INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

16 April 2021 (am)

Subject CS1 - Actuarial Statistics
Core Principles

Paper B

Time allowed: One hour and forty-five minutes

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator.

If you encounter any issues during the examination please contact the Assessment team at T. 0044 (0) 1865 268 873.

1 A researcher records the levels of caffeine present in their bloodstream at various time intervals after drinking a cup of coffee.

T: Time elapsed (30-minute intervals)	1	2	3	4	5	6	7	8	9	10
C: Caffeine in blood (milligrams)	52.2	39.3	28.3	19.03	13.96	11.46	8.78	6.55	5.43	5.03

You can enter the values, for *T* (time) and *C* (caffeine), into R using:

```
time = c(1,2,3,4,5,6,7,8,9,10)

caffeine = c(52.2,39.3,28.3,19.03,13.96,11.46,8.78,6.55,5.43,5.03)
```

(i) Plot a scatterplot of the data. [2]

(ii) Comment on the relationship between *C* and *T* based on the plot in part (i). [3]

(iii) Justify why a logarithmic transformation of *C* is appropriate. [3]

(iv) Perform the transformation suggested in part (iii). [1]

(v) (a) Plot a scatterplot of the transformed data.

(b) Calculate the Pearson correlation coefficient for the transformed data.

(c) Comment on the scatterplot and Pearson correlation coefficient produced in parts (v)(a) and (v)(b).

[6]

[Total 15]

2 An analysis was carried out to investigate the fairness of two exam markers. They both marked the same 150 exam papers, with 10 questions and total possible marks of 100 for each exam paper. The data were collected and arranged into 10 equally spaced groups, with marks rounded to the nearest whole number.

Below are the frequencies of the marks given by each of the exam markers:

Exam marks	0–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
Marker 1	1	8	14	22	33	34	21	9	6	2
Marker 2	0	4	16	25	27	42	23	4	9	0

You can enter the exam marks for Marker 1 and Marker 2 into R using:

```
marks_1 = c(1,8,14,22,33,34,21,9,6,2)

marks_2 = c(0,4,16,25,27,42,23,4,9,0)
```

(i) (a) Plot two graphs, one for each marker, for the distribution of the exam marks given by the two markers.

[**Hint:** You may find the R command barplot() useful.]

(b) Comment on the graphs produced in part (i)(a).

[7]

One of the marked exam papers is selected at random and the scores given by each of the markers are analysed further by question. Below are the scores given by each of the exam markers for this selected exam paper for each of the 10 questions:

Question	1	2	3	4	5	6	7	8	9	10
Marker 1	4	1	5	1	4	6	4	5	3	6
Marker 2	3	2	4	0	3	4	2	3	3	6

You can enter the scores for Marker 1 and Marker 2 into R using:

```
marker_1 = c(4,1,5,1,4,6,4,5,3,6)
marker_2 = c(3,2,4,0,3,4,2,3,3,6)
```

(ii) Perform a suitable test to determine whether the difference in the mean scores of the two markers is zero or not, at the 5% confidence level, taking into account that the two markers have marked the same exam paper. [5]

(iii) Perform the test specified in part (ii), ignoring that the two markers have marked the same exam paper. [4]

(iv) Comment on your conclusions from parts (ii) and (iii). [2]

(v) Comment on the issues arising when analysing paired data as independent samples, and independent samples as though they were paired. [4]

[Total 22]

3 A study was carried out to estimate the proportion, p, of workers that commute by train to work. A total of $n = 200$ workers were sampled at random and were asked the question: 'Do you take the train to work?' The workers' answers were recorded as a binary outcome, y_i, for worker i, with 1 for yes and 0 for no. The data are available in the file `BinaryTrain.RData`.

Two commuters, Alice and Norman, were interested in the study and proposed different prior distributions for the proportion p.

Alice assumed a discrete prior distribution $g(p)$ given in the following table:

p	0.1	0.2	0.3	0.4	0.5
$g(p)$	0.5	0.2	0.2	0.05	0.05

Norman chose to use a beta prior distribution for p, with parameters 3 and 12.

(i) (a) Calculate the mean and the standard deviation for Alice's prior distribution. [4]

(b) Generate 10,000 random values from Norman's prior distribution. [1]

(c) Calculate the mean and standard deviation of the values generated in part (i)(b). [2]

(d) Comment on whether or not Alice and Norman have similar prior beliefs for p. [2]

Norman's beta prior distribution for p is adopted for the remainder of the question.

The likelihood of the model in the study is given by:

$$L(p) \propto p^{\Sigma y_i} (1-p)^{n-\Sigma y_i} .$$

The posterior density of p is given by:

$$f(p|\boldsymbol{y}) \propto p^{2+\Sigma y_i}(1-p)^{11+n-\Sigma y_i} ,$$

where $\sum y_i$ is the total sum of all the binary data.

(ii) Plot the shape of the posterior density of p without identifying it. [4]

(iii) Plot the density of Norman's prior distribution by setting `ylim = c(0,14)`. [3]

The posterior distribution of p is beta with parameters $3 + \sum y_i$ and $12 + n - \sum y_i$.

(iv) (a) Plot the posterior density of p by adding it to the plot in part (iii). [3]

(b) Compare the two densities using your answer in part (iv)(a). [1]

(c) Comment on the extent to which the posterior distribution is affected by the prior distribution. [1]

(v) Determine a 90% interval estimate for p based on its posterior distribution. [2]

(vi) Determine the exact posterior probability that p exceeds 0.25. [2]

(vii) (a) Generate 10,000 samples from the posterior distribution of p.

(b) Calculate the proportion of sampled values of p that exceed 0.25.

(c) Compare your answer in part (vii)(b) with your answer in part (vi). [3]

[Total 28]

4 A statistician wants to model the number of passengers boarding a bus from a bus stop close to a student residential area. They can think of three explanatory variables: which `route` it is (at 8 am or 9 am), if it is during the `semester` or not, and the temperature (`temp`) in degrees Celsius. The statistician has data for 20 days, given in the file named `CS1passenger.RData`, and believes that the response variable (`Passengers`) follows a Poisson distribution. After loading the data into R, the data frame `data_passenger` with all variables (`Passengers, route, semester, temp`) will be available.

(i) State the linear predictor corresponding to models specified with the following R code, explaining all relevant terms:

(a) `temp+semester`

(b) `temp*semester`

(c) `temp*semester + route`

[6]

(ii) (a) Fit a Poisson Generalised Linear Model (GLM) to the data set for the model in part (i)(c). Label this model as Model1. Your answer should include a summary of the fitted model.

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

[6]

(iii) (a) Fit an improved model for the model in part (ii)(a), using your answer in part (ii)(b). Label this model as Model2.

(b) Justify why Model2 improves Model1 by referring to the R output.

[4]

You are given a new model (Model3), specified by the following R code:

```
Model3 <- glm(Passengers~temp+temp:semester,family=poisson(link="log"))
```

(iv) (a) Demonstrate that Model3 outperforms the models defined in parts (i)(a) and (i)(b). [8]

(b) Comment on your answer in part (iv)(a). [1]

(v) (a) Draw a suitable plot, for the residuals of Model3, for checking the model's validity.

(b) Comment on the plot in part (v)(a).

[6]

(vi) Calculate the predicted number of passengers for an 8 am route during the semester at a temperature of 0 degree Celsius, using Model3. [4]

[Total 35]

END OF PAPER