Computational Statistics Assignment 1

by James D. Wilson (University of San Francisco)

1. Revisit the change point problem with text data in Chapter 1 of the *Probabilistic Programming and Bayesian Methods* book. An important aspect of Bayesian analysis is properly choosing a prior distribution, which (often) involves the choice of hyperparameters. Notice that below equation (13), the author uses the identity to provide a reasonable value for the hyperparameter α:
$$\alpha = \left(\frac{1}{N}\sum_{i=0}^{N} C_i\right)^{-1}.$$
Here α is the rate of texts before *and* after the change as seen by equations (12) and (13). This choice is reasonable, but I want you to investigate what happens to your posterior inference when you change this hyperparameter and hence the prior distribution. For each scenario below, re-run the analysis and plot the posterior distributions of λ_1, λ_2, and τ:

   (a) Try two values of α for λ_1 and λ_2 which incorporates the belief that we will have a change at $\tau = 40$. That is, try
$$\alpha_1 = \left(\frac{1}{40}\sum_{i=0}^{39} C_i\right)^{-1}.$$
$$\alpha_2 = \left(\frac{1}{N-40}\sum_{i=40}^{N} C_i\right)^{-1}.$$

   (b) Repeat (a) above but now with incorporating the belief that the change is at $\tau = 25$.

   (c) Keep α the same as in the book but change the prior on τ to be DiscreteUniform(1,40) and probability 0 for all other time points. Note that this is putting in the belief that the change could have occurred anywhere from day 1 to 40 with equal probability but could *not* have occurred on any other day.

   Comment on how the prior hyperparameter(s) choices affects the posterior belief in what τ is and compare this to what happens in the book.

2. **The Monty Hall Problem**: Suppose you're given the choice of three doors: behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to switch to door No. 2?" The question is whether or not it to your advantage to switch your choice.

   (a) Simulate an experiment that tests this question and calculate your "best guess" of what your probability of winning is if you switch doors. Note that whether or not you win is a random variable, so histograms are helpful here. Is this surprising?

   (b) Analytically calculate the probability of winning if you switch doors.

3. **The Birthday Problem**: Our class has around 40 students. Assume that all birthdays are equally likely and you may assume no one in the class was born on February 29th. You are interested in calculating the probability that there are at least two people with the same birthday.

(a) Simulate an experiment that enables you can estimate the probability that there are at least two people with the same birthday in a class of 40. What probability do you obtain?

(b) Generalize the above simulation for a class of size n, for any $n > 1$. By estimating the probabilities for a range of n, what is the minimum size of class required for there to be a probability of 0.50 or higher of having two students with the same birthday?

4. The Golden Gate Transit bus travels between the San Anselmo and San Francisco, which are 20 miles apart. If the bus has a breakdown, the distance from the breakdown to San Anselmo has a uniform distribution over (0, 20). Currently, there is a bus service station in San Anselmo, San Francisco and in the center of the route between the two cities. It is suggested that it would be more efficient to have the three stations located 5, 10, and 15 miles, respectively, from San Anselmo. Conduct a simulation study that compares these two competing strategies and discuss which option you think is advisable.