OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE

MATHEMATICS (MEI) 4768

Statistics 3

Thursday 23 June 2011
Morning

Duration: 1 hour 30 minutes

INSTRUCTIONS TO CANDIDATES

*   Write your name, centre number and candidate number in the spaces provided on the answer booklet. Please write clearly and in capital letters.
*   Use black ink. 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

*   The number of marks is given in brackets **[ ]** at the end of each question or part question.
*   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**.
*   This document consists of **4** pages. Any blank pages are indicated.

OCR is an exempt Charity

Turn over

1   Gerry runs 5000-metre races for his local athletics club. His coach has been monitoring his practice times for several months and he believes that they can be modelled using a Normal distribution with mean 15.3 minutes. The coach suggests that Gerry should try running with a pacemaker in order to see if this can improve his times. Subsequently a random sample of 10 of Gerry's times with the pacemaker is collected to see if any reduction has been achieved. The sample of times (in minutes) is as follows.

    14.86   15.00   15.62   14.44   15.27   15.64   14.58   14.30   15.08   15.08

   (i) Why might a *t* test be used for these data?   **[2]**

   (ii) Using a 5% significance level, carry out the test to see whether, on average, Gerry's times have been reduced.   **[9]**

   (iii) What is meant by 'a 5% significance level'? What would be the consequence of decreasing the significance level?   **[3]**

   (iv) Find a 95% confidence interval for the true mean of Gerry's times using a pacemaker.   **[4]**

2   Scientists researching into the chemical composition of dust in space collect specimens using a specially designed spacecraft. The craft collects the particles of dust in trays that are made up of a large array of cells containing aerogel. The aerogel traps the particles that penetrate into the cells.

   (i) For a random sample of 100 cells, the number of particles of dust in each cell was counted, giving the following results.

Number of particles	0	1	2	3	4	5	6	7	8	9	10+
Frequency	4	7	10	20	17	15	10	9	5	3	0

   It is thought that the number of particles collected in each cell can be modelled using the distribution Poisson(4.2) since 4.2 is the sample mean for these data.

   Some of the calculations for a χ^2 test are shown below. The cells for 8, 9 and 10+ particles have been combined.

Number of particles	5	6	7	8+
Observed frequency	15	10	9	8
Expected frequency	16.33	11.44	6.86	6.39
Contribution to X^2	0.1083	0.1813	0.6676	0.4056

   Complete the calculations and carry out the test using a 10% significance level to see whether the number of particles per cell may be modelled in this way.   **[12]**

   (ii) The diameters of the dust particles are believed to be distributed symmetrically about a median of 15 micrometres (μm). For a random sample of 20 particles, the sum of the signed ranks of the diameters of the particles smaller than 15 μm (W_-) is found to be 53. Test at the 5% level of significance whether the median diameter appears to be more than 15 μm.   **[6]**

3 The time, in hours, until an electronic component fails is represented by the random variable X. In this question two models for X are proposed.

   (i) In one model, X has cumulative distribution function

$$G(x) = \begin{cases} 0 & x \leqslant 0, \\ 1 - \left(1 + \dfrac{x}{200}\right)^{-2} & x > 0. \end{cases}$$

      (A) Sketch $G(x)$.   **[3]**

      (B) Find the interquartile range for this model. Hence show that a lifetime of more than 454 hours (to the nearest hour) would be classed as an outlier.   **[6]**

  (ii) In the alternative model, X has probability density function

$$f(x) = \begin{cases} \frac{1}{200} e^{-\frac{1}{200}x} & x > 0, \\ 0 & \text{elsewhere.} \end{cases}$$

      (A) For this model show that the cumulative distribution function of X is

$$F(x) = \begin{cases} 0 & x \leqslant 0, \\ 1 - e^{-\frac{1}{200}x} & x > 0. \end{cases}$$
      [3]

      (B) Show that $P(X > 50) = e^{-0.25}$.   **[2]**

      (C) It is observed that a particular component is still working after 400 hours. Find the conditional probability that it will still be working after a further 50 hours (i.e. after a total of 450 hours) given that it is still working after 400 hours.   **[4]**

4 The weights of Avonley Blue cheeses made by a small producer are found to be Normally distributed with mean 10 kg and standard deviation 0.4 kg.

   (i) Find the probability that a randomly chosen cheese weighs less than 9.5 kg.   **[3]**

One particular shop orders four Avonley Blue cheeses each week from the producer. From experience, the shopkeeper knows that the weekly demand from customers for Avonley Blue cheese is Normally distributed with mean 35 kg and standard deviation 3.5 kg. In the interests of food hygiene, no cheese is kept by the shopkeeper from one week to the next.

  (ii) Find the probability that, in a randomly chosen week, demand from customers for Avonley Blue will exceed the supply.   **[4]**

Following a campaign to promote Avonley Blue cheese, the shopkeeper finds that the weekly demand for it has increased by 30% (i.e. the mean and standard deviation are both increased by 30%). Therefore the shopkeeper increases his weekly order by one cheese.

 (iii) Find the probability that, in a randomly chosen week, demand will now exceed supply.   **[4]**

 (iv) Following complaints, the cheese producer decides to check the mean weight of the Avonley Blue cheeses. For a random sample of 12 cheeses, she finds that the mean weight is 9.73 kg. Assuming that the population standard deviation of the weights is still 0.4 kg, find a 95% confidence interval for the true mean weight of the cheeses and comment on the result. Explain what is meant by a 95% confidence interval.   **[7]**

4

THERE ARE NO QUESTIONS PRINTED ON THIS PAGE

GCE

Mathematics (MEI)

Advanced GCE

Unit **4768:** Statistics 3

Mark Scheme for June 2011

Q1				
(i)	t test might be used because • population variance is unknown • background population is Normal	E1 E1	Allow "sample is small" as an alternative.	**2**
(ii)	H_0: $\mu = 15.3$ H_1: $\mu < 15.3$	B1	Both hypotheses. Hypotheses in words only must include "population". Do NOT allow "$\overline{X} = ...$" or similar unless $\overline{X}$ is clearly and explicitly stated to be a <u>population</u> mean.	
	where μ is the mean of Gerry's times.	B1	For adequate verbal definition. Allow absence of "population" if correct notation μ is used.	
	$\overline{x} = 14.987$ $s_{n-1} = 0.4567(5)$	B1	$s_n = 0.4333$ but do <u>NOT</u> allow this here or in construction of test statistic, but FT from there.	
	Test statistic is $\dfrac{14.987 - 15.3}{\dfrac{0.45675}{\sqrt{10}}}$	M1	Allow c's $\overline{x}$ and/or s_{n-1}. Allow alternative: $15.3 + ($c's $-1.833)$ $\times \dfrac{0.45675}{\sqrt{10}}$ (= 15.035) for subsequent comparison with $\overline{x}$. (Or $\overline{x} - ($c's $-1.833) \times \dfrac{0.45675}{\sqrt{10}}$ (= 15.252) for comparison with 15.3.)	
	$= -2.167(0)$.	A1	c.a.o. but ft from here in any case if wrong. Use of $\mu - \overline{x}$ scores M1A0, but ft.	
	Refer to t_9. Single-tailed 5% point is -1.833.	M1 A1	No ft from here if wrong. Must be minus 1.833 unless absolute values are being compared. No ft from here if wrong. $P(t < -2.167(0)) = 0.0292$.	
	Significant. Seems that Gerry's times have been reduced on average.	A1 A1	ft only c's test statistic. ft only c's test statistic. Conclusion in context to include "average" o.e.	**9**
(iii)	A 5% significance level means that the probability of rejecting H_0 given that it is true is 0.05. Decreasing the significance level would make it less likely that a true H_0 would be rejected. Evidence for rejecting H_0 would need to be stronger.	E1 E1 E1	Or equivalent. Allow answers that relate to the context of the question.	**3**
(iv)	CI is given by $14.987 \pm$	M1	ZERO/4 if not same distribution as test. Same wrong distribution scores maximum M1B0M1A0. Recovery to t_9 is OK.	
	$2 \cdot 262$ $\times \dfrac{0.45675}{\sqrt{10}}$	B1 M1		
	$= 14.987 \pm 0.3267 = (14.66(0), 15.31(3))$	A1	c.a.o. Must be expressed as an interval.	**4**
				18

Q2									

(i)

No. particles	0	1	2	3	4	5
Obs fr	4	7	10	20	17	
Prob'y	0.0150	0.0630	0.1322	0.1852	0.1944	
Exp fr	1.50	6.30	13.22	18.52	19.44	
Contrib to X^2	(4.1667)	(0.0778)	0.7843	0.1183	0.3063	

Combined	11
	7.80
	1.3128

$X^2 = 1.3128 + 0.7843 + 0.1183 + 0.3063 + 0.1083 +$ $\quad 0.1813 + 0.6676 + 0.4056$ $\quad = 3.884(5)$	M1 M1 A1 M1 M1 A1	Probs correct to 3d.p. or better. $\times\,100$ for expected frequencies. All correct. Merge first 2 cells. Calculation of X^2. c.a.o. (For ungrouped cells $X^2 = 6.816$.)
H_0: The Poisson model fits the data. H_1: The Poisson model does not fit the data.	B1 B1	Ignore any reference to the parameter. Do not accept "data fit model" oe.
Refer to χ_6^2.	M1	Allow correct df (= cells – 2) from wrongly grouped table and ft. Otherwise, no ft if wrong.
Upper 10% point is 10.64.	A1	No ft from here if wrong. ($\chi_7^2 = 12.02$) $P(X^2 > 3.884) = 0.6924$.
Not significant. Evidence suggests that the model fits the data.	A1 A1	ft only c's test statistic. ft only c's test statistic. Do not accept "data fit model" oe.

12

(ii)

H_0: $m = 15$ $\qquad$ H_1: $m > 15$ where m is the population median diameter (in μm).	B1 B1	Both. Accept hypotheses in words. Adequate definition of m to include "population".
Given $W_- = 53$ ($\therefore W_+ = 157$)		
Refer to tables of Wilcoxon paired (/single sample) statistic for $n = 20$. Lower 5% point is 60 (or upper is 150 if W_+ used).	M1 A1	No ft from here if wrong. i.e. a 1-tail test. No ft from here if wrong.
Result is significant. Evidence suggests that the median diameter appears to be more than 15 μm.	A1 A1	ft only c's test statistic. ft only c's test statistic. Conclusion in context to include "average" o.e.

6

18

Q3			
(i) (A)		M1 A1 A1	Increasing curve, through (0, 0), in first quadrant only. Asymptotic behaviour. Asymptote labelled; condone absence of axis labels. **3**
(B)	For the UQ G(u) = 0.75 $\therefore \left(1 + \dfrac{u}{200}\right)^{-2} = \dfrac{1}{4}$ $\therefore u = 200$ For the LQ G(l) = 0.25 $\therefore \left(1 + \dfrac{l}{200}\right)^{-2} = \dfrac{3}{4}$ $\therefore l = 200\left(\dfrac{2}{\sqrt{3}} - 1\right) = 30.94...$ $\therefore$ IQR = 200 − 30.94 = 169(.06) For an outlier x > UQ +1.5 × IQR = 200 + 1.5 × 169 =453(.58) ≈ 454 (nearest hour)	M1 A1 A1 M1 M1 E1	Use of G(x) for either quartile. c.a.o. c.a.o. UQ − LQ UQ +1.5 × IQR. Answer given; must be obtained genuinely. **6**
(ii) (A)	$F(x) = \int_0^x \dfrac{1}{200} e^{\frac{-t}{200}} \, dt$ $= \left[-e^{\frac{-t}{200}} \right]_0^x = \left(-e^{\frac{-x}{200}} \right) - \left(-e^{\frac{-0}{200}} \right) = 1 - e^{\frac{-x}{200}}$	M1 A1 E1	Correct integral, including limits (which may be implied subsequently). Correctly integrated. Limits used. Answer given; must be shown convincingly. Condone the omission of x < 0 part. Allow use of "+ c" with F(0) = 0. **3**
(B)	P(X > 50) = 1 − F(50) $= e^{\frac{-50}{200}} = e^{-0.25}$	M1 E1	Use of 1 − F(x) Answer given: must be convincing. (= 0.7788(0)) **2**
(C)	P(X > 400) $= e^{\frac{-400}{200}} = 0.1353(35)$ P(X > 450) $= e^{\frac{-450}{200}} = 0.1053(99)$ $P(X > 450 \mid X > 400) = \dfrac{P(X > 450)}{P(X > 400)}$ $= \dfrac{e^{\frac{-450}{200}}}{e^{\frac{-400}{200}}} = e^{\frac{-50}{200}} = e^{-0.25} \left(= 0.7788\right)$	B1 B1 M1 A1	Accept any form. Accept any form. Conditional probability. Not P(X > 50) × P(X > 400) unless <u>clearly</u> justified. Accept division of decimals, 3dp or better. Accept a.w.r.t. 0.778 or 0.779. **4**
			18

Q4	$C \sim$ N(10, 0.4^2), $\quad D \sim$ N(35, 3.5^2) 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($C < 9.5$) = $P\left(Z < \dfrac{9.5 - 10}{0.4} = -1.25 \right)$ $= 1 - 0.8944 = 0.1056$	M1 A1 A1	For standardising. Award once, here or elsewhere. c.a.o.	3
(ii)	$D - S = D - (C_1 + C_2 + C_3 + C_4) \sim$ N(-5, $\sigma^2 = 3.5^2 + (0.4^2 + 0.4^2 + 0.4^2 + 0.4^2) = 12.89$) Want P($D > S$) = P($D - S > 0$) $= 1 - \Phi\left(\dfrac{0 - (-5)}{3.59} = 1.39(27) \right)$ $= 1 - 0.9182 = 0.0818$	B1 B1 M1 A1	Mean. Accept +5 for $S - D$. Variance. Accept sd (= 3.590…). Formulation of requirement. Accept $S - D < 0$. This mark could be awarded in (iii) if not earned here. c.a.o.	4
(iii)	*New* ($D - S$) = ($D \times 1.3$) $- (C_1 + ... + C_5) \sim$ N(-4.5, $\sigma^2 = (3.5^2 \times 1.3^2) + (0.4^2 + ... + 0.4^2) = 21.5025$) Again want P($D > S$) = P($D - S > 0$) $= 1 - \Phi\left(\dfrac{0 - (-4.5)}{4.637} = 0.9704 \right)$ $= 1 - 0.8341 = 0.1659$	B1 M1 A1 A1	Mean. Accept +4.5 for $S - D$. Correct use of $\times 1.3^2$ for variance. c.a.o. Accept sd (= 4.637…) Or $S - D < 0$. M1 for formulation in (ii) available here. c.a.o.	4

(iv)	CI is given by 9.73 $\pm$
	 1.96
	
	 $\times \dfrac{0.4}{\sqrt{12}}$
	 $= 9.73 \pm 0.2263 = (9.50(37), 9.95(63))$
	
	 Since 10 lies above this interval, it seems that the cheeses are underweight.
	
	 In repeated sampling,
	  95% of all confidence intervals
	  constructed in this way will contain the true mean.
		M1
		 B1
		 M1
		
		 A1
		
		 E1
		
		 E1
		 E1
			
			 1.96 seen.
			
			 c.a.o. Must be expressed as an interval.
			
			 Ft c's interval.
				7

				18

4768: Statistics 3

General Comments

There were 462 candidates from 88 centres (June 2010: 428 from 83) for this sitting of the paper. The overall standard of many of the scripts seen remains high. Once again it was pleasing to note that most candidates remembered to state the hypotheses (Questions 1 and 2) despite not been instructed to do so.

There was a noticeable improvement in respect of the use of clear and accurate notation, concise and accurate computation and the overall presentation of the work. However, there were two parts in Question 3 that required the recall and application of topics from Statistics 1 that were rather badly answered.

Invariably all four questions were attempted. There was no evidence to suggest that candidates found themselves unable to complete the paper in the available time.

Comments on Individual Questions

1 (i)   Most candidates knew exactly what was required here. It helps to clarify matters if they can be explicit about whether they are referring to population or sample properties.

 (ii)   There was much good work seen here. In particular, candidates remembered to state their hypotheses correctly and they appeared to be making good, efficient use of their calculators in obtaining the test statistic. There was just one widespread issue: the final conclusion, in context, is expected to include an explicit reference to "on average" or equivalent.

 (iii)   This part was not at all well answered. Few candidates appeared to have a clear and correct understanding of the meaning of a significance level.

 (iv)   On the whole this part was usually answered correctly, although there was the familiar problem of remembering to remain with the *t* distribution for the relevant percentage point in the construction of the confidence interval.

2   There were many good and fully or nearly fully correct solutions to this question.
 (i)   There was a noticeable improvement in the statement of hypotheses – there were far fewer instances of "the data fit the model" (or equivalent) than in the past. The expected frequencies were almost always found correctly, but then many candidates forgot to combine the first two classes and so ended up with the wrong test statistic. There was less success when it came to the critical value. Despite being told in the question that the mean for the model had been obtained from the data, some candidates forgot to make allowance for that in their number of degrees of freedom.

 (ii)   Here too there was a pleasing improvement in the way that the hypotheses were stated, though there still seems to be a reluctance by some to use the symbol "*m*" for the median and then define it as the <u>population</u> median. The subsequent test was carried out with little difficulty by most candidates.

3   This question seemed to cause the most problems for candidates, and yet the issues over which they struggled were largely to do with topics from Statistics 1 and with solving quadratic equations. Those candidates who could recall the relevant facts were able to obtain the required results quickly and easily and were well rewarded as a consequence.

(i) (A)   Sketches were fairly poor. Many simply drew a straight line from the origin to some (unspecified) point on $y = 1$. There was often no attempt to show the asymptotic behaviour of G(x).

  (B)   Candidates made very heavy weather of this part and clearly spent a lot of time making little, if any, progress. Time was wasted in finding the median (unnecessarily) and in trying to integrate the CDF. It was surprising how many seemed to have forgotten about quartiles and interquartile range. To find each quartile it was necessary to solve a quadratic equation which, with a little thought, could be done quickly and easily. Many just pressed ahead regardless and often ended up making things more complicated for themselves. Candidates were then expected to apply the well-known criteria for outliers in sample data. A very large proportion did not appear to have remembered the rule of thumb that outliers are more than $1.5 \times$ the interquartile range from the nearer quartile. "Median + 2 × the interquartile range" was a very common, but incorrect, alternative to this.

(ii) (A)   Generally, this was answered satisfactorily by most candidates, though some may not have fared so well had they not known the result they were aiming for. Notation was often sloppy.

  (B)   As in the previous part, the answer, given in the question, may well have helped some candidates, but having said that the vast majority of responses were perfectly adequate.

  (C)   This part seemed to confound most candidates. Many had forgotten how to find a conditional probability. Furthermore, they overlooked the clues, given by the structure of the question, about how they might find P($X > 400$) and P($X > 450$).

4 (i)   This part was almost invariably answered correctly.

  (ii)   Many candidates tackled this part well and successfully. However, it was somewhat surprising to see just how many candidates did not think to combine the distributions for supply and demand: they attempted to find the probability that the demand exceeded 40 kg. Thus for these candidates this part was trivialised with costly consequences.

  (iii)   The same was true in this part. There were many correct solutions but those whose attempts were flawed in part (ii) repeated the same mistake here. A further complication was that some candidates forgot to increase the number of cheeses supplied to 5.

  (iv)   There was much correct work in the construction of the confidence interval, although there were many candidates who appeared not to trust the information given in the question about the population standard deviation and so used the percentage point 2.201 from t_{11} instead of 1.96.

  A large proportion of candidates neglected to make any comment about the interval. Explanations of the meaning of "a 95% confidence interval" were fairly mixed in quality.

GCE Mathematics (MEI)

		Max Mark	a	b	c	d	e	u
4751/01 (C1) MEI Introduction to Advanced Mathematics	Raw	72	55	49	43	37	32	0
	UMS	100	80	70	60	50	40	0
4752/01 (C2) MEI Concepts for Advanced Mathematics	Raw	72	53	46	39	33	27	0
	UMS	100	80	70	60	50	40	0
4753/01 (C3) MEI Methods for Advanced Mathematics with Coursework: Written Paper	Raw	72	54	48	42	36	29	0
4753/02 (C3) MEI Methods for Advanced Mathematics with Coursework: Coursework	Raw	18	15	13	11	9	8	0
4753/82 (C3) MEI Methods for Advanced Mathematics with Coursework: Carried Forward Coursework Mark	Raw	18	15	13	11	9	8	0
4753 (C3) MEI Methods for Advanced Mathematics with Coursework	UMS	100	80	70	60	50	40	0
4754/01 (C4) MEI Applications of Advanced Mathematics	Raw	90	63	56	50	44	38	0
	UMS	100	80	70	60	50	40	0
4755/01 (FP1) MEI Further Concepts for Advanced Mathematics	Raw	72	59	52	45	39	33	0
	UMS	100	80	70	60	50	40	0
4756/01 (FP2) MEI Further Methods for Advanced Mathematics	Raw	72	55	48	41	34	27	0
	UMS	100	80	70	60	50	40	0
4757/01 (FP3) MEI Further Applications of Advanced Mathematics	Raw	72	55	48	42	36	30	0
	UMS	100	80	70	60	50	40	0
4758/01 (DE) MEI Differential Equations with Coursework: Written Paper	Raw	72	63	57	51	45	39	0
4758/02 (DE) MEI Differential Equations with Coursework: Coursework	Raw	18	15	13	11	9	8	0
4758/82 (DE) MEI Differential Equations with Coursework: Carried Forward Coursework Mark	Raw	18	15	13	11	9	8	0
4758 (DE) MEI Differential Equations with Coursework	UMS	100	80	70	60	50	40	0
4761/01 (M1) MEI Mechanics 1	Raw	72	60	52	44	36	28	0
	UMS	100	80	70	60	50	40	0
4762/01 (M2) MEI Mechanics 2	Raw	72	64	57	51	45	39	0
	UMS	100	80	70	60	50	40	0
4763/01 (M3) MEI Mechanics 3	Raw	72	59	51	43	35	27	0
	UMS	100	80	70	60	50	40	0
4764/01 (M4) MEI Mechanics 4	Raw	72	54	47	40	33	26	0
	UMS	100	80	70	60	50	40	u
4766/01 (S1) MEI Statistics 1	Raw	72	53	45	38	31	24	0
	UMS	100	80	70	60	50	40	0
4767/01 (S2) MEI Statistics 2	Raw	72	60	53	46	39	33	0
	UMS	100	80	70	60	50	40	0
4768/01 (S3) MEI Statistics 3	Raw	72	56	49	42	35	28	0
	UMS	100	80	70	60	50	40	0
4769/01 (S4) MEI Statistics 4	Raw	72	56	49	42	35	28	0
	UMS	100	80	70	60	50	40	0
4771/01 (D1) MEI Decision Mathematics 1	Raw	72	51	45	39	33	27	0
	UMS	100	80	70	60	50	40	0
4772/01 (D2) MEI Decision Mathematics 2	Raw	72	58	53	48	43	39	0
	UMS	100	80	70	60	50	40	0
4773/01 (DC) MEI Decision Mathematics Computation	Raw	72	46	40	34	29	24	0
	UMS	100	80	70	60	50	40	0
4776/01 (NM) MEI Numerical Methods with Coursework: Written Paper	Raw	72	62	55	49	43	36	0
4776/02 (NM) MEI Numerical Methods with Coursework: Coursework	Raw	18	14	12	10	8	7	0
4776/82 (NM) MEI Numerical Methods with Coursework: Carried Forward Coursework Mark	Raw	18	14	12	10	8	7	0
4776 (NM) MEI Numerical Methods with Coursework	UMS	100	80	70	60	50	40	0
4777/01 (NC) MEI Numerical Computation	Raw	72	55	47	39	32	25	0
	UMS	100	80	70	60	50	40	0