Università degli Studi di Pavia
Dipartimento di Ingegneria Industriale e dell'Informazione

Corso di Identificazione dei Modelli e Analisi dei Dati

Random Variables (part 2)

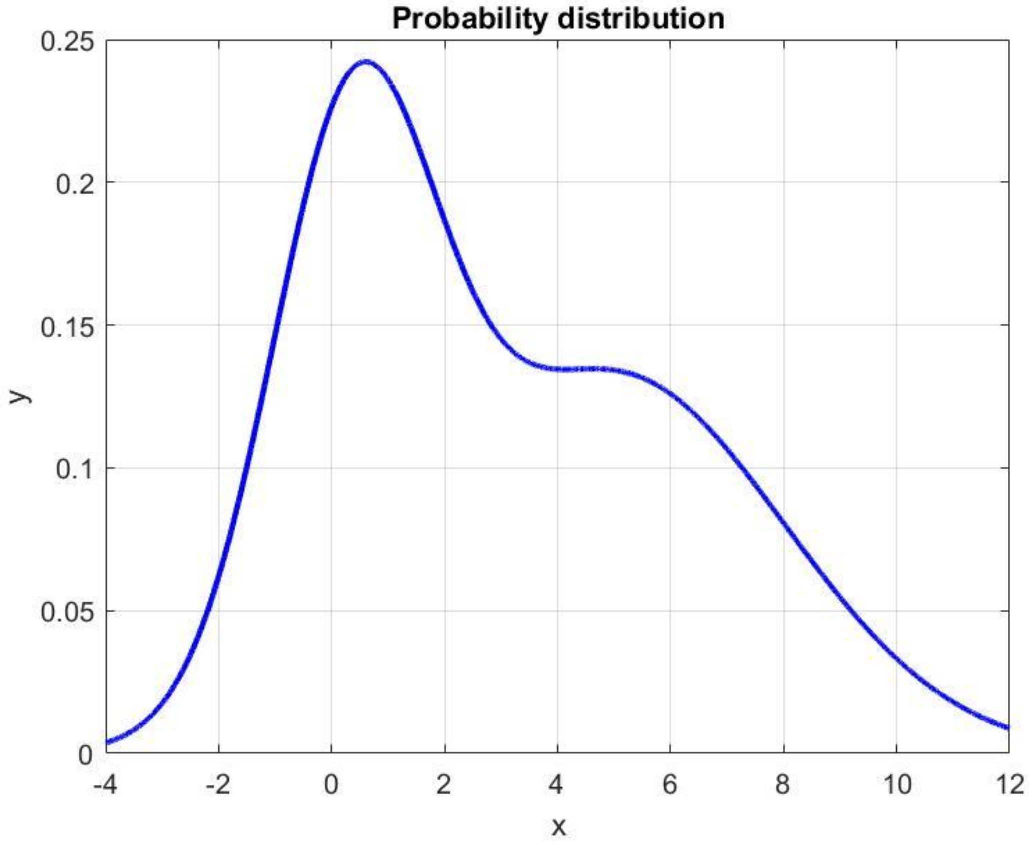
Prof. Giuseppe De Nicolao, Federica Acerbi, Alessandro Incremona

Outline

1. Theoretical parameters of a random variable
2. Sample parameters of a random variable
3. Sample mean and sample median
4. Functions of random variables

Parameters of a random variable

- Mode
- Mean
- Median
- Quantiles

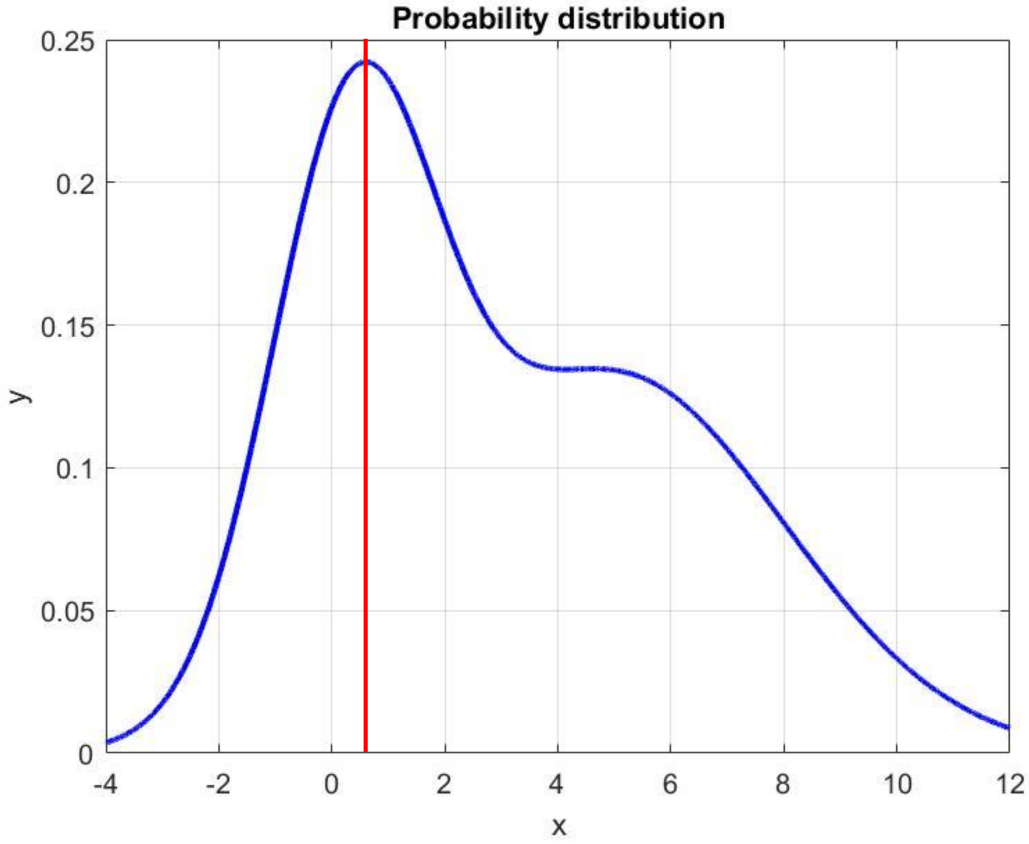


Parameters of a random variable

- **Mode**

$$\underset{x}{\operatorname{argmax}}(f_X(x))$$

Cons: 1) It may not be unique.
2) It may not be significant.

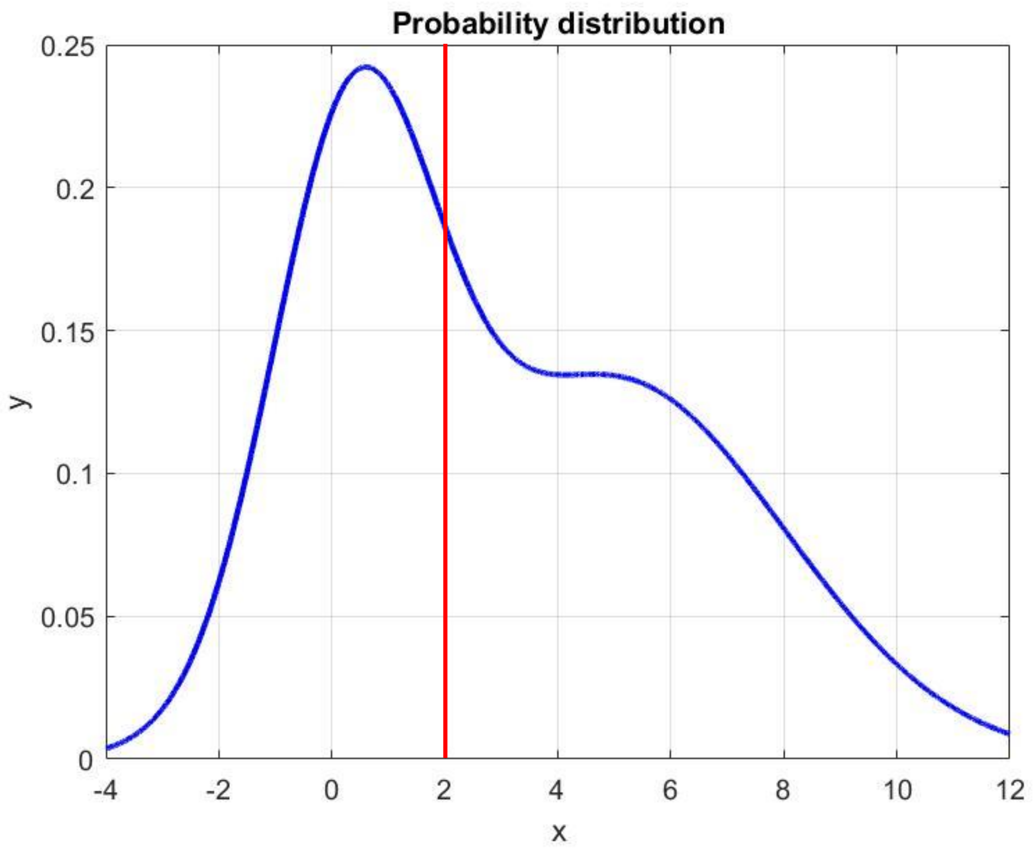


Parameters of a random variable

- **Mean**

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

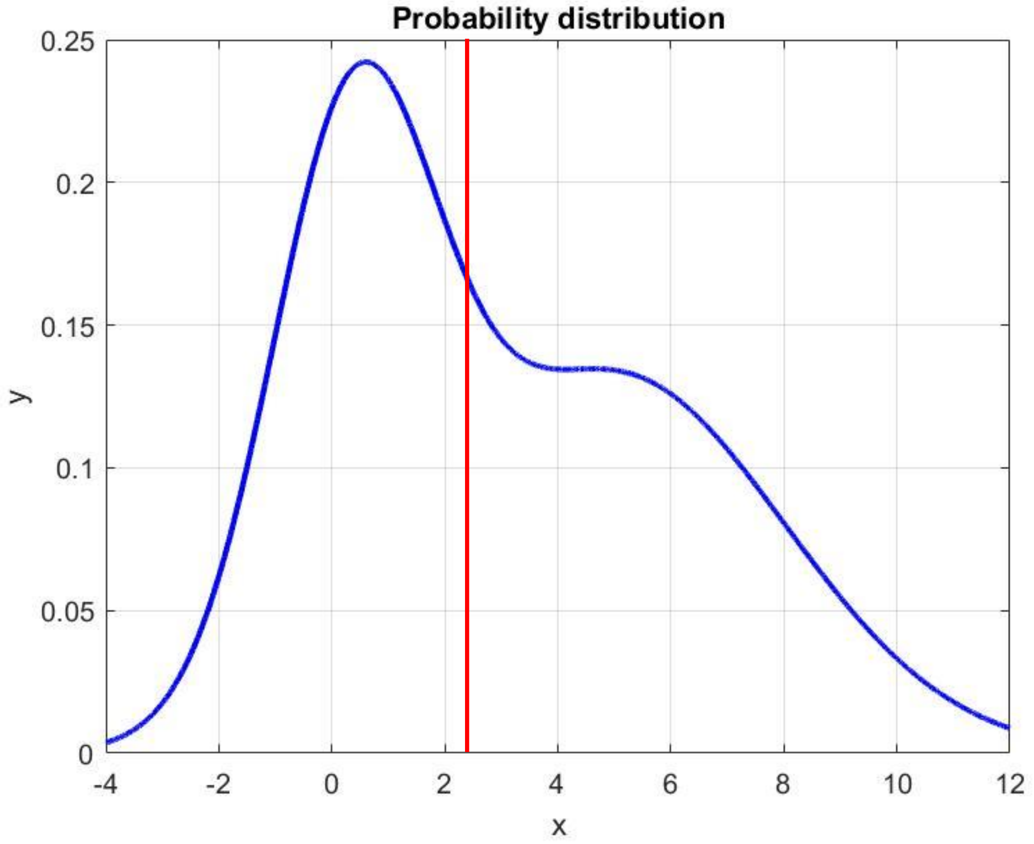
It is the centroid of the distribution.



Parameters of a random variable

- **Median**

$$F_X^{-1}(0.5)$$



Cons: It may not be unique.

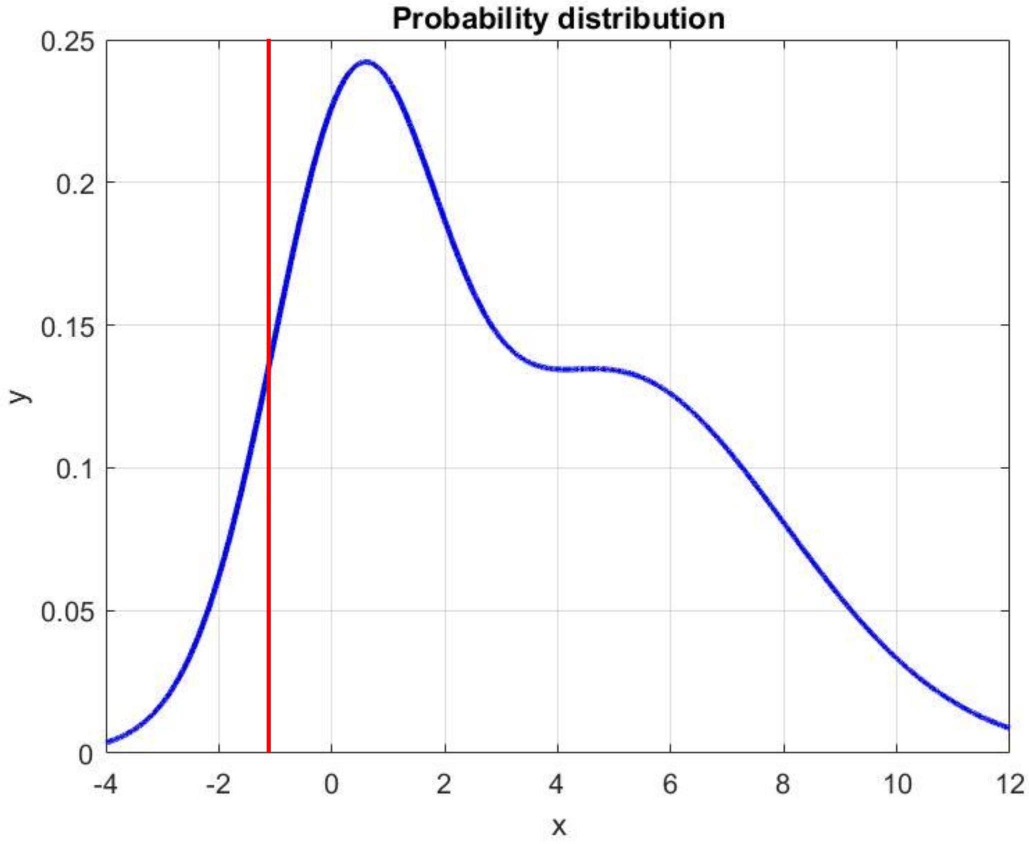
Parameters of a random variable

- **Quantiles**

e.g.: the 0.25-quantile
(i.e. the first quartile)

$$F_X^{-1}(0.25)$$

The median is the quantile $x_{0.5}$



Calculate quantiles with MATLAB

$$\underbrace{F_X}_{\text{Cumulative Distribution Function}}(x_\alpha) = \alpha \implies x_\alpha = \underbrace{F_X^{-1}}_{\text{Inverse Cumulative Distribution Function}}(\alpha)$$

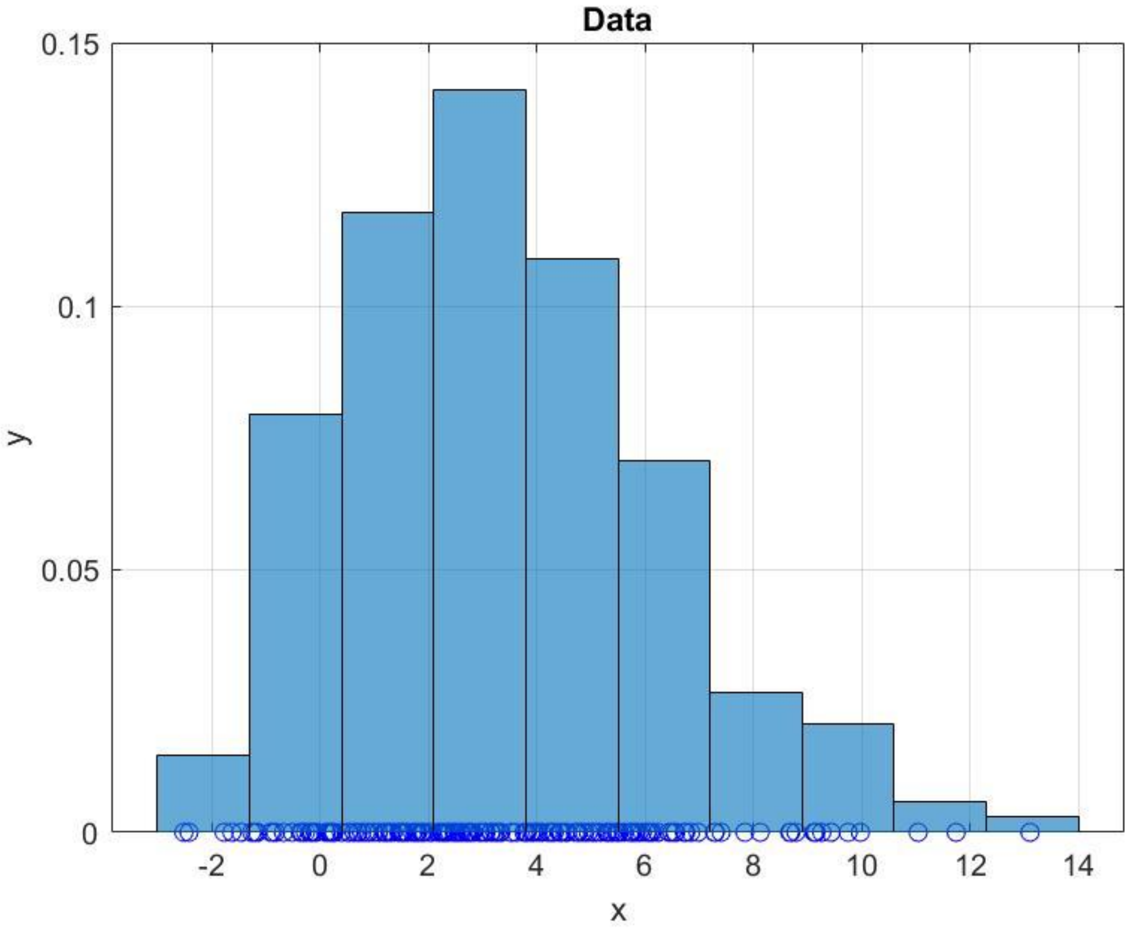
```
doc icdf
```

Exercise 1

1. Create a Normal distribution object with m = 1 and sigma = 2.
2. Calculate the theoretical median, first quartile and third quartile using `icdf` command.
3. Open a figure and create the default Cartesian axes using `hax = axes`.
4. Plot the theoretical distribution (use `linspace` to create a grid of x-values and `pdf` to obtain the y-values).
5. Plot, in the same graph, 4 vertical asymptote in correspondence of the theoretical mean, median, first quartile and third quartile.→ `plot([m m], get(hax, 'YLim'), 'r-')`

Estimate the parameters of a distribution from the data

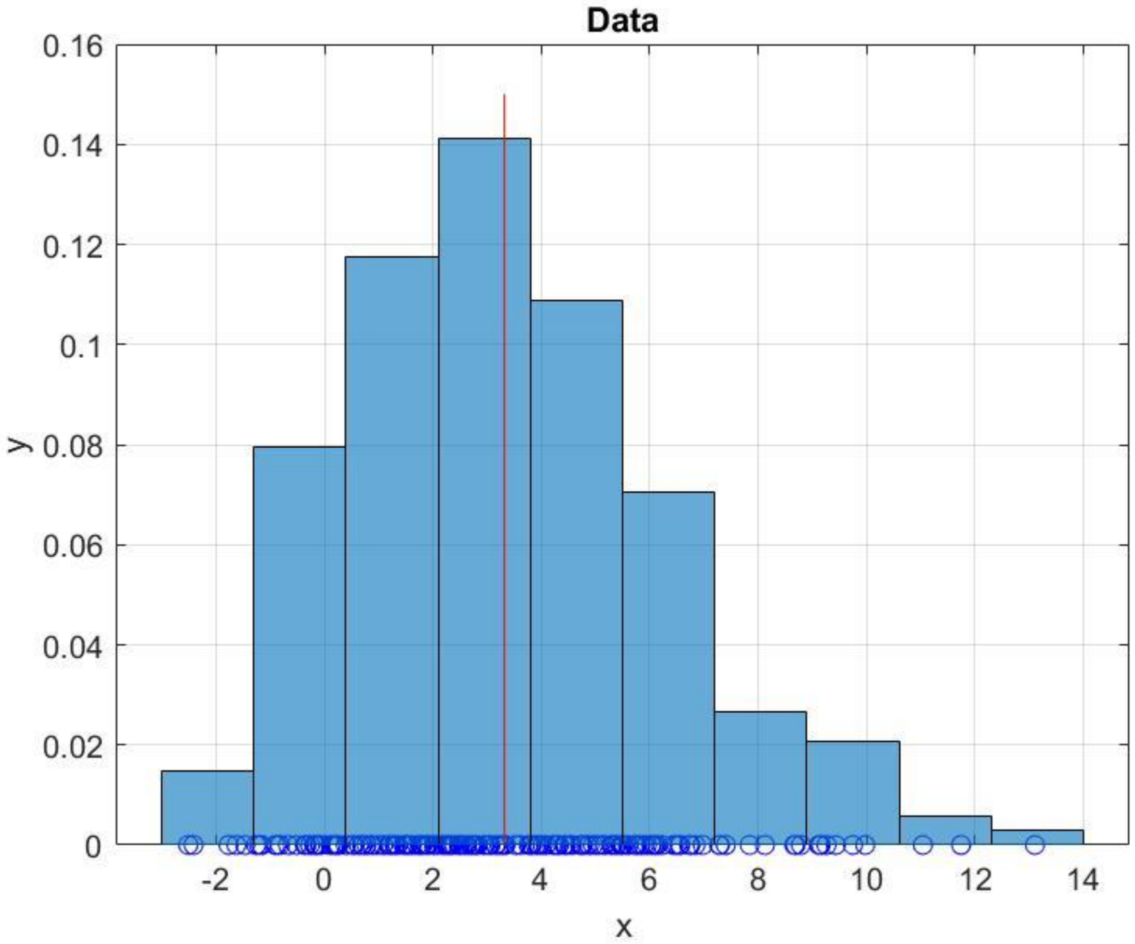
- Sample mean
- Sample median
- Sample quantiles



Estimate the parameters of a distribution from the data

- Sample mean

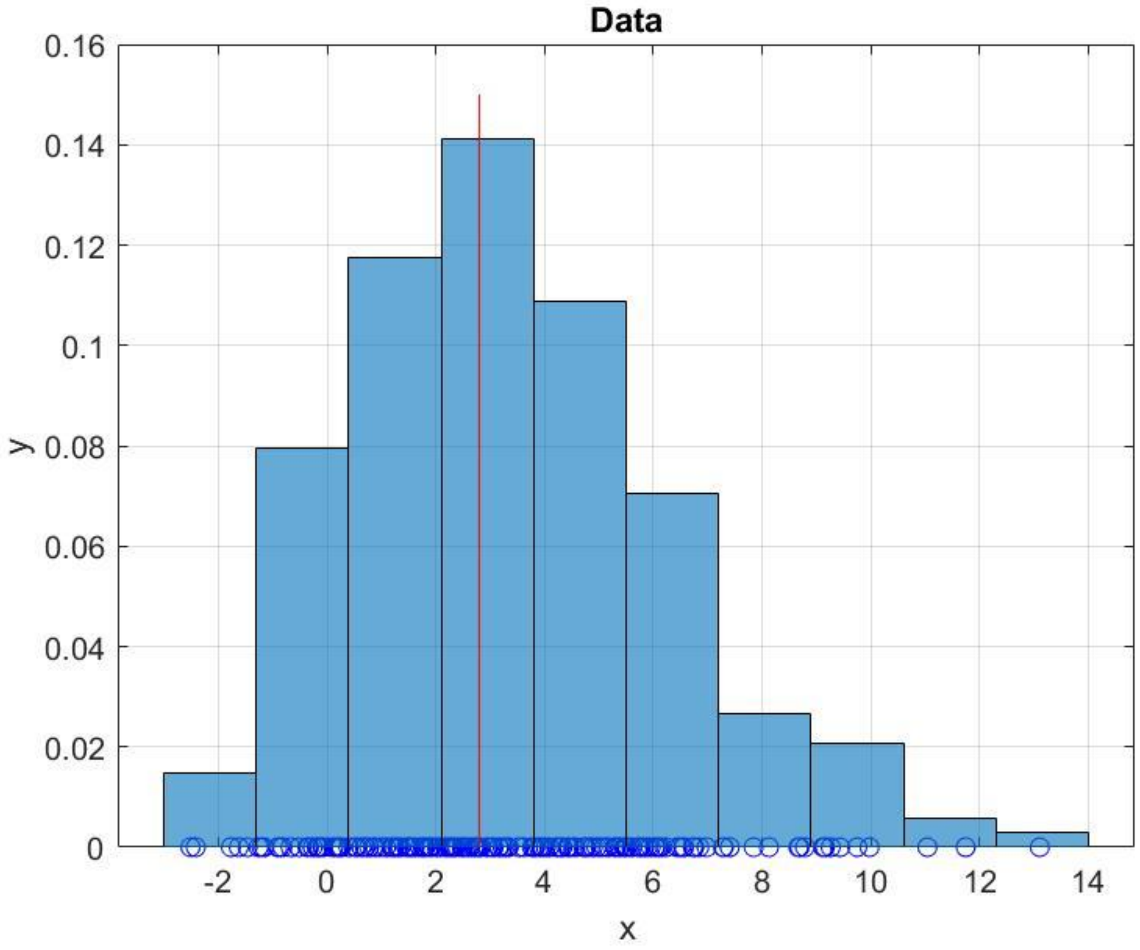
$$\frac{1}{N}\sum_{i=1}^{N} x_i$$



Estimate the parameters of a distribution from the data

- **Sample median**

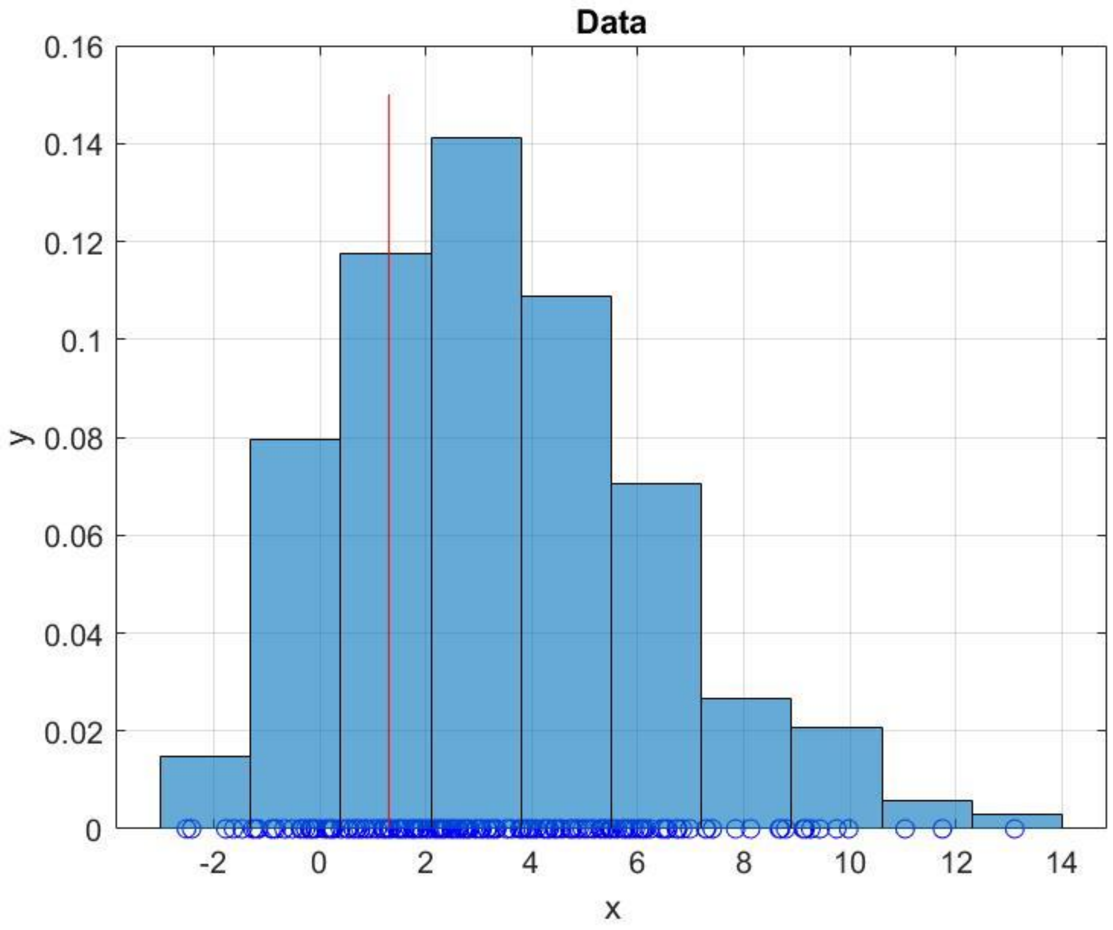
The ''middle'' value



Estimate the parameters of a distribution from the data

- Sample quantiles

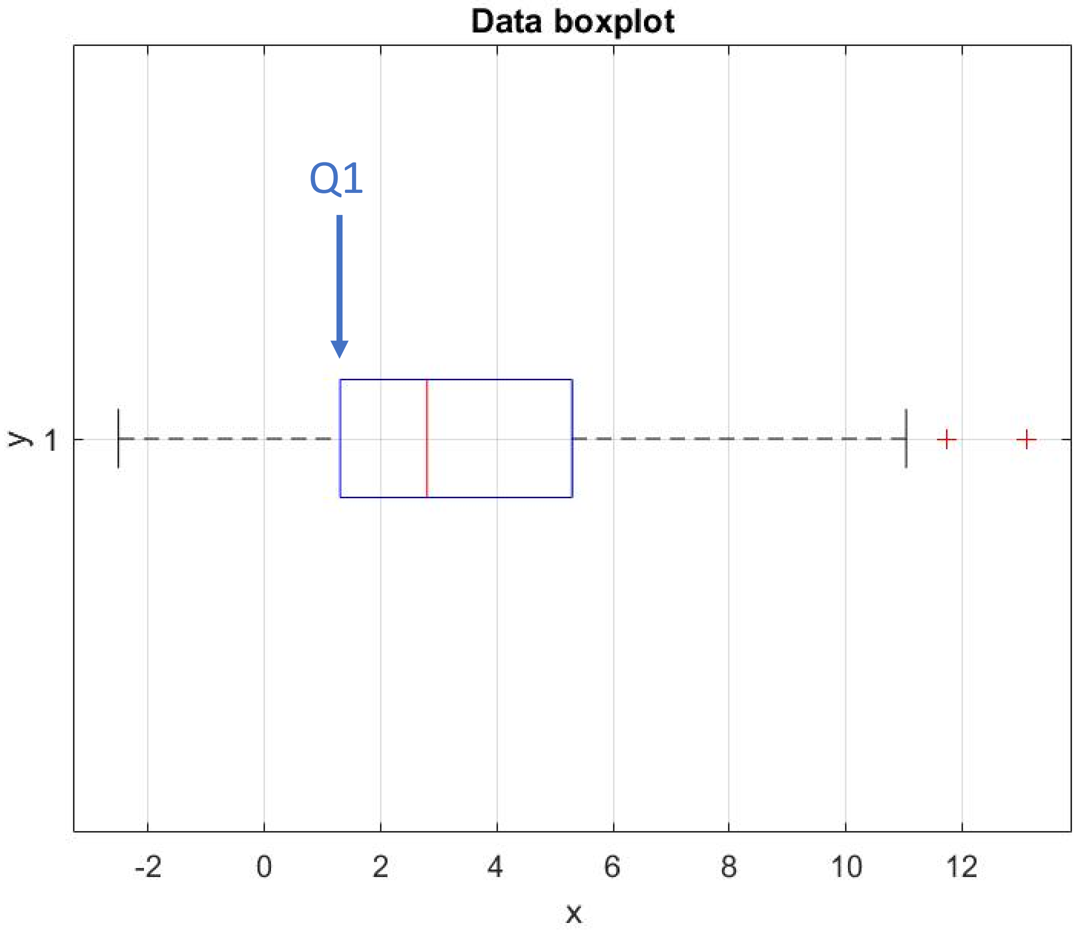
Sorting-based algorithm (see documentation for `quantile`)



Estimate the parameters of a distribution from the data

- **Sample quantiles**

Sorting-based algorithm (see documentation for `quantile`)



Calculate sample parameters with MATLAB

Sample mean: → `doc mean`

Sample median: → `doc median`

Sample quantile: → `doc quantile`

Exercise 2

1. Create a Log-normal distribution object with m = 1 and sigma = 0.5.
2. Calculate the theoretical mean, median, first quartile and third quartile.
3. Plot the theoretical distribution (use `linspace` to create a grid of x-values and `pdf` to obtain the y-values)→ for the grid use: `x_grid = linspace(m - 4*sigma, m + 16*sigma, 10000)`.
4. Plot, in the same graph, the theoretical mean as a vertical asymptote.
5. Simulate 1000 random values from the theoretical distribution and compute the sample mean.
6. Plot, in the same graph, the sample mean as a vertical asymptote.
7. Repeat the exercise considering the median instead of the mean.

Sample mean and sample median

The sample mean accuracy increases with the number of samples (without outliers).

The sample median is robust against outliers. If you throw away the largest and smallest values in a data set then the **median** does not change but **the sample mean** does.

Exercise 3

1. Repeat the Exercise 2 with a bigger dataset and then with a smaller dataset.

2. Add an outlier to the simulated data (use the square brackets to concatenate a new value to the data vector generated with `random`) and recompute and plot the sample mean.

3. Repeat the exercise considering the median.

Functions of random variables

Example:

$$Y = 2X + 10 \quad \text{with} \quad X \sim \mathcal{N}(0,\, 1)$$

- The value of $f(X)$ depends on the value of X (random variable)
- ... and therefore on the result of an experiment!

In this case, Y is called **stochastic function**:

- It is a random variable
- It can be evaluated running an experiment
- It has its own probability distribution

Exercise 4

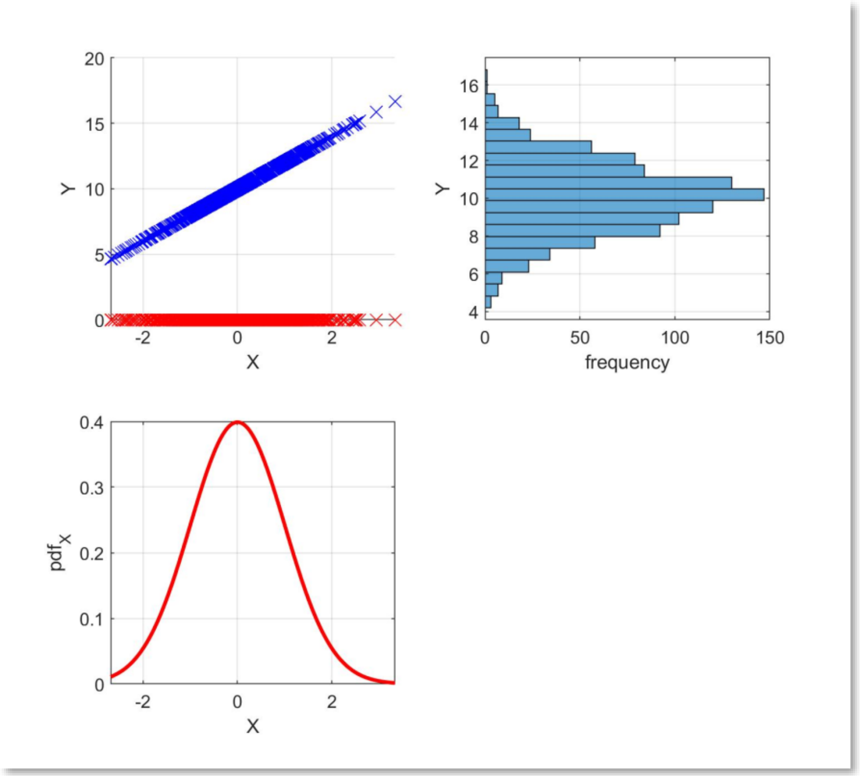
Example:

$$Y = 2X + 10 \qquad \text{with} \quad X \sim \mathcal{N}(0,\, 1)$$

1. Generate X as a [1000x1] vector of random numbers sampled from a `normal` distribution with `mean = 0` and `variance =1`

2. Draw the theoretical probability distribution of X

3. Compute $Y = f(X)$ and approximate its probability distribution using the `histogram` function



Exercise 5

1. Replicate the exercise n° 4 with

   – X : a vector (dimension [1000x1]) of random numbers sampled from a `lognormal` distribution

   – $Y = 2\log(X) + 10$

Reference Documentation:

- https://it.mathworks.com/
- http://sisdin.unipv.it/labsisdin/teaching/courses/imadlt/esercitazioni