Notes FSML II[*]

Tobías Chavarría

DSTI | DSBD2-001

Contents

Introduction

Statistics notation:

1. If $X_1, \ldots, X_n$ are random variables (r.v).
2. $x_1, \ldots, x_n$ are observations.
3. If we write *i.i.d* means that the r.v are independent and identically distributed.

First aim: To propose a model for a random variable.

Generalization to multi-dimensional case:

- Y: response variable.
- $X^{(1)}, \ldots, X^{(p)}$: explanatory variables.

Aim: To find a functional link between Y and the explanatory variables.

To find this functional link , the method to apply depends on the nature of the r.v's.

Y	Model
Numeric	Linear model
Qualitative (labels)	Classification

Linear model

A linear model is given by:

$Y_i = \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \varepsilon_i$

where:

- $\beta_0, \ldots, \beta_p$ are unknown *fixed* parameters that can be estimated by two methods:
  - Point estimation
  - Confidence interval
- ε is the noise and also a random variable.

[*]Replication files are available on the author's Github account (http://github.com/svmiller/svm-r-markdown-templates).

Chapter 1: Estimation for one parameter

Previous Knowledge

- Random Variable:
- The notion of distribution.
- The expectation and variance
- The distribution function
- The classical distributions (in particular the Gaussian)
- The Law of Large numbers and the Central Limit theorem

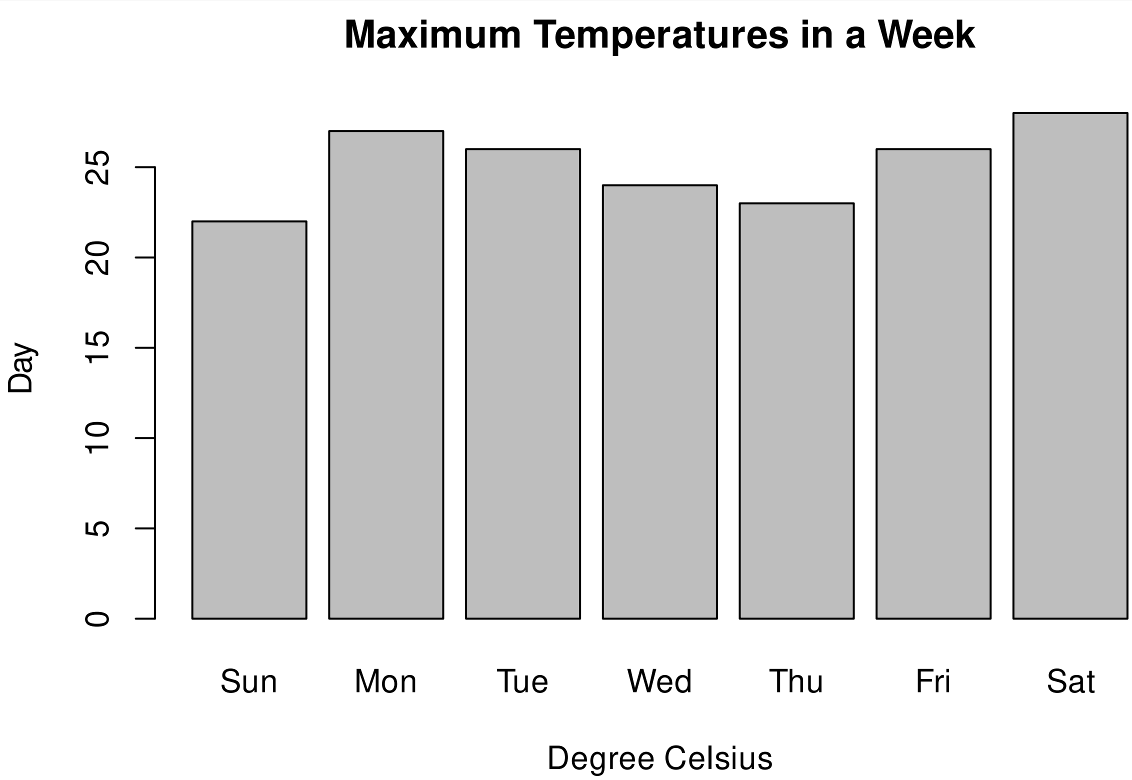
Introduction

Given $x_1, \ldots, x_n$ numeric observations, to try to find a correct parametric model, we can use 2 graphs:

- Barplot for discrete variables.

density = count/n

```
max.temp <- c(22, 27, 26, 24, 23, 26, 28)

barplot(max.temp,
main = "Maximum Temperatures in a Week",
xlab = "Degree Celsius",
ylab = "Day",
names.arg = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"))
```

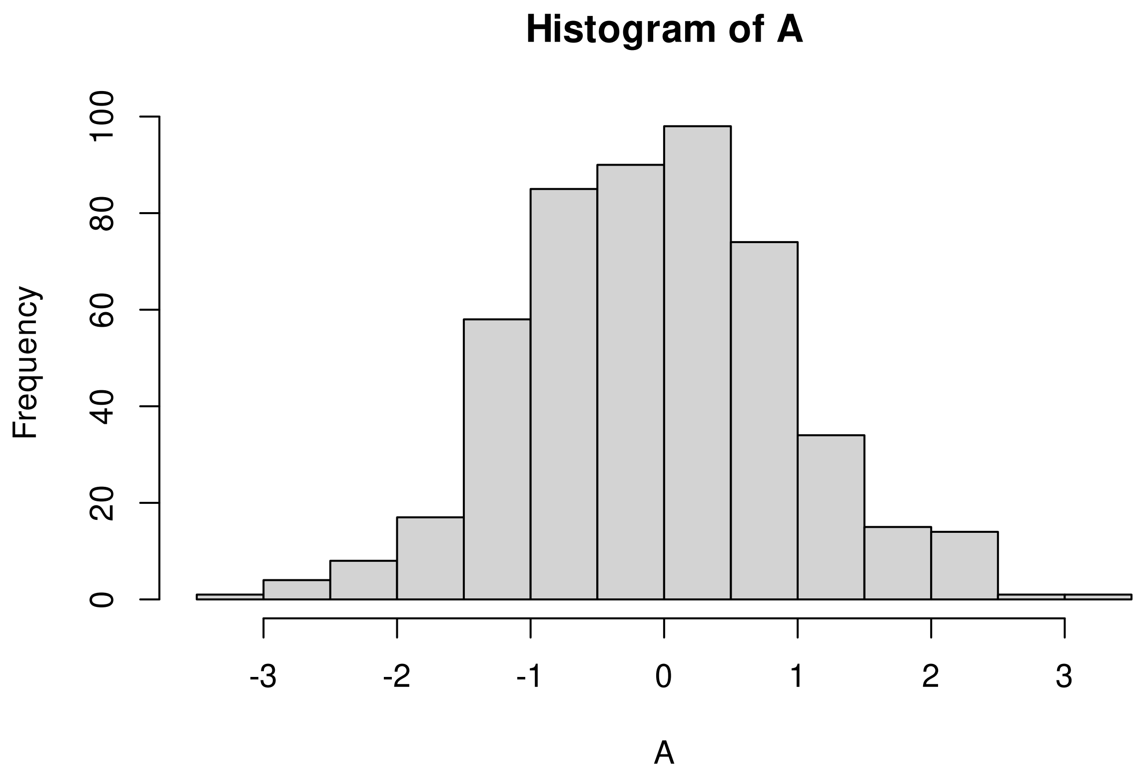


- Histogram for continuous variables

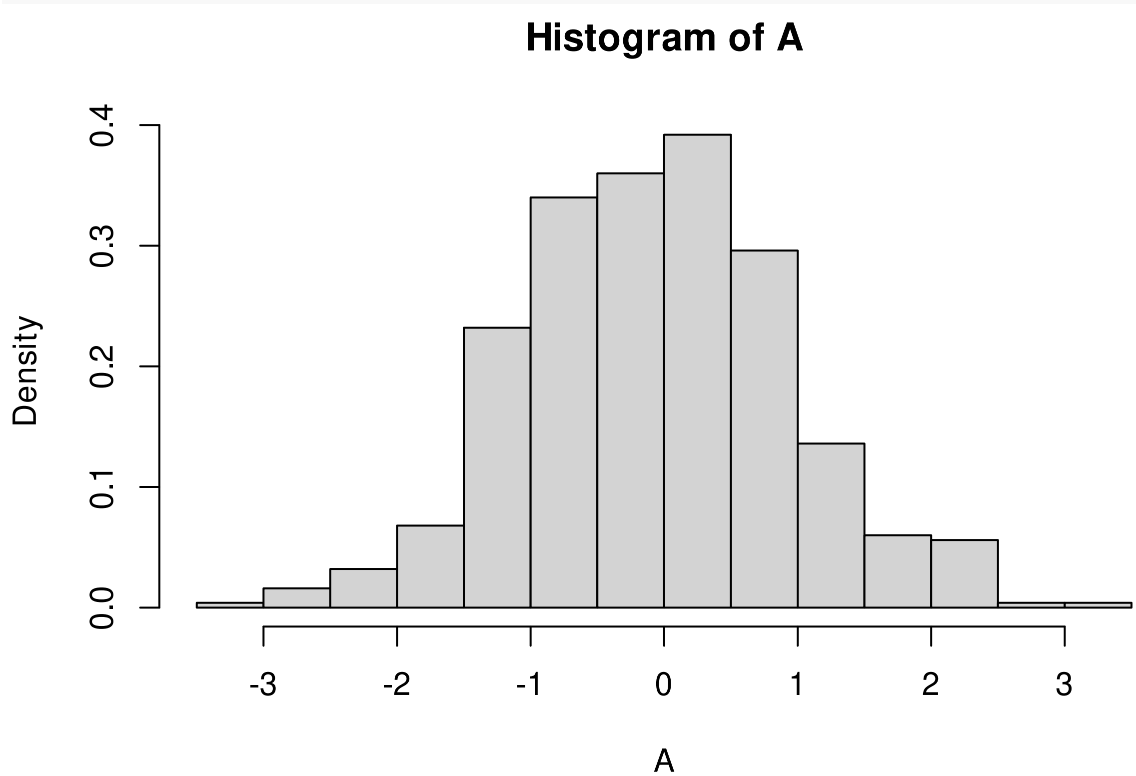
density = count for a bin/ n x length of the bin

```
A <- rnorm(500, 0, 1)
```

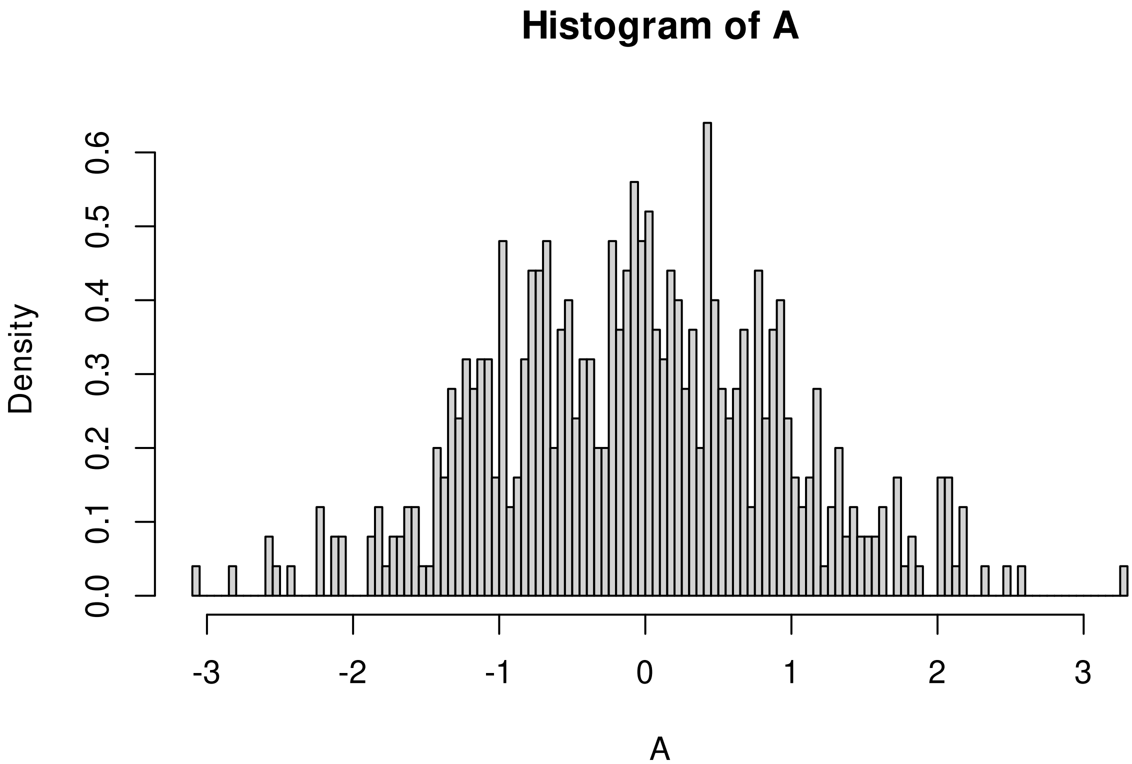
```
hist(A)
```



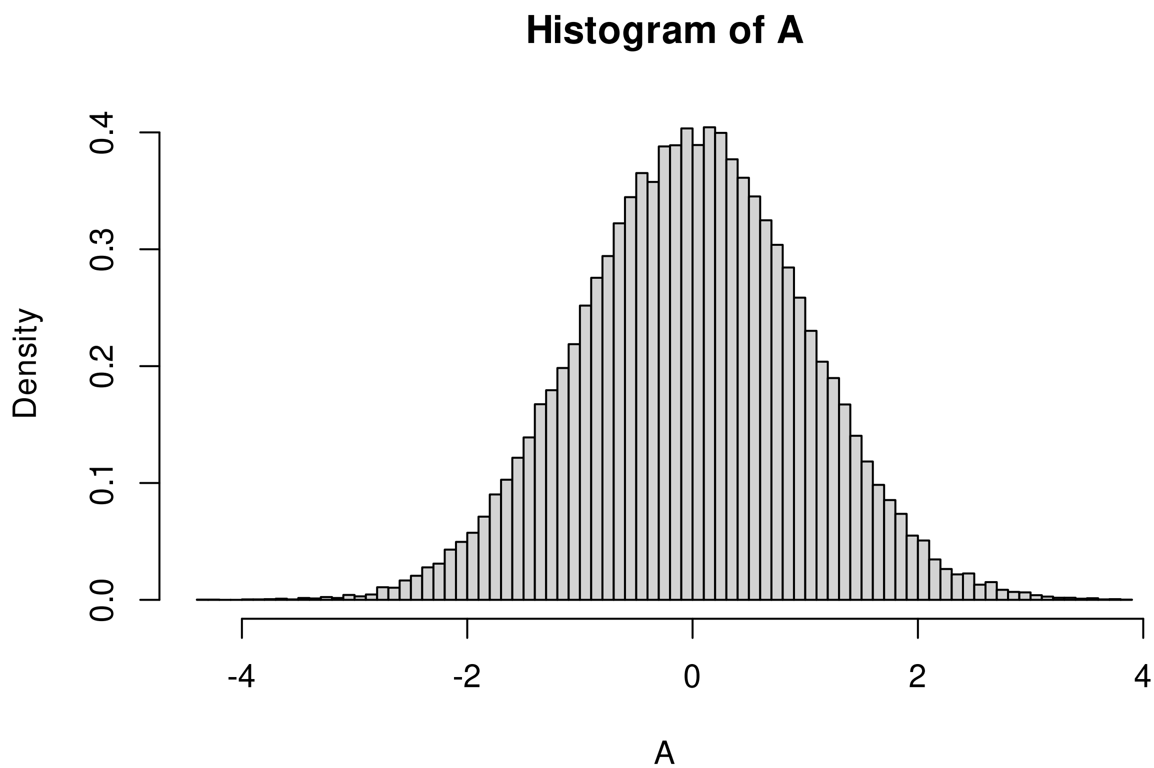

```
hist(A, freq = FALSE)

hist(A, freq = FALSE, breaks = 20)
```



```r
hist(A, freq = FALSE, breaks = 100)
```




```r
# In order to create more breaks, you need to increase the
# numbers of observations


A <- rnorm(50000, 0, 1)
hist(A, freq = FALSE, breaks = 100)
```



To propose a parametric model:

1. We make a graphical representation of the observations.
2. We guess a theoretical model by looking the previous graphic.

Question: with a representation, we can guess a parametric family of models, denoted by $\{P_\Theta, \theta \in \Theta\}$. How to guess a correct value for θ thanks to the observations?

Answer: *Estimation*.

Point estimation

x_i an observation of a r.v X_i we assume that $x_1, ..., x_n$ are *i.i.d* with common distribution P_Θ.

Def: Estimator An estimator of Θ is just a function of $X_1, ...X_n$ that **does not depend onto others unknown parameters.**

Rk: An estimator is a random variable!

Def: Estimation An estimation is the value of an estimator computed thanks to the observations.

Example

Consider $X_1, ..., X_n$ exponential distributed and *i.i.d*, an *estimator* of λ is $\hat{\lambda}_n = \dfrac{n}{\sum X_i}$ an *estimation* is $\hat{\lambda}_n = \dfrac{n}{\sum x_i}$.

Def: Bias (for univariate parameter)

Let consider $\hat{\Theta}_n$ an estimator of Θ.

The bias of $\hat{\Theta}_n$ is defined by:

$b(\hat{\Theta}_n) := \mathbb{E}(\hat{\Theta}_n) - \Theta$

We say that $\hat{\Theta}_n$ is an unbiased estimator if $\forall n \in \mathbb{N}^+ \quad b(\hat{\Theta}_n) = 0$

We say that $\hat{\Theta}_n$ is asymptotic unbiased estimator if:

$b(\hat{\Theta}_n) \to 0$ as $n \to \infty+$

How to construct estimator?

- Method of moments
  - less computations
  - based on the Law of large numbers
- Maximum likelihood

Method of moments Let Θ a parameter to estimated, parameter which is associate to $X_1, ..., X_n$ *i.i.d* r.v.

Let consider $k \in \mathbb{N}^*$:

- the moment of order k : $\mathbb{E}[x^k]$
- the centered moment of order k: $\mathbb{E}[x - \mathbb{E}[x]]^k$

If there exist a value k such that:

- (a) $\mathbb{E}[x^k] = g(\Theta)$
- (b) $\mathbb{E}[x - \mathbb{E}[x]]^k = h(\Theta)$

Applications

Let consider $X_1, ..., X_n$ exponential distributed and *i.i.d*

Solution:

…

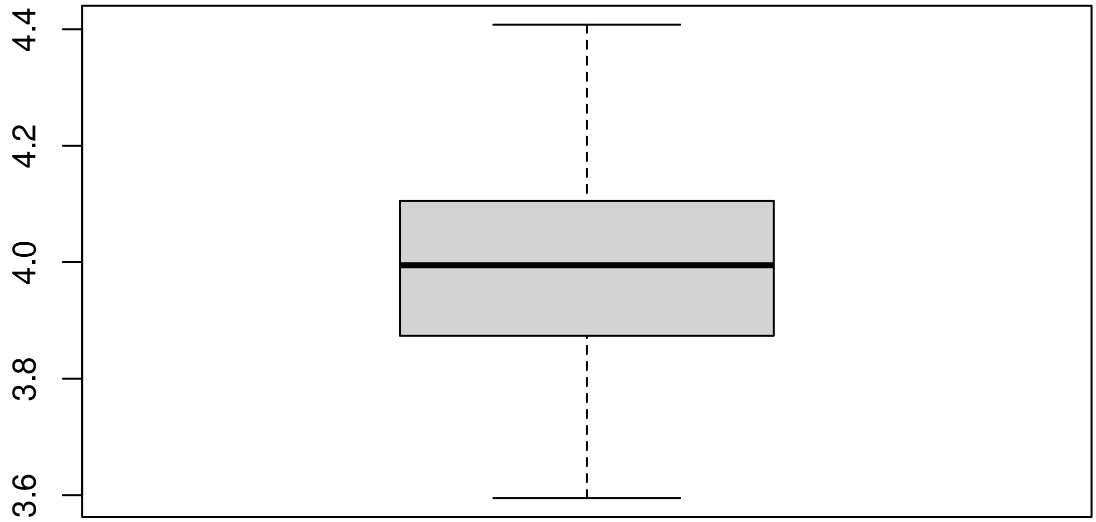
```
A = rexp(500, 4)

1/mean(A)
```

```
## [1] 3.891916
```

```
m = c()

for (i in 1:50) {
  A = rexp(500, 4)
  m[i] <- 1/mean(A)
}


mean(m)
```

```
## [1] 3.998223
```

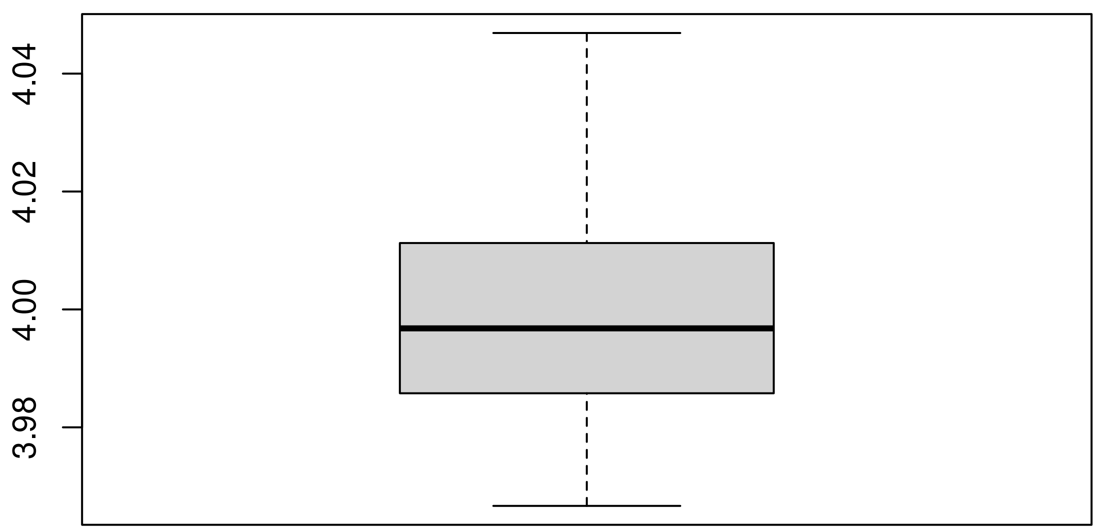
```r
boxplot(m)
```




```r
## With more observations we got less variation (500 -> 5000)

## Law of large numbers

m = c()

for (i in 1:50) {
  A = rexp(50000, 4)
  m[i] <- 1/mean(A)
}


mean(m)
```

```
## [1] 3.997474
```

```r
boxplot(m)
```



The Maximum Likelihood **Def:** likelihood

Let $X_1, \dots, X_n$ independent random variables, whose distributions are all depending on the same parameter Θ.

Let $x_1, \dots, x_n$ observations of those r.v

$$\mathcal{L}(x_1, \dots, x_n, \Theta) = \begin{cases} \Pi \\ \Pi \end{cases}$$

Def: Estimator thanks to the maximum likelihood.

$\hat{\Theta}_n$, an estimator for Θ, due to the maximum likelihood, is solution of:

$$\mathcal{L}(x_1, \dots, x_n, \Theta) = \max_{\theta} \mathcal{L}(x_1, \dots, x_n, \Theta)$$

Applications

Let consider $X_1, \dots, X_n$ $\xi(\lambda)$ *i.i.d.*

Compute the maximum likelihood estimator.

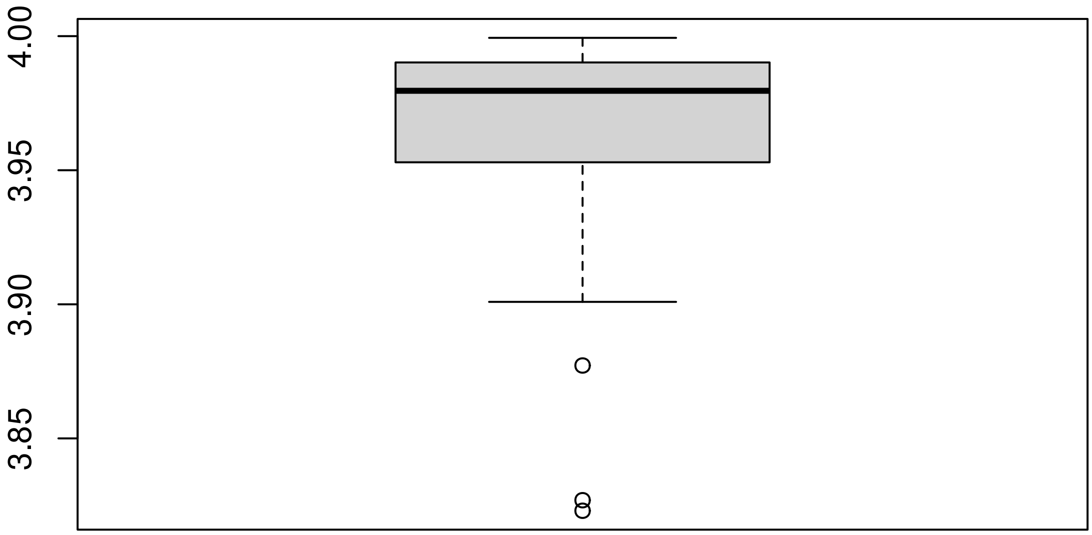
Solution:

…

```
n= 100

U = runif(n, 0, 4)

theta = max(U)


for (i in 1:50) {
  U = runif(n, 0, 4)
  theta = c(theta, max(U))
}


mean(theta)
```

```
## [1] 3.965163
```

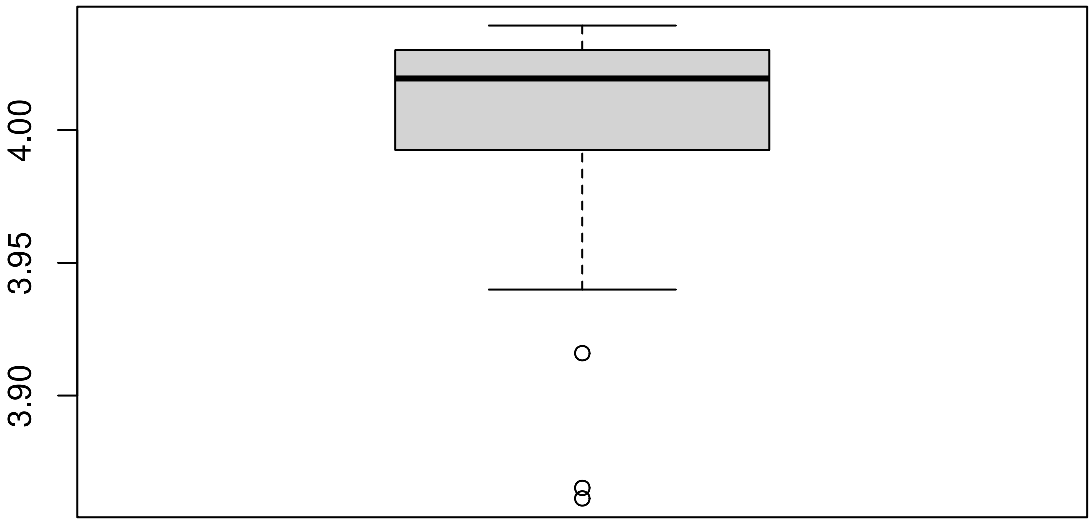
```
boxplot(theta)
```




```r
## Adjust to make the estimator unbiased

thetab = (n+1)/n*theta

mean(thetab)
```

```
## [1] 4.004815
```

```r
boxplot(thetab)
```

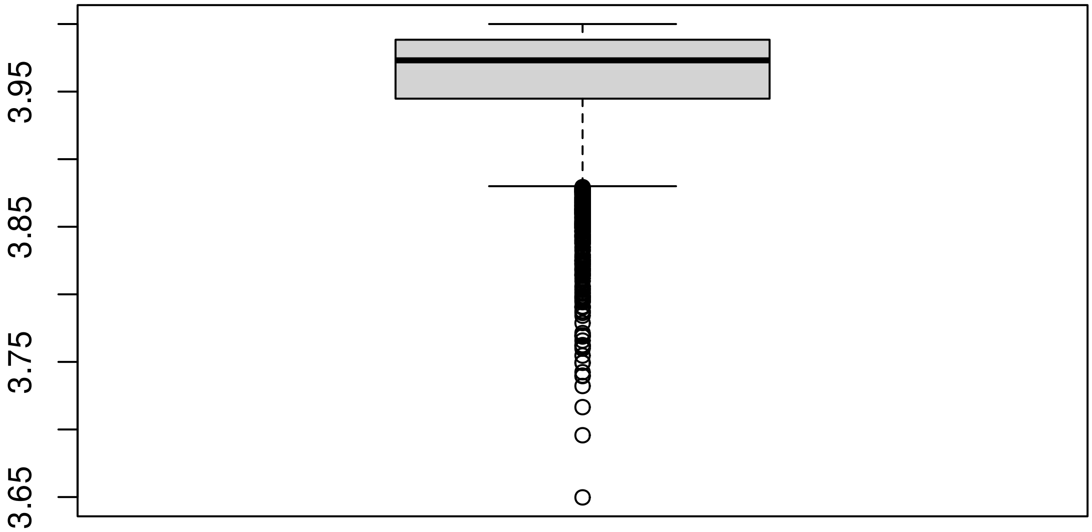



```r
## With more observations

n= 100

U = runif(n, 0, 4)

theta = max(U)


for (i in 1:5000) {
  U = runif(n, 0, 4)
```

```r
  theta = c(theta, max(U))
}


mean(theta)
```

```
## [1] 3.960279
```

```r
boxplot(theta)
```



Property

Let $X_1, \ldots, X_n$ *i.i.d* r.v. Let $\mu = \mathbb{E}[X_1]$ (unknown) Let $\sigma^2 = V(X_1)$ (unknown)

A classical estimator is:

- μ is $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum X_i$
- σ^2 is $\hat{\sigma}^2_n = \frac{1}{n}\sum(X_i - \overline{X}_n)^2$

Exercise

Show that:

1. $\hat{\mu}$ is unbiased.
2. $\hat{\sigma}^2_n$ is biased and that $\dfrac{n}{n-1}\hat{\sigma}^2_n$ is unbiased.