OCR
Oxford Cambridge and RSA

Friday 17 June 2022 – Afternoon

A Level Further Mathematics B (MEI)

Y432/01 Statistics Minor

Time allowed: 1 hour 15 minutes

9452616730

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.

© OCR 2022 [M/508/5596]
DC (ST/SG) 300832/3

OCR is an exempt Charity

stemup.co.uk

Answer **all** the questions.

1 In a quiz a contestant is asked up to four questions. The contestant's turn ends once the contestant gets a question wrong or has answered all four questions. The probability that a particular contestant gets any question correct is 0.6, independently of other questions. The discrete random variable X models the number of questions which the contestant gets correct in a turn.

(a) Show that $P(X = 4) = 0.1296$. **[1]**

The probability distribution of X is shown in **Fig. 1.1**.

r	0	1	2	3	4
$P(X = r)$	0.4	0.24	0.144	0.0864	0.1296

Fig. 1.1

(b) Find each of the following.

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

The number of points that a contestant scores is as shown in **Fig. 1.2**.

Number of questions correct	Number of points scored
0 or 1	0
2	2
3	3
4	5

Fig. 1.2

The discrete random variable Y models the number of points which the contestant scores.

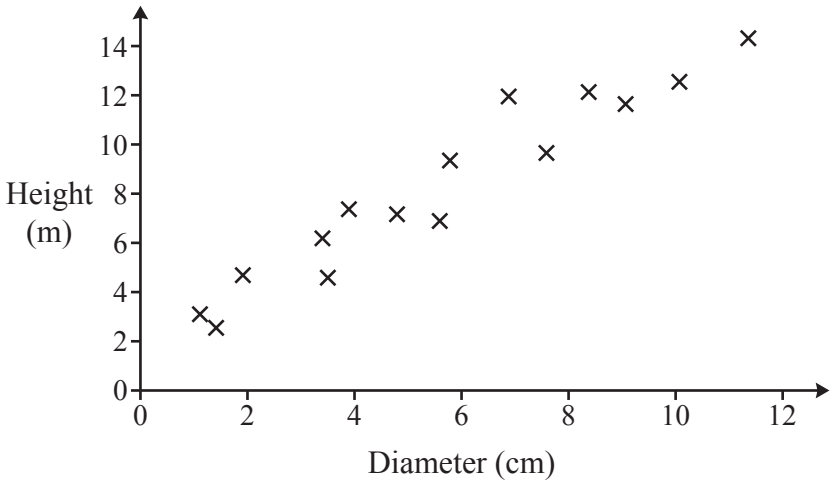
(c) Without doing any working, explain whether each of the following will be less than, equal to or greater than the corresponding value for X.

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

© OCR 2022

stemup.co.uk

2 A forester is investigating the relationship between the diameter and the height of young beech trees. She selects a random sample of 15 young beech trees in a forest and records their diameters, d cm, and their heights, h m. The data are illustrated in the scatter diagram.



(a) State whether either or both of the variables d and h are random variables. **[1]**

Summary data for the diameters and heights are as follows.

$n = 15 \quad \Sigma d = 84.9 \quad \Sigma h = 124.7 \quad \Sigma d^2 = 624.55 \quad \Sigma h^2 = 1230.57 \quad \Sigma dh = 866.63$

(b) Find the equation of the regression line of h on d. Give your answer in the form $h = ad + b$, giving the values of a and b correct to **2** decimal places. **[4]**

(c) Use the regression line to predict the heights of beech trees with the following diameters.

- 7.5 cm
- 20.0 cm

[2]

(d) Comment on the reliability of your predictions. **[3]**

(e) There are many mature beech trees with diameter of 60 cm or greater. However, there are no beech trees with a height of more than 50 m.

Comment on this in relation to your regression line. **[2]**

(f) State the coordinates of the point at which the regression line of d on h meets the line which you calculated in part **(b)**. **[1]**

stemup.co.uk

3 Jane wonders whether the number of wasps entering a wasp's nest per 5 second interval can be modelled by a Poisson distribution with mean μ. She counts the number of wasps entering the nest over 60 randomly selected 5-second intervals. The results are shown in **Fig. 3.1**.

Number of wasps	0	1	2	3	4	5	6	7	8	9	⩾10
Frequency	0	2	5	5	12	10	10	11	1	4	0

Fig. 3.1

(a) Show that a suitable estimate for the value of μ is 5.1. **[2]**

Fig. 3.2 shows part of a screenshot for a χ^2 test to assess the goodness of fit of a Poisson model. The sample mean has been used as an estimate for the population mean. Some of the values in the spreadsheet have been deliberately omitted.

	A	B	C	D	E
1	Number of wasps	Observed frequency	Poisson probability	Expected frequency	Chi-squared contribution
2	⩽2	7	0.1165	6.9887	0.0000
3	3	5		8.0874	1.1786
4	4	12			0.2765
5	5	10			0.0255
6	6	10	0.1490	8.9400	0.1257
7	7	11	0.1086	6.5134	3.0904
8	⩾8	5	0.1440	8.6414	
9					

Fig. 3.2

(b) Determine the missing values in each of the following cells, giving your answers correct to **4** decimal places.

- C3
- D5
- E8

[4]

(c) Explain why some of the frequencies have been combined into the categories ⩽2 and ⩾8. **[1]**

(d) **In this question you must show detailed reasoning.**

Carry out the hypothesis test at the 5% significance level. **[6]**

(e) Jane also carries out a χ^2 test for the number of wasps leaving another nest. As part of her calculations, she finds that the probability of no wasps leaving the nest in a 5-second period is 0.0053. She finds that a Poisson distribution is also an appropriate model in this case.

Find a suitable estimate for the value of the mean number of wasps leaving the nest per 5-second period. **[2]**

© OCR 2022

stemup.co.uk

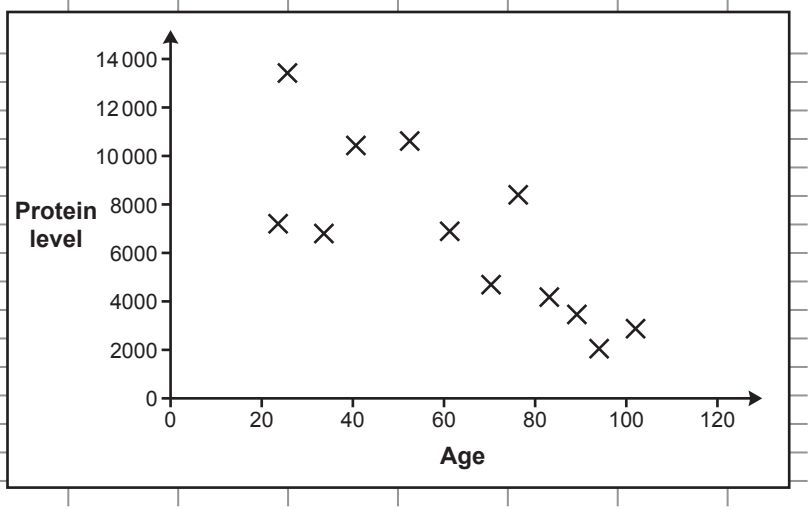
4 Alex is practising bowling at a cricket wicket. Every time she bowls a ball, she has a 30% chance of hitting the wicket.

(a) Assuming that successive bowls are independent, determine the probability that Alex first hits the wicket on her third attempt. **[2]**

(b) Determine the probability that Alex hits the wicket for the fourth time on her tenth attempt. **[3]**

5 A medical researcher is investigating whether there is any relationship between the age of a person and the level of a particular protein in the person's blood. She measures the levels of the protein (measured in suitable units) in a random sample of 12 hospital patients of various ages (in years). The spreadsheet shows the values obtained, together with a scatter diagram which illustrates the data.

	A	B
1	**Age**	**Protein level**
2	23	7200
3	25	13400
4	33	6800
5	40	10400
6	52	10600
7	61	6900
8	70	4700
9	76	8400
10	83	4200
11	89	3500
12	94	2100
13	102	2900



(a) The researcher decides that a test based on Pearson's product moment correlation coefficient may **not** be valid. Explain why she comes to this conclusion. **[2]**

(b) Calculate the value of Spearman's rank correlation coefficient. **[3]**

(c) Carry out a test based on this coefficient at the 5% significance level to investigate whether there is any association between age and protein level. **[5]**

(d) Explain why the researcher chose a sample that was random. **[2]**

(e) The researcher had originally intended to use a sample size of 6 rather than the 12 that she actually used.

Explain what advantage there is in using the larger sample size. **[2]**

© OCR 2022

6 The random variable X has a uniform distribution over the values $\{1, 4, 7, \ldots, 3n - 2\}$, where n is a positive integer.

(a) Determine $\text{Var}(X)$ in terms of n. **[3]**

(b) Given that $n = 100$, find the probability that X is within one standard deviation of the mean. **[4]**

END OF QUESTION PAPER

© OCR 2022

stemup.co.uk

BLANK PAGE

© OCR 2022

stemup.co.uk

OCR
Oxford Cambridge and RSA

Copyright Information

OCR is committed to seeking permission to reproduce all third-party content that it uses in its assessment materials. OCR has attempted to identify and contact all copyright holders whose work is used in this paper. To avoid the issue of disclosure of answer-related information to candidates, all copyright acknowledgements are reproduced in the OCR Copyright Acknowledgements Booklet. This is produced for each series of examinations and is freely available to download from our public website (www.ocr.org.uk) after the live examination series.

If OCR has unwittingly failed to correctly acknowledge or clear any third-party content in this assessment material, OCR will be happy to correct its mistake at the earliest possible opportunity.

For queries or further information please contact The OCR Copyright Team, The Triangle Building, Shaftesbury Road, Cambridge CB2 8EA.

OCR is part of Cambridge University Press & Assessment, which is itself a department of the University of Cambridge.

© OCR 2022

stemup.co.uk

Surname	StemUp Model Solution
Answer booklet	Write the question number in the left-hand margin at the start of each answer.

Question number

1 (a)

$X = 4$ requires all four questions correct

$P(X = 4) = 0.6^4 = 0.1296$ as required

1 (b)

$E(X) = 0(0.4) + 1(0.24) + 2(0.144) + 3(0.0864) + 4(0.1296)$

$E(X) = 1.3056 = 1.31$ (3 s.f.)

$E(X^2) = 1(0.24) + 4(0.144) + 9(0.0864) + 16(0.1296) = 3.6672$

$\text{Var}(X) = 3.6672 - 1.3056^2 = 1.9626... = 1.96$ (3 s.f.)

1 (c)

$E(Y)$ is less than $E(X)$: $r = 1$ becomes 0 points and $r = 4$

becomes 5 points, but $P(X = 1) = 0.24 > P(X = 4) = 0.1296$,

so the loss of 1 outweighs the gain of 1 on average.

$\text{Var}(Y)$ is greater than $\text{Var}(X)$: the values are moved further

apart (0 and 5 are more extreme), so Y is more spread out.

2 (a)

Both d and h are random variables (neither was controlled).

2 (b)

$$S_{dh} = 866.63 - \frac{84.9 \times 124.7}{15} = 160.828$$

$$S_{dd} = 624.55 - \frac{84.9^2}{15} = 144.016$$

$$a = \frac{S_{dh}}{S_{dd}} = \frac{160.828}{144.016} = 1.1167...$$

$$\bar{d} = \frac{84.9}{15} = 5.66, \quad \bar{h} = \frac{124.7}{15} = 8.3133...$$

$$b = \bar{h} - a\bar{d} = 8.3133... - 1.1167... \times 5.66 = 1.9926...$$

Question number

$h = 1.12d + 1.99$

2 (c)

$d = 7.5$: $h = 1.1167... \times 7.5 + 1.9926... = 10.4$ m

$d = 20.0$: $h = 1.1167... \times 20.0 + 1.9926... = 24.3$ m

2 (d)

$d = 7.5$ is within the range of the data (interpolation) and the points lie close to the line (strong linear correlation), so the prediction of 10.4 m is likely to be reliable.

$d = 20.0$ is well beyond the data (extrapolation, max $d \approx 11$), so the prediction of 24.3 m is unlikely to be reliable.

2 (e)

For $d = 60$: $h = 1.1167... \times 60 + 1.9926... = 69.0$ m, which is more than 50 m, so it is impossible.

So the regression line is not valid for mature trees: the relationship is not linear for large diameters.

2 (f)

Both regression lines pass through $(\bar{d}, \bar{h})$

(5.66, 8.31)

3 (a)

$$\bar{x} = \frac{1\times2 + 2\times5 + 3\times5 + 4\times12 + 5\times10 + 6\times10 + 7\times11 + 8\times1 + 9\times4}{60}$$

$$= \frac{306}{60} = 5.1 \text{ as required}$$

3 (b)

$$C3 = P(X = 3) = e^{-5.1} \times \frac{5.1^3}{3!} = 0.1348$$

$$DS = 60 \times P(X = 5) = 60 \times e^{-5.1} \times \frac{5.1^5}{5!} = 10.5177$$

Question number	
	$E8 = \frac{(5 - 8.6414)^2}{8.6414} = 1.5344$
3 (c)	Otherwise some expected frequencies would be less than 5, so the χ^2 test would not be valid.
3 (d)	H_0: a Poisson model is a good fit for the number of wasps entering the nest per 5-second interval H_1: a Poisson model is not a good fit $X^2 = 0.0000 + 1.1786 + 0.2765 + 0.0255 + 0.1257 + 3.0904 + 1.5344$ $X^2 = 6.231$ (3 d.p.) Cells = 7, constraints: total and mean estimated degrees of freedom = 7 − 2 = 5 Critical value χ^2_5 at 5% = 11.07 6.231 < 11.07, so do not reject H_0 There is insufficient evidence to suggest that the Poisson model is not a good fit for the number of wasps entering the nest.
3 (e)	$P(X = 0) = e^{-\lambda} = 0.0053$ $\lambda = -\ln 0.0053 = 5.24$ (3 s.f.)
4 (a)	X = number of attempts to first hit, $X \sim Geo(0.3)$ $P(X = 3) = 0.7^2 \times 0.3 = 0.147$
4 (b)	Need exactly 3 hits in the first 9 attempts, then a hit on the 10th Hits in first 9 $\sim B(9, 0.3)$ $P(3 \text{ hits in } 9) = {}^9C_3 \times 0.3^3 \times 0.7^6 = 0.26682...$ $P = 0.26682... \times 0.3 = 0.0800$ (3 s.f.)

Question number

5 (a) The scatter diagram is not elliptical (it is more funnel shaped), so the data are unlikely to come from a bivariate Normal distribution, which a test based on Pearson's pmcc requires.

5 (b)

Rank age:	1	2	3	4	5	6	7	8	9	10	11	12
Rank protein:	8	12	6	10	11	7	5	9	4	3	1	2
d^2:	49	100	9	36	36	1	4	1	25	49	100	100

$\Sigma d^2 = 510$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 510}{12 \times 143}$$

$r_s = -0.783$ (3 s.f.)

5 (c) H_0: there is no association between age and protein level in the population

H_1: there is some association between age and protein level in the population

$n = 12$, 2-tailed test at 5%: critical value = 0.5874

$|-0.783| = 0.783 > 0.5874$, so reject H_0

There is sufficient evidence to suggest that there is association between age and protein level in the population.

5 (d) A random sample is needed so that valid inferences can be made about the population (it avoids bias, so the sample is representative and the test is valid).

5 (e) With a larger sample, the random variation in the sample tends to be smaller, so the sample value of r_s tends to be

Question number

closer to the population coefficient (a more reliable test).

6 (a) $X = 3W - 2$ where W is uniform on $\{1, 2, ..., n\}$

$\text{Var}(W) = \frac{1}{12}(n^2 - 1)$

$\text{Var}(X) = 3^2 \text{Var}(W) = 9 \times \frac{1}{12}(n^2 - 1)$

$\text{Var}(X) = \frac{3}{4}(n^2 - 1)$

6 (b) $n = 100$: $E(X) = 3E(W) - 2 = 3 \times 50.5 - 2 = 149.5$

$SD(X) = \sqrt{\frac{3}{4}(100^2 - 1)} = \sqrt{7499.25} = 86.598...$

Require $149.5 - 86.598... \leq x \leq 149.5 + 86.598...$

$62.90... \leq x \leq 236.09...$

$x = 3w - 2 \Rightarrow 21.63... \leq w \leq 79.36...$

so $w = 22, 23, ..., 79$: that is 58 values out of 100

$\text{Probability} = \frac{58}{100} = 0.58$