INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

26th November 2024

Subject CS1B – Actuarial Statistics (Paper B)

Time allowed: 1 Hours 45 Minutes (09.30 – 11.15 Hours)

Total Marks: 100

Q. 1) Consider a random sample $X_1, X_2, \ldots\ldots.., X_n$ from a Chi-Square distribution with 2 degrees of freedom and define $Y = \sum_{i=1}^{n} X_i$

i) State the distribution of Y, giving all the parameters of the distribution.

[**Hint:** *If W ~ Gamma (α,λ) then 2λW has* χ^2 *distribution with 2α degrees of freedom.*] (3)

Following 15 random numbers have been generated from a U(0,1) distribution using R.

```
0.07991847,    0.82064314,    0.33683219,    0.93005953,    0.31919393,
0.92695533,0.76263949, 0.51740370, 0.49224880, 0.46354694, 0.89832157,
0.21920729, 0.20471780, 0.19055074, 0.69137537
```

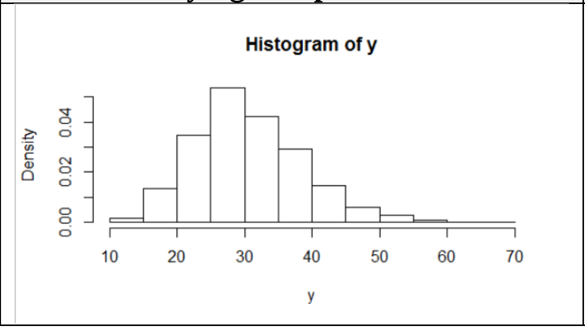
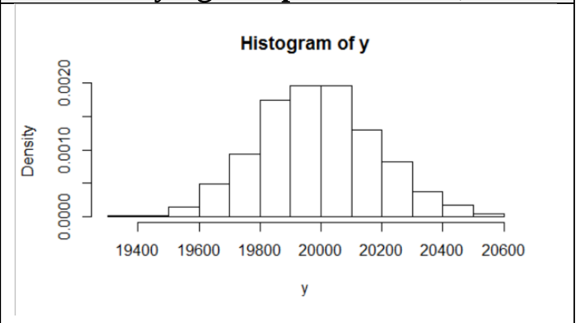
ii) Using the 15 random numbers provided above, simulate a sample $x_1, x_2, \ldots\ldots., x_{15}$ from a chi-square distribution with 2 degrees of freedom. Based on this sample, calculate the value of Y.

Note: *You can directly copy these random numbers from* ***Codes.docx*** *file in your R Script.* (3)

iii) Using the values of $x_1, \ldots\ldots., x_n$ generated in part (ii), test whether the standard deviation of X is equal to 2.5 from scratch using `qchisq()` to determine the critical values and `pchisq()` function to determine the p-value. Clearly state the null and alternate hypothesis, the p-value and the conclusion of the test. Perform the test at 5% level of significance. (6)

iv) Write and execute R script to generate 1,000 samples of x of size 15. Then, calculate the sum of 1,000 corresponding values of Y based on these samples. Use `set.seed(47)` and print the sum of 1,000 y values. (4)

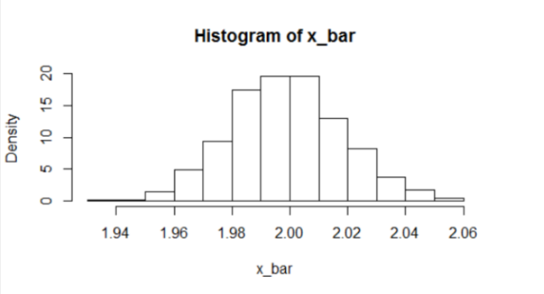
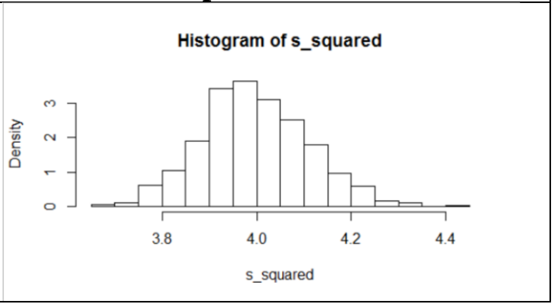
v) Histograms have been plotted showing the relative frequencies of y_1 to y_{1000} with underlying sample sizes of 15 and 10,000 respectively. They are given below:

Underlying Sample Size of 15	**Underlying Sample Size of 10,000**
	

Comment on the difference between the two histograms of Y for sample sizes 15 and 10,000, particularly in relation to the central limit theorem and how the sample size affects the shape of the distribution. (3)

vi) By appropriately modifying the code written in part (iv) or otherwise, write code in R to generate sample means $\overline{X}$ and sample variances S^2 for 1,000 random samples of Y each having 10,000 values of x. Use `set.seed(47).` You are NOT required to execute the code. (5)

vii) Histograms have been plotted for sample means $\overline{X}$ and sample variannces S^2 using the code written in part (vi). They have been given overleaf:

Sample means	**Sample variances**
	

By doing visual inspection of these histograms or otherwise, check whether –

(a) $\overline{X}$ is an unbiased estimator of the population mean μ
(b) S^2 is an unbiased estimator of the population variance δ^2 (4)

viii) Also comment on the normality of the distributions of $\overline{X}$ and S^2 using the above histograms. (2)

[30]

Q. 2) There are three life insurance companies having below summary of claims for last 4 years:

Aggregate claims in millions:

Company	**Year**			
	1	**2**	**3**	**4**
A	14.2	15.8	22.7	19
B	58.6	63.1	81	64.2
C	123	132	161	133

No of claims:

Company	**Year**			
	1	**2**	**3**	**4**
A	163	189	252	199
B	4435	4761	5576	4581
C	16184	17443	20102	18000

We want to calculate credibility premiums and estimate risk premiums under the assumptions of the Empirical Bayes Credibility Theory (EBCT) Model 2 using the following code in R:

```
claims<-data.frame(
  Year1 = c(14.2,58.6,123),
   Year2 = c(15.8,63.1,132),
   Year3 = c(22.7,81.0,161),
   Year4 = c(19,64.2,133)
 )

nopols<-data.frame(
  Year1 = c(163,4435,16184),
  Year2 = c(189,4761,17443),
  Year3 = c(252,5576,20102),
  Year4 = c(199,4581,18000)
)

n <- ncol(claims)

N <-nrow(claims)
```

```
X <- claims/nopols

Xibar <-rowSums(claims) / rowSums(nopols)

Pibar <- rowSums(nopols) #..................... A

Pbar <-sum(Pibar)

Pstar <-sum(Pibar * (1-Pibar/Pbar))/(N*n-1) #..................... B

m <-sum(claims) / Pbar #........................ C

s <- mean(rowSums(nopols *(X-Xibar)^2)/(n-1))  #.................. D

v <- (sum(rowSums(nopols*(X-m)^2))/(n*N-1)-s)/Pstar #..................... E
```

i) Briefly explain the meaning of the quantities calculated in code lines A to E in simple terms. (5)

ii) Use the code provided to calculate the estimates of E[m(θ)], E[s^2(θ)] and Var[m(θ)] under EBCT Model 2.

Note: *You can directly copy these code lines from* ***Codes.docx*** *file in your R Script.* (3)

iii) Calculate the credibility factors Z_i. (2)

iv) If the number of claims for the next year are 5000, 4800 and 4200 respectively for insurers A, B and C, then estimate the risk premiums for the next year using the credibility factors determined in part (iii). (3)

v) Based on your calculations in part (iii), comment on whether more emphasis is given to direct data or collateral data when calculating risk premiums. (2)

[15]

Q. 3) A National Sports Meet was organised and 27 States participated in the meet. There are three categories of sports:

1. Running : 100m, 400m and 110m hurdle
2. Jumping : High, Long and Pole Vault
3. Throwing: Shot put, Javeline and Discuss throw

Data, named `Sports.csv`, was collected of the event and shared with National Academy to analyse and prepare for International events.

i)

(a) Fit a multiple regression model between winning points (as the response variable) and the scores in the nine sports (as explanatory variables). Display the summary of the fitted model and identify which explanatory variables are significant at 5% level of significance. (6)

(b) Prepare a plot of the residuals of the multiple regression model. (2)

ii) Fit a Generalised Linear Model (GLM) to the data using winning points as the response variable and scores obtained in the nine sports as the explanatory variables, assuming a Poisson distribution for the response variable. Your answer should include the estimated coefficients and the Akaike's Information Criteria (AIC) of the fitted model. (4)

iii) Explain why scaled deviance cannot be used to compare the fit of the models in parts (i) (2) and (ii).

iv) Fit, by choosing a suitable argument for `family` in the `glm` command, a GLM to the data that would be equivalent to the model fitted in part (i). Your answer should include the estimated coefficients and the AIC of the fitted model. (4)

v) Compare the fit of the models fitted in parts (i) and (ii) using AIC for comparison. (2)

vi) Perform Principal Components Analysis (PCA) separately for the above three categories of sports viz. Running Sports, Jumping Sports and Throwing Sports and share how much variation is captured by first principal component (PC_1) for each category.

[**Hint:** *Use* `prcomp` *and do scaling by using* `scale.=TRUE` *parameter.*] (7)

vii) Under which sports category, principal components will be most useful in reducing dimensionality of the dataset while capturing 90% of variance? (3)

viii) The organisers informed that pole vault sports winning points are not properly captured. Test whether there is any correlation between pole vault points and overall winning points by calculating a 95% confidence interval for the Slope Coefficient Beta. Clearly state the null and alternative hypotheses and the conclusion of the test. (5)

ix) Players for racing sports have now been shortlisted for the international event and trials have started for the upcoming international event. However due to unforeseen circumstances, trials for 110 metres hurdle race got cancelled for last 5 runners. An analyst suggests that their scores can be predicted using their 100m trial scores.

Below are the 100 metres scores of last 5 runners.

1. 10.68
2. 10.42
3. 11.68
4. 11.62
5. 10.54

Fit a linear regression model to predict the scores in 110 metres hurdle race using scores in 100 metres race as the explanatory variable using the national sports meet data. Clearly state the equation of this linear regression model. Using this equation, predict the expected 110m hurdle race scores of the last five runners. (5)

[40]

Q. 4) A researcher has collected the following data on a group of students, regarding whether they passed or failed an exam and whether or not they attended tutorials:

Number of students	Exam passed	Exam failed
Attended tutorials	132	27
Did not attend tutorials	120	51

The data can be entered into R in matrix form using the following code:

```
exam.success = matrix(c(132,120,27,51),ncol=2,nrow=2)
```

Note: *You can directly copy the above code from* ***Codes.docx*** *file in your R Script.*

The researcher wants to establish whether tutorial attendance is independent of exam success, using a chi-square test.

Load the data in R and check for errors by displaying it.

i) State the hypothesis of this test. (1)

ii) Calculate the expected frequencies for the data under the null hypotheses in part (i) using an appropriate function in R. (2)

iii) Perform the test with continuity correction

Clearly state your conclusions at 1% level of significance. (3)

iv) What would be the conclusion of the test at 1% level of significance if Fisher's test is used instead of chi-square test? The null and alternative hypotheses need NOT be re-stated.

[**Hint:** *Use* `fisher.test` *function in R.*] (3)

v) In the context of the two tests performed in parts (iii) and (iv) above –

(a) Which one is an exact test and which one is an approximation?

(b) Which test is suitable for only 2×2 datasets and which test is suitable for any N×N dataset? (2)

vi) Using binomial test, test whether the proportion of students passing the examination is equal to 60%. You are required to clearly state the null and alternative hypotheses, p-value of the test and your conclusion at 5% level of significance. (4)

[15]

****