OCR

RECOGNISING ACHIEVEMENT

OXFORD CAMBRIDGE AND RSA EXAMINATIONS

Advanced Subsidiary General Certificate of Education
Advanced General Certificate of Education

MEI STRUCTURED MATHEMATICS 2617

Statistics 5

Thursday **5 JUNE 2003** Morning 1 hour 20 minutes

Additional materials:
Answer booklet
Graph paper
MEI Examination Formulae and Tables (MF12)

TIME 1 hour 20 minutes

INSTRUCTIONS TO CANDIDATES

- Write your Name, Centre Number and Candidate Number in the spaces provided on the answer booklet.
- Answer any **three** questions.
- You are permitted to use a graphical calculator in this paper.

INFORMATION FOR CANDIDATES

- The 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 you show sufficient detail of the working to indicate that a correct method is being used.
- Final answers should be given to a degree of accuracy appropriate to the context.
- The total number of marks for this paper is 60.

This question paper consists of 5 printed pages and 3 blank pages.

1 **(i)** The discrete random variable X has the geometric distribution with

$$P(X = x) = p(1 - p)^{x-1} \qquad x = 1, 2, 3, \ldots ,$$

where $0 < p < 1$. Show that the probability generating function of X is

$$G(t) = pt[1 - t(1 - p)]^{-1}$$

and hence that the mean and variance of X are

$$\mu = \frac{1}{p} \qquad \text{and} \qquad \sigma^2 = \frac{1-p}{p^2}$$

respectively. [9]

(ii) The random variable X in part **(i)** may be interpreted as the number of trials up to and including the first "success" in a sequence of independent trials on each of which the probability of "success" is p. Explain why the random variable Y which gives the number of such trials up to and including the nth success may be written as

$$Y = X_1 + X_2 + \ldots + X_n$$

where the X_i are independent random variables each distributed as X in part **(i)**. Hence write down the probability generating function, the mean and the variance of Y. [5]

(iii) State an approximation to the distribution of Y for large n. [1]

(iv) A university has bought a large batch of cheap felt pens for its stationery stores. Unfortunately some of these are of faulty manufacture and their ink supply dries up immediately on use. Suppose that 75% of the batch are satisfactory.

A department needs 100 such pens for use by its staff. Find the probability that it needs to order at least 125 from the stationery stores to get 100 satisfactory ones. [5]

2 You are given the distributional result

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

with S^2 defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

where $X_1, X_2, \ldots, X_n$ are independent N(0, 1) random variables and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(i) The following are 8 observations from a Normal distribution.

66.4 64.8 67.6 70.6 68.2 66.8 63.8 66.0

Test at the 10% level of significance whether its variance can be taken to be 14.8. [8]

(ii) The moment generating function of the random variable having the χ^2_m distribution is

$$\mathrm{M}(\theta) = (1 - 2\theta)^{-\frac{1}{2}m}.$$

Use this to show that S^2 is an unbiased estimator of σ^2. [6]

(iii) The probability density function $\mathrm{f}(y)$ of the random variable having the χ^2_m distribution is

$$\mathrm{f}(y) = Ky^{\frac{1}{2}(m-2)}\mathrm{e}^{-\frac{1}{2}y}$$

for $y \geqslant 0$, where K is a constant (dependent on the value of m). Use this to obtain the moment generating function quoted in part (ii). [6]

[*Hint.* Use the substitution $y(\frac{1}{2} - \theta) = \frac{1}{2}u$ and then reconsider the form of $\mathrm{f}(y)$.]

[Turn over

3 New machinery has been installed on a production line.

(a) It is claimed that the new machinery will reduce the variability of a critical dimension in the items produced.

This dimension had been carefully measured for a random sample of 11 items produced using the old machinery. The results, in centimetres, were as follows.

8.63 8.45 8.48 8.64 8.53 8.58 8.55 8.69 8.59 8.51 8.53

A random sample of 9 items produced using the new machinery gives the following measurements, in centimetres, for this dimension.

8.54 8.55 8.61 8.57 8.49 8.56 8.54 8.58 8.57

Carry out a two-sided test, at the 5% level of significance, of whether the underlying variances for this dimension with the old and new machinery may be assumed equal. State carefully the required distributional assumptions. [10]

(b) It is also claimed that the new machinery will reduce the proportion of unsatisfactory items that are produced.

It had been noted that there were 8 unsatisfactory items in a batch of 100 produced using the old machinery. It is now noted that a batch of 150 produced using the new machinery has only 6 unsatisfactory items.

Construct a 95% confidence interval for the difference in the true proportions of unsatisfactory items produced using the old and new machinery. Interpret this interval in terms of the claim made for the new machinery. [10]

4 A manufacturer of perfume has both a standard range and a luxury range. The luxury range is more profitable, so the manufacturer is investigating whether improved marketing and advertising can persuade a greater proportion of customers to choose it in preference to the standard range. Records show that currently 23% of customers have been choosing the luxury range.

After a sustained advertising campaign, it is found that 28 out of a sample of 100 customers (regarded as a random sample from the underlying population) choose the luxury range.

(i) Test the null hypothesis $p = 0.23$ against the alternative hypothesis $p > 0.23$ at the $2\frac{1}{2}\%$ level of significance, where p is the proportion of customers now choosing the luxury range. [8]

(ii) Show that, if in the sample of 100 customers there had been 32 choosing the luxury range, the test in part (i) would have rejected the null hypothesis. [3]

(iii) Consider now the test procedure, still with a sample size of $n = 100$, and regard the observed proportion of customers choosing the luxury range as a *continuous* variable. Denote this observed proportion by $\hat{p}$, take the acceptance region as $\hat{p} < 0.3125$, and now ignore any continuity correction. Show that an expression for the operating characteristic of the test is

$$\Phi\left(\frac{3.125 - 10p}{\sqrt{p(1-p)}}\right)$$

where Φ denotes the standard Normal cumulative distribution function.

Evaluate this expression for $p = 0.20, 0.23, 0.30, 0.40$. [5]

(iv) Explain why the operating characteristic is important to the manufacturer in analysing the behaviour of the test. Discuss whether you think that the test is behaving well. [4]

Mark Scheme

Marking Instructions

Some marks in the mark scheme are explicitly designated as 'M', 'A', 'B' or 'E'.

'M' marks ('method') are for an *attempt to use* a correct method (not merely for stating the method).

'A' marks ('accuracy') are for accurate answers and can only be earned if corresponding 'M' mark(s) have been earned. Candidates are expected to give answers to a sensible level of accuracy in the context of the problem in hand. The level of accuracy quoted in the mark scheme will sometimes deliberately be greater than is required, when this facilitates marking.

'B' marks are independent of all others. Typically they are available for correct quotation of points such as 1.96 from tables.

'E' marks ('explanation') are for explanation and/or interpretation. These will frequently be sub-dividable depending on the thoroughness of the candidate's answer.

Follow-through marking should normally be used wherever possible – there will however be an occasional designation of 'c.a.o.' for 'correct answer only'.

Full credit **MUST** be given when correct altemative methods of solution are used. If errors occur in such methods, the marks awarded should correspond as nearly as possible to equivalent work using the method in the mark scheme.

All queries about the mark scheme should have been resolved at the standardisation meeting. Assistant Examiners should telephone the Principal Examiner (or Team Leader if appropriate) if further queries arise during the marking.

Assistant Examiners may find it helpful to use shorthand symbols as follows:

FT	Follow-through marking
√^	Correct work after error
✗^	Incorrect work after error
C	Condonation of a minor slip
BOD	Benefit of doubt
NOS	Not on scheme (to be used *sparingly*)
I	Work of no value

Q1		$P(X=x)=p(1-p)^{x-1}\quad x=1,2,\ldots$		
	(i)	Pgf $G(t)=E[t^X]=pt^x(1-p)^{x-1}$	M1	
		$=pt\{1+t(1-p)+[t(1-p)]^2+\ldots\}$	1	
		$=\frac{pt}{1-t(1-p)}$ [provided $\lvert t(1-p)\rvert<1$] (consideration of this not required)	1	
		$\mu=G'(1),\ \sigma^2=G''(1)+\mu-\mu^2$		
		For attempt to find $G'(t)$ and/or $G''(t)$	M1	
		$G'(t)=pt(-1)[1-t(1-p)]^{-2}[-(1-p)]+p[1-t(1-p)]^{-1}$ $=p(1-p)t[1-t(1-p)]^{-2}+p[1-t(1-p)]^{-1}$		
		or equivalent expression; $\frac{p}{[1-(1-p)t]^2}$ is useful	1	
		$\therefore\ \mu=G'(1)=p(1-p)[p]^{-2}+p[p]^{-1}=\frac{p-p^2}{p^2}+1=\frac{1}{p}$ beware printed answer	1	
		$G''(t)=p(1-p)t(-2)[1-t(1-p)]^{-3}[-(1-p)]$ $+p(1-p)[1-t(1-p)]^{-2}+p(-1)[1-t(1-p)]^{-2}[-(1-p)]$ or equivalent expression	1	
		$\therefore\ G''(1)=2p(1-p)^2p^{-3}+p(1-p)p^{-2}+p(1-p)p^{-2}$ $=\frac{2(1-p)^2}{p^2}+\frac{2(1-p)}{p}=\frac{2(1-p)}{p^2}\{1-p+p\}$	1	
		$\therefore\ \sigma^2=\frac{2(1-p)}{p^2}+\frac{1}{p}-\frac{1}{p^2}=\frac{1}{p^2}-\frac{1}{p}=\frac{1-p}{p^2}$ beware printed answer	1	9
	(ii)	X_1 = no of trials to 1st success X_2 = no of trials to next success X_n = no of trials to nth success } $\therefore\ Y=X_1+X_2+\ldots+X_n$ = total no of trials to the nth success	E2	
		The X_i are independent because the original sequence of trials is.	E1	
		$\therefore$ pgf of Y = (pgf of $X)^n=p^nt^n[1-t(1-p)]^{-n}$	1	
		$\mu_Y=n\mu_X=\frac{n}{p}\quad \sigma_Y^2=n\sigma_X^2=\frac{n(1-p)}{p^2}$ For both	1	5
	(iii)	$N(\mu_n,\sigma_n^2)$ [ft candidate's mean and variance]		1
	(iv)	Y = (unknown) number of pens ordered to get 100 good ones		
		~ random variable as in (ii) [-ve binomial] with $n=100$ and $p=0.75$	1	
		$\div N\left(\frac{100}{0.75}=133.\dot{3},\ \frac{100\times0.25}{0.75^2}=44.\dot{4}\right)$	1	
		$P(Y\geqslant125)\ \approx P\left(N(133.3, 44.4)>124\frac{1}{2}\right)$	M1	
		Do not award if cty corr absent or wrong, but FT if 125 used ↓		
		$=P(N(0,1)>-1.325)$ >-1.25	A1	
		$=0.9074$ 0.8944 cao	A1	5

2.	(i)	From the data, $\bar{x} = 66.775$ [not needed] $\Sigma(x_i - \bar{x})^2 = 30.835$ $s_{n-1} = 2.099$, $s_{n-1}^2 = 4.405$ — Need one of these $s_n = 1.963$, $s_n^2 = 3.854$ [allow only if correctly used in sequel]	B1	
		Test statistic (for test of $\sigma^2 = 14.8$ against $\sigma^2 \neq 14.8$) is		
		$\frac{(n-1)s^2}{\sigma^2} = \frac{7\times 4.405}{14.8} = \frac{30.835}{14.8} = 2.083(45)$	M1 A1	
		Refer to χ_7^2	1	
		Lower 5% point of χ_7^2 is 2.167 (upper is 14.07 – need not be quoted) M1 for quotation of any reasonably sensible lower pt from cand's χ^2 B1 CAO if correct	M1 B1	
		$(2.083 < 2.167)$ significant	1	
		Appears variance is not equal to (and smaller than) 14.8	1	8
	(ii)	$M(\theta) = (1 - 2\theta)^{-m/2}$ Use this to find μ for χ_m^2	M1	
		$\mu = M'(0)$		
		$M'(\theta) = -\frac{m}{2}(1 - 2\theta)^{-\frac{m}{2}-1}(-2) = m(1 - 2\theta)^{-\frac{m}{2}-1}$	1	
		$\therefore \mu = M'(0) = m$ [or by series expansion of $M(\theta)$]	1	
		$\therefore$ we have $E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1$	M1	
		$\therefore E[s^2] = \sigma^2$	1	
		i.e. unbiased	1	6
	(iii)	χ_m^2 : pdf is $f(y) = Ky^{\frac{m-2}{2}}e^{-\frac{y}{2}}$		
		$M(\theta) = E[e^{\theta Y}] = \int_0^\infty Ky^{\frac{m-2}{2}}e^{-y(\frac{1}{2}-\theta)}dy$	M1	
		$= \int_0^\infty K\left(\frac{\frac{1}{2}u}{\frac{1}{2}-\theta}\right)^{\frac{m-2}{2}} e^{-\frac{1}{2}u} \frac{\frac{1}{2}du}{\frac{1}{2}-\theta}$ put $y(\frac{1}{2}-\theta) = \frac{1}{2}u$ For correct substitution, may subdivide	2	
		$= \left(\frac{\frac{1}{2}}{\frac{1}{2}-\theta}\right)^{\frac{m}{2}} \underbrace{\int_0^\infty Ku^{\frac{m-2}{2}}e^{-\frac{u}{2}}du}_{=1}$ for achieving this form	1 1	
		$= \left(\frac{1}{1-2\theta}\right)^{\frac{m}{2}}$ [or various longer methods]	1	6

3	(a)	$n_1 = 11 (\bar{x} = 8.562)$ $s_{n-1} = 0.07236$ $s_{n-1}^2 = 0.005236$ $[s_n = 0.06899(5), s_n^2 = 0.00476]$	B1	
		$n_2 = 9 (\bar{x} = 8.55\dot{6})$ $s_{n-1} = 0.03317$ $s_{n-1}^2 = 0.0011$ $[s_n = 0.01327, s_n^2 = 0.0009\dot{7}]$	B1	
		Test statistic (for test of $\sigma_1^2 = \sigma_2^2$) is $\frac{0.005236}{0.0011} = 4.76$	1	
		Refer to $F_{10,8}$ (1, 1) FT incorrect F distribution	1 1	
		Test is 5% two-sided – need upper 2½% point which is 4.30 (No FT if wrong)	1	
		Significant	1	
		Seems variances are different (and smaller with new machinery)	1	
		Needs Normality	1	
		of both underlying distributions	1	10
	(b)	We have $\hat{p}_1 = \frac{8}{100} = 0.08$, $\hat{p}_2 = \frac{6}{150} = 0.04$	B1	
		95% CI for $p_1 - p_2$ is given by		
		$(0.08 - 0.04) \pm 1.96\sqrt{\frac{(0.08)(0.92)}{100} + \frac{(0.04)(0.96)}{150}}$ M1 B1 M1 two terms, M1 both correct	M1 B1 M1 M1	
		$= 0.04 \pm 1.96\sqrt{0.000992}$		
		$= 0.04 \pm 1.96 \times 0.031496$	A1	
		$= 0.04 \pm 0.0617$		
		$= (-0.022, 0.102)$	A1	
		zero is in this interval, so we would accept, at the 5% level of significance, (Likely key phrases) the null hypothesis that $p_1 - p_2 = 0$ ie that there is no difference between the proportions unsatisfactory	E3	10

4	(i)	We have $\hat{p}=\frac{28}{100}=0.28$		
		Test statistic (for test of $H_0: p = 0.23$ against $H_1: p > 0.23$) is		
		$\dfrac{0.28-\frac{1}{200}-0.23}{\sqrt{\frac{(0.23)(0.77)}{100}}}$ Numerator [award M1 if no cty corr, but FT – value comes to 1.188]	M2	
		Denominator	M1	
		$=\dfrac{0.045}{\sqrt{0.001771}}=\dfrac{0.045}{0.042083}=1.069$	A1	
		Refer to N(0, 1) — No FT if wrong	1	
		Upper single-tailed 2½% point is 1.96 — No FT if wrong	1	
		Not significant	1	
		Seems proportion has not increased	1	8
	(ii)	32 out of 100 would give test statistic $\frac{0.085}{0.042083}=2.02$ [allow this without cty corr: get $\frac{0.09}{0.042083}=2.14$]	M1	
		this is > 1.96	1	
		- so H_0 would be rejected	1	3
	(iii)	OC = P(accept H_0, as a function of p)	M1	
		$= P\left(\hat{p}<0.3125 \mid \hat{p}\sim N\left(p,\frac{p(1-p)}{n}\right)\right)$	M1	
		$= P\left(N(0,1)<\dfrac{0.3125-p}{\sqrt{p(1-p)/100}}\right)=\Phi\left(\dfrac{3.125-10p}{\sqrt{p(1-p)}}\right)$ beware printed answer	1	
		$p = 0.20$ gives Φ (2.8125) = 0.9975 $p = 0.23$ gives Φ (1.96[04] = 0.975 $p = 0.30$ gives Φ (0.2728) = 0.6075 $p = 0.40$ gives Φ (-1.786) = 1 – 0.9630 = 0.0370 If all correct, award A1 if any two are correct	A2	5
	(iv)	OC measures, in a particular way, the 'sensitivity' of the test – it gives the probability that H_0 will be accepted for any value of p. In particular, it gives the probability of failing to detect an increase in the proportion that has actually occurred.	E2	
		Reward sensible discussion here, but it does not look very good that the OC is as high as 0.6075 for p as high as 0.3 (remember H_0 is $p = 0.23$)	E2	4

Examiner’s Report

2617 Statistics 5

General Comments

There were 35 candidates from 10 centres, a further reduction in numbers taking this paper.

Though there was, as usual, a lot of good work, quite a few candidates had unwonted difficulties with the algebra in questions 1 and 2. This is disappointing, and begs some questions. The algebra was *not* difficult, and some of it was standard bookwork. Candidates entering at this level really ought to be able to cope.

Comments on Individual Questions

Q.1 Probability generating functions, based on geometric and negative binomial distributions

This was a reasonably popular question. Some candidates produced very good solutions, and virtually everyone who attempted it was able to get started. In some cases, however, the aforementioned difficulties with algebra eventually overwhelmed the work. The quoted probability generating function (pgf) and mean and variance for the geometric distribution in part (i) were usually found satisfactorily, with a wide variety of techniques being used to find the derivative and second derivative of the pgf. In part (ii), the required explanation was not always completely secure; many candidates did not explain the importance of the independence of the trials properly (or at all). The resulting pgf was usually written down correctly [the nth power of that obtained in part (i)] but some candidates then used it, at considerable length, to derive the mean and variance instead of following the instruction in the question to *write down* these quantities; they should have been immediately obvious as the Y variable is the sum of n independent X variables. Sadly, a few candidates did the write-down wrongly, usually with a variance of something like σ^2/n instead of $n\sigma^2$. In part (iii), the Normal approximation was usually an immediate write-down (though the mean and variance were required to be given as well, otherwise there are (bivariately) infinitely many possibilities!). Part (iv) caused difficulties. Despite having been led all the way through the question to get to this part, several candidates resorted to some form of binomial analysis, and other errors occurred too. The correct analysis is based on a *negative binomial* random variable as in part (ii) with $n = 100$ and $p = 0.75$, with the Normal approximation from part (iii) used to find the probability of being greater than or equal to 125 (with a continuity correction, leading to 124.5); the final answer is 0.9074.

Q.2 Chi-squared test for variance, with underlying theoretical work based on the moment generating function

The test in part (i) was usually well done [value of test statistic is 2.083, refer to chi-squared with 7 degrees of freedom, consider lower 5% point which is 2.167, result is significant]. In part (ii), candidates knew that moments (mean, variance, etc) can be obtained from the moment generating function by differentiating (or by considering the power series expansion), but some candidates did not appreciate that it was, simply, the *mean* of the underlying distribution that was required, not the variance. Presumably these candidates were confused by the fact that we are dealing with the distribution of the sample *variance*, but it is disturbing that they did not realise that it was the *mean* of that distribution that was required. Part (iii) was often well done, and it was pleasing to see a lot of good careful work – though there were diverse errors too!

Q.3 *F* test for comparing variances and a confidence interval for a difference in proportions

This question combined these two techniques in the context of installation of new machinery on a production line. The F test was generally well done [value of test statistic is 4.76, refer to F with 10 and 8 degrees of freedom], most candidates knowing that the upper 2½% point [4.30] was needed to give a 5% test in the two-sided situation. The confidence interval for the difference in proportions was also often well done [answer is (–0.022, 0.102)], but some candidates had an incorrect standard error term. At the end of the question, an interpretation was asked for *in terms of the claim made for the new machinery*. Thus the usual general interpretative statement, to the effect that 95% of all such intervals would contain the true difference, was inadequate on this occasion – an interpretation *in* *context* was explicitly asked for. Not all candidates appreciated this. Those who did try to answer the question as posed expressed themselves in a variety of ways, but their intentions were generally clear. The published mark scheme suggests key phrases about zero being in the interval so that a null hypothesis of no difference would have been accepted at the 5% level of significance. This is of course only a summary for marking purposes; all relevant intelligent comments were accepted.

Q.4 Test for a binomial p parameter; operating characteristic based on this test

The test in part (i) was usually well done [value of test statistic is 1.069 (with use of continuity correction), refer to N(0,1)], and in part (ii) candidates quickly worked through the same analysis with different data to get a test statistic of 2.02 (or 2.14 without continuity correction) which, unlike that in part (i), was significant at the 5% level. In part (iii), candidates had to derive a quoted expression for the operating characteristic of the test. As is so often the case, it appeared in many scripts that the quoted answer was *extremely* helpful in guiding candidates' progress, especially with the clear importance of the number 3.125 which could readily be seen to be the 0.3125 given in the text of the question multiplied by 10 (and there were plenty of 10s, often as $\sqrt{100}$, available in the surrounding work). It should at once be said that there were candidates who clearly knew exactly what they were doing and would undoubtedly have derived the correct expression even if it had not been quoted. Marks were of course awarded if explanations were convincing. It was surprising that not everyone found the numerical evaluations of the expression to be trivial [the probabilities are 0.9975, 0.975, 0.6075, 0.0370]. The interpretations of the operating characteristic (OC) in part (iv) were on the whole fairly good without being perfect. Many candidates talked about the OC being the probability of wrongly accepting the null hypothesis (i.e. the probability of a Type II error), but this is not complete: the OC gives the probability of accepting the null hypothesis for *all* values of the parameter, including those for which the null hypothesis *ought to be* accepted, i.e. there is not an error at all. Finally, whether the test is behaving well is a matter of opinion, and all reasonably justified opinions were accepted; however, a value of the OC as high as 0.6075 for $p = 0.3$ ought probably to ring alarm bells.