

# 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

```
1 #####
2 # Problem 1 = My Answer
3 #####
4
5 # Because the sample or  $n < 30$  a t-distribution can be used as opposed to a
   normal distribution
6 # t-distribution is a type of probability distribution. used where the sample
   size is small.
7
8 # The sample size is 25
9 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)
10
11 # get the sample mean ( x bar) of y which is 98.4
12 n <- length(y)
13 # sample mean
14 mean(y)
15 # standard deviation
16 sd(y)
17 (sd(y)/sqrt(n))
18 # CI of 90%
19 # From tables look for Degree of Freedom or DF of 24 and alpha level of 0.05
20 # this gives 1.71 or a t-score of 1.71
21 # 1. 25 - 1 = 24
22 # 2. 1 - .90 = .05 or the alpha level
23 # use alpha level of 0.05
24 c90 <- qt(.05, 24, lower.tail = FALSE)
25 c90
26 # CI tells us to take the mean of 98.44 and plus or minus 4.4778
27 lower_lvl <- mean(y) - c90*(sd(y)/sqrt(n))
28 upper_lvl <- mean(y) + c90*(sd(y)/sqrt(n))
29 # get the s or standard deviation of y which is 13.09287
30 #sd(y)
31 # Using a 90% confidence level and need to find the confidence interval
32 # Confidence Level or CI is the margin of error and written as
33 # CI is to do with how reliable the estimation is
34 # 1.71 * 13.09287 / square root of 25 = 4.4778
35 # 98.44 - 4.778 = 93.96 is the lower level
36 # 98.44 + 4.778 = 102.92 is the upper level
37 c(lower_lvl, upper_lvl)
38
39
40 # t test can be used for small samples
41 t.test(y)
```

## Problem 1 - My Answer - R console output

```
1
2 > # The sample size is 25
3 > 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)
4 >
5 > # get the sample mean ( x bar) of y which is 98.4
6 > n <- length(y)
7 > # sample mean
8 > mean(y)
9 [1] 98.44
10 > # standard deviation
11 > sd(y)
12 [1] 13.09287
13 > (sd(y)/sqrt(n))
14 [1] 2.618575
15 > # CI of 90%
16 > # From tables look for Degree of Freedom or DF of 24 and alpha level of 0.05
17 > # this gives 1.71 or a t-score of 1.71
18 > # 1. 25 - 1 = 24
19 > # 2. 1 - .90 = .05 or the alpha level
20 > # use alpha level of 0.05
21 > c90 <- qt(.05, 24, lower.tail = FALSE)
22 > c90
23 [1] 1.710882
24 > # CI tells us to take the mean of 98.44 and plus or minus 4.4778
25 > lower_lvl <- mean(y) - c90*(sd(y)/sqrt(n))
26 > upper_lvl <- mean(y) + c90*(sd(y)/sqrt(n))
27 > # get the s or standard deviation of y which is 13.09287
28 > #sd(y)
29 > # Using a 90% confidence level and need to find the confidence interval
30 > # Confidence Level or CI is the margin of error and written as
31 > # CI is to do with how reliable the estimation is
32 > # 1.71 * 13.09287 / square root of 25 = 4.4778
33 > # 98.44 - 4.778 = 93.96 is the lower level
34 > # 98.44 + 4.778 = 102.92 is the upper level
35 > c(lower_lvl, upper_lvl)
36 [1] 93.95993 102.92007
37 > # t test can be used for small samples
38 > t.test(y)
39 One Sample t-test
40 data: y
41 t = 37.593, df = 24, p-value < 2.2e-16
42 alternative hypothesis: true mean is not equal to 0
43 95 percent confidence interval:
44 93.03553 103.84447
45 sample estimates:
46 mean of x
47 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.