Nature of Data Assignment

**DECLARATION**

Name: Hari Krishnan Suresh Varma

Student ID: 22056724

Subject Code: MATH 7016

I, the author 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.

rides <- read.csv("rides.csv")  
head(rides)

## TripId DriverAge DriverGender PassengerGender DriverRating PassengerRating  
## 1 1 56 Male Male 4 3  
## 2 2 62 Female Male 4 5  
## 3 3 50 Male Male 4 4  
## 4 4 28 Male Non-binary 5 4  
## 5 5 45 Non-binary Male 4 4  
## 6 6 54 Male Female 3 3  
## PickupLoc DropoffLoc Fare TripDist Duration Weather VehicleType  
## 1 Bondi Olympic Park 97.0 24.1 40.3 Cloudy SUV  
## 2 Redfern Coogee 30.0 8.8 22.7 Rainy Sedan  
## 3 Bondi Redfern 43.7 7.6 24.6 Sunny Sedan  
## 4 Surry Hills Kings Cross 83.9 20.4 28.1 Clear Economy  
## 5 Kings Cross Manly 50.8 18.7 42.5 Cloudy Van  
## 6 Darlinghurst Parramatta 105.1 33.6 47.6 Clear Sedan  
## PickupTime Tip DayofWeek  
## 1 Early Morning 5 Tuesday  
## 2 Night 2 Friday  
## 3 Morning 3 Thursday  
## 4 Afternoon 2 Sunday  
## 5 Morning 1 Saturday  
## 6 Afternoon 4 Saturday

# Question 1

**Test if there is a significant association between the passenger’s rating and the passenger’s gender for the rides taken in the mornings.**

**What does being associated mean in this context? Interpret your findings.**

## ANSWER:

From the data set, and the question statement, we have to find the association as described in the question. And, as the data we need to analyse is categorical, we can use chi-squared analysis to find the association between the two.

### STEP 1: Set the Hypothesis

H0: There is no Association between passenger rating and passenger Gender for rides taken in mornings.

H1: There is significant association between passenger rating and passenger Gender in mornings.

As significance level is not mentioned, we can take it as 0.05

### STEP 2: Create the subsets for Morning rides

Let’s create subsets of passenger rating and gender where Pickup time is ‘Morning’ and ‘Early Morning’.

passengerRating <- subset(rides, PickupTime == "Morning" | PickupTime == "Early Morning", PassengerRating, drop = T)  
passengerGender <- subset(rides, PickupTime == "Morning" | PickupTime == "Early Morning", PassengerGender, drop = T)

### STEP 3: Create the contingency table

We create a contingency table so that we can identify patterns and relationships and calculate expected frequencies.

tab <- table(passengerGender, passengerRating)  
tab

## passengerRating  
## passengerGender 2 3 4 5  
## Female 2 32 23 5  
## Male 7 37 43 6  
## Non-binary 0 5 10 0

### STEP 4:

Let’s check whether we have at least 5 counts in each cells in the expected values for 10000 samples

chisq.test(tab, B = 10000)$expected

## Warning in chisq.test(tab, B = 10000): Chi-squared approximation may be  
## incorrect

## passengerRating  
## passengerGender 2 3 4 5  
## Female 3.2823529 26.988235 27.717647 4.0117647  
## Male 4.9235294 40.482353 41.576471 6.0176471  
## Non-binary 0.7941176 6.529412 6.705882 0.9705882

As we can see, there are cells which doesn’t have 5 counts. This will cause chi-square test to give incorrect results.

Therefore, we have to set simulate.p.value = TRUE to do the chi-sq test. this allows us to obtain p-value based on a simulated distribution instead of asymptotic chi-square approximation.

### STEP 5: conduct chi square test.

Let’s do the chi-square test after setting simulate p value to TRUE.

chisq.test(tab, simulate.p.value = T, B = 10000)

##   
## Pearson's Chi-squared test with simulated p-value (based on 10000  
## replicates)  
##   
## data: tab  
## X-squared = 7.4433, df = NA, p-value = 0.2684

### STEP 6: Conclusion.

From the test, we have a p-value of **0.27**. Assuming the alpha value is **0.05**, the p-value is much higher than the significance level.

From this, we can say that we do not have significant evidence to reject the null hypothesis. That is, There is no association with Passenger rating and Passenger gender for rides taken in mornings.

The meaning of being associated in this context is that whether a person’s gender influences them in their rating. Passenger Rating is given considering factors like **how well they behave with the driver, do they pay fares correctly**, etc. Therefore, we are **trying to find the influence of gender in these factors**.

# Question 2

**Test whether the mean fare charge for rides on weekends are less than those in weekdays.**

## ANSWER

### STEP 1: Set the Hypothesis

H0: Fares on weekends are greater than or equal to weekdays.

H1: Fares on weekends are less than those in weekdays.

As significance level is not mentioned, we can take it as 0.05

### STEP 2: Define subsets

Now that we have set the hypotheses, lets create subsets for which we have to do the test.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

weekendFares <- rides %>% filter(DayofWeek == c("Sunday", "Saturday")) %>% select(Fare)  
weekdayFares <- rides %>% filter(DayofWeek != c("Sunday", "Saturday")) %>% select(Fare)

### STEP 3: Sample size.

Let’s check the sample size.

nrow(weekdayFares)

## [1] 412

nrow(weekendFares)

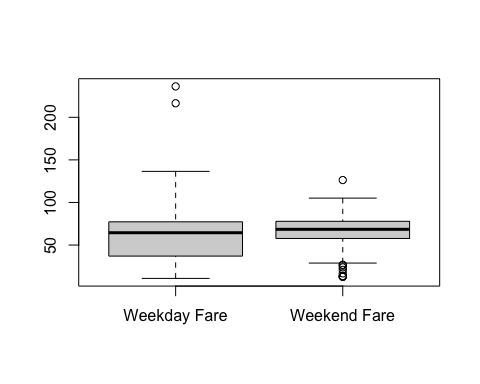
## [1] 64

The sample size is large enough to do an approximation test and as it is large enough, we can assume the sample data is normally distributed.

But we also have to check whether the variances are equal. So, let’s plot a box plot and do a var test to find that out.

### STEP 4: Check the variance.

boxplot(weekdayFares$Fare, weekendFares$Fare, names = c("Weekday Fare", "Weekend Fare"))



From the box plot, we can see **the variance are not equal**.

Therefore, we have to use the **Welch Two sample t-test**. Let’s do that.

### STEP 5: Conduct t test.

t.test(x = weekendFares$Fare, y = weekdayFares$Fare, var.equal = F, paired = F, alternative = 'less')

##   
## Welch Two Sample t-test  
##   
## data: weekendFares$Fare and weekdayFares$Fare  
## t = 0.98857, df = 91.355, p-value = 0.8373  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 8.838962  
## sample estimates:  
## mean of x mean of y   
## 63.20938 59.91238

The p-value is found to be **0.83**.

### STEP 6: Conclusion.

If we assume the null hypothesis is true, **the range of 95 % confidence interval should include 0 and positive values of difference in mean of weekend and weekday fares**. The test has shown the confidence interval from **-inf to 8.83**, meaning **it includes 0 and positive values.**

Considering the p-value and confidence interval, **we do not have enough evidence to reject the null hypothesis, That is, There is no evidence that mean weekend fare is less than mean weekday fares.**

# Question 3

**Compute the 96% confidence interval for the difference in the mean fare charged of rides taken by passengers who provided a tip versus those who did not provide a tip.**

1. **First, use bootstrapping to compute the confidence interval.**
2. **Then approximate the confidence interval based on a t-distribution.**
3. **How do the results compare? Justify your answer.**

## ANSWER:

### 1. First, use bootstrapping to compute the confidence interval.

We have to create subsets of fares of passengers who provided a tip and who did not.

tipped <- rides %>% filter(Tip != 0.0) %>% select(Fare, Tip)  
notTipped <- rides %>% filter(Tip == 0.0) %>% select(Fare, Tip)

Now lets do bootstrapping.

sim <- replicate(1000, {  
 tipSample <- sample(tipped$Fare, replace = T, size = nrow(tipped))  
 notipSample <- sample(notTipped$Fare, replace = T, size = nrow(notTipped))  
 mean(tipSample) - mean(notipSample)  
})

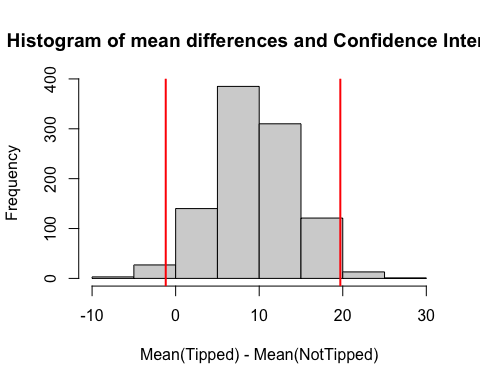
Find the 96% confidence interval.

CI <- quantile(sim, c(0.020, 0.980))  
CI

## 2% 98%   
## -1.185268 19.708446

Let’s plot the histogram to see where the CI falls among the samples.

hist(sim,  
 xlab = "Mean(Tipped) - Mean(NotTipped)",  
 main = "Histogram of mean differences and Confidence Interval")  
abline(v = CI, col = 'red', lwd = 2)



## 2. Then approximate the confidence interval based on a t-distribution.

Now, lets do the two samples t test to approximate the confidence interval.

t.test(x = tipped$Fare, y = notTipped$Fare, var.equal = T, conf = 0.96)

##   
## Two Sample t-test  
##   
## data: tipped$Fare and notTipped$Fare  
## t = 1.7416, df = 474, p-value = 0.08223  
## alternative hypothesis: true difference in means is not equal to 0  
## 96 percent confidence interval:  
## -1.735692 20.755522  
## sample estimates:  
## mean of x mean of y   
## 60.89510 51.38519

## 3. How do the results compare? Justify your answer.

From the two methods that I have performed above, one is Bootstrapping method and another is two samples t-test. Both tests gave me almost same values for 96% confidence interval.

Bootstrapping: **-1.18 to 19.70.**

T-Test: **-1.73 to 20.75.**

By comparing the results, we can conclude it like given below,

* There are some assumptions T-test have made
  1. The data is normally distributed.
  2. The variances of both sample are equal.
* as t.test and bootstrapping methods gave almost same confidence intervals, **it is safe to assume that the data is normally distributed and variances of both sample are equal.**
* We can say this by comparing it with the bootstrapping result because, bootstrapping does not do assumptions while doing the test. Therefore, the results are more robust and precise compared to an approximation test. As the t.test gave us similar values, we can conclude like given in the previous bullet.

# Question 4

**Test if the mean fare charged for rides taken is different for each time of the day. If so, find which day time has the highest fare charged.**

## ANSWER:

### STEP 1: Set the Hypothesis.

H0: The mean fare for each Pickup Time are equal (µ0)

HA: The mean fare for at least one Pickup Time are not equal. (µA)

Significance Level: 0.05

### STEP 2: Check the Normality

To check the normality, let’s check the sample size.

temp\_data <- rides %>% select(PickupTime, Fare)  
table(temp\_data$PickupTime)

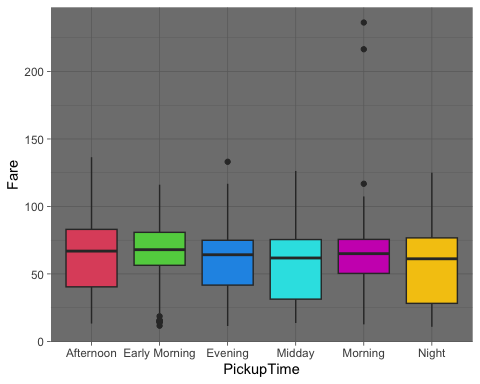
##   
## Afternoon Early Morning Evening Midday Morning   
## 65 86 102 79 84   
## Night   
## 60

Since the sample size is large enough (>30), we can safely assume that **The sample data is normally distributed.**

### STEP 3: Plot a box plot.

Lets plot a box plot between Pickup Time vs Fare.

library(ggplot2)  
ggplot(data = rides, aes(x = PickupTime, y = Fare)) +  
 geom\_boxplot(fill = 2:7) +  
 theme\_dark()



From the above Box plot here are the findings.

* The **Variances seems to be different**.

Let’s do one way test to find whether for all Pickup Times, is the fare same or not. Here we give var.equal = Fbecause from the box plot, we understood the variance is not equal.

### STEP 4: One way test.

oneway.test(formula = Fare ~ PickupTime,  
 data = rides,  
 var.equal = F)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Fare and PickupTime  
## F = 1.3189, num df = 5.00, denom df = 208.04, p-value = 0.2573

From the test, We have **a p-value of 0.25**.

As the p-values from both the tests are higher than the significance level of 0.05, we can say that **we do not have significant evidence to reject the null hypothesis**. That is, **There is no statistically significant difference between the mean fare in at-least two different Pickup Times.**

As there is no statistically significant difference between mean fares of at-least 2 groups, we cannot find which Pickup time gives a higher fare to other.

# Question 5

1. **Draw an appropriate plot to show the relationship between the distance of the ride and the duration of the ride taken on rainy days. Interpret your plot.**
2. **Test if there is a linear relationship between the distance of the ride and the duration of the ride taken on rainy days.**
3. **Can we predict the distance of the ride based on the duration of the time taken on rainy days?**
4. **If so predict the distance of the ride when duration of the time is 27.7 minutes.**
5. **How good is your estimate? Discuss the suitability and/or strength of your model.**

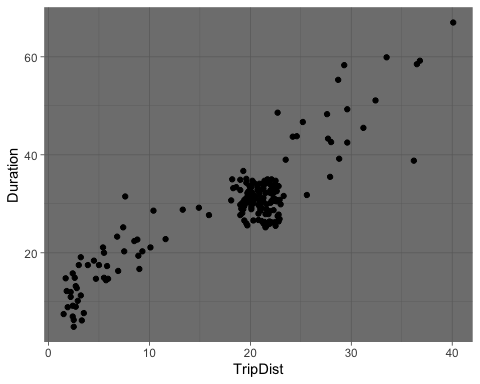
## ANSWER:

### 1. Draw an appropriate plot to show the relationship between the distance of the ride and the duration of the ride taken on rainy days. Interpret your plot.

As per the question, the independent variable here is distance and duration depends on distance. so we can plot the scatter plot as shown below.

distNDuration <- rides %>% filter(Weather == "Rainy") %>% select(TripDist, Duration)

ggplot(data = distNDuration, aes(x = TripDist, y = Duration)) +  
 geom\_point() +  
 theme\_dark()



Let’s find the correlation between them.

obsPearsonCor <- cor(x = distNDuration$TripDist, y = distNDuration$Duration)  
print(paste("Pearson Correlation:", obsPearsonCor))

## [1] "Pearson Correlation: 0.88801367051676"

In the above scatter plot, Trip Duration is plotted against Trip Distance when weather is rainy. After analyzing the plot and correlation value,

* We can see that generally, the **trip duration increases with Trip Distance.**
* Most trip durations taken on rainy days are **around** **20 km** and it takes **around** **30 mins** to complete the trip.
* From the Pearson correlation value, the relationship between trip distance and duration, **shows a positive linear relationship**.

### 2. Test if there is a linear relationship between the distance of the ride and the duration of the ride taken on rainy days.

### STEP 1: Set the hypothesis

* H0: There is no significant linear relationship between Trip duration and Trip Distance
* H1: There is a linear relationship between Trip duration and Trip Distance.

As significance level is not mentioned, we can take it as 0.05

### STEP 2: Sample correlation

We have already found the sample correlation in the previous question.

obsPearsonCor

## [1] 0.8880137

Its clear that the sample follows a strong linear relationship. But what about the population ?

Let’s find it using simulation and replicate 10,000 samples.

### STEP 3: Simulation to find p-value.

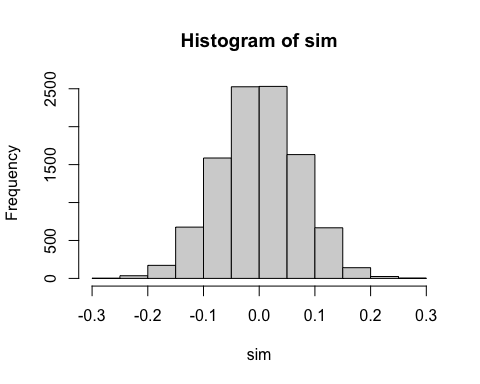
The reason we use simulation is to find the probability (**p-value**) of getting a correlation value as strong as or greater than the actual sample.

Here, we assume the null hypothesis is true.

sim <- replicate(10000, {  
 sample <- sample(distNDuration$TripDist)  
 cor(x = sample, y = distNDuration$Duration)  
})

Now, using sim, lets plot a histogram to show the range of correlation and its frequency.

hist(sim)



From this simulated cor values, we can find the p-value.

pVal <- mean(abs(sim) > abs(obsPearsonCor))  
pVal

## [1] 0

**The P-value is 0.**

### STEP 4: Bootstrapping

Now, using the bootstrapping method, let’s find the confidence interval. Here we will sample it without breaking the relationship between the pairs.

n <- nrow(distNDuration)  
bootstrap <- replicate(10^4, {  
 sample <- sample(1:n, size = n, replace = T)  
 cor(x = distNDuration$TripDist[sample], y = distNDuration$Duration[sample])  
})

now that we have correlation for 10,000 samples, let’s find the 95% confidence interval.

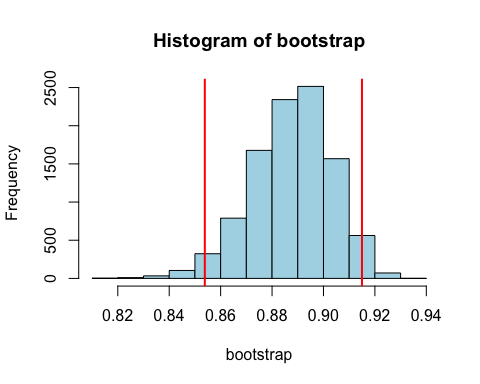
### STEP 5: Find the 95% confidence interval.

CI <- quantile(bootstrap, c(0.025, 0.975))  
CI

## 2.5% 97.5%   
## 0.8538448 0.9149729

now let’s use this confidence interval and plot it on the bootstrap histogram.

hist(bootstrap, col = 'lightblue')  
abline(v = CI, col = 'red', lwd = 2)



### STEP 6: Conclusion

From the test two results are known,

1. **The P-value is 0**
2. **The confidence interval ranges from 0.854 to 0.914.**

From this we can conclude that,

* The probability of getting a correlation value equal to or higher than the actual sample correlation assuming null hypothesis is true is 0.
* The confidence interval does not contain a correlation value of 0, which is consistent with the alternate hypothesis.

Therefore, we **have the evidence to reject the null hypothesis**. That is, **There is significant relationship between Trip duration and the trip distance.**

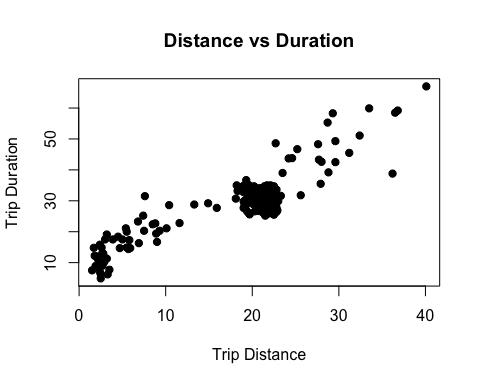
## 3. Can we predict the distance of the ride based on the duration of the time taken on rainy days ?

ANS:

Yes we can predict the distance of ride based on the duration of the time taken on rainy days. As **the sample data follows a strong linear relationship**, we can plot a line which best suits the data using **Linear Model function.**

To plot a line, first we have to create a scatter plot.

plot(x = distNDuration$TripDist, y = distNDuration$Duration, pch = 19,  
 xlab = "Trip Distance",  
 ylab = "Trip Duration",  
 main = "Distance vs Duration")



On this scatter plot, we can plot the line.

(fit <- lm(Duration ~ TripDist, data = distNDuration))

##   
## Call:  
## lm(formula = Duration ~ TripDist, data = distNDuration)  
##   
## Coefficients:  
## (Intercept) TripDist   
## 8.825 1.123

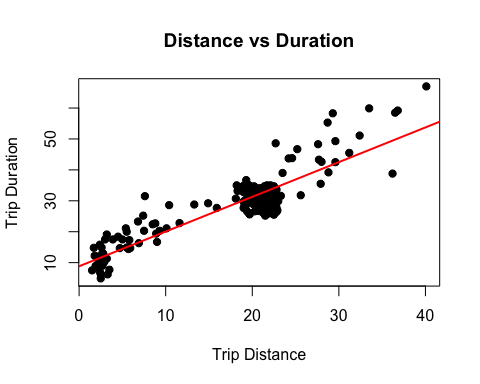
The above line of code gave us,

* **y - intercept of line = 8.82**
* **Slope of the line = 1.12**

From this we can write the equation of line as,

Now, let’s plot this line on the scatter plot.

plot(x = distNDuration$TripDist, y = distNDuration$Duration, pch = 19,  
 xlab = "Trip Distance",  
 ylab = "Trip Duration",  
 main = "Distance vs Duration")  
abline(fit, col = 'red', lwd = 2)



With this line we can predict the Trip Duration using trip distance.

### 4. If so predict the distance of the ride when duration of the time is 27.7 minutes.

ANS:

Calculating it manually,

We have been given the duration of time as 27.7 mins

That is, **y = 27.7**

Therefore, from the line that we have derived, **Trip distance (x)** is,

x <- (27.7 - 8.82)/1.12  
x

## [1] 16.85714

**Therefore, the trip distance for a trip duration of 27.7 mins is 16.8 Kms.**

### 5. How good is your estimate? Discuss the suitability and/or strength of your model.

The strength or suitability of the model can be understood from the summary of the linear model.

summary(fit)

##   
## Call:  
## lm(formula = Duration ~ TripDist, data = distNDuration)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.6734 -3.0408 -0.3337 1.9620 16.5745   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.82532 0.84494 10.45 <2e-16 \*\*\*  
## TripDist 1.12287 0.04252 26.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.752 on 187 degrees of freedom  
## Multiple R-squared: 0.7886, Adjusted R-squared: 0.7874   
## F-statistic: 697.4 on 1 and 187 DF, p-value: < 2.2e-16

From the summary for the linear model,

* The **Multiple R-squared value is 0.78** on a scale of 0 to 1. T**his says that the 78% variance of Trip Duration are explained by the Trip Distance info** and the remaining 22% is explained by other factors.
* The **p-value is 0.0**, which is an **evidence to reject the null hypothesis**, which says that ‘there is no relationship between predictor and response variables’.

From all these values, we can estimate that **this a strong linear model which has strong relationship between the predictor and response variables.**