Oxford Cambridge and RSA

A Level Further Mathematics B (MEI)

Y422/01 Statistics Major

Practice Paper – Set 1

Time allowed: 2 hours 15 minutes

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.
- Complete the boxes provided on the Printed Answer Booklet with your name, centre number and candidate number.
- 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.
- Do **not** write in the barcodes.
- 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 number of marks 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 (34 marks)

Answer **all** the questions.

1 In a field, there is a large number of dandelion seedlings.

(i) Write down the conditions required for the number of dandelion seedlings in a randomly chosen plot of area $1\,\text{m}^2$ to be modelled by a Poisson distribution. **[2]**

You should now assume that these conditions do apply, and that the mean number of dandelion seedlings in a randomly chosen plot of area $1\,\text{m}^2$ is 1.3.

(ii) Find the probability that there are at least 3 dandelion seedlings in a randomly chosen plot of area $1\,\text{m}^2$. **[2]**

(iii) Find the probability that there are at most 10 dandelion seedlings in a randomly chosen plot of area $10\,\text{m}^2$. **[2]**

In the same field there is also a large number of yarrow seedlings. The mean number of yarrow seedlings in a randomly chosen plot of area $1\,\text{m}^2$ is 0.6 and the number of yarrow seedlings is also modelled by a Poisson distribution. You should assume that the two types of weed, dandelion and yarrow, occur independently.

(iv) Find the probability that there are at least 20 seedlings in total of the two types of weed in a randomly chosen plot of area $10\,\text{m}^2$. **[2]**

2 The score, X, obtained on a spin of a biased five-sided spinner is given by

$$\mathrm{P}(X = r) = k(r + 3)(9 - r) \qquad \text{for } r = 1, 2, 3, 4, 5, \text{ where } k \text{ is a constant.}$$

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

r	1	2	3	4	5
$\mathrm{P}(X = r)$	$32k$				

(ii) Show that $k = \frac{1}{170}$. **[2]**

(iii) Find each of the following.

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

3 A ferry operator is investigating the punctuality of ferries. The random variable T represents the difference in hours between the actual arrival time and the scheduled arrival time of a ferry. A positive value of T means that the ferry is late. The probability density function of T is given by

$$f(t) = \begin{cases} 1 - \frac{16}{9}\left(t - \frac{1}{4}\right)^2 & -\frac{1}{2} \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

The graph of this probability density function is shown in Fig. 3.

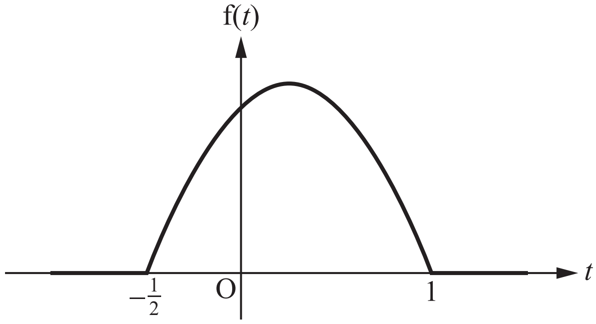

Fig. 3

(i) Show that the cumulative distribution function of T is given by

$$F(t) = \begin{cases} 0 & t < -\frac{1}{2}, \\ t + \frac{1}{4} - \frac{16}{27}\left(t - \frac{1}{4}\right)^3 & -\frac{1}{2} \leqslant t \leqslant 1, \\ 1 & t > 1. \end{cases}$$

[3]

(ii) Find $P\left(T > \frac{1}{2}\right)$. **[2]**

(iii) Find the probability that the ferry arrives early. **[2]**

(iv) Verify that the 90th percentile of T is 0.7063, approximately. **[2]**

(v) Find

- the mean of T,
- the standard deviation of T.

[4]

4 The time in hours taken for the train journey from town A to town B is modelled by a Normal distribution with mean 2.95 and standard deviation 0.12. The time, in hours, taken for the train journey from B to A is modelled by an independent Normal distribution with mean 3.03 and standard deviation 0.17. All journey times are assumed to be independent of one another.

(i) Find the probability that the total time for a return journey from A to B and back is less than 6 hours. **[2]**

(ii) Find the probability that the time for a journey from A to B is greater than the time for a journey from B to A. **[3]**

(iii) Find the probability that the total time for ten journeys from A to B exceeds the total time for ten journeys from B to A by at least half an hour. **[3]**

Section B (86 marks)

Answer **all** the questions.

5 Water accumulates underground in layers of water-bearing rocks called aquifers. The distance below ground level at which the aquifer starts is called the depth of the water table. The spreadsheet in Fig. 5 shows the depth of the water table and the thickness of the aquifer below it (both measured in metres) for a number of sites in a region. It also has a scatter diagram to illustrate the data. You should assume that these sites form a random sample of sites in the region.

	A	B
1	Water table depth	Aquifer thickness
2	25.4	15.4
3	29.3	5.5
4	36.0	41.1
5	39.8	14.8
6	41.6	9.4
7	44.6	33.5
8	51.3	11.6
9	54.5	6.7
10	61.5	18.8
11	64.8	11.0
12	73.1	22.2
13	94.5	3.4
14		
15		
16		

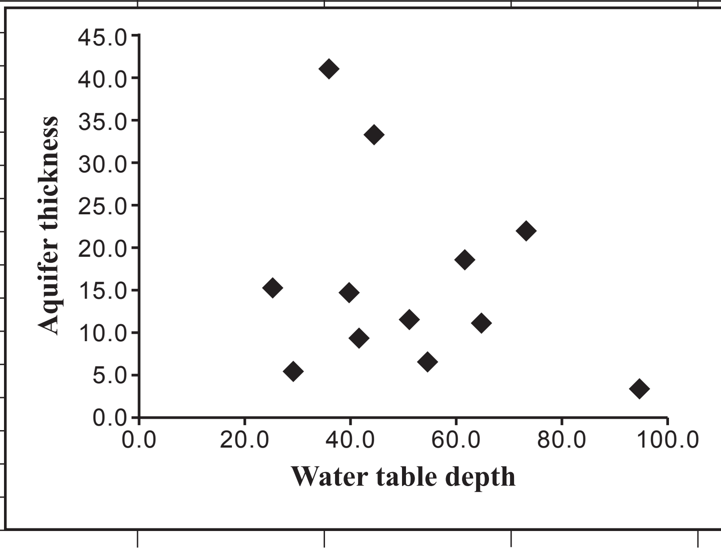

Fig. 5

A researcher is investigating whether there is any relationship between water table depth and aquifer thickness.

(i) Explain why, in the light of the scatter diagram, it might not be valid to carry out a test based on the product moment correlation coefficient. **[2]**

(ii) Calculate the value of Spearman's rank correlation coefficient for these data. **[3]**

(iii) Carry out a hypothesis test at the 5% significance level to examine whether there is any association between the depth of the water table and the thickness of the aquifer in this region. **[5]**

(iv) Explain why, if both tests are valid, it is preferable to use a test based on the product moment correlation coefficient rather than on Spearman's rank correlation coefficient. **[2]**

6 A particular brand of marine distress flare should have a mean burn time of at least 63 seconds. The burn times, t seconds, of a random sample of 32 flares from a batch are measured, giving the following results.

$$\sum t = 2112 \qquad \sum t^2 = 141\,969$$

(i) Calculate estimates of

- the population mean,
- the population variance. **[3]**

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

Find a 95% confidence interval for the mean burn time of this batch of flares. **[4]**

(iii) Explain whether the confidence interval which you have calculated suggests that the mean burn time is at least 63 seconds. **[2]**

A random sample of 40 flares is selected from another batch. A 99% confidence interval for the mean burn time for this batch is [62.7, 63.2].

(iv) Find

- the sample mean,
- the sample variance. **[3]**

7 At a clinic at a hospital, there are 11 appointments each morning. However, people often fail to arrive for their appointments. Such people are described as 'no-shows'. For a random sample of 120 days the numbers of no-shows are as follows.

Number of no-shows, x	0	1	2	3	4	5	6	>6
Frequency	36	34	26	10	4	6	4	0

A manager wonders whether a binomial distribution would be an appropriate model for the number of no-shows. You are given that the mean number of no-shows out of the 11 appointments each morning, calculated from the sample, is 1.55.

(i) Estimate the value of the parameter p for this binomial model. **[2]**

(ii) Using this value of p, calculate the expected number of mornings in the 120 days on which there would be at least 5 no-shows if a binomial model is appropriate. **[2]**

The screenshot in Fig. 7 shows part of a spreadsheet for a test to assess the goodness of fit of a binomial model, using the value of p calculated in part **(i)**. Some values in the spreadsheet have been deliberately omitted.

	A	B	C	D	E
1	Number of no-shows, x	Observed frequency	Binomial probability	Expected frequency	Chi-squared contribution
2	0	36	0.1881	22.5742	7.9849
3	1	34	0.3394	40.7290	1.1117
4	2	26			1.6404
5	3	10	0.1370	16.4360	
6	≥ 4	14	0.0572	6.8587	7.4355
7					

Fig. 7

(iii) State the null and alternative hypotheses for the test. **[1]**

(iv) Calculate the missing values for each of the following cells.

- C4
- E5 **[3]**

(v) Complete the test at the 1% significance level. **[5]**

(vi) Briefly explain why some of the x-values were combined in order to carry out the test. **[1]**

8 A researcher is investigating a claim that wearing an activity monitor that counts the steps that a person takes in a day increases the amount of walking that people do. It is known that when not wearing an activity monitor, the average number of steps that a person living in a particular area takes in a day is 5500. The researcher chooses a random sample of 12 people from this area who are each fitted with an activity monitor which they wear for a few weeks. The numbers of steps taken on a randomly chosen day towards the end of these few weeks by these 12 people are as follows.

3258	5909	6912	6410	6895	7979	6153	6204	4710	7743	6047	1105

The researcher wishes to carry out a test to investigate the claim. She first uses software to draw the Normal probability plot for these data, as shown in Fig. 8.

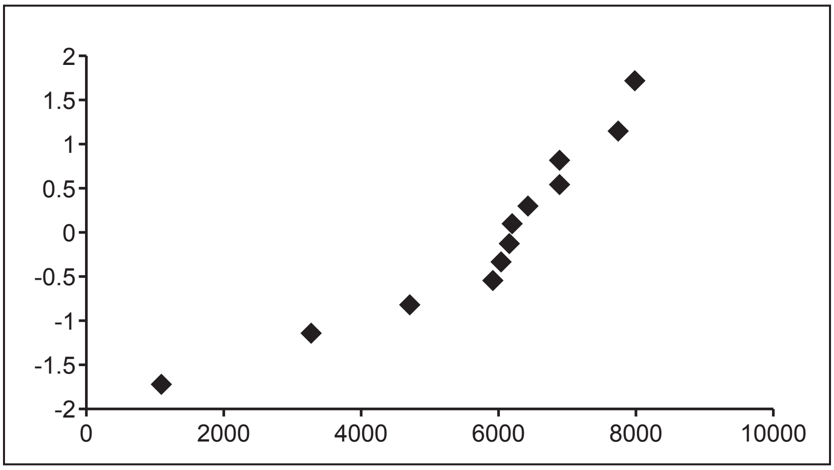

Fig. 8

The researcher concludes from the plot that a test based on the *t*-distribution might not be valid.

(i) Explain why she comes to this conclusion. **[2]**

(ii) Explain why, in order to carry out a test, the sample must be random. **[1]**

(iii) Carry out a suitable test at the 5% significance level to investigate the claim. **[8]**

(iv) Under what circumstances, with a sample of size 12, would it have been valid to carry out a test based on the Normal distribution? **[2]**

9 The legal minimum size of bass (a species of fish) which can be caught is 42 cm. In one area, most of the bass were below this legal minimum and so fishing for bass was stopped entirely to allow the stocks to recover.

Two years later, the lengths in cm of a random sample of 10 bass caught in this area were as follows.

29.7	43.8	45.9	45.8	44.1	45.3	45.0	36.5	36.9	43.8

(i) What assumption do you need to make in order to construct a confidence interval based on the *t* distribution for the mean length of bass in this area? **[1]**

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

Given that the assumption in part **(i)** is valid, find a 95% confidence interval for the mean length of bass in this area. **[7]**

(iii) Explain what is meant by a 95% confidence interval. **[2]**

10 **(a)** Prove that the variance of a uniform distribution over $\{1, 2, ..., n\}$ is $\frac{1}{12}(n^2 - 1)$. **[6]**

A fair eight-sided dice has faces labelled 1, 2, 3, 4, 5, 6, 7, 8.

(b) **(i)** Calculate an estimate of the probability that, when the dice is rolled 50 times, the mean score is less than 4. **[5]**

(ii) The distribution of the mean score when the dice is rolled 50 times is not known. Explain why the answer to part **(i)** can be found despite this. **[2]**

11 The random variable X has the binomial distribution B(10, 0.1).

(i) State each of the following.

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

Fatima is investigating whether the sample mean of n values of X can be well approximated by a Normal distribution. She designs a simulation, shown in the spreadsheet in Fig. 11, for the case $n = 5$. Each of the 20 rows below the heading row consists of 5 values of X together with the value of the sample mean Y.

	A	B	C	D	E	F
1	X_1	X_2	X_3	X_4	X_5	Y
2	2	0	2	2	1	1.40
3	0	2	2	0	1	1.00
4	1	3	2	2	0	1.60
5	1	2	2	2	1	1.60
6	2	0	3	1	1	1.40
7	0	4	0	1	1	1.20
8	0	1	1	1	0	0.60
9	1	0	3	1	2	1.40
10	1	2	2	2	0	1.40
11	0	1	0	3	2	1.20
12	1	1	1	0	1	0.80
13	3	2	0	1	0	1.20
14	1	0	0	1	0	0.40
15	0	1	2	0	0	0.60
16	2	1	2	0	1	1.20
17	1	2	1	1	2	1.40
18	1	0	0	2	4	1.40
19	0	1	1	1	1	0.80
20	1	0	0	1	0	0.40
21	3	2	1	2	1	1.80

Fig. 11

(ii) Use the spreadsheet output to estimate each of the following.

- P($Y < 0.6$)
- P($Y > 1.4$) **[2]**

Fatima finds the mean and variance of Y and then considers the Normal random variable Z with this mean and variance.

Fatima wishes to compare $P(Y > 1.4)$ with an appropriate probability for Z.

(iii) Explain why she should calculate $P(Z > 1.5)$, rather than $P(Z > 1.4)$. **[2]**

(iv) Find $P(Z > 1.5)$. **[2]**

Fatima now constructs a new spreadsheet which calculates the sample mean W of 25 values of X rather than the 5 she used in her first spreadsheet. She again uses 20 rows in her simulation. Of the 20 values of W obtained, she finds that 4 are greater than 1.1.

(v) Write down an estimate of $P(W > 1.1)$. **[1]**

(vi) Use an appropriate Normal distribution to calculate an approximation for $P(W > 1.1)$. **[3]**

END OF QUESTION PAPER

BLANK PAGE

BLANK PAGE