

# EXAMINATION

18 September 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 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 mean of a sample of 30 claim amounts arising from a certain kind of insurance policy is £5,200. Six of these claim amounts have mean £8,000 while ten others have mean £3,100.

Calculate the mean of the remaining claim amounts in this sample. [3]

- 2** Five years ago a financial institution issued a specialised type of investment bond and investors had the option to cash in after 1, 2, 3, 4 or 5 years. The following table gives a frequency distribution showing the numbers of those investors who cashed in at each stage.

<i>duration (length of time held before being cashed in)</i>				
<i>1 year</i>	<i>2 years</i>	<i>3 years</i>	<i>4 years</i>	<i>5 years</i>
130	151	97	64	98

Calculate the sample mean and standard deviation of the duration of these bonds before being cashed in. [4]

- 3** (i) Let  $Y$  be the sum of two independent random variables  $X_1$  and  $X_2$ , that is,

$$Y = X_1 + X_2.$$

Show that the moment generating function (mgf) of  $Y$  is the product of the mgfs of  $X_1$  and  $X_2$ . [2]

- (ii) Let  $X_1$  and  $X_2$  be independent gamma random variables with parameters  $(\alpha_1, \lambda)$  and  $(\alpha_2, \lambda)$ , respectively.

Use mgfs to show that  $Y = X_1 + X_2$  is also a gamma random variable and specify its parameters. [2]

[Total 4]

- 4** A random sample of 15 observations is taken from a normally distributed population of values. The sample mean is 94.2 and the sample variance is 24.86.

Calculate a 99% confidence interval for the population mean. [3]

- 5** The number of claims,  $X$ , arising on each policy in a certain portfolio depends on another random variable  $Y$ .  $X$  is considered to follow a Poisson distribution with mean  $Y$ . The variable  $Y$  itself is assumed to have a gamma distribution with parameters  $(a, b)$ .

Find expressions for the unconditional moments  $E[X]$  and  $E[X^2]$  using appropriate conditional moments. [4]

**6** Suppose that the time  $T$ , measured in days, until the next claim arises under a portfolio of non-life insurance policies, follows an exponential distribution with mean 2.

(i) Find the probability that no claim is made in the next one day period. [2]

(ii) The median of a random variable is defined as the value for which the cumulative distribution function of the variable is equal to 0.5.

Find the median time until the next claim arises. [2]

(iii) Now let  $T_1, T_2, \dots, T_{30}$  be the times (in days) until the next claim arises under each one of 30 similar portfolios of non-life insurance policies, and assume that each  $T_i, i = 1, \dots, 30$ , follows an exponential distribution with mean 2, independently of all others.

Calculate, approximately, the probability that the total of all 30 times which elapse until a claim arises on each of the portfolios exceeds 45 days. [4]

[Total 8]

**7** Let  $N$  be the number of claims arising on a group of policies in a period of one week and suppose that  $N$  follows a Poisson distribution with mean 60.

Let  $X_1, X_2, \dots, X_N$  be the corresponding claim amounts and suppose that, independently of  $N$ , these are independent and identically distributed with mean £500 and standard deviation £400.

Let  $S = \sum_{i=1}^N X_i$  be the total claim amount for the period of one week.

(i) Determine the mean and the standard deviation of  $S$ . [2]

(ii) Explain why the distribution of  $S$  can be taken as approximately normal, and hence calculate, approximately, the probability that  $S$  is greater than £40,000.

[3]

[Total 5]

- 8 (i) Use the following uniform(0,1) random numbers

0.9236 , 0.2578

and a suitable table of probabilities to simulate two observations of the random variable  $X$ , where  $X \sim N(200,100)$ . [3]

- (ii) Use the following uniform(0,1) random numbers

0.3287 , 0.9142

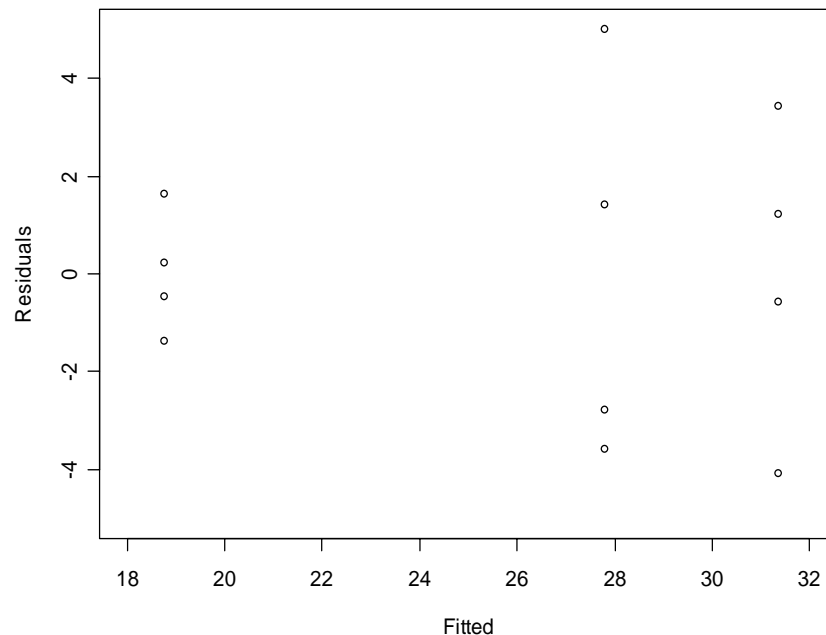
to simulate two observations of the random variable  $Y$ , where  $Y$  has an exponential distribution with mean 100. [3]

[Total 6]

- 9 A random sample of four insurance policies of a certain type was examined for each of three insurance companies and the sums insured were recorded. An analysis of variance was then conducted to test the hypothesis that there are no differences in the means of the sums insured under such policies by the three companies.

The total sum of squares was found to be  $SS_T = 420.05$  and the between-companies sum of squares was found to be  $SS_B = 337.32$ .

- (i) Perform the analysis of variance to test the above hypothesis and state your conclusion. [4]
- (ii) State clearly any assumptions that you made in performing the analysis in (i). [2]
- (iii) The plot of the residuals of this analysis of variance against the associated fitted values, is given below.



Comment briefly on the validity of the test performed in (i), basing your answer on the above plot. [2]

[Total 8]

**10** When a new claim comes into an office it is screened at a first stage and has a probability  $\theta$  of being cleared for progress, otherwise it is rejected. If it clears the first stage, it is then independently screened at a second stage and has the same probability  $\theta$  of being cleared for progress, otherwise it is rejected.

(i) Explain clearly why the probability of a claim being rejected at the first stage is  $1 - \theta$ , of being rejected at the second stage is  $\theta(1 - \theta)$  and of progressing after the two stages is  $\theta^2$ . [3]

(ii) For a sample of  $n$  independent claims which came into the office  $x_1$  were rejected at the first stage,  $x_2$  were rejected at the second stage and  $x_3$  progressed after the two stages ( $x_1 + x_2 + x_3 = n$ ).

(a) Write down the likelihood  $L(\theta)$  for this sample and hence show that the derivative of the log-likelihood is given by

$$\frac{\partial}{\partial \theta} \log L(\theta) = \frac{x_2 + 2x_3}{\theta} - \frac{x_1 + x_2}{1 - \theta}.$$

(b) Show that the maximum likelihood estimator (MLE) is given by

$$\hat{\theta} = \frac{x_2 + 2x_3}{x_1 + 2x_2 + 2x_3}.$$

[7]

(iii) (a) Determine the second derivative  $\frac{\partial^2}{\partial \theta^2} \log L(\theta)$  of the log-likelihood in part (ii) above and hence show that the Cramer-Rao lower bound (CRlb) is given by  $\frac{\theta(1 - \theta)}{n(1 + \theta)}$ .

(b) Use the asymptotic distribution for the MLE  $\hat{\theta}$  with the CRlb evaluated at  $\hat{\theta}$  to obtain an approximate large-sample 95% confidence interval for  $\theta$  expressing it simply in terms of  $\hat{\theta}$  and  $n$ .

[7]

(iv) For a sample of 1,000 independent claims, 110 were rejected at the first stage, 96 were rejected at the second stage and 794 progressed after the two stages.

Calculate the MLE  $\hat{\theta}$  together with an approximate 95% confidence interval for  $\theta$ . [3]

[Total 20]

- 11** A study was conducted to investigate lengths of stay, in days, of short-term stay patients in a particular hospital. Independent random samples of 40 male patients and 35 female patients were selected and the lengths of stay of these patients are given in the following tables:

<i>Male</i>								<i>Female</i>						
4	8	2	6	9	6	4	10	2	7	5	1	9	1	3
6	6	8	7	3	5	6	1	8	6	4	1	5	4	4
10	7	5	7	7	6	9	2	3	5	4	6	5	1	8
1	7	2	8	11	5	6	2	7	2	4	4	11	6	8
1	8	1	3	9	3	1	3	3	6	5	9	6	2	3

Male:  $\Sigma x = 215$     $\Sigma x^2 = 1,481$       Female:  $\Sigma x = 168$     $\Sigma x^2 = 1,026$

The male observations are assumed to be normally distributed with mean  $\mu_1$  and standard deviation  $\sigma_1$ , and independently the female observations are assumed to be normally distributed with mean  $\mu_2$  and standard deviation  $\sigma_2$ .

- (i) Suppose that it is known that  $\sigma_1 = 3.0$  days and  $\sigma_2 = 2.5$  days.
- (a) Construct a 95% confidence interval for the difference between the mean length of stay for males and the mean length of stay for females, that is for  $\mu_1 - \mu_2$ .
- (b) Comment briefly on any implications of this confidence interval. [6]
- (ii) Suppose now that  $\sigma_1$  and  $\sigma_2$  are unknown.
- (a) Perform a two-sample  $t$ -test to investigate whether there is a difference between the mean length of stay for males and the mean length of stay for females, assuming that  $\sigma_1$  and  $\sigma_2$  are equal.
- (b) Show that the variances in the male and female samples are not significantly different at the 5% level, and comment briefly with reference to the validity of the test conducted in (ii)(a).
- (c) Suppose you are not prepared to assume more than you feel is absolutely necessary – in particular you do not want to assume that  $\sigma_1$  and  $\sigma_2$  are equal, nor that the observations necessarily come from normal populations.

Perform an alternative (large-sample) test to that conducted in part (ii)(a), “to investigate whether there is a difference between the mean length of stay for males and the mean length of stay for females”, and compare the results of the test with the results of the test obtained in part (ii)(a).

[11]

[Total 17]

- 12** Consider a situation in which the data consist of two responses at each of five values of an explanatory variable ( $x = 1, 2, 3, 4, 5$ ), so we have a data set with ten responses ( $y$ ), as in the following table:

$x$	1	1	2	2	3	3	4	4	5	5
$y$	12	19	18	35	19	44	32	53	44	65

For these data  $\Sigma x = 30$ ,  $\Sigma y = 341$ ,  $\Sigma x^2 = 110$ ,  $\Sigma y^2 = 14,345$ ,  $\Sigma xy = 1,211$

- (i) You are asked to carry out a linear regression analysis using these data.
- Draw a plot of the data to show the relationship between the response and explanatory values.
  - Calculate the total, regression, and residual sums of squares for a least-squares linear regression analysis of  $y$  on  $x$ , and hence calculate the value of  $R^2$ , the coefficient of determination.
  - Determine the equation of the fitted regression line.
  - Calculate a 95% confidence interval for the slope of the underlying regression line.
- [12]
- (ii) A colleague suggests that it will be simpler and will produce the same results if we use the following reduced data, in which the two responses at each  $x$  value are replaced by their mean:

$x$	1	2	3	4	5
$y$	15.5	26.5	31.5	42.5	54.5

The details of the regression analysis for these data are given in the box below.

Regression equation: $y = 5.90 + 9.40x$					
	<i>Coef</i>	<i>Stdev</i>	<i>t-ratio</i>	<i>p-val</i>	
<i>Intercept</i>	5.900	2.233	2.64	0.078	
<i>x</i>	9.400	0.673	13.96	0.001	
$s = 2.129$		$R\text{-sq} = 98.5\%$			
Analysis of Variance					
<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p-val</i>
Regression	1	883.60	883.60	194.91	0.001
Error	3	13.60	4.53		
Total	4	897.20			

Discuss the similarities and the differences between the two approaches and their results, in particular addressing the claim by the colleague that the two analyses will produce “the same results”.

[6]  
[Total 18]

**END OF PAPER**