OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE **4734/01**

MATHEMATICS

Probability & Statistics 3

TUESDAY 15 JANUARY 2008 Morning

Time: 1 hour 30 minutes

Additional materials: Answer Booklet (8 pages)
List of Formulae (MF1)

INSTRUCTIONS TO CANDIDATES

- Write your name, 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.
- Give non-exact numerical answers correct to 3 significant figures unless a different degree of accuracy is specified in the question or is clearly appropriate.
- You are permitted to use a graphical calculator in this paper.

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 reminded of the need for clear presentation in your answers.**

This document consists of **4** printed pages.

[Turn over

1 A blueberry farmer increased the amount of water sprayed over his berries to see what effect this had on their weight. The farmer weighed each of a random sample of 80 berries of the previous season's crop and each of a random sample of 100 berries of the new crop. The results are summarised in the following table, in which $\bar{x}$ denotes the sample mean weight in grams, and s^2 denotes an unbiased estimate of the relevant population variance.

	Sample size	$\bar{x}$	s^2
Previous season's crop (P)	80	1.24	0.003 56
New crop (N)	100	1.36	0.003 40

(i) Calculate an estimate of $\text{Var}(\bar{X}_N - \bar{X}_P)$. [2]

(ii) Calculate a 95% confidence interval for the difference in population mean weights. [3]

(iii) Give a reason why it is unnecessary to use a t-distribution in calculating the confidence interval. [1]

2 The times taken for customers' phone complaints to be handled were monitored regularly by a company. During a particular week a researcher checked a random sample of 20 complaints and the times, x minutes, taken to handle the complaints are summarised by $\Sigma x = 337.5$. Handling times may be assumed to have a normal distribution with mean μ minutes and standard deviation 3.8 minutes.

(i) Calculate a 98% confidence interval for μ. [4]

During the same week two other researchers each calculated a 98% confidence interval for μ based on independent samples.

(ii) Calculate the probability that at least one of the three intervals does not contain μ. [2]

(iii) State two ways in which the calculation in part **(i)** would differ if the standard deviation were unknown. [2]

3 A transport authority wished to compare the performance of two rail companies, Western and Northern. They noted that the number of 'on-time' arrivals for a random sample of 80 Western trains over a particular route was 71. The corresponding number for a random sample of 90 Northern trains over a similar route was 73.

(i) Test, at the 5% significance level, whether the population proportion of on-time Western trains exceeds the population proportion of on-time Northern trains. [7]

(ii) Ranjit wishes to test whether the population proportion of on-time Western trains exceeds the population proportion of on-time Northern trains by more than 0.01. What variance estimate should she use? [2]

4 Eezimix flour is sold in small bags of weight S grams, where $S \sim N(502.1, 0.31^2)$. It is also sold in large bags of weight L grams, where $L \sim N(1004.9, 0.58^2)$.

(i) Find the probability that a randomly chosen large bag weighs at least 1 gram more than two randomly chosen small bags. [6]

(ii) Find the probability that a randomly chosen large bag weighs less than twice the weight of a randomly chosen small bag. [5]

5 Of two brands of lawnmower, *A* and *B*, brand *A* was claimed to take less time, on average, than brand *B* to mow similar stretches of lawn. In order to test this claim, 9 randomly selected gardeners were each given the task of mowing two regions of lawn, one with each brand of mower. All the regions had the same size and shape and had grass of the same height. The times taken, in seconds, are given in the table.

Gardener	1	2	3	4	5	6	7	8	9
Brand *A*	412	386	389	401	396	394	397	411	391
Brand *B*	422	394	385	408	394	399	397	410	397

(i) Test the claim using a paired-sample t-test at the 5% significance level. State a distributional assumption required for the test to be valid. [10]

(ii) Give a reason why a paired-sample t-test should be used, rather than a 2-sample t-test, in this case. [1]

[Questions 6 and 7 are printed overleaf.]

6 The Research and Development department of a paint manufacturer has produced paint of three different shades of grey, G_1, G_2 and G_3. In order to find the reaction of the public to these shades, each of a random sample of 120 people was asked to state which shade they preferred. The results, classified by gender, are shown in Table 1.

		Shade		
		G_1	G_2	G_3
Gender	Male	11	24	23
	Female	18	13	31

Table 1

Table 2 shows the corresponding expected values, correct to 2 decimal places, for a test of independence.

		Shade		
		G_1	G_2	G_3
Gender	Male	14.02	17.88	26.10
	Female	14.98	19.12	27.90

Table 2

(i) Show how the value 17.88 for Male, G_2 was obtained. [2]

(ii) Test, at the 5% significance level, whether gender and preferred shade are independent. [7]

(iii) Determine the smallest significance level obtained from tables or calculator for which there is evidence that not all shades are equally preferred by people in general, irrespective of gender. [6]

7 The continuous random variable T has probability density function given by

$$f(t) = \begin{cases} 4t^3 & 0 < t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Obtain the cumulative distribution function of T. [2]

(ii) Find the cumulative distribution function of H, where $H = \frac{1}{T^4}$, and hence show that the probability density function of H is given by $g(h) = \frac{1}{h^2}$ over an interval to be stated. [7]

(iii) Find $E(1 + 2H^{-1})$. [3]

4734 Probability & Statistics 3

1(i)	$s^2 = 0.00356/80+0.00340/100$ $= 7.85 \times 10^{-5}$	M1 A1 **2**	Sum of variances Or pooled, giving 7.81×10^{-5}
(ii)	---	---	---
	$(1.36-1.24) \pm zs$ z=1.96 (0.103, 0.137)	M1 B1 A1 **3**	Must be s, accept t
	---	---	---
(iii)	Not necessary since sample sizes are large	B1 **1 (6)**	Or equivalent. Nothing wrong
2 (i)	Use $\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$	M1	
	$\bar{x} = 337.5 / 20$	**B1** B1	
	z =2.326 (14.9,18.9)	A1 4	3 or 4 SF
	---	---	---
(ii)	$1- 0.98^3$ 0.0588	M1 A1 **2**	Use B(3,0.02) or B(3,0.98) for M.
	---	---	---
(iii)	Unbiased estimate of σ^2 required t – distribution used to obtain CV	B1 B1 2 **(8)**	
3 (i)	H_0: $p_W = p_N$, H_1: $p_W > p_N$	B1	For both hypotheses. Or π.
	Pooled $\hat{p} = \dfrac{71+73}{80+90}$ $\left(=\dfrac{144}{170}\right)$	B1	SR: from $p_1q_1/n_1 + p_2q_2/n_2 = 0.00295$ z = 1.406 B1M1A1M1A1 Max 5/7
	s^2 = (144/170)(26/170)(1/80+1/90)	B1	
	z = (71/80-73/90)/s	M1	
	=1.381	A1	If no explicit comparison and correct conclusion then M1A0.
	1.381 < 1.645 Do not reject H_0, there is insufficient evidence that the proportion of on-time Western trains exceeds the proportion of on-time Northern trains	M1 A1 **7**	Or use P-value or CR In context, not too assertive
(ii)	---	---	---
	s^2= 71×9/80^3+73×17/90^3 = 0.00295	M1 A1 **2 (9)**	AEF Allow one error Accept 0.0029
4(i)	Use $L - S_1 - S_2$	M1	Or equivalent, or implied
	μ = 0.7	B1	
	σ^2 = 0.58^2+ 0.31^2+0.31^2 = 0.5286	M1 A1	May be implied later
	(1-0.7)/σ	M1	Correct numerator
	0.340	A1 **6**	
	---	---	---
(ii)	Use $L - 2S$ with μ=0.7	M*1	M0 if as (i) unless correct
	σ^2 = 0.58^2 + 4$(0.31)^2$	B1	
	- 0.7/ σ	Dep*M1	Accept +
	- 0.824(5)	A1	
	0.2048	A1 **5** **(11)**	0.205 (3SF)

5(i)	Population of differences is normal H_0:$\mu_A = \mu_B$, H_1: $\mu_A < \mu_B$ where μ_A and μ_B denote the population means $\bar{x}_D = 3.222$ $s_D = 5.019$ $t = 3.222/(5.019/3)$ $=1.926$ CV = 1.860 1.926 > 1.860 Reject H_0 , there is evidence that brand *A* takes less time than brand *B* ----------	B1 B1 B1 M1A1 M1 A1 B1 M1 A1 **10** ----------	*Not "independent"* Or $\mu_D = 0$, $\mu_D > 0$ From formula ,or B2 from calculator Accept 1.93. M1A0 if t = - 1.926 ----------
(ii)	One valid reason	B1 **1 (11)**	Data are clearly paired Data not independent
6(i)	37×58/120 17.883.. , 17.88 AG ----------	M1 A1 **2** ----------	Or equivalent ----------
(ii)	H_0: Gender and shade are independent (H_1:--are not independent $3.02^2(14.02^{-1}+14.98^{-1})$ + $6.12^2(17.88^{-1}+19.12^{-1})$ $+3.1^2(26.1^{-1}+27.9^{-1})$ =6.03 EITHER: CV 5.991 6.03 > 5.991, reject H_0 and accept that gender and shade are not independent OR: $P(\chi^2 > 6.03) = 0.049$ < 0.05 , reject H_0 and accept that gender and shade are not independent ----------	B1 M1 A1 A1 B1 M1 A1√ **7** B1 M1 A1√ ----------	 At least two correct All correct Ft X^2 . Can be assertive. Ft X^2 ----------
(iii)	G1 G2 G3 O 29 37 54 E 40 40 40 121/40 + 9/40+196/40 = 8.15 Using df = 2 2.5% tables, 1.7% calculator	 M1 A1 M1 A1 M1 A1 **6 (15)**	 For combining

7(i)	$F(t) = \begin{cases} 0 & t \le 0, \\ t^4 & 0 < t \le 1, \\ 1 & \text{otherwise.} \end{cases}$	B1 B1 **2**	For t^4 For rest
	---	---	---
(ii)	$G(h) = P(H \le h)$ $= P(T \ge 1/h^{1/4})$ $= 1 - F((1/h^{1/4})$ $= 1 - 1/h$ $g(h) = G'(h)$ $= 1/h^2$ $h \ge 1$,(0 otherwise)	M1 A1 A1 A1 M1 A1 **B**1 **7**	Accept < With attempt at differentiation Only from G obtained correctly
	---	---	---
(iii)	EITHER: $\int_1^\infty (h^{-2} + 2h^{-3})\mathrm{d}h$ $= \left[-h^{-1} - h^{-2}\right]_1^\infty$ $= 2$ OR: $= 1 + 2\int_1^\infty \frac{1}{h^3}\mathrm{d}h$ $= 1 + 2\left[-\frac{1}{2h^2}\right]_1^\infty$ $= 2$ OR: $E(1+2T^4) = 1 + \int_0^1 8t^7\,\mathrm{d}t$ $= 1 + [t^8]$ $= 2$	M1 B1 A1 M1 B1 A1 M1 B1 A1 **3 (12)**	For integrating $(1+2h^{-1})g(x)$, with limits from (ii) Limits not required Limits not required Limits not required