Problem Set 1

Applied Stats/Quant Methods 1

Due: October 1, 2021

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 in `.pdf` form.
- This problem set is due before 8:00 on Friday October 1, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 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.

   First, sort the data and calculate mean, standard deviation and sample size in RStudio. Code is as follows...

- Sort the data: sort(y) = 69 80 82 86 87 87 89 90 94 95 95 97 98 98 100 104 105 108 110 111 111 112 113 114 126

- Find the sample mean: mean(y) = 98.44

- Find the standard deviation: sd(y) = 13.09287

- Find the sample size: length(y) = 25

Then, find the z score using qnorm function. ANS equals 1.644854 and -1.644854 - qnorm(.95) = 1.644854

- qnorm(.05) = -1.644854

Use function to find 90 CI.

- Confidence interval function = 1.644854*13.09287/sqrt(25)

- Confidence interval = 4.307172

Add and subtract 4.307172 from the sample mean to find 90 Confidence Interval.

- Sample mean (98.44) minus the confidence interval (4.307172) = 94.13283

- Sample mean (98.44) plus the confidence interval (4.307172) = 102.7472

ANS!!!

90 Confidence Interval for the population average IQ score is between 94.13283 and 102.7472 range.

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 $\alpha = 0.05$.

   Conduct Hypothesis Test on whether School/Population average IQ is higher than national average (IQ 100)

   5 steps: 1) Data assumptions, 2) Null and alt hypothesis, 3) Calculate T-stat, 4) Calculate P-value, 5) Conclusion

   1) Data Assumptions:

   - mean = 98.44,

   - sd = 13.09287,

   - sample size = 25,

   - standord error = 2.618575,

   - level of significance = 0.05.

   2) Formulate null and alt hypotheses.

   - Null Hypothesis: The avg IQ score of the school will be greater than or equal to 100.

   - Alt Hypothesis: The avg IQ score of the school will be less than 100.

3) Calculate test statistic by subtracting avg IQ nationally from sample avg IQ. ANS = -1.56

- national IQ avg = 100
- Test stat = sample mean minus national IQ avg = 98.44 minus 100
- Test stat = -1.56

Calculate P-value using pnorm function.

- p value = 2*pnorm(-abs(-1.56))
- p value = 0.1187599
- p value (0.1187599) is greater than 0.05

P value is greater than level of significance. Null hypothesis is therefore rejected.

Question 2 (50 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`.

```
qnorm(.05)
```

- 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.

  PART 1: PLOT RELATIONSHIP BETWEEN Y, X1, X2, AND X3

  Explore data and plot relationships between Y, X1, X2 and X3

  Y = output varibale.

  X1, 2 and 3 = input variables

  Plot correlation between Y and X1. Subset function in R.

  - expenditure[, 2:3]

  - plot() function used to plot subsetted data

  No significant relationship or correlation between Y (housing exp) and X1 (personal income). Data very dispersed.

  Plot correlation between Y and X2 in R.

- expenditure[, c(2, 4)]

- plot function used to plot subsetted data

No significant relationship or correlation between Y (housing exp) and X2 (financial insecurity). Data quite dispersed.

Plot correlation between Y and X3

- expenditure[, c(2, 5)]

- plot() function used to plot subsetted data

No significant relationship or correlation between Y (housing exp) and X2 (urban residents). Data quite dispersed.

PART 2 PLOT RELATIONSHIP BETWEEN Y and REGION

Which region has highest per capita expenditure on housing assistance?

- expenditure[, c(2, 6)] subsets expenditure dataframe into regions and expenditure on housing

- plot() function used to plot subsetted data

Subset subsetted dataset on expenditure and regions into sub-regions to clarify data more

- northeastexp = regionexp[1:9, 1]

- northcentralexp = regionexp[10:21, 1]

- southexp = regionexp[22:37, 1]

- northexp = regionexp[38:50, 1]

Find the mean expenditure on housing assiatnce for each region. Calculate which is highest.

- mean(northeastexp)

- mean(northcentralexp)

- mean(southexp)

- mean(northexp)

ANS!!!!

The North spent most on housing assistance at 88.30769, followed by North Central , Northeast and lastly the South

PART 3: PLOT AND COMMENT UPON RELATIONSHIP BETWEEN Y AND X1

Include new variable 'Region'. Use colors and symbols for each region

- regincomeexp = expenditure[, c(2, 3, 6)]

- plot(regincomeexp)

Rename Region Values from Integers 1:4 to 'Northeast', 'North-Central', etc

- regincomeexp["Region"] [regincomeexp["Region"] == 1] ¡- "Northeast"

- regincomeexp["Region"] [regincomeexp["Region"] == 2] ¡- "North-Central"

- regincomeexp["Region"] [regincomeexp["Region"] == 3] ¡- "South"

- regincomeexp["Region"] [regincomeexp["Region"] == 4] ¡- "West"

regincomeexp is now a dataset containing regions, personal income and expenditure on housing assistance.

Regions are now named rather than using 1, 2, 3, 4 to represent region names.

This will make the final plot with color more clear

- plot(regincomeexp)