Exercise A

A gardener caught in the midst of the tulip mania of the Dutch Golden Age bought 100 bags of tulip bulbs, each containing more or less 30 bulbs. He planted them in his garden in 100 rows, using a different bag for each row. Come springtime, he expected to see around 80% of the bulbs grow into tulips (as this is the expected fertility rate for high quality bulbs).

However, some of the rows contained much fewer tulips than he predicted. He later hears a rumor at the market that the bulb seller filled some of the bags with bulbs of lesser quality. For simplicity, let's assume that each bag is filled with bulbs of the same quality and that the gardener counts the number of sprouted tulips without error.

Your main task is to determine if the gardener was cheated by the bulb supplier.

You can find the data in tulips.txt. It contains a vector of size 100 in which each element corresponds to the number of tulips that sprouted in a specific row (out of roughly 30). You can read in the data in R with the read.table R function:

```
read.table('your_path/tulips.txt', sep="\t")
```

Assumptions for the tasks:

1. There are only two types of bulbs, and each bag only contained exclusively bulbs from one type. Let's say A bags contain quality bulbs and B bags contain lesser quality bulbs.
2. The number of tulip bulbs in each bag (and therefore in each row) can be modeled by a normal distribution with standard deviation 1.5. Use a rounded Gaussian since the number of tulips in a bag must be an integer. The way to do this is to use the `round` function.
3. We believe that bulbs in A bags have approximately 80% fertility rate. This can be modeled by beta(40,10). We don't have any specific prior assumption on the fertility rate of the bulbs from B bags.

Tasks:

1. Based on the gardener's data, we would like to know whether the gossip is true. Specifically, we would like to know what is the ratio of A and B bags among the 100 bags purchased. Further, we should estimate what percentage of bulbs sprouted tulips in A and B bags.

   a) Construct a model to answer the questions above. Please denote the fertility/sprouting rates (for A and B bags) with p[1] and p[2], and the proportion of A bags with theta. We don't have any prior beliefs about the value of theta (so choose an appropriate prior function that does not favor any value of theta).
   Please submit the following plots and corresponding explanations:

   - The histogram of the samples of theta values with the 95% HDI. What do you conclude based on the estimated theta value? How confident are you in the estimate?
   - Three histograms (with a marked 95% HDI) corresponding to the posteriors of p[1] , p[2] , and their difference, p[1]-p[2]. (You can use plotPost.R as in the previous assignment). Again, explain your conclusions.

b) Repeat the task with the assumption that the actual value of theta is close to 0.5.

Use beta(15,15) as a prior for theta. Report if/how your conclusions have changed.

c) Perform a model comparison, and decide whether the assumption about theta in task a) or in task b) is more suitable. Let the program jump from one model to the other. Plot which model index (1 or 2) was selected on each iteration (or report the % of model visits). Also include the histograms of p[1], p[2] and theta posterior samples.

2. Extra credit task on power analysis:

Write a piece of code/function which generates datasets like the one you were given.

a) Simulate the case when the true values of p[1] and p[2] are 0.8 and 0.5, and the theta is 0.3. The number of bulbs in the bags is still coming from a Gaussian with mean 30 and standard deviation 1.5. With the help of this simulation, estimate the number of bags that you need to buy, if you want to show a significant difference between p[1] and p[2] (meaning their HDIs do not overlap) with 0.8 power.

- Run 10 simulations to estimate the power (at each sample size = number of bags).

  Use a dbeta(1,1) prior for theta and the true distribution for the number of bulbs in a bag (Gaussian with mean 30 and sd = 1.5). You're welcome to run more simulations, but then you will need to let the program run for longer.

- Your initial number of bags should be 4 and you should increment this value by 1 at each iteration until you reach .8 power

- Please be mindful of the initial values you provide for your parameters because otherwise the sampler might fail to initialize in the middle of your simulations and you'll have to start over (alternatively, more experienced R users can choose to use the tryCatch function to allow the program to procced with execution after it encounters an error).

b) Repeat task a) but first set the p[2] value in the generator to 0.7 and keep the p[1] value equal to 0.8.

- First increment the number of bags by 10 (so do a greedy estimation) until you reach the .8 power. Let's call the value where the iteration stops N*. Then, start again the iteration from N=(N*-10) but now increment the value of N by 1 at each step to find a better approximation to the bag number.

Please submit the code alongside a short summary of your results.

Exercise B

The caschool.csv file contains a random sample of California elementary school districts. The data consists of test scores (Y: testscr) and class sizes (X: stratio). The test score is a district-wide average of reading and math scores on the Stanford achievement test, a test utilized by school districts in the USA. The student-teacher ratio, i.e. the total number of students in the district divided by the number of teachers, is used as a measure of the (overall) class size in the district.
Policy makers collected this data to determine if reducing class size, for instance by hiring more teachers, improves education quality. Skeptics worry that reducing class size will increase costs without producing substantial benefits.

We are going to study the relationship between these two variables by fitting a simple linear model, specifically, for each observation $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$, where the error ε is iid Gaussian with mean zero and standard deviation σ.

You can read in the data by using the read.table function:

```
data = read.table('your path/caschools.csv', header = TRUE,  sep = ',')
```

You can explore it but running:

```
str(data)
summary(data)
```

And you can reference the variables of interest:

```
data$testscr
data$stratio
```

If you want, you are free to construct a hierarchical model using the other variables provided, but this is not required.

Use a vague Gaussian prior for the betas, and gamma for the inverse variance. You can use the classroom regression code as a starting point.

Tasks:

1. Report your estimates for the regression coefficient (β_1), the intercept (β_0), and the error standard deviation (σ) alongside their HDIs. Plot the histograms of their posteriors.

2. What should the researchers conclude? Does the student to teacher ratio have a significant effect on test scores?

3. What is the predicted distribution of test scores given your parameter posteriors for a student to teacher ratio of 20?

4. Extra credit task on illustrating your results:

   Create the following plots:
   - The independent variable on the X axis (selected values of student-teacher ratios) and your mean posterior predictions for the dependent variable (with 95% HDI) on the Y axis.
     - You can overlay your actual data points to visualize if there are any discrepancies with the linear-model predictions (are there outliers? non-linear trends?)
   - The credible regression slopes overlaid on observations (there are a lot of observations, so you can randomly select a subset if you think your plot looks too busy)