2nd-Year Master of Statistics and Data Science Computer Intensive Methods: Final projects (2024/2025)

Project 2

Mikita Bisliuk (2364811), Edmond Sacla Aide (2159278)

26 January, 2025

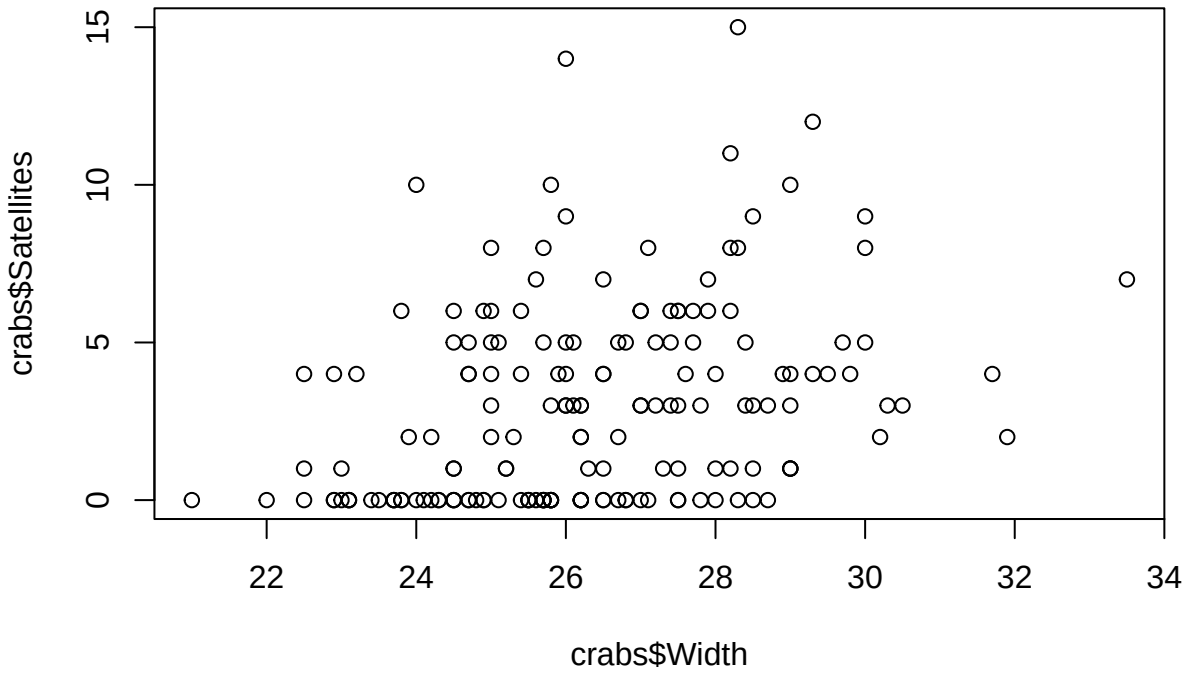
Project 2

In this question we use the horseshoe crab dataset. The data is available in R (crabs) as a part of the R package glm2. To get the data, install the package glm2 and use the code below to access the data.

```
library(glm2)
data(crabs)
names(crabs)
```

```
## [1] "Satellites" "Width"      "Dark"       "GoodSpine"  "Rep1"
## [6] "Rep2"
```

```
plot(crabs$Width, crabs$Satellites)
```



The dataset contains information about of 173 female horseshoe crab. You can find more details about this dataset in the book of Alen Agresti (An Introduction to Categorical Data Analysis, Section 3.3.2). The first 6 lines are given below.

```
head(crabs)
```

```
##   Satellites Width Dark GoodSpine Rep1 Rep2
## 1          8  28.3   no        no    2    2
## 2          0  22.5  yes        no    4    5
## 3          9  26.0   no       yes    5    6
```

```
## 4          0  24.8  yes        no    6    6
## 5          4  26.0  yes        no    6    8
## 6          0  23.8   no        no    8    8
```

Each female horseshoe crab in the study had a male crab attached to her nest. The study investigated factors that affect whether the female crab had any other males, **called satellites**, residing nearby her. The response outcome for each female crab is her number of satellites (*Satellites*). In this question, possible explanatory variables are the female crab's shell width (*Width*), which is a summary of her size and a binary factor indicating whether the female has good spine condition (yes or no, in R: *GoodSpine*)

Part 1

Question 1.1

Let Y_i be the number of satellites,we assume that Y_i $Poisson(\mu_i)$ where μ_i denotes the expected number of satellites for the i-th female crab. We consider the following linear predictor:

$$g(\mu_i) = \beta_0 + \beta_1 * Width_i + \beta_2 * GoodSpine_i.$$

Here, $g()$ is the link function. Formulate an appropriate model for the number satellites. Fit the model and use the likelihood ratio test in order to test the null hypothesis $H_0 : \beta_2 = 0$ against a two sided alternative.

The response variable Y_i (number of satellites) follows a Poisson distribution:

$$Y_i \sim \text{Poisson}(\mu_i)$$

The expected number of satellites (μ_i) is modeled using a log-link function:

$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{Width}_i + \beta_2 \times \text{GoodSpine}_i$$

Here:

- β_0: Intercept term
- β_1: Effect of shell width
- β_2: Effect of spine condition (GoodSpine_i, a binary factor)

Likelihood Ratio Test for $H_0 : \beta_2 = 0$

Hypotheses:

- $H_0 : \beta_2 = 0$ (GoodSpine has no effect on the number of satellites)
- $H_a : \beta_2 \neq 0$ (GoodSpine has an effect)

```
## [1] "Likelihood Ratio Test (p-value): 0.560"
```

Reject H_a, there is no significant difference between the two models. Therefore we can omit *good spine condition* as a predictor of number of satellites.

Question 1.2

Use parametric and non parametric bootstrap to test the null hypothesis in Q1.1. Compare the distribution of the likelihood ratio statistic obtained for the two bootstrap procedures to the theoretical distribution of the likelihood ratio test, what is you conclusion ?

- Test statistic: D=−2 (loglikelihood of reduce model−loglikelihood of full model)
- Theoretical distribution under H_0:$D \sim \chi^2_1$

We will now obtain the bootstrap distribution of D using parametric and non-parametric methods.

Parametric Bootstrap:

- Simulate data under the null hypothesis (reduced model: satell Poisson( ) with only Width as an explanatory variable) -Refit the full and reduced models for each bootstrap sample -Compute the likelihood ratio statistic D* for each sample.

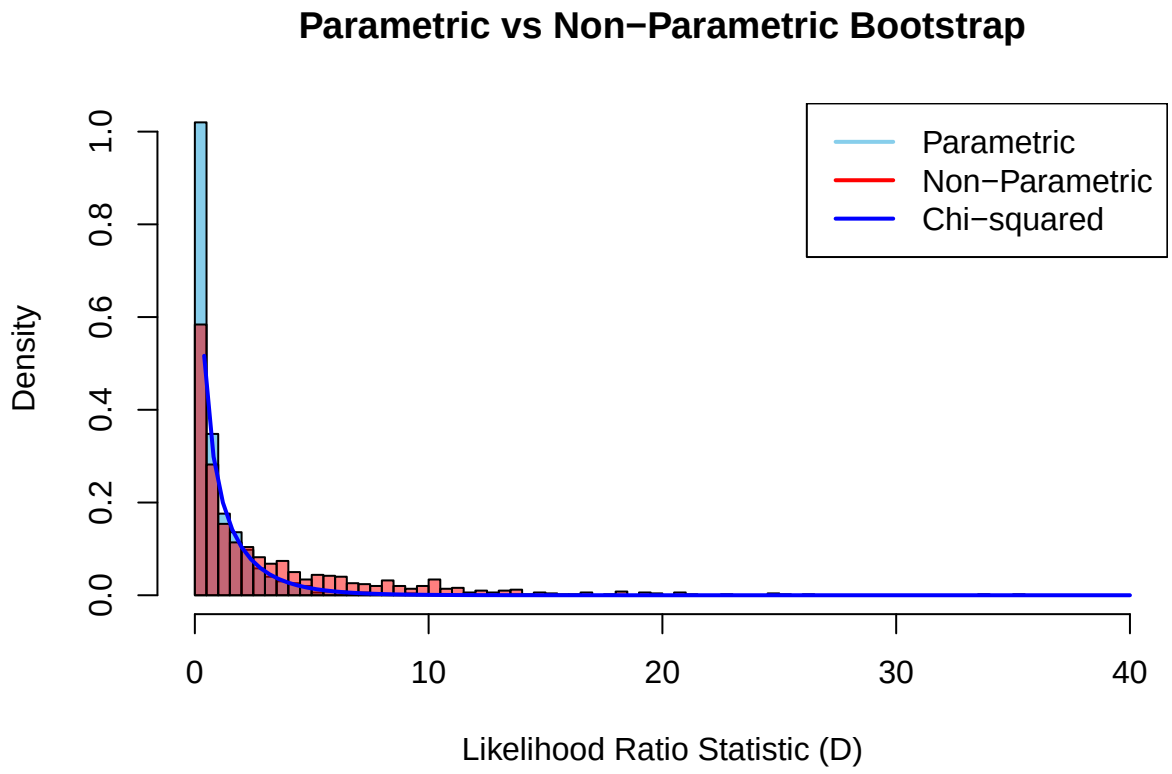
Non-Parametric Bootstrap:

-Resample the data (with replacement) from the observed dataset -Fit the full and reduced models for each bootstrap sample. -Compute the likelihood ratio statistic D* for each sample.

```
## [1] "Likelihood Ratio Test using parametric bootstrap (p-value): 0.556"
```

```
## [1] "Likelihood Ratio Test using non-parametric bootstrap (p-value): 0.754"
```

At a significance level of 5%, both bootstrap methods indicate no evidence of a significant effect of *good spine condition*.



The parametric and non-parametric bootstrap distributions align with the theoretical ^2_1 distribution, this suggest that the theoretical distribution captures the data behavior under the null hypothesis.

Question 1.3

Use permutations test to test the null hypothesis formulated in Q1.1

```
## [1] "Likelihood Ratio Test using permutations (p-value): 0.735"
```

At a significance level of 5%, the permutations test does not identify a significant effect of *good spine condition*.

Part 2

The data we use for this question is the sleep data. The study was conducted to show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients. The variable extra is the response variable , represents the increase in hours of sleep due to the treatment, and the variable group is the grouping factor. The data is given below.

```
extra<-c(0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0,
         1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,4.3)
group<-c(rep(1,10),rep(2,10))
ID<-c(1:20)
sleep <- data.frame(extra, group, ID)
```

let μ_1 and μ_2 be the means of the first and the second treatment group, respectively. We wish to test the null hypothesis

$$H_0 : \mu_1 = \mu_2,$$

against a two sided alternative.

Question 2.1

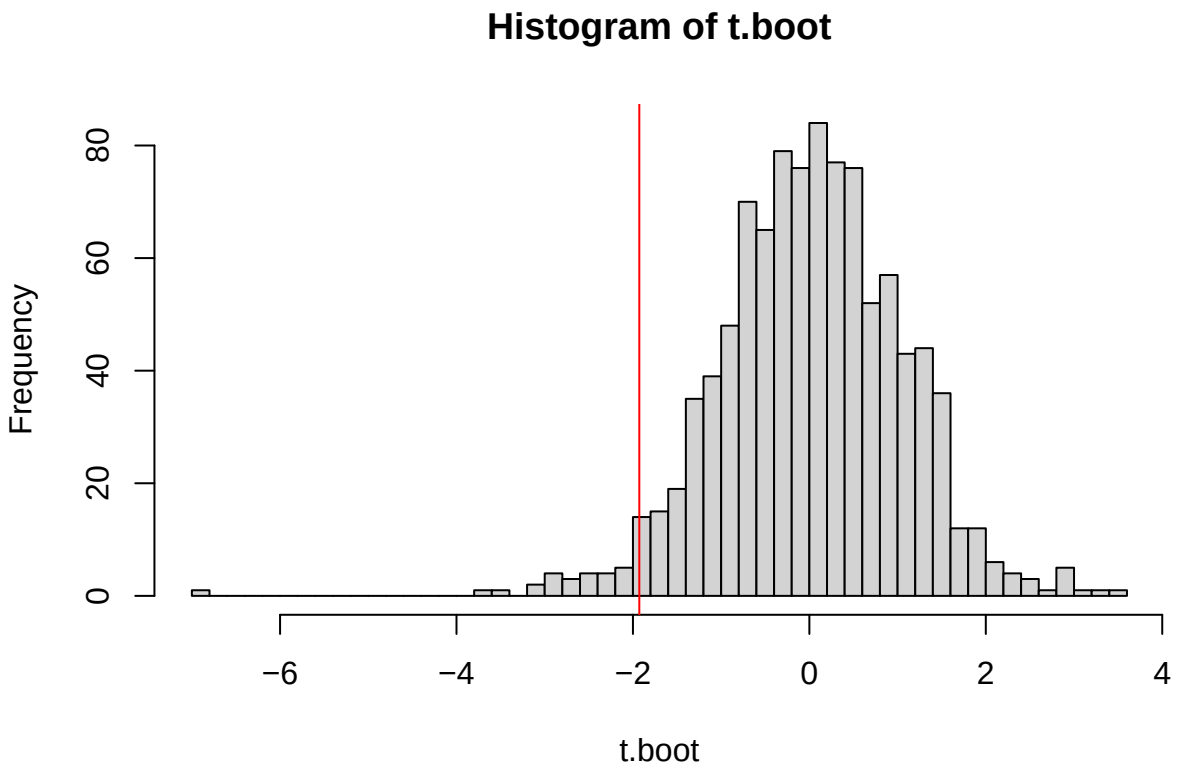
Use the classical two-samples t-test for *two independent samples*.

```
##
##  Welch Two Sample t-test
##
## data:  extra by group
## t = -1.9278, df = 17.619, p-value = 0.07016
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -3.4928289  0.1528289
## sample estimates:
## mean in group 1 mean in group 2
##            0.75            2.42
```

The associate p-value >0.05. So at a significance level of 5/%, the group of soporific drugs does not influence significantly increase in hours of sleep.

Question 2.2

```
## [1] "Non-parametric t-test (p-value): 0.057"
```



The non-parametric bootstrap reveals that, p-value >0.05. So at a significance level of 5%, the group of soporific drugs does not influence significantly increase in hours of sleep.

Question 2.3

To use a parametric bootstrap for testing the null hypothesis H_0 : M1=M2 with the test statistic:

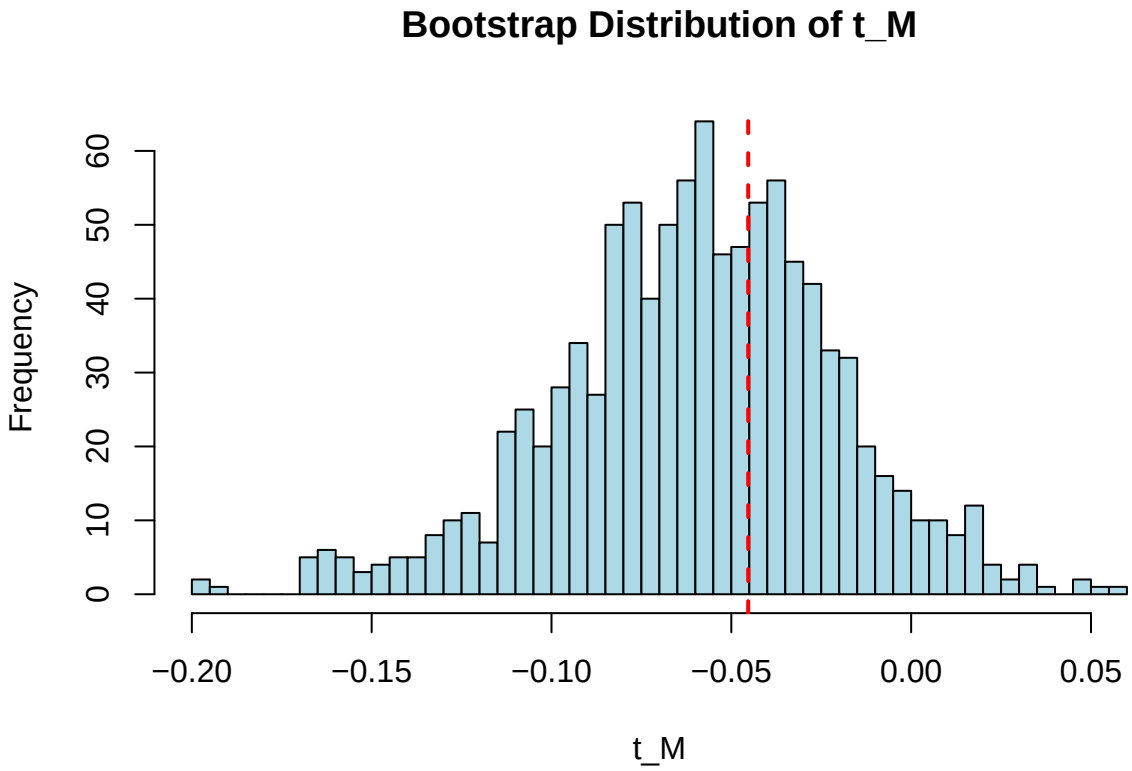
-Write a function to calculate tm

-Use parametric bootstrap to simulate the null hypothesis (we simulate the the null hypothesis by combining data from the 2 groups to indicate no difference in medians); Generate bootstrap samples and calculate tm.

```
## [1] "tM.obs: 0.057"
```

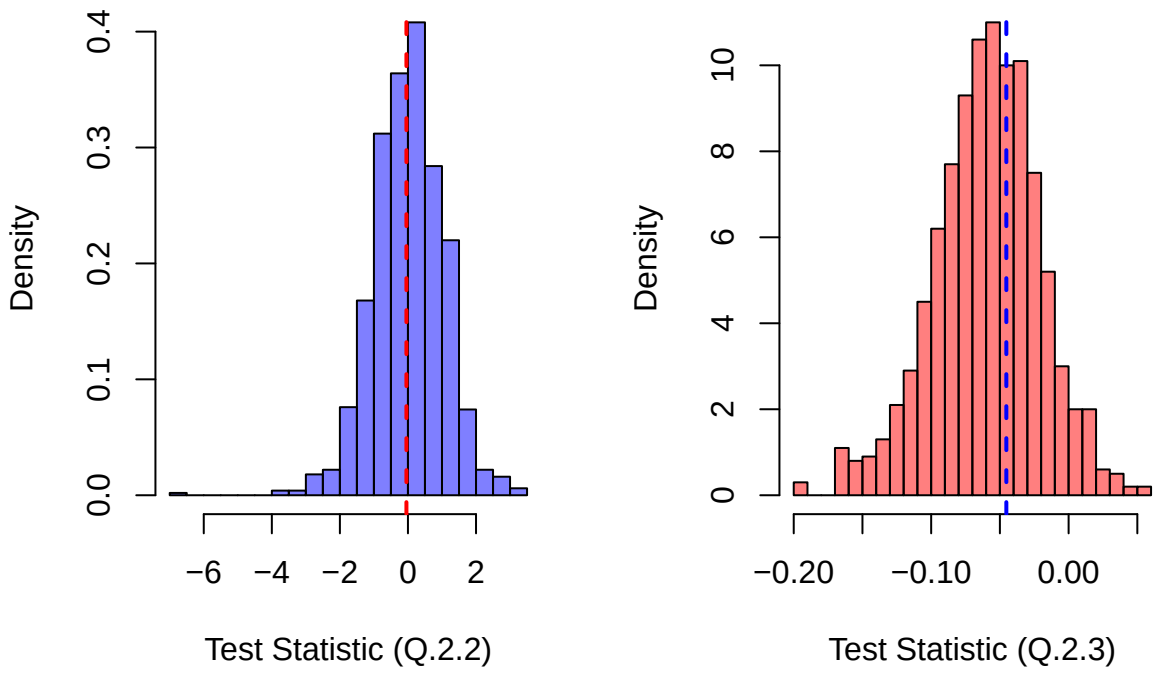
```
## [1] " Parametric bootstrap for testing tm: 0.636"
```

p-value=0.636 ==> no significant difference between the two group of treatment.



Question 2.4

Compare the distribution of the test statistics in Q2.2 and Q2.3




```
## Bootstrap t_M Mean: -0.05970169

## Bootstrap t_M SD: 0.03879186

## Theoretical t-distribution Mean: 0.01941521

## Theoretical t-distribution SD: 1.04305

## Bootstrap Tail Proportion: 0

## Theoretical Tail Proportion: 0.047
```

Comparing the two bootstrap distributions helps highlight how robust tM is compared to the classical t-statistic, especially in the presence of outliers or non-normality. If the distributions are similar, the choice of test may not matter much. However, differences would suggest that tM is less sensitive to the assumptions underlying the classical t-test.

Part 3

We consider the following dataset with three variables and 10 observations.

```r
ID<-c(1:10)
x1<-c(0.8,-1.23,1.25,-0.28,-0.03,0.61,1.43,-0.54,-0.35,-1.60)
x2<-c(0.64,-1.69,1.47,-0.14,-0.18,0.43,1.61,-0.31,-0.38,-1.82)
data.frame(ID,x1,x2)
```

```
##    ID    x1    x2
## 1   1  0.80  0.64
## 2   2 -1.23 -1.69
## 3   3  1.25  1.47
## 4   4 -0.28 -0.14
## 5   5 -0.03 -0.18
## 6   6  0.61  0.43
## 7   7  1.43  1.61
## 8   8 -0.54 -0.31
## 9   9 -0.35 -0.38
## 10 10 -1.60 -1.82
```

Note that there two observations per subject: (X_{1i}, X_{2i}) which represent a measurement of the same variable before and after a treatment. The statistic of primary interest in this question is the ratio between the means, that is

$$\hat{\theta} = \frac{\overline{X}_1}{\overline{X}_2}$$

Question 3.1

Estimate the ratio statistic.

```
## [1] -0.1621622
```

Question 3.2

Estimate the standard error of the ratio using non parametric bootstrap and Jackknife. For the bootstrap procedure use: B=10,20,50,100,250,500,1000,2500,5000,7500,10000. Which value of B you recommend to use?

```
## [1] "Jackknife non parametric bootstrap: 1.60e-30"
```

During bootstrap iteration we need to add a small epsilon=1e-6 to the denominator values to avoid infinite ratio results.

Bootstrap SE for /*theta* estimates using different number of bootstrap samples:

```
##          B=10          B=20          B=50         B=100         B=250         B=500
## 3.538347e-01 3.102986e+00 1.088935e+04 1.658502e+00 1.051534e+00 1.181820e+00
##        B=1000        B=2500        B=5000        B=7500       B=10000
## 2.972512e+03 3.085174e+03 2.511114e+03 1.607912e+03 2.916045e+03
```

Given the small sample size of the initial dataset, it's important to adjust the number of bootstrap samples to avoid overfitting and ensure meaningful results. Larger bootstrap samples have a higher likelihood of including outliers, so we recommend using B=100.

Question 3.3

Construct a 95% bootstrap confidence interval for the ratio:

```
##      2.5%     97.5%
## -1.399557  4.415674
```

Question 3.4

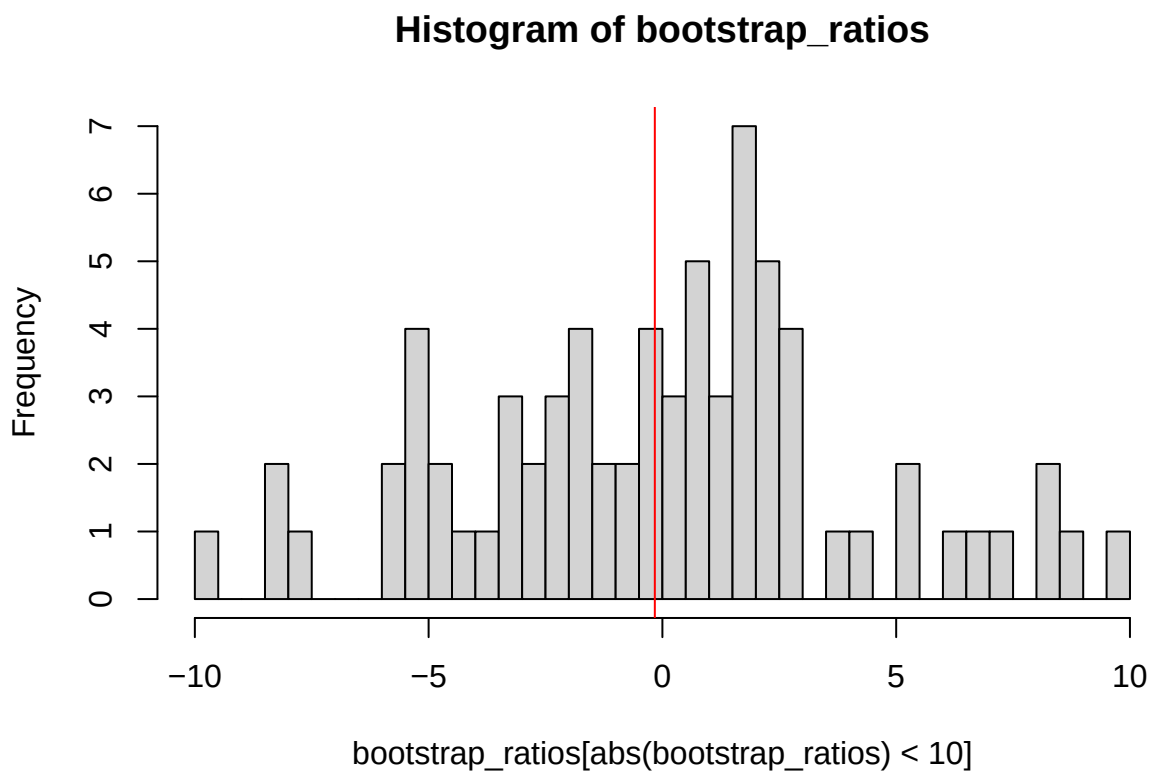
Use a bootstrap procedure to test the hull hypothesis $H_0 : \theta = 1$ against a one sided alternative. **Do not** use a two-samples paired t-test for the mean difference to test the null hypothesis

To test the null hypothesis H_0 : $\theta = 1$, against a one-sided alternative using a bootstrap procedure, we proceed as follows:

-Define the Test Statistic -Bootstrap Under the Null Hypothesis:

- Adjust the data under H_0 so that the null hypothesis holds true ($\theta = 1$). This can be done by scaling one variable so that the ratio of means equals 1. x2_adjusted <- x2*
- Compute Bootstrap Samples (non-parametric): Resample the adjusted data B times and calculate the test statistic for each bootstrap sample.
- Compare Observed Statistic: Compute the p-value by comparing the observed test statistic to the bootstrap distribution of the statistic under H_0

```
## [1] "Non-parametric for testing H_0 (p-value) : 0.584"
```



No significant treatment effect at 5% significance level.

Part 4

Consider the data in Q3, let Let μ_1 and μ_2 the mean of the subjects' first and the second measurements, respectively. Let the mean deference $\mu_d = \mu_1 - \mu_2 = E(X_{1i}) - (X_{2i})$.

Question 4.1

Construct a 95% C.I for μ_d using the classical method.

```
## [1] -0.9713644  1.0573644
```

Question 4.2

Use non parametric bootstrap to construct a 95% C.I for μ_d. Use the percentile, bootstrap t and BCa methods to construct the C.I.

-Percentile Method: The CI is directly obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution of μ_d

-Bootstrap t-Method : Bootstrap mean estimates. Bootstrap standard error estimates for each resample. A t-statistic for each resample Use the quantiles of the bootstrap t-statistics to construct the CI - Bias-Corrected and Accelerated (BCa) Method

Bootstrap percentile intervals

```
##      2.5%     97.5%
## -0.083025  0.171175
```

Bootstrap t-Method

```
##        2.5%       97.5%
## -0.08998098  0.33128911
```

BCa

```
##      alpha bca point
## [1,] 0.025    -0.098
## [2,] 0.975     0.235
```

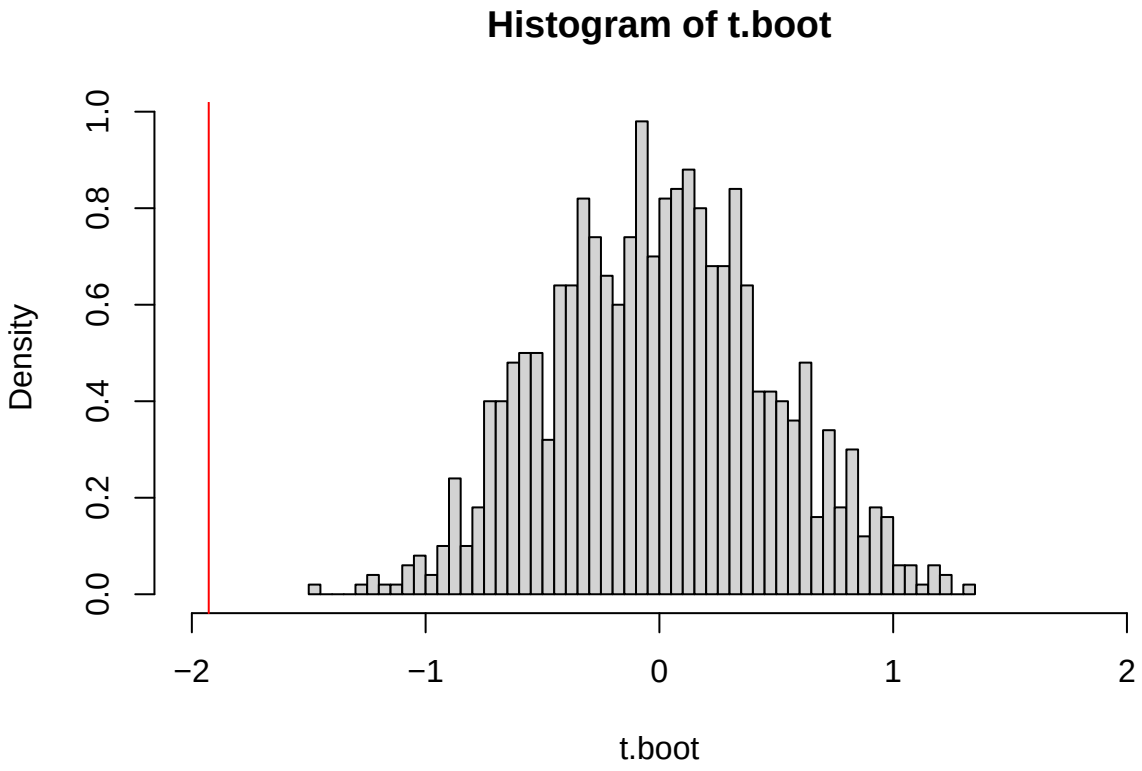
Question 4.3

Test the hypothesis $H_0 : \mu_d = 0$ using a non parametric bootstrap procedure.

Null hypothesis: $H_0 : \mu_d = 0$

-Compute the observed mean difference: _d -Generate bootstrap resamples from the original data, calculate the mean difference for each resample, and build the bootstrap distribution under the null hypothesis. -Compute the p-value as the proportion of bootstrap samples where the bootstrap statistic is more extreme than the observed _d under the null hypothesis

```
## [1] "non parametric bootstrap (p-value: 0.84715"
```



p-value>0.05, The difference in the means of the two treatments is not significant at 5% significance level.