Problem Set 1

Applied Stats/Quant Methods 1

Due: October 9, 2025

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Thursday October 9, 2025. No late assignments will be accepted.

Question 1: Education

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:

```r
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.

```r
#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
```

```r

#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

#I now calculate the variance and the S

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

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

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

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

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

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

ME <- tcrit*SE
ME

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

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

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 $\alpha = 0.05$.

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

t <- (mean_y - 100)/SE
t

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

#Now, I must compare t (observed) and tcrit

t
tcrit
```

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.

Question 2: Political Economy

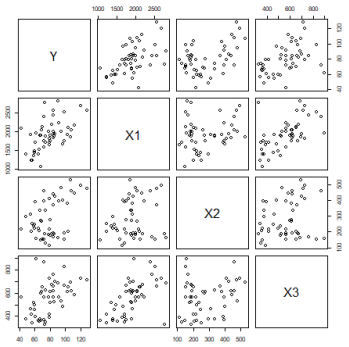
Researchers are curious about what affects the amount of money communities spend on addressing homelessness. The following variables constitute our data set about social welfare expenditures in the USA.

`State`	*50 states in US*
`Y`	*per capita expenditure on shelters/housing assistance in state*
`X1`	*per capita personal income in state*
`X2`	*Number of residents per 100,000 that are "financially insecure" in state*
`X3`	*Number of people per thousand residing in urban areas in state*
`Region`	*1=Northeast, 2= North Central, 3= South, 4=West*

Explore the `expenditure` data set and import data into `R`.

```
#I now calculate the variance and the S
```

- 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)?



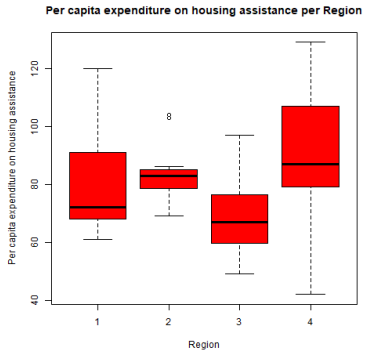

```
#I start selecting the variables I'm interested in:


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

#And now I can plot the relationship amongst them:

pdf("pdf1.pdf")
plot(data_sel)


```

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


table_correlation <- cor(data_sel)
```

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.

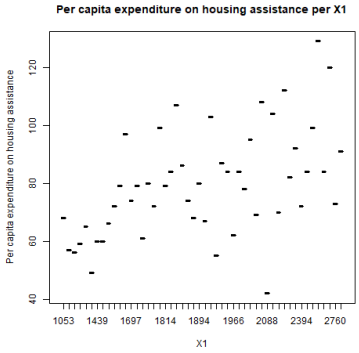
- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?




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

boxplot(Y ~ Region, data = expenditure,
        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(expenditure$Y, expenditure$Region, mean)
mean_by_region
```

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

- 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.

pdf("pdf2.pdf")
boxplot(Y ~ X1, data = expenditure,
        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)
pdf("pdf3.pdf")
ggplot(data = expenditure, aes(x=X1, y=Y, color=as.factor(Region), shape=
                               as.factor(Region)))+geom_point(size=4)+labs(
    title = "Correlation between oer capita expenditure on housing
    assistance and X1 per Region",
                                                                    x =
     "X1",
                                                                    y =
     "Per capita expenditure on housing assistance",
```

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).