Statistics for Data Science - 2

Week 9 graded Assignment

Bayesian estimation

1. Suppose that the number of buses reaching a particular stop in an one-hour time period follows the Poisson distribution with an unknown parameter λ. Previous records suggest that the prior probabilities of λ are $P(\lambda = 0.25) = 0.3$ and $P(\lambda = 0.20) = 0.7$. If in a particular one-hour time period seven buses reach the bus stop, find the posterior mode of λ. Write your answer correct to two decimal places.

   Solution:

   Prior probabilities of λ are $P(\lambda = 0.25) = 0.3$ and $P(\lambda = 0.20) = 0.7$.

   The posterior probabilities of λ will be

$$P(\lambda = 0.25|X = 7) = \frac{P(X = 7|\lambda = 0.25).P(\lambda = 0.25)}{P(X = 7)} = \frac{e^{-0.25}(0.25)^7(0.3)}{7!P(X = 7)} \qquad ...(1)$$

$$P(\lambda = 0.20|X = 7) = \frac{P(X = 7|\lambda = 0.20).P(\lambda = 0.20)}{P(X = 7)} = \frac{e^{-0.20}(0.20)^7(0.7)}{7!P(X = 7)} \qquad ...(2)$$

   Dividing equation (2) by (1), we get

$$\frac{P(\lambda = 0.20|X = 7)}{P(\lambda = 0.25|X = 7)} = \frac{e^{-0.20}(0.20)^7(0.7)}{e^{-0.25}(0.25)^7(0.3)} = \frac{7e^{0.05}4^7}{3(5^7)} < 1$$

   It implies that $P(\lambda = 0.20|X = 7) < P(\lambda = 0.25|X = 7)$

   Therefore, $\lambda = 0.25$ is the posterior mode.

---

2. Outcomes on rolling a die ten times are:
   1, 3, 4, 3, 2, 5, 4, 6, 4, 1
   Use the Uniform$[0,1]$ prior to find the posterior mean of p, which denotes the probability of getting an even number.

   Solution:
   Let $\mathbf{p}$ denote the probability of getting an even number.
   Prior distribution of $\mathbf{p}$ is $f_{\mathbf{p}} \sim$ Uniform$[0,1]$.
   It implies that $f_{\mathbf{p}}(p) = 1$

   Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \boldsymbol{p} = p) f_{\boldsymbol{p}}(p)$
   $\Rightarrow$ posterior density $\propto p^5(1-p)^5(1)$
   $\Rightarrow$ posterior density $\propto p^5(1-p)^5$
   $\Rightarrow$ posterior density $=$ Beta$(6,6)$
   $\Rightarrow$ posterior mean $= \dfrac{6}{6+6} = \dfrac{6}{12} = 0.5$

---

3. Let $X_1, X_2, \ldots, X_n \sim$ i.i.d. Exp(λ), where λ is an unknown parameter. Find the posterior mean of λ assuming the prior distribution of λ to be Exp(μ).

   (a) $\dfrac{n}{X_1 + X_2 + \ldots + X_n}$
   (b) $\dfrac{n}{\mu + X_1 + X_2 + \ldots + X_n}$
   (c) $\dfrac{n+1}{X_1 + X_2 + \ldots + X_n}$
   (d) $\dfrac{n+1}{\mu + X_1 + X_2 + \ldots + X_n}$

   Solution:
   Let Λ be the prior distribution of λ.
   From the given information, $f_\Lambda(\lambda) \sim$ Exp(μ).

   It implies that $f_\Lambda(\lambda) = \mu e^{-\mu\lambda}$.

   Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \Lambda = \lambda) f_\Lambda(\lambda)$
   $\Rightarrow$ posterior density $\propto \lambda^n e^{-\lambda(X_1+X_2+\ldots+X_n)}(\mu e^{-\mu\lambda})$
   $\Rightarrow$ posterior density $\propto \lambda^n e^{-\lambda(X_1+X_2+\ldots+X_n+\mu)}$
   $\Rightarrow$ posterior density $=$ Gamma$(n+1, X_1 + X_2 + \ldots + X_n + \mu)$
   $\Rightarrow$ posterior mean $= \dfrac{n+1}{X_1 + X_2 + \ldots + X_n + \mu}$

---

4. Call duration of daily stand up meetings of employees of a certain company follows the exponential distribution with an unknown parameter λ. Duration (in minutes) of last ten meetings are 20, 30, 35, 30, 25, 25, 20, 28, 34, 30. Find the Bayesian estimate (posterior mean) of λ using the prior distribution of Exp$(\frac{1}{15})$ for λ. Write your answer correct to two decimal places.

   Solution:
   Let Λ be the prior distribution of λ.
   From the given information, $f_\Lambda(\lambda) \sim \text{Exp}(\frac{1}{15})$.

   It implies that $f_\Lambda(\lambda) = \frac{1}{15}e^{-\lambda/15}$.

   Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n|\Lambda = \lambda)f_\Lambda(\lambda)$
   $\Rightarrow$ posterior density $\propto \lambda^n e^{-\lambda(X_1+X_2+\ldots+X_n)}(\frac{1}{15}e^{-\lambda/15})$
   $\Rightarrow$ posterior density $\lambda^n e^{-\lambda(X_1+X_2+\ldots+X_n+\frac{1}{15})}$
   $\Rightarrow$ posterior density $=$ Gamma$(n+1, X_1 + X_2 + \ldots + X_n + \frac{1}{15})$

   $$\Rightarrow \text{posterior mean} = \frac{n+1}{X_1 + X_2 + \ldots + X_n + \frac{1}{15}} = \frac{11}{277 + \frac{1}{15}}$$

   $$\Rightarrow \text{posterior mean} = \frac{11 \times 15}{15 \times 277 + 1} = 0.03$$

---

5. Marks of tenth class students of a school follow the normal distribution with an unknown mean μ and variance 25. Marks of 10 students of the tenth class are 50, 45, 70, 60, 75, 90, 45, 60, 80, 75. Find the Bayesian estimate (posterior mean) of μ assuming the Normal$(50, 25)$ prior distribution. Write your answer correct to two decimal places.

   Solution:
   We know that normal distribution is conjugate to the normal distribution. That is if prior distribution of μ is normal(μ_0, σ_0^2) and sample is taken from Normal(μ, σ), then posterior distribution of the μ will be Normal with mean $\overline{X}\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \frac{\mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$

   Here, $\overline{X} = 65$, $n = 10$, $\mu_0 = 50$, $\sigma_0^2 = 25$, $\sigma^2 = 25$

Therefore,

$$\begin{aligned}
\text{Posterior mean} &= \frac{65 \times 10 \times 25}{10(25)+25} + \frac{50 \times 25}{10(25)+25} \\
&= \frac{16250}{275} + \frac{1250}{275} \\
&= 59.09 + 4.545 = 63.63
\end{aligned}$$

---

6. The outcomes on tossing a coin ten times are: H T T H T H H H T H. Let p be the probability of heads. Previous records show that heads appear on an average 40% of the time. Find the posterior mean of p using the Beta$(2, \beta)$ prior. Write your answer correct to two decimal places.

Solution:
Let $\boldsymbol{p}$ denote the probability of heads.
Given that prior of $\boldsymbol{p}$ is Beta$(2, \beta)$ with an average of 0.4.
It implies that $E[\text{Beta}(2, \beta)] = 0.4$
$\Rightarrow \dfrac{2}{2+\beta} = 0.4$
$\Rightarrow \beta = 3$

Therefore, prior distribution of $\boldsymbol{p}$ is Beta$(2, 3)$
It implies that $f_{\boldsymbol{p}}(p) \propto p^1(1-p)^2$

Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \boldsymbol{p} = p) f_{\boldsymbol{p}}(p)$
$\Rightarrow$ posterior density $\propto p^6(1-p)^4(p^1(1-p)^2)$
$\Rightarrow$ posterior density $\propto p^7(1-p)^6$
$\Rightarrow$ posterior density $=$ Beta$(8, 7)$
$\Rightarrow$ posterior mean $= \dfrac{8}{8+7} = \dfrac{8}{15} = 0.53$

---

7. One out of the last ten candidates wins a treasure hunt game. Previous record shows fraction of winners follows the Beta$(20, b)$ distribution with an average of 20%. Estimate the long-term fraction of winners of the treasure hunt game. Write your answer correct to two decimal places.

Solution:
Let the long-term fraction of winners (probability of winning) be denoted by $\boldsymbol{p}$.
Previous data shows that fraction of winners follows the Beta$(20, b)$ distribution with an average of 20%.
It implies that $E[\text{Beta}(20, b)] = 0.2$

$\Rightarrow \dfrac{20}{20+b} = 0.2$
$\Rightarrow b = 80$

Therefore, prior distribution of $\boldsymbol{p}$ is Beta$(20, 80)$
It implies that $f_{\boldsymbol{p}}(p) \propto p^{19}(1-p)^{79}$

Now, posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \boldsymbol{p} = p) f_{\boldsymbol{p}}(p)$
$\Rightarrow$ posterior density $\propto p^1(1-p)^9(p^{19}(1-p)^{79})$
$\Rightarrow$ posterior density $\propto p^{20}(1-p)^{88}$
$\Rightarrow$ posterior density $=$ Beta$(21, 89)$
$\Rightarrow$ posterior mean $= \dfrac{21}{21+89} = \dfrac{21}{110} = 0.19$

---

8. Rainfall in the monsoon season in Delhi follows normal distribution with mean μ and variance 200 mm. Rainfall (in mm) registered in the 2021 monsoon are 600, 300, 450, 700, 850, 150, 200, 750. Prior information about the average rainfall is that it has mean 600 mm and variance 225 mm. Use the normal prior that matches your prior information and find the posterior mean.

Solution:
Prior distribution of μ is given Normal with mean 600 and variance 225.
We know that normal distribution is conjugate to the normal distribution. That is if prior distribution of μ is normal(μ_0, σ_0^2) and sample is taken from Normal(μ, σ^2), then posterior distribution of the μ will be Normal with mean $\overline{X}\dfrac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \dfrac{\mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$

Here, $\overline{X} = 500$, $n = 8$, $\mu_0 = 600$, $\sigma_0^2 = 225$, $\sigma^2 = 200$

Therefore,

$$\begin{aligned}\text{Posterior mean} &= \frac{500 \times 8 \times 225}{8(225) + 200} + \frac{600 \times 200}{8(225) + 200} \\ &= \frac{900000}{2000} + \frac{120000}{2000} \\ &= 450 + 60 = 510\end{aligned}$$

---

9. Following frequency data shows the number of patients (n) arriving in an emergency room between 12:00 AM and 6:00 AM.

n	frequency	n	frequency
0	1	6	14
1	4	7	4
2	17	8	4
3	17	9	1
4	17	10+	0
5	21		

(i) Fit the data into Poisson distribution (Find the parameter). Write your answer correct to two decimal places.

Solution:

We know that $\hat{\lambda} = \overline{X}$ is an estimate of λ.

$$\begin{aligned}\text{Sample mean, } \overline{X} &= \frac{\sum_i f_i n_i}{\sum_i f_i} \\ &= \frac{0+4+34+51+68+105+84+28+32+9}{1+4+17+17+17+21+14+4+4+1} \\ &= \frac{415}{100} = 4.15\end{aligned}$$

Therefore, $\hat{\lambda} = 4.15$

(ii) Find an approximate 95% confidence interval using a normal approximation for the error distribution.
(Use the following information:
sample variance $S^2 = 3.40$ and $P(-0.36 < \mathrm{N}(0, 0.034) < 0.36) = 0.95$)

(a) $[3.87, 5, 134]$
(b) $[3.79, 4.51]$
(c) $[3.12, 5.21]$
(d) $[4.01, 5.23]$

Solution:

Error, e is given by
$e = \hat{\lambda} - \lambda$
Now, $E[\hat{\lambda} - \lambda] = E[\hat{\lambda}] - \lambda = \lambda - \lambda = 0$
$\mathrm{Var}(\hat{\lambda} - \lambda) = \mathrm{Var}(\lambda) = \dfrac{\sigma^2}{n} \approx \dfrac{s^2}{n} = \dfrac{3.4}{100} = 0.034$

It implies that error follows Normal$(0, 0.034)$.

Let 95% confidence interval be $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$. Now,

$$\begin{aligned}&P(|\text{error}| < \delta) = 0.95 \\ \Rightarrow &P(|\text{Normal}(0, 0.034)| < \delta) = 0.95\end{aligned}$$

It is given that $P(-0.36 < \mathrm{N}(0, 0.034) < 0.36) = 0.95$
Therefore $\delta = 0.36$

So, 95% confidence interval will be $[3.79, 4.51]$.

---

10. Following frequency table shows the number of bankruptcies (n) filed by customers in a time period of one month. The data consists of last 200 months.

n	frequency	n	frequency
0	13	6	8
1	26	7	4
2	48	8	1
3	44	9	1
4	39	10+	0
5	16		

(i) Fit the data into Poisson distribution (Find the parameter). Write your answer correct to two decimal places.

Solution:

We know that $\hat{\lambda} = \overline{X}$ is an estimate of λ.

$$\text{Sample mean, } \overline{X} = \frac{\sum_i f_i n_i}{\sum_i f_i}$$

$$= \frac{0 + 26 + 96 + 132 + 156 + 80 + 48 + 28 + 8 + 9}{13 + 26 + 48 + 44 + 39 + 16 + 8 + 4 + 1 + 1}$$

$$= \frac{583}{200} = 2.91$$

Therefore, $\hat{\lambda} = 2.91$

(ii) Find an approximate 95% confidence interval using a normal approximation for the error distribution.
(Use the following information:
sample variance $S^2 = 2.852$ and $P(-0.23 < \mathrm{N}(0, 0.0142) < 0.23) = 0.95$)

(a) $[1.97, 4.14]$
(b) $[2.08, 3.34]$
(c) $[2.68, 3.14]$
(d) $[2.01, 4.232]$

Solution:

Error, e is given by
$e = \hat{\lambda} - \lambda$

Now, $E[\hat{\lambda} - \lambda] = E[\hat{\lambda}] - \lambda = \lambda - \lambda = 0$
$\mathrm{Var}(\hat{\lambda} - \lambda) = \mathrm{Var}(\lambda) = \dfrac{\sigma^2}{n} \approx \dfrac{s^2}{n} = \dfrac{2.852}{200} = 0.0142$

It implies that error follows Normal$(0, 0.0142)$.

Let 95% confidence interval be $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$. Now,

$$\begin{aligned} &P(|\text{error}| < \delta) = 0.95 \\ \Rightarrow &P(|\text{Normal}(0, 0.0142)| < \delta) = 0.95 \end{aligned}$$

It is given that $P(-0.23 < \mathrm{N}(0, 0.0142) < 0.23) = 0.95$
Therefore $\delta = 0.23$

So, 95% confidence interval will be $[2.68, 3.14]$