Applied Statistics - Problem Set 1

2022-10-02

Problem set 1 – Cianna Devitt (17321885)

Question 1

Part One : Finding a 90% confidence interval for the average of the sample.

I first find the mean, stadard deviation and size of the sample ‘y’.

```
y <- c(105, 69, 86, 100, 82, 111, 104, 110, 87, 108, 87, 90, 94, 113, 112, 98
, 80, 97, 95, 111, 114, 89, 95, 126, 98)

sd <- sd(y)
sd

## [1] 13.09287

sample.mean <- mean(y)
sample.n <- length(y)
```

Using these objects, I then find the standard error, degrees of freedom, t-stat and margin of error.

```
se <- sd / sqrt(sample.n)
degrees.freedom = sample.n - 1
t.score = qt(p=0.05/2, df=degrees.freedom, lower.tail = F)
margin.error <- t.score * se
```

I then find the two confidence interval figures

```
CI_lower <- sample.mean - margin.error
CI_upper <- sample.mean + margin.error

print(c(CI_lower, CI_upper))

## [1]  93.03553 103.84447
```

With a 90% confidence interval we can say the mean for these IQ scores lies between 93.03553 and 103.84447.

Part Two: Conducting a hypothesis test to determine if the IQ average from this sample is greater than the national average (100)

- **Null Hypothesis: The average IQ score from this sample is not different from the national average (100).**

- **Alternative Hypothesis: The average IQ score from this sample is greater than the national average.**

To conduct this test I will run a one-sided (just investigating if sample average is greater than national figure) t-test (due to small sample size).

Firstly, finding the t-statistic.

```
t.stat <- (sample.mean - 100 / se)
```

I will then find p-value

```
p <- pt(q=t.stat, df=degrees.freedom,lower.tail = F)
```

Using t.test() function, I run the one sample t-test.

```
t.test(y, mu=100, alternative="greater", conf.level = .90)

##
##  One Sample t-test
##
## data:  y
## t = -0.59574, df = 24, p-value = 0.7215
## alternative hypothesis: true mean is greater than 100
## 90 percent confidence interval:
##  94.98915      Inf
## sample estimates:
## mean of x
##     98.44
```

The p value is not low enough to reject our null hypothesis (that the school's average IQ scores are not different from the national average).

Question Two

Part (1/3) :

Firstly, reading in the table from GitHub

```
expenditure <- read.table("~/GitHub/StatsI_Fall2022/datasets/expenditure.txt"
, header = TRUE)
```

Next, plotting X1, X2, X3 against Y in three graphs and naming them.

Graph One could be described as a linear relationship between Personal income per capita (x1) and expenditure on housing assistance (y)

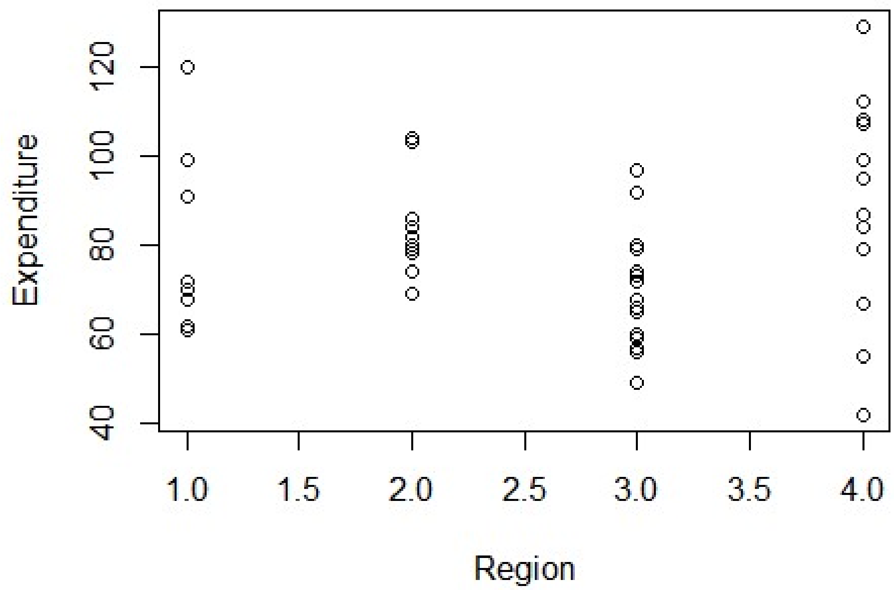
Graph Two could be described as a linear relationship between number of finacially insecure residents (x2) and expenditure on housing assistance (y)

Graph Three could be described as a linear relationship between Number of Urban residents (x3) and expenditure on housing assistance (y)

Part (2/3)

Plotting relationship of Y against Region

```
plot(expenditure$Region, expenditure$Y, xlab= "Region", ylab="Expenditure")
```



#

On average the west region (4) has the highest per capita expenditure on housing assistance.

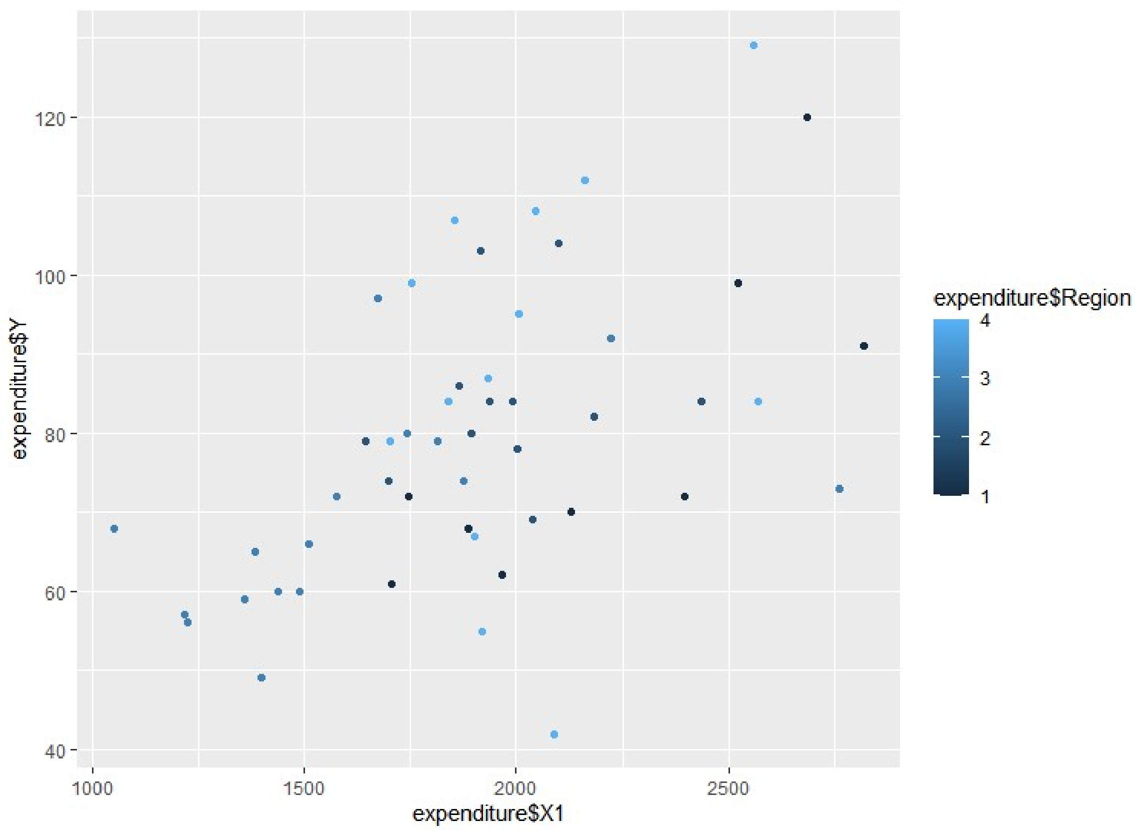
I Installed the "ggplot" package to add and alter a third (non-axis) variable to Graph One

```
install.packages("ggplot2")
```

```
library(ggplot2)
```

I then used the ggplot function (with X1 on the x- axis, Y on the y-axis) to plot region to Graph One. With a little more time I would change the colours (to differentiate regions better) and alter symbols.

```
ggplot(data= expenditure, mapping= aes(x=expenditure$X1, y= expenditure$Y)) +
geom_point(aes(color= expenditure$Region)))
```


expenditure$Region
4
3
2
1
120
100
80
60
40
expenditure$Y
1000
1500
2000
2500
expenditure$X1