HW 1

Your name

Exercise 1

The French mathematician Pierre-Simon Laplace (1749-1827) was the first person to show definitively that the proportion of female births in the French population was less then 0.5, in the late 18th century, using a Bayesian analysis based on a uniform prior distribution.

Suppose you were doing a similar analysis but you had more definite prior beliefs about the ratio of male to female births. In particular, if θ represents the proportion of **female** births in a given population, you are willing to place a Beta(100,100) prior distribution on θ.

- Show that this prior implies that you are more than 95% sure that θ is between 0.4 and 0.6, although you are ambivalent as to whether it is greater or less than 0.5.
- Now you observe that out of a random sample of $1,000$ births, 511 are boys. What is your posterior probability that $\theta > 0.5$?

Exercise 2

Suppose we are modeling the egocentricity of a person as measured by their speech. A researcher theorizes that one can get an idea of a speaker's egocentricity by measuring the number of times they produce a first-person pronoun in a sentence per day (e.g., starting a sentence with "I", or using a word like "me", "my', etc., anywhere in a sentence).

The number of times x that such a pronoun is used can be modeled by a Poisson distribution.

We are going to first define a Gamma prior on the λ parameter (representing the rate of occurrences of the pronouns) in the Poisson distribution. We have prior data suggesting that the mean number of occurrences of a first person personal pronoun in a population is 75 and the variance is 185. Using the mean and variance, figure out the hyperparameters of the Gamma prior, i.e., find out what a and b will be in Gamma(a,b). Hint: a/b is the mean of a Gamma distribution, and a/b^2 is the variance. Once you have figured out the Gamma hyperparameters, plot the Gamma distribution representing the priors.

Now suppose that the observed number of pronoun utterances per day over four days is 200,87,99,121. What is the posterior distribution of the λ parameter? This distribution will be a Gamma distribution with updated hyperparameters a^* and b^*. Your task is to figure out what these updated values a^* and b^* are.

Now, suppose that you get additional data for *two* further days, and the number of produced utterances over the two days taken together is 300. [Hint: this means that the data 300 is from n=2 days].

i. Use the posterior you just obtained (Gamma(a^*,b^*)) as a prior for this new data, and write down the posterior distribution for λ as a Gamma(a^{**},b^{**}) distribution, where a^{**},b^{**} are the updated hyperparameters.
ii. Now, start with your original prior Gamma(a,b), and use all the data available (200,87,99,121,300) and derive the posterior distribution of λ, expressed as a Gamma distribution with some hyperparameters a^{***} and b^{***}. (Hint: think carefully about what n is. Is it 5 or 6?)

Is there any difference in the posterior between the above two analyses (i) and (ii)?

Bonus Exercise 3 (in case you finish early and/or are bored)

Here we apply Bayes' rule for the case of discrete events.

Suppose that 1 in 1000 people in a population is expected to get HIV. Suppose a test is administered on a suspected HIV case, where the test has a true positive rate (the proportion of positives that are actually HIV positive) of 95% and true negative rate (the proportion of negatives are actually HIV negative) 98%. Use Bayes' theorem to find out the probability that a patient testing positive actually has HIV.