

**Friday 17 June 2016 – Afternoon**

**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.** If additional space is required, you should use the lined page(s) at the end of the Printed Answer Book. The question number(s) must be clearly shown.
- 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.

- 1 A game consists of 20 rounds. Each round is denoted as either a starter, middle or final round. The times taken for each round are independently and Normally distributed with the following parameters (given in seconds).

Type of round	Mean	Standard deviation
Starter	200	15
Middle	220	25
Final	250	20

The game consists of 4 starter, 12 middle and 4 final rounds. Find the probability that

- (i) the mean time per round for the 4 final rounds will exceed 260 seconds, [3]
- (ii) all 20 rounds will be completed in a total time of 75 minutes or less, [5]
- (iii) the 12 middle rounds will take at least 3.5 times as long in total as the 4 starter rounds, [5]
- (iv) the mean time per round for the 12 middle rounds will be at least 25 seconds less than the mean time per round for the 4 final rounds. [5]

- 2 (a) A genetic model involving body colour and eye colour of fruit flies predicts that offspring will consist of four phenotypes in the ratio 9:3:3:1.

A random sample of 200 such offspring is taken. Their phenotypes are found to be as follows.

Phenotype	Brown body Red eye	Brown body Brown eye	Black body Red eye	Black body Brown eye
Frequency	125	37	32	6
Relative proportion from model	9	3	3	1

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

- (b) The median length of European fruit flies is 2.5 mm. South American fruit flies are believed to be larger than European fruit flies. A random sample of 12 South American fruit flies is taken. The flies are found to have the following lengths (in mm).

1.7   1.4   3.1   3.5   3.8   4.2   2.2   2.9   4.4   2.6   3.9   3.2

Carry out a Wilcoxon signed rank test, using a 5% level of significance, to test this belief. [9]

- 3 The random variable  $X$  has the following probability density function:

$$f(x) = \begin{cases} k(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

where  $k$  is a positive constant.

- (i) Calculate the value of  $k$ . [3]
  - (ii) Sketch the probability density function. [3]
  - (iii) Calculate  $\text{Var}(X)$ . [3]
  - (iv) Find a cubic equation satisfied by the upper quartile  $q$ , and hence verify that  $q = 0.35$  to 2 decimal places. [5]
  - (v) A random sample of 40 values of  $X$  is taken. Using a suitable approximating distribution, calculate the probability that the mean of these values is greater than 0.125. Justify your choice of distribution. [4]
- 4 An insurance company is investigating a new system designed to reduce the average time taken to process claim forms. The company has decided to use 10 experienced employees to process claims using the old system and the new system.

Two procedures for comparing the systems are proposed.

*Procedure A* There are two sets of claim forms, set 1 and set 2. Each contains the same number of forms. Each employee processes set 1 on the old system and set 2 on the new system. The times taken are compared.

*Procedure B* There is just one set of claim forms which each employee processes firstly on the old system and then on the new system. The times taken are compared.

- (i) State one weakness of each of these procedures. [2]

In fact a third procedure which avoids these two weaknesses is adopted. In this procedure each employee is given a randomly selected set of claim forms. Each set contains the same number of forms. The employees each process their set of claim forms on both systems. The times taken, in minutes, are shown in the table.

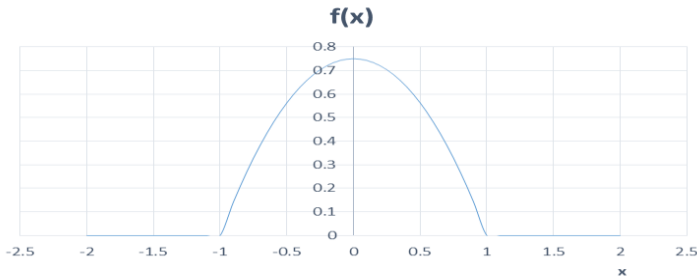
Employee	1	2	3	4	5	6	7	8	9	10
Old system	40.5	42.9	52.8	51.7	77.2	66.7	65.2	49.2	55.6	58.3
New system	39.2	40.7	50.6	50.7	71.4	70.5	71.1	47.7	52.1	55.5

- (ii) Carry out a paired  $t$  test at the 5% level of significance to investigate whether the mean length of time taken to process a set of forms has reduced using the new system. [10]
- (iii) State fully the usual conditions for a paired  $t$  test. [3]
- (iv) Construct a 99% confidence interval for the mean reduction in time taken to process a set of forms using the new system. [3]

**END OF QUESTION PAPER**

Question		Answer	Marks	Guidance	
1	i	$F \sim N(250, 20^2)$ $P(\bar{F}_4 > 260) = P\left(Z > \frac{260-250}{10}\right)$ $= P(Z > 1)$ $= 0.1587$	M1 M1 A1  [3]	standardisation including division by $\sqrt{n}$ correct tail (probability < 0.5) cao (to 3 or 4 sf)	
1	ii	$(F_1 + F_2 + F_3 + F_4) = F' \sim N(1000, 1600)$ $(M_1 + M_2 + \dots + M_{12}) = M' \sim N(2640, 7500)$ $(S_1 + S_2 + S_3 + S_4) = S' \sim N(800, 900)$  $\rightarrow T \sim N(4440, 100^2)$  $P(T < 4500) = P\left(Z < \frac{4500-4440}{100}\right) = P(Z < 0.6)$ $= 0.7258$	M1 A1 B1  M1 A1 [5]	for variances: at least one of $4 \times 15^2$ etc. seen allow ' $4 \times 15 + 12 \times 25 + 4 \times 20$ ', but not $4^2$ etc. for 10,000 (or 2.778 in minutes) for 4440 (or 74 in minutes)  correct tail (probability > 0.5) and $\sqrt{\text{(their variance)}}$ art 0.726 (given to 3 or 4 sf)	
1	iii	Looking for $P((M' - 3.5S') > 0)$  $[M' \sim N(2640, 7500)]$ $3.5S' \sim N(2800, 11025)$ $(M' - 3.5S') \sim N(-160, 18525)$ $= P\left(Z > \frac{160}{\sqrt{18525}}\right) = P(Z > 1.1755) = 0.1199$	M1  M1 B1, A1  A1 [5]	interpret the question correctly; e.g. ' $12M - 3.5 \times 4S$ ' or ' $12M > 3.5 \times 4S$ ' seen  their $\text{Var}(S') \times 12.25$ mean and variance  cao (0.1198 to 0.120)	
1	iv	Looking for $P(\bar{F}_4 - \bar{M}_{12} > 25)$ $\bar{M}_{12} \sim N\left(220, \frac{625}{12}\right)$ $\bar{F}_4 \sim N(250, 100)$ $(\bar{F}_4 - \bar{M}_{12}) \sim N(30, 152.08)$  $P\left(Z > \frac{25-30}{\sqrt{152.08}}\right) = P(Z > -0.4054) = 0.6574$	M1  M1 A1 B1 A1 [5]	interpret the question correctly; e.g. $P(\bar{F}_4 > \bar{M}_{12} + 25)$ seen  variance: at least one of $\frac{25^2}{12}$ or $\frac{20^2}{4}$ seen  correct variance correct mean answer rounds to 0.657 or 0.658	

Question		Answer	Marks	Guidance	
2	a	$H_0$ : The (genetic) model fits the data. $H_1$ : The (genetic) model does not fit the data Observed    125        37        32        6 Expected    112.5    37.5    37.5    12.5  Cont's        1.3889   0.0067   0.8067   3.38  $X^2 = 5.582$ Degrees of freedom = 3 Critical value = 9.348 $5.582 < 9.348 \rightarrow$ cannot reject $H_0$ The data give no reason to doubt the genetic model	B1   <		

Question		Answer	Marks	Guidance	
		Suggests population median length of South American fruit flies exceeds 2.5cm	M1 A1 [9]	FT their $W$ including median (or 'on average') and context	
3	i	$k \int_{-1}^1 (1 - x^2) dx = 1 \quad (\rightarrow k \left[ x - \frac{x^3}{3} \right]_{-1}^1 = 1)$ $\rightarrow \frac{4k}{3} = 1$ $\rightarrow k = \frac{3}{4}$	M1 M1 A1 [3]	Correct integral including limits  (const) $\times k = 1$  cao	
3	ii		B1 B1 B1 [3]	General shape between -1 and +1 Axes labelled with scales and intercepts (FT their $k$ ) Nothing outside $ x  < 1$	
3	iii	$E(X) = 0 \rightarrow V(X) = E(X^2)$ $V(X) = \frac{3}{4} \int_{-1}^1 (x^2 - x^4) dx = \frac{3}{4} \left[ \frac{x^3}{3} - \frac{x^5}{5} \right]_{-1}^1$ $= \frac{1}{5}$	B1 M1 A1 [3]	for $E(X) = 0$  for correct integral including limits  cao (ignore mistakes in working)	

Question	Answer	Marks	Guidance	
3 iv	$\frac{3}{4} \int_0^q (1 - x^2) dx = \frac{1}{4}$ $\text{integration} = \frac{3}{4} \left[ x - \frac{x^3}{3} \right]$ $\rightarrow q - \frac{q^3}{3} = \frac{1}{3} \quad \text{or} \quad \rightarrow q^3 - 3q + 1 = 0$ $g(0.345) = 0.006$ $g(0.355) = -0.02$ Change of sign $\rightarrow 0.345 < q < 0.355$ So upper quartile = 0.35 to 2 dp	M1 B1 A1  M1  E1 [5]	Correct limits and equality f/t their $k$  any correct simplified (3-term) cubic  (allow correct alternative) If solving using calculator: state all three solutions  must be explained clearly If solving by calculator: explain why only one works	
3 v	$\sum X_i > 5 \rightarrow \overline{X}_{40} > 0.125n$ large and so can use Central Limit Theorem $\overline{X}_{40} \sim N\left(0, \frac{0.2}{40}\right)$ $P(\overline{X}_{40} > 5) = P\left(Z > \frac{0.125 - 0}{0.0707}\right) = P(Z > 1.768)$ $P(\overline{X}_{40} > 0.125) = P\left(Z > \frac{0.125 - 0}{0.0707}\right) = P(Z > 1.768)$ $= 0.0385$	B1 M1 A1  A1 [4]	both (or $\bar{X}$ normal) for $\frac{\text{Var}}{40}$ or $\frac{\text{Var}}{\sqrt{40}}$ correct mean and variance (ft any positive variance from iii)  cao 0.0385 or 0.0386	
4 i	A) One set of claim forms could be more difficult to process. B) A form would be more familiar on second processing.	B1 B1 [2]	Allow suitable alternatives	

Question	Answer	Marks	Guidance
4 ii	<p>Values of <math>d</math> 1.3, 2.2, 2.2, 1.0, 5.8, -3.8, -5.9, 1.5, 3.5, 2.8  <math>\bar{d} = 1.06, s = 3.4378</math></p> <p><math>H_0: \mu_D = 0</math>  <math>H_1: \mu_D &gt; 0</math></p> <p>Where <math>\mu_D</math> is the population mean difference between the times taken to process claims using the old system and the new system</p> <p><math>t = \frac{1.06 - 0}{3.4378/\sqrt{10}} = 0.975</math></p> <p>9 degrees of freedom  Critical value = 1.833  <math>0.975 &lt; 1.833</math> so cannot reject <math>H_0</math>  Insufficient evidence to suggest that the mean time for processing forms has reduced using the new system</p>	<p>M1 A1  B1 B1  M1 A1 B1 B1 M1  A1 [10]</p>	<p>calculating differences (at least 3 correct)  both. Allow <math>s^2 = 11.818</math>  Do not allow <math>s_n = 3.2613</math> or <math>s_n^2 = 10.636</math>  hypotheses (allow <math>&lt;0</math> if consistent)  definition including difference, mean, and context.  Allow other symbols only if they are defined as the population mean difference.  Hypotheses in words must include 'population'.  Not 'difference of means'  including <math>\sqrt{10}</math>. FT their <math>s</math> or <math>s_n</math>  cao  no FT if wrong  no FT if wrong  sensible comparison using their test statistic if the previous M1 awarded.  including mean and context FT; not assertive but accept 'Evidence suggests mean time has reduced'.</p>
4 iii	<p>Differences should be:  Normally distributed in population  With unknown variance  Sample (of differences) must be random</p>	<p>B1 B1 B1  [3]</p>	<p>or 'underlying distribution of differences is normal'</p> <p>NB: candidates may say the sample should be small. This is incorrect, but should be ignored for the purposes of marking.  Also ignore 'paired'</p>
4 iv	<p><math>1.06 \pm \frac{3.25(3.4378)}{\sqrt{10}}</math>  <math>= (-2.473, 4.593)</math></p>	<p>M1 B1 A1 [3]</p>	<p>for their <math>\bar{d}</math> and <math>s/\sqrt{10}</math> in correct position  for 3.25  cao    SC: Answers from calculator with no working  3sf or 4sf gets 3/3, &gt;4sf gets 1/3  (-4.593, 2.473) gets 2/3</p>



## 4768 Statistics 3

### General Comments:

This paper was generally very well answered. A vast majority of candidates attempted all the questions and were able to show what they could do.

The solutions to the hypothesis testing questions (2 and 4) were correctly structured, with both hypotheses and conclusions stated clearly and in the context of the question. Occasionally they lacked sufficient context, or the conclusions were overly assertive, but it was pleasing to see that such cases were in a minority. Learners should be reminded that it is good practice to state the hypotheses before carrying out any calculations, even though marks in the examination are awarded for the hypotheses seen anywhere within the solution.

Generally the solutions contained sufficient detail to make the methods clear, even when graphical calculators were used to find, for example, probabilities from the Normal distribution.

The notation for random variables was sometimes unclear, both in Question 1 and when explaining the Central Limit Theorem in Question 3(v). Insisting on clearer notation here might have helped candidates avoid confusion between a random variable and the corresponding sample mean, and enable them to score more marks.

It was pleasing to see that most candidates attempted to answer the 'wordy' questions. There were many good explanations, but sometimes the sentences were not clearly structured which made it difficult to follow the ideas. Bad handwriting on occasion made it difficult to award marks.

Candidates should be reminded to give their answers to an appropriate degree of accuracy, which is generally three or four significant figures. Over-specifying answers can lead to a loss of marks.

### Comments on Individual Questions:

#### Question No. 1

This question was about combinations of Normal variables. Calculations using the Normal distribution were generally done well, with sufficient detail and to an appropriate accuracy.

The random variables in question were not always clearly defined. This was not penalised in itself, but often led to the wrong variance and hence the wrong answer. There were two common mistakes: using the variable instead of its sample mean (and thus failing to divide the variance by  $n$ ); and writing, for example,  $4M$  when what was meant was  $M_1 + M_2 + M_3 + M_4$  (which would lead to the variance  $16\sigma^2$  instead of  $4\sigma^2$ ). Teachers are therefore advised to insist on correct notation for random variables in this type of question.

In part (i) a large number of candidates divided by 20 instead of 10 in the standardisation. In parts (ii) and (iii) careless notation, as described above, often led to the incorrect variance. Where this was avoided, part (ii) in particular was very well done. In part (iv) many either misinterpreted the question ( $M - F - 25$  was often seen) or forgot that it was about the means.

#### Question No. 2

This question consisted of two parts, the first requiring a chi-squared test and the second a Wilcoxon test. Both parts were very well done overall, with the calculations mostly being correct, hypothesis and conclusions being given in sufficient detail, and critical values correctly identified.

In part (a), it was good to see that the hypotheses were generally stated correctly (in the past we saw ‘data fits model’ much more often). Candidates are expected to understand that ‘result insignificant’ means that we have insufficient evidence to reject  $H_0$ ; although this was not always clearly stated, a majority managed to score full marks by correctly interpreting the conclusion in context.

In part (b), the most common loss of marks was for forgetting to include ‘population’ in the definition of the parameter in the hypotheses, and concluding that the South American fruit flies were larger than the European ones without referring to the ‘median’ or ‘average’ length. A more subtle error was to omit the value 2.5 from the hypotheses (stating only that ‘the median length of the South American flies is larger than the median length of the European ones’); the value 2.5 is in fact needed in testing the hypotheses. However, it was good to see that most hypotheses included a definition of the parameter and that the conclusion was nearly always contextualised.

### Question No. 3

Early parts of this question were usually well done. In part (ii) a few candidates drew a triangle instead of a curve, and some parabolas had “tails” (looking more like a normal distribution curve).

In part (iii), since  $E(X) = 0$ , we expected to see ‘ $-0^2$ ’ or ‘ $-E(X)^2$ ’ in the calculation of the variance.

In part (iv) most candidates reached the correct 3-term cubic, although some stopped too soon. Some candidates equated an integral with limits 0 and  $q$  to 0.25 which is correct, but it was not clear that they were using the symmetry of the pdf. For the final part, most were content to demonstrate that substituting 0.35 gave an approximately correct right-hand side of the equation, rather than showing that the value of  $q$  was 0.35 to two decimal places. The latter requires looking for a sign change between 0.345 and 0.355.

Part (v) asked for an application of the Central Limit Theorem. The numerical answer was mostly correct, but the justification for the use of the Normal distribution revealed some misconceptions about the CLT. Many candidates seem to think that, when a sample is large,  $X$  itself can be approximated by a Normal distribution. Many correct references to the CLT forgot to mention the large sample, which is required for the application of the Theorem to be valid.

### Question No. 4

In part (i) most candidates made at least one sensible comment. However, the language was often unclear so it was difficult to tell exactly what point was being made. For Procedure A there needed to be some mention of the two systems, rather than just the two sets of forms. Procedure B weakness was generally well explained, although some candidates resorted to the need for a larger sample.

Part (ii) was largely well done. In stating the hypotheses quite a few described  $\mu$  as the difference between the two sets of times, omitting the word ‘mean’. A minority described  $\mu$  as the ‘difference of the means’, which is incorrect for a paired test. Again, the final conclusion often failed to mention ‘average’ reduction in time.

Part (iii) revealed that the conditions for using a t-test are not very well understood. Most knew that something needed to be Normally distributed, but many simply said ‘the data’ or ‘the underlying variable’, not specifying that it is in fact the population of differences. Most knew that the sample variance should be unknown. Many candidates stated a requirement for a small sample, which is not in fact necessary. On the other hand, the requirement of a random sample was often forgotten.

Finally, part (iv) asked for a confidence interval for the mean reduction in time. Some candidates used one of the single samples instead of the differences, and it was quite common to see the interval for the increase, rather than the reduction. Quite a high proportion used a wrong  $t$  value. Overall, however, most candidates knew the correct method and scored at least one mark on this question.