Bayesian Data Science: Assignment 3

Instructions:

- Please complete all questions and parts.
- Please show all relevant working and use R for all statistical programming and analysis. This assignment will cover more of part 2 of the course: Chapters 6-13 "*Doing Bayesian Data Analysis*", second edition by JK. The fourth assignment will cover part 3.
- Note that we will not be covering chapter 14 since STAN is still in its infancy. We will use it in later courses. Of course, you are free to read and understand the chapter.
- This assignment will be graded on 100 points. This second part of the course is very important because it …

… covers all the crucial ideas of modern Bayesian data analysis while using the simplest possible type of data, namely dichotomous data such as agree/disagree, remember/forget, male/female, etc. Because the data are so simplistic, the focus can be on Bayesian techniques. In particular, the modern techniques of "Markov chain Monte Carlo" (MCMC) are explained thoroughly and intuitively. Because the data are kept simple in this part of the book, intuitions about the meaning of hierarchical models can be developed in glorious graphic detail. This second part of the book also explores methods for planning how much data will be needed to achieve a desired degree of precision in the conclusions, broadly known as "power analysis."

- Use "R Mark down" to construct your assignment answers. Use appropriate Latex formulae (https://en.wikibooks.org/wiki/LaTeX/Mathematics) and r code chunks.
    - Please submit your Rmd document, html and pdf knit documents.
    - See http://rmarkdown.rstudio.com/index.html for help with rmarkdown
    - Make sure all of the files pertaining to the assignment are in the same directory.
    - I will run your Rmd document and check that your code works.
    - The functions you make must have arguments suitable to solve the problem and produce the desired output. ***It is up to YOU to determine what arguments are needed***.
    - You may use base R functions and/or other packages like ggplot2
    - ***Make sure your name is on the titles of all plots you make.***
    - ***You may consult google, youtube etc for help on R,ggplot2, JAGS etc***
    - All plots should be large (at least ¼ a page)
- This assignment assumes that you have already installed the following software (all of which are free)
    - R
    - R Studio
    - Latex (best to have a full distribution)
    - Jags
- Your finished assignment should be readable and intelligible.
- ***You are expected to complete ALL questions and parts!!!***
    - ***For those who did not do them all please give a list of all questions and parts you DID NOT DO!! This will be the last thing you do before uploading the documents.***
    - ***If you did all questions and parts please say: "All questions and parts done completely"***
    - ***Put this statement at the top of the document.***

Questions:

1. Consider the high-level script, Jags-Ydich-XnomSsubj-MbernBeta-Example.R. For this exercise you will use that script with a new data file, and notice that you only need to change a single line, namely the one that loads the data file. Use the following R commands to create a csv file in the working directory:

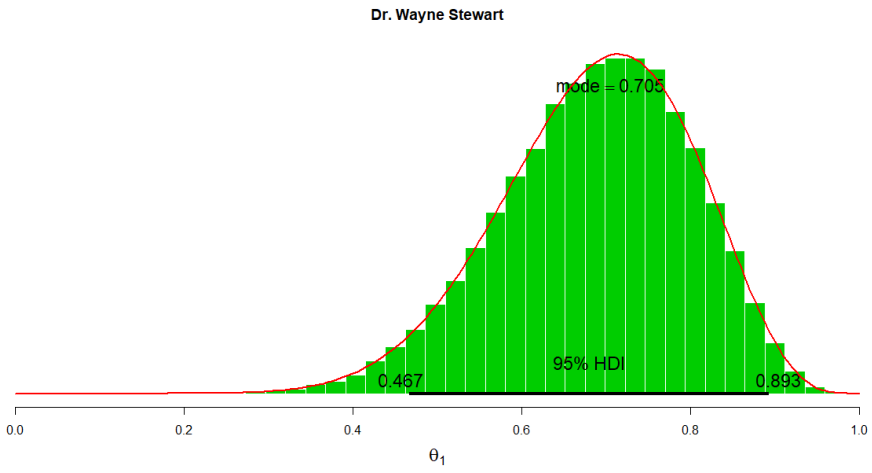
   ```
   y = c( rep(1,9),rep(0,3) , rep(1,45),rep(0,15) , rep(1,3),rep(0,9) )
   s = c( rep("A",12) , rep("B",60) , rep("C",12) )
   write.csv( data.frame(y=y,s=s) , file="Ass3.1.csv" , row.names=FALSE )
   ```

   Then, in the script, use the above file name in the read.csv command. In your report:
   a. include the head of the data file
   b. include the graphical output of the analysis
   c. Are the estimates reasonable? Say why or why not!
   d. What is the effect of different sample sizes for the estimates of different subjects?

Next, our goal is to superimpose an analytically correct beta distribution curve on the MCMC histogram produced by JAGS. After the end of the script used above, run the R code below, with the boxes □ replaced with the appropriate values:

```
openGraph()
plotPost( mcmcCoda[,"theta[1]"] , xlim=c(0,1) )
a = □ ; b = □ # constants from prior in Jags-Ydich-XnomSsubj-MbernBeta.R
H = □ ; T = □ # heads and tails from your data
thetaGrid=seq(0,1,length=201)
lines( thetaGrid , dbeta( thetaGrid , a+H , b+T ) )
```

   e. Explain how you determined the a and b values.
   f. Explain the meaning of the lines() command, being sure to relate it to Equation 6.8.
   g. Include the resulting graph in your write-up. (The superimposed curve should closely match the histogram.)



h. Examine the utility file "DBDA2E-utilities.R" to see what options you can utilize to change or better present the plot (titles etc) notice the … ellipsis. For the plotPost function:
    i. What option – value pair did I use to obtain the label on the x axis above?
    ii. What option – value pair did I use to make the colour green?
    iii. What option -value pair did I use to make the title?
i. For the lines() function – what other options would produce the curve shown above?

2. In this exercise, you'll use R to compute the likelihood value in Equation 9.10, p. 247. Consider four coins, each flipped 4 times, with these outcomes:
$\{y_{i|s=1}\}$=1,0,0,0 $\{y_{i|s=2}\}$=1,1,0,0 $\{y_{i|s=3}\}$=1,1,0,0 $\{y_{i|s=4}\}$=1,1,1,0
    a. Write down equation 9.10 p. 247 using Latex and explain the notation.
    b. What are the proportions of heads for each coin?
    c. What is the value of the likelihood when ω=0.5, κ=2, θ_1=0.25, θ_2=0.50, θ_3=0.50, and θ_4=0.75?
    d. Do these parameter values constitute any shrinkage relative to the data proportions?
    e. Is the shape of the beta distribution flat or peaked?
    f. What is the value of the likelihood when ω=0.5, κ=20, θ_1=0.35, θ_2=0.50, θ_3=0.50, and θ_4=0.65?
    g. Do these parameter values constitute any shrinkage relative to the data proportions?
    h. Is the shape of the beta distribution flat or peaked?
    i. Which set of parameter values, these or the previous part, yield a higher likelihood value for the data?
    j. What does shrinkage do to likelihood?

3. We have a six-sided die, and we want to know whether the probability that the six-dotted face comes up is fair. Thus, we are considering two possible outcomes: six-dots or not six-dots. If the die is fair, the probability of the six-dotted face is 1/6. Suppose we roll the die N = 45 times, intending to stop at that number of rolls. Suppose we get 3 six-dot rolls.
    a. What is the two-tailed p value?

    Hints: Use Equation 11.5 (p. 303) to compute the tail probability of the binomial sampling distribution in R. R has various relevant functions built in, such as factorial(), choose(), and even dbinom() . To maintain correspondence with Equation 11.5, I will not use dbinom(). Try this script:

```
N = 45 ; z = 3 ; theta = 1/6
lowTailZ = 0:z
sum( choose(N,lowTailZ) * thetaˆlowTailZ * (1-theta)ˆ(N-lowTailZ) )
```

    b. Explain carefully what each line of the script does.
    c. Why does it consider the low tail and not the high tail?
    d. Explain the meaning of the final result.
    e. Suppose that instead of stopping at fixed N, we stop when we get 3 six-dot outcomes. It takes 45 rolls. (Notice this is the same result as the previous part.) What is the two-tailed p value?

    Hint: Use Equation 11.6 (p. 306). Try this:

```
sum( (lowTailZ/N) * choose(N,lowTailZ) * thetaˆlowTailZ * (1-theta) ˆ(N-lowTailZ) )
```

    f. Explain carefully what that code does and what its result means.

4. We continue with the scenario of the previous question: A dichotomous outcome, with N = 45 and z = 3.
    a. If the intention is to stop when N = 45, what is the 95% CI?

Hints: Try this continuation of the R script from the previous question:

```
for ( theta in seq( 0.170 , 0.190 , 0.001) ) {
show( c(
theta ,
2*sum( choose(N,lowTailZ) * thetaˆlowTailZ * (1-theta)ˆ(N-lowTailZ) )
))
}

highTailZ = z:N
for ( theta in seq( 0.005 , 0.020 , 0.001) ) {
show( c(
theta ,
2*sum( choose(N,highTailZ) * thetaˆhighTailZ * (1-theta)ˆ(N-highTailZ) )
))
}
```

    b. Explain carefully what the code does and what it means!
    c. If the intention is to stop when z = 3, what is the 95% CI?
    d. Is the CI the same as for stopping when N = 45? Hint: Modify the R script of the previous part for use with stopping at z, like the second part of the previous question.

5. The script called "Jags-Ydich-XnomSsubj-MbinomBetaOmegaKappa-Power.R" carries out an example of this analysis. It is used in the problem relating to "therapeutic touch" where practitioners were tested for the ability to sense the presence of the experimenter's hand near to their own. The following picture shows the flow of information in a power analysis. It is important that you master this picture and understand the methodology and code that performs the analysis.

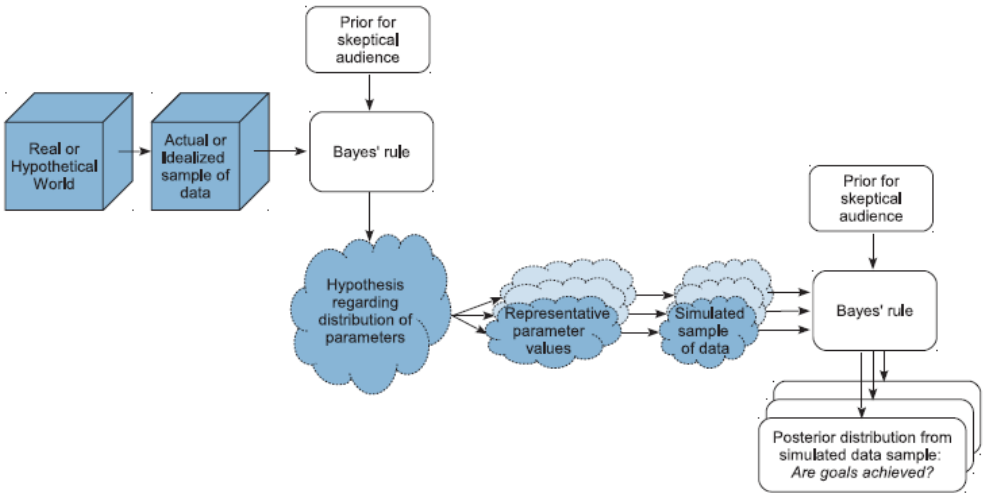


Figure 13.2 Flow of information in a power analysis when the hypothesis regarding the distribution of parameters is a posterior distribution from a Bayesian analysis on real or idealized previous data. Compare with Figure 13.1, p. 363.

a. Explain the method of power analysis by interpreting each part of the above picture (Fig. 13.2)

The following portions of the script mentioned above will now be given and you will need to explain what they do. Do NOT simply reproduce # comments. You MUST understand the method first and then the meaning of the code :

b. Interpret each of the following lines of R code:

```
source("Jags-Ydich-XnomSsubj-MbinomBetaOmegaKappa.R")

idealGroupMean = 0.65
idealGroupSD = 0.07
idealNsubj = 100
idealNtrlPerSubj = 100
```

c. Interpret each of the following lines of R code:

```
betaAB = betaABfromMeanSD( idealGroupMean , idealGroupSD )
theta = rbeta( idealNsubj , betaAB$a , betaAB$b )
```

d. Interpret each of the following lines of R code:

```
theta = ((theta-mean(theta))/sd(theta))*idealGroupSD + idealGroupMean
theta[ theta >= 0.999 ] = 0.999
theta[ theta <= 0.001 ] = 0.001
z = round( theta*idealNtrlPerSubj )
```

e. Interpret each of the following lines of R code:

```
dataMat=matrix(0,ncol=2,nrow=0,dimnames=list(NULL,c("y","s")))
for ( sIdx in 1:idealNsubj ) {
  yVec = c(rep(1,z[sIdx]),rep(0,idealNtrlPerSubj-z[sIdx]))
  dataMat = rbind( dataMat , cbind( yVec , rep(sIdx,idealNtrlPerSubj) ) )
}
```

f. Interpret each of the following lines of R code:

```
idealDatFrm = data.frame(dataMat)

mcmcCoda = genMCMC( data=idealDatFrm , saveName=NULL ,
          numSavedSteps=2000 , thinSteps=20 )
mcmcMat = as.matrix(mcmcCoda)
```

g. Run the script using nSimulatedDataSets of 50. Show which line of code you changed to accomplish this. Report the final power estimates. How do your results compare with the results shown in Section 13.2.5? *Hint:* The power estimates should be about the same, but because you used a smaller number of simulated data sets, the bounds on your power estimate should be wider (less certain).

h. Now you will run the power simulation starting with idealized data that mimic the actual data. Refer to the posterior distribution from the analysis of the actual data in Figure 9.10, p. 243. Notice the central tendency and HDI on the group-level mode. We will use those characteristics for the idealized data generating hypothesis. Specifically, near the beginning of the script, set idealGroupMean = 0.44, idealGroupSD = 0.04, idealNsubj = 28, and idealNtrlPerSubj = 10. Explain what each of those settings does and explain why those values were chosen.

i. Because the idealized data have central tendency near chance performance, we cannot have high hopes for rejecting the null, and therefore our goal might be high precision. In the function goalAchievedForSample, set the HDImaxwid to 0.1. Also, for high precision, we will need more data than was obtained in the original experiment, so try setting NSubj to 40 and NtrlPerSubj to 100. Because this is an exercise, not real research, change the number of simulated data sets to only 20. Report the lines of code you changed (and any you deleted or commented out). Now run the simulation and report the final estimated power for each of the goals. Why does the goal omegaNarrowHDI have high power but the goal thetaNarrowHDI have low power?

j. For those who want a simple programming exercise in R, try this: Instead of using idealized data to create hypothetical data-generating parameter values, use the actual data from the original experiment. In the first part of the script, just comment out or delete the lines that create idealized data. Instead, use the actual data in the genMCMC function. Then repeat the previous part. Are the power estimates about the same?

6. In this problem I want you to create a discrete ***two state*** sampler in R by answering the following questions:
    a. Make a proposal function – this will propose a "0.3" or "0.6", where the "0.3" corresponds to a head and "0.6" the tail. Make the head proposed with probability "p". Call the function myprop(). Hint: You may wish to use the R function sample() – essentially you are mimicking a number "n" tosses of a biased coin but instead of a "H" you get a "0.3" and instead of a "T" you get a "0.6".
    b. Give the outcome of
        i. myprop(n=10,p=0.5)
        ii. myprop(n=20,p=0.1)
        iii. myprop(n=100, p=0.4)
    c. Make a barplot of the output of myprop(10000,0.3) with two colors and title with your name on it.
    d. Now suppose that an experiment results in x = 4 successes after N=10 trials (the trials were fixed from the beginning of the experiment). What is the likelihood equation $f(x|\theta)$(Latex) where θ is the probability of a success.
    e. Suppose that we use a Beta prior of the form $\theta \sim Beta(2,2)$ and we define $h(\theta) = p(\theta)f(x|\theta)$. Give the equation for $h(\theta)$ (Latex)
    f. What is the acceptance equation you will use here? (Latex) Hint: You will use $h(\theta)$
    g. Now make a function that will create an MCMC sample from the posterior using the proposal made above. Call the function mymcmc(), it should do the following:
        i. Create a single plot interface divided with width ratio 1:1:2 so that a barplot of the proposal is on the left, a barplot of the posterior in the middle and an MCMC trace is on the right. The plot should be suitably labelled and colored with your name on the title.
        ii. The command-line output should be a list containing the proposals, accepted, rejected and the posterior samples.
    h. Call the function using a proposal with p=0.5, n = 50000 iterations obj = mymcmc(…). Give the head(obj) and show the plot.