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.

Steps for calculating CIs:

1. Calculate ȳ

2. Calculate S and then σ^y=S√n

3. How much area do we need under the curve to the right?I(1-ConfidenceCoecient)/2How much area do we need under the curve to the left?I(ConfidenceCoecient)/2

4.Find the z-score associated with that number

5.Use these values to calculateȳ±Z×σ^y

Sample size =25

T-test as the properties of the t distribution are met: Code for histogram did not work, therefore normal distribution was assumed for the purpose of this assignment. – Histogram was formed after several attempts. Described by me as

Mean=98.44

SD= 13.09287

~~Since *n* = 25, our test statistic *t** has *n* – 1~~

90% CI

Rstudio output:

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

Looking for 90% CI???

Do manually in this case as not sure how to get 90% instead of 95% CI in Rstudio? Possible? Not today.

Formula (A&Fp118):

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 = 0:05.

Look at hypothesis testing slides, H0 etc.

Tutorial example:

```
#4.
Formulate
our null
hypothesis
             #NULL HYPOTHESIS: There is no difference in means between IVR/Online polls
             and
             #Other kinds of polling.

             #5. Test our hypothesis
             length(Trump$mode[Trump$mode == "IVR/Online"]) #how many samples do we
             have?

             Trump_IVR <- subset(Trump, Trump$mode == "IVR/Online") #make a subset of
             IVR
             Trump_other <- subset(Trump, Trump$mode != "IVR/Online") #make a subset of
             everything else

             Trump_null <- t.test(Trump_IVR$Approve, Trump_other$Approve, mu = 0)
             Trump_null
             #What does it all mean?!
```

What is my Null hypothesis for this question: There is no difference between the average student IQ in her school and the average IQ score among all the schools in the country.

```
length(Trump$mode[Trump$mode == "IVR/Online"]) #how many samples do we have?
```

```
Trump_IVR <- subset(Trump, Trump$mode == "IVR/Online") #make a subset of IVR
Trump_other <- subset(Trump, Trump$mode != "IVR/Online") #make a subset of everything
else

Trump_null <- t.test(Trump_IVR$Approve, Trump_other$Approve, mu = 0)
Trump_null
```

That hypothesis doesn't indicate direction though. However, I will use it for this example.

Testing my hypothesis:

Question 2 (50 points): Political Economy

Researchers are curious about what ects 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

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.

Import data into R achieved.

(i)• 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)?

Multiple graphs or multiple regression? Probably multiple regression. Just graphing though? And can you describe the relationships among variables in the same graph, usually this is done after multiple regression, picking apart the relationship between two variables out of many using graphs and t-tests?

So ~~individual line graphs~~ in this case. Scatterplots and describe the appropriate relationship. Isn't that an excessive amount of graphs: Y and X1. X1 and X2. X2 and X3. X3 and Y. Four graphs, Not that excessive actually. In the softwares we are using there is actually the ability to graph the four variables on the same scatterplot.

Watch Youtube video on to create these graphs as editing code from examples is proving extremely beyond my understanding at this point. – name of video: Latex Tutorial- Creating graphs from data with Tiks and Pgfplots in Overleaf.

Therefore, I need to create one scatterplot of plotting the relationships among Y, X1, X2, and X3. Then describe the graph and the relationships as was done in undergrad (negative, positive relationship, etc.).

Resources I used: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

If I need to describe the relationship, then a regression line will be useful to add. Look at the heading 'Scatter plots with multiple groups' and 'Change the point color/shape/size automatically'. I have highlighted 'color=Region' as this isn't asked in this question, however it is asked in the next.

```
# Basic scatter plot ggplot(expenditure, aes(x=Y, y=X1)) + geom_point() #
Change the point size, and shape ggplot(expenditure, aes(x=Y, y=X1)) +
geom_point(size=2, shape=23)# Change point shapes, colors and sizes
ggplot(expenditure, aes(x=Y, y=X1, shape=X2, color=Region, size=X3)) +
geom_point()
```

```
# Add the regression line ggplot(expenditure, aes(x=Y, y=X1)) +
geom_point()+ geom_smooth(method=lm)
```

(ii)• Please plot the relationship between Y and Region? On average, which region has the highest per capita expenditure on housing assistance?

Linear regression, simple bivariate scatter plot?

Possibly code lines about 80 through to 140 of the tutorial 2 slides, from aggregates to barplot

```
#Here, we
use the
aggregate()
function.
It takes as
an argument
the vector
you want
              #to apply another function to, here the Trump$Approve vector. We supply
              the function
              #we want to apply to Trump$Approve using the FUN = argument, and we use
              the by =
              #argument to tell R what we want to group our operation by - here, the
              survey house.
              #We need to supply a list to the by = argument, so we coerce the
              survey_house variable
              #using the list() function.

              Trump_means <- aggregate(Trump$Approve, by = list(Trump$survey_house), FUN
              = mean)
              class(Trump_means) #aggregate returns a data.frame

              Trump_means[Trump_means$Group.1 %in% c("Gallup", "Pew", "FOX"),]
              subset(Trump_means, Group.1 %in% c("Gallup", "Pew", "FOX"))

              #Once we have an object with all the approval means according to survey
              company, we
              #can easily subset that by passing a vector of the companies we want. We
              use the
              #logical operator %in% to check a list of characters - handy for checking
              names of
              #things.

              #### Exercise
```

```
#Let's try applying what we've learned using aggregate() to another column
in the
#Trump data.frame. Polling companies use different methods to survey
voters. Let's
#see if these different methods result in different means.

#Let's find the right variable
ls.str()

#We can call the unique() function on our variable to see the different
methods
unique()

#Now use aggregate() to get the mean of the approval rate as grouped by
this
#variable. Remember to assign it to an object.
object1 <- aggregate(data.frame$var, by = list(data.frame$other), FUN =
fun)

######################
# Visualising our Data
######################

#We can use R's base plotting functions to see the results of our
analysis. Let's
#try this for the polling companies.

barplot(Trump_means$x)
#What kind of a distribution do we see here?

barplot(Trump_means$x,
        main = "Trump Approval Ratings by Survey Company", #Our title
        names.arg = Trump_means$Group.1, #The vector with our axis names.
Could use Trump_means[,2]
        horiz = TRUE, #Flipping our axes
        las = 1) #Rotating our axis labels
#We can use additional arguments to barplot() to edit our axis labels,
etc. There's
#no really good way of fitting all our labels on here, but it gives you an
idea.

#### Exercise
#Try the same with your survey method means object.
barplot(data.frame$x, #change the data.frame name
        main = "Your Title Here", #change the title
        names.arg = data.frame$Group.1, #change the data.frame name
```

```
horiz = TRUE,
las = 1)
```

However, there are other ways to do this. How exactly do I do that in R though????? Is there multiple ways of graphing this data?

I could try a box plot or "boxplot R". Box and whisker plots are **very effective and easy to read**, as they can summarize data from multiple sources and display the results in a single graph. Box and whisker plots allow for comparison of data from different categories for easier, more effective decision-making.

Resources: https://www.datamentor.io/r-programming/box-plot/#:~:text=In%20R%2C%20boxplot%20

https://www.datamentor.io/r-programming/box-plot/#:~:text=In%20whisker,numeric%20vectors%20as%20its%20components

See 'different boxplots for each month' section, in this case it would be 'different boxplots for each region.

```
boxplot(Y~Region,

data=expenditure,

main="Different boxplots for each region",

xlab="Region Number",

ylab="per capita expenditure on shelters/housing assistance in state",

col="orange",

border="brown"

)
```

Region is our grouping variable, with region in the form of a number (1,2,3,4).

I tried running this in Rstudio, error message. This must be one of the ways to illustrate this relationship. Meaning I have not imported my dataset correctly to Rstudio from Github?????? I cannot view the data and view which region has the highest per capita expenditure on housing assistance? So I guess it is 1=Northeast.

(iii)• 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 colours.

Three variables graph, stratified or grouped by variable Region?

From http://www.sthda.com/english/wiki/creating-and-saving-graphs-r-base-graphs

```
plot(x = my_data$wt, y = my_data$mpg, pch = 16, frame = FALSE, xlab = "wt", ylab = "mpg", col = "#2E9FDF")
```

```
plot(x = Y, y = X1, pch = 16, frame = FALSE, xlab = "wt", ylab = "mpg", col = "#2E9FDF")
```

Did not work^^

Need to determine exactly what is the relationship between these variables first and then choose the graph based on this.

Answer: https://stackoverflow.com/questions/39110770/graphing-3-variable-scatterplot-r

Basically a 2D scatter plot with colour indicating the 3 variable, in this case Region. Four regions, 4 colours.

Since gender is a binary variable (usually, otherwise ternary), I would plot a 2D scatterplot with color encoding the gender.

Dummy data:

```
a = data.frame(x=runif(100), y = runif(100)+2, group = round(runif(100))+1 )
```

Now I would plot `y` against `x` using `a$group` to select the color:

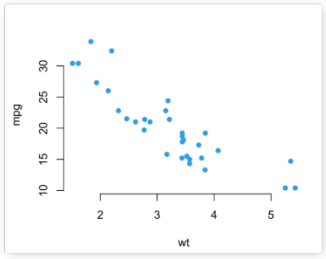
```
plot(a$y, a$x, pch = 16, col = c('cornflowerblue', 'springgreen')[a$group])
```

This is how it is described in sthda.com:

Creating graphs

The R base function **plot**() can be used to create graphs.

```
plot(x = my_data$wt, y = my_data$mpg,
     pch = 16, frame = FALSE,
     xlab = "wt", ylab = "mpg", col = "#2E9FDF")
```



Why would the question contain four colours if the example describes how to do one? Maybe there is just multiple ways to do this? Actually see part (ii) and the stdha.com example seems to fit with that, so this is another step perhaps showing independent work? So try out the four colour scatterplot for this part of the q.

My attempt:

Replace a$ with Region, Replace X with X1.

a=data.frame(x1=runif(100), y = runif(100)+2, group = round(runif(100)+1

plot(Regiony,Regionx1, pch = 16, col = c('cornflowerblue', 'springgreen', 'red', 'yellow')[Regiongroup])

Run this and see what happens.

Error : unexpected symbol.

Check colours as I made up two for Microsoft example.

Shouldn't R be suggesting colour examples for me as I begin to type?

Tried replacing Red with = #FF0000 and yellow with =#000000 which is black.

Error message: Error: unexpected input in:

"a=data.frame(x1=runif(100), y = runif(100)+2, group = round(runif(100)+1

+ plot(Regiony,Regionx1, pch = 16, col = c('"

Why does R give me an error message and not tell me what exactly is the error???????

Isn't this what Latex is for?

Other resources: https://r-charts.com/correlation/scatter-plot-group/, http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

So essentially the type of graph as the first part?

This is my answer for the first part:

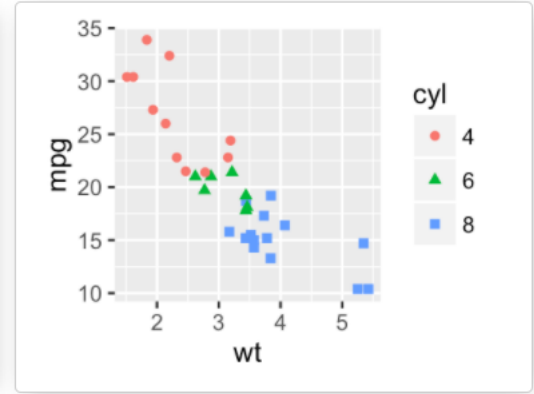
```
# Basic scatter plot ggplot(expenditure, aes(x=Y, y=X1)) + geom_point() #
Change the point size, and shape ggplot(expenditure, aes(x=Y, y=X1)) +
geom_point(size=2, shape=23)# Change point shapes, colors and sizes
ggplot(expenditure, aes(x=Y, y=X1, shape=X2, color=Region, size=X3)) +
geom_point()

# Add the regression line ggplot(expenditure, aes(x=Y, y=X1)) +
geom_point()+ geom_smooth(method=lm)
```

This is the question: 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 colours.

So something like this:



But perhaps with four colours/symbols and x and y are continuous.