Problem Set 1

Applied Stats/Quant Methods 1

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:

```
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 1: Education

1. *Solution for Part 1*

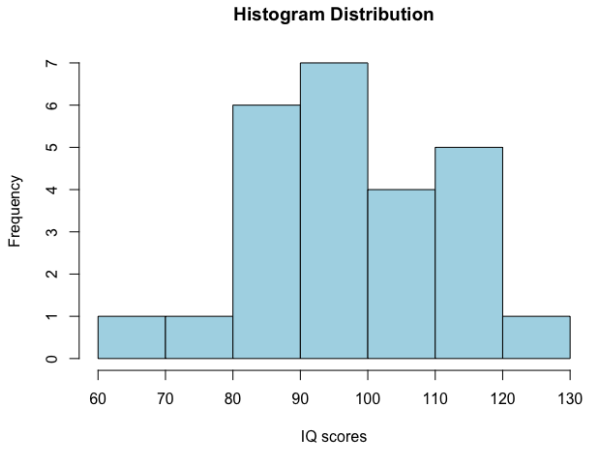

Figure 1: Histogram plot

We use the formula for computing the confidence interval

$$CI = (\bar{x}–EBM, \bar{x} + EBM)$$

where EBM =

$$EBM = t_(\left(\frac{\alpha}{2}, df = 24\right)) \times \left(\frac{s}{sqrt(n)}\right)$$

We will use the *t-distribution* in the above case as we have:

(a) n <30

(b) standard deviation of the population is unknown

(c) the data is approximately normally distributed as shown in Figure 1

To find the critical t value, we will compute alpha/2 and degrees of freedom from the t distribution table.

alpha = 1 - Confidence Interval = 1 - 0.90

alpha/2 = 0.10/2 = 0.05

degrees of freedom = n - 1 = 25 - 1 = 24

Using the above values of alpha and degrees of freedom, the t-value becomes: 1.71 (from the t-distribution table, calculated in R)

Given the standard deviation $s = 13.09$ (computed from the data) and n = 25, $\bar{x} = 98.44$, the confidence intervals become

where,

upper bound for the mean =

$$98.44 + 1.71 \times \left(\frac{13.09}{\sqrt{25}}\right)$$

= **102.92**
lower bound for the mean =

$$98.44 - 1.71 \times \left(\frac{13.09}{\sqrt{25}}\right)$$

= **93.95**

Therefore, we can say with **90% confidence** that the mean lies within the range of **93.95 and 102.92**

Code for computing the above:

```
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)
3
4 ### histogram to check if normally distributed
5 hist(y, col = "lightblue", main = "Histogram Distribution", xlab = "IQ
    scores")
6
7 ## finding input parameters for computing confidence intervals
8
9 length_y <- length(y)
10 sample_mean <- mean(y)
11 sample_std_dev <- sd(y)
12 alpha <- 1 - 0.90
13 alpha_by_2 <- alpha/2 ## as confidence interval is a two tail test
14
15 ## calculate t critical value and error bounds
16
17 t <- qt(1 - (alpha_by_2), df = length_y - 1)
18 error_bound <- t * (sample_std_dev/sqrt(length_y))
19
20 ## upper and loewr bounds for confidence intervals
21
22 upper_error_bound <- sample_mean + error_bound
23 lower_error_bound <- sample_mean - error_bound
```

2. Solution for part 2

For this problem, our null hypothesis H_0 and alternative hypothesis H_a are as follows:

- H_0: $\mu \leq 100$
- H_a: μ>100

Since, μ falls on only one side, we are going to use a one tail test, or in this case a right sided one-tail test. Additionally, since the below two conditions are met we are going to use the t test to reject or accept the null hypothesis.

(a) n <30

(b) standard deviation of the population is unknown

We are going to perform the hypothesis test using the following steps:

- Step 1: Compute the t critical value
  With $alpha = 0.05$ and degrees of freedom = 24, the t critical value = 1.71.
- Step 2: Compute the t-statistic value.

  To compute this, we use the formula:

  $$\frac{\bar{x} - \mu_0}{s/\sqrt{(n)}}$$

  where, sample mean $\bar{x} = 98.44$,
  sample std dev, $s = 13.09$,
  hypothesized mean, $\mu_0 = 100$ and
  length of y, $n = 25$

  the t statistic value therefore equals = -0.595
- Step 3: Compare the two values to accept or reject the hypothesis.

  Since the t critical value is greater than the t statistic value, we do not reject the null hypothesis. Therefore, we cannot say with 95% confidence that our average student IQ is higher than 100.

Code for above computations:

```
hypothesized_mean = 100
alpha_2 = 0.05
t_value = (sample_mean - hypothesized_mean) / (sample_std_dev / sqrt(
    length_y))

# we use 1 - alpha as this is a one-tail test

t_critical = qt(1 - alpha_2, df = length_y - 1)

```

```
9  if (t_value > t_critical) {
10   print("Reject the null hypothesis in favor of the alternative
       hypothesis")
11 } else {
12   print("Do not reject the null hypothesis.")
13         }
```

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`.

- 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.

Solution 1: Political Economy

1. Part 1

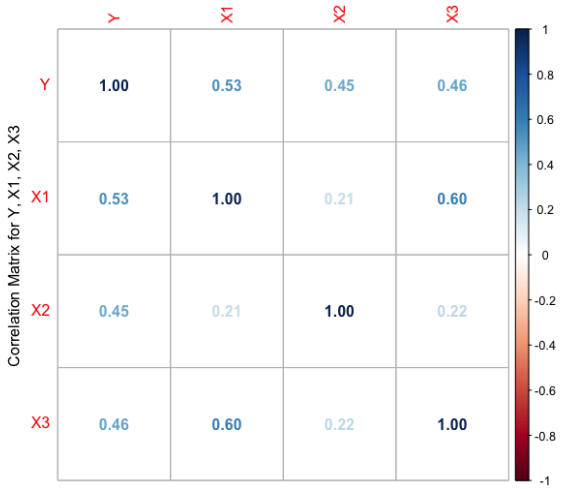

Figure 2: Correlation Plot bw Variables

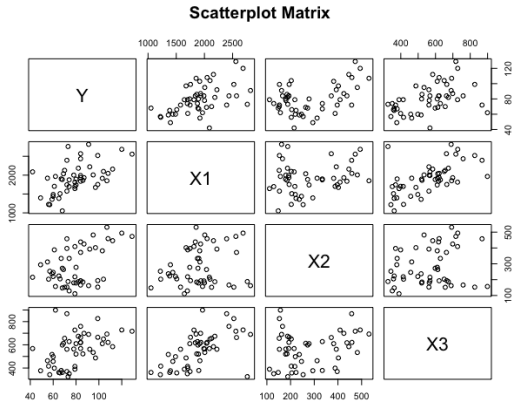

Figure 3: Scatter Plot Matrix

The most obvious correlation a variable has with itself, and as we can see from the correlation matrix table (figure 2), each feature variable is highly correlated with itself, indicating a 1:1 linear relationship.

Moving on to relationships between other variables in figure 2, the first column shows that variable Y (per capita expenditure on shelters/housing in state) has a relatively strong linear relationship with all the features X1, X2 and X3, Y having the strongest relationship with X1.

In the second row, for X1, the variable is highly correlated with X3, but has a very weak relationship (closer to zero) with X2.

Additionally, in the same row we can also see that X1 and X3 are the most highly correlated variables in the entire matrix with the highest value of 0.60. This indicates that we have a relationship worth investigating between X1 (per capital personal income in state) with X3 (Number of people/1000 residing in urban areas in state).

The scatter plots between all variables (figure 3) also show similar patterns to the ones we have in the correlation table. For the relationships that are not strongly dependent (i.e X3 and X2), the data points are scattered, indicating either a non-linear or a non-significant relationship between the two. All other variables that have a decent amount of linear dependence, have data points that cluster alongside (or can be approximated by) a linear line.

2. Part 2

To plot the relationship between Y and Region, I used two different plots. The first figure shows a scatterplot matrix between the variables whereas the second figure shows a boxplot (to be able to find the highest per capita expenditure (on average))

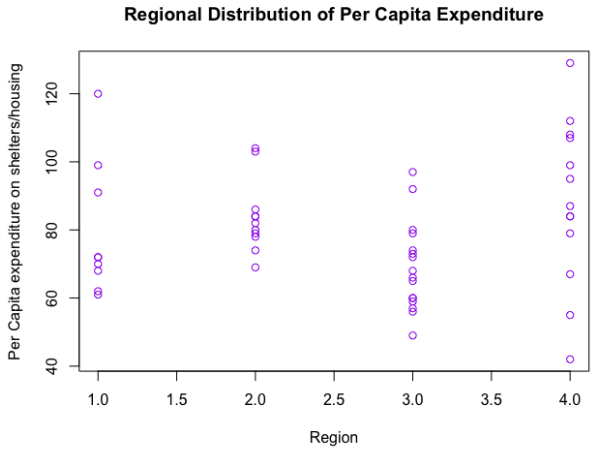

Figure 4: Scatter Plot Matrix of Regional Distribution of Expenditures

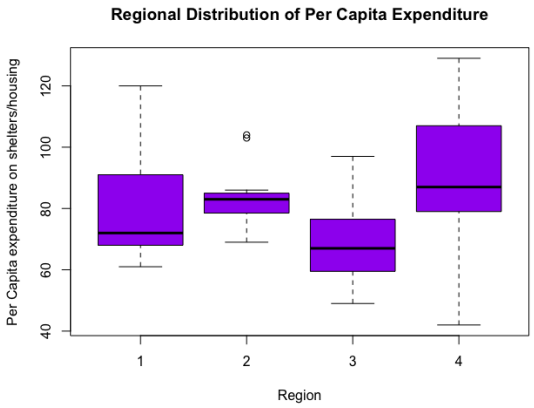

Figure 5: Box Plot Matrix of Regional Distribution of Expenditures

On average, region 4 (West) has the highest per capital expenditure on housing assistance. This information is extracted from the fact that Region 4 has the highest median value among all other regions.

3. Part 3

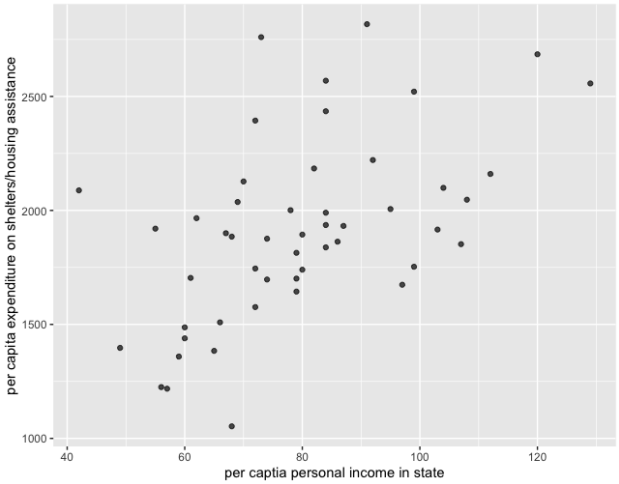

Figure 6: Scatter Plot of Y and X1

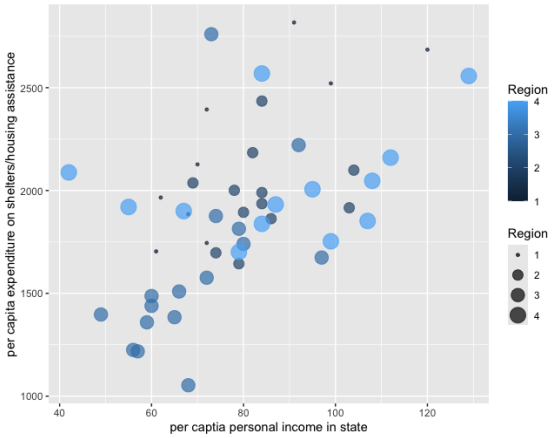

Figure 7: Bubble Graph for Y and X1

As the scatterplot matrix (figure 5) shows Y and X1 have a cluster of data points that can be approximated by a linear line. More points are clustered towards the center diagnally. Its not an ideal 1:1 linear relationship, as we saw in part 1 as well where the two had a correlation factor of 0.53, but it still does appear to be significant.

Figure 6 shows the same plot enhanced with a bubble plot where the Region determines the color and the size of the data points.