OCR
Oxford Cambridge and RSA

Thursday 6 June 2019 – Afternoon

A Level Further Mathematics B (MEI)

Y422/01 Statistics Major

Time allowed: 2 hours 15 minutes

7768896723

You must have:
- Printed Answer Booklet
- Formulae Further Mathematics B (MEI)

You may use:
- a scientific or graphical calculator

INSTRUCTIONS
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Answer **all** the questions.
- **Write your answer to each question in the space provided in the Printed Answer Booklet.** If additional space is required, you should use the lined page(s) at the end of the Printed Answer Booklet. The question number(s) must be clearly shown.
- You are permitted to use a scientific or graphical calculator in this paper.
- Final answers should be given to a degree of accuracy appropriate to the context.

INFORMATION
- The total mark for this paper is **120**.
- The marks for each question are shown in brackets **[ ]**.
- You are advised that an answer may receive **no marks** unless you show sufficient detail of the working to indicate that a correct method is used. You should communicate your method with correct reasoning.
- The Printed Answer Booklet consists of **16** pages. The Question Paper consists of **12** pages.

Section A (29 marks)

Answer **all** the questions.

1 A fair six-sided dice is rolled three times.
The random variable X represents the lowest of the three scores.
The probability distribution of X is given by the formula

$$P(X = r) = k(127 - 39r + 3r^2) \text{ for } r = 1, 2, 3, 4, 5, 6.$$

(a) Complete the copy of the table in the Printed Answer Booklet. **[1]**

r	1	2	3	4	5	6
$P(X = r)$	$91k$	$61k$	$37k$			

(b) Show that $k = \frac{1}{216}$. **[2]**

(c) Draw a graph to illustrate the distribution. **[2]**

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

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

Find each of the following.

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

2 A special railway coach detects faults in the railway track before they become dangerous.

(a) Write down the conditions required for the numbers of faults in the track to be modelled by a Poisson distribution. **[2]**

You should now assume that these conditions do apply, and that the mean number of faults in a 5 km length of track is 1.6.

(b) Find the probability that there are at least 2 faults in a randomly chosen 5 km length of track. **[2]**

(c) Find the probability that there are at most 10 faults in a randomly chosen 25 km length of track. **[2]**

(d) On a particular day the coach is used to check 10 randomly chosen 1 km lengths of track. Find the probability that exactly 1 fault, in total, is found. **[3]**

3 The weights of bananas sold by a supermarket are modelled by a Normal distribution with mean 205 g and standard deviation 11 g.

(a) Find the probability that the total weight of 5 randomly selected bananas is at least 1 kg. **[2]**

When a banana is peeled the change in its weight is modelled as being a reduction of 35%.

(b) Find the probability that the weight of a randomly selected peeled banana is at most 150 g. **[3]**

Andy makes smoothies. Each smoothie is made using 2 peeled bananas and 20 strawberries from the supermarket, all the items being randomly chosen. The weight of a strawberry is modelled by a Normal distribution with mean 22.5 g and standard deviation 2.7 g.

(c) Find the probability that the total weight of a smoothie is less than 700 g. **[4]**

Section B (91 marks)

Answer **all** the questions.

4 Shellfish in the sea near nuclear power stations are regularly monitored for levels of radioactivity. On a particular occasion, the levels of caesium-137 (a radioactive isotope) in a random sample of 8 cockles, measured in becquerels per kilogram, were as follows.

2.36 2.97 2.69 3.00 2.51 2.45 2.21 2.63

Software is used to produce a 95% confidence interval for the level of caesium-137 in the cockles. The output from the software is shown in Fig. 4. The value for 'SE' has been deliberately omitted.

T Estimate of a Mean ▼

Confidence Level 0.95

Sample

Mean 2.6025

s 0.2793

N 8

Result

T Estimate of a Mean

Mean	2.6025
s	0.2793
SE	
N	8
df	7
Interval	2.6025 ± 0.2335

Fig. 4

(a) State an assumption necessary for the use of the t distribution in the construction of this confidence interval. **[1]**

(b) State the confidence interval which the software gives in the form $a < \mu < b$. **[1]**

(c) In the software output shown in Fig. 4, SE stands for standard error.
Find the standard error in this case. **[2]**

(d) Show how the value of 0.2335 in the confidence interval was calculated. **[2]**

(e) State how, using this sample, a wider confidence interval could be produced. **[1]**

5 In an investigation into the possible relationship between smoking and weight in adults in a particular country, a researcher selected a random sample of 500 adults.
The adults in the sample were classified according to smoking status (non-smoker, light smoker or heavy smoker, where light smoker indicates less than 10 cigarettes per day) and body weight (underweight, normal weight or overweight).

Fig. 5 is a screenshot showing part of the spreadsheet used to calculate the contributions for a chi-squared test. Some values in the spreadsheet have been deliberately omitted.

	A	B	C	D	E	F
1	**Observed frequencies**					
2		**Underweight**	**Normal**	**Overweight**	**Totals**	
3	**Non-smoker**	8	52	178	238	
4	**Light smoker**	10	40	68	118	
5	**Heavy smoker**	5	47	92	144	
6	**Totals**	23	139	338	500	
7						
8	**Expected frequencies**					
9	**Non-smoker**	10.9480	66.1640	160.8880		
10	**Light smoker**	5.4280		79.7680		
11	**Heavy smoker**		40.0320	97.3440		
12						
13	**Contributions to the test statistic**					
14	**Non-smoker**	0.7938		1.8200		
15	**Light smoker**	3.8510	1.5785	1.7361		
16	**Heavy smoker**	0.3982	1.2129	0.2934		
17						

Fig. 5

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

- B11
- C10
- C14

[4]

(b) Complete the hypothesis test at the 1% level of significance. **[6]**

(c) For each smoking status, give a brief interpretation of the largest of the three contributions to the test statistic. **[3]**

6 (a) A researcher is investigating the date of the 'start of spring' at different locations around the country.
A suitable date (measured in days from the start of the year) can be identified by checking, for example, when buds first appear for certain species of trees and plants, but this is time-consuming and expensive. Satellite data, measuring microwave emissions, can alternatively be used to estimate the date that land-based measurements would give.

The researcher chooses a random sample of 12 locations, and obtains land-based measurements for the start of spring date at each location, together with relevant satellite measurements. The scatter diagram in Fig. 6.1 shows the results; the land-based measurements are denoted by x days and the corresponding values derived from satellite measurements by y days.

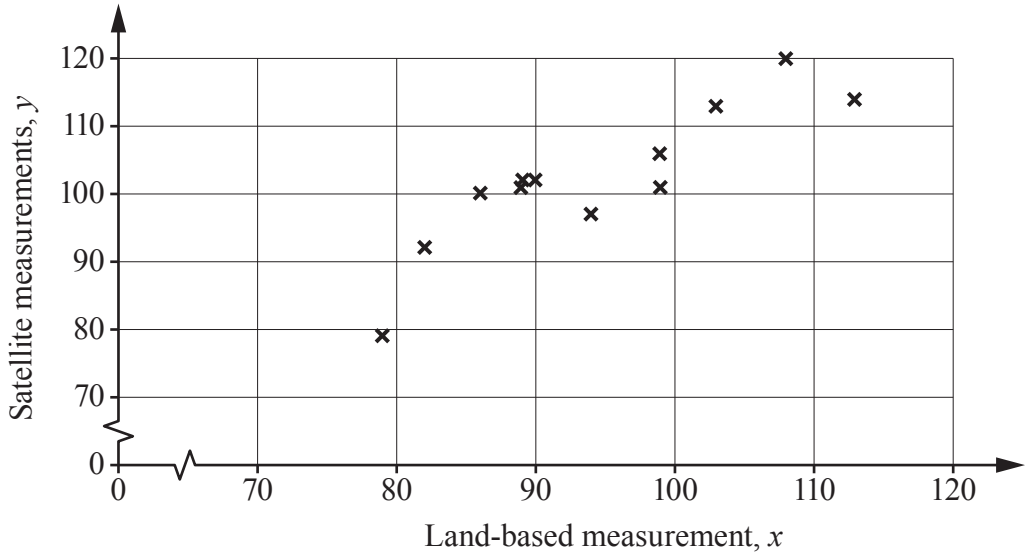


Fig. 6.1

Fig. 6.2 shows part of a spreadsheet used to analyse the data. Some rows of the spreadsheet have been deliberately omitted.

	A	B	C	D	E	F
1		***x***	***y***	***x*²**	***y*²**	***xy***
2		90	102	8100	10404	9180
3						
10						
11						
12		94	97	8836	9409	9118
13		99	101	9801	10201	9999
14	**Sum**	1131	1227	107783	126725	116724
15						

Fig. 6.2

(i) Calculate the equation of a regression line suitable for estimating the land-based date of the start of spring from satellite measurements. **[5]**

(ii) Using this equation, estimate the land-based date of the start of spring for the following dates from satellite measurements.

- 95 days
- 60 days **[2]**

(iii) Comment on the reliability of each of your estimates. **[2]**

(b) The researcher is also investigating whether there is any correlation between the average temperature during a month in spring and the total rainfall during that month at a particular location. The average temperatures in degrees Celsius and total rainfall in mm for a random selection, over several years, of 10 spring months at this location are as follows.

Temperature	4.2	7.1	5.6	3.5	8.6	6.5	2.7	5.9	6.7	4.1
Rainfall	18	26	42	76	15	43	84	53	66	36

The researcher plots the scatter diagram shown in Fig. 6.3 to check which type of test to carry out.

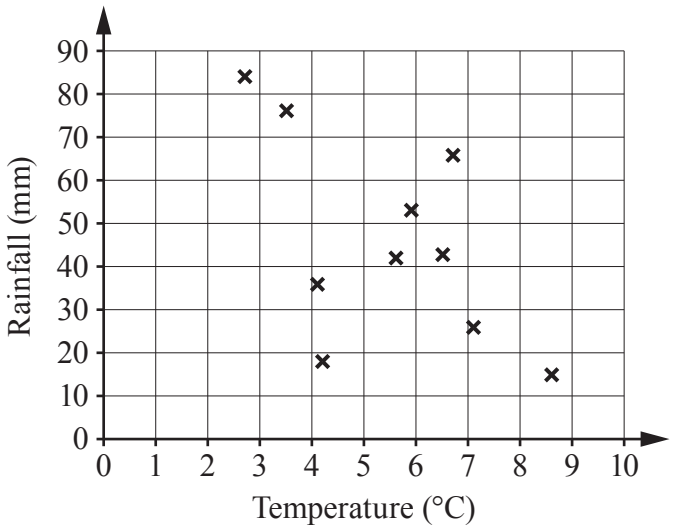


Fig. 6.3

(i) Explain why the researcher might come to the conclusion that a test based on Pearson's product moment correlation coefficient may be valid. **[2]**

(ii) Find the value of Pearson's product moment correlation coefficient. **[2]**

(iii) Carry out a test at the 5% significance level to investigate whether there is any correlation between temperature and rainfall. **[5]**

7 A swimming coach believes that times recorded by people using stopwatches are on average 0.2 seconds faster than those recorded by an electronic timing system.
In order to test this, the coach takes a random sample of 40 competitors' times recorded by both methods, and finds the differences between the times recorded by the two methods. The mean difference in the times (electronic time minus stopwatch time) is 0.1442 s and the standard deviation of the differences is 0.2580 s.

(a) Find a 95% confidence interval for the mean difference between electronic and stopwatch times. **[4]**

(b) Explain whether there is evidence to suggest that the coach's belief is correct. **[2]**

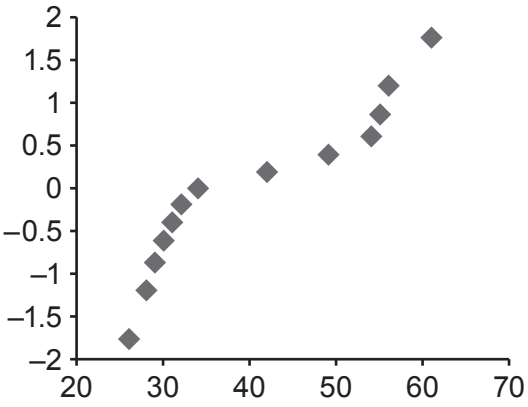
(c) Explain how you can calculate the confidence interval in part **(a)** even though you do not know the distribution of the parent population of differences. **[2]**

(d) If the coach wanted to produce a 95% confidence interval of width no more than 0.12 s, what is the minimum sample size that would be needed, assuming that the standard deviation remains the same? **[3]**

8 A student doing a school project wants to test a claim which she read in a newspaper that drinking a cup of tea will improve a person's arithmetic skills.
She chooses 13 students from her school and gets each of them to drink a cup of tea. She then gives each of them an arithmetic test. She knows that the average score for this test in students of the same age group as those she has chosen is 33.5.
The scores of the students she tests, arranged in ascending order, are as follows.

26 28 29 30 31 32 34 42 49 54 55 56 61

The student decides to use software to draw a Normal probability plot for these data, and to carry out a Normality test as shown in Fig. 8.



Kolmogorov-Smirnov Test for Normality

p-value = 0.048

Fig. 8

(a) The student uses the output from the software to help in deciding on a suitable hypothesis test to use for investigating the claim about drinking tea.
Explain what the student should conclude. **[3]**

(b) The student's teacher agrees with the student's choice of hypothesis test, but says that even this test may not be valid as there may be some unsatisfactory features in the student's project.
Give three features that the teacher might identify as unsatisfactory. **[3]**

(c) Assuming that the student's procedures can be justified, carry out an appropriate test at the 5% significance level to investigate the claim about drinking tea. **[7]**

9 Every weekday Jonathan takes an underground train to work. On any weekday the time in minutes that he has to wait at the station for a train is modelled by the continuous uniform distribution over [0, 5].

(a) Find the probability that Jonathan has to wait at least 3 minutes for a train. **[1]**

The total time that Jonathan has to wait on two days is modelled by the continuous random variable X with probability density function given by

$$f(x) = \begin{cases} \frac{1}{25}x & 0 \leqslant x \leqslant 5, \\ \frac{1}{25}(10-x) & 5 < x \leqslant 10, \\ 0 & \text{otherwise.} \end{cases}$$

(b) Find the probability that Jonathan has to wait a total of at most 6 minutes on two days. **[3]**

Jonathan's friend suggests that the total waiting time for 5 days, T minutes, will almost certainly be less than 18 minutes. In order to investigate this suggestion, Jonathan constructs the simulation shown in Fig. 9. All of the numbers in the simulation have been rounded to 2 decimal places.

	A	B	C	D	E	F
1	Mon	Tue	Wed	Thu	Fri	Total T
2	1.78	4.36	2.74	3.88	4.64	17.41
3	0.95	1.30	4.83	4.29	1.81	13.18
4	4.27	4.90	4.57	1.41	3.66	18.81
5	0.80	0.06	3.20	1.76	0.35	6.17
6	0.03	4.82	1.26	3.53	0.13	9.77
7	3.88	4.73	1.19	3.75	1.29	14.84
8	4.11	3.54	4.33	0.77	4.50	17.25
9	3.54	0.11	3.85	2.86	1.58	11.94
10	1.87	1.82	3.00	3.53	1.83	12.05
11	4.00	2.98	4.59	1.73	1.76	15.06
12	1.91	3.85	2.08	1.72	2.82	12.38
13	0.10	4.86	2.51	0.52	2.17	10.15
14	1.24	4.26	0.95	1.33	1.78	9.57
15	2.99	0.69	3.85	3.41	2.42	13.36
16	4.67	1.76	2.13	3.48	3.10	15.14
17	1.94	1.07	0.91	0.63	3.34	7.89
18	0.11	2.29	0.71	4.21	0.86	8.18
19	0.43	4.58	4.89	1.86	2.84	14.60
20	4.23	0.88	2.71	4.88	4.20	16.91
21	3.72	4.58	3.11	4.89	3.18	19.49

Fig. 9

(c) Use the simulation to estimate $P(T > 18)$. **[1]**

(d) Explain how Jonathan could obtain a better estimate. **[1]**

Jonathan thinks that he can use the Central Limit Theorem to provide a very good approximation to the distribution of T.

(e) Find each of the following.

- $\mathrm{E}(T)$
- $\mathrm{Var}(T)$ **[3]**

(f) Use the Central Limit Theorem to estimate $\mathrm{P}(T > 18)$. **[2]**

(g) Comment briefly on the use of the Central Limit Theorem in this case. **[1]**

Jonathan travels to work on 200 days in a year.

(h) Find the probability that the total waiting time for Jonathan in a year is more than 510 minutes. **[3]**

10 The probability density function of the continuous random variable X is given by

$$f(x) = \begin{cases} kx^m & 0 \leqslant x \leqslant a, \\ 0 & \text{otherwise,} \end{cases}$$

where a, k and m are positive constants.

(a) Show that $k = \dfrac{m+1}{a^{m+1}}$. **[3]**

(b) Find the cumulative distribution function of X in terms of x, a and m. **[4]**

(c) Given that $\mathrm{P}\left(\frac{1}{4}a < X < \frac{1}{2}a\right) = \frac{1}{10}$,

(i) show that $2p^2 - 10p + 5 = 0$, where $p = 2^m$, **[4]**

(ii) find the value of m. **[3]**

END OF QUESTION PAPER