ST440/540 – Mid-term in-class exam

SOLUTIONS Febuary 25, 2019

The exam is open notes but you are not allowed to use a phone or laptop. GIVING OR RECEIVING ASSISTANCE FROM OTHER STUDENTS IS NOT PERMITTED!

1. Let $Y_1, ..., Y_n$ be independent with $Y_i|\theta \sim \text{Gamma}(a_i, \theta)$[1]. Identify a conjugate prior distribution for θ and give a step-by-step mathematical derivation of the posterior distribution. The answer should be a distribution, such as $\theta \sim \text{Beta}(Y_2, 1)$.

   Solution: The likelihood is

   $$p(Y_1, ..., Y_n|\theta) = \prod_{i=1}^{n} f(Y_i|\theta) \propto \prod_{i=1}^{n} \theta^{a_i} \exp(-Y_i\theta) \propto \theta^A \exp(-B\theta).$$

   where $A = \sum_{i=1}^n a_i$ and $B = \sum_{i=1}^n Y_i$. This is the kernel of a gamma PDF and so the conjugate prior is $\theta \sim \text{Gamma}(c, d)$. The posterior is then

   $$p(\theta|Y_1, ..., Y_n) \propto [\theta^A \exp(-B\theta)][\theta^{c-1} \exp(-d\theta)] \propto \theta^{A+c-1} \exp(-(B+d)\theta)$$

   and therefore the posterior is $\text{Gamma}(c + \sum_{i=1}^n a_i, d + \sum_{i=1}^n Y_i)$

[1] That is, the PDF of Y_i is $\frac{\theta^{a_i}}{\Gamma(a_i)} y_i^{a_i-1} \exp(-\theta y_i)$.

2. Denote the probability that a part is defective as θ. The industry standard is that no more than 0.1% of parts can be defective, i.e., $\theta < 0.001$. Your company has purchased a new machine, generated 10,000 parts, and tested each to determine if it is defective. You are now tasked with testing the null hypothesis that $\theta \leq 0.001$ versus the alternative hypothesis that $\theta > 0.001$.

   (a) Describe an *objective* Bayesian analysis of the problem including the likelihood and prior and posterior distributions of θ.
   Solution: Let $Y \in \{0, 1, ..., n\}$ be the number of defective parts in the sample of size $n = 10,000$. The likelihood is $Y|\theta \sim \text{Binomial}(n, \theta)$. In class we found the Jeffreys' prior for this case is $\theta \sim \text{Beta}(0.5, 0.5)$ and the posterior was then $\theta|Y \sim \text{Beta}(Y + 0.5, n - Y + 0.5)$.

   (b) Explain how you would compute the posterior probability that the null hypothesis is true.
   Solution: I would compute $\text{Prob}(\theta < 0.0001|Y)$ in `R` using the command

   ```
   > pbeta(0.0001,Y+0.5,10000-Y+0.5)
   ```

   (c) Explain how a frequentist would compute the probability that the null hypothesis is true.
   Solution: From the frequentist viewpoint this probability is not well defined because the parameters and thus the hypotheses are not random variables. So they might say "NA."

3. Your daily commute is distributed uniformly[2] between 15 and 20 minutes if there no convention downtown. However, conventions are scheduled for roughly 1 in 4 days, and your commute time is distributed uniformly from 15 to 30 minutes if there is a convention. Let Y be your commute time this morning.

   (a) What is the probability that there was a convention downtown given $Y = 18$? **Solution**: Let $C = 1$ if there is a convention downtown and $C = 0$ otherwise. The likelihood for the commute time, Y, is $Y|C = 0 \sim \text{Uniform}(15, 20)$ and $Y|C = 1 \sim \text{Uniform}(15, 30)$ and the prior is $\text{Prob}(C = 1) = 1/4$. Bayes' rule gives

$$\begin{aligned}\text{Prob}(C = 1|Y = 18) &= \frac{f(18|C = 1)\pi(C = 1)}{f(18|C = 1)\pi(C = 1) + f(18|C = 0)\pi(C = 0)} \\ &= \frac{(1/15)(1/4)}{(1/15)(1/4) + (1/5)(3/4)} = 0.1.\end{aligned}$$

   (b) What is the probability that there was a convention downtown given $Y = 28$? **Solution**: $Y = 28$ is only consistent with a convention, so $\text{Prob}(C = 1|Y = 28) = 1$. You can arrive at this via Bayes' rule by inserting $f(28|C = 0) = 0$.

[2] If Y is distributed uniformly on (a, b) then its PDF is $1/(b - a)$ for $y \in (a, b)$ and 0 if y is outside (a, b).

4. The Major League Baseball player Reggie Jackson is known as "Mr. October" for his outstanding performances in the World Series (which takes place in October). Over his long career he hit 563 home runs in 2820 regular-season games, and 10 home runs in 27 World Series games. Denote λ_1 and λ_2 as the home run rate (i.e., the mean number of home runs per game) in the regular season and world series, respectively.

   (a) Describe a Bayesian analysis (likelihood, prior and posterior) for λ_1 using only his regular-season data and improper prior $\pi(\lambda_1) = 1$ for all $\lambda_1 > 0$. Is the posterior proper?
   Solution: Let Y_1 and N_1 be the number of home runs and games, respectively, in the regular season. Since $Y_1 \in \{0, 1, 2, ...\}$ we use the likelihood $Y_1|\lambda_1 \sim \text{Poisson}(N_1\lambda_1)$. Assuming the likelihood $\pi(\lambda_1) = 1$ gives posterior

$$p(\lambda_1|Y_1) \propto \exp(-N_1\lambda_1)(N_1\lambda_1)^{Y_1} \propto \lambda_1^{(Y_1+1)-1} \exp(-N_1\lambda_1)$$

   and therefore $\lambda_1|Y_1 \sim \text{Gamma}(Y_1 + 1, N_1)$.

   (b) Explain how you would test whether his home run rate is higher in the world series than the regular season. Include all details of the statistical model (likelihood, prior, posterior) and computational algorithm.
   Solution: I would apply the same methods for world series games as for regular season games (and assuming they are independent) to arrive at the posterior

$$\lambda_j|Y_j \sim \text{Gamma}(Y_j + 1, N_j)$$

   for $j = 1$ (regular season) and $j = 2$ (world series). To approximate the posterior probability that $\lambda_2 > \lambda_1$ I would sample $S = 1,000,000$ draws from each posterior using the `rgamma` function in `R` and compute the sample proportion that have $\lambda_2 > \lambda_1$. (It turns out to be...)

```
> Y1 <- 563
> N1 <- 2820
> Y2 <- 10
> N2 <- 27
> S  <- 1000000
> l1 <- rgamma(S,Y1+1,N1)
> l2 <- rgamma(S,Y2+1,N2)
> mean(l2>l1)
[1] 0.976839
```