Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2023

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 Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): 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:

```
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.

   This question is asking to use statistics from a sample (25 random students) to estimate population parameters (the whole school) - in this case, a confidence interval, which is a range of numbers within which the parameter is believed to fall 90% of the time with repeated sampling. This requires us to assume a normal sampling distribution.

First I find the point estimate - or sample mean - which is the sum of all the observation values divided by the total number of observations. I also find the sample's n-size.

```r
n_y <- length(y)
mean(y) == sum(y)/n_y # is TRUE
mean_y <- mean(y)
```

Now I find the sample standard deviation - which is a measure of the spread/dispersion within the sample. This is calculated by first calculating the variance (the sum of the squared distances of each observation from the sample mean (ensuring observations above and below the mean don't cancel out) divided by the total number of observations less one ie n-1). The SD is the square root of the variance, returning the value from squared to original (and more easily comparable) units. I further divide this by the square root of total observations to find the standard error for the sampling distribution ie an estimator for the deviation of sample means with repeated sampling).

```r
sqrt(var(y)) == sd(y) # is TRUE
sd_y <- sd(y)
se_y <- sd_y/sqrt(n_y)
```

I now define the corresponding p-value for 90% confidence. Will the calculated values, I can manually approximate the 90% confidence intervals by multiplying the corresponding Z-score (aka critical value) to the alpha value for 90% confidence (ie 0.1 including both tails of the distribution, or 0.05 each tail) with the standard error, and then adding/subtracting this figure from the sample mean. In other words, I am adding/subtracting the 'margin for error' to calculate the confidence intervals.

```r
p = (1-0.9)
z_score <- qnorm(p/2, lower.tail = FALSE) # finds corresponding Z-score to the p-value 0.05.
lower_90 <- mean_y - z_score*se_y
upper_90 <- mean_y + z_score*se_y
```

Alternatively, I can use the qnorm() function to calculate the lower and upper values. This can be more precise if some figures such as z-score or standard deviation/error have been rounded.

```r
lower_90 == qnorm(p/2, mean = mean_y, sd = se_y) # TRUE confirms the values are the same
upper_90 == qnorm(1-(p/2), mean = mean_y, sd = se_y) # TRUE confirms the values are the same
```

These bounds indicate that, with 90% confidence with repeated sampling, the population mean (ie the school average IQ) is within the values 94.13 and 102.75 (rounded to two decimal places).

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 $\alpha = 0.05$.

This question requires a significance test for a mean using the t.test() function.

H0: *The average IQ score among all schools (mean_all) is NOT less than or equal to the average IQ score of our sample (mean_y).*

HA: *The average IQ score among all schools (mean_all) IS less than or equal to the sample (mean_y).*

```r
mean_all <- 100
t_test_result <- t.test(x = y, mu = mean_all, alternative = "greater")
```

The result indicates that p is greater than 0.05, failing to reject the null hypothesis. In full, this means I cannot reject the null hypothesis that the counsellor's school's average IQ score is higher than the average IQ score for all schools.

Question 2 (40 points): 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`.

```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)

str(expenditure)
summary(expenditure)
attributes(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)?

  I plot the variables accordingly. Most plots appear to indicate some positive correlation between the two variables. I highlight the correlations between Y and X1, as well as X1 and X2 as being relatively clear. Meanwhile, the (positive) relationship between Y and X3 appears relatively weak or non-existent. I use the png() function to print the plots.

  ```r
  png(filename="scatter_y_x1_x2_x3.png")
  plot(expenditure[c("Y","X1","X2","X3")],
        main = "Correlations between Y, X1, X2, and X3")
  dev.off()
  ```

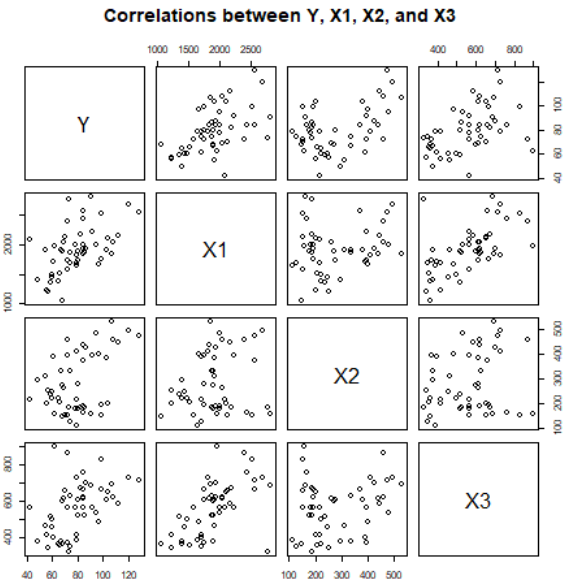

Figure 1: Relationships across variables: Expenditure on shelters/housing, Personal Income, Number of ’financially insecure’ residents, and Number of urban residents

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

  First I need to coerce the variable ”Region” from integer type to factor type, and then rename the levels to the region names for clarity.

```
expenditure$Region <- as.factor(expenditure$Region)
levels(expenditure$Region) <- c(”Northeast”, ”North Central”, ”South”, ”
    West”)
```

  Now I create a box plot showing expenditure on shelters/housing assistance across region. I then export it as a png.

```
png(filename=”boxplot_Expenditure_by_Region.png”)
Expenditure_by_Region <- boxplot(expenditure$Y ~ expenditure$Region,
          main = ”Boxplot of expenditure on shelters/housing assistance”,
          ylab = ”Dollars per capita”,
          xlab = ”Region”)
dev.off()
```

  The thick black horizontal bars indicate the mean average for each region, indicating the West region has the highest average per capita spending on shelters/housing assistance.

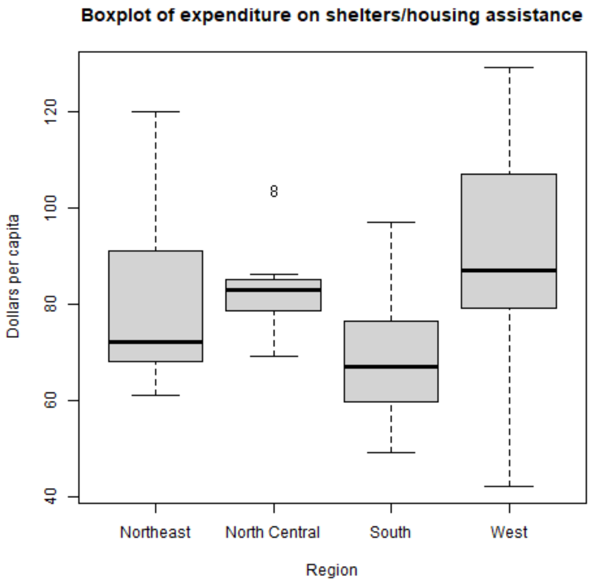

Figure 2: The Expenditure on shelters/housing assistance by states, across Regions

- 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 produce the first graph, adding a title and axis labels for clarity. The scatter plot indicates a positive relationship between state expenditure on shelters/housing assistance, and personal incomes. I print this first graph as a png.

  Using the function cor(), I find a 0.53 correlation coefficient, indicating a positive correlation and supporting with my visual assessment.

```
png(filename="expenditure_by_personal_income.png")
plot(expenditure$Y, expenditure$X1,
     main = "Expenditure on shelters/housing assistance by personal
   income",
     ylab = "State expenditure per capita",
     xlab = "Personal Income per capita")
dev.off()
cor(expenditure$Y, expenditure$X1)
```

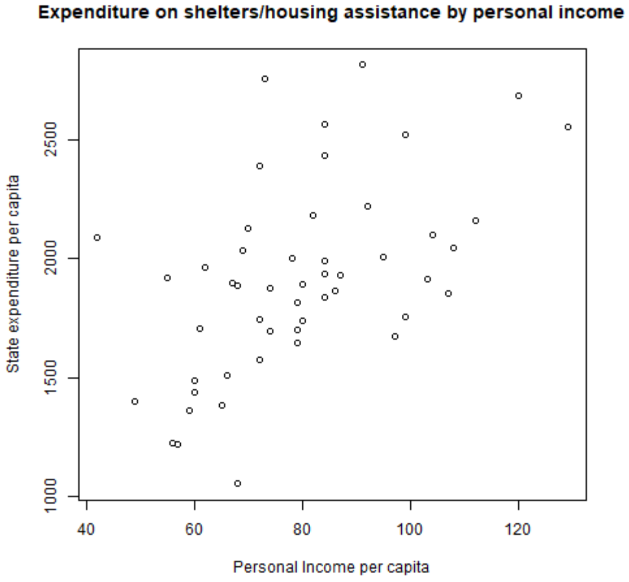

Figure 3: Scatter Plot of Expenditure by Personal Income

Now I reproduce the plot including a legend, with Region indicated by unique symbols and colours.

```
png(file="expenditure_by_personal_income_with_legend.png")
plot(expenditure$Y, expenditure$X1,
     main = "Expenditure on shelters/housing assistance by personal
    income",
     ylab = "State expenditure per capita",
     xlab = "Personal Income per capita",
     col = expenditure$Region,
     pch = c(1,2,3,4)[expenditure$Region])
legend(x = "bottomright",
        legend = levels(expenditure$Region),
        col = c(1,2,3,4),
        pch = c(1,2,3,4)
        )
dev.off()
```

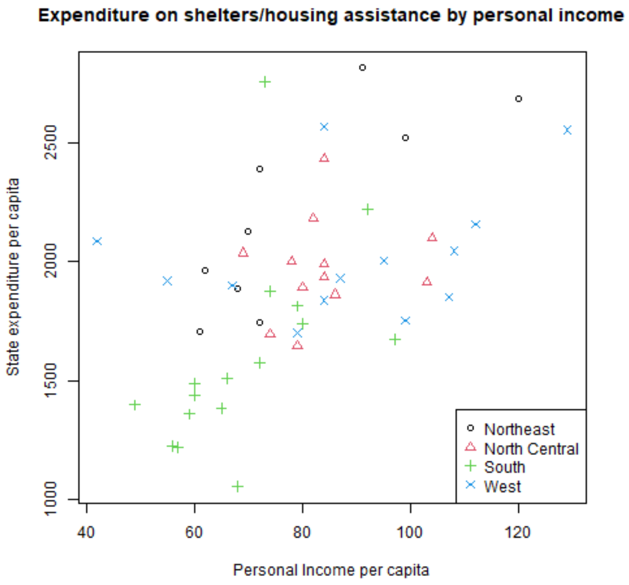

Figure 4: Scatter Plot of Expenditure by Personal Income with Legend by Region

Visual indications suggests North Central-states cluster (aka are more similar) with higher state spending AND personal incomes compared to South-states. Meanwhile, West states may have variation in state expenditure compared to personal incomes. No clear associations are indicated for Northeast-states.