INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

16 September 2022 (am)

Subject CS1 – Actuarial Practice
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.

1 A Systems Actuary is developing an automated application to replace a time-consuming manual process. It is assumed that the number of errors, X, under this automated application process follows a Poisson distribution with mean 6. The Actuary wants to perform an analysis on the error rate for the automated process, using the sample mean $\bar{X}$.

Use the command `set.seed(2022)` to initialise the random number generator.

(i) Determine an estimate for the mean and variance of the sample mean $\bar{X}$ by implementing 5,000 Monte Carlo repetitions, each involving a sample of size 150 from the assumed Poisson distribution. You should save the Monte Carlo $\bar{X}$ values for later use. [7]

The Actuary recalls that a Normal approximation can be used, by referring to the central limit theorem.

(ii) Write down the approximate distribution of $\bar{X}$, by using the central limit theorem. [1]

(iii) Compare the approximation in part (ii) with your answer to part (i). [1]

The Actuary wants to justify using the Normal approximation by comparing all the quantiles in one go of $\bar{X}$ and the Normal distribution, using a QQ plot.

(iv) Construct a QQ plot for $\bar{X}$ and the Normal distribution, using the Monte Carlo $\bar{X}$ values produced in part (i). [3]

(v) Comment on the plot from part (iv). [4]

[Total 16]

2 An insurance company designed a new product and wanted to assess its clients' responses to the product. A survey was carried out giving an opportunity to each participating client to give a positive or negative response to the product, independently of other clients. Let X be the random variable representing the positive responses to the new product.

(i) Identify the distribution of X and its parameters. [1]

Out of 160 clients who responded independently to the survey, 101 gave a positive response for the new product.

The probability of obtaining a positive response for the product is denoted by θ and a Beta prior distribution with parameters (α, β) is assumed for θ. The posterior distribution of θ is proportional to:

$$f(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1},$$

where x is the number of positive responses obtained out of n clients surveyed.

(ii) Specify the posterior distribution of θ with its parameters. [2]

(iii) Comment on the prior distribution of θ in relation to the posterior distribution. [1]

(iv) State the parameter values for which the prior is a Uniform(0, 1) distribution. [1]

(v) (a) Plot the prior density of θ with the parameters obtained in part (iv). Set the maximum limit of the y axis to 12. [2]

(b) Plot the posterior distribution of θ on the same graph as above. [2]

[**Hint**: you may find the `lines` function useful.]

An Analyst consulted by the company suggests that based on previous experience, a Beta prior with parameters (40, 24) is more appropriate.

(vi) Plot the new prior and posterior distributions of θ on the same graph from part (v). [3]

(vii) Comment on the plots obtained in parts (v) and (vi). [2]

The company will put the new product on the market only if there is a high probability that θ is higher than 60%.

(viii) (a) Calculate the probability $P(\theta > 0.6 \mid X)$ in the case of both priors; that is, Uniform(0, 1) and Beta with parameters (40, 24). [4]

(b) Comment on your answer to part (viii)(a). [2]

[Total 20]

3 A male athlete ran 1 mile in 254.4 seconds on 31 May 1913. This was a world record at the time. The data file `mile_records.Rdata` contains the dates measured in days since 31 May 1913 and the times (in seconds) of all new world records for males over the 1-mile distance. The data set `mile_records.Rdata` contains 32 records. The variables are called `record.date` and `record.time` in the Rdata file. You can load the file with `load("mile_records.Rdata")`.

(i) Plot `record.time` as a function of `record.date`. [2]

(ii) Calculate the Pearson's correlation coefficient between the two data sets. [2]

(iii) Fit a linear regression model to the data using `record.time` as the response variable and `record.date` as the only explanatory variable. State the estimated intercept and slope of the regression line. [3]

(iv) Plot the regression line by adding it to the plot from part (i). [2]

(v) Perform a statistical test in order to test the null hypothesis that the slope of the regression line in part (iv) is zero, against a suitable alternative using the output of the fitted model from part (iii). [4]

(vi) Comment on the relationship between the two variables. [4]

For simplicity, you can assume that 1 year has 365 days.

(vii) Estimate the expected time in seconds of the world record 100 years after the most recent record in this data set. [4]

(viii) Calculate the number of years (from the most recent record) in which you expect the world record to be 2 minutes, based on your fitted model from part (iii). [4]

(ix) Comment on the suitability of the linear regression model for modelling `record.time` as a function of `record.date`. [2]

[Total 27]

4 An Analyst is asked to produce a report on the existing imbalance in salary in jobs related to Science, Technology, Engineering and Maths (STEM). The Analyst considers a sample of 100 employees in STEM-related jobs. For each of these employees, information is provided on starting and current salary (in units of £5,000), gender, type of job and job location, the employee's age and their relevant experience. The data is given in the file named `employee.RData`. After loading the data into R, the data frame `data_employee`, with its variables listed below, will be available.

Variable	*Variable definition*
`salary.current`	Current yearly salary
`salary.start`	Starting yearly salary
`gender`	Male = 0, female = 1
`job.type`	1 = geneticist, 2 = civil engineer, 3 = statistician, 4 = biophysicist, 5 = pathologist
`job.location`	Big city = 0, small city = 1
`age`	Age in years
`experience`	Relevant job experience in years

(i) Write down the categorical and the numerical variables in the data. [2]

(ii) Plot a scatter graph between each pair of the numerical variables using your answer to part (i). [3]

(iii) Comment on the relationship between the current salary and the remaining numerical variables. [2]

(iv) (a) Calculate the lower quartile, median, upper quartile and the mean for the current yearly salary. [3]

(b) Test whether the proportion of male employees with current salary below 9.86 is significantly different from the proportion of female employees with current salary below 9.86. [11]

[**Hint**: `salary.current[gender==0]` gives a vector of current salary for males.]

(v) Determine the median, mean and variance of the current yearly salary for each of the job types in `job.type`. [6]

[**Hint**: `salary.current[job.type==1]` gives a vector of current salary for geneticists.]

(vi) (a) Test at the 5% level whether the mean starting salary and the mean current salary are significantly different.

(b) Test at the 5% level whether the mean current salary for big-city employees is greater than the mean current salary for small-city employees.

[10]
[Total 37]

END OF PAPER