Problem Set 1

Athena Rodrigues

Due: September 30, 2024

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 Monday September 30, 2024. 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.

```
90% CI = [94, 103]
```

Code:

```r
# 90% Confidence Interval: point estimate +/- margin of error
ymean <- mean(y) # mean 98.44
df <- length(y)-1 # degrees of freedom 24
ystderr <- sd(y)/sqrt(length(y)) # standard error 2.618575
confidence <- qt((.9)+ (1-.9)/2, df) # confidence 1.710882
upper_90 <- (ymean + (confidence)*ystderr)  # upper bounds 102.9201
lower_90 <- (ymean - (confidence)*ystderr)  # lower bounds 93.95993
```

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 $\alpha = 0.05$.

```
Step 1: Assumptions
        The data y is a random, continous, numeric sample.
Code:
```

```r
class(y)
length(y)
```

```
Step 2: Hypothesis
        The Null Hypothesis is that the mean is <= 100
        The Alternative Hypothesis is that the mean > 100
Step 3: Calculate the Test Statistic
        The test statistic is -0.5957439.
Code:
```

```r
test_stat <- ((ymean - 100)/ystderr)
```

```
Step 4: Calculate the P-Value
        The p-value is  0.7215383.
Code:
```

```r
pvalue <- (1-pt(test_stat, df))
```

```
Step 5:
        Based on sufficient evidence, the p-value (0.7215) is
        greater than the alpha (0.05),
        we fail to reject the null hypothesis that the
        average IQ score is less than or equal to 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`.

```
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/expenditure.txt", header=T)
head(expenditure)
```

- 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)?

Figure 1: Relationships Between Variables (Y, X1, X2, and X3)

Relationships:
Housing Expenditure vs. Personal Income (top left) shows
a linear relationship with a strong, positive correlation.
As personal income per capita in state increases
so does the expenditure per capita on
housing assistance in state.

Housing Expenditure vs. Financial Insecurity (top middle) has a
nonlinear relationship with a weak correlation.

Housing Expenditure vs. Urban Residents (top right) showcases a
linear relationship with a weak, positive correlation.
This shows there is an effect in expenditure on
housing assistance as urban areas grow.

Financial Insecurity vs. Personal Income (bottom left) has a
nonlinear relationship with little to no correlation.
The variables of personal income and financial insecurity
have little impact on eachother.

Urban Residents vs. Personal Income (bottom middle) has a linear relationship
with a strong, positive correlation. There is a possible connection between
increased personal income in state and increasing urban populations.

Urban Residents vs. Financial Insecurity (bottom right) has a nonlinear
with little to no correlation. This shows that these
variables have little impact in relation to each other.

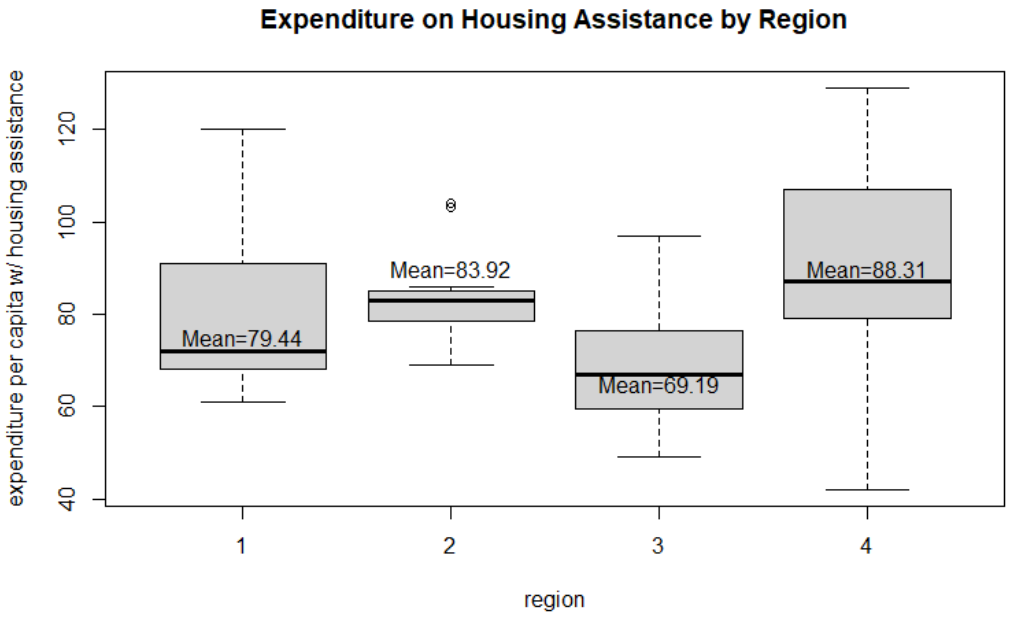
Code:

```
par(mfrow =c(2,3))
plot(expenditure$X1,expenditure$Y, col='red',
     xlab = "per capita personal income in state",
     ylab = "expenditure per capita on housing assistance",
     main = "Housing Expenditure vs. Personal Income",
     text(1450,120 , sprintf("Corr=%s", round(cor(expenditure$X1,
   expenditure$Y), 4))))
plot(expenditure$X2,expenditure$Y, col='blue',
     xlab = "# residents 'financially insecure' per 100,000",
     ylab = "expenditure per capita on housing assistance",
     main = "Housing Expenditure vs. Financial Insecurity",
     text(200, 120, sprintf("Corr=%s", round(cor(expenditure$X2,
   expenditure$Y), 4))))
plot(expenditure$X3,expenditure$Y, col='orange',
     xlab = "# residents residing in urban areas per thousand",
     ylab = "expenditure per capita on housing assistance",
```

```
      main = "Housing Expenditure vs. Urban Residents",
      text(500, 120, sprintf("Corr=%s", round(cor(expenditure$X3,
    expenditure$Y), 4))))
plot(expenditure$X1, expenditure$X2, col='pink',
      xlab = "per capita personal income in state",
      ylab = "# residents 'financially insecure' per 100,000",
      main = "Financial Insecurity vs. Personal Income",
      text(1400, 500, sprintf("Corr=%s", round(cor(expenditure$X1,
    expenditure$X2), 4))))
plot(expenditure$X1, expenditure$X3, col='purple',
      xlab = "per capita personal income in state",
      ylab = "# residents residing in urban areas per thousand",
      main = "Urban Residents vs. Personal Income",
      text(1400, 800, sprintf("Corr=%s", round(cor(expenditure$X1,
    expenditure$X3), 4))))
plot(expenditure$X2, expenditure$X3, col='magenta',
      xlab = "# residents 'financially insecure' per 100,000",
      ylab = "# residents residing in urban areas per thousand",
      main = "Urban Residents vs. Financial Insecurity",
      text(250, 800, sprintf("Corr=%s", round(cor(expenditure$X2,
    expenditure$X3), 4))))
```

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

Figure 2: Relationship between Y and Region



On average, the West (Region 4) has the highest per capita expenditure on housing assistance at 88.31 followed by the North Central (Region 2), the Northeast (Region 1), and the South (Region 3).

Code:

```
boxplot(expenditure$Y ~ expenditure$Region,
        main = "Expenditure on Housing Assistance by Region",
        xlab = "region",
        ylab = "expenditure per capita w/ housing assistance")
text(1, 75, sprintf("Mean=%s", round(mean(expenditure$Y[expenditure$
    Region == "1"]), 2)))
text(2, 90, sprintf("Mean=%s", round(mean(expenditure$Y[expenditure$
    Region == "2"]), 2)))
text(3, 65, sprintf("Mean=%s", round(mean(expenditure$Y[expenditure$
    Region == "3"]), 2)))
text(4, 90, sprintf("Mean=%s", round(mean(expenditure$Y[expenditure$
    Region == "4"]), 2)))
```

- 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.

Region 1 and 3 have a positive linear relationship and a wide data spread with strong correlations.
Region 2 and 4 have a nonlinear relationship with a central distribution of data and little to no correlation between variables.
This means, according to these basic correlation tests, that states in the Northeast and the South generally see an increase in expenditure per capita on housing assistance when personal income per capita increases as well, however, this cannot be said for the North Central and West.

Code:

```
plot(expenditure$X1,expenditure$Y,
     col=expenditure$Region, pch = expenditure$Region,
     xlab = "per capita personal income in state",
     ylab = "expenditure per capita w/ housing assistance",
     main = "Housing Expenditure vs. Personal Income")
legend("bottomright", legend = c("1", "2", "3", "4"),
       col = unique(expenditure$Region),
       pch = unique(expenditure$Region),
       title = "Region")
cor(expenditure$X1[expenditure$Region == "1"],expenditure$Y[expenditure$
    Region == "1"])
```

```
cor(expenditure$X1[expenditure$Region == "2"],expenditure$Y[expenditure$
    Region == "2"])
cor(expenditure$X1[expenditure$Region == "3"],expenditure$Y[expenditure$
    Region == "3"])
cor(expenditure$X1[expenditure$Region == "4"],expenditure$Y[expenditure$
    Region == "4"])
```

Figure 3: Relationship between Y and X1 based on Region