Model Answers

OCR
Oxford Cambridge and RSA

Thursday 15 October 2020 – Afternoon

AS Level Further Mathematics B (MEI)

Y412/01 Statistics a

Time allowed: 1 hour 15 minutes

You must have:
- the Printed Answer Booklet
- the Formulae Booklet for Further Mathematics B (MEI)
- a scientific or graphical calculator

INSTRUCTIONS
- Use black ink. You can use an HB pencil, but only for graphs and diagrams.
- Write your answer to each question in the space provided in the **Printed Answer Booklet**. If you need extra space use the lined pages at the end of the Printed Answer Booklet. The question numbers must be clearly shown.
- Fill in the boxes on the front of the Printed Answer Booklet.
- Answer **all** the questions.
- Where appropriate, your answer should be supported with working. Marks might be given for using a correct method, even if your answer is wrong.
- Give your final answers to a degree of accuracy that is appropriate to the context.
- Do **not** send this Question Paper for marking. Keep it in the centre or recycle it.

INFORMATION
- The total mark for this paper is **60**.
- The marks for each question are shown in brackets **[ ]**.
- This document has **8** pages.

ADVICE
- Read each question carefully before you start your answer.

DC (NF/CGW) 190277/2

Answer **all** the questions.

1 The random variable X represents the number of cars arriving at a car wash per 10-minute period. From observations over a number of days, an estimate was made of the probability distribution of X. Table 1 shows this estimated probability distribution.

r	0	1	2	3	4	>4
$P(X = r)$	0.30	0.38	0.19	0.08	0.05	0

Table 1

(a) In this question you must show detailed reasoning.

Use Table 1 to calculate estimates of each of the following.

- $E(X)$
- $Var(X)$ **[5]**

(b) Explain how your answers to part **(a)** indicate that a Poisson distribution may be a suitable model for X. **[1]**

You should now assume that X can be modelled by a Poisson distribution with mean equal to the value which you calculated in part **(a)**.

(c) Find each of the following.

- $P(X = 2)$
- $P(X > 3)$ **[3]**

(d) Given that the probability that there is at least 1 car arriving in a period of k minutes is at least 0.99, find the least possible value of k. **[3]**

1a)

r	0	1	2	3	4	>4
$P(X = r)$	0.30	0.38	0.19	0.08	0.05	0

$E(X) = (0 \times 0.30) + (1 \times 0.38) + (2 \times 0.19) + (3 \times 0.08) + (4 \times 0.05)$

$= 0 + 0.38 + 0.38 + 0.24 + 0.2$

$= 1.2$

$E(X^2) \rightarrow (0^2 \times 0.30) + (1^2 \times 0.38) + (2^2 \times 0.19) + (3^2 \times 0.08) + (4^2 \times 0.05)$

$= 0 + 0.38 + 0.76 + 0.72 + 0.8$

$= 2.66$

$\therefore \text{var}(X) = 2.66 - 1.2^2$

$= 1.22$

b) Because the variance ≈ the mean, which suggests that the poisson distribution may be suitable

c) $X \sim Po(1.2)$

$P(X = 2) = 0.216859832 6$

≈ 0.2169 (4sf)

$P(X > 3) = 1 - P(X \leq 3)$

$= 1 - 0.966231032 4$

$= 0.033768967 62$

≈ 0.03377 (4sf)

d) $P(X \geq 1) \geq 0.99$

$\therefore P(X = 0) \leq 0.01$

$\therefore e^{\lambda} \leq 0.01$

$\lambda \leq \ln 0.01$

$\lambda \geq 4.605$ (4sf)

$k = (4.605 / 1.2) \times 10$

$= 38.38$ (4sf)

≈ 38.4 minutes

It's over a 10-minute period

2 A researcher is investigating the concentration of bacteria and fungi in the air in buildings. The researcher selects a random sample of 12 buildings and measures the concentrations of bacteria, x, and fungi, y, in the air in each building. Both concentrations are measured in the same standard units. Fig. 2 illustrates the data collected. The researcher wishes to test for a relationship between x and y.

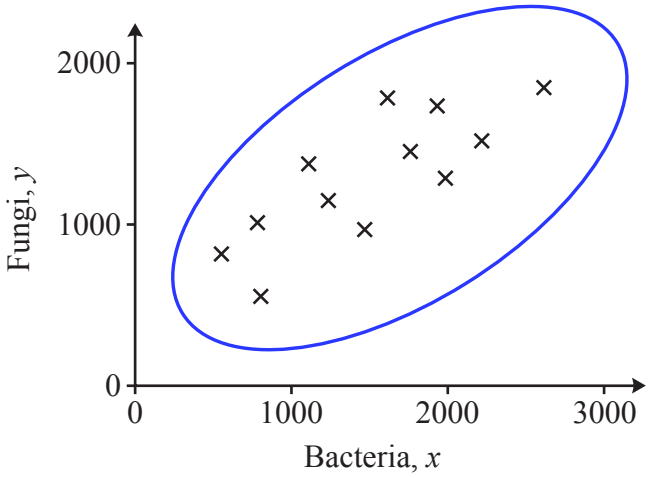

Fig. 2

(a) Explain why a test based on the product moment correlation coefficient is likely to be appropriate for these data. **[2]**

Summary statistics for the data are as follows.

$n = 12 \quad \Sigma x = 18\,030 \quad \Sigma y = 15\,550 \quad \Sigma x^2 = 31\,458\,700 \quad \Sigma y^2 = 21\,980\,500 \quad \Sigma xy = 25\,626\,800$

(b) In this question you must show detailed reasoning.

Calculate the product moment correlation coefficient between x and y. **[4]**

(c) Carry out a test at the 5% significance level based on the product moment correlation coefficient to investigate whether there is any correlation between concentrations of bacteria and fungi. **[5]**

(d) Explain why, in order for proper inference to be undertaken, the sample should be chosen randomly. **[1]**

2 a) A test based on the product moment correlation coefficient is likely to be appropriate as the data is in an elliptical shape. This suggests data is drawn from a bivariate normal population

$n = 12 \quad \Sigma x = 18\,030 \quad \Sigma y = 15\,550 \quad \Sigma x^2 = 31\,458\,700 \quad \Sigma y^2 = 21\,980\,500 \quad \Sigma xy = 25\,626\,800$

b) $PMCC = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n} \rightarrow 31458700 - \dfrac{(18030)^2}{12} = 4368625$

$S_{yy} = \Sigma y^2 - \dfrac{(\Sigma y)^2}{n} \rightarrow 21980500 - \dfrac{(15550)^2}{12} = 1830291.667$

$S_{xy} = \Sigma xy - \dfrac{\Sigma x \Sigma y}{n} \rightarrow 25626800 - \dfrac{(18030)(15550)}{12} = 2262925$

$PMCC = \dfrac{2262925}{\sqrt{4368625 \times 1830291.667}} = 0.800272007$

$r \approx 0.8003$ (usf)

c) @ 5%

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$ (two-tailed)

where ρ is the correlation coefficient between concentrations of bacterial and fungi in the background population

$r = 0.8003$

@ 5% Sig level and $n = 12$

the CV is 0.5760

$0.8003 > 0.5760$, result is significant

Therefore there is significant evidence to reject the null hypothesis. would suggest there is a positive correlation between concentrations of fungi and bacteria.

---

d) Because then the probability basis on which the sample has been selected is known

3 A child is trying to draw court cards from an ordinary pack of 52 cards (court cards are Kings, Queens and Jacks; there are 12 in a pack). She draws cards, one at a time, with replacement, from the pack.

Find the probabilities of the following events.

(a) She draws a court card for the first time on the sixth try. **[2]**

(b) She draws a court card at least once in the first six tries. **[2]**

(c) She draws a court card for the second time on the sixth try. **[2]**

(d) She draws at least two court cards in the first six tries. **[2]**

no. of not court cards = 40

3a) $\left(\frac{40}{52}\right)^5\left(\frac{12}{52}\right) = \left(\frac{10}{13}\right)^5\left(\frac{3}{13}\right)$ → (not court card 5 times)(court card once)

$= 0.0621528631$

≈ 0.06215 (4sf)

b) $1 - \left(\frac{10}{13}\right)^6 = 0.7928237889$

≈ 0.7928 (4sf)

1 − P(not court card in first 6 tries)

c) $5 \times \left(\frac{10}{13}\right)^4\left(\frac{3}{13}\right)^2$ → The first court card can be drawn any time in cards 1–5 hence the "×5".

$= 0.0932297949 6$

≈ 0.09323 (4sf)

$X \sim B(6, 3/13)$

d) $P(X \geq 2) = 1 - P(X \leq 1)$

$= 1 - 0.5800933909$

$= 0.4199066091$

≈ 0.4199 (4sf)

4 A fair 8-sided dice has faces labelled 1, 2, …, 8. The random variable X represents the score when the dice is rolled once.

(a) State the distribution of X. **[2]**

(b) Find $P(X < 4)$. **[1]**

(c) Find each of the following.

- $E(X)$
- $Var(X)$ **[2]**

(d) The random variable Y is defined by $Y = 10X + 5$. Find each of the following.

- $E(Y)$
- $Var(Y)$ **[3]**

4 a) Uniform distribution

r	1	2	3	4	5	6	7	8
P(X = r)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

b) $P(X < 4) = 1/8 + 1/8 + 1/8$ (x = 1, x = 2, x = 3)

$= 3/8 = 0.375$

c) $E(X) = \frac{1}{8} + \frac{2}{8} + \frac{3}{8} + \frac{4}{8} + \frac{5}{8} + \frac{6}{8} + \frac{7}{8} + \frac{8}{8}$

$= 4.5$

$E(X^2) = \frac{1}{8} + \frac{4}{8} + \frac{9}{8} + \frac{16}{8} + \frac{25}{8} + \frac{36}{8} + \frac{49}{8} + \frac{64}{8}$

$= 25.5$

$\therefore Var(X) = 25.5 - 4.5^2$

$= 5.25$

d) $Y = 10(X) + 5$

$E(Y) = 10\,E(X) + 5 = 45 + 5 = 50$

$Var(Y) = 10^2 (Var(X)) = 100 \times 5.25$

$= 525$

5 A doctor is investigating the relationship between the levels in the blood of a particular hormone and of calcium in healthy adults. The levels of the hormone and of calcium, each measured in suitable units, are denoted by x and y respectively.

The doctor selects a random sample of 14 adults and measures the hormone and calcium levels in each of them. The spreadsheet in Fig. 5 shows the values obtained, together with a scatter diagram which illustrates the data. The equation of the regression line of y on x is shown on the scatter diagram, together with the value of the square of the product moment correlation coefficient.

	A	B
1	**Hormone level *x***	**Calcium level *y***
2	110	10.6
3	130	9.6
4	135	8.9
5	155	8.4
6	162	9.6
7	165	9.1
8	170	8.6
9	188	8.9
10	190	7.8
11	200	7.9
12	210	8.2
13	225	7.4
14	182	8.1
15	174	8.6

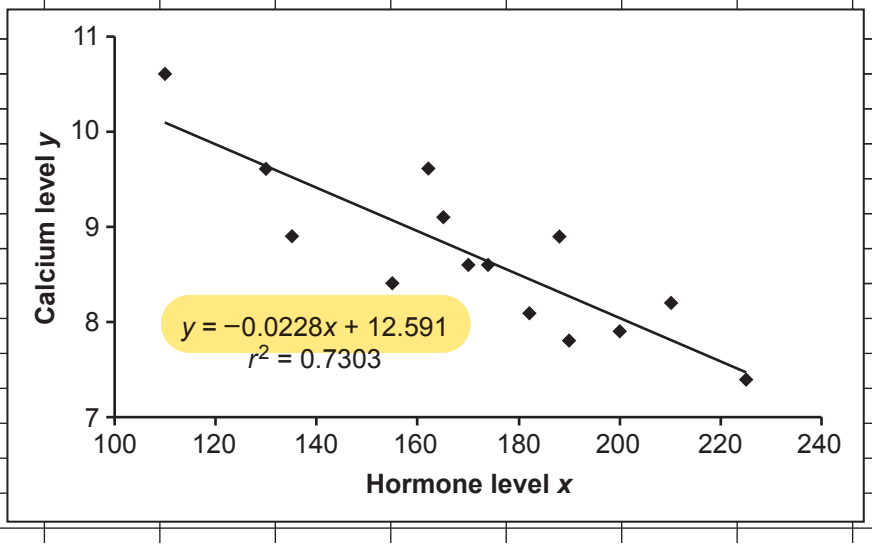

Fig. 5

(a) Use the equation of the regression line to estimate the mean calcium level of people with the following hormone levels.

- 150
- 250

[2]

(b) Explain which of your two estimates is likely to be more reliable. **[1]**

(c) Comment on the goodness of fit of the regression line. **[2]**

(d) Explain whether it would be appropriate to plot the scatter diagram the other way around with calcium level on the horizontal axis and hormone level on the vertical axis. **[1]**

(e) Calculate the equation of a regression line which would be suitable for estimating the mean hormone level of people with a known calcium level. **[2]**

5a. $y = -0.0228x + 12.591$

$150 \rightarrow y = -0.0228(150) + 12.591$

$= 9.171$

$250 \rightarrow y = -0.0228(250) + 12.591$

$= 6.891$

b) 150 is likely to be more reliable than 250 as 250 is extrapolation, and 150 is interpolation.

c) $r^2 = 0.7303$ and the points are close to a straight line, the fit of the regression line is fairly good

d) It would be appropriate as the variables are random on random

e) $a = 499.8254345 \approx 499.8$ (usf)

$b = -32.05880101 \approx -32.06$ (usf)

$\therefore x = -32.06y + 499.8$

note: to get this swap the x and y columns in columns A and B, then put it into your calculator.

6 A researcher is investigating whether there is any relationship between whether a cyclist wears a helmet and the distance, x m, the cyclist is from the kerb (the edge of the road). Data are collected at a particular location for a random sample of 250 cyclists.

The researcher carries out a chi-squared test. Fig. 6 is a screenshot showing part of a spreadsheet used to analyse the data. Some values in the spreadsheet have been deliberately omitted.

	A	B	C	D	E	F	G
1			**Observed frequency**				
2			**$x \leq 0.3$**	**$0.3 < x \leq 0.5$**	**$0.5 < x \leq 0.8$**	**$x > 0.8$**	**Totals**
3	**Wears helmet**	**Yes**	26	27	23	46	122
4		**No**	45	31	21	31	128
5		**Totals**	71	58	44	77	250
6							
7			**Expected frequency**				
8			**$x \leq 0.3$**	**$0.3 < x \leq 0.5$**	**$0.5 < x \leq 0.8$**	**$x > 0.8$**	
9	**Wears helmet**	**Yes**	34.6480			37.5760	
10		**No**	36.3520			39.4240	
11							
12			**Contribution to the test statistic**				
13			**$x \leq 0.3$**	**$0.3 < x \leq 0.5$**	**$0.5 < x \leq 0.8$**	**$x > 0.8$**	
14	**Wears helmet**	**Yes**	2.1585	0.0601	0.1087	1.8885	
15		**No**	2.0573	0.0573		1.8000	
16							

Fig. 6

(a) Showing your calculations, find the missing values in each of the following cells.

- E10
- E15 **[3]**

(b) In this question you must show detailed reasoning.

Carry out a hypothesis test at the 10% significance level to investigate whether there is any association between helmet wearing and distance from the kerb. **[6]**

(c) Discuss briefly what the data suggest about helmet wearing for different distances from the kerb. **[3]**

6 a) E10 : $\frac{44}{250} \times \frac{128}{250} \times 250 = 22.528$

E15 : $\frac{(21 - 22.528)^2}{22.528} = 0.103639205$

≈ 0.1036 (4sf)

b) 10% sig level

H_0 : there is no association between helmet wearing and distance from the kerb

H_1 : there is some association between helmet wearing and distance from the kerb

$\chi^2 = 2.1585 + 0.0601 + 0.1087 + 1.885 + 2.0573 + 0.0573 + 0.1036 + 1.800$

$= 8.2305 \approx 8.234$ (4sf)

$\nu = (4-1)(2-1) = 3$

CV at $\nu = 3$ and 10% sig level = 6.251

$8.234 > 6.251$, result is significant

Sufficient evidence to reject H_0, would suggest there is some association between helmet wearing and distance from the kerb and they are not independent.

c) 1. for distances of 0.3 and below the contributions of 2.1585 and 2.0573, more people do not wear helmets and fewer do wear helmets than expected.

2. for distances greater than 0.8 the contributions of 1.8885 and 1.8000 suggest that more people do wear helmets and fewer do not wear helmets than expected

3. for distances between 0.3 and 0.8, it is as expected if there was no association

END OF QUESTION PAPER

BLANK PAGE