OCR
Oxford Cambridge and RSA

Tuesday 20 May 2025 – 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

QP

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.

1 A student is investigating what level of exercise is given to dogs in the UK by their owners.

(a) When trying to collect suitable information, give **two** reasons why it might be advantageous to use a sample of dog owners, rather than a census of dog owners. **[2]**

(b) Explain why it is preferable that the size of the sample is large. **[1]**

2 The probability distribution of a discrete random variable X is given in the table.

x	0	1	2	3
$\mathrm{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{3}$	a	$\frac{1}{10}$

(a) Find the value of a. **[1]**

The probability function of another discrete random variable, Y, is defined below.

$\mathrm{P}(Y = y) = \frac{y}{595}$ for $y = 1, 2, ..., n$,

where n is a positive integer.

(b) Determine the value of n. **[3]**

(c) Find $\mathrm{P}(Y < n)$. **[1]**

(d) Given that X and Y are independent, determine $\mathrm{P}(X + Y \leqslant 2)$. **[2]**

3 The discrete random variable Q has the distribution Geo(0.3).

(a) Find $\mathrm{P}(Q = 2)$. **[1]**

(b) Determine $\mathrm{P}(Q \leqslant 5)$. **[2]**

The discrete random variable R has a uniform distribution on $\{4, 5, ..., 8\}$.

The discrete random variable S has the distribution $\mathrm{B}(10, 0.1)$.

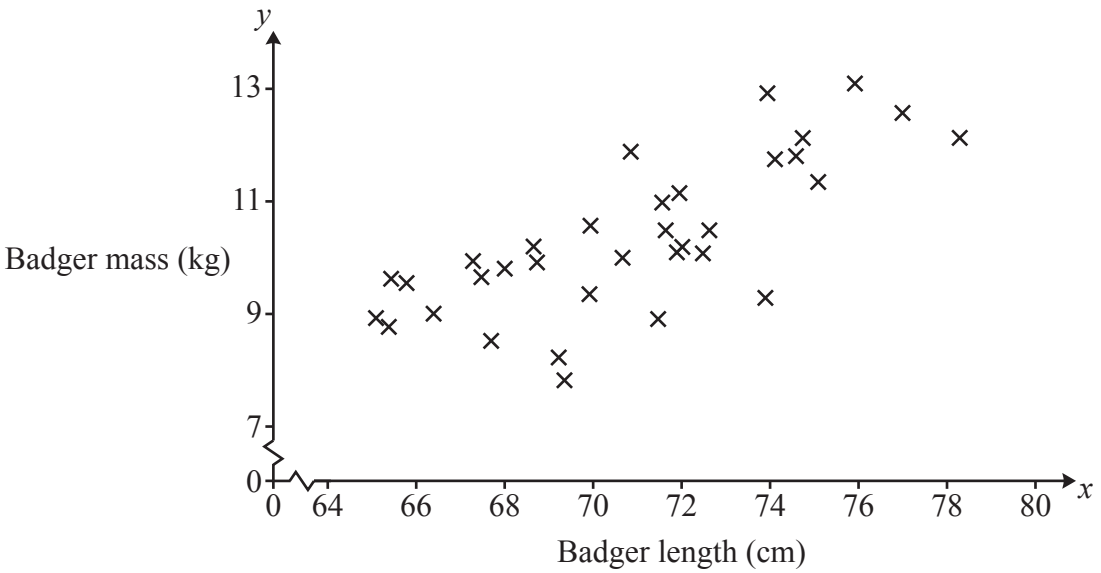
Q, R and S are mutually independent of each other.

(c) Find $\mathrm{P}(Q + R + S = 5)$ **[3]**

The discrete random variable T is defined by $T = R - Q + 2S$

(d) Determine the value of $\mathrm{E}(T)$. **[4]**

4 A scientist measures the length, x cm, and mass, y kg, of a random sample of 34 adult female badgers. A scatter diagram for the data is shown in the diagram.



(a) Give **two** reasons why it would be appropriate for a hypothesis test based on the product-moment correlation coefficient to be carried out in this situation. **[2]**

The scientist analyses the collected data and produces the following summary statistics.

$n = 34 \quad \Sigma x = 2409.4 \quad \Sigma y = 350.87 \quad \Sigma x^2 = 171140.28 \quad \Sigma y^2 = 3681.851 \quad \Sigma xy = 24\,981.486$

(b) Find and use the equation of a suitable regression line to determine an estimate of the length of an adult female badger of mass 10 kg. **[4]**

5 Two judges, Judge A and Judge B, independently rank 8 entries in a jam-making competition, awarding each jam a rank from 1 (best) to 8 (worst). There are no tied ranks. The ranks are shown in the table below but unfortunately two of the rankings awarded by Judge B have been lost.

	Jam 1	Jam 2	Jam 3	Jam 4	Jam 5	Jam 6	Jam 7	Jam 8
Judge A	4	8	7	1	5	6	3	2
Judge B	4		5	2	6		1	3

(a) Explain which jam is likely to be the best. **[1]**

Before Judge B's ranks were lost, Spearman's rank correlation coefficient was calculated for all the data and found to be $\frac{6}{7}$.

(b) Determine the missing ranks that Judge B gave, specifying which rank was given to which jam. **[3]**

The two judges want to assess whether there is positive association between their rankings of jam.

(c) Carry out a hypothesis test to make the assessment using a significance level of 1%. **[5]**

6 The number of misprints in every 100 000 words of a book is denoted by the random variable X. The probability of any given word being misprinted is 0.000 11. You may assume that misprints occur randomly and independently of each other.

(a) (i) Explain why a certain binomial distribution could be used to model the distribution of X. **[2]**

(ii) Hence explain why X could also be modelled by a certain Poisson distribution. **[1]**

(iii) State the mean of this Poisson distribution. **[1]**

For the remainder of this question, you should use a **Poisson** distribution to model the number of misprints in a book.

(b) For a book containing exactly 100 000 words, find the probability that it contains exactly 9 misprints. **[1]**

A proof-reader is paid £20 for each misprint identified in this book. You may assume that the proof-reader identifies every misprint in the book. The payment in pounds made to the proof-reader is denoted by the random variable P.

(c) By considering the mean and variance of P, show that P does **not** follow a Poisson distribution. **[2]**

The proof-reader is paid for proof-reading another book containing exactly 50 000 words, again identifying every misprint in the book.

(d) Determine the probability that the proof-reader is paid a total amount which is more than £300 and less than £500 for proof-reading **both** books. **[3]**

7 The manager of a Health Service Trust receives data from two hospitals, Hospital A and Hospital B, about the lengths of stays in hospital for patients with a particular condition. The data are in the form of a list of the number of days that each patient with that condition has spent at each hospital in one particular year.

The manager decides to categorise the length of each stay according to **Table 7.1** before putting the data into a contingency table.

Table 7.1

Length of Stay (days)	Category
1–3	Short Stay
4–6	Medium Stay
$\geqslant 7$	Long Stay

(a) State **one** advantage and **one** disadvantage of categorising the data in this way. **[2]**

The contingency table created by the manager is shown in **Table 7.2**.

Table 7.2

	Short Stay	Medium Stay	Long Stay	Total
Hospital A	9	20	15	44
Hospital B	29	25	40	94
Total	38	45	55	138

The manager wishes to use a chi-squared test to assess whether there is any association between the length of stay of the patient at the hospital and which hospital is treating the patient. You may assume that the lengths of stay are independent of each other and the data is a random sample from the underlying population of patients with this condition.

(b) Write down suitable null and alternative hypotheses for the test. **[2]**

Table 7.3 shows some of the expected frequencies. The table is incomplete.

Table 7.3

Expected frequency	Short Stay	Medium Stay	Long Stay
Hospital A	12.116		
Hospital B			37.464

(c) Complete the copy of **Table 7.3** in the **Printed Answer Booklet**. Your entries should be given to **3** decimal places. **[2]**

Table 7.4 shows some of the chi-squared contributions. Entries are given to **4** decimal places. The table is incomplete.

Table 7.4

Chi-squared contribution	Short Stay	Medium Stay	Long Stay
Hospital A			
Hospital B	0.3751		0.1717

(d) Complete the copy of **Table 7.4** in the **Printed Answer Booklet** and carry out the test at the 10% significance level. You do not need to restate your hypotheses. **[6]**

(e) Use the chi-squared contributions found in part **(d)** to explain which length stay, at which hospital, gives most evidence of any association. **[1]**

The manager receives a new instruction that all hypothesis tests carried out by the trust should actually have a significance level of 5%.

(f) Explain whether following this new instruction would have affected the result of the test. **[2]**

END OF QUESTION PAPER