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.

Solution to question 1.1

```r
#Question1.1
#Sample size
n <- length(y)
# Sample mean and standard deviation
mean_iq <- mean(y)
```

```r
sd_iq   <- sd(y)
# Standard error of the mean
se <- sd_iq / sqrt(n)
#Confidence Intervals
# t critical value for 90% CI (two-tailed, df = n-1)
t_crit <- qt(0.90, df = n - 1)
# Confidence interval
lower <- mean_iq - t_crit * se
upper <- mean_iq + t_crit * se
cat("90% Confidence Interval for mean IQ: [",lower,",",upper,"]\n")
```

The 90 percent confidence intervals are [94.99 , 101.89]

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 $\alpha = 0.05$.

Solution to question 1.2

```r
#Question1.2
# One-sided test: is the mean greater than 100?
mu0 <- 100 #hypothesized mean
test_statistic <- (mean_iq - mu0)/se
test_statistic
#finding t_crit using alpha = 0.05
t_crit <- qt(0.05, df = n - 1, lower.tail = FALSE)
t_crit
# comparison and decision
if(test_statistic > t_crit){
  decision <- "Reject H0: The school's mean IQ is significantly higher
   than 100."
} else {
  decision <- "Fail to Reject H0: No evidence that the school's mean IQ
   is higher than 100."
}
list(
  decision = decision
)
```

According to my hypothesis test, the test statistic is smaller than the critical value so there is no evidence that the school's mean IQ is higher than 100.

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*

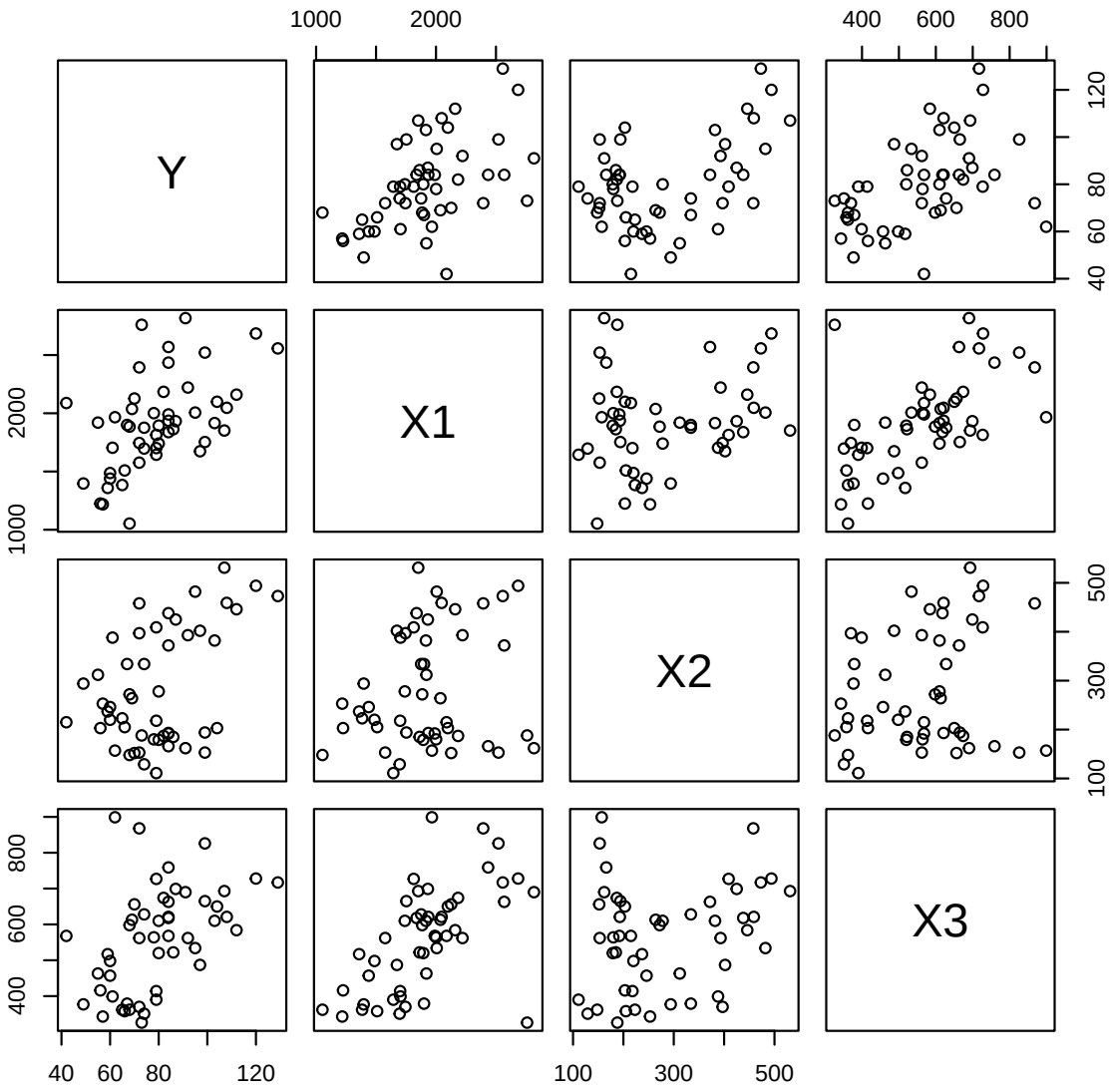
Explore the `expenditure` data set and import data into `R`.

- 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)?

Solution to question 2.1

```r
#Question 2.1
#read expenditure data
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_2025/main/datasets/expenditure.txt", header=T)
#Explore Data
str(expenditure)          # Check structure and variable types
summary(expenditure)      # Get descriptive statistics
head(expenditure)         # Preview first few rows
# Select only the variables of interest
exp_data <- expenditure[, c("Y", "X1", "X2", "X3")]
# Create the pairs plot
pairs(exp_data)
```

Here, I first inspected the data using str, sumamry and head. "expdata" is a subset of the original data set; that includes only the necessary variables. Using the 'pairs' function I plotted the data. From the graphs below we can observe the following: Most graphs have a positive correlation among them, however, X1 and X2, X2 and X3 don't really have a strong trend. Whereas, Y has a positive trend with all the graphs.

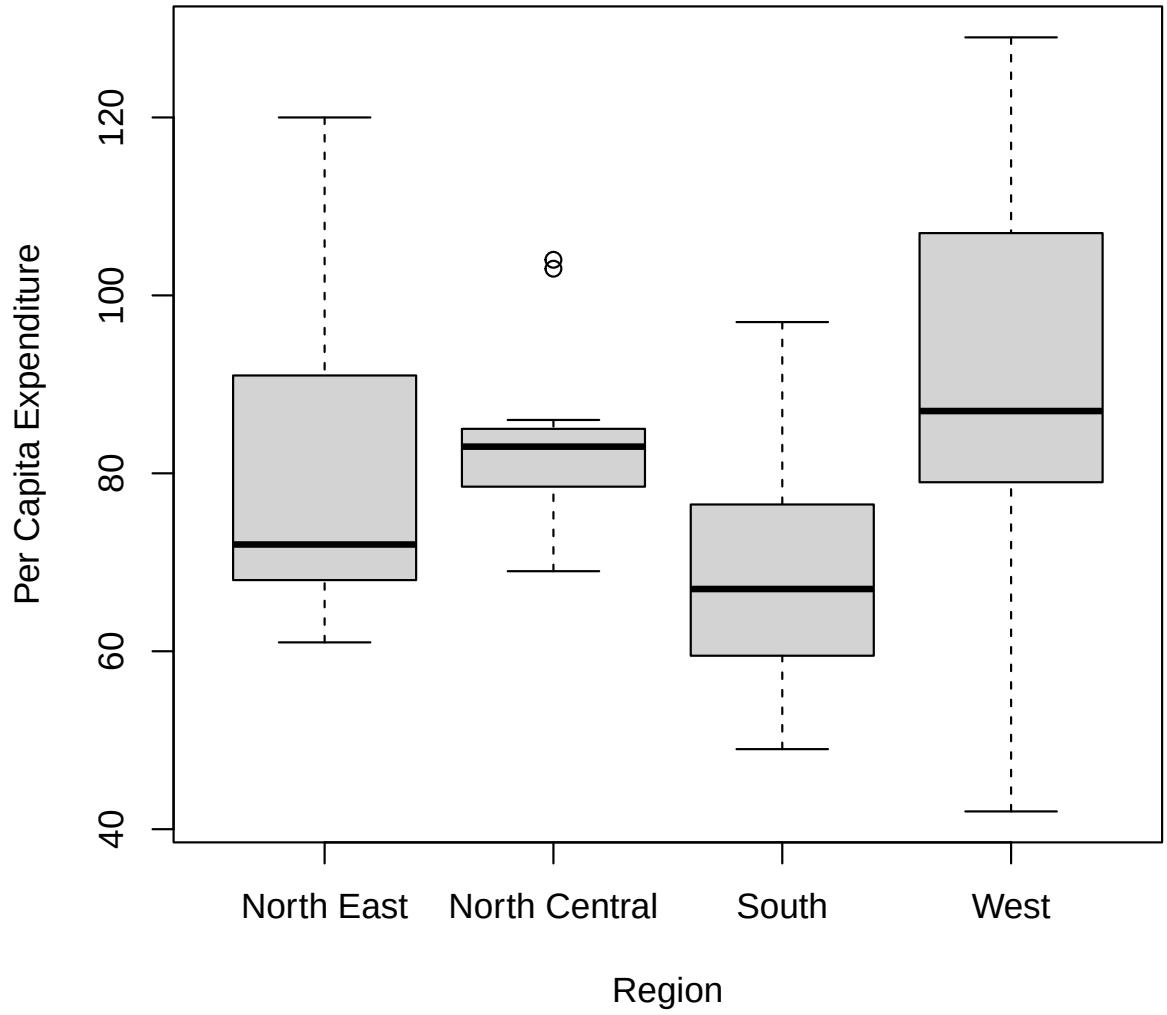


- Please plot the relationship between Y and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

Solution to question 2.2

```
#region and plot
boxplot(Y ~ Region, data = expenditure,
        xlab = "Region",
        ylab = "Per Capita Expenditure",
        names= c("North East", "North Central","South", "West"))
```

```
#Question2.3
```



In this plot, the region which has the highest per capita expenditure on housing asistance on average is West.

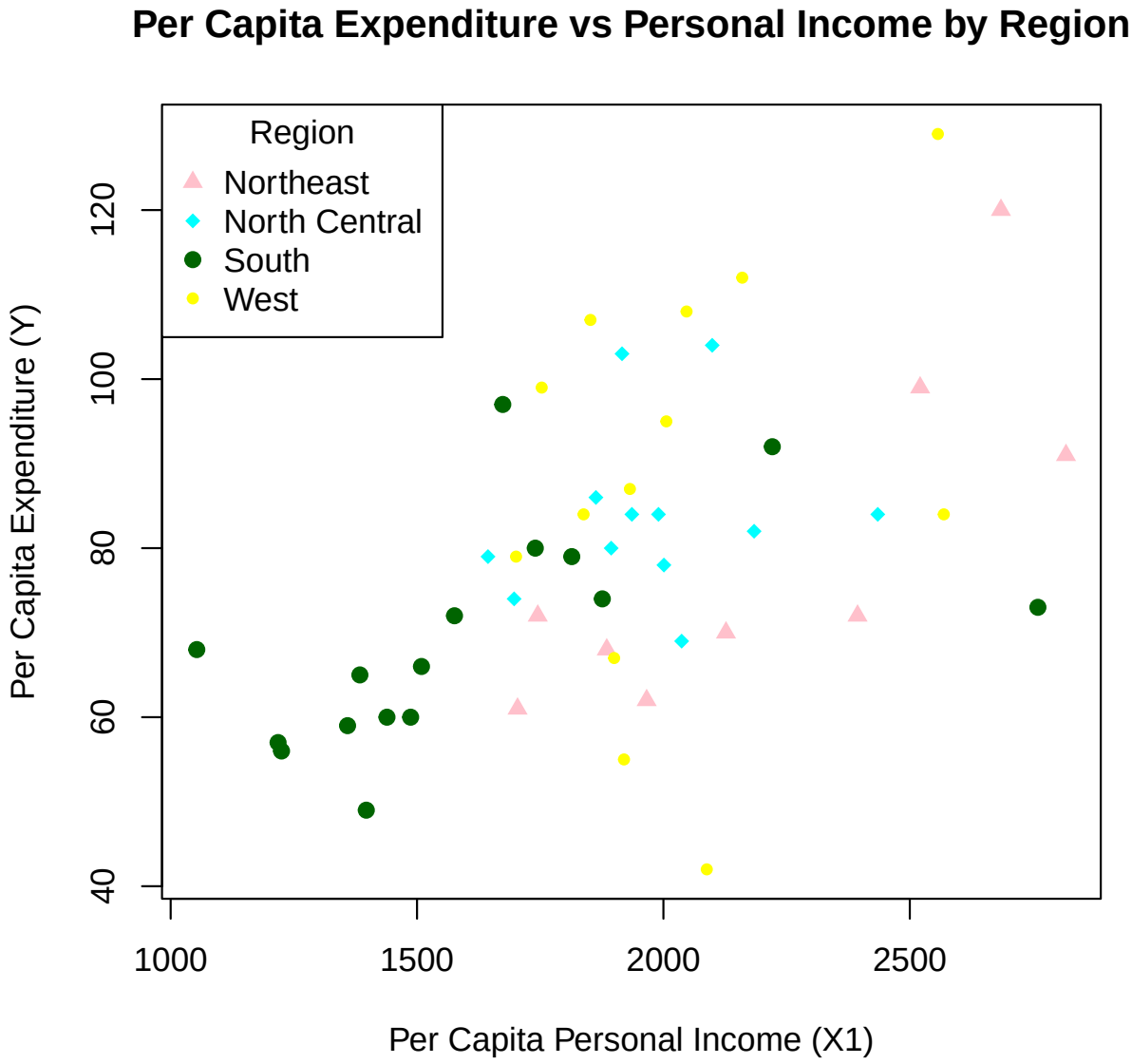
- Please plot the relationship between Y and $X1$? Describe this graph and the relationship. Reproduce the above graph including one more variable *Region* and display different regions with different types of symbols and colors.

Solution to question 2.3

```
region_colors <- c("pink", "cyan", "darkgreen", "yellow")
region_symbols <- c(17, 18, 19, 20)

#plot
expenditure$Region <- factor(expenditure$Region,
                      levels = c(1,2,3,4),
                      labels = c("Northeast","North Central","South","
    West"))
plot(expenditure$X1, expenditure$Y,
     main = "Per Capita Expenditure vs Personal Income by Region",
     xlab = "Per Capita Personal Income (X1)",
     ylab = "Per Capita Expenditure (Y)",
     col = region_colors[expenditure$Region],
     pch = region_symbols[expenditure$Region])
#legend
legend("topleft",
       legend = levels(expenditure$Region),
       col = region_colors,
       pch = region_symbols,
       title = "Region")
```

From the graph below we can infer that, there is a a noticeable positive linear relationship between the per capita expenditure and the personal income per capita for all regions.


Per Capita Expenditure vs Personal Income by Region
Region
Northeast
North Central
South
West
Per Capita Expenditure (Y)
40
60
80
100
120
1000
1500
2000
2500
Per Capita Personal Income (X1)