INSTITUTE OF ACTUARIES OF INDIA

Subject CS1B– Actuarial Statistics

(Paper B)

May 2024 Examination

INDICATIVE SOLUTION

Introduction

The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

Solution 1:

(i)

a) `> pbeta(0.8,5,1) - pbeta(0.2,5,1)` (1)

`[1] 0.32736` (1)

b) `> qbeta(0.65,5,1,lower= FALSE) or Alternate: qbeta(0.35,5,1)`

(1)

`[1] 0.8106131` (1)

(ii)

```
> a=c(5,1,3)
> b=c(1,5,3)
> skew=2*((b-a)/(a+b+2))*sqrt((a+b+1)/(a*b))          (2)
> skew
[1] -1.183216  1.183216  0.000000                     (1)
```

(3)

Or Alternatively

```
> skew1=2*((1-5)/(5+1+2))*sqrt((5+1+1)/5)
> skew1
[1] -1.183216

> skew2=2*((5-1)/(1+5+2))*sqrt((5+1+1)/5)
> skew2
[1] 1.183216

> skew3=2*((3-3)/(3+3+2))*sqrt((3+3+1)/9)
> skew3
[1] 0
```

(3)

(iii)

```
> set.seed(421967)              (0.5)
> x1=rbeta(12000,5,1)           (1)
> hist(x1)                      (0.5)
> x2=rbeta(12000,1,5)
> hist(x2)                      (0.5)
> x3=rbeta(12000,3,3)
> hist(x3)                      (0.5)
```

(3)

(iv)

As Alpha is greater than 1 and Beta is equal to 1, the histogram is heavily negatively skewed and strictly increasing as evident from the result obtained in (ii) above and from the shape of the graph.

(1)

As Alpha is equal to 1 and Beta is greater than 1, the histogram is heavily positively skewed and strictly decreasing as evident from the result obtained in (ii) above and from the shape of the graph.

(1)

As both the parameters alpha and beta are equal, the graph is roughly symmetrical as evident from the graph and the value of the skewness obtained in (ii) above.

(1)

(3)

(v)

```
> set.seed(421967)
> x1_bar <-replicate (1200, mean(rbeta (12000,5,1)))
```

Or alternatively

```
> set.seed(421967)
> x1_bar=rep(0,1200)
> for(i in 1:1200){x1<-rbeta(12000,5,1);x1_bar[i]<-mean(x1)}
```

(3)

```
> set.seed(421967)
> x2_bar <-replicate (1200, mean (rbeta (12000,1,5)))
```

Or alternatively

```
> set.seed(421967)
> x2_bar=rep(0,1200)
> for(i in 1:1200){x2<-rbeta(12000,1,5);x2_bar[i]<-mean(x2)}
```

(1)

```
> set.seed(421967)
> x3_bar <-replicate (1200, mean(rbeta (12000,3,3)))
```

Or alternatively

```
> set.seed(421967)
> x3_bar=rep(0,1200)
> for(i in 1:1200){x3<-rbeta(12000,1,5);x3_bar[i]<-mean(x3)}
```

(1)

(5)

(vi) The distribution of sample mean is roughly symmetrical in all the three cases irrespective of the values of the shape parameters (alpha and beta).These shape parameters(alpha and beta) do not significantly affect the sample mean of large sample size, which is in line with the central limit theorem. Irrespective of the population distribution of the random variable from which the sample is selected, for a large sample size the distribution of the sample means is approximately normal.

(2)

[20]

<u>Solution 2:</u>

(i)

The given equation is

Maximum Systolic Blood Pressure = 100 + Age (in years)

It can be written as

y = 100 + x

y = 100 + 1*(x)

y = α + βx

$\alpha = 100 \quad and \ \beta = 1$

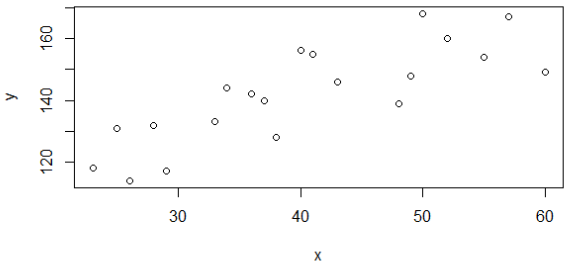
(2)

(ii)

```
> x=c(28, 37, 41, 52, 57, 49, 38, 25, 23, 48, 60, 55, 29, 43, 36, 50, 34, 40, 26, 33)
> y=c(132, 140, 155, 160, 167, 148, 128, 131, 118, 139, 149, 154, 117, 146, 142, 168, 144, 156, 114, 133)

> plot(x,y)
```

(1)



(1)

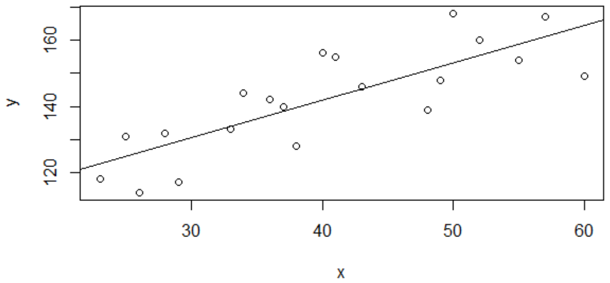
The age in years(x) and the systolic blood pressure(y) are positively correlated.

(1)
(3)

(iii)

```
> lm.result=lm(y~x)
```
(1)

```
> abline(lm(y~x))
```
(1)



(2)
(4)

(iv)

```
>anova(lm.result)
```

(1)

```
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 3082.9 3082.94  33.591 1.717e-05 ***
Residuals 18 1652.0   91.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(1)

From the above, it is clear that the slope parameter is significant.

(1)
(3)

(v)

```
>summary(lm.result)
```
(1)

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.485  -6.504   1.177   5.979  14.846 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  96.4994     8.1460  11.846 6.21e-10 ***
x             1.1331     0.1955   5.796 1.72e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.58 on 18 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6317 
F-statistic: 33.59 on 1 and 18 DF,  p-value: 1.717e-05
```
(1)

The value of the estimates of the co-efficient are

Alpha(α) = 96.4994   Beta(β) = 1.1331 (1)

(3)

(vi) The values of alpha and beta are expected to be 100 and 1 respectively as per (i). But empirical test results fetch the values as 96.4994 and 1.1331 respectively, which is close to the expected values of 100 and 1 respectively. Hence when empirically tested, we find that the claim made by the research about the maximum systolic blood pressure is valid.

(2)

(vii)

```
> cor(x,y,method="pearson")
[1] 0.8069094
```
(1)

```
> cor(x,y,method="spearman")
[1] 0.8180451
```
(1)

```
> cor(x,y,method="kendall")
[1] 0.6105263
```
(1)
(3)

(viii)

Pearson’s correlation co-efficient measures the strength of the linear relationship between the two variables, whereas Spearman Correlation method measures the strength of monotonic but not necessarily linearity between two variables.

Since Spearman considers the rank than the actual values, the value of the coefficient is less affected by extreme values/outliers in the data than Pearson’s Correlation Coefficient. Hence it is more robust.

Kendall’s correlation coefficient is considered to have better statistical properties when the data set is small and have more tied ranks, though it considers the relative values between the data set and not actual values.

Generally, the value of Kendall's coefficient is lower than the Spearman's rank coefficient.

Based on the sample correlation coefficients calculated in part (vii), we conclude Spearman Rank Coefficient > Pearson's Coefficient > Kendall's Coefficient

(4)

(ix)

H_0 : ρ = 1

H_1 : ρ ≠ 1 (1)

```
> cor.test(x,y,method="pearson")
```
(1)

```
        Pearson's product-moment correlation

data:  x and y
t = 5.7958, df = 18, p-value = 1.717e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5667661 0.9206793
sample estimates:
      cor
0.8069094
```
(1)

p-value is 1.717e-05 (1)

Since 95% confidence interval (0.5667661, 0.9206793) does not include the value 1, there is sufficient evidence to reject the hypothesis that there is perfect correlation between Age and Systolic Blood Pressure though there is strong positive correlation between the age and systolic Blood Pressure.

(2)

(6)

[30]

Solution 3:

(i)
```
> Firepolicies<-read_csv("Firepolicies.csv")
```
(1)

(ii)
a)

```
>  Maha<-Firepolicies[Firepolicies$Location=="M",]
> Maha_0_Claims<-Maha[Maha$Claimed == 0,]
> ProporMaha_0_Claims<-nrow(Maha_0_Claims)/nrow(Maha)
> ProporMaha_0_Claims
[1] 0.7346939
```
(3)

b)

```
>  Gujarat<-Firepolicies[Firepolicies$Location=="G",]
>  Gujarat_0_Claims<-Gujarat[Gujarat$Claimed == 0,]
>  ProporGuj_0_claims <-nrow(Gujarat_0_Claims)/nrow(Gujarat)
> ProporGuj_0_claims
[1] 0.5909091
```
(1)

(iii)

H_0(Null Hypothesis):

Proportion of No claims in the past one year in Maharashtra is equal to Proportion of NO Claims in the past one year in Gujarat

H_1(Alternative Hypothesis):
Proportion of No claims in the past one year in Maharashtra is NOT equal to Proportion of NO Claims in the past one year in Gujarat

(1)

```
> prop.test (c(nrow( Maha_0_Claims),nrow(Gujarat_0_Claims)),c(nrow(Maha),nrow(Gujarat)),correct = FALSE)
```

(1)

```
2-sample test for equality of proportions without continuity correction

data:  c(nrow(Maha_0_Claims), nrow(Gujarat_0_Claims)) out of c(nrow(Maha), nrow(Gujarat))
X-squared = 2.1568, df = 1, p-value = 0.1419
alternative hypothesis: two. sided
95 percent confidence interval:
 -0.04696637  0.33453595
sample estimates:
   prop 1    prop 2
0.7346939 0.5909091
```

(1)

p-value is 0.1419 (1)

At 95% confidence interval (-0.04696637, 0.33453595), which contains "0", we have insufficient evidence to reject null hypothesis and can conclude that there is no significant difference between Maharashtra and Gujarat in respect of the proportion of No claims in the previous year. (1)

(5)

(iv)

H_0(Null Hypothesis):
Population mean of Textile Mills Claims is equal to Population mean of Transporters' Godowns Claims

H_1(Alternative Hypothesis):
Population mean of Textile Mills Claims is NOT equal to Population mean of Transporters' Godowns Claims

(1)

```
> Textile<-Firepolicies[Firepolicies$Occupancy=="TM",]
> Transporter<-Firepolicies[Firepolicies$Occupancy=="TG",]
```

(1)

```
>  t.test(Textile$Claim.Size,Transporter$Claim.Size,var.equal=TRUE)
```

(1)

```
        Two Sample t-test

data:  Textile$Claim.Size and Transporter$Claim.Size
t = 7.877, df = 103, p-value = 3.586e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 24.27889 40.61871
sample estimates:
mean of x mean of y
 98.07843  65.62963
```

(1)

p-value is 3.586e-12 (1)

At 95% Confidence interval as the confidence interval (24.27889 40.61871) does not contain the value 0, we have sufficient evidence to reject Null Hypothesis and conclude that there is significant difference between the average claim size of Textile Mills and Transporters' Godowns.

(1)

(6)

(v)

(a)

```
> model1=glm(Firepolicies$Claimed~Firepolicies$Claim.Size+Firepolicies$Location,family = binomial())
```

(2)

```
> summary(model1)
```

(1)

```
Call:
glm(formula = Firepolicies$Claimed ~ Firepolicies$Claim.Size +
    Firepolicies$Location, family = binomial())

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.840330   0.453954  -1.851   0.0642 .
Firepolicies$Claim.Size  0.003711   0.004935   0.752   0.4521
Firepolicies$LocationG   0.198437   0.494674   0.401   0.6883
Firepolicies$LocationK   0.205169   0.506205   0.405   0.6853
Firepolicies$LocationM  -0.418141   0.498094  -0.839   0.4012
Firepolicies$LocationT   0.448865   0.519037   0.865   0.3871
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.91   on 191  degrees of freedom
Residual deviance: 247.65  on 186  degrees of freedom
AIC: 259.65

Number of Fisher Scoring iterations: 4
```

(1)

Or Alternatively

```
> model1=glm(Firepolicies$Claimed~Firepolicies$Claim.Size+Firepolicies$Location,family = binomial(link=logit))
> model1

Call:  glm(formula = Firepolicies$Claimed ~ Firepolicies$Claim.Size +
    Firepolicies$Location, family = binomial (link = logit))

Coefficients:
            (Intercept)  Firepolicies$Claim.Size   Firepolicies$LocationG
              -0.840330                 0.003711                 0.198437
 Firepolicies$LocationK   Firepolicies$LocationM   Firepolicies$LocationT
               0.205169                -0.418141                 0.448865

Degrees of Freedom: 191 Total (i.e. Null);  186 Residual
```

```
Null Deviance:      251.9
Residual Deviance: 247.6       AIC: 259.6
```

(4)

b)

```
> model2=glm(Firepolicies$Claimed~Firepolicies$Claim.Size+Firepolicies$Occupancy,family = binomial())
> model2
```

(1)

```
Call:  glm(formula = Firepolicies$Claimed ~ Firepolicies$Claim.Size + 
    Firepolicies$Occupancy, family = binomial())

Coefficients:
             (Intercept)   Firepolicies$Claim.Size  Firepolicies$OccupancyDW  
               -0.492623                 -0.009519                 -0.266777  
Firepolicies$OccupancyHG  Firepolicies$OccupancyTG  FirepoliciesOccupancyTG F
irepolicies$OccupancyTM  
                0.100853                  0.817776                  1.063207  

Degrees of Freedom: 191 Total (i.e. Null);  186 Residual
Null Deviance:      251.9
Residual Deviance: 247.3       AIC: 259.28
```

(1)

Or Alternatively

```
>  model2=glm(Firepolicies$Claimed~Firepolicies$Claim.Size+Firepolicies$Occupancy,family = binomial(logit))
> model2

Call:  glm(formula = Firepolicies$Claimed ~ Firepolicies$Claim.Size + 
    Firepolicies$Occupancy, family = binomial(logit))

Coefficients:
             (Intercept)   Firepolicies$Claim.Size  Firepolicies$OccupancyDW  
               -0.492623                 -0.009519                 -0.266777  
Firepolicies$OccupancyHG  Firepolicies$OccupancyTG  Firepolicies$OccupancyTM  
                0.100853                  0.817776                  1.063207  

Degrees of Freedom: 191 Total (i.e. Null);  186 Residual
Null Deviance:      251.9
Residual Deviance: 247.3       AIC: 259.3
```

(2)

(vi)

AIC for Model 1 = 259.65 and AIC for Model 2 = 259.28 (1)

The AIC is smaller for the model2 as compared to model1 (1)

So Claim Size and Occupancy model2 seems to be a better predictor than Claim Size and Location, and we would choose the model2. (1)

Alternatively,

Since both the models have a very minor difference in AICs, one can conclude that both models 1 and 2 are equally good.

(3)

(vii)

Claimed is a numerical variable

Claim.Size is a numerical variable

Location is a factor variable

Occupancy is a factor variable

(1)

Numerical variables are continuous variables which can take numerical values. Claim size and Claimed (0 for no claims in the last year or 1 for claim in t he last year) are examples in the context of this GLM.

(2)

Factor / categorical variables are variables which only take categories. Location is a factor variable which takes values of 5 states – M, G, T, K and A. Occupancy is also a factor variable which takes 5 values viz.TM, TG, DW, CS and HG.

(2)

(5)

[30]

Solution 4:

(i)

a)
```
> m <- mean(rowMeans(Claims))
> m
[1] 278542.3
```
(1) (1)

(2)

b)
```
> s<-mean(apply(Claims,1,var))
> s
[1] 846425572
```
(1) (1)

(2)

c)
```
> n <-ncol(Claims)
> n
[1] 5
```
(1)

```
> v<-var(rowMeans(Claims))-mean(apply(Claims,1,var))/n
> v
[1] 2842778626
```
(2) (1)

(4)

(ii)

```
> Z <- n/(n+s/v)
> Z
[1] 0.9437976
```
(1) (1)

```
> PurePremium <-Z*rowMeans(Claims)+(1-Z)*m
> PurePremium
[1] 259773.5 258770.4 249940.8 383415.5 268150.2 251203.4
```
(1) (1)

Credibility premium for Delhi is 38341.50 and for Kerala is 251203.40 (1)

(5)

(iii)

Z is an increasing function of n. In the formula for credibility factor Z = n / (n + s/v), with an increasing value of n, Z will tend to increase. Intuitive ly also, it is true because as the number of observations for the particular r isk under consideration are more, more reliable is the specific data from that particular risk and hence credibility factor Z would be higher indicative of m ore weightage to the specific data(mean for the specific risk) rather than the collateral data (overall mean for all risks)

(2)

(iv)

Based on the graph, the approximate maximum likelihood estimate i.e. the value at which the log likelihood is maximum is around 1.8 to 1.9.

(2)

(v)

Exact Maximum Likelihood Estimate λ is

$$\lambda = \Sigma x_i / n$$

$$= 280/150$$

$$= 1.87$$

(2)

So, the actual maximum likelihood estimate calculated using first principles is close to the approximate maximum likelihood determined based on the graph.

(1)

(3)

[20]

****