OCR
Oxford Cambridge and RSA

Thursday 7 October 2021 – Afternoon

A Level Further Mathematics B (MEI)

Y432/01 Statistics Minor

Time allowed: 1 hour 15 minutes

9050968036

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 2021 [M/508/5596]
DC (CE/CB) 300212/3

OCR is an exempt Charity

stemup.co.uk

Answer **all** the questions.

1 The probability distribution of a discrete random variable X is given by the formula

$\mathrm{P}(X = r) = k\left((r-1)^2 + 1\right)$ for $r = 1, 2, 3, 4, 5$.

(a) Show that $k = \frac{1}{35}$. **[2]**

The distribution of X is shown in the table.

r	1	2	3	4	5
$\mathrm{P}(X = r)$	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{17}{35}$

(b) Comment briefly on the shape of the distribution. **[1]**

(c) Find each of the following.

- $\mathrm{E}(X)$
- $\mathrm{Var}(X)$ **[2]**

The random variable Y is given by $Y = 5X - 10$.

(d) Find each of the following.

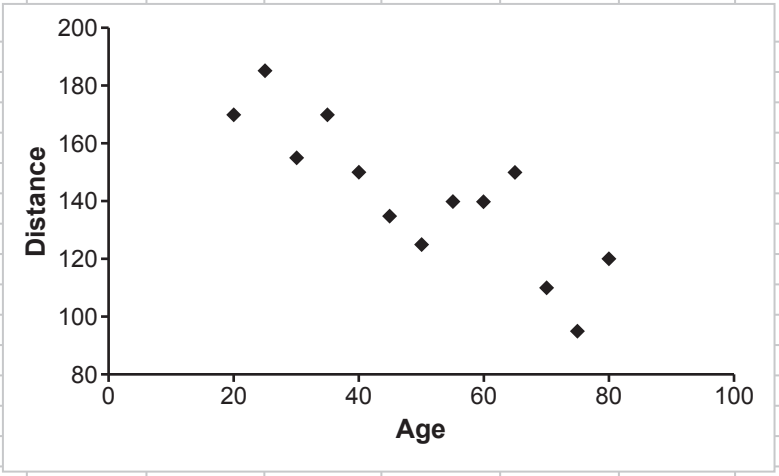
- $\mathrm{E}(Y)$
- $\mathrm{Var}(Y)$ **[2]**

© OCR 2021

stemup.co.uk

2 A road transport researcher is investigating the link between the age of a person, a years, and the distance, d metres, at which the person can read a large road sign. The researcher selects 13 individuals of different ages between 20 and 80 and measures the value of d for each of them. The spreadsheet below shows the data which the researcher obtained, together with a scatter diagram which illustrates the data.

	A	B
1	Age (years)	Distance (metres)
2	20	170
3	25	185
4	30	155
5	35	170
6	40	150
7	45	135
8	50	125
9	55	140
10	60	140
11	65	150
12	70	110
13	75	95
14	80	120



(a) Explain which of the two variables a and d is the independent variable. **[1]**

(b) Find the equation of the regression line of d on a. **[2]**

(c) Use the regression line to predict the average distance at which a 60-year-old person can read the road sign. **[1]**

(d) Explain why it might not be sensible to use the regression line to predict the average distance at which a 5-year-old child can read the road sign. **[2]**

(e) Determine the value of the residual for $a = 40$. **[2]**

(f) Explain why it would not be useful to find the equation of the regression line of a on d. **[1]**

© OCR 2021

stemup.co.uk

3 A student wants to know whether there is any association between age and whether or not people smoke. The student takes a sample of 120 adults and asks each of them whether or not they smoke. Below 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
1			Observed frequency		
2			Age		
3			16-34	35-59	60 and over
4	Smoking status	Smoker	13	7	3
5		Non-smoker	28	43	26
6					
7			Expected frequency		
8			7.8583		
9			33.1417		
10					
11			Contributions to the test statistic		
12			3.3642	0.6964	1.1775
13				0.1651	0.2792

(a) The student wants to carry out a chi-squared test to analyse the data.

State a requirement of the sample if the test is to be valid. **[1]**

For the rest of this question, you should assume that this requirement is met.

(b) Determine the missing values in each of the following cells.

- E8
- C13

[3]

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

Carry out a hypothesis test at the 5% significance level to investigate whether there is any association between age and smoking status. **[6]**

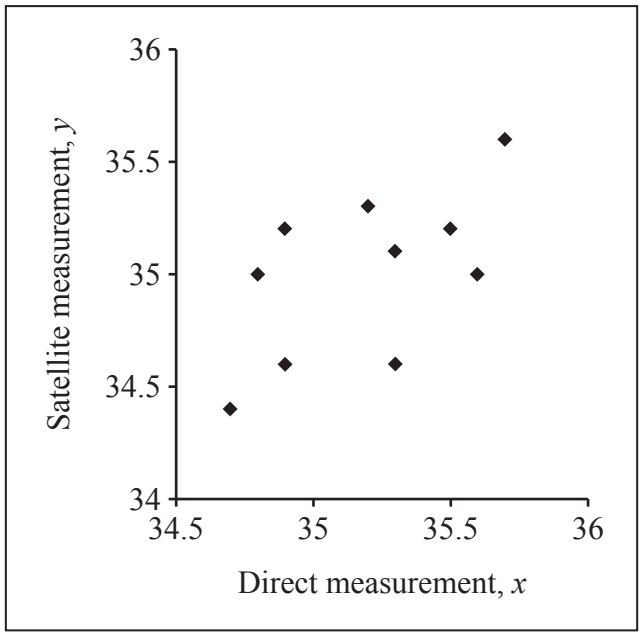
(d) Discuss what the data suggest about the smoking status for each different age group. **[3]**

© OCR 2021

stemup.co.uk

4 A scientist is investigating sea salinity (the level of salt in the sea) in a particular area. She wishes to check whether satellite measurements, y, of salinity are similar to those directly measured, x. Both variables are measured in parts per thousand in suitable units.

The scientist obtains a random sample of 10 values of x and the related values of y. Below is a screenshot of a scatter diagram to illustrate the data. She decides to carry out a hypothesis test to check if there is any correlation between direct measurement, x, and satellite measurement, y.



(a) Explain why the scientist might decide to carry out a test based on the product moment correlation coefficient. **[2]**

Summary statistics for x and y are as follows.

$n = 10$ $\quad \Sigma x = 351.9$ $\quad \Sigma y = 350.0$ $\quad \Sigma x^2 = 12\,384.5$ $\quad \Sigma y^2 = 12\,251.2$ $\quad \Sigma xy = 12\,317.2$

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

Calculate the product moment correlation coefficient. **[4]**

(c) Carry out a hypothesis test at the 5% significance level to investigate whether there is positive correlation between directly measured and satellite measured salinity levels. **[5]**

(d) Explain why it would be preferable to use a larger sample. **[1]**

The scientist is also interested in whether there is any correlation between salinity and numbers of a particular species of shrimp in the water. She takes a large sample and finds that the product moment correlation coefficient for this sample is 0.165. The result of a test based on this sample is to reject the null hypothesis and conclude that there is correlation between salinity and numbers of shrimp.

(e) Comment on the outcome of the hypothesis test with reference to the effect size of 0.165. **[2]**

© OCR 2021

stemup.co.uk

5 Biological cell membranes have receptor molecules which perform various functions. It is known that the number of receptor molecules of a particular type can be modelled by a Poisson distribution with mean 6 per area of 1 square unit.

(a) (i) Determine the probability that there are at least 10 of these receptor molecules in an area of 1 square unit. **[2]**

(ii) Determine the probability that there are fewer than 50 of these receptor molecules in an area of 10 square units. **[2]**

(b) A scientist is looking at areas of 1 square unit of cell membrane in order to find one which has at least 10 receptor molecules.

Find the probability that she has to look at more than 20 to find such an area. **[2]**

It is known that the number of receptor molecules of another type in an area of 1 square unit can be modelled by the random variable X which has a Poisson distribution with mean μ. It is given that $\mathrm{E}(X^2) = 12$.

(c) Determine $\mathrm{P}(X < 5)$. **[4]**

6 A lottery has tickets numbered 1 to n inclusive, where n is a positive integer. The random variable X denotes the number on a ticket drawn at random.

(a) Determine $\mathrm{P}(X \leqslant \frac{1}{4}n)$ in each of the following cases.

(i) n is a multiple of 4. **[1]**

(ii) n is of the form $4k+1$, where k is a positive integer. Give your answer as a single fraction in terms of n. **[3]**

(b) Given that $n = 101$, find the probability that X is within one standard deviation of the mean. **[3]**

END OF QUESTION PAPER

© OCR 2021

stemup.co.uk

BLANK PAGE

© OCR 2021

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 the Cambridge Assessment Group; Cambridge Assessment is the brand name of University of Cambridge Local Examinations Syndicate (UCLES), which is itself a department of the University of Cambridge.

© OCR 2021

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) $P(X = r) = k((r-1)^2 + 1)$ for $r = 1, 2, 3, 4, 5$

$k(1) + k(2) + k(5) + k(10) + k(17) = 1$

$35k = 1$

$k = \frac{1}{35}$ as required

1 (b) The distribution has strong negative skew.

1 (c) $E(X) = \frac{1(1) + 2(2) + 3(5) + 4(10) + 5(17)}{35} = \frac{145}{35}$

$E(X) = \frac{29}{7} = 4.14$ (3 s.f.)

$E(X^2) = \frac{1(1) + 4(2) + 9(5) + 16(10) + 25(17)}{35} = \frac{639}{35}$

$Var(X) = \frac{639}{35} - (\frac{29}{7})^2 = \frac{268}{245} = 1.09$ (3 s.f.)

1 (d) $E(Y) = 5E(X) - 10 = 5 \times \frac{29}{7} - 10 = \frac{75}{7} = 10.7$ (3 s.f.)

$Var(Y) = 5^2 Var(X) = 25 \times \frac{268}{245} = \frac{1340}{49} = 27.3$ (3 s.f.)

2 (a) a is the independent variable, since the values of a were chosen by the researcher (controlled), so a is not subject to random variation, whereas d is.

Question number	
2 (b)	$\Sigma a = 650$, $\Sigma d = 1845$, $\Sigma a^2 = 37050$, $\Sigma ad = 87225$, $n = 13$
	$S_{ad} = 87225 - \frac{650 \times 1845}{13} = -5025$
	$S_{aa} = 37050 - \frac{650^2}{13} = 4550$
	$\text{gradient} = \frac{-5025}{4550} = -1.1044...$
	$\text{intercept} = \frac{1845}{13} - (-1.1044...)(50) = 197.14...$
	$d = -1.104a + 197.1$
2 (c)	$d = -1.1044... \times 60 + 197.14... = 130.9$ m
2 (d)	$a = 5$ is well outside the range of the data (20 to 80), so this would be extrapolation. The relationship may be different for young children (eg a 5-year-old child may not be able to read yet).
2 (e)	predicted d at $a = 40$: $-1.1044... \times 40 + 197.14... = 152.97...$
	residual = observed − predicted = $150 - 152.97...$
	residual = -3.0
2 (f)	The values of a are not random (they were chosen), so it makes no sense to try to predict a from d.
3 (a)	The sample must be a random sample (of adults).

Question number

3 (b) $E8 = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{23 \times 29}{120} = 5.5583$

$C13 = \frac{(28 - 33.1417)^2}{33.1417} = 0.7977$

3 (c) H_0: there is no association between age and smoking status

H_1: there is some association between age and smoking status

$X^2 = 3.3642 + 0.6964 + 1.1775 + 0.7977 + 0.1651 + 0.2792$

$X^2 = 6.4801$

degrees of freedom $= (2 - 1)(3 - 1) = 2$

critical value at 5%: $\chi^2_2 = 5.991$

$6.4801 > 5.991$, so reject H_0

There is sufficient evidence at the 5% level to suggest that there is association between age and smoking status.

3 (d) 16-34: the contribution 3.3642 is large and $O = 13 > E = 7.86$, so more people in this group smoke than would be expected.

35-59: the contributions (0.6964, 0.1651) are small, so the smoking status is roughly as expected if no association.

60 and over: the contribution 1.1775 is fairly large and $O = 3 < E = 5.56$, so fewer smoke than would be expected.

4 (a) The scatter diagram is roughly elliptical in shape, so the data may come from a bivariate Normal distribution, which is the condition for a test based on the pmcc.

4 (b) $S_{xy} = 12317.2 - \frac{351.9 \times 350.0}{10} = 0.7$

Question number	
	$S_{xx} = 12384.5 - \frac{351.9^2}{10} = 1.139$
	$S_{yy} = 12251.2 - \frac{350.0^2}{10} = 1.2$
	$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{0.7}{\sqrt{1.139 \times 1.2}} = 0.5987...$
	$r = 0.599$ (3 s.f.)
4 (c)	H_0: $\rho = 0$, H_1: $\rho > 0$
	where ρ is the population product moment correlation
	coefficient between x and y
	$n = 10$, one-tailed 5% critical value = 0.5494
	$0.599 > 0.5494$, so the result is significant: reject H_0
	There is sufficient evidence at the 5% level to suggest that
	there is positive correlation between directly measured
	and satellite measured salinity levels.
4 (d)	With a larger sample one can be more confident that the
	correlation is genuine rather than just the result of
	random variation.
4 (e)	The test shows there is almost certainly some real
	correlation in the population.
	However, the effect size of 0.165 is very small, so the
	correlation is weak and the result is of little practical use.
5(a)(i)	$X \sim Po(6)$
	$P(X \geq 10) = 1 - P(X \leq 9) = 1 - 0.91608...$

Question number	
	$P(X \geqslant 10) = 0.0839$ (3 s.f.)
5(a)(ii)	Y = number in 10 square units, $Y \sim Po(60)$
	$P(Y < 50) = P(Y \leqslant 49) = 0.0844$ (3 s.f.)
5 (b)	More than 20 needed means the first 20 areas all have
	fewer than 10 receptor molecules.
	$P = (P(X \leqslant 9))^{20} = 0.91608...^{20}$
	$P = 0.173$ (3 s.f.)
5 (c)	For a Poisson distribution, $Var(X) = E(X) = \mu$
	$Var(X) = E(X^2) - (E(X))^2$
	$\mu = 12 - \mu^2$
	$\mu^2 + \mu - 12 = 0 \Rightarrow (\mu + 4)(\mu - 3) = 0$
	$\mu = 3$ (since $\mu > 0$)
	$X \sim Po(3)$: $P(X < 5) = P(X \leqslant 4) = 0.815$ (3 s.f.)
6(a)(i)	X is uniform on $\{1, 2, ..., n\}$
	$n/4$ is an integer, so there are $n/4$ values with $X \leqslant n/4$
	$P(X \leqslant \frac{n}{4}) = \frac{n/4}{n} = \frac{1}{4}$
6(a)(ii)	$n = 4k + 1 \Rightarrow \frac{n}{4} = k + \frac{1}{4}$
	so $X \leqslant \frac{n}{4}$ means $X = 1, 2, ..., k$
	$P(X \leqslant \frac{n}{4}) = \frac{k}{4k + 1} = \frac{k}{n}$, where $k = \frac{n - 1}{4}$

Question number

$P(X \leq \frac{n}{4}) = \frac{n-1}{4n}$

6 (b)

$E(X) = \frac{n+1}{2} = \frac{102}{2} = 51$

$Var(X) = \frac{n^2-1}{12} = \frac{101^2-1}{12} = 850$

$SD(X) = \sqrt{850} = 29.15...$

Require $51 - 29.15... < X < 51 + 29.15...$

$21.85... < X < 80.15... \Rightarrow X = 22, 23, ..., 80$

that is 59 values, so $P(22 \leq X \leq 80) = \frac{59}{101}$