OCR
Oxford Cambridge and RSA

# Thursday 08 October 2020 – Afternoon

## A Level Further Mathematics B (MEI)

### Y422/01 Statistics Major

**Time allowed: 2 hours 15 minutes**

*8179021430*

**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 **120**.
- The marks for each question are shown in brackets **[ ]**.
- This document has **12** pages.

**ADVICE**
- Read each question carefully before you start your answer.

© OCR 2020 [H/508/5594]
DC (ST/CT) 194739/4
OCR is an exempt Charity

**Section A** (32 marks)

Answer **all** the questions.

**1** In a game at a fair, players choose 4 countries from a list of 10 countries. The names of all 10 countries are then put in a box and the player selects 4 of them at random. The random variable $X$ represents the number of countries that match those which the player originally chose.

**(a)** Show that the probability that a randomly selected player matches all 4 countries is $\frac{1}{210}$. **[2]**

Table 1 shows the probability distribution of $X$.

| $r$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $P(X = r)$ | $\frac{1}{14}$ | $\frac{8}{21}$ | $\frac{3}{7}$ | $\frac{4}{35}$ | $\frac{1}{210}$ |

**Table 1**

**(b)** Find each of the following.

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

**(c)** A player has to pay £1 to play the game. The player gets 40 pence back for every country which is matched.

Find the mean and standard deviation of the player's loss per game. **[3]**

**(d)** In order to try to attract more customers, the rules will be changed as follows.

The game will still cost £1 to play. The player will get 25 pence back for every country which is matched, plus an additional bonus of £100 if all four countries are matched.

Find the player's mean gain or loss per game with these new rules. **[2]**

© OCR 2020

**2** On average 1 in 4000 people have a particular antigen in their blood (an antigen is a molecule which may cause an adverse reaction).

**(a) (i)** A random sample of 1200 people is selected. The random variable $X$ represents the number of people in the sample who have this antigen in their blood. Explain why you could use either a binomial distribution or a Poisson distribution to model the distribution of $X$. **[3]**

**(ii)** Use either a binomial or a Poisson distribution to calculate each of the following probabilities.

- $P(X = 3)$
- $P(X > 3)$ **[3]**

**(b)** A researcher needs to find 2 people with the antigen. Find the probability that at most 5000 people have to be tested in order to achieve this. **[3]**

**3** A supermarket sells cashew nuts in three different sizes of bag: small, medium and large. The weights in grams of the nuts in each type of bag are modelled by independent Normal distributions as shown in Table 3.

| Bag size | Mean | Standard deviation |
|---|---|---|
| Small | 51.5 | 1.1 |
| Medium | 100.7 | 1.6 |
| Large | 201.3 | 1.7 |

**Table 3**

**(a)** Find the probability that the mean weight of two randomly selected large bags is at least 200 g. **[3]**

**(b)** Find the probability that the total weight of eight randomly selected small bags is greater than the total weight of two randomly selected medium bags and one randomly selected large bag. **[5]**

**4** An amateur meteorologist records the total rainfall at her home each day using a traditional rain gauge. This means that she has to go out each day at 9 am to read the rain gauge and then to empty it. She wants to save time by using a digital rain gauge, but she also wants to ensure that the readings from the digital gauge are similar to those of her traditional gauge. Over a period of 100 days, she uses both gauges to measure the rainfall.

The meteorologist uses software to produce a 95% confidence interval for the difference between the two readings (the traditional gauge reading minus the digital gauge reading). The output from the software is shown in Fig. 4. Although rainfall was measured over a period of 100 days, there was no rain on 40 of those days and so the sample size in the software output is 60 rather than 100.

Z Estimate of a Mean ▾

Confidence Level 0.95

Sample

Mean 0.1173

σ 0.5766

N 60

Result

Z Estimate of a Mean

| | |
|---|---|
| Mean | 0.1173 |
| σ | 0.5766 |
| SE | 0.07444 |
| N | 60 |
| Lower Limit | −0.0286 |
| Upper Limit | 0.2632 |
| Interval | 0.1173 ± 0.1459 |

**Fig. 4**

**(a)** Explain why this confidence interval can be calculated even though nothing is known about the distribution of the population of differences. **[2]**

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

**(c)** Show how the value 0.07444 (labelled SE) was calculated. **[1]**

**(d)** Comment on whether you think that the confidence interval suggests that the two different methods of measurement are broadly in agreement. **[2]**

© OCR 2020

**Section B** (88 marks)

Answer **all** the questions.

**5** A hearing expert is investigating whether web-based hearing tests can be used instead of hearing tests in a hearing laboratory. The expert selects a random sample of 16 people with normal hearing. Each of them is given two hearing tests, one in the laboratory and one web-based. The scores in the laboratory-based test, $x$, and the web-based test, $y$, are both measured in the same suitable units.

**(a)** Half of the participants do the laboratory-based test first and the other half do the web-based test first.

Explain why the expert adopts this approach. **[2]**

The scatter diagram in Fig. 5 shows the data that the expert collected.

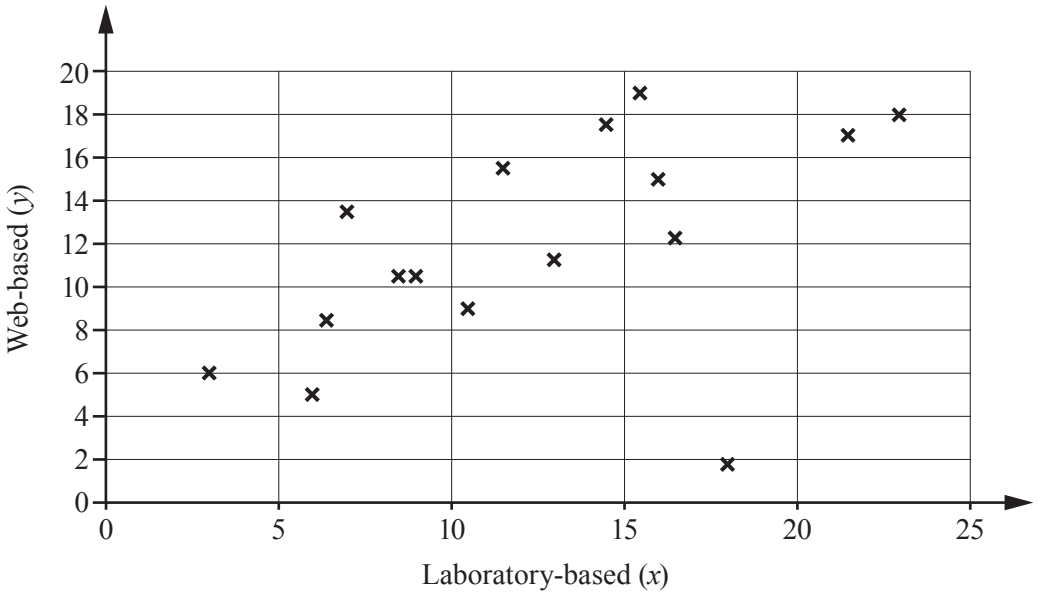



**Fig. 5**

Summary statistics for these data are as follows.

$\Sigma x = 198.0$ $\Sigma x^2 = 2936.92$ $\Sigma y = 188.7$ $\Sigma y^2 = 2605.35$ $\Sigma xy = 2554.87$

**(b)** Calculate the equation of the regression line suitable for estimating web-based scores from laboratory-based scores. **[5]**

**(c)** Estimate the web-based scores of people whose laboratory-based scores were as follows.

- 12
- 25

**[2]**

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

**(e)** A colleague of the expert suggests that the regression line is not valid because one of the data values is an outlier.

Stating the approximate coordinates of the outlier, suggest what the expert should do. **[2]**

© OCR 2020

**6** A pollution control officer is investigating a possible link between the levels of various pollutants in the air and the speed of the wind at various sites. A random sample of 60 values of the wind-speed together with the levels of a variety of pollutants is taken at a particular site. The product moment correlation coefficient between wind-speed and nitrogen dioxide level is 0.3231.

**(a)** Carry out a hypothesis test at the 10% significance level to investigate whether there is any correlation between wind-speed and nitrogen dioxide level. **[5]**

**(b)** State the condition required for the test carried out in part **(a)** to be valid. **[1]**

Table 6.1 shows the values of the product moment correlation coefficient between 5 different measures of pollution and also wind-speed for a very large random sample of values at another site. Those correlations that are significant at the 10% level are denoted by a * after the value of the correlation.

| **Correlations** | **PM10** | **SPEED** | **$NO_2$** | **$O_3$** | **PM25** | **$SO_2$** |
|---|---|---|---|---|---|---|
| **PM10** | 1.00 | | | | | |
| **SPEED** | 0.08* | 1.00 | | | | |
| **$NO_2$** | 0.59* | 0.25* | 1.00 | | | |
| **$O_3$** | −0.05* | −0.04* | −0.30* | 1.00 | | |
| **PM25** | 0.85* | −0.01 | 0.56* | −0.02 | 1.00 | |
| **$SO_2$** | 0.42* | 0.15* | 0.73* | −0.63* | 0.40* | 1.00 |

**Table 6.1**

Table 6.2 shows standard guidelines for effect sizes.

| **Product moment correlation coefficient** | **Effect size** |
|---|---|
| 0.1 | Small |
| 0.3 | Medium |
| 0.5 | Large |

**Table 6.2**

The officer analyses these data for effect size.

**(c)** Explain how the very large sample size relates to the interpretation of the correlation coefficients shown in Table 6.1. **[2]**

**(d)** Comment briefly on what the pollution control officer might conclude from these tables, relevant to her investigation into wind-speed and pollutant levels. **[2]**

© OCR 2020

**7** The lengths in mm of a random sample of 6 one-year-old fish of a particular species are as follows.

271 293 306 287 264 290

**(a)** State an assumption required in order to find a confidence interval for the mean length of one-year-old fish of this species. **[1]**

Fig. 7 shows a Normal probability plot for these data.

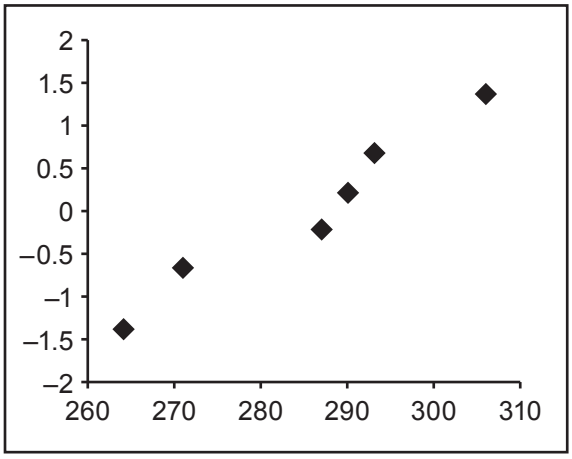



**Fig. 7**

**(b)** Explain why the Normal probability plot suggests that the assumption in part **(a)** may be valid. **[1]**

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

Assuming that this assumption is true, find a 95% confidence interval for the mean length of one-year-old fish of this species. **[7]**

**8 In this question you must show detailed reasoning.**

On the manufacturer's website, it is claimed that the average daily electricity consumption of a particular model of fridge is 1.25 kWh (kilowatt hours). A researcher at a consumer organisation decides to check this figure.

A random sample of 40 fridges is selected. Summary statistics for the electricity consumption $x$ kWh of these fridges, measured over a period of 24 hours, are as follows.

$\Sigma x = 51.92$ $\Sigma x^2 = 70.57$

Carry out a test at the 5% significance level to investigate the validity of the claim on the website. **[10]**

**9** A supermarket sells trays of peaches. Each tray contains 10 peaches. Often some of the peaches in a tray are rotten. The numbers of rotten peaches in a random sample of 150 trays are shown in Table 9.1.

| Number of rotten peaches | 0 | 1 | 2 | 3 | 4 | 5 | 6 | ⩾7 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 39 | 39 | 33 | 19 | 8 | 8 | 4 | 0 |

**Table 9.1**

A manager at the supermarket thinks that the number of rotten peaches in a tray may be modelled by a binomial distribution.

**(a)** Use these data to estimate the value of the parameter $p$ for the binomial model B(10, $p$). **[2]**

The manager decides to carry out a goodness of fit test to investigate further. The screenshot in Fig. 9.2 shows part of a spreadsheet to assess the goodness of fit of the distribution B(10, $p$), using the value of $p$ estimated from the data.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Number of rotten peaches | Observed frequency | Binomial probability | Expected frequency | Chi-squared contribution |
| 2 | 0 | 39 | | | |
| 3 | 1 | 39 | | | 1.4229 |
| 4 | 2 | 33 | 0.2941 | 44.1167 | 2.8012 |
| 5 | 3 | 19 | 0.1629 | 24.4383 | 1.2102 |
| 6 | ⩾4 | 20 | 0.0769 | 11.5311 | 6.2199 |
| 7 | | | | | |

**Fig. 9.2**

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

- C2
- D2
- E2

**[4]**

**(c)** Explain why the numbers for 4, 5, 6 and at least 7 rotten peaches have been combined into the single category of at least 4 rotten peaches, as shown in the spreadsheet. **[1]**

**(d)** Carry out the test at the 1% significance level. **[6]**

**(e)** Using the values of the contributions, comment on the results of the test. **[3]**

© OCR 2020

**10** The discrete random variables $X$ and $Y$ have distributions as follows: $X \sim \text{B}(20, 0.3)$ and $Y \sim \text{Po}(3)$.

The spreadsheet in Fig. 10 shows a simulation of the distributions of $X$ and $Y$. Each of the 20 rows below the heading row consists of a value of $X$, a value of $Y$, and the value of $X - 2Y$.

| | A | B | C |
|---|---|---|---|
| 1 | $X$ | $Y$ | $X-2Y$ |
| 2 | 6 | 6 | −6 |
| 3 | 5 | 4 | −3 |
| 4 | 8 | 1 | 6 |
| 5 | 6 | 5 | −4 |
| 6 | 6 | 3 | 0 |
| 7 | 8 | 1 | 6 |
| 8 | 6 | 4 | −2 |
| 9 | 5 | 4 | −3 |
| 10 | 7 | 4 | −1 |
| 11 | 8 | 3 | 2 |
| 12 | 6 | 2 | 2 |
| 13 | 5 | 1 | 3 |
| 14 | 6 | 1 | 4 |
| 15 | 5 | 4 | −3 |
| 16 | 7 | 2 | 3 |
| 17 | 5 | 2 | 1 |
| 18 | 4 | 4 | −4 |
| 19 | 5 | 0 | 5 |
| 20 | 5 | 1 | 3 |
| 21 | 4 | 2 | 0 |

**Fig. 10**

**(a)** Use the spreadsheet to estimate each of the following.

- $\text{P}(X - 2Y > 0)$
- $\text{P}(X - 2Y > 1)$ **[2]**

**(b)** How could the estimates in part **(a)** be improved? **[1]**

The mean of 50 values of $X - 2Y$ is denoted by the random variable $W$.

**(c)** Calculate an estimate of $\text{P}(W > 1)$. **[9]**

**11** The length of time in minutes for which a particular geyser erupts is modelled by the continuous random variable $T$ with cumulative distribution function given by

$$\mathrm{F}(t) = \begin{cases} 0 & t \leqslant 2, \\ k(8t^2 - t^3 - 24) & 2 < t < 4, \\ 1 & t \geqslant 4, \end{cases}$$

where $k$ is a positive constant.

**(a)** Show that $k = \frac{1}{40}$. **[1]**

**(b)** Find the probability that a randomly selected eruption time lies between 2.5 and 3.5 minutes. **[2]**

**(c)** Show that the median $m$ of the distribution satisfies the equation $m^3 - 8m^2 + 44 = 0$. **[2]**

**(d)** Verify that the median eruption time is 2.95 minutes, correct to 2 decimal places. **[2]**

The mean and standard deviation of $T$ are denoted by $\mu$ and $\sigma$ respectively.

**(e)** Find $\mathrm{P}(\mu - \sigma < T < \mu + \sigma)$. **[7]**

**(f)** Sketch the graph of the probability density function of $T$. **[2]**

**(g)** A Normally distributed random variable $X$ has the same mean and standard deviation as $T$.

By considering the shape of the Normal distribution, and without doing any calculations, explain whether $\mathrm{P}(\mu - \sigma < X < \mu + \sigma)$ will be greater than, equal to or less than the probability that you calculated in part **(e)**. **[2]**

**END OF QUESTION PAPER**

© OCR 2020

**BLANK PAGE**

© OCR 2020

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 2020