EXAMINATION

10 April 2008 (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 from the approved list.

1 The number of claims which arose during the calendar year 2005 on each of a group of 80 private motor policies was recorded and resulted in the following frequency distribution:

Number of claims x	0	1	2	3	4
Number of policies f	64	12	3	0	1

For these data $\Sigma fx = 22$, $\Sigma fx^2 = 40$

Calculate the sample mean and standard deviation of the number of claims per policy. [3]

2 Data on a sample of 29 claim amounts give a sample mean of £461.5 and a sample standard deviation of £618.8.

One claim amount of £3,657.50 is identified as an outlier and after investigation is found to be in error. Calculate the revised sample mean and standard deviation if this erroneous amount is removed. [4]

3 The following sample contains claim amounts (£) on a particular class of insurance policies:

1,717	1,595	1,764	1,464	1,854	1,560	1,698
1,614	1,524	4,320	1,626	1,440	1,602	

(i) Determine the mean and the median of the claim amounts. [2]

(ii) State, with reasons, which of the two measures considered above you would prefer to use to estimate the central point of the claim amounts. [1]

[Total 3]

4 Consider two events A and B, such that $P(A) = 0.3$ and $P(A \cap B) = 0.1$.

Find the minimum and maximum possible values of the conditional probability $P(A | B)$. [4]

5 An insurance company covers claims from four different non-life portfolios, denoted as G_1, G_2, G_3 and G_4. The number of policies included in each portfolio is given below:

Portfolio	G_1	G_2	G_3	G_4
No. of policies	4,000	7,000	13,000	6,000

It is estimated that the percentages of policies that will result in a claim in the following year in each of the portfolios are 8%, 5%, 2% and 4% respectively.

Suppose a policy is chosen at random from the group of 30,000 policies comprising the four portfolios after one year and it is found that a claim did arise on this policy during the year. Calculate the probability that the selected policy comes from portfolio G_3. [3]

6 Consider two random variables X and Y with joint probability density function (pdf)

$$f(x,y) = \frac{4}{3}(1 - xy), \quad 0 < x < 1, 0 < y < 1.$$

The marginal pdf of X is given by

$$f(x) = \frac{2}{3}(2 - x), \quad 0 < x < 1$$

with a corresponding marginal pdf for Y by symmetry (*you are not asked to verify these marginal densities*).

(i) Show that the conditional pdf of Y given $X = x$ is given by

$$f(y \mid x) = 2\frac{(1 - xy)}{(2 - x)}, \quad 0 < y < 1.$$ [2]

(ii) (a) Determine the conditional expectation $E(Y \mid X = x)$ as a function of x and hence determine $E(Y)$.

(b) Verify your answer in part (a) by determining $E(Y)$ directly from the marginal pdf of Y. [5]

[Total 7]

7 The claim amount X in units of £1,000 for a certain type of industrial policy is modelled as a gamma variable with parameters $\alpha = 3$ and $\lambda = ¼$.

(i) Use moment generating functions to show that $\frac{1}{2}X \sim \chi_6^2$. [3]

(ii) Hence use tables to find the probability that a claim amount exceeds £20,000. [2]

[Total 5]

8 A woodcutter has to cut 100 fence posts of a standard length and he has a metal bar of the required length to act as the standard. The woodcutter decides to vary his procedure from post to post – he cuts the first post using the metal standard, then uses this post as his standard for the cut of the next post. He continues in a similar manner, each time using the most recently cut post as the standard for the next cut.

Each time the woodcutter cuts a post there is an error in the length cut relative to the standard being employed for that cut – you should assume that the errors are independent observations of a random variable with mean 0 and standard deviation 3mm.

Calculate, approximately, the probability that the length of the final post differs from the length of the original metal standard by more than 15mm.

[5]

9 A researcher wishes to investigate whether a coin is balanced or not, that is if $P(heads) = 0.5$. She throws the coin four times and decides to accept the hypothesis $H_0 : P(heads) = 0.5$ in a test against the alternative $H_1 : P(heads) \neq 0.5$, if the number of times that the coin lands "heads" is 1, 2, or 3.

(i) Calculate the probability of the type I error of this test. [3]

(ii) Calculate the probability of the type II error of this test, if the true probability that the coin lands "heads" is 0.7. [3]

[Total 6]

10 Pressure readings are taken regularly from a meter. It transpires that, in a random sample of 100 such readings, 45 are less than 1, 35 are between 1 and 2, and 20 are between 2 and 3.

Perform a χ^2 goodness of fit test of the model that states that the readings are independent observations of a random variable that is uniformly distributed on (0, 3).

[5]

11 In an investigation about the duration of insurance policies of a certain type, a sample of n policies is studied. All n policies have been initiated at the same time, which is also the time of the start of the investigation. For each policy, the time T (in months) until the policy expires can be modelled as an exponential random variable with parameter λ, independently of the times for all other policies.

(i) Suppose that the investigation is terminated as soon as k policies have expired, where k is a known (predetermined) constant. The observed policy expiry times are denoted by $t_1, t_2, \ldots, t_k$ with $0 < k \le n$ and $t_1 < t_2 < \ldots < t_k$.

(a) Show that the probability that any randomly selected policy is still in force at the time of the termination of the investigation is $e^{-\lambda t_k}$.

(b) Show that the likelihood function of the parameter λ, using information from all n policies, is given by

$$L(\lambda) = \lambda^k e^{-\lambda \sum_{i=1}^{k} t_i} e^{-(n-k)\lambda t_k}.$$

Hence find the maximum likelihood estimate (MLE) of λ.

(c) Consider an investigation on 20 policies which is terminated when five policies have expired, giving the following observed expiry times (in months):

1.03 6.67 12.70 12.88 21.54

Calculate the MLE of λ based on this sample.

[9]

(ii) Suppose instead that the investigation is terminated after a fixed length of time t_0. The number of policies that have expired by time t_0 is considered to be a random variable, denoted by K.

(a) Explain clearly why the distribution of K is binomial and determine its parameters.

(b) Hence find the MLE of λ in this case.

(c) Consider an investigation on 20 policies that is terminated after 24 months. By the time of termination five policies have expired.

Use this information to calculate the MLE of λ in this case.

[9]
[Total 18]

12 The members of the computer games clubs of three neighbouring schools decide to take part in a light-hearted competition. Each club selects five of its members at random under a procedure agreed and supervised by the clubs. Each selected student then plays a particular game at the end of which the score he/she has attained is displayed and recorded – the standard set by the games designers is such that reasonably competent players should score about 100.

The results are as follows:

School 1	105	134	96	147	116
School 2	103	81	91	100	110
School 3	137	115	105	123	149

(i) An analysis of variance is conducted on these results and gives the following ANOVA table:

Source of variation	*d.f.*	*SS*	*MSS*
Between schools	2	2,298	1,149
Residual	12	3,468	289
Total	14	5,766	

(a) Test the hypothesis that there are no school effects against a general alternative.

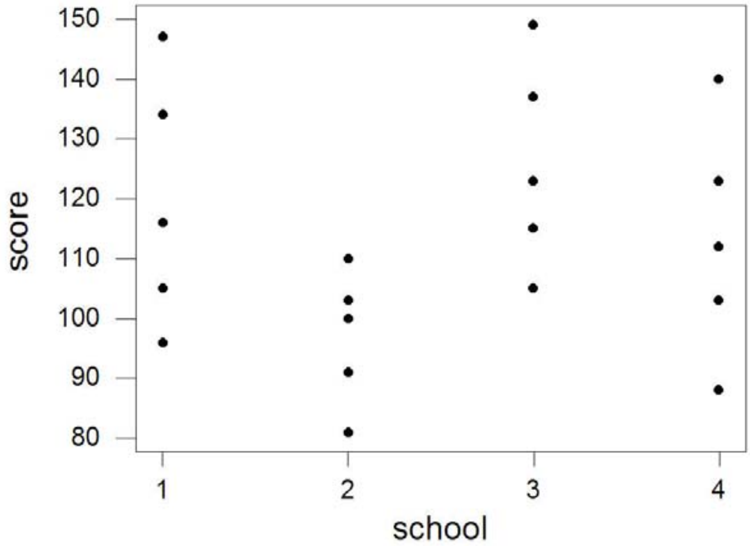
You should quote a narrow range of values within which the probability-value of the data lies, and state your conclusion clearly.

(b) Calculate a 95% confidence interval for the underlying mean score for club members in School 1, using the information available from all three schools. [7]

(ii) The members of the computer games club of a nearby fourth school hear about the competition and ask to be included in the overall comparison. Scores for a random sample of five of the club members at this school (School 4) are obtained and are:

112 140 88 103 123.

The scores obtained by all twenty students are shown in the display below:



(a) Carry out an analysis of variance on the results for all four schools together − you should construct the ANOVA table and test the hypothesis that there are no school effects against a general alternative.

You should quote an approximate value for the probability-value of the data, and state your conclusion clearly.

(b) Comment briefly on the comparison of the results of the analysis involving Schools 1−3 only conducted in part (i)(a) and the results here of the analysis involving all four schools.

(c) Calculate a 95% confidence interval for the difference in the underlying mean scores for club members in Schools 1 and 2, using the information available from all four schools.

[13]

[Total 20]

13 In a medical experiment concerning 12 patients with a certain type of ear condition, the following measurements were made for blood flow (y) and auricular pressure (x):

x:	8.5	9.8	10.8	11.5	11.2	9.6	10.1	13.5	14.2	11.8	8.7	6.8
y:	3	12	10	14	8	7	9	13	17	10	5	5

$$\Sigma x = 126.5 \quad \Sigma x^2 = 1{,}381.85 \quad \Sigma y = 113 \quad \Sigma y^2 = 1{,}251 \quad \Sigma xy = 1{,}272.2$$

(i) Construct a scatterplot of blood flow against auricular pressure and comment briefly on any relationship between them. [3]

(ii) Calculate the equation of the least-squares fitted regression line of blood flow on auricular pressure. [4]

(iii) (a) Use a suitable pivotal quantity with a t distribution to show how to derive the usual 95% confidence interval for the slope coefficient of the underlying regression line, and calculate the interval.

(b) Use your calculated confidence interval to comment on the hypothesis that the true underlying slope coefficient is equal to 1.5. [5]

(iv) (a) Use a suitable pivotal quantity with a χ^2 distribution to derive a 95% confidence interval for the underlying error variance σ^2, and calculate the interval.

(b) Hence calculate a 95% confidence interval for the error standard deviation σ. [5]

[Total 17]

END OF PAPER