STAT 578 – Advanced Bayesian Modeling – Spring 2020

Assignment 1

1. The film review aggregator website `rottentomatoes.com` publishes ranked lists of movies based on the number of positive critical reviews out of a total number counted for each movie. See, for example, `https://www.rottentomatoes.com/top/bestofrt/?year=2019`. Because the site uses an "Adjusted Score," a movie with a higher approval percentage sometimes ranks lower on the list.

   Consider the following hypothetical scenario:

   Movie 1: 425 positive reviews out of 500 (85%)
   Movie 2: 9 positive reviews out of 10 (90%)

   Assume that reviews of Movie i are independent with a common probability p_i of being positive (depending on the movie). Assume a $\mathrm{U}(0,1)$ prior on each p_i.

   (a) [4 pts] Determine the posterior distribution of p_1 and of p_2 (separately). (Name the type of distribution and give the values of its defining constants.)

   (b) [3 pts] Which movie ranks higher according to posterior mean? According to posterior median? According to posterior mode? Show your computations. (For median, use R function `qbeta`. For mean and mode, use formulas in BDA3, Table A.1. Do *not* use simulation, as it may not be sufficiently accurate.)

2. File `randomwikipedia.txt` contains the ID number and number of `bytes` in length for 20 randomly selected English Wikipedia articles.

   (a) (i) [2 pts] Draw a histogram of article length, and describe the distribution.

   (ii) [2 pts] Transform article length to the (natural) log scale. Then re-draw the histogram and describe the distribution.

   (iii) [1 pt] Based on your histograms, explain why the log scale would be better to use for the remainder of the analysis. (Read below.)

   (b) [2 pts] Let y_i be length of article i on the *log* scale (i.e., the natural logarithm of the number of bytes). Compute the sample mean and sample variance of $y_1, \ldots, y_{20}$.

   In the remaining parts, assume the y_is have a normal sampling distribution with mean μ and variance σ^2.

   (c) Assume σ^2 is <u>known</u> to equal the sample variance. Consider a flat prior for μ. Use it to:

   (i) [3 pts] Compute the posterior mean, posterior variance, and posterior precision of μ.

   (ii) [2 pts] Plot the prior density and the posterior density of μ together in a single plot. Label which is which.

   (iii) [2 pts] Compute a 95% central posterior interval for μ.

   (d) Now let μ and σ^2 have prior

$$p(\mu, \sigma^2) \propto \left(\sigma^2\right)^{-1} \qquad \sigma^2 > 0$$

Use it to:

(i) [3 pts] Compute the posterior mean, posterior variance, and posterior precision of μ. (If you cannot compute explicitly, use a good computational approximation.)

(ii) [2 pts] Approximate a 95% central posterior interval for μ.

(iii) [2 pts] Approximate a 95% central posterior interval for σ^2.

(e) Assume the prior of the previous part. Use simulation in R to answer the following, based on 1,000,000 draws from the posterior.

(i) [2 pts] Approximate a 95% central posterior predictive interval for the length (in bytes) of a single (new) randomly selected article. (Note that this is on the *original* scale, not the log scale.)

(ii) [2 pts] Approximate the posterior predictive probability that the length of a single (new) randomly selected article will exceed the maximum article length in the data.

(iii) [2 pts] Approximate the posterior predictive probability that the maximum length of 20 (new) randomly selected articles will exceed the maximum article length in the data. (Be careful! All 20 randomly selected articles have the *same* value for μ and for σ^2.)

Reminder: Show the R code you used and also a summary of the approximate inference results that you used to answer the preceding parts.

Total: 34 pts