Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper B)

November 2020 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)

```
> claims <- read.csv("MotorClaims.csv")
> mean = mean(claims$Claims)
> mean
[1] 18672.76                                              [1]
> stddev = sd(claims$Claims)
> variance = stddev ^ 2
> variance
[1] 161323921                                             [1]
> lambda <- mean/variance
> lambda
[1] 0.000115747                                           [2]
> alpha <- mean * lambda
> alpha
[1] 2.161316                                              [2]
```

X ~ Gamma (2.16, 0.0001) [2]

[8]

ii)

```
> set.seed(100)
> samples <- rgamma(1000,alpha,lambda)                     [2]
> head(samples,6)                                          [1]
[1]  9305.461  2125.292 25926.442 15685.099 18120.436  8605.442    [2]
```

[5]

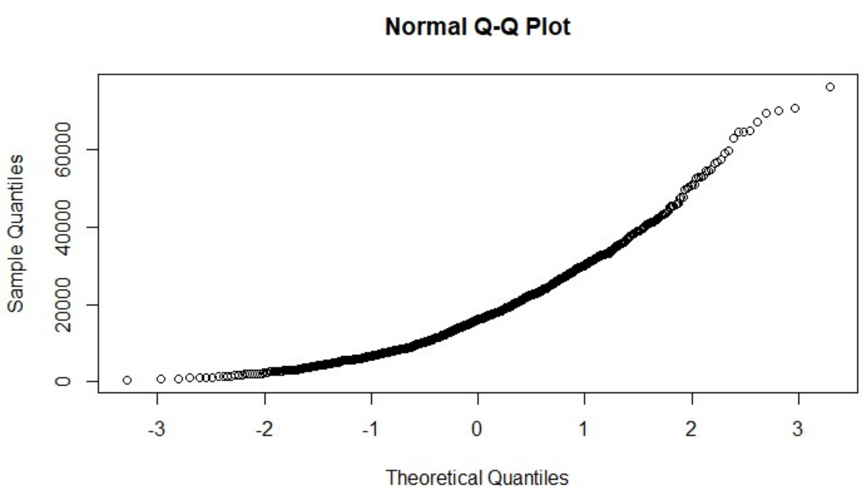
iii)

```
> mean(samples)
[1] 18423.47
> variance <- sd(samples) ^ 2
> variance
[1] 153958637
```

[2]

iv)

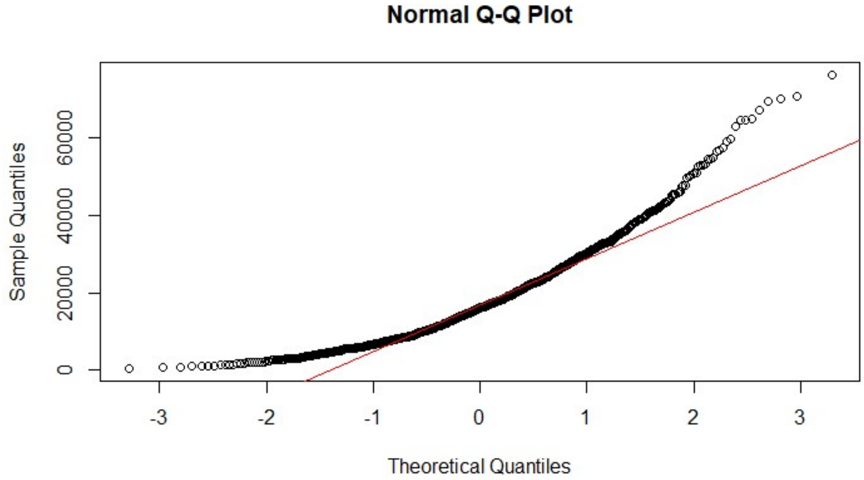
```
> qqnorm(samples)
```



[4]

v)

```
> qqline(samples,col="red")
```



[2]

vi)

Close to normal…*(1 mark)* in the middle values…*(1 mark)*.
'Banana-shaped' indicates positively skewed…*(1 mark)*. **[3]**

[24 Marks]

Solution 2:

i)

```
data("mtcars")
> str(mtcars)
'data.frame':	32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

There are 32 observations (car models) and 11 variables (car features) in the dataset. **[4]**

ii)

```
summary(mtcars)
      mpg             cyl             disp             hp             drat      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930  
       wt             qsec             vs               am              gear      
 Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000  
 1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000  
 Median :3.325   Median :17.71   Median :0.0000   Median :0.0000   Median :4.000  
 Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688  
 3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000  
```

```
 Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000
      carb
 Min.   :1.000
 1st Qu.:2.000
 Median :2.000
 Mean   :2.812
 3rd Qu.:4.000
 Max.   :8.000
```

The two variables 'vs' and 'am' are categorical variables. (This can be identified using str or summary function)

```
mtcars1 <- mtcars[,c(1:7,10,11)]
```

[5]

iii)

```
> str(mtcars1)
'data.frame':	32 obs. of  9 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

There are 32 observations (car models) and 9 variables (car features) in the dataset. **[2]**

iv)

```
mtcars1.pca <- prcomp(mtcars1,center = TRUE,scale=TRUE)
```
[2]

```
> summary(mtcars1.pca)
```
[1]

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.3782	1.4429	0.71008	0.51481	0.42797	0.35184	0.32413	0.2419	0.14896
Proportion of Variance	0.6284	0.2313	0.05602	0.02945	0.02035	0.01375	0.01167	0.0065	0.00247
Cumulative Proportion	0.6284	0.8598	0.91581	0.94525	0.96560	0.97936	0.99103	0.9975	1.00000

[2]

[5]

v)

The R analysis shows that the proportion of variance explained by first three principal components is 91.5% and by first four variables is 94.5%.

Thus, it will be appropriate to retain the first three (or four) principal components. **[3]**

[19 Marks]

<u>Solution 3:</u>

i)

```
> BMI <- read.csv("BMIClaims.csv")
> n <- length(BMI$BMI)
> alpha <- 0.05                                                        …   [2]
> sqrt(c((n-1)*var(BMI$BMI)/qchisq(1-alpha/2,df=n-1),(n-1)*var(BMI$BMI)/qchisq(alpha/2
,df=n-1)))                                                                 [2]
[1] 5.920028 7.434763                                                      [2]
```

[6]

ii)

```
> sigma <- 4
> statistic <- (n-1)*var(BMI$BMI)/sigma^2                                    [1]
> statistic
[1] 404.5421
> qchisq(alpha/2,n-1)
[1] 117.098
> qchisq(alpha/2,n-1,lower=FALSE)
[1] 184.687
> 2*(pchisq((n-1)*var(BMI$BMI)/sigma^2,df=n-1,lower.tail=FALSE))           [2]
[1] 3.564503e-25                                                             [1]
```

Since p-value is less than 5%, there is sufficient evidence to reject the hypothesis, i.e. the standard deviation of BMI is not equal to 4. [2]

[6]

iii)

```
> x <- nrow(BMI[BMI$BMI>30,])                                                [1]
> binom.test(x,n,conf.level = 0.99)                                          [2]

        Exact binomial test

data:  x and n
number of successes = 10, number of trials = 150, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.02522882 0.13728337                                                       [1]
sample estimates:
probability of success
            0.06666667
```

Since 99% CI for p doesn’t contain p=0.2 [1]

it is unlikely that the proportion of obese policyholders is more than 20%.... [1]

[6]

iv)

```
> table(BMI$BMI>30,BMI$ClaimCount)

         0   1
  FALSE 133   7
  TRUE    7   3

> y <- c(3,7)
> m <- c(10,140)                                                             [2]
> poisson.test(y,m)                                                          [1]

        Comparison of Poisson rates

data:  y time base: m
count1 = 3, expected count1 = 0.66667, p-value = 0.02493
alternative hypothesis: true rate ratio is not equal to 1
95 percent confidence interval:
  1.001171 26.282304
sample estimates:
rate ratio
         6
```

Since p-value is less than 5% i.e. 2.5%, there is sufficient evidence to reject the hy pothesis, i.e. Claim frequency is different between obese and others.

[2]

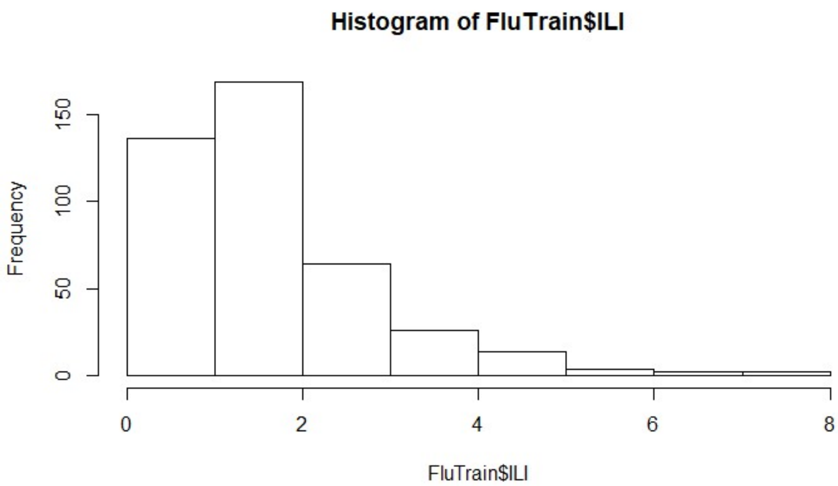
(Alternatively, can use prop.test) **[6]**

[24 Marks]

Solution 4:

i)

```
setwd("C:/Users/shrey/Downloads")
FluTrain <- read.csv("FluTrain.csv")
> str(FluTrain)
'data.frame':	417 obs. of  3 variables:
 $ Week   : Factor w/ 417 levels "2004-01-04 - 2004-01-10",..: 1 2 3 4 5 6 7 8 9 10 ..
.
 $ ILI    : num  2.42 1.81 1.71 1.54 1.44 ...
 $ Queries: num  0.238 0.22 0.226 0.238 0.224 ...
hist(FluTrain$ILI)
```

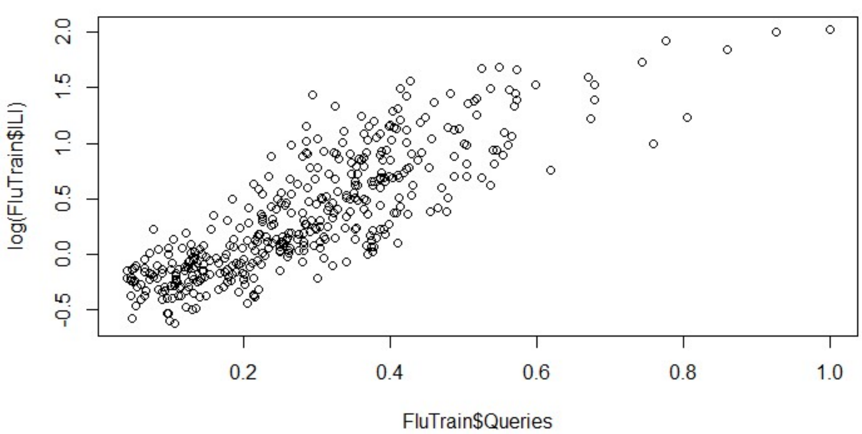


The data is positively skewed. Most of the ILI values are small, with a relatively small number of much larger values.

[3]

ii)

```
plot(FluTrain$Queries,log(FluTrain$ILI))
```



There is a positive linear relationship between log(ILI) and Queries.

i.e. more the number of the Google search queries, higher the number of ILI-related physician visits.

[4]

iii)

```
FluTrend1 = lm(log(ILI) ~ Queries, data = FluTrain)
> summary (FluTrend1)
```
[3]
[1]

```
Call:
lm(formula = log(ILI) ~ Queries, data = FluTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76003 -0.19696 -0.01657  0.18685  1.06450 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.49934    0.03041  -16.42   <2e-16 ***
Queries      2.96129    0.09312   31.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2995 on 415 degrees of freedom
Multiple R-squared:  0.709,    Adjusted R-squared:  0.7083 
F-statistic:  1011 on 1 and 415 DF,  p-value: < 2.2e-16
```
[2]

[6]

iv)

ln y = -0.49934 +2.96129x [2]

where x is the google search queries and y is the percentage of ILI related physician visits. [1]

[3]

v)

From the R output, R-squared value is 0.709. [1]

```
correlation <- cor(FluTrain$Queries,log(FluTrain$ILI))
> correlation
[1] 0.8420333
> correlation ^ 2
[1] 0.7090201
```
[1]
[1]

Hence, R-squared = Correlation ^ 2 [2]

[5]

vi)

```
which.max(FluTrain$ILI)
[1] 303
> FluTrain$Week[303]
[1] 2009-10-18 - 2009-10-24
417 Levels: 2004-01-04 - 2004-01-10 2004-01-11 - 2004-01-17 ... 2011-12-25 - 2011-12-3
1
```

Week of 18th October 2009 to 24th October 2009 corresponds to the highest percentage of ILI-related physician visits.

[4]

vii)

```
PredTest1 = exp(predict(FluTrend1,newdata = FluTrain))
> PredTest1[303]
     303 
11.72765
```
[2]
[2]

[4]

viii)

```
FluTrain$ILI[303]
[1] 7.618892
```
[2]

```
(7.618892-11.72765)/7.618892
[1] -0.5392855
```
[2]

[4]

[33 Marks]

****