practice 6_test for difference of two means

Practice: Hypothesis test for 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 thelineof theobserved mean differenceusing theabline()‘ function.

```
# your code here
```

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

```
# your code here
```

Step 5: Make decision/Judgment

```
#your code here
```

Answers:

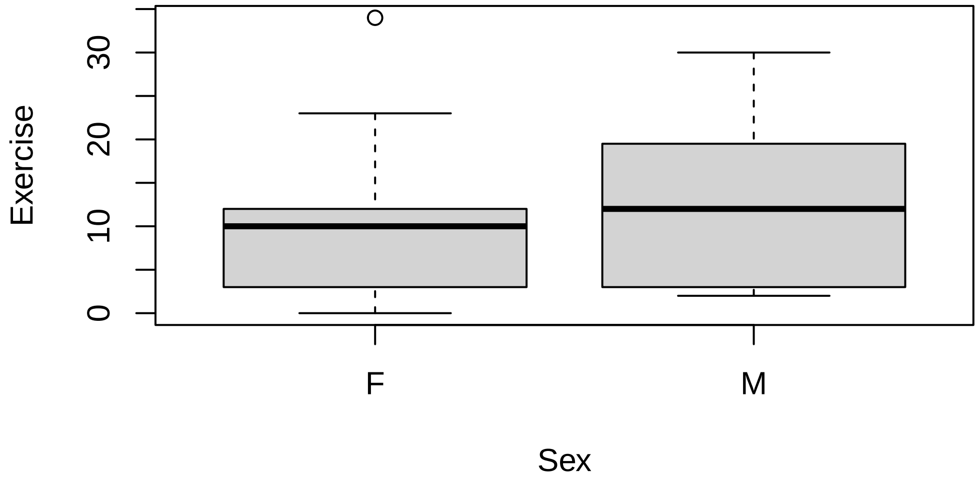
```
library(Lock5Data)
library(SDS1000)
data(ExerciseHours)
```

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] 1.666667
```

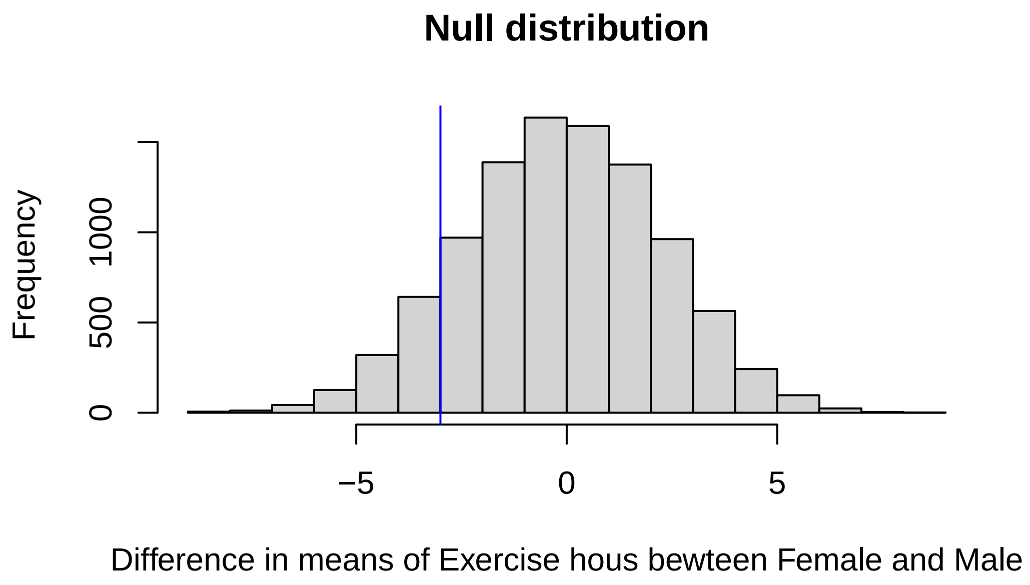
```
# answers may vary
```

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] 0.1149
```

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

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

```
#There are no enough evidence to conclude that there is  a mean difference in Exercise Hours
```