

**ADVANCED SUBSIDIARY GCE  
MEI STATISTICS**

**G243/01**

Statistics 3 (Z3)

**MONDAY 2 JUNE 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 (45 marks)

1 The manager of a large shopping precinct has commissioned a survey of customers. A sample of customers leaving the precinct one morning are to be interviewed and asked how far they have travelled to get to the precinct and how much money they have spent there.

(i) Each interviewer is instructed to interview twenty adult males (age between 20 and 60), twenty adult females, ten teenage males, ten teenage females and fifteen senior citizens (age 60 or above). What form of sampling is this? Give an advantage and a disadvantage of carrying out the sampling in this way. [3]

(ii) In addition, each person in a simple random sample of 20 customers in the precinct's cafeteria is asked the same questions. The results are as follows.

Distance travelled (miles)	0.2	0.3	0.8	1.0	1.1	1.8	2.3	2.4	2.9	3.6
Money spent (£)	4.6	66.2	4.0	25.5	20.6	4.5	55.8	8.0	7.5	36.4
Distance travelled (miles)	4.1	4.8	5.0	5.1	5.3	8.0	9.4	11.6	13.8	15.2
Money spent (£)	12.0	30.0	6.3	46.6	68.3	60.0	72.5	34.2	65.7	65.0

(A) Draw a scatter diagram to illustrate these data. [3]

(B) Calculate the value of Spearman's rank correlation coefficient and hence test at the 5% level of significance whether there appears to be any association between distance travelled and money spent for the underlying population of such customers. [7]

(C) The manager suggests that the test should instead be based on the product moment correlation coefficient. Explain whether or not you agree with this suggestion. [2]

2 Industrial quality control engineers are studying the output of steel rods made by two machines. It is accepted that, for each machine, there is some variability in the lengths of the rods, but the two machines should make rods of the same length on average. The lengths  $x$ , in metres, of large random samples of rods made by each machine are measured; the results are summarised as follows.

Machine A	Sample size $n_1 = 90$	$\Sigma x = 184.5$	$\Sigma x^2 = 396.94$
Machine B	Sample size $n_2 = 75$	$\Sigma x = 156.0$	$\Sigma x^2 = 334.19$

(i) Test, at the 10% level of significance, whether the two machines are making rods with the same length on average, stating carefully your null and alternative hypotheses and your conclusion. [13]

(ii) Explain briefly why it is not necessary to make any assumption of Normality for the underlying populations. [2]

- 3 Two airlines compete on a route from one city to another. Two businessmen who both make the journey on Tuesday of each week agree to compare the two airlines by noting their total journey times, including checking-in, waiting in the departure lounge and collecting luggage after arrival. One businessman uses one airline, and the other uses the other. Their total journey times, in minutes, for a random sample of 10 weeks are as follows.

Week	A	B	C	D	E	F	G	H	I	J
First airline	121	114	138	110	92	98	92	104	101	115
Second airline	107	98	132	118	116	86	90	119	91	116

- (i) Use the Wilcoxon signed rank test, at the 5% level of significance, to examine whether the underlying median journey times may be assumed equal. [9]
- (ii) Explain why it is sensible for a 'paired comparisons' experiment to have been carried out by the businessmen. [2]
- (iii) Briefly discuss **two** other features of the situation that the businessmen might take into consideration. [4]

[Question 4 is printed overleaf.]

**Section B (27 marks)**

- 4** Scientists at an agricultural research station are developing two new varieties, A and B, of a particular crop. They wish to compare these new varieties with each other and with the existing standard variety, V.

The research station has an experimental field that is divided into plots, all of the same size, on which the crop can be grown under controlled conditions. However, the natural fertility of the soil in the field varies somewhat from place to place and depending on its previous use, so some plots may be naturally more fertile than others.

- (i) Explain why it is sensible to allocate the varieties A, B and V to the plots in the field in a random way. [2]
- (ii) Explain why each of the varieties should appear more than once in the allocation. [2]

Initially the scientists compare the two new varieties, A and B, with each other. 6 of the plots in the field had received variety A and 5 had received variety B. Their yields, in kg per plot, were as follows.

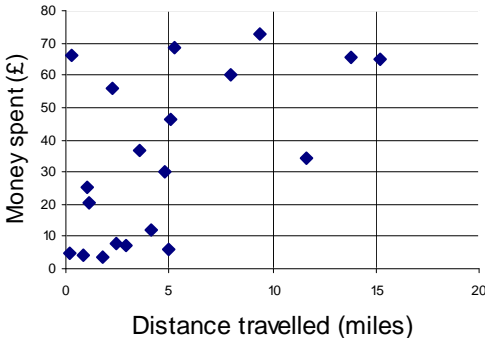
Variety A	24.1	23.3	21.8	24.6	23.7	23.5
Variety B	20.8	22.4	21.6	23.1	21.8	

- (iii) Stating carefully the required assumptions and the null and alternative hypotheses, use a  $t$  test to examine, at the 5% level of significance, whether the mean yields of the two varieties may be assumed to be the same. [15]

Development of variety B is discontinued. The scientists design a further experiment to compare A with V, in which plots of A and V are used alongside each other at each of a number of locations around the field.

- (iv) Explain why this design should be expected to achieve a more precise comparison of A with V than using plots of A and V in random locations around the field. [2]
- (v) This experiment has 5 locations where A and V are used alongside each other. Making the appropriate distributional assumption, the scientists examine whether the overall mean yield of A appears to be greater than that of V by calculating the value of the usual  $t$  statistic for this situation; they find this value to be 1.91. Complete the test, using a 5% level of significance, and state your conclusion. State the distributional assumption the scientists made. [6]

# G243 Statistics 3

1		(i)	<p>Quota sampling.  <b>Advantage</b> – probably the only realistic way to get a reasonably ‘representative’ sample in these circumstances.  <b>Disadvantage</b> – non-random, so statistical analysis is complicated.</p>	B1 E1 E1	<p>Or other sensible comments.                  Eg cost or time effective as an advantage</p>	3
	(a)	(ii)		G1 G1 G1	<p>Axes, including labels.                  Correct zero.                  All points correct (allow 2 errors).</p>	3
	(b)		<p>Ranks:                  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20                  3 18 1 9 8 2 14 6 5 12 7 10 4 13 19 15 20 11 17 16  <math> d </math> 2 16 2 5 3 4 7 2 4 2 4 2 9 1 4 1 3 7 2 4</p> $r_s = 1 - \frac{6 \times 584}{20 \times 399} = 1 - 0.4391 = 0.5609$ <p>Critical point for <math>n=20</math> at two-sided 5% level is 0.4466                  Significant.                  Seems there is an association between distance travelled and money spent.</p>	B2 M1 A1 B1 E1 E1	<p>Allow B1 if one or two errors.                  CAO.                  No FT if wrong.                  No access to these marks if value of <math>r_s</math> is nonsense.</p>	7
	(c)		<p>Some sensible explanation of “no”.                  Scatter diagram does not suggest bivariate Normality.</p>	M1 A1	<p>SC1. Allow 1 out of 2 if “bivariate” missing.                  SC2. Allow 1 out of 2 for sensible comment re “outliers”.                  No marks for “data not linear”.</p>	2
2						
		(i)	<p><math>H_0 : \mu_A = \mu_B</math>  <math>H_1 : \mu_A \neq \mu_B</math></p> <p>Where <math>\mu_A, \mu_B</math> are the population mean lengths for the machines.</p>	B1 B1 B1	<p>Do not allow <math>\bar{A}, \bar{B}</math> or similar unless they are clearly and explicitly stated to be <u>population</u> means.                  Hypotheses in words must include “population”.</p>	

		$\left. \begin{aligned} \bar{x}_1 &= \frac{184.5}{90} = 2.05 \\ \bar{x}_2 &= \frac{156.0}{75} = 2.08 \end{aligned} \right\}$ $\left. \begin{aligned} s_1^2 &= \frac{1}{89} \left( 396.94 - \frac{184.5^2}{90} \right) = \frac{18.715}{89} = 0.2103 \\ s_2^2 &= \frac{1}{74} \left( 334.19 - \frac{156.0^2}{75} \right) = \frac{9.71}{74} = 0.1312 \end{aligned} \right\}$ <p>Because the samples are large, the values of <math>s_1^2</math> and <math>s_2^2</math> are taken as <math>\sigma_1^2</math> and <math>\sigma_2^2</math>.</p> <p>Two-sample test based on <math>N(0,1)</math>.</p> <p>Test statistic is:</p> $\frac{2.05 - 2.08}{\sqrt{\frac{0.2103}{90} + \frac{0.1312}{75}}} = -\frac{0.03}{\sqrt{0.004086}} = -\frac{0.03}{0.0639} = -0.46 \text{ (93)}$ <p>Double-tailed 10% point of <math>N(0,1)</math> is 1.645. Not significant. No reason to suppose mean lengths differ.</p>	<p>B1</p> <p>M1</p> <p>A1</p> <p>M1</p> <p>M1</p> <p>M1</p> <p>A1</p> <p>A1</p> <p>E1</p> <p>E1</p>	<p>For adequate verbal definition. Allow absence of "population" here if correct notation <math>\mu</math> has been used.</p> <p>M0 A0 for divisor <math>n</math>, but FT.</p> <p>Accept as implicit if <math>s_1^2</math> and <math>s_2^2</math> are <u>correctly</u> used in sequel.</p> <p>Accept usual alternatives.</p> <p>No FT if wrong.</p>	<p>13</p>
	(ii)	<p>Samples are large, so by the Central Limit Theorem the underlying distribution of the sample means will be approximately Normal.</p>	E2	(2, 1, 0)	2
3					
	(i)	<p>Differences are:</p> <p style="text-align: center;">14 16 6 -8 -24 12 2 -15 10 -1</p> <p>Ranks of <math> d </math> 7 9 3 4 10 6 2 8 5 1</p> <p>Test statistic is <math>4+10+8+1=23</math> (or <math>7+9+3+6+2+5=32</math>)</p> <p>Refer to paired Wilcoxon table with <math>n=10</math>.</p> <p>Need lower <math>2\frac{1}{2}\%</math> point which is 8 (or, if 32 used, upper <math>2\frac{1}{2}\%</math> point which is 47).</p> <p>Not significant.</p> <p>Seems underlying median total journey times may be assumed equal.</p>	<p>B1</p> <p>M1</p> <p>A1</p> <p>M1</p> <p>A1</p> <p>M1</p> <p>A1</p> <p>E1</p> <p>E1</p>	<p>FT if M1 earned or if <math>d</math> (not <math> d </math>) ranked.</p> <p>FT if M1 earned.</p> <p>No FT if wrong.</p> <p>No FT if wrong.</p>	9
	(ii)	<p>The "pairing" will eliminate any differences between the weeks - so can compare the two airlines.</p>	E1		2

		(iii)	Two sensible comments such as: <ul style="list-style-type: none"> <li>- check-in and waiting times not in airlines' control</li> <li>- time for collecting luggage not in airlines' control</li> <li>- other journey criteria might be of importance (e.g. departure time, on-board service, fares).</li> </ul>	E2 E2	Reward any two sensible comments for (E2 each) (2,1,0). For 2 marks there must be some comment to the effect of <u>comparison</u> , not merely that a factor might affect both airlines.	4
4						
		(i)	Randomisation: to guard against possible unsuspected sources of bias - caused by fertility patterns among the plots.	E1 E1	Or equivalent comments.	2
		(ii)	Replication: so that natural variation can be measured, so that any observed inter-variety variation can be compared with it.	E1 E1	Or equivalent comments.	2
		(iii)	Normality of <u>both populations</u> , equal <u>population</u> variances.  $H_0 : \mu_A = \mu_B$ $H_1 : \mu_A \neq \mu_B$  where $\mu_A, \mu_B$ are the population means for varieties A and B.  A : $\bar{x} = 23.50$ , $s_{n-1} = 0.9529$ B : $\bar{y} = 21.94$ , $s_{n-1} = 0.8649$  $\text{Pooled } s^2 = \frac{(5 \times 0.908) + (4 \times 0.748)}{9} = \frac{7.532}{9} = 0.836\dot{8}$  $\text{Test statistic is } \frac{23.50 - 21.94(-0)}{\sqrt{0.836\dot{8}} \sqrt{\frac{1}{6} + \frac{1}{5}}}$  $= \frac{1.56}{0.5539(5)} = 2.816$  Refer to $t_9$ . Double-tailed 5% point is 2.262.	B1 B1  B1 B1  B1  M1 A1  M1 M1  M1  A1	Do not allow $\bar{A}, \bar{B}$ or similar unless they are clearly and explicitly stated to be <u>population</u> means. Hypotheses in words must include "population".  Do not allow $s_n$ : 0.8699, 0.7736  For any reasonable attempt at pooling. If correct.  For numerator. For $\sqrt{0.8368}$ (or cand's value). For $\sqrt{\frac{1}{6} + \frac{1}{5}}$ .  FT from here if all M marks earned.	

		Significant. Appears that population mean yields are different.	M1 A1 E1 E1	No FT if wrong. [accept usual No FT if wrong. [alternatives.]	15
	(iv)	The pairing will eliminate differences around the field. - can compare the plots within the pairs.	E1 E1		2
	(v)	Refer to $t_4$ Single-tailed 5% point is 2.132. Not significant.  No evidence to reject $H_0$ that population mean yields of A and V are the same. Normality of underlying population of differences.	M1 A1 E1  E1 B1 B1	No FT if wrong. No FT if wrong.	6