OCR
Oxford Cambridge and RSA

Friday 7 June 2024 – Afternoon

A Level Further Mathematics A

Y542/01 Statistics

Time allowed: 1 hour 30 minutes

You must have:
- the Printed Answer Booklet
- the Formulae Booklet for A Level Further Mathematics A
- 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 non-exact numerical answers correct to **3** significant figures unless a different degree of accuracy is specified in the question.
- The acceleration due to gravity is denoted by $g\,\mathrm{m\,s^{-2}}$. When a numerical value is needed use $g = 9.8$ unless a different value is specified in the question.
- Do **not** send this Question Paper for marking. Keep it in the centre or recycle it.

INFORMATION
- The total mark for this paper is **75**.
- 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 discrete random variable X has the following distribution, where a, b and c are constants.

x	0	1	2	3
$\mathrm{P}(X = x)$	a	b	c	0.1

It is given that $\mathrm{E}(X) = 1.25$ and $\mathrm{Var}(X) = 0.8875$.

(a) Determine the values of a, b and c. **[5]**

(b) The random variable Y is defined by $Y = 7 - 2X$.

Write down the value of $\mathrm{Var}(Y)$. **[1]**

(c) Twenty independent observations of X are obtained. The number of those observations for which $X = 3$ is denoted by T.

Find the value of $\mathrm{Var}(T)$. **[2]**

2 A newspaper article claimed that "taller dog owners have taller dogs as pets". Alex investigated this claim and obtained data from a random sample of 16 fellow students who owned exactly one dog. The results are summarised as follows, where the height of the student, in cm, is denoted by h and the height, in cm, of their dog is denoted by d.

$n = 16 \quad \Sigma h = 2880 \quad \Sigma d = 660 \quad \Sigma h^2 = 519276 \quad \Sigma d^2 = 30000 \quad \Sigma hd = 119425$

(a) Calculate the value of Pearson's product moment correlation coefficient for the data. **[2]**

(b) State what your answer tells you about a scatter diagram illustrating the data. **[1]**

(c) Use the data to test, at the 5% significance level, the claim of the newspaper article. **[5]**

(d) Explain whether the answer to part **(a)** would be likely to be different if the dogs' weights had been used instead of their heights. **[1]**

3 Research suggests that the mean reading age of a child about to start secondary school is 10.75. The reading ages, X years, of a random sample of 80 children who were about to start secondary school in a particular district were measured, and the results are summarised as follows.

$n = 80 \qquad \Sigma x = 893 \qquad \Sigma x^2 = 10267$

(a) Test at the 5% significance level whether the mean reading age of children about to start secondary school in this district is **not** 10.75. **[10]**

(b) A student wrote: "Although we do not know that the distribution of X is normal, the central limit theorem allows us to assume that it is, as the sample size is large." This statement is incorrect. Give a corrected version of the student's statement. **[1]**

4 **(a)** Write down the number of ways of choosing 5 objects from 12 distinct objects. **[1]**

(b) Each possible set of 5 different integers selected from the integers 1, 2, …, 12 is obtained, and for each set, the sum of the 5 integers is found. The sum S can take values between 15 and 50 inclusive. Part of the frequency distribution of S is shown in the following table, together with the cumulative frequencies.

S	15	16	17	18	19	20	21	22	23
Frequency	1	1	2	3	5	7	10	13	17
Cumulative Frequency	1	2	4	7	12	19	29	42	59

Use these numbers to determine the critical region for a 1-tail Wilcoxon rank-sum test at the 2% significance level when $m = 5$ and $n = 7$. **[2]**

(c) A student says that, for a Wilcoxon rank-sum test on samples of size m and n, where m and n are large, the mean and variance of the test statistic R_m are 200 and $616\frac{2}{3}$ respectively.

Show that at least one of these values must be incorrect. **[3]**

5 Some bird-watchers study the song of chaffinches in a particular wood. They investigate whether the number, N, of separate bursts of song in a 5 minute period can be modelled by a Poisson distribution. They assume that a burst of song can be considered as a single event, and that bursts of song occur randomly.

(a) State **two** further assumptions needed for N to be well modelled by a Poisson distribution. **[2]**

The bird-watchers record the value of N in each of 60 periods of 5 minutes. The mean and variance of the results are 3.55 and 5.6475 respectively.

(b) Explain what this suggests about the validity of a Poisson distribution as a model in this context. **[2]**

The complete results are shown in the table.

n	0	1	2	3	4	5	6	7	8	$\geqslant 9$
Frequency	10	3	7	8	13	6	6	2	5	0

The bird-watchers carry out a χ^2 goodness of fit test at the 5% significance level.

(c) State suitable hypotheses for the test. **[1]**

(d) Determine the contribution to the test statistic for $n = 3$. **[3]**

(e) The total value of the test statistic, obtained by combining the cells for $n \leqslant 1$ and also for $n \geqslant 6$, is 9.202, correct to 4 significant figures.

Complete the goodness of fit test. **[3]**

(f) It is known that chaffinches are more likely to sing in the presence of other chaffinches.

Explain whether this fact affects the validity of a Poisson model for N. **[1]**

6 A bag contains 6 identical blue counters and 5 identical yellow counters.

(a) Three counters are selected at random, without replacement.

Find the probability that at least two of the counters are blue. **[2]**

All 11 counters are now arranged in a row in a random order.

(b) Find the probability that all the yellow counters are next to each other. **[2]**

(c) Find the probability that no yellow counter is next to another yellow counter. **[3]**

(d) Find the probability that the counters are arranged in such a way that **both** of the following conditions hold.

- Exactly three of the yellow counters are next to one another.
- Neither of the other two yellow counters is next to a yellow counter. **[3]**

(e) Explain whether the answer to part **(d)** would be different if the yellow counters were numbered 1, 2, 3, 4 and 5, so that they are not identical. **[1]**

7 The coordinates of a set of 10 points are denoted by (x_i, y_i) for $i = 1, 2, \ldots, 10$.
For a particular set of values of (x_i, y_i) and any constants a and b it can be shown that

$$\Sigma(y_i - a - bx_i)^2 = 10(11 - a - 6b)^2 + 126\left(b - \frac{83}{42}\right)^2 + \frac{139}{14}.$$

(a) **(i)** Explain why $\Sigma(y_i - a - bx_i)^2$ is minimised by taking $b = \frac{83}{42}$ and $a = 11 - 6b$. **[1]**

(ii) Hence explain why the equation of the regression line of y on x for these points is given by the corresponding values of a and b (so that the equation is $y = \frac{83}{42}x - \frac{6}{7}$). **[1]**

(b) State which of the following terms **cannot** apply to the variable X if the regression line of y on x can be used for estimating values of Y.

Dependent Independent Controlled Response **[1]**

(c) Use the regression line to estimate the value of y corresponding to $x = 8$. **[1]**

(d) State what must be true of the value $x = 8$ if the estimate in part **(c)** is to be reliable. **[1]**

(e) Variables u and v are related to x and y by the following relationships.

$u = 2 + 4x \quad v = 8 - 2y$

Show that the gradient of the regression line of v on u is very close to -1. **[3]**

8 A random sample of 100 students were given a task and the time taken by each student to complete the task was recorded. The maximum time allowed to complete the task was one minute and all students completed the task within the maximum time. The times, T minutes, for the random sample of students are summarised as follows.

$n = 100 \qquad \sum t = 61.88$

A researcher proposes that T can be modelled by the continuous random variable with probability density function

$$f(t) = \begin{cases} \alpha t^{\alpha - 1} & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

where α is a positive constant.

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

By finding $E(T)$ according to the researcher's model, determine an approximation for the value of α. Give your answer correct to **3** significant figures. **[6]**

Further information about the times taken for the sample of 100 students to complete the task is given in the table.

Time t	$0 \leqslant t < \frac{1}{3}$	$\frac{1}{3} \leqslant t < \frac{2}{3}$	$\frac{2}{3} \leqslant t \leqslant 1$
Frequency	18	37	45

(b) Using the value of α found in part **(a)**, determine the extent to which the proposed model is a good model. (Do not carry out a goodness of fit test.) **[4]**

END OF QUESTION PAPER

BLANK PAGE