Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper B)

March 2021 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)

```
> data = c(1990,2400,2150,2090,2300,2100,2180,2150,2030,2100,2180,2010,2060,2
160,2120)…
```
[1]

ii)

```
> quantile(data)
  0%  25%  50%  75% 100%
1990 2075 2120 2170 2400
> IQR(data)
[1] 95
```
Q1 – INR 2,075; Q2 – INR 2,120; Q3 – INR 2,170 and IQR – INR 95

[1 MARK EACH FOR QUARTILE AND IQR]

[4]

iii)

```
> mean(data)
[1] 2134.667
> var(data)
[1] 11469.52
```
[1 MARK EACH FOR MEAN & VARIANCE]

[2]

iv)

Ho: The mean claim amount is INR 2000
H1: Mean claim amount is not equal to INR 2000

```
> t.test(data,mean=2000)

        One Sample t-test

data:  data
t = 77.197, df = 14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2075.359 2193.974
sample estimates:
mean of x
 2134.667
```

Since the p-value is 2.2*10^-16 is less than 5%, there is sufficient evidence to reject Ho of mean equal to INR 2000.

[1 MARK FOR HYPOTHESIS, 1 MARK FOR T-TEST, 1 MARK FOR RESULTS, 1 MARK FOR MENTIONING P-VALUE & 1 MARK FOR CONCLUSION]

[5]

v)

```
> pnorm(2300,mean,sqrt(var),lower.tail = FALSE)
[1] 0.06131982
```
[3 MARKS FOR CODE, 1 MARK FOR RESULT]

[4]

vi)

```
> data1 = data[data < max(data)]
> data1
 [1] 1990 2150 2090 2300 2100 2180 2150 2030 2100 2180 2010 2060 2160 2120
> data2 = data1[data1 < max(data1)]
> data2
 [1] 1990 2150 2090 2100 2180 2150 2030 2100 2180 2010 2060 2160 2120
> mean(data2)
[1] 2101.538
> median(data2)
[1] 2100
```

Impact of removing outliers from the data has led to the mean and median being almost equal. Earlier the mean was higher than median, which shows that the mean is more likely to be affected by outliers.

[2 MARKS FOR CALCULATING REVISED DATA, 1 MARK FOR MEAN, 1 MARK FOR MEDIAN, 2 MARKS FOR COMMENTS]

[6]

[22 Marks]

Solution 2:

i)

```
> priormean <- 120
> priorsd <- 10
> x <- seq(80,160,len = 100)
> x
  [1]  80.00000  80.80808  81.61616  82.42424  83.23232  84.04040  84.84848
85.65657  86.46465
 [10]  87.27273  88.08081  88.88889  89.69697  90.50505  91.31313  92.12121
92.92929  93.73737
 [19]  94.54545  95.35354  96.16162  96.96970  97.77778  98.58586  99.39394 1
00.20202 101.01010
 [28] 101.81818 102.62626 103.43434 104.24242 105.05051 105.85859 106.66667 1
07.47475 108.28283
 [37] 109.09091 109.89899 110.70707 111.51515 112.32323 113.13131 113.93939 1
14.74747 115.55556
 [46] 116.36364 117.17172 117.97980 118.78788 119.59596 120.40404 121.21212 1
22.02020 122.82828
 [55] 123.63636 124.44444 125.25253 126.06061 126.86869 127.67677 128.48485 1
29.29293 130.10101
 [64] 130.90909 131.71717 132.52525 133.33333 134.14141 134.94949 135.75758 1
36.56566 137.37374
 [73] 138.18182 138.98990 139.79798 140.60606 141.41414 142.22222 143.03030 1
43.83838 144.64646
 [82] 145.45455 146.26263 147.07071 147.87879 148.68687 149.49495 150.30303 1
51.11111 151.91919
 [91] 152.72727 153.53535 154.34343 155.15152 155.95960 156.76768 157.57576 1
58.38384 159.19192
[100] 160.00000
```

[2 MARKS FOR CODE, 1 MARK FOR RESULTS]

[3]

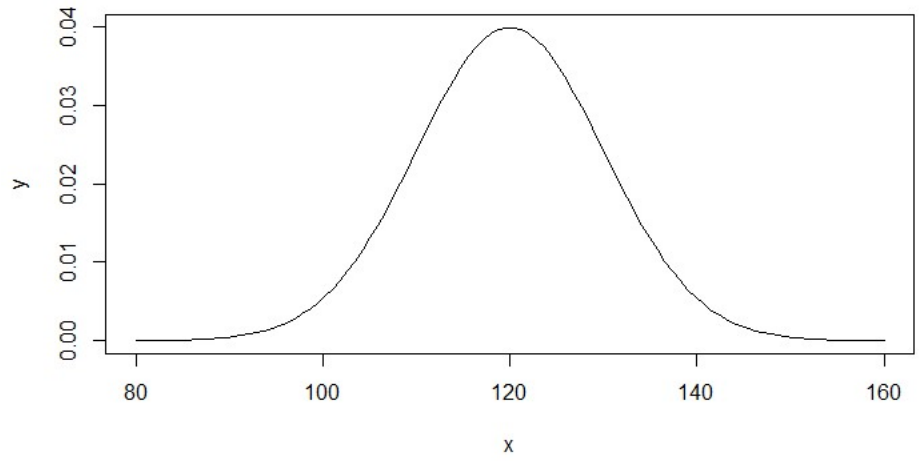
ii)

```
> y<-dnorm(x,mean = priormean,sd=priorsd)….
```

[4]

```
> plot(x,y,type = "l")….
```

[1]



[1]
[6]

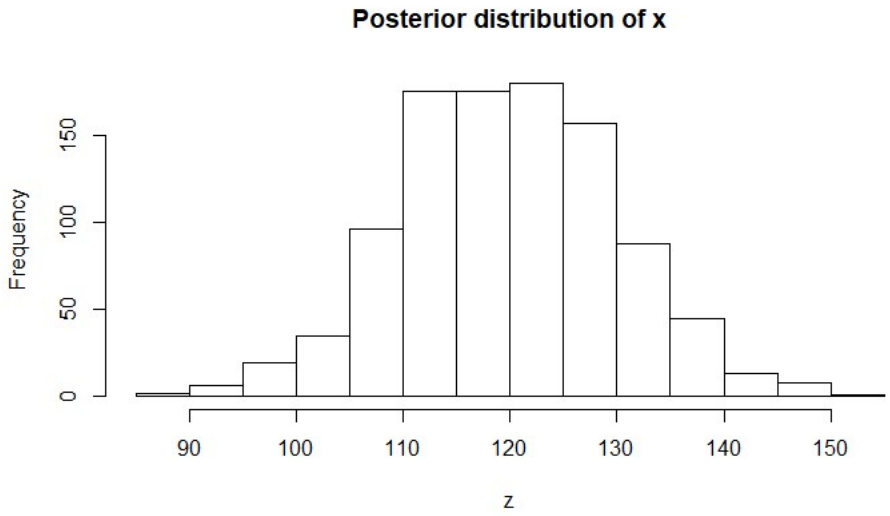
iii)

```
> z <- rnorm(1000,priormean,priorsd)…
```

[2]

iv)

```
> hist(z,main = "Posterior distribution of x")
```



[2 MARKS FOR CODE, 1 MARK FOR GRAPH]
[3]

v)

```
> mean(z)
[1] 119.8336
> sd(z)
[1] 10.06614
```

[1 MARK FOR MEAN, 1 MARK FOR SD]
[2]

vi)

```
> sbp <- mean(z) + qnorm(c(0.025,0.975)) * sd(z)
> sbp
[1] 100.1044 139.5629
```

[2 MARKS FOR CODE, 2 MARKS FOR CONFIDENCE INTERVAL]

[4]

[20 Marks]

Solution 3:

i)

```
> city1 = c(9150,9418,9218,9539,9179,8907,9472,8921)
> city2 = c(8919,9095,9046,9321,9719,9704,9107,9275)…
```

[1]

ii)

Ho: There is no difference in the average number of monthly COVID19 cases between two cities.

H1: There is a difference in the average number of monthly COVID19 cases between two cities.

```
> t.test(x=city1,y=city2,var.equal = TRUE,conf.level = 0.95)

        Two Sample t-test

data:  city1 and city2
t = -0.35359, df = 14, p-value = 0.7289
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -337.3886  241.8886
sample estimates:
mean of x mean of y
  9225.50   9273.25
```

Since the p-value is 0.7289 is significantly greater than 5%, there is insufficient evidence to reject Ho. Thus, we have no evidence to suggest that the means are different between the two samples.

[2 MARKS FOR HYPOTHESIS, 2 MARKS FOR T-TEST, 1 MARK FOR RESULTS, 1 MARK FOR MENTIONING P-VALUE, 2 MARKS FOR CONCLUSION]

[8]

iii)

```
> var.test(x=city1,y=city2,conf.level = 0.95)

        F test to compare two variances

data:  city1 and city2
F = 0.63907, num df = 7, denom df = 7, p-value = 0.5691
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1279433 3.1920724
sample estimates:
ratio of variances
         0.6390651
```

The p-value is 0.5691 > 5%, so there is insufficient evidence to reject the assumption of equal variance.

[2 MARKS FOR VAR TEST, 1 MARK FOR RESULTS, 1 MARK FOR P-VALUE, 1 MARK FOR CONCLUSION]

[5]

iv)

Confidence interval can be read from Part (b) or can be derived as below:

```
> t.test(x=city1,y=city2,var.equal = TRUE,conf.level = 0.95)$conf.int
[1] -337.3886  241.8886
attr(,"conf.level")
[1] 0.95
i.e. 95% CI is (-337.39, 241.89)
```

[1 MARKS FOR CODE, 2 MARKS FOR CONFIDENCE INTERVAL]

[3]

v)

The confidence interval (-337,241) contains 0, therefore the assumption of equal means holds.

[1 MARK FOR MENTIONING CONTAINS 0, 1 MARK FOR CONCLUSION]

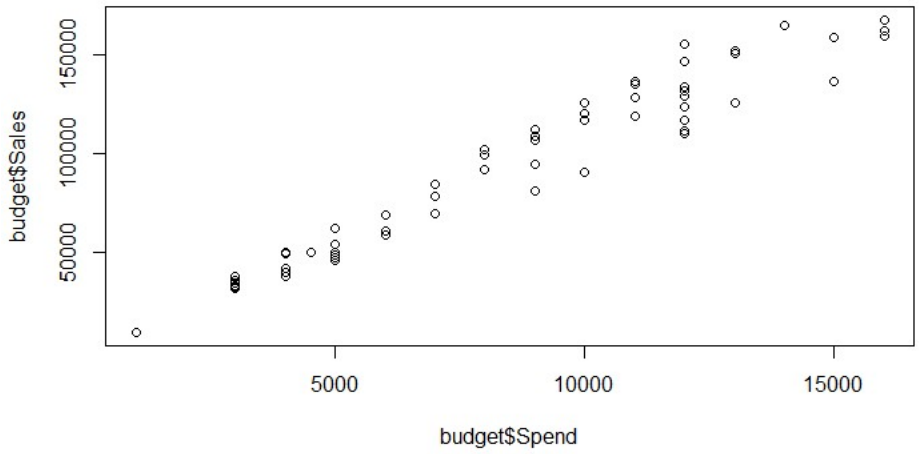
[2]

[19 Marks]

Solution 4:

i)

```
> budget = read.csv("marketingbudget.csv")
> plot(budget$Spend,budget$Sales)
```



The above scatter plot shows a positive linear relationship between marketing Spend and Sales data.

[1 MARK FOR CODE, 1 MARK FOR SCATTER PLOT, 1 MARK FOR MENTIONING POSITIVE, 1 MARK FOR LINEAR]

[4]

ii)

```
> cor = cor(budget$Sales,budget$Spend)

> cor
[1] 0.9701669
```

[1 MARK FOR CODE, 1 MARK FOR RESULT]

[2]

iii)

```
> cor.test(budget$Spend,budget$Sales,method="pearson",alternative = "greater"
)

        Pearson's product-moment correlation

data:  budget$Spend and budget$Sales
t = 30.476, df = 58, p-value < 2.2e-16
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.9542479 1.0000000
sample estimates:
      cor
0.9701669
```

The p-value is 2.2 X 10^-16, showing very strong evidence against the null hypothesis. Thus, we reject that the Pearson's correlation coefficient is equal to 0 and conclude that it is positive.

[2 MARKS FOR COR TEST, 1 MARK FOR RESULTS, 1 MARK FOR P-VALUE, 1 MARK FOR CONCLUSION]

[5]

iv)

```
> reg = lm(Sales ~ Spend, data = budget)
> summary(reg)

Call:
lm(formula = Sales ~ Spend, data = budget)

Residuals:
     Min       1Q   Median       3Q      Max
-25331.9  -6783.1   -844.5   7965.9  25320.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3431.5592  3245.9169   1.057    0.295
Spend         10.5310     0.3455  30.476   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10650 on 58 degrees of freedom
Multiple R-squared:  0.9412,   Adjusted R-squared:  0.9402
F-statistic: 928.8 on 1 and 58 DF,  p-value: < 2.2e-16
```

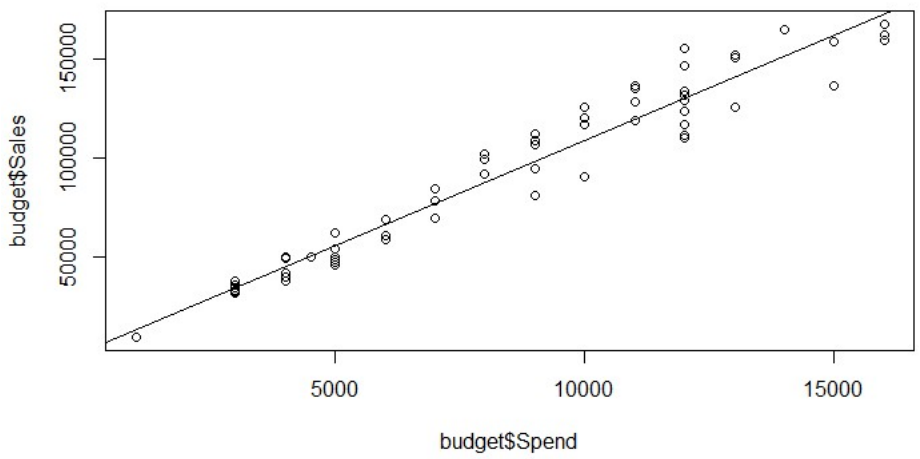
From the output, the estimate of parameter sigma is 10,650.

[3 MARKS FOR REG CODE, 2 MARKS FOR RESULTS, 1 MARK FOR ESTIMATE OF SIGMA]

[6]

v)

```
> abline(reg)
```



[1 MARK FOR CODE, 1 MARK FOR GRAPH]

[2]

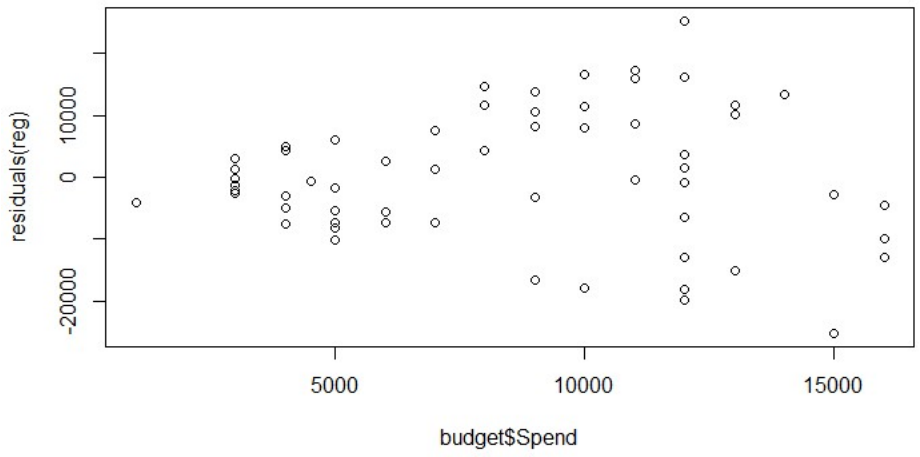
vi)

From the R output, the proportion of total variability of the responses explained by the model is 94.12%.

[1]

vii)

```
> plot(budget$Spend,residuals(reg))
```



[1 MARK FOR CODE, 1 MARK FOR GRAPH]

[2]

viii)

```
es = resid(reg)
> t.test(es,conf.level = 0.99)$conf.int
[1] -3630.146  3630.146
attr(,"conf.level")
[1] 0.99
```

From the above, the confidence interval for parameter sigma is (-3630.15, 3630.15)

[1 MARK FOR EVALUATING RESIDUALS, 1 MARK FOR T-TEST, 2 MARKS FOR CONFIDENCE INTERVAL]

[4]

ix)

Based on the results in both part (vii) and part (viii), the errors seem to be close to zero and the confidence interval of residuals also contains 0. Hence the model seems to be a good fit.

[2]

x)

```
Let Ho: Beta = 10 and H1: Beta not equal to 10

> b1 = (coef(reg))[['Spend']]…                                    [1]
> n = 60
> s = sqrt(sum(es^2)/(n-2))
> SE = s/sqrt(sum((budget$Spend-mean(budget$Spend))^2))..         [2]
> t = (b1-10)/SE …                                                [1]
> pt(t,58,lower.tail = FALSE)…                                    [1]
[1] 0.06489565
> pvalue = 2*pt(t,58,lower.tail = FALSE)..                        [1]
> pvalue
[1] 0.1297913..                                                   [1]
```

[7]

xi)

There is insufficient evidence to reject the null hypothesis at 5% level of significance. The slope is equal to 10 for this data.

[2]

xii)

```
> y = 3431.5592 + (b1*4500)
> y
[1] 50821.17
```

With a marketing spend of INR 4,500, the Sales would be INR 50,821.

[1 MARK FOR CODE, 1 MARK FOR RESULTS, DEDUCT ½ MARK FOR NOT MENTIONING INR]

[2]

[39 Marks]

****