INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

20 September 2021 (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 In a small country, a political election was held recently to decide on a political party to govern the country. A survey was conducted to monitor the approval rating for the winning political party.

A sample of 15 voters were asked to complete the same survey twice: once before the election (Approval before) and a second time 6 months after the election (Approval after). The survey asked each of the 15 voters to record their approval for the winning political party on a scale from 1 to 10, where

1 = Strongly disapprove and 10 = Strongly approve.

The results are shown in the table below:

Voter	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Approval before	8	6	8	7	7	4	2	10	8	7	10	8	8	9	6
Approval after	5	6	2	3	4	1	4	7	4	2	10	1	5	3	7

The values can be entered in R using:

```
approval_before <- c(8, 6, 8, 7, 7, 4, 2, 10, 8, 7, 10,
8, 8, 9, 6)

approval_after <- c(5, 6, 2, 3, 4, 1, 4, 7, 4, 2, 10, 1,
5, 3, 7)
```

(i) Calculate the means for the Approval before and Approval after results. [2]

(ii) Outline why Kendall's Tau would be a suitable correlation coefficient to use for these results. [2]

(iii) Calculate Kendall's Tau coefficient between the Approval before and Approval after results. [2]

(iv) Comment on your results in parts (i) and (iii). [3]

[Total 9]

2 Use the command `set.seed(2021)` to initialise the random number generator. When you execute any R code in this question, make sure you run the entire R script including the line `set.seed(2021)` at the start.

(i) (a) Generate a sample of size $n = 180$ from an Exponential distribution with parameter $\lambda = 5$.

(b) Calculate the sample median of the sample in part (i)(a).

(c) Calculate the median of the Exponential distribution with $\lambda = 5$.

(d) Comment on your results in parts (i)(b) and (i)(c).

[8]

The maximum likelihood estimator for λ is given by $\hat{\lambda} = 1/\overline{X}$ where $\overline{X}$ is the sample mean.

(ii) (a) Estimate the parameter λ using the maximum likelihood estimator and the sample in part (i)(a).

(b) Generate another sample of size $n = 180$ from an Exponential distribution with parameter $\lambda = 5$ and estimate the parameter λ using the maximum likelihood estimator and the new sample.

(c) Comment on the estimated values in parts (ii)(a) and (ii)(b).

[7]
[Total 15]

3 An insurer's marketing team has developed a new lottery that gives each new customer the chance to win a cash prize. The insurer states that any new customer can win with probability 0.36 independently of all other customers.

Let Y denote the number of winners in a random sample of 900 new customers.

(i) State a suitable distribution for the random variable Y. [1]

(ii) Calculate, to four decimal places, the probabilities $P[k < Y \le k + 20]$ for all values $k = 220, 240, 260, \ldots, 400$. [5]

It is suggested that the true distribution of Y can be approximated with a Normal distribution with expectation $E[Y] = 324$ and a standard deviation of 14.4.

(iii) Calculate, to four decimal places, the probabilities $P[k < Y \le k + 20]$ for all values $k = 220, 240, 260, \ldots, 400$ based on the suggested Normal distribution. [4]

(iv) Plot the values calculated in parts (ii) and (iii) on the same graph. [4]

[**Hint**: you may find the R command lines(…) useful.]

(v) Discuss your answers in parts (i)–(iv). [5]

[Total 19]

4 An actuarial modeller in a health insurance company is constructing a Generalised Linear Model (GLM) to analyse claim numbers for its critical illness policies. For every policy over the past year, the company has collected the number of reported claims (`Claim_number`) and data on the following covariates:

`Age:`	Age of policyholder, a number between 18 and 80
`Gender:`	Gender of the policyholder (Male or Female)
`Region:`	A description of the region where the policyholder lives
`Pre_existing_health_condition:`	A categorical value representing whether the policyholder has a pre-existing health condition, 0 = no and 1 = yes

The data given in the file named `Claims_Experience.RData` show the past year's reported claims for this company's critical illness policies. After loading the data into R, using the command `load("Claims_Experience.RData")`, the data frame `data_claims` with its variables listed above will be available.

The modeller wants to fit a GLM with the `Claim_number` as the response variable and is deciding whether to fit the GLM using a Poisson or Normal distribution.

(i) Fit a GLM, using a Normal distribution, that treats age as a numerical variable and the remaining three covariates as factors, calling this `model_g`. Your answer should include the estimated value, standard error and *p*-value of each parameter in the model. [7]

(ii) (a) Fit a GLM, using the Poisson distribution, that treats age as a numerical variable and the remaining three covariates as factors, calling this `model_p`.

Your answer should include the estimated value, standard error and *p*-value of each parameter in the model.

(b) Justify which model the modeller should use, using your answers to parts (i) and (ii)(a).

(c) Comment on the dependence of the number of reported claims on `Pre_existing_health_condition` based on your answers to parts (ii)(a) and (ii)(b).

[7]

The modeller has fitted the GLM using a Poisson distribution and now wants to investigate which of the three factors should be selected in the model, by using a method similar to the backward selection approach. The following models are considered:

`model_p2`: removes `Gender` from `model_p`

`model_p3`: removes `Gender` and `Pre_exisiting_health_condition` from `model_p`

`model_p4`: removes `Gender`, `Pre_exisiting_health_condition` and `Region` from `model_p`

(iii) Compare the values of the Akaike's Information Criterion for the four models `model_p`, `model_p2`, `model_p3` and `model_p4`. [6]

(iv) (a) Comment on your answer to part (iii).

(b) Comment on the differences between the process in part (iii) and the full backward selection method.

[3]

The modeller suggests refining the fitted GLM (`model_p`) by also including the interactions between the variables: `Age`, `Region` and `Pre_existing_health_condition`.

(v) (a) Fit an appropriate model, to include the interactions between the variables `Age`, `Region` and `Pre_existing_health_condition`.

(b) Justify whether the refined model in part (v)(a) improves the model (`model_p`) fitted in part (ii)(a).

(c) Comment on whether any of the interactions are significantly associated with claim numbers.

[7]
[Total 30]

5 An analyst collected samples of the prices (in £000s) of 25 one-bedroom flats in each of two different cities. The data from these two samples are saved in the file `onebedflat.Rdata`. After loading the data into R, using the command `load("onebedflat.Rdata")`, the data frame `onebedflat` with its variables listed below will be available.

City1: one-bedroom flat prices in the first city.
City2: one-bedroom flat prices in the second city.

It is assumed that the two samples come from Normal populations with equal variances.

(i) Calculate an appropriate test statistic for the hypothesis of equal means in the two corresponding populations using these data. [2]

(ii) Test at a 1% significance level whether the mean flat prices are the same in the two cities, against the alternative that they are different, based on your answer to part (i). [4]

The standard assumptions for a two-sided test of the hypothesis of equal means in two populations are that the populations:

- (A1) follow Normal distributions
- (A2) have equal variances.

The analyst suspects that those assumptions may not be satisfied, and they are therefore interested in learning about the true significance level for such tests when assumptions (A1) and (A2) do not necessarily hold.

With two independent samples, given a particular choice of the significance level, α, the true significance level for a two-sided test of the hypothesis of equal means can be calculated as

$$\alpha_{\text{true}} = P(|\text{test statistic}| \geq \text{ critical point}),$$

where the critical point is the $\alpha/2$ quantile of an appropriate distribution.

An estimate of the true significance level for this two-sided test, when the chosen level of significance is $\alpha = 0.1$, can be computed using the R Monte Carlo simulation code provided below:

```
set.seed(123) #
alpha = 0.1
m=20
n=20
N=8000
nrej=0
for (i in 1:N) {  #
  x=rnorm(m,mean=0,sd=1); y=rnorm(n,mean=0,sd=1) #
  ts=t.test(x, y, var.equal=TRUE)$stat #
  df = n+m-2  #
  if (abs(ts)>=qt(1-alpha/2, df)) #
  nrej=nrej+1  #
}
alpha_est=nrej/N #
```

Note that the code provided above lacks best practice of including appropriate comments.

(iii) Explain the provided code by placing comments after each '#' sign in the code. [8]

(iv) Estimate the true significance level, by using the code provided above, when the two populations are from a standard Normal distribution. [1]

(v) Estimate the true significance level, by adjusting the code provided above to reflect the following pairs of populations:

(a) The first population is from a standard Normal distribution and the second population is from a Normal distribution with mean 0 and standard deviation 10.

(b) The first population is from a Normal distribution with mean 10 and standard deviation 2 and the second population is from an Exponential distribution with mean 10.

[8]

(vi) Compare your estimates of the true significance level under all scenarios in parts (iv), (v)(a) and (v)(b), stating your conclusions. [4]

[Total 27]

END OF PAPER