EXAMINATION

9 October 2009 (am)

Subject CT3 — Probability and Mathematical Statistics Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is not required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 In a sample of 100 households in a specific city, the following distribution of number of people per household was observed:

Number of people x	1	2	3	4	5	6	7
Number of households f_x	7	f_2	20	f_4	18	10	5

The mean number of people per household was found to be 4.0. However, the frequencies for two and four members per household (f_2 and f_4 respectively) are missing.

(i) Calculate the missing frequencies f_2 and f_4. [2]

(ii) Find the median of these data, and hence comment on the symmetry of the data. [2]

[Total 4]

2 Two tickets are selected at random, one after the other and without replacement, from a group of six tickets, numbered 1, 2, 3, 4, 5, and 6.

(i) Calculate the probability that the numbers on the selected tickets add up to 8. [2]

(ii) Calculate the probability that the numbers on the selected tickets differ by 3 or more. [2]

[Total 4]

3 Let X be a random variable with moment generating function $M_X(t)$ and cumulant generating function $C_X(t)$, and let $Y = aX + b$, where a and b are constants. Let Y have moment generating function $M_Y(t)$ and cumulant generating function $C_Y(t)$.

(i) Show that $C_Y(t) = bt + C_X(at)$. [2]

(ii) Find the coefficient of skewness of Y in the case that $M_X(t) = (1 - t)^{-2}$ and $Y = 3X + 2$ (you may use the fact that $C_Y'''(0) = E[(Y - \mu_Y)^3]$). [5]

[Total 7]

4 Let the random variables (X,Y) have the joint probability density function

$$f_{X,Y}(x,y) = \exp\{-(x+y)\}, \quad x>0, y>0.$$

(i) Derive the marginal probability density functions of X and Y, and hence determine (giving reasons) whether or not the two variables are independent. [3]

(ii) Derive the joint cumulative distribution function $F_{X,Y}(x,y)$. [2]

[Total 5]

5 Let X be a random variable with probability density function given by

$$f(x) = 2x\theta^{-2}, \quad 0<x<\theta.$$

Find an unbiased estimator of θ, based on a single observation of X. [4]

6 A random sample of size n is taken from an exponential distribution with parameter λ, that is, with probability density function

$$f(x) = \lambda e^{-\lambda x}, \quad 0<x<\infty.$$

(i) Determine the maximum likelihood estimator (MLE) of λ. [3]

Claim sizes for certain policies are modelled using an exponential distribution with parameter λ. A random sample of such claims results in the value of the MLE of λ as $\hat{\lambda} = 0.00124$.

A large claim is defined as one greater than £4,000 and the claims manager is particularly interested in p, the probability that a claim is a large claim.

(ii) Determine $\hat{p}$, the MLE of p, explaining why it is the MLE. [3]

[Total 6]

7 A scientific investigation involves a linear regression with the usual assumptions that the response variable y follows a normal distribution with mean $\alpha + \beta x$ and variance σ^2. Twenty data points were recorded, corresponding to four observations of y at $x = 1$, three observations of y at $x = 2$, six observations of y at $x = 3$, and seven observations of y at $x = 4$. The resulting means of these sets of y observations are given in the table below.

x	1	2	3	4
no. of y's	4	3	6	7
mean of y's	18.6	21.7	23.2	27.1

(i) Determine the fitted regression line of y on x. [5]

(ii) Suppose that you have been asked to provide a 95% confidence interval for the slope coefficient.

(a) Comment briefly on any problems you might encounter in the computation of the required confidence interval.

(b) Indicate briefly any further information that you would need in order to overcome these problems.

[3]
[Total 8]

8 The table below shows a bivariate probability distribution for two discrete random variables X and Y:

	$X = 0$	$X = 1$	$X = 2$
$Y = 1$	0.15	0.20	0.25
$Y = 2$	0.05	0.15	0.20

Find the value of $E[X|Y = 2]$. [3]

9 In a group of motor insurance policies issued by a company, 80% of claims are made on comprehensive policies and 20% are made on third-party-only policies.

(i) Calculate the average amount paid out on a claim, given that the average amount paid out by the company on a comprehensive policy claim is £1,650, and the average amount paid out on a third-party-only policy claim is £625. [1]

(ii) Calculate the total expected amount paid out in claims by the company in one year, given that the total number of policies is 150,000 and, on average, the claim rate is 0.15 claims per policy per year. [2]
[Total 3]

10 Consider a population in which a proportion θ of members have some specified characteristic. Let P denote the corresponding proportion of members in a random sample of size n from the population.

(i) Explain clearly why the mean and standard error of P are given by

$$E[P] = \theta, \quad s.e.[P] = \sqrt{\frac{\theta(1-\theta)}{n}}\,. \qquad [3]$$

An insurance company uses a questionnaire to monitor the satisfaction of its customers.

In one part customers are asked to answer "yes" or "no" to a particular question.

Suppose that a random sample of 200 responses is examined.

(ii) Calculate the approximate probability that at least 150 "yes" answers are found in the sample, on the assumption that the true (population) proportion of "yes" answers is 0.7. [4]

Suppose the true (population) proportion of "yes" answers (θ) is unknown, and for a random sample of 200 responses, the number of "yes" answers is found to be 146.

(iii) (a) Calculate an upper (one-sided) 95% confidence interval of the form $(0, L)$ for θ.

(b) Calculate a lower (one-sided) 95% confidence interval of the form $(L, 1)$ for θ.

(c) A test of the hypotheses:

$$H_0: \theta = 0.7 \text{ v } H_1: \theta > 0.7$$

results in a P-value of 0.198.

Comment on how this result relates to the confidence interval in part (iii)(b). [9]

[Total 16]

11 Three insurance company colleagues had just completed an investigation which involved the application of a two-sample *t*-test to compare two independent samples, each of size 11. They were concerned about the validity of the equal variance assumption required for this test. Their data were as follows.

A:	21	22	28	27	20	23	26	32	25	21	30
B:	19	18	38	33	24	39	22	20	28	26	30

$\Sigma x_A = 275, \quad \Sigma x_A^2 = 7,033, \quad \Sigma x_B = 297, \quad \Sigma x_B^2 = 8,559$

(i) One of the colleagues suggested a graphical approach for the comparison of the variances.

Draw a suitable diagram to represent these data so that the variability of the samples can be compared, and comment briefly on that comparison. [3]

(ii) Another colleague suggested using an *F*-test for the comparison of variances.

(a) Perform this *F*-test at the 5% level to compare the variances and express your conclusion clearly.

(b) In addition obtain an approximate value of the *P*-value for this test by linearly interpolating between suitable entries in the tables.

[6]

(iii) The third colleague suggested another procedure using a two-sample *t*-test in the following way:

> "For each sample calculate the absolute values of the deviations of the observations from the mean of that sample; then apply a two-sample *t*-test to the two sets of absolute deviations."

(a) Discuss the possible reasoning behind this suggested procedure by considering the potential values of such absolute deviations when the assumption of equal variances is valid and when it is not valid.

(b) (1) Calculate the required sets of absolute deviations for the given data.

(2) Perform the suggested two-sample *t*-test at the 5% level stating your conclusion clearly.

(3) Obtain, in addition, an approximate *P*-value for this test by linearly interpolating between suitable entries in the tables.

[11]

(iv) Comment briefly on the conclusions that may have been reached by the three colleagues. [2]

[Total 22]

12 A bank has a free telephone number for its customer services department. Often the call volume is heavy and customers are placed on hold until a staff member is available to answer. The bank hopes that a caller remains on hold until the call is answered, so as not to upset or lose an existing or potential customer.

A survey was conducted to analyse whether callers would remain on hold longer (on average), if they heard a recorded message containing: (*A*) an advertisement about the bank's products; (*B*) "easy listening" music; or (*C*) classical music. The bank randomly selected a sample of five unanswered calls under each recorded message, and the length of time (in minutes) that the caller remained on hold before hanging up is given in the table below.

Recorded message	Time					Total
A: advertisement	5	1	11	2	8	27
B: *easy listening music*	0	1	4	6	3	14
C: *classical music*	13	9	8	15	7	52

For these data $\Sigma y = 93$, $\Sigma y^2 = 865$

Let μ_A, μ_B, μ_C denote the mean telephone holding times under recorded message *A*, *B* and *C* respectively.

(i) (a) Perform an analysis of variance to test the hypothesis that the nature of the recorded message has no effect on the length of time that callers remain on hold. You should construct an appropriate ANOVA table and state your conclusion clearly.

(b) Calculate a 95% confidence interval for $\mu_A - \mu_C$, using information available from all three samples.

[11]

An equivalent approach for analysing the effects of the recorded messages on holding time is the following:

consider the regression model $E[Y_i] = a + b_1 x_{i1} + b_2 x_{i2}$, $i = 1, 2, \ldots, 15$, where Y_i is the telephone holding time and x_{i1}, x_{i2} are indicator variables such that $x_{i1} = 1$ if the message for caller *i* contains an advertisement (and 0 otherwise), and $x_{i2} = 1$ if the message contains easy listening music (and 0 otherwise).

The results from fitting this model are given below:

	Coef.	*Std. Error*	*t-value*	*p-value*
Intercept	10.400	1.523	6.828	$1.8 * 10^{-5}$
x_1	−5.000	2.154	−2.321	0.039
x_2	−7.600	2.154	−3.528	0.004

$s = 3.406$ R-sq = 51.7%

(ii) Using the fitted model:

(a) Calculate the predicted value for the telephone holding time when the message contains classical music.

(b) Test the hypothesis H_0: $b_1 = 0$ against H_1: $b_1 \neq 0$ at the 5% level of significance.

(c) Derive an expression relating b_1 with μ_A and μ_C, and hence verify your result from the test in (ii)(b) using the confidence interval obtained in (i)(b).

[7]
[Total 18]

END OF PAPER