

**ADVANCED SUBSIDIARY GCE
MATHEMATICS (MEI)**

4766/01

Statistics 1

TUESDAY 15 JANUARY 2008

Morning
Time: 1 hour 30 minutes

Additional materials: Answer Booklet (8 pages)
Graph paper
MEI Examination Formulae and Tables (MF2)

INSTRUCTIONS TO CANDIDATES

- Write your name in capital letters, your Centre Number and Candidate Number in the spaces provided on the Answer Booklet.
- Read each question carefully and make sure you know what you have to do before starting your answer.
- Answer **all** the questions.
- You are permitted to use a 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.
- The total number of marks for this paper is 72.
- 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.

This document consists of **4** printed pages.

Section A (36 marks)

- 1 Alice carries out a survey of the 28 students in her class to find how many text messages each sent on the previous day. Her results are shown in the stem and leaf diagram.

0	0	0	1	1	3	5	7	7	7	8	8
1	0	1	2	3	3	4	4	6	9		
2	0	1	3	3	7						
3	5	7									
4											
5	8										

Key: 2 | 3 represents 23

- (i) Find the mode and median of the number of text messages. [2]
- (ii) Identify the type of skewness of the distribution. [1]
- (iii) Alice is considering whether to use the mean or the median as a measure of central tendency for these data.
- (A) In view of the skewness of the distribution, state whether Alice should choose the mean or the median. [1]
- (B) What other feature of the distribution confirms Alice's choice? [1]
- (iv) The mean number of text messages is 14.75. If each message costs 10 pence, find the total cost of all of these messages. [2]
- 2 Codes of three letters are made up using only the letters A, C, T, G. Find how many different codes are possible
- (i) if all three letters used must be different, [3]
- (ii) if letters may be repeated. [2]
- 3 Steve is going on holiday. The probability that he is delayed on his outward flight is 0.3. The probability that he is delayed on his return flight is 0.2, independently of whether or not he is delayed on the outward flight.
- (i) Find the probability that Steve is delayed on his outward flight but not on his return flight. [2]
- (ii) Find the probability that he is delayed on at least one of the two flights. [3]
- (iii) Given that he is delayed on at least one flight, find the probability that he is delayed on both flights. [3]

- 4 A company is searching for oil reserves. The company has purchased the rights to make test drillings at four sites. It investigates these sites one at a time but, if oil is found, it does not proceed to any further sites. At each site, there is probability 0.2 of finding oil, independently of all other sites.

The random variable X represents the number of sites investigated. The probability distribution of X is shown below.

r	1	2	3	4
$P(X = r)$	0.2	0.16	0.128	0.512

- (i) Find the expectation and variance of X . [5]
- (ii) It costs £45 000 to investigate each site. Find the expected total cost of the investigation. [1]
- (iii) Draw a suitable diagram to illustrate the distribution of X . [2]
- 5 Sophie and James are having a tennis competition. The winner of the competition is the first to win 2 matches in a row. If the competition has not been decided after 5 matches, then the player who has won more matches is declared the winner of the competition.

For example, the following sequences are two ways in which Sophie could win the competition. (S represents a match won by Sophie; J represents a match won by James.)

SJSS SJSJS

- (i) Explain why the sequence **SSJ** is not possible. [1]
- (ii) Write down the other three possible sequences in which Sophie wins the competition. [3]
- (iii) The probability that Sophie wins a match is 0.7. Find the probability that she wins the competition in no more than 4 matches. [4]

Section B (36 marks)

- 6 The maximum temperatures x degrees Celsius recorded during each month of 2005 in Cambridge are given in the table below.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
9.2	7.1	10.7	14.2	16.6	21.8	22.0	22.6	21.1	17.4	10.1	7.8

These data are summarised by $n = 12$, $\Sigma x = 180.6$, $\Sigma x^2 = 3107.56$.

- (i) Calculate the mean and standard deviation of the data. [3]
- (ii) Determine whether there are any outliers. [3]
- (iii) The formula $y = 1.8x + 32$ is used to convert degrees Celsius to degrees Fahrenheit. Find the mean and standard deviation of the 2005 maximum temperatures in degrees Fahrenheit. [3]
- (iv) In New York, the monthly maximum temperatures are recorded in degrees Fahrenheit. In 2005 the mean was 63.7 and the standard deviation was 16.0. Briefly compare the maximum monthly temperatures in Cambridge and New York in 2005. [2]

The total numbers of hours of sunshine recorded in Cambridge during the month of January for each of the last 48 years are summarised below.

Hours h	$70 \leq h < 100$	$100 \leq h < 110$	$110 \leq h < 120$	$120 \leq h < 150$	$150 \leq h < 170$	$170 \leq h < 190$
Number of years	6	8	10	11	10	3

- (v) Draw a cumulative frequency graph for these data. [5]
- (vi) Use your graph to estimate the 90th percentile. [2]
- 7 A particular product is made from human blood given by donors. The product is stored in bags. The production process is such that, on average, 5% of bags are faulty. Each bag is carefully tested before use.
- (i) 12 bags are selected at random.
- (A) Find the probability that exactly one bag is faulty. [3]
- (B) Find the probability that at least two bags are faulty. [2]
- (C) Find the expected number of faulty bags in the sample. [2]
- (ii) A random sample of n bags is selected. The production manager wishes there to be a probability of one third or less of finding any faulty bags in the sample. Find the maximum possible value of n , showing your working clearly. [3]
- (iii) A scientist believes that a new production process will reduce the proportion of faulty bags. A random sample of 60 bags made using the new process is checked and one bag is found to be faulty. Write down suitable hypotheses and carry out a hypothesis test at the 10% level to determine whether there is evidence to suggest that the scientist is correct. [8]

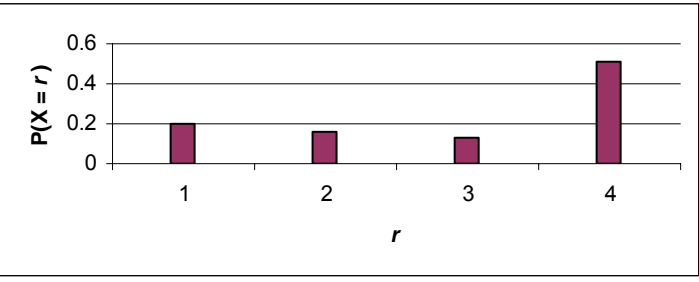
Permission to reproduce items where third-party owned material protected by copyright is included has been sought and cleared where possible. Every reasonable effort has been made by the publisher (OCR) to trace copyright holders, but if any items requiring clearance have unwittingly been included, the publisher will be pleased to make amends at the earliest possible opportunity.

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.

4766

Statistics 1

Q1 (i)	Mode = 7 Median = 12.5	B1 cao B1 cao	2
(ii)	Positive or positively skewed	E1	1
(iii)	(A) Median (B) There is a large outlier or possible outlier of 58 / figure of 58. Just 'outlier' on its own without reference to either 58 or large scores E0 Accept the large outlier affects the mean (more) E1	E1 cao E1indep	2
(iv)	There are $14.75 \times 28 = 413$ messages So total cost = 413×10 pence = £41.30	M1 for 14.75×28 but 413 can also imply the mark A1cao	2
		TOTAL	7
Q2 (i)	$\binom{4}{3} \times 3! = 4 \times 6 = 24$ codes or ${}^4P_3 = 24$ (M2 for 4P_3) Or $4 \times 3 \times 2 = 24$	M1 for 4 M1 for $\times 6$ A1	3
(ii)	$4^3 = 64$ codes	M1 for 4^3 A1 cao	2
		TOTAL	5
Q3 (i)	Probability = $0.3 \times 0.8 = 0.24$	M1 for 0.8 from (1-0.2) A1	2
(ii)	Either: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $= 0.3 + 0.2 - 0.3 \times 0.2$ $= 0.5 - 0.06 = 0.44$ Or: $P(A \cup B) = 0.7 \times 0.2 + 0.3 \times 0.8 + 0.3 \times 0.2$ $= 0.14 + 0.24 + 0.06 = 0.44$ Or: $P(A \cup B) = 1 - P(A' \cap B')$ $= 1 - 0.7 \times 0.8 = 1 - 0.56 = 0.44$	M1 for adding 0.3 and 0.2 M1 for subtraction of (0.3 \times 0.2) A1 cao M1 either of first terms M1 for last term A1 M1 for 0.7 \times 0.8 or 0.56 M1 for complete method as seen A1	3
(iii)	$P(A B) = \frac{P(A \cap B)}{P(B)} = \frac{0.06}{0.44} = \frac{6}{44} = 0.136$	M1 for numerator of their 0.06 only M1 for 'their 0.44' in denominator A1 FT (must be valid p)	3
		TOTAL	8

Q4 (i)	$E(X) = 1 \times 0.2 + 2 \times 0.16 + 3 \times 0.128 + 4 \times 0.512 = 2.952$ Division by 4 or other spurious value at end loses A mark $E(X^2) = 1 \times 0.2 + 4 \times 0.16 + 9 \times 0.128 + 16 \times 0.512 = 10.184$ $\text{Var}(X) = 10.184 - 2.952^2 = 1.47 \text{ (to 3 s.f.)}$	M1 for $\sum rp$ (at least 3 terms correct) A1 cao M1 for $\sum x^2p$ at least 3 terms correct M1 for $E(X^2) - E(X)^2$ Provided ans > 0 A1 FT their $E(X)$ but not a wrong $E(X^2)$	5
(ii)	Expected cost = $2.952 \times \text{£}45000 = \text{£}133000$ (3sf)	B1 FT (no extra multiples / divisors introduced at this stage)	1
(iii)		G1 labelled linear scales G1 height of lines	2
		TOTAL	8
Q5 (i)	Impossible because the competition would have finished as soon as Sophie had won the first 2 matches	E1	1
(ii)	SS, JSS, JSJSS	B1, B1, B1 (-1 each error or omission)	3
(iii)	$0.7^2 + 0.3 \times 0.7^2 + 0.7 \times 0.3 \times 0.7^2 = 0.7399 \text{ or } 0.74(0)$ $\{ 0.49 + 0.147 + 0.1029 = 0.7399 \}$	M1 for any correct term M1 for any other correct term M1 for sum of all three correct terms A1 cao	4
		TOTAL	8

Section B																			
Q6 (i)	$\text{Mean} = \frac{180.6}{12} = 15.05 \text{ or } 15.1$ $S_{xx} = 3107.56 - \frac{180.6^2}{12} \text{ or } 3107.56 - 12(\text{their } 15.05)^2 = (389.53)$ $s = \sqrt{\frac{389.53}{11}} = 5.95 \text{ or better}$ NB Accept answers seen without working (from calculator)	B1 for mean M1 for attempt at S_{xx} A1 cao	3																
(ii)	$\bar{x} + 2s = 15.05 + 2 \times 5.95 = 26.95$ $\bar{x} - 2s = 15.05 - 2 \times 5.95 = 3.15$ So no outliers	M1 for attempt at either M1 for both A1 for limits and conclusion FT their mean and sd	3																
(iii)	New mean = $1.8 \times 15.05 + 32 = 59.1$ New s = $1.8 \times 5.95 = 10.7$	B1FT M1 A1FT	3																
(iv)	New York has a higher mean or 'is on average' higher (oe) New York has greater spread /range /variation or SD (oe)	E1FT using $^{\circ}F$ (\bar{x} dep) E1FT using $^{\circ}F$ (σ dep)	2																
(v)	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tr> <td>Upper bound</td> <td>(70)</td> <td>100</td> <td>110</td> <td>120</td> <td>150</td> <td>170</td> <td>190</td> </tr> <tr> <td>Cumulative frequency</td> <td>(0)</td> <td>6</td> <td>14</td> <td>24</td> <td>35</td> <td>45</td> <td>48</td> </tr> </table> 	Upper bound	(70)	100	110	120	150	170	190	Cumulative frequency	(0)	6	14	24	35	45	48	B1 for all correct cumulative frequencies (may be implied from graph). Ignore cf of 0 at this stage G1 for linear scales (linear from 70 to 190) ignore $x < 70$ vertical: 0 to 50 but not beyond 100 (no inequality scales) G1 for labels G1 for points plotted as (UCB, their cf). <u>Ignore (70,0)</u> at this stage. No mid – point or LCB plots.	5
Upper bound	(70)	100	110	120	150	170	190												
Cumulative frequency	(0)	6	14	24	35	45	48												
(vi)	NB all G marks dep on attempt at cumulative frequencies. NB All G marks dep on attempt at cumulative frequencies Line on graph at cf = 43.2(soi) or used 90th percentile = 166	G1 for joining all of 'their points'(line or smooth curve) AND now including (70,0) M1 for use of 43.2 A1FT but dep on 3rd G mark earned	2																
		TOTAL	18																

<p>Q7 (i)</p>	<p>$X \sim B(12, 0.05)$</p> <p>(A) $P(X = 1) = \binom{12}{1} \times 0.05 \times 0.95^{11} = 0.3413$</p> <p>OR from tables $0.8816 - 0.5404 = 0.3412$</p> <p>(B) $P(X \geq 2) = 1 - 0.8816 = 0.1184$</p> <p>(C) Expected number $E(X) = np = 12 \times 0.05 = 0.6$</p>	<p>M1 0.05×0.95^{11}</p> <p>M1 $\binom{12}{1} \times pq^{11} (p+q) = 1$</p> <p>A1 cao</p> <p>OR: M1 for 0.8816 seen and M1 for subtraction of 0.5404</p> <p>A1 cao</p> <p>M1 for $1 - P(X \leq 1)$</p> <p>A1 cao</p> <p>M1 for 12×0.05</p> <p>A1 cao (= 0.6 seen)</p>	<p>3</p> <p>2</p> <p>2</p>
<p>(ii)</p> <p>(iii)</p>	<p><i>Either:</i> $1 - 0.95^n \leq \frac{1}{3}$ $0.95^n \geq \frac{2}{3}$ $n \leq \log \frac{2}{3} / \log 0.95$, so $n \leq 7.90$ Maximum $n = 7$</p> <p><i>Or:</i> (using tables with $p = 0.05$): $n = 7$ leads to $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.6983 = 0.3017 (< \frac{1}{3})$ or $0.6983 (> \frac{2}{3})$ $n = 8$ leads to $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.6634 = 0.3366 (> \frac{1}{3})$ or $0.6634 (< \frac{2}{3})$ Maximum $n = 7$ (total accuracy needed for tables)</p> <p><i>Or:</i> (using trial and improvement): $1 - 0.95^7 = 0.3017 (< \frac{1}{3})$ or $0.95^7 = 0.6983 (> \frac{2}{3})$ $1 - 0.95^8 = 0.3366 (> \frac{1}{3})$ or $0.95^8 = 0.6634 (< \frac{2}{3})$ Maximum $n = 7$ (3 sf accuracy for calculations)</p> <p>NOTE: $n = 7$ unsupported scores SC1 only</p> <p>Let $X \sim B(60, p)$ Let p = probability of a bag being faulty $H_0: p = 0.05$ $H_1: p < 0.05$</p> <p>$P(X \leq 1) = 0.95^{60} + 60 \times 0.05 \times 0.95^{59} = 0.1916 > 10\%$</p> <p>So not enough evidence to reject H_0</p> <p>Conclude that there is not enough evidence to indicate that the new process reduces the failure rate or scientist incorrect/wrong.</p>	<p>M1 for equation in n</p> <p>M1 for use of logs</p> <p>A1 cao</p> <p>M1 indep</p> <p>M1 indep</p> <p>A1 cao dep on both M's</p> <p>M1 indep (as above)</p> <p>M1 indep (as above)</p> <p>A1 cao dep on both M's</p> <p>B1 for definition of p</p> <p>B1 for H_0</p> <p>B1 for H_1</p> <p>M1 A1 for probability</p> <p>M1 for comparison</p> <p>A1</p> <p>E1</p>	<p>3</p> <p>8</p>
TOTAL			18

4766: Statistics 1 (G241 Z1)

General Comments

The level of difficulty and accessibility of the paper seemed appropriate for the majority of candidates. The range of marks was fairly wide although some centres had few high scoring scripts. There were a small number of very weak scripts mainly restricted to those centres with a large number of candidates.

Almost all candidates were able to score some marks throughout the paper although there remain a significant minority who seem unprepared for questions at this level. The better scripts produced answers which were very well presented with methods and working clear. Arithmetic accuracy was generally good although there was little appreciation of the consequences of using rounded answers in subsequent calculations. Some weaker candidates were reluctant to provide reasons to support answers, thus losing valuable marks when a wrong answer appeared.

It does appear that not all centres had covered the specification in sufficient depth or detail as an occasional topic was poorly answered by a majority of the candidates of that centre. Hypothesis testing remains an example of this. Common errors included the use of point probabilities in hypothesis testing, the failure to define the parameter, p explicitly before trying to establish the hypotheses and the lack of full logical reasoning in coming to the final conclusions.

Comments on Individual Questions

Section A

- 1) A very mixed set of answers except from very good candidates. Almost all candidates identified the mode correctly as 7 or sometimes as 07 but many made errors with the median with 13 or 2.5 (forgetting to add on the stem value of 10) being common mistakes. Some just calculated the location of the median $(28 + 1)/2 = 14.5$, believing this was the median value.

A significant minority of candidates thought that the skewness of the distribution was negative. Most selected the median as the appropriate measure of central tendency although in (iii) part B several candidates referred to an outlier but did not specify what the outlier was, or whether it was a large or small value, preferring to state that the distribution was bimodal, unimodal or had a large range.

Many attempts at the total cost of the messages (even amongst better candidates) failed because of a reluctance to multiply by 28 with popular answers being £1.48 or £1.50. Some omitted the units altogether giving an answer of 4130 whilst a couple of scripts contrived to have a daily mobile text bill of £4130.
- 2) This question produced the weakest response overall by a wide margin with full marks being scored very rarely. Answers of 4 for part (i) followed by 24 for part (ii) were very frequent. Other errors seen included $3!$, 4C_3 or some multiple of 24 in part (i); 4^4 , 4P_3 , or some multiple of 4P_3 , ${}^{12}P_3$ and ${}^{12}C_3$ in part (ii).
- 3) (i) Virtually all candidates obtained 0.24 as their answer and scored two marks.

Report on the Units taken in January 2008

- (ii) Answers to this part were much less successful with a large number of candidates giving an answer of $0.3 \times 0.8 + 0.2 \times 0.7 = 0.38$, forgetting about the 'both' term of 0.06. Other common wrong answers were $0.3 + 0.2 = 0.5$ and $0.24 + 0.14 + 0.56 = 0.94$ although both were much less frequent than 0.38.
 - (iii) There were many correct attempts at the conditional probability although the usual error of $(0.06 \times 0.44)/0.44$ was often seen. A small number of candidates quoted a formula for conditional probability correctly but were then confused by the terms included in the formula often resulting in multiplication of 0.06 by 0.44 or similar.
- 4)
- (i) Most of the better candidates scored very highly on this question with their likely source of error being arithmetic or a misread of a probability. Weaker candidates were less successful although usually obtaining a correct answer for $E(X)$. Some candidates still insisted in dividing their $E(X)$ by 4 for which a penalty was incurred. Errors for $\text{Var}(X)$ included a failure to square $E(X)$ or the quoting of an incorrect formula. A few candidates tried to use $\sum p(X - \mu)^2$ often then making an arithmetic error.
 - (ii) Many candidates were confused by the expected total cost often writing 4 x £45000 as their answer.
 - (iii) Most diagrams scored some credit with the most frequent error being a lack of labelling of the axes. A small minority of candidates produced a pie chart or a tree diagram.
- 5)
- This question was very well answered with many candidates scoring full marks.
- (i) Almost all candidates explained why the sequence **SSJ** was impossible.
 - (ii) The majority scored well on this part with the most common error being the omission of **SS**.
 - (iii) Good candidates used the symbolic information given to move directly to the answer of 0.7399; weaker candidates ignored that information and attempted to use Binomial probabilities or a method of subtracting probabilities from 1. The answer of $1 - 0.7^5 = 0.8319$ was common

Section B

- 6)
- Most candidates scored some marks on this question although totally correct answers were rare.
- (i) A small minority divided by 12 (divisor n) thus finding the RMSD. Most knew the method to find outliers; errors included the use of 1.5s and 3s in place of 2s.
 - (iii) Some candidates started from scratch and converted all 12 temperatures to °F before calculating mean and standard deviation often correctly, but then possibly finding difficulty in completing all questions in time. Others found the new mean quickly and correctly but wrote $1.8 \times 5.95 + 32 = 42.7$ for the standard deviation.

- (iv) Comments were usually correct although some candidates made no reference to any mean or average temperature. Some weaker candidates believed they could compare $^{\circ}\text{C}$ with $^{\circ}\text{F}$ without any conversion.
- (v) The cumulative frequency graph, was surprisingly poorly attempted. Only a few managed all 7 marks. The vertical axis was often labelled as frequency or number of years; there was confusion as to how to scale or label the horizontal axis between 70 and 190, the horizontal scale was sometimes shown in intervals as $70 \leq x \leq 100$, $100 \leq x \leq 110$ etc instead of a linear scale. A lot of candidates drew histograms or cumulative frequency histograms or frequency polygons. Some even calculated 'fx' and then calculated a 'cumulative fx'. In drawing the graph the point (70, 0) was often omitted and the cumulative frequency curve was left 'hanging' or was taken back to the origin or some other random point between 0 and 100. In finding the 90th percentile there was use of 90% of 50 or 90% of 190 as a method. Some weaker candidates looked at 90 on the horizontal axis and gave a value from the cumulative frequency axis as their answer. The major error was the use of mid-points rather than the upper class boundaries in plotting the points, an error made by a very large proportion of the candidates. Some candidates, even after they had plotted the correct points, made the fatal error of trying to draw a 'curve of best fit' rather than join their points with a smooth curve.

- 7) Few candidates scored very highly in this question. In part (i), $p(X = 1)$ was usually well answered although a number omitted the ${}^{12}C_1$ term. In part (B) $p(X \geq k)$ was often answered as $1 - P(X \leq k) = 1 - 0.9978 = 0.0022$ instead of $1 - p(X \leq k-1) = 0.1184$ or was omitted by the weaker candidates. There were in general good answers to the expected number of faulty bags although many candidates rounded their answer of 0.6 to 1 and a few thought that the question meant finding the most likely number of faulty bags.

The majority of candidates did not seem to understand what was meant by "finding any faulty bags in the sample". Some thought that it meant no faulty bags leading to $0.95^n < 1/3$; others used the probability of one faulty bag leading to a trial and error method using tables. The few who reached $0.95^n < 2/3$ often then obtained the correct answer of $n = 7$. A very small number of attempts failed because 0.6634 was deemed to be greater than $2/3$ or similar.

The hypothesis test was poorly answered except by the best candidates. Common errors initially included a failure to define the parameter p , the writing of $H_0 = 0.05$, and the use of $p = 1/60$ or even 0.1. In performing the test candidates often wrote $p(X = 1) = 0.1455$, reject H_1 or similar without ever stating a critical region or indicating that they should be considering $p(X \leq 1)$.