Practice Session_answers _7

Intro

The focus of this practice session will be to perform hypothesis tests for the difference of two or more means. We will look at various test statistics that can be used. We will also conduct hypothesis testing for correlations.

Question 1: Exercise Hypothesis test for the difference of two means

A study is interested to check if the mean exercise hours for female are less than the mean exercise hours for male students. Use data `ExerciseHours` and the two variables `Exercise` and `Sex`.

1.) **Step 1**: Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

a.) Create a boxplot to describe hours of exercise for `female` versus `male`.

```
# your code here
```

b.) Find some favorites statistics of `Exercise hours` for female and male students. You might find the function: `mosaic::favstats` useful. *Note*: you can search online for this function arguments.

```
#your code here
```

c.) Subset the data `ExerciseHours` to two groups: `F` and `M`.

```
#your code here
```

2.) **Step 2**: Compute the observed statistic (mean difference of exercise hours for Female and Male).

```
#your code here
```

3.) **Step 3**: Create null hypothesis distribution

a.) Shuffle the two groups of `female` and `Male` into two samples, and find the mean difference of the two shuffled samples.

b.) Create the Null hypothesis Distribution using `do_it()` function.

c.) Plot a `histogram' of the null distribution and show the`line`of the`observed mean difference`using the`abline()‘ function.

```
# your code here
```

4.) **Step 4:** Calculate p-value

```
# your code here
```

Step 5: Make decision/Judgment

```
#your code here
```

Answers:

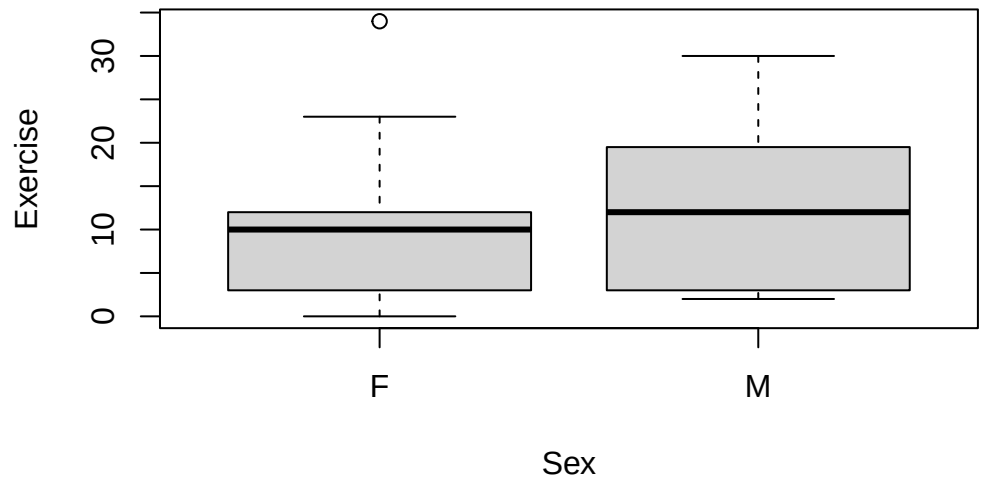
```
library(Lock5Data)
library(SDS1000)
data(ExerciseHours)
```

1.) **Step 1**: Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

$H_0 : \mu_f = \mu_m$ vs $H_a : \mu_f < \mu_m$

a.) Create a boxplot to describe hours of exercise for `female` versus `male`.

```
boxplot(Exercise ~ Sex , data = ExerciseHours)
```



b.) Find some favorites statistics of `Exercise hours` for female and male students. You might find the function: `mosaic::favstats` useful.

Note: you can search online for this function arguments.

```
mosaic::favstats( Exercise ~ Sex, data = ExerciseHours)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2

  Sex min Q1 median    Q3 max mean       sd  n missing
1   F   0  3     10 12.00  34  9.4 7.407359 30       0
2   M   2  3     12 19.25  30 12.4 8.798325 20       0
```

c.) Subset the data `ExerciseHours` to two groups: `F` and `M` using `subset()` function.

```
# we will use the function `subset`
excercise_fem<- subset(  ExerciseHours$Exercise, ExerciseHours$Sex == "F")
excercise_fem
```

```
 [1]  2 10 14 10 12 10  0 10 12  5  3 23  2  3 10 10  1  2 20 15  1 10  3 34  8
[26]  7 10  6 17 12
```

```r
excercise_mal<- subset(  ExerciseHours$Exercise, ExerciseHours$Sex == "M")
excercise_mal
```

```
 [1] 15 20  8 14  2  3  3  2 10 30 19 20  8  2  3 24 27 14 10 14
```

```r
length(excercise_fem)
```

```
[1] 30
```

```r
length(excercise_mal)
```

```
[1] 20
```

```r
## 30
## 20
```

2.) **Step 2**: Compute the observed statistic (mean difference of exercise hours for Female and Male).

```r
obs_stat <- mean(excercise_fem) - mean(excercise_mal)
obs_stat
```

```
[1] -3
```

```r
## -3
```

3.) **Step 3:** Create null hypothesis distribution

a.) Shuffle the two groups of `female` and `Male` into two samples, and find the mean difference of the two shuffled samples.

```r
combined_sample <- c(excercise_fem, excercise_mal)
shuffled_sample <- sample(combined_sample )

shuff_fem <- shuffled_sample[1:30]
shuff_mal <- shuffled_sample[31:50]

shuff_stat <- mean(shuff_fem) - mean(shuff_mal)
shuff_stat
```

```
[1] 2.583333
```

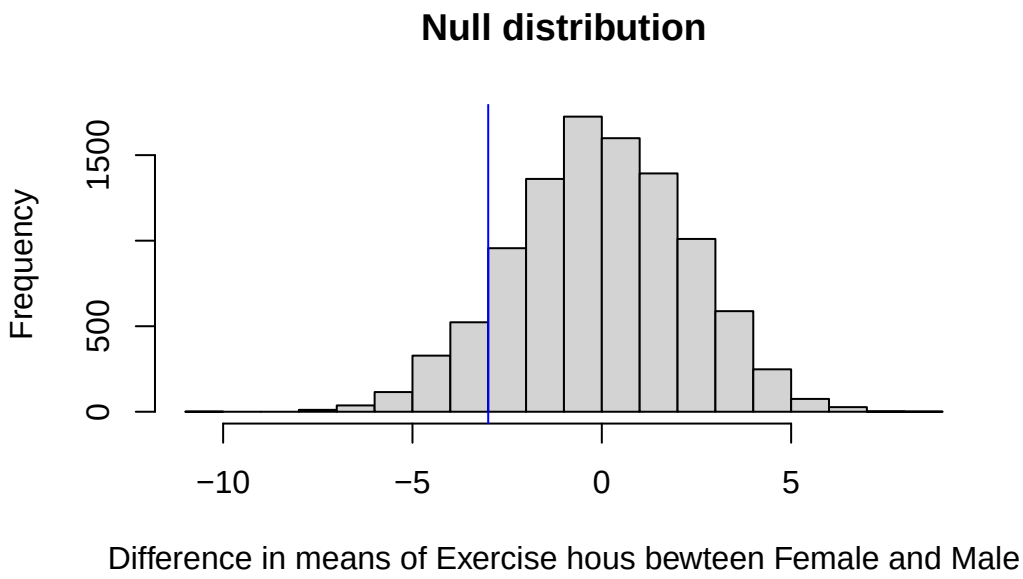
```
# answers may vary
```

b.) Create the Null hypothesis Distribution

```
null_dist <- do_it(10000) * {
shuffled_sample <- sample(combined_sample )

shuff_fem <- shuffled_sample[1:30]
shuff_mal <- shuffled_sample[31:50]

shuff_stat <- mean(shuff_fem) - mean(shuff_mal)
}
```

c.) Plot histogram of the null distribution and show the line of the observed mean difference

```
hist(null_dist , xlab = "Difference in means of Exercise hous bewteen Female and Male", main

abline(v = obs_stat, col = "blue")
```



4.) **Step 4:** Calculate p-value

```
p_value <- pnull(obs_stat, null_dist, lower.tail = T)
p_value
```

```
[1] 0.1015
```

```
#0.1038 (# answers may vary)
```

5.) **Step 5:** Make decision/Judgment

We fail to reject the **null hypothesis**. There are no enough evidence to conclude that there is a mean difference in Exercise Hours between Female and Male.

Question 2: Caffeine Taps

A sample of male college students were asked to tap their fingers at a rapid rate. The sample was then divided at random into two groups of ten students each. Each student drank the equivalent of about two cups of coffee, which included about 200 mg of caffeine for the students in one group but was decaffeinated coffee for the second group. After a two hour period, each student was tested to measure finger tapping rate (taps per minute). The goal of the experiment was to determine whether caffeine produces an increase in the average tap rate.

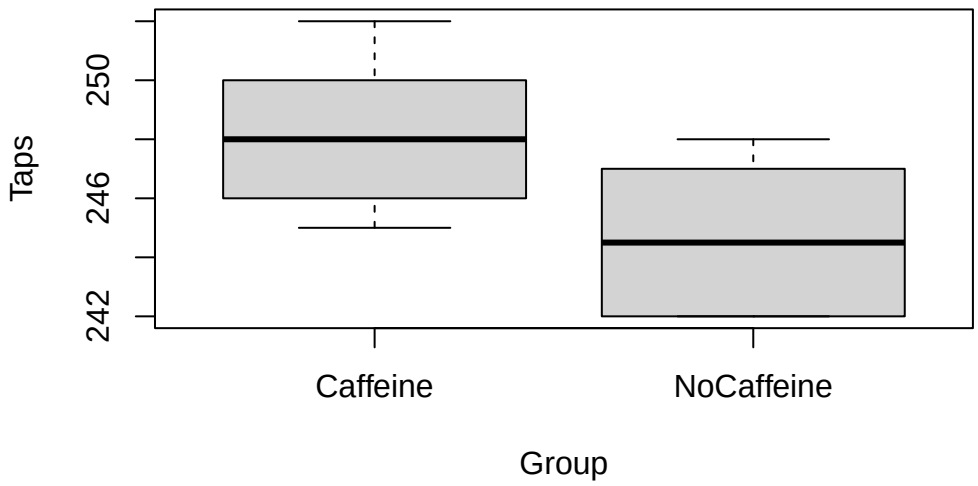
1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

- Null: There is no difference in the average taping rate between the caffeinated and decaffeinated groups
- Alternative: The average taping rate is greater for the caffeinated group than for the decaffeinated group

- $H_0 : \mu_c - \mu_{nc} = 0$
- $H_A : \mu_c - \mu_{nc} > 0$

2.) Create a boxplot to describe tap rates for `Caffeine` versus `No Caffeine`.

```
library(Lock5Data)
library(SDS1000)
boxplot(data = CaffeineTaps, Taps ~ Group)
```



3.) Find some favorite statistics to visualize the number of taps for the `Caffeine` and `No Caffeine` group.

```
mosaic::favstats(Taps ~ Group, data = CaffeineTaps)
```

```
       Group min    Q1 median     Q3 max  mean       sd  n missing
1   Caffeine 245 246.5  248.0 250.00 252 248.3 2.213594 10       0
2 NoCaffeine 242 242.5  244.5 246.75 248 244.8 2.394438 10       0
```

4.) Subset the data `CaffeineTaps` to two groups: `Caffeine` and `NoCaffeine`.

```
Caffeine = subset(CaffeineTaps$Taps, CaffeineTaps$Group == "Caffeine")
NoCaffeine = subset(CaffeineTaps$Taps, CaffeineTaps$Group == "NoCaffeine")
```

5.) Compute the observed statistic (mean difference of tap number for the two groups).

```
diff_caffeine = mean(Caffeine) - mean(NoCaffeine)
```

6.) Create null hypothesis distribution

- a.) Shuffle the two groups of `Caffeine` and `No Caffeine` into two samples, and find the mean difference of the two shuffled samples.

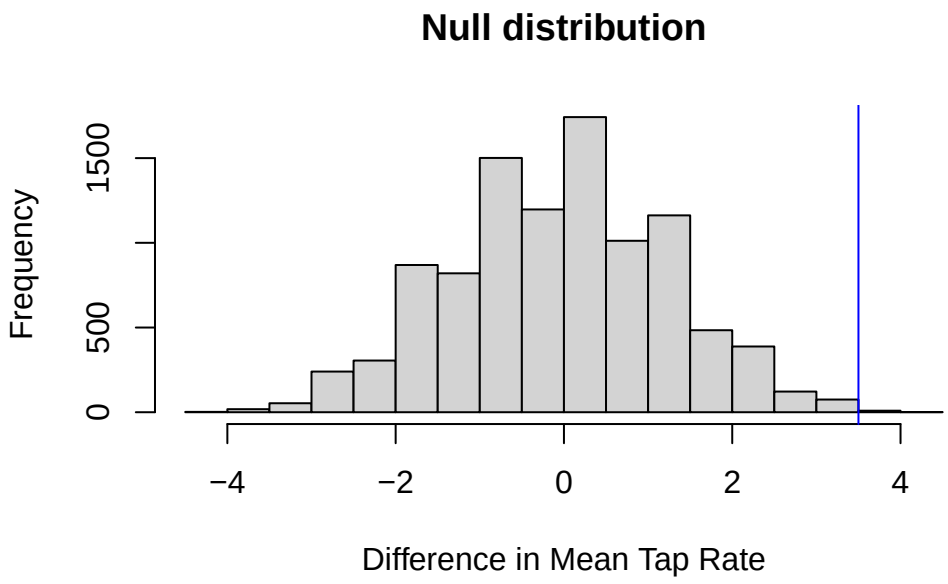
```r
combined_sample <- c(Caffeine, NoCaffeine)
```

- b.) Create the Null hypothesis Distribution using `do_it()` function.

```r
null_dist <- do_it(10000) * {
  shuffled_sample <- sample(combined_sample)

  shuff_caff <- shuffled_sample[1:10]
  shuff_nocaff <- shuffled_sample[11:20]

  shuff_stat <- mean(shuff_caff) - mean(shuff_nocaff)
}
```

- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed mean difference` using the `abline()` function.

```r
hist(null_dist ,
     xlab = "Difference in Mean Tap Rate", main = "Null distribution")
abline(v = diff_caffeine, col = "blue")
```



7.) Calculate p-value

```r
p_value <- pnull(diff_caffeine, null_dist, lower.tail = F)
p_value
```

```
[1] 0.0023
```

8.) Make a decision and state your conclusion:

Since our p-value is less than 0.05, we will reject the null hypothesis. We therefore have evidence that drinking caffeinated coffee produces an increase in the average tap rate compared to those that drink decaffeinated coffee.

Question 3: Hypothesis Testing for Correlations

Do more home runs mean more wins? In other words, is there a positive correlation between the number of home runs a team hits and the number of wins? Test this hypothesis using a permutation test. Data from the 2019 Major League Baseball (MLB) season is available in the `BaseballHits2019` dataset in the `Lock5Data` library. Make sure to extract the appropriate columns from the data.

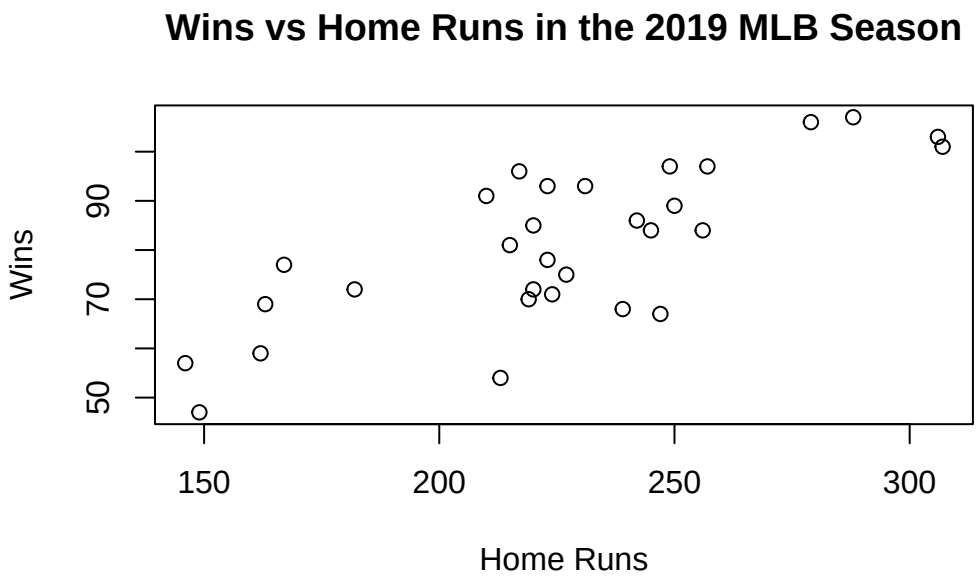
1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

- Null: There is no linear association between home runs and wins
- Alternative: There is a positive, linear association between home runs and wins

- $H_0 : \rho = 0$
- $H_A : \rho > 0$

2.) Create a scatterplot to visualize the association between `HomeRuns` and `Wins`.

```r
plot(BaseballHits2019$HomeRuns, BaseballHits2019$Wins,
     xlab = "Home Runs", ylab = "Wins",
     main = "Wins vs Home Runs in the 2019 MLB Season")
```



3.) Compute the observed statistic (correlation between `HomeRuns` and `Wins`).

```
wins_hr_corr = cor(BaseballHits2019$HomeRuns, BaseballHits2019$Wins)
wins_hr_corr
```

```
[1] 0.763656
```

4.) Create null hypothesis distribution

- a.) Shuffle the variables `HomeRuns` and `Wins` into two new variables, and find the correlation between these two new shuffled variables.

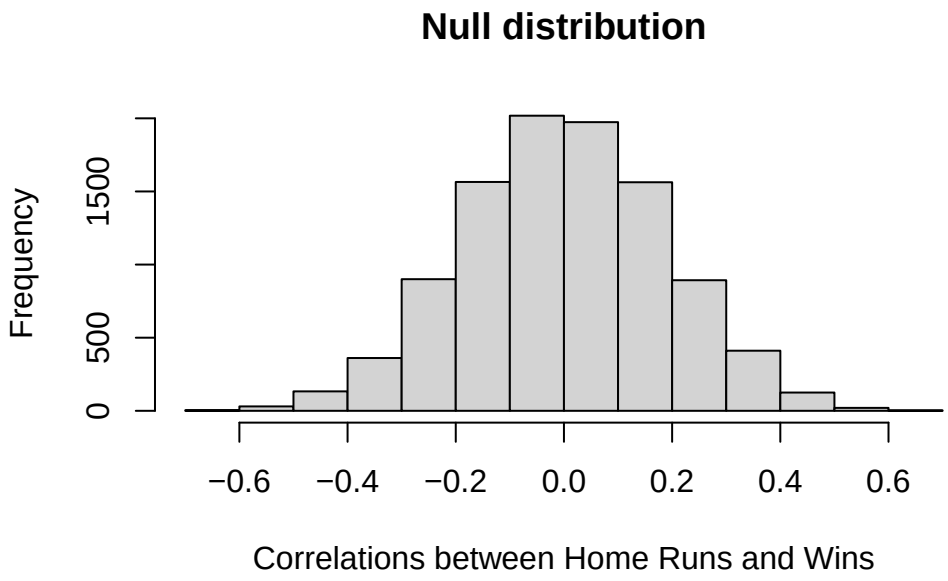
```
combined_sample <- c(BaseballHits2019$HomeRuns, BaseballHits2019$Wins)
```

- b.) Create the Null hypothesis Distribution using `do_it()` function.

```
null_dist <- do_it(10000) * {
  shuffled_sample <- sample(combined_sample)

  shuff_hr <- shuffled_sample[1:30]
  shuff_wins <- shuffled_sample[31:60]

  shuff_stat <- cor(shuff_hr, shuff_wins)
}
```

- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed correlation` using the `abline()` function.

```
hist(null_dist,
     xlab = "Correlations between Home Runs and Wins",
     main = "Null distribution")
abline(v = wins_hr_corr, col = "blue")
```



5.) Calculate p-value

```
p_value <- pnull(wins_hr_corr, null_dist, lower.tail = F)
p_value
```

```
[1] 0
```

6.) Make a decision and state your conclusion:

Since our p-value is less than 0.05, we will reject the null hypothesis. We therefore have evidence that there is a positive correlation between wins and home runs.

Question 4: Comparing Multiple Samples: The Mean Absolute Deviation (MAD) Statistic

The mean absolute deviation (MAD) statistic is a statistic that we can calculate to compare differences among more than two groups. Suppose we had 4 groups. Then the MAD statistic is given by:

$$\text{MAD} = \frac{|\bar{x}_1 - \bar{x}_2| + |\bar{x}_1 - \bar{x}_3| + |\bar{x}_1 - \bar{x}_4| + |\bar{x}_2 - \bar{x}_3| + |\bar{x}_2 - \bar{x}_4| + |\bar{x}_3 - \bar{x}_4|}{6}$$

The Lock5Data library contains the data set `TextbookCosts`. Conduct a permutation test to see if the mean textbook cost differs among subjects. *Hint:* Follow the steps that you did for the previous exercises, except now use the `get_MAD_stat()` function to compute your observed statistic and generate your null distribution.

1.) Write the `null hypothesis` and `alternative hypothesis` in words and in symbols.

- Null: There is no difference in average textbook costs among subjects
- Alternative: There is a difference in average textbook costs among subjects

2.) Compute the observed MAD statistic. Use the `get_MAD_stat()` function.

```
mad_stat = get_MAD_stat(TextbookCosts$Cost, TextbookCosts$Field)
```

```
arts = subset(TextbookCosts$Cost, TextbookCosts$Field == "Arts")
human = subset(TextbookCosts$Cost, TextbookCosts$Field == "Humanities")
ns = subset(TextbookCosts$Cost, TextbookCosts$Field == "NaturalScience")
ss = subset(TextbookCosts$Cost, TextbookCosts$Field == "SocialScience")
```

3.) Create null hypothesis distribution

- a.) Shuffle the variables `HomeRuns` and `Wins` into two new variables, and find the correlation between these two new shuffled variables.

```
combined_sample <- c(arts, human, ns, ss)
```

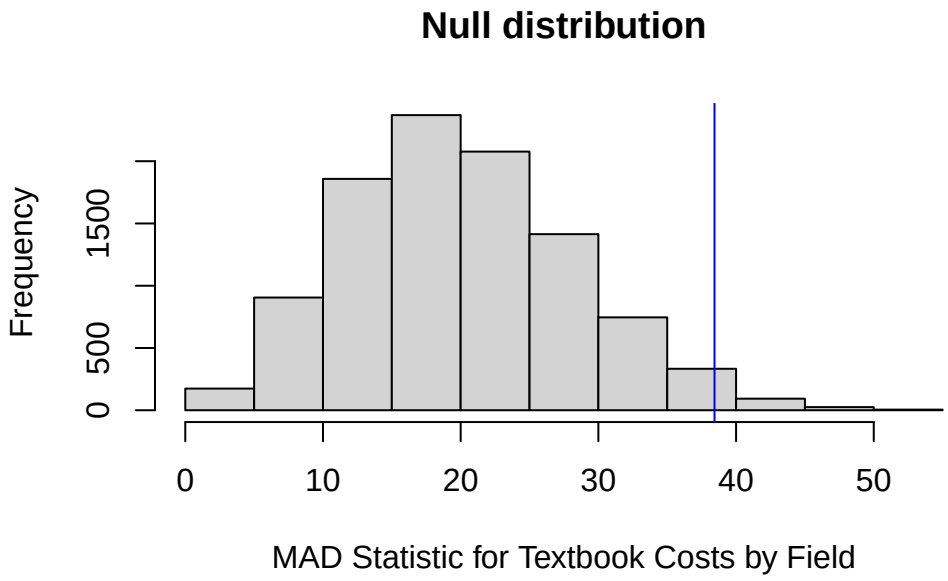
- b.) Create the Null hypothesis Distribution using `do_it()` function.

```
null_dist <- do_it(10000) * {
  shuffled_sample <- sample(combined_sample)

  shuff_data_frame = data.frame(
    Cost = shuffled_sample,
    Field = c(
      rep("Arts", 10), rep("Humanities", 10),
```

```
      rep("NaturalScience", 10), rep("SocialScience", 10)
      )
    )

  shuff_stat = get_MAD_stat(shuff_data_frame$Cost, shuff_data_frame$Field)
}
```

- c.) Plot a `histogram` of the null distribution and show the `line` of the `observed correlation` using the `abline()` function.

```
hist(null_dist,
     xlab = "MAD Statistic for Textbook Costs by Field",
     main = "Null distribution")
abline(v = mad_stat, col = "blue")
```



4.) Calculate p-value

```
p_value <- pnull(mad_stat, null_dist, lower.tail = F)
p_value
```

```
[1] 0.0189
```

5.) Make a decision and state your conclusion:

Since our p-value is less than 0.05, we will reject the null hypothesis. We therefore have evidence that there is a difference in textbook cost among subjects.

Question 5: Non-parameteric test using vacccine antibodies (Kruskal-Wills test)

We will use data on `Antibodies` (in g/ml) production after receiving a `Vaccine` (Vaccine A, Vaccine B, Vaccine C). A hospital administered three different vaccines to 6 individuals each and measured the antibody presence in their blood after a chosen time period. The data is saved in `patient_vaccine.csv`.

We walk you through testing for the difference between the three groups of vaccines using a different method than in class, it is called the Kruskal-Wills test.

1.) Create a boxplot to show the three vaccines variation in terms of the antibodies.

2.) Write in words the `null hypothesis` and the `alternative hypothesis`.

3.) Let prepare your data. Rank your data from all groups together in one column, name it `ranks`. *hint*: you can use function `rank`.

4.) Sum the ranks for each group of the `Vaccine`. Reports those sums results.

5.) Calculate the test statistic, `H` of the Kruskal-Wills test given by the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Where :

- N is the total sample size
- k is the number of groups we are comparing.
- $ {n_i} $ is the sum of ranks for group i.
- $ {R_i} $ is the sample size of group i.

6.) Find at significance level $\alpha = 0.05$, the critical value, which is the cutoff determined by `chi-square` distribution with `df= k-1` (degrees of freedom).

hint1: from the chi-square table, find the chi-square `critical value` with `df= k-1`.

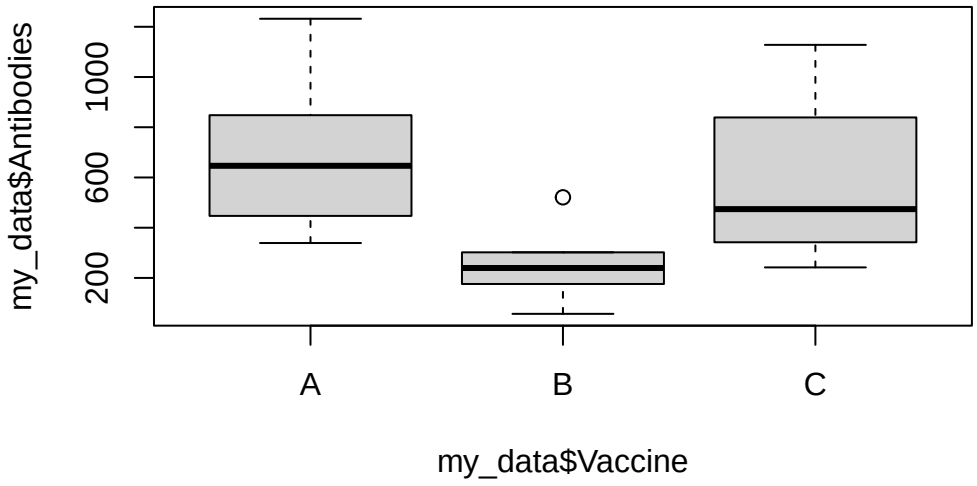
hint2: or you can use `R` function: `qchisq(pth, df, lower.tail= "F" )`.

7.) Make Judgement about your hypothesis within the context.

Answers:

1.) Create a boxplot to show the three vaccines variation in terms of the antibodies.

```
library(SDS1000)

my_data<- read.csv("PatientVaccine.csv")
boxplot( my_data$Antibodies ~ my_data$Vaccine )
```



2.)

Null Hypothesis: the vaccines cause the same amount of antibodies to be produced (all three groups originate from the same distribution and have the same median)

Alternative Hypothesis: At least one of the vaccines causes a different amount of antibodies to be produced (at least one group originates from a different distribution and has a different median)

3.) Let prepare your data. Rank your data from all groups together in one column, name it `ranks`. *hint*: you can use function `rank`.

```r
# Calculate overall ranks for the Antibodies variable and create a column named ` ranks`
my_data$rank <- rank(my_data$Antibodies)

# View the data frame with the new ranks column
print(my_data)
```

```
   X1 Vaccine Antibodies rank
1   2       A       1232   18
2   3       A        751   14
3   4       A        339    7
4   5       A        848   16
5   6       A        447    9
6   7       A        542   13
7   8       B        302    6
8   9       B         57    1
9  10       B        521   12
10 11       B        278    5
11 12       B        176    2
12 13       B        201    3
13 14       C        839   15
14 15       C        342    8
15 16       C        473   10
16 17       C       1128   17
17 18       C        242    4
18 19       C        475   11
```

4.) Sum the ranks for each group of the `Vaccine`. Reports those sums results.

```r
## Find the sum of the ranks under each vaccine group:
#R1 <- sum(subset(my_data, Vaccine == "A")$rank)

R1 <- sum(my_data$rank[my_data$Vaccine == "A"])

R2 <- sum(my_data$rank[my_data$Vaccine == "B"])

R3 <- sum(my_data$rank[my_data$Vaccine == "C"])

## 77, 29, 65
```

5.) Calculate the test statistic, `H` of the Kruskal-Wills test given by the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Where :

- N is the total sample size
- k is the number of groups we are comparing.
- $ {n_i} $ is the sum of ranks for group i.
- $ {R_i} $ is the sample size of group i.

```
#Kruskal-Wallis H statistic (manual formula)

N <- 18                # total number of observations
k <- 3                 # number of groups
R <- c(77, 29, 65)     # sum of ranks for each group
n <- c(6, 6, 6)        # sample sizes for each group

# Compute H statistic
H <- (12 / (N * (N + 1))) * sum((R^2) / n) - 3 * (N + 1)
H
```

```
[1] 7.298246
```

```
## 7.2982
```

6.) Find at significance level $\alpha = 0.05$, the critical value, which is the cutoff determined by `chi-square` distribution with `df= k-1` (degrees of freedom).

hint1: from the chi-square table, find the chi-square `critical value` with `df= k-1`.

hint2: or you can use `R` function: `qchisq(pth, df, lower.tail= "F" )`.

```
# the critical value Kurskal-Wills test is :
cvU<- qchisq(0.05, 2, lower.tail= F )

# Another way to calculate it
cvL<- qchisq(0.95, 2, lower.tail= T )
```

7.) Make Judgement about your hypothesis within the context.

Since $\texttt{H} = 7.2982 > \texttt{CV} = 5.9914$, thus, our test statistics `H` is in the rejection region, so, we **reject the null** and **conclude the alternative**, that the data do provide enough evidence to say, there **is difference between the vaccines antibodies production**.

Note: You can use the function `kruskal.test` to answer the quetion 5.You will get the same conclusion.

```
# To Perform the Kruskal-Wallis test you can use this function :
kruskal_result <- kruskal.test(Antibodies ~ Vaccine, data = my_data)

# View the full test results
print(kruskal_result)
```

```
    Kruskal-Wallis rank sum test

data:  Antibodies by Vaccine
Kruskal-Wallis chi-squared = 7.2982, df = 2, p-value = 0.02601
```

```
## Kruskal-Wallis chi-squared = 7.2982, df = 2, p-value = 0.02601
```

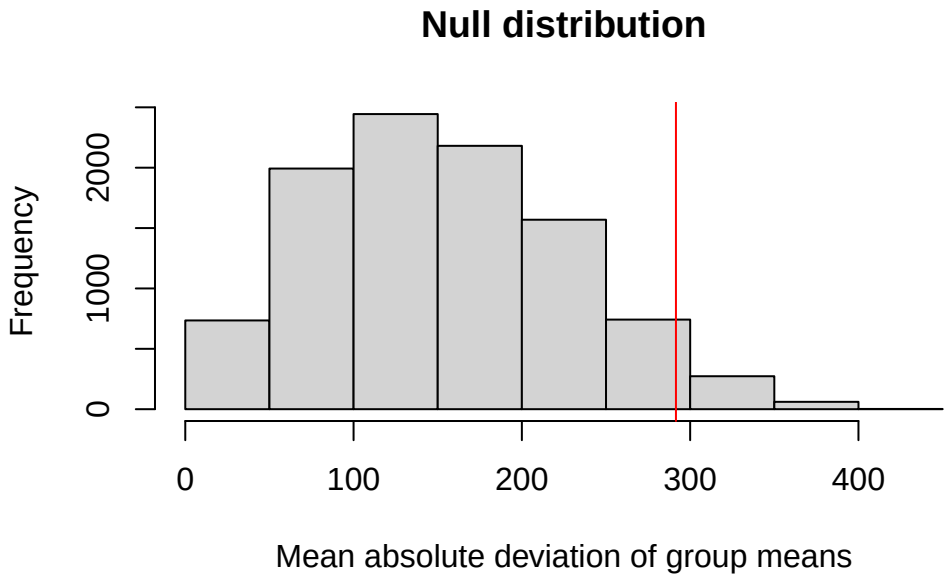
Question 6:

Repeat Question 5 with the randomization method `MAD` and compare your results. What is your reflection.

```
library(SDS1000)
my_data <- read.csv("PatientVaccine.csv")

# calculate the observed statistic
obs_stat <- get_MAD_stat(my_data$Antibodies, my_data$Vaccine)
obs_stat
```

```
[1] 291.5556
```

```r
# create the null distribution
null_dist <- do_it(10000) * {

  # shuffle the completion times
  shuffled_Vaccine <- sample(my_data$Vaccine )

  # calculate the simulated statistic
  get_MAD_stat(my_data$Antibodies, shuffled_Vaccine)

}

# plot the null distribution
hist(null_dist, #breaks = 200,#
     main = "Null distribution",
     xlab = "Mean absolute deviation of group means")


# add a red vertical line at the observed statistic
abline(v = obs_stat, col = "red")
```




```r
# P.Value
p_value <- pnull(obs_stat, null_dist, lower.tail = FALSE)
```

```
p_value
```

```
[1] 0.0434
```

```
### pv= 0.0433
```

Answers:

*) At significance level 0.05, we can **reject the null** and conclude the **alternative**, that there is difference between the three vaccines in terms of the amounts of antibodies they procedure. Which is the same conclusion as the method used in **question5** .