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:

1. Find a 90% confidence interval for the average student IQ in the school.

```r
mean_y <- mean(y) # Point Estimate
std_error <- sd(y)/sqrt(length(y)) # Standard Error

# T-Score (since n<30, use T-score instead of Z-score)
t_score <- qt(0.95, df=length(y)-1)

# Upper Bound
upper_t <- mean(y)+(t_score)*(std_error)
upper_t
```

```r

# Lower Bound
lower_t <- mean(y)-(t_score)*(std_error)
lower_t
```

```
90% CI = [93.95,102.92]
```

A 90% confidence interval indicates that if we were to take many random samples from the population and calculate a confidence interval for each of them, we would expect the true population value to be contained in 90% of the confidence intervals.

In the case above, we can say that we are 90% certain that the true average IQ for students in the counselor's school is between 93.95 and 102.92.

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 $\alpha = 0.05$.

```r
# Step 1: Assumptions
class(y)
# Data are quantitative and randomly sampled
length(y)
# N < 30, which is below the CLT threshold.
# Hence, assume that data are approximately t-distributed

# Step 2: Hypotheses
# H0: The average student IQ  in the counselor's school is less than or equal to 100
# H1: The average student IQ in the counselor's school is greater than 100

# Step 3: Calculating Test Statistic
t_score <- (mean_y-100)/(std_error)
t_score

# Step 4: Calculate P-value
p_value <- (1-pt(manual_t, length(y)-1))
p_value
```

```
P-Value = 0.72
```

Conclusion: The P-value is greater than 0.05, therefore there is not enough evidence to reject the null hypothesis that the average student IQ in the counselor's class is less than or equal to 100 when alpha = 0.05. We therefore conclude that the average student IQ in the counselor's class 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`.

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

# Summary of Data Set
summary(expenditure)
```

```
       Y               X1            X2              X3
 Min.   : 42.00   Min.   :1053   Min.   :111.0   Min.   :326.0
 1st Qu.: 67.25   1st Qu.:1698   1st Qu.:187.2   1st Qu.:426.2
 Median : 79.00   Median :1897   Median :241.5   Median :568.0
 Mean   : 79.54   Mean   :1912   Mean   :281.8   Mean   :561.7
 3rd Qu.: 90.00   3rd Qu.:2096   3rd Qu.:391.8   3rd Qu.:661.2
 Max.   :129.00   Max.   :2817   Max.   :531.0   Max.   :899.0
```

- 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:

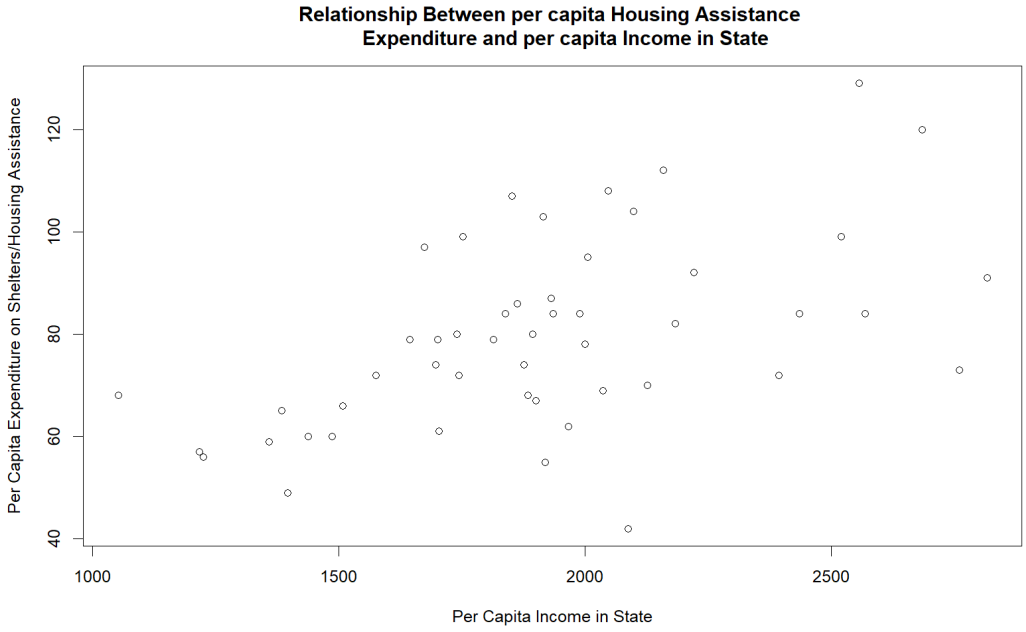

Figure 1 Analysis: There is a strong positive linear relationship between per capita income in state and per capita expenditure on shelter/housing assistance in state. Higher income per capita in state tends to be associated with higher per capita expenditure on housing assistance and vice versa.

Figure 2:

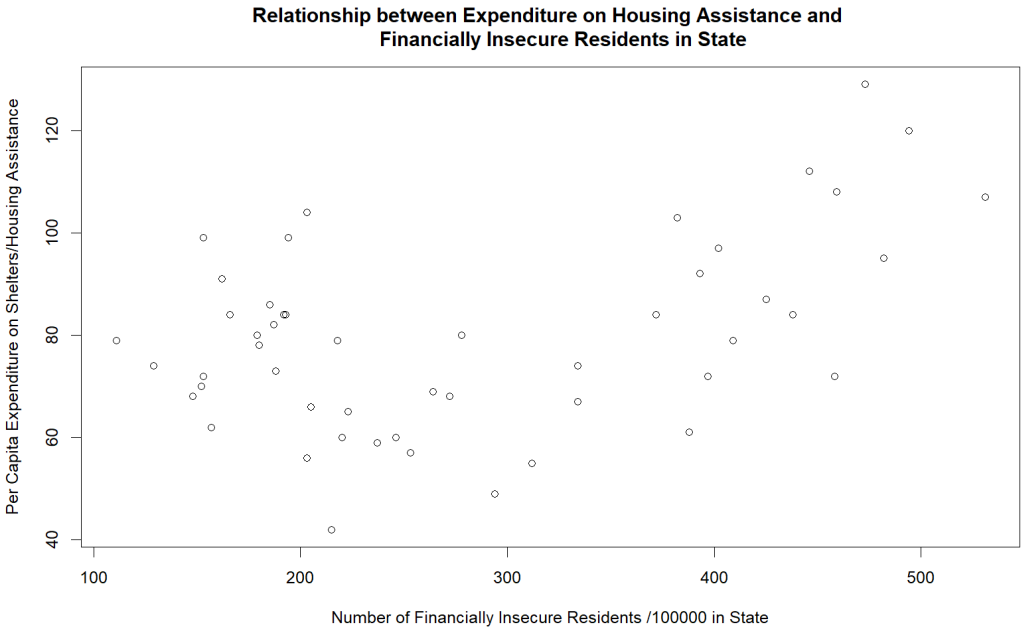

Figure 2 Analysis: There is no linear relationship between the number of financially insecure residents per 100000 in state and per capita expenditure on housing assistance.

Figure 3:

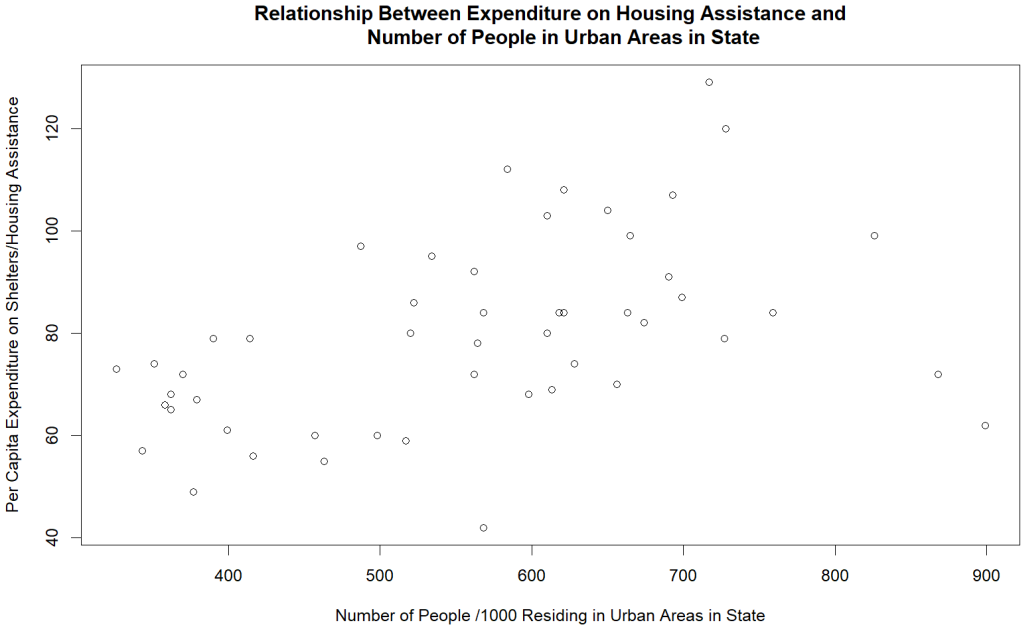

Figure 3 Analysis: There is a weak positive linear relationship between the number of people residing in urban areas and per capita expenditure on housing assistance in state.

Figure 4:

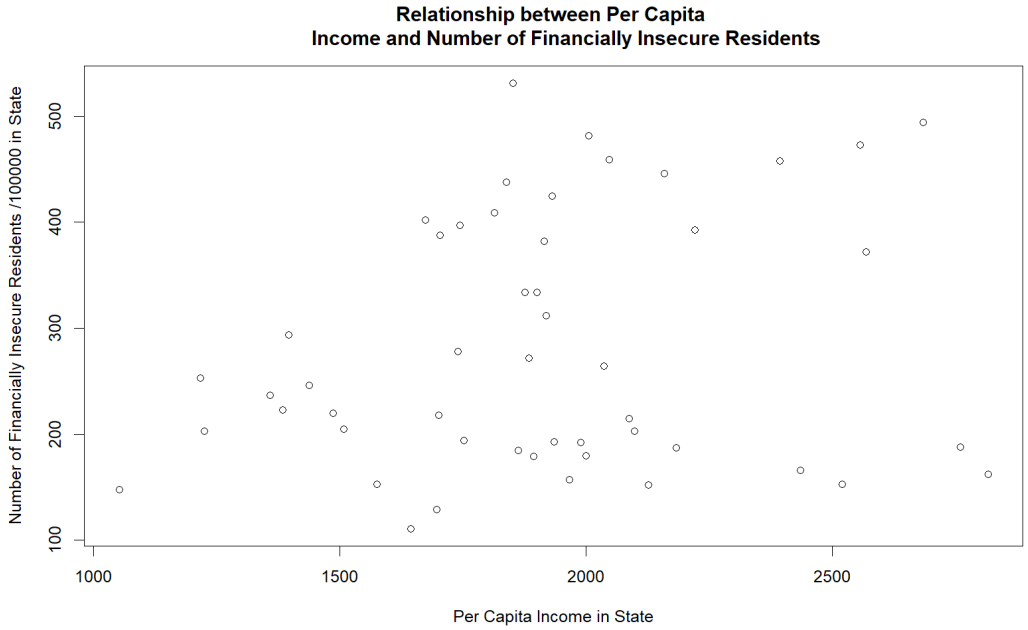

Figure 4 Analysis: There is no correlation between per capita income and the number of financially insecure residents per 100000 in state.

Figure 5:

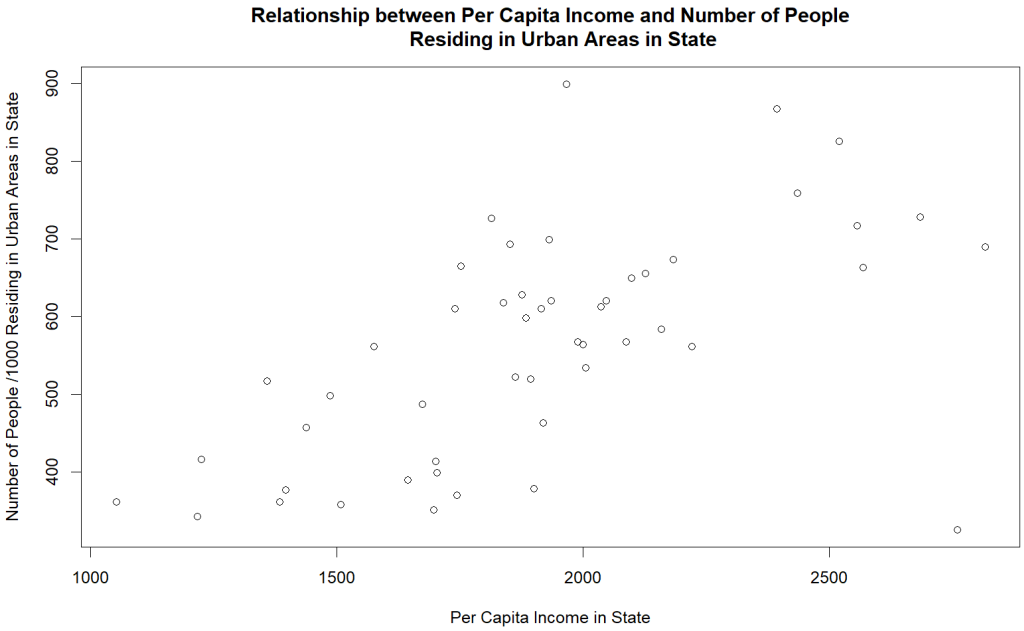

Figure 5 Analysis: There is a strong positive linear correlation between per capita income in state and the number of people per 1000 residing in urban areas in state. There is one outlier point where a relatively high per capita income state corresponds to a low number of people living in urban areas in the state.

Figure 6:

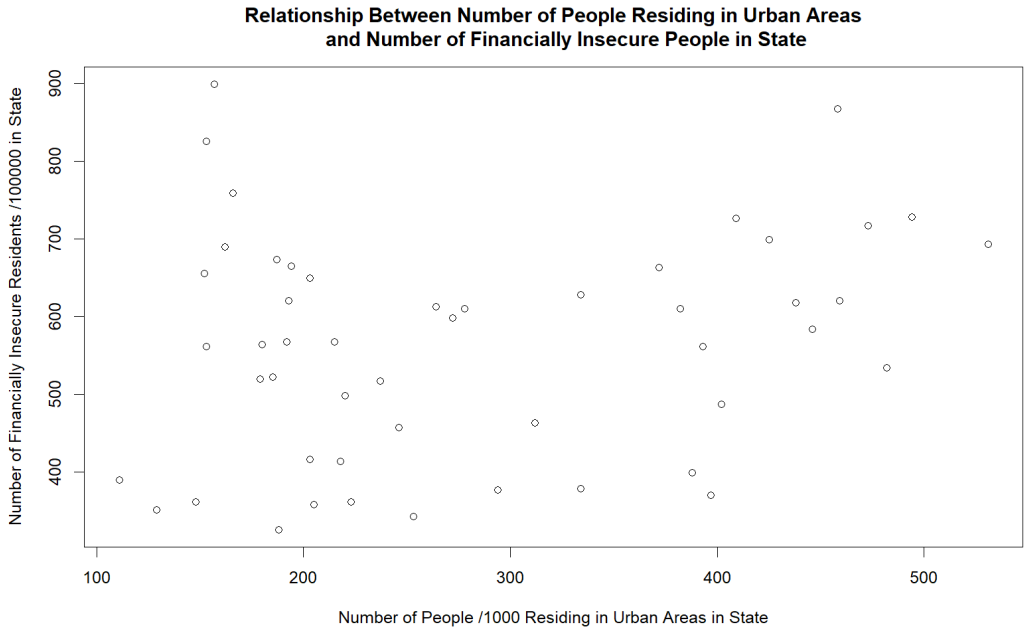

Figure 6 Analysis: There is no correlation between the number of people in urban areas (per 1000) and the number of financially insecure residents (per 100000) in state.

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

Figure 7:

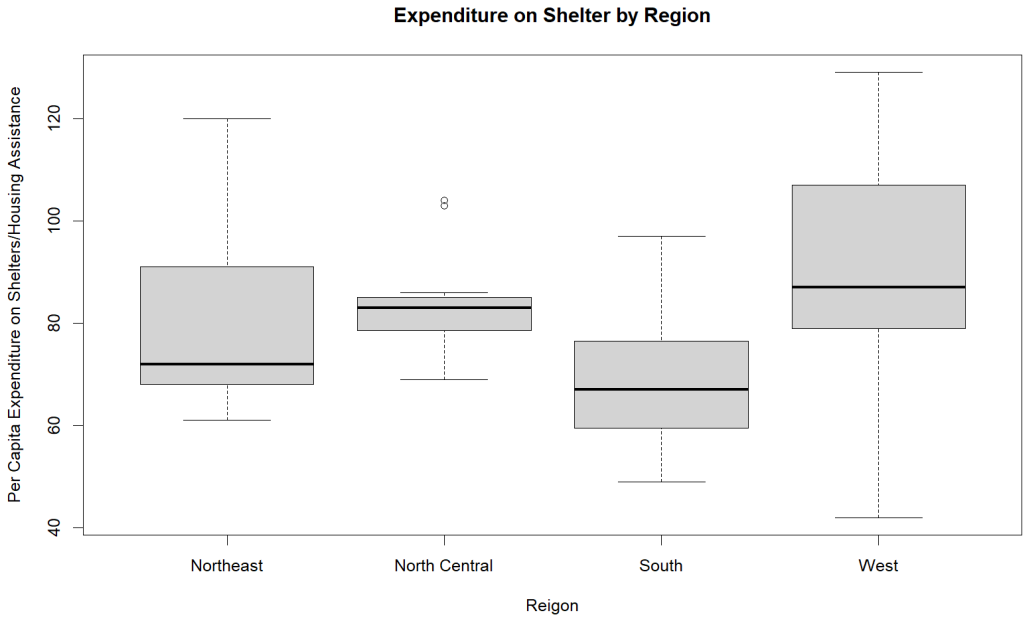

Figure 7 Analysis: On average, the West region has the highest per capita spending on housing assistance relative to the other three regions in the sample.

- 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.

Figure 8:

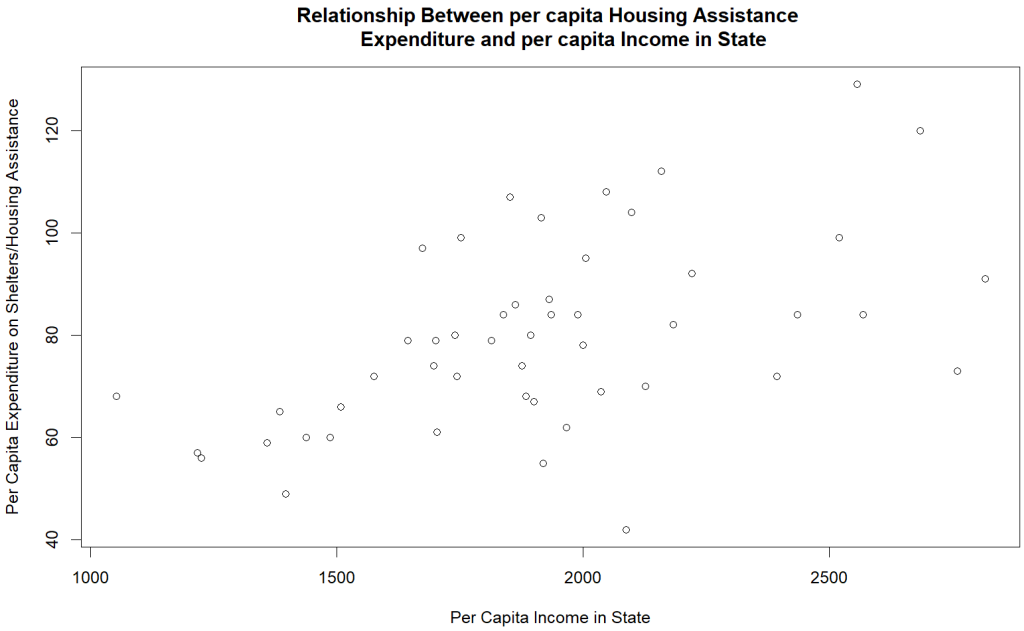

Figure 8 Analysis: Refer to Figure 1.1

Figure 9:

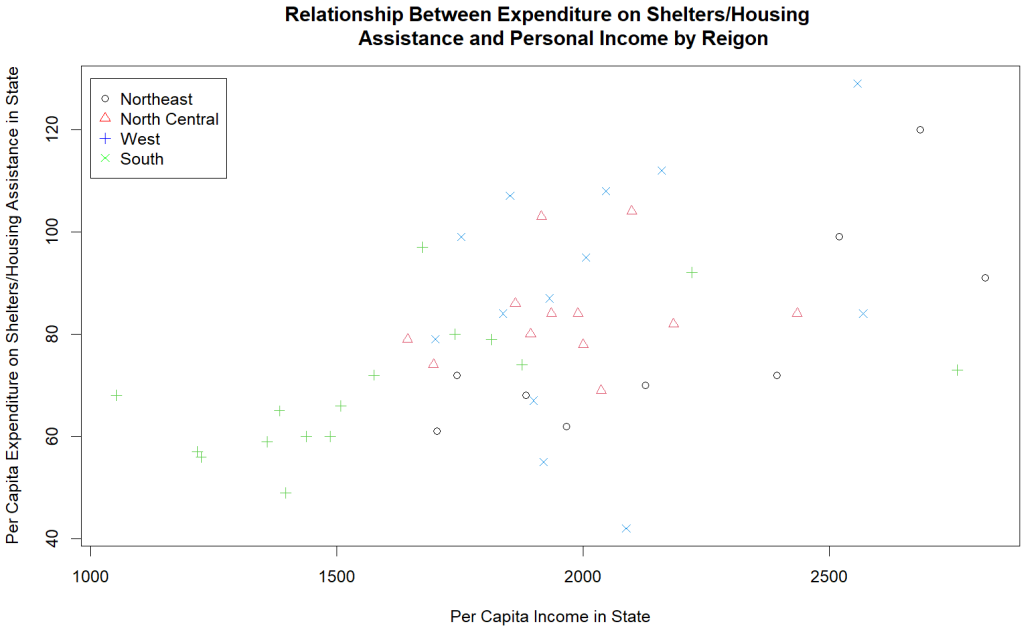

Figure 9 Analysis: Figure 9 is consistent with Figure 7 in that it confirms that, on average, both per capita expenditure on housing assistance and per capita personal income tends to

be higher in the West relative to the other regions in the data set. The South appears to have the lowest average spending on housing assistance and lowest income per capita relative to the other three regions. Figure 9 also re-emphasizes the positive linear correlation between per capita personal income in state, and per capita expenditure on housing assistance in state.