BST413 spring 2023: Homework 1 (due Tues. Jan. 31, 2023)

You are welcome to discuss the homework with others, but please write your own programs and write up the answers individually.

Please upload your homework **and** R or Rmarkdown code to Blackboard. I'll let you know later how to hand in paper and pencil questions.

1. **Short questions:**
   (a) About how long did you spend on the homework?
   (b) Was the homework length "reasonable", "too time-consuming", or "too short"?
   (c) Who did you work with on this homework?

2. **Identical twin**: A female family member, "Melissa", has a twin sister. Approximately 1/125 of all births are fraternal twins, and approximately 1/300 of all births are identical twins, and we will assume that 1/2 of all births are girls. What is the probability that "Melissa" is an identical twin? NOTE: be sure to define some events, such as G=girl, I=identical twin, etc, then think carefully about what"data" you need to condition on. We can take for granted that Melissa has a sibling.

3. **Monty Hall**: Gelman chapter 1, problem 7. Please do this by explicitly defining all key events and carrying out the calculations using Bayes rule (no shortcut answers!).

4. **Sensitivity and specificity in R**: We will examine 2 by 2 tables of the following form, where $\theta = 1$ denotes true disease and $y = 1$ denotes the test is positive for the disease. (As per the table, $\theta = 0$ or 1 and $y = 0$ or 1.)

	$y = 1$	$y = 0$
$\theta = 1$		
$\theta = 0$		

   (a) Suppose $P(\theta = 1) = 0.001$, $P(y = 1 \mid \theta = 1) = 0.95$, and $P(y = 1 \mid \theta = 0) = 0.05$. Show your hand calculation of $P(\theta = 1 \mid y = 1)$. (As a check, you should get 0.0187).
   (b) Write an R function that takes 4 arguments: $P(\theta = 1)$, $P(y = 1 \mid \theta = 1)$, $P(y = 1 \mid \theta = 0)$, and y, and returns $P(\theta = 1 \mid y)$.
   (c) Call your R function to calculate $P(\theta = 1 \mid y = 1)$ under the scenario you considered above..
   (d) Explain in terms a non-statistician might understand what $P(y = 1 \mid \theta = 1) = 0.95$ and $P(\theta = 1 \mid y = 1) = 0.0187$ mean, and why these numbers are not contradictory.

(e) Call your R function again to calculate $P(\theta = 1 \mid y = 1)$ under the following three conditions:

   i. $P(\theta = 1) = 0.001$, $P(y = 1 \mid \theta = 1) = 0.80$, $P(y = 1 \mid \theta = 0) = 0.05$
   ii. $P(\theta = 1) = 0.01$, $P(y = 1 \mid \theta = 1) = 0.95$, and $P(y = 1 \mid \theta = 0) = 0.05$
   iii. $P(\theta = 1) = 0.01$, $P(y = 1 \mid \theta = 1) = 0.80$, and $P(y = 1 \mid \theta = 0) = 0.05$

(f) Compare the four sets of results in some sensible way (e.g. some pairs of results may make particular sense to compare, so be sure to talk about them).

5. **M&M data**: I have entered the M&M data we collected in class, where the last two rows are data from the two TAs. Download the data from blackboard (the data is "mmdark2023.txt") and read it into R. You can input the data with
`mmdark.dat = read.table("mmdark2023.txt",header=TRUE)`

   Your data should contain 7 rows, where each row corresponds to one person. Note that you can sum over the rows or columns by using either the apply command, or the sum command (for example `sum(mmdark.dat[1,])` sums over the first row, and `apply(mmdark.dat,2,sum)` will give column sums). For the moment we will use information only about "blue" or "not blue", i.e. you will want to sum over the non-blue columns to create "not blue" counts.

   (a) In the data from class (including from the TAs), how many blue and how many non-blue M&Ms were there? For the rest of the question we will use Y to denote the number of blue M&Ms out of n total in our class data. We will model Y as $Y \sim \text{Binomial}(n, \theta)$.

   (b) In 2017 I collected data from dark M&M's, and had 5 blue, and 27 non-blue M&Ms. Use this data as the basis of an informative Beta prior (do not downweight this information). What are the hyperparameters of your prior?

   (c) Using your informative prior, obtain $p(\theta \mid Y)$.

   (d) Now read in the entire data on dark M&M's from the 2017 class, which is in "mmdark2017.txt", and use this data as an informative prior. What values of α and β would you use for this Beta prior, assuming you don't downweight the data?

   (e) Obtain $p(\theta \mid Y)$ using the prior in the last part. Comment on whether or not it is reasonable to use this 2017 data as an informative prior.

   (f) Using all the data, obtain a 95% CI for θ using (i) the normal approximation, and (ii) a random sample from the beta posterior.

   (g) Make a plot, similar to the class notes, to show the likelihood, prior and posterior for both priors we considered above. Comment on the relative informativeness of the prior versus the likelihood in the two cases. There is no need to do any formal test here; a sentence or two will suffice.

Question 6:

Posteriors etc for some 1-parameter families w/ conjugate priors: Part I

Please fill in the blanks in this sheet, AND SHOW YOUR DERIVATIONS (on a separate piece of paper) for all parts. By "blanks" I mean the posterior distributions, and in some cases the marginal distributions (and for one, the posterior predictive distribution). For the posterior distribution of the normal distribution with known variance, I would like to see your calculations the long way, e.g. completing the square.

Normal with known variance

Sampling model: $p(y_i \mid \theta) \sim N(\theta, \sigma^2)$, σ^2 known, for $i = 1, \cdots, n$.

Conjugate prior: $p(\theta) \sim N(\theta_0, \tau^2)$

Posterior: $p(\theta \mid y_1, \cdots, y_n)$

Marginal: (we will skip this)

Posterior predictive: $p(\tilde{y} \mid y_1, \cdots, y_n)$

Binomial

Sampling model: $p(y \mid \theta) \sim \text{Binomial}(n, \theta)$

Conjugate prior: $p(\theta) \sim \text{Beta}(\alpha, \beta)$

Posterior: $p(\theta \mid y)$

Marginal: $p(y)$