Homework 5

Due Wednesday, 4/3

PUBH 8442: Bayes Decision Theory and Data Analysis

Include any code used to generate answers at the end of your assignment, and submit electronically via Canvas.

1. *Pump failure*
   Table 1 gives data (from Gaver & O'Muircheartaigh, 1987) on the number of pump failures Y_i observed in t_i thousands of hours for 10 nuclear power systems $i = 1, \ldots, 10$. A natural model for these data is $Y_i \overset{iid}{\sim} \text{Poisson}(\theta_i t_i)$, $\theta_i \overset{iid}{\sim} \text{Gamma}(a, b)$. Here θ_i represents the failure rate per thousand hours for the i^{th} system, and a, b are shared hyperparameters. [Credit: this is based on exercises 5.9 and 5.10 in *Carlin & Louis*. However, solve it using the parameterization of the Gamma distribution that we use in class, not that used in the Carlin & Louis book.]

   (a) Find the marginal distribution of $\mathbf{Y} = (Y_1, \ldots, Y_{10})^T$, $P(\mathbf{Y} \mid a, b)$.

   (b) Use the method of moments to obtain closed form expressions for hyperparameter estimates $\hat{a}$ and $\hat{b}$. (Define the rates $r_i = Y_i/t_i$, and equate their first two moments $\bar{r}$ and s_r^2 to the corresponding theoretical moments in terms of a and b).

   (c) Calculate $\hat{a}$ and $\hat{b}$ for the pump failure data.

   (d) Based on $\hat{a}$ and $\hat{b}$, compute point estimates for the failure rate of each pump.

i	Y_i	t_i	r_i
1	5	94.320	0.053
2	1	15.720	0.064
3	5	62.880	0.080
4	14	125.760	0.111
5	3	5.240	0.573
6	19	31.440	0.604
7	1	1.048	0.954
8	1	1.048	0.954
9	4	2.096	1.910
10	22	10.480	2.099

Table 1: Pump failure data

2. *Stratified sampling*
   Stratified random sampling is a popular survey sampling scheme where the population of N units is first divided into J non-overlapping subpopulations, or *strata*, of $N_1, N_2, \ldots, N_J$ units so that $\sum_{j=1}^{J} N_j = N$. Once the strata have been determined, a simple random sample (here we assume without replacement) is drawn from *within* each stratum, the drawings being made independently across strata. The sample sizes within each strata are given by $n_1, \ldots, n_J$ and the total sample size is $n = \sum_{j=1}^{J} n_j$.

   Let Y_{ij} denote unit i of the population in stratum j and let $\mathbf{I} = \{I_{ij}\}$ be the collection of inclusion indicators, where $I_{ij} = 1$ if the i-th population unit in stratum j is included

Stratum 1	Stratum 2		
324	180	130	101
797	314	172	121
507	238	153	116
748	296	163	119
457	235	138	113
381	192	132	104

Table 2: Inhabitants, in thousands, from a stratified random sample of 24 US cities.

in the sample, and $I_{ij} = 0$ otherwise. We also denote by $\mathbf{Y}_I = \{y_{ij}\}$ the collection of sampled units, where y_{ij} represents unit i of the random sample from stratum j. Finally, let $D = (\mathbf{Y}_I, \mathbf{I})$ be the *observed data* conditional upon which all posterior distributions will be evaluated.

Define $\bar{y_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ and $\bar{Y} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} Y_{ij}$, and assume

$$Y_{ij} \mid \mu_j, \sigma_j^2 \overset{iid}{\sim} N(\mu_j, \sigma_j^2),\ j = 1, \ldots, J.$$

with a flat (uniform) prior on each μ_j. Then,

$$\mu_j \mid D, \sigma_j^2 \overset{indep}{\sim} N(\bar{y}_j, \frac{\sigma_j^2}{n_j}),$$

and

$$P(\bar{Y}_j \mid D, \sigma_j^2) = N\left(\bar{y}_j, \frac{\sigma_j^2}{n_j} - \frac{\sigma_j^2}{N_j}\right)$$

(a) Now assume that the σ_j^2's are all unknown and consider the non-informative prior $P(\mu_j, \sigma_j^2) \propto 1/\sigma_j^2$ for each j. Describe clearly an exact (i.e., direct) posterior sampling algorithm that will yield samples from the posterior distribution of the population stratum means $\bar{Y}_j$.

(b) Explain how the posterior samples of $\bar{Y}_j$ for $j = 1, \ldots, J$ from part (a) can be used to obtain exact posterior samples of $\bar{Y}$.

(c) Table 1 presents data from one of the first stratified sampling exercises in the United States, from 1920 to estimate the number of inhabitants in the $N = 64$ largest cities at the time. These 64 cities were divided into $J = 2$ strata, where the first strata comprised $N_1 = 16$ of the largest cities and the second strata comprised the remaining $N_2 = 48$ cities. Within the first stratum $n_1 = 6$ cities were chosen, while $n_2 = 18$ cities were chosen from stratum 2.

Using the data in Table 1, use `R` to compute the posterior mean and 95% credible intervals for the *total* (not average) number of inhabitants in the 64 cities. Also present posterior estimates (mean, median and 95% credible intervals) for the *total* number of inhabitants in each stratum.