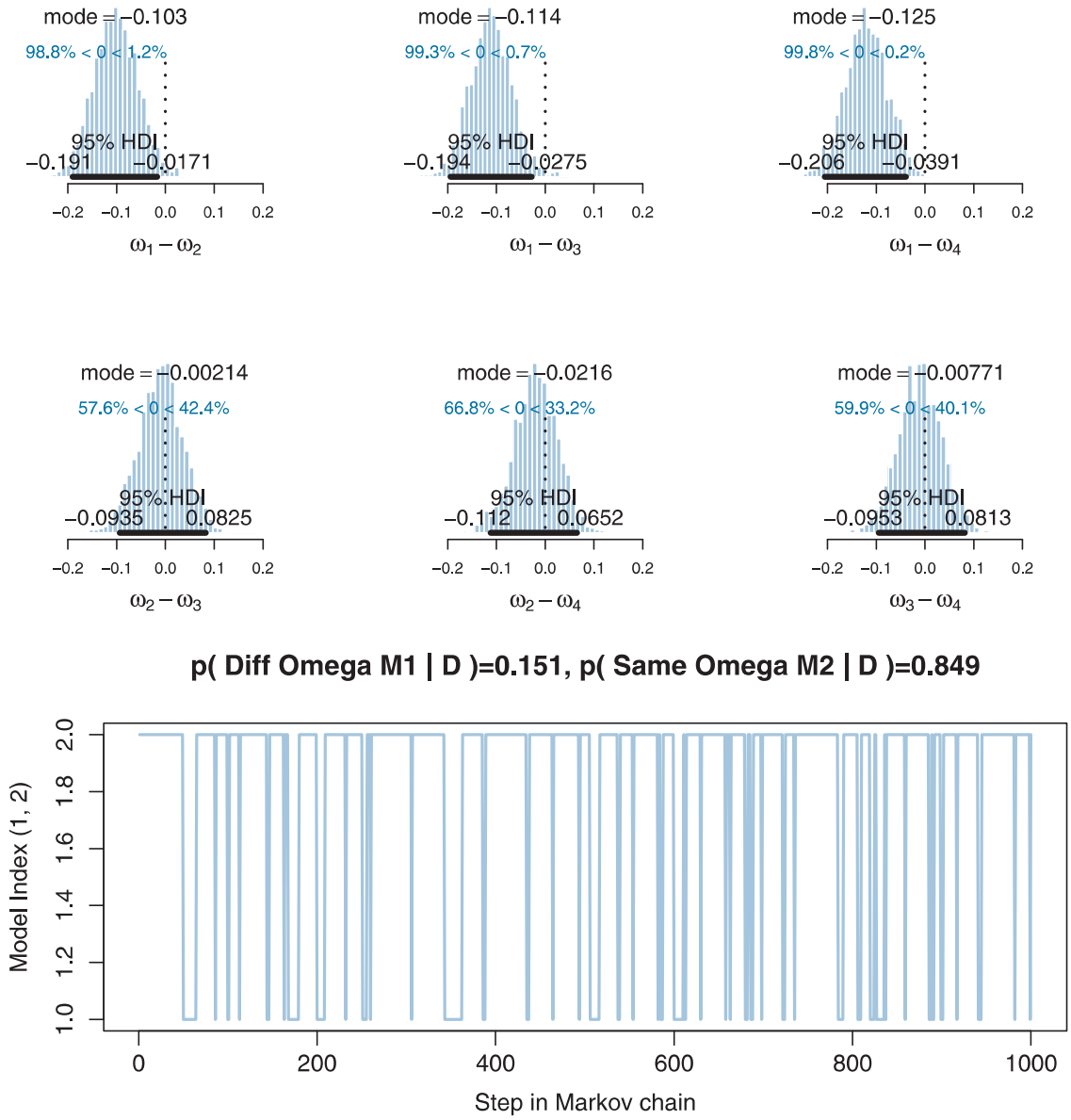

Figure 12.5 Top: Differences of posterior ω_j values for the four groups in the different-omega model. Notice that ω_1 *is credibly different from* ω_3 *and* ω_4, and possibly different from ω_2. The histograms are a bit choppy because the MCMC chain visits the different-omega model relatively rarely. Bottom: Trace plot of the model index shows that the model with a *single* omega parameter ("Same Omega M2") is *preferred* to a model with a separate omega parameter for each group ("Diff Omega M1").

with all different means.) We may still want to estimate the magnitude of those small differences, even if they are small. An explicit posterior estimate will reveal the magnitude and uncertainty of those estimates. Thus, unless we have a viable reason to believe that different group parameters may be literally identical, an estimation of distinct group parameters will tell us what we want to know, without model comparison.

12.2.2.1 Model specification in JAGS

Although it is somewhat tangential to the conceptual points of this section, here is the complete model specification for the results presented in Figure 12.5. These implementation details are relevant for Exercise 12.2 and they provide a review of pseudopriors that were introduced in Section 10.3.2.1.

The data structure has one row per subject, with the number of trials (words) for subject `s` denoted `nTrlOfSubj[s]`, the number correctly recalled for subject `s` denoted `nCorrOfSubj[s]`, and the condition of subject `s` denoted `CondOfSubj[s]`. The model specification begins with saying that each subject has an individual ability `theta[s]` from a condition-specific beta distribution:

```
model {
  for ( s in 1:nSubj ) {
    nCorrOfSubj[s] ~ dbin( theta[s] , nTrlOfSubj[s] )
    theta[s] ~ dbeta( aBeta[CondOfSubj[s]] , bBeta[CondOfSubj[s]] )
  }
```

The shape parameters of the beta distribution are then re-written in terms of the mode and concentration. Model 1 uses condition-specific `omega[j]`, while Model 2 uses the same `omega0` for all conditions. The JAGS function `equals(mdlIdx,...)` is used to select the appropriate model for index `mdlIdx`:

```
for ( j in 1:nCond ) {
  # Use omega[j] for model index 1, omega0 for model index 2:
  aBeta[j] <-       ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  )   * (kappa[j]-2)+1
  bBeta[j] <- ( 1 - ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  ) ) * (kappa[j]-2)+1
  omega[j] ~ dbeta( a[j,mdlIdx] , b[j,mdlIdx] )
}
omega0 ~ dbeta( a0[mdlIdx] , b0[mdlIdx] )
```

The priors on the concentration parameters are then specified:

```
for ( j in 1:nCond ) {
  kappa[j] <- kappaMinusTwo[j] + 2
  kappaMinusTwo[j] ~ dgamma( 2.618 , 0.0809 ) # mode 20 , sd 20
}
```

Notice that the groups have distinct concentration parameters under either model, even the single-mode model. This is merely a simplification for purposes of presentation. The concentration parameters could be structured differently across groups, analogous to the mode parameters.

Finally, the true and pseudoprior constants are set for the condition-level mode and concentration priors. (Pseudopriors were discussed in Section 10.3.2.1, p. 279.) The pseudoprior constants were selected to mimic the posterior reasonably well:

```
  # Constants for prior and pseudoprior:
  aP <- 1
  bP <- 1
  # a0[model] and b0[model]
  a0[1] <- 0.48*500       # pseudo
  b0[1] <- (1-0.48)*500   # pseudo
  a0[2] <- aP             # true
  b0[2] <- bP             # true
  # a[condition,model] and b[condition,model]
  a[1,1] <- aP            # true
  a[2,1] <- aP            # true
  a[3,1] <- aP            # true
  a[4,1] <- aP            # true
  b[1,1] <- bP            # true
  b[2,1] <- bP            # true
  b[3,1] <- bP            # true
  b[4,1] <- bP            # true
  a[1,2] <- 0.40*125      # pseudo
  a[2,2] <- 0.50*125      # pseudo
  a[3,2] <- 0.51*125      # pseudo
  a[4,2] <- 0.52*125      # pseudo
  b[1,2] <- (1-0.40)*125  # pseudo
  b[2,2] <- (1-0.50)*125  # pseudo
  b[3,2] <- (1-0.51)*125  # pseudo
  b[4,2] <- (1-0.52)*125  # pseudo
  # Prior on model index:
  mdlIdx ~ dcat( modelProb[] )
  modelProb[1] <- 0.5
  modelProb[2] <- 0.5
}
```

The pseudoprior constants used above are not uniquely correct. Other values might reduce autocorrelation even better.

12.3. RELATIONS OF PARAMETER ESTIMATION AND MODEL COMPARISON

We have now seen several examples of Bayesian approaches to assessing a null value, using a model-comparison approach or a parameter-estimation approach. In the model-comparison approach, a decision is made by putting a threshold on the Bayes' factor. In the parameter-estimation approach, a decision is made by putting a threshold on the parameter (involving the HDI and ROPE). In other words, both approaches involve decision rules applied to some aspect of a Bayesian posterior distribution.

One key relationship between the two approaches was emphasized in the introduction to model comparison, back in the hierarchical diagram of Figure 10.1, p. 267. Recall that model comparison is really a single hierarchical model in which the submodels

fall under a top-level indexical parameter. The model-comparison approach to null assessment focuses on the top-level model-index parameter. The parameter-estimation approach to null assessment focuses on the parameter distribution within the alternative model that has a meaningful prior. Thus, both approaches are logically coherent and can be applied simultaneously. Their conclusions about the null value do not have to agree, however, because they are assessing the null value in different ways, posing the question at different levels of the model. Neither level is the uniquely "correct" level, but one or the other level can be more or less meaningful in different applications.

A second relationship between the two approaches is that the Bayes' factor in the model comparison approach can be discerned in the estimation approach by noting how much the credibility of the null value increases or decreases in the parameter estimation. Consider, for example, the left side of Figure 12.3, p. 345, which shows parameter estimation using a Haldane prior. The top-left plot shows a fairly small ROPE around the null value, indicating that 0.08% of the prior distribution falls within the ROPE. The bottom-left plot shows that the posterior distribution has 1.67% of the distribution within the ROPE. *The ratio of proportions is (approximately) the Bayes' factor in favor of the null value.* The intuition for this fact is straightforward: The prior puts a certain probability on the null value, and Bayes' rule re-allocates probability, producing greater or lesser probability on the null value. If the null value gets more probability than it had in the prior, then the null hypothesis is favored, but if the null value gets less than it had in the prior, then the alternative is favored. In the present example case, the ratio of posterior probability in the ROPE to prior probability in the ROPE is $1.67/0.08 = 20.9$. Compare that with the analytically computed Bayes' factor from Equation 12.4: $1/0.0481 = 20.8$.

The right side of Figure 12.3 shows another example. The ratio of posterior probability in the ROPE to prior probability in the ROPE is $1.38/5.00 = 0.28$. Compare that with the analytically computed Bayes' factor from Equation 12.4: $1/3.7227 = 0.27$. Figure 12.4, p. 347, provides more examples. In its left side, the ratio of posterior to prior probabilities in the ROPE is $4.02/0.08 \approx 50$, which is nearly the same as the analytically computed Bayes' factor of 51. In the right side of Figure 12.4, the ratio of posterior to prior probabilities in the ROPE is $12.52/4.00 = 3.13$, which is nearly the same as the analytically computed Bayes' factor of 3.14.

Visualizing the Bayes' factor as the ratio of posterior to prior probabilities within a narrow ROPE helps us intuit the apparent contradictions between the conclusions of the model comparison and the parameter estimation. In the left side of Figure 12.3, p. 345, the proportion of the distribution inside the ROPE increases by a large ratio, even though most of the posterior distribution falls outside the ROPE. Similarly in the right side of Figure 12.4, p. 347, which again shows a large increase in the proportion of the distribution inside the ROPE, but with most of the distribution outside the ROPE. Thus, the model comparison focuses on the null value and whether its local probability increases from prior to posterior. The parameter estimation considers the entire posterior

distribution, including the uncertainty (i.e., HDI) of the parameter estimate relative to the ROPE.

The derivation of the Bayes' factor by considering the null value in parameter estimation is known as the Savage-Dickey method. A lucid explanation is provided by Wagenmakers, Lodewyckx, Kuriyal, and Grasman (2010), who also provide some historical references and applications to MCMC analysis of hierarchical models.

12.4. ESTIMATION OR MODEL COMPARISON?

As mentioned above, neither method for null value assessment (parameter estimation or model comparison) is uniquely "correct." The two approaches merely pose the question of the null value in different ways. In typical situations, I find the estimation approach to be more transparently meaningful and informative because parameter estimation provides an explicit posterior distribution on the parameter, while the Bayes' factor by itself does not provide that information. As I have emphasized, the two methods can be applied together in an integrated hierarchical model, but when their conclusions agree, the parameter estimation provides the information I typically want, and when their conclusions disagree, the parameter estimation still usually provides the more meaningful information.

The model-comparison approach to null-value assessment has two requirements for it to be meaningful. First, it must be theoretically meaningful for the parameter value to be exactly the null value. Second, the alternative-hypothesis prior must be meaningfully informed. These two requirements simply demand that both priors should be genuinely meaningful and viable, and they are merely an instance of the general requirement made back in Section 10.6.1 (p. 294) that the priors of both models should be meaningful and equally informed. Regarding the null-hypothesis prior, if it is not really plausible for two different treatment groups to be exactly the same, is it meaningful to give credibility to a null model that describes them as exactly the same? Maybe only as an approximate description, but perhaps not literally. Regarding the alternative-hypothesis prior, if it does not represent a plausible theory, is it meaningful to say it is more or less credible than any other prior? Much of the effort in pursuing the model-comparison approach to null-hypothesis testing goes into justifying an "automatic" prior for the alternative model that has desirable mathematical properties (e.g., in the psychological methods literature, Dienes, 2008, 2011; Edwards, Lindman, & Savage, 1963; Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007). But, in my opinion, a default prior is only useful to the extent that it happens to express a meaningful informed theory.

The estimation approach to null-value assessment uses a decision rule that involves the HDI and ROPE. Unfortunately, the ROPE has no automatic default limits, and the decision maker must make a case for a reasonable ROPE. In some applications, the ROPE can be specified with respect to the magnitude of conventionally "small" effect

sizes for the domain of research (Cohen, 1988). We have not yet had the need to discuss a technical definition of effect size, but the topic will arise in Chapter 16. But just as the labeling of effect sizes as "small" depends on conventional practice, the setting of a ROPE must be made in the context of current theory, measurement abilities, and the practical purpose of the decision. Recall that Serlin and Lapsley (1985, 1993) argued that research should affirm quantitative predictions, not merely reject null values, and affirmation of a theory is always relative to currently competing theories and the current state of measurement noise. Thus, the ROPE should be argued reasonably with the understanding that it may subsequently change.

12.5. EXERCISES

Look for more exercises at https://sites.google.com/site/doingbayesiandataanalysis/

Exercise 12.1. [Purpose: To make sure you understand the Bayes' factors regarding a single coin in Figure 12.3 and Equation 12.4, including the Savage-Dickey method.] Find the file BernBeta.R in the programs that accompany this book. Open RStudio with the folder of that file as R's working directory. Source the file so that R knows about the function `BernBeta`:

```
source("BernBeta.R")
```

Now, suppose we have a coin that is flipped 24 times and shows 7 heads. Enter these data into R, like this:

```
z=7 ; N=24
```

(A) According to the spike null hypothesis, for which the only credible value of θ is 0.5, what is the probability of the data? *Hint:* It is $\theta^z(1-\theta)^{(N-z)}$. Compute the value.

(B) Verify the result of the previous part by approximating a spike prior with a narrow beta distribution. Use the `BernBeta` function with a beta$(\theta|2000, 2000)$ prior, like this:

```
a=2000 ; b=2000
openGraph(width=5,height=7)
BernBeta( c(a,b) , c(rep(0,N-z),rep(1,z)) ,  ROPE=c(0.48,0.52) ,
          plotType="Bars" , showCentTend="Mode" , showHDI=TRUE , showpD=TRUE )
```

Include the resulting graph in your report. What is the value of $p(D)$ for this prior? Is it very close to the value computed for the exact spike prior in the previous part of this exercise? (It should be.) Explain why they are not exactly equal.

(C) Show the result when using a nearly Haldane prior, like this:

```
a=0.01 ; b=0.01
openGraph(width=5,height=7)
BernBeta( c(a,b) , c(rep(0,N-z),rep(1,z)) ,  ROPE=c(0.48,0.52) ,
          plotType="Bars" , showCentTend="Mode" , showHDI=TRUE ,showpD=TRUE )
```

Include the resulting graph in your report. What is the value of $p(D)$ for this prior? Compute and report the Bayes' factor of this prior relative to the spike (null) prior, using the formula $p(D|\text{Haldane})/p(D|\text{null})$.

(D) Continuing with the Haldane prior from the previous part, compute the approximate Bayes' factor using the Savage-Dickey method. That is, compute and report the ratio of percentage of prior within the ROPE over percentage of posterior with the ROPE.

(E) Suppose we have previous knowledge that in this application there tend to be more tails than heads. Show the result when using a mildly informed prior, like this:

```
a=2 ; b=4
openGraph(width=5,height=7)
BernBeta( c(a,b) , c(rep(0,N-z),rep(1,z)) ,  ROPE=c(0.48,0.52) ,
          plotType="Bars" , showCentTend="Mode" , showHDI=TRUE , showpD=TRUE )
```

Include the resulting graph in your report. What is the value of $p(D)$ for this prior? Compute and report the Bayes' factor of this prior relative to the spike (null) prior, using the formula $p(D|\text{informed})/p(D|\text{null})$.

(F) Continuing with the mildly informed prior from the previous part, compute the approximate Bayes' factor using the Savage-Dickey method. That is, compute and report the ratio of percentage of prior within the ROPE over percentage of posterior with the ROPE.

(G) Report the 95% HDIs when starting with the Haldane prior and the mildly informed prior. Are the HDIs very different? Were the Bayes' factors very different?

(H) Which approach, model comparison or estimation, seems most informative? Why? Within the model-comparison approach, which prior, uninformed Haldane or mildly informed, seems most meaningful? Why?

Exercise 12.2. [Purpose: Model comparison for different partitions of group modes, using the script of Section 12.2.2.1.] Open the script `OneOddGroupModelComp2E.R`, making sure that R's working directory includes the various utility programs used with this book.

(A) For this part of the exercise, the goal is to reproduce the findings presented in Section 12.2.2.1. First, be sure that the prior probabilities on the models are set to 50/50:

```
modelProb[1] <- 0.5
modelProb[2] <- 0.5
```

Run the script and report the results, including the graphs for the model index and the modes and differences of modes. State what the two models are, and state which

model is preferred and by how much. (*Hint:* Model 2 is the single-mode model, and it is preferred.)

(B) Continuing with the previous part, consider the graphs of differences of modes. What do they imply about differences between groups? Does this conclusion agree or disagree with the conclusion from the model comparison? How do you reconcile the conclusions? (*Hint:* The model index and the groups modes are all parameters being simultaneously estimated, so there is no contradiction. The parameters answer different questions; which questions?)

(C) For this part of the exercise, the goal is to compare the single-mode model against a different partitioning of the group modes. Instead of letting each group have its own distinct mode, we will allow a distinct mode for the first group, but restrict groups 2 through 4 to use a single mode. One way to accomplish this is to change this part of the model specification:

```
for ( j in 1:nCond ) {
  # Use omega[j] for model index 1, omega0 for model index 2:
  aBeta[j] <-       ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  )   * (kappa[j]-2)+1
  bBeta[j] <- ( 1 - ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  ) ) * (kappa[j]-2)+1
  omega[j] ~ dbeta( a[j,mdlIdx] , b[j,mdlIdx] )
}
```

to this:

```
for ( j in 1:nCond ) {
  # Use omega[j] for model index 1, omega0 for model index 2:
  aBeta[j] <-       ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  )   * (kappa[j]-2)+1
  bBeta[j] <- ( 1 - ( equals(mdlIdx,1)*omega[j]
                    + equals(mdlIdx,2)*omega0  ) ) * (kappa[j]-2)+1
}
for ( j in 1:2 ) {
  omega[j] ~ dbeta( a[j,mdlIdx] , b[j,mdlIdx] )
}
omega[3] <- omega[2]
omega[4] <- omega[2]
```

In your report, *carefully explain what the change does*. Make the change, and run the script (with the prior probabilities on the models set to 50/50). Report the results, including the graphs for the model index and the modes and differences of modes. State what the two models are, and state which model is preferred and by how much. (*Hint:* Model 2 is the single-mode model, and it is *not* preferred.)

(D) Continuing with the previous part, consider the graphs of differences of modes. What do they imply about differences between groups? Does this conclusion agree or

disagree with the conclusion from the model comparison? Even though Model 1 is preferred, is it really meaningful?

(E) Considering the results of the previous parts, what seems to be the most meaningful approach to analyzing differences of groups? (I'm hoping you'll say parameter estimation, not model comparison. But you might give arguments to the contrary.) Can you think of other applications in which model comparison might be more useful?