Statistics Refresher I

18th JRC Annual training on Composite Indicators and Scoreboards

Elizabeth Casabianca

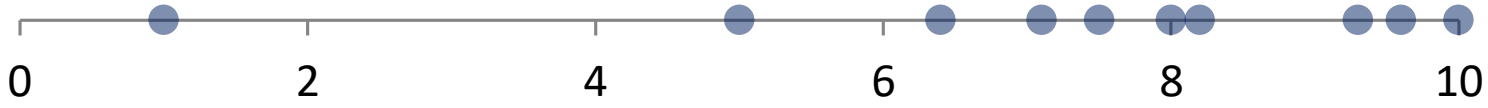
Notation

• An indicator:	x	e.g. GDP
• jth value of an indicator:	x_j	e.g. GDP for country j
• ith indicator (in a group):	x_i	e.g. group is GDP, life expectancy, median income, …
• jth value, ith indicator:	$x_{i,j}$	e.g. GDP for country j
• Sample size	N	e.g. number of countries

Some data

x_i
9.6
10
9.3
7.1
5
8
7.5
8.2
6.4
1

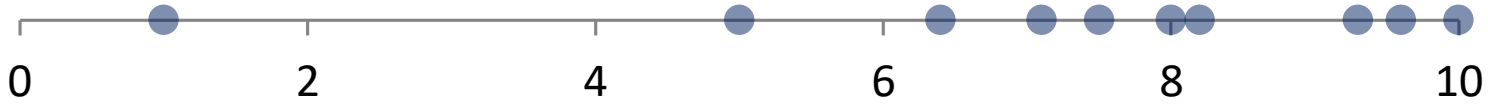
COIN Training 2019 satisfaction scores



Rank

x_i	Rank
9.6	2
10	1
9.3	3
7.1	7
5	9
8	5
7.5	6
8.2	4
6.4	8
1	10

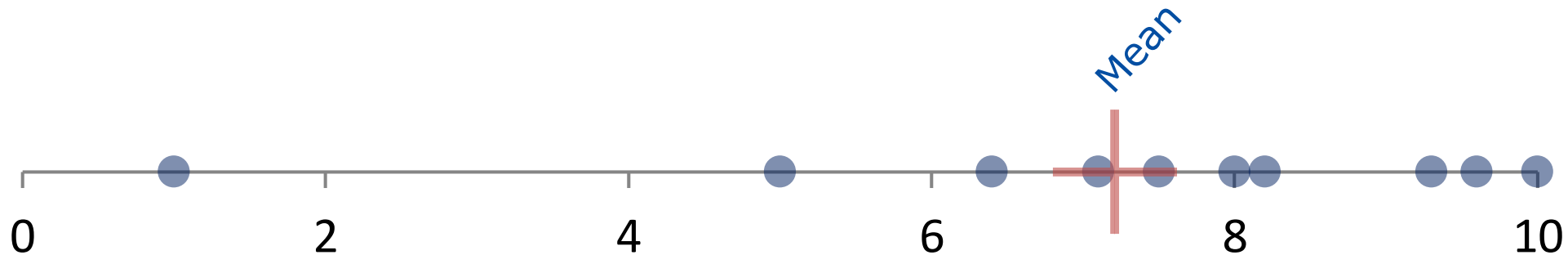
COIN Training 2019 satisfaction scores



Mean

$$\mu_i = E(x_i) = \frac{1}{N}(x_{i,1} + x_{i,2} + x_{i,3} + \cdots + x_{i,N}) = \frac{1}{N}\sum_{j=1}^{N} x_{i,j}$$

$$\mu_i = \frac{1}{10}(9.6 + 10 + 9.3 + 7.1 + 5 + 8 + 7.5 + 8.2 + 6.4 + 1)$$



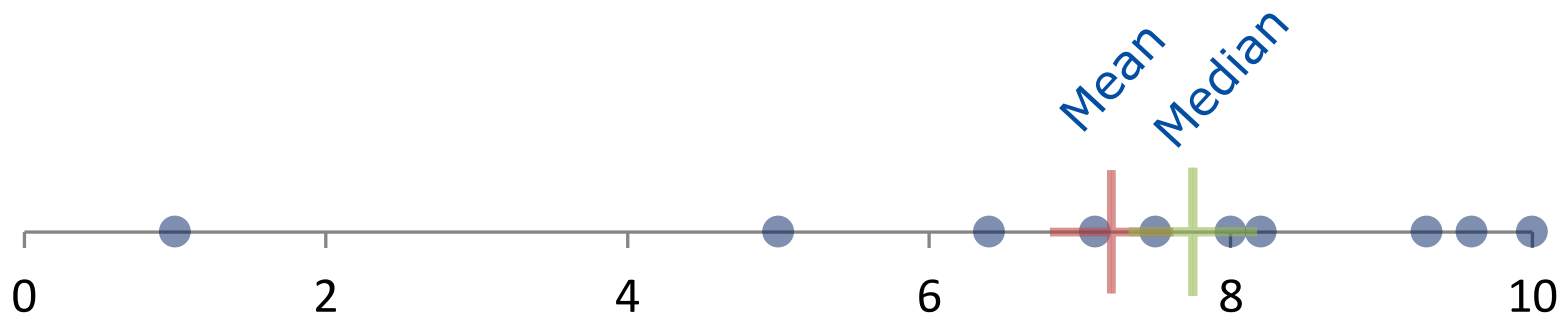
Median

x_i
1
5
6.4
7.1
7.5
8
8.2
9.3
9.6
10

The **middle value of the list of numbers**. Half of the values should be higher, half should be lower.

For an even number of values, it is the midway point between the two middle values.

The median is not equal to the mean unless it is a symmetric distribution.



Percentiles

x_i
1
5
6.4
7.1
7.5
8
8.2
9.3
9.6
10

20th percentile

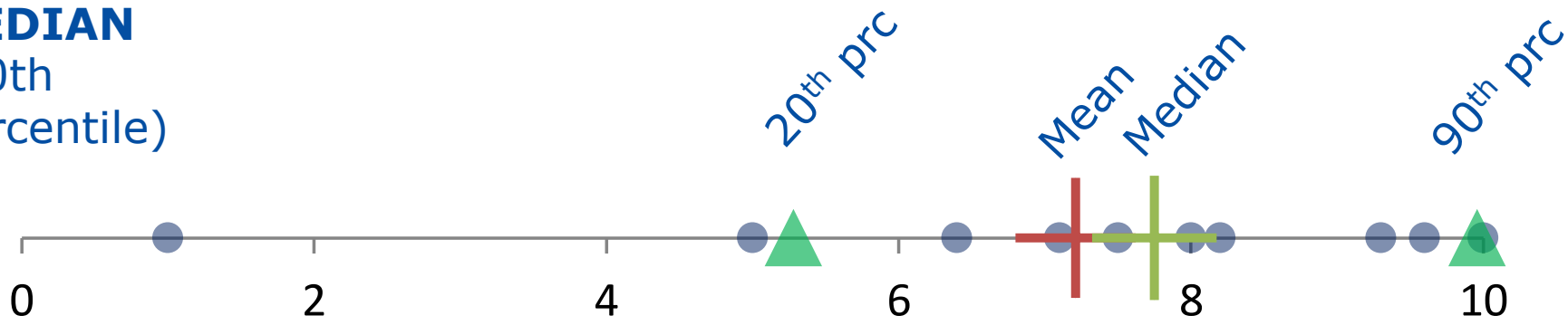
MEDIAN (50th percentile)

90th percentile

Percentiles are a generalisation of the median.

Median: value such that 50% of values are below it

xth percentile: value such that X% of values are below it



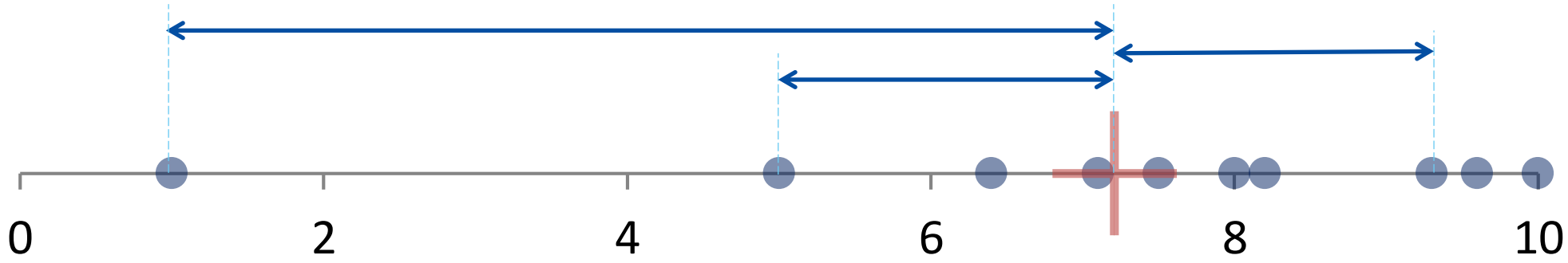
European Commission

Variance

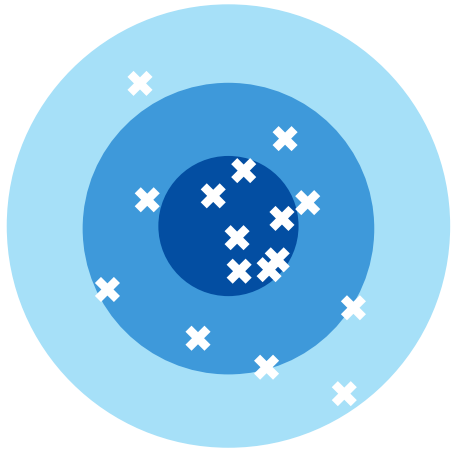
A measure of dispersion: how different are values from one another?

$$\mathrm{var}(x_i) = \sigma_i^2 = \frac{1}{N}\sum_{j=1}^{N}\underbrace{\left(\mu_i - x_{i,j}\right)^2}$$

The average squared distance of data points from the mean

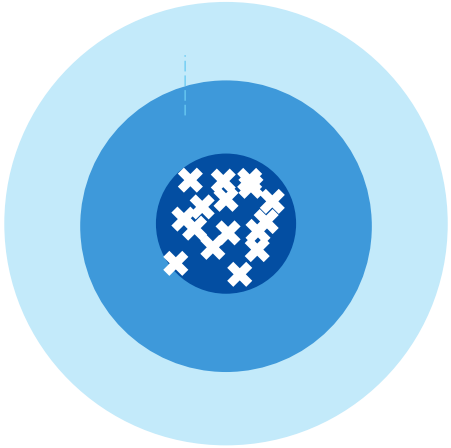


Variance and standard deviation



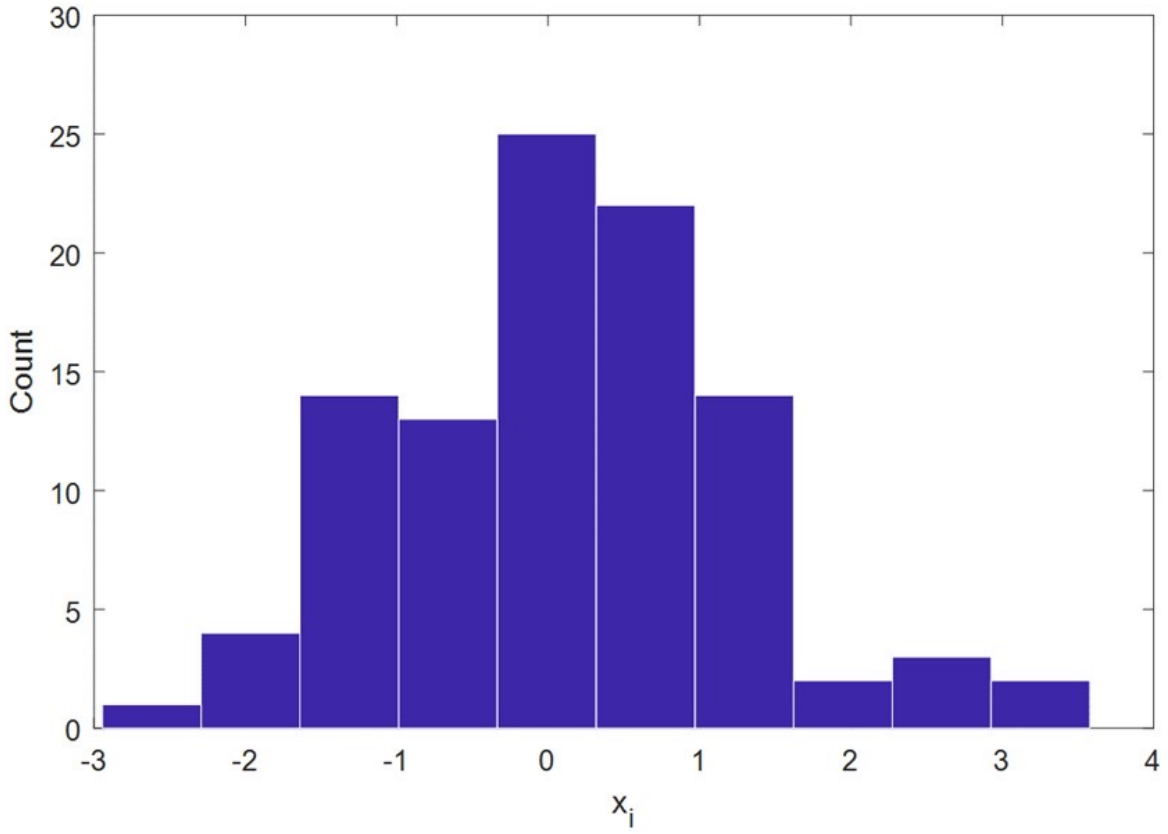
High variance

σ_i is the **standard deviation:** the square root of the variance



Low variance

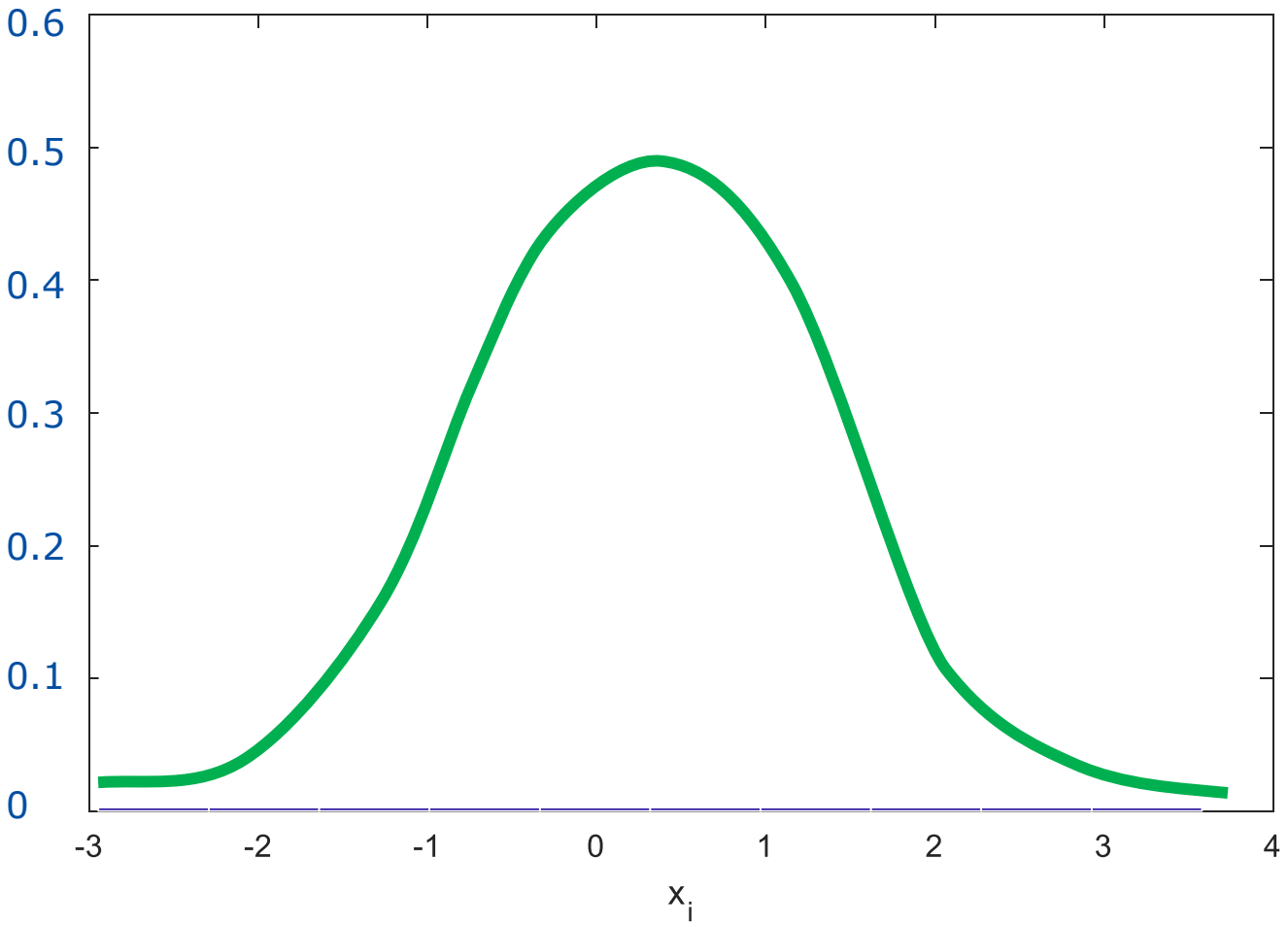
Probability distributions



A histogram shows how data is distributed into equally-spaced "bins".

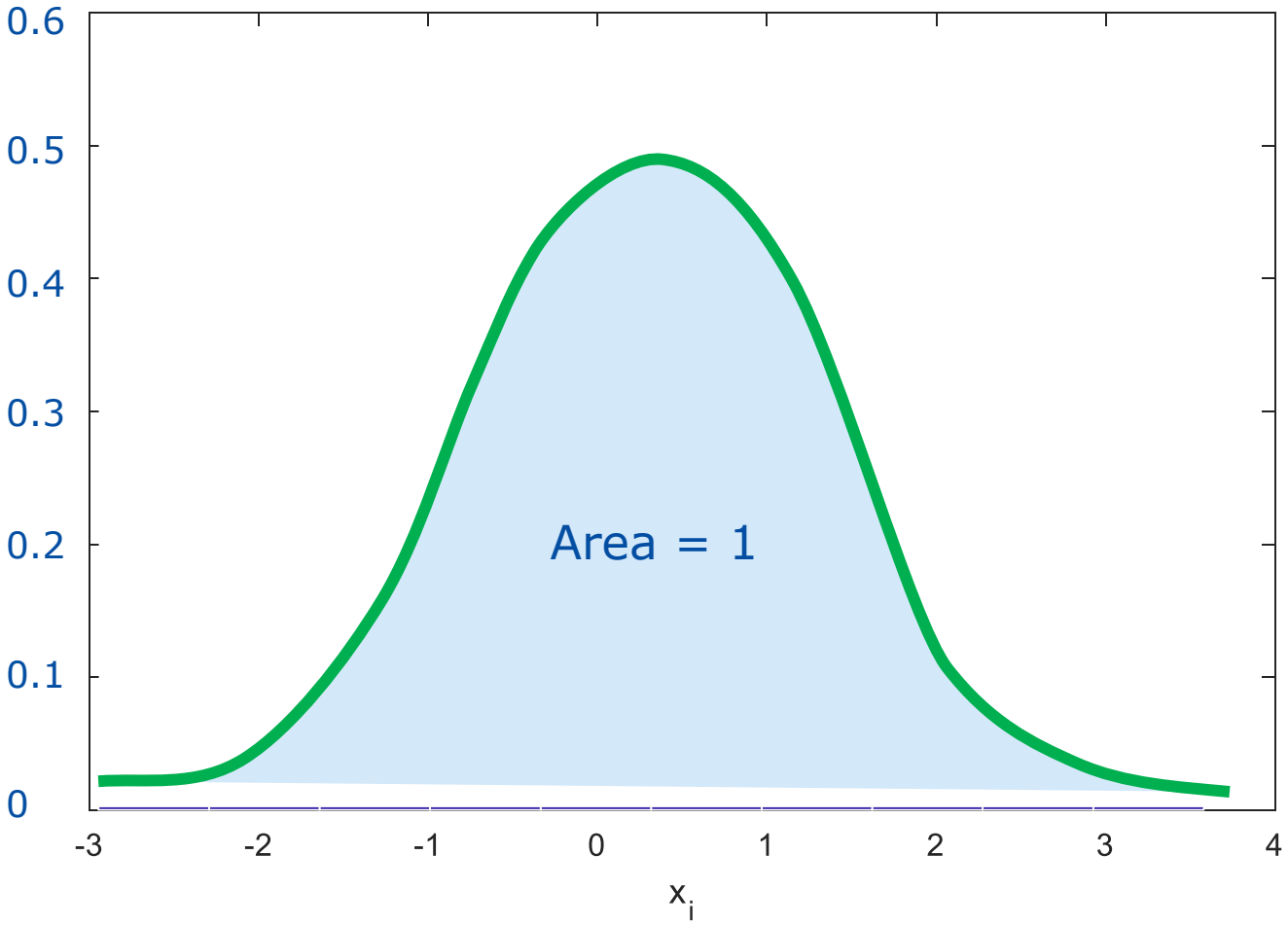
Each bar gives an indication of how frequently it is that a value will fall inside that range.

Probability distributions



A probability density function (pdf) shows the relative probability of different values occurring.

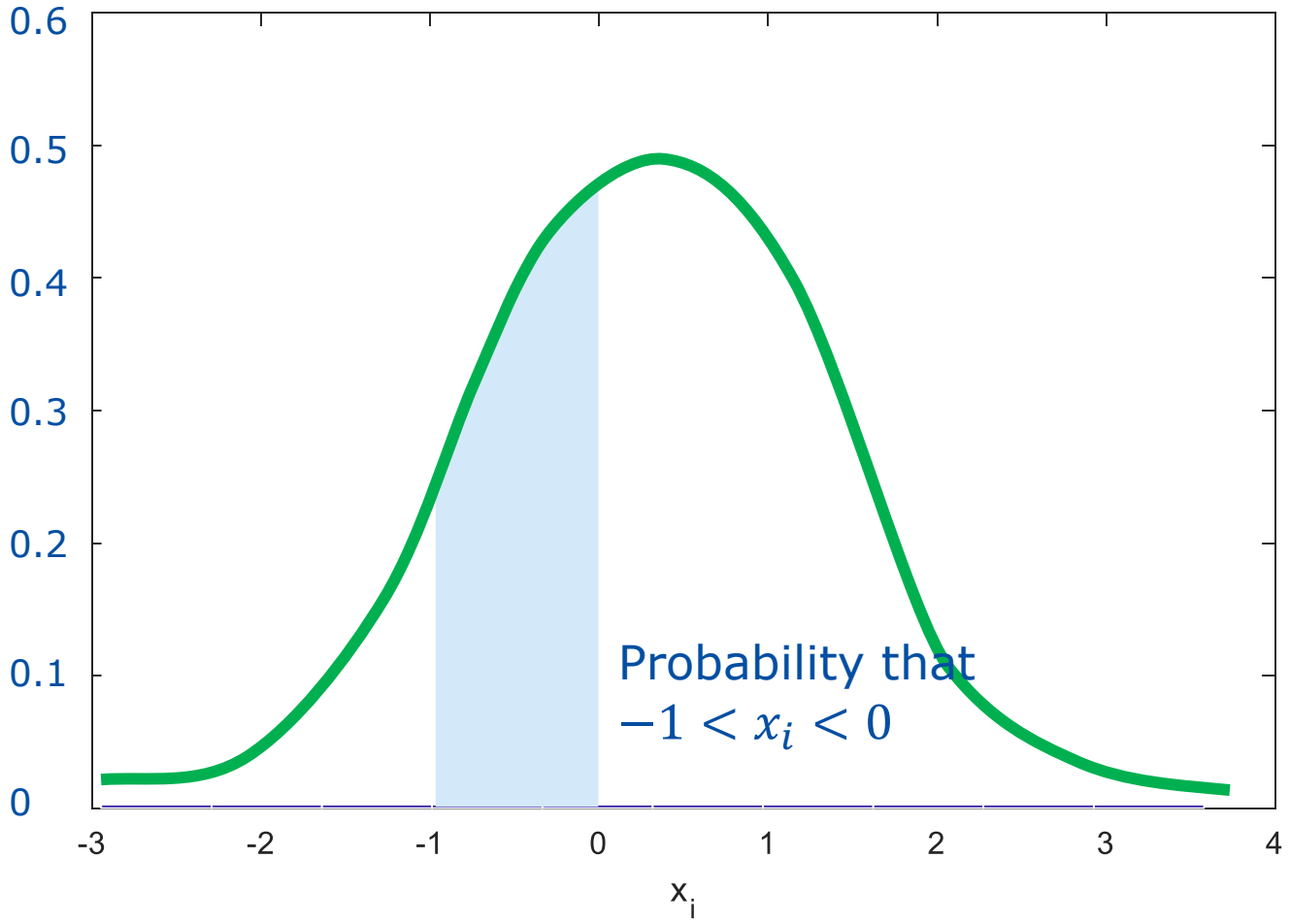
Probability distributions



A probability density function (pdf) shows the relative probability of different values occurring.

The total area under the pdf (integral) is equal to 1 by definition.

Probability distributions

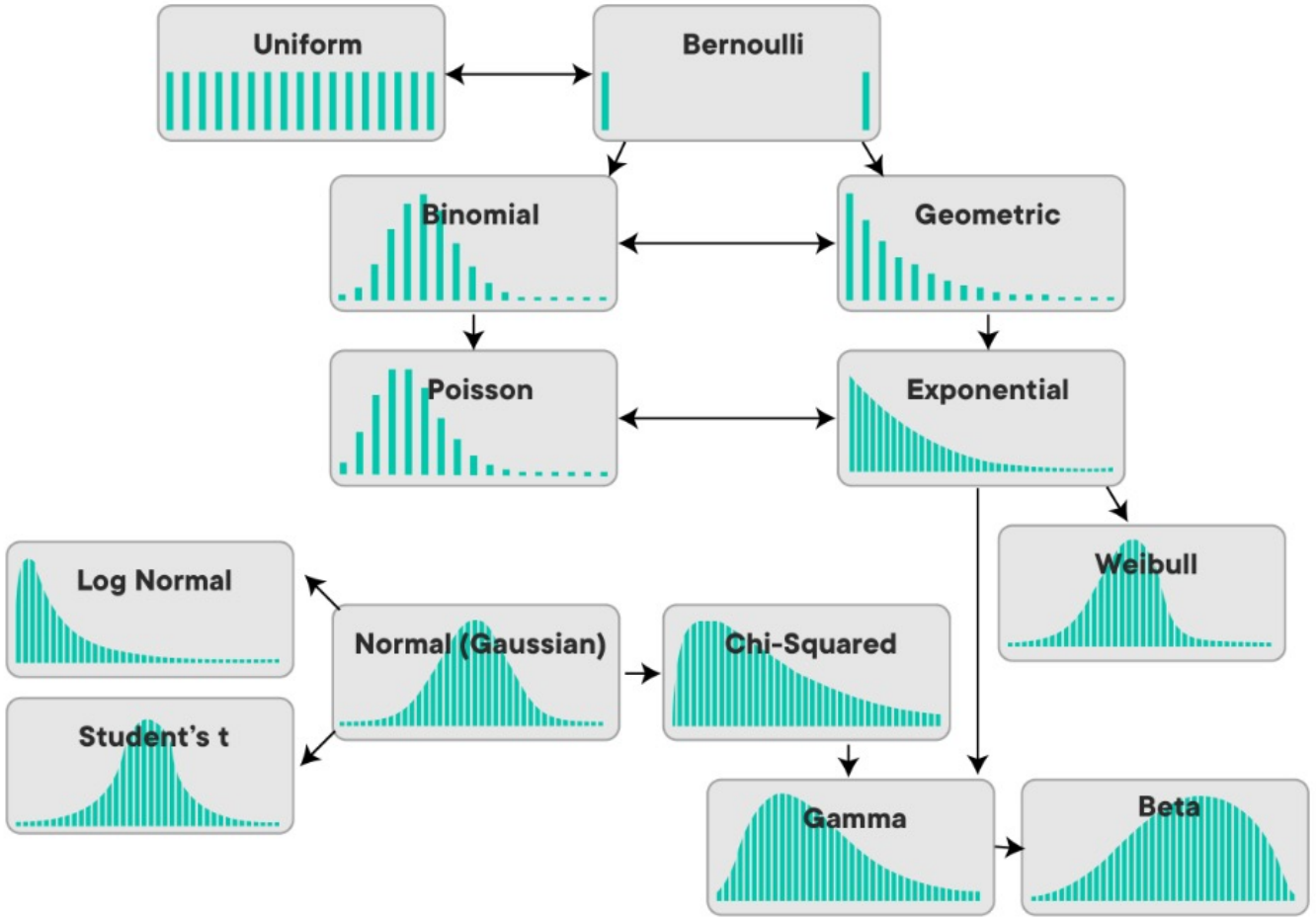


A probability density function (pdf) shows the relative probability of different values occurring.

The total area under the pdf (integral) is equal to 1 by definition.

The area between two points shows the probability of a value falling in that range.

Probability distributions



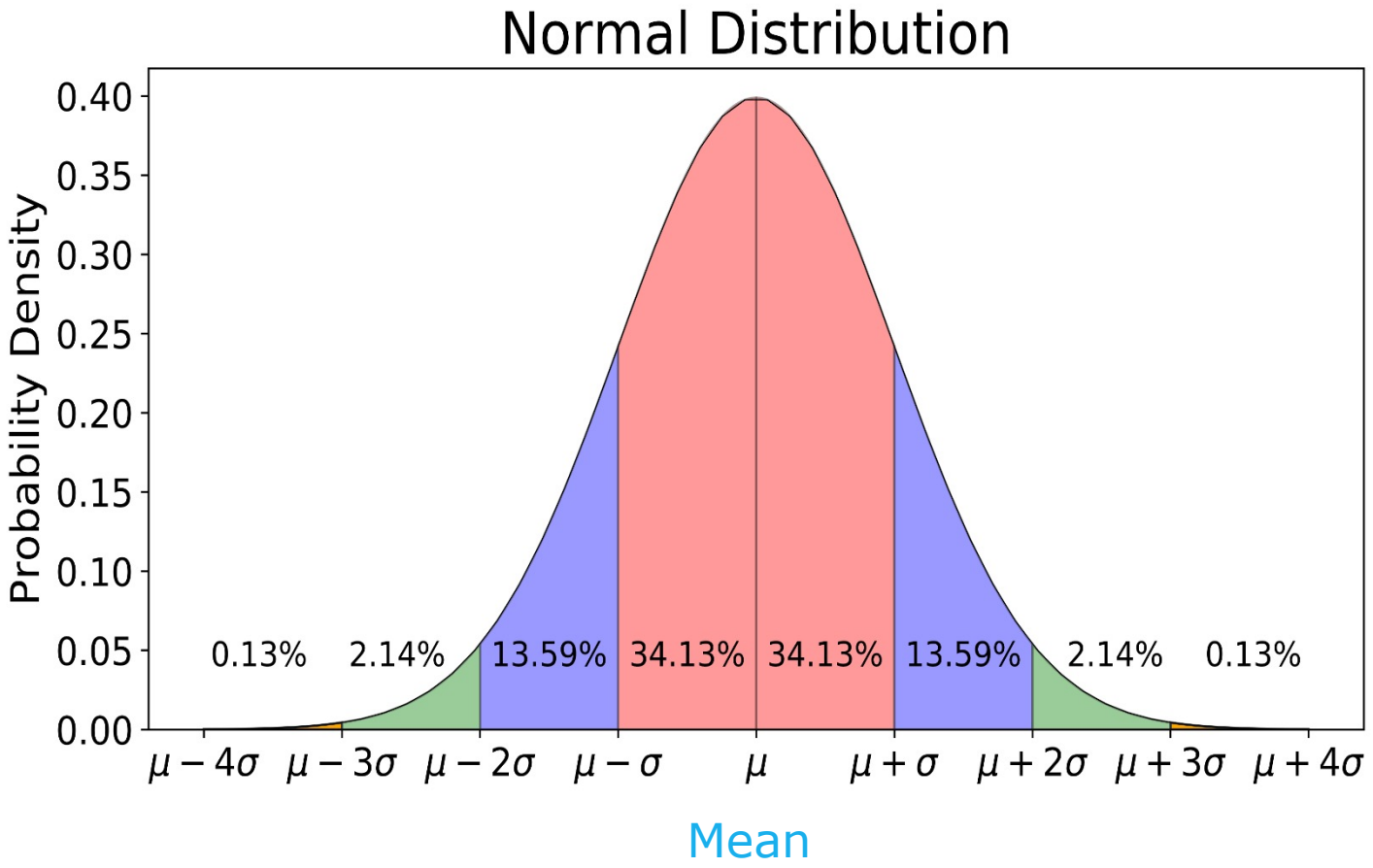
There are many theoretical types of probability distribution.

They aim to mimic real distributions observed in nature, economics, physics, etc.

Discrete distributions can only take a finite amount of values.

Continuous distributions can take an infinite number of values (in a finite or infinite range).

Normal distribution

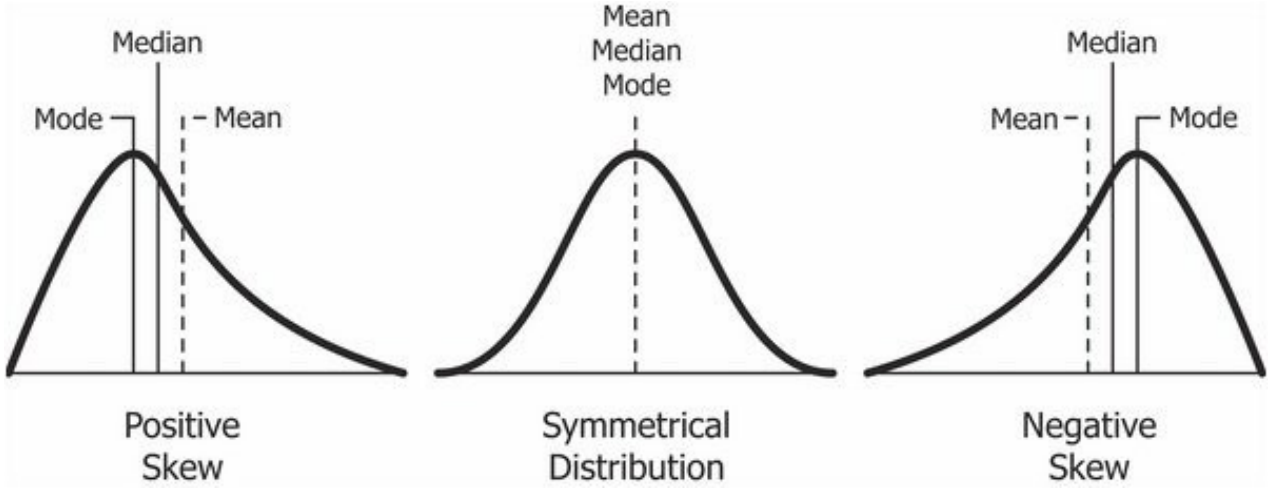


The Normal (Gaussian) distribution is the simplest type of distribution!

- Values of a random variable fall into a (smooth) continuous curve with a **bell shaped**, **symmetric pattern**

- Main characteristics:
  - Mean=mode=median
  - Symmetric
  - Asymptotic

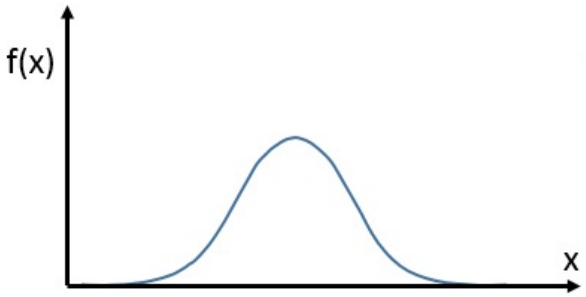
Skew and kurtosis



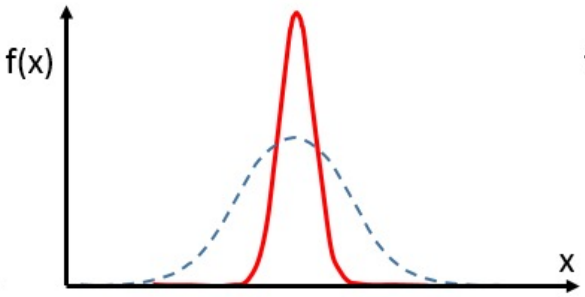
Skew shows the extent to which values are clustered towards one end of the scale. Measure of lack of symmetry.

Kurtosis measures how heavy the tails of the distribution are, compared to a normal distribution.

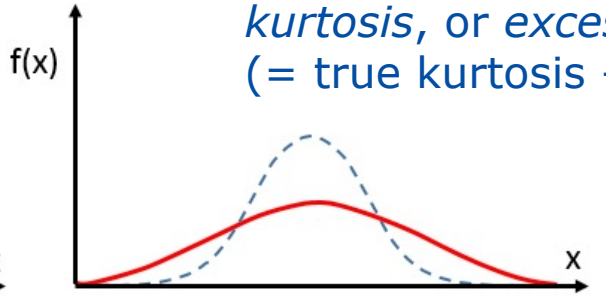
Be careful: "kurtosis" may mean *true kurtosis*, or *excess kurtosis* (= true kurtosis – 3)



Normal
Kurtosis = 3
Excess kurtosis = 0



Leptokurtic
Kurtosis > 3
Excess kurtosis > 0



Platykurtic
Kurtosis < 3
Excess kurtosis < 0

Thank you

European Commission