INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

19 April 2012 (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 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.

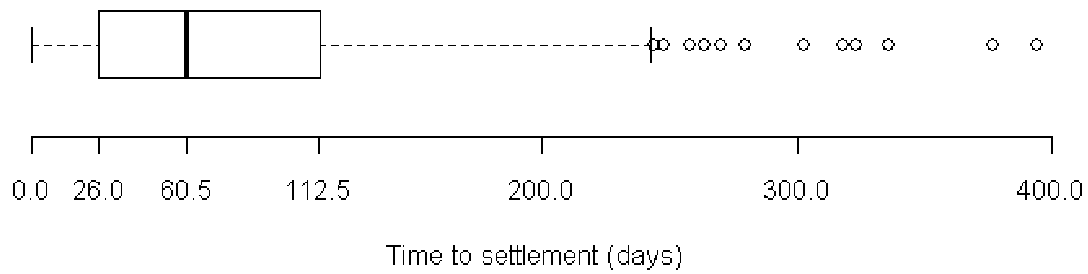
1 The following 24 observations give the length of time (in hours, ordered) for which a specific fully charged laptop computer will operate on battery before requiring recharging.

1.2	1.4	1.5	1.6	1.7	1.7	1.8	1.8	1.9	1.9	2.0	2.0
2.1	2.1	2.1	2.2	2.3	2.4	2.4	2.5	3.1	3.6	3.7	4.5

The owner of this computer is about to watch a film on his fully charged laptop.

Calculate from these data the longest showing time for a film that he can watch, so that the probability that the battery's lifetime will be sufficient for watching the entire film is 0.75. [3]

2 Data were collected on the time (in days) until each of 200 claims is settled by the insurer in a certain insurance portfolio. A boxplot of the data is shown below.



(i) Calculate the median and the interquartile range of the data using the plot. [2]

(ii) Comment on the distribution of the data as shown in the plot. [2]
[Total 4]

3 Two students are selected at random without replacement from a group of 100 students, of whom 64 are male and 36 are female.

Calculate the probability that the two selected students are of different genders. [3]

4 Claim amounts arising under a particular type of insurance policy are modelled as having a normal distribution with standard deviation £35. They are also assumed to be independent from each other.

Calculate the probability that two randomly selected claims differ by more than £100.
[4]

5 Claims on a group of policies arise randomly and independently of each other through time at an average rate of 2 per month.

(i) Calculate the probability that no claims arise in a particular month. [2]

(ii) Calculate the probability that more than 30 claims arise in a period of one year. [2]
[Total 4]

6 In a random sample of 200 people taken from a large population of adults, 70 people intend to vote for party A at the next election.

(i) Calculate an approximate equal-tailed 95% confidence interval for θ, the true proportion of this population who intend to vote for party A at the next election. [3]

(ii) Give a brief interpretation of the interval calculated in part (i). [1]
[Total 4]

7 A coin has two sides, "heads" and "tails". Such a coin with $P(\text{heads}) = p$ is tossed repeatedly until it lands "heads" for the first time. Let X be the number of tosses required.

Suppose the process is repeated independently a total of n times, producing values of the variables $X_1, X_2, \dots, X_n$, where each X_i has the same distribution as X.

Let $Y = \min(X_1, X_2, \dots, X_n)$, so Y is the smallest number of tosses required to produce a "heads" in the n repetitions of the experiment.

(i) Explain why, for each $i = 1, 2, \dots, n$, $P(X_i \geq x)$ is given by

$$P(X_i \geq x) = (1-p)^{x-1}, \quad x = 1, 2, \dots \quad .$$ [2]

(ii) (a) Find an expression for $P(Y \geq y)$.

(b) Hence deduce the probability function of Y.

[5]
[Total 7]

8 In an analysis of variance investigation four treatments are compared using random samples each of size 10. The total sum of squares is calculated as $SS_T = 673.5$ and the between-treatments sum of squares as $SS_B = 148.3$.

(i) (a) Calculate an unbiased estimate of the error variance σ^2.

(b) State the number of degrees of freedom associated with the estimate in part (i)(a). [3]

(ii) Suggest an unbiased estimator of σ^2 that is different from the one used in part (i). [1]

(iii) Comment on which of the two estimators should be used. [2]

[Total 6]

9 A random sample of 200 email messages was selected from all messages delivered through an internet provider company. Each message is monitored for the presence of computer viruses. It is assumed that each message contains a virus with the same probability p, independently from all other messages.

Let Y_i, $i = 1,\ldots,200$ be indicator random variables taking the value 1 if message i contains a virus, and 0 otherwise. Also, let Y denote the total number of messages in this sample found to contain viruses, i.e. $Y = \sum_{i=1}^{200} Y_i$.

(i) Derive expressions for the expected value and the variance of Y in terms of the parameter p, using the indicator variables $Y_1, Y_2, \ldots, Y_{200}$. [4]

(ii) Explain why the approximate distribution of Y is $N(200p, 200p(1-p))$, using the indicator variables $Y_1, Y_2, \ldots, Y_{200}$. [3]

It is found that 38 email messages in this sample contained viruses.

(iii) Calculate an equal-tailed 90% confidence interval for the probability p using the approximate normal distribution from part (ii). [3]

[Total 10]

10 In a portfolio of car insurance policies, the number of accident-related claims, N, made by a policyholder in a year has the following distribution:

No. of claims, n	0	1	2
Probability	0.4	0.4	0.2

The number of cars, X, involved in each accident that results in a claim is distributed as follows:

No. of cars, x	1	2
Probability	0.7	0.3

It can be assumed that the occurrence of a claim and the number of cars involved in the accident are independent. Furthermore, claims made by a policyholder in any year are also independent of each other. Let S be the total number of cars involved in accidents related to such claims by a policyholder in a year.

(i) (a) Determine the probability function of S.
(b) Hence find $E(S)$.

[4]

The expectation $E(S)$ can also be calculated using the formula

$$E(S) = \sum_{n=0}^{2} E(S|N=n)\Pr(N=n).$$

(ii) (a) Find $E(S|N=n)$ for $n = 0,1,2$.
(b) Hence calculate $E(S)$.

[4]
[Total 8]

11 An experiment has three possible outcomes (A, B, C) and a model states that the probabilities of these outcomes are θ, θ^2, and $1 - \theta - \theta^2$ respectively, for some suitable value of $\theta > 0$.

Let n_A, n_B, and n_C be the number of occurrences of outcomes A, B, and C respectively in n (= $n_A + n_B + n_C$) repetitions of the experiment. Let $\ell(\theta)$ represent the log-likelihood function, and let $U(\theta) = \dfrac{\partial \ell(\theta)}{\partial \theta}$.

(i) (a) Show that

$$U(\theta) = \frac{n_A + 2n_B}{\theta} - \frac{n_C(1+2\theta)}{1-\theta-\theta^2}.$$

(b) Hence find a quadratic equation whose solution gives the maximum likelihood estimate of θ.

[5]

(ii) (a) Find an expression for $\dfrac{\partial U(\theta)}{\partial \theta}$.

(b) Hence show that

$$E\left[-\frac{\partial U(\theta)}{\partial \theta}\right] = \frac{n\left(1+4\theta-\theta^2\right)}{\theta\left(1-\theta-\theta^2\right)}.$$ [4]

The results of 100 repetitions of the experiment show that outcome A occurred 51times, outcome B occurred 16 times, and outcome C occurred 33 times.

(iii) (a) Show that the maximum likelihood estimate of θ is $\hat{\theta}$ = 0.4525.

(b) Calculate an estimate of the asymptotic standard error of $\hat{\theta}$.

(c) Find an approximate 95% confidence interval for θ.

[6]
[Total 15]

12 Consider a random sample $X_1,\ldots,X_k$ of size $k = 400$. Statistician A wants to use a χ^2-test to test the hypothesis that the distribution of X_i is a binomial distribution with parameters $n = 2$ and unknown p based on the following observed frequencies of outcomes of X_i:

Possible realisation of X_i	0	1	2
Frequency	90	220	90

(i) Estimate the parameter p using the method of moments. [2]

(ii) Test the hypothesis that X_i has a binomial distribution at the 0.05 significance level using the data in the above table and the estimate of p obtained in part (i). [5]

Statistician B assumes that the data are from a binomial distribution and wants to test the hypothesis that the true parameter is $p_0 = 0.5$.

(iii) Explain whether there is any evidence against this hypothesis by using the estimate of p in part (i) and without performing any further calculations. [2]

Statistician C wants to test the hypothesis that the random variables X_i have a binomial distribution with known parameters $n=2$ and $p = 0.5$.

(iv) Write down the null hypothesis and the alternative hypothesis for the test in this situation. [2]

(v) Carry out the test at the significance level of 0.05 stating your decision. [3]

(vi) Explain briefly the relationship between the test decisions in parts (ii), (iii) and (v), and in particular whether there is any contradiction. [3]

[Total 17]

13 The quality of primary schools in eight regions in the UK is measured by an index ranging from 1 (very poor) to 10 (excellent). In addition the value of a house price index for these eight regions is observed. The results are given in the following table:

Region i	1	2	3	4	5	6	7	8	Sum
School quality index x_i	7	8	5	8	4	9	6	9	56
House price index y_i	195	195	170	190	150	190	200	210	1500

The last column contains the sum of all eight columns.

From these values we obtain the following results:

$$\sum x_i y_i = 10{,}695; \quad \sum x_i^2 = 416; \quad \sum y_i^2 = 283{,}750$$

(i) Calculate the correlation coefficient between the index of school quality and the house price index. [4]

You can assume that the joint distribution of the two random variables is a bivariate normal distribution.

(ii) Perform a statistical test for the null hypothesis that the true correlation coefficient between the school quality index and the house price index is equal to 0.8 against the alternative that the correlation coefficient is smaller than 0.8, by calculating an approximate p-value. [6]

(iii) Fit a linear regression model to the data, by considering the school quality index as the explanatory variable. You should write down the model and estimate all parameters. [3]

(iv) Calculate the coefficient of determination R^2 for the regression model obtained in part (iii). [1]

(v) Provide a brief interpretation of the slope of the regression model obtained in part (iii). [1]

[Total 15]

END OF PAPER