Problem Set 1

Jacqueline Bouvier Applied Stats/Quant Methods 1

Due: October 1, 2023

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 Sunday October 1, 2023. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 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) length(y) n ¡- length(y) df ¡- (n-1)

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

```r
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)
length(y)
# y = 25
n <- length(y)
df <- (n-1)
```

```r
# df = 24
# Find Mean of y
mean(y)
mean <- mean(y)
# mean = 98.44
# Find Standard Deviation of y
sd(y)
sd <- sd(y)
# sd = 13.0928733795654
# Since n < 30 we use t-statistic
t <-1.318
# CI_90 = mean +_ t (sd/sqrt(n))
Upper_CI_90 <- (mean + t(sd/sqrt(n)))
Lower_CI_90 <- (mean - t(sd/sqrt(n)))

# CI 90% [95.8, 101] - We are 90% confident that the average student IQ
    is
# between [95.8, 101]
```

]

2. Next, the schoolounselor 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
#Problem 1 Part 2 (https://www.adamnsmith.com/MSIN0010/hypothesis-tests.
    html)
#(http://courses.washington.edu/psy315/tutorials/t_test_tutorial.pdf)
# Using one tailed t.test as there is one parameter
#Standard Error
Sx <- (sd/sqrt(25))
a <- 0.05
# t value from t.table from https://www.sjsu.edu/faculty/gerstman/
    StatPrimer/t-table.pdf
t_tab <- 1.711

t_obs <- ((mean - 100)/(Sx))
pt(t_obs,df,lower.tail= TRUE)
p <- pt(t_obs,df,lower.tail= TRUE)


t.test(y, mu = 100) # to double check?

# Our Ho would be that the average IQ of students in the counselors
    school was
# lower than that of the average IQ score of 100 from all schools across
    the country
# we would accept the Ho as the t.table value for 90% CI is 1.711 and is
    larger than our
# t_obs value of 0.59
```

]

Question 2 (40 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)?

```r
expenditure <- read.table("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/expenditure.txt", header=T)
expenditure
str(expenditure)
# part 1 (Please Refer to Figure 1 & 2)
# Code to get the plots to be shown together from https://www.statmethods.net/advgraphs/layout.html
attach(expenditure)
par(mfrow = c(2,2))
plot(expenditure$X1, expenditure$Y, main = "Personal Income - Shelter Assistance",
     xlab = "Personal Income", ylab = "$Shelter Assistance",
)
plot(expenditure$X2, expenditure$Y, main = "Financially Insecure - Shelter Assistance",
     xlab = "Financially Insecure", ylab = "$Shelter Assistance",
)
plot(expenditure$X3, expenditure$Y, main = "Urban Area - Shelter Assistance",
     xlab = "Residing Urban", ylab = "$Shelter Assistance"
)
# Create scatter plot of (X1,Y),(X2,Y),(X3,Y)
pdf("PS01_Q2_pt1")
# We see that in the graph mapping personal income, there is somewhat of a positive
```

```
# correlation between state spending on shelter assistance and the per capita personal
# income in the state, with greater variation in the middle of the graph.
# In the graph mapping financial insecurity, there is a vague "U" shape of the
# relationship, suggesting a somewhat postive correlation with a slight bump with the
# states with the highest numbers of financially insecure residents spending more on
# shelter assistance than a strong positive correlation would suggest.
# In the graph considering urban area, there is a very weak positive correlation
# between the amount states spend on urban assistance and the number of residents
# living in urban areas.
# Create scatter plot of (X1,X2),(X1,X3),(X2,X3)
pdf("JBouvier_X1X2X3plots")
attach(expenditure)
par(mfrow = c(2,2))
plot(expenditure$X1, expenditure$X2, main = "Personal Income vs Financially Insecure",
     xlab = "Personal Income", ylab = "Financially Insecure"
     )
plot(expenditure$X1, expenditure$X3, main = "Personal Income vs Urban Residing",
     xlab = "Personal Income", ylab = "Urban Residing"
)
plot(expenditure$X2, expenditure$X3, main = "Financially Insecure vs Urban Residing ",
     xlab = "Financially Insecure", ylab = "Urban Residing"
)


# I would have thought there would be more correlations between these. In the plot of
# personal income vs urban residing we can assume a positive correlation where the
# higher the personal income the more likely to reside in an urban area. Which then if you look
# at financially insecure vs urban residing, we can see that there is very little correlation
# as this could be because we could assume that more people living in urban areas would perhaps make more income
# Though what we don't see with these is the states where the states/regions could tell us a lot more about
# any relationships that could be there.
```

]

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

```r
# part 2 (Please Refer to Figure 3)

plot(expenditure$Region, expenditure$Y, main = "Region vs Expenditure",
     xlab = "Region", ylab = "Expenditure")
# It looks like that Region 4 has the highest expenditure per capita on housing assistance
# Though the data also is much more spread out than say Region 2 or 3. Looking at the Data we can see that
# California is the highest expenditure. Region 4 also has the lowest expenditure as well with Wyoming.
#
```

]

- 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.

```r
# part 3 (Please Refer to Figure 4)

attach(expenditure)
par(mfrow = c(2,2))
plot(expenditure$X1, expenditure$Y, main = "Personal Income - Shelter Assistance",
     xlab = "Personal Income", ylab = "$Shelter Assistance",
)
# We can see how personal income vs how much is spent on shelter/housing assistance. As mentioned in part
# one of this question that perhaps with higher income comes higher taxes - and depending on the
colors <- c('green', 'blue','red', 'black')
plot(expenditure$X1, expenditure$Y, main = "Personal Income - Expenditure",
     xlab = "Personal Income", ylab = "$Shelter Assistance",
     pch = expenditure$Region, col = colors[expenditure$Region])

# Green Region 1, Blue Region 2, Red Region 3, Black Region 4
```

]

Figure 1: Variable Relationships.

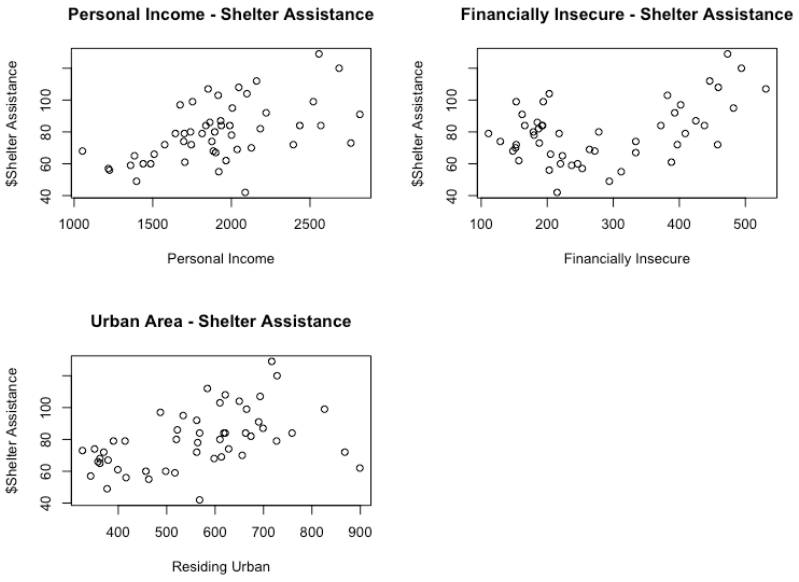

Figure 2: X Variable Relationships.

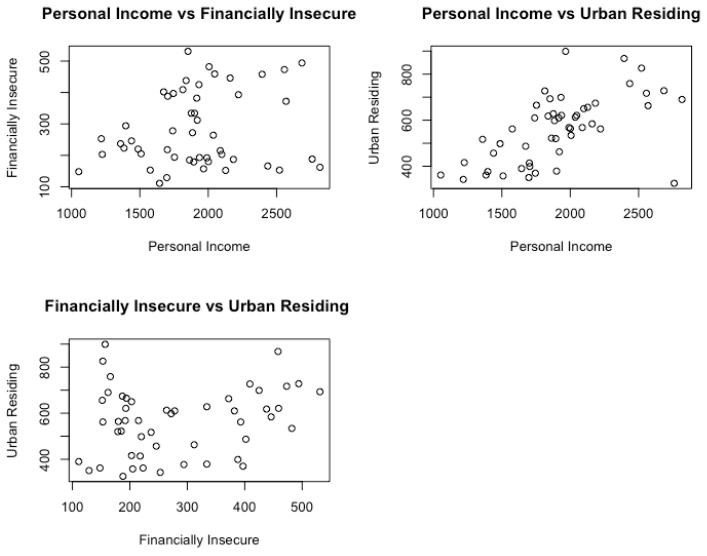

Figure 3: Expenditure vs Region.

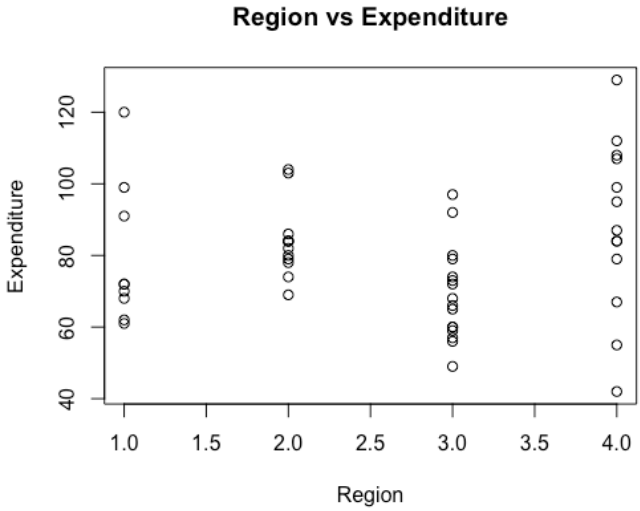

Figure 4: Expenditure vs Peronal income with Regional Variable.