

# Lab 5 Vaso

Vasileia Kampouraki

20/2/2020

## Question 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (Y=Draft No) sorted by day of year (X=Day of year) are given in the file lottery.xls.

1. Make a scatterplot of Y versus X and conclude whether the lottery looks random.

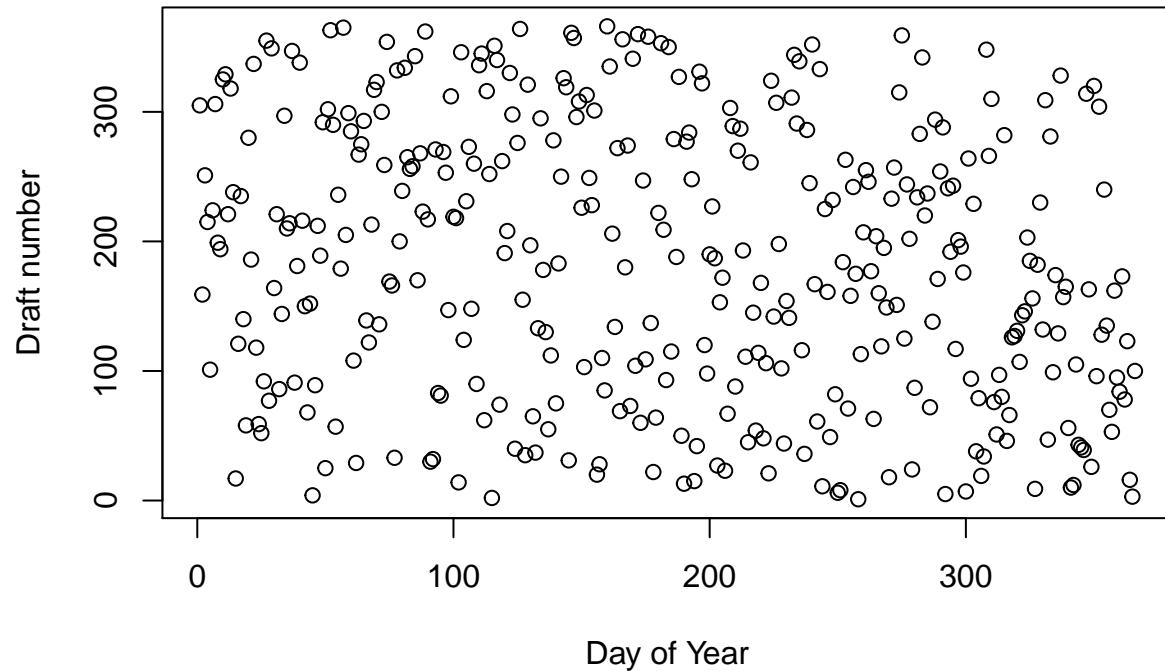
```
RNGversion("3.5.1")
```

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used

set.seed(12345)
library(readxl)

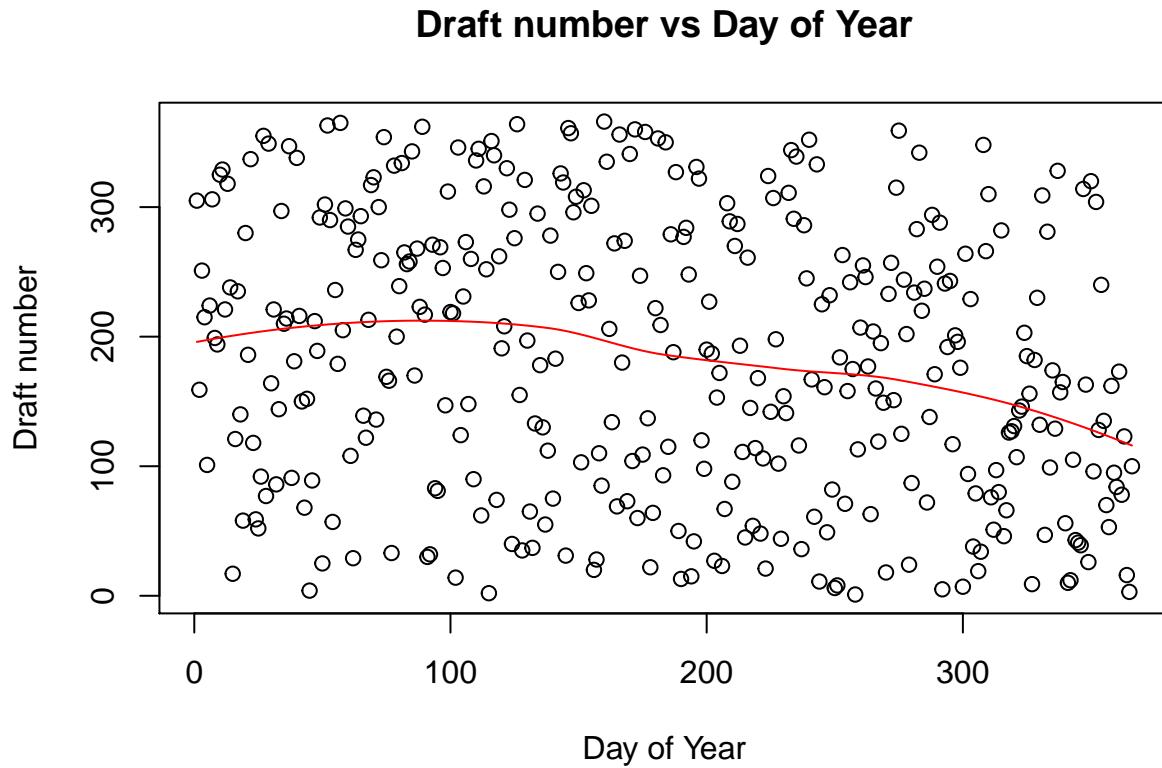
data <- read_excel("C:/Users/Vaso/Desktop/data_computational/lottery.xls")
df <- data.frame("X"=data$Day_of_year, "Y"=data$Draft_No)
```

## Draft number vs Day of Year



As we can see from the scatterplot, it seems that there isn't any specific pattern of the points so we can conclude that the lottery looks random and there's no relationship between the date and the draft number.

2. Compute an estimate  $\hat{Y}$  of the expected response as a function of  $X$  by using a loess smoother (use `loess()`), put the curve  $\hat{Y}$  versus  $X$  in the previous graph and state again whether the lottery looks random.



Plotting the **Expected** Draft Number, this time, against  $X$  we observe that the lottery doesn't look random as there is a downward trend. More precisely, it seems that the highest draft numbers are obtained for the first days of the year and as days go by, the expected draft numbers become smaller, steadily. So it seems as there is a reverse relation between  $X$  and  $\hat{Y}$ .

3. To check whether the lottery is random, it is reasonable to use test statistics  $T = \frac{Y(\hat{X}_b) - Y(\hat{X}_a)}{\hat{X}_b - \hat{X}_a}$  where  $X_b = \operatorname{argmax}_X \hat{Y}(X)$ ,  $X_a = \operatorname{argmin}_X \hat{Y}(X)$ . If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of  $T$  by using a non-parametric bootstrap with  $B = 2000$  and comment whether the lottery is random or not. What is the p-value of the test?

We will use the test statistics  $T = \frac{Y(\hat{X}_b) - Y(\hat{X}_a)}{\hat{X}_b - \hat{X}_a}$  where

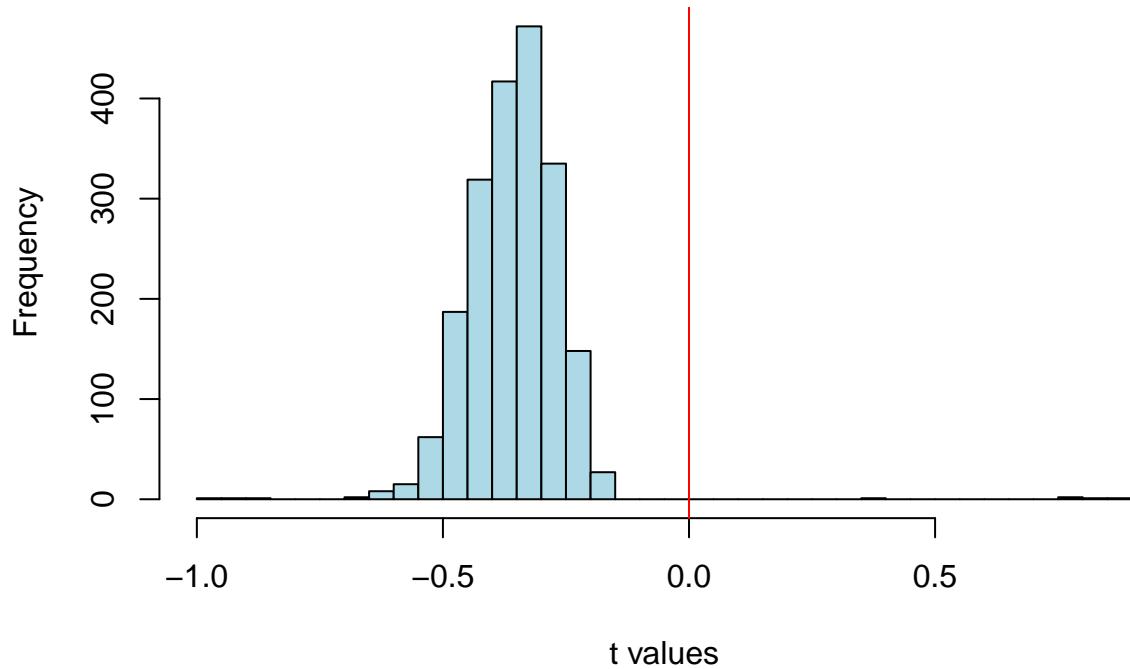
$X_b = \operatorname{argmax}_X \hat{Y}(X)$ ,  $X_a = \operatorname{argmin}_X \hat{Y}(X)$  to check whether the lottery is random.

So the hypothesis test we want to conduct is:

$H_0 : T \leq 0$  lottery is random vs.  $H_1 : T > 0$  : lottery is not random.

For this, we estimate the distribution of  $T$  using a non-parametric bootstrap with  $B=2000$ . That means that from the original data we generate 2000 different samples with replacement.

## Histogram of T distribution



Observing the histogram of T, we can see that almost all values of the T statistics is below 0. Thus, we wouldn't say that the T statistics is significantly greater than zero, so we reject the null hypothesis, the lottery is not random.

We now want to compute the p-value which is the probability of an observed or very unlikely(because of the tail) result under the assumption that the null hypothesis is true.

p-value=

$$\frac{\#|T(X_i)| > T(X)}{B}$$

In our case, p-value=0.0025 <0.05 which means that we reject the null hypothesis. The lottery is not random.

```
## [1] 0.0025
```

4. Implement a function depending on data and B that tests the hypothesis H<sub>0</sub>: Lottery is random versus H<sub>1</sub>: Lottery is non-random by using a permutation test with statistics T. The function is to return the p-value of this test. Test this function on our data with B = 2000.

We want to test the hypothesis:

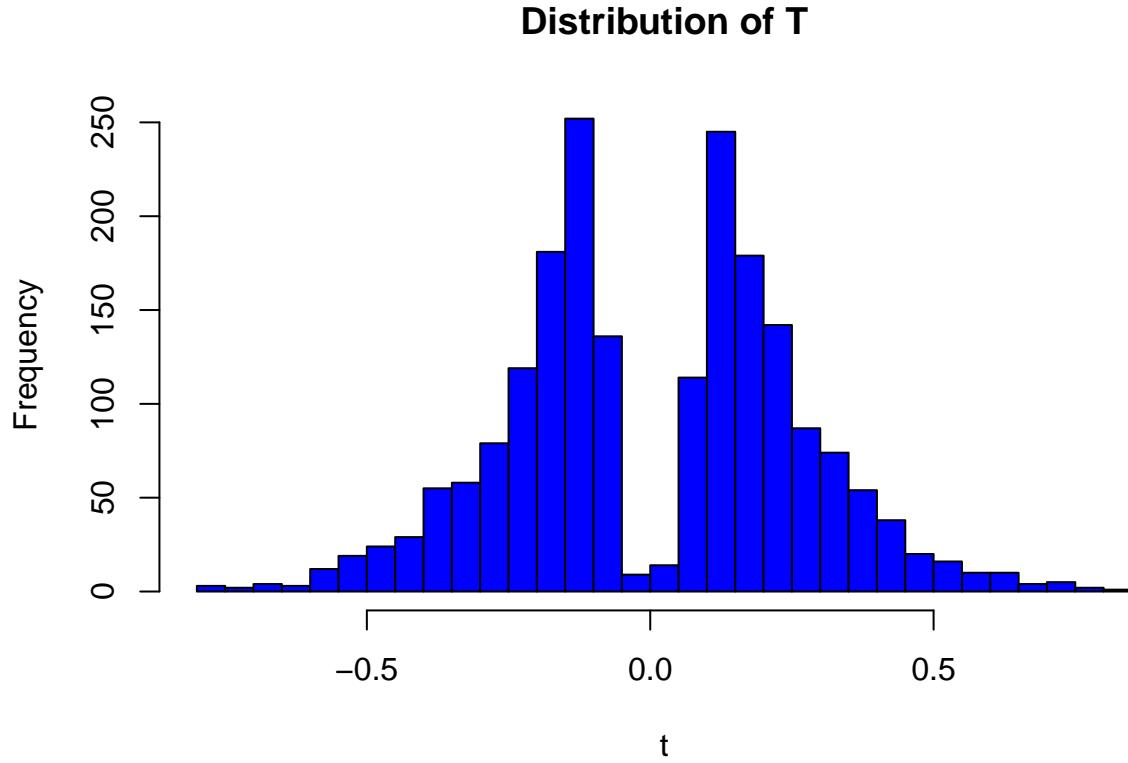
$H_0: T=0$  Lottery is random vs.  $H_1: T \neq 0$  Lottery is non-random by using a

permutation test and T statistics  $T = \frac{Y(\hat{X}_b) - Y(\hat{X}_a)}{X_b - X_a}$ , where

$X_b = \text{argmax}_X \hat{Y}(X)$ ,  $X_a = \text{argmin}_X \hat{Y}(X)$ .

The difference between bootstrap and permutation tests is that in bootstrap resamples are drawn with replacement whereas in permutation test resamples are drawn without replacement.

```
## p-value=
## [1] 0.159
```



Based on our hypothesis test the further the values of the T-statistics are from 0 the stronger is the evidence against the null hypothesis. Looking at the histogram of T we would say that the majority of values are around zero but then we have to check also the p-value to make a conclusion. Based on the p-value which is equal to  $0.159 > 0.05$  we fail to reject the null hypothesis, so based on this test Lottery is random.

##### 5. Make a crude estimate of the power of the test constructed in Step 4:

- (a) Generate (an obviously non-random) dataset with  $n = 366$  observations by using same  $X$  as in the original data set and  $Y(x) = \max(0, \min(\alpha x + \beta, 366))$ , where  $\alpha = 0.1$  and  $\beta \sim N(183, sd = 10)$ .
- (b) Plug these data into the permutation test with  $B = 200$  and note whether it was rejected.

```
## p-value=
## [1] 0
```

Using the new dataset that we constructed we get  $0 < 0.05$  as p-value and thus we reject the null hypothesis, so based on this test Lottery is not random.

- (c) Repeat Steps 5a-5b for  $\alpha = 0.2, 0.3, \dots, 10$ . What can you say about the quality of your test statistics considering the value of the power?

```
## The number of rejections is:
```

```
## [1] 99
```

```
## The power of this test is:
```

```
## [1] 1
```

First, we observe that for all values of  $\alpha$  the null hypothesis is rejected and we would say that this was expected as the data generated are not random.

The power of a hypothesis test is defined as the probability that the test rejects the null hypothesis when, actually, the null hypothesis is not true, i.e. correctly reject  $H_0$ .

It can be written mathematically as: **power= P(reject  $H_0$  |  $H_0$  is not true)**.

Type II error in statistics is defined as  $\beta = P(\text{accept } H_0 | H_0 \text{ is not true})$  and thus the power of the test can be written as  $1-\beta = P(\text{reject } H_0 | H_0 \text{ is not true})$  and because the power is a probability its values range from 0 to 1.

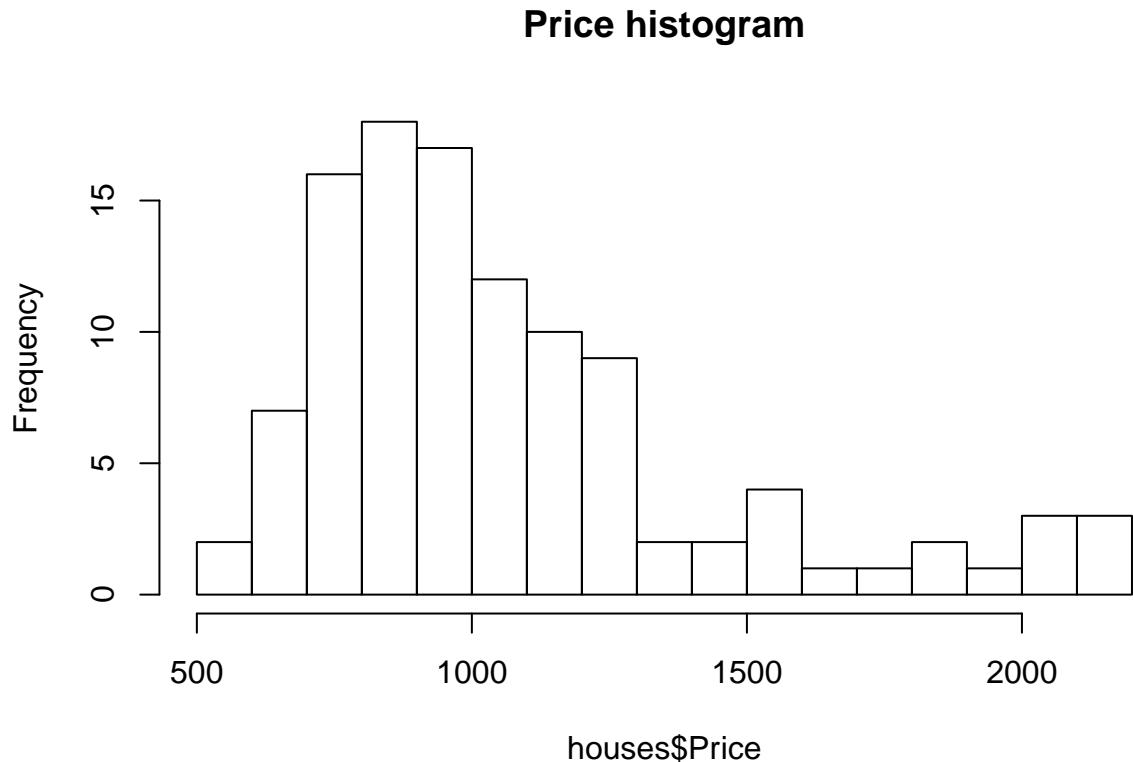
So, from the definition of the power, as the power increases the probability of making a type II error( $\beta$ ) decreases.

To find the value of the power, we did the permutations tests for all the different values of  $\alpha$  and then computed the percent of correct rejections. In the end we got a value of 1 which means that our test statistics is good.

## Question 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993. The variables present are Price; SqFt: the area of a house; FEATS: number of features such as dishwasher, refrigerator and so on; Taxes: annual taxes paid for the house. Explore the file **prices1.xls**.

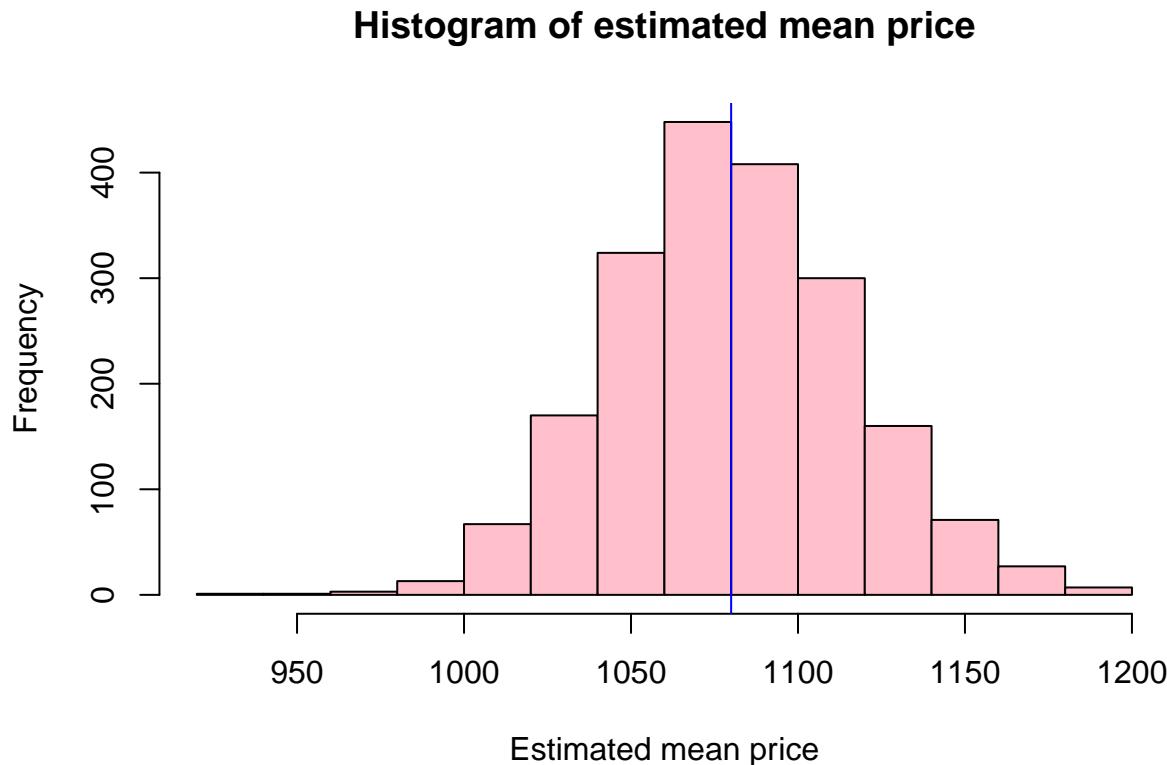
1. Plot the histogram of Price. Does it remind any conventional distribution? Compute the mean price.



Looking at the histogram of Prices we wouldn't say that it reminds us any of the conventional distributions as the shape of the histogram is not clear.

The mean price is 1080.473.

2. Estimate the distribution of the mean price of the house using bootstrap. Determine the bootstrap bias-correction and the variance of the mean price. Compute a 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first-order normal approximation (Hint: use `boot()`,`boot.ci()`,`plot.boot()`,`print.bootci()`).



Based on the above histogram we could say that the mean price is normally distributed. The blue vertical line represents the mean of the estimated mean prices, which is 1079.983. We could also get this result and a qq-plot by using the `plot.boot()` function but this looks nicer in terms of appearance and both give the same histogram.

The bias-correction estimator is

$$2T(D) - \frac{1}{B} \sum_{i=1}^B T_i^*$$

Using R to calculate this we obtain:

```
## The bias-corrected bootstrap mean is: 1080.963
```

The bias-correction mean seems good, as it's close to the real mean(1080.473) and thus we expect bootstrap mean and bias-corrected bootstrap mean to be around this price. Also, we want the bias to be small.

```
## The variance of the estimator is:
```

```
## [1] 1284.358
```

Another measure of uncertainty is the confidence intervals. We will compute the different types of confidence intervals using the `boot.ci()` function. Let's see some theory first. (<http://users.stat.umn.edu/~helwig/notes/bootci-Notes.pdf>)

#### **Bootstrap Confidence Interval via Percentiles:**

For this type of CI the  $100(1-\alpha)$ -th percentiles of  $T(D_i^*)$  are used. So, for a 95% confidence interval, in our case that we have  $B=2000$ , the CI will be:  $[T_{50}^*, T_{1950}^*]$ . *Note:* All the  $T(D_i^*)$  must be sorted first.

#### **Bootstrap BCa CI:**

Those intervals use percentiles of bootstrap distribution, but they do not necessarily use the  $100(1-\alpha)$ -th percentiles.

#### **First-order normal approximation Bootstrap CI:**

Those intervals are computed using  $\hat{\sigma}_B$  which is the sd of the bootstrap samples and  $\bar{T}$  which is the mean. So, the 95% CI is:  $\bar{T} \pm z_{\frac{\alpha}{2}} \hat{\sigma}_B$ .

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = res1)
##
## Intervals :
## Level      Normal          Basic
## 95%    (1011, 1151 )  (1009, 1148 )
##
## Level      Percentile       BCa
## 95%    (1013, 1152 )  (1018, 1162 )
## Calculations and Intervals on Original Scale
```

Running the `boot.ci()` function we got those results:

**Bootstrap Confidence Interval via Percentiles:** (1013, 1152)

**Bootstrap BCa CI:** (1018, 1162)

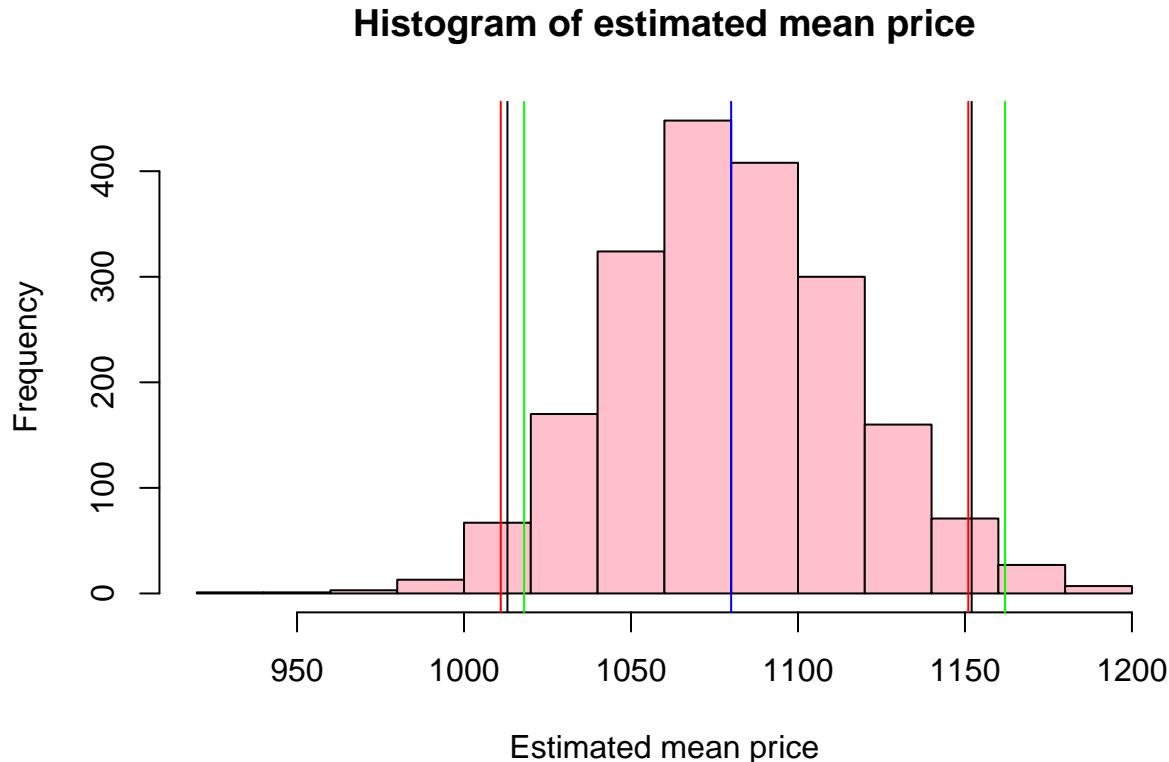
**First-order normal approximation Bootstrap CI:** (1011, 1151)

**3. Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate**

```
## [1] 1320.911
```

Using the jackknife algorithm the variance of the mean price is 1320.911 and it's a bit bigger than the variance obtained using the bootstrap estimate( 1284.358).

4. Compare the confidence intervals obtained with respect to their length and the location of the estimated mean in these intervals.



Above, we have the histogram of estimated mean prices using bootstrap and the blue line represents the mean of all those estimated mean prices.

**Bootstrap Confidence Interval via Percentiles:** (1013, 1152) and it's represented by the **black** lines.

**Bootstrap BCa CI:** (1018, 1162) and it's represented by the **green** lines.

**First-order normal approximation Bootstrap CI:** (1011, 1151)and it's represented by the **red** lines.

First of all we observe that the estimated mean(1079.983) falls inside all of the three different confidence intervals. Considering the length of the CIs the wider one is the BCa(length=144) and the narrowest one is the percentiles CI(length=139). The First-order normal CI has length 140.

However, it seems that all the three confidence intervals have more or less the same length and no significant differences can be detected between them.