

Friday 20 January 2012 – Afternoon

A2 GCE MATHEMATICS (MEI)

4768 Statistics 3

QUESTION PAPER



Candidates answer on the Printed Answer Book.

OCR supplied materials:

- Printed Answer Book 4768
- 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 in the centre of 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.
- Answer **all** the questions.
- Read each question carefully. Make sure you know what you have to do before starting your answer.
- 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.

- 1 (a) Define simple random sampling. Describe briefly one difficulty associated with simple random sampling. [4]

- (b) Freeze-drying is an economically important process used in the production of coffee. It improves the retention of the volatile aroma compounds. In order to maintain the quality of the coffee, technologists need to monitor the drying rate, measured in suitable units, at regular intervals. It is known that, for best results, the mean drying rate should be 70.3 units and anything substantially less than this would be detrimental to the coffee. Recently, a random sample of 12 observations of the drying rate was as follows.

66.0 66.1 59.8 64.0 70.9 71.4 66.9 76.2 65.2 67.9 69.2 68.5

- (i) Carry out a test to investigate at the 5% level of significance whether the mean drying rate appears to be less than 70.3. State the distributional assumption that is required for this test. [10]
- (ii) Find a 95% confidence interval for the true mean drying rate. [4]
- 2 In a particular chain of supermarkets, one brand of pasta shapes is sold in small packets and large packets. Small packets have a mean weight of 505 g and a standard deviation of 11 g. Large packets have a mean weight of 1005 g and a standard deviation of 17 g. It is assumed that the weights of packets are Normally distributed and are independent of each other.
- (i) Find the probability that a randomly chosen large packet weighs between 995 g and 1020 g. [3]
- (ii) Find the probability that the weights of two randomly chosen small packets differ by less than 25 g. [3]
- (iii) Find the probability that the total weight of two randomly chosen small packets exceeds the weight of a randomly chosen large packet. [4]
- (iv) Find the probability that the weight of one randomly chosen small packet exceeds half the weight of a randomly chosen large packet by at least 5 g. [4]
- (v) A different brand of pasta shapes is sold in packets of which the weights are assumed to be Normally distributed with standard deviation 14 g. A random sample of 20 packets of this pasta is found to have a mean weight of 246 g. Find a 95% confidence interval for the population mean weight of these packets. [4]

- 3 (a) A medical researcher is looking into the delay, in years, between first and second myocardial infarctions (heart attacks). The following table shows the results for a random sample of 225 patients.

Delay (years)	0 –	1 –	2 –	3 –	4 – 10
Number of patients	160	40	13	9	3

The mean of this sample is used to construct a model which gives the following expected frequencies.

Delay (years)	0 –	1 –	2 –	3 –	4 – 10
Number of patients	142.23	52.32	19.25	7.08	4.12

Carry out a test, using a 2.5% level of significance, of the goodness of fit of the model to the data. [9]

- (b) A further piece of research compares the incidence of myocardial infarction in men aged 55 to 70 with that in women aged 55 to 70. Incidence is measured by the number of infarctions per 10 000 of the population. For a random sample of 8 health authorities across the UK, the following results for the year 2010 were obtained.

Health authority	A	B	C	D	E	F	G	H
Incidence in men	47	56	15	51	45	54	50	32
Incidence in women	36	30	30	47	54	55	27	27

A Wilcoxon paired sample test, using the hypotheses $H_0: m = 0$ and $H_1: m \neq 0$ where m is the population median difference, is to be carried out to investigate whether there is any difference between men and women on the whole.

- (i) Explain why a paired test is being used in this context. [1]
 (ii) Carry out the test using a 10% level of significance. [8]

[Question 4 is printed overleaf.]

- 4 At the school summer fair, one of the games involves throwing darts at a circular dartboard of radius a lying on the ground some distance away. Only darts that land on the board are counted. The distance from the centre of the board to the point where a dart lands is modelled by the random variable R . It is assumed that the probability that a dart lands inside a circle of radius r is proportional to the area of the circle.

(i) By considering $P(R < r)$ show that $F(r)$, the cumulative distribution function of R , is given by

$$F(r) = \begin{cases} 0 & r < 0, \\ \frac{r^2}{a^2} & 0 \leq r \leq a, \\ 1 & r > a. \end{cases} \quad [3]$$

(ii) Find $f(r)$, the probability density function of R . [2]

(iii) Find $E(R)$ and show that $\text{Var}(R) = \frac{a^2}{18}$. [7]

The radius a of the dartboard is 22.5 cm.

- (iv) Let \bar{R} denote the mean distance from the centre of the board of a random sample of 100 darts. Write down an approximation to the distribution of \bar{R} . [3]
- (v) A random sample of 100 darts is found to give a mean distance of 13.87 cm. Does this cast any doubt on the modelling? [3]

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 Copyright Team, First Floor, 9 Hills Road, Cambridge CB2 1GE.

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.

Question		Answer	Marks	Guidance
1	(a)	<p>Simple random sampling is when every <u>sample</u> of the required size ... has an equal chance of being chosen. eg One needs access to the entire population in order to establish the sampling frame.</p>	E1 E1 E2 [4]	<p>SC Allow E1 for “Every member of the population has an equal chance of being chosen”.</p> <p>E2, 1, 0. Reward any sensible point.</p> <p>SC1 “Sample may not be representative of the population” oe.</p>
1	(b)	<p>$H_0 : \mu = 70.3$ $H_1 : \mu < 70.3$</p> <p>Where μ is the (population) mean drying rate.</p> <p>$\bar{x} = 67.675$, $s_{n-1} = 4.129$ ($s_{n-1}^2 = 17.049$)</p> <p>Test statistic is $\frac{67.675 - 70.3}{\sqrt{\frac{4.129}{12}}} = -2.202(2)$</p> <p>Refer to t_{11}. Lower 5% point is -1.796.</p> <p>$-2.202 < -1.796$, \therefore Result is significant. Seems mean drying rate has reduced.</p> <p>Underlying population is Normal.</p>	B1 B1 B1 M1 A1 M1 A1 A1 A1 B1 [10]	<p>Both. Hypotheses in words only must include “population”. Do NOT allow “$\bar{X} = \dots$” or similar unless \bar{X} is clearly and explicitly stated to be a <u>population</u> mean. For adequate verbal definition. Allow absence of “population” if correct notation μ is used. Do not allow $s_n = 3.953$ ($s_n^2 = 15.629$).</p> <p>Allow c’s \bar{x} and/or s_{n-1}. Allow alternative: $70.3 + (c's - 1.796) \times \frac{4.129}{\sqrt{12}}$ (= 68.15(9)) for subsequent comparison with \bar{x}. (Or $\bar{x} - (c's - 1.796) \times \frac{4.129}{\sqrt{12}}$ (= 69.81(6)) for comparison with 70.3) cao but ft from here in any case if wrong. Use of $70.3 - \bar{x}$ scores M1A0, but ft.</p> <p>No ft from here if wrong. Allow any t_{11} value from tables. c.a.o. No ft from here if wrong. Must be -1.796 unless it is clear that absolute values are being used. $P(t < -2.202) = 0.0249$. ft only c’s test statistic. ft only c’s test statistic. “Non-assertive” conclusion in context to include “on average” oe.</p>

Question			Answer	Marks	Guidance
1	(b)	(ii)	CI is given by $67.675 \pm 2.201 \times \frac{4.129}{\sqrt{12}}$ $= 67.675 \pm 2.623 = (65.05(2), 70.29(8))$	M1 B1 M1 A1 [4]	ZERO if not same distribution as test. Same wrong distribution scores max M1B0M1A0. Recovery to t_{11} is OK. ft c's $\bar{x} \pm$. ft c's s_{n-1} . cao Must be expressed as an interval. $(t=1.796$ gives $(65.534, 69.816)$ (M1B0M1A0)).
2			$S \sim N(505, 11^2)$ $L \sim N(1005, 17^2)$		When a candidate's answers suggest that (s)he appears to have neglected to use the difference columns of the Normal distribution tables, penalise the first occurrence only.
	(i)		$P(995 < L < 1020)$ $= P\left(\frac{995 - 1005}{17} < Z < \frac{1020 - 1005}{17}\right)$ $= P(-0.5882 < Z < 0.8824)$ $= 0.8113 - (1 - 0.7218)$ $= 0.5331$	M1 A1 A1 [3]	For standardising. Award once, here or elsewhere. cao
2	(ii)		$S_1 - S_2 \sim N(0, 11^2 + 11^2 = 242)$ $P(-25 < S_1 - S_2 < 25)$ $= P\left(\frac{-25 - 0}{\sqrt{242}} < Z < \frac{25 - 0}{\sqrt{242}}\right)$ $= P(-1.607 < Z < 1.607)$ $= 2 \times (0.9459 - 0.5)$ $= 0.8918$	B1 M1 A1 [3]	Mean and variance. Accept sd = $\sqrt{242} = 15.55\dots$ Formulate the problem. cao

Question		Answer	Marks	Guidance												
2	(iii)	<p>Want $P(S_1 + S_2 > L)$ i.e. $P(S_1 + S_2 - L > 0)$</p> $S_1 + S_2 - L \sim N(505 + 505 - 1005 = 5, 11^2 + 11^2 + 17^2 = 531)$ $P(\text{this} > 0) = P(Z > \frac{0 - 5}{\sqrt{531}}) = -0.2170$ $= 0.5859$	M1 B1 B1 A1 [4]	Allow $L - (S_1 + S_2)$ provided subsequent work is consistent. Mean Variance. Accept sd = $\sqrt{531} = 23.04\dots$ cao												
2	(iv)	<p>Want $P(S > \frac{1}{2}L + 5)$ i.e. $P(S - \frac{1}{2}L > 5)$</p> $S - \frac{1}{2}L \sim N(505 - 1005/2 = 2.5, 11^2 + 17^2/2^2 = 193.25)$ $P(\text{this} > 5) = P(Z > \frac{5 - 2.5}{\sqrt{193.25}}) = 0.1798$ $= 1 - 0.5714 = 0.4286$	M1 B1 B1 A1 [4]	Allow $\frac{1}{2}L - S$ provided subsequent work is consistent. Mean. Variance. Accept sd = $\sqrt{193.25} = 13.90\dots$ cao												
2	(v)	<p>CI is given by</p> $246 \pm \frac{1.96}{\sqrt{20}} \times \frac{14}{\sqrt{20}}$ $= 246 \pm 6.1358 = (239.8(6), 252.1(3))$	M1 B1 M1 A1 [4]	Must be 1.96. Anything else can get M1B0M1A0 max. cao Must be expressed as an interval.												
3	(a)	<p>H_0: The model for the delay fits the data. H_1: The model for the delay does not fit the data.</p> <table border="1"> <tr> <td>Obs'd frequency</td> <td>160</td> <td>40</td> <td>13</td> <td>9</td> <td>3</td> </tr> <tr> <td>Exp'd frequency</td> <td>142.23</td> <td>52.32</td> <td>19.25</td> <td>7.08</td> <td>4.12</td> </tr> </table> <p>Merge last 2 cells: Obs 12 Exp 11.2 $X^2 = 2.2202 + 2.9010 + 2.0292 + 0.0571 = 7.207(5)$ Refer to χ^2. Upper 2.5% point is 7.378.</p>	Obs'd frequency	160	40	13	9	3	Exp'd frequency	142.23	52.32	19.25	7.08	4.12	B1 B1 M1 M1 A1 M1 A1 A1 [4]	Do not allow hypotheses of the form "Data fit model" o.e. Calculation of X^2 . cao. If not merged, $X^2 = 7.975(5\dots)$ No ft if wrong. Allow correct dof (= cells - 2) from wrongly grouped table and ft. Allow any value from tables for c's dof. c.a.o. Upper 2.5% point for c's dof. No ft from here if wrong. $P(X^2 > 7.2075) = 0.0272$.
Obs'd frequency	160	40	13	9	3											
Exp'd frequency	142.23	52.32	19.25	7.08	4.12											

Question		Answer	Marks	Guidance																			
		7.207 < 7.378 ∴ Not Significant. Suggests it is reasonable to suppose the model fits the data.	A1 A1 [9]	ft only c's test statistic. ft only c's test statistic. "Non-assertive" conclusion in words (+ context). Do not allow "Data fit model" o.e.																			
3	(b)	(i) A paired test is used in this context in order to eliminate differences between health authorities.	E1 [1]	oe																			
3	(b)	(ii)	<table border="1" style="margin-bottom: 10px;"> <tr> <td>Diff</td><td>11</td><td>26</td><td>-15</td><td>4</td><td>-9</td><td>-1</td><td>23</td><td>5</td></tr> <tr> <td>Rank</td><td>5</td><td>8</td><td>6</td><td>2</td><td>4</td><td>1</td><td>7</td><td>3</td></tr> </table> <p> $W_- = 1 + 4 + 6 = 11$ Refer to tables of Wilcoxon paired (/single sample) statistic for $n = 8$. Lower 5% tail is 5 (or upper is 31 if 25 used). $11 > 5 \therefore$ Result is not significant. No evidence to suggest a difference between the incidences of myocardial infarction in men and women on the whole. </p>	Diff	11	26	-15	4	-9	-1	23	5	Rank	5	8	6	2	4	1	7	3	M1 M1 A1 B1 M1 A1 A1 A1 [8]	For differences. ZERO in this section if differences not used. For ranks. ft from here if ranks wrong. (or $W_+ = 2 + 3 + 5 + 7 + 8 = 25$) No ft from here if wrong. ie a 2-tail test. No ft from here if wrong. ft only c's test statistic. ft only c's test statistic. "Non-assertive" conclusion in context to include "on the whole" oe.
Diff	11	26	-15	4	-9	-1	23	5															
Rank	5	8	6	2	4	1	7	3															
4	(i)	$P(R < r) = k\pi r^2$ $P(R < a) = 1 \therefore k = \frac{1}{\pi a^2}$ $\therefore P(R < r) = \frac{\pi r^2}{\pi a^2} = \frac{r^2}{a^2}.$ Thus $F(r) = \frac{r^2}{a^2}$ (for $0 \leq r \leq a$).	M1 M1 A1 [3]	Formulate probability proportional to area in terms of "k". Find k. IF M0M0, allow SC B1 for $P(R < r) = \frac{\pi r^2}{\pi a^2} = \dots$ www Convincingly shown; ANSWER GIVEN. Condone omission of $r < 0$ and/or $r > a$.																			
4	(ii)	For $0 \leq r \leq a$, $f(r) = \frac{d}{dr} F(r) = \frac{2r}{a^2}.$	M1 A1 [2]	Condone omission of $r < 0$ and/or $r > a$.																			

Question		Answer	Marks	Guidance
4	(iii)	$\begin{aligned} E(R) &= \int_0^a r \frac{2r}{a^2} dr \\ &= \left[\frac{2r^3}{3a^2} \right]_0^a \\ &= \frac{2a}{3} \end{aligned}$ $\begin{aligned} E(R^2) &= \int_0^a r^2 \frac{2r}{a^2} dr \\ &= \left[\frac{2r^4}{4a^2} \right]_0^a = \frac{a^2}{2} \end{aligned}$ $\text{Var}(R) = \frac{a^2}{2} - \left(\frac{2a}{3} \right)^2 = \frac{9a^2 - 8a^2}{18} = \frac{a^2}{18}$	M1 A1 A1 M1 A1 M1 A1 [7]	<p>Correct integral with limits (which may be implied subsequently). Correctly integrated.</p> <p>Limits used. Accept unsimplified form.</p> <p>Correct integral with limits (which may be implied subsequently). Correctly integrated and limits used. Accept unsimplified form.</p> <p>Use of $\text{Var}(R) = E(R^2) - E(R)^2$ Convincingly shown; ANSWER GIVEN. Require sight of both terms expressed with a common denominator.</p>
4	(iv)	$\bar{R} \sim (\text{approx}) N\left(\frac{2}{3} \times 22.5 = 15, \frac{22.5^2}{18 \times 100} = 0.28125\right)$	B1 B1 B1 [3]	<p>Normal. Mean. ft c's $E(R) (>0)$ with $a = 22.5$. Variance. cao ($= 0.5303(3)^2$) Accept unsimplified form.</p>
4	(v)	EITHER can argue that 13.87 is more than 2 SD's from the Mean (15). $15 - 2\sqrt{0.28125} = 13.93(9)$ <u>must</u> refer to $SD(\bar{R})$, not $SD(R)$ i.e. outlier \Rightarrow doubt. OR more formally like a significance test: refer to $N(0,1)$ $\frac{13.87 - 15}{\sqrt{0.28125}} = -2.131$, sig at (eg) 5% \Rightarrow doubt.	M1 M1 A1 [3] M1 M1 A1	Allow 1.96 SD's, but not 1.984. 1.96 gives 13.96(1). A 95% C.I. is (12.831, 14.909). Must see explicit evidence for this. ft c's mean. Could imply first M. $P(Z > 2.131) = 0.0332$. ft c's mean.