MA2261 Linear Statistical Models - DLI, Year 2022-2023

Coursework 1

INSTRUCTIONS AND DEADLINE:

Please submit *electronically* one piece of written/typed work per person in a single PDF file by **Friday 24/Mar/2023 at 4pm UK time/23:59 China time**.

Please use this page as the cover page for your submission. Write below your student ID and sign it.

Student ID:

Signature:

MARKING CRITERIA:

- This problem sheet is worth 100 points. Scores for each main question are indicated at the beginning of each.
- Clearly justify and explain your answers. If you are using the R software for calculations, a printout of your answers without a full explanation of the formulas you are using and your reasoning will not score full marks.
- A true/false question answered without justification will get zero marks.
- Computational mistakes will be penalized more in coursework than in exam marking, since you have plenty of time and tools to check your calculations when doing the coursework.

Question 1 [15 marks]

Decide if the following statements are true or false. Justify your answers.

i) A power set must be a σ-algebra, vice versa.

ii) There is a 50% chance of rain on Earth and a 10% chance of rain on Mars. Therefore, there is a 45% chance that it will rain in neither place.

iii) Let $f(x)$ be a density function of a continuous random variable. Then it must have $\int_{-\infty}^{\infty} x\, f(x) dx = 1$.

iv) The central limit theorem ensures that the sampling distribution of the sample mean approaches normal as the sample size increases.

v) The chi-square distribution is left-skewed and can take on values between $-\infty$ and ∞.

Question 2 [15 marks]

A factory production line is manufacturing bolts using three machines, A, B and C. Of the total output, machine A is responsible for 25%, machine B for 35% and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B and 2% from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from

i) machine A?

ii) machine B?

iii) machine C?

Question 3 [15 marks]

Among 10000 random digits, find the probability P that the digit 3 appears:

i) 9998 times.

ii) at least 975 and no more than 1025 times.

iii) at most 950 times.

Question 4 [25 marks]

A continuous random variable X has the following density function:

$$f_X(x) = \begin{cases} (C-\alpha)x^{1-\alpha}, & 0 \leqslant x \leqslant 1 \\ 0, & \text{otherwise.} \end{cases}$$

where α is an unknown parameter satisfying $\alpha < 2$ and $\alpha \neq 1$.

i) Determine the value of the constant C.

ii) Find $E[X]$ and $Var(X)$.

iii) For independent observations $x_1, \ldots, x_n$ of X, write down the likelihood function, where the aim is to find an estimate for the parameter α.

iv) Find the log-likelihood function.

v) Find the maximum likelihood estimate for α in terms of x_i, $i = 1, \ldots, n$.

Question 5 [20 marks]

Assume that the weight of a product follows a normal distribution with unknown mean and variance. Now we randomly selected 16 pieces from a batch of products. The sample average weight is 820g, and the sample standard deviation is 60g. Decide whether to reject $H_0 : \mu = 800$g when the significance level is $\alpha = 0.05$.

Question 6 [10 marks]

Suppose that $X_1, \ldots, X_{10}$ are random variables, each an observation from the same distribution $X \sim N(\mu, 4)$, where μ is unknown. Suppose that the following values $x_1, x_2, \ldots, x_{10}$ are recorded:

$$12.4, \quad 3.5, \quad 9.8, \quad 11.9, \quad 16.5, \quad 9.6, \quad 10.9, \quad 11.1, \quad 2.9, \quad 5.8$$

i) Find a 95% confidence interval for μ.

ii) Find a 99% confidence interval for μ.

Solution to Question 1

i) **[3 marks]** False. Power set is just a special σ-algebra.

ii) **[3 marks]** True. Let E be the event that it rains here and M be the event it rains on Mars. The two events are independent. $P(E^c \cap M^c) = P(E^c)P(M^c) = (1 - 0.5)(1 - 0.1) = 0.45$.

iii) **[3 marks]** False. For a density function, it satisfies $\int_{-\infty}^{\infty} f(x)dx = 1$.

iv) **[3 marks]** True. According to the central limit theorem, as the sample size approaches infinity, the sampling distribution of the mean approaches a normal distribution.

v) **[3 marks]** False. The chi-square distribution is right-skewed and can take on only positive values.

Solution to Question 2

Let

- $D = \{\text{bolt is defective}\}$,
- $A = \{\text{bolt is from machine } A\}$,
- $B = \{\text{bolt is from machine } B\}$,
- $C = \{\text{bolt is from machine } C\}$.

We know that $P(A) = 0.25$, $P(B) = 0.35$ and $P(C) = 0.4$.
Also, $P(D|A) = 0.05$, $P(D|B) = 0.04$, $P(D|C) = 0.02$.

i) **[5 marks]** Then,

$$\begin{aligned} P(A|D) &= \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.362 \end{aligned}$$

ii) **[5 marks]**

$$\begin{aligned} P(B|D) &= \frac{P(D|B)P(B)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.406 \end{aligned}$$

iii) **[5 marks]**

$$\begin{aligned} P(C|D) &= \frac{P(D|C)P(C)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.02 \times 0.4}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4} \\ &= 0.232 \end{aligned}$$

Solution to Question 3

i) **[5 marks]** Let X denote the number of times that 3 appears. This is a binomial experiment $X \sim Bin(n, p)$ with $n = 10000$ and $p = 0.1$. Then,

$$\begin{aligned} \mu &= np = 10000(0.1) = 1000 \\ \sigma^2 &= np(1-p) = 10000(0.1)(0.9) = 900, \quad \sigma = 30 \end{aligned}$$

Then, $P(X = 9998) = \binom{10000}{9998} 0.1^{9998}(1 - 0.1)^2 = 0.4049595 \times 10^{-9990} \approx 0$.

Alternatively, by central limit theorem, let $Z = \frac{X+0.5-\mu}{\sigma} \sim N(0, 1)$, we have

$$\begin{aligned} P(X = 9998) &= P(X \leqslant 9998) - P(X \leqslant 9997) \\ &= P\left(Z \leqslant \frac{9998.5 - 1000}{30}\right) - P\left(Z \leqslant \frac{9997.5 - 1000}{30}\right) \\ &= \Phi(299.95) - \Phi(299.916) \\ &\approx 0 \end{aligned}$$

Students need only use either method.

ii) **[5 marks]** We seek $P(975 \leq X \leq 1025)$. By central limit theorem, we have $Z = \frac{X+0.5-\mu}{\sigma} \sim N(0, 1)$, and

$$\begin{aligned} P(975 \leq X \leq 1025) &\approx P\left(\frac{974.5 - \mu}{\sigma} \leq Z \leq \frac{1025.5 - \mu}{\sigma}\right) \\ &= P\left(\frac{974.5 - 1000}{30} \leq Z \leq \frac{1025.5 - 1000}{30}\right) \\ &= P(-0.85 \leq Z \leq 0.85) \\ &= \Phi(0.85) - \Phi(-0.85) \\ &= 2\Phi(0.85) - 1 = 2(0.80234) - 1 = 0.60468 \end{aligned}$$

iii) **[5 marks]** We seek the probability $P(X \leq 950)$.

$$\begin{aligned} P(X \leq 950) &\approx P\left(Z \leq \frac{950.5 - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{950.5 - 1000}{30}\right) \\ &= P(Z \leq -1.65) \\ &= \Phi(-1.65) = 1 - \Phi(1.65) = 1 - 0.95053 = 0.04947 \end{aligned}$$

Solution to Question 4

i) **[5 marks]** Since $f_X(x)$ is the density of X, it must satisfy that

$$\int_{\mathbb{R}} f_X(x)dx = \int_{-\infty}^{\infty} f_X(x)dx = 1$$

Hence,

$$\int_0^1 (C-\alpha)x^{1-\alpha}dx = 1$$

Solving

$$\left[\frac{C-\alpha}{2-\alpha}x^{2-\alpha}\right]_0^1 = \frac{C-\alpha}{2-\alpha} = 1 \quad \Rightarrow \quad C = 2$$

ii) **[6 marks]**

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x(2-\alpha)x^{1-\alpha}dx = \frac{2-\alpha}{3-\alpha}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 x^2(2-\alpha)x^{1-\alpha}dx = \frac{2-\alpha}{4-\alpha}$$

$$Var(X) = E[X^2] - (E[X])^2 = \frac{2-\alpha}{4-\alpha} - \left(\frac{2-\alpha}{3-\alpha}\right)^2 = \frac{2-\alpha}{(3-\alpha)^2(4-\alpha)}$$

iii) **[5 marks]** The likelihood function is

$$\mathcal{L}(\alpha, x_1, \ldots, x_n) = \prod_{k=1}^{n} f_X(x_k, \alpha) = \prod_{k=1}^{n}(2-\alpha)x_k^{1-\alpha} = (2-\alpha)^n \prod_{k=1}^{n} x_k^{1-\alpha}$$

iv) **[4 marks]** Hence, the log likelihood:

$$\log \mathcal{L} = n \log(2-\alpha) + (1-\alpha)\sum_{k=1}^{n} \log(x_k)$$

v) **[5 marks]** Set

$$\frac{d \log \mathcal{L}}{d\alpha} = -\frac{n}{2-\alpha} - \sum_{k=1}^{n} \log(x_k) = 0$$

We then have

$$\hat{\alpha} = 2 + \frac{n}{\sum_{k=1}^{n} \log(x_k)}$$

Check the second-order derivative,

$$\frac{d^2 \log \mathcal{L}}{d\alpha^2} = -\frac{n}{(2-\alpha)^2} < 0$$

for all $\alpha < 2$. Hence, $\hat{\alpha}$ is the MLE for α.

Solution to Question 5

Step a) **[5 marks]** We want to test the null hypothesis $H_0 : \mu = 800$ against $H_1 : \mu \neq 800$. (Here we use the two-tailed hypothesis test.)

Step b) **[5 marks]** As σ is unknown, the statistic is $t = \frac{\overline{x}-\mu_0}{S/\sqrt{n}} = \frac{820-800}{60/\sqrt{16}} = 1.333 \sim t_{15}$. (Here d.f.$= n - 1 = 15$)

Step c) **[5 marks]** When the significance level is $\alpha = 0.05$, the rejection region is $(-\infty, -t_{0.025,15}) \cup (t_{0.025,15}, +\infty) = (-\infty, -2.131) \cup (2.131, +\infty)$.

Alternatively, p-value $= P(T < -1.333) + P(T > 1.333) = 0.2023$.

Step d) **[5 marks]** $t = 1.333$ is outside the rejection region, and p-value is greater than the significance level $\alpha = 0.05$. Thus, we accept H_0.

Solution to Question 6

i) **[5 marks]** When $n = 10$ and $\sigma = 2$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{2/\sqrt{10}} \sim N(0, 1)$$

From the data, we get $\overline{x} = 9.44$. Hence the 95% CI for μ is

$$(\overline{x} - Z_{0.025}\frac{\sigma}{\sqrt{n}},\ \overline{x} + Z_{0.025}\frac{\sigma}{\sqrt{n}}) = (9.44 - 1.96\frac{2}{\sqrt{10}},\ 9.44 + 1.96\frac{2}{\sqrt{10}}) \approx (8.2004, 10.6796)$$

ii) **[5 marks]** The 99% CI for μ is

$$(\overline{x} - Z_{0.005}\frac{\sigma}{\sqrt{n}},\ \overline{x} + Z_{0.005}\frac{\sigma}{\sqrt{n}}) = (9.44 - 2.576\frac{2}{\sqrt{10}},\ 9.44 + 2.576\frac{2}{\sqrt{10}}) \approx (7.8108, 11.0692)$$