



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

Name: Edmond SACLA AIDE (2159278) Lecturer: Prof. Ziv

2024-12-08

## 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.

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.

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  $Y_i \sim \text{Poisson}(\mu_i)$  where  $\mu_i$  denotes the expected number of satellites for the  $i$ -th female crab. We consider the following linear predictor:

$$g(\mu_i) = \beta_0 + \beta_1 * \text{Width}_i + \beta_2 * \text{GoodSpine}_i.$$

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 ( $\mu_i$ ) 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:

- $\beta_0$ : Intercept term
- $\beta_1$ : Effect of shell width
- $\beta_2$ : Effect of spine condition ( $\text{GoodSpine}_i$ , a binary factor)

```
##
## Call:
## glm(formula = Satellites ~ Width + GoodSpine, family = "poisson",
##      data = crabs)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.27908    0.54195  -6.050 1.44e-09 ***
## Width         0.16243    0.02008   8.089 6.02e-16 ***
## GoodSpineyes  0.05595    0.09579   0.584  0.559
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 567.54  on 170  degrees of freedom
## AIC: 928.84
##
## Number of Fisher Scoring iterations: 6
```

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"
```

$RH_0$ , there is significant different between the two models. Therefore it is important to keep *good spine condition* as a predictor of number of satellites.

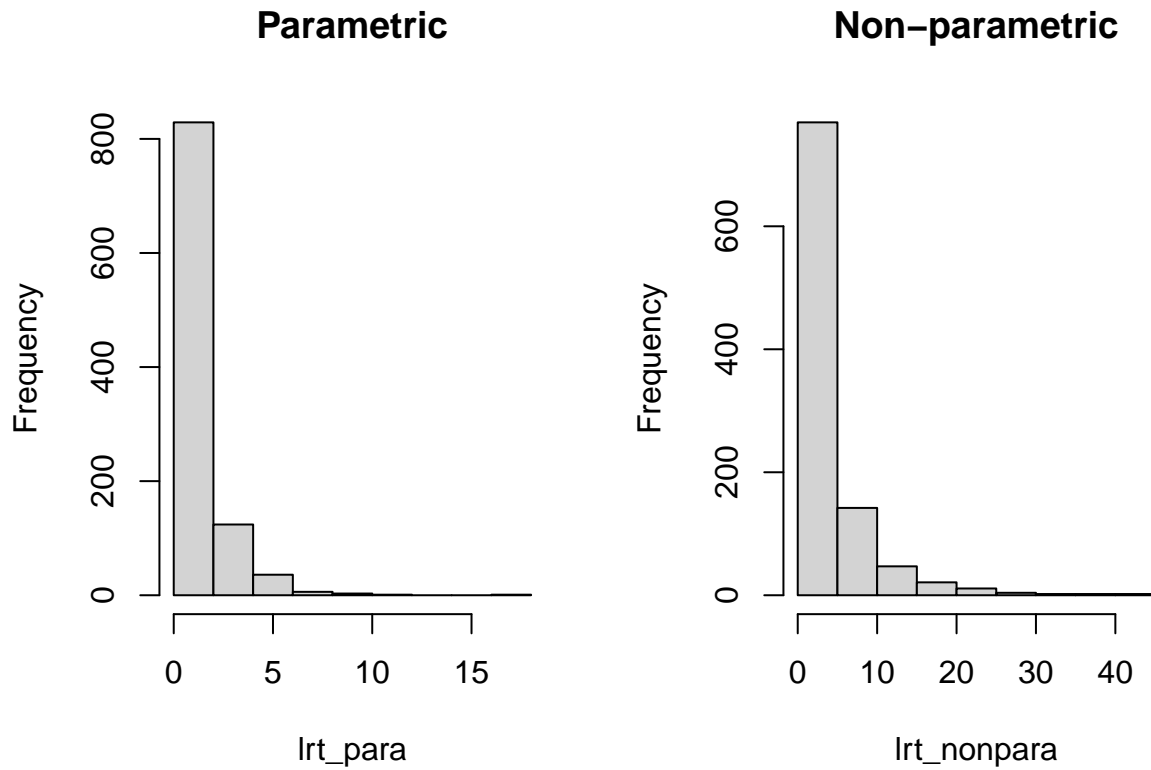
## Question 1.2

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 your conclusion ?

- Test statistic:  $D = -2 (\log\text{likelihood of reduced model} - \log\text{likelihood of full model})$  -Theoretical distribution under  $H_0: D \sim \chi_1^2$  We will now obtain the bootstrap distribution of 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.

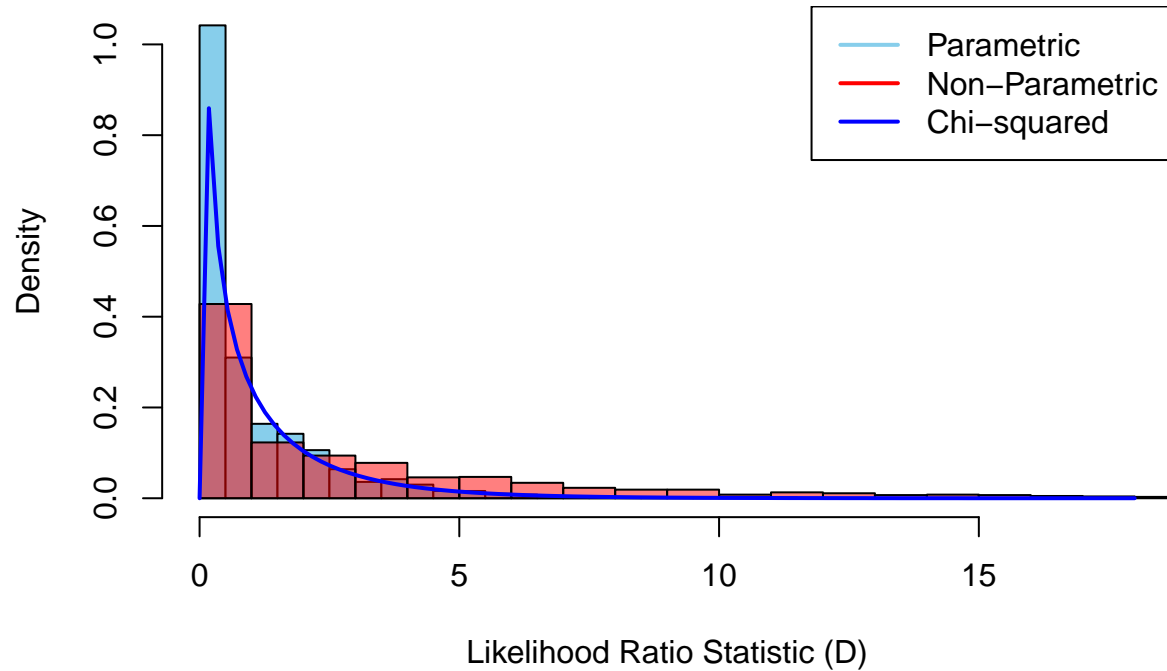


```
## [1] "likelihood ratio test using parametric bootstrap (p-value): 0.055"
```

```
## [1] "likelihood ratio test using non-parametric bootstrap (p-value): 0.286"
```

At a significance level of 5%, only the parametric bootstrap seems to identify a significant effect of *good spine condition*.

## Parametric vs Non-Parametric Bootstrap



### Compare Distributions

The non-parametric bootstrap distributions deviate significantly theoretical  $\chi^2_1$  distribution, this suggest that the theoretical distribution does not fully capture the data's behavior under the null hypothesis.

The parametric bootstrap distributions aligns somehow the theoretical  $\chi^2_1$  distribution, this suggest that the theoretical distribution more or less y capture the data's 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 permutation (p-value): 0.746"
```

At a significance level of 5%, only 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.

let  $\mu_1$  and  $\mu_2$  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

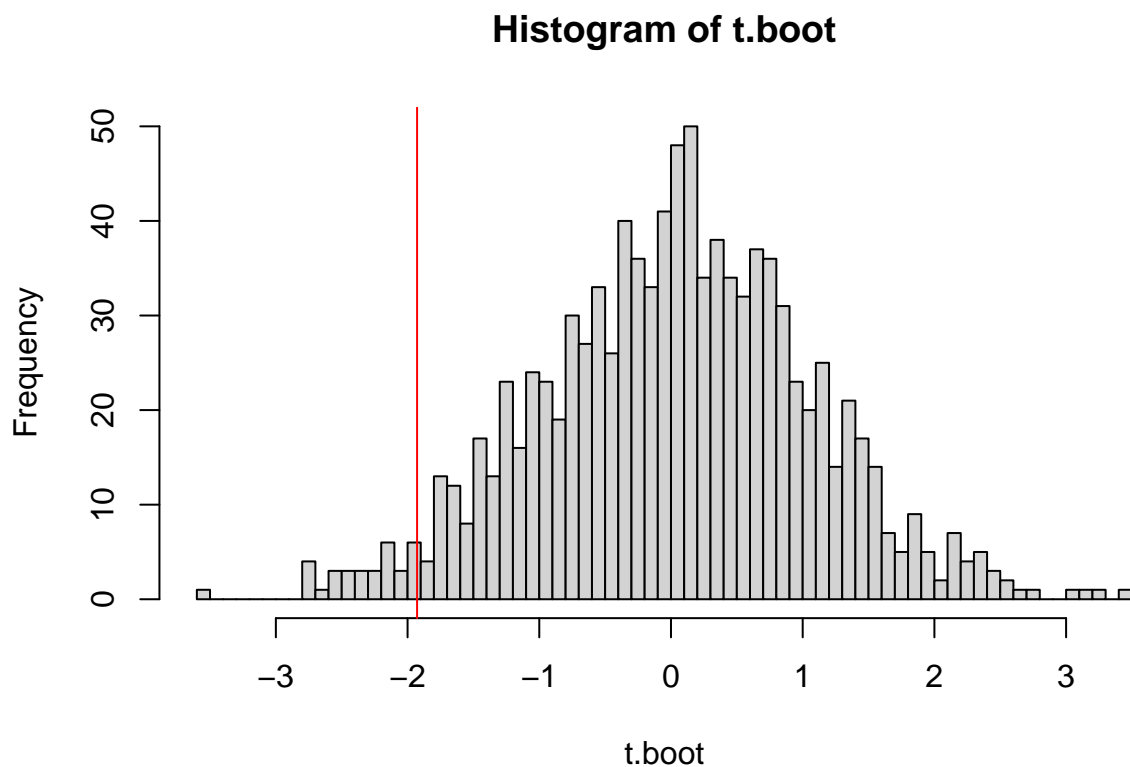
Use the classical two-samples t-test for *two independent samples*.

```
##
##  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): 0.064"
```



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

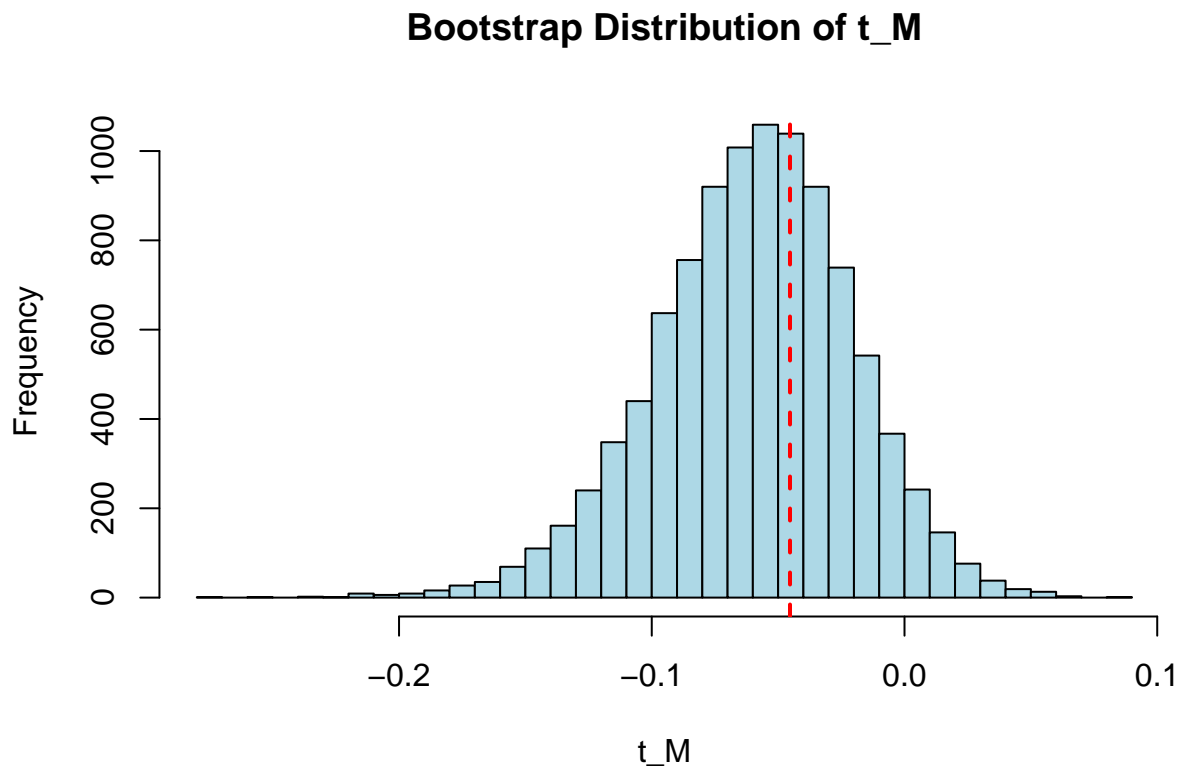
### Question 2.3

To use a parametric bootstrap for testing the null hypothesis  $H_0 : M1=M2$  with the test statistic, -Write a function to calculate  $t_m$  -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  $t_m$ .

```
## [1] "tM.obs: 0.064"
```

```
## [1] " Parametric bootstrap for testing tm: 0.638"
```

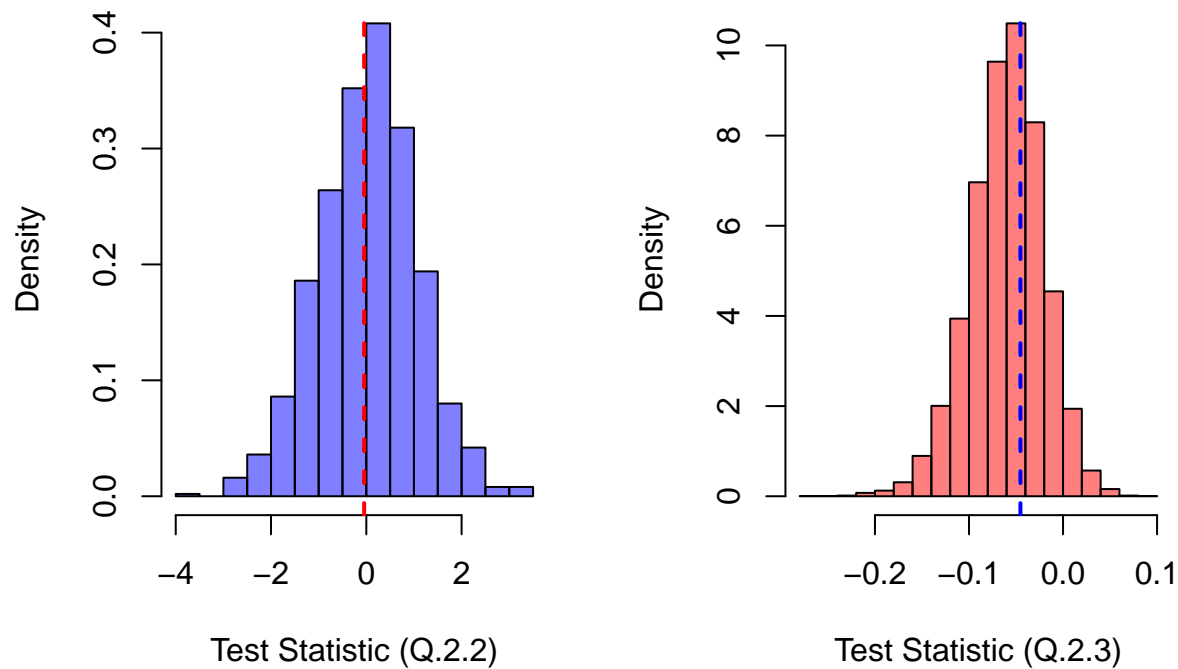
$p\text{-value}=0.638 \implies$  no significant difference between the two group of treatment.



### Question 2.4

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

## Comparison of Bootstrap Distributi



```
## Bootstrap t_M Mean: -0.06008697

## Bootstrap t_M SD: 0.03937468

## Theoretical t-distribution Mean: 0.03113136

## Theoretical t-distribution SD: 1.032813

## Bootstrap Tail Proportion: 0

## Theoretical Tail Proportion: 0.056
```

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.

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: 0.0000000000000000000000000000002"
```

##	B=10	B=20	B=50	B=100	B=250	B=500
##	1.252084e-01	9.622803e+00	3.855668e+30	2.751601e+00	1.105635e+00	1.396660e+00
##	B=1000	B=2500	B=5000	B=7500	B=10000	
##	7.966424e+27	6.422383e+28	NaN	NaN	NaN	

We recommend to use B=2500

### Question 3.3

Construct a 95% bootstrap confidence interval for the ratio:

```
##      2.5%      97.5%
## -1.657188  3.601020
```

### 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:  $x_2$
- 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) : 0.514"
```

No significant treatment effect at 5% significance level.

## Part 4

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

### Question 4.1

Construct a 95% C.I for  $\mu_d$  using the classical method.

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

### Question 4.2

Use non parametric bootstrap to construct a 95% C.I for  $\mu_d$ . 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  $\mu_d$  -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

*Non-parametric bootstrap*

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

*Bootstrap t-Method*

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

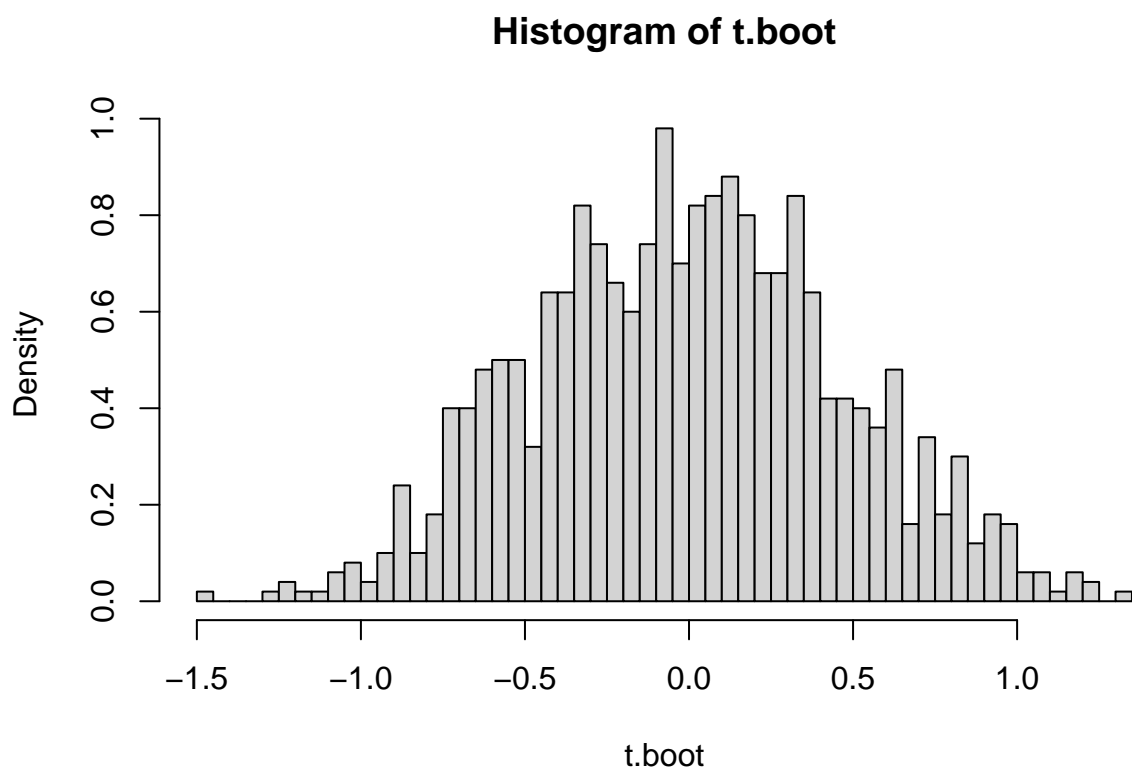
*BCa*

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

### 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:  $\bar{d}$  -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  $\bar{d}$  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.