This is a Rscript converted to a pdf file.

```
# Loading the dataset
setwd("C:/Users/shiva/Documents/MS Data Science/631-Foundations of Data Analysis/Class12")
load("hr.RData")
View(hr)
#*************************************************************************
```

Paired t test

```
# x(diff) = -3.33   s(diff) = 4.88
x_diff <- -3.33
s_diff <- 4.88
n <- 30
```

1. Hypothesis test using t-test

```
# H0: Mu_post – Mu_pre = 0
# HA: Mu_post – Mu_pre ≠ 0

# Checking for conditions
# 1. Each patient is independent of each other.
# 2. By plotting a histogram of the difference between post and pre heart rates we can eliminate the strong skew problem.
hist(hr$postHR-hr$preHR)

# Applying paired t-test
#Standard Error
(SE <- s_diff/sqrt(30))
#T-Score
(t <- x_diff - 0/SE)
```

#Degress of freedom

(df <- n-1)

#p-value

(p_value <- pt(t,df,lower.tail = TRUE)*2)

Conclusion:

p-value is 0.002 which is less than aplha 0.05. Hence we can reject the null hypothesis.

We have enough evidence to say that there is difference between heart rates of the patient before and after the drug.

Since, the sample difference is negative (after-before) which means heart rate before consuming the drug was high and after consuming the drug heart rate was lowered.

2.Hypothesis test using t.test() function

?t.test

t.test(hr$postHR, hr$preHR, paired = TRUE, alternative = "two.sided")

```
        Paired t-test

data:  hr$postHR and hr$preHR
t = -3.7398, df = 29, p-value = 0.000807
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.151131 -1.508869
sample estimates:
mean of the differences
                 -3.33
```

Conclusion:

p-value = 0.0008 which is less than 0.05, so null hypothesis is rejected.

We have enough evidence that there is a difference between heart rates before and after consuming the drug.Results are close with manual calculations and t.test() function.

#**********#

2-sample t-test

x_f_hr <- 75.2 # Mean female heart rate before drug treatment

s_f_hr <- 12.9 # Standard deviation for female's heart rate

n_f = 15 # Number of females

x_m_hr <- 67.8 # Mean male heart rate before drug treatment

s_m_hr <- 14.3 # Standard deviation for female's heart rate

n_m <- 15 # Number of males

3. Hypothesis Test without using t.test() function

H0: Mu__pre_male – Mu_pre_female = 0

HA: Mu_pre_male – Mu_pre_female ≠ 0

Checking for conditins

1. Observations are independent of each other.

2. By plotting box plots for heart rate of males and females before drug treatment we can say that, there is no strong skew for males although for females data is slightly skewed.

Since, sample size is sample size < 30 we assume the skewness to be normal.

boxplot(hr$preHR ~ hr$gender)

#Standard error

(SE <- sqrt((s_f_hr^2/n_m)+(s_m_hr^2/n_f)))

#T_df

(t <- ((x_f_hr-x_m_hr)-0)/SE)

#Degrees of freedom

(df <- n_f + n_m - 2)

#p-value

(p_value <- pt(t,df,lower.tail = FALSE)*2)

Conclusion:

p-value is 0.14 which is less than alpha 0.05. Hence we can reject the null hypothesis.

We have enough evidence to say that the males and females have different average heart rates before drug treatment.

4. 95% condidence interval

point_estimate +/- ME

ME = t_df * SE

#t*

(t <- qt(0.025,df,lower.tail = FALSE))

Margin of Error(ME)

(ME <- t*SE)

(upper_bound <- (x_f_hr-x_m_hr) + ME)

(lower_bound <- (x_f_hr-x_m_hr) - ME)

The average difference lies between (-2.78,17.58)

#5.Hypothesis Test using t.test() function

t.test(hr$preHR ~ hr$gender, paired = FALSE, alternative = "two.sided")

```
        Welch Two Sample t-test

data:  hr$preHR by hr$gender
t = 1.4868, df = 27.693, p-value = 0.1484
alternative hypothesis: true difference in means between group f and group m is not equal to 0
95 percent confidence interval:
 -2.800379 17.600398
sample estimates:
mean in group f mean in group m
           75.2            67.8
```

Conclusion:

p-value is 0.14 which is less than aplha 0.05. Hence we can the reject the null hypothesis.

We have enough evidence to say that the males and females have different average heart rates before

drug treatment.

The results from manual calculation and t.test() are very close.

#**#

ANOVA

6. Hypothesis test

H0: Mu_pre_young = Mu_pre_middle = Mu_pre_elderly

HA: At least mean for one group is different.

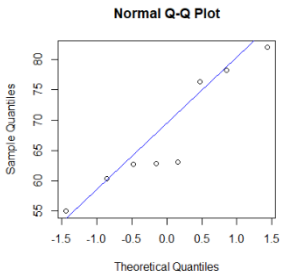
#Checking for conditions:

1.Observations are independent of each other for all age groups.

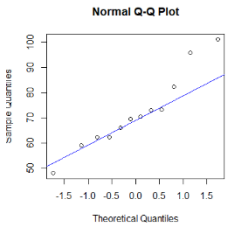
2.By plotting qq plots we can check for normal distribution

```
library(tidyverse)
```

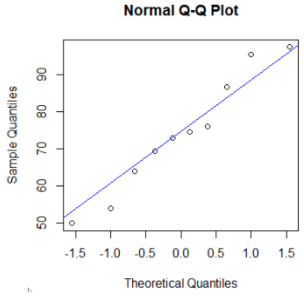
```
# Checking normality for Elderly age group
preHR_el <- select(filter(hr,hr$age=="elderly"),preHR)
qqnorm(preHR_el$preHR)
qqline(preHR_el$preHR, col = "blue", lwd = 1)
```




```
# Checking normality for Middle group
preHR_mi <- select(filter(hr,hr$age=="middle"),preHR)
qqnorm(preHR_mi$preHR)
qqline(preHR_mi$preHR, col = "blue", lwd = 1)
```

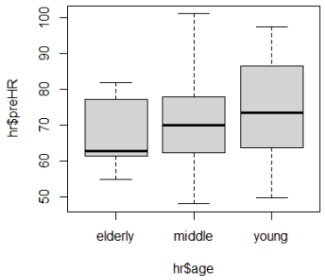



```
# Checking normality for Young group
preHR_yo <- select(filter(hr,hr$age=="young"),preHR)
qqnorm(preHR_yo$preHR)
qqline(preHR_yo$preHR, col = "blue", lwd = 1)
```



#3. Checking for constant variance by plotting box plot all ages.

boxplot(hr$preHR ~ hr$age)



the variance is different for all the age groups.

Total sample size

n<- 30

Total groups

k <-3

Determine df for groups, error and total.

(dfg <- k-1)

(dft <- n-1)

(dfe <- dft-dfg)

#Summary statistics

```
df_stat <- hr %>%
 group_by(age) %>%
 summarise(n_prehr = n(),mean_prehr = mean(preHR),sd_preHr=sd(preHR))
```

```
  age      n_prehr mean_prehr sd_preHr
  <fct>      <int>      <dbl>    <dbl>
1 elderly        8       67.6     9.81
2 middle        12       72.0    15.0
3 young         10       74.0    15.9
> 
```

(mean <- mean(df_stat$mean_prehr))

(Total_sample <- sum(df_stat$n_prehr))

```
# Determine Sum of squares total

(SST <- sum((hr$preHR-Total_mean)^2))

5615.529

# Determine Sum of squares between groups

(SSG <- sum(df_stat$n_prehr * (df_stat$mean_prehr - Total_mean)^2))

191.0765

# Determine sum of squares error

(SSE <- SST - SSG)

5424.453

# Determine mean square error

(MSE <- SSE/dfe)

200.9057

# Determine mean square for groups

(MSG <- SSG/dfg)

95.53824

# F-statistics

(f <- MSG/MSE)

0.4755378

# Determine p-value

(pf(f,dfg,dfe,lower.tail = FALSE))

0.6266582

#Conclusion:

#p-value is 0.6266582 which is greater than 0.05. Hence we will fail to reject

#the null hypothesis.

#We don't have enough evidence to say that at hear rate for before drug treatment is different for at least one age group.
```

#**********#

Simple Linear Regression

7.Determine slope and intercept

#a. slope = b1

#Equation of a regresison line is given by y-y0 = b1(x-x0)

x0 <- 112.22

y0 <- 73.60

sx <- 17.96

sy <- 66.06

R <- 0.374

y-73.60 = b1(x-112.22)

(b1 <- (sy/sx)*R)

#The slope is 1.375.Final equation of the model can be written as y-73.60 = 1.375(x-112.22)

#On simplifying, y = 1.375x - 80.7

#For each additional minute of runtime the budget will be different by 1.375 factor.

#b. intercept = b0

y-y0=b0 + b1(x-x0)

b0= -80.7

#When the runtime of a movie is zero minutes the budget will be -80.7.

8. Interpreting slope and intercept

The slope of 1.375 means that for each additional min of runtime the budget will increase by a factor of 1.375.

#The -80.7 intercept is negative. Since, the budget can't be negative when the runtime is zero, so it is not meaningful in this scenario.

9. Using equation y = 1.375x - 80.7, substitute the value of x as 120 mins.

(y <- (1.375 * 120) - 80.7)

For runtime of 120 mins the budget will be 84.3 million.

10. R-squared = R^2

(R_squared <- R^2)

The R-squared is 0.139. This linear model accounts for 14% variability in the explanation of budget values.