Problem Sheet 4 (Solutions)

MATH1710 Probability and Statistics I

University of Leeds, 2023-24

A: Short questions

A1. Let $X \sim \text{Bin}(20, 0.4)$. Calculate

(a) $\mathbb{P}(X = 8)$

Solution.

$$\mathbb{P}(X = 8) = \binom{20}{8} 0.4^8 \times 0.6^{12} = 0.180.$$

(b) $\mathbb{P}(8 \leq X \leq 11)$

Solution.

$$\begin{aligned}\mathbb{P}(8 \leq X \leq 11) &= \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10) + \mathbb{P}(X = 11) \\ &= \binom{20}{8} 0.4^8 \times 0.6^{12} + \binom{20}{9} 0.4^9 \times 0.6^{11} + \binom{20}{10} 0.4^10 \times 0.6^{10} + \binom{20}{11} 0.4^8 \times 0.6^{11} \\ &= 0.180 + 0.160 + 0.117 + 0.071 \\ &= 0.528.\end{aligned}$$

(c) $\mathbb{E}X$

Solution. $\mathbb{E}X = 20 \times 0.4 = 8.$

A2. Let $X \sim \text{Geom}(0.2)$. Calculate

(a) $\mathbb{P}(X = 2)$

Solution. $\mathbb{P}(X = 2) = 0.8^1 \times 0.2^1 = 0.16.$

(b) $\mathbb{P}(X \geq 3)$

Solution. $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - 0.2 - 0.8 \times 0.2 = 0.64.$

(c) $\text{Var}(X)$

Solution. $\text{Var}(X) = \dfrac{1 - 0.2}{0.2^2} = 20.$

A3. Let $X \sim \text{Po}(2.5)$. Calculate

(a) $\mathbb{P}(X = 3)$

Solution. $\mathbb{P}(X = 3) = \mathrm{e}^{-2.5}\dfrac{2.5^3}{3!} = 0.214.$

(b) $\mathbb{P}(X \geq \mathbb{E}X)$

Solution. First, $\mathbb{E}X = 2.5$. So

$$\begin{aligned}\mathbb{P}(X \geq \mathbb{E}X) &= \mathbb{P}(X \geq 2.5)\\ &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2)\\ &= 1 - \mathrm{e}^{-2.5} - 2.5\mathrm{e}^{-2.5} - \frac{2.5^2}{2}\mathrm{e}^{-2.5}\\ &= 1 - 0.082 - 0.204 - 0.257\\ &= 0.456.\end{aligned}$$

A4. Consider the following joint PMF:

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$
$x=0$	$2k$	$2k$	k	0
$x=1$	k	$3k$	k	k
$x=2$	0	k	k	$2k$

(a) Find the value of k that makes this a joint PMF.

Solution. The total of the joint PMF is

$$2k + 2k + k + k + 3k + k + k + k + k + 2k = 15k$$

which must be 1, so $k = \frac{1}{15}$.

(b) Find the marginal PMFs of X and Y.

Solution. By summing across the rows and down the columns, respectively, we get this:

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=0$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	0	$\frac{5}{15}$
$x=1$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{6}{15}$
$x=2$	0	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{4}{15}$
$p_Y(y)$	$\frac{3}{15}$	$\frac{6}{15}$	$\frac{3}{15}$	$\frac{3}{15}$	

(c) What is the conditional distribution of Y given $X = 1$?

Solution. We get this by taking the $x = 1$ row of the table, than normalising it by dividing through by $p_X(1) = \frac{6}{15}$. This gives

$$p_{Y|X}(0 \mid 1) = \frac{1}{6} \qquad p_{Y|X}(1 \mid 1) = \frac{3}{6} \qquad p_{Y|X}(2 \mid 1) = \frac{1}{6} \qquad p_{Y|X}(3 \mid 1) = \frac{1}{6}.$$

(d) Are X and Y independent?

Solution. No. For one example, $p_{X,Y}(0,0) = \frac{2}{15}$, while $p_X(0)\,p_Y(0) = \frac{5}{15} \times \frac{3}{15} = \frac{1}{15}$, so they are not equal.

B: Long questions

B1. Calculate the CDF $F(x) = \mathbb{P}(X \leq x)$ of the geometric distribution...

(a) ... by summing the PMF;

Solution. We have,using the standard formula for the sum of a finite geometric progression,

$$\begin{aligned} F(x) &= \sum_{y=1}^{x} p(y) \\ &= \sum_{y=1}^{x} (1-p)^{y-1} p \\ &= \frac{p\big(1-(1-p)^x\big)}{1-(1-p)} \\ &= \frac{p\big(1-(1-p)^x\big)}{p} \\ &= 1-(1-p)^x. \end{aligned}$$

(b) ... by explaining how the "number of trials until success" definition tells us what $1 - F(x) = \mathbb{P}(X > x)$ must be.

Solution. Note that $1 - F(x) = \mathbb{P}(X > x)$ is precisely the probability that the first x trials are failures, and hence that the first success comes strictly after the xth trial. The probability that the first x trials are failures is $(1-p)^x$. So $F(x) = 1 - (1-p)^x$.

(c) A gambler rolls a pair of dice until he gets a double-six. What is the probability that this takes between 20 and 40 double-rolls?

Solution. Let $X \sim \text{Geom}(\frac{1}{36})$. Then

$$\begin{aligned}\mathbb{P}(20 \le X \le 40) &= \mathbb{P}(X \le 40) - \mathbb{P}(X \le 19) \\ &= F(40) - F(19) \\ &= \left(1 - \left(1 - \tfrac{1}{36}\right)^{40}\right) - \left(1 - \left(1 - \tfrac{1}{36}\right)^{19}\right) \\ &= 0.676 - 0.414 \\ &= 0.261.\end{aligned}$$

B2. Let $X \sim \text{Geom}(p)$. Recall that X represents the number of trials up to and including the first success. Recall also that $\mathbb{E}X = 1/p$ and $\text{Var}(X) = (1-p)/p^2$.

Let Y be a geometric distribution with parameter p according to the alternative "number of failures *before* the first success" definition.

(a) Write down the PMF for Y.

Solution. Having $Y = y$ requires y consecutive failures immediately followed by a success. So $p_Y(y) = (1-p)^y p$.

(b) Explain why the expectation of Y of

$$\mathbb{E}Y = \frac{1}{p} - 1 = \frac{1-p}{p}.$$

Solution. If $X \sim \text{Geom}(p)$ under the standard definition, then (as we saw in the notes) Y has the same distribution as $X - 1$. Therefore,

$$\mathbb{E}Y = \mathbb{E}(X-1) = \mathbb{E}X - 1 = \frac{1}{p} - 1 = \frac{1-p}{p}.$$

(c) What is the variance of Y?

Solution.

$$\text{Var}(Y) = \text{Var}(X-1) = \text{Var}(X) = \frac{1-p}{p^2}.$$

B3 Let $X \sim \text{Po}(\lambda)$.

(a) Show that $\mathbb{E}X(X-1) = \lambda^2$. You may use the Taylor series for the exponential,

$$\mathrm{e}^{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}.$$

Solution. We follow exactly the method used to calculate $\mathbb{E}X$ in the notes. We have

$$\begin{aligned}\mathbb{E}X(X-1) &= \sum_{x=0}^{\infty} x(x-1)\,\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \\ &= \lambda^2\mathrm{e}^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} \\ &= \lambda^2\mathrm{e}^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^{y}}{y!} \\ &= \lambda^2\mathrm{e}^{-\lambda}\,\mathrm{e}^{\lambda} \\ &= \lambda^2.\end{aligned}$$

In the second line, we took a λ^2 and a $\mathrm{e}^{-\lambda}$ outside the brackets; cancelled the x and $x-1$ out of the $x!$; and removed the $x=0$ and $x=1$ terms from the sum, since they were 0 anyway. In the third line, we re-indexed the sum by setting $y=x-2$. In the fourth line, we used the Taylor series for the exponential

(b) Hence show that $\mathrm{Var}(X)=\lambda$. You may use the fact that $\mathbb{E}X=\lambda$.

Solution. We know from part (a) that

$$\mathbb{E}X(X-1) = \mathbb{E}(X^2-X) = \mathbb{E}X^2 - \mathbb{E}X = \mathbb{E}X^2 - \lambda = \lambda^2,$$

which gives $\mathbb{E}X^2 = \lambda^2+\lambda$. We can then use the computational formula for the variance to get

$$\mathrm{Var}(X) = \mathbb{E}X^2 - \lambda^2 = \lambda^2+\lambda-\lambda^2 = \lambda.$$

B4. Each week in the UK about 15 million Lotto tickets are sold. As we saw in Lecture 6, the probability of each ticket winning is about 1 in 45 million. Estimate the proportion of weeks when there is **(a)** a roll-over (no jackpot winners), **(b)** a unique jackpot winner, or **(c)** when multiple winners share the jackpot. State any modelling assumptions you make and the approximation that you use.

Solution. We assume that each ticket is uniformly randomly chosen from all possible tickets, independent of all other tickets. Then the number of winners is $X \sim \mathrm{Bin}(15\text{ million}, 1/(45\text{ million}))$. It will be convenient to use a Poisson approximation with rate

$$\lambda = 15\text{ million} \times \frac{1}{45\text{ million}} = \tfrac{1}{3}.$$

The probability there is a roll-over is

$$\mathbb{P}(X=0) \approx \mathrm{e}^{-1/3} = 0.72.$$

The probability there is a unique jackpot winner is

$$\mathbb{P}(X=1) \approx \tfrac{1}{3}\mathrm{e}^{-1/3} = 0.24.$$

The probability there are multiple winners is

$$\mathbb{P}(X\geq 2) = 1-\mathbb{P}(X=0)-\mathbb{P}(X=1) = 0.04.$$

B5. Let X and Y be Bernoulli($\frac{1}{2}$) random variables.

(a) Write down the table for the joint PMF of X and Y if X and Y are independent.

Solution. For all these questions, we need to fill in a table for the joint PMF, where the columns sum to $p_X(0) = p_X(1) = \frac{1}{2}$ and the rows sum to $p_Y(0) = p_Y(1) = \frac{1}{2}$.

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$			$\frac{1}{2}$
$y=1$			$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

If X and Y are independent, we have $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$; so, for example, $p_{X,Y}(0,0) = p_X(0)\,p_Y(0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. In fact, all the entries in the joint PMF table are $\frac{1}{4}$.

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$y=1$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

(b) Write down a table for a joint PMF of X and Y that is consistent with their marginal distributions but that leads to X and Y having a positive correlation.

Solution. We still need the rows and columns to add up to $\frac{1}{2}$, but we want low values of X (that is, 0) to be more likely to occur alongside low values of Y (that is, 0), and high values of X (that is, 1) alongside high values of Y (that is, 1). One was to do this is

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	$\frac{1}{2}$	0	$\frac{1}{2}$
$y=1$	0	$\frac{1}{2}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

A single table like that is a perfectly sufficient answer. But, in fact, any table of the form

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	a	$\frac{1}{2}-a$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}-a$	a	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

for $\frac{1}{4} < a \leq \frac{1}{2}$ will do. This has

$$\mathbb{E}XY = \sum_{x,y} xy\, p_{X,Y}(x,y) = p_{X,Y}(1,1) = a,$$

as $x = y = 1$ is the only nonzero term in the sum. This means the covariance is, by the computational formula,

$$\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mu_X\mu_Y = a - \tfrac{1}{2} \times \tfrac{1}{2} = a - \tfrac{1}{4}.$$

So the covariance is positive for $a > \frac{1}{4}$, so the correlation is too.

(c) Write down a table for a joint PMF of X and Y that is consistent with their marginal distributions but that leads to X and Y having a negative correlation.

Solution. For example

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	0	$\frac{1}{2}$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}$	0	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

Alternatively, any table of the form

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	a	$\frac{1}{2}-a$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}-a$	a	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

for $0 \leq a < \frac{1}{4}$ will have negative covariance $a - \frac{1}{4}$, so negative correlation.

C: Assessed questions

C1. A collector wants to collect football stickers to fill an album. There are n unique stickers to collect. Each time the collector buys a sticker, it is one of the n stickers chosen independently uniformly at random. Unfortunately, it is likely the collector will end up having "swaps", where he has received the same sticker more than once, so he will likely need to buy more than n stickers in total to fill his album. But how many?

(a) Suppose the collector has already got j unique stickers (and some number of swaps), for $j = 0, 1, 2, \ldots, n-1$. Let X_j be the the number of extra stickers he buys until getting a new unique sticker. Explain why X_j is geometrically distributed, and state the parameter $p = p_j$ of the geometric distribution.

Hint. To get a new sticker you need to "succeed" in an experiment, where "failure" is "receiving a sticker you already have a copy of" and "success" is "receiving a new sticker you haven't already got". What is the probability of success?

(b) Hence, show that the expected number of stickers the collector must buy to fill his album is

$$n \sum_{k=1}^{n} \frac{1}{k}.$$

Hint. Recall the expectation of the geometric distribution, and use linearity of expectation.

(c) The World Cup 2020 sticker album required $n = 670$ unique stickers to complete it, and stickers cost 18p each. Using the expression from (b), calculate the expected amount of money needed to fill the album. You should do this calculation in R and include the command you used in your answer.

Hint. In R, `sum(1 / (1:678))` sum the reciprocals of the integers from 1 to 678.

(d) By approximating the sum in part (b) by an integral, explain why the expected number of stickers required is approximately $n \log n$, where log denotes the natural logarithm to base e.

Hint. A detailed proof is not required – just informally explain why the sum is approximately the area under the function (ie the integral).

C2. Let X and Y be random variables, and let a and b be constants.

(a) Starting from the definition of covariance, show that $\mathrm{Cov}(aX, Y) = a\ \mathrm{Cov}(X, Y)$. You may find it helpful to remember that if $\mathbb{E}X = \mu_X$, then $\mathbb{E}aX = a\mu_X$.

(b) Show that $\mathrm{Cov}(X + b, Y) = \mathrm{Cov}(X, Y)$.

Hint. You may find it helpful to look back at Theorem 7.4 in the notes.

Now let X, Y, Z be *independent* random variables with common variance σ^2.

(c) Find the value of $\mathrm{Corr}(2X - 3Y + 4, 2Y - Z - 1)$. You may use any facts about covariance from the notes, including those from parts (a) and (b) of this question, provided you state them clearly.

Hint. Start by calculating the covariance. Try to sort out the $2X - 3Y + 4$ part first, using the rules above. Once that's done you can then deal with the $2Y - Z - 1$ part with the same rules, because $\mathrm{Cov}(U, V) = \mathrm{Cov}(V, U)$, so anything you can do to the first term you can also do to the second.