STAT 535 – Spring 2022 – Homework 3

1. The Gibbs sampler can take a long time to converge if the target distribution is multimodal. Suppose we are trying to sample from a density for a parameter θ that is the mixture of three normal densities: θ has density $f(\theta) = 0.45f_1(\theta|\mu = -3, \sigma^2 = 1/3) + 0.10f_2(\theta|\mu = 0, \sigma^2 = 1/3) + 0.45f_3(\theta|\mu = 3, \sigma^2 = 1/3)$, where $f_j(\cdot|\mu, \sigma^2)$ here represents the normal density with mean μ and variance σ^2. The marginal distribution of θ is plotted on the last page.

   Note that we could easily sample from this distribution using ordinary Monte Carlo methods, but consider using Gibbs sampling to sample from it.

   (a) If the indicator $\delta \in \{1, 2, 3\}$, argue that the full conditional distribution for θ is

$$\theta|\delta \sim N(\mu_\delta, \sigma^2_\delta).$$

   What are μ_δ and σ^2_δ here, for $\delta = 1, 2, 3$?

   $\mu_1 = -3$, $\sigma^2_1 = 1/3$, $\mu_2 = 0$, $\sigma^2_2 = 1/3$, $\mu_3 = 3$, $\sigma^2_3 = 1/3$.

   (b) Use Bayes' theorem to show that the full conditional for δ is

$$P[\delta = k|\theta] = \frac{P[\delta = k] \times \text{dnorm}(\theta, \mu_k, \sigma_k)}{\sum_{m=1}^{3} P[\delta = m] \times \text{dnorm}(\theta, \mu_m, \sigma_m)}, \text{ for } k \in \{1, 2, 3\}$$

   where `dnorm` represents the normal density function (R shorthand).

   By Bayes' rule,

$$P[\delta = k|\theta] = \frac{p(\delta = k)p(\theta|k)}{p(\theta)}$$

   and then in the denominator,

$$p(\theta) = \sum_{m=1}^{3} P[\delta = m] \times p(\theta|m) = \sum_{m=1}^{3} P[\delta = m] \times \text{dnorm}(\theta, \mu_m, \sigma_m)$$

   by the law of total probability.

   (c) Write a Gibbs sampling algorithm (you can use the helpful R commands given on the course web page) to sample from the joint density of (θ, δ). Begin the chain with the initial values $\delta^{[0]} = 2$ and $\theta^{[0]} = 0$, and generate 1000 values of θ. Give a plot of a relative frequency histogram of the θ values – using a command like `hist(theta.values, freq=F)` – and comment on how it compares to the true marginal density of θ plotted on the last page. (You could try repeating part (c) a few times to get a sense for the variability in the Gibbs sampler results, as well.)

   The R code is given on the course webpage. The histogram may look something like this (I've added the true density on top of the histogram, just for a reference). Unless you got lucky, this histogram may not do a good job of approximating the distribution of θ.

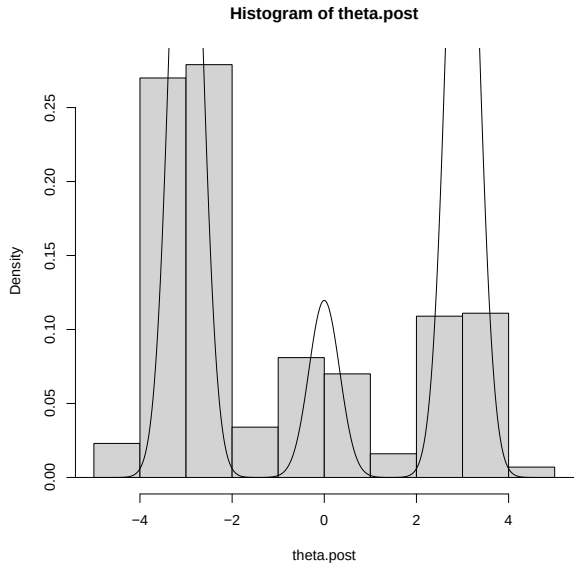

Figure 1: Problem 1: Histogram that approximates the distribution for θ, based on 1000 iterations.

(d) Repeat part (c), but generate 40000 values of θ. Again give a plot of a relative frequency histogram of the θ values and comment on how it compares to the true marginal density of θ plotted on the last page. (You could try repeating part (d) a few times to get a sense for the variability in these Gibbs sampler results, as well.)

The R code is given on the course webpage. The histogram may look something like this (I've added the true density on top of the histogram, just for a reference). This histogram should do a better job of approximating the distribution of θ.

(e) Try trace plots for the θ values and plots of the autocorrelation functions for parts (c) and (d). Comment on what these MCMC diagnostics tell you.

The trace plot shows the MCMC algorithm has not stabilized after 1000 iterations. The ACF plot shows severe autocorrelation between sampled values.

The trace plot shows the MCMC algorithm has not exactly converged to one spot, but that is just because of the nature of the multimodal distribution we are sampling from. All in all it seems to be sampling from the true distribution pretty well. The ACF plot still shows severe autocorrelation — some thinning may be useful.

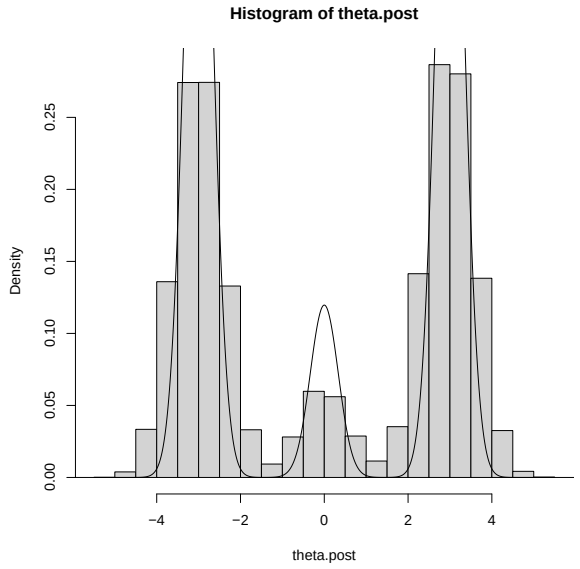

Figure 2: Problem 1: Histogram that approximates the distribution for θ, based on 40000 iterations.

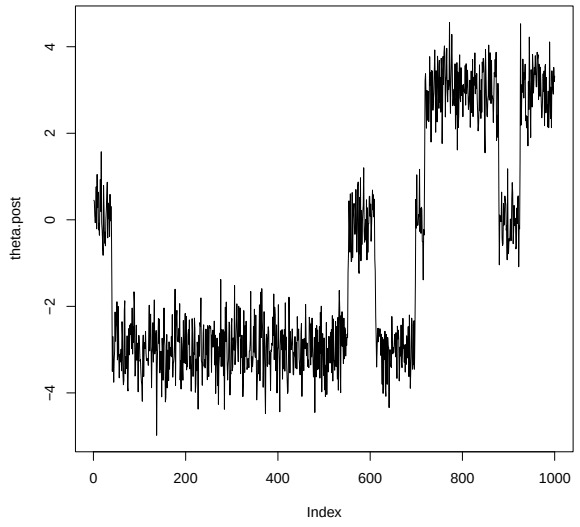

Figure 3: Problem 1: Trace plot based on 1000 iterations.

2. Two separate test developers have created different IQ tests. Each of these tests is designed so that scores follow a normal distribution with population mean 100 and standard deviation 15. Hence any set of test scores can be standardized to so that the

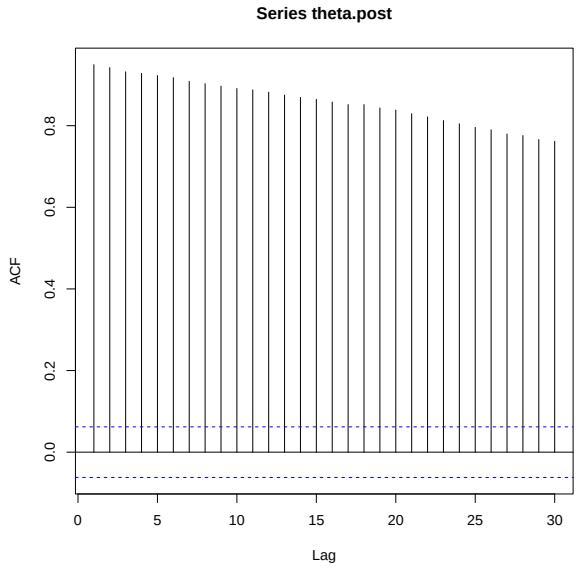

Figure 4: Problem 1: ACF plot based on 1000 iterations.

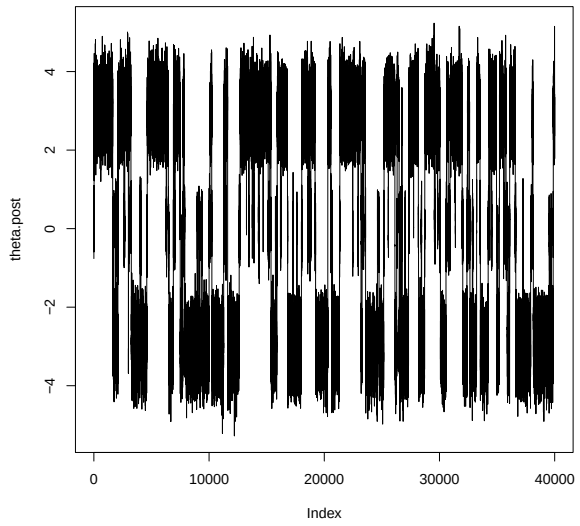

Figure 5: Problem 1: Trace plot based on 40000 iterations.

standardized scores follow a $N(0, 1)$ distribution. We wish to investigate the coefficient of correlation ρ between test-takers' scores on the two tests. Consider a random sample of n test-takers, who each take both IQ tests. Let $X_1, \ldots, X_n$ be the test-takers' standardized scores on the first test, and let $Y_1, \ldots, Y_n$ be the corresponding standardized

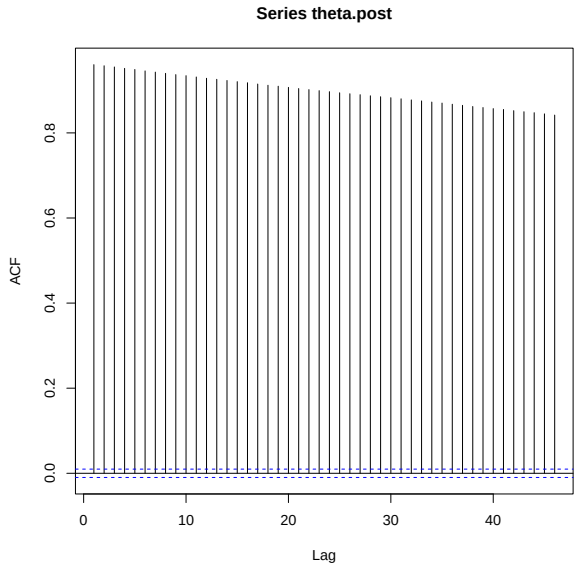

Figure 6: Problem 1: ACF plot based on 40000 iterations.

scores on the second test. The likelihood is

$$L(\rho|\mathbf{x},\mathbf{y}) = [2\pi]^{-n}[1-\rho^2]^{-n/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right]\right\}.$$

(a) Suppose we use the prior $p(\rho) = 1$, $0 < \rho < 1$, for ρ. Show that the posterior for ρ is

$$p(\rho|\mathbf{x},\mathbf{y}) \propto [(1-\rho^2)]^{-n/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right]\right\}$$

$$\begin{aligned} p(\rho|\mathbf{x},\mathbf{y}) &\propto L(\rho|\mathbf{x},\mathbf{y})p(\rho) \\ &= [2\pi]^{-n}[1-\rho^2]^{-n/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right]\right\} \quad (1) \\ &\propto [(1-\rho^2)]^{-n/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\sum x_i^2 - 2\rho\sum x_i y_i + \sum y_i^2\right]\right\} \end{aligned}$$

(b) We gather data on 13 test-takers. Their standardized scores on the two tests are:

```
X_i:  0.92  0.42  3.62  0.89 -0.69  0.45 -0.11 -0.14 -0.47  1.09 -0.34  0.62  0.27
Y_i:  0.26  1.65  2.10  0.62 -1.16  1.29 -0.82 -0.36 -0.29  0.86  0.19  1.25  0.33
```

Plugging in the necessary summary statistics, write and simplify the posterior (up to a constant of proportionality).

(c) We will use the following proposal density to generate values of ρ:

(i) Given current value $\rho^{[t]}$, sample $\rho^* \sim \text{Uniform}(\rho^{[t]} - 0.2, \rho^{[t]} + 0.2)$.

(ii) If the sampled $\rho^* < 0$, then set $\rho^* = |\rho^*|$.

(iii) If the sampled $\rho^* > 1$, then set $\rho^* = 2 - \rho^*$.

Argue that this is a symmetric proposal density.

This is quite of tricky and technical, but here is an example argument:

If the current value $\rho^{[t]} \in [0.2, 0.8]$, then the uniform density centered at $\rho^{[t]}$ is clearly symmetric.

Suppose $\rho^{[t]} < 0.2$. For example, say $\rho^{[t]} = 0.1$, so that a sampled ρ^* of 0.03 or -0.03 each yield a candidate $\rho^* = 0.03$. The density value associated with a Unif$(0.1-0.2, 0.1+0.2)$ distribution at locations -0.03 and 0.03 is the same as the density value associated with a Unif$(0.03 - 0.2, 0.03 + 0.2)$ distribution at locations -0.1 and 0.1.

Now suppose $\rho^{[t]} > 0.8$. For example, say $\rho^{[t]} = 0.9$, so that a sampled ρ^* of 0.97 or 1.03 each yield a candidate $\rho^* = 0.97$. The density value associated with a Unif$(0.9 - 0.2, 0.9 + 0.2)$ distribution at locations 0.97 and 1.03 is the same as the density value associated with a Unif$(0.97-0.2, 0.97+0.2)$ distribution at locations 0.9 and 1.1 (each of which yield a next candidate ρ^* of 0.9.

(d) Explain carefully and completely how the provided R code on the course web page performs the Metropolis-Hastings algorithm.

(e) Using the code on the course web page, sample from the posterior distribution of ρ. Perform diagnostics to check convergence and check autocorrelation, and perform remedial action if needed. Summarize the posterior distribution of ρ, including giving an estimated density plot, point estimate, and 95% interval estimate for ρ.

My acceptance rate was around 0.64, which is a little high, but not terrible. This could be changed by adjusting the width of the proposal density.

The trace plot shows the MCMC has stabilized well. The ACF plot shows high autocorrelation. We can remedy this by thinning; I will take every 10th value of the chain.

The trace plot still looks good. The thinned chain shows no autocorrelation, so the problem has been fixed.

Based on the thinned chain: The posterior median is around 0.71. A quantile based 95% credible interval is $(0.38, 0.86)$. An HPD 95% credible interval is $(0.43, 0.88)$.

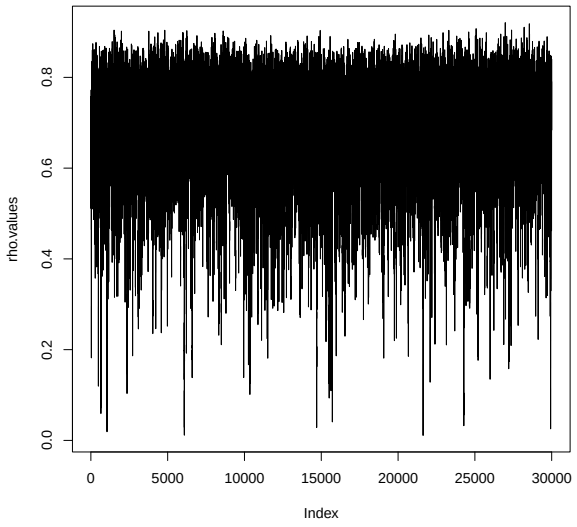

Figure 7: Problem 2: Trace plot based on 30000 iterations.

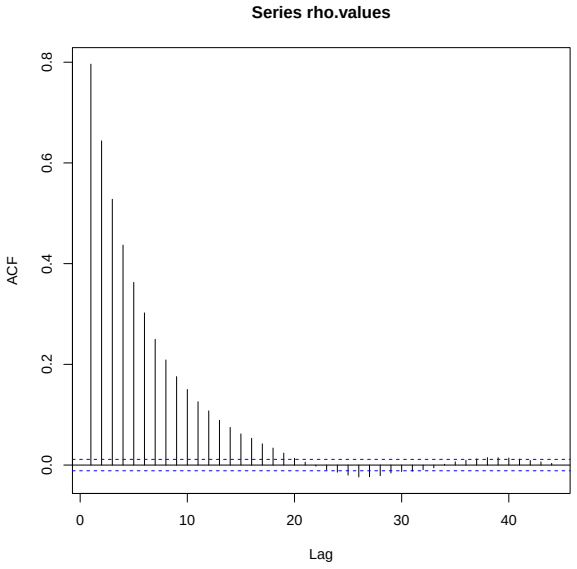

Figure 8: Problem 2: ACF plot based on 30000 iterations.

3. In class (for the Prussian cavalry example) we derived the posterior predictive distribution in the case where $Y_1, \ldots, Y_n \sim$ i.i.d. Poisson(λ), with the Gamma prior $\lambda \sim G(2, 4)$. Instead of deriving the posterior predictive distribution analytically, we could have sampled from it using Monte Carlo methods.

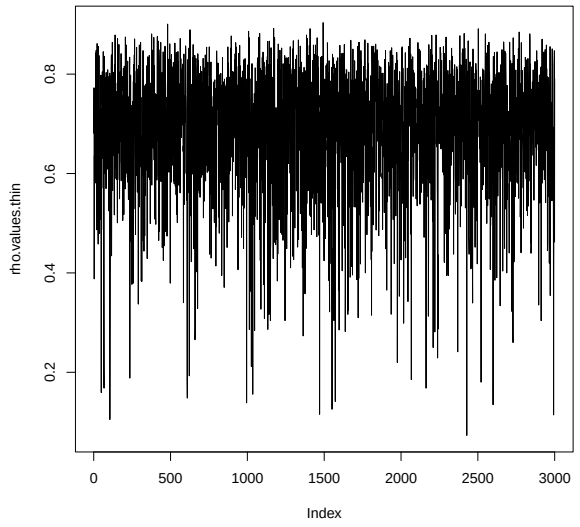

Figure 9: Problem 2: Trace plot based on thinned chain.

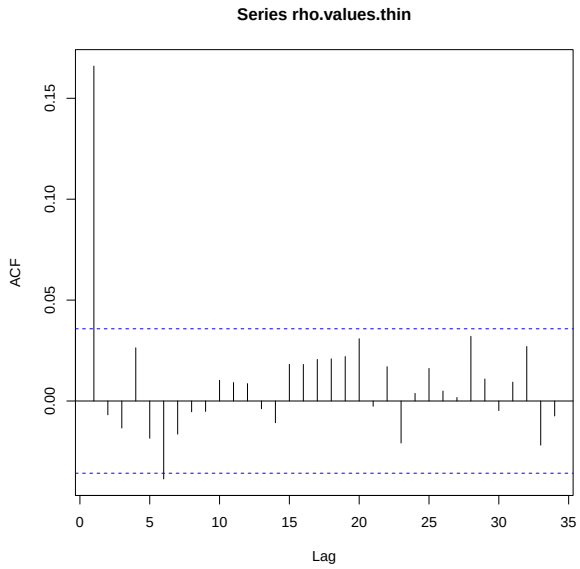

Figure 10: Problem 2: ACF plot based on thinned chain.

(a) Randomly sample $\lambda^{[1]}, \ldots, \lambda^{[J]}$ from a Gamma(shape $= \sum y_i + 2$, rate $= n + 4$) distribution. Using these, sample $Y^{[1]}, \ldots, Y^{[J]}$ from Poisson($\lambda^{[j]}$) distributions, $j = 1, \ldots, J$.

See R code on course webpage.

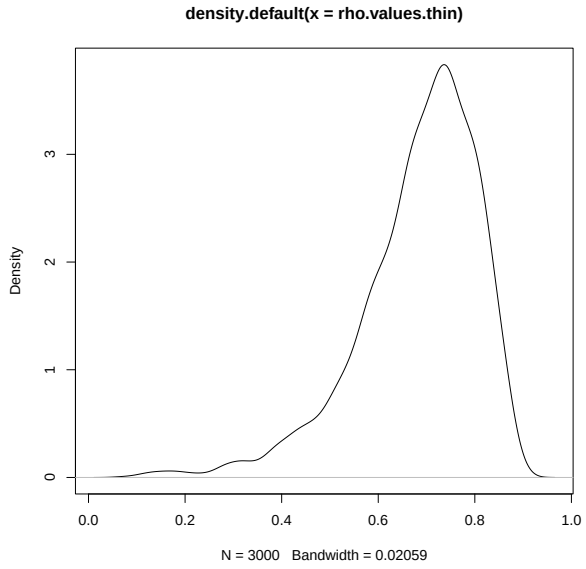

Figure 11: Problem 2: Density plot based on thinned chain.

(b) Plot the approximate posterior predictive distribution, similarly as our in-class example. How does it appear to compare to the observed Prussian army data distribution?

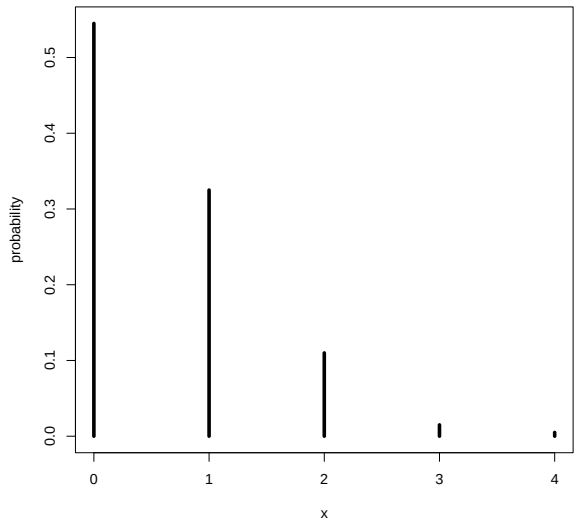

Figure 12: Problem 3: Plot of observed proportions of values in the Prussian horse kick example.

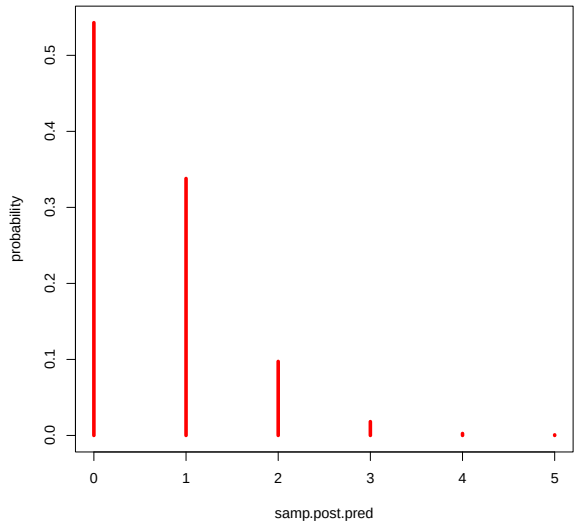

Figure 13: Problem 3: Plot of posterior predictive distribution.

The posterior predictive distribution looks very similar to the observed proportions. This indicates the Poisson model is a great fit to these data.

4. In a NASA experiment, 14 male rats were sent into space. When they returned, the red blood cell mass (in ml) of each rat was measured. In addition, 14 other male rats were kept on earth during the same period of time. Those rats also had their red blood cell mass measured. Assume the red blood cell masses for the two groups can be modeled with a normal distribution, with equal variances across the two groups. The data are:

```
Space rats: 8.59 8.64 7.43 7.21 6.87 7.89 9.79 6.85 7.00 8.80 9.30 8.03 6.39 7.24
Earth rats: 8.65 6.99 8.40 9.66 7.62 7.44 8.55 8.70 7.33 8.58 9.88 9.94 7.14 9.04
```

(a) Suppose the research question of interest was to test whether the mean red blood cell mass differed for the two groups. Answer this question based on a Bayesian hypothesis test. Clearly state your prior specifications (use $\mu_\Delta = 0, \sigma^2_\Delta = 1/5$). Give a posterior probability for each hypothesis being true.

The prior beliefs do not favor either hypothesis, since I choose $\mu_\Delta = 0$. My Bayes Factor was 0.83 for the null model that the mean red blood cell mass is the same for each group. So there is some evidence that the means differ, but it is not especially strong evidence. The posterior probability for H_0 (means are same) is 0.455, so the posterior probability for H_a (means differ) is 0.545.

(b) Suppose the research question of interest was to test whether the mean red blood cell mass for the space group was lower than the mean red blood cell mass for the control group. Answer this question based on a Bayesian hypothesis test. Clearly state your

prior specifications. (The researcher believed *a priori* that the rat population as a whole might have average red blood cell mass somewhere around 7 ml, but was not at all sure about the effect of the space travel.) Give a posterior probability for each hypothesis being true.

I used a Gibbs sampling approach with a normal prior on μ (with mean 7) and a normal prior on τ (with mean 0, not favoring either group *a priori*). The posterior probability of H_a (mean is lower for space group) was about 0.87. So the posterior probability of H_0 was about 0.13. I'd tend to believe H_a (mean is lower for space group), based on that. But that is a subjective decision of mine.

5. A physician is interested in determining whether the mean systolic blood pressure of a certain set of patients is less than 130. She takes a random sample of 17 patients and measured their systolic blood pressure. Assume the measurements follow a $N(\mu, \sigma^2)$ distribution, with μ unknown and known $\sigma^2 = 225$. (You can use normal-normal results from Chapter 5 to obtain the posterior distribution for μ.) The physician says *a priori* that she is 95% sure that the true mean systolic blood pressure is between 120 and 140. The data are:

```
118 140 90 150 128 112 134 140 112 126 112 148 124 130 142 105 125
```

(a) Conduct a Bayesian hypothesis test of $H_0 : \mu \geq 130$ vs. $H_a : \mu < 130$, basing your conclusions on the posterior distribution for μ. Clearly state your prior specifications.

Based on the expert opinion, I use a normal prior on μ with mean 130 and standard deviation 5. The true σ is assumed to be 15. The conjugate analysis gives a normal posterior with mean 127.1539 and standard deviation 2.941742 (see R code for how to get this). Based on this, the posterior probability that H_0 is true is $P(\mu \geq 130|\mathbf{y}) = 0.167$. Based on this, I might conclude that $\mu < 130$, but this is my own judgment call.

(b) Conduct a classical t-test using $\alpha = 0.05$. Are the substantive conclusions any different from those in part (a)?

The classical t-test of $H_0 : \mu \geq 130$ vs. $H_a : \mu < 130$ has a P-value of 0.142. So using $\alpha = 0.05$, we would fail to reject H_0 and we would not conclude that $\mu < 130$. Answers will vary about whether this is different from the conclusion in part (a).

6. Do Problem 6.3 from the *Bayes Rules!* textbook. [Read Section 6.3 for more details about "mixing slowly."]

Mixing too slowly means that the chain is taking a long time to become a true sample from the posterior distribution. You would have to run the chain for many, many iterations in order to get valid results.

High autocorrelation means that the values drawn from the posterior are not independent, so it is not a random sample from the posterior. This could affect the posterior inference.

If the chain gets stuck, it is not exploring the entire posterior distribution; it may be getting stuck in areas of high probability. Again, the sampled values will not be representative of the whole posterior.

7. Do Problem 7.6 from the *Bayes Rules!* textbook. [Hint: Use R functions like `rnorm` and `runif` to draw the requested value, and **(just for the purposes of this exercise!)**, remember to enter `set.seed(84735)` before each random draw, so that the random number that you draw will match the random number that I draw.]

```
> #(a)
> set.seed(84735)
> rnorm(n=1,mean=4.6,sd=2)
[1] 5.934465
>
> #(b)
> set.seed(84735)
> rnorm(n=1,mean=2.1,sd=7)
[1] 6.770629
>
> #(c)
> set.seed(84735)
> runif(n=1,min=8.9-2,max=8.9+2)
[1] 9.890753
>
> #(d)
> set.seed(84735)
> runif(n=1,min=1.2-0.5,max=1.2+0.5)
[1] 1.447688
>
> #(e)
> set.seed(84735)
> runif(n=1,min=7.7-3,max=7.7+3)
[1] 9.18613
>
```

8. Do Problem 7.7 from the *Bayes Rules!* textbook. [Hint: Use R functions like `dnorm`, `dunif`, and `dexp` to calculate the value of the proposal density for the specified current value and proposed value of the parameter. These proposal density values are the q parts that go in the numerator and denominator of the Metropolis-Hastings acceptance ratio probability.]

```
> proposed<-2.1
> current<-2
>
> #(a)
>
> top<-((proposed)^(-2))*dnorm(current,mean=proposed, sd=1)
> bottom<-((current)^(-2))*dnorm(proposed,mean=current, sd=1)
> accept.ratio <- top/bottom
> print(accept.ratio)
[1] 0.9070295
>
> #(b)
>
> top<-(exp(proposed))*dnorm(current,mean=proposed, sd=0.2)
> bottom<-(exp(current))*dnorm(proposed,mean=current, sd=0.2)
> accept.ratio <- top/bottom
> print(accept.ratio)
[1] 1.105171
>
> #(c)
>
> top<-(exp(-10*proposed))*dunif(current,min=proposed-0.5, max=proposed+0.5)
> bottom<-(exp(-10*current))*dunif(proposed,min=current-0.5, max=current+0.5)
> accept.ratio <- top/bottom
> print(accept.ratio)
[1] 0.3678794
>
> #(d)
>
> top<-(exp(-(proposed^4)))*dexp(current,rate=proposed)
> bottom<-(exp(-(current^4)))*dexp(proposed,rate=current)
> accept.ratio <- top/bottom
> print(accept.ratio)
[1] 0.03339631
>
```

We can see that in part (b), we are DEFINITELY going to accept the proposed value, since the acceptance ratio is greater than 1.

9. Do Problem 8.8 from the *Bayes Rules!* textbook. [You do not have to provide the sketches of the intervals on the posterior pdf. And you can use the `hpd` function in the `TeachingDemos` package to get the HPD intervals.]

```
> # HPD:
>
> library(TeachingDemos)
>
> hpd(qgamma,shape=1,rate=5,conf=0.95)
[1] 0.000000001268843 0.599146480087658
>
> # middle 95%:
>
> c(qgamma(0.025,shape=1,rate=5), qgamma(0.975,shape=1,rate=5) )
[1] 0.005063562 0.737775891
>
> # HPD:
>
> library(TeachingDemos)
>
> hpd(qnorm,mean=-13,sd=2,conf=0.95)
[1] -16.919928  -9.080072
>
> # middle 95%:
>
> c(qnorm(0.025,mean=-13,sd=2), qnorm(0.975,mean=-13,sd=2) )
[1] -16.919928  -9.080072
>
```

We see the two methods produce different intervals with the gamma posterior density, since the gamma density is skewed. The HPD interval is much shorter and is probably to be preferred here.

With the normal posterior density, the two methods produce the same interval since the normal density is symmetric.

10. Do Problem 8.9 from the *Bayes Rules!* textbook.

```
> # posterior probability for Ha:
>
> post.prob.Ha <- 1-pbeta(0.4,4,3)
> print(post.prob.Ha)
[1] 0.8208
>
> # posterior odds for Ha:
>
> post.odds.Ha <- (post.prob.Ha)/(1-post.prob.Ha)
> print(post.odds.Ha)
```

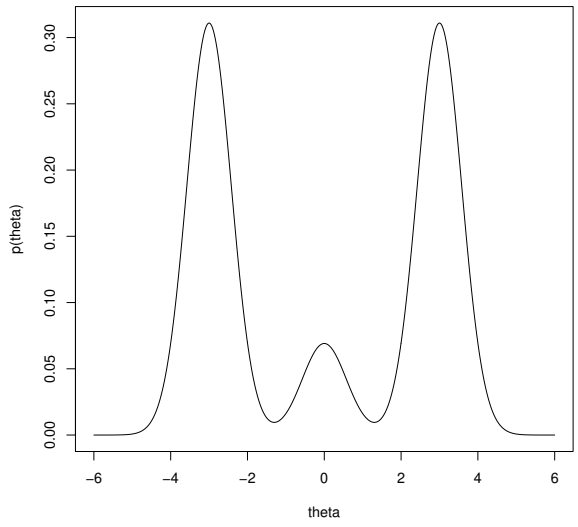

Figure 14: True distribution for θ.

```
[1] 4.580357
>
> # prior odds for Ha:
>
> prior.prob.Ha <- 1-pbeta(0.4,1,0.8)
> prior.odds.Ha <- (prior.prob.Ha)/(1-prior.prob.Ha)
> print(prior.odds.Ha)
[1] 1.98098
>
> # Bayes factor for Ha:
>
> (post.odds.Ha)/(prior.odds.Ha)
[1] 2.312168
>
> # Note the Bayes Factor for H0 would be the reciprocal of this.
>
```

After seeing the data, the odds that H_a is true are 2.3 times as great as they were before seeing the data.

Before seeing the data, I believed H_a was about twice as likely as H_0. After seeing the data, I now believe H_a is about 4.6 times as likely as H_0.