Homework 5

PUBH 8442: Bayes Decision Theory and Data Analysis

Include any code used to generate answers at the end of your assignment.

1. Pump Failure

   (a) The marginal pmf of $\mathbf{Y}$ is the product $P(\mathbf{y} \mid a, b) = \prod_{i=1}^{10} P(y_i \mid a, b)$, where

$$\begin{aligned} P(y_i \mid a, b) &= \int P(y_i|\theta_i) p(\theta_i \mid a, b) \mathrm{d}\theta_i \\ &= \int \frac{(\theta_i t_i)^{y_i}}{y_i!} e^{-\theta_i t_i} \times \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-b\theta_i} \mathrm{d}\theta_i \\ &= \frac{t_i^{y_i}}{y_i!} \frac{b^a}{\Gamma(a)} \frac{\Gamma(y_i + a)}{(t_i + b)^{y_i + a}} \int \frac{(t_i + b)^{y_i + a}}{\Gamma(y_i + a)} \theta_i^{y_i + a - 1} e^{-(t_i + b)\theta_i} \mathrm{d}\theta_i \\ &= \binom{y_i + a - 1}{y_i} \left(\frac{t_i}{t_i + b}\right)^{y_i} \left(\frac{b}{t_i + b}\right)^{a} \end{aligned}$$

$$y_i \sim NB\left(a, \frac{t_i}{t_i + b}\right) \text{ (Negative Binomial)}$$

   (b) Using the law of total expectation,

$$E(r_i) = E_{\theta_i} E_{Y|\theta_i} r_i = E_{\theta_i} \theta_i = \frac{a}{b}.$$

   Using the law of total variance, and results for the mean and variance of a Poisson or a Gamma distribution,

$$\begin{aligned} V(r_i) &= E_{\theta_i} V_{Y|\theta_i}(r_i) + V_{\theta_i} E_{Y|\theta_i}(r_i) \\ &= E_{\theta_i}(\theta_i / t_i) + V_{\theta_i} \theta_i \\ &= \frac{a}{b} \cdot \frac{1}{t_i} + \frac{a}{b^2}. \end{aligned}$$

   Then, solving

$$\bar{r} = E(r_i) = \frac{a}{b} \tag{1}$$

$$s_r^2 = \frac{1}{10} \sum_{i=1}^{10} V(r_i) = \frac{a}{b^2} + \frac{a}{10b} \sum_{i=1}^{10} \frac{1}{t_i} \tag{2}$$

   gives $\hat{b} = \hat{a}/\bar{r}$ and $\hat{a} = \frac{\bar{r}^2}{s_r^2 - (\bar{r}/10) \sum_{i=1}^{10} t_i^{-1}}$.

   (c) If the student uses $s_r^2 = \frac{1}{10-1} \sum_{i=1}^{10} (r_i - \bar{r})^2$, $\hat{a} = 1.52$; $\hat{b} = 2.05$. If the student uses $s_r^2 = \frac{1}{10} \sum_{i=1}^{10} (r_i - \bar{r})^2$, $\hat{a} = 1.80$; $\hat{b} = 2.43$.

   (d) The posterior distribution is

$$\begin{aligned} p(\theta_i|y_i) &\propto p(\theta_i) \cdot p(y_i|\theta_i) \\ &\propto \frac{b^a}{\Gamma(a)} \theta_i^{a-1} e^{-b\theta_i} \times \frac{(\theta_i t_i)^{y_i}}{y_i!} e^{-\theta_i t_i} \\ &\propto \frac{(b + t_i)^{y_i + a}}{\Gamma(y_i + a)} \theta_i^{y_i + a - 1} e^{-(b + t_i)\theta_i} \\ &\sim \text{Gamma}(y_i + a, t_i + b) \end{aligned}$$

We use the posterior mean

$$\hat{\theta}_i = \frac{y_i + \hat{a}}{t_i + \hat{b}}$$

If the student uses $s_r^2 = \frac{1}{10-1}\sum_{i=1}^{10}(r_i - \bar{r})^2$, the point estimates for the failure rate of each pump are 0.067, 0.142, 0.100, 0.121, 0.620, 0.613, 0.813, 0.813, 1.330, 1.877. If the student uses $s_r^2 = \frac{1}{10}\sum_{i=1}^{10}(r_i - \bar{r})^2$, the point estimates for the failure rate of each pump are 0.070, 0.154, 0.104, 0.123, 0.626, 0.614, 0.805, 0.805, 1.281, 1.843.

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 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.

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

a.) First, draw samples from the marginal posterior distribution $P(\sigma_j^2|D)$. Then, draw samples from the conditional posterior distribution $P(\bar{Y}_j|D, \sigma_j^2)$ using σ_j^2 sampled in the previous step.
We can get the posterior distribution for σ^2 as follows:

$$
\begin{aligned}
P(\mu_j, \sigma_j^2) &\propto \frac{1}{\sigma_j^2} \\
P(\mu_j, \sigma_j^2|D) &\propto (\sigma_j^2)^{-n_j/2-1} \exp\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2\} \\
P(\sigma_j^2|D) &\propto \int P(\mu_j, \sigma_j^2|D)\mathrm{d}\mu_j \\
&\propto \left(\sigma_j^2\right)^{-\frac{n_j-1}{2}-1} \exp\{-\frac{\sum(y_{ij}-\bar{y}_j)^2}{2\sigma_j^2}\} \\
&\sim IG\left(\frac{n_j-1}{2}, \frac{\sum(y_{ij}-\bar{y}_j)^2}{2}\right)
\end{aligned}
$$

b.) Note that $\bar{Y} = \sum_{j=1}^{J} \frac{N_j}{N}\bar{Y}_j$. We can obtain exact posterior samples of $\bar{Y}$ by the weighted average of $\bar{Y}_j$ sampled from $P(\bar{Y}_j|D, \sigma_j^2)$.

c.) The posterior mean and the 95% credible intervals for the total number of inhabitants in the 64 cities: 16613.36 (13746.80, 19402.04)
The posterior estimates for stratum 1: Mean 8565.13; median 8571.94; CI (5946.19, 11108.49)
The posterior estimates for stratum 2: Mean 8048.23; median 8044.57; CI (6789.25, 9264.15)

Following is the R program:

```
y1 <- c(324, 797, 507, 748, 457, 381)
y2 <- c(180, 130, 101, 314, 172, 121,
        238, 153, 116, 296, 163, 119,
        235, 138, 113, 192, 132, 104)
N <- 64; N1 <- 16; N2 <- 48
n1 <- 6; n2 <- 18

ybar1 <- mean(y1)
ybar2 <- mean(y2)
ssq1 <- var(y1)
ssq2 <- var(y2)

n.samp <- 10000
```

```
samp.ybar1 <- samp.ybar2 <- c()
for (i in 1:n.samp)
{
   samp.ssq1 <- 1/rgamma(1, (n1-1)/2, (n1-1)*ssq1/2)
   samp.ssq2 <- 1/rgamma(1, (n2-1)/2, (n2-1)*ssq2/2)
   samp.ybar1 <- c(samp.ybar1, rnorm(1, ybar1, sqrt(samp.ssq1/n1-samp.ssq1/N1)))
   samp.ybar2 <- c(samp.ybar2, rnorm(1, ybar2, sqrt(samp.ssq2/n2-samp.ssq2/N2)))
}

tot.Y <- N1*samp.ybar1 + N2*samp.ybar2
tot.y1 <- N1*samp.ybar1
tot.y2 <- N2*samp.ybar2

mean(tot.Y)
quantile(tot.Y, c(0.025, 0.975))
mean(tot.y1)
quantile(tot.y1, c(0.025, 0.500, 0.975))
mean(tot.y2)
quantile(tot.y2, c(0.025, 0.500, 0.975))
```