Problem Set 1 Response

Applied Statistics/Quantitative Methods 1

Victor Gomez

September 30^{th} 2024

Problem 1

Preliminary code

Here is first a succint presentation of dataset. This one used is a sample y of 25 IQ values of students.
Loaded as follows:

```
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)
```

Which can be summarised with

```
summary(y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  69.00   89.00   98.00   98.44  110.00  126.00
```

Question 1

We assume that the sample follow a Sutend distribution of $\#y-1$ degrees of freedom, since its size is small ($\lesssim 30$). Therefore, the confidence interval CI at 90% confidence (or at risk $\alpha := 0.1$) is

$$CI_{90\%} = [\bar{y} - T(0.95).se(y), \bar{y} + T(0.95).se(y)]$$

where se states for standard error, and
T is the Student cumulative distribution function
Which leads to the following code:

```
n <- length(y) # capture the number of observations
sgm <-sd(y) # Calculate empirical standard deviation
mu <- mean(y) # calculate empirical mean
```

```r

# Assuming y values follow a student law, we obtain the following confidence
    interval CI at 90% of confidence:

alpha <-0.1
CI_standard <- c(qt(alpha/2,df=n-1),qt(1-alpha/2,df=n-1))
CI <- CI_standard*sgm/sqrt(n) + mu
CI
```

Which result is :

```
[1]  93.95993 102.92007
```

for lower and upper bound respectively. I.e. $CI_{90\%} = [93.95993, 102.92007]$

Question 2

We consider Hypothesis $\mathcal{H}_0$ as "Average student IQ is lower or equal than the average IQ score among all the schools int he country."
We denote $\mu_{general}$ the mean of general population ($\mu_{general} = 100$) and $\mu_{student}$ the IQ mean of the students.
I.e.

$$\mathcal{H}_0 := \{\mu_{general} \geqslant \mu_{students}\}$$

So, the test statistic T is

$$T := \frac{\bar{y} - \mu_{general}}{\sigma_y/\sqrt{n}}$$

where σ_y is the standard deviation of the sample y and $\sqrt{n} := \#y$. We apply a one-sidded T-test to this hypothesis, with risk $\alpha := 5\%$

```r
alpha<-0.05
test<-t.test(y, mu=100, alternative = 'greater',conf.level = 1-alpha)
test
test$p.value
```

```
> test

 One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
```

```
 93.95993      Inf
sample estimates:
mean of x
    98.44

> test$p.value
[1] 0.7215383
```

I.e the p-value is 0.7215383. $\mathcal{H}_0$ is not rejected.

Problem 2

Preliminary considerations

We load the required data set as follow:

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
    Fall2024/main/datasets/expenditure.txt", header=T)
```

and summarise it.

```
summary(expenditure)
```

```
   STATE                  Y                X1             X2       
 Length:50          Min.   : 42.00   Min.   :1053   Min.   :111.0  
 Class :character   1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2  
 Mode  :character   Median : 79.00   Median :1897   Median :241.5  
                    Mean   : 79.54   Mean   :1912   Mean   :281.8  
                    3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8  
                    Max.   :129.00   Max.   :2817   Max.   :531.0  

       X3            Region    
 Min.   :326.0   Min.   :1.00  
 1st Qu.:426.2   1st Qu.:2.00  
 Median :568.0   Median :3.00  
 Mean   :561.7   Mean   :2.66  
 3rd Qu.:661.2   3rd Qu.:3.75  
 Max.   :899.0   Max.   :4.00  
```

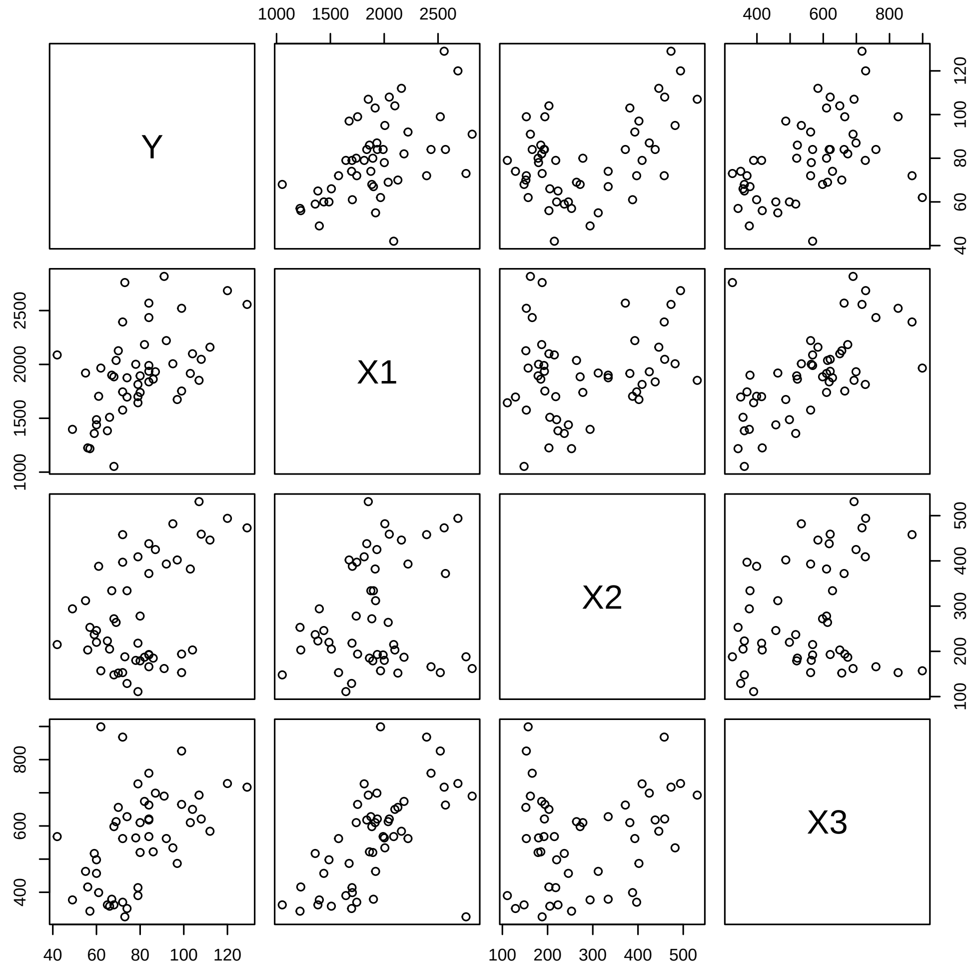
For a further introduction to dataset cf. problem's wording.

Question 1

The following code produce the next figure.

```
pairs(expenditure[,2:5])
```

Figure 1: Pair plot between $Y, X1, X2, X3$



The plot above shows the bivariate relationship between $Y, X1, X2, X3$ variables. Relationships between $X1$ and $X2$, and $X2$ and $X3$ seem aleatory. However, $X1$ and $X3$ seem to have a linear relation.
Then Y seem to have some positive correlation (in Spearman's correlation coefficient sens, i.e. here, linearity is not obvious) with $X1$ $X3$, a linear relationship with $X1$, and Y-$X2$ relationship seems parabolic. For further analysis, a covariance matrix would be appropriated (for linear relationships only).

Question 2

The folowing plot is generated with:

```
#Create box plot graph with values
pdf("regions_boxplot.pdf")
```

```r
boxplot(expenditure$Y ~ expenditure$Region,xlab="Region", ylab="Y",xaxt = "n")
axis(1, at=c(1,2,3,4), labels=c("Northeast","North Central","South","West"),
    las=0)
# Add data points with jitter for limiting points overlap.
mylevels <- as.numeric(levels(factor(expenditure$Region)))
levelProportions <- summary(expenditure$Region)/nrow(expenditure)
for(i in 1:length(mylevels)){
  thislevel <- mylevels[i]
  thisvalues <- expenditure[expenditure$Region==thislevel, "Y"]

  # take the x-axis indices and add a jitter, proportional to the N in each
    level
  myjitter <- jitter(rep(i, length(thisvalues)), amount=levelProportions[i]/2)
  points(myjitter, thisvalues, pch=1, col=rgb(0,0,0,.9))

}
dev.off()
```

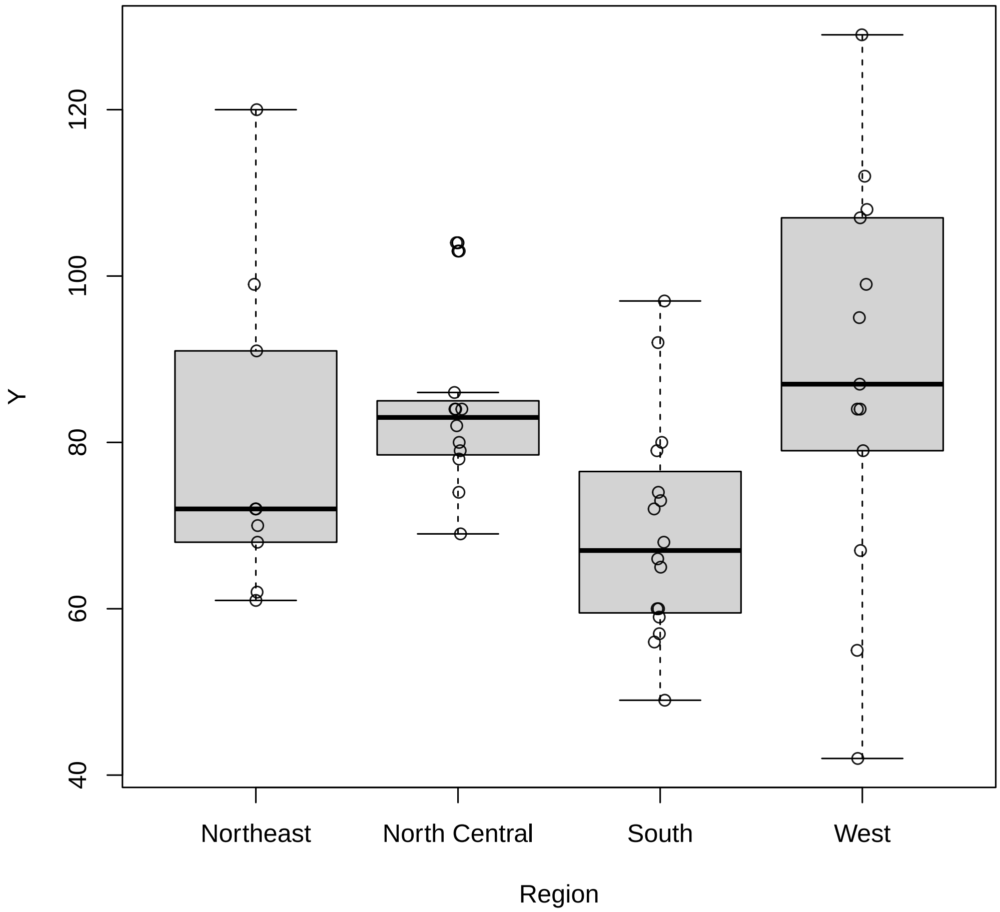
Considering the graph above, the region with, in average, the highest per capita expenditure on housing assistance is *West*. The box plot that *West* has the highest median (90$/capita) more or less centered between min (45$/capita) and max (110$/capita) were at the lower third between the1^{st} and the 3^{rd} quantile. Thus, average for *West* region should be higher than the median. On the other hand, the 3^{rd} quantile of *Northcentral* and *South* are below the median of *West* region and their maxima are under the 3^{rd} quantile of *West*. This leads to consider that their respective means should be under the mean of *West* region. Finally, *Northeast* has more dispersion, while its median is below 1^{rst} quantile of *West*. Its 3^{rd} quantile is more or less equal to *West*'s median and its maximum is below *West*'s maximum. I.e. , mean of *Northeast*'s region is below *West* average.
To confirm this visual analysis, it suffice to calculate the actual means of each region sample.

Question 3

The following graph is a reproduction of one of the plot of Fig.1 where are added *Region* repartition by colours, simply generated by:

```r
plot(expenditure$X1,
     expenditure$Y,
     col=expenditure$Region,
     pch=expenditure$Region,
     xlab="per capita personal income in state ($/pers)",
     ylab="per capita expenditure on shelters/housing assistance in state ($/
   pers)",
     main="Relationship between shelters/housing assistance and personal
   income")
# Add legend
legend(1000, 130, # x and y position of legend
        legend=c("Northeast", "North Central", "South", "West"),
        col= as.integer(names(table(expenditure$Region))),
        pch= as.integer(names(table(expenditure$Region))))
```

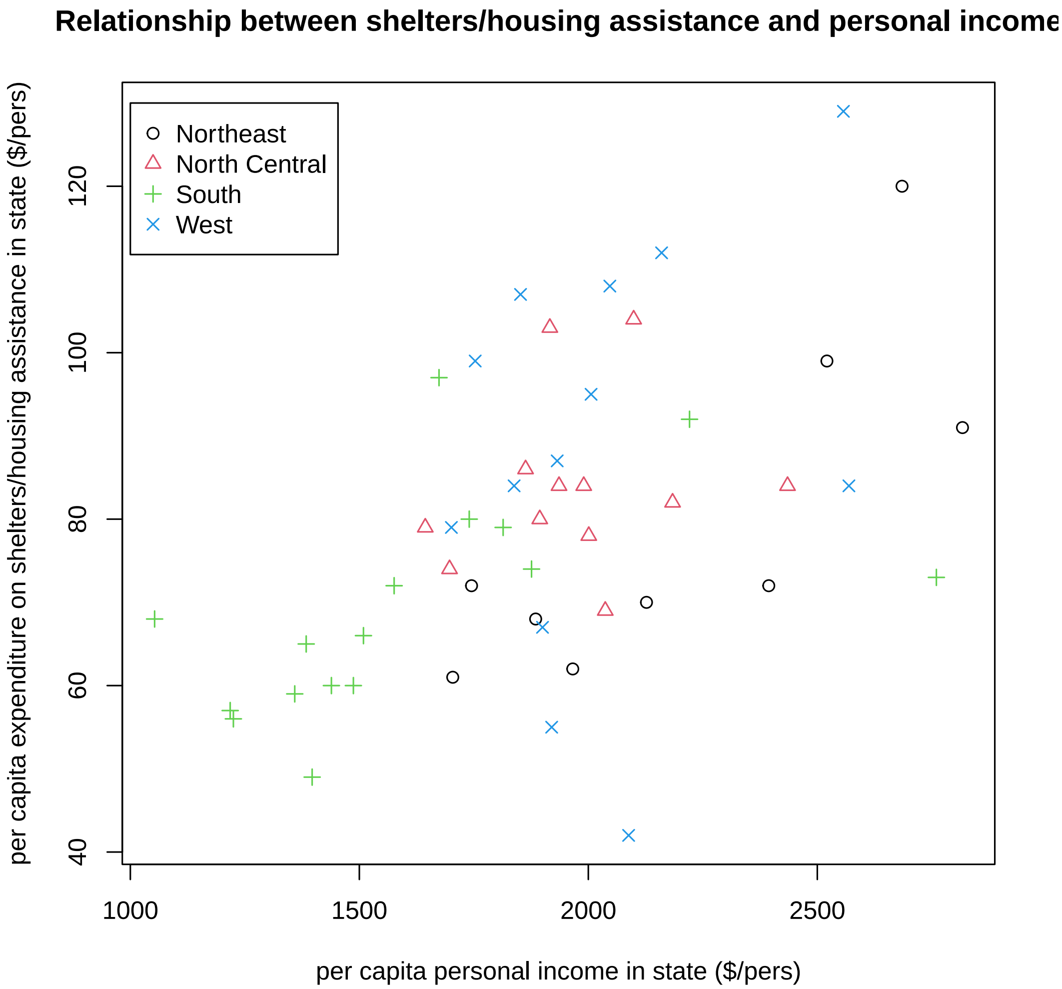
Figure 2: Boxplot of Y according to *Region*



The Fig.3 shows that *South* region's people have mainly lower income than others (mainly between 1200 and 1600 \$/pers.) and use generally give between 50 and 70\$/pers. for shelters or housing assistance.
NorthCentral follows the smae pattern of a centered group , but with higher incomes and higher expenditure (respectively between 1700 and 2200\$/pers. and 70 and 100\$/pers.).
At the contrary, the expenditure givien by *West* region's sample seems not to be correlated to $X1$ variable, since the main income is between 1700 and 2100\$/pers. were as expenditure are spread from 60\$/pers. to 110\$/pers.
Concerning *Northeast* region, it is difficult to give any comment, since sample values are scarce and scattered.

In order to be compleate, it has to be mentioned that the samples for each region are little,

Figure 3: Relationship between Y and $X1$ showing *Regions*



thus, intervals given above are unprecise and have to be considered as rough approximations.

To conclude regional analysis, *South* and *NorthCentral* region seem to from punctual clusters, the former below the latter for two edges. In addition, *West* expenditure seems to be indifferent to personal income, and it is not possible to conclude concerning *Northeast* region.

Finally, concerning the general relation between per capita expenditure on shelters and housing and per capita personal income, it might appear a linear relation, without homoscedasticity, i.e. with a growing dispersion along per capita personal income variable.