INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

20 April 2016 (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 11 questions, beginning your answer to each question on a new page.*

5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT 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 An university director of studies records the number of students failing the examinations of several courses. The data are presented in the following stem-and-leaf plot where the stems are with units 10 and the leaves with units 1:

```
0 | 224
0 | 59
1 | 03
1 | 57889
2 |
2 |
3 | 144
3 | 5
```

(i) Determine the range of the data. [1]

(ii) Determine the median number of students failing the examinations of these courses. [1]

(iii) Determine the mean number of students failing the examinations of these courses. [1]
[Total 3]

2 Consider two random variables X and Y.

(i) Write down the precise mathematical definition for the correlation coefficient $\rho(X, Y)$ between X and Y. [1]

Assume now that $Y = aX + b$ where $a < 0$ and $-\infty < b < \infty$.

(ii) Determine the value of the correlation coefficient $\rho(X, Y)$. [3]
[Total 4]

3 A random variable Y has probability density function

$$f(y) = \frac{\theta}{y^{\theta+1}}, \qquad y > 1$$

where $\theta > 0$ is a parameter.

(i) Show that the probability density function of $Z = \ln(Y)$ is given by $\theta e^{-\theta z}$ and determine its range. [3]

(ii) State the distribution of Z identifying any parameters involved. [1]
[Total 4]

4 A manufacturing company is analysing its accident record. The accidents fall into two categories:

- Minor – dealt with by first aider. Average cost £50.
- Major – hospital visit required. Average cost £1,000.

The company has 1,000 employees, of which 180 are office staff and the rest work in the factory.

The analysis shows that 10% of employees have an accident each year and 20% of accidents are major. It is assumed that an employee has no more than one accident in a year.

(i) Determine the expected total cost of accidents in a year. [2]

On further analysis it is discovered that a member of office staff has half the probability of having an accident relative to those in the factory.

(ii) Show that the probability that a given member of office staff has an accident in a year is 0.0549. [3]

(iii) Determine the probability that a randomly chosen employee who has had an accident is office staff. [2]
[Total 7]

5 Players A and B play a game of "heads or tails", each throwing 50 fair coins. Player A will win the game if she throws 5 or more heads than B; otherwise, B wins. Let the random variables X_A and X_B denote the numbers of heads scored by each player and $D = X_A - X_B$.

(i) Explain why the approximate asymptotic distribution of D is normal with mean 0 and variance 25. [3]

(ii) Determine the approximate probability that player A wins any particular game, based on your answer in part (i). [2]
[Total 5]

6 A statistician is sent a summary of some data. She is told that the sample mean is 9.46 and the sample variance is 25.05. She decides to fit a continuous uniform distribution to the data.

(i) Estimate the parameters of the distribution using the method of moments. [4]

The full data are sent later and are given below:

3.5 5.4 7.3 8.5 9.2 10.3 11.4 20.1

(ii) Comment on the results in part (i) in the light of the full data. [2]
[Total 6]

7 A random sample is taken from an exponential distribution with parameter λ. The sample contains some censored observations for which we only know that the value is greater than 3. The observed values are given in the following table:

i	1	2	3	4	5	6	7	8	9	10
x_i	1.3	1.8	2.1	2.2	2.2	2.4	>3	>3	>3	>3

Estimate the parameter λ using the method of maximum likelihood. You are not required to verify that your answer corresponds to the maximum. [4]

8 A scientist is comparing how productive three new strains of wheat are. Thirty widely spread plots of equal size are chosen randomly. Ten plots are planted with each strain and the weight of wheat produced in each plot is measured. The scientist wishes to compare the strains using an analysis of variance and produces the following calculations:

Sum of squares between strains	55.672
Sum of squares within strains	13.332

(i) Perform an analysis of variance on the data. [5]

(ii) (a) Determine the width of a 95% confidence interval for the difference between any two of the mean weights produced for each strain. [4]

The mean weight produced for each sample is:

Strain	1	2	3
Mean	2.03	4.42	5.24

(b) State which means are significantly different. [1]

The scientist discovers that in practice only ten plots in total were chosen, with a third of each planted with each strain.

(iii) Comment on what effect this discovery might have on the scientist's original analysis above. [2]

[Total 12]

9 An insurance company is examining its claims and reserving history over the last ten years.

Ten years ago the data in one year gave the following number of policies with a given number of claims:

Number of Claims	0	1	2	3	4
Number of Policies	871	117	5	5	2

The company assumes that claims occur independently of each other and at a constant rate.

(i) Estimate the rate of claims for an individual policy. [2]

In the most recent year the following data were obtained:

Number of Claims	0	1	2	3	4
Number of Policies	1850	140	5	3	2

(ii) Perform a goodness-of-fit test to investigate if the number of claims in the new data follow a Poisson distribution with the same rate as the rate estimated from the old data. [6]

To date the company has believed that the size of claims, x_i for claim i, was independent of the policyholders' claims history. It now wishes to investigate that belief. It splits the policyholders who have made claims in the most recent year into those with no claims in the preceding five years and those with at least one claim. The total amount of claims in the most recent year is given below:

	Number of policies	Σx_i	Σx_i^2
No claims in previous 5 years	70	6.42m	8.76×10^{11}
Claim in previous 5 years	80	9.22m	1.52×10^{12}

(iii) Perform a test of the null hypothesis that the mean claim amount per policy in each group is equal against the alternative that the mean claim amount per policy is not the same. Use a 5% significance level. [8]

The company wishes to budget for next year. It estimates it will write 2,200 policies, of which half will be for policies with no claims in the last 5 years. It assumes that claims occur at a rate of 0.1 per policy per annum. It also assumes that claim amounts will average £94,000 for policies with no claims in the last 5 years and £120,000 otherwise. Assume that the standard deviation of a claim from a randomly chosen policy is £70,000.

(iv) Determine the estimated mean and variance of the total amount of claims next year. [5]

[Total 21]

10 Consider a large portfolio of insurance policies and denote the claim size (in £) per claim by X. A random sample of policies with a total of 20 claims is taken from this portfolio and the claims made for these policies are reported in the following table:

Claim i	1	2	3	4	5	6	7	8	9	10
Claim size x_i	130	164	170	173	173	175	177	183	183	184

Claim i	11	12	13	14	15	16	17	18	19	20
Claim size x_i	185	186	197	202	208	213	215	229	233	272

For these data: $\Sigma x_i = 3{,}852$ and $\Sigma x_i^2 = 759{,}348$.

(i) Calculate the mean, the median and the standard deviation of the claim size per claim in this sample. [3]

(ii) Determine a 95% confidence interval for the expected value $E[X]$ based on the above random sample, stating any assumptions you make. [3]

(iii) Determine a 95% confidence interval for the standard deviation of X based on the above random sample. [2]

(iv) Explain briefly why the confidence interval in part (iii) is not symmetric around the estimated value of the standard deviation. [1]

An actuary assumes that the number of claims from each policy has a Poisson distribution with an unknown parameter λ. In a new sample of 50 policies the actuary has observed a total of 80 claims yielding an estimated value of $\hat{\lambda} = 1.6$ for the parameter λ.

(v) Determine a 95% confidence interval for λ using a normal approximation. [2]

(vi) Determine the smallest required sample size n for which a 95% confidence interval for λ has a width of less than 0.5. You should use the same normal approximation as in part (v), and assume that the estimated value of λ will not change. [3]

Now assume that the true value of λ is 1.6 and the values calculated in part (i) are the true values. Also assume that all claims in the portfolio are independent and the claim sizes are independent of the number of claims.

(vii) Determine the expected value and the standard deviation of the total amount of all claims from a portfolio of 5,000 insurance policies. [3]

[Total 17]

11 A car magazine published an article exploring the relationship between the mileage (in units of 1,000 miles) and the selling price (in units of £1,000) of used cars. The following data were collected on 10 four year old cars of the same make.

Car	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mileage, x	42	29	51	46	38	59	18	32	22	39
Price, y	5.3	6.1	4.7	4.5	5.5	5.0	6.9	5.7	5.8	5.9

$\Sigma x = 376$; $\Sigma x^2 = 15{,}600$; $\Sigma y = 55.4$; $\Sigma y^2 = 311.44$; $\Sigma xy = 2{,}014.5$

(i) (a) Determine the correlation coefficient between x and y.
(b) Comment on its value.
[5]

A linear model of the form $y = \alpha + \beta x + \varepsilon$ is fitted to the data , where the error terms (ε) independently follow a $N(0, \sigma^2)$ distribution, with σ^2 being an unknown parameter.

(ii) Determine the fitted line of the regression model. [3]

(iii) (a) Determine a 95% confidence interval for β.

The article suggests that there is a "clear relationship" between mileage and selling price of the car.

(b) Comment on this suggestion based on the confidence interval obtained in part (iii)(a).
[7]

(iv) Calculate the estimated difference in the selling prices for cars that differ in mileage by 5,000 miles. [2]
[Total 17]

END OF PAPER