Problem Sheet 3 (Solutions)

MATH1710 Probability and Statistics I

University of Leeds, 2022-23

A: Short questions

A1. Consider the continuous random variable X with PDF

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 \le x \le 1 \\ \frac{1}{2} & \text{for } 1 < x \le 2 \\ \frac{3}{2} - \frac{1}{2}x & \text{for } 2 < x \le 3 \end{cases}$$

and $f(x) = 0$ otherwise.

(a) Calculate the CDF for X.

Solution. We treat the different cases separately.

For $x < 0$, we have $F(x) = 0$.

For $0 \le x \le 1$, we have

$$F(x) = \int_0^x \tfrac{1}{2}y \,\mathrm{d}y = \left[\tfrac{1}{4}y^2\right]_0^x = \tfrac{1}{4}x^2.$$

In particular, $F(1) = \frac{1}{4}$.

For $1 < x \le 2$, we have

$$F(x) = \int_0^x f(y)\,\mathrm{d}y = F(1) + \int_1^y \tfrac{1}{2}\,\mathrm{d}y = \tfrac{1}{4} + \left[\tfrac{1}{2}y\right]_1^x = \tfrac{1}{2}x - \tfrac{1}{4}.$$

In particular, $F(2) = \frac{3}{4}$.

For $2 < x \le 3$, we have

$$F(x) = \int_0^x f(y)\,\mathrm{d}y = F(2) + \int_2^y \left(\tfrac{3}{2} - \tfrac{1}{2}y\right)\,\mathrm{d}y = \tfrac{3}{4} + \left[\tfrac{3}{2}y - \tfrac{1}{4}y^2\right]_2^x = \tfrac{3}{2}x - \tfrac{1}{4}x^2 - \tfrac{5}{4}.$$

In particular, $F(3) = 1$.

For $x > 3$, we have $F(x) = 1$.

Hence,

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{4}x^2 & \text{for } 0 \le x \le 1 \\ \frac{1}{2}x - \frac{1}{4} & \text{for } 1 < x \le 2 \\ \frac{3}{2}x - \frac{1}{4}x^2 - \frac{5}{4} & \text{for } 2 < x \le 3 \\ 1 & \text{for } x > 3. \end{cases}$$

(b) What is $\mathbb{P}(\frac{3}{2} \le X \le \frac{5}{2})$?

Solution. This is

$$F\left(\tfrac{5}{2}\right) - F\left(\tfrac{3}{2}\right) = \tfrac{15}{16} - \tfrac{1}{2} = \tfrac{7}{16}.$$

(Here, it was useful to note that $x = \frac{5}{2}$ is in the $2 < x \leq 3$ range and $x = \frac{3}{2}$ is in the $1 < x \leq 2$ range.)

(c) Calculate the expectation $\mathbb{E}X$.

Solution. We have

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x \\
&= \int_0^1 x \times \tfrac{1}{2}x\, \mathrm{d}x + \int_1^2 x \times \tfrac{1}{2}\, \mathrm{d}x + \int_2^3 x \times \left(\tfrac{3}{2} - \tfrac{1}{2}x\right)\, \mathrm{d}x \\
&= \left[\tfrac{1}{6}x^3\right]_0^1 + \left[\tfrac{1}{4}x^2\right]_1^2 + \left[\tfrac{3}{4}x^2 - \tfrac{1}{6}x^3\right]_2^3 \\
&= \tfrac{1}{6} - 0 + 1 - \tfrac{1}{4} + \tfrac{9}{4} - \tfrac{5}{3} \\
&= \tfrac{3}{2}.
\end{aligned}$$

A2. Let X be a continuous random variable with PDF

$$f(x) = \frac{k}{x^3} \qquad \text{for } x \geq 1$$

and $f(x) = 0$ otherwise.

(a) What value of k makes this into a true PDF?

Solution. We need the PDF to integrate to 1. So

$$1 = \int_{-\infty}^{\infty} f(x)\, \mathrm{d}x = \int_1^{\infty} kx^{-3}\, \mathrm{d}x = \left[-\tfrac{1}{2}kx^{-2}\right]_1^{\infty} = -0 + \tfrac{1}{2}k.$$

So $k = 2$.

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

Solution. This is

$$\mathbb{P}(X \geq 3) = \int_3^{\infty} 2x^{-3}\, \mathrm{d}x = \left[-x^{-2}\right]_3^{\infty} = \tfrac{1}{9}.$$

(c) What is the expected value $\mathbb{E}X$?

Solution. This is

$$\mathbb{E}X = \int_1^{\infty} x \times 2x^{-3} \mathrm{d}x = \left[2x^{-1}\right]_1^{\infty} = 2.$$

A3. Let $X \sim \mathrm{Exp}(\frac{1}{2})$.

(a) What is $\mathbb{E}X$?

Solution. $\mathbb{E}X = \dfrac{1}{\frac{1}{2}} = 2$

(b) What is $\mathbb{P}(1 \leq X \leq 3)$?

Solution. We have

$$\mathbb{P}(1 \leq X \leq 3) = F(3) - F(1) = (1 - \mathrm{e}^{-3/2}) - (1 - \mathrm{e}^{-1/2}) = 0.383.$$

A4. Let $Z \sim \mathrm{N}(0, 1)$. Calculate the following **(a)** using statistical tables; **(b)** using R. (For part (a), you should show enough working to convince a reader that you really did use the tables.)

(i) $\mathbb{P}(Z \leq -1.2)$

Solution. Using statistical tables,

$$\Phi(-1.2) = 1 - \Phi(1.20) = 1 - 0.8849 = 0.1151.$$

Using R: `pnorm(-1.2)` gives `0.1150697`.

(ii) $\mathbb{P}(-1.2 \leq Z \leq 0.8)$

Solution. Using statistical tables, and part (i),

$$\Phi(0.80) - \Phi(-1.2) = 0.7781 - 0.1151 = 0.6730.$$

Using R: `pnorm(0.8) - pnorm(-1.2)` gives `0.6730749`.

(iii) $\mathbb{P}(Z \leq 0.27)$ (using interpolation for part (a))

Solution. We can interpolate between $\Phi(0.25) = 0.5987$ and $\Phi(0.30) = 0.6179$, to get

$$\Phi(0.27) \approx 0.6\,\Phi(0.25) + 0.4\,\Phi(0.30) = 0.6064.$$

Using R: `pnorm(0.27)` gives `0.6064199`.

A5. Let $X \sim \mathrm{Po}(25)$. Calculate the following **(a)** exactly, using R; **(b)** approximately, using a normal approximation with a continuity correction and statistical tables. (For part (b), you should show enough working to convince a reader that you really did use the tables.)

(i) $\mathbb{P}(X \leq 27)$

Solution. Using R: `ppois(27, 25)` gives `0.7001861`.

The approximation is $X \approx Y \sim \mathrm{N}(25, 25) = \mathrm{N}(25, 5^2)$. With a continuity correction, we

expand the interval $(-\infty, 27]$ outwards to $(-\infty, 27.5]$, and get

$$\mathbb{P}(X \leq 27) \approx \mathbb{P}(Y \leq 27.5) \approx \mathbb{P}\left(\frac{Y-25}{5} \leq \frac{27.5-25}{5}\right) = \Phi(0.50) = 0.692.$$

(ii) $\mathbb{P}(X \geq 28 \mid X \geq 27)$

Solution. By the definition of conditional probability, we have

$$\mathbb{P}(X \geq 28 \mid X \geq 27) = \frac{\mathbb{P}(X \geq 28 \text{ and } X \geq 27)}{\mathbb{P}(X \geq 27)} = \frac{\mathbb{P}(X \geq 28)}{\mathbb{P}(X \geq 27)},$$

since if $X \geq 28$ it's automatically the case that $X \geq 27$.

Using R, we need to remember that `lower.tail = FALSE` gives $\mathbb{P}(X > x)$ with strict inequality, which for discrete random variables is equivalent to $\mathbb{P}(X \geq x + 1)$. So we actually want

```
ppois(27, 25, lower.tail = FALSE) / ppois(26, 25, lower.tail = FALSE)
```

which gives `0.8089648`.

The approximations are

$$\mathbb{P}(Z \geq 28) \approx \mathbb{P}(Y \geq 27.5) = 1 - \Phi(0.50) = 0.3085$$
$$\mathbb{P}(Z \geq 27) \approx \mathbb{P}(Y \geq 26.5) = 1 - \Phi(0.30) = 0.3821,$$

where again we used the continuity correct to expand $[28, \infty)$ outwards to $[27.5, \infty)$ and the same for $[27, \infty)$. This gives the answer $0.3085/0.3821 = 0.807$.

B: Long questions

B1. (a) Let $X \sim \text{Exp}(\lambda)$. Show that

$$\mathbb{P}(X > x + y \mid X > y) = \mathbb{P}(X > x).$$

Solution.
Using the definition of conditional probability, we have

$$\mathbb{P}(X > x + y \mid X > y) = \frac{\mathbb{P}(X > x + y \text{ and } X > y)}{\mathbb{P}(X > y)} = \frac{\mathbb{P}(X > x + y)}{\mathbb{P}(X > y)},$$

since is $X > x + y$ then we automatically have $X > y$. Note also that, for an exponential distribution we have

$$\mathbb{P}(X > x) = 1 - F(x) = 1 - (1 - \mathrm{e}^{-\lambda x}) = \mathrm{e}^{-\lambda x}.$$

So the left-hand side of the statement in the question is

$$\frac{\mathrm{e}^{-\lambda(x+y)}}{\mathrm{e}^{-\lambda y}} = \mathrm{e}^{-\lambda x - \lambda y + \lambda y} = \mathrm{e}^{-\lambda x},$$

which equals the right-hand side, by the above.

(b) The result proved in part (a) is called the "memoryless property". Why do you think it's called that?

Solution. Think of X as a waiting time. The result tells us that, given that we've already waited y minutes, the probability that we have to wait at least another x minutes is exactly the same as the probability we had to wait at least x minutes starting from the beginning. In other words, *no matter when we start timing from*, the probability we have to wait more than x minutes remains the same.

This is called the "memoryless property" because it's as if the process has no memory of how long we've already been waiting for.

(This property also holds for the geometric distribution. The expected number of rolls of a dice until you get a six is always 6 rolls, no matter how many times you've already rolled the dice.)

(c) When you get to certain bus stop, the average amount of time you have to wait for a bus to arrive is 20 minutes. Specifically, the time until the next bus arrives is modelled as an exponential distribution with expectation $1/\lambda = 20$ minutes. Suppose you have already been waiting at the bus stop for 15 minutes. What is the expected further amount of time you still have to wait for a bus to arrive?

Solution. By the memoryless property, it's irrelevant how long we've been waiting for: the average time until a bus arrives is always $1/\lambda = 20$ minutes.

B2. The main dangerous radioactive material left over after the Chernobyl disaster is Caesium-137. The amount of time it takes a Caesium-137 particle to decay is known to follow an exponential distribution with rate $\lambda = 0.023$ years^{-1}.

(a) What is the average amount of time it takes a Caesium-137 particle to decay?

Solution. The expectation is $1/\lambda = 43.5$ years.

(b) The "half-life" of a radioactive substance is the amount of time it takes for half of the substance to decay. Using the information in the question, calculate the half-life of Caesium-137.

Solution. The half-life is the median of the distribution; that is, the solution x to

$$F(x) = 1 - \mathrm{e}^{-0.023x} = \tfrac{1}{2}.$$

So

$$x = \frac{\log \frac{1}{2}}{-0.023} = \frac{\log 2}{0.023} = 30.1 \text{ years.}$$

(c) It is estimated that roughly 24 kg of Caesium-137 was released during the Chernobyl disaster, which happened roughly 35.6 years ago. Estimate the mass of Caesium-137 that has still not decayed?

Solution. The proportion of Caesium-137 still remaining is

$$\mathbb{P}(X > 35.6) = \mathrm{e}^{-0.023 \times 35.6} = 0.441,$$

so roughly $24 \times 0.441 = 10.6$ kg of Caesium-137 has still not decayed.

(The Chernobyl disaster was actually 36.6 *years ago – I forgot to update the question this year. The amount of Caesium-137 will have gone down about another 250g in the last year.)*

B3. Consider the pair of random variables (X, Y) with joint PDF

$$f_{X,Y}(x, y) = 2 \qquad \text{for } 0 \le x \le y \le 1$$

and $f_{X,Y}(x, y) = 0$ otherwise. (In particular, note that the joint PDF is only nonzero when $x \le y$.)

(a) Draw a picture of the range of (X, Y) in the xy-plane.

(b) Describe the conditional distribution of X given $Y = y$, for $0 \le y \le 1$.

Solution. Fix y. The conditional distribution is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \propto f_{X,Y}(x, y).$$

We know that $f_{X,Y}(x, y) = 2$ when $0 \le x \le y$ and is 0 otherwise. So the conditional distribution of X given $Y = y$ is continuous uniform on the interval $[0, y]$.

If we want to check the denominator $f_Y(y)$ formally, we can check that

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x = \int_0^y 2 \,\mathrm{d}y = 2y,$$

so the conditional PDF is indeed $f_{X|Y}(x \mid y) = 2/2y = 1/y$ for $0 \le x \le y$ and 0 otherwise.

(c) What is the marginal PDF f_X of X?

Solution. Again the key is that the joint PDF is only nonzero when $y \ge x$ but $y \le 1$. So

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \,\mathrm{d}y = \int_x^1 2 \,\mathrm{d}y = 2(1 - x)$$

for $0 \le x \le 1$ and 0 otherwise.

(d) Are X and Y independent?

Solution. No. Take, for example, $x = \frac{3}{4}$ and $y = \frac{1}{4}$. It's clear that this $f_{X,Y}(\frac{3}{4}, \frac{1}{4}) = 0$, while $f_X(\frac{3}{4})$ and $f_Y(\frac{1}{4})$ are nonzero, just by looking at the picture from part (a).

We can check it formally too, if we want. Since $x > y$, this point has joint PDF $f_{X,Y}(\frac{3}{4}, \frac{1}{4}) = 0$. We know the marginal PMFs, though are

$$f_X\left(\tfrac{3}{4}\right) = 2\left(1 - \tfrac{3}{4}\right) = \tfrac{1}{2}$$
$$f_Y\left(\tfrac{1}{4}\right) = 2 \times \tfrac{1}{4} = \tfrac{1}{2}.$$

(we used $f_Y(y) = 2y$) based on symmetry with $f_X(1 - x)$, or alternatively by calculating it "long-hand".) So $f_X(x)f_Y(y) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq 0$. So X and Y are not independent.

B4. *(Optional)* Engineers and scientists often use the rule of thumb "Only 5% of data is more than two sample standard deviations away from the sample mean." Carefully justify this rule, using concepts from the module.

Solution. By the central limit theorem, and other related approximation arguments, it is reasonable that lots of real life data – especially that which is affected by the accumulation of numerous small effects – is approximately normally distributed.

Write μ for the *true* expectation and σ^2 for the *true* variance of the population distribution X. Then the proportion of data that is within two true-standard-deviations of the true-expectation will, be the law of large numbers, tends to

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$$

Using standardisation, this is

$$\mathbb{P}\left(\frac{(\mu - 2\sigma) - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{(\mu + 2\sigma) - \mu}{\sigma}\right) = \mathbb{P}(-2 \leq Z \leq 2).$$

Using R or statistical tables, this is

$$\Phi(2) - \Phi(-2) = 2\,\Phi(2) - 1 = 0.9545.$$

So only $1 - 0.9545 = 0.0455$, or approximately 5%, of data is more than two true-standard-deviations away from the true-expectation.

Finally, the law of large numbers also tells us that, provided a large number of datapoints n are collected, the sample mean $\bar{x}$ and the sample standard deviation s_x will be very close to the true expectation μ and the true standard deviation σ respectively, so we can replace the latter with the former in our calculations.

B5. Let $X_1, X_2, \ldots, X_n$ be IID random variable with common expectation μ and common variance σ^2, and let $\overline{X} = (X_1 + \cdots + X_n)/n$ be the mean of these random variables. We will be considering the random variable S^2 given by

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

(a) By writing

$$X_i - \overline{X} = (X_i - \mu) - (\overline{X} - \mu)$$

or otherwise, show that

$$S^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Solution. Using the suggestion in the question, we have

$$\begin{aligned}
S^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X} - \mu)\right)^2 \\
&= \sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\right) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}2(X_i - \mu)(\overline{X} - \mu) + \sum_{i=1}^{n}(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2\left(\sum_{i=1}^{n}X_i - \sum_{i=1}^{n}\mu\right)(\overline{X} - \mu) + (\overline{X} - \mu)^2\sum_{i=1}^{n}1 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(n\overline{X} - n\mu)(\overline{X} - \mu) + n(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\overline{X} - \mu)^2 + n(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

This is mostly manipulation of sums as we have seen before, although note that going from the fifth to sixth lines we used the definition of $\overline{X}$ to write $\sum_{i=1}^{n} X_i$ as $n\overline{X}$.

(b) Hence or otherwise, show that

$$\mathbb{E}S^2 = (n-1)\sigma^2.$$

You may use facts about $\overline{X}$ from the notes provided you state them clearly. (You may find it helpful to recognise some expectations as definitional formulas for variances, where appropriate.)

Solution. Starting with the linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}S^2 &= \mathbb{E}\left(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right) \\
&= \sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2 - n\mathbb{E}(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}\operatorname{Var}(X_i) - n\operatorname{Var}(\overline{X}).
\end{aligned}$$

The last line follows because $\mathbb{E}X_i = \mu$ for all i by assumption, and we showed in the notes that $\mathbb{E}\overline{X} = \mu$ also; hence, as hinted, the expectations are precisely definitional formulas for the variances. We then also know that $\operatorname{Var}(X_i) = \sigma^2$ by assumption, and we showed Lecture 18 that $\operatorname{Var}(\overline{X}) = \sigma^2/n$. Hence

$$\mathbb{E}S^2 = \sum_{i=1}^{n}\sigma^2 - n\,\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2,$$

as required.

(c) At the beginning of this module, we defined the sample variance of the values $x_1, x_2, \ldots, x_n$ to be

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Explain one reason why we might consider it appropriate to use $1/(n-1)$ as the factor at the beginning of this expression, rather than simply $1/n$.

Solution. We often model a data set $x_1, x_2, \ldots, x_n$ as being realisations of an IID sequence of random variables $X_1, X_2, \ldots, X_n$. In this case, we are using the summary statistic of the sample variance s_x^2 to "estimate" the variance $\mathrm{Var}(X_1) = \sigma^2$. Using the factor $1/(n-1)$ ensures that this estimator is correct "in expectation", because

$$\mathbb{E}s_X^2 = \mathbb{E}\frac{1}{n-1}S^2 = \frac{1}{n-1}\mathbb{E}S^2 = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

This property of being correct in expectation is called being an "unbiased" estimator, and its usually considered beneficial for an estimator to be unbiased.

Note that we already know that the sample mean $\bar{x}$ is an unbiased estimator for the expectation $\mathbb{E}X = \mu$, as we already know that $\mathbb{E}\overline{X} = \mu$.

(You may learn more about estimation and "unbiasedness" in MATH1712 Probability and Statistics II.)

B6. *(New)* Roughly how many times should I toss a coin for there to be a 95% chance that between 49% and 51% of my coin tosses land Heads?

Solution. The number of Heads in n coin tosses is $X \sim \mathrm{Bin}(n, \frac{1}{2})$, which is approximately $Y \sim \mathrm{N}(\frac{n}{2}, \frac{n}{4})$. We want to choose n such that

$$\mathbb{P}(0.49n \leq Y \leq 0.51n) = 0.95.$$

Standardising, this is

$$\mathbb{P}\left(\frac{0.49n - 0.5n}{0.5\sqrt{n}} \leq \frac{Y - 0.5n}{0.5\sqrt{n}} \leq \frac{0.51n - 0.5n}{0.5\sqrt{n}}\right) = \mathbb{P}(-0.02\sqrt{n} \leq Z \leq 0.02\sqrt{n})$$

Since the normal distribution is symmetric, we want

$$\mathbb{P}(X \leq 0.02\sqrt{n}) = 0.975.$$

From Table 2 of the statistical tables, or by the R command `qnorm(0.975)`, this requires $0.01\sqrt{n} = 1.960$, which is $n \approx 9600$.

So if we toss 10,000 coins, there's about a 95% chance we get between 4900 and 5100 Heads.

(I meant to delete Question B4 from this problem sheet, to make it shorter, but I accidentally deleted Question B6 instead. I've now added B6 and marked B4 as "optional".)

C: Assessed questions

C1. Let X be a continuous random variable with PDF

$$f(x) = \tfrac{2}{9}(2 - x) \qquad \text{for } -1 \leq x \leq c$$

and $f(x) = 0$ otherwise.

(a) Explaining your work, find the value of the constant c.

> *Hint.* Remember that the integral under a PDF must equal 1.

(b) What is $\mathbb{P}(X > 1)$?

> *Hint.* This is standard.

(c) Calculate the expectation of$X$.

> *Hint.* This is standard.

(d) Calculate the variance of X.

> *Hint.* This is standard. I recommend using the computational formula, so start by finding $\mathbb{E}X^2$.

C2. For each of the following, **(a)** calculate the *exact* value using R; **(b)** get an approximate value using an appropriate approximation and *without* using R. (Statistical tables are available.)

(i) $\mathbb{P}(X \leq 3)$, where $X \sim \text{Bin}(1000, 0.005)$.

> *Hint.* For the approximation, note that n is large and p is small.

(ii) $\mathbb{P}(296 \leq Y \leq 307)$, where $Y \sim \text{Bin}(1200, 0.25)$.

> *Hint.* For the approximation, note that n is large and p is *not* small.

(iii) $\mathbb{P}(Z \geq 398)$, where $Z \sim \text{Bin}(400, 0.995)$.

> *Hint.* This one will require you to think for yourself! I have not told you how to do this question. You might notice it looks a bit like the Poisson 'small p' case, but mirror-imaged. How can you use this to your advantage.