EXAMINATION

4 October 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 Data collected on claim amounts (£) for two postcode regions give the following values for n (the number of claims), $\bar{x}$ (the mean claim amount) and s (the sample standard deviation) of the claim amounts.

	Region 1	*Region 2*
n	25	18
$\bar{x}$	120.2	142.7
s	58.1	62.2

Calculate, to one decimal place, the mean and sample standard deviation of the claim amounts for both regions combined. [4]

2 The random variable X has probability density function

$$f(x) = \frac{2}{x^3}, \quad x > 1$$

and cumulative distribution function

$$F(x) = \begin{cases} 0, & x < 1 \\ 1 - \frac{1}{x^2}, & x \geq 1 \end{cases}.$$

Use the following uniform(0,1) random numbers

0.5719, 0.8612, 0.3028

to simulate three observations of X, explaining your method and calculations clearly. [4]

3 It is known that 24% of the customers in a bank holding a current account also have another type of account with the bank.

Calculate an approximate value for the probability that fewer than 50 customers in a random sample of 250 customers with a current account also have another type of account. [3]

4 In a random sample of 200 policies from a company's private motor business, there are 68 female policyholders and 132 male policyholders.

Calculate an approximate 99% confidence interval for the proportion of policyholders who are female in the corresponding population of all policyholders. [3]

5 For a particular insurance company a sample of eight claim amounts (in units of £1,000) on household contents is taken. The data give $\sum_{i=1}^{8} x_i = 56.7$ and $\sum_{i=1}^{8} x_i^2 = 403.95$. The claim amounts are assumed to follow a normal distribution.

(i) Calculate a 90% confidence interval for the true mean claim amount. [3]

(ii) Use the confidence interval calculated in (i) above to comment on an expert's assessment that the average claim amount for the company is £6,500. [1]

[Total 4]

6 In an investigation into the relationship between two normally distributed variables, X and Y, based on a sample of 15 points, it is desired to perform the following test concerning the true underlying correlation coefficient ρ

$$H_0 : \rho = 0 \quad \text{v} \quad H_1 : \rho > 0.$$

Use the t distribution to determine an upper critical value for the sample correlation coefficient r for this test at the 1% level. [4]

7 Let N denote the number of claims which arise in a portfolio of business and let X_i be the amount of the ith claim. Let each of the X_i's be independently modelled as a normal variable with mean £10,000 and standard deviation £2,000 and let N be independently modelled as a Poisson variable with parameter 20.

Calculate the mean and standard deviation of the total claim amount $S = X_1 + \ldots + X_N$. [3]

8 Claim sizes in a certain insurance situation are modelled by a normal distribution with mean μ = £30,000 and standard deviation σ = £4,000 The insurer defines a claim to be a *large claim* if the claim size exceeds £35,000.

(i) Calculate the probabilities that the size of a claim exceeds

(a) £35,000, and
(b) £36,000

[2]

(ii) Calculate the probability that the size of a *large claim* (as defined by the insurer) exceeds £36,000. [2]

(iii) Calculate the probability that a random sample of 5 claims includes 2 which exceed £35,000 and 3 which are less than £35,000. [2]

[Total 6]

9 For a certain class of policies issued by a large insurance company it is believed that the probability of each policy giving rise to any claims is 0.5, independently of all other policies. A random sample of 250 such policies is selected.

(i) Determine approximately the probability that at least 139 of the policies in the sample will each give rise to any claims. [4]

(ii) Suppose we do observe that 139 policies in our sample give rise to at least one claim. Use your answer to part (i) to determine whether this suggests at the 1% level of significance that the probability of any claims arising from a policy of this certain class is greater than initially believed. [3]

[Total 7]

10 A chi-square test of association for the frequency data in the following 2 × 3 table

		Factor A		
		A1	*A2*	*A3*
Factor B	*B1*	40	30	50
	B2	80	30	70

produces a chi-square statistic with value 4.861 and associated *P*-value 0.089.

Consider a chi-square test of association for the data in the following 2 × 3 table, in which all frequencies are twice the corresponding frequencies in the first table:

		Factor A		
		A1	*A2*	*A3*
Factor B	*B1*	80	60	100
	B2	160	60	140

(i) State, or calculate, the value of the chi-square test statistic for the second table. [2]

(ii) Find the *P*-value associated with the test statistic in (i). [1]

(iii) Comment on the results. [2]

[Total 5]

11 Suppose that the random variable X follows an exponential distribution with probability density function

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

Define a new random variable $Y = X^{\frac{1}{3}}$.

(i) (a) Show that the cumulative density function of Y is given by

$$F_Y(y) = \begin{cases} 1 - \exp(-\lambda y^3), & y \geq 0 \\ 0, & y < 0 \end{cases}$$

and hence, or otherwise, find the probability density function of Y.

(b) Explain how you would simulate a value of Y given a value u from the uniform $U(0,1)$ distribution.

[7]

(ii) (a) Find an expression for the maximum likelihood estimator of the parameter λ, using a sample $y_1, y_2, \ldots, y_n$, from the distribution of Y.

(b) Eight observed values of the random variable Y are given below:

0.72 1.15 1.26 1.03 1.69 1.30 1.42 1.15

Calculate the maximum likelihood estimate of λ using these values. [6]

(iii) (a) The hazard function of a continuous random variable T is defined as $h(t) = \frac{f(t)}{S(t)}$, where $f(t)$ denotes the probability density function and $S(t)$ denotes the survival function defined as $S(t) = P(T > t)$.

Derive the hazard functions of the random variables X and Y defined above.

(b) If a random variable T represents the lifetime of an individual, then the hazard function $h(t)$, as defined in part (iii)(a), gives the instantaneous mortality rate (that is, the force of mortality) at time t for that individual.

State (with reasons) which of the two random variables (X and Y) you would use to model the lifetime of pensioners for a period of time longer than one year, basing your answer on the form of the corresponding hazard functions derived in part (iii)(a).

[5]
[Total 18]

12 A series of n geomagnetic readings are taken from a meter, but the readings are judged to be approximate and unreliable. The chief scientist involved does know however that the true values are all positive and she suggests that an appropriate model for the readings is that they are independent observations of a random variable which is uniformly distributed on $(0, \theta)$, where $\theta > 1$.

(i) Suppose that the chief scientist knows only that the number, M, of the readings which are less than 1 is m, with the remaining $n - m$ being greater than 1 and that she adopts the model as suggested above.

(a) Show that the probability that a single reading is less than 1 is $\frac{1}{\theta}$.

(b) Demonstrate that the maximum likelihood estimate of θ is $\hat{\theta} = \frac{n}{m}$.

(c) Demonstrate that the Cramer-Rao lower bound (CRlb) for estimating θ is $\frac{\theta^2(\theta - 1)}{n}$ and hence state the large sample distribution of $\hat{\theta}$.

[10]

(ii) Suppose that exactly 45 readings in a random sample of 100 readings are less than 1.

(a) Calculate an estimate of the standard error of $\hat{\theta}$ and hence calculate an approximate two-sided 95% confidence interval for θ.

(b) Use the large sample distribution of $\hat{\theta}$ to test the hypotheses H_0: $\theta = 3$ v H_1: $\theta < 3$.

[9]
[Total 19]

13 In a laboratory experiment a response variable (yield, y) is thought to be affected by a quantitative factor (percentage of catalyst, x). The experiment involved making four observations of y at each of four values of x, being 12%, 14%, 16% and 18%, and resulted in the following observed response data.

12%	*14%*	*16%*	*18%*
46	56	56	47
51	57	63	51
47	63	60	54
42	60	64	55

These data are analysed by two statisticians, *A* and *B*, who use an analysis of variance approach and a linear regression approach, respectively.

(i) Statistician *A*'s approach:

You are given the following data summaries:

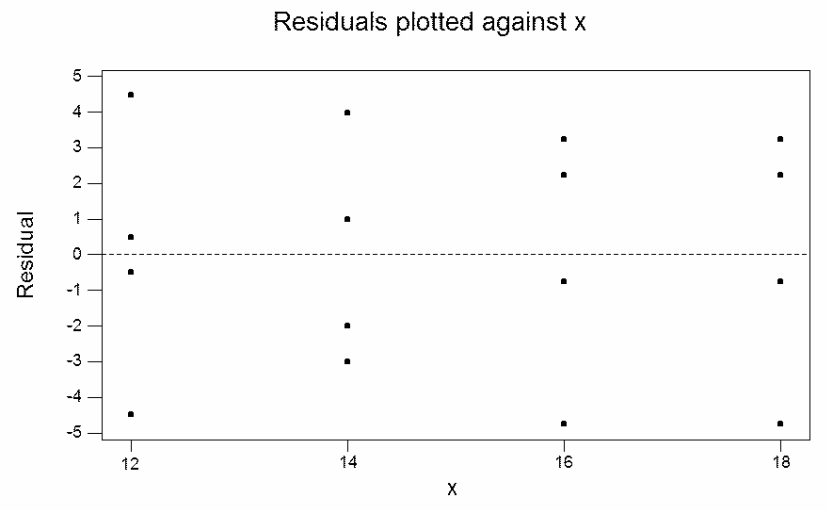
sub-totals $\Sigma y = 186,\ 236,\ 243$ and 207 at $x = 12, 14, 16$ and 18, respectively, and overall totals $\Sigma y = 872$ and $\Sigma y^2 = 48{,}196$.

(a) Apply a one-way analysis of variance to these data and obtain the resulting F-value for the usual test.

(b) Show that the P-value for the test is substantially less than 0.01, by referring to tables of percentage points for the F distribution.

(c) The result of part (b) above shows that there is very strong evidence of an effect on y due to the quantitative factor x. Suggest a suitable diagram that statistician *A* could now use to describe the effect of x on y. Draw this diagram and hence comment on the effect of x on y.

(d) The graph below shows the residuals plotted against the values of x:



Comment briefly on any implications of this graph. [10]

(ii) Statistician *B*'s approach:

You are given the following data summaries:

$$\Sigma x = 240 \quad \Sigma y = 872 \quad \Sigma x^2 = 3{,}680 \quad \Sigma y^2 = 48{,}196 \quad \Sigma xy = 13{,}150.$$

(a) Perform a linear regression analysis on these data to show that the fitted line is given by $y = 41.4 + 0.875x$.

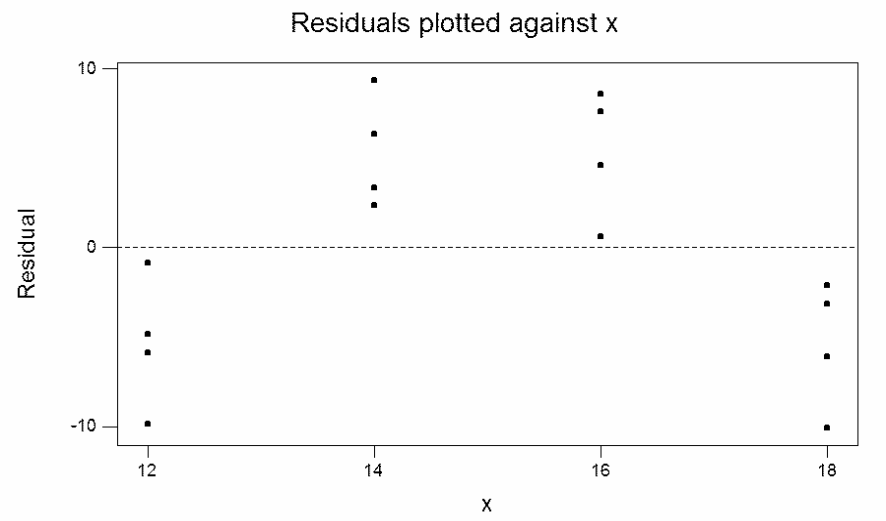
(b) Perform the hypothesis test on the slope coefficient

$$H_0 : \beta = 0 \quad \text{v} \quad H_1 : \beta \neq 0$$

showing that the *P*-value is greater than 0.20.

Comment on what this implies about the relationship between *x* and *y*.

(c) The graph below shows the residuals plotted against the values of *x*:



Comment briefly on what this graph implies about the effect of *x* on *y*.

(d) Suggest an additional analysis statistician *B* could now use to describe the effect of *x* on *y*.

[10]
[Total 20]

END OF PAPER