Math 152 - Statistical Theory - Homework 2

write your name here

Due: Thursday, September 8, 11:59pm

Important Note:

You should work to turn in assignments that are clear, communicative, and concise. Part of what you need to do is not print pages and pages of output. Additionally, you should remove these exact sentences and the information about HW scoring below.

Click on the *Knit to PDF* icon at the top of R Studio to run the R code and create a PDF document simultaneously. [PDF will only work if either (1) you are using R on the network, or (2) you have LaTeX installed on your computer. Lightweight LaTeX installation here: https://yihui.name/tinytex/]

> Either use the college's RStudio server (https://rstudio.pomona.edu/) or install R and R Studio on to your personal computer. See: https://research.pomona.edu/johardin/math152f20/ for resources.

Assignment

Goals: In this assignment, the fun will include:

- practice using and creating prior distributions.
- practice using and creating posterior distributions.
- using R to visualize how different parameter values change the prior and posterior distributions.

Book problems

- Feel free to do the book problems with a pencil or in LaTeX (RMarkdown supports writing mathematics using LaTeX).

- If you use a pencil, you can take a picture of the problem(s), and include the image(s) using (remove the tick marks to make it work):

```
![](myimage.jpeg)
```

- Note that myimage.jpeg needs to live in the same folder as the relevant .Rmd file (maybe you called the folder "math 152 hw" and put it on your desktop?)
- Saving as jpg, jpeg, png, or pdf should work, but make sure to specify the exact name of the file.
- If you have the 3rd edition of the book, the problems will be the same unless they don't exist – that is, the 4th edition *added* problems but didn't change the order of them. Ask me if you want to see the 4th edition problems.

1: Community Q Describe one thing you learned (not **during** class) from a member of the class (student, mentor, professor) – it could be: content, logistical help, background material, R information, etc. 1-3 sentences.

2: 7.2.6 Suppose that the proportion θ of defective items in a large manufactured lot is unknown, and the prior distribution of θ is the uniform distribution on the interval [0, 1]. When eight items are selected at random from the lot, it is found that exactly three of them are defective. Determine the posterior distribution of θ.

3: 7.2.9 Consider again the problem described in Exercise 6, and assume the same prior distribution of θ. Suppose now, however, that instead of selecting a random sample of eight items from the lot, we perform the following experiment: Items from the lot are selected at random one by one until exactly three defectives have been found. If we find that we must select a total of eight items in this experiment, what is the posterior distribution of θ at the end of the experiment?

4: 7.2.10 Suppose that a single observation X is to be taken from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, the value of θ is unknown, and the prior distribution of θ is the uniform distribution on the interval [10, 20]. If the observed value of X is 12, what is the posterior distribution of θ?

5: 7.2.11 Consider again the conditions of Exercise 10, and assume the same prior distribution of θ. Suppose now, however, that six observations are selected at random from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, and their values are 11.0, 11.5, 11.7, 11.1, 11.4, and 10.9. Determine the posterior distribution of θ.

6: 7.3.7 Suppose that the heights of the individuals in a certain population have a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2 inches. Suppose also that the prior distribution of θ is a normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. If 10 people are selected at random from the population, and their average height is found to be 69.5 inches, what is the posterior distribution of θ?

7: 7.3.8 Consider again the problem described in Exercise 7.
a. Which interval 1-inch long had the highest prior probability of containing the value of θ?
b. Which interval 1-inch long has the highest posterior probability of containing the value of θ?
c. Find the values of the probabilities in parts (a) and (b).

8: 7.3.9 Suppose that a random sample of 20 observations is taken from a normal distribution for which the value of the mean θ is unknown and the variance is 1. After the sample values have been observed, it is found that $\bar{x} = 10$, and that the posterior distribution of θ is a normal distribution for which the mean is 8 and the variance is 1/25. What was the prior distribution of θ?

9: 7.3.11 Suppose that a random sample of 100 observations is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution. Show that no matter how large the standard deviation of the prior distribution is, the standard deviation of the posterior distribution will be less than 1/5.

10: R - beta-binomial family Consider the beta-binomial family (i.e., beta prior, binomial likelihood (with parameter θ), beta posterior).
That is, the parameter of interest is θ, and both the prior and posterior distributions of θ are from the beta family.

(a) Write down the posterior distribution of θ given the data as a function of prior α, prior β, n, and $\hat{p}$ = **proportion of successes**.

(b) How does the posterior expected value of θ change as a function of each of the values above?

(c) Using simulations, histograms, and means, **discuss the role of sample size** when using a prior and Bayesian inference. For the discussion:

i. give posterior histogram and sample means for the following combinations (12 histograms):

- (α, β): (4,4); (4,10) [these are the prior values]
- $\hat{p}$: 0.2, 0.5

- n: 10, 100, 1000

ii. Using your histograms and means above, discuss the role of sample size in determining the posterior distribution of the parameter.

Some R code that might be helpful. If you do not understand the code, please ask the professor, the mentor, or your peers.

Note that I wrote "eval = FALSE". The FALSE allows the .Rmd file to compile with missing information (see the parts below that are blanked out: _____). Once you figure out the blanks, fill them in, and change the code chunk to "eval = TRUE" (you'll need to do this for all the chunks below).

```
library(tidyverse)

# below is a function that generates random beta data and then adds columns (`cbind`)
# to the output so that you can track the input values.

# after you fill in the missing parts to `rbeta`, run the function a few times (in the console)
# by typing:  postbeta_data(1,4,.3,47, 15)
# what does the `postbeta_data` function output?
postbeta_data <- function(a, b, phat, samp_size, reps){
  data <- data.frame(obs = rbeta(reps, ___ , ___)) %>%
    cbind(a = rep(a, reps), b= rep(b, reps),
          phat = rep(phat, reps), samp_size = rep(samp_size, reps))
}

# below I've created an object (called `params`) which contains all the
# different combinations we are interested in
# fill in the missing blanks
params <- list(
  a = c(___),
  b = c(___, ___),
  phat = c(___, ___),
  samp_size = c(___, ___, ___),
  reps = c(___)
)

# the following lines run the `postbeta_data` function for each of the parameter combinations
# don't worry about the exact syntax, but do run the code and **look** at the object called `sim_beta`.
# Describe the object `sim_beta` to yourself:  how many rows does it have? how many columns?
#        what are the columns?  what are the rows?
sim_beta <- params %>%
  cross() %>%
  map_dfr(lift(postbeta_data))
```

Plotting

We will use `ggplot()` to plot the data. Note that the main column of interest (in the data object called `sim_beta`) is called `obs`. The pipe function (`%>%`) is used in data wrangling to chain commands. The plus function (`+`) is used only in plotting with `ggplot` to add layers of graphics. In the code below I've given the English equivalent of what is happening with the code, please ask if you don't follow.

```
sim_beta %>%  # the dataset sim_beta is sent to
  ggplot(aes(x = obs)) + # the plot function called ggplot where the x-axis is defined as the obs varia
  geom_histogram() +  # the plotting activity we want is a histogram
  # broken down ("faceted") with samp_size on the x part of the grid and
  # both p-hat and b on the y part of the grid
  facet_grid(samp_size ~ phat + b)
```

Summary statistics

Using the simulated data, we can find the mean and standard deviation of each of the sets of random observations. Note that before calculating the mean and the standard deviation, the observations are **grouped** by the parameter variables. Also note that the series of steps are chained using the pipe (`%>%`) operator.

```
sim_beta %>%
  group_by(b, phat, samp_size) %>%
  summarize(mean = mean(obs), sd = sd(obs))
```