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:

```
# lapply(c(”stringr”),  pkgTest)
```

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 $\alpha = 0.05$.

Answers Q1 (part 1)

To find the 90 per cent Confidence Interval, we look for the Point estimate +/- Margin of error, where margin of error is a multiple of the standard error.

What do we need?

1. Central tendency, mean - a point estimate: mean(y)

2. Variability, the standard error: sd(y) / sqrt(length(y))

Looking at the normal distribution, we see that 90 per cent of observations lie within +/-1.64 standard errors of the point estimate

```
qnorm(0.05) # value for first 5%
qnorm(0.95) # value last 5%
```

Solution method 1: The **approximate** solution for 90 per cent confidence level

```
upper_90 = mean(y)+(1.64*sd(y) / sqrt(length(y)))
lower_90 = mean(y)-(1.64*sd(y) / sqrt(length(y)))
```

90 per cent confidence level calculated as 94.14554 to 102.7345

Solution method 2: The **precise** solution for 90 per cent confidence level

```
lower_90_n <- qnorm(0.05,
mean = mean(y),
sd = (sd(y) / sqrt(length(y))))

upper_90_n <- qnorm(0.95,
mean = mean(y),
sd = (sd(y) / sqrt(length(y))))
```

On this basis, we can confirm that the confidence interval lies between 94.13283 and 102.7472

Answers Q1 (part 2)

As the sample size is less than 30 we conduct a T-test. It will be a one-sided test as we are looking to see if the student IQ in the school is higher than the population average.

The NULL Hypothesis is that there is no difference.

The Alternative Hypothesis is that the IQ is higher in the school

The significance level is set at 0.05

First step: calculate the standard error, T-statistic, and probability score:

```
standard_error <-  sd(y) / sqrt(length(y))
t_statistic <- abs((mean(y) - 100) / standard_error)
prob <- pt(t_statistic, (length(y)-1))
```

P VALUE = 0.7215383

Therefore we have evidence to accept the null hypothesis, as the p value is greater than 0.05.

Alternatively, we can find the same result using the below method:

```
t.test(y, mu = 100, conf.level = 0.05, alternative = "greater")
```

This gives the same P VALUE

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
sd(y) # Variability, standard deviation
```

- 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)?

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

- 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.

Answers Q2 (part 1)

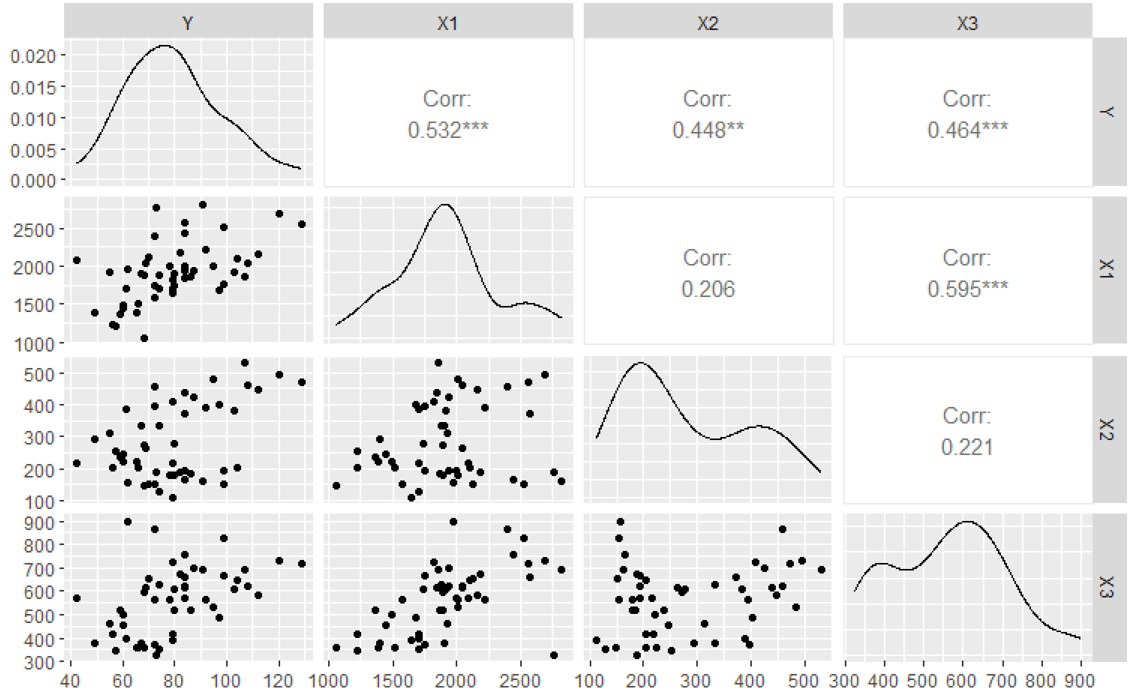

Figure 1: Scatter Plots with Correlation

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*

The strongest associations observed are between per capital expenditure on shelters/housing assistance in the state (Y) and per capita personal income in state (X1) and also the number of people living in urban areas in the state (X3).

The weakest associations are between the per capital personal income in state (X1) and number of residents per 100,00 that are ”financially secure” in state (X2), as well as between number of residents that are financially secure (X2) and the number of people residing in urban areas of the state (X3).

Answers Q2 (part 2)

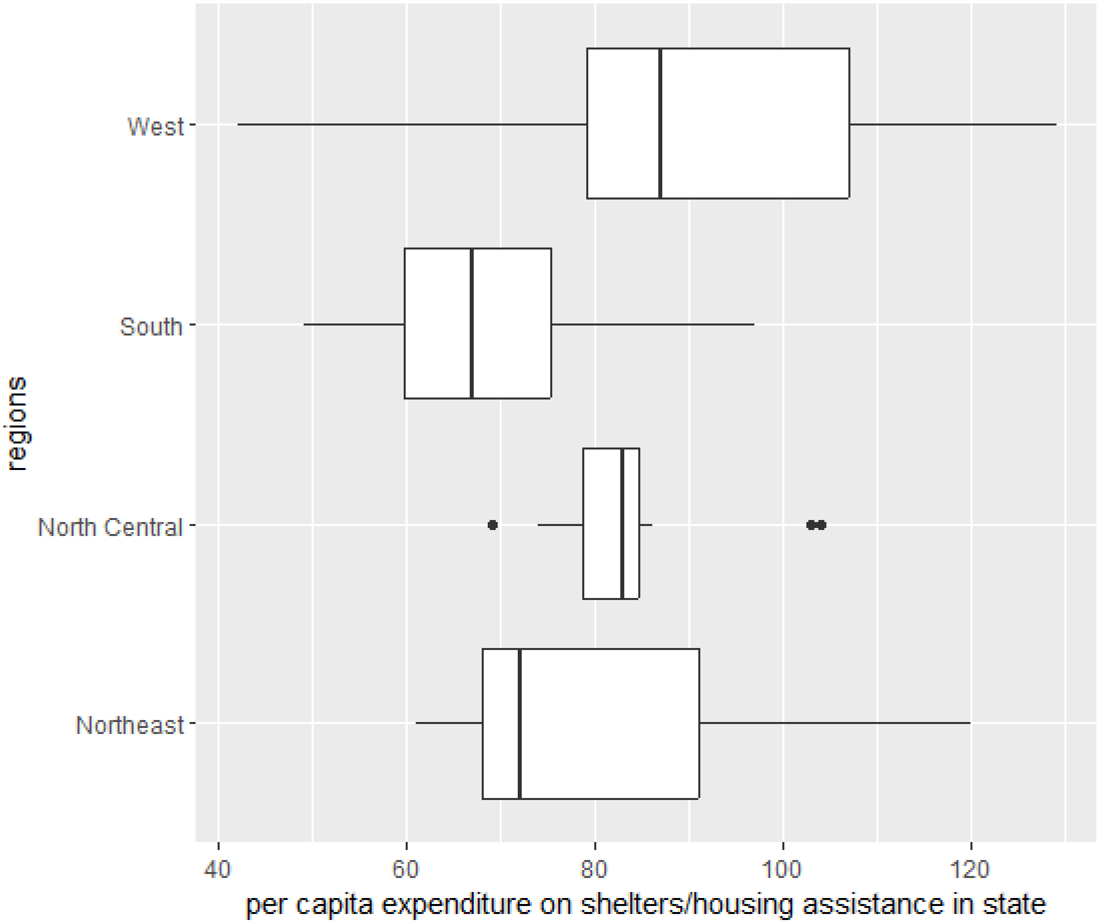

Figure 2: Comparison of State Expenditure on shelters by regions

Comment: the West region appears to have the highest expenditure on shelters / housing assistance.

Answers Q2 (part 3)

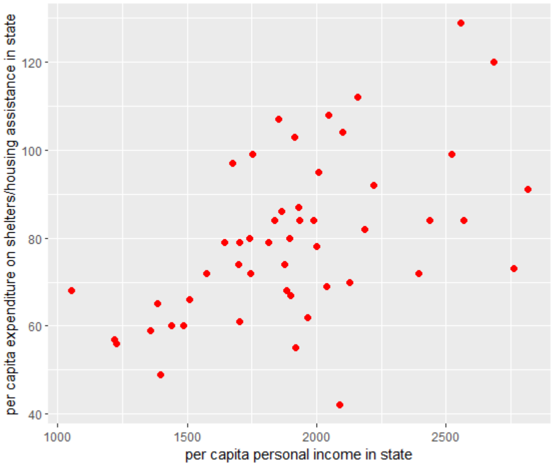

Figure 3: Comparison of State Expenditure with Personal Income

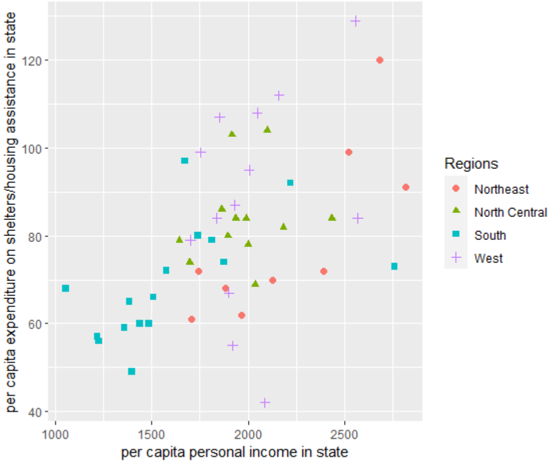

Figure 4: Association across Regions

There is a strong correlation in Figure 3 (calculated to be 0.532). In Figure 4, you can see a strong correlation in particular across the Southern region and the North Central region. The Northeast and Western regions appear to have a less strong correlation.