INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

18 September 2023 (am)

Subject CS1 – Actuarial Statistics
Core Principles

Paper B

Time allowed: One hour and fifty 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 on T. 0044 (0) 1865 268 873.

© Institute and Faculty of Actuaries

1 (i) Simulate 10,000 samples from an Exp(3) distribution, each of size $n = 10$, using the inverse transform method. You should save the generated values in R for later use.

You must use the command `set.seed(3202)` to initialise the random number generator, before you start the simulation. [3]

(ii) Plot a histogram of the means of the samples generated in part (i), using an appropriate option in R for plotting the histogram on the probability density scale. [3]

(iii) Plot the Probability Density Function (PDF) of the sampling distribution of the sample mean, under the Central Limit Theorem (CLT), corresponding to the samples generated in part (i). The graph of the PDF should be superimposed on the histogram produced in part (ii). [7]

(iv) Comment on the sampling distribution of the sample mean and the application of the CLT in this case based on your answers to parts (ii) and (iii). [3]

[Total 16]

2 A researcher wants to investigate the proportion of electric cars among all registered cars in two large populations (denoted as populations A and B). Two samples are considered, one from each population. The first sample consists of $n_A = 900$ registered cars, while the second sample consists of $n_B = 1{,}200$ registered cars. The type of each car (electric or not) is assumed to be independent of the type of other cars in the samples.

(i) Simulate the two samples in R assuming that the proportion of electric cars in population A is 0.02, while in population B it is 0.025, and save the simulated samples for later use.

You must use the command `set.seed(12345)` to initialise the random number generator, before you start the simulation. [2]

The true proportions of electric cars in the two populations are unknown to the researcher, and the researcher wants to estimate them.

(ii) Determine an equal-tailed 99% confidence interval for the difference in the true proportions of electric cars in the two populations A and B, using your sample data. [5]

Based on previous knowledge, the researcher believes that the proportion of electric cars in population A is lower than the proportion of electric cars in population B.

(iii) Perform a hypothesis test to investigate this belief using your sample data. In doing so, you should include the following steps, in addition to any other necessary steps:

- Compute the value of a suitable test statistic under an appropriate normal approximation.
- Compute the *p*-value of the test.
- State your conclusion based on the calculated *p*-value.

[9]
[Total 16]

3 A data science actuary wants to build a linear regression model for cars that will predict the stopping distance based on their speed.

The actuary will work on the dataset that already exists in R known as ‘Cars’, which can be initialised using the following code:

```
data("cars")

attach(cars)
```

This dataset has 50 observations of two variables:

- The first variable ‘speed’ is speed, in miles per hour (mph).
- The second variable ‘dist’ is stopping distance, in feet.

(i) Fit a linear regression model to the data using stopping distance as the response variable and speed as the only explanatory variable, stating the estimated intercept and slope of the regression line. [2]

The actuary wants to predict the stopping distances for the following ten speed values:

6	7	16	22	28	33	40	42	57	64

(ii) Estimate the ten expected stopping distances, based on your fitted model from part (i). [5]

The actuary is concerned about the uncertainty around the mean predictions.

(iii) Calculate ten confidence intervals, one for each of the expected stopping distances in part (ii), using a 90% confidence level. [4]

(iv) (a) State the interpretation of the 90% confidence interval for the stopping distances for cars with speed 64 mph, using your answer to part (iii).

(b) Comment on the suitability of the linear regression model, based on the confidence intervals obtained in part (iii).

[2]
[Total 13]

4 A coin used by a referee before the start of a football game is suspected to be unfair. The hypothesis of a fair coin is tested by tossing the coin 200 times. The coin is then declared fair if between 85 and 115 tosses (both inclusive) are heads, otherwise the hypothesis of fair coin is rejected.

(i) (a) Calculate the exact probability, to two decimal places, of rejecting the hypothesis of fair coin when it is actually true, using the test above. [5]

(b) Calculate an approximate probability, to two decimal places, of rejecting the hypothesis of fair coin when it is actually true, using the test above and a normal approximation. [5]

(c) Comment on your results to parts (i)(a) and (i)(b). [1]

(ii) Calculate the exact probability, to four decimal places, of not rejecting the hypothesis of the coin being fair when the actual probability of heads is equal to 0.7. [4]

(iii) Calculate the power of the test, to four decimal places, for the probability of heads taking values from 0.1 to 0.9 with a 0.01 step. You are not required to print the results. [4]

[**Hint:** you might find the `seq()` command useful.]

(iv) Plot the power of the test against the probability of heads values in part (iii). [3]

(v) Comment on the plot in part (iv). [2]

[Total 24]

5 A medical study examines the dependence of Intensive Care Unit (ICU) admission rates on a numbers of risk factors, for patients with a certain medical condition. The data are available in the file ‘medical_data.RData’, which contains the following variables for 16 groups of patients:

- Rate: the rate of admission to ICU in a particular group of patients.
- History: the medical history of the patients in a particular group (two levels; 0: short; 1: long).
- Operation: indicator showing if patients in a particular group have had a recent surgical operation (two levels; 0: no; 1: yes).
- Comorbidity: an index of existing comorbidities in a particular group (four levels; 1–4 for comorbidities A–D).
- Age: the median age of patients in a particular group.

(i) Plot the admission rates against Age, while also distinguishing between different History levels on your graph. [5]

[**Hint:** you can use the `points(…)` command in R. Also, make sure that all 16 rates are shown in your graph, by adjusting the axes limits if necessary, e.g. use `xlim`, `ylim` in R.]

(ii) Plot the admission rates against Age, while also distinguishing between different Operation levels on your graph. [5]

[**Hint:** See hint in part (i).]

(iii) Comment on the impact of the variables History, Operation and Age on the admission rates, based on your graphs from parts (i) and (ii). [2]

(iv) Fit a normal Generalised Linear Model (GLM) to the data to investigate the dependence of admission rates on the three following variables: History, Operation and Age (when these variables are only included as main effects in the model).

You should use the logarithmic link function in the model. [3]

(v) Comment on the impact of the variables History, Operation and Age on the admission rates, based on the output of the model fitted in part (iv). [3]

(vi) Fit a normal GLM to the data, including the variable Comorbidity as a main effect, in addition to the other three variables included in part (iv). You should use the logarithmic link function in the model. [3]

(vii) Comment on the impact of all four variables on the admission rates, based on the output of the model fitted in part (vi). [4]

(viii) Determine which of the two models fitted in parts (iv) and (vi), should be preferred for investigating the dependence of admission rates on the available variables. [2]

(ix) Calculate the predicted rate of admission to ICU for a group of patients with:

- short medical history.
- no recent surgical operation.
- comorbidity B.
- median age 62.

The prediction should be based on the preferred model from part (viii). [2]

(x) Comment on a disadvantage of the GLMs used in parts (iv) and (vi), in terms of predicting admission rates. [2]

[Total 31]

END OF PAPER