OCR
RECOGNISING ACHIEVEMENT

**ADVANCED GCE**

# MATHEMATICS (MEI) 4767

Statistics 2

Candidates answer on the Answer Booklet

**OCR Supplied Materials:**
- 8 page Answer Booklet
- Graph paper
- MEI Examination Formulae and Tables (MF2)

**Other Materials Required:**
None

**Monday 15 June 2009**
**Afternoon**

**Duration:** 1 hour 30 minutes

*4767*

**INSTRUCTIONS TO CANDIDATES**

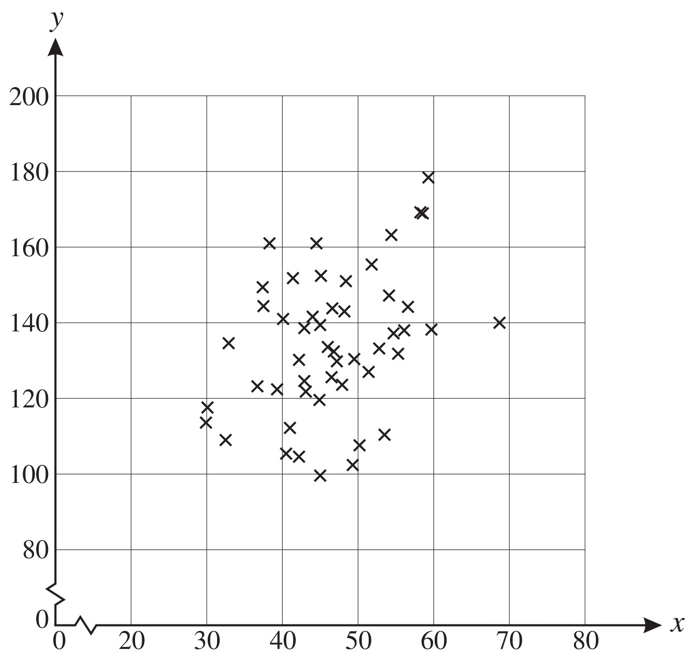
- Write your name clearly in capital letters, your Centre Number and Candidate Number in the spaces provided on the Answer Booklet.
- Use black ink. Pencil may be used for graphs and diagrams only.
- Read each question carefully and make sure that you know what you have to do before starting your answer.
- Answer **all** the questions.
- Do **not** write in the bar codes.
- You are permitted to use a 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.

**1** An investment analyst thinks that there may be correlation between the cost of oil, $x$ dollars per barrel, and the price of a particular share, $y$ pence. The analyst selects 50 days at random and records the values of $x$ and $y$. Summary statistics for these data are shown below, together with a scatter diagram.

$$\Sigma x = 2331.3 \quad \Sigma y = 6724.3 \quad \Sigma x^2 = 111\,984 \quad \Sigma y^2 = 921\,361 \quad \Sigma xy = 316\,345 \quad n = 50$$




**(i)** Calculate the sample product moment correlation coefficient. **[5]**

**(ii)** Carry out a hypothesis test at the 5% significance level to investigate the analyst's belief. State your hypotheses clearly, defining any symbols which you use. **[6]**

**(iii)** An assumption that there is a bivariate Normal distribution is required for this test to be valid. State whether it is the sample or the population which is required to have such a distribution. State, with a reason, whether in this case the assumption appears to be justified. **[3]**

**(iv)** Explain why a 2-tail test is appropriate even though it is clear from the scatter diagram that the sample has a positive correlation coefficient. **[2]**

**2** Jess is watching a shower of meteors (shooting stars). During the shower, she sees meteors at an average rate of 1.3 per minute.

**(i)** State conditions required for a Poisson distribution to be a suitable model for the number of meteors which Jess sees during a randomly selected minute. **[2]**

You may assume that these conditions are satisfied.

**(ii)** Find the probability that, during one minute, Jess sees

($A$) exactly one meteor,

($B$) at least 4 meteors. **[4]**

**(iii)** Find the probability that, in a period of 10 minutes, Jess sees exactly 10 meteors. **[3]**

**(iv)** Use a suitable approximating distribution to find the probability that Jess sees a total of at least 100 meteors during a period of one hour. **[5]**

**(v)** Jess watches the shower for $t$ minutes. She wishes to be at least 99% certain that she will see one or more meteors. Find the smallest possible integer value of $t$. **[5]**

**3** Intensity of light is measured in lumens. The random variable $X$ represents the intensity of the light from a standard 100 watt light bulb. $X$ is Normally distributed with mean 1720 and standard deviation 90. You may assume that the intensities for different bulbs are independent.

**(i)** Show that $\text{P}(X < 1700) = 0.4121$. **[4]**

**(ii)** These bulbs are sold in packs of 4. Find the probability that the intensities of exactly 2 of the 4 bulbs in a randomly chosen pack are below 1700 lumens. **[3]**

**(iii)** Use a suitable approximating distribution to find the probability that the intensities of at least 20 out of 40 randomly selected bulbs are below 1700 lumens. **[5]**

A manufacturer claims that the average intensity of its 25 watt low energy light bulbs is 1720 lumens. A consumer organisation suspects that the true figure may be lower than this. The intensities of a random sample of 20 of these bulbs are measured. A hypothesis test is then carried out to check the claim.

**(iv)** Write down a suitable null hypothesis and explain briefly why the alternative hypothesis should be $\text{H}_1 : \mu < 1720$. State the meaning of $\mu$. **[3]**

**(v)** Given that the standard deviation of the intensity of such bulbs is 90 lumens and that the mean intensity of the sample of 20 bulbs is 1703 lumens, carry out the test at the 5% significance level. **[5]**

**4** In a traffic survey a random sample of 400 cars passing a particular location during the rush hour is selected. The type of car and the sex of the driver are classified as follows.

| | | Sex | | Row totals |
|---|---|---|---|---|
| | | Male | Female | |
| Type of car | Hatchback | 96 | 36 | 132 |
| | Saloon | 77 | 35 | 112 |
| | People carrier | 38 | 44 | 82 |
| | 4WD | 19 | 8 | 27 |
| | Sports car | 22 | 25 | 47 |
| Column totals | | 252 | 148 | 400 |

**(i)** Carry out a test at the 5% significance level to examine whether there is any association between type of car and sex of driver. State carefully your null and alternative hypotheses. Your working should include a table showing the contributions of each cell to the test statistic. **[12]**

**(ii)** For each type of car, comment briefly on how the number of drivers of each sex compares with what would be expected if there were no association. **[5]**

# 4767 Statistics 2

**Question 1**

| | | | |
|---|---|---|---|
| **(i)** | **EITHER:**<br>$S_{xy} = \Sigma xy - \frac{1}{n}\Sigma x\Sigma y = 316345 - \frac{1}{50} \times 2331.3 \times 6724.3$<br>$= 2817.8$<br>$S_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 111984 - \frac{1}{50} \times 2331.3^2 = 3284.8$<br>$S_{yy} = \Sigma y^2 - \frac{1}{n}(\Sigma y)^2 = 921361 - \frac{1}{50} \times 6724.3^2 = 17036.8$<br>$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{2817.8}{\sqrt{3284.8 \times 17036.8}} = 0.377$<br>**OR:**<br>$\text{cov}(x,y) = \frac{\sum xy}{n} - \bar{x}\bar{y} = 316345/50 - 46.626 \times 134.486$<br>$= 56.356$<br>$\text{rmsd}(x) = \sqrt{\frac{S_{xx}}{n}} = \sqrt{(3284.8/50)} = \sqrt{65.696} = 8.105$<br>$\text{rmsd}(y) = \sqrt{\frac{S_{yy}}{n}} = \sqrt{(17036.8/50)} = \sqrt{340.736} = 18.459$<br>$r = \frac{\text{cov(x,y)}}{rmsd(x)rmsd(y)} = \frac{56.356}{8.105 \times 18.459} = 0.377$ | M1 for method for $S_{xy}$<br>M1 for method for at least one of $S_{xx}$ or $S_{yy}$<br>A1 for at least one of $S_{xy}$, $S_{xx}$ or $S_{yy}$ correct<br>M1 for structure of *r*<br>A1 (AWRT 0.38)<br>M1 for method for cov (*x*,*y*)<br>M1 for method for at least one msd<br>A1 for at least on of cov(*x*,*y*), rmsd(*x*) or rmsd(*y*) correct<br>M1 for structure of *r*<br>A1 (AWRT 0.38) | **5** |
| **(ii)** | $H_0$: $\rho = 0$<br>$H_1$: $\rho \neq 0$ (two-tailed test)<br>where $\rho$ is the population correlation coefficient<br>For *n* = 50, 5% critical value = 0.2787<br>Since 0.377 > 0.2787 we can reject $H_0$:<br>There is sufficient evidence at the 5% level to suggest that there is correlation between oil price and share cost | B1 for $H_0$, $H_1$ in symbols<br>B1 for defining $\rho$<br>B1FT for critical value<br>M1 for sensible comparison leading to a conclusion<br>A1 for result<br>B1 FT for conclusion in context | **6** |
| **(iii)** | Population<br>The scatter diagram has a roughly elliptical shape, hence the assumption is justified. | B1<br>B1 elliptical shape<br>E1 conclusion | **3** |
| **(iv)** | Because the alternative hypothesis should be decided without referring to the sample data and there is no suggestion that the correlation should be positive rather than negative. | E1<br>E1 | **2** |
| | | **TOTAL** | **16** |

**Question 2**

| | | | |
|---|---|---|---|
| **(i)** | Meteors are seen randomly and independently<br>There is a uniform (mean) rate of occurrence of meteor sightings | B1<br>B1 | **2** |
| **(ii)** | (*A*) *Either* P(*X* = 1) = 0.6268 – 0.2725 = 0.3543<br>*Or* P(*X* = 1) = $e^{-1.3}\frac{1.3^1}{1!}$ = 0.3543<br>(*B*) Using tables: P(*X* ≥ 4) = 1 – P(*X* ≤ 3)<br>= 1 – 0.9569<br>= 0.0431 | M1 for appropriate use of tables or calculation<br>A1<br>M1 for appropriate probability calculation<br>A1 | **4** |
| **(iii)** | $\lambda$ = 10 × 1.3 = 13<br>P(*X* = 10) = $e^{-13}\frac{13^{10}}{10!}$ = 0.0859 | B1 for mean<br>M1 for calculation<br>A1 CAO | **3** |
| **(iv)** | Mean no. per hour = 60 × 1.3 = 78<br>Normal approx. to the Poisson, $X \sim N(78, 78)$<br>P(*X* ≥ 100) = $P\left(Z > \frac{99.5-78}{\sqrt{78}}\right)$<br>= P(*Z* >2.434) = 1 – Φ(2.434)<br>= 1 – 0.9926 = 0.0074 | B1 for Normal approx.<br>B1 for correct parameters (SOI)<br>B1 for continuity corr.<br>M1 for correct Normal probability calculation using correct tail<br>A1 CAO, (but FT wrong or omitted CC) | **5** |
| **(v)** | *Either*<br>P(At least one) = 1 - $e^{-\lambda}\frac{\lambda^0}{0!}$ = 1 - $e^{-\lambda}$ ≥ 0.99<br>$e^{-\lambda}$ ≤ 0.01<br>$-\lambda$ ≤ ln 0.01, so $\lambda$ ≥ 4.605<br>1.3 *t* ≥ 4.605, so *t* ≥ 3.54<br>Answer *t* = 4<br>*Or*<br>*t* = 1, $\lambda$ = 1.3, P(At least one) = 1 - $e^{-1.3}$ = 0.7275<br>*t* = 2, $\lambda$ = 2.6, P(At least one) = 1 - $e^{-2.6}$ = 0.9257<br>*t* = 3, $\lambda$ = 3.9, P(At least one) = 1 - $e^{-3.9}$ = 0.9798<br>*t* = 4, $\lambda$ = 5.2, P(At least one) = 1 - $e^{-5.2}$ = 0.9944<br>Answer *t* = 4 | M1 formation of equation/inequality using P(*X* ≥ 1) = 1 – P(X = 0) with Poisson distribution.<br>A1 for correct equation/inequality<br>M1 for logs<br>A1 for 3.54<br>A1 for *t* (correctly justified)<br>M1 at least one trial with any value of *t*<br>A1 correct probability.<br>M1 trial with either *t* = 3 or *t* = 4<br>A1 correct probability of *t* = 3 and *t* = 4<br>A1 for *t* | **5** |
| | | **TOTAL** | **19** |

**Question 3**

| | | | |
|---|---|---|---|
| **(i)** | $X \sim N(1720, 90^2)$<br>$P(X < 1700) = P\left(Z < \frac{1700-1720}{90}\right)$<br>$= P(Z < -0.2222)$<br>$= \Phi(-0.2222) = 1 - \Phi(0.2222)$<br>$= 1 - 0.5879$<br>$= 0.4121$ | M1 for standardising<br>A1<br><br>M1 use of tables (correct tail)<br>A1CAO<br><br>**NB ANSWER GIVEN** | **4** |
| **(ii)** | P(2 of 4 below 1700)<br>$= \binom{4}{2} \times 0.4121^2 \times 0.5879^2 = 0.3522$ | M1 for coefficient<br>M1 for $0.4121^2 \times 0.5879^2$<br>A1 FT (min 2sf) | **3** |
| **(iii)** | Normal approx with<br>$\mu = np = 40 \times 0.4121 = 16.48$<br>$\sigma^2 = npq = 40 \times 0.4121 \times 0.5879 = 9.691$<br><br>$P(X \geq 20) = P\left(Z \geq \frac{19.5-16.48}{\sqrt{9.691}}\right)$<br>$= P(Z \geq 0.9701) = 1 - \Phi(0.9701)$<br>$= 1 - 0.8340 = 0.1660$ | B1<br><br>B1<br>B1 for correct continuity corr.<br><br>M1 for correct Normal probability calculation using correct tail<br>A1 CAO**, (**but FT wrong or omitted CC) | **5** |
| **(iv)** | $H_0$: $\mu = 1720$;<br>$H_1$ is of this form since the consumer organisation suspects that the mean is below 1720<br>$\mu$ denotes the mean intensity of 25 Watt low energy bulbs made by this manufacturer. | B1<br>E1<br><br>B1 for definition of $\mu$ | **3** |
| **(v)** | Test statistic = $\frac{1703-1720}{90/\sqrt{20}} = \frac{-17}{20.12}$<br>$= -0.8447$<br><br>Lower 5% level 1 tailed critical value of $z = -1.645$<br><br>$-0.8447 > -1.645$ so not significant.<br>There is not sufficient evidence to reject $H_0$<br><br>There is insufficient evidence to conclude that the mean intensity of bulbs made by this manufacturer is less than 1720 | M1 must include $\sqrt{20}$<br><br>A1FT<br><br>B1 for –1.645 No FT from here if wrong. Must be –1.645 unless it is clear that absolute values are being used.<br>M1 for sensible comparison leading to a conclusion.<br>FT only candidate's test statistic<br><br>A1 for conclusion in words in context | **5** |
| | | **TOTAL** | **20** |

**Question 4**

| | | | |
|---|---|---|---|
| **(i)** | $H_0$: no association between type of car and sex;<br>$H_1$: some association between type of car and sex; | B1 | |
| | <table><tr><td>EXPECTED</td><td>Male</td><td>Female</td></tr><tr><td>Hatchback</td><td>83.16</td><td>48.84</td></tr><tr><td>Saloon</td><td>70.56</td><td>41.44</td></tr><tr><td>People carrier</td><td>51.66</td><td>30.34</td></tr><tr><td>4WD</td><td>17.01</td><td>9.99</td></tr><tr><td>Sports car</td><td>29.61</td><td>17.39</td></tr></table> | M1 A2 for expected values (to 2 dp)<br><br>(allow A1 for at least one row or column correct) | |
| | <table><tr><td>CONTRIBUTION</td><td>Male</td><td>Female</td></tr><tr><td>Hatchback</td><td>1.98</td><td>3.38</td></tr><tr><td>Saloon</td><td>0.59</td><td>1.00</td></tr><tr><td>People carrier</td><td>3.61</td><td>6.15</td></tr><tr><td>4WD</td><td>0.23</td><td>0.40</td></tr><tr><td>Sports car</td><td>1.96</td><td>3.33</td></tr></table> | M1 for valid attempt at $(O-E)^2/E$<br>A1 for all correct<br>NB These M1A1 marks cannot be implied by a correct final value of $X^2$ | |
| | $X^2 = 22.62$ | M1 for summation<br>A1 for $X^2$ CAO | **12** |
| | Refer to $\chi_4^2$<br>Critical value at 5% level = 9.488 | B1 for 4 deg of f<br>B1 CAO for cv | |
| | 22.62 > 9.488<br>Result is significant<br>There is evidence to suggest that there is some association between sex and type of car.<br><br>NB if $H_0$ $H_1$ reversed, or 'correlation' mentioned, do not award first B1or final A1 | M1 sensible comparison leading to a conclusion<br>A1 | |
| **(ii)** | • In hatchbacks, male drivers are more frequent than expected.<br>• In saloons, male drivers are slightly more frequent than expected.<br>• In people carriers, female drivers are much more frequent than expected.<br>• In 4WDs the numbers are roughly as expected<br>• In sports cars, female drivers are more frequent than expected. | E1<br>E1<br>E1<br>E1<br>E1 | **5** |
| | | **TOTAL** | **17** |

Deleted: ¶

# 4767 Statistics 2

## General Comments

Once again, the general performance for most candidates taking this paper was high. It is pleasing to see continued improvement in the handling of hypothesis tests. One aspect that many candidates seem not to fully grasp is the difference between 'sample' and 'population'; this often leads to loss of marks when stating hypotheses.

## Comments on Individual Questions

### Section A

1) (i) Well answered. A small minority of candidates lost marks through minor slips or by mixing methods e.g. using $S_{xy}$ in the numerator with rmsd(x) and rmsd(y) in the denominator.

(ii) Well answered. In tests involving the product moment correlation coefficient, candidates should be encouraged to write hypotheses in terms of ρ and define ρ as the population correlation coefficient. Most candidates obtained the correct critical value, made a sensible comparison and provided a conclusion in context. Many candidates scored 5 of the available 6 marks; in most cases the lost mark was due to failure to accurately define ρ.

(iii) What should have been an easy mark for stating that the ‘population’ is required to have a bivariate Normal distribution was missed by many candidates. Most candidates picked up the remaining marks for commenting on the elliptical shape required and making a relevant comment regarding the given case.

(iv) Few candidates gained the mark for pointing out that the alternative hypothesis should be decided before referring to the sample data. Most picked up the other available mark. A large number of candidates commented that ‘correlation does not imply causation’, gaining no credit on this occasion.

2) (i) Well answered. Most candidates obtained the mark for explaining that some element of randomness or independence was needed. Candidates should learn to use the phrase ‘uniform mean rate’ in such questions, as other attempts to word this phrase rarely describe what is needed. Fortunately, only a few candidates quoted ‘n is large and p is small’.

(ii) (*A*) Well answered.

(ii) (*B*) Well answered.

(iii) Well answered.

(iv) Most realised that a Normal approximation was appropriate and used the correct parameters. Many candidates failed to apply the correct continuity correction. Otherwise, the handling of the Normal probability calculation was good.

(v) Poorly answered. Many unsuccessful attempts to use an inappropriate Normal approximation were seen. Of the few that managed to proceed as required, many missed the final mark through failing to properly justify their final answers.

3) (i) Mostly well answered. With 4 marks available, candidates were expected to demonstrate how to obtain the given answer; many managed this, but a lot of unconvincing attempts were seen. In some cases, candidates simply wrote 'my calculator tells me this is the answer', or words to that effect; this was not taken as 'sufficient detail'.

(ii) Well answered.

(iii) A similar success rate to question 2 (iv). Again, continuity corrections were frequently omitted or incorrectly applied. Otherwise, the probability calculations were handled well. Common errors include dividing by variance when standardising, and obtaining the wrong 'tail'.

(iv) Well answered, apart from the definition of μ as the population mean.

(v) Well answered. A variety of approaches seen, with many leading to full marks. In most cases, marks were lost for either using the wrong distribution (treating the observed value as a single value rather than the mean of a sample of 20) or by sloppy handling of the comparison of the test statistic with the critical value. Common mistakes involved comparing the test statistic with (commonly) 5%, and comparing a negative test statistic with a positive critical value.

4) (i) Well answered. A small number omitted the context from their hypotheses. Very few mentioned correlation or tried to use parameters in their hypotheses. Most candidates obtained the correct $X^2$ value and provided a table of individual contributions as requested. Most candidates obtained the last 4 marks for carrying out the test using their $X^2$ value.

(ii) Quite well answered. However, it was unclear in many cases whether candidates appreciated the connection between the size of the individual contributions and the strength of the association. Simple comments could score 3 of the 5 available marks quite easily. To gain full credit, candidates needed to display a deeper understanding by interpreting the contributions to the test statistic.