Homework 2

PSTAT 115, Summer 2023

Due on August 30th, 2023 at 11:59 pm

1. Knowing someone who is transgender

A September 2016 Pew Research survey found that 30% of U.S. adults are aware that they know someone who is transgender. It is now the 2020s, and Sylvia believes that the current percent of people who know someone who is transgender, π, has increased to somewhere between 35% and 60%.

1a. (4pts) Identify and plot a Beta model that reflects Sylvia's prior ideas about π.

1b. (4pts) Sylvia wants to update her prior, so she randomly selects 200 US adults and 80 of them are aware that they know someone who is transgender. Specify and plot the posterior model for π?

1c. (5pts) What is the mean, mode, and standard deviation of the posterior model?

1d. (7pts) Describe how the prior and posterior Beta models compare. Hint: in class we showed a special way in which we can write the posterior mean in a Beta-Binomial model. How can this help? Check the lectures notes.

2. Sample survey

Suppose we are going to sample 100 individuals from a county (of size much larger than 100) and ask each sampled person whether they support policy Z or not. Let $Y_i = 1$ if person i in the sample supports the policy, and $Y_i = 0$ otherwise.

2a. (5pts) Assume $Y_1, \ldots, Y_{100}$ are, conditional on θ, iid binary random variables with expectation θ. Write down the joint distribution of $P(Y_1 = y_1, \ldots, Y_{100} = y_{100}) \mid \theta)$ in a compact form. Also write down the form of $P(\sum_{i=1}^{100} Y_i = y \mid \theta)$.

2b. (5pts) For the moment, suppose you believed that $\theta \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Given that the results of the survey were $\sum_{i=1}^{100} y_i = 57$, compute $P(\sum_{i=1}^{100} Y_i = 57 \mid \theta)$ for each of the 11 values of θ and plot these probabilities.

2c. (5pts) Now suppose you originally had no prior information to believe one of these θ-values over another, and so $P(\theta = 0.0) = P(\theta = 0.1) = \ldots = P(\theta = 1.0)$. Use Baye's rule to compute $p(\theta \mid \sum_{i=1}^{100} Y_i = 57)$ for each θ value. Make a plot of this posterior distribution as a function of θ.

2d. (5pts) Now suppose you allow θ to be any value in the interval $[0, 1]$. Using the uniform prior density for θ, so that $p(\theta) = 1$, plot the *kernel* of the posterior density $p(\theta) \times P(\sum_{i=1}^{100} Y_i = 57 \mid \theta)$ as a function of θ.

3. Soccer World cup

Let λ be the average number of goals scored in a Women's World Cup game. We'll analyze λ by the following a Y_i is the observed number of goals scored in a sample of World Cup games:

$$\begin{aligned} Y_i|\lambda &\overset{ind}{\sim} \text{Pois}(\lambda) \\ \lambda &\sim \text{Gamma}(1, 0.25) \end{aligned}$$

3a. (5pts) Plot and summarize our prior understanding of λ.

3b. (5pts) Why is the Poisson model a reasonable choice for our data Y_i?

3c. (5pts) The `wwc_2019_matches` data in the *fivethirtyeight* package includes the number of goals scored by the two teams in each 2019 Women's World Cup match. Define, plot, and discuss the total number of goals scored per game:

```r
library(fivethirtyeight)
```

```
## Warning: package 'fivethirtyeight' was built under R version 4.2.3
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```r
data("wwc_2019_matches")
wwc_2019_matches <- wwc_2019_matches %>%
  mutate(total_goals = score1 + score2)
```

3d. (5pts) Identify the posterior model, ie, what is the posterior distribution (including its parameters).

3e. (5pts) Plot the prior, likelihood and posterior for λ. Describe the evolution in your understanding of λ from the prior to the posterior.

4. A Mixture Prior for Heart Transplant Surgeries

A hospital in the United States wants to evaluate their success rate of heart transplant surgeries. We observe the number of deaths, y, in a number of heart transplant surgeries. Let $y \sim \text{Pois}(\nu\lambda)$ where λ is the rate of deaths/patient and ν is the exposure (total number of heart transplant patients). When measuring rare events with low rates, maximum likelihood estimation can be notoriously bad. We'll tak a Bayesian approach. To construct your prior distribution you talk to two experts. The first expert thinks that $p_1(\lambda)$ with a gamma$(3, 2000)$ density is a reasonable prior. The second expert thinks that $p_2(\lambda)$ with a gamma$(7, 1000)$ density is a reasonable prior distribution. You decide that each expert is equally credible so you combine their prior distributions into a mixture prior with equal weights: $p(\lambda) = 0.5 * p_1(\lambda) + 0.5 * p_2(\lambda)$

4a. (10pts) What does each expert think the mean rate is, *a priori*? Which expert is more confident about the value of λ a priori (i.e. before seeing any data)?

4b. (5pts) Plot the mixture prior distribution.

4c. (10pts) Suppose the hospital has $y = 8$ deaths with an exposure of $\nu = 1767$ surgeries performed. Write the posterior distribution up to a proportionality constant by multiplying the likelihood and the prior density. Plot this unnormalized posterior distribution and add a vertical line at the MLE. *Warning:* be very careful about what constitutes a proportionality constant in this example.

4e. (10pts) Let $K = \int L(\lambda; y)p(\lambda)d\lambda$ be the integral of the proportional posterior. Then the proper posterior density, i.e. a true density integrates to 1, can be expressed as $p(\lambda \mid y) = \frac{L(\lambda;y)p(\lambda)}{K}$. Compute this posterior density and clearly express the density as a mixture of two gamma distributions.