EXAMINATION

13 September 2005 (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 14 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 The table below gives the number of thunderstorms reported in a particular summer month by 100 meteorological stations.

Number of thunderstorms:	0	1	2	3	4	5
Number of stations:	22	37	20	13	6	2

(a) Calculate the sample mean number of thunderstorms.

(b) Calculate the sample median number of thunderstorms.

(c) Comment briefly on the comparison of the mean and the median.

[3]

2 In an opinion poll, each individual in a random sample of 275 individuals from a large population is asked which political party he/she supports. If 45% of the population support party A, calculate (approximately) the probability that at least 116 of the sample support A. [3]

3 Claim amounts on a certain type of policy are modelled as following a gamma distribution with parameters $\alpha = 120$ and $\lambda = 1.2$.

Calculate an approximate value for the probability that an individual claim amount exceeds 120, giving a reason for the approach you use. [3]

4 Calculate a 99% confidence interval for the percentage of claims for household accidental damage which are fully settled within six months of being submitted, given that in a random sample of 100 submitted claims of this type, exactly 83 were fully settled within six months of being submitted. [3]

5 A random sample of 500 claim amounts resulted in a mean of £237 and a standard deviation of £137.

Calculate an approximate 95% confidence interval for the true underlying mean claim amount for such claims, explaining why the normal distribution can be used.

[3]

6 A random sample of 16 observations was selected from each of four populations. Given that the between treatments mean square is 280 and the total sum of squares is $SS_T = 1{,}500$, construct an analysis of variance table and test the null hypothesis that the means of the four populations are equal. [3]

7 A sample of 20 claim amounts (£) on a group of household policies gave the following data summaries:

$\Sigma x = 3{,}256$ and $\Sigma x^2 = 866{,}600$.

(a) Calculate the sample mean and standard deviation for these claim amounts.

(b) Comment on the skewness of the distribution of these claim amounts, giving reasons for your answer. [4]

8 A simple procedure for incorporating a "no claims discount" into an annual insurance policy is as follows:

a premium of £400 is payable for the first year;

if no claim is made in the first year, the premium for the second year is £400k, where k is a constant such that $0 < k < 1$;

if no claim is made in the first and second years, the premium for the third year is £400k^2;

if no subsequent claims are made in future years, the premium remains as £400k^2;

if a claim is made, the premium the following year reverts to £400 and the procedure starts again as above.

(i) Show that the probability distribution of the premium for the fourth year, that is, for the year following the third year, is given by

£400	with probability	p
£400k	with probability	$p(1-p)$
£400k^2	with probability	$(1-p)^2$

where p is the probability of a claim being made in any year. [3]

(ii) Obtain an expression for the expected premium for the fourth year under this procedure. [1]

(iii) If it is desired that this expected premium should equal £300, determine the required value of k for the case where $p = 0.1$. [3]

[Total 7]

9 Random samples are taken from two independent normally distributed populations (with means μ_1 and μ_2 respectively) with the following results.

Sample from population 1:

sample size $n_1 = 11$, sample mean $\bar{x}_1 = 124$, sample variance $s_1^2 = 59$

Sample from population 2:

sample size $n_2 = 15$, sample mean $\bar{x}_2 = 105$, sample variance $s_2^2 = 42$

Calculate a 95% confidence interval for $\mu_1 - \mu_2$, the difference between the population means (you may assume that the population variances are equal). [5]

10 Let X denote a random variable with a continuous uniform (0, 1000) distribution. Define a random variable Y as the minimum of X and 800.

(i) Show that the conditional distribution of X given $X < 800$ is a continuous uniform (0, 800) distribution. [2]

(ii) Verify (giving clear reasons) that the expectation of the random variable Y is 480. [3]

(iii) Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed, each with the same distribution as Y.

In the case that n is large, determine the approximate distribution of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, stating its expectation. (You are **not** required to derive or state the variance of $\bar{Y}$.) [1]

(iv) Comment on the comparison of the conditional expectation of X given $X < 800$ with the expectation of Y. [2]

[Total 8]

11 Consider the following simple model for the number of claims, N, which occur in a year on a policy:

n	0	1	2	3
$P(N = n)$	0.55	0.25	0.15	0.05

(a) Explain how you would simulate an observation of N using a number r, an observation of a random variable which is uniformly distributed on (0, 1).

(b) Illustrate your method described in (i) by simulating three observations of N using the following random numbers between 0 and 1:

0.6221, 0.1472, 0.9862.

[4]

12 A certain type of insurance policy has a claim rate of λ per year and the cover ceases and the policy expires after the first claim. Accordingly the duration of a policy is modelled by an exponential distribution with density function $\lambda e^{-\lambda x} : 0 < x < \infty$.

A company has data on $(m + n)$ policies which have expired and which may be assumed to be independent. Of these, m policies had duration less than 5 years and n policies had duration greater than or equal to 5 years.

(i) An investigator makes note of the actual durations, $x_1, \ldots, x_n$, of the latter group of n policies, but ignores the former group without even noting the value of m.

(a) Explain why the x_i's come from a truncated exponential distribution with density function

$$f(x) = k.\lambda e^{-\lambda x} \ , \ 5 < x < \infty$$

and show that $k = e^{5\lambda}$.

(b) Write down the likelihood for the data from the point of view of this investigator and hence show that the maximum likelihood estimate (MLE) of λ is given by

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i - 5n}.$$

(c) The data yield the values: $n = 10$ and $\Sigma x_i = 71$. Calculate this investigator's MLE of λ.

[8]

(ii) A second investigator ignores the actual policy durations and simply notes the values of m and n.

(a) Write down the likelihood for this information and hence show that the resulting MLE of λ is given by

$$\hat{\lambda} = \frac{1}{5}\log\left(\frac{m+n}{n}\right).$$

(b) The same data as in part (i) yield the values: $m = 120$ and $n = 10$. Calculate this investigator's MLE of λ.

[5]

(iii) The two investigators decide to pool their data, and so have the information that there are m policies with duration less than 5 years, and n policies with actual durations $x_1, \ldots, x_n$.

(a) Explain why the likelihood for this joint information is given by

$$L(\lambda) = (1-e^{-5\lambda})^m . \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

and determine an equation, the solution of which will lead to the MLE of λ.

(b) Given that this leads to an MLE of λ equal to 0.508, comment on the comparison of the three MLE's.

[5]
[Total 18]

13 A survey, carried out at a major flower and gardening show, was concerned with the association between the intention to return to the show next year and the purchase of goods at this year's show. There were 220 people interviewed and of these 101 had made a purchase; 69 of these people said they intended to return next year. Of the 119 who had not made a purchase, 68 said they intended to return next year.

(i) Suppose one of the 220 people surveyed is selected at random.

Calculate the probabilities that the selected person:

(a) intends to return next year, given that he/she has made a purchase
(b) intends to return next year, given that he/she has not made a purchase
(c) has made a purchase, given that he/she intends to return next year

[3]

(ii) By testing the difference between the proportions of purchasers and non-purchasers who intend to return next year, examine whether there is sufficient evidence to justify concluding that the intention to return depends on whether or not a purchase was made. [7]

(iii) Present the data as a contingency table and perform a χ^2 test of association between the attributes "intention to return" and "purchasing status". [5]

(iv) Discuss briefly the connection between the comparison of proportions carried out in part (ii) and the test of association performed in part (iii). [2]

[Total 17]

14 The data given in the following table are the numbers of deaths from AIDS in Australia for 12 consecutive quarters starting from the second quarter of 1983.

Quarter (i):	1	2	3	4	5	6	7	8	9	10	11	12
Number of deaths (n_i):	1	2	3	1	4	9	18	23	31	20	25	37

(i) (a) Draw a scatterplot of the data.

(b) Comment on the nature of the relationship between the number of deaths and the quarter in this early phase of the epidemic.

[4]

(ii) A statistician has suggested that a model of the form

$$E\left[N_i\right] = \gamma i^2$$

might be appropriate for these data, where γ is a parameter to be estimated from the above data. She has proposed two methods for estimating γ, and these are given in parts (a) and (b) below.

(a) Show that the least squares estimate of γ, obtained by minimising $q = \sum_{i=1}^{12}(n_i - \gamma i^2)^2$, is given by

$$\hat{\gamma} = \frac{\sum_{i=1}^{12} i^2 n_i}{\sum_{i=1}^{12} i^4}.$$

(b) Show that an alternative (weighted) least squares estimate of γ, obtained by minimising $q^* = \sum_{i=1}^{12} \frac{\left(n_i - \gamma i^2\right)^2}{i^2}$ is given by

$$\tilde{\gamma} = \frac{\sum_{i=1}^{12} n_i}{\sum_{i=1}^{12} i^2}.$$

(c) Noting that $\sum_{i=1}^{12} i^4 = 60,710$ and $\sum_{i=1}^{12} i^2 = 650$, calculate $\hat{\gamma}$ and $\tilde{\gamma}$ for the above data.

[8]

(iii) To assess whether the single parameter model which was used in part (ii) is appropriate for the data, a two parameter model is now considered. The model is of the form

$$E\left[N_i\right] = \gamma i^{\theta}$$

for $i = 1, \ldots, 12$.

(a) To estimate the parameters γ and θ, a simple linear regression model

$$E\left[Y_i\right] = \alpha + \beta x_i$$

is used, where $x_i = \log(i)$ and $Y_i = \log(N_i)$ for $i = 1, \ldots, 12$. Relate the parameters γ and θ to the regression parameters α and β.

(b) The least squares estimates of α and β are -0.6112 and 1.6008 with standard errors 0.4586 and 0.2525 respectively (you are **not** asked to verify these results).

Using the value for the estimate of β, conduct a formal statistical test to assess whether the form of the model suggested in (ii) is adequate.

[7]

[Total 19]

END OF PAPER