Bayesian Cognitive Modeling: A Practical Course

MICHAEL D. LEE AND ERIC-JAN WAGENMAKERS

ANSWERS AS OF August 11, 2013

Contents

Preface

This document contains answers to the exercises from the book 'Bayesian Cognitive Modeling: A Practical Course" (for updates see the book's website `www.bayesmodels.com`). Contrary to popular belief, statistical modeling is rarely a matter of right or wrong; instead, the overriding concerns are reasonableness and plausibility. Therefore, you may find yourself disagreeing with some of our intended solutions. Please let us know if you believe your answer is better or more informative than ours, and—if we agree with you—we will adjust the text accordingly.

Michael D. Lee
Irvine, USA
Eric-Jan Wagenmakers
Amsterdam, The Netherlands
August 2013

1 The basics of Bayesian analysis

1.1 General principles

Exercise 1.1.1 The famous Bayesian statistician Bruno de Finetti published two big volumes entitled *Theory of Probability* (de Finetti, 1974). Perhaps surprisingly, the first volume starts with the words "probability does not exist." To understand why de Finetti wrote this, consider the following situation: someone tosses a fair coin, and the outcome will be either heads or tails. What do you think the probability is that the coin lands heads? Now suppose you are a physicist with advanced measurement tools, and you can establish relatively precisely both the position of the coin and the tension in the muscles immediately before the coin is tossed in the air—does this change your probability? Now suppose you can briefly look into the future (Bem, 2011), albeit hazily. Is your probability still the same?

These different situations illustrate how the concept of probability is always conditional on background knowledge, and does not exist in a vacuum. This idea is central to the subjective Bayesian school, a school that stresses how inference is, in the end, dependent on personal beliefs.

Exercise 1.1.2 On his blog, prominent Bayesian Andrew Gelman wrote (March 18, 2010) "Some probabilities are more objective than others. The probability that the die sitting in front of me now will come up '6' if I roll it ... that's about 1/6. But not exactly, because it's not a perfectly symmetric die. The probability that I'll be stopped by exactly three traffic lights on the way to school tomorrow morning: that's well, I don't know exactly, but it is what it is." Was de Finetti wrong, and is there only one clearly defined probability of Andrew Gelman encountering three traffic lights on the way to school tomorrow morning?

A detailed knowledge of the layout of the traffic signs along the route will influence your assessment of this probability, as well as your knowledge of how busy traffic will be tomorrow morning, how often the traffic signs malfunction, whether traffic will be rerouted because of construction, and so on. When you can look one day into the future, the probability of Andrew Gelman encountering

three traffic lights on the way to school is either zero or one. As before, probability statements are conditional on your background knowledge. Both this exercise and the previous one get at exactly the same issue.

Exercise 1.1.3 Figure 1.1 shows that the 95% Bayesian credible interval for θ extends from 0.59 to 0.98. This means that one can be 95% confident that the true value of θ lies between 0.59 and 0.98. Suppose you did an orthodox analysis and found the same confidence interval. What is the orthodox interpretation of this interval?

The orthodox interpretation is that if you repeat the experiment very many times, and every time determine the confidence interval in the same way as you did for the observed data, then the true value of θ falls inside the computed intervals for 95% of the replicate experiments. Note that this says nothing about the confidence for the current θ, but instead refers to the long-run performance of the confidence interval method across many hypothetical experiments. According to the Wikipedia entry for confidence intervals (retrieved April 2012), "A confidence interval with a particular confidence level is intended to give the assurance that, if the statistical model is correct, then taken over all the data that *might* have been obtained, the procedure for constructing the interval would deliver a confidence interval that included the true value of the parameter the proportion of the time set by the confidence level. (...) A confidence interval does not predict that the true value of the parameter has a particular probability of being in the confidence interval given the data actually obtained." As an example, consider a uniform distribution with range 1. Suppose we draw two numbers from this distribution, and assess whether the mean μ falls in the interval from the lowest number y_l to the highest number y_h. When we repeat this procedure very many times, we find that this happens in 50% of the cases. Hence, when we draw two samples from a uniform distribution with range 1, (y_l, y_h) constitutes a 50% confidence interval for the mean μ. However, this interval does not condition on the data that were actually observed. For instance, assume that $y_l = 10.2$ and $y_h = 10.8$. Because the range is 1, we can be 100% confident that the mean lies within a 50% confidence interval. Additional anomalies are described in Berger and Wolpert (1988).

Exercise 1.1.4 Suppose you learn that the questions are all true or false questions. Does this knowledge affect your prior distribution? And, if so, how would this prior in turn affect your posterior distribution?

With true or false questions, zero ability corresponds to guessing, that is, $\theta = 0.5$. Because negative ability is implausible (unless the questions are deliberately misleading), it makes sense to have a uniform prior that ranges from

0.5 to 1, and hence has zero mass below 0.5. Because there is no prior mass below 0.5, there will also be no posterior mass below 0.5.

1.2 Prediction

Exercise 1.2.1 Instead of "integrating over the posterior," orthodox methods often use the "plug-in principle." In this case, the plug-in principle suggests that we predict $p(k^{\text{rep}})$ solely based on $\hat{\theta}$, the maximum likelihood estimate. Why is this generally a bad idea? Can you think of a specific situation in which this may not be so much of a problem?

The plug-in principle ignores uncertainty in θ, and therefore lead to predictions that are overconfident, that is, predictions that are less variable than they should be (Aitchison & Dunsmore, 1975). The overconfidence increases with the width of the posterior distribution. This also means that when the posterior is very peaked, that is, when we are very certain about θ (for instance because we have observed many data), the plug-in principle will only result in very little overconfidence.

1.3 Sequential updating

No exercises.

1.4 Markov chain Monte Carlo

Exercise 1.4.1 Use Google and list some other scientific disciplines that use Bayesian inference and MCMC sampling.

Bayesian inference is used in almost all scientific disciplines (but we wanted you to discover this yourself).

Exercise 1.4.2 The text reads: "Using MCMC sampling, posterior distributions can be approximated to any desired degree of accuracy." How is this possible?

By drawing more and more MCM samples, the discrepancy between the true distribution and the histogram can be made arbitrarily small. Or, in other words, longer chains yield better approximations.

3 Inferences with binomials

3.1 Inferring a rate

Exercise 3.1.1 Carefully consider the posterior distribution for θ given $k = 5$ successes out of $n = 10$ trials. Based on a visual impression, what is your estimate of the probability that the rate θ is higher than 0.4 but smaller than 0.6? How did you arrive at your estimate?

In the case of θ, the posterior is a *continuous* distribution, meaning that θ can take on any value in a certain interval; in this case, θ can take on any value from 0 to 1. Hence, the probability of $\theta \in (.4, .6)$ simply equals the area under the curve (i.e., the integral) from 0.4 to 0.6. Taking into account that the total area under the curve (from 0 to 1) equals by definition 1, you can guestimate that the probability of $\theta \in (.4, .6)$ equals about 50%.

Exercise 3.1.2 Consider again the posterior distribution for θ given $k = 5$ successes out of $n = 10$ trials. Based on a visual impression, what is your estimate of how much more likely the rate θ is to equal to 0.5 rather than 0.7? How did you arrive at your estimate?

Because θ is continuous, the probability that it equals any specific value x exactly is simply 0 (the integral from x to x is zero, see previous answer). However, the relative posterior plausibility of two specific values of θ is defined through the ratio of their posterior heights. In this case, the height at $\theta = 0.5$ is about 2.5, and the height at $\theta = 0.7$ is about 1. Hence, the value $\theta = 0.5$ is about $2.5/1 = 2.5$ times more likely than the value $\theta = 0.7$.

Exercise 3.1.3 Alter the data to $k = 50$ and $n = 100$, and compare the posterior for the rate θ to the original with $k = 5$ and $n = 10$.

When you have more information (i.e., high n) the posterior becomes more peaked. This means that you are more certain about what values are plausible, and what values are not.

Exercise 3.1.4 For both the $k = 50$, $n = 100$ and $k = 5$, $n = 10$ cases just considered, re-run the analyses with many more samples (e.g., 10 times as

many) by changing the `nsamples` variable in Matlab, or the `n.iter` variable in R. This will take some time, but there is an important point to understand. What controls the width of the posterior distribution (i.e., the expression of uncertainty in the rate parameter θ)? What controls the quality of the approximation of the posterior (i.e., the smoothness of the histograms in the figures)?

The width of the posterior distribution, expressing the uncertainty in the single true underlying rate, is controlled by the available information in the data. Thus, higher n leads to narrower posterior distributions. The quality of the estimate, visually evident by the smoothness of the posterior histogram, is controlled by how many samples are collected to form the approximation. Note that these two aspects of the analysis are completely independent. It is possible to have many data but just collect a few samples in a quick data analysis, to get a crude approximation to a narrow posterior. Similarly, it is possible to have only a few data, but collect many samples, to get a very close approximation to a very broad posterior.

Exercise 3.1.5 Alter the data to $k = 99$ and $n = 100$, and comment on the shape of the posterior for the rate θ.

The posterior distribution is not symmetric, because of the "edge effect" given by the theoretical upper bound of one for the rate. This goes some way to demonstrating how a Bayesian posterior distribution can take any form, and certainly does not have to be symmetric, or Gaussian, or any other simple form.

Exercise 3.1.6 Alter the data to $k = 0$ and $n = 1$, and comment on what this demonstrates about the Bayesian approach.

The fact that a posterior distribution exists at all shows that Bayesian analysis can be done even when there are very few data. The posterior distribution is very broad, reflecting the large uncertainty following from the lack of information, but nonetheless represents (as always) everything that is known and unknown about the parameter of interest.

3.2 Difference between two rates

Exercise 3.2.1 Compare the data sets $k_1 = 8$, $n_1 = 10$, $k_2 = 7$, $n_2 = 10$ and $k_1 = 80$, $n_1 = 100$, $k_2 = 70$, $n_2 = 100$. Before you run the code, try to predict the effect that adding more trials has on the posterior distribution for δ.

When you have more information (i.e., high n) the posteriors–for the individual

rates, as well as for the difference between them that is of interest—become more peaked. This means that you are more certain about what values for the difference are plausible, and what values are not.

Exercise 3.2.2 Try the data $k_1 = 0$, $n_1 = 1$ and $k_2 = 0$, $n_2 = 5$. Can you explain the shape of the posterior for δ?

The key to understanding the posterior is that you can be relatively sure that θ_2 is small, but you cannot be so sure about the value of θ_1. This means $\theta_1 - \theta_2$ could be a large positive value, because θ_1 could be large and θ_2 small. But $\theta_1 - \theta_2$ cannot be a large negative value, since θ_2 is small. The asymmetry in the uncertainty about θ_1 and θ_2 creates the asymmetry evident in the posterior for the difference.

Exercise 3.2.3 In what context might different possible summaries of the posterior distribution of δ (i.e., point estimates, or credible intervals) be reasonable, and when might it be important to show the full posterior distribution?

In general, point estimates (usually mean, median, or mode) and credible intervals are appropriate when they convey much the same information as would be gained from examining the whole posterior distribution. For example, if the posterior distribution is symmetric and with a small variance, its mean is a good summary of the entire distribution.

3.3 Inferring a common rate

Exercise 3.3.1 Try the data $k_1 = 14$, $n_1 = 20$, $k_2 = 16$, $n_2 = 20$. How could you report the inference about the common rate θ?

One reasonable reporting strategy here might be to use a measure for central tendency, such as a mean, median, or mode, together with a credible interval, for instance a 95% credible interval.

Exercise 3.3.2 Try the data $k_1 = 0$, $n_1 = 10$, $k_2 = 10$, $n_2 = 10$. What does this analysis infer the common rate θ to be? Do you believe the inference?

The analysis wants you to believe that the most plausible value for the common rate is around 0.5. This example highlights that the posterior distributions generated by a Bayesian analysis are conditional on the truth of the observed data, and of the model. If the model is wrong in an important way, the posteriors will be correct for that model, but probably not useful for the real problem. If a single rate really did underly $k_1 = 0$ and $k_2 = 10$ then the rate

must be near a half, since it is the most likely way to generate those data. But the basic assumption of a single rate seems problematic. The data suggest that a rate of 0.5 is one of the least plausible values. Perhaps the data are generated by two different rates, instead of one common rate.

Exercise 3.3.3 Compare the data sets $k_1 = 7$, $n_1 = 10$, $k_2 = 3$, $n_2 = 10$ and $k_1 = 5$, $n_1 = 10$, $k_2 = 5$, $n_2 = 10$. Make sure, following on from the previous question, that you understand why the comparison works the way it does.

The results for these data sets will be exactly the same. Because the model assumes a common rate, both data sets can in fact be re-described as having $k = k_1 + k_2 = 10$, $n = n_1 + n_2 = 20$.

3.4 Prior and posterior prediction

Exercise 3.4.1 Make sure you understand the prior, posterior, prior predictive, and posterior predictive distributions, and how they relate to each other (e.g., why is the top panel of Figure 3.9 a line plot, while the bottom panel is a bar graph?). Understanding these ideas is a key to understanding Bayesian analysis. Check your understanding by trying other data sets, varying both k and n.

Line plots are for continuous quantities (e.g., rate parameter θ) and bar plots are for discrete quantities (e.g., success counts of data).

Exercise 3.4.2 Try different priors on θ, by changing $\theta \sim \text{Beta}(1, 1)$ to $\theta \sim \text{Beta}(10, 10)$, $\theta \sim \text{Beta}(1, 5)$, and $\theta \sim \text{Beta}(0.1, 0.1)$. Use the figures produced to understand the assumptions these priors capture, and how they interact with the same data to produce posterior inferences and predictions.

One of the nice properties of using the $\theta \sim \text{Beta}(\alpha, \beta)$ prior distribution for a rate θ, is that it has a natural interpretation. The α and β values can be thought of as counts of "prior successes" and "prior failures", respectively. This means, using a $\theta \sim \text{Beta}(3, 1)$ prior corresponds to having the prior information that 4 previous observations have been made, and 3 of them were successes. Or, more elaborately, starting with a $\theta \sim \text{Beta}(3, 1)$ is the same as starting with a $\theta \sim \text{Beta}(1, 1)$, and then seeing data giving two more successes (i.e., the posterior distribution in the second scenario will be same as the prior distribution in the first). As always in Bayesian analysis, inference starts with prior information, and updates that information—by changing the probability distribution representing the uncertain information—as more information becomes available. When a type of likelihood function (in this case, the Binomial) does not change the type of distribution (in this case, the Beta) going from the posterior to

the prior, they are said to have a "conjugate" relationship. This is valued a lot in analytic approaches to Bayesian inference, because it makes for tractable calculations. It is not so important for that reason in computational approaches, because sampling methods can handle easily much more general relationships between parameter distributions and likelihood functions. But conjugacy is still useful in computational approaches because of the natural semantics it gives in setting prior distributions.

Exercise 3.4.3 Predictive distributions are not restricted to exactly the same experiment as the observed data, but for any experiment where the inferred model parameters make predictions. In the current simple binomial setting, for example, predictive distributions could be found by an experiment that is different because it has $n' \neq n$ observations. Change the graphical model, and Matlab or R code, to implement this more general case.

The script `Rate_4_answer.txt` implements the modified graphical model.

```
# Prior and Posterior Prediction
model{
   # Observed Data
   k ~ dbin(theta,n)
   # Prior on Rate Theta
   theta ~ dbeta(1,1)
   # Posterior Predictive
   postpredk ~ dbin(theta,npred)
   # Prior Predictive
   thetaprior ~ dbeta(1,1)
   priorpredk ~ dbin(thetaprior,npred)
}
```

Exercise 3.4.4 In October 2009, the Dutch newspaper *Trouw* reported on research conducted by H. Trompetter, a student from the Radboud University in the city of Nijmegen. For her undergraduate thesis, Trompetter had interviewed 121 older adults living in nursing homes. Out of these 121 older adults, 24 (about 20%) indicated that they had at some point been bullied by their fellow residents. Trompetter rejected the suggestion that her study may have been too small to draw reliable conclusions: "If I had talked to more people, the result would have changed by one or two percent at the most." Is Trompetter correct? Use the code `Rate_4.m` or `Rate_4.R`, by changing the `dataset` variable (Matlab) or changing the values for `k` and `n` (R), to find the prior and posterior predictive for the relevant rate parameter and bullying counts. Based on these distributions, do you agree with Trompetter's claims?

The 95% credible interval on the predicted number of bullied elderly (out of a total of 121) ranges from approximately (depending on sampling) 13 to 38. This means that the percentage varies from $13/121 \approx 10.7\%$ to $38/121 \approx 31.4\%$. This is about a 20% spread, considerably more than Trompetter estimated.

The key to understanding this exercise is that there are at least two sources of uncertainty; first, there is uncertainty about the true rate of bullying—observing 24 out of 121 does not allow one to identify the true rate with pinpoint precision. But even if this were the case, the second source of uncertainty comes into play, namely that of sampling variability. Just as a perfectly fair coin does not always produce 5 heads when you throw it 10 times, even perfect knowledge about the true rate of bullying necessarily leads to variable predictions. In order to make reasonable predictions you need to take into account both your lack of knowledge about the true rate of bullying (as given by the posterior distribution) *and* the uncertainty related to sampling variability (by simulating the sampling process based on draws from the posterior distribution). Of course, there are more sources of variability; most importantly, the present analysis ignores variability in nursing homes. The moral of this story is that are often more sources of uncertainty than you'd expect, it is important to take into account as many as you can.

3.5 Posterior prediction

Exercise 3.5.1 Why is the posterior distribution in the left panel inherently one-dimensional, but the posterior predictive distribution in the right panel inherently two-dimensional?

There is only one parameter, the rate θ, but there are two data, the success counts k_1 and k_2.

Exercise 3.5.2 What do you conclude about the descriptive adequacy of the model, based on the relationship between the observed data and the posterior predictive distribution?

The posterior predictive mass, shown by the squares, is very small for the actual outcome of the experiment, shown by the cross. The posterior prediction is concentrated on outcomes (around $k_1 = k_2 = 5$) that are very different from the data, and so the model does not seem descriptively adequate.

Exercise 3.5.3 What can you conclude about the parameter θ?

If the model is a good one, the posterior distribution for θ indicates that it is somewhere between about 0.2 and 0.8, and most likely around 0.5. *But*, it seems unlikely the model is a good one, and so it is not clear anything useful can be concluded about θ.

3.6 Joint distributions

Exercise 3.6.1 The basic moral of this example is that it is often worth thinking about joint posterior distributions over model parameters. In this case the marginal posterior distributions are probably misleading. Potentially even more misleading are common (and often perfectly appropriate) point estimates of the joint distribution. The cross in Figure 3.13 shows the expected value of the joint posterior, as estimated from the samples. Notice that it does not even lie in a region of the parameter space with any posterior mass. Does this make sense?

In general, it seems unhelpful to have a point summary that is not a plausible estimate of the true underlying value. One way to think about this result is in terms of the goal of the point estimate. The mean in this example is trying to minimize squared loss to the true value, and the possible values follow a curved surface, causing it to lie in the interior. Another way to think about the location of the mean is physically. It is the center of mass of the joint posterior (i.e., the place where you would put your finger to make the curved scatter plot balance). More mundanely, the expectation of the joint posterior is (by mathematical fact) the combination of the expectations for each parameter taken independently. Looking at the marginal posteriors, it is clear why the cross lies where it does.

Exercise 3.6.2 The circle in Figure 3.13 shows an approximation to the mode (i.e., the sample with maximum likelihood) from the joint posterior samples. Does this make sense?

This estimate seems to be more useful, at least in the sense that it falls on values that are plausible. In fact, it falls on the values with the highest density in the (estimated) posterior. Think of it as sitting on top of the hill surface traced out by the scatter plot. Nonetheless, it still seems unwise to try and summarize the complicated and informative curved pattern shown by the joint posterior scatterplot by a single set of values.

Exercise 3.6.3 Try the very slightly changed data $k = \{16, 18, 22, 25, 28\}$. How does this change the joint posterior, the marginal posteriors, the expectation, and the mode? If you were comfortable with the mode, are you still comfortable?

The minor change to the data hardly affects the mean, but greatly shifts the mode. This shows that the mode can be very sensitive to the exact information available, and is a non-robust summary in that sense. Metaphorically, the hill traced out by the joint density scatter plot has a "ridge" running along the top that is very flat, and the single highest point can move a long way if the

data are altered slightly.

Exercise 3.6.4 If you look at the sequence of samples in the trace plot, some autocorrelation is evident. The samples "sweep" through high and low values in a systematic way, showing the dependency of a sample on those immediately preceding. This is a deviation from the ideal situation in which posterior samples are independent draws from the joint posterior. Try thinning the sampling, taking only every 100th sample, by setting `nthin=100` in Matlab or `n.thin=100` in R. To make the computational time reasonable, reduce the number of samples collected after thinning to just 500 (i.e., run 50,000 total samples, so that 500 are retained after thinning). How is the sequence of samples visually different with thinning?

With thinning, the sequence of samples (i.e., the trace plots that WinBUGS shows) no longer shows the visual pattern of autocorrelation, as resembles more of a "block" than a "curve." One colorful description of the ideal visual appearance of samples is as a "fat hairy caterpillar." Thinning is needed in this example to achieve that type of visual appearance.

4 Inferences With Gaussians

4.1 Inferring means and standard deviations

Exercise 4.1.1 Try a few data sets, varying what you expect the mean and standard deviation to be, and how many data you observe.

As usual, posterior distributions become more peaked the more data you observe. The posterior distribution for μ should be located around the sample average. Highly variable numbers lead to a low precision λ, that is, a high standard deviation σ. Note that with many data points, you may estimate the standard deviation σ quite accurately (i.e., the posterior for σ can be very peaked). In fact, with an infinite number of data, the posterior distribution converges to a single point. This happens independently of whether the standard deviation σ is large or small; for instance, after observing a large sequence of highly variable data you can be relatively certain that the standard deviation is very high.

Exercise 4.1.2 Plot the *joint* posterior of μ and σ. That is, plot the samples from μ against those of σ. Interpret the shape of the joint posterior.

There is a tendency for the joint posterior to be U-shaped (as seen with μ on the x-axis). This is because extreme values of μ are only plausible when σ is high.

Exercise 4.1.3 Suppose you knew the standard deviation of the Gaussian was 1.0, but still wanted to infer the mean from data. This is a realistic question: For example, knowing the standard deviation might amount to knowing the noise associated with measuring some psychological trait using a test instrument. The x_i values could then be repeated measures for the same person, and their mean the trait value you are trying to infer. Modify the WinBUGS script and Matlab or R code to do this. What does the revised graphical model look like?

The script can be adjusted in several ways. The easiest is probably just to replace the statement `x[i] ~ dnorm(mu,lambda)` with `x[i] ~ dnorm(mu,1)`. In the graphical model. This change means that the node for σ is now shaded, because σ is no longer an unknown quantity that needs to be inferred.

Exercise 4.1.4 Suppose you knew the mean of the Gaussian was zero, but wanted to infer the standard deviation from data. This is also a realistic question: Suppose you know that the error associated with a measurement is unbiased, so its average or mean is zero, but you are unsure how much noise there is in the instrument. Inferring the standard deviation is then a sensible way to infer the noisiness of the instrument. Once again, modify the WinBUGS script and Matlab or R code to do this. Once again, what does the revised graphical model look like?

Again, the script can be adjusted in several ways. Again, the easiest is probably just to replace the statement `x[i]` $\sim$ `dnorm(mu,lambda)` with `x[i]` $\sim$ `dnorm(0,lambda)`. In the graphical model, this change means that the node for μ is now shaded, because μ is no longer an unknown quantity that needs to be estimated. Follow-up question: if you set μ to zero as suggested above, WinBUGS still provides a trace plot for μ. Why?

4.2 The seven scientists

Exercise 4.2.1 Draw posterior samples using the Matlab or R code, and reach conclusions about the value of the measured quantity, and about the accuracies of the seven scientists.

The posterior distributions for most standard deviations are very skewed. As a result, the posterior mean will be dominated by relatively low proportion of extreme values. For this reason, it is more informative to look at the posterior median. As expected, the first two scientists are pretty inept measurers and have high estimates of sigma. The third scientist does better than the first two, but also appears more inept than the remaining four.

Exercise 4.2.2 Change the graphical model in Figure 4.2 to use a uniform prior over the standard deviations, as was done in Figure 4.1. Experiment with the effect the upper bound of this uniform prior has on inference.

This exercise requires you to put a uniform distribution on sigma, so that the code needs to read (for an upper bound of 100): `sigma[i]`$\sim$ `dunif(0,100)`. Then `lambda[i] <-1/pow(sigma[i],2)`. Note that this change also requires that you change the Matlab or R code to assign initial values to sigma instead of lambda, because now sigma is assigned a prior and `lambda` is calculated deterministically from `sigma`, instead of the other way around.

When you make these changes you can see that the difference between the scientists is reduced. To get a more accurate idea of what is going on you

may want to set the number of MCMC samples to 100,000 (and, optionally, set a thinning factor to 10, so that only every tenth sample is recorded – this reduces the autocorrelation in the MCMC chains). As before, posterior median for the first scientist is largest, followed by that of numbers two and three.

4.3 Repeated measurement of IQ

Exercise 4.3.1 Use the posterior distribution for each person's μ_i to estimate their IQ. What can we say about the precision of the IQ test?

The posterior means for the `mu` parameters are very close to the sample means. The precision is $1/\sigma^2$, and because the posterior for `sigma` is concentrated around 6 the posterior precision is concentrated around $1/36 \approx 0.03$.

Exercise 4.3.2 Now, use a more realistic prior assumption for the μ_i means. Theoretically, IQ distributions should have a mean of 100, and a standard deviation of 15. This corresponds to having a prior of `mu[i] ~ dnorm(100,.0044)`, instead of `mu[i] ~ dunif(0,300)`, because $1/15^2 = 0.0044$. Make this change in the WinBUGS script, and re-run the inference. How do the estimates of IQ given by the means change? Why?

Parameter `mu[3]` is now estimated to be around 150, which is 5 points lower than the sample mean. Extremely high scores are tempered by the prior expectation. That is, an IQ of 150 is much more likely, according to the prior, than an IQ of 160. The same strong effect of the prior on inference is not evident for the other people, because their IQ scores have values over a range for which the prior is (relatively) flat.

Exercise 4.3.3 Repeat both of the above stages (i.e., using both priors on μ_i) with a new, but closely related, data set that has scores of $(94, 95, 96)$, $(109, 110, 111)$, and $(154, 155, 156)$. How do the different prior assumptions affect IQ estimation for these data. Why does it not follow the same pattern as the previous data?

The tempering effect of prior expectation has now disappeared, and even under realistic prior assumptions the posterior means for `mu` are close to the sample means. This happens because the data suggest that the test is very accurate, and accurate data are more robust against the prior. One helpful way to think about this is that the IQ test is now more informative (because it measures more accurately), and that extra information now overwhelms the prior. Notice

how this example shows that it is not necessarily more *data* that is needed to remove the influence of priors, but rather more *information*. Often, of course, the best way to get more information is to collect more data. But, another way is to develop data that are more precisely measured, or in some other way that is more informative.

5 Some examples of data analysis

5.1 Pearson correlation

Exercise 5.1.1 The second data set in the Matlab and R code is just the first data set from Figure 5.2 repeated twice. Set `dataset=2` to consider these repeated data, and interpret the differences in the posterior distributions for r.

With more data the posterior distribution for r becomes more peaked, showing that there is less uncertainty about the true correlation coefficient when more information is available.

Exercise 5.1.2 Do you find the priors on μ_1 and μ_2 to be reasonable?

We know that IQ is a positive number with mean 100 and standard deviation 15. We know that response times in semantic verification are faster than 0.2 seconds and slower than, say, 5 seconds. This knowledge is at odds with the priors on μ_1 and μ_2, priors that allow negative values and do not prefer specific ranges of positive values. Nevertheless, from a practical point of view, the priors for μ_1 and μ_2 are relatively flat across a large range that includes the plausible regions, and hence it is unlikely that more plausible priors affect the inference much.

Exercise 5.1.3 The current graphical model assumes that the values from the two variables—the $\boldsymbol{x}_i = (x_{i1}, x_{i2})$—are observed with perfect accuracy. When might this be a problematic assumption? How could the current approach be extended to make more realistic assumptions?

Very often in psychology, as with all empirical sciences, data are not measured with arbitrary precision. Other than nominal or ordinal variables (gender, color, occupation, and so on), most variables are measured imperfectly. Some, like response time, might be quite precise, consistent with measurement in the physical sciences. Others, like IQ, or personality traits, are often very imprecise. The current model makes the assumption that these sorts of measurements are perfectly precise. Since they are the basis for the correlation coefficient, the inference understates the uncertainty, and could lead to conclusions that are too confident, or otherwise inappropriate. The next section shows one approach to extending the model to address this problem.

5.2 Pearson correlation with uncertainty

Exercise 5.2.1 Compare the results obtained in Figure 5.4 with those obtained earlier using the same data, in Figure 5.2, for the model without any account of uncertainty in measurement.

The posterior distributions are (surprisingly, perhaps) quite similar.

Exercise 5.2.2 Generate results for the second data set, which changes $\sigma_2^e = 10$ for the IQ measurement. Compare these results with those obtained assuming $\sigma_2^e = 1$.

These results are very different. Allowing the large (but perhaps plausibly large, depending on the measurement instrument) uncertainty in the IQ data introduces large uncertainty into inference about the correlation coefficient. Larger values are more likely, but all possible values, including positive correlations, remain plausible. Note also that the expectation of this posterior is the same as in the case where the uncertainty of measurement is low or non-existent. This is a good example of the need to base inference on posterior distributions, rather than point estimates.

Exercise 5.2.3 The graphical model in Figure 5.3 assumes the uncertainty for each variable is known. How could this assumption be relaxed to the case where the uncertainty is unknown?

Statistically, it is straightforward to extend the graphical model, making the $\boldsymbol{\sigma}^e$ variables into parameters with prior distributions, and allowing them to be inferred from data. Whether the current data would be informative enough about the uncertainty of measurement to allow helpful inference is less clear. It might be that different sorts of data, like repeated measurements of the same people's IQs, are needed for this model to be effective. But it is straightforward to implement.

Exercise 5.2.4 The graphical model in Figure 5.3 assumes the uncertainty for each variable is the same for all observations. How could this assumption be relaxed to the case where, for example, extreme IQs are less accurately measured than IQs in the middle of the standard distribution?

The basic statistical idea would be to model the σ_{i2}^e variables, representing the ith persons error of measurement in their IQ score as a functions of μ_{i2}, representing their IQ itself. This would express a relationship between where people lie on the IQ scale, and how precisely their IQ can be measured. Whatever

relationship is chosen is itself a statistical model, formalizing assumptions about this relationship, and so can have parameters that are given priors and inferred from data.

5.3 The kappa coefficient of agreement

Exercise 5.3.1 *Influenza Clinical Trial.* Poehling, Griffin, and Dittus (2002) reported data evaluating a rapid bedside test for influenza using a sample of 233 children hospitalized with fever or respiratory symptoms. Of the 18 children known to have influenza, the surrogate method identified 14 and missed 4. Of the 215 children known not to have influenza, the surrogate method correctly rejected 210 but falsely identified 5. These data correspond to $a = 14$, $b = 4$, $c = 5$, and $d = 210$. Examine the posterior distributions of the interesting variables, and reach a scientific conclusion. That is, pretend you are a consultant for the clinical trial. What would your two- or three-sentence "take home message" conclusion be to your customers?

The surrogate method does a better job detecting the absence of influenza than it does detecting the presence of influenza. The 95% Bayesian confidence interval for kappa is $(.51, .84)$, suggesting that the test is useful.

Exercise 5.3.2 *Hearing Loss Assessment Trial.* Grant (1974) reported data from a screening of a pre-school population intended to assess the adequacy of a school nurse assessment of hearing loss in relation to expert assessment. Of those children assessed as having hearing loss by the expert, 20 were correctly identified by the nurse and 7 were missed. Of those assessed as not having hearing loss by the expert, 417 were correctly diagnosed by the nurse but 103 were incorrectly diagnosed as having hearing loss. These data correspond to $a = 20$, $b = 7$, $c = 103$, $d = 417$. Once again, examine the posterior distributions of the interesting variables, and reach a scientific conclusion. Once again, what would your two- or three-sentence "take home message" conclusion be to your customers?

Compared to the expert, the nurse displays a bias to classify children as having hearing loss. In addition, the nurse misses 7 out of 27 children with hearing loss. The nurse is doing a poor job, and this is reflected in the 95% credible interval for kappa of (approximately, up to sampling) $(.12, .29)$.

Exercise 5.3.3 *Rare Disease.* Suppose you are testing a cheap instrument for detecting a rare medical condition. After 170 patients have been screened, the test results show that 157 did not have the condition, but 13 did. The expensive ground truth assessment subsequently revealed that, in fact, none

of the patients had the condition. These data correspond to $a = 0$, $b = 0$, $c = 13$, $d = 157$. Apply the kappa graphical model to these data, and reach a conclusion about the usefulness of the cheap instrument. What is special about this data set, and what does it demonstrate about the Bayesian approach?

The posterior mean for `kappa` is approximately .05, with a 95% credible interval of approximately $(0, .24)$. The data are noteworthy because the disease has never been observed, so there are two zero cells, and a zero column sum. This poses a challenge for frequentist estimators. In order to deal with the problem of zero counts a frequentist may add a "1" to each cell in the design, but this amounts to fabricating data. An attractive property of the Bayesian approach is that it is always possible to do the analysis.

5.4 Change detection in time series data

Exercise 5.4.1 Draw the posterior distributions for the change point, the means, and the common standard deviation.

When you look at the trace plots, you may see that it takes a few samples for the chains to lose their dependence on the initial value that was used as a starting point. These initial values are non-representative outliers, and they also stretch out the y-axis of the trace plots. In the call to bugs, set burn-in to 10 and observe the change.

With respect to the posterior distributions, it is worthwhile to note that the key parameter `tau` is estimated relatively precisely around 732. One of the reasons for this is that μ_1 and μ_2 are relatively easy to tell apart.

Exercise 5.4.2 Figure 5.7 shows the mean of the posterior distribution for the change point (this is the point in time where the two horizontal lines meet). Can you think of a situation in which such a plotting procedure can be misleading?

One case in which this procedure may be misleading is when the posterior distribution is relatively wide (i.e., not peaked around its mean); in such a situation, there is a lot of uncertainty about the location of the change-point, and the plotting procedure, based on a point estimate, falsely suggests that the location is determined precisely.

Exercise 5.4.3 Imagine that you apply this model to a data set that has two change points instead of one. What could happen?

In this case the model is seriously misspecified. The model assumes that there are two regimes, but in reality there are three. One thing that could happen is that the model groups together the two adjacent regimes that are most similar to each other and treats them as one. The problems that result should be visible from mismatches between the posterior predictive distribution and the data.

5.5 Censored data

Exercise 5.5.1 Do you think Cha Sa-soon could have passed the test by just guessing?

It is unlikely that Cha Sa-soon was just guessing. First, the posterior distribution for `theta` is relatively peaked around .40, whereas chance performance in a four-choice task is only 0.25. Second, the probability of scoring 30 or more correct answers when guessing equals .00000016 (in R: `1-pbinom(29,50,.25)`). With this success probability, the number of attempts to pass the exam follows a geometric distribution. Therefore we know that when guessing, the average number of attempts equals $1/.00000016 \approx 6,097,561$, considerably more than Cha Sa-soon required. The probability of guessing and "only" needing 950 attempts is a relatively low .00016 (in R: `pgeom(950, prob=.00000016)`).
In contrast, with a `theta` of .4 the the probability of scoring 30 or more correct answers equals .0034 (in R: `1-pbinom(29,50,.40)`). With this probability, the associated expected number of attempts until success is 294, and the probability of passing the exam and "only" needing 950 attempts is a relatively high 0.96.

Exercise 5.5.2 What happens when you increase the interval in which you know the data are located, from 15–25 to something else?

Increasing the interval increases the posterior uncertainty for `theta`.

Exercise 5.5.3 What happens when you decrease the number of failed attempts?

When the number of failed attempts becomes low (say 20), the posterior for `theta` becomes wider and shifts to values that are somewhat higher.

Exercise 5.5.4 What happens when you increase Cha Sa-soon's final score from 30?

Not that much! Apparently, the extra information about Cha Sa-soon's final score is much less informative than the knowledge that she had failed 949

times (and with scores ranging from 15 to 25).

Exercise 5.5.5 Do you think the assumption that all of the scores follow a binomial distribution with a single rate of success is a good model for these data?

It seems to be a poor model. The chance of the same underlying rate generating all of the censored scores below 25, and then producing the 30, can be calculated according to the model, and is tiny. Alternative models would assume some sort of change in the underlying rate. This could psychologically correspond to learning, for at least some of the problems in the test, at some point in the sequence of 950 attempts.

5.6 Recapturing planes

Exercise 5.6.1 Try changing the number of planes seen again in the second sample from $k = 4$ to $k = 0$. What inference do you draw about the population size now?

The inferred population size is now much larger. This makes sense, because none of the same planes were seen again on the second viewing, suggesting a bigger fleet. The single most likely answer if 50 planes, at the upper limit of the possibilities allowed by prior information, and the expected value of the posterior is around 40 planes.

Exercise 5.6.2 How much impact does the upper bound $t^{\max} = 50$ have on the final conclusions when $k = 4$ and when $k = 0$? Develop your answer by trying both the $k = 4$ and $k = 0$ cases with $t^{\max} = 100$.

When $k = 4$ planes are seen again on the second viewing, the upper bound makes little difference. The posterior distributions are similar, and their expected value is around 17 or 18 planes for both $t^{\max} = 50$ and $t^{\max} = 100$. When $k = 0$, however, the posterior distribution in general, and any summary inference likely to be drawn about the number of planes from this distribution, depends heavily on the upper bound. When $t^{\max} = 50$, the mean of the posterior is about 40 planes, and the mode is 50 planes. When $t^{\max} = 100$, the mean of the posterior is about 71 planes, and the mode is somewhere near 100 planes. In this situation, it is important we are sure about the upper bound is accurate information, since it has a strong influence on the inference that will be drawn.

Exercise 5.6.3 Suppose, having obtained the posterior mass in Figure 5.11, the same fleet of planes was subjected to a new sighting at a different airport at a later day. What would be an appropriate prior for t?

The posterior distribution from the first analysis represents what we do and do not know about the number of planes based on the original prior, and the first data set. It is therefore the appropriate prior distribution for the second analysis. This is a good example of "today's posterior is tomorrow's prior", and the logical cumulative way Bayesian analysis incorporates additional information into an analysis as it becomes available.

6 Latent mixture models

6.1 Exam scores

Exercise 6.1.1 Draw some conclusions about the problem from the posterior distribution. Who belongs to what group, and how confident are you?

Inspection of the `z[i]` nodes confirms that the first five people are confidently assigned to the guessing group, and the remaining ten people are confidently assigned to the knowledge group. The high confidence is clear from the fact that the posteriors for `z[i]` are approximately located either at 0 or 1.

Exercise 6.1.2 The initial allocations of people to the two groups in this code is random, and so will be different every time you run it. Check that this does not affect the final results from sampling.

This is easily done by running the code again a few times.

Exercise 6.1.3 Include an extra person in the exam, with a score of 28 out of 40. What does his posterior for z tell you? Now add four extra people, all with the score 28 out of 40. Explain the change these extra people make to the inference.

The performance of the new participant is completely ambiguous. The posterior mean for z is approximately 0.5, indicating that this participant is as likely to belong to the guessing group as they are to belong to the knowledge group. When four extra people are added, all with the score of 28 out of 40, things change: now all five people with 28 out of 40 are classified to be in the knowledge group with a probability of about .82. This happens because five people with a score of 28 out of 40 drive down the estimated rate of answering correctly. When the questions are estimated to be difficult, performance that is only a little better than chance will already land you in the knowledge group; in contrast, when the questions are estimated to be easy, performance that is only a little better than chance will put you firmly in the guessing group.

Exercise 6.1.4 What happens if you change the prior on the success rate of the second group to be uniform over the whole range from 0 to 1, and so allow

for worse-than-guessing performance?

Nothing much. While the original prior assumption makes more sense, since it captures information we have about the problem, the data are sufficiently informative that inference is not significantly affected by this change.

Exercise 6.1.5 What happens if you change the initial expectation that everybody is equally likely to belong to either group, and have an expectation that people generally are not guessing, with (say), $z_i \sim \text{Bernoulli}(0.9)$?

This expectation does not change the classification of the first 15 participants much, because these participants are unambiguous in terms of their performance. However, the new participant with a score of 28 is inferred to be in the knowledge group with probability 0.9, whereas this was 0.5 before. Because the data for this participant are ambiguous it is the prior expectation that largely determines how this participant is classified.

6.2 Exam scores with individual differences

Exercise 6.2.1 Compare the results of the hierarchical model with the original model that did not allow for individual differences.

For the first 15 participants the results are essentially unchanged. The new participant with a score of 28 is now inferred to be in the knowledge group with probability 0.8, compared to the original 0.5. This happens because the new participant is more likely to be a low-knowledge member of the knowledge group than a member of the guessing group. The fact that the current model allows for individual differences helps it account for the relatively low score of 28.

Exercise 6.2.2 Interpret the posterior distribution of the variable `predphi`. How does this distribution relate to the posterior distribution for `mu`?

The variable `predphi` is based on a draw from a Gaussian distribution with mean `mu` and standard deviation `sigma`. This predictive distribution indicates what we can expect about the success rate of a new, as yet unobserved participant from the knowledge group. If there were no individual differences, `sigma` would be zero and draws for `predphi` are effectively draws from the posterior of `mu`. But because there are individual differences, this adds uncertainty to what we can expect for a new participant and hence the posterior distribution for `predphi` is wider than that of `mu`.

Exercise 6.2.3 In what sense could the latent assignment of people to groups in this case study be considered a form of model selection?

There are two rival explanations, specified as statistical models, for the data. These are the guessing or knowledge-based responding accounts. When a participant is assigned to the guessing group this means that, for that particular participant, we believe the guessing model is a better explanation for the observed data than is the knowledge-based model.

6.3 Twenty questions

Exercise 6.3.1 Draw some conclusions about how well the various people listened, and about the difficulties of the various questions. Do the marginal posterior distributions you are basing your inference on seem intuitively reasonable?

This question is best answered by tallying the total number of correct responses separately for each participant and for each item. The result show that, first, participants who answer most items correctly have the highest estimated values for `p`, and, second, items answered correctly most often have the highest estimated values for `q`. These results are consistent with intuition.

Exercise 6.3.2 Now suppose that three of the answers were not recorded, for whatever reason. Our new data set, with missing data, now takes the form shown in Table 6.2. Bayesian inference will automatically make predictions about these missing values (i.e., "fill in the blanks") by using the same probabilistic model that generated the observed data. Missing data are entered as `nan` ("not a number") in Matlab, and `NA` ("not available") in R or WinBUGS. Including the variable `k` as one to monitor when sampling will then provide posterior values for the missing values. That is, it provides information about the relative likelihood of the missing values being each of the possible alternatives, using the statistical model and the available data. Look through the Matlab or R code to see how all of this is implemented in the second data set. Run the code, and interpret the posterior distributions for the three missing values. Are they reasonable inferences?

The estimates are reasonable. One of the nice things about Bayesian inference is that, given that the model is appropriate, the estimates are *always* reasonable. Sometimes the reasonableness may be hidden from your intuition, but this just means that your intuition was faulty. Consider person 1 on item M. We know that person 1 is relatively attentive, because he answers relatively many questions correctly; we also know that item M is relatively easy, because many other people answer item M correctly. This item-person combination looks like it

could have resulted in a correct answer. The inferred probability for M-1 being correct is approximately 0.74. For item-person combination E-8 the reverse holds. The item is difficult and the participant is inattentive. Consequently, the inferred probability for E-8 being correct is approximately 0.01. Finally, combination R-10 is middle-of-the-road on both dimensions, and the inferred probability for it being correct is 0.41.

These inferred probabilities are directly related to the knowledge about each participant and item. In fact, if you multiply the estimated p's and q's you can recover the inferred probabilities. For instance, `p[1]` is approximately 0.88, and `q[13]` (the M) is approximately 0.84. The multiplication of these probabilities yields .74. The same calculations may be performed for the other missing item-participant calculations.

Exercise 6.3.3 The definition of the accuracy for a person on a question in terms of the product $\theta_{ij} = p_i q_j$ is very simple to understand, but other models of the interaction between person ability and question difficulty are used in psychometric models. For example, the Rasch model (e.g., Andrich, 1988) uses $\theta_{ij} = \exp(p_i - q_j) / (1 + \exp(p_i - q_j))$. Change the graphical model to implement the Rasch model.

The script `TwentyQuestions_Rasch.txt` implements the modified graphical model.

```
# Twenty Questions Using Rasch Model
model{
  # Correctness Of Each Answer Is Bernoulli Trial
  for (i in 1:np){
    for (j in 1:nq){
       k[i,j] ~ dbern(theta[i,j])
    }
  }
  # Probability Correct Is Product Of Question By Person Rates
  for (i in 1:np){
    for (j in 1:nq){
      theta[i,j] <- exp(p[i]-q[j])/(1+exp(p[i]-q[j]))
    }
  }
  # Priors For People and Questions
  for (i in 1:np){
    p[i] ~ dbeta(1,1)
  }
  for (j in 1:nq){
    q[j] ~ dbeta(1,1)
  }
}
```

6.4 The two country quiz

Exercise 6.4.1 Interpret the posterior distributions for `x[i]`, `z[j]`, `alpha`, and `beta`. Do the formal inferences agree with your original intuitions?

Yes, people 1, 2, 5, and 6 form one group, and people 3, 4, 7, and 8 form the other group. The model also groups together questions A, D, E, and H versus questions B, C, F, and G. And the rates of correct decisions for matched and mismatched groups make sense, too.

Exercise 6.4.2 The priors on the probabilities of answering correctly capture knowledge about what it means to match and mismatch, by imposing an order constraint $\alpha \geq \beta$. Change the code so that this information is not included, by using priors `alpha~dbeta(1,1)` and `beta~dbeta(1,1)`. Run a few chains against the same data, until you get an inappropriate, and perhaps counter-intuitive, result. The problem that is being encountered is known as model indeterminacy or label switching. Describe the problem, and discuss why it comes about.

The result you get from the analysis with uniform priors can change from chain to chain, switching the inferences about `alpha` and `beta`. This is a basic and common problem for mixture models, known as *model indeterminacy* or label switching. The probability α is used whenever $x_i = z_j$. If this corresponds to a Thai person answering a Thai question, then α should be high, as we expect. But there is nothing stopping the model, without the order constraint, from coding Thai people as $x_i = 1$ and Moldovan questions as $z_j = 1$, in which case α will be low. Effectively, with this coding, α and β will swap roles. Overall, there are four possibilities (two ways people can be encoded, by two ways questions can be encoded). Our semantics of α being knowledge-based and β being ignorance-based will apply for 2 of these 4 possible encodings, but will be reversed for the other two. The core problem is that α and β are statistically defined the same way in the revised model. This is the indeterminacy. A practical but inelegant way to solve this problem is by being flexible in interpretation. A better way is, as per the original model and code, by defining the statistical model itself more carefully, introducing the order constraint, and removing the indeterminacy.

Exercise 6.4.3 Now suppose that three extra people enter the room late, and begin to take the quiz. One of them (Late Person 1) has answered the first four questions, the next (Late Person 2) has only answered the first question, and the final new person (Late Person 3) is still sharpening their pencil, and has not started the quiz. This situation can be represented as an updated

data set, now with missing data, as in Table 6.4. Interpret the inferences the model makes about the nationality of the late people, and whether or not they will get the unfinished questions correct.

Late person 1 is confidently placed in the same category as people 1, 2, 5, and 6. This is also reflected in the probabilities of answering the remaining four questions correctly: 0.88, 0.07, 0.05, 0.90, predicting a "1 0 0 1" pattern that was also observed for people 1, 2, 5, and 6.

Late person 2 only answered a single question, but this information suffices to assign this person with probability .89 to the same category as persons 3, 4, 7, and 8. This is reflected in the probabilities of answering the remaining seven questions correctly: .80, .80, .15, .15, .80, .80, .13 predicting a "1 1 0 0 1 1 0" pattern that was relatively typical for people 3, 4, 7, and 8.

Late person 3 did not answer a single question and is equally likely to belong to either group. Because each group has an opposite pattern of answering any particular question correctly, the model predicts that the performance of late person 3 will be around chance (slightly worse than chance because not all questions are answered correctly even if the question matches the nationality).

Exercise 6.4.4 Finally, suppose that you are now given the correctness scores for a set of 10 new people, whose data were not previously available, but who form part of the same group of people we are studying. The updated data set is shown in Table 6.5. Interpret the inferences the model makes about the nationality of the new people. Revisit the inferences about the late people, and whether or not they will get the unfinished questions correct. Does the inference drawn by the model for the third late person match your intuition? There is a problem here. How could it be fixed?

The new people are all classified in the same group as people 1, 2, 5, and 6. However, late person 3 is still equally likely to be classified in either group. This is a problem in the sense that the model is insensitive to changes in baseline proportions: if we know that there are many more people in one category than another, this knowledge should affect our prediction for late person 3. In this case, late person 3 is likely to belong to the same category as the new persons. The model can be extended to deal with baseline proportions by changing the line `x[i] ~ dbern(0.5)`, which assumes equal baselines, to `x[i] ~ dbern(phi)`, and `phi ~ dbeta(1,1)`, which now estimates the baseline.

6.5 Assessment of malingering

Exercise 6.5.1 What are your conclusions about group membership? Did all of the participants follow the instructions?

The expectation of the posterior for the indicator variable z shows everybody to have (essentially) the value 0 or 1. This means that the classification into the bona fide and malingering groups is virtually a certainty, with 0 corresponding to bona fide and 1 to malingering. The first 10 people, as expected, are bona fide. The rest, with two exceptions, are inferred to be malingerers. The two exceptions are the 14th and 15th people, who scored 44. It seems likely these people did not follow the instruction to malinger.

6.6 Individual differences in malingering

Exercise 6.6.1 Is the inferred rate of malingering consistent with what is known about the instructions given to participants?

The mean of the posterior for ϕ is a little below a half, consistent with 10 bona fide participants out of 22.

Exercise 6.6.2 Assume you know that the base rate of malingering is 10%. Change the WinBUGS script to reflect this knowledge. Do you expect any differences?

The change can be made by changing the definition of the indicator variables to `z[i]~dbern(0.1)`, This base rate is different from the one that is inferred for ϕ, which has posterior expectation of about 0.5, so it is reasonable to expect inference to be affected. Specifically, when the base rate of malingering is very low we should be less confident about the classification of participants who performed poorly.

Exercise 6.6.3 Assume you know for certain that participants 1, 2, and 3 are bona fide. Change the code to reflect this knowledge.

This change can be made by defining `z[1] <- 0`, `z[2] <- 0` and `z[3] <- 0`. It is important, once this change is made, to be sure that initial values are not set for these three parameters, since they are no longer stochastic variables to be inferred. This is not always straightforward, in terms of the data structures being passed to and from WinBUGS from Matlab or R. An effective solution is to define `z[4] <- ztmp[1]` to `z[22] <- ztmp[19]`, and set the initial values

directly on the complete `ztmp` vector.

Exercise 6.6.4 Suppose you add a new participant. What number of questions answered correctly by this participant would lead to the greatest uncertainty about their group membership?

A little trial-and-error experimentation finds that an extra score of 41 questions correct leads to a posterior expectation of about 0.45. This is more uncertain than the neighboring possibilities of 40 questions correct, with posterior expectation about 0.8, and 42 questions correct, with posterior expectation about 0.15.

Exercise 6.6.5 Try to solve the label switching problem by using the `dunif(0,mubon)` approach instead of the logit transform.

TBA

Exercise 6.6.6 Why are the priors for λ_{b} and λ_{m} different?

Conceptually, it might be reasonable to expect less variability for the bona fide group. Doing the task correctly seems more constraining than the freedom to simulate amnesia and malinger. Computationally, the lack of variability in the bona fide scores can cause under- and over-flow issues in sampling, and limiting the priors to reasonable and computational ranges is a practical approach to address this issue.

6.7 Alzheimer's recall test cheating

Exercise 6.7.1 Suppose the utilities are very different, so that a false-alarm costs \$100, because of the risk of litigation in a false accusation, but misses are relatively harmless, costing \$10 in wasted administrative costs. What decisions should be made about bona fide and cheating now?

If it is 10 times more important not to make false-alarms (i.e., \$100 vs \$10), then you should only treat someone as a cheater if that is 10 times more likely. This means the expectation of z_i should be less than 0.1 before the decision is made to classify them as having cheated. It is clear from the figure that his applies only to the few people who scored 39 or 40 on the test.

Exercise 6.7.2 What other potential information, besides the uncertainty about classification, does the model provide? Give at least one concrete example.

The model provides information about the base rate of cheating, as well

as information about the levels of performance, and variability in that performance, for the different groups of people. By providing a complete model of the data-generating process, Bayesian inference is able to provide a much more complete analysis of the data than a simple set of classifications.

7 Bayesian model comparison

7.1 Marginal likelihood

Exercise 7.1.1 Suppose you construct a second model for the same data D. This model, $\mathcal{M}_y$, has a parameter ζ that can take on two values, ζ_1 and ζ_2. You assign prior probability mass $p(\zeta_1) = 0.3$ and $p(\zeta_2) = 0.7$. For these two values, the likelihoods are 0.002 and 0.003, respectively. Compute the marginal likelihood for $\mathcal{M}_y$.

The marginal likelihood for for $\mathcal{M}_y$ is

$$\begin{aligned} p(D \mid \mathcal{M}_y) &= p(\zeta_1)p(D \mid \zeta_1) + p(\zeta_2)p(D \mid \zeta_2) \\ &= .3 \times .002 + .7 \times .003 \\ &= .0027. \end{aligned}$$

Exercise 7.1.2 What is the relative support of the data D for $\mathcal{M}_y$ versus $\mathcal{M}_x$?

The evidence for $\mathcal{M}_y$ is $p(D \mid \mathcal{M}_y) = .0027$, whereas the evidence for $\mathcal{M}_x$ is $p(D \mid \mathcal{M}_x) = .0015$. Hence, the evidence for $\mathcal{M}_y$ is almost twice as strong as it is for $\mathcal{M}_x$, that is, the evidence ratio is $.0027/.0015 = 1.8$. This ratio is called the Bayes factor, and it is the topic of the next section.

Exercise 7.1.3 Suppose you construct a third model for the same data D. This model, $\mathcal{M}_z$, has a parameter μ that can take on 5 values, $\mu_1 = \mu_2 = \ldots = \mu_5$ with equal prior probability. The likelihoods are $p(D \mid \mu_1) = 0.001$, $p(D \mid \mu_2) = 0.001$, $p(D \mid \mu_3) = 0.001$, $p(D \mid \mu_4) = 0.001$, and $p(D \mid \mu_5) = 0.006$. Note that the likelihood for μ_5 is twice as high as the best possible likelihood for $\mathcal{M}_y$ and $\mathcal{M}_x$. Calculate the marginal likelihood for $\mathcal{M}_z$. Do you prefer it over $\mathcal{M}_y$ and $\mathcal{M}_x$? What is the lesson here?

The marginal likelihood for for $\mathcal{M}_z$ is

$$\begin{aligned} p(D \mid \mathcal{M}_z) &= p(\mu_1)p(D \mid \mu_1) + p(\mu_2)p(D \mid \mu_2) + p(\mu_3)p(D \mid \mu_3) \\ &\quad + p(\mu_4)p(D \mid \mu_4) + p(\mu_5)p(D \mid \mu_5) \\ &= .2 \times .001 + .2 \times .001 + .2 \times .001 + .2 \times .001 + .2 \times .006 \\ &= .002. \end{aligned}$$

The evidence for $\mathcal{M}_z$ is higher than that for $\mathcal{M}_x$, but slightly lower than that for $\mathcal{M}_y$. Even though the best parameter value in model $\mathcal{M}_z$ yields a much better goodness-of-fit than any of the parameter values in model $\mathcal{M}_y$, model $\mathcal{M}_z$ suffers because it also harbors parameter values that produce a relatively poor fit to the data. Again, the moral of the story is that complexity (i.e., the size of the parameter space or the number of different predictions a model can make) comes at a cost.

Exercise 7.1.4 Consider Bart and Lisa, who each get 100 euros to bet on the winner of the world cup soccer tournament. Bart decides to divide his money evenly over 10 candidate teams, including those from Brazil and Germany. Lisa divides her money over just two teams, betting 60 euros on the team from Brazil and 40 euros on the team from Germany. Now if either Brazil or Germany turn out to win the 2010 world cup, Lisa wins more money than Bart. Explain in what way this scenario is analogous to the computation of marginal likelihood.

By betting all her money on just two teams, Lisa was willing to take a risk, whereas Bart was just trying to keep his options open. For Bart, this means that even if his prediction of Brazil winning turns out to be correct, he will still lose the 90 euros he betted on the other countries to win (i.e., Bart is punished for making many false predictions). The point of the story is that, both at the betting office and in Bayesian inference, hedging your bets is not necessarily the best option, because this requires you to spread your resources – be it money or prior probability mass – thinly over the alternative options.

Exercise 7.1.5 Holmes and Watson are involved in a rather simple game of darts, in which, with each dart, the player tries to score as many points as possible. The maximum score per dart is 60, and the minimum score is 0 (when the dart lands outside the board). After 5 darts, Holmes scored {38, 10, 0, 0, 0} and Watson scored {20,20,20,18,16}. How do you determine who is the better player? Consider another game of darts, but now one of the players gets to throw 50 times instead of 5. Explain how this scenario shows the importance of averaging instead of maximizing in order to penalize complexity.

It appears that Holmes just got lucky with the one dart, and Watson is the far better player. Most people agree that the best way to determine the

superior darts player in the above scenario is to average (or sum) the points from all five darts that were thrown. Think of Holmes and Watson as two competing models, and think of each dart as a prediction about the observed data. The procedure of maximum likelihood suggest Holmes is the better player (his maximum score is higher than that of Watson), whereas the procedure of marginal likelihood suggests that Watson is the better player. When one of the players gets more darts, this is analogous to a model making more predictions and being more complex. This means that with more darts, a poor player is more likely to get lucky and score a high number of points with one of the darts. Hence, maximum likelihood does not correct for model complexity. Instead, by simply averaging the scores, marginal likelihood allows us to compare the abilities of two players fairly, even if one player throws more darts than the other: with more darts, the poor player will be more likely to occasionally score a high number, but this advantage is entirely undone by the large number of darts that will hit the wall, fly out of the window, or injure the cat.

7.2 The Bayes factor

Exercise 7.2.1 Suppose you entertain a set of three models, x, y, and z. Assume you know $BF_{xy} = 4$ and $BF_{xz} = 3$. What is BF_{zy}?

$BF_{zy} = BF_{zx} \times BF_{xy} = 1/BF_{xz} \times BF_{xy} = 4/3$. Bayes factors are transitive.

Exercise 7.2.2 Suppose the $BF_{ab} = 1,000,000$, such that the data are one million times more likely to have occurred under $\mathcal{M}_a$ than under $\mathcal{M}_b$. Give two arguments for why you may still believe that $\mathcal{M}_a$ provides an inadequate or incorrect account of the data.

Argument 1: $\mathcal{M}_a$ may be strongly supported by the data (vis-a-vis $\mathcal{M}_b$), but it could be a wildly implausible model on a priori grounds. Extraordinary claims require extraordinary evidence (see the box in the next section).

Argument 2: The Bayes factor is a measure of relative evidence; $\mathcal{M}_a$ may provide a very poor account of the data, but it may still outperform $\mathcal{M}_b$, whose account of the data may be truly terrible.

7.3 Posterior model probabilities

Exercise 7.3.1 In this book, you have now encountered two qualitatively different kinds of priors. Briefly describe what they are.

One prior is a single number that quantifies the a priori plausibility of a given model; the other prior is a distribution that describes the uncertainty about a given model's parameters. And finally there is the prior predictive distribution, the distribution of simulated data that a model generates from its prior parameter distribution.

Exercise 7.3.2 Consider one model, $\mathcal{M}_1$, that predicts a post-surgery survival rate by gender, age, weight, and history of smoking. A second model, $\mathcal{M}_2$, includes two additional predictors, namely body-mass index and fitness. We compute posterior model probabilities and find that $p(\mathcal{M}_1 \mid D) = .6$ and consequently $p(\mathcal{M}_2 \mid D) = .4$. For a patient Bob, $\mathcal{M}_1$ predicts a survival rate of 90%, and $\mathcal{M}_2$ predicts a survival rate of 80%. What is your prediction for Bob's probability of survival?

It is perhaps tempting to base your prediction solely on $\mathcal{M}_1$, which is after all the preferred model. However, this would ignore the uncertainty inherent in the model selection procedure, and it would ignore the very real possibility that the best model is $\mathcal{M}_2$, according to which the survival rate is 10% lower than it is for $\mathcal{M}_1$. The Bayesian solution is to weight the two competing predictions with their associated posterior model probabilities, fully taking into account the uncertainty in the model selection procedure. This procedure is called Bayesian model averaging (Madigan & Raftery, 1994). In our example, the model-averaged prediction for survival rate would be $.6 \times 90\% + .4 \times 80\% = 86\%$.

7.4 Advantages of the Bayesian approach

There are no exercises for this section.

7.5 Challenges for the Bayesian approach

There are no exercises for this section.

7.6 The Savage–Dickey method

Exercise 7.6.1 The Bayes factor is relatively sensitive to the width of the prior distributions for the model parameters. Use Equation 7.9 to argue why this is the case.

The denominator of the Savage-Dickey equation features the height of the prior for ϕ at $\phi = \phi_0$. This means that the choice of prior can greatly influence the Bayes factor. The choice of prior will also influence the shape of the posterior, of course, but this influence quickly diminishes as the data accumulate. Consider again a test for a Gaussian mean μ, with $\mathcal{H}_0 : \mu = 0$ and $\mathcal{H}_1 : \mu \neq 0$. Suppose the prior for μ is a uniform distribution that ranges from $-a$ to a, and suppose that the number of observations is reasonably large. In this situation, the data will have overwhelmed the prior, so that the posterior for μ is relatively robust against changes in a. In contrast, the height of the prior at $\mu = 0$ varies directly with a: if a is doubled, the height of the prior at $\mu = 0$ becomes twice as small, and according to the Savage-Dickey equation this would about double the Bayes factor in favor of H_0. In the limit, as a grows very large, the height of the prior at $\mu = 0$ goes to zero, which means that the Bayes factor will go to infinity, indicating decisive support for the null hypothesis.

Exercise 7.6.2 The Bayes factor is is relatively sensitive to the width of the prior distributions, but only for the parameters that differ between the models under consideration. Use Equation 7.9 to argue why this is the case.

In contrast to the prior for the parameter of interest ϕ, the Savage-Dickey equation indicates that the prior for the nuisance parameters ψ is not critical. Hence, priors on the nuisance parameters can be vague or even "improper". Intuitively, the prior vagueness of nuisance parameters is present in both models and cancels out in the computation of the Bayes factor.

Exercise 7.6.3 What is the main advantage of the Savage–Dickey procedure?

The main advantages are computational efficiency, ease of use, and conceptual transparancy. The Savage-Dickey equation shows that in nested models, under plausible assumptions on the prior structure for the nuisance parameters, computation of the Bayes factor is relatively straightforward. All that is needed is an estimate of posterior and prior ordinates under the alternative hypothesis $\mathcal{H}_1$. This computational shortcut is often much less involved than the more general solution that involves integrating out nuisance parameters ψ for $\mathcal{H}_0$, and

parameters ψ and ϕ for $\mathcal{H}_1$, as follows:

$$BF_{01} = \frac{p(D \mid \mathcal{H}_0)}{p(D \mid \mathcal{H}_1)} = \frac{\int p(D \mid \phi = \phi_0, \psi) p(\phi = \phi_0, \psi) \mathrm{d}\psi}{\int p(D \mid \psi, \phi) p(\psi, \phi) \mathrm{d}\psi \mathrm{d}\phi}. \tag{7.1}$$

7.7 Disclaimer and summary

Exercise 7.7.1 Browse the empirical literature of your subfield of study. Do you find the null hypotheses plausible? That is, could they ever be exactly true?

Cognitive scientists may browse the content of any issue of *Psychological Science* in the period from 2009–2012 and find several point null hypotheses that are perfectly plausible (even though they were rejected by conventional p-value hypothesis testing).

8 Comparing Gaussian means

8.1 One-sample comparison

Exercise 8.1.1 Here we assumed a half-Cauchy prior distribution on the standard deviation `sigma`. Other choices are possible and reasonable. Can you think of a few?

One common choice is to assign precision $\lambda = 1/\sigma^2$ a Gamma distribution, e.g., `lambda` $\sim$ `dgamma(.001,.001)`. Another common choice is to assign `sigma` a distribution that is uniform across a wide range. Yet another choice is based on background knowledge or prior quantitative insight about the problem at hand, perhaps obtained through the results from earlier, similar studies.

Exercise 8.1.2 Do you think the different priors on `sigma` will lead to substantially different conclusions? Why or why not? Convince yourself by implementing a different prior and studying the result.

The parameter `sigma` is common to all models. Changing the prior on `sigma` may affect model complexity, but it will do so for all models to the same extent and hence cancel out in the Bayes factor. The same conclusion follows from the Savage-Dickey density ratio test: all that is important are the prior and posterior distributions for the parameter under test, that is, the parameter that is present in one model and absent in the other.

Exercise 8.1.3 We also assumed a Cauchy prior distribution on effect size `delta`. Other choices are possible and reasonable. One such choice is the standard Gaussian distribution. Do you think this prior will lead to substantially different conclusions? Why or why not? Convince yourself by implementing the standard Gaussian prior and studying the result.

This change will affect the Bayes factor, although not by much. Compared to the Cauchy, the standard Normal distribution has more mass near 0. Because near zero is where the data lie, the evidence in favor of $\mathcal{H}_0$ is somewhat less pronounced with the standard Normal prior than with the Cauchy prior. Specifically, for the two-sided test involving $\mathcal{H}_1$, the analytical result from

Jeff Rouder's website yields $BF_{01} = 4.75$ with the standard Normal prior and $BF_{01} = 6.06$ with the Cauchy prior.

8.2 Order-restricted one-sample comparison

Exercise 8.2.1 For completeness, estimate the Bayes factor for the summer and winter data between $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_3 : \text{Cauchy}\big(0, 1\big)_{\mathcal{I}(0,\infty,)}$, involving the order-restricted alternative hypothesis that assumes the effect is positive.

TBA.

Exercise 8.2.2 In this example, it matters whether the alternative hypothesis is unrestricted, order-restricted to negative values for δ, or order-restricted to positive values for δ. Why is this perfectly reasonable? Can you think of a situation where the three versions of the alternative hypothesis yield exactly the same Bayes factor?

The results are reasonable because, in this case, $\mathcal{H}_2 : \delta < 0$ appears inconsistent with the data; after all, Dr. Smith observed a positive effect. Hence, the order-restricted $\mathcal{H}_2$ competes with $\mathcal{H}_0 : \delta = 0$ much less successfully than did $\mathcal{H}_1 : \delta \neq 0$. Similar reasoning explains why $\mathcal{H}_3 : \delta > 0$ fares better than $\mathcal{H}_1$. There is only one situation where all three versions of the alternative hypothesis yield exactly the same Bayes factor: the posterior distribution for δ needs to be perfectly symmetric about 0.

Exercise 8.2.3 From a practical standpoint, we do not need a new graphical model and WinBUGS script to compute the Bayes factor for $\mathcal{H}_0$ versus the order-restricted $\mathcal{H}_2$. Instead, we can use the original graphical model in Figure 8.1 that implements the unrestricted Cauchy distribution and discard those prior and posterior MCMC samples that are inconsistent with the $\delta < 0$ order-restriction. The Savage–Dicky density ratio test still involves the height of the prior and posterior distributions at $\delta = 0$, but now the samples from these distributions are truncated, respecting the order-restriction, such that they range only from $\delta = -\infty$ to $\delta = 0$. Implement this method in Matlab or R, and check that the same conclusions are drawn from the analysis.

TBA.

Exercise 8.2.4 Wagenmakers and Morey (2013) describe yet another method to obtain the Bayes factor for order-restricted model comparisons. This method is perhaps the most reliable because it avoids the numerical complications associated with having to estimate the posterior density at a boundary. Go to

`http://www.ejwagenmakers.com/papers.html`, download the Wagenmakers and Morey paper, and read the introduction with a focus on Equation 1. Implement their suggested method and compare the results to those obtained earlier.

TBA.

8.3 Two-sample comparison

Exercise 8.3.1 The two-sample comparison of means outlined above assumes that the two groups have equal variance. How can you extend the model when this assumption is not reasonable?

This is the Behrens-Fisher problem. See Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009, p. 757), available at `http://www.ejwagenmakers.com/2009/WetzelsEtAl2009Ttest.pdf`. In a nutshell, the idea is to define two `sigma` parameters, one per group, and then use a single, *pooled* `sigma` parameter in the computation of $\alpha \leftarrow \delta * \sigma_{pooled}$.

9 Comparing binomial rates

9.1 Equality of proportions

Exercise 9.1.1 Because the rate parameters θ_1 and θ_2 both have a uniform prior distribution, the prior distribution for the difference parameter δ can be found analytically as a triangular distribution. What are the advantages of using this result, rather than relying on computational sampling? What are the disadvantages?

TBA.

Exercise 9.1.2 In the current analysis, we put independent priors on θ_1 and θ_2. Do you think this is plausible? How would you change the model to take into account the possible dependence? How would this affect the outcome of the Bayesian test?

TBA.

Exercise 9.1.3 This example corresponds to a rare case in which the Bayes factor is available analytically. $BF_{01} = p\left(D \mid \mathcal{H}_0\right) / p\left(D \mid \mathcal{H}_1\right)$ is given by

$$BF_{01} = \frac{\binom{n_1}{s_1}\binom{n_2}{s_2}}{\binom{n_1+n_2}{s_1+s_2}} \frac{(n_1+1)(n_2+1)}{n_1+n_2+1}.$$

Calculate the Bayes factor analytically, and compare it to the result obtained using the Savage–Dickey method.

TBA.

Exercise 9.1.4 For the pledger data, a frequentist test for equality of proportions indicates that $p \approx 0.006$. This tells us that when $\mathcal{H}_0$ is true (i.e., the proportions of condom users are equal in the two groups), then the probability is about 0.006 that we would encounter a result at least as extreme as the one that was in fact observed. What conclusions would you draw based on this information? Discuss the usefulness of the Bayes factor and the p-value in answering the scientific question of whether pledgers are less likely than

non-pledgers to use a condom.

TBA.

9.2 Order-restricted equality of proportions

Exercise 9.2.1 Consider an order-restricted test of $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_3 : \delta > 0$. What do you think the result will be? Check your intuition by implementing the appropriate graphical model, and estimating the Bayes factor.

TBA.

9.3 Comparing within-subject proportions

Exercise 9.3.1 The Zeelenberg data can also be analyzed using the Bayesian t-test discussed in Chapter 8. Think of a few reasons why this might not be such a good idea. Then, despite your reservations, apply the Bayesian t-test anyway. How do the results differ? Why?

TBA.

9.4 Comparing between-subject proportions

Exercise 9.4.1 A between-subjects frequentist t-test on the proportion of correctly sorted cards does not allow one to reject the null hypothesis, $t\,(40.2) = 0.37$, $p = 0.72$. In what way does the Bayesian approach improve upon the frequentist inference?

TBA.

Exercise 9.4.2 In what way is the model of the data in Figure 9.10 superior to the statistical model assumed by the t-test?

This t-test does not quantify the evidence in favor of the null hypothesis and ignores the fact that trials are nested in participants. This design calls for a hierarchical or multi-level analysis.

9.5 Order-restricted between-subjects comparison

Exercise 9.5.1 For the order-restricted comparison of $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_2 : \delta > 0$, what is the maximum support in favor of $\mathcal{H}_0$ that could possibly be obtained, given the present number of subjects, and given that the average rate of correct card sorts is 65%?

TBA.

Exercise 9.5.2 What is the maximum support for the earlier unrestricted test of $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_1 : \delta \neq 0$?

TBA.

10 Memory retention

10.1 No individual differences

Exercise 10.1.1 Why is the posterior predictive distribution for all four subjects the same? Are there any (real or fabricated) data that could make the model predict different patterns of retention for different subjects? What if there were massive qualitative differences, such as one subject remembering everything, and the other two remembering nothing?

It is a basic assumption of the model that there are no individual differences. While the model parameters are inferred from multiple subjects, those subjects are assumed to be using exactly one parameterization of the memory retention model. Thus, there are no data for which the posterior parameter inferences and the posterior predictive distributions will not be identical for all subjects.

10.2 Full individual differences

Exercise 10.2.1 What are the relative strengths and weaknesses of the full individual differences model and the no individual differences model? Think about this, because the hierarchical approach we consider next could be argued to combine the best features of both of these approaches.

The no individual differences model is able to use all of the available subjects' data to make inferences and, by assumption, can make predictions about new subjects (because they are assumed to be the same.) The full individual differences model allows for any sort of variability between subjects, but requires that subject to provide data to make inferences and predictions about them, because each subject is a completely new and surprising instantiation of the memory retention model.

10.3 Structured individual differences

Exercise 10.3.1 Think of a psychological model and data, in a different context from the current memory retention example, where the hierarchical approach might be useful.

The basic balance between modeling how people differ and how they are the same is broadly applicable in psychological modeling. It is hard to think or cognitive processes or psychological phenomena that *do not* seem likely to have structured individual differences. Degenerate cases like ESP—in which nobody has the ability and so everybody is identical—might be a counter-example.

Exercise 10.3.2 Develop a modified model that does not require you to truncate the rate scale when sampling the α_i decay rates and β_i baselines for each subject. The truncation is not only theoretically inelegant, but technically problematic as it is implemented in WinBUGS, which does not distinguish between censoring and truncation. Implement your modified model and see whether it leads to different conclusions than the ones presented here.

TBA.

11 Signal detection theory

11.1 Signal detection theory

Exercise 11.1.1 Do you feel that the priors on discriminability and bias are plausible, a priori? Why or why not? Try out some alternative priors and study the effect that this has on your inference for the data sets discussed above.

The priors imply a uniform prior on the hit and false alarm rates. This is a nice statistical result, but one could argue that we expect hit rates generally to be high and false alarm rates generally to be low.

Exercise 11.1.2 Lehrner, Kryspin-Exner, and Vetter (1995) report data on the recognition memory for odors of three groups of subjects. Group I had 18 subjects, all with positive HIV antibody tests, and CD-4 counts of 240–700/mm^3. Group II had 19 subjects, all also with positive HIV antibody tests, but with CD-4 counts of 0–170/mm^3. The CD-4 count is a measure of the strength of the immune system, with a normal range being 500–700/mm^3, so Group II subjects had weaker immune systems. Group III had 18 healthy subjects and functioned as a control group. The odor recognition task involved each subject being presented with 10 common household odors to memorize, with a 30 sec interval between each presentation. After an interval of 15 min, a total of 20 odors were presented to subjects. This test set comprised the 10 previously presented odors, and 10 new odors, presented in a random order. Subjects had to decide whether each odor was "old" or "new." The signal detection data that resulted are shown in Table 11.2. Analyze these three data sets using signal detection theory to infer the discriminability and bias of the recognition performance for each group. What conclusions do you draw from this analysis? What, if anything, can you infer about individual differences between the subjects in the same groups?

TBA.

11.2 Hierarchical signal detection theory

Exercise 11.2.1 Of key interest for testing the Rips (2001) conjecture is how the group-level means for bias and (especially) discriminability differ between the induction and deduction conditions. What conclusion do you draw about the Rips (2001) conjecture base on the current analysis of the Heit and Rotello (2005) data?

TBA.

Exercise 11.2.2 Heit and Rotello (2005) used standard significance testing methods on their data to reject the null hypothesis that there was no difference between discriminability for induction and deduction conditions. Their analysis involved calculating the mean discriminabilities for each participant, using edge-corrections where perfect performance was observed. These sets of discriminabilities gave means of 0.93 for the deduction condition and 1.68 for the induction condition. By calculating via the t statistic, and so assuming associated Gaussian sampling distributions, and observing that the p-value was less than 0.01, Heit and Rotello (2005) rejected the null hypothesis of equal means. According to Heit and Rotello (2005), this finding of different discriminabilities provided evidence against the criterion-shifting uni-dimensional account offered by SDT. Is this consistent with your conclusions from the Bayesian analysis?

TBA.

Exercise 11.2.3 Re-run the analysis without discarding burn-in samples. This can be done by setting `nburnin` to 0 in the code `SDT_2.m` or `SDT_2.R`. The result should look something like Figure 11.6. Notice the strange set of samples leading from zero to the main part of the sampled distribution. Explain why these samples exist, and why they suggest burn in is important in this analysis.

TBA.

11.3 Parameter expansion

with Dora Matzke

Exercise 11.3.1 Experiment with different priors for the unscaled precision, such as `dgamma(0.1,0.1)`, `dgamma(0.01,0.01)`, or `dgamma(0.001,0.001)`, and for the scaling parameters, such as `dunif(0,1)`, `dunif(0,2)`, or `dnorm(0,1)`. How does the prior for the scaled standard deviations change when you

change the scaling factor?

Increasing the range of the uniform prior for the scaling factor (e.g., from `dunif(0,1)` to `dunif(0,2)`) results in a flatter prior for the scaled standard deviation. Using a standard normal prior for the scaling factor and a gamma prior for the unscaled precision results in a folded non-central t-prior for the scaled standard deviation (Gelman & Hill, 2007, p. 428).

12 Psychophysical functions

with Joram van Driel

12.1 Psychophysical functions

Exercise 12.1.1 What do you think is the function of the `thetalim` construction in the WinBUGS script?

The `thetalim` construction prevents the logistic function from taking on extreme values that cause numerical overflow or underflow errors ("undefined real result"). Any time you work with logit or probit models and you obtain an "undefined real result", consider bounding the function away from extreme values.

Exercise 12.1.2 The sigmoid curves in Figure 12.4 are single lines derived from point estimates. How can you visualize the uncertainty in the psychometric function?

The psychometric functions from Figure 12.4 were generated from a very concise summary of the posterior distributions for α and β, as given by their posterior modes. Consequently, Figure 12.4 does not reflect the uncertainty in α and β, and does not show the associated uncertainty in the estimated psychometric functions. This uncertainty can be visualized by drawing samples from the joint posterior distribution of α and β repeatedly, and every time plotting the psychometric function associated with those samples. Overlaying these functions in a single plot visualizes the variability. In principle, one could use all of the samples from the joint posterior. In practice, plotting 1000 or so psychometric functions is overkill. Instead, the variability can be visualized by first selecting a random subset of, say, 20 draws from the joint posterior for α and β, and then plotting the associated psychometric functions. This is accomplished in the code `PsychometricFunction1_Answers.m` or `PsychometricFunction1_Answers.R`,. The visualization generated makes clear the additional information that this brings, with subjects 5 and 6 having a fairly thin band of curves, but subject 4 having a few divergent lines towards extreme data points, and subject 8 having a thicker band of curves.

Exercise 12.1.3 Figure 12.4 shows the PSE for each subject. Compare subject 2 with subject 8. How do they differ in their perception of the intervals?

The standard comparison interval in this experiment was 300 ms. For an unbiased subject, the 50% point of subjective equivalence coincides with a test interval duration of 300 ms. Note that this holds true even for a subject with a large JND. The steepness of the psychometric function reflects the ability to discriminate between the standard and test interval, but the position along the x-axis reflects bias. In 50% of the cases, subject 2 classifies test intervals around 330 ms as shorter than the standard interval. Thus, subject 2 has a bias to say "shorter". The results for subject 8 show the opposite pattern. Subject 8 has a PSE of approximately 250 ms, indicating a bias to say "longer".

Exercise 12.1.4 One of the aims of the analysis is to use the psychometric function to infer the JND. In Figure 12.4 the JND is indicated by the difference on the x-axis between the dashed lines corresponding to the 50% and 84% points on the y-axis. The JNDs from Figure 12.4 are point estimates. Plot posterior distributions for the JND, and interpret the results. Which subjects are better at perceiving differences in time, and how certain are your conclusions?

The code for plotting posterior distributions for the JND is provided in `PsychometricFunction1_Answers.m` or `PsychometricFunction1_Answers.R`. These distributions are computed from the inverse of Equation 12.1 (i.e., providing the value on the x-axis that corresponds to a y-value of 0.50 and 0.84, and subtracting the result) with parameters α and β sampled from their joint posterior distribution. One could interpret the results as follows: better subjects are those with a JND distribution more to the left on the x-axis (they are better capable of perceiving smaller differences in duration); more variable subjects are those with a broader distribution (from their data a JND could be estimated with less certainty.) Subject 7 is clearly the worst subject.

Exercise 12.1.5 Look closely at the data points that are used to fit the psychometric functions. Are all of them close to the sigmoid curve? How do you think possible outliers would influence the function, and the inferred JND?

The data from subjects 4 and 7 illustrate the problem well. Subject 4 has a nice, clean psychometric function with almost all data points located close to it. However, for the extreme test interval of 400 ms (the value that the staircase-procedure started with, and that was clearly distinguishable from 300 ms), subject 4 did not always respond correctly. This is surprising, because subject 4 did better when confronted with several intervals that were more difficult (i.e., closer to 300 ms). Thus, the result for the 400 ms. test interval is likely to be an outlier, one that does not reflect the psychological process of time perception. Inclusion of the outlier distorts the estimate of the psychometric

function and produces a estimated JND larger and more variable than it should be. Note that subject 7 has data points lying at a similar distance from the curve. However, there are more of these points, with varying distances, and most of these lie around 300 ms (i.e., harder to distinguish from the standard). These points are thus not necessarily outliers, but indicate a less accurate and more variable subject.

12.2 Psychophysical functions under contamination

Exercise 12.2.1 How did the inclusion of the contaminant process change the inference for the psychophysical functions, and the key JND and PSE properties?

The changes are most pronounced for subject 4, the subject with a single clear outlier. For subject 4, inclusion of the contaminant process resulted in a different psychometric function, one that is further away from the outlier. A comparison of Figure 12.4 with Figure 12.2 reveals that, for subject 4, inclusion of the contaminant process decreased the JND (the distance between the two vertical dotted lines). Inclusion of the contaminant process did not affect the PSE; at the 50% point, the old and the new psychometric functions are almost identical, even for subject 4.

13 Extrasensory perception

13.1 Evidence for optional stopping

Exercise 13.1.1 What does the Bayesian analysis tell you about the association between sample size and effect size in the Bem (2011) studies?

The data are about 21.5 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$; this is strong evidence in favor of an association.

Exercise 13.1.2 Section 5.2 considered extending the correlation model in Figure 13.2 to incorporate uncertainty about the measures being related. Could that extension be usefully applied here?

TBA.

Exercise 13.1.3 A classical p-value test on the Pearson product-moment correlation coefficient yields $r = -0.87$, 95% CI $= [-0.97, -0.49]$, $p = 0.002$. What conclusions would you draw from this analysis, and how do they compare to the conclusions you drew from the Bayesian analysis?

TBA.

Exercise 13.1.4 We do not need to compute the Savage–Dickey density ratio on the original scale. For example, there are good arguments first to transform the posterior samples using the Fisher z-transform, so that $z = \operatorname{arctanh}(r)$. Try using this transformation. What difference do you observe?

The code `StopBem_Answer1.m` or `StopBem_Answer1.R` produces Figure 13.1; the posterior is now approximately normally distributed, and so it is the prior. One could even fit a normal distribution to the transformed posterior, and estimate the height of the posterior in that way.

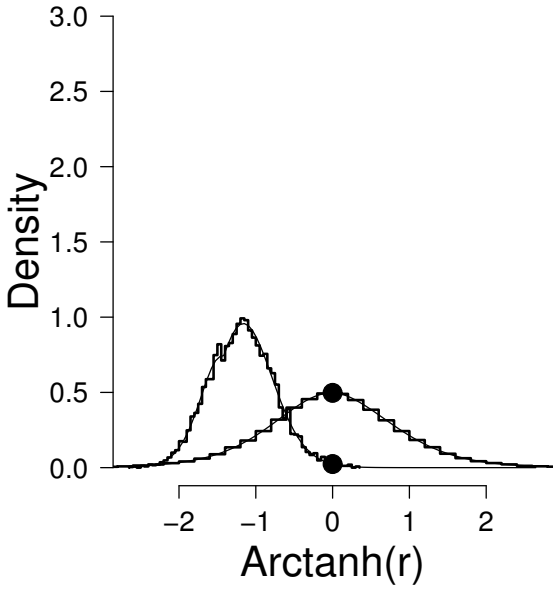

Fig. 13.1 Prior and posterior distributions for the Fisher z-transformed correlation coefficient between sample size and effect size in the Bem (2011) experiments.

13.2 Evidence for differences in ability

Exercise 13.2.1 Suppose that the alternative hypothesis does not assume a positive correlation between the abilities of subjects over the two sessions, but instead allows for any correlation, so that the prior is $r \sim \text{Uniform}(-1, 1)$. Intuitively, what is the value of the Bayes factor in this case?

Most of the posterior distribution for r lies to the right of zero, and this is consistent with the restriction that $r > 0$. Therefore, the restricted hypothesis $\mathcal{H}_1 : \texttt{r} \sim \texttt{dunif(0,1)}$ will be more competitive with $\mathcal{H}_0$ than the unrestricted hypothesis $\mathcal{H}_1 : \texttt{r} \sim \texttt{dunif(-1,1)}$. This is confirmed by numerical integration that yields $BF_{10} = 0.26$, so the odds in favor of $\mathcal{H}_0$ are $BF_{01} = 1/0.26 = 3.88$.

Exercise 13.2.2 A classical analysis yields $r = 0.12$, 95% CI $= [-0.08, 0.31]$, $p = 0.23$. This non-significant p-value, however, fails to indicate whether the data are ambiguous or whether there is evidence in favor of $\mathcal{H}_0$. How does the Bayes factor resolve this ambiguity?

TBA.

13.3 Evidence for the impact of extraversion

Exercise 13.3.1 What do you conclude about whether or not the correlation is zero, based on the Bayes factor?

TBA.

Exercise 13.3.2 Try more extreme assumptions about the accuracy with which extraversion is measured, by setting $\lambda^x = 1$ and $\lambda^x = 1/100$. How does the Bayes factor change in response to this change in available information?

TBA.

14 Multinomial processing trees

with Dora Matzke

14.1 Multinomial processing model of pair-clustering

Exercise 14.1.1 What do you conclude from the posterior distributions in Figure 14.3 about learning over the course of the trials?

All three parameters tend to increase with practice. However, the group shows a steeper increase in the r parameter over the trials than in either the c or the u parameter. Hence, recall of clustered pairs benefits the most from practice. For this particular data set, the conclusions are the same regardless if we fit the data with the latent-trait model or if we aggregate the data over participants and fit it with the non-hierarchical model.

Exercise 14.1.2 Because the u parameter corresponds to both the storage and retrieval of unclustered words, it is typically regarded as a nuisance parameter. In an approach to inference that is not fully Bayesian, the lack of interest in the posterior distribution of u might lead to the short-cut of a reasonable value being substituted, rather than assigning a prior distribution. Modify the graphical model so that u is set to a constant for each trial, given by the expected value of the posterior from the fully Bayesian analysis. How does this change affect the posterior distributions of c and r, the parameters of interest?

TBA.

14.2 Latent-trait MPT model

Exercise 14.2.1 What do you conclude from the posterior distributions in Figure 14.5 about learning over the course of the trials? Compare your conclusions from the latent-trait model to the conclusions from the original MPT model.

All three parameters tend to increase with practice. However, the group

shows a steeper increase in the r parameter over the trials than in either the c or the u parameter. Hence, recall of clustered pairs benefits the most from practice. For this particular data set, the conclusions are the same regardless if we fit the data with the latent-trait model or if we aggregate the data over participants and fit it with the non-hierarchical model.

Exercise 14.2.2 Extend the WinBUGS script to collect samples from the prior distributions for the standard deviation and correlation parameters. This will involve including variables `SigmaInvprior`, `Sigmaprior`, `rhoprior`, `sigmacprior`, `sigmarprior`, `sigmauprior`. Examine the prior and posterior distributions for the standard deviations and correlations. What can you conclude about the usefulness of including the correlation parameters in the latent-trait approach?

The standard deviations are estimated relatively precisely, while the correlations are not. We need more observations (e.g., by adding singletons) or some sort of parameter constraint across the trials to be able to obtain reliable posterior distributions for the correlation parameters.

Exercise 14.2.3 The latent-trait approach deals with parameter heterogeneity as a result of individual differences between subjects and relies on data that is aggregated over items. In many applications, however, it is reasonable to assume that the model parameters do not only differ between subjects but also between items. For example, it might be easier to cluster some pairs of semantically related words than others. This suggests using MPT models that incorporate both subject and item variability. Develop the graphical model that incorporates this extension. What is preventing the model from being applied to the current data?

We may model the probit transformed θ_{ijp} parameters as additive combinations of participant and item effects (e.g., Rouder & Lu, 2005; Rouder et al., 2007; Rouder, Lu, Morey, Sun, & Speckman, 2008), where the item effects are drawn from zero-centered (multivariate) normal distributions.

15 The SIMPLE model of memory

15.1 The SIMPLE model

Exercise 15.1.1 Modify the graphical model so that the same parameter values are used to account for all of the data sets. You will also need to modify the Matlab or R code that produces the graphs.

TBA.

15.2 A hierarchical extension of SIMPLE

Exercise 15.2.1 Why are empirical tests of generalization potentially more powerful or compelling evaluations of a model than fitting to existing data?

TBA.

16 The BART model of risk taking

16.1 The BART model

Exercise 16.1.1 Apply the model to data from a different subject, Bill, provided in the file `BillSober.txt`. Compare the estimated parameters for George and Bill. Who has the greater propensity for risk?

The posterior mean for parameter γ^+ is 0.70 for George and 1.58 for Bill. Since higher values of parameter γ^+ indicate a higher propensity for risk taking, Bill is the riskier decision taker.

Exercise 16.1.2 What happens if two pumps are added to each trial for George's data? Make this change to the `npumps` variable in Matlab or R, and examine the new results. Which of the two parameters changed the most?

Change `nPumps = Data[,6]` to `nPumps = Data[,6] + 2`. This will cause George's parameter estimate of γ^+, and therefore his propensity for risk taking, to increase. Note also, that the estimate of George's β parameter increases as well, suggesting that his pumping behavior is now more consistent.

Exercise 16.1.3 Modify George's data in a different way to affect the behavioral consistency parameter.

There are many ways of doing this, such as adding either -1 or 1 randomly to each number of pumps, alternately add and subtract one pump, or change each number of pumps to 4. In the end, the first two measures lead to a smaller estimate of the β parameter and the last measure leads to a substantially larger estimate of the β parameter.

16.2 A hierarchical extension of the BART model

Exercise 16.2.1 Apply the model to the data from the other subject, Bill. Does alcohol have the same effect on Bill as it did on George?

Participant Bill's behavior does not appreciably change from the sober to the tipsy condition. Only in the drunk condition is his behavior affected: his behavior becomes riskier andless consistent.

Exercise 16.2.2 Apply the non-hierarchical model in Figure 16.2 to each of the six data files independently. Compare the results for the two parameters to those obtained from the hierarchical model, and explain any differences.

The experimental effects are smaller for the hierarchical model, in particular for participant Bill. This is because in hierarchical modeling, parameters shrink towards the group mean.

Exercise 16.2.3 The hierarchical model in Figure 16.4 provides a structured relationship between the drinking conditions, but is still applied independently to each subject. Many of the applications of hierarchical modeling considered in our case studies, however, involve structured relationships between subjects, to capture individual differences. Develop a graphical model that extends Figure 16.4 to incorporate hierarchical structure both for drinking conditions and subjects. How could interactions between these two factors be modeled?

TBA.

17 The GCM model of categorization

with Ruud Wetzels

17.1 The GCM model

Exercise 17.1.1 Setting $b = 0.5$ to make the decision rule unbiased seems reasonable, since there are two alternatives with equal numbers of equally-often presented stimuli. But, the assumption can be easily examined. Change the model so that the bias parameter b is given a uniform prior over the range 0 to 1, and is inferred from the data. Summarize the findings from this model, and compare them to the results from the original model.

TBA.

Exercise 17.1.2 Figure 17.4 shows a posterior predictive analysis of the modeling. The average y_i counts are shown for each of the 8 stimuli, overlaid on gray violin plots showing the posterior predictive distributions. Also shown, by the broken lines, are individual participant data. These are linearly scaled from the individual count of 8, to the group count of 320, to allow visual comparison of the number of times. From this figure, what do you conclude about the ability of the GCM to describe the group data? What do you conclude about the adequacy of the group data as a summary of human performance?

TBA.

17.2 Individual differences in the GCM

Exercise 17.2.1 Compare the inferences about the attention parameter based on an individual subject analysis in Figure 17.7 with those based on a no individual differences analysis in Figure 17.3. Give a psychological interpretation of these differences, in terms of how the subjects selectively attended to the stimulus dimensions.

TBA.

Exercise 17.2.2 Three of the individual subject joint posterior distributions—for Subjects 3, 31, and 33—in Figure 17.7 are labeled. These three subjects lie in different areas of the parameter space, and so possibly correspond to different sorts of categorization behavior. Look at the individual data in Figure 17.5 for these three subjects, and give a short description of the differences in their categorization decisions.

TBA.

17.3 Latent groups in the GCM

Exercise 17.3.1 The analyses presented focus on the inferred group membership, and the posterior predictive distributions for each group. These might be two of the most important inferences, but they are not the only available or useful ones. Extend the analysis by considering the posterior distributions of the inferred proportion of contaminant subjects ϕ^c, and the difference in group mean attention δ, giving psychological interpretations for both.

TBA.

Exercise 17.3.2 Construct the posterior distribution for the probability that a subject is in the attend position group rather than the attend height group. This is not simply the posterior for the ϕ^g parameter.

TBA.

18 Heuristic decision-making

18.1 Take-the-best

Exercise 18.1.1 What is the posterior expectation of γ, and how can it be interpreted?

TBA.

18.2 Stopping

Exercise 18.2.1 Plot and interpret the posterior distribution of ϕ. Approximate and interpret the Bayes factor comparing $\mathcal{H}_0 : \phi = 0$ versus $\mathcal{H}_1 : \phi \neq 0$.

TBA.

18.3 Searching

Exercise 18.3.1 Look through the search order posterior distribution summaries for all of the subjects. How would you characterize the uncertainty they represent? Think about how many search orders are sampled, how many could be sampled, and how similar those sampled are to one another.

TBA.

Exercise 18.3.2 Do you expect to be able to make inferences about the full search order if a subject is using a one-reason stopping rule like TTB? What consequences for analysis does this issue have?

TBA.

Exercise 18.3.3 Are the inferred search orders for all of the subjects, and not just subjects 12 and 13, consistent with the individual differences observed

in the answers to question 17?

TBA.

18.4 Searching and stopping

Exercise 18.4.1 How do the posterior expectations of the ϕ_i parameters for the searching and stopping model in Figure 18.8, compare to the posterior expectations of the z_i parameters for the stopping model in Figure 18.3? Interpret the similarities and differences.

TBA.

Exercise 18.4.2 In what sense does the current model incorporate, and not incorporate, structured individual differences over searching and stopping? Suggest hierarchical extensions to the model in Figure 18.8 that could add structure to the individual differences in searching and stopping parameters.

TBA.

19 Number concept development

19.1 Knower level model for Give-N

Exercise 19.1.1 Report the posterior for the evidence parameter v.

TBA.

Exercise 19.1.2 Interpret the distinctive visual patterns of posterior prediction in Figure 19.4 for each child, explaining how they combine knower level knowledge, and the task base-rate.

TBA.

Exercise 19.1.3 What do you think of relying on the maximum a posteriori summary of the posterior uncertainty about knower levels to classify children? What might be a justifiable alternative?

TBA.

Exercise 19.1.4 The model currently assumes that the final decision is sampled from the distribution over all possible responses, in proportion to the mass associated with each response. Does this probability matching strategy seem psychologically plausible? What is an alternative model of this part of the decision-making process?

TBA.

Exercise 19.1.5 Explain why the distribution shown in Figure 19.2 is not exactly the posterior distribution of the base-rate $\boldsymbol{\pi}$. What distribution is shown, how is it related to the posterior of $\boldsymbol{\pi}$, and what are the advantages of presenting the distribution in Figure 19.2?

TBA.

19.2 Knower level model for Fast Cards

Exercise 19.2.1 Report the posterior for the evidence parameter v, and compare it to the value found in the Give-N analysis.

TBA.

Exercise 19.2.2 Compare the posterior distributions over knower levels shown in Figure 19.8 with those inferred using the Give-N data in Figure 19.3.

TBA.

Exercise 19.2.3 The uncertainty in the posterior distribution in Figure 19.8 always involves adjacent knower levels (e.g., it is uncertain whether child 2 is a one-knower or two-knower). Does this follow necessarily from the statistical definition of the z knower level parameter? If so, how? If not, how do the patterns of uncertainty in Figure 19.8 arise?

TBA.

19.3 Knower level model for Give-N and Fast Cards

Exercise 19.3.1 The behavioral data for Child 18 are detailed in Table 19.1, showing their answers to every question in both tasks. Explain why the inference based on only the Give-N data in Figure 19.3 has most posterior mass on four-knower, but there is some uncertainty, with three-knower also a possibility. Explain why the inference based on only the Fast-Cards data in Figure 19.8 has posterior mass almost entirely on the three-knower possibility. Explain why the combined inference in Figure 19.11 favors three-knower, and why the possible four-knower inference in the Give-N analysis could be viewed as arising from the nature of the task itself, rather than from the actual number knowledge of the child that is of primary interest developmentally.

TBA.

References

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge, MA: Cambridge University Press.

Andrich, D. (1988). *Rasch Models for Measurement*. London: Sage.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Berger, J. O., & Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

de Finetti, B. (1974). *Theory of Probability, Vol. 1 and 2*. New York: John Wiley.

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

Grant, J. A. (1974). Evaluation of a screening program. *American Journal of Public Health*, *64*, 66–71.

Heit, E., & Rotello, C. (2005). Are there two kinds of reasoning? In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 923–928). Mahwah, NJ: Erlbaum.

Lehrner, J. P., Kryspin-Exner, I., & Vetter, N. (1995). Higher olfactory threshold and decreased odor identification ability in HIV-infected persons. *Chemical Senses*, *20*, 325–328.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.

Poehling, K. A., Griffin, M. R., & Dittus, R. S. (2002). Bedside diagnosis of influenza virus infections in hospitalized children. *Pediatrics*, *110*, 83–88.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*, 129–134.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.

Wagenmakers, E.-J., & Morey, R. D. (2013). Simple relation between one–sided and two–sided Bayesian point–null hypothesis tests. *Manuscript submitted for publication*.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t-test. *Psychological Bulletin & Review*, *16*, 752–760.