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.

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

1. The 90% confidence interval of the random sample of 25 students' IQ scores lies between 94.99 and 101.89.

   My code:

```r
n <- length(y)
y_mean <- mean(y)
y_sd <- sd(y)
t90 <- qt((1 - .90), df = (n - 1), lower.tail = FALSE)
t90

lower_bound <- y_mean - (t90 * (y_sd/sqrt(n)))
upper_bound <- y_mean + (t90 * (y_sd/sqrt(n)))

ci <- c(lower_bound, upper_bound)
ci
```

2. There is not enough statistical evidence to conclude that the average IQ of students at this school is higher than the national average of 100, as the calculated t-statistic (-0.596) is not greater than the critical t-value (1.711)

   My code:

```r
#calculating test statistic
mu <- 100
se <- y_sd / sqrt(n)
t_statistic <- (y_mean - mu) / se
t_statistic

#comparing t-statistic to the critical t-value
t_crit <- qt((1 - .95), df = (n - 1), lower.tail = FALSE)
is_significant <- t_statistic > t_crit
is_significant
```

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`.

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_2025/main/datasets/expenditure.txt", header=T)
```

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)?

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

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.

Solution to Question 2

1. The scatterplot matrix reveals no strong or clear-cut correlations among the variables. However, a few moderate linear relationships are noteworthy:

   - Y and $X1$ show a moderately positive linear relationship, suggesting that states with higher per capita income tend to spend more on shelters and housing assistance.
   - $X1$ and $X3$ also appear moderately positively correlated, indicating that higher-income states may have more urbanised populations.
   - $X2$ does not exhibit any obvious linear relationship with the other variables, implying that financial insecurity may not be directly associated with expenditure, income, or urbanisation in a consistent way.
   - $X3$ and Y show a weak positive linear trend, suggesting that more urbanised states might spend slightly more on housing assistance, though the relationship is not strong.

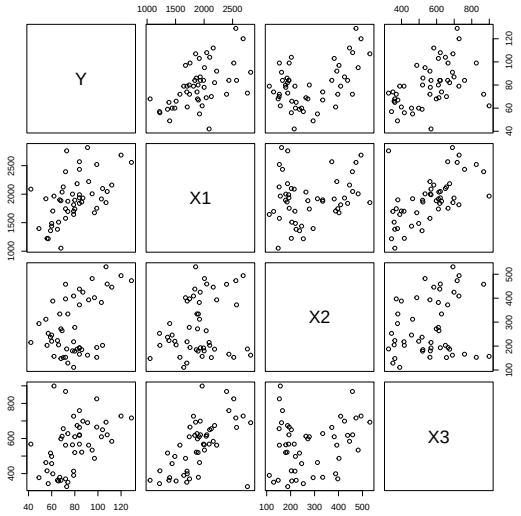


Figure 1: Scatterplot matrix of selected variables

My code:

```
pdf("pairs_plot_2_1.pdf")
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/
    StatsI_2025/main/datasets/expenditure.txt", header=T)
head(expenditure)
str(expenditure)
#?pairs
select_vars <- expenditure[,c("Y","X1","X2","X3")]
pairs(select_vars)
dev.off()
```

2. On average, the region *West* (4) has the highest expenditure on housing assistance.

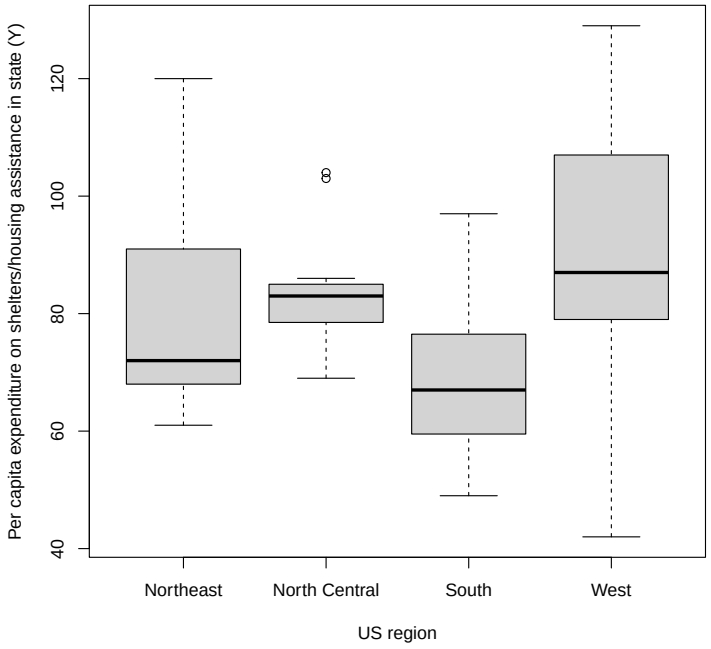
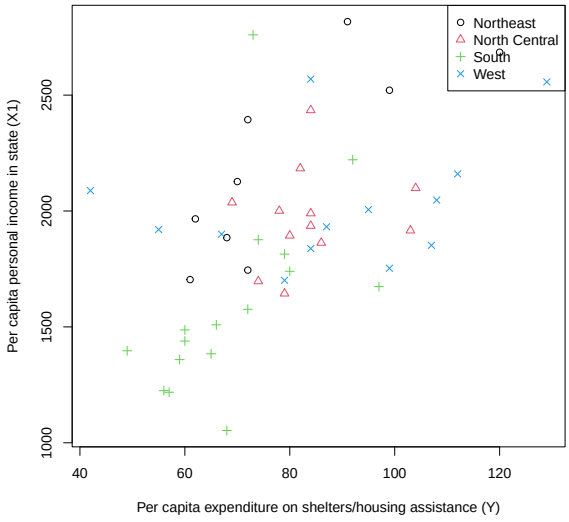


Figure 2: Boxplot showing Y by *Region*

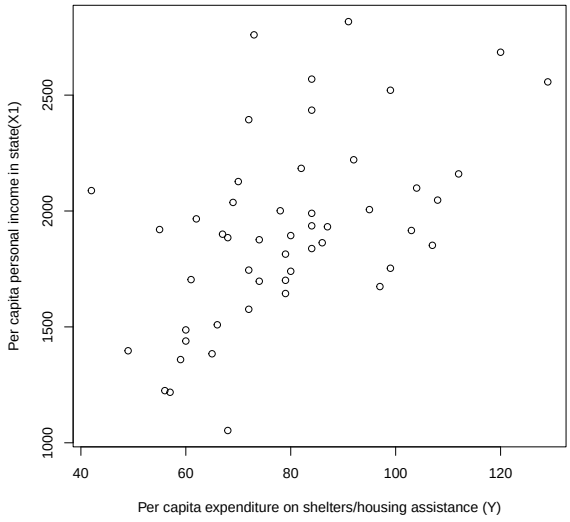
My code:

```
pdf("box_plot_2_2.pdf")
boxplot(Y ~ Region, data = expenditure,
        ylab="Per capita expenditure on shelters/housing assistance in
    state (Y)",
        xlab="US region",
        names=c("Northeast", "North Central", "South", "West"))
dev.off()
```

3. Y and $X1$ show a moderately positive linear relationship, suggesting that states with higher per capita income tend to spend more on shelters and housing assistance.



(a) Scatterplot of $X1$ versus Y, with points coloured and shaped by *Region*.



(b) Scatterplot showing the relationship between $X1$ and Y.

My code:

```
pdf("scatter_plot_2_3a.pdf")
plot(expenditure$Y, expenditure$X1,
     xlab="Per capita expenditure on shelters/housing assistance (Y)",
     ylab="Per capita personal income in state(X1)")
dev.off()
pdf("scatter_plot_2_3b.pdf")
plot(expenditure$Y, expenditure$X1,
     xlab="Per capita expenditure on shelters/housing assistance (Y)",
     ylab="Per capita personal income in state (X1)",
     col=expenditure$Region,
     pch=expenditure$Region)
legend("topright",
       legend = c("Northeast", "North Central", "South", "West"),
       col = c(1, 2, 3, 4),
       pch = c(1, 2, 3, 4))
dev.off()
```