Sample solutions to exam 22-05-31

Fråga 1

a) We have $p(2) = 1 - 0.1 - 0.3 - 0.2 = 0.4$, so $E[X] = 0.1 \cdot 1 + 0.4 \cdot 2 + 0.3 \cdot 3 + 0.2 \cdot 4 = 2.6$, $E[X^2] = 0.1 \cdot 1^2 + 0.4 \cdot 2^2 + 0.3 \cdot 3^2 + 0.2 \cdot 4^2 = 7.6$, so $V[X] = E[X^2] - E[X]^2 = 7.6 - 2.6^2 = 0.84$

b) $P(X \leqslant 3 \mid X \geqslant 3) = \frac{P(X \leqslant 3 \text{ and } X \geqslant 3)}{P(X \geqslant 3)} = \frac{P(X=3)}{P(X \geqslant 3)} = \frac{0.3}{0.3+0.2} = 0.6$

c) No, they are not independent: If they were, we would have $P(X \leqslant 3 \mid X \geqslant 3) = P(X \geqslant 3)$, but $P(X \leqslant 3) = 0.1 + 0.4 + 0.3 = 0.8$, and we calculated in b) that $P(X \geqslant 3 \mid X \leqslant 3) = 0.6$.

d) $C(X, 1/X) = E[X \cdot 1/X] - E[X]E[1/X] = 1 - E[X]E[1/X]$, so we need to calculate $E[1/X] = 0.1 \cdot \frac{1}{1} + 0.4 \cdot \frac{1}{2} + 0.3 \cdot \frac{1}{3} + 0.2 \cdot \frac{1}{4} = 0.45$. This gives $C(X, 1/X) = 1 - 2.6 \cdot 0.45 = -0.17$

Fråga 2

a) $F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y = \int_0^x \frac{1}{4}y^3 \mathrm{d}y = [\frac{1}{16}y^4]_0^x = \frac{1}{16}x^4$ for $0 \leqslant x \leqslant 2$ (and $F(x) = 0$ for $x < 0$, $F(x) = 1$ for $x > 2$). $E[X] = \int_0^2 x \cdot \frac{1}{4}x^3 \mathrm{d}x = \int_0^2 \frac{1}{4}x^4 \mathrm{d}x = [\frac{1}{20}x^5]_0^2 = \frac{32}{20} - \frac{0}{20} = 1.6$

b) 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. We have from part a) that $F(x) = \frac{1}{16}x^4$, seen as a function from $[0,2]$ to $[0,1]$. To find the inverse, set $y = F(x)$ and solve for x:

$$y = \frac{1}{16}x^4, \text{ so } x^4 = 16y, \text{ so } x = 2y^{\frac{1}{4}}$$

(As $0 \leqslant y \leqslant 1$ and $0 \leqslant x \leqslant 2$, $x = 2y^{\frac{1}{4}}$ is the only solution.)

Applying $F^{-1}(x) = 2x^{\frac{1}{4}}$ to the pseudorandom numbers from Re[0, 1] gives the following five simulated pseudorandom numbers.

$$t_1 = 2 \cdot 0.5755^{1/4} \approx 1.813, \quad t_2 = 2 \cdot 0,1438^{1/4} \approx 1.232, \quad t_3 = 2 \cdot 0.8224^{1/4} \approx 1.905$$

Fråga 3

Let $X_1, X_2, \ldots, X_{400}$ be the amount of money (in SEK) that customers $1, \ldots, 400$ spend at Anna's ice cream stand. By our assumptions these random variables are independent. Furthermore,

$$E[X_i] = 0.3 \cdot 10 + 0.5 \cdot 20 + 0.2 \cdot 30 = 19, \quad E[X_i^2] = 0.3 \cdot 10^2 + 0.5 \cdot 20^2 + 0.2 \cdot 30^2 = 410,$$

$$V[X_i] = E[X_i^2] - E[X_i]^2 = 410 - 19^2 = 49.$$

The total amount of money spent at the icecream shop is $Y = X_1 + X_2 + \cdots + X_{400}$, so we want to approximate $P(Y \geqslant 7500)$. Note that $E[Y] = 400 \cdot E[X_1] = 400 \cdot 19 = 7600$ and, by independence, $V[Y] = 400 \cdot V[X_1] = 19600$. By the central limit theorem, we can approximate the distribution of Y with a normal distribution: $Y \approx N(7600, 19600)$. Therefore, letting $Z \sim N(0,1)$ be a standard normal random variable,

$$P(Y \geqslant 7500) = P\left(\frac{Y - 7600}{\sqrt{19600}} \geqslant \frac{7500 - 7600}{\sqrt{19600}}\right) \approx P\left(Z \geqslant -\frac{5}{7}\right) = 1 - P\left(Z \leqslant -\frac{5}{7}\right)$$

$$= 1 - \Phi\Big(-\frac{5}{7}\Big) = 1 - \Big(1 - \Phi\Big(\frac{5}{7}\Big)\Big) = \Phi\Big(\frac{5}{7}\Big) \approx \Phi(0.71) \approx 0.7611,$$

where we looked up the last value in the table.

Fråga 4

a) Let T_1, T_2 be the corresponding estimators, then $E[T_1] = 2E[X_1] - E[X_2] = 2\mu - \mu = \mu$, so T_1 is unbiased. Furthermore recall from the lectures that $E[\bar{X}] = E[\bar{Y}] = \mu$, and so $E[T_2] = (E[\bar{X}] + E[\bar{Y}])/2 = (\mu + \mu)/2 = \mu$, so T_2 is unbiased.

b) As all the random variables X_i and Y_j are independent, we can calculate the variances $V[T_1] = 2^2 V[X_1] + (-1)^2 V[X_2] = 5$ and $V[T_2] = \frac{1}{4}(V[\bar{X}] + V[\bar{Y}]) = \frac{1}{4}\left(\frac{1}{n} + \frac{100}{n}\right) = 25.25/n$ (recalling from the lectures that $V[\bar{X}] = \sigma^2/n$ if $V[X_i] = \sigma^2$). So T_1 is more effective than T_2 if $5 < 25.25/n$, or equivalently $n < 25.25/5 = 5.05$, so this is the case for all integers $n \leqslant 5$ (and $n \geqslant 2$).

Fråga 5

a) Using the model with the estimated parameters, and letting $x_0 = 55$, the prediction for Y_0 is

$$y_0 = \hat{m} + \hat{k}x_0 = 240 + 10.1 \cdot 55 = 795.5 \text{ (milliseconds)}$$

b) Since $\hat{k} = 10.1 > 0$, we are looking for a scatter plot where the points follow an 'upwards slope', which is either Figure 1 or Figure 3. As $R^2 = 88.2\%$ is quite high, Figure 1 looks more suitable — the points are closer together and look more like a line, while they are quite spread out on Figure 3. So Figure 1 is the correct scatter plot.

(Remark: In Figure 3, we have $R^2 = 43.1\%$.)

c) In a linear regression model for the new server, the (new) random variables Y_i' corresponding to an input of length x_i' would follow $Y_i' = m_{\text{new}} + k_{\text{new}} x_i' + \epsilon_i'$, where $\epsilon_i' \sim N(0, \sigma^2_{\text{new}})$. We can estimate the parameters m_{new} and k_{new} from their given relationship to m, k and the estimators for these: As $m_{\text{new}} = 2m$ and $k_{\text{new}} = 0.5k$,

$$\hat{m}_{\text{new}} = 2\hat{m} = 480, \qquad \hat{k}_{\text{new}} = 0.5\hat{k} = 5.05.$$

A request of length x should be sent to the new server if the (estimated) expected latency is smaller there than for the old server. This gives

$$480 + 5.05x < 240 + 10.1x$$

or equivalently $5.05x > 240$, that is, $x > 240/5.05 \approx 47.52$. So all requests of length at least 48 characters should be sent to the new server, and shorter ones to the old server.

Fråga 6

a) The numbers on the arrows coming out of each state have to sum up to 1, so $* = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$.

b) The transition matrix is

$$\begin{pmatrix} 1/4 & 0 & 3/4 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}$$

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

$$p^{(1)} = p^{(0)}\mathbf{P} = \begin{pmatrix} 0.5 & 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 1/4 & 0 & 3/4 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{pmatrix} = \begin{pmatrix} 0.25 & 0.25 & 0.5 \end{pmatrix}$$

so $P(X_1 = 2) = 0.5$ (the third element of the vector).

d) $P(X_0 = 0 \mid X_1 = 2) = \frac{P(X_0=0, X_1=2)}{P(X_1=2)} = \frac{0.5 \cdot 0.75}{0.5} = 0.75$.

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

$$\begin{pmatrix} a & b & c \end{pmatrix} \cdot \begin{pmatrix} 1/4 & 0 & 3/4 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{pmatrix} = \begin{pmatrix} a & b & c \end{pmatrix}$$

This gives $\frac{1}{4}a + b + \frac{1}{4}c = a$, $\frac{1}{2}c = b$, $\frac{3}{4}a + \frac{1}{4}c = c$. From the third equation we get $\frac{3}{4}a = \frac{3}{4}c$, so $a = c$. Plugging this and the second equation into $a + b + c = 1$ gives $c + \frac{1}{2}c + c = 1$, which solves to $c = \frac{2}{5}$, $a = c = \frac{2}{5}$, $b = \frac{1}{2}c = \frac{1}{5}$. So the stationary distribution is represented by the vector

$$\begin{pmatrix} \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix}$$

Fråga 7

a) If some emails are spam with probability $p = 0.2$ independently, we can view this as randomly deleting the Poisson events with probability $p = 0.2$, independently. By the thinning property, the result is still a Poisson process, with intensity $(1-p)\lambda = 0.8 \cdot 2 = 1.6$. The number of non-spam emails in an interval of length 8 (hours) is Poisson distributed with parameter $8 \cdot 1.6 = 12.8$. So the probability that you get 12 emails is $e^{-12.8} \cdot 12.8^{12}/12! \approx 11.1\%$.

b) The numbers of emails in each interval are distributed according to Poisson distributions with with parameters $\lambda_1 = 3\lambda = 6$, $\lambda_2 = 1 \cdot \lambda = 2$ and $\lambda_3 = 2\lambda = 4$, respectively. Furthermore, as the intervals are disjoint, the numbers are independent. Thus, by the addition theorem for the Poisson distribution, the total number of emails during the outages has the distribution $\text{Po}(\lambda_1 + \lambda_2 + \lambda_3) = \text{Po}(12)$.

c) The number of spam emails over time is given by a Poisson process with intensity $\lambda \cdot p = 2 \cdot 0.2 = 0.4$ (to see this, note that we can view this as the number of all emails with the non-spam emails deleted, which happens with probability 0.8, so by the thinning property this is a Poisson process with intensity $(1-(1-0.8))\lambda = 0.2\lambda$). Thus the waiting time for the first event has the distribution $\mathrm{Exp}(1/0.4) = \mathrm{Exp}(2.5)$, which has expectation 2.5 (hours).