EXAMINATION

4 October 2010 (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 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 marks of a sample of 25 students from a large class in a recent test have sample mean 57.2 and standard deviation 7.3. The marks are subsequently adjusted: each mark is multiplied by 1.1 and the result is then increased by 8.

Calculate the sample mean and standard deviation of the adjusted marks. [2]

2 In a survey, a sample of 10 policies is selected from the records of an insurance company. The following data give, in ascending order, the time (in days) from the start date of the policy until a claim has arisen from each of the policies in the sample.

297 301 312 317 355 379 404 419 432+ 463+

Some of the policies have not yet resulted in any claims at the time of the survey, so the times until they each give rise to a claim are said to be censored. These values are represented with a plus sign in the above data.

(i) Calculate the median of this sample. [2]

(ii) State what you can conclude about the mean time until claims arise from the policies in this sample. [2]

[Total 4]

3 Suppose that in a group of insurance policies (which are independent as regards occurrence of claims), 20% of the policies have incurred claims during the last year. An auditor is examining the policies in the group one by one in random order until two policies with claims are found.

(i) Determine the probability that exactly five policies have to be examined until two policies with claims are found. [2]

(ii) Find the expected number of policies that have to be examined until two policies with claims are found. [1]

[Total 3]

4 For a certain class of business, claim amounts are independent of one another and are distributed about a mean of μ = £4,000 and with standard deviation σ = £500.

Calculate an approximate value for the probability that the sum of 100 such claim amounts is less than £407,500. [3]

5 A random sample of 200 travel insurance policies contains 29 on which the policyholders made claims in their most recent year of cover.

Calculate a 99% confidence interval for the proportion of policyholders who make claims in a given year of cover. [3]

6 The random variable X has a Poisson distribution with mean Y, where Y itself is considered to be a random variable. The distribution of Y is lognormal with parameters μ and σ^2.

Derive the unconditional mean $E[X]$ and variance $V[X]$ using appropriate conditional moments. (You may use any standard results without proof, including results from the book of Formulae and Tables.) [4]

7 Let X be a discrete random variable with the following probability distribution:

X:	0	1	2	3
$P(X = x)$:	0.4	0.3	0.2	0.1

(i) Simulate three observations of X using the following three random numbers from a uniform distribution on (0,1) (you should explain your method briefly and clearly).

Random numbers: 0.4936, 0.7269, 0.1652 [3]

Let X be a random variable with cumulative distribution function:

$$F_X(x) = \frac{1}{1-e^{-1}}\left(1-e^{-x^2}\right), \quad 0 < x < 1 \quad (F_X(x) = 0 \text{ for } x \le 0 \text{ and } F_X(x) = 1 \text{ for } x \ge 1).$$

(ii) Derive an expression for a simulated value of X using a random number u from a uniform distribution on (0,1) and hence simulate an observation of X using the random number $u = 0.8149$. [3]

[Total 6]

8 A certain type of claim amount (in units of £1,000) is modelled as an exponential random variable with parameter $\lambda = 1.25$. An analyst is interested in S, the total of 10 such independent claim amounts. In particular he wishes to calculate the probability that S exceeds £10,000.

(i) (a) Show, using moment generating functions, that:

(1) S has a gamma distribution, and

(2) $2.5S$ has a χ^2_{20} distribution.

(b) Use tables to calculate the required probability.

[5]

(ii) (a) Specify an approximate normal distribution for S by applying the central limit theorem, and use this to calculate an approximate value for the required probability.

(b) Comment briefly on the use of this approximation and on the result.

[3]

[Total 8]

9 Let the random variable X have the Poisson distribution with probability function:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) Show that $P(X = k+1) = \frac{\lambda}{k+1} P(X = k), \quad k = 0, 1, 2, \ldots$ [2]

It is believed that the distribution of the number of claims which arise on insurance policies of a certain class is Poisson. A random sample of 1,000 policies is taken from all the policies in this class which have been in force throughout the past year. The table below gives the number of claims per policy in this sample.

No. of claims, k:	0	1	2	3	4	5	6	7	8 or more
No. of policies, f_k:	310	365	202	88	26	6	2	1	0

For these data the maximum likelihood estimate (MLE) of the Poisson parameter λ is $\hat{\lambda} = 1.186$.

(ii) Calculate the frequencies expected under the Poisson model with parameter given by the MLE above, using the recurrence formula of part (i) (or otherwise). [3]

(iii) Perform an appropriate statistical test to investigate the assumption that the numbers of claims arising from this particular class of policies follow a Poisson distribution. [5]

[Total 10]

10 In the collection of questionnaire data, randomised response sampling is a method which is used to obtain answers to sensitive questions. For example a company is interested in estimating the proportion, p, of its employees who falsely take days off sick. Employees are unlikely to answer a direct question truthfully and so the company uses the following approach.

Each employee selected in the survey is given a fair six-sided die and asked to throw it. If it comes up as a 5 or 6, then the employee answers yes or no to the question "have you falsely taken any days off sick during the last year?". If it comes up as a 1, 2, 3 or 4, then the employee is instructed to toss a coin and answer yes or no to the question "did you obtain heads?". So an individual's answer is either yes or no, but it is not known which question the individual has answered.

For the purpose of the following analysis you should assume that each employee answers the question truthfully.

(i) Show that the probability that an individual answers yes is $\frac{1}{3}(p+1)$. [2]

Suppose that 100 employees are surveyed and that this results in 56 yes answers.

(ii) (a) Show that the likelihood function $L(p)$ can be expressed in the form:

$$L(p) \propto (p+1)^{56}(2-p)^{44}.$$

(b) Hence show that the maximum likelihood estimate (MLE) of p is $\hat{p} = 0.68$.

[5]

Let $\theta = \frac{1}{3}(p+1)$ and note that, using binomial results, the MLE of θ is $\hat{\theta} = \frac{56}{100}$.

(iii) Explain why $\hat{p}$ can be obtained as the solution of $\frac{1}{3}(\hat{p}+1) = \hat{\theta}$, and hence verify that $\hat{p} = 0.68$. [2]

(iv) (a) Determine the second derivative of the log likelihood for p and evaluate it at $\hat{p} = 0.68$.

(b) State an approximate large-sample distribution for the MLE $\hat{p}$.

(c) Hence calculate approximate 95% confidence limits for p.

[5]

Now suppose that the same numerical estimate, that is $\hat{p} = 0.68$, had been obtained from a sample of the same size, that is 100, without using the randomised response method but relying on truthful answers. So the number of yes answers was 68 and $\hat{p} = \frac{68}{100}$ using binomial results.

(v) (a) Calculate approximate 95% confidence limits for p for this situation.

(b) Suggest why the confidence limits in part (iv)(c) are wider than these limits.

[3]
[Total 17]

11 A life insurance company issuing critical illness insurance wants to compare the delay times from the date when a claim is made until it is settled, for different causes of illness covered. Random samples of 12 claims associated with two types of illness (A and B) related to heart disease have been collected. The logarithms of the delay times are given below (where the original times were measured in days):

Cause A, y_A:	4.0	5.4	4.6	3.5	4.2	4.5	4.2	4.9	5.1	5.2	5.1	5.4
Cause B, y_B:	5.7	5.6	4.2	5.1	4.4	5.9	5.4	3.9	5.7	4.5	4.8	3.9

For these data: $\sum y_A = 56.1, \quad \sum y_A^2 = 266.33, \quad \sum y_B = 59.1, \quad \sum y_B^2 = 297.03$

(i) Use a suitable *t*-test to investigate the hypothesis that the mean delay time is the same for claims related to the two causes of illness and state clearly your conclusion. [6]

(ii) Give a possible reason why the logarithms of the original delay time observations are used in this analysis. [2]

(iii) (a) Calculate an equal-tailed 95% confidence interval for σ_A^2 / σ_B^2, the ratio of the variances of the delay times for the two causes of illness.

(b) Comment on the validity of the test in part (i) based on this confidence interval.

[4]

The company collects a third sample of 12 claims associated with an illness (C) related to brain disease, and the logarithms of the delay times are given below:

Cause C, y_C:	5.6	6.2	6.0	5.6	7.1	5.0	4.5	6.4	4.6	6.0	5.5	5.3

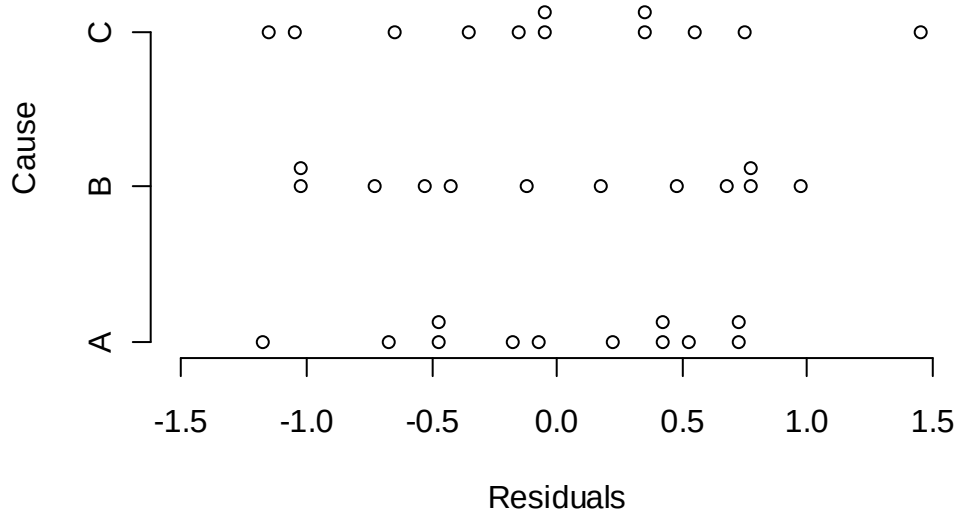
For these data: $\sum y_C = 67.8, \quad \sum y_C^2 = 389.28$

For data in all three samples: $\sum\sum y = 183.0, \quad \sum\sum y^2 = 952.64$

(iv) Use analysis of variance to test the hypothesis that the mean delay times are the same for all three causes of illness. [6]

(v) State the assumptions made for this analysis of variance. [2]

(vi) Comment briefly on the validity of the test in (iv), using the plot of the residuals of the analysis given below. [2]



[Total 22]

12 An investigation concerning the improvement in the average performance of female track athletes relative to male track athletes was conducted using data from various international athletics meetings over a period of 16 years in the 1950s and 1960s. For each year and each selected track distance the observation y was recorded as the average of the ratios of the twenty best male times to the corresponding twenty best female times.

The data for the 100 metres event are given below together with some summaries.

year t:	1	2	3	4	5	6	7	8
ratio y:	0.882	0.879	0.876	0.888	0.890	0.882	0.885	0.886

year t:	9	10	11	12	13	14	15	16
ratio y:	0.885	0.887	0.882	0.893	0.878	0.889	0.888	0.890

$$\Sigma t = 136, \quad \Sigma t^2 = 1496, \quad \Sigma y = 14.160, \quad \Sigma y^2 = 12.531946, \quad \Sigma ty = 120.518$$

(i) Draw a scatterplot of these data and comment briefly on any relationship between ratio and year. [3]

(ii) Verify that the equation of the least squares fitted regression line of ratio on year is given by:

$$y = 0.88105 + 0.000465t.$$ [4]

(iii) (a) Calculate the standard error of the estimated slope coefficient in part (ii).

(b) Determine whether the null hypothesis of "no linear relationship" would be accepted or rejected at the 5% level.

(c) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model.

[5]

Corresponding data for the 200 metres event resulted in an estimated slope coefficient of:

$\hat{\beta} = 0.000487$ with standard error 0.000220.

(iv) (a) Determine whether the "no linear relationship" hypothesis would be accepted or rejected at the 5% level.

(b) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model and comment on whether or not the underlying slope coefficients for the two events, 100m and 200m, can be regarded as being equal.

(c) Discuss why the results of the tests in parts (iii)(b) and (iv)(a) seem to contradict the conclusion in part (iv)(b).

[6]

[Total 18]

END OF PAPER