Statistics for Data Science - 2

Week 9 Practice Assignment Solution

1. Let p be the proportion of students in IITM online degree programme who approve the online proctored exams. The students' committee is going to take a random sample of $n = 40$ students from IITM online degree programme and ask if they approve the online proctored exams. Suppose 10 out of the 40 students answered yes.
   i) Calculate the posterior distribution if we use a continuous Uniform[0, 1] prior.

   a) Beta(10, 30)
   b) Beta(11, 31)
   c) Beta(10, 40)
   d) Beta(11, 40)

   Solution:
   Let $f_{\mathbf{p}}(p)$ denote the prior distribution of $\mathbf{p}$.
   Then, by given information $f_{\mathbf{p}}(p) = 1$, since, $\mathbf{p} \sim$ Uniform[0, 1].
   If $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli($\mathbf{p}$)
   $\Rightarrow$ posterior density = Beta$(w + 1, n - w + 1)$ where w is the number of success.
   Here $n = 40, w = 10 \Rightarrow$ posterior density = Beta(11, 31)
   ii) Find the Bayesian estimate (posterior mean) of p. Enter your answer correct to two decimals accuracy.
   Solution:
   posterior mean $= \dfrac{11}{11 + 31} = 0.26$

   iii) Find the Bayesian estimate (posterior mean) with Beta(5, 5) prior. Enter your answer correct to two decimals accuracy.
   Solution:
   Given the prior distribution is Beta(α, β)
   $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli($\mathbf{p}$)
   $\Rightarrow$ posterior density = Beta$(w + \alpha, n - w + \beta)$
   Here $n = 40, w = 10, \alpha = 5, \beta = 5$
   $\Rightarrow$ posterior density = Beta(15, 35)
   posterior mean $= \dfrac{15}{15 + 35} = 0.30$

   iv) Find the Bayesian estimate (posterior mean) with Beta(10, 10) prior. Enter your answer correct to two decimals accuracy.
   Solution:
   Here $n = 40, w = 10, \alpha = 10, \beta = 10$

$\Rightarrow$ posterior density = Beta(20, 40)

posterior mean $= \dfrac{20}{20+40} = 0.33$

2. The new method of screening for a disease fails to detect the presence of the disease in 20% of the patients from prior experience. A new random sample of $n = 100$ patients who are known to have the disease is screened using the new method. Out of these 100 patients, the new method failed to detect the disease in 20 cases. Use a Beta(2, β) with a suitable β to estimate the failure fraction. Enter your answer correct to two decimals accuracy.

   Solution:

   The prior is given as Beta(2, β) with the information that the new method failed to detect the disease in 20 cases.

   $\Rightarrow \dfrac{2}{2+\beta} = 0.20$

   $\Rightarrow \beta = 8.$

   If $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli($\mathbf{p}$)

   $\Rightarrow$ posterior density = Beta$(w+\alpha, n-w+\beta)$

   Here $n = 100, w = 20, \alpha = 2, \beta = 8$

   $\Rightarrow$ posterior density = Beta(22, 88)

   posterior mean $= \dfrac{22}{22+88} = 0.20$

3. Suppose that the number of customers arriving in a restaurant in a one day time period follows the Poisson distribution with unknown parameter λ. Previous records suggest that the prior probabilities of λ are $P(\lambda = 10) = 0.4$ and $P(\lambda = 8) = 0.6$. If on a particular day 15 people arrive at the restaurant, find the posterior mode of λ.

   Solution:

$$\begin{aligned} P(X=15) &= P(X=15 \mid \lambda=10)P(\lambda=10) + P(X=15 \mid \lambda=8)P(\lambda=8) \\ &= \frac{e^{-10}10^{15}}{15!} \times 0.4 + \frac{e^{-8}8^{15}}{15!} \times 0.6 \end{aligned}$$

   Now,

$$\begin{aligned} P(\lambda=10 \mid X=15) &= P(X=15 \mid \lambda=10)P(\lambda=10)/P(X=15) \\ &= \frac{e^{-10}10^{15} \times 0.4}{e^{-10}10^{15} \times 0.4 + e^{-8}8^{15} \times 0.6} \end{aligned}$$

And

$$P(\lambda = 8 \mid X = 15) = P(X = 15 \mid \lambda = 8)P(\lambda = 8)/P(X = 15)$$
$$= \frac{e^{-8}8^{15} \times 0.6}{e^{-10}10^{15} \times 0.4 + e^{-8}8^{15} \times 0.6}$$

$$\Rightarrow \frac{P(\lambda = 10 \mid X = 15)}{P(\lambda = 8 \mid X = 15)} = \frac{e^{-10}10^{15} \times 0.4}{e^{-8}8^{15} \times 0.6}$$
$$= 2.56$$
$$\Rightarrow P(\lambda = 10 \mid X = 15) > P(\lambda = 8 \mid X = 15)$$

Hence, the posterior mode of λ is 10.

4. Consider a Bayesian estimation problem, with $X_1, \ldots, X_n \sim$ iid Normal$(\mu, 1)$, and a Normal(0, 1) prior. Letting $Y_n = \sum_{i=1}^{n} X_i$, the posterior mean is

   a) $\frac{Y_n}{n}$

   b) $\frac{Y_n}{n+1}$

   c) $\frac{nY_n}{n+1}$

   d) $\frac{nY_n}{n+2}$

   Solution:
   If $X_1, \ldots, X_n \sim$ iid Normal(μ, σ^2), and the prior is Normal(μ_0, σ_0^2)
   then posterior density $= \overline{X}\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}$
   Here $\sigma = 1, \mu_0 = 0, \sigma_0^2 = 1$
   $\Rightarrow$ posterior mean $= \frac{X_1 + \ldots + X_n}{n}\left(\frac{n \times 1}{n \times 1 + 1}\right)$
   $\Rightarrow$ posterior mean $= \frac{Y_n}{n+1}$

5. The marks distribution of IITM students in the end semester exam follows normal distribution with unknown mean μ and variance 20. A random sample of marks of 8 students are:
   60, 60, 65, 65, 70, 70, 72, 75.
   i) Assume that the prior distribution is Normal$(50, 5)$. Find the posterior mean of μ . Enter your answer correct to two decimals accuracy.

Solution:
$\overline{X} = \dfrac{60+60+65+65+70+70+72+75}{8} = 67.125$
Here $\overline{X} = 67.125, \sigma^2 = 20, n = 8, \mu_0 = 50, \sigma_0^2 = 5$

$$\begin{aligned}\text{posterior mean} &= 67.125\left(\frac{8\times 5}{8\times 5+20}\right) + 50\left(\frac{20}{8\times 5+20}\right)\\ &= 67.125\left(\frac{40}{60}\right) + 50\left(\frac{20}{60}\right)\\ &= 61.416\end{aligned}$$

ii) Assume that the prior distribution is Normal$(50, 25)$. Find the posterior mean of μ . Enter your answer correct to two decimals accuracy.
Solution:
$\overline{X} = \dfrac{60+60+65+65+70+70+72+75}{8} = 67.125$
Here $\overline{X} = 67.125, \sigma^2 = 20, n = 8, \mu_0 = 50, \sigma_0^2 = 25$

$$\begin{aligned}\text{posterior mean} &= 67.125\left(\frac{8\times 25}{8\times 25+20}\right) + 50\left(\frac{20}{8\times 25+20}\right)\\ &= 67.125\left(\frac{200}{220}\right) + 50\left(\frac{20}{220}\right)\\ &= 65.56\end{aligned}$$

6. Suppose X is a discrete random variable taking values $\{1, 2, 3\}$ with respective probabilities $\{p, 2(1-p)/3, (1-p)/3\}$, where $0 \le p \le 1$ is a parameter. Consider the samples 1, 1, 3, 1, 3, 2, 1, 2, 3, 2 taken from X.
Use a Uniform[0, 1] prior on p to find the posterior mean. Enter your answer correct to two decimals accuracy.
Solution:
Let $f_{\mathbf{p}}(p)$ denote the prior distribution of $\mathbf{p}$.
Then, by given information $f_{\mathbf{p}}(p) = 1$, since, $\mathbf{p} \sim$ Uniform[0, 1].

Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \mathbf{p} = p) f_{\mathbf{p}}(p)$
$\Rightarrow$ posterior density $\propto p^{n_1}(1-p)^{n_2+n_3}$ where n_i denotes the number of i in the samples.
Here $n_1 = 4, n_2 = 3, n_3 = 3$
$\Rightarrow$ posterior density $\propto p^4(1-p)^{3+3}$
$\Rightarrow$ posterior density = Beta(5, 7)
$\Rightarrow$ posterior mean $= \dfrac{5}{5+7} = 0.416$

7. The following ten samples are taken from the Geometric(p):
   2, 4, 10, 8, 12, 6, 14, 6, 3, 5.
   Find the posterior mean of p using Uniform$[0, 1]$ prior. Enter your answer correct to two decimals accuracy.
   Solution:
   If $X_1, \ldots, X_n \sim$ iid Geometric($\mathbf{p}$), and the prior is Uniform[0, 1],
   then posterior density = Beta$(n + 1, x_1 + x_2 + \ldots + x_n - n + 1)$
   Here $n = 10, x_1 + x_2 + \ldots + x_n = 2 + 4 + 10 + 8 + 12 + 6 + 14 + 6 + 3 + 5 = 70$
   $\Rightarrow$ posterior density = Beta(11, 61).
   $\Rightarrow$ posterior mean $= \dfrac{11}{11 + 61} = 0.15$

8. Consider the samples 7, 5, 0, 2, 10, 4, 9, 8, 3 taken from Poisson(λ), where λ is unknown. Using a Gamma(4, 11) prior, find the posterior mean of λ. Enter your answer correct to one decimal accuracy.
   Solution:
   If $X_1, \ldots, X_n \sim$ iid Poisson(λ), and the prior is Gamma(α, β)
   then posterior density = Gamma$(x_1 + x_2 + \ldots + x_n + \alpha, \beta + n)$
   Here $n = 9, x_1 + x_2 + \ldots + x_n = 7 + 5 + 0 + 2 + 10 + 4 + 9 + 8 + 3 = 48, \alpha = 4, \beta = 11$
   $\Rightarrow$ posterior density = Gamma(52, 20)
   $\Rightarrow$ posterior mean $= \dfrac{52}{20} = 2.6$

9. The number of defects per 10 meters of cloth produced by a weaving machine has the Poisson distribution with mean λ. You examine 100 meters of cloth produced by the machine and observe 61 defects. Your prior belief about λ is that it has mean 6 and standard deviation 2. Use a Gamma(α, β) prior that matches your prior belief and find the posterior distribution.

   a) Gamma(70, 111.5)
   b) Gamma(70, 11.5)
   c) Gamma$(61, 11.5)$
   d) Gamma$(61, 13)$

   Solution:
   Prior is Gamma(α, β) with mean 6 and standard deviation 2.

$$\Rightarrow \frac{\alpha}{\beta} = 6 \quad \text{and} \quad \frac{\alpha}{\beta^2} = 4$$
$$\Rightarrow \alpha = 6\beta \quad \text{and} \quad \alpha = 4\beta^2$$

   Solving these two equations we will get
   $\alpha = 9$ and $\beta = 1.5$.
   Also $n = 10$ and $x_1 + x_2 + \ldots + x_n = 61$
   $\Rightarrow$ posterior distribution = Gamma(70, 11.5)

10. Assume that the time that elapses from one call to the next at a 911 call center has the exponential distribution with parameter λ. The time elasped between ten calls (in minutes) are: $3, 4, 6, 1, 7, 8, 2, 5, 1$. Your prior belief about λ is that it has mean 3.5 and standard deviation 1. Use a Gamma(α, β) prior that matches your prior belief and find the posterior mean. Enter your answer correct to two decimals accuracy.
Solution:
Prior is Gamma(α, β) with mean 6 and standard deviation 2.

$$\Rightarrow \frac{\alpha}{\beta} = 3.5 \text{ and } \frac{\alpha}{\beta^2} = 1$$
$$\Rightarrow \alpha = 3.5\beta \text{ and } \alpha = \beta^2$$

Solving these two equations we will get
$\alpha = 12.25$ and $\beta = 3.5$.
If $X_1, \ldots, X_n \sim$ iid Exponential(λ), and the prior is Gamma(α, β)
then posterior density = Gamma$(n + \alpha, \beta + x_1 + x_2 + \ldots + x_n)$
Here $n = 9, x_1 + x_2 + \ldots + x_n = 3+4+6+1+7+8+2+5+1 = 37, \alpha = 12.25, \beta = 3.5$
$\Rightarrow$ posterior density = Gamma(21.25, 40.5)
$\Rightarrow$ posterior mean $= \frac{21.25}{40.5} = 0.52$

11. The frequency data on number of deaths per month due to a certain disease is given below:

No. of deaths per month	Frequency
0	224
1	102
2	23
3	5
4	1
5+	0

Table 9.1.P

(i) Fit a Poisson distribution to the given frequency table and find the parameter. Write your answer correct to two decimal places.

Solution:

Let $\hat{\lambda} = \bar{X}$ be an estimate of λ.

$$\begin{aligned}\text{Sample mean}(\bar{X}) =& \frac{\sum n_i f_i}{\sum f_i} \\ =& \frac{(0 \times 224) + (1 \times 102) + (2 \times 23) + (3 \times 5) + (4 \times 1)}{224 + 102 + 23 + 5 + 1} \\ =& \frac{167}{355} \\ =& 0.47\end{aligned}$$

Therefore, $\hat{\lambda} = 0.47$.

Therefore, the distribution is Poisson(0.47).

n	Frequency	Poisson fit
0	224	$(e^{-0.47})355 = 221.87$
1	102	$(e^{-0.47}(0.47)/1!)355 = 104.28$
2	23	$(e^{-0.47}(0.47)^2/2!)355 = 24.05$
3	5	$(e^{-0.47}(0.47)^3/3!)355 = 3.83$
4	1	$(e^{-0.47}(0.47)^4/4!)355 = 0.45$
5+	0	$(e^{-0.47}(0.47)^5/5!)355 = 0.04$

As we can observe from the table that the actual count is close to the expected count, therefore, Poisson(0.47) is a reasonable fit for the given data.

(ii) Find an approximate 95% confidence interval using a normal approximation for the sampling distribution.
(Use the following information:
sample variance $S^2 = 0.498$ and $P(-0.07 < \mathrm{N}(0, 0.0014) < 0.07) = 0.95$)

(a) $[0.40, 0.47]$
(b) $[0.40, 0.54]$
(c) $[0.44, 0.54]$
(d) $[0.44, 0.52]$

Solution:

Error: $\hat{\lambda} - \lambda$

$E[\hat{\lambda} - \lambda] = 0$

$\mathrm{Var}(\hat{\lambda} - \lambda) = \mathrm{Var}(\hat{\lambda}) = \dfrac{\sigma^2}{n} \approx \dfrac{S^2}{n}$

Therefore, we will assume the sampling distribution to be Normal$\left(0, \frac{s^2}{n}\right)$.

Given that, sample variance $(s^2) = 0.498$.
Therefore, the sampling distribution is Normal$(0, 0.0014)$.
Now, 95% confidence interval for λ is $[\hat{\lambda} - \delta_1, \hat{\lambda} - \delta_2]$.
It is given that $P(-0.07 < \text{N}(0, 0.0014) < 0.07) = 0.95$, therefore,

$$\delta_1 = 0.07 \text{ and } \delta_2 = -0.07$$

Hence the 95% confidence interval for λ is $[0.40, 0.54]$.

12. The number of emails received by Neeti in intervals of one hour is given in Table 9.2.P.

No. of emails per hour	Frequency
0	5
1	15
2	22
3	22
4	17
5	10
6	5
7	3
8	1
9+	0

Table 9.2.P: Emails received by Neeti in one-hour interval for the last 100 hours.

(i) Fit a Poisson distribution to the given frequency table and find the parameter. Write your answer correct to two decimal places.

Solution:
Let $\hat{\lambda} = \bar{X}$ be an estimate of λ.

$$\text{Sample mean}(\bar{X}) = \frac{\sum n_i f_i}{\sum f_i}$$

$$\Rightarrow \bar{X} = \frac{(1 \times 15) + (2 \times 22) + (3 \times 22) + (4 \times 17) + (5 \times 10) + (6 \times 5) + (7 \times 3) + (8 \times 1)}{5 + 15 + 22 + 22 + 17 + 10 + 5 + 3 + 1}$$

$$\Rightarrow \bar{X} = \frac{302}{100} = 3.02$$

Therefore, $\hat{\lambda} = 3.02$.
Therefore, the distribution is Poisson(3.02).
We can check the fit, the same way we did in the previous question.

(ii) Find an approximate 95% confidence interval using a normal approximation for the sampling distribution.
(Use the following information:
sample variance $S^2 = 3.05$ and $P(-0.34 < \mathrm{N}(0, 0.0305) < 0.34) = 0.95$)

(a) $[1.89, 4.15]$
(b) $[2.08, 4.34]$
(c) $[2.68, 3.36]$
(d) $[1.89, 3.35]$

Solution:
Error: $\hat{\lambda} - \lambda$
$E[\hat{\lambda} - \lambda] = 0$
$\mathrm{Var}(\hat{\lambda} - \lambda) = \mathrm{Var}(\hat{\lambda}) = \dfrac{\sigma^2}{n} \approx \dfrac{S^2}{n}$
Therefore, we will assume the sampling distribution to be $\mathrm{Normal}\left(0, \dfrac{s^2}{n}\right)$.
Given that, sample variance $(s^2) = 3.05$.
Therefore, the sampling distribution is $\mathrm{Normal}(0, 0.0305)$.
Now, 95% confidence interval for λ is $[\hat{\lambda} - \delta_1, \hat{\lambda} - \delta_2]$.
It is given that $P(-0.34 < \mathrm{N}(0, 0.0305) < 0.34) = 0.95$, therefore,

$$\delta_1 = 0.34 \text{ and } \delta_2 = -0.34$$

Hence the 95% confidence interval for λ is $[2.68, 3.36]$.