INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

4 April 2019 (am)

Subject CS1B – Actuarial Statistics
Core Principles

Time allowed: One hour and forty-five minutes

INSTRUCTIONS TO THE CANDIDATE

1. *You are given this question paper. Data files relating to this exam were given as part of the pre work material.*
2. *Mark allocations are shown in brackets.*
3. *Attempt all questions. Begin your answer to each question on a new page.*

If you encounter any issues during the examination, please contact the Examinations Team at T. +44 (0) 1865 268 255

1 The following data represent the average total number of marks obtained for a particular exam, observed over seven exam sessions that had been administered by a professional examination body:

87 53 72 90 78 85 83

(i) Enter these data into R and compute their sample mean and variance. [3]

(ii) Investigate whether the Poisson model is appropriate for these data, by calculating the sample mean and sample variance of 10 Poisson samples having the same size and mean as the sample given above. [6]

[Total 9]

2 Consider the $n = 30$ independent and identically distributed observations $(y_1, y_2, \ldots, y_n)$ given below from a random variable Y with probability distribution function $f(y, \theta) = \dfrac{\theta^y e^{-\theta}}{y!}$.

You can enter the y values into R by using:

```
y = c(5,5,6,2,4,10,2,5,5,2,5,3,7,4,4,5,4,6,7,2,8,4,6,4,3,
6,6,6,5,7)
```

By assuming a prior distribution proportional to $e^{-\alpha\theta}$, we can show that the posterior distribution of θ is:

$$f(\theta \mid y_1, y_2, \ldots, y_n) \propto \theta^{\sum_{i=1}^{n} y_i} e^{-(n+\alpha)\theta}$$

We can observe that the posterior distribution of θ is Gamma with parameters $\sum_{i=1}^{n} y_i - 1$ and $n + \alpha$.

(i) (a) Plot the posterior probability density function of θ for values of θ in the interval [3.2, 6.8] and assuming $\alpha = 0.01$.
[Hint: the range of values of θ can be obtained in R by `seq(3.2, 6.8, by = 0.01)`.]

(b) Carry out a simulation of $N = 5{,}000$ posterior samples for the parameter θ.

[8]

(ii) Plot the histogram of the posterior distribution of θ. [2]

(iii) Calculate the mean, median and standard deviation of the posterior distribution of θ. [3]

Two possible values for the true value of parameter θ are θ =15 and θ = 5.

(iv) Comment on these two values based on the posterior distribution of θ plotted in part (ii) and summarised in part (iii). [3]

[Total 16]

3 In a small empirical study, data are recorded on the number of waves per hour and the average wave height per hour at a location just off the coast of Scotland. The data are given in the file named `CS1waves.Rdata`. Loading the data into R will create two vectors in your R workspace, called `Wn` (number of waves per hour) and `Wheight` (average wave height in cm during the hour).

(i) Generate an appropriate plot to visually inspect the relationship between wave height and number of waves per hour. [2]

(ii) Comment on the plot in part (i). [2]

(iii) Calculate Pearson's correlation coefficient between the number of waves per hour and the average wave height. [2]

(iv) Calculate Spearman's rank correlation coefficient between the number of waves per hour and the average wave height. [2]

(v) Comment on your findings in parts (iii) and (iv). [3]

We now model the number of waves per hour, X, as a random variable with a Poisson distribution with unknown parameter λ. The log likelihood function for estimating λ is given by $l(\lambda) = \log(\lambda)\left(\sum_{i=1}^{n} x_i\right) - \lambda n$ where n is the number of observations.

(vi) Plot the log likelihood function for values of $\lambda = 220, 221, \ldots, 280$. [5]

(vii) Determine an approximate maximum likelihood estimate for λ using the plot in part (vi). [3]

(viii) Calculate the exact maximum likelihood estimate of λ. [3]

[Total 22]

4 Actuaries in an insurance company have data which have been collected separately in two different samples. The actuaries are concerned about the validity of the equal variance assumption between the two underlying populations when carrying out a two-sample *t*-test. The two samples are given below:

Sample 1:	21	22	28	27	20	23	26	32	25	21	30
Sample 2:	19	18	38	33	24	39	22	20	28	26	30

(i) Enter the data into R. [1]

(ii) Perform an *F*-test at a 5% significance level to compare the population variances. You should state your conclusion clearly. [4]

(iii) (a) Perform a suitable *t*-test for the null hypothesis that the two population means are equal at a 5% significance level. [4]

(b) Calculate a two-sided 95% confidence interval for the difference in the population means. [2]

(c) Comment on the answers in parts (iii)(a) and (iii)(b). [2]

A different approach for checking the equal variance assumption between the two underlying populations is suggested. It involves a two-sample *t*-test: for each sample, calculate the absolute deviations defined as the absolute value of the difference between each observed value and the mean of the sample. Apply a two-sample *t*-test to the two sets of absolute deviations. The idea is that if the samples have equal variances, then the absolute deviations will have the same mean for both samples.

(iv) (a) Calculate the two sets of absolute deviations.
[Hint: `abs(x)` gives the absolute value of *x*.] [2]

(b) Perform a two-sample *t*-test on the two sets of absolute deviations at a 5% level, stating your conclusion. [4]

(v) Comment on your conclusions in parts (ii) and (iv)(b). [3]

(vi) Comment on whether or not the equality of the population means can still be tested. [2]

[Total 24]

5 A statistician is carrying out an exercise to analyse a dataset that describes the failure times of outdoor telephone cables, with respect to the cable material quality (graded 1 to 4) and level of rainfall in centimetres that the cable is exposed to.

The data given in the file “Cables_dataset.csv” show failure times in years for 20 different cables.

(i) Fit a linear model to the data with the failure time as the response, including both cable material quality and level of rainfall as the two covariates. Your answer should include a summary of the fitted model. [5]

(ii) (a) State the formula of the model fitted in part (i), clearly explaining the notation that you use.

(b) Comment on the significance of the parameters of the model fitted in part (i).

[6]

(iii) (a) Plot the residuals of the model in part (i).

(b) Comment on the plot created in (iii)(a).

[4]

An analyst suggests that the 6th row of the original data should be removed.

(iv) (a) Construct a new data set from the original data “Cables dataset.csv” with the 6th row removed. [2]

(b) Justify the removal of the 6th row from the original data. [2]

(v) (a) Fit a linear model to the new data set constructed in part (iv)(a). [1]

(b) Comment on the fit of the model from part (v)(a) compared to the model fitted in part (i), by comparing suitable statistics from the R outputs. [3]

(vi) (a) Fit a generalised linear model (GLM) to the data set constructed in part (iv)(a) using a Gamma distribution.

(b) State the formula of the model fitted in part (vi)(a), clearly explaining the notation that you use.

(c) Comment on the significance of the parameters of the model fitted in part (vi)(a).

[6]

[Total 29]

END OF PAPER