OCR
RECOGNISING ACHIEVEMENT

General Certificate of Education
Advanced Supplementary (AS) and Advanced Level
former Oxford and Cambridge modular syllabus

MEI STRUCTURED MATHEMATICS 5516

Statistics 4

Tuesday **16 JANUARY 2001** Afternoon 1 hour 20 minutes

Additional materials:
Answer paper
Graph paper
Students' Handbook

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

Write your name, Centre number and candidate number in the spaces provided on the answer paper/ answer booklet.

Answer any **three** questions.

Write your answers on the separate answer paper provided.

If you use more than one sheet of paper, fasten the sheets together.

INFORMATION FOR CANDIDATES

The approximate allocation 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 sufficient detail of the working is shown on the answer paper to indicate that a correct method is being used.

This question paper consists of 4 printed pages.

[Turn over

1 The continuous random variable X has probability density function

$$f(x) = xe^{-\frac{1}{2}x^2} \quad \text{for } x \geqslant 0.$$

(i) By first explaining why

$$\int_0^\infty x^2 e^{-\frac{1}{2}x^2}\,dx = \tfrac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2}\,dx$$

and then considering the Normal distribution with mean 0 and variance 1, show that

$$E(X) = \tfrac{1}{2}\sqrt{2\pi}.$$

[You are reminded that the probability density function of $Z \sim N(0,1)$ is $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.] [7]

(ii) Show that the cumulative distribution function of X, $F(x) \equiv P(X \leqslant x)$, is

$$F(x) = 1 - e^{-\frac{1}{2}x^2}.$$ [2]

The random variable Y is defined by $Y = X^2$.

(iii) Explain why, for $y = x^2$,

$$P(Y \leqslant y) = P(X \leqslant x)$$

and deduce that the cumulative distribution function of Y, $G(y)$, is

$$G(y) = 1 - e^{-\frac{1}{2}y}.$$ [5]

(iv) Deduce the probability density function of Y. [1]

(v) Find the mean of Y and hence obtain the variance of X. [5]

2 Two different designs for a large open-plan office are being compared in respect of the amount of light available at locations where employees will be working. The amount of light is measured by photoelectric cells at 12 randomly selected locations for one design and, independently, at 10 randomly selected locations for the other design. The data, in a standard unit of light, are summarised as follows.

First design:	$n_1 = 12$	$\bar{x} = 9.85$	$\sum(x_i - \bar{x})^2 = 23.410.$
Second design:	$n_2 = 10$	$\bar{y} = 8.76$	$\sum(y_i - \bar{y})^2 = 23.058.$

It is desired to examine whether, overall, the mean amount of light delivered is the same in the two designs.

(i) State the null and alternative hypotheses and the required assumptions for the use of a t test. [4]

(ii) Carry out the test, at the 10% significance level. [9]

(iii) Provide a two-sided 99% confidence interval for the true mean difference. [4]

(iv) Suppose that the underlying variances of the amount of light delivered, σ_1^2 for the first design and σ_2^2 for the second, could be taken as known (perhaps from analysis of many other designs). Describe briefly [no calculations are required] how the test procedure in part **(ii)** should be modified. [3]

3 Records have been kept, during a large number of working days, of the numbers of heavy lorries per hour travelling eastbound and westbound on a certain stretch of main road. It is anticipated that there will be some variation from hour to hour, and it is thought that these variations might not be well modelled by Normal distributions. Therefore the Wilcoxon paired sample test is to be used to examine whether, overall, the distributions of eastbound and westbound numbers can be assumed to have the same location parameter. The data for a random sample of 12 hours are as follows.

Eastbound	89	94	79	70	86	68	73	76	85	75	57	66
Westbound	71	90	58	46	94	55	51	92	84	77	71	73

(i) Carry out the test, at the 5% significance level. [8]

(ii) Suppose that the null hypothesis that the location parameters are equal is true, and that all the other conditions for the correct applicability of the test are satisfied. Show that the expected value of the test statistic is $\frac{1}{4}n(n+1)$, where n is the number of pairs of observations. [5]

(iii) Find the level of significance of the above data, using the Normal approximation

$$\mathrm{N}\left(\tfrac{1}{4}n(n+1), \tfrac{1}{24}n(n+1)(2n+1)\right).$$ [7]

[Turn over

4 The marketing manager at a theme park undertakes a survey of a random sample of 200 visitors. As part of the analysis, he categorises them as local people, people who have come a medium distance or people who have come a long distance, with a separate category of people in coach parties. He also categorises them according to the amount of money they spend in the park, as light, medium or heavy spenders. A table displaying the results is as follows.

		Amount spent		
		Light	Medium	Heavy
Distance	Local	17	23	16
	Medium distance	15	25	34
	Long distance	4	16	12
	Coach party	8	22	8

Stating carefully your null and alternative hypotheses, examine whether or not there is any association between 'distance' and 'amount spent'. Use a 10% significance level. [14]

Discuss your conclusions. [6]

Mark Scheme

1	(i)	[Rayleigh dist with $\theta = 2$] $f(x) = xe^{-\frac{x^2}{2}}$ $x \geqslant 0$; $Y = X^2$			
		Explanation $\left(\text{of } \int_0^\infty x^2 e^{-\frac{x^2}{2}}\,dx = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx \right)$ to the effect that the integrand is an even function, etc	E2		
		$E[x] = \int_0^\infty x^2 e^{-\frac{x^2}{2}}\,dx = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\,dx$	M1		
		$= \frac{1}{2}\sqrt{2\pi}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 . e^{-\frac{x^2}{2}}\,dx$	M1	To introduce $\sqrt{2\pi}$ in denominator pdf of N(0, 1)	
		Integral = $E[x^2]$ for N(0, 1) = Var[N(0, 1)]	M1 M1	Any or all of these M marks may be implicit – but beware printed answer	
		$= \frac{1}{2}\sqrt{2\pi}.1$	1		7
	(ii)	$F(x) = \int_0^x te^{-\frac{t^2}{2}}\,dt = \left[-e^{-\frac{t^2}{2}} \right]_0^x$	1		
		$= 1 - e^{-\frac{x^2}{2}}$	1		2
	(iii)	Explanation	E2	Formality is NOT expected, but to the effect that: $y = x^2$ for $x \geqslant 0$ is one-one monotonic increasing; so, for $y = x^2$, we have that the event $\{Y \leqslant y\}$ is the same as the event $\{X \leqslant x\}$; etc	
		$\therefore\ G(y) = P(Y \leqslant y) = P(X \leqslant x) = 1 - e^{-\frac{x^2}{2}}$	M1		
		Re-expressed as a function of y	M1		
		$= 1 - e^{-\frac{y}{2}}$	1	Beware printed answer; must be convincing	5
	(iv)	pdf of Y is $g(y) = \frac{d}{dy}G(y) = \frac{1}{2}e^{-\frac{y}{2}}$	1		1
	(v)	$E[Y] = \int_0^\infty \frac{1}{2} y e^{-\frac{y}{2}}\,dy = \frac{1}{2}\left\{ \left[\frac{y e^{-\frac{y}{2}}}{-\frac{1}{2}} \right]_0^\infty + 2\int_0^\infty e^{-\frac{y}{2}}\,dy \right\}$	M1	The M1 + 1 are available for recognition of the mean of the exponential dist with $\lambda = \frac{1}{2}$, provided this is explained.	
		$= \frac{1}{2}\left\{ 0 + 2\left[\frac{e^{-\frac{y}{2}}}{-\frac{1}{2}} \right]_0^\infty \right\} = \frac{1}{2}\{-4[0-1]\} = 2$	1		
		$\text{Var}(x) = E[x^2] - (E[x])^2$	M1		
		We now have $E[x^2] = 2$	1		
		$\therefore\ \text{Var}(x) = 2 - \left(\frac{1}{2}\sqrt{2\pi}\right)^2 = 2 - \frac{\pi}{2}$	1		5

2	(i)	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	1 1	Deduct 1 from any marks awarded here if μ_1, μ_2 are not defined in words	
		Where μ_1, μ_2 are the population mean amounts of light delivered for the two designs			
		Normality of <u>both</u> populations Same variance	1 1		4
	(ii)	$n_1 = 12 \quad \bar{x} = 9.85 \quad \sum(x-\bar{x})^2 = 23.410$ $n_2 = 10 \quad \bar{y} = 8.76 \quad \sum(x-\bar{y})^2 = 23.058$			
		Pooled $s^2 = \frac{23.410+23.058}{20} = 2.3234$	M1 A1	Must be correct method; but f.t. any reasonable attempt into the test and CI	
		Test statistic is Numerator: Denominator: $\frac{9.85-8.76(-0)}{\sqrt{2.3234}\sqrt{\frac{1}{12}+\frac{1}{10}}} = \frac{1.09}{0.6526(5)} = 1.67$	 M1 M1 A1		
		Refer to t_{20}	1	May be awarded even if test statistic is wrong	
		Dt 10% pt is 1.725 Not significant Seems designs are the same in this respect	1 1 1	No f.t. if wrong	9
	(iii)	CI given by 1.09 $\pm$ 2.85 $\times$ 0.6526(5) $= 1.09 \pm 1.8568$ $= -0.76(68), 2.94(68)$	 M1 B1 M1 A1	 cao	4
	(iv)	Use denominator of $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ and refer to N(0, 1)	1 1 1	For (almost!) any form involving the two separate σ^2 If correct	3

3	(i)	Eastbound 89 94 79 70 86 68 73 76 85 75 57 66 Westbound 71 90 58 46 94 55 51 92 84 77 71 73 Difference 18 4 21 24 -8 13 22 -16 1 -2 -14 -7 M1 Rank of |d| 9 3 10 12 5 6 11 8 1 2 7 4 M1 A1 $T = \begin{cases} 9+3+10+12+6+11+1=53 \\ \text{or} \\ 5+8+2+7+4 \quad =26 \end{cases}$ 1 Refer smaller value to appropriate table M1 d.t. 5% pt for $n = 12$ is 13 1 Note for examiner – s.t. 5% pt is 17 Result is not significant 1 Can assume Eastbound and Westbound location parameters are the same 1	8
	(ii)	If locations are the same, each difference is equally likely to be positive or negative E2 So E[T] is simply $\frac{1}{2}(1 + 2 + \ldots + n)$ M2 Must refer to E[T], not merely to T $= \frac{1}{4}n(n+1)$ 1	5
	(iii)	T ~ approx N $\frac{1}{4}\times12\times13=39$, $\frac{1}{24}\times12\times13\times25=162.5$ Significance level of data = 2 2 f.t. if absent $\times$ P(T $\leqslant$ 26) M1 $\approx 2 \times P(N(0, 1) < \frac{26\frac{1}{2}-39}{\sqrt{162.5}}$ 2 CT7 CORN, f.t. if 26 used; get – 1.0198 and 0.1539 $= -0.9806$) 1 $= 2 \times 0.1633$ 1 $= 0.3266$ (32.66%)	7

4		H_0 : no association H_1 : association	1 1		

O_i:	Light	Medium	Heavy	
Local	17	23	16	56
Medium	15	25	34	74
Long	4	16	12	32
Coach	8	22	8	38
	44	86	70	200

E_i:			
	12.32	24.08	19.6
	16.28	31.82	25.9
	7.04	13.76	11.2
	8.36	16.34	13.3

A4 — deduct 1 for every 2 errors; deduct 1 if only 1 d.p.; deduct 2 if integers

Contributions to x^2:

1.7778	0.0484	0.6612
0.1006	1.4617	2.5332
1.3127	0.3647	0.0571
0.0155	1.9606	2.1120

	Refer to χ^2_6	3	or zero f.t. if df wrong, unless ≈ 200	
	Upper 10% point is 10.64	1		
	Significant	1		
	Seems there is association	1	Zero if $H_0 \leftrightarrow H_1$	14
	Perhaps there is no overwhelming feature here, but there are lots of 'little points' to bring out: Locals tend to be light spenders and not heavy Mediums tend to be heavy spenders and not medium Longs tend to be medium spenders but not light Coach people tend to be medium spenders but not heavy	E6		6

Examiner’s Report

Statistics 4 (5516)

General Comments

There were 54 candidates from 13 centres. This was a distinctly smaller entry than in January 2000 (94 from 23 centres). The size of the winter entry for this module has varied quite considerably in recent years. There was some very good work, with many candidates deservedly scoring highly. On the other hand, there was a higher proportion of very poor work than is usual, several candidates being clearly extremely uncomfortable. Overall, the marks awarded covered the full range of the scale.

Comments on Individual Questions

Question 1 (Expectation algebra)

This was by a long way the least popular of the four questions on the paper. Candidates for this module often seem to avoid the more mathematical work. This is not a wise strategy; usually plenty of intermediate results are given, and it is only necessary to proceed carefully through the question a step at a time. Though there is commonly some integration to be done, candidates for statistics modules at this level certainly ought *not* to be in any way frightened of the technical calculus that is an inherent part of the theory of statistics. Turning now to this particular question, it was based on what is known as Rayleigh distribution, exploring its relationships with the Normal and exponential distributions. Part (i) required the mean of the given Rayleigh distribution to be found, by first explaining that the required integral is that of an even function and then relating it to the variance of the N(0, 1) distribution. The phrase 'even function' was not necessarily expected by name; most candidates spotted the symmetry and could describe it convincingly. Candidates were however less happy with relating the integral to that giving the variance of N(0, 1). The next three parts of the question guided candidates through the procedure for finding the probability density function (pdf) of X^2 from that of X, using what is often called the 'cumulative distribution function (cdf)' method. There were some candidates who could not obtain the quoted cdf of X, which is very poor as this work should be thoroughly known from the Statistics 3 module. However, the manipulations following this, leading to the cdf for X^2, were usually explained fairly convincingly, though it was surprising that a few candidates could not differentiate the cdf to get the pdf. The final parts, leading to $\mathrm{E}(X^2) = 2$ and $\mathrm{Var}(X) = 2 - \frac{\pi}{2}$, remained a mystery to some candidates, but others were reasonably successful here.

Question 2 (unpaired *t* test and confidence interval)

The most disturbing feature here was that a large number of candidates had difficulty in forming the pooled estimate of σ^2 from the given values of $\sum(x_i - \bar{x})^2$ and $\sum(y_i - y)^2$. Rather more fundamentally, they did not know how these quantities relating to the sample variances. This is work from Statistics 1 and it is disappointing that so many candidates could not handle it. The correct value here is 2.3234. Most candidates then knew how to find the test statistic, using whatever value for the pooled estimate they had calculated, but there were several errors with the factor representing the sample sizes. Sometimes t distributions with other than the correct number [20] of degrees of freedom turned up, and sometimes the double-tailed 10% point was not correctly found. [Value of test statistic is 1.67, critical point is 1.725.] Proceeding to the confidence interval, the error of not using the same distribution as for the test appeared from time to time – but, that apart, this work was usually correct [$1.09 \pm 2.845 \times 0.6526 = (-0.767, 2.947)$.] In the last part of the question, candidates were invited to describe the N(0, 1) test to be used if the population variances were known. Most knew roughly what to do, but errors appeared in the standard deviation of the test statistic. Finally, to return to the first part of the question, where hypotheses and assumptions had to be stated, here candidates were expected to be *careful* and *complete*, particularly in ensuring that the hypotheses referred to *population* quantities and that the assumption of Normality referred to *both* populations.

However, it must also be said that many candidates did the question very well.

Question 3 (Paired Wilcoxon test)

The test was nearly always carried out correctly. [Value of test statistic is 26; critical point is 13.] A few candidates had ranking systems that were wrong in principle (not merely minor slips in carrying out the correct method). The explanation of the expected value of the test statistic in part (ii) required candidates to grasp that, under the null hypothesis, positive and negative differences are *equi-probable* and so the *expected value* of the test statistic is simply half the sum of all the ranks from 1 to n. Many candidates seemed to have some idea about these points; some clearly grasped them fully, others were not wholly secure. The last part of the question referred to the Normal approximation and its use in finding the level of significance of the data in the question. The correct answer here was 0.3266 (or 32.66%) and this was often achieved; but the familiar mistakes, of not incorporating a continuity correction (or incorporating an incorrect one) and/or not doubling the tail-area probability as it is a two-sided test, appeared as usual.

Question 4 (Chi-squared analysis of contingency table)

Overwhelmingly popular as always, and generally the arithmetic was well done [value of test statistic = 12.41] and with the correct number of degrees of freedom [6; critical point is 10.64]. Sometimes the discussion at the end was rather thin.

However many candidates had their null and alternative hypotheses the wrong way round, averring that the null is that there *is* association. This is a serious mistake and in fact it completely invalidates the whole of the rest of the solution. Despite the severity of the error, it has been the practice in marking to allow the 'arithmetic' marks and the marks for the 'mechanics' of the test to be earned. This is perhaps too lenient. In future examination sessions the practice may be adopted of awarding zero for the whole question to candidates who have their hypotheses the wrong way round.

Decision and Discrete Mathematics (5519)

General Comments

Candidates were well prepared for this paper. This was particularly true of the simulation question. This was not any easier than previous simulation questions, but most candidates seemed to avoid the problems which have caused difficulties in the past.

Comments on Individual Questions

Question 1 (Simulation)

Part (a) was almost uniformly correct, with the exception of the odd arithmetic mistake. Part (b)(i) was divided between the majority with the correct solution (or a common error in which the numbers 00-96 were used), and those who tried to divide the entire range to fit the required proportions. A very small number provided odd (but workable) rules using 60, 72 or 84 random numbers rather than 96. In part (b)(ii) there were many possible misreadings of the instructions, and most were seen at some point. Some candidates read the random numbers in columns rather than rows; some read columns and omitted columns with an out of range random number; some read in rows, but simulated five questions instead of four; some used all seven numbers in a row; on one script there was a seemingly random selection of the random numbers which had been provided. Follow-through was applied as far as was possible. Part (b)(iii) was straightforward, and found so by candidates. Part (b)(iv), was often the part that really tested the understanding of the process of simulation.