Assignment (thinking about data)

Quang Dong Nguyen

2022-05-27

By including this statement, we the authors of this work, verify that:

- I hold a copy of this assignment that we can produce if the original is lost or damaged.
- I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
- I am aware that this work may be reproduced and submitted to plagiarism detection software programs to detect possible plagiarism (which may retain a copy on its database for future plagiarism checking).
- I hereby certify that we have read and understood what the School of Computing, Data and Mathematical Sciences defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

Question 1.1:

Test if the mean number of cylinder is different between Mazda and Isuzu vehicles

Since the question asks specifically if there is any different in the mean number of cylinders between Mazda and Isuzu.

To find the evidence of differences in the mean number of cylinders between Mazda and Isuzu, I will specifically use the t.test techniques with the method: alternative = "two.sided" and compute a hypothesis test, as I only want to find if there is any difference in means between two pairs of data. If the result gives out a p-value of less than 0.05, it means there is evidence of differences between Mazda and Isuzu's mean number of cylinders.

```
vehicles= read.csv("light_vehicles.csv")
head(vehicles)
```

```
##   Year   Make Colour         Fuel.type Number.of.cylinders Number.of.seats
## 1 2007   Ford   Blue      Petrol - Gas                  14               7
## 2 2015 Toyota  Beige Diesel - Electric                   5               9
## 3 1995 Holden  Black Diesel - Electric                   8               9
## 4 2016  Honda  Black Petrol - Electric                  10              13
## 5 1996 Suziki Silver            Diesel                   6              11
## 6 1990  Isuzu Silver Petrol - Electric                   3               9
##   GVM.weight Tare.weight
## 1       1189        4062
## 2       2866        3497
## 3       5778        1936
## 4       2609        3074
## 5       3395        4966
## 6       1817        4972
```

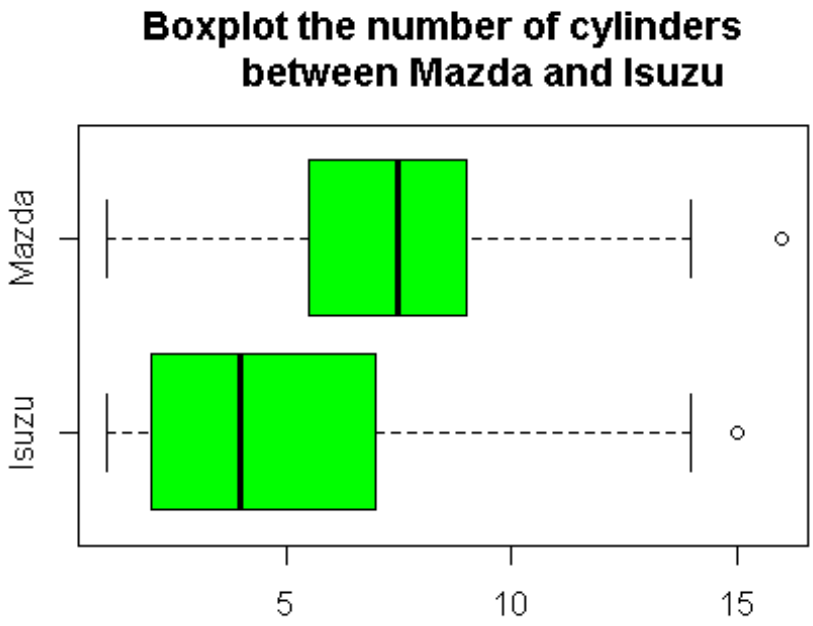
```
#T.test computically
#h0: There is no difference between number of cylinders of Mazda and Isuzu
#hA: There is a difference between number of cylinders of Mazda and Isuzu
#data pulled
vehicles.Isuzu=subset(vehicles, Make=="Isuzu", Number.of.cylinders, drop=TRUE
)
mean(vehicles.Isuzu) #mean of vehicle Isuzu
```

```
## [1] 4.927536
```

```
vehicles.Mazda= subset(vehicles, Make=="Mazda", Number.of.cylinders, drop=TRU
E)
mean(vehicles.Mazda) #mean of vehicle Mazda
```

```
## [1] 7.3
```

```
label=c("Isuzu", "Mazda")
boxplot(vehicles.Isuzu, vehicles.Mazda, names=label, col="green", horizontal=
TRUE, main="Boxplot the number of cylinders
        between Mazda and Isuzu")
```



- The box plot displays an overview of all number of cylinders of Mazda and Isuzu from the max value to the min value, the mean, the range, the quartile and quartile data.
- This graphing techniques on two groups of value allows viewers to have a better overview of the overall data within the two groups.
- We can see the max and min number of cylinders between Mazda and Isuzu is the same, as they came from the same population, which gives them an equal variance. And for the overview, we can see the mean number of cylinders of Mazda is generally higher than Isuzu.

Then the t.test is proceeded on the two extracted data of Number of cylinders from Mazda and Isuzu, to find the probability of the occurrence of any differences in the number of cylinders. if p is smaller than 0.05, then it will reject the null hypothesis and accept that there is an evidence of difference.

```
vehicles.t.test=t.test(vehicles.Isuzu,vehicles.Mazda, var.equal=TRUE, alterna
tive="two.sided")
vehicles.t.test$p.value

## [1] 0.0005053977
```

```
#p value is less than 5%
#since p-value is small, reject h0.
#hA:There is a difference in mean N.cylinder of Isuzu and Mazda
```

However, since the size of data is small, to prevent error in measurement to happen, we need to simulate more data from the sample population of the data. Meaning, we need to create more samples of Isuzu and Mazda 's number of cylinder to reduce bias in the approximation. Therefore, I will replicate data over 1000 times with sampled data from both Isuzu and Mazda's number of cylinder and perform a t.test to obtain t-statistics after each replication.

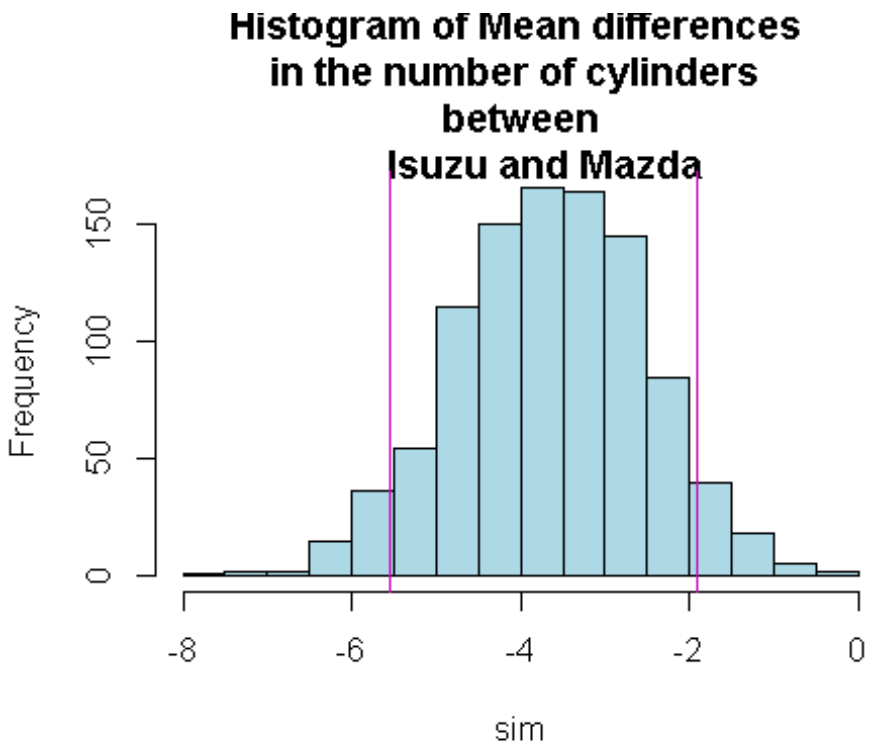
```
#Since the sample size is small, We can create sample to test assure the quan
tile interval:
#from statistics
sim= replicate(1000, {
  Isuzu.resamp= sample(vehicles.Isuzu, replace=TRUE)
  Mazda.resamp= sample(vehicles.Mazda, replace=TRUE)
  t.test(Isuzu.resamp, Mazda.resamp, var.equal=TRUE)$statistic
})
hist(sim, breaks=20, main= "Histogram of Mean differences
in the number of cylinders
between
     Isuzu and Mazda", col="lightblue")
quantile(sim, c(0.05, 0.95))

##        5%       95%
## -5.551902 -1.918636

quantile (sim, 0.50) #highest frequency of mean differences

##       50%
## -3.603018

abline(v=quantile(sim, c(0.05, 0.95)), col=6)
```



- In graphing techniques, I used the distribution histogram to represent the data because, the data is replicated 1000 times. Therefore, the distribution histogram can attribute the occurrence frequency of data in a most observable way when replicating
- In fact, from the graph alone, we can see the mean differences in number of cylinders between Isuzu and Mazda is already centralised at approximate -3.6030181.This means Mazda has almost 4 more numbers of cylinders than Isuzu. And from the function quantile(sim, c(0.05, 0.95)), We are 95% confident that the mean difference in number of Cylinders between Isuzu and Mazda is approximately between -5.5519025 to -1.9186363.

Question 1.2:

Test if the mean number of seats is different for each colour. If so, determine which colour has a statistically different mean

The question is asking for evidences of different of the mean number of seats for each colour and then shows which colour is statistically different. Therefore, in this example, I will use oneway test and then TukeyHSD test to find the different for each colour. As one way test is specifically used for finding any significant differences in means between two or more groups, we set the threshold of 0.05, if p-value is less than 0.05, meaning there is a significant differences between means of two groups or more. Meanwhile, TukeyHSD test enables users to see the differences in means, the p adjacent between every possible paired groups. If p-value is less than 0.05, there is a different between means of that paired groups.

```
#t.test compares the differences in mean between two groups
#since populations is large we dont need sample simulation
F.test=oneway.test(vehicles$Number.of.seats~vehicles$Colour, data=vehicles)
F.test #since p. value is large and is F-statistics is small, shows there is
not much of  a difference in groups means
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  vehicles$Number.of.seats and vehicles$Colour
## F = 0.70236, num df = 8.00, denom df = 181.56, p-value = 0.6892
```

```
fit=aov(vehicles$Number.of.seats~vehicles$Colour, data=vehicles)
fit
```

```
## Call:
##    aov(formula = vehicles$Number.of.seats ~ vehicles$Colour, data = vehicl
es)
##
## Terms:
##                 vehicles$Colour Residuals
## Sum of Squares           87.139  7443.829
## Deg. of Freedom               8       491
##
## Residual standard error: 3.893655
## Estimated effects may be unbalanced
```

- Since we want an anova table to proceed to the next step of using TukeyHSD test, we will call the function aov().

Then proceeds to call the function TukeyHSD() to summarise the F-data.

```
summary(fit)
```

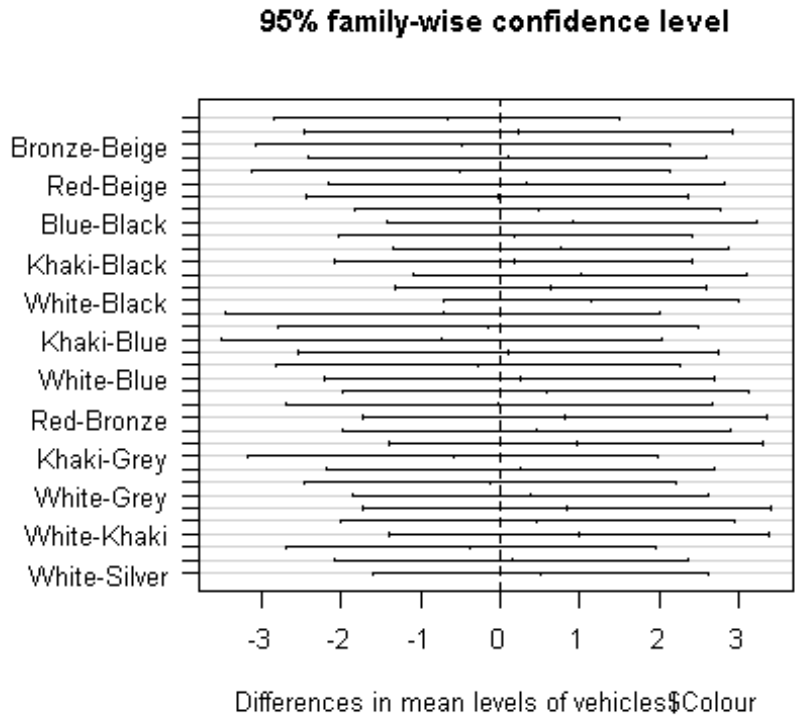
```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## vehicles$Colour    8     87   10.89   0.718  0.675
## Residuals        491   7444   15.16

Tukey.test=TukeyHSD(fit)
Tukey.test

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = vehicles$Number.of.seats ~ vehicles$Colour, data = vehi
cles)
##
## $`vehicles$Colour`
##                         diff       lwr      upr     p adj
## Black-Beige    -0.66730769 -2.831926 1.497311 0.9890902
## Blue-Beige      0.23783784 -2.454384 2.930060 0.9999990
## Bronze-Beige   -0.48095238 -3.083747 2.121842 0.9997096
## Grey-Beige      0.09387755 -2.410909 2.598664 1.0000000
## Khaki-Beige    -0.50000000 -3.136236 2.136236 0.9996474
## Red-Beige       0.34000000 -2.152767 2.832767 0.9999711
## Silver-Beige   -0.03333333 -2.425679 2.359012 1.0000000
## White-Beige     0.48219178 -1.817049 2.781432 0.9992598
## Blue-Black      0.90514553 -1.417073 3.227364 0.9531873
## Bronze-Black    0.18635531 -2.031566 2.404277 0.9999993
## Grey-Black      0.76118524 -1.340860 2.863231 0.9697678
## Khaki-Black     0.16730769 -2.089766 2.424381 0.9999998
## Red-Black       1.00730769 -1.080400 3.095016 0.8538434
## Silver-Black    0.63397436 -1.332737 2.600686 0.9854129
## White-Black     1.14949947 -0.702836 3.001835 0.5907337
## Bronze-Blue    -0.71879022 -3.454054 2.016473 0.9962946
## Grey-Blue      -0.14396029 -2.786134 2.498214 1.0000000
## Khaki-Blue     -0.73783784 -3.504943 2.029267 0.9959009
## Red-Blue        0.10216216 -2.528620 2.732944 1.0000000
## Silver-Blue    -0.27117117 -2.807003 2.264661 0.9999956
## White-Blue      0.24435394 -2.203836 2.692544 0.9999975
## Grey-Bronze     0.57482993 -1.976162 3.125822 0.9987529
## Khaki-Bronze   -0.01904762 -2.699224 2.661129 1.0000000
## Red-Bronze      0.82095238 -1.718239 3.360143 0.9851308
## Silver-Bronze   0.44761905 -1.993061 2.888299 0.9997255
## White-Bronze    0.96314416 -1.386348 3.312636 0.9375979
## Khaki-Grey     -0.59387755 -3.178982 1.991227 0.9985669
## Red-Grey        0.24612245 -2.192506 2.684751 0.9999972
## Silver-Grey    -0.12721088 -2.463092 2.208670 1.0000000
## White-Grey      0.38831423 -1.852117 2.628745 0.9998203
## Red-Khaki       0.84000000 -1.733460 3.413460 0.9841923
## Silver-Khaki    0.46666667 -2.009646 2.942979 0.9996638
## White-Khaki     0.98219178 -1.404294 3.368678 0.9362017
## Silver-Red     -0.37333333 -2.696321 1.949654 0.9998985
```

```
## White-Red      0.14219178 -2.084793 2.369176 0.9999999
## White-Silver   0.51552511 -1.598450 2.629500 0.9978060

par(mar=c(5.1,10,4.1,2.1), cex=0.8)
plot(TukeyHSD(fit), las=1)
```



- I used plot(TukeyHSD()) graphing since it can visually show the comparison between paired data.However, the data is too large, so before that, I need to set a default parameter with margin size that will enable to compact all the data of fit to the plot(TukeyHSD()).
- As we can see from this graph, there is not much of differences in means number of seats for each colour (these means are close to zero) . And combining with the Tukey.test listing, we can see white-black, might potentially be different since p adjacent is much lower than other paired colours. However, the p-value is still high, where p adj of white-black is 0.5907337 (>0.05, the threshold to reject). In assumption, there is not much of a different between white and black also.

Question 1.3:

1/ Use Bootstraping to compute a 88% confidence interval for the difference between GVM and TareWeights for Volkswagen vehicles?

2/ Compute a 88% cI for the difference between GVM and Tare weights for Volkswagen vehicle by using approximation.

3/ Can we conclude that GVM weights are different than Tare weights for Volkswagen vehicles (Dont do a hypothesis test)?

4/ Test the hypothesis that GVM weights are greater than Tare weights for Volkswagen vehicles

```
#1/
#since there is not much record on VOlkswagen's properties, we need to simula
te boot to approximate data:
#this is the evidences of the lack of Volkswagen data:
count.Volkswagen.GVM.weight= length(subset(vehicles, Make=="Volkswagen", GVM.
weight, drop=TRUE))
count.Volkswagen.GVM.weight
```

```
## [1] 36
```

```
count.Volkswagen.Tare.weight= length(subset(vehicles, Make=="Volkswagen", Tar
e.weight, drop=TRUE))
count.Volkswagen.Tare.weight
```

```
## [1] 36
```

- There are only 36 recorded data of the GVM.weight of Volkswagen.
- There are only 36 recorded data of the Tare.weight of Volkswagen.

Now, I do the simulation by bootstrapping the shuffled data from GVM and Tare weight of Volkswagen vehicles.

```
Volkswagen.GVM= subset(vehicles, Make=="Volkswagen", GVM.weight, drop=TRUE)
Volkswagen.GVM
```

```
##  [1] 6727 3852 4307 4143 2259 1890 5521 2232 2843 3315 6709 1958 3285 2760
1989
## [16] 5921 5718 6602 2060 2878 1613 4173  561 5227 6091 4783 2881 3311 6073
5417
## [31] 4097 5146 3006 1758 4156 2811
```

```
Volkswagen.Tare= subset(vehicles, Make=="Volkswagen", Tare.weight, drop=TRUE)
Volkswagen.Tare
```

```
##  [1] 1411 1744 1030 2842 2046 3608 3767 3392 1138 2044 4454 1543 4895 1115
1662
## [16] 3619 1699 5266 4173 4805 3863 3958 2050 1526 2421 4623 2724 4201 2444
4070
## [31] 1045 2060 2834 3266 1969 4147
```

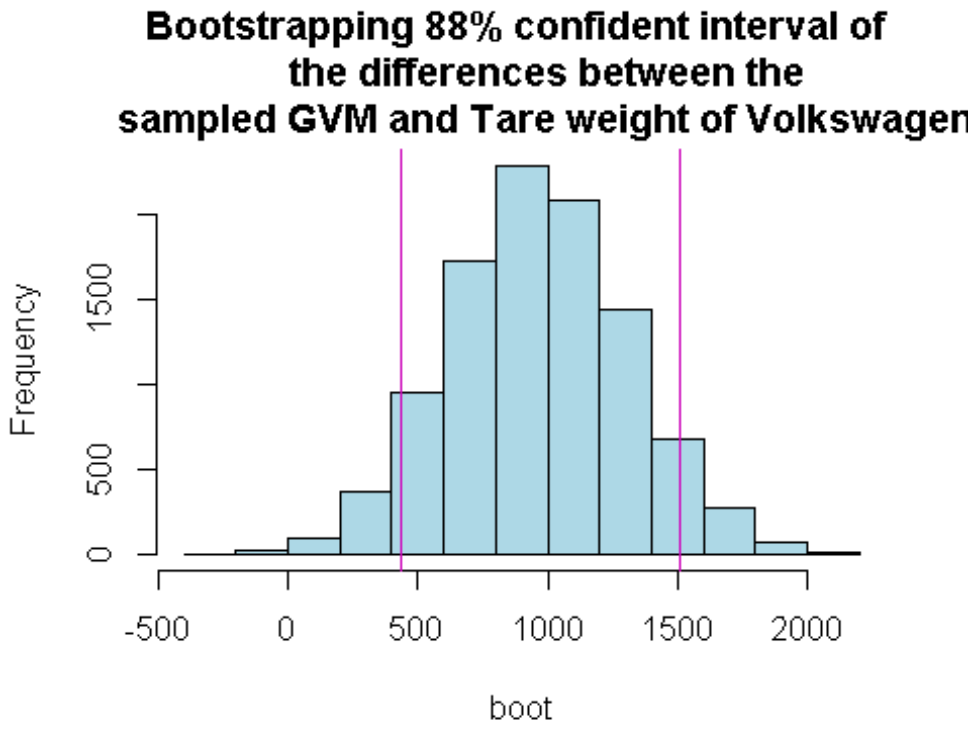
```
d0= mean(Volkswagen.GVM-Volkswagen.Tare)
d0
```

```
## [1] 961.6389
```

```
boot= replicate(10000, {
  Volkswagen.GVM.sampled= sample(Volkswagen.GVM, replace=TRUE)
  Volkswagen.Tare.sampled= sample(Volkswagen.Tare, replace=TRUE)
  mean(Volkswagen.GVM.sampled)-mean(Volkswagen.Tare.sampled)
})
hist(boot, main= "Bootstrapping 88% confident interval of
     the differences between the
     sampled GVM and Tare weight of Volkswagen", col="lightblue")
quantile(boot, c(0.06, 0.94))
```

```
##        6%       94%
##  437.3872 1505.5328
```

```
abline(v=quantile(boot, c(0.06, 0.94)), col=6)
```



- Again, I used distribution histogram to represent the bootstrapping data since it can show the frequency of occurrence of the bootstrapping data. And then the function

abline shows the horizontal vector to highlight the range of 88% confident interval. In this boot distribution, the abline lines up between 437.3872222, 1505.5327778. And the highest frequency occurrence of the difference between sampled GVM and Tare weight of Volkswagen is at 962.5138889.

The Wilcox.test is used show the difference between two pairs of data. In this case, we use Wilcoxon-Mann-Whitney test to find the range of difference between GVM and Tare Weight of Volkswagen vehicles that has 88% confident level.

```
#2/
#or using wilcox.test (alternative=two.sided) #dont do hypothesis test
CI_88_approx=wilcox.test(Volkswagen.GVM,Volkswagen.Tare, conf.int= TRUE, conf
.level=0.88)

## Warning in wilcox.test.default(Volkswagen.GVM, Volkswagen.Tare, conf.int =
## TRUE, : cannot compute exact p-value with ties

## Warning in wilcox.test.default(Volkswagen.GVM, Volkswagen.Tare, conf.int =
## TRUE, : cannot compute exact confidence intervals with ties

CI_88_approx

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Volkswagen.GVM and Volkswagen.Tare
## W = 869, p-value = 0.01301
## alternative hypothesis: true location shift is not equal to 0
## 88 percent confidence interval:
##   347 1480
## sample estimates:
## difference in location
##               910.1808

CI_88_approx$conf.int #shows the approximate conf level of 88%CI

## [1]  347 1480
## attr(,"conf.level")
## [1] 0.88

#3/
#Yes, we are 88% confident to conclude that GVM weights are different than Ta
re weights of Volkswagen vehicles since 88% of data
#lies between approximate 347 and 1480 (observed from Wilcoxon-Mann Whitney t
est and from graph data) and a very small percentage lying on 0
#which means the difference is zero between GVM and Tare weight for Volkswage
n
```

- So 88% confident level locates between 346.9999928, 1480.0000158.

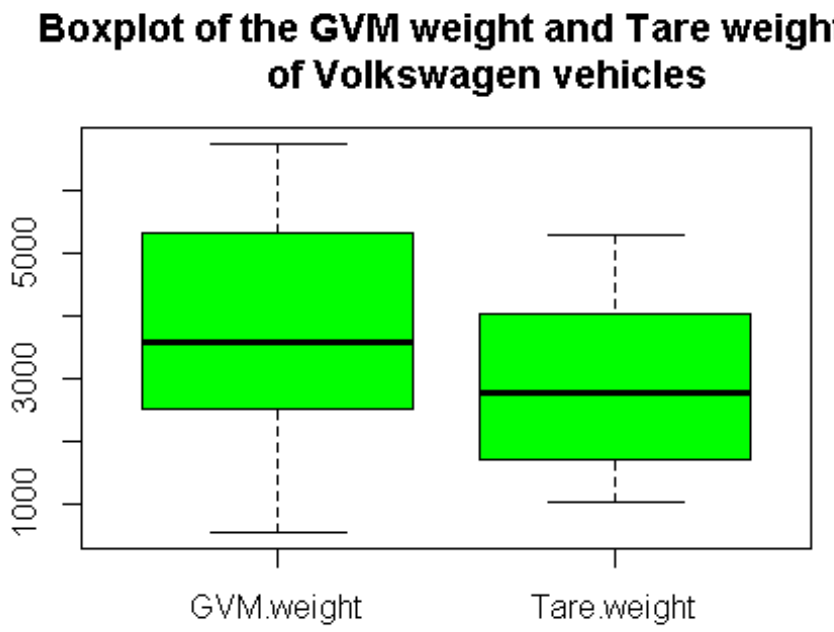
Again, I will use the Wilcoxon-Mann-Whitney test to find if GVM weights are greater than Tare weights for the vehicle Volkswagen. Before that, I need to conduct the hypothesis.

```
#4/
##Use Wilcoxon-Mann Whitney test to find if GVM weights are greater than Tare Weights for Volkswagen
#h0: GVM.weight is not greater to Tare.Weights
#hA; GVM.weight is greater than Tare.weights
wilcox.test(Volkswagen.GVM, Volkswagen.Tare, alternative="greater")

## Warning in wilcox.test.default(Volkswagen.GVM, Volkswagen.Tare, alternativ
e =
## "greater"): cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Volkswagen.GVM and Volkswagen.Tare
## W = 869, p-value = 0.006507
## alternative hypothesis: true location shift is greater than 0

label= c("GVM.weight","Tare.weight")
boxplot(Volkswagen.GVM,Volkswagen.Tare, names=label, main="Boxplot of the GVM weight and Tare weight
        of Volkswagen vehicles", col="green") #Observation from boxplot
```



```
#since p-value is small, therefore reject the null hypothesis
#hA: GVM.weight is greater than Tare.weights
```

- The box plot displays an overview of all GVM.weight and Tare.weight of Volkswagen vehicles from the max value to the min value, the mean, the range, the quartile data.
- I used box plot technique to summarise both data of GVm and Tare weight of Volkswagen in to two box plot techniques, so that we observe the general differences between GVM and Tare weight of the vehicles Volkswagen. And we can see the GVM weight is generally higher than the Tare weight but at the same time, it has a higher range from max value to min value than the Tare weight of the vehicle Volkswagen
- Since p-value is small ($p = 0.0065072$), therefore reject the null hypothesis, and we can conclude that GVM.weight is greater than Tare.weights

Question 1.4:

Test if there is a difference in proportions of the Blue vehicles between Landrover and Mercedes (The question is requested by Dr Gizem Intepe through email on Tuesday 5/24/2022 at 9:54 pm, to be replaced into the below):

Test if there is a difference in proportions of the Silver vehicles between Landrover and Mercedes

```
a=length(subset(vehicles, Make=="Landrover", Colour, drop=TRUE)) #total numbe
r of Landrover was made
a
```

```
## [1] 16
```

```
b=length(subset(vehicles,Make=="Mercedes", Colour, drop=TRUE)) #total number
of Mercedes was made
b
```

```
## [1] 24
```

```
Colour.vehicles.table=table(vehicles$Make[vehicles$Colour=="Silver"])
Colour.vehicles.table
```

```
##
##        BMW       Ford     Holden      Honda      Isuzu        Kia  Landrov
er
##          2          6          3          4          2          9
6
##      Mazda   Mercedes Mitsubishi     Nissan      Skoda     Suziki     Toyo
ta
##          7          3          1          1          3          4
5
## Volkswagen
##          4
```

```
Silver.Landrover=Colour.vehicles.table[7]
Silver.Landrover
```

```
## Landrover
##         6
```

```
Silver.Mercedes=Colour.vehicles.table[9]
Silver.Mercedes
```

```
## Mercedes
##        3
```

```
diff.proportion=(Silver.Landrover/a) - (Silver.Mercedes/b)
diff.proportion #Difference in proportion of Silver vehicles between Landrove
r and Mercedes
```

```
## Landrover
##      0.25
```

- Therefore, the difference in proportion between Silver Landrover and Mercedes would be 0.25. Meaning the proportion of Silver Landrover is 0.25 higher than Silver Mercedes.

Question 1.5:

The recent trend shows that people tend to buy more powerful vehicles. We would like to investigate whether there is a linear relationship between the registration year and the mean of the number of cylinders

a/ Decide if the mean numbers of cylinders and the registration year are linearly related?

b/ If so, compute the equation to predict mean number of cylinders by using the registration year and discuss the significance of this equation? What is your estimate of the population mean number of cylinders when the year is 1984?

```
#a/
min(vehicles$Year)
```

```
## [1] 1985
```

```
max(vehicles$Year)
```

```
## [1] 2021
```

```
#1985 to 2021
```

```
#plotting the mean in N.cylinders depended on registration year
mean.calculated=tapply(vehicles$Number.of.cylinders,vehicles$Year, mean)
mean.calculated#this calculate the mean of cylinders of vehicles which has th
e same registration year
```

```
##     1985     1986     1987     1988     1989     1990     1991     1992
## 6.833333 6.416667 7.875000 8.214286 9.111111 7.000000 8.538462 6.700000
##     1993     1994     1995     1996     1997     1998     1999     2000
## 7.400000 7.500000 7.647059 7.125000 7.666667 9.000000 5.916667 7.888889
##     2001     2002     2003     2004     2005     2006     2007     2008
## 6.000000 8.000000 7.733333 5.818182 8.000000 6.500000 7.133333 8.125000
##     2009     2010     2011     2012     2013     2014     2015     2016
## 6.363636 6.428571 6.466667 5.333333 7.333333 7.928571 6.176471 8.142857
##     2017     2018     2019     2020     2021
## 7.000000 6.000000 5.545455 5.047619 4.440000
```

```
year.registrated= c(1985:2021)
```

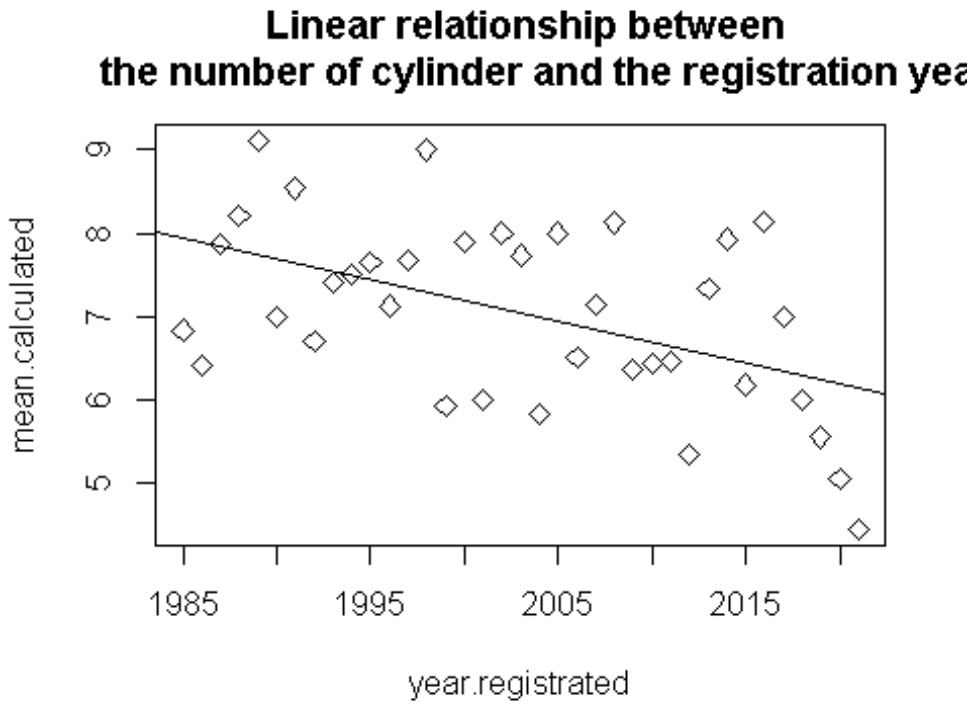
```
#h0: r  = 0
#hA: r != 0
r.Correlation=cor(mean.calculated,year.registrated, method="pearson")
r.Correlation # Correlation
```

```
## [1] -0.4919101
```

As we can see here r is a “some decreasing linear relationship”, as r is -0.4919101, which has intermediate downward slope.

Eventhough, It is a intermediate decreasing linear relationship, we can still plot the linear equation to predict the next value of y or x. And we can also find the slope coefficient and intercept value through the function lm() as below:

```
plot(y=mean.calculated, x=year.registrated, pch=5, main=" Linear relationship between
     the number of cylinder and the registration year") #mean number of cylinders against the registration year
abline(lm(mean.calculated~year.registrated))# there seems to have a down trend pattern but very weak correlation coefficient
```



- The graph above shows the linear relationship between the mean number of cylinders and the registration year, where the vertical vertical is the mean number of cylinders of the registration year and the horizontal is the registration year.
- The line across x and y is the line of best fit which is used to predict the mean number of cylinder of the registration year.

```r
#b/
#if so, compute the equation to predict mean number of cylinders by using the
#registration year and discuss the significance of this equation? what is you
r
#estimate of the population mean number of cylinders when the year is 1984?
#final equation would be:
fittest=lm(mean.calculated~year.registrated) #summarise linear model of the m
ean number of cylinder and the registration year
fittest

##
## Call:
## lm(formula = mean.calculated ~ year.registrated)
##
## Coefficients:
##      (Intercept)  year.registrated
##        107.01013          -0.04991

m= fittest[[1]][2]
m #Coefficient

## year.registrated
##      -0.04991196

b=fittest[[1]][1]
b #intercept

## (Intercept)
##    107.0101

x=1984
y=m*x+b
```

- Therefore, y is the predict number of cylinder when choosing the registration year.
- Therefore, y = -0.049912x + 107.0101277 is the new equation to find the predict number of cylinder depending on registration year. -when the year is 1984 then the number of cylinder will equal to 7.9848002

Question 2.1: Can a genuine causal relationship be established from this study? Justify your answer

This test genuinely cannot be concluded yet, multiple reasons are drawn:

-Small sample sizes: In the given data, it tells there is only a small proportion of marijuana users, which is around 30 people, tasked and compared to another small, proportional group of drug-free people (about 20 people). Therefore, when conducting a quantitative experiment, a smaller sample size will generalise the observed data, with less frequency of occurrence measured. As a result, the probability of drug effects on its users will vary. And it will be harder to draw out the final assumption about the drug effects. However, this data was complemented with qualities since both samples were collected from the same city and same group of age. This reduced errors in population, age group variations. Meaning that applying other populations or age groups rather than the same population group or age group may result differently. Causing the lack of accuracy, precision on the data analysis of the experiment.

Question 2.2: Can the results be generalised to other 14- to 16-year-olds? Justify your answer

This experiment cannot be generalised to other 14 to 16 years old because:

Until when the experiment is tested with large sample sizes and data is recorded from a variety of populations with other groups of age 14 to 16, which then gives out a common result on the effect of marijuana users. Then, the results can be generalised to other 14 to 16 years old in a professional manner with an accurate state of the level of confidence that marijuana causes short-term memory in adolescents. Because there are variabilities occurred in the experiment:

Variabilities:

Firstly, the temporal variation in kids' brain's development. Even though, the experiment was conducted amongst 14-16 years old kids, some kids may have their brain developing slower than others of the same ages. It means that some are smarter than the others. This causal event coincided with the small sub-populations may cause the bad test's results to happen by chances. Therefore, the drug users' group that was held drug-free for the next six weeks, could be one of the small proportions of a large population group of 14 -16 years old drug users that has their neuropsychological tests resulted badly. Meanwhile, the other group of non-drug users may be the advantaged group that has their brain developing faster, which is why their result was better than the drug users' group.

Secondly, variation in populations. The test cannot be generalised to others 14 to 16 years old without conducting the experiments on different groups of ethics, genders. The tested sample population is not representative to all other populations. Particularly, this

experiment was conduct based on two sub-population samples from a same population, without further specifying what genders and ethnicity they are, leading to a poor design for the drug effects' experiment. Because drug effects may affect differently to genders and others ethnicity.

Thirdly, the average dosage per user with their age ranging from 14 to 16 years old. The consequences of using drug vary depending on the average dosage of users and their ages. It seems likely that the younger they are, the more vulnerable and sensitive they are to the effects of drug. And the more surmountable dosage of drug they use, the more it affects the brain in term of memory. If these factors are mentioned in the quantitative analysis, the data recorded will lead to measurement errors.

Question 2.3: What are some potential confounding factors

There are some potential confounding factors occurred in the experiments, such as:

The designation of the sample sizes of the drug effects' experiment:

-The effect of small sample size would likely to cause the sub-populations of interest to coincide with features by chances, meaning that the Marijuana's drug effect tested on the two initial groups happened due to chances. Therefore, the differences could be the random choosing of groups from the same population, that becomes confounding factors to the drug test.

The average dosage per users depending on their current ages:

-As mentioned before in the previous answer, the drug dosage can affect differently on its users and ages can also involve in causing those effects. But it is not clear, which ones is the main factor. Therefore, both ages and drug dosages are the confounding factors.

Population variations and gender types:

-As mentioned before in the previous answer, different populations may have varied reactions to drug effects, some will react violently to the drug effects and some will not react at all, as well as for different genders. Thus, the neuropsychological test on drug users from 14 to 16 years old, compared to other non-drug users of the same ages in the same population, has happened due to chances. But it may not be the same to other populations and their gender types. Therefore, both population variations and gender types are the confounding factors.

THE END