Take-Home Problems

Instructions.

- These problems are due via Gradescope on **on Wednesday, March 10, 2021**.
- You may use your notes and the class text, but you may not consult other resources or other people. Send me an email if you have questions or difficulties of any sort.
- Prepare your solutions using R Markdown and turn in a printed copy. Also email me your Rmd file.

1 A number of years ago I gave students a survey on which was one of the following two questions.

- Version 1:

  Suppose that you have decided to see a play for which the admission charge is $20 per ticket. As you prepare to purchase the ticket, you discover that you have lost a $20 bill. Would you still pay $20 for a ticket to see the play?

  Yes No

- Version 2:

  Suppose that you have decided to see a play for which the admission charge is $20 per ticket. As you prepare to enter the theater, you discover that you have lost your ticket. Would you pay $20 to buy a new ticket to see the play?

  Yes No

Here are the results.

```
        version
response  v1  v2
     no   61  69
     yes 103  44
```

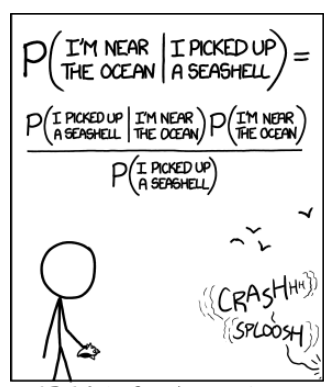
In this problem you will use the grid method to fit and interpret a model that has two parameters:

- p_1: the proportion of college students who say they will buy a ticket if they are given version 1.
- p_2: the proportion of college students who say they will buy a ticket if they are given version 2.

a) What is the likelihood when $p_1 = 0.5$ and $p_2 = 0.5$?

b) What is the likelihood when $p_1 = 0.4$ and $p_2 = 0.6$?

c) What is the likelihood when $p_1 = 0.6$ and $p_2 = 0.4$?



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

d) Use the grid method to fit the model. Use a uniform prior for each parameter.

Hints:

i. You might start with a course grid until you get everything working and then use a finer grid for your final work.

ii. The previous 3 items are intended to help you determine how you will compute your likelihood. Check your grid to make sure those values match.

e) Provide a plot of the posterior distribution for each parameter. What do these plots tell you about the model?

f) Use posterior sampling to estimate p_1 and p_2. Provide a point estimate and a 92% credible interval for each parameter. (Why 92%? Because we can.)

g) Does this model suggest that the form of the question leads to different responses? Explain.

h) Now fit a model using a "triangular prior" for each parameter that peaks at 0.5 and falls off to 0 at each end of of the interval $[0, 1]$.

i) How do the results of this model compare to those of the previous model? Why?

j) Which model thinks there is a larger difference between the two versions of the question? Why?

2 This problem combines what you know about quantitative and categorical predictors in a single model. We will use the `iris` data set, which includes measurements of 150 iris plants, 50 each from three species. The length measurements are in cm.

```
data(iris)
```

a) Use `quap()` to fit a model that predicts sepal length from sepal width and species.

b) Discuss your choice of priors.

c) What does your model say about how length and width of sepals are related?

d) Does your model think that species matters?

e) Is this model a good fit to the data? Explain how you arrive at your conclusion. (Feel free to comment on both good and bad features of the fit.)

f) Suppose you measure a new iris plant of the *setosa* species and find that the sepal width is 2.8 cm. What does your model predict for the sepal length of this iris?

g) How precise is the estimate?

(Quanitfy this: I'm not looking for an answer like "pretty precise" or "not precise at all". But you have some flexibility about how you express the precision.)

h) Now fit a model that uses only the *setosa* data. You can get just the *setosa* data using

```
Setosa <- iris %>% filter(Species == "setosa")
```

How does this model compare with your previous model?