Oxford Cambridge and RSA

# Tuesday 9 June 2015 – Morning

## A2 GCE MATHEMATICS (MEI)

**4768/01** Statistics 3

**QUESTION PAPER**

Candidates answer on the Printed Answer Book.

**OCR supplied materials:**
- Printed Answer Book 4768/01
- MEI Examination Formulae and Tables (MF2)

**Other materials required:**
- Scientific or graphical calculator

**Duration:** 1 hour 30 minutes

### INSTRUCTIONS TO CANDIDATES

These instructions are the same on the Printed Answer Book and the Question Paper.

- The Question Paper will be found inside the Printed Answer Book.
- Write your name, centre number and candidate number in the spaces provided on the Printed Answer Book. Please write clearly and in capital letters.
- **Write your answer to each question in the space provided in the Printed Answer Book**. Additional paper may be used if necessary but you must clearly show your candidate number, centre number and question number(s).
- Use black ink. HB pencil may be used for graphs and diagrams only.
- Read each question carefully. Make sure you know what you have to do before starting your answer.
- Answer **all** the questions.
- Do **not** write in the bar codes.
- 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 FOR CANDIDATES

This information is the same on the Printed Answer Book and the Question Paper.

- The number of marks is given in brackets **[ ]** at the end of each question or part question on the Question Paper.
- 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 being used.
- The total number of marks for this paper is **72**.
- The Printed Answer Book consists of **12** pages. The Question Paper consists of **4** pages. Any blank pages are indicated.

### INSTRUCTION TO EXAMS OFFICER / INVIGILATOR

- Do not send this Question Paper for marking; it should be retained in the centre or recycled. Please contact OCR Copyright should you wish to re-use this document.

© OCR 2015 [R/102/2658]
DC (ST) 75067/3

OCR is an exempt Charity

**1** **(a)** A stratified sample of pupils at secondary schools in a particular local authority is to be chosen in order to collect information on absenteeism. In the local authority there are 4 secondary schools, A, B, C and D, with 1310, 1453, 843 and 1110 pupils respectively.

**(i)** How many pupils should be chosen from each school in a stratified sample of 500 so that each school is represented proportionally? **[3]**

**(ii)** Suggest two possible criteria for stratification other than by school. **[2]**

**(iii)** State one advantage of choosing a stratified sample. **[1]**

**(b)** At a large secondary school, the median number of half days absent per pupil per year (based on several years' records) was known to be 23. Last year the school carried out a drive to lower the number of absences. A random sample of 12 pupils had been absent for the following numbers of half days during the year.

14 10 15 13 35 9 24 19 30 26 29 8

A Wilcoxon single sample test is to be carried out to see if the drive has been successful.

**(i)** Why might a Wilcoxon test be appropriate? **[1]**

**(ii)** What distributional assumption is needed for the test? **[1]**

**(iii)** Carry out the test, using a 5% level of significance. **[10]**

**2** The distribution of the random variable $X$ is thought to be well modelled by the following probability density function:

$$\mathrm{f}(x) = \begin{cases} k(1+x) & \text{for } 0 \leqslant x < 5, \\ 0 & \text{elsewhere,} \end{cases}$$

where $k$ is a positive constant.

**(i)** Find the value of $k$. **[3]**

**(ii)** Show that $\mathrm{P}(a \leqslant X < a+1) = \frac{1}{35}(2a+3)$ for $0 \leqslant a \leqslant 4$. **[2]**

A random sample of 50 observations of $X$ is summarised as follows.

| $x$ | $0 \leqslant x < 1$ | $1 \leqslant x < 2$ | $2 \leqslant x < 3$ | $3 \leqslant x < 4$ | $4 \leqslant x < 5$ |
|---|---|---|---|---|---|
| Frequency | 1 | 5 | 7 | 20 | 17 |

**(iii)** Test at the 10% level of significance whether the distribution of $X$ is well modelled by $\mathrm{f}(x)$. **[10]**

**(iv)** With reference to your calculations in part **(iii)** discuss briefly the outcome of the test. **[2]**

© OCR 2015

**3** In agricultural research the oil content, as a percentage of the whole grain, of a cereal can be measured using near infra-red spectroscopy. An investigation into the effect of a particular treatment on the oil content of a certain cereal is being carried out. A sample of 10 plots of land is chosen and each plot is divided in half. In one half of each plot the cereal is grown with the treatment and in the other half the cereal is grown without the treatment. Subsequently the percentage oil content of the cereal for each half of each plot is measured and the results are as follows.

| Plot | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| With treatment | 41.1 | 44.3 | 42.4 | 48.2 | 52.4 | 54.6 | 35.9 | 33.6 | 51.1 | 47.0 |
| Without treatment | 42.5 | 37.7 | 42.1 | 32.4 | 42.7 | 41.5 | 36.9 | 31.7 | 52.6 | 41.2 |

A paired *t* test with a 5% level of significance is to be used to see if the treatment appears to make any difference to the mean percentage oil content of the cereal.

**(i)** Explain what is meant by a 5% level of significance in a hypothesis test. **[2]**

**(ii)** State the conditions necessary for the test to be carried out. **[3]**

**(iii)** Assuming the conditions stated in part **(ii)** are met, carry out the test. **[10]**

**(iv)** Find a 90% confidence interval for the population mean difference in the percentage oil content with and without the treatment. **[4]**

**4** Paul has been trying a new route to work in the mornings. He collects a large random sample of times, in minutes, and calculates a 95% confidence interval for the population mean time by this route. The confidence interval is $(45.369, 47.231)$ and the sample variance is 20.3.

**(i)** Explain what is meant by a 95% confidence interval for a population mean. **[1]**

**(ii)** Calculate the sample mean and the sample size. **[4]**

Paul reverts to his usual route and the time, in minutes, to travel to work each morning is modelled by a random variable which is Normally distributed with mean 41.3 and variance 11.7. The time, in minutes, for Paul to travel home each evening is modelled by a random variable which is Normally distributed with mean 44.8 and variance 14.2. In the rest of this question all journeys are by Paul's usual route and may be assumed to be independent of each other.

**(iii)** Calculate the probability that, on a randomly chosen day, Paul's total travelling time will be less than 90 minutes. **[3]**

**(iv)** Calculate the probability that, on a randomly chosen day, the time for Paul to travel home will be more than 5 minutes longer than the time to travel to work. **[4]**

**(v)** Calculate the probability that, in a randomly chosen five-day week, the mean time for Paul to travel to work on Monday and Tuesday will be more than 3 minutes longer than his mean time to travel to work on Wednesday, Thursday and Friday. **[6]**

**END OF QUESTION PAPER**

© OCR 2015

<table>
<tr><th colspan="3">Question</th><th>Answer</th><th>Marks</th><th>Guidance</th></tr>
<tr><td>1</td><td>(a)</td><td>(i)</td><td>$\frac{500}{4716} \times ...$<br>= 138.88…, 154.05…, 89.37…, 117.68…<br>= 139, 154, 89, 118</td><td>M1<br>A1<br>A1<br>[3]</td><td>Correct factor used for at least one school.<br>All correct and given to at least 1 dp.<br>FT any errors in the previous line provided that sum = 500.</td></tr>
<tr><td>1</td><td>(a)</td><td>(ii)</td><td>e.g. Sex (gender)<br>Year group</td><td>B1<br>B1<br>[2]</td><td>Allow reasonable alternatives including ethnicity, birth date, distance from school</td></tr>
<tr><td>1</td><td>(a)</td><td>(iii)</td><td>e.g. Provides information on each stratum (as well as the population).</td><td>B1<br>[1]</td><td>Or representative</td></tr>
<tr><td>1</td><td>(b)</td><td>(i)</td><td>We have no information about the background population.</td><td>E1<br>[1]</td><td>o.e. Must include “population” o.e.</td></tr>
<tr><td>1</td><td>(b)</td><td>(ii)</td><td>Symmetry.</td><td>B1<br>[1]</td><td></td></tr>
<tr><td>1</td><td>(b)</td><td>(iii)</td><td>$H_0$: $m = 23$ $H_1$: $m < 23$<br>where $m$ is the population median number of days absent.<br>
<table>
<tr><th>Absences</th><th>−23</th><th>Rank of |diff|</th></tr>
<tr><td>14</td><td>–9</td><td>7</td></tr>
<tr><td>10</td><td>–13</td><td>10</td></tr>
<tr><td>15</td><td>–8</td><td>6</td></tr>
<tr><td>13</td><td>–10</td><td>8</td></tr>
<tr><td>35</td><td>12</td><td>9</td></tr>
<tr><td>9</td><td>–14</td><td>11</td></tr>
<tr><td>24</td><td>1</td><td>1</td></tr>
<tr><td>19</td><td>–4</td><td>3</td></tr>
<tr><td>30</td><td>7</td><td>5</td></tr>
<tr><td>26</td><td>3</td><td>2</td></tr>
<tr><td>29</td><td>6</td><td>4</td></tr>
<tr><td>8</td><td>–15</td><td>12</td></tr>
</table>
</td><td>B1<br>B1<br>M1<br>M1<br>A1</td><td>Both. Accept hypotheses in words, but must include “population”. Do NOT allow symbols other than $m$ unless clearly and explicitly stated to be a population median.<br>Adequate definition of $m$ to include “population”.<br>for subtracting 23.<br>for ranks.<br>ft if ranks wrong.</td></tr>
</table>

| Question | | | Answer | Marks | Guidance |
|---|---|---|---|---|---|
| | | | $W_+ = 1 + 2 + 4 + 5 + 9 = 21$ | B1 | ($W_- = 3 + 6 + 7 + 8 + 10 + 11 + 12 = 57$) |
| | | | Refer to Wilcoxon single sample tables for $n = 12$. | M1 | No ft from here if wrong. |
| | | | Lower 5% point is 17 (or upper is 61 if 57 used). | A1 | i.e. a 1-tail test. No ft from here if wrong. |
| | | | Result is not significant. | A1 | ft only c's test statistic. Dependent on all 3 M marks |
| | | | Insufficient evidence to suggest that the median number of days absent has been reduced. | A1 | ft only c's test statistic. Dependent on all 3 M marks. Conclusion in context to include "on average" o.e. |
| | | | | **[10]** | |
| **2** | **(i)** | | Require $\int_0^5 k(1+x)\mathrm{d}x = 1$ | | |
| | | | $\int_0^5 k(1+x)\mathrm{d}x = k\left(x+\frac{x^2}{2}\right)\Big\|_0^5 = k\left(5+\frac{25}{2}\right) - k\times 0 = \frac{35k}{2}$ | M1<br>A1 | Set up correct integral, including limits which may appear later.<br>Allow method based on area, e.g., a trapezium.<br>Integral correctly evaluated, or correct area obtained, in terms of $k$. |
| | | | $\therefore \frac{35k}{2} = 1 \quad \therefore k = \frac{2}{35}$ Not 0.057…. | A1 | Set equal to 1 and rearranged for $k$. |
| | | | | **[3]** | |
| **2** | **(ii)** | | $\mathrm{P}(a \le X < a+1) = \int_a^{a+1} \frac{2}{35}(1+x)\mathrm{d}x$ | M1 | Set up correct integral, including limits which may appear later.<br>Allow method based on area, e.g., a trapezium.<br>Allow candidate's value of $k$. |
| | | | $= \frac{2}{35}\left((a+1)+\frac{(a+1)^2}{2} - a - \frac{a^2}{2}\right)$ | | |
| | | | $= \frac{2}{35}\left(1+\frac{2a+1}{2}\right) = \frac{1}{35}(2a+3)$ | A1 | AG. Must be shown convincingly. |
| | | | | **[2]** | |
| **2** | **(iii)** | | $H_0$: The model is suitable / fits the data.<br>$H_1$: The model is not suitable / does not fit the data. | B1 | Both hypotheses. Must be the right way round.<br>Do not accept "data fit model" oe. |
| | | | Expected frequencies are: $50\times\left(\frac{3,\ 5,\ 7,\ 9,\ 11}{35}\right)$ | M1 | |
| | | | $= \left(\frac{30,\ 50,\ 70,\ 90,\ 110}{7}\right)$<br>= 4.2857, 7.1428/9, 10, 12.8571, 15.7142/3 | A1 | Accept either fractions or decimals. |

| Question | | | Answer | Marks | Guidance | |
|---|---|---|---|---|---|---|
| | | | Merge first 2 cells: Obs f = 6, Exp f = 11.4285 | M1 | Merge first 2 cells | |
| | | | $X^2 = 2.5786 + 0.9 + 3.9683 + 0.1052$ | M1 | Calculation of $X^2$. Independent of previous mark. | |
| | | | $= 7.552$ | A1 | Awrt 7.55 | |
| | | | Refer to $\chi_3^2$. | M1 | Allow correct df (= cells – 1) from differently grouped table and ft. critical value only. Otherwise, no ft if wrong. | |
| | | | Upper 10% point is 6.251. | A1 | No ft from here if wrong. $P(X^2 > 7.552) = 0.0562$.<br>If cells not merged $\chi_4^2$ 10% point is 7.779; $P(X^2 > 8.135) = 0.0867$. | |
| | | | Significant. | A1 | ft only c's test statistic. | |
| | | | Sufficient evidence to suggest that the pdf of $X$ is not well modelled by f($x$).. | A1<br>**[10]** | ft only c's test statistic.<br>Do not accept "data do not fit model" oe. | |
| **2** | **(iv)** | | e.g. The model overestimates for $0 \le x < 2$.<br>The model underestimates for $3 \le x < 4$.<br>"Large discrepancy" but no direction E1 max | E1<br>E1<br>**[2]** | Any 2 points relating to or explaining the outcome of the test.<br>Other possibilities might include:<br>The test would not have been significant at 5%.<br>The sample is a bit small making it difficult to assess. | |
| **3** | **(i)** | | 5% represents the probability of rejecting the null hypothesis …<br>… when it is, in fact, true. | E1<br>E1<br>**[2]** | | |
| **3** | **(ii)** | | Must assume: Normality of population …<br>… of <u>differences</u>.<br>Sample is random. | B1<br>B1<br>B1<br>**[3]** | Ignore references to unknown variance and/or sample size. | |
| **3** | **(iii)** | | $H_0$: $\mu_D = 0$<br>$H_1$: $\mu_D \ne 0$ | B1 | Both. Accept alternatives e.g. $\mu_A - \mu_B$ etc provided adequately defined.<br>Hypotheses in words only must include "population". Do NOT allow "$\bar{X} = \ldots$" or similar unless $\bar{X}$ is clearly and explicitly stated to be a <u>population</u> mean. | |
| | | | Where $\mu_D$ is the (population) mean difference in percentage oil content. | B1 | For adequate verbal definition. Allow absence of "population" if correct notation $\mu$ is used. | |

| Question | | | Answer | Marks | Guidance | |
|---|---|---|---|---|---|---|
| | | | <u>MUST</u> be PAIRED COMPARISON *t* test.<br>Differences (with – without) are:<br>–1.4 6.6 0.3 15.8 9.7 13.1 –1.0 1.9 –1.5 5.8 | M1 | Allow “without – with” if consistent with alternatives for hypotheses above. | |
| | | | $\bar{x} = 4.93$ $s_{n-1} = 6.310(4)$ $(s_{n-1}{}^2 = 39.822(3))$ | A1 | Do not allow $s_n$ = 5.9886 ($s_n{}^2$ = 35.8401). | |
| | | | Test statistic is $\dfrac{4.93-0}{\dfrac{6.310}{\sqrt{10}}}$ | M1 | Allow c’s $\bar{x}$ and/or $s_{n-1}$.<br>Allow alternative: 0 + (c’s 2.262) × $\dfrac{6.3104}{\sqrt{10}}$ (= 4.514) for subsequent comparison with $\bar{x}$ .<br>(Or $\bar{x}$ – (c’s 2.262) × $\dfrac{6.3104}{\sqrt{10}}$ (= 0.416) for comparison with 0.) | |
| | | | = 2.470(4). | A1 | c.a.o. but ft from here in any case if wrong.<br>Use of 10 – $\bar{x}$ scores M1A0, but ft. | |
| | | | Refer to $t_9$. | M1 | No ft from here if wrong. | |
| | | | Double-tailed 5% point is ±2.262. | A1 | Must be minus 2.262 for “without – with” unless absolute values are being compared. No ft from here if wrong. P(\|$t$\| > 2.4704) = 0.03554. | |
| | | | Significant. | A1 | ft only c’s test statistic as long as it includes their $s/\sqrt{n}$ | |
| | | | Sufficient evidence to suggest that the treatment appears to make a difference to the mean percentage oil content of the cereal. | A1 | ft only c’s test statistic as above<br>Conclusion in context to include “on average” o.e. | |
| | | | | **[10]** | | |
| **3** | **(iv)** | | CI is given by 4.93 ± | M1 | ZERO/4 if not same distribution as test. Same wrong distribution scores maximum M1B0M1A0. Recovery to $t_9$ is OK.<br>Allow c’s $\bar{x}$ . | |
| | | | 1.833 | B1 | 1.833 seen. | |
| | | | × $\dfrac{6.3104}{\sqrt{10}}$ | M1 | Allow c’s $s_{n-1}$. | |
| | | | = 4.93 ± 3.6577= (1.271(9), 8.588(1)) | A1 | c.a.o. Must be expressed as an interval. | |
| | | | | **[4]** | | |

| Question | | | Answer | Marks | Guidance |
|---|---|---|---|---|---|
| 4 | (i) | | In repeated sampling, 95% of all confidence intervals constructed in this way will contain the true mean. | E1<br>**[1]** | |
| 4 | (ii) | | Mean = (45.369 + 47.231)/2=46.3<br>$47.231 = 46.3 + \sqrt{20.3} \times 1.96 / \sqrt{n}$<br>$n = \dfrac{1.96^2 \times 20.3}{0.931^2} \approx 89.97(2)$ $\therefore n = 90$ | B1<br>B1<br>M1<br>A1<br>**[4]** | cao<br>Sight of 1.96.<br>Or equivalent.<br>Must be an integer. FT candidate's mean. |
| 4 | (iii) | | Time to work $X \sim \text{N}(41.3,\ 11.7)$<br>Time to home $Y \sim \text{N}(44.8,\ 14.2)$<br>$X + Y \sim \text{N}(86.1,\ 25.9)$<br>$\text{P}(X + Y < 90) = \Phi\left(\dfrac{90 - 86.1}{\sqrt{25.9}} = 0.7663\right) = 0.7783$ | B1<br>B1<br>B1<br>**[3]** | Mean<br>Variance<br>cao |
| 4 | (iv) | | Require $\text{P}(Y - X > 5)$<br>$Y - X \sim \text{N}(3.5,\ 25.9)$<br>$\text{P}(Y - X > 5) = 1 - \Phi\left(\dfrac{5 - 3.5}{\sqrt{25.9}} = 0.2947\right) = 1 - 0.6159 = 0.3841$ | M1<br>B1<br>B1<br>A1<br>**[4]** | Allow equivalent alternatives, e.g. $Y > X + 5$ or $X - Y < -5$<br>Mean<br>Variance<br>cao |
| 4 | (v) | | Require $\text{P}\left(\dfrac{X_1 + X_2}{2} > \dfrac{X_3 + X_4 + X_5}{3} + 3\right)$<br>Mean = 3(41.3 + 41.3) – 2(41.3 + 41.3 + 41.3) = 0<br>Variance = $\dfrac{1}{4}$ (11.7 + 11.7) + $\dfrac{1}{9}$(11.7 + 11.7 + 11.7) = 9.75 | M1<br>A1<br>B1<br>M1<br>A1 | For considering some $\overline{X}_2$ and $\overline{X}_3$<br>For $\overline{X}_2$ - $\overline{X}_3$ > 3<br>For 0<br>For1/4, 1/9, and 11.7 seen o.e.<br>For 9.75 |

| Question | | | Answer | Marks | Guidance | |
|---|---|---|---|---|---|---|
| | | | P( $Z > \dfrac{3-0}{\sqrt{9.75}}$ ) = 1 – 0.8317 = 0.1683 | A1 | cao | |
| | | | | **[6]** | | |