Bayesian Statistics - Exercise 1

Objective

In this first exercise you will practice working with rjags, a package that enables you to work with JAGS (Just Another Gibbs Sampler) in R for Bayesian data analysis. JAGS is a program for analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulation. You will get hands on experience with applying the key concepts of Bayesian analysis.

Help files

If you need help, look at theJAGS user manual or the rjags reference manual. If you are stuck with a specific error message, a Google search may turn up possible causes and solutions.

Setting up your system

To set up your system for using JAGS, there are two steps:
1. Install the current version of JAGS.
2. Install the current rjags package from CRAN.

The data

Several studies suggest that cognitive behavioral therapy is an effective treatment for Post-Traumatic Stress Disorder (PTSD) in male veterans. Suppose that you did a study to compare Prolonged Exposure (PE), a type of cognitive behavioral therapy, with Present-Centered therapy (PC), a supportive intervention. It was a randomized controlled trial, where 284 veterans suffering from PTSD were assigned randomly to receive either PE or PC. The outcome measure of interest was loss of diagnosis (LD), a dichotomous variable. The resulting data are displayed in Table **??**.

Table 1: PTSD data

	Type of Intervention	
	PE	PC
Loss of Diagnosis	58	40
Total Men Treated	141	143

Your goal now is to analyze this data in OpenBUGS to examine whether PE is an effective treatment for PTSD in veterans (i.e., more effective than the 'control condition' of supportive intervention).

A. Step 1: Input your data.

In R, create a new file in the directory where you saved the model file. In this file, specify the numbers of 'successes' (LD) and numbers of subjects in both treatment conditions with the following statement:

```
# You can specify the data in this manner:
dat <-list(y.PE=58, n.PE=141, y.PC=40, n.PC=143)
#Alternatively, you can source the data from a txt file,
#this will be useful when using bigger datasets later on.
source('Exercise 1 - Data.txt')
```

B. Step 2: Specify your model.

Running a model with the RJAGS package consists of several steps. The first step is loading in the data, as you did in the previous step. The second step is creating a model object with the `jags.model()` function. The `jags.model()` function needs as input a file containing a description of the model in JAGS language. Therefore, in this step, we describe the model in JAGS language in a .txt file that we can give as input to `jags.model()` in a later step.

Open the file `Exercise 1 - Model.txt` in notepad. This is a template for the model that you will need to specify. Using `#` you can also add your own comments. Since you are interested in the proportion of the groups that has Lost their Diagnosis, you can model the number of 'successes' (y) in each condition by a binomial distribution:

$$y_{\text{PE}} \sim Binomial(\theta_{\text{PE}}, n_{\text{PE}})$$

$$y_{\text{PC}} \sim Binomial(\theta_{\text{PC}}, n_{\text{PC}}),$$

where θ_{PE} is the 'success probability' in the PE condition, and n_{PE} is the total number of persons in the PE condition, and similarly for the PC condition. In order for JAGS to understand the model, you can specify the two likelihood functions in the model txt-file by including the following statements:

```
# likelihood of the data
y.PE ~ dbin(theta.PE, n.PE)
y.PC ~ dbin(theta.PC, n.PC)
```

A natural and conjugate prior for each of the parameters θ_{PE} and θ_{PC} is the beta distribution:

$$\theta_{\text{PE}} \sim Beta(\alpha_{\text{PE}}, \beta_{\text{PE}})$$

$$\theta_{\text{PC}} \sim Beta(\alpha_{\text{PC}}, \beta_{\text{PE}}),$$

which can be made non-informative by setting $\alpha_{\text{PE}} = \beta_{\text{PE}} = 1$ and $\alpha_{\text{PC}} = \beta_{\text{PC}} = 1$. You can specify this as follows:

```
# prior distributions
theta.PE ~ dbeta(1,1)
theta.PC ~ dbeta(1,1)
```

Using the statements above, you have a complete model file that estimates separate models for the two groups. However, because you are interested in the contrast between the groups, you should also specify some measure of comparison. One such measure is the *relative risk* (risk of staying ill, in this case), defined as $RR = \theta_{\text{PC}}/\theta_{\text{PE}}$. With the following statements you can let JAGS calculate the RR in each iteration of the sampler:

```
# contrast
RR <- theta.PC/theta.PE
```

When you have finished the model file, you need to save it (**Ctrl+S or File: Save As...**).

C. Step 3: Obtain initial values.

For this particular model, it is not necessary to provide any initial values manually. JAGS automatically generates initial values when the model is specified, and no initial vales are provided. These are chosen to be a typical value from the prior distribution.

D. Step 4: Obtain samples from the posterior distribution of the parameters.

For the next steps in the analysis you will run JAGS from R using the rjags package. First load rjags:

```
library(rjags)
```

Next, create a model object in R by means of the `jags.model` function. You need to specify your model .txt-file, data and number of chains:

```
model.def <- jags.model(file = 'Exercise 1 - Model.txt', data = dat, n.chains = 2)
```

Subsequently, use the `update()` function to run a large number of burn-in iterations (for example 1000 iterations) for your model:

```
# burn-in period :
update(object = model.def, n.iter = )
```

Then, use the `coda.samples()` function to set monitors on the parameters of interest and draw a large number of samples from the posterior distribution, (for example 10000):

```
# obtain samples from the posterior distribution of the parameters and monitor these:
parameters <- c('theta.P', 'theta.PC', 'RR')
res <- coda.samples(model = model.def, variable.names = parameters, n.iter = )
```

E. Step 5: Inspecting convergence.

At this point, we have not covered the topic of convergence so we skip this step. This is just a reminder that in a real Bayesian analysis this step always comes before interpretation of the results.

F. Step 6: Substantive interpretation.

Use the `summary()` function to inspect the results. Look at the posterior means, medians, and 95% credible intervals. How convinced are you that the PE type of cognitive behavioral therapy is an effective treatment for PTSD (compared to the baseline of PC)?

G. Deriving the posteriors analytically.

Rjags has provided you with the posterior results for the RR, and for each of the proportions θ_{PE} and θ_{PC} separately. However, for this simple model it is also possible to derive these results analytically.
The posterior distribution for each θ_i is given by

$$p(\theta_i|y_i, n_i) \propto \theta_i^{\alpha_i+y_i-1}(1-\theta_i)^{\beta_i+n_i-y_i-1},$$

and the mean of this distribution for $\hat{\theta}_i$ is given by

$$\hat{\theta}_i = \frac{(\alpha_i + y_i)}{(\alpha_i + y_i) + (\beta_i + n_i - y_i)}. \tag{1}$$

Using equation (1), you can calculate the posterior mean for both proportions θ_{PE} and θ_{PC}. Then you can calculate the relative risk with

$$RR = \frac{\theta_{\text{PC}}}{\theta_{\text{PE}}}. \tag{2}$$

Compare these analytical results with the results obtained using RJAGS. Are they similar?

H. Evaluating historical data.

Suppose that, on an international conference on PTSD, you meet two fellow researchers who also evaluated the use of PE versus PC. Jessica recently executed a randomized clinical trial with 520 *female* veterans to evaluate the use of PE versus PC. Ronald did a comparable trial with 235 WWII *male* veterans in 1946. After the conference, you look up their articles to study the findings. The results from the two studies are given in Tables 2 and 3.

Table 2: PTSD data from Jessica

	Type of Intervention	
	PE	PC
Loss of Diagnosis	120	80
Total Women Treated	245	275

Table 3: PTSD data from Ronald

	Type of Intervention	
	PE	PC
Loss of Diagnosis	40	45
Total Men Treated	105	130

As discussed during the lecture on informative prior specification, it can be worthwhile to make use of data obtained in previous studies in the analysis of new data. Would you be interested in including the results obtained in either of the trials in the prior distribution for the analysis of your own data? How relevant do you think the two datasets are, is one more relevant than the other, and why?

I. Specifying informative priors.

Your goal is to re-run the analysis with informative prior distributions based on the relevant historical datasets (you can pick one or both). For this purpose, which hyperparameters α and β do you want to use for the two groups? Hint: When the beta distribution is used as a prior distribution for a probability parameter, you can think of the hyperparameters α and β as the previously observed number of successes+1 and the previously observed number of failures+1, respectively.

J. Assessing prior influence.

Run a new analysis with the informative prior distributions (follow the same steps, starting with checking the new model file). Look at the estimates for both proportions θ_{PE} and θ_{PC} and for the Relative Risk. Compare them with the estimates obtained with the uninformative priors. To what degree are the results influenced by the informative priors? Do you think this is a desirable effect?

K. Assessing prior influence, part 2 (optional).

Rerun the analysis with informative prior distributions based on the data that you deemed less relevant (or based on a different weighting of the two data sets). Then compare the results from all three of your analyses. What happens when you include conflicting data, and what happens when your prior is based on more data than available in the current study?

L. Analytical results with informative priors (optional).

Can you obtain the results from part J (or K) analytically? Look at the formulas from exercise G. Are the results similar?