

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)

```
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 98.44
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 <- mean(y) - c90*(sd(y)/sqrt(n))
28 upper <- 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, upper)
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 98.44
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.

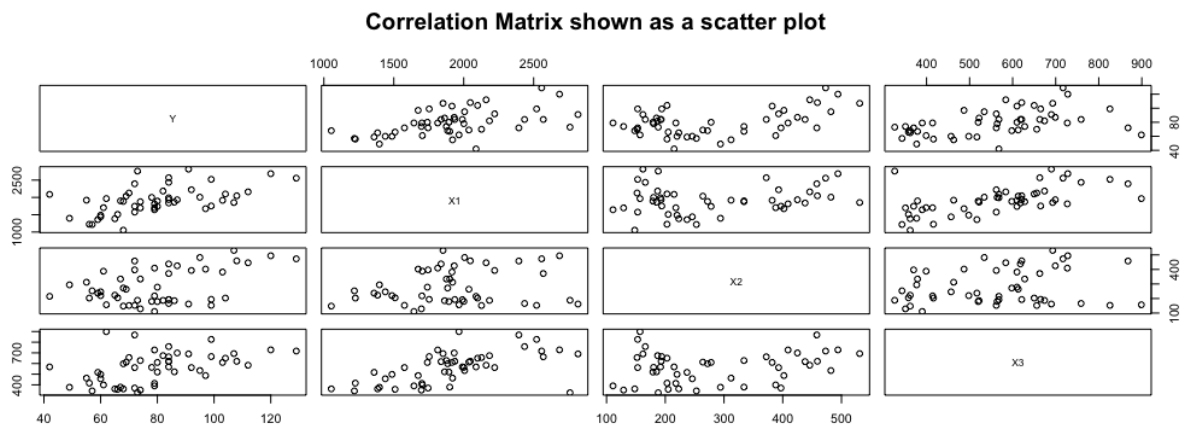
State	<i>50 states in US</i>
Y	<i>per capita expenditure on shelters/housing assistance in state</i>
X1	<i>per capita personal income in state</i>
X2	<i>Number of residents per 100,000 that are "financially insecure" in state</i>
X3	<i>Number of people per thousand residing in urban areas in state</i>
Region	<i>1=Northeast, 2= North Central, 3= South, 4=West</i>

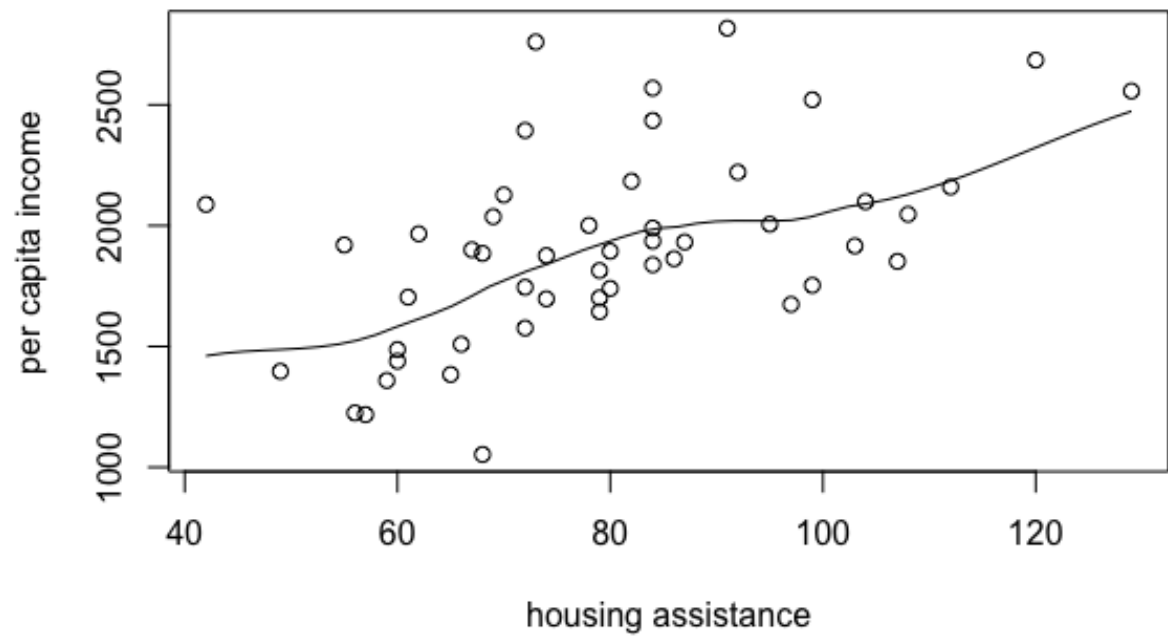
Explore the `expenditure` data set and import data into R.

- 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)

```
1 #####
2 # Problem 2 = My Answer
3 #####
4
5
6 expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
  Fall2021/main/datasets/expenditure.txt", header=T)
7 str(expenditure)
8 # rows and columns 2 to 5
9 expenditure[2:5]
10
11 # This plots the relationships among Y, X1, X2, and X3
12 # using pairs() a matrix of scatter plots is produced
13 pairs(expenditure[2:5], main = "Correlation Matrix shown as a scatter plot")
14 # This shows that it is not correlated and is random
15
16 # Housing assistance and per capita expenditure / income
17 scatter.smooth(expenditure$Y, expenditure$X1, ylab="per capita income", xlab =
  "housing assistance")
18 # cor function calculates correlation among the vectors
19 # Region X1
20 # has the highest per capita on housing assistance
21 cor(expenditure$Y, expenditure$X1)
22 scatter.smooth(expenditure$Y, expenditure$X2)
23 # regions X2
24 cor(expenditure$Y, expenditure$X2)
25 scatter.smooth(expenditure$Y, expenditure$X3)
26 # Region X3
27 cor(expenditure$Y, expenditure$X3)
28 scatter.smooth(expenditure$X1, expenditure$X2)
```





Problem 2 - My Answer - R console output

```
1
2 > expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI
  _Fall2021/main/datasets/expenditure.txt", header=T)
3 > str(expenditure)
4 'data.frame': 50 obs. of 6 variables:
5 $ STATE : chr  "ME" "NH" "VT" "MA" ...
6 $ Y      : int   61 68 72 72 62 91 120 99 70 82 ...
7 $ X1     : int  1704 1885 1745 2394 1966 2817 2685 2521 2127 2184 ...
8 $ X2     : int   388 272 397 458 157 162 494 153 152 187 ...
9 $ X3     : int   399 598 370 868 899 690 728 826 656 674 ...
10 $ Region: int    1 1 1 1 1 1 1 1 1 2 ...
11 > # rows and columns 2 to 5
12 > expenditure[2:5]
13      Y    X1  X2  X3
14 1    61 1704 388 399
15 2    68 1885 272 598
16 3    72 1745 397 370
17 4    72 2394 458 868
18 5    62 1966 157 899
19 6    91 2817 162 690
20 7   120 2685 494 728
21 8    99 2521 153 826
22 9    70 2127 152 656
23 10   82 2184 187 674
24 11   84 1990 192 568
25 12   84 2435 166 759
26 13  104 2099 203 650
27 14   84 1936 193 621
28 15  103 1916 382 610
29 16   86 1863 185 522
30 17   69 2037 264 613
31 18   74 1697 129 351
32 19   79 1644 111 390
33 20   80 1894 179 520
34 21   78 2001 180 564
35 22   73 2760 188 326
36 23   92 2221 393 562
37 24   97 1674 402 487
38 25   66 1509 205 358
39 26   65 1384 223 362
40 27   57 1218 253 343
41 28   60 1487 220 498
42 29   74 1876 334 628
43 30   49 1397 294 377
44 31   60 1439 246 457
45 32   59 1359 237 517
46 33   68 1053 148 362
47 34   56 1225 203 416
48 35   72 1576 153 562
```

```

49 36 80 1740 278 610
50 37 79 1814 409 727
51 38 55 1920 312 463
52 39 79 1701 218 414
53 40 42 2088 215 568
54 41 108 2047 459 621
55 42 84 1838 438 618
56 43 87 1932 425 699
57 44 99 1753 194 665
58 45 84 2569 372 663
59 46 112 2160 446 584
60 47 95 2006 482 534
61 48 129 2557 473 717
62 49 67 1900 334 379
63 50 107 1852 531 693
64 >
65 > # This plots the relationships among Y, X1, X2, and X3
66 > # using pairs() a matrix of scatter plots is produced
67 > pairs(expenditure[2:5], main = "Correlation Matrix shown as a scatter plot")
68 > # This shows that it is not correlated and is random
69 >
70 > # Housing assistance and per capita expenditure / income
71 > scatter.smooth(expenditure$Y, expenditure$X1, ylab="per capita income", xlab
    = "housing assistance")
72 > # cor function calculates correlation among the vectors
73 > # Region X1
74 > # has the highest per capita on housing assistance
75 > cor(expenditure$Y, expenditure$X1)
76 [1] 0.5317212
77 > scatter.smooth(expenditure$Y, expenditure$X2)
78 > # regions X2
79 > cor(expenditure$Y, expenditure$X2)
80 [1] 0.4482876
81 > scatter.smooth(expenditure$Y, expenditure$X3)
82 > # Region X3
83 > cor(expenditure$Y, expenditure$X3)
84 [1] 0.4636787
85 > scatter.smooth(expenditure$X1, expenditure$X2)

```