Sample solutions to exam 22-05-31

Question 1: Let X be a random variable with probability density function $f_X(x) = 3x^2$ for $0 \leqslant x \leqslant 1$, and $f_X(x) = 0$ otherwise.

a) Calculate $E[X]$ and $V[X]$. (2p)

b) Determine $P(X = \frac{1}{2})$ and $P(X \geqslant \frac{1}{2})$. (2p)

c) Calculate $C(X, X^2)$, the covariance between the random variables X and X^2. (2p)

Solution:

a) $E[X] = \int_0^1 x \cdot 3x^2 \mathrm{d}x = [\frac{3}{4}x^4]_{x=0}^1 = \frac{3}{4} - 0 = \frac{3}{4} = 0.75$
$E[X^2] = \int_0^1 x^2 \cdot 3x^2 \mathrm{d}x = [\frac{3}{5}x^5]_{x=0}^1 = \frac{3}{5} - 0 = \frac{3}{5} = 0.6$
$V[X] = E[X^2] - E[X]^2 = 0.6 - 0.75^2 = 0.0375$

b) $P(X = \frac{1}{2}) = \int_{0.5}^{0.5} 3x^2 \mathrm{d}x = 0$; $P(X \geqslant \frac{1}{2}) = \int_{0.5}^{1} 3x^2 \mathrm{d}x = [x^3]_{x=0.5}^1 = 1^3 - 0.5^3 = 0.875$

c) $C(X, X^2) = E[X \cdot X^2] - E[X]E[X^2] = E[X^3] - E[X]E[X^2]$. We already calculated $E[X]$ and $E[X^2]$ above, so we are only missing $E[X^3]$.
$E[X^3] = \int_0^1 x^3 \cdot 3x^2 \mathrm{d}x = [\frac{3}{6}x^6]_{x=0}^1 = \frac{1}{2} - 0 = 0.5$
$C(X, X^2) = E[X^3] - E[X]E[X^2] = 0.5 - 0.75 \cdot 0.6 = 0.05$

Question 2: Let X be a random variable with probability density function

$$f(x) = \frac{2}{9}x \qquad \text{for } 0 \leqslant x \leqslant 3$$

and $f(x) = 0$ otherwise. Simulate three pseudorandom numbers from the distribution of X with the help of the following pseudorandom numbers from Re$[0, 1]$.

$$u_1 = 0.5644 \qquad u_2 = 0.2375 \qquad u_3 = 0.9252$$

(5p)

Solution: We use the inversion method, using the fact that if U is uniform from $[0, 1]$, then $F_X^{-1}(U)$ has the same distribution as X. First we find the distribution function F:

$$F(x) = \int_{-\infty}^{x} f(x)\mathrm{d}y = \begin{cases} 0 & \text{if } x < 0 \\ \int_0^x \frac{2}{9}y\,\mathrm{d}y = [\frac{1}{9}y^2]_{y=0}^x = \frac{1}{9}x^2 & \text{if } 0 \leqslant x \leqslant 3 \\ \int_0^1 \frac{2}{9}y\,\mathrm{d}y = 1 \end{cases}$$

Viewing F as a function from $[0, 3]$ to $[0, 1]$, to find the inverse, set $y = F(x)$ and solve for x:

$$y = \frac{1}{9}x^2, \text{ so } x^2 = 9y, \text{ so } x = 3\sqrt{y}$$

(As $0 \leqslant y \leqslant 1$ and $0 \leqslant x \leqslant 3$, $x = 3\sqrt{y}$ is the only solution.)

Applying $F^{-1}(x) = 3\sqrt{x}$ to the pseudorandom numbers from Re$[0, 1]$ gives the following three simulated pseudorandom numbers.

$$t_1 = 3\sqrt{0.5644} \approx 2.254 \qquad t_2 = 3\sqrt{0.2375} \approx 1.462 \qquad t_3 = 3\sqrt{0.9252} \approx 2.886$$

Question 3: Let $X_1 \sim N(\mu, 4)$ and $X_2 \sim N(2\mu, 9)$ be two independent random variables, where μ is an unknown parameter.

a) Show that $T_1 = \frac{1}{2}X_1 + \frac{1}{4}X_2$ is an unbiased estimator for μ. (2p)

b) Calculate the standard deviation of T_1. (1p)

c) Let $T_2 = 2X_1 - cX_2$, where c is a constant. How do we need to choose c so that T_2 is an unbiased estimator for μ? For this c, which estimator is more efficient, T_1 or T_2? (3p)

Solution:

a) $E[T_1] = \frac{1}{2}E[X_1] + \frac{1}{4}E[X_2] = \frac{1}{2}\mu + \frac{1}{4}2\mu = \mu$, so T_1 is unbiased.

b) As X_1, X_2 are independent, $V[\frac{1}{2}X_1 + \frac{1}{4}X_2] = \frac{1}{4}V[X_1] + \frac{1}{16}V[X_2] = 1 + \frac{9}{16} = \frac{25}{16}$, so $D[T_1] = \sqrt{V[T_1]} = \frac{5}{4}$.

c) $E[T_2] = 2E[X_1] - cE[X_2] = 2\mu - 2c\mu$, which is supposed to be μ for T_2 to be unbiased. So $2\mu - 2c\mu = \mu$, which gives $2 - 2c = 1$, which gives $c = 0.5$.

For $T_2 = 2X_1 - 0.5X_2$, as X_1, X_2 are independent, we can calculate

$$V[T_2] = 4V[X_1] + (-0.5)^2V[X_2] = 4 \cdot 4 + 0.25 \cdot 9 = 18.25 > \frac{25}{16} = V[T_1],$$

so T_1 is more efficient.

Question 4: The latency of the response time of a company's server is assumed to be normally distributed with unknown expectation and variance. The company wants to measure the latency and makes 20 requests to the server. Each time the latency is measured in milliseconds, giving data points $x_1, \ldots, x_{20}$ which are assumed to be independent measurements. From this, the sample mean is calculated to be $\bar{x} = 42$ (milliseconds), and the sample standard deviation is calculated as $s = 10$ (milliseconds).

a) Give a confidence interval for the latency of the server at confidence level 95%. (2p)

b) Now assume that the standard deviation is known to be exactly $\sigma = 10$. How many observations would you need to give a confidence interval for the latency at confidence level 95% of length at most 5? (3p)

Solution

a) The confidence interval for a normal distribution of unknown variance is

$$I = [\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}],$$

where $n = 20$, $\alpha = 0.05$, $\bar{x} = 42$, $s = 10$ and we look up $t_{0.025}(19) \approx 2.09$ in the quantile table for Student's t-distribution. Plugging in, we get

$$I = [37.35, 46.67].$$

b) The central confidence interval for a normal distribution of known variance is

$$I = [\bar{x} \pm \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}],$$

which has length $L = 2\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We have $\alpha = 0.05$, $\sigma = 10$, and $\lambda_{0.025} \approx 1.96$ from the quantile table of the normal distribution. If we want $L \leqslant 5$, we need

$$n = \left(\frac{2\lambda_{0.025}\sigma}{L}\right)^2 \geqslant \left(\frac{2\lambda_{0.025}\sigma}{5}\right)^2 \approx 61.5,$$

so we need at least $n \geqslant 62$ observations.

Question 5: In a social network, most users are real humans, but about 5% of all accounts are bot accounts. The average number of interactions that a human user makes per day is assumed to be normally distributed with expectation 18 and standard deviation 5, independently from all other accounts. Bot accounts have considerably more interactions; their daily average is assumed to be normally distributed with expectation 28 and standard deviation 5, independently from all other accounts.

Introduce a suitable model, defining appropriate random variables, to answer the following questions:

a) What is the probability that a given human user has an average daily interaction rate of more than 30? What the probability that a given bot account has an average daily interaction rate of more than 30? (3p)

b) The company behind the social network reviews a radical suggestion to reduce the number of bot accounts: Delete all accounts which on average have more than 30 interactions per day. What is the probability that an account with an average daily interaction rate of over 30 belongs to a human? (3p)

Solutions:

a) Let X_1 be the daily interaction rate of a given human user, and X_2 the daily interaction rate of a given bot account. Then we have $X_1 \sim N(18, 25)$ and $X_2 \sim N(28, 25)$. So, as $Y = (X_1 - 16)/5$ has the distribution $N(0, 1)$,

$$\begin{aligned} P(X_1 > 30) &= P\left(\frac{X_1 - 18}{5} > \frac{30 - 18}{5}\right) = P(Y > 2.4) = 1 - \Phi(2.4) \approx 1 - 0.9918 \\ &= 0.0082 = 0.82\%, \end{aligned}$$

and analogously

$$P(X_2 > 30) = 1 - \Phi\left(\frac{30 - 28}{5}\right) = 1 - \Phi(0.4) \approx 1 - 0.6554 = 0.3446 = 34.46\%.$$

b) Choose an account uniformly at random from all accounts on the social network, and let X be its average daily interaction rate. Let H be the event that the account is human, and B the event that it is a bot account. Then

$$P(H) = 0.95 \quad \text{and} \quad P(B) = 0.05.$$

From part a), we have

$$P(X > 30|H) \approx 0.0082 \quad \text{and} \quad P(X > 30|B) \approx 0.3446.$$

We are looking for $P(H|X > 30)$. We use Bayes' Theorem, together with the law of total probability:

$$\begin{aligned} P(H|X > 30) &= \frac{P(H)P(X > 30|H)}{P(X > 30)} = \frac{P(H)P(X > 30|H)}{P(H)P(X > 30|H) + P(B)P(X > 30|B)} \\ &\approx \frac{0.95 \cdot 0.0082}{0.95 \cdot 0.0082 + 0.05 \cdot 0.3446} \approx 0.31 = 31\% \end{aligned}$$

Question 6: Let $X_0, X_1, X_2, ...$ be a Markov chain with state space $E = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 2/3 & 0 & * \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

The Markov chain starts at either state 0 or state 2 with equal probability.

a) What is the value $*$ which is missing from the matrix, and why? (1p)

b) Calculate $P(X_2 = 1)$. (2p)

c) It can be shown that there is a unique stationary distribution for the transition matrix $\mathbf{P}$. Determine this stationary distribution. (3p)

Solution:

a) The elements in each row of the matrix sum to 1, so for row 2, we have $\frac{2}{3} + 0 + * = 1$, so $* = \frac{1}{3}$.

b) The initial distribution of X_0 is represented by the vector $p^{(0)} = \begin{pmatrix} 0.5 & 0 & 0.5 \end{pmatrix}$. Then

$$p^{(1)} = p^{(0)}\mathbf{P} = \begin{pmatrix} 0.5 & 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 1 \\ 2/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 0 & 0.25 & 0.75 \end{pmatrix}$$

and

$$p^{(2)} = p^{(1)}\mathbf{P} = \begin{pmatrix} 0 & 0.25 & 0.75 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 1 \\ 2/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

and so the second element of $p^{(2)}$ — which is $P(X_2 = 1)$ — is $0.75 \cdot \frac{1}{2} = 0.375$.

c) We need to find $a, b, c \in [0, 1]$ with $a + b + c = 1$ so that

$$\begin{pmatrix} a & b & c \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 1 \\ 2/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} a & b & c \end{pmatrix}$$

This gives $\frac{2}{3}b = a$, $\frac{1}{2}c = b$ and $a + \frac{1}{3}b + \frac{1}{2}c = c$. Plugging the first two relations into $a+b+c = 1$ gives $\frac{2}{3} \cdot \frac{1}{2}c + \frac{1}{2}c + c = 1$, which solves to $c = \frac{6}{11}$, $b = \frac{1}{2}c = \frac{3}{11}$, $a = \frac{2}{3}b = \frac{2}{11}$. So the stationary distribution is represented by the vector

$$\begin{pmatrix} \frac{2}{11} & \frac{3}{11} & \frac{6}{11} \end{pmatrix}$$

Question 7: In a computer network, there are two servers A and B where security incidents may occur. The security incidents over time can be modeled as two independent Poisson processes $\{N_A(t) : t \geqslant 0\}$ and $\{N_B(t) : t \geqslant 0\}$ with intensities $\lambda_A = 0.001$ per hour (for server A) and $\lambda_B = 0.003$ per hour (for server B).

a) What is the probability that within one day, there is no security incident at server A? (2p)

b) What is the probability that there are at least three security incidents in total (from both servers A and B together) within ten days? (2p)

c) The network had exactly one security incident today. What is the probability that the security incident happened at server A? (2p)

Solution:

a) We want $P(N_A(24) = 0)$, where $N_A(24)$ is the number of security incidents in 24 hours. We have $N_A(24) \sim \text{Po}(24 \cdot 0.001) = \text{Po}(0.024)$, and so

$$P(N_A(24) = 0) = e^{-0.024} 0.024^0/0! = e^{-0.024} \approx 0.976$$

b) As $\{N_A(t) : t \geqslant 0\}$ and $\{N_B(t) : t \geqslant 0\}$ are independent Poisson processes, by the superposition property $M(t) = N_A(t) + N_B(t)$ is also a Poisson process, with intensity $\lambda = \lambda_A + \lambda_B = 0.004$. We are interested in the number of incidents in ten days, which is $M(240) \sim \text{Po}(240 \cdot \lambda) = \text{Po}(0.96)$. So

$$\begin{aligned} P(M(240) \geqslant 3) &= 1 - P(M(240) \leqslant 2) = 1 - e^{-0.96}\left(\frac{0.96^0}{0!} + \frac{0.96^1}{1!} + \frac{0.96^2}{2!}\right) \\ &\approx 0.0731 \end{aligned}$$

c) We know that $M(24) = N_A(24) + N_B(24) = 1$, and want, conditional on this, the probability that $N_A(24) = 1$. We have $N_A(24) \sim \text{Po}(24 \cdot 0.001) = \text{Po}(0.024)$, $N_B(24) \sim \text{Po}(24 \cdot 0.003) = \text{Po}(0.072)$, $M(24) \sim \text{Po}(0.096)$, and $N_A(24)$ and $N_B(24)$ are independent, so

$$\begin{aligned} P(N_A(24) = 1 | M(24) = 1) &= \frac{P\big(\{N_A(24) = 1\} \cap \{M(24) = 1\}\big)}{P(M(24) = 1)} \\ &= \frac{P\big(\{N_A(24) = 1\} \cap \{N_B(24) = 0\}\big)}{P(M(24) = 1)} \\ &= \frac{P\big(N_A(24) = 1\big) P\big(N_B(24) = 0\big)}{P(M(24) = 1)} \\ &= \frac{e^{-0.024}\frac{0.024^1}{1!} e^{-0.072}\frac{0.072^0}{0!}}{e^{-0.096}\frac{0.096^1}{1!}} = \frac{0.024}{0.096} = 0.25 \end{aligned}$$