PS01 Response

Sania Suneeth

Applied Stats/Quant Methods 1

A school counselor was curious about the average IQ of the students in her school and took a random sample of 25 students' IQ scores. The following is the data set:

```
1
2 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 percent confidence interval for the average student IQ in the school*

```
1 mean(y)
2 sd(y)
3 s.e <-sd(y)/sqrt(length(y))
4 s.e
5 df <-length(y)-1
6 t_score <- qt(0.95,df =df)
7 upper_90_t <- mean(y)+(t_score*s.e)
8 upper_90_t
9 lower_90_t <-mean(y)-(t_score*s.e)
10 lower_90_t
```

```
> mean(y)
[1] 98.44
> sd(y)
[1] 13.09287
> s.e = sd(y)/sqrt(length(y))
> s.e
[1] 2.618575
> df = length(y)-1
> t_score = qt(0.95,df =df)
> upper_90_t = mean(y)+(t_score*s.e)
> upper_90_t
[1] 102.9201
> lower_90_t = mean(y)-(t_score*s.e)
```

```
> lower_90_t
[1] 93.95993
```

the result shows that we are 90 percent confident that the true mean value of our sample is between the range 93.95 to 102.920

2.Next, the school counselor was curious whether the average student IQ in her school is higher than the average IQ score among all the schools in the country.Using the same sample, conduct the appropriate hypothesis test with 0:05.

the major steps in conducting the hypothesis test includes

- **State the hypothesis:**
  Null hypothesis: the average IQ of the students on the school is less than or equal to 100
  Alternative Hypothesis: the average IQ of the students on the school is higher than 100

- **Choose the significance level:**
  the significance level for this study is at 0.05 level, we are 5 percent risk of concluding that there exist a difference in average where there is no actual difference

- **Calculate the t test-statistics**
  we are running a one sample one tailed t- test, to determine whether the mean of the sample is greater than or less than the known population mean, the test is one sided as we want to check if the sample mean is greater than the known population mean we first find the critical t- score then we find the lower and upper bound of the t-distribution we find the t-score of the one sample one sided t-test and finally we find the p-value

- **Determining the critical t-value and the p-value**
  here we need to compare the critical t-score with the t-test score, if the t-test score is greater than the critical t- value we can reject the null hypothesis, and the p-value tells us the probability of obtaining the result when the null hypothesis is true

```r
## finding t-test
t_score_critical
df2 <-length(y)-1
t_score_critical <-qt(0.95,df =df2)
t_score_critical

upper_bound_95<- mean(y)+(t_score_critical+s.e)
upper_bound_95
lower_bound_95 <- mean(y) - (t_score_critical * s.e)
lower_bound_95

population_mean<-100
t_score<- (mean(y)-population_mean)/s.e
```

```r
t_score

p_value <- pt(abs(t_score), df2, lower.tail = FALSE)
p_value

## now running t- test
t.test(y,mu = 100, conf.level = 0.95,alternative = 'greater')
```

```
t_score_critical = qt(0.95,df =df2)
> t_score_critical
[1] 1.710882
upper_bound_95<- mean(y)+(t_score_critical+s.e)
> upper_bound_95
[1] 102.7695
> lower_bound_95 <- mean(y) - (t_score_critical * s.e)
> lower_bound_95
[1] 93.95993
t_score<- (mean(y)-population_mean)/s.e
> t_score
[1] -0.5957439
p_value <- pt(abs(t_score), df2, lower.tail = FALSE)
> p_value
[1] 0.2784617
\newline

 One Sample t-test

data:  y
t = -0.59574, df = 24, p-value = 0.7215
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
93.95993      Inf
sample estimates:
mean of x
98.44
```

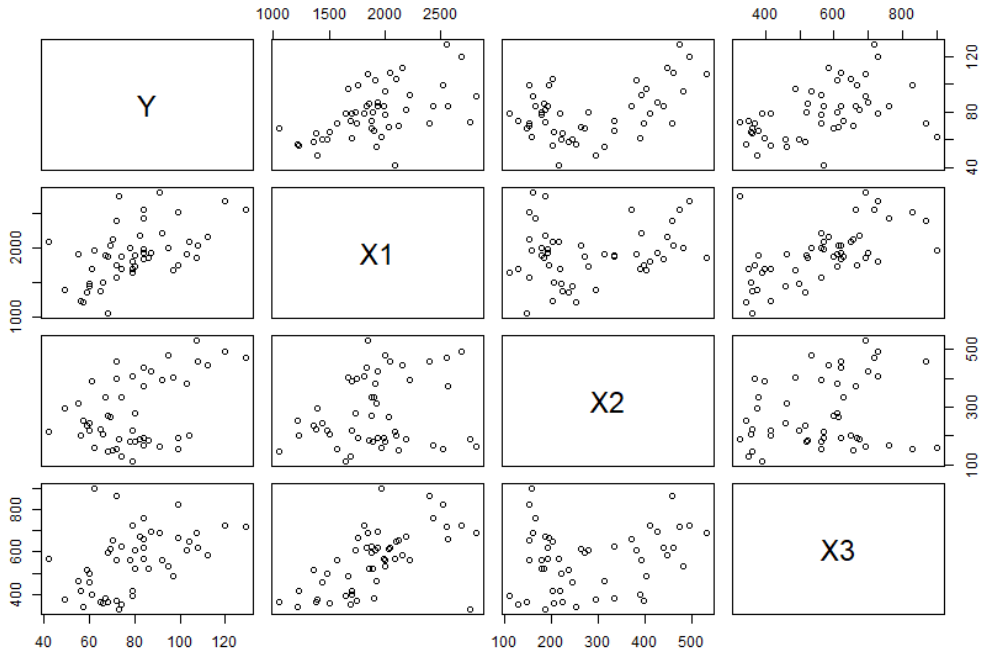
- **Interpreting the result:**
  based on the result the p value is greater than the significance level , indicating that we cannot reject the null hypothesis at 0.05 significance level, moreover the confidence level of 95percent suggest that the true mean value of the student population falls between 93.95 and infinity, also the t- scores is negative indicating the true mean of our sample is less than the population mean

Question 2

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?

```
str(expenditure)
plot = pairs(expenditure[,c('Y','X1','X2','X3')])
```

Figure 1: relationship between Y,X1,X2,X3



a pair plot graph is plotted to visualize the relationships among the variables(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 ”

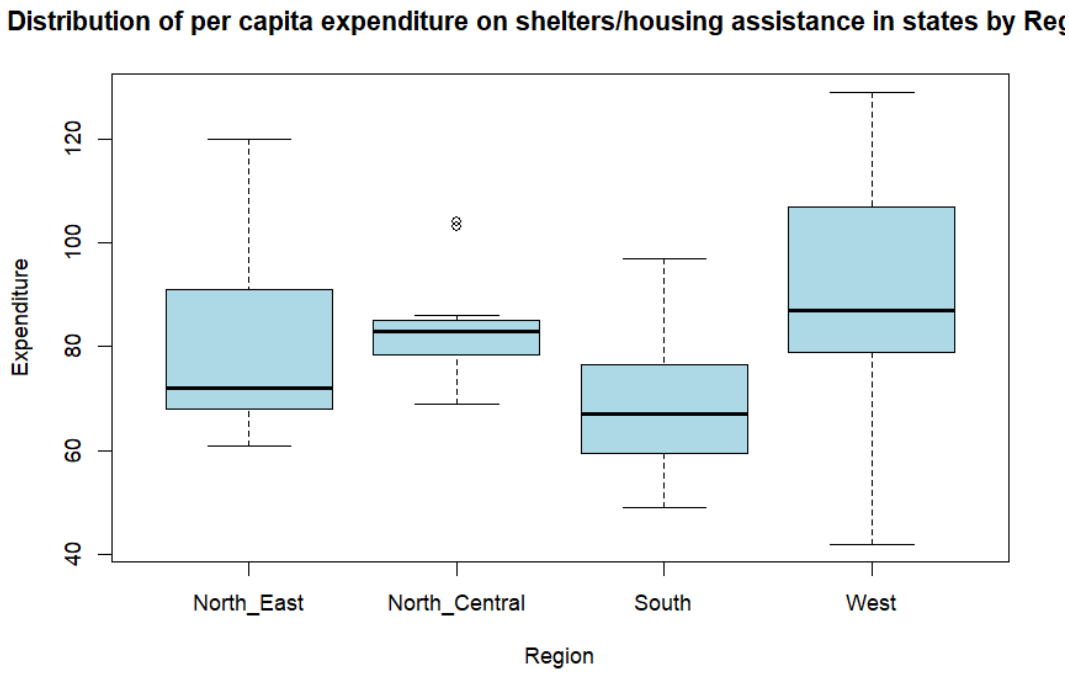
financially insecure” in state,X3: Number of people per thousand residing in urban areas in state)

The pair plot reveals that Y and X1 have a positive correlation, indicating a direct relationship where Y increases as X1 increases. In contrast, the relationship between Y and X2 appears weak or non-existent, and a similar lack of correlation is observed between Y and X3. Additionally, there is no strong correlation between X1 and X2, as well as between X2 and X3. However, X1 and X3 exhibit a negative correlation, although this relationship is quite weak, suggesting that these variables do not strongly influence one another.

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

```
boxplot(expenditure$Y ~ expenditure$Region,names = c('North_East','North_
    Central','South','West'),
        main ='Distribution of per capita expenditure on shelters/housing
    assistance in states by Region',
        xlab = 'Region', ylab = 'Expenditure',col='lightblue')
```

Figure 2: Distribution of per capita expenditure on shelters/housing assistance in states by Region



here note that the median of the box plot of the west region is of approx 80 units, where we can conclude that 50 percent of the social welfare expenditure falls within the interquartile range of 110 and 80, with the maximum expenditure value going beyond 120 units while the lowest reaching up-to approx 40 units, though west having relatively higher expenditure values with the long whiskers, based on the interquartile range and the spread of the data,we can infer that west having higher expenditure value over other regions on average.

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 diffrent types of symbols and colors.

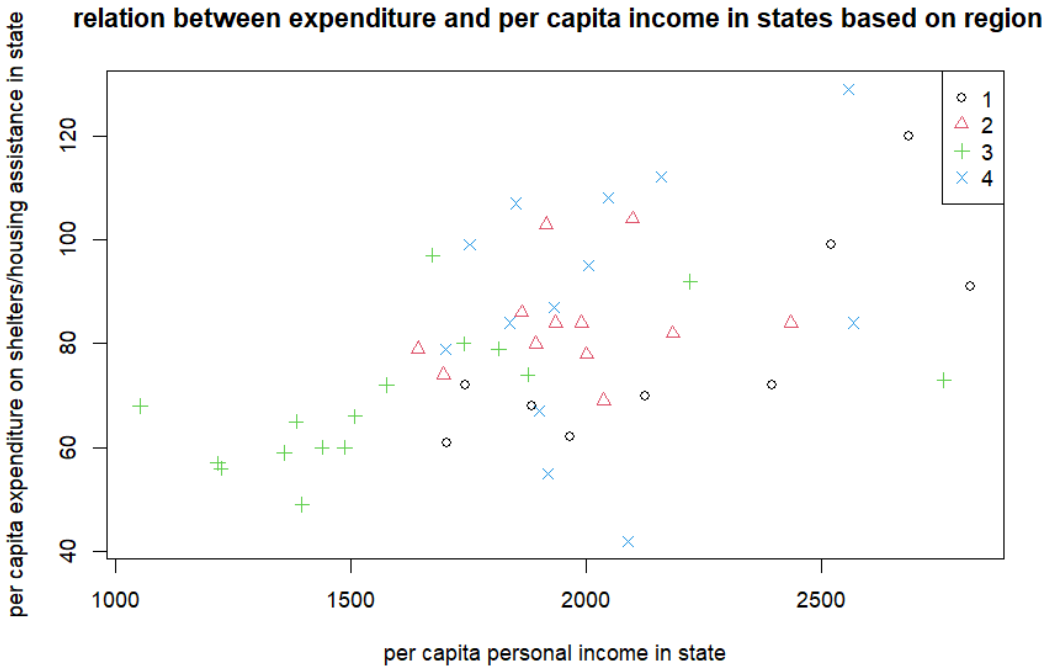
```
plot(expenditure$X1,expenditure$Y,col=as.factor(expenditure$Region),pch=as.
    numeric(as.factor(expenditure$Region)),
```

```
     main = ' Relationship between expenditure and per capita income in states
    based on region',
     xlab = 'per capita personal income in state',ylab ='per capita
   expenditure on shelters/housing assistance in state')

legend('topright',legend = levels(as.factor(expenditure$Region)),
       col = 1:length(levels(as.factor(expenditure$Region))),
       pch = 1:length(levels(as.factor(expenditure$Region))))

# the scatterplot shows evidently that there exists a positive association
   between personal income76
```

Figure 3: relationship between per capita expenditure on shelters/housing assistance in state and per capita personal income in state



the scatterplot shows evidently that there exists a positive association between per capita personal income and housing assistance expenditure, the strength of the association varies across different region, the West shows most diversity in terms of income and expenditure indicating that the states in the region are affluent which directly influences the housing expenditure, while the South shows the least diversity over both the variables,meanwhile the Northeast and North Central shows a narrower range of display of data points majorly taking the center portion of the graph, yet some states in the region show higher expenditure but within less income limits compared to the West. Overall we can assume that the social welfare expenditures in the US is influenced by the per capita personal income of the people residing in a particular region.