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:

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 $\alpha = 0.05$.

Problem 1 - My Answer (R code)

```r
######################
# Problem 1 = My Answer
######################

# Because the sample or n < 30 a t-distribution can be used as opposed to a normal distribution
#  t-distribution is a type of probability distribution. used where the sample size is small.

# The sample size is 25
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)

# get the sample mean ( x bar) of y which is 98.4
n <- length(y)
# sample mean
mean(y)
# standard deviation
sd(y)
(sd(y)/sqrt(n))
# CI of 90%
# From tables look for Degree of Freedom or DF of 24 and alpha level of 0.05
# this gives 1.71 or a t-score of 1.71
# 1. 25 - 1 = 24
# 2. 1 - .90 = .05 or the alpha level
# use alpha level of 0.05
c90 <- qt(.05, 24, lower.tail = FALSE)
c90
# CI tells us to take the mean of 98.44 and plus or minus 4.4778
lower_lvl <- mean(y) - c90*(sd(y)/sqrt(n))
upper_lvl <- mean(y) + c90*(sd(y)/sqrt(n))
# get the s or standard deviation of y which is 13.09287
#sd(y)
# Using a 90% confidence level and need to find the confidence interval
# Confidence Level or CI is the margin of error and written as
# CI is to do with how reliable the estimation is
# 1.71 * 13.09287 / square root of 25 = 4.4778
# 98.44 - 4.778 = 93.96 is the lower level
# 98.44 + 4.778 = 102.92 is the upper level
c(lower_lvl, upper_lvl)


# t test can be used for small samples
t.test(y)
```

Problem 1 - My Answer - R console output

```
> # The sample size is 25
> 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)
>
> # get the sample mean ( x bar) of y which is 98.4
> n <- length(y)
> # sample mean
> mean(y)
[1] 98.44
> # standard deviation
> sd(y)
[1] 13.09287
> (sd(y)/sqrt(n))
[1] 2.618575
> # CI of 90%
> # From tables look for Degree of Freedom or DF of 24 and alpha level of 0.05
> # this gives 1.71 or a t-score of 1.71
> # 1. 25 - 1 = 24
> # 2. 1 - .90 = .05 or the alpha level
> # use alpha level of 0.05
> c90 <- qt(.05, 24, lower.tail = FALSE)
> c90
[1] 1.710882
> # CI tells us to take the mean of 98.44 and plus or minus 4.4778
> lower_lvl <- mean(y) - c90*(sd(y)/sqrt(n))
> upper_lvl <- mean(y) + c90*(sd(y)/sqrt(n))
> # get the s or standard deviation of y which is 13.09287
> #sd(y)
> # Using a 90% confidence level and need to find the confidence interval
> # Confidence Level or CI is the margin of error and written as
> # CI is to do with how reliable the estimation is
> # 1.71 * 13.09287 / square root of 25 = 4.4778
> # 98.44 - 4.778 = 93.96 is the lower level
> # 98.44 + 4.778 = 102.92 is the upper level
> c(lower_lvl, upper_lvl)
[1]  93.95993 102.92007
> # t test can be used for small samples
> t.test(y)
  One Sample t-test
data:  y
t = 37.593, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  93.03553 103.84447
sample estimates:
mean of x
    98.44
```

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`.

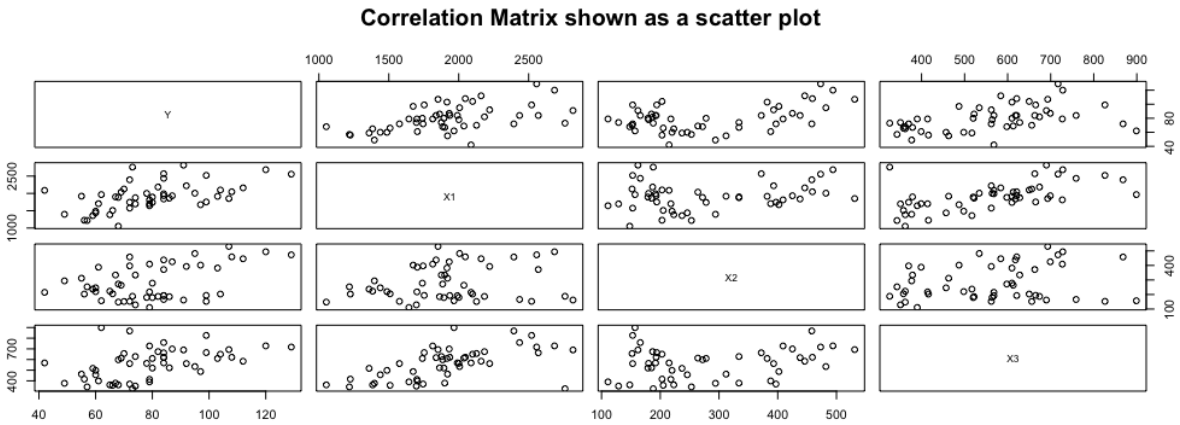
- Please plot the relationships among *Y*, *X1*, *X2*, and *X3*? What are the correlations among them (you just need to describe the graph and the relationships among them)?

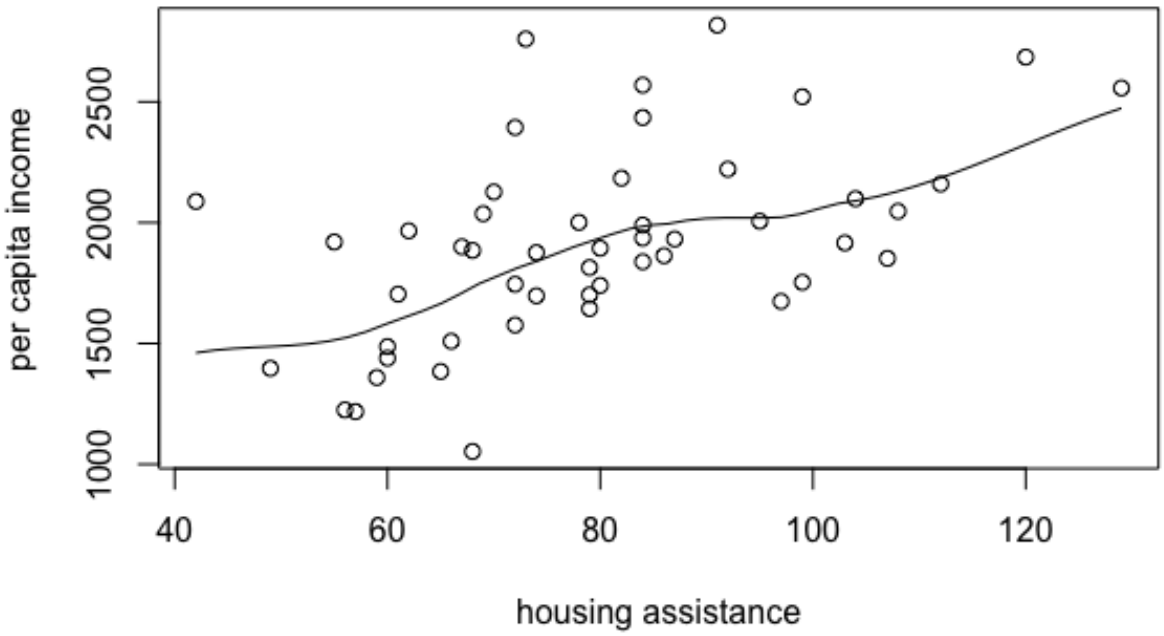
- Please plot the relationship between *Y* and *Region*? On average, which region has the highest per capita expenditure on housing assistance?

- 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.

Problem 2 - My Answer (R code)

```r

######################
# Problem 2 = My Answer
######################

expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2021/main/datasets/expenditure.txt", header=T)
str(expenditure)
# rows and columns 2 to 5
expenditure[2:5]

# using pairs() a matrix of scatter plots is produced


pairs(expenditure[2:5], main = "Correlation Matrix shown as a scatter plot")
# This shows that it is not correlated and is random

# Housing assistance and per capita expenditure / income
scatter.smooth(expenditure$Y, expenditure$X1, ylab="per capita income", xlab = "housing assistance")
# cor function calculates correlation among the vectors
cor(expenditure$Y, expenditure$X1)
```




per capita income
1000
1500
2000
2500
40
60
80
100
120
housing assistance