EXAMINATION

23 April 2007 (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 13 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is 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.

1 Consider the following two random samples of ten observations which come from the distributions of random variables which assume non-negative integer values only.

Sample 1: 7 4 6 11 5 9 8 3 5 5
sample mean = 6.3, sample variance = 6.01

Sample 2: 8 3 5 11 2 4 6 12 3 9
sample mean = 6.3, sample variance = 12.46

One sample comes from a Poisson distribution, the other does not.

State, with brief reasons, which sample you think is likely to be which. [2]

2 A random sample of 200 policy surrender values (in units of £1,000) yields a mean of 43.6 and a standard deviation of 82.2.

Determine a 99% confidence interval for the true underlying mean surrender value for such policies. [3]

3 It is assumed that claims on a certain type of policy arise as a Poisson process with claim rate λ per year.

For a group of 150 independent policies of this type, the total number of claims during the last calendar year was recorded as 123.

Determine an approximate 95% confidence interval for the true underlying annual claim rate for such a policy. [4]

4 The sample correlation coefficient for the set of data consisting of the three pairs of values

$$(-1,-2)\ ,\ (0,0)\ ,\ (1,1)$$

is 0.982. After the x and y values have been transformed by particular linear functions, the data become:

$$(2,2)\ ,\ (6,-4)\ ,\ (10,-7).$$

State (or calculate) the correlation coefficient for the transformed data. [2]

5 The number of claims arising in one year from a group of policies follows a Poisson distribution with mean 12. The claim sizes independently follow an exponential distribution with mean £80 and they are independent of the number of claims.

The current financial year has six months remaining.

Calculate the mean and the standard deviation of the total claim amount which arises during this remaining six months. [4]

6 Consider the discrete random variable X with probability function

$$f(x) = \frac{4}{5^{x+1}}, \quad x = 0, 1, 2, ...$$

(i) Show that the moment generating function of the distribution of X is given by

$$M_X(t) = 4(5 - e^t)^{-1},$$

for $e^t < 5$. [3]

(ii) Determine $E[X]$ using the moment generating function given in part (i). [3]
[Total 6]

7 A charity issues a large number of certificates each costing £10 and each being repayable one year after issue. Of these certificates, 1% are randomly selected to receive a prize of £10 such that they are repaid as £20. The remaining 99% are repaid at their face value of £10.

(i) Show that the mean and standard deviation of the sum repaid for a single purchased certificate are £10.1 and £0.995 respectively. [2]

Consider a person who purchases 200 of these certificates.

(ii) Calculate approximately the probability that this person is repaid more than £2,040 by using the Central Limit Theorem applied to the total sum repaid. [3]

(iii) An alternative approach to approximating the probability in (ii) above is based on the number of prize certificates the person is found to hold. This number will follow a binomial distribution.

Use a Poisson approximation to this binomial distribution to approximate the probability that this person is repaid more than £2,040. [3]

(iv) Comment briefly on the comparison of the two approximations above given that the exact probability using the binomial distribution is 0.0517. [1]
[Total 9]

8 A random sample of size n is taken from a distribution with probability density function

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}}, \quad 0 < x < \infty$$

where α is a parameter such that $\alpha > 0$.

(i) Show by evaluating the appropriate integral that, in the case $\alpha > 1$, the mean of this distribution is given by $\frac{1}{\alpha - 1}$.

[*Hint:* when integrating, write $x = (1 + x) - 1$ and exploit the fact that the integral of a density function is unity over its full range.] [3]

(ii) Determine the method of moments estimator of α. [2]
[Total 5]

9 Consider three random variables X, Y, and Z with the same variance $\sigma^2 = 4$. Suppose that X is independent of both Y and Z, but Y and Z are correlated, with correlation coefficient $\rho_{YZ} = 0.5$.

(i) Calculate the covariance between X and U, where $U = Y+Z$. [1]

(ii) Calculate the covariance between Z and V, where $V = 3X - 2Y$. [2]

(iii) Calculate the variance of W, where $W = 3X - 2Y + Z$. [2]
[Total 5]

10 A random sample of insurance policies of a certain type was examined for each of four insurance companies and the sums insured (y_{ij}, for companies $i = 1, 2, 3, 4$) under each policy are given in the table below (in units of £100):

Company						*Total*
1	58.2	57.2	58.4	55.8	54.9	284.5
2	56.3	54.5	57.0	55.3		223.1
3	50.1	54.2	55.5			159.8
4	52.9	49.9	50.0	51.7		204.5

For these data, $\sum_i \sum_j y_{ij} = 871.9$ and $\sum_i \sum_j y_{ij}^2 = 47{,}633.53$

Consider the ANOVA model $Y_{ij} = \mu + \tau_i + e_{ij}, i = 1,...,4, j = 1,...,n_i$, where Y_{ij} is the jth sum insured for company i, n_i is the number of responses for company i, $e_{ij} \sim N(0, \sigma^2)$ are independent errors, and $\sum_{i=1}^{4} n_i \tau_i = 0$.

(i) Calculate estimates of the parameters μ and τ_i, $i = 1, 2, 3, 4$. [3]

(ii) Test the hypothesis that there are no differences in the means of the sums insured under such policies by the four companies. [5]

[Total 8]

11 The number of claims, X, which arise in a year on each policy of a particular class is to be modelled as a Poisson random variable with mean λ. Let $\underline{X} = (X_1, X_2, \ldots, X_n)$ be a random sample of size n from the distribution of X, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Suppose that it is required to estimate λ, the mean number of claims on a policy.

(i) Show that $\hat{\lambda}$, the maximum likelihood estimator of λ, is given by $\hat{\lambda} = \bar{X}$. [3]

(ii) Derive the Cramer-Rao lower bound (CRlb) for the variance of unbiased estimators of λ. [4]

(iii) (a) Show that $\hat{\lambda}$ is unbiased for λ and that it attains the CRlb.

(b) Explain clearly why, in the case that n is large, the distribution of $\hat{\lambda}$ can be approximated by

$$\hat{\lambda} \sim N\left(\lambda, \frac{\lambda}{n}\right).$$

[3]

(iv) (a) Show that, in the case $n = 100$, an approximate 95% confidence interval for λ is given by

$$\bar{x} \pm 0.196\sqrt{\bar{x}}.$$

(b) Evaluate the confidence interval in (iv)(a) based on a sample with the following composition:

observation	0	1	2	3	4	5	6	7
frequency	11	28	19	28	9	2	2	1

[6]

[Total 16]

12 An insurance company is investigating past data for two household claims assessors, A and B, used by the company. In particular claims resulting from similar types of water damage were extracted. The following table shows the assessors' initial estimates of the cost (in units of £100) of meeting each claim.

A:	4.6	6.6	2.8	5.8	2.1	5.2	5.9	3.4	7.8	3.5	1.6	8.6	2.7
B:	5.7	3.4	4.7	3.6	6.5	3.3	3.8	2.4	7.0	4.0	4.4		

for the A data: $n_A = 13$, $\Sigma x = 60.6$ and $\Sigma x^2 = 340.92$

for the B data: $n_B = 11$, $\Sigma x = 48.8$ and $\Sigma x^2 = 236.80$

(i) Draw a suitable diagram to represent these data so that the initial estimates of the two assessors can be compared. [2]

(ii) You are required to perform an appropriate test to compare the means of the assessors' initial estimates for this type of water damage.

(a) State your hypotheses clearly.

(b) Use your diagram in part (i) to comment briefly on the validity of your test.

(c) Calculate your test statistic and specify the resulting P-value.

(d) State your conclusion clearly.

[10]

(iii) You are required to perform an appropriate test to compare the variances of the assessors' initial estimates.

(a) State your hypotheses clearly.

(b) Use your diagram in part (i) to comment briefly on the validity of your test.

(c) Calculate your test statistic and specify the resulting P-value.

(d) State your conclusion clearly and hence comment further on the validity of your test in part (ii).

[7]

(iv) Use your answers in parts (i) to (iii) to comment on the overall comparison of the two assessors as regards their initial estimates for this type of water damage. [1]

[Total 20]

13 In a study of the relation between the amount of information available and use of buses in eight comparable test cities, bus route maps were given to residents of the cities at the beginning of the test period. The increase in average daily bus use during the test period was recorded. The numbers of maps and the increase in bus use are given in the table below (both in thousands).

Number of maps (x)	80	220	140	120	180	100	200	160
Increase in bus use (y)	0.60	6.70	5.30	4.00	6.55	2.15	6.60	5.75

For these data:

$$\sum x = 1{,}200, \ \sum x^2 = 196{,}800, \ \sum y = 37.65, \ \sum y^2 = 213.4875, \ \sum xy = 6{,}378$$

(i) Construct a scatterplot of the data and comment on the relationship between the increase in bus use and the number of maps distributed. [4]

(ii) The equation of the fitted linear regression is given by $y = -1.816 + 0.04348x$. Perform an appropriate statistical test to assess the hypothesis that the slope in this fitted model suggests no relationship between the increase in bus use and the number of maps distributed. Any assumptions made should be clearly stated. [6]

(iii) The fitted responses and the residuals from the linear regression model fitted in part (ii) are given below:

Fitted values ($\hat{y}$)	1.66	7.75	4.27	3.40	6.01	2.53	6.88	5.14
Residuals ($\hat{e}$)	-1.06	-1.05	1.03	0.60	0.54	-0.38	-0.28	0.61

Plot the residuals against the values of the fitted responses and comment on the adequacy of the model. [4]

(iv) A new city is added to the study, and 250,000 maps are distributed to its citizens.

Calculate the prediction of the increase in bus use in this city according to the model fitted in part (ii) and comment on the validity of this prediction. [2]
[Total 16]

END OF PAPER