PS1

Rubèn Llorens Poblador

2025-10-08

A school counselor was curious about the average of IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
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)
```

1. Find a 90% confidence interval for the average student IQ in the school.

We will start finding the mean of our data set to then be able to calculate the S:

```
mean_y <- sum(y)/length(y)
mean_y
```

```
## [1] 98.44
```

To find the sum of errors, we need to create a vector with the sum of the demeaned values of y

```
demeaned <- y - mean_y
```

And then, the sum of squared errors

```
squaredError <- sum(demeaned^2)
squaredError
```

```
## [1] 4114.16
```

I now calculate the variance and the S

```
variance <- squaredError/(length(y)-1)
S <- sqrt(variance)
S
```

```
## [1] 13.09287
```

Now we need to calculate the mean standard error (SE), thus, S/sqrt(n)

```r
SE <- S/sqrt(length(y))
SE
```

```
## [1] 2.618575
```

Now I'll find the t value, consdiering that the DF = 24 (as length(y) = 25)

```r
tcrit <- qt(0.95, df=24)
tcrit
```

```
## [1] 1.710882
```

Now I will get the Margin of Error (ME), this is, "tcrit · SE"

```r
ME <- tcrit*SE
ME
```

```
## [1] 4.480072
```

Now I just need to define the interval limits according to mean_y

```r
lower <- mean_y - ME
upper <- mean_y + ME
ci_90_y <- c(lower, upper)
ci_90_y
```

```
## [1]  93.95993 102.92007
```

This means that the 90% confidence interval for the average student IQ in the school is [93.96, 103.92], meaning that, with 90% of confidence, the real average IQ for this school's students is between both values.

2. Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country. Using the same sample, conduct the appropriate hypothesis test with alfa = 0.05.

To do this, we have a H0 that mu = 100 and a H1 that mu > 100. I need to do a T test.

```r
t <- (mean_y - 100)/SE
t
```

```
## [1] -0.5957439
```

```r
alfa <- 0.05
df <- length(y)-1
tcrit <- qt(1-alfa, df)
tcrit
```

```
## [1] 1.710882
```

Now, I must compare t (observed) and tcrit

```
t
```

```
## [1] -0.5957439
```

```
tcrit
```

```
## [1] 1.710882
```

As t is clearly lower than t, I can't reject the H0. Therefore, I don't have strong evidence to tell the school counselor whether the average student IQ in her school is higher than the average IQ score (100) among all the schools in the country.

I now start the second question. First, I import the data set (in .txt format) into R:

```
data <- read.delim("C:/Users/Usuario/Documents/GitHub/StatsI_2025/datasets/expenditure.txt",
                   header = TRUE, sep = "\t")
View(data)
```

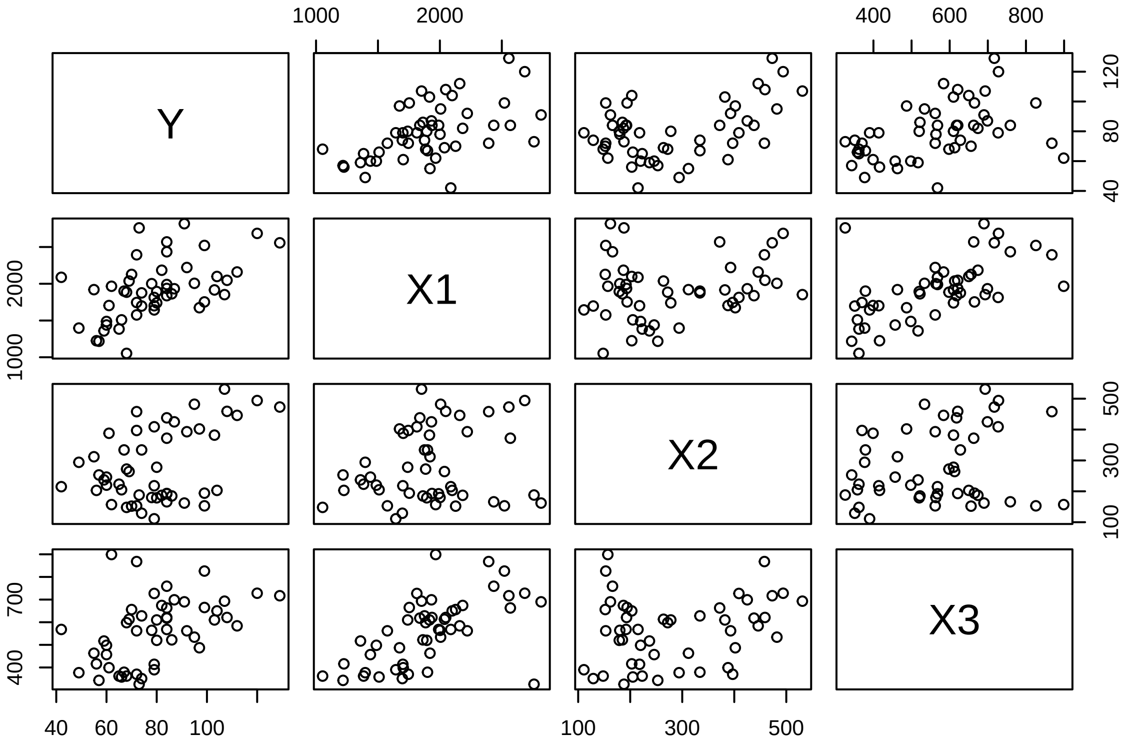
#2.1. Please plot the relationships among Y, X1, X2, and X3? What are the correlations among them (you just need to describe the graph and the relationships among them)?

#I start selecting the variables I'm interested in:

```
data_sel <- data[c("Y","X1","X2","X3")]
```

#And now I can plot the relationship amongst them:

```
plot(data_sel)
```



#Now, it’s time to calculate the correlations between these variables (using the function cor():

```
table_correlation <- cor(data_sel)
table_correlation
```

```
##            Y        X1        X2        X3
## Y  1.0000000 0.5317212 0.4482876 0.4636787
## X1 0.5317212 1.0000000 0.2056101 0.5952504
## X2 0.4482876 0.2056101 1.0000000 0.2210149
## X3 0.4636787 0.5952504 0.2210149 1.0000000
```

#Interpretation:

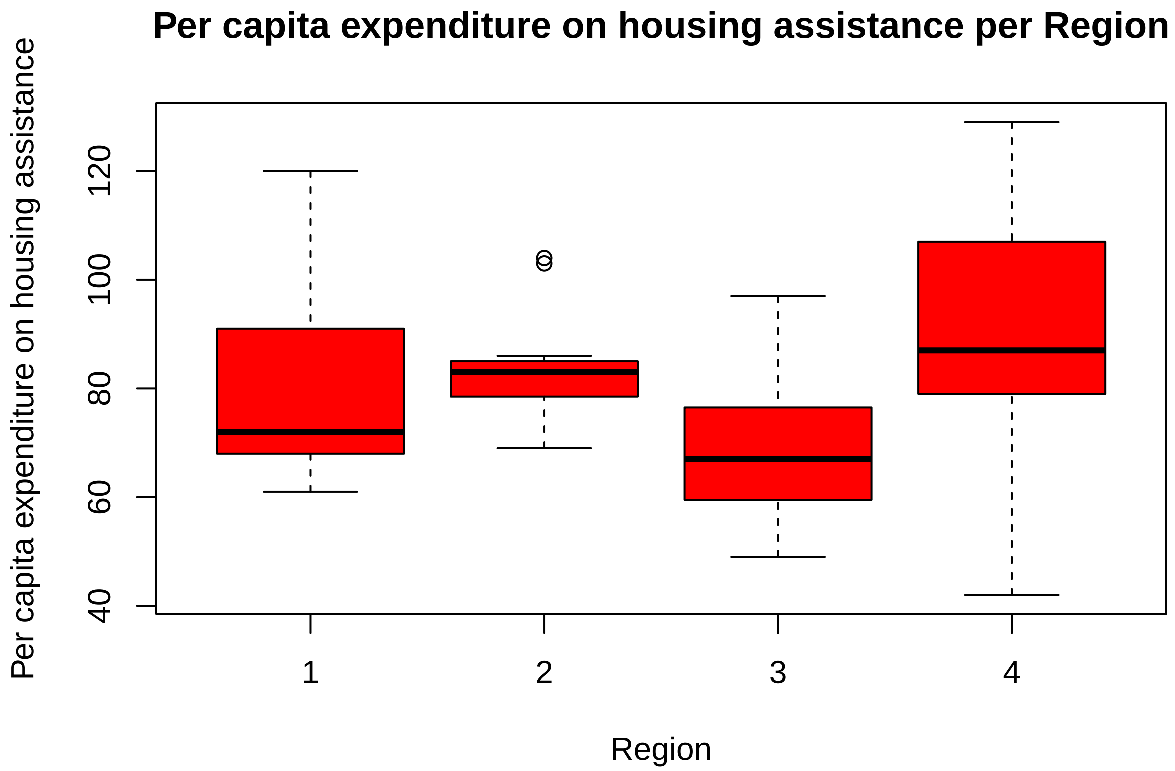
#As seen in the graphs, the correlations between all four variables are always positive, also their strength varies.

#Y is moderately correlated with all other three, but especially with X1. X1, however, is more strongly correlated with X3, while really weakly with X2. Finally, X2 and X3 are also weakly correlated.

#2.2. Please plot the relationship between Y and Region? On average, which region has the highest per capita expenditure on housing assistance?

#I will use the function boxplot(), asking r to use data from my dataset “data” and to use “Region” as the X and Y as the Y.

```
boxplot(Y ~ Region, data = data,
        main = "Per capita expenditure on housing assistance per Region",
        xlab = "Region",
        ylab = "Per capita expenditure on housing assistance",
        col = "red")
```



#According to the graph, region 4 has the highest per capita expenditure on housing assistance.

#But we can also calculate the mean per region:

```
mean_by_region <- tapply(data$Y, data$Region, mean)
mean_by_region
```

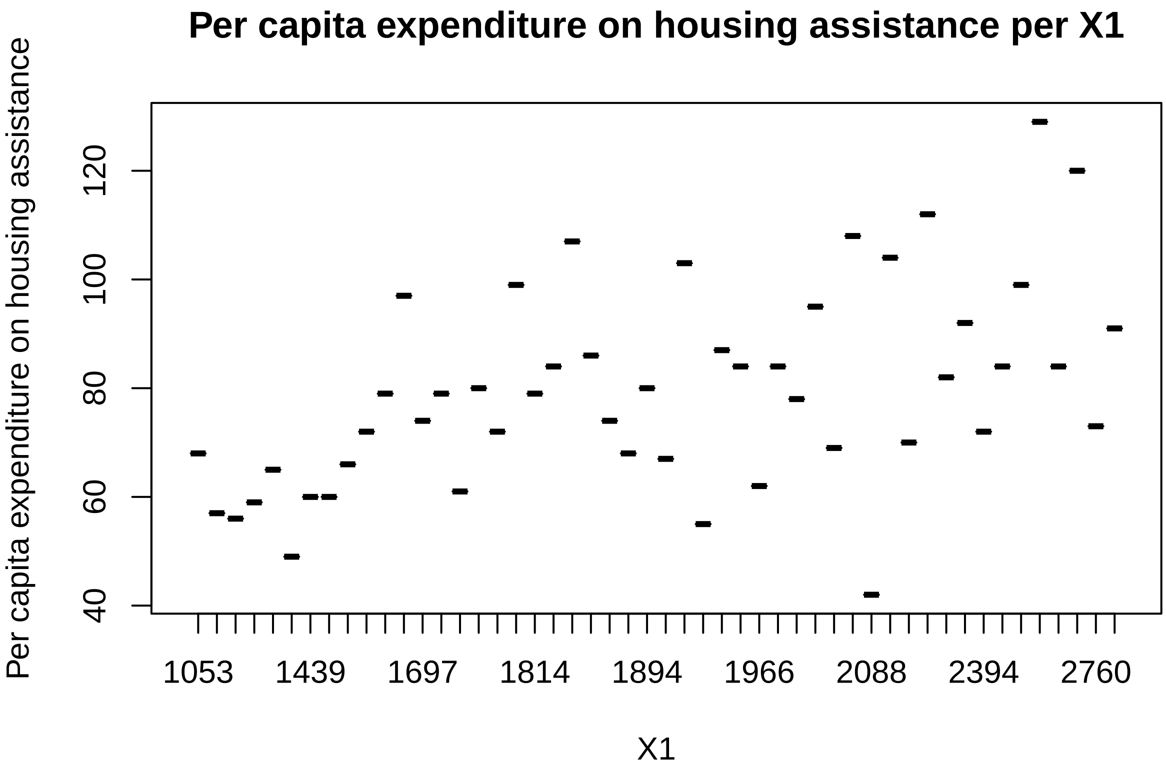
```
##        1        2        3        4
## 79.44444 83.91667 69.18750 88.30769
```

#Now, I confirm that my observation of the graph was correct.

#2.3. Please plot the relationship between Y and X1? Describe this graph and the relationship. Reproduce the above graph including one more variable Region and display different regions with different types of symbols and colors.

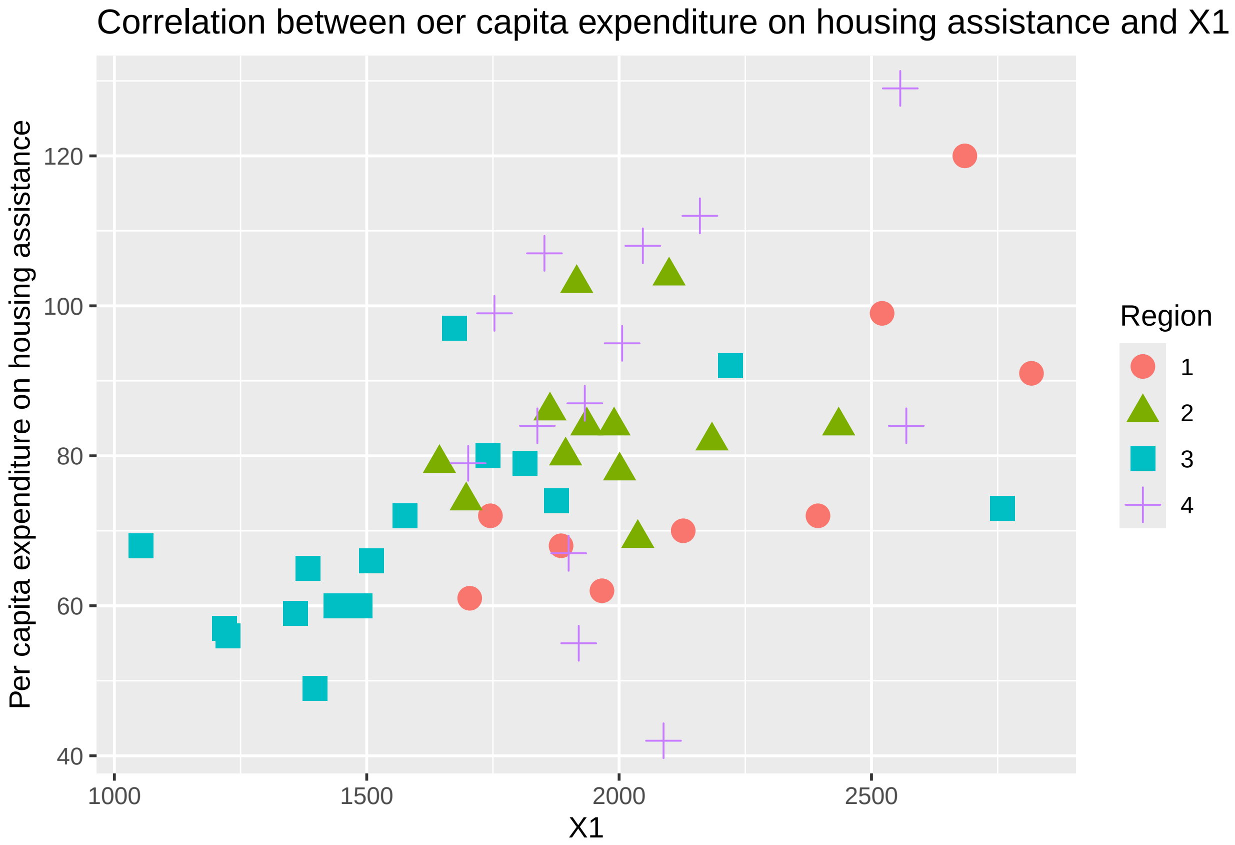
#I will use the function boxplot(), asking r to use data from my dataset "data" and to use "X1" as the X and Y as the Y.

```
boxplot(Y ~ X1, data = data,
        main = "Per capita expenditure on housing assistance per X1",
        xlab = "X1",
        ylab = "Per capita expenditure on housing assistance",
        col = "green")
```



#In general, we can see that the per capita expenditure on housing assistance increases as the value of X1 does. This is coherent with the correlation or r=0.53 that we have seen previously.

```
library(ggplot2)
ggplot(data = data, aes(x=X1, y=Y, color=as.factor(Region), shape=
                          as.factor(Region)))+geom_point(size=4)+labs(title = "Correlation between oer
       x = "X1",
       y = "Per capita expenditure on housing assistance",
       color = "Region",
       shape = "Region")
```



#We can now see that there are important regional differences. The graphic shows, as we have previously seen, that the per capita expenditure on housing assistance is, indeed, highest in Region 4 and lowest in Region 3. Also, Regions 3 and 4 seem to be quite homogeneous in both dimensions (variables X1 and Y).