STATS 205P Assignment 1
Name: Chuqi Wang
Student ID: 79167724
Date: April 22nd, 2024

1. A patient is diagnosed with a disease that affects 0.05% of the population. The doctor warns the patient however that the test is not very accurate: the false positive (FP) rate is 0.08 and false negative (FN) rate is 0.04.

   (a) What is the probability that the patient actually has the disease.

   (b) To be more certain, the patient is advised to take a more thorough test (which is of course expensive) with $FP = 0.03$ and $FN = 0.01$. What is the probability that the patient actually has the disease if he is tested positive in the second test?

   (c) What is this probability if he is tested negative in the second test?

(a) Let D represents the event that the patient has the disease.

Let T represents the event that the patient has positive test result for the disease.

By Baye's theorem, $P(D|T) = \dfrac{P(T|D)\cdot P(D)}{P(T)}$

where $P(T|D) = P(TP) = 1 - P(FN) = 1 - 0.04 = 0.96$

$P(D) = 0.05\%$

and $P(T) = P(T|D)\cdot P(D) + P(T|D^c)\cdot P(D^c)$

$P(T|D^c) = P(FP) = 0.08$ and $P(D^c) = 1 - P(D) = 99.95\%$

$\Rightarrow P(T) = 0.96 \times 0.0005 + 0.08 \times 0.9995 = 0.08044$

$\Rightarrow P(D|T) = \dfrac{0.96 \times 0.0005}{0.08044} \approx 0.005967$

Therefore, the probability that the patient actually has the disease is 0.5967%.

(b) Given that $P(FP) = 0.03$ and $P(FN) = 0.01$, then we have

$P(T|D) = P(TP) = 1 - P(FN) = 0.99$

$P(T|D^c) = P(FP) = 0.03$

Again by Baye's Theorem we have

$$P(D|T) = \frac{P(T|D)\cdot P(D)}{P(T)} = \frac{P(T|D)\cdot P(D)}{P(T|D)\cdot P(D) + P(T|D^c)\cdot P(D^c)}$$

$$= \frac{0.99 \times 0.0005}{0.99\times 0.0005 + 0.03 \times 0.9995} \approx 0.01624$$

Therefore, the probability that the patient actually has the disease if he is tested positive in the second test is about 1.624%.

(c) By Baye's Theorem,

$$P(D|T^c) = \frac{P(T^c|D)\cdot P(D)}{P(T^c)} \quad \text{where} \quad P(T^c|D) = P(FN) = 0.01$$

and $P(T^c) = P(T^c|D)\cdot P(D) + P(T^c|D^c)\cdot P(D^c)$

$= P(FN)\cdot P(D) + P(TN)\cdot P(D^c)$

$= P(FN)\cdot P(D) + [1-P(FP)]\cdot[1-P(D)]$

Thus, $P(D|T^c) = \frac{0.01 \times 0.0005}{0.01 \times 0.0005 + 0.99 \times 0.9995} \approx 0.000005053$

Therefore, if he is tested negative in the second test, the probability that he has the disease is 0.0005053%

2. Consider a simple normal model for the height (y) of students where $y|\theta \sim N(\theta, 4^2)$. Assume a conjugate $\theta \sim N(\mu_0, 3^2)$ prior. We measure the height of two students and observe $y_1 = 68$ and $y_2 = 71$.

   (a) Find the posterior distribution of θ given y when the data come sequentially (i.e., we first observe y_1, we update our prior, then observe y_2 and observe our prior again).

   (b) Find the posterior distribution of θ given y when data come simultaneously (i.e., we update our prior after observing both y_1 and y_2). Plot the prior, the likelihood and all the different posterior distributions.

   (c) Discuss your findings (one paragraph).

(a) Given that $y|\theta \sim N(\theta, 4^2)$ and $\theta \sim N(\mu_0, 3^2)$

The posterior distribution is $\theta|y \sim N(\mu_1, \tau_1^2)$ after we observe $y_1 = 68$ where

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y_1}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} = \frac{\frac{\mu_0}{3^2} + \frac{68}{4^2}}{\frac{1}{3^2} + \frac{1}{4^2}} \quad \text{and} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2} = \frac{1}{3^2} + \frac{1}{4^2}$$

Thus, $\mu_1 = 0.64\mu_0 + 24.48$ and $\tau_1^2 = 5.76$

After observe $y_1 = 68$, the posterior distribution is given by $\theta|y \sim N(0.64\mu_0 + 24.48, 5.76)$ and we update the prior to be $\theta \sim N(\mu_1, \tau_1^2)$ where $\mu_1 = 0.64\mu_0 + 24.48$ and $\tau_1^2 = 5.76$.

After we observe $y_2 = 71$, the posterior distribution is $\theta|y \sim N(\mu_2, \tau_2^2)$ where $\frac{25}{34}\mu_1 + \frac{639}{34}$

$$\mu_2 = \frac{\frac{\mu_1}{\tau_1^2} + \frac{y_2}{\sigma^2}}{\frac{1}{\tau_1^2} + \frac{1}{\sigma^2}} = \frac{\frac{\mu_1}{5.76} + \frac{71}{16}}{\frac{1}{5.76} + \frac{1}{16}} = \frac{25}{34}\mu_1 + \frac{639}{34}$$

$$= 0.4706\mu_0 + 36.79$$

and $\frac{1}{\tau_2^2} = \frac{1}{5.76} + \frac{1}{16} = 0.2361 \Rightarrow \tau_2^2 = 4.2353$

Therefore, after observe $y_2 = 71$, $\theta|y \sim N(0.4706\mu_0 + 36.79, 4.2353)$

(b) When the data come simultaneously, the posterior distribution can be written as: $\theta|y \sim N(\mu_n, \tau_n^2)$ where

$$\mu_n = \frac{\frac{\mu_0}{3^2} + \frac{y_1+y_2}{4^2}}{\frac{1}{3^2} + \frac{2}{4^2}} = \frac{\frac{\mu_0}{9} + \frac{139}{16}}{\frac{1}{9} + \frac{1}{8}} = 0.4706\mu_0 + 36.79$$

and $\frac{1}{\tau_n^2} = \frac{1}{9} + \frac{2}{16} \Rightarrow \tau_n^2 = 4.2353$

Thus, the posterior distribution is $\theta|y \sim N(0.4706\mu_0 + 36.79, 4.2353)$

(c) From the results of (a) and (b), we found the final posterior distributions are the same for data come sequentially and data come simultaneously. This reflects the consistency and cumulative nature of Bayesian inference, where the order of data incorporation does not affect the final inference.

3. Suppose the number of solar flares within a week has a Poisson(θ) distribution. Based on some previous studies, we believe that the mean and standard deviation of θ are 1.6 and 0.4 respectively.

   (a) Find an appropriate distribution that reflects our prior belief regarding the value of θ. Plot your prior distribution.

   (b) We measure the number of solar flares per week over one month. The observed values are 2, 4, 1 and 2. What is the posterior distribution of θ? Plot the posterior distribution and write down the posterior mean and variance.

(a) Suppose the number of solar flares is $y = (y_1, y_2, \ldots, y_n)$ Given that $y|\theta \sim \text{Poisson}(\theta)$. From class we found the conjugate prior would have the following form $P(\theta) \propto \exp(-\beta\theta)\theta^{\alpha-1}$ which is $\text{Gamma}(\alpha, \beta)$, so for poisson model, we can choose a prior is Gamma distribution.

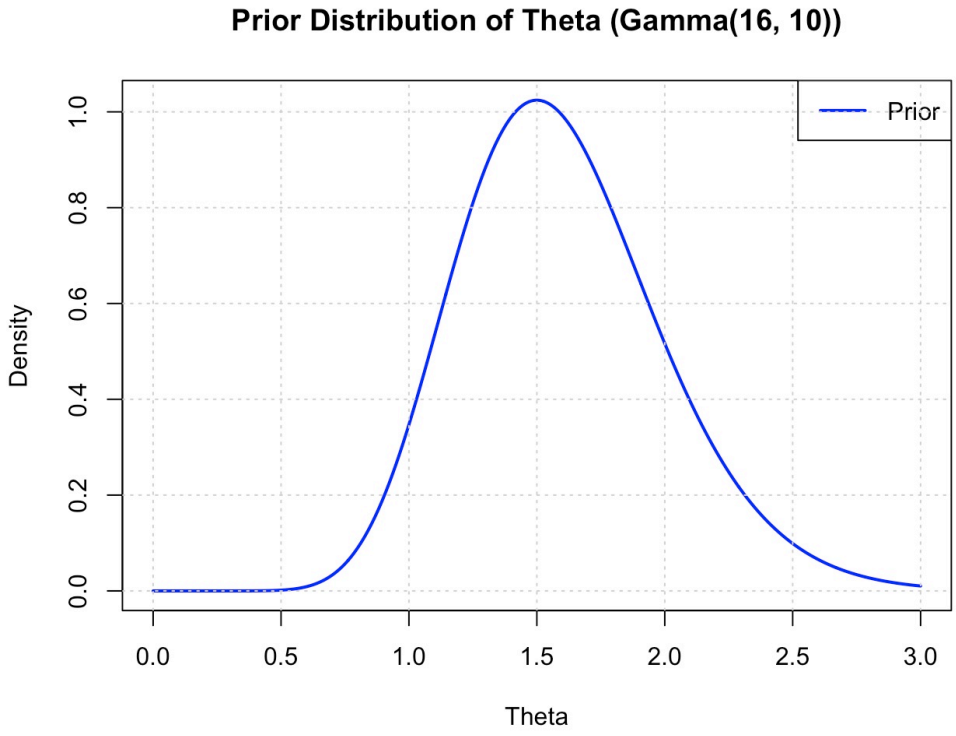
Let $\theta \sim \text{Gamma}(\alpha, \beta)$, Given the mean = 1.6 and variance = 0.4^2 for prior θ.

So we have $\frac{\alpha}{\beta} = 1.6$ and $\frac{\alpha}{\beta^2} = 0.16$

By solving $\begin{cases} \frac{\alpha}{\beta} = 1.6 \\ \frac{\alpha}{\beta^2} = 0.16 \end{cases}$, we have $\alpha = 16$ and $\beta = 10$

Therefore, an appropriate prior should be $\theta \sim \text{Gamma}(16, 10)$

The plot of prior distribution are shown below.



(b) Since $P(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} \exp(-\theta)}{y_i!} \propto \exp(-n\theta)\,\theta^{\sum_{i=1}^{n} y_i}$

and $P(\theta) \propto \exp(-\beta\theta)\,\theta^{\alpha-1}$, so posterior is given by:

$$P(\theta|y) \propto P(y|\theta)\,P(\theta) \propto \theta^{\sum_{i=1}^{n} y_i} \exp(-n\theta)\,\theta^{\alpha-1} \exp(-\beta\theta)$$

$$= \theta^{\alpha+\sum_{i=1}^{n} y_i - 1} \exp(-(\beta+n)\theta)$$

which gives $\theta|y \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} y_i,\ \beta+n)$

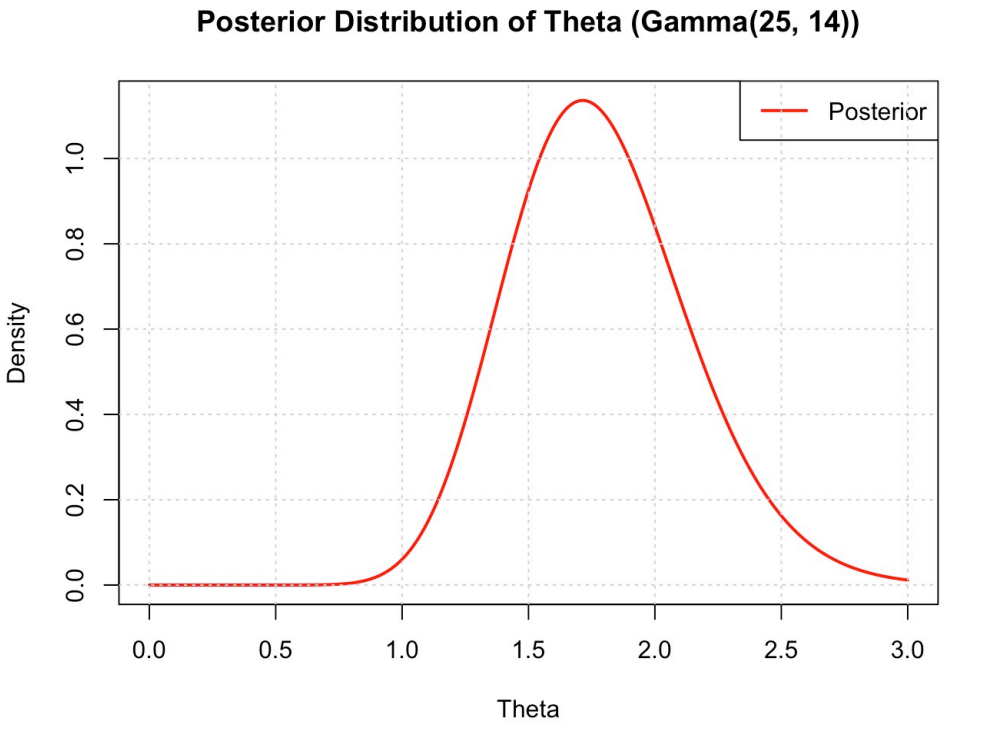
In our example, $\alpha = 16$, $\beta = 10$. $y = (y_1, y_2, y_3, y_4) = (2, 4, 1, 2)$

and $n = 4$.

Therefore, $\theta|y \sim \text{Gamma}(16+2+4+1+2,\ 10+4) = \text{Gamma}(25, 14)$

the mean is $\frac{\alpha}{\beta} = \frac{25}{14} \approx 1.786$ and variance $= \frac{\alpha}{\beta^2} \approx 0.128$

The plot of posterior distribution are shown below.



4. We are rolling a 6-sided die, which might not be a fair die. A random variable X represents the outcome, where $\theta_j = P(X = j)$.

   (a) Choose an appropriate model for X and assume prior for $\theta = (\theta_1, \ldots, \theta_6)$.

   (b) We roll the die 8 times and observe two 1's, zero 2's, one 3's, two 4's, zero 5's, and three 6's. Use the data to update the prior and obtain the posterior distribution of θ_j (for $j = 1, 2, ..., 6$).

   (c) Now assume that we are only interested to know whether the outcome is 6. How would you modify the model?

   (d) This time use a more informative prior (we strongly believe that the die is "fair") for the model parameter and obtain its posterior distribution given the observed data.

(a) Random variable X has a multinomial distribution with 6 groups, the sampling distribution would have the following form $P(x|\theta) \propto \prod_{j=1}^{6} \theta_j^{x_j}$. This is a multinomial model, so the conjugate prior for this model is the Dirichlet distribution.

$P(\theta|\alpha) \propto \prod_{j=1}^{6} \theta_j^{\alpha_j - 1}$, $\theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$

and the posterior distribution is also a $\text{Dirichlet}(x_1+\alpha_1, \ldots, x_6+\alpha_6)$

$P(\theta|\alpha, x) \propto \prod_{j=1}^{6} \theta_j^{x_j + \alpha_j - 1}$

(b) Suppose our prior is $\theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$

After we observe $x_1 = 2$, $x_2 = 0$, $x_3 = 1$, $x_4 = 2$, $x_5 = 0$, $x_6 = 3$ in 8 times rolling the die.

We update the prior and obtain the posterior is

$\theta|x \sim \text{Dirichlet}(\alpha_1+2, \alpha_2, \alpha_3+1, \alpha_4+2, \alpha_5, \alpha_6+3)$

(c) If we are only interested to know whether the outcome is 6, then we can just modify the model to a Binomial distribution. Let Y represents the number of 6's in n rolls of the die.

Then $Y|\theta_6 \sim \text{Binomial}(n, \theta_6)$ where θ_6 is the probability of rolling 6.

And the prior will be $\theta_6 \sim \text{Beta}(\alpha, \beta)$ where α is the number of rolling 6 and β is the number of not rolling 6.

$$P(\theta_6|y) \propto P(y|\theta_6) \cdot P(\theta_6) \quad \text{where} \quad P(\theta_6) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_6^{\alpha-1}(1-\theta_6)^{\beta-1}$$

$$P(y|\theta_6) = \binom{n}{y} \theta_6^{y}(1-\theta_6)^{n-y}$$

The posterior distribution simplifies to

$$P(\theta_6|y) \propto \theta_6^{y+\alpha-1}(1-\theta_6)^{\beta+n-y-1}$$

Thus, $\theta_6|y \sim \text{Beta}(\alpha+y, \beta+n-y)$

(d) If we strongly believe that the die is "fair", then we may use $\theta \sim \text{Dirichlet}(1,1,1,1,1,1)$ for (a) and $\theta_6 \sim \text{Beta}(1,5)$ since $E[\theta_6] = \frac{1}{1+5} = \frac{1}{6}$.

For part (b), the posterior distribution will be

$\theta|x \sim \text{Dirichlet}(3,1,2,3,1,4)$

R code :

```r
# Q3(a)

theta_values <- seq(0, 3, length.out = 1000)

prior_density <- dgamma(theta_values, shape = 16, rate = 10)

plot(theta_values, prior_density, type = 'l', col = 'blue', lwd = 2,
     main = 'Prior Distribution of Theta (Gamma(16, 10))',
     xlab = 'Theta', ylab = 'Density')
legend("topright", legend = sprintf("Prior", 16, 10),
       col = 'blue', lty = 1, lwd = 2)
grid()

#Q3(b)
posterior_density <- dgamma(theta_values, shape = 25, rate = 14)

plot(theta_values, posterior_density, type = 'l', col = 'red', lwd = 2,
     main = 'Posterior Distribution of Theta (Gamma(25, 14))',
     xlab = 'Theta', ylab = 'Density')
legend("topright", legend = sprintf("Posterior", 25, 14),
       col = 'red', lty = 1, lwd = 2)
grid()
```