The European Commission's science and knowledge service

Joint Research Centre

European Commission

Statistics Refresher

William Becker

COIN 2019 - 16th JRC Annual Training on Composite Indicators & Scoreboards
04-06/11/2019, Ispra (IT)

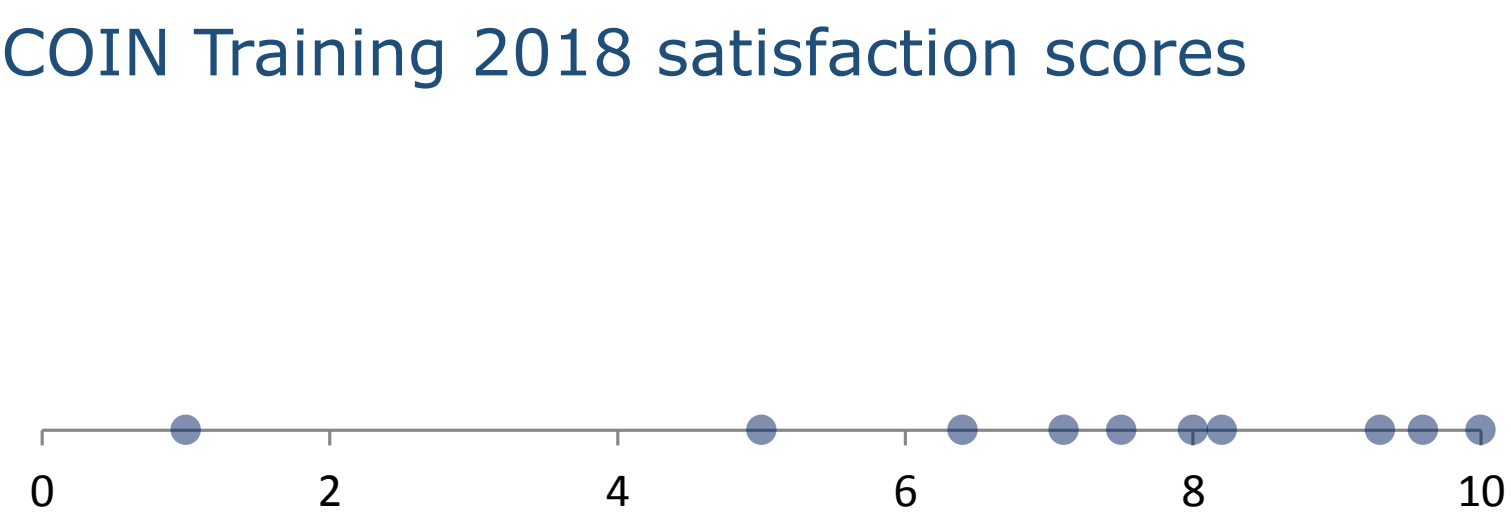
European Commission

Notation

An indicator:	x	e.g. GDP
*j*th value of an indicator:	x_j	e.g. GDP for country j
*i*th indicator (in a group):	x_i	e.g. group is GDP, life expectancy, median income, …
*j*th value, *i*th indicator:	$x_{i,j}$	e.g. GDP for country j
Number of countries	N	

Some data

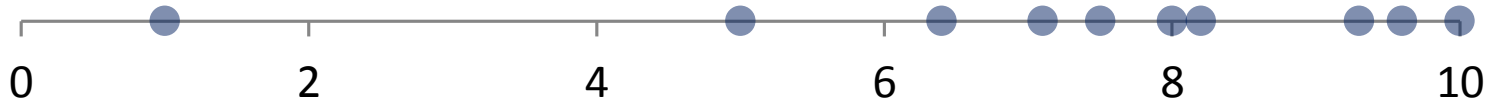
x_i
9.6
10
9.3
7.1
5
8
7.5
8.2
6.4
1



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

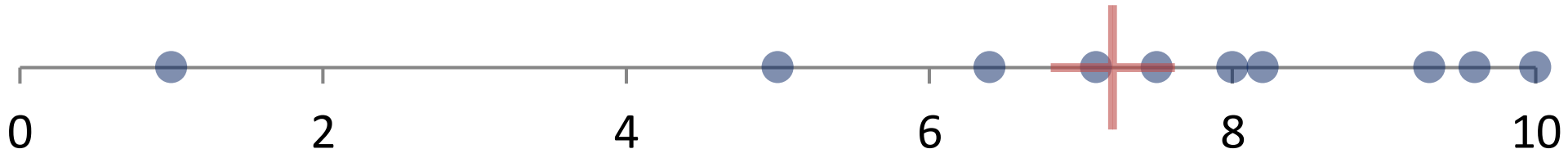
COIN Training 2018 satisfaction scores



Mean

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

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



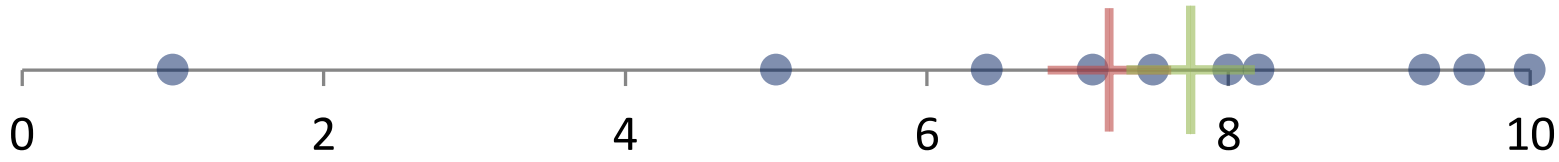
Median

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

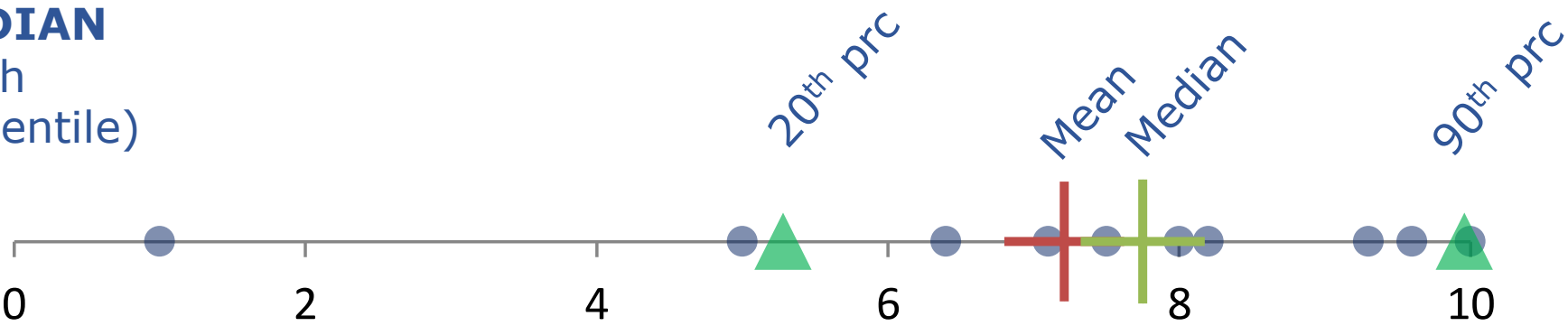
MEDIAN (50th percentile)

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

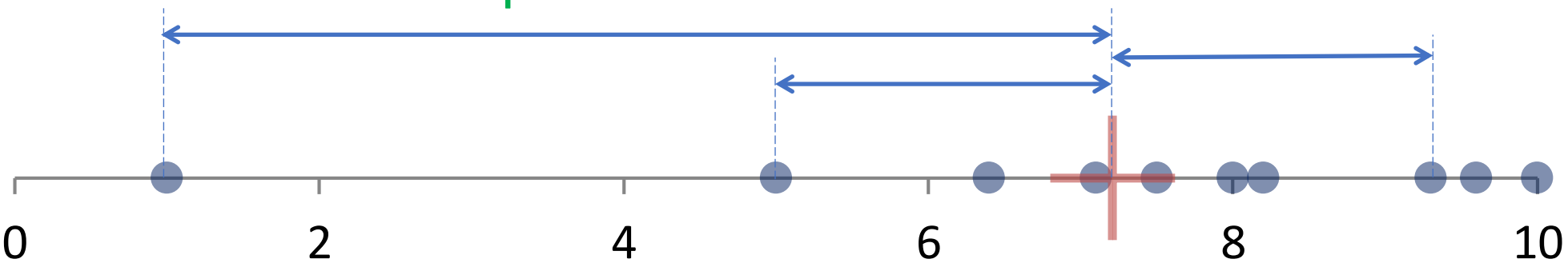


Variance

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

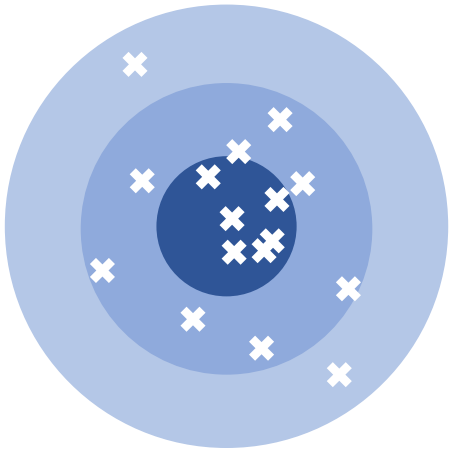
$$\mathrm{var}(x_i) = \sigma_i^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\mu_i - x_{i,j}\right)^2$$

σ_i is the standard deviation

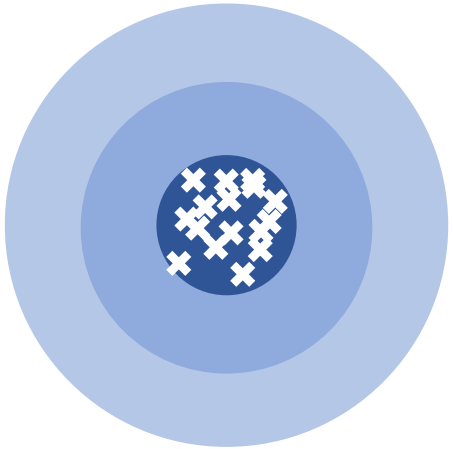


The average squared distance of data points from the mean

Variance and standard deviation



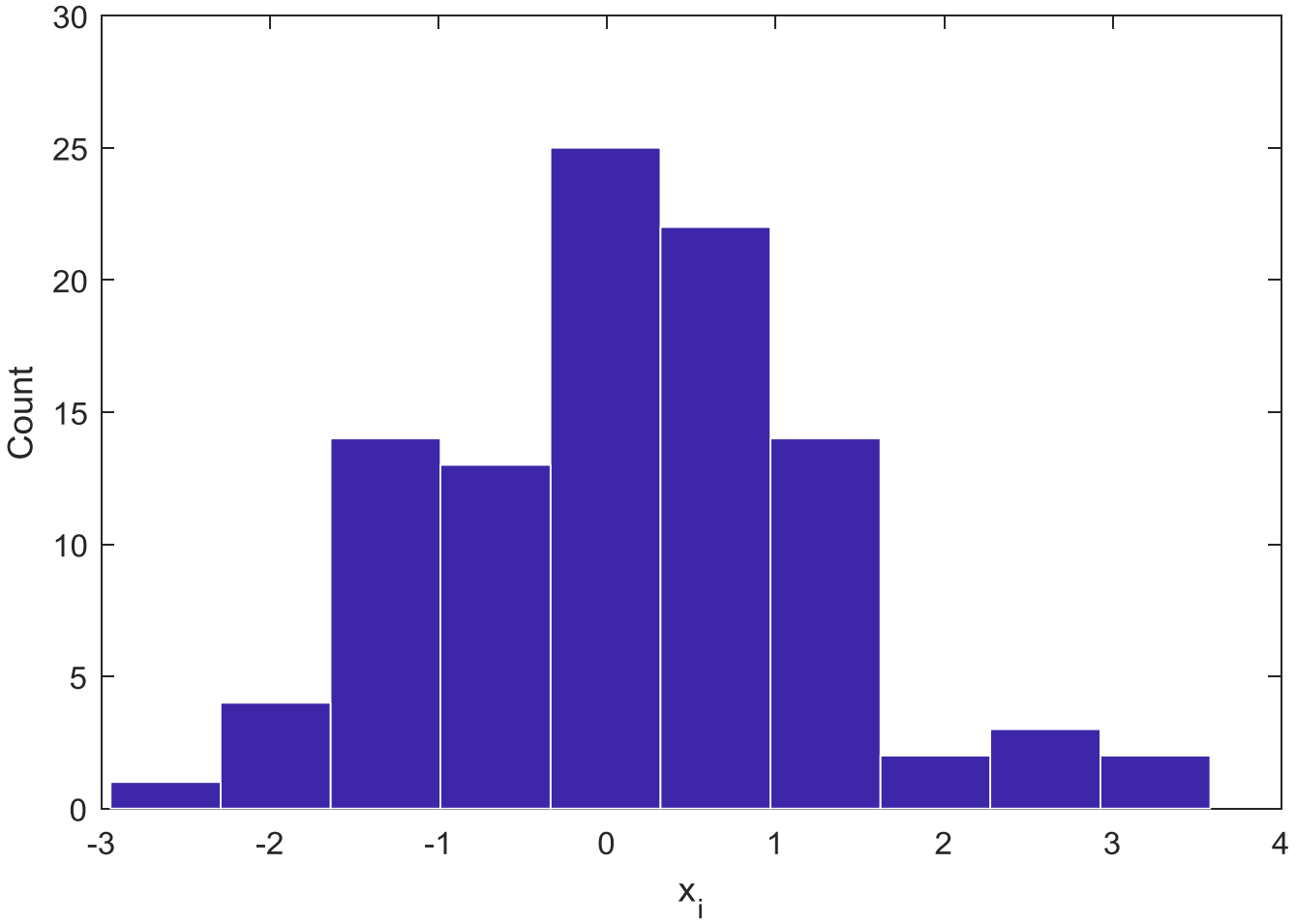
High variance



Low variance

Standard deviation is the square root of the variance.

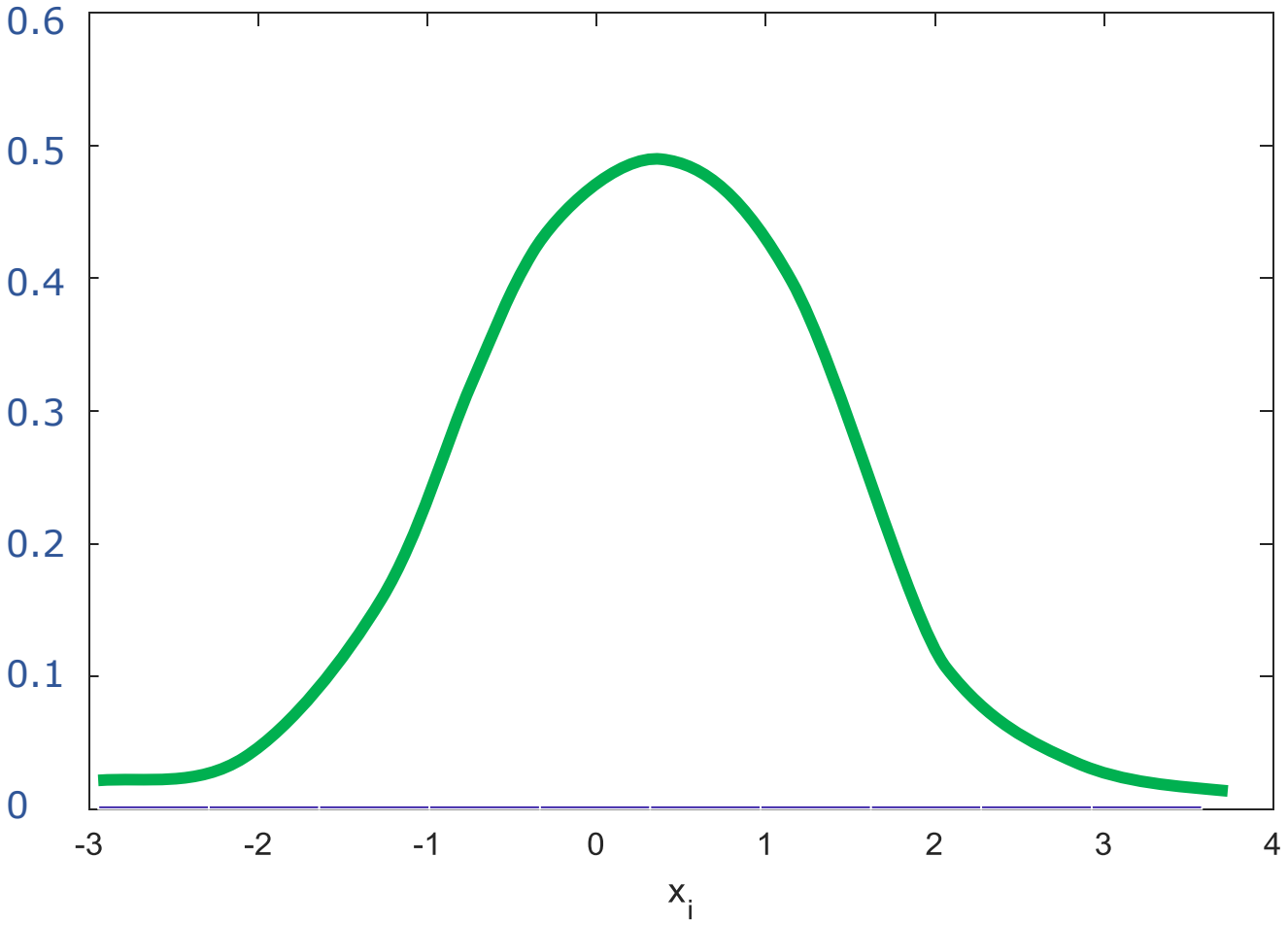
Probability distributions



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

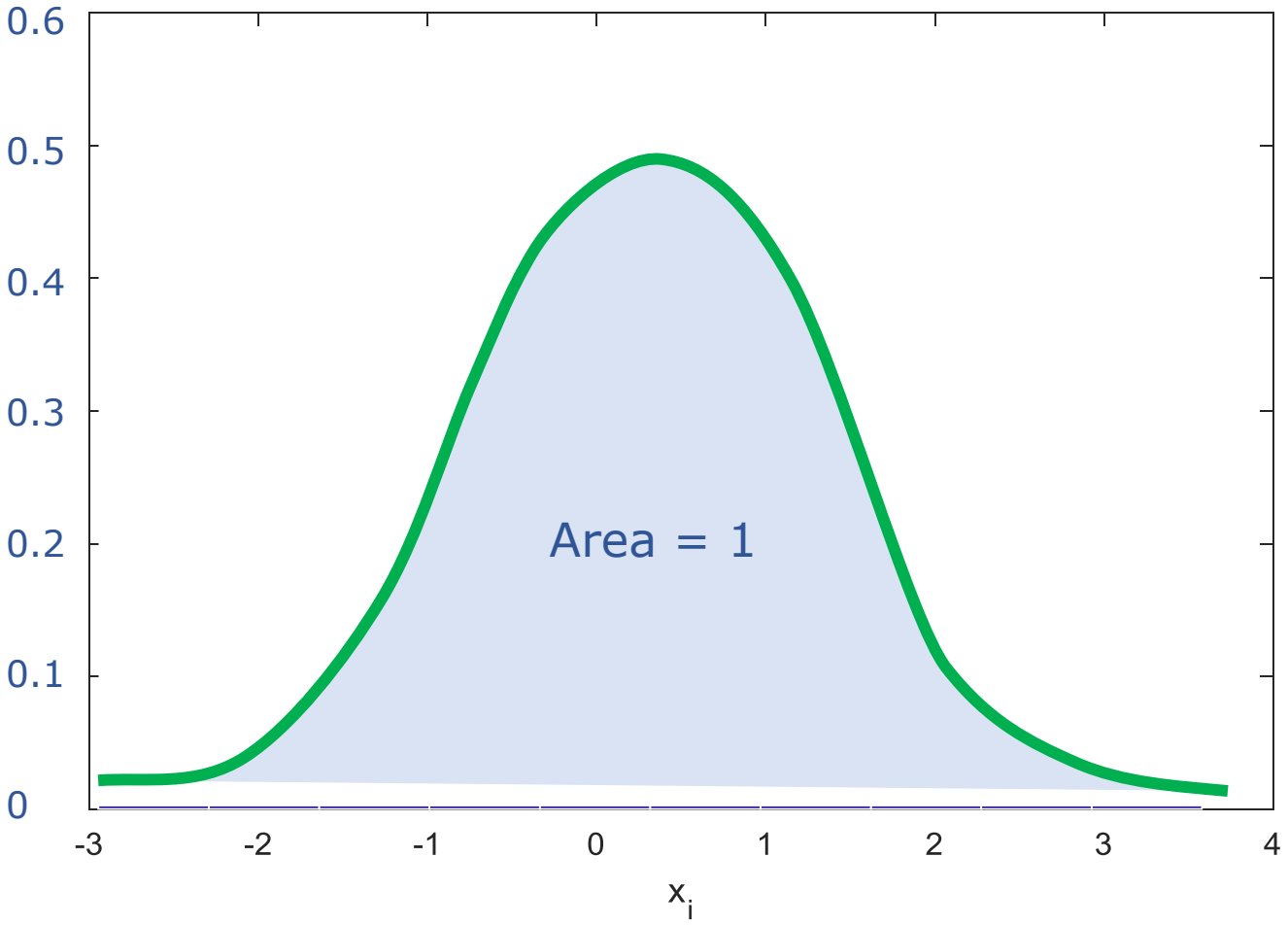
Each bar gives an indication of how frequently (how probable) it is that a value will fall inside that range.

Probability distributions



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

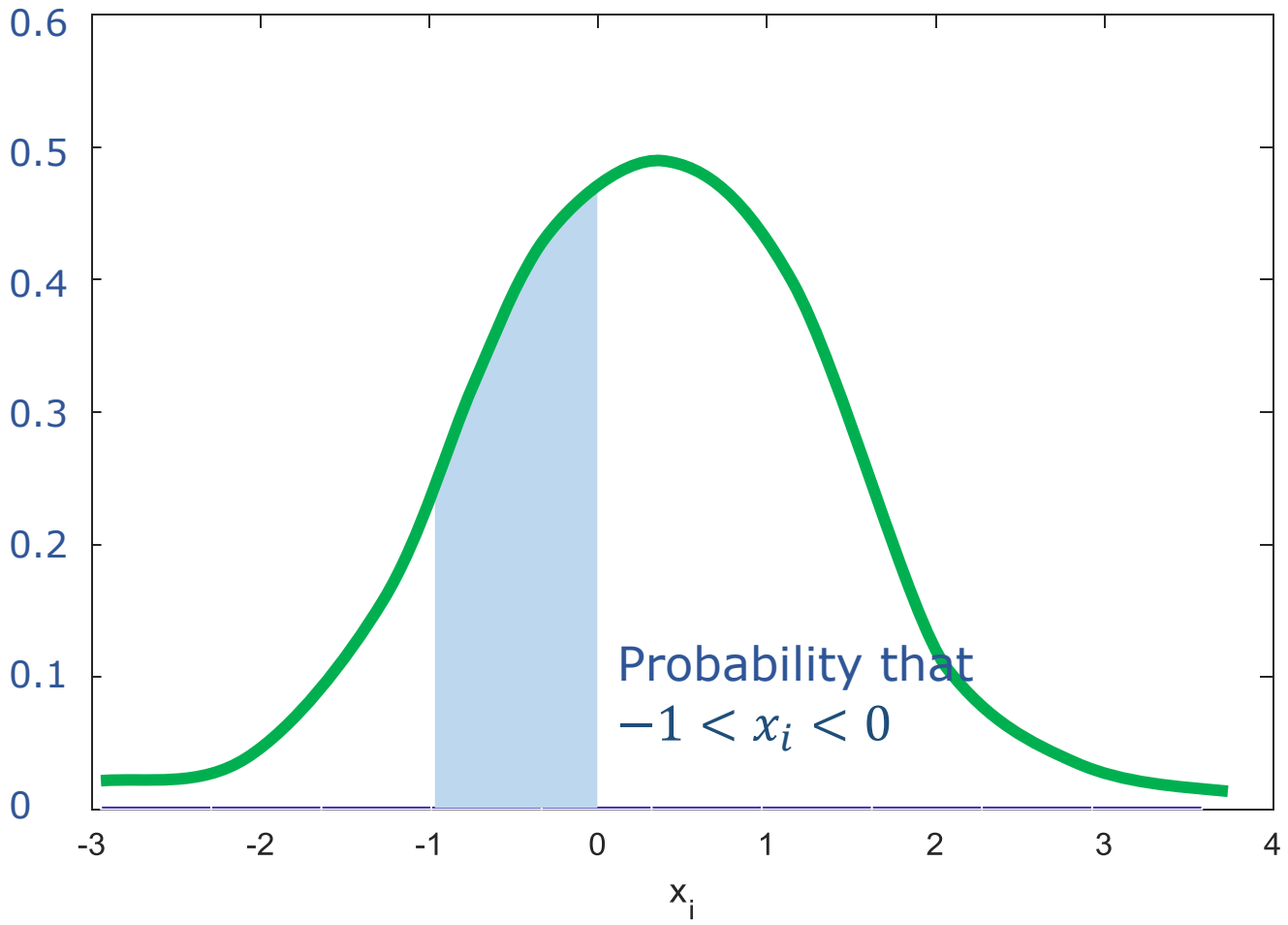
Probability distributions



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

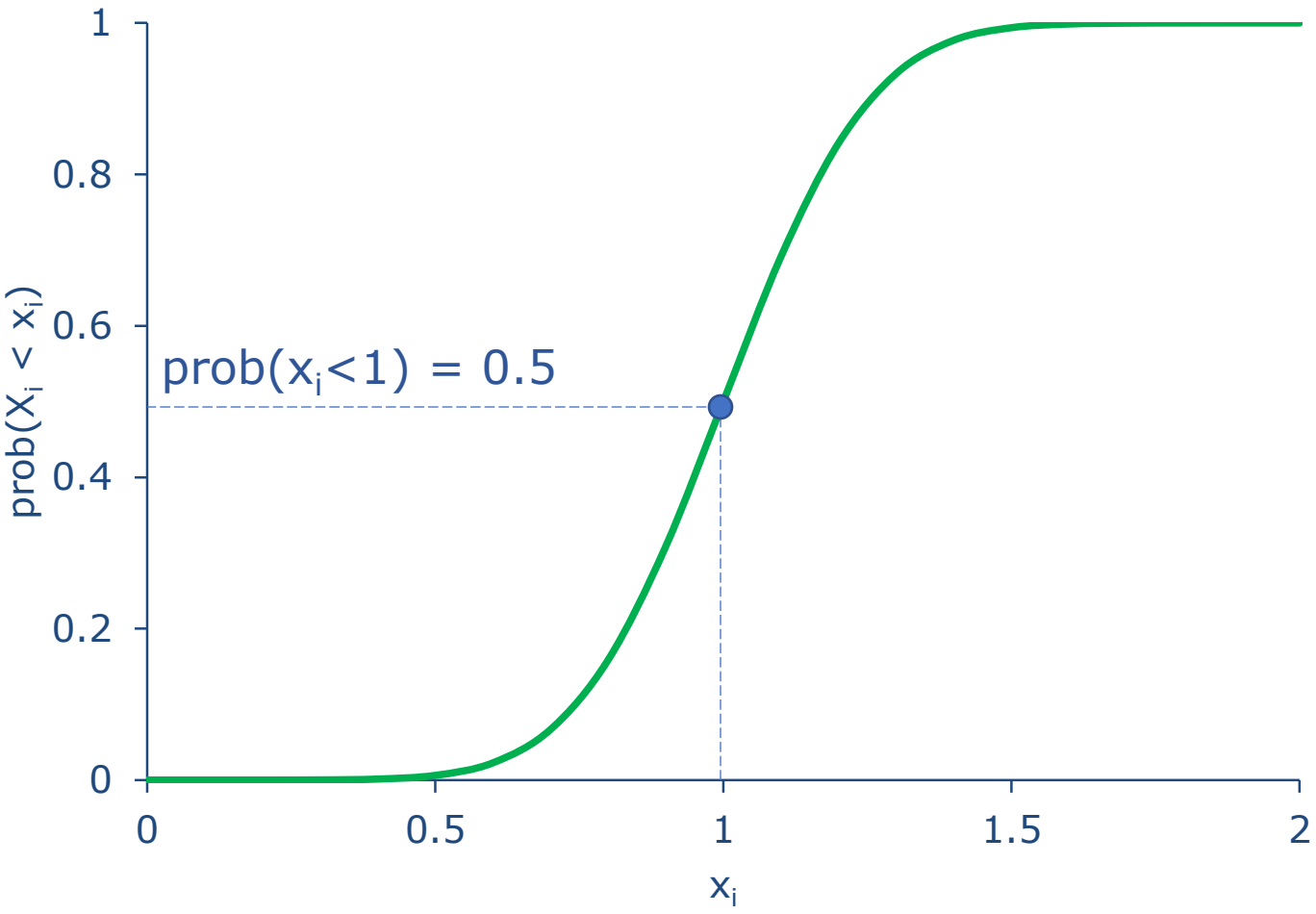


A probability density function (pdf) shows the relative probability of different values occurring.

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

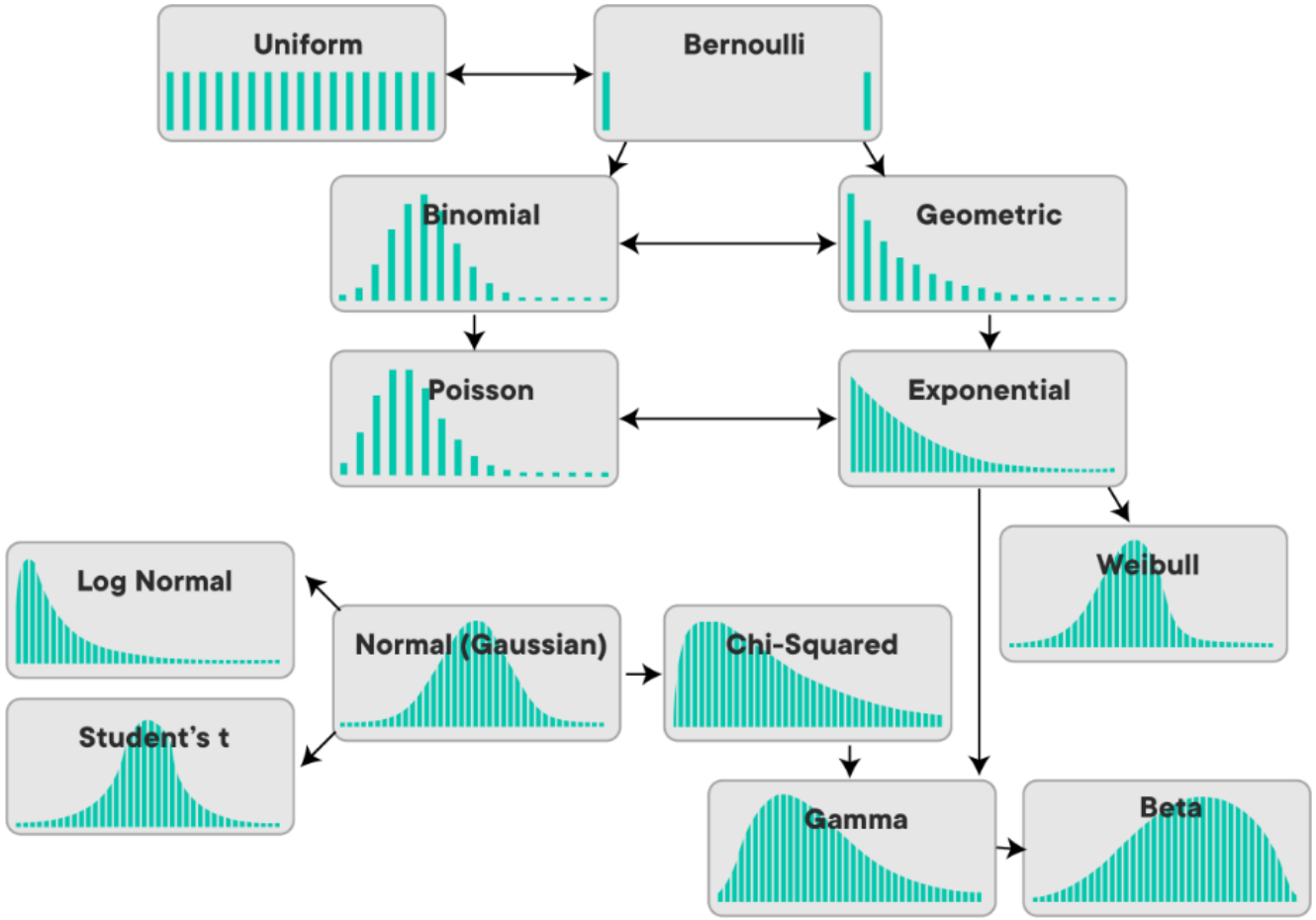
The area between two points shows the probability of a value falling in that range.

Probability distributions



A cumulative density function (cdf) shows the probability that the variable is less than a given value.

Probability distributions



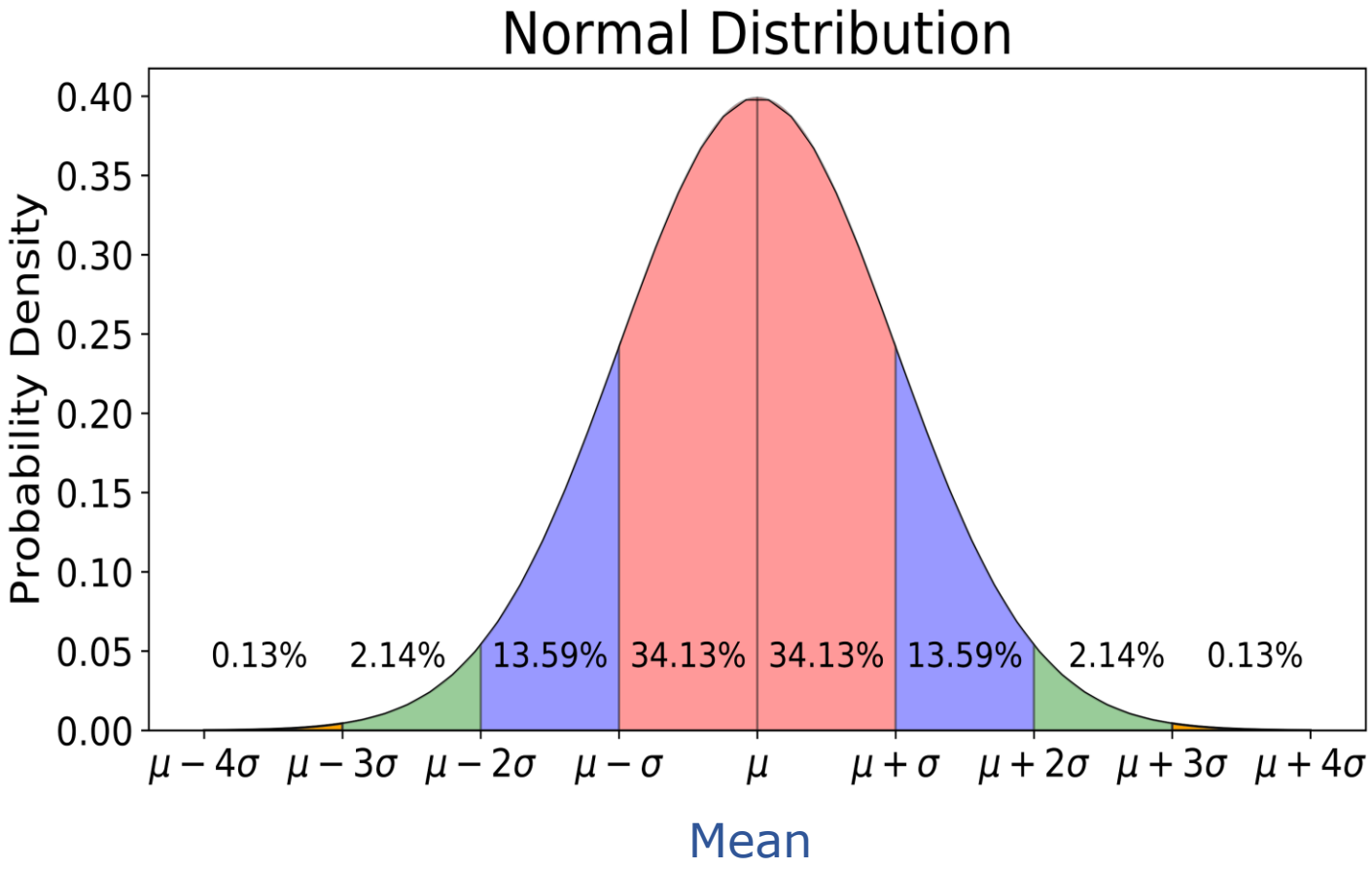
There are many theoretical types of probability distribution.

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).

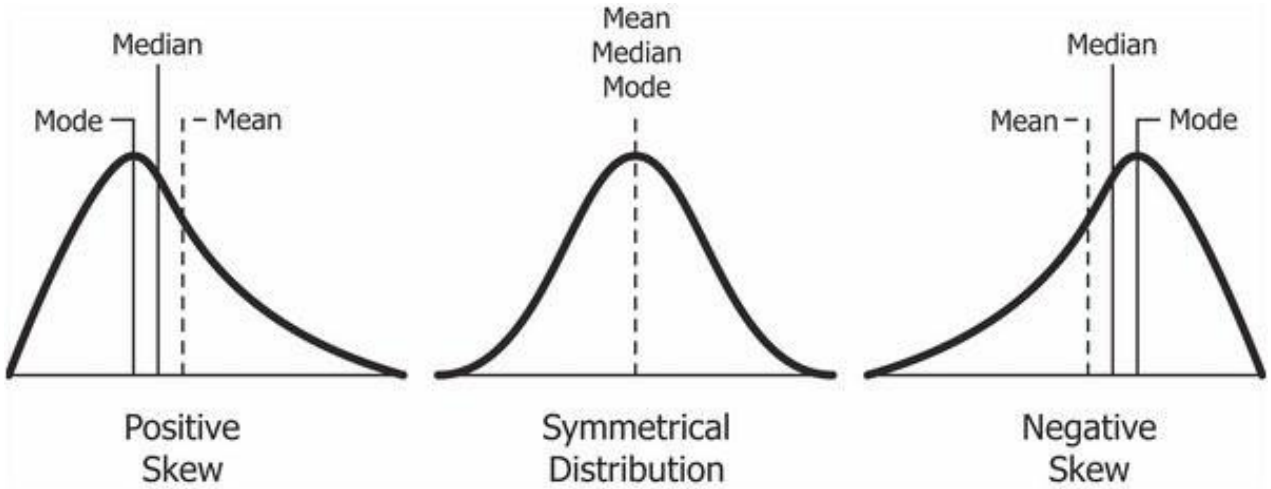
Normal distribution



The Normal (Gaussian) distribution is the best distribution!

- Very common due to central limit theorem: sum of lots of random variables converges to normal
- Makes things easy:
  - Linear
  - Analytical
  - No skew/kurtosis
  - etc.

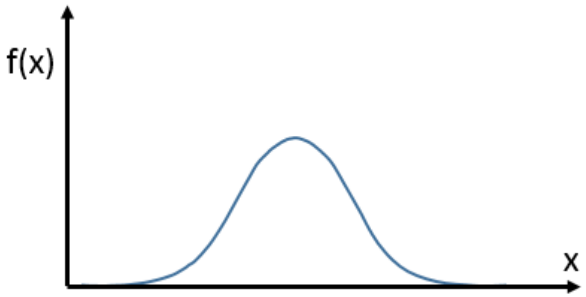
Skew and kurtosis



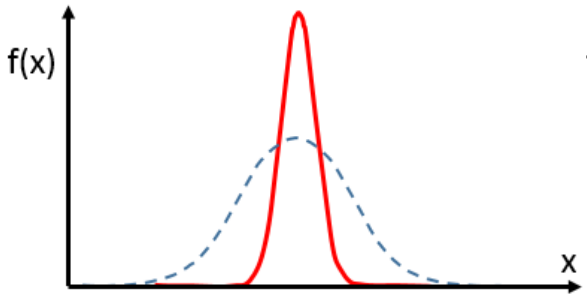
Skew shows the extent to which values are clustered towards one end of the scale. Measure of lack of symmetry.

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

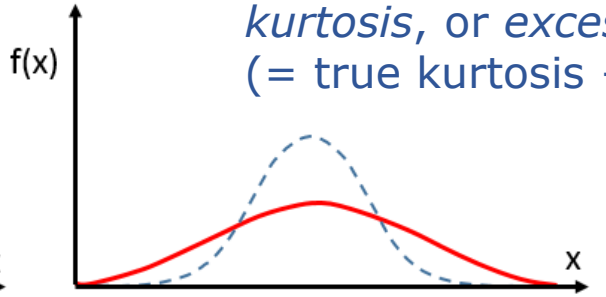
Be careful: "kurtosis" may mean *true kurtosis*, or *excess kurtosis* (= true kurtosis – 3)



Normal
Kurtosis = 3
Excess kurtosis = 0



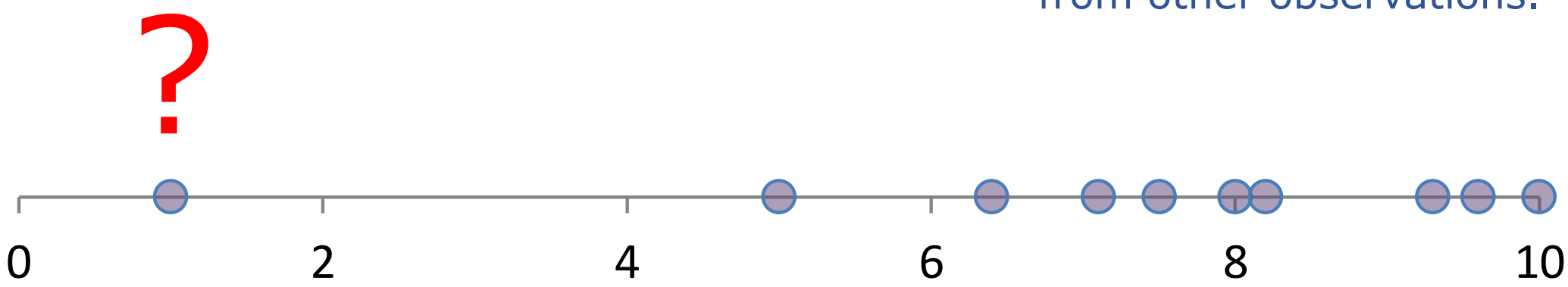
Leptokurtic
Kurtosis > 3
Excess kurtosis > 0



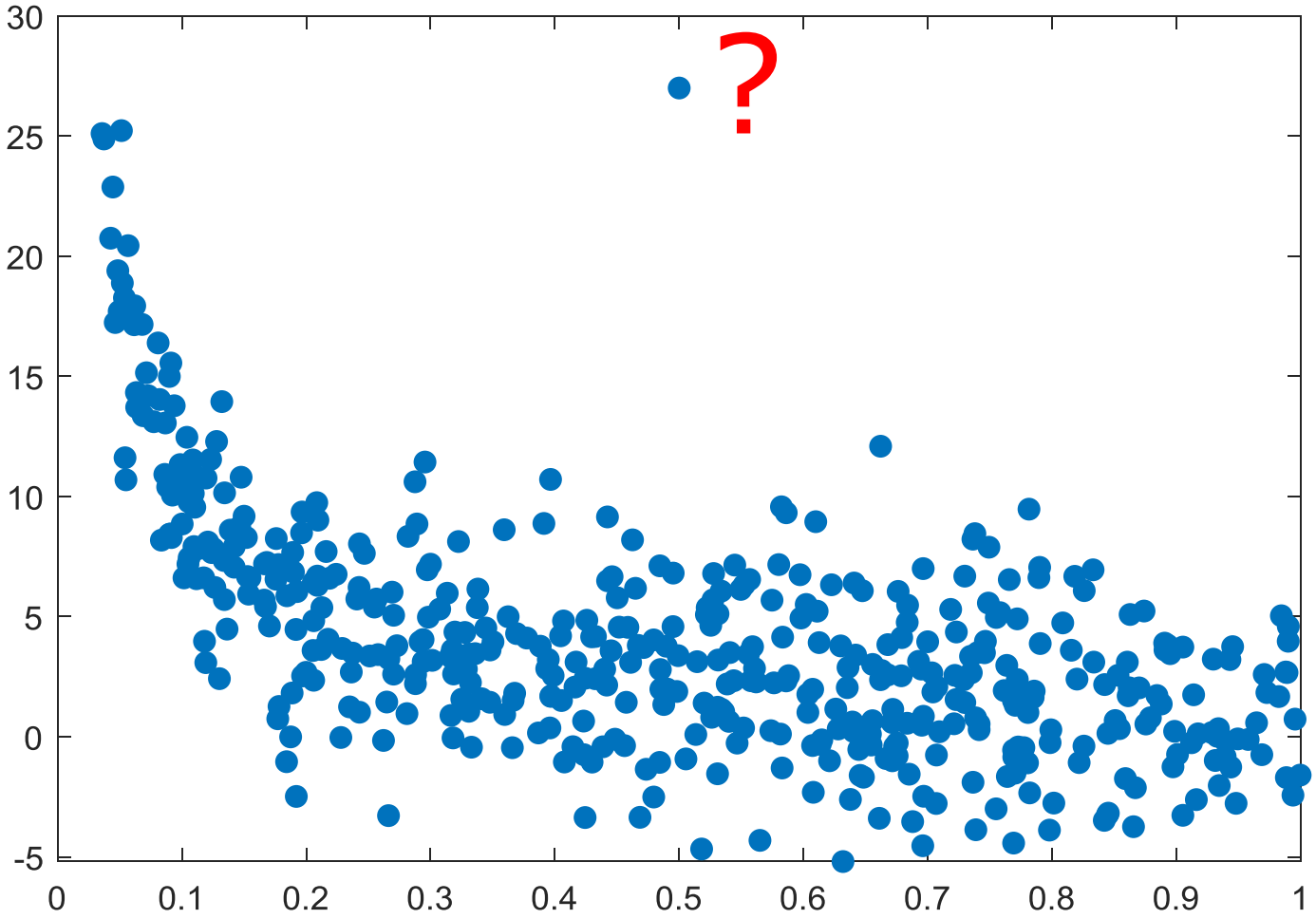
Platykurtic
Kurtosis < 3
Excess kurtosis < 0

Outliers

An outlier is an observation that differs significantly from other observations.



Outliers



An outlier is an observation that differs significantly from other observations.

Outlier identification can be performed via:

- Inter-quartile range
- Skew/kurtosis measures
- Multivariate analysis
- Etc.

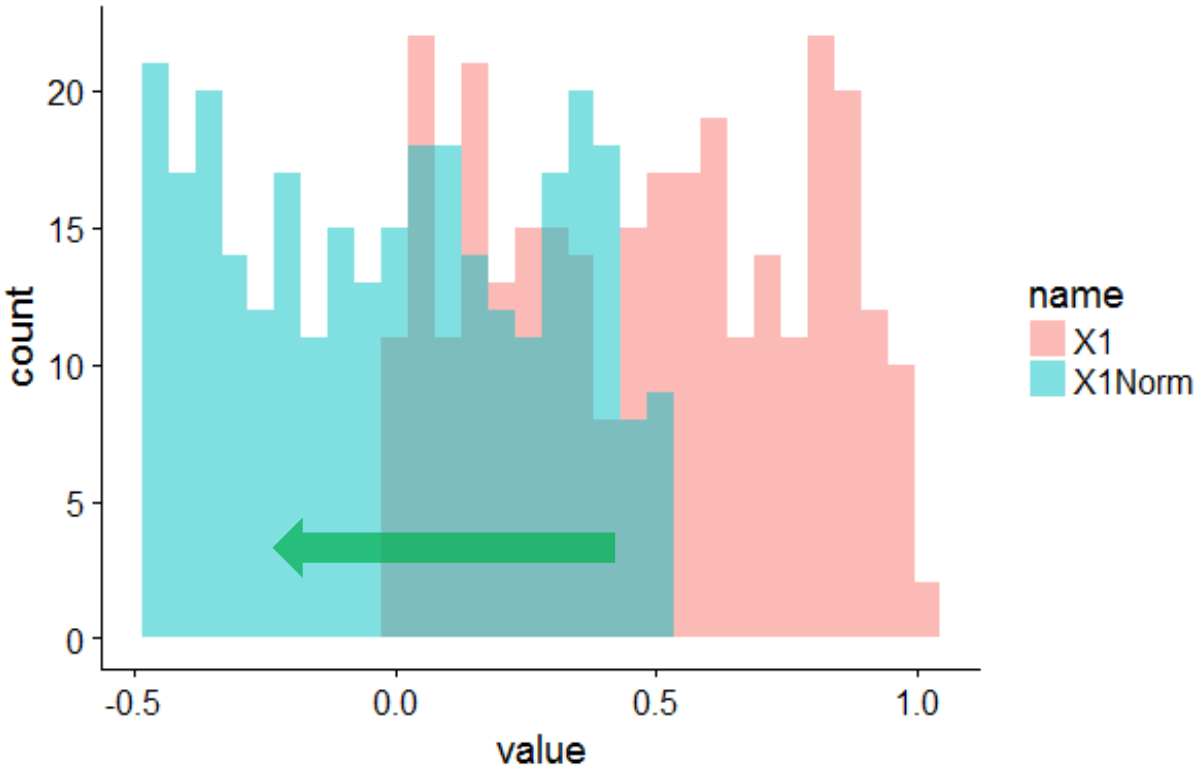
Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

$(x_2 - \mu_2)$ is shifting x_2 so that it has a mean of zero



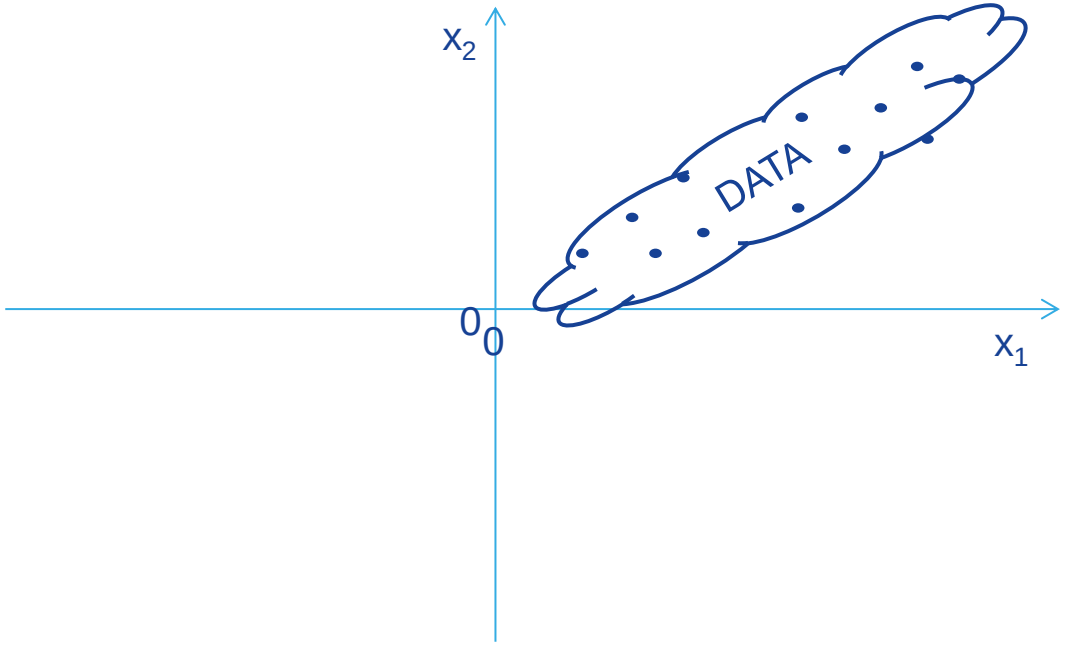
Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

$(x_2 - \mu_2)$ is shifting x_2 so that it has a mean of zero



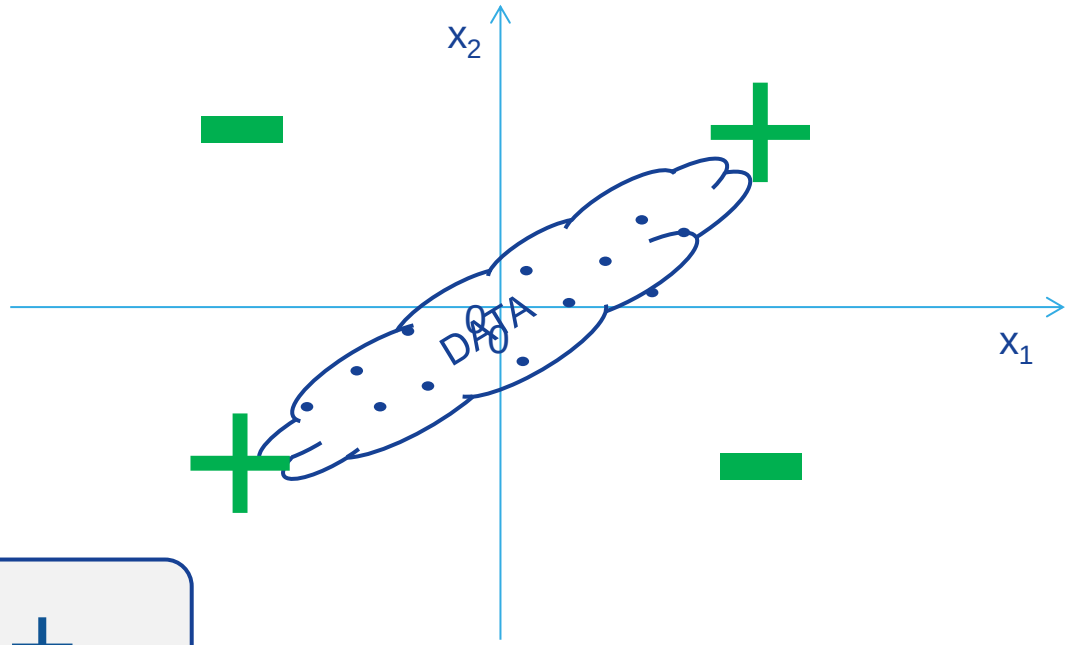
Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

$(x_2 - \mu_2)$ is shifting x_2 so that it has a mean of zero



2+

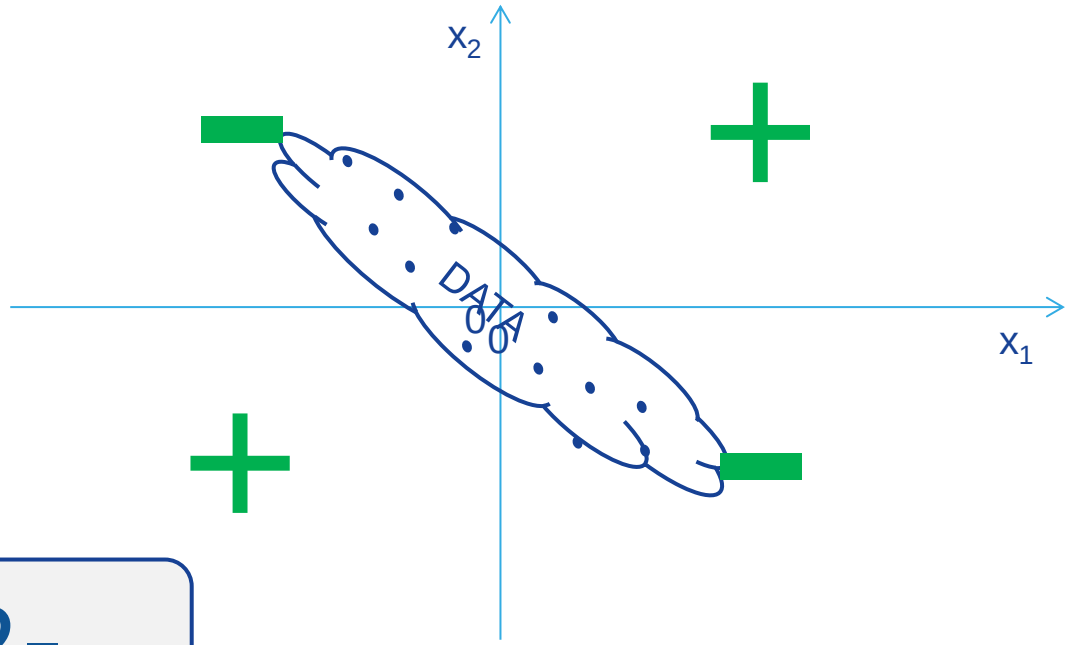
Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

$(x_2 - \mu_2)$ is shifting x_2 so that it has a mean of zero



2-

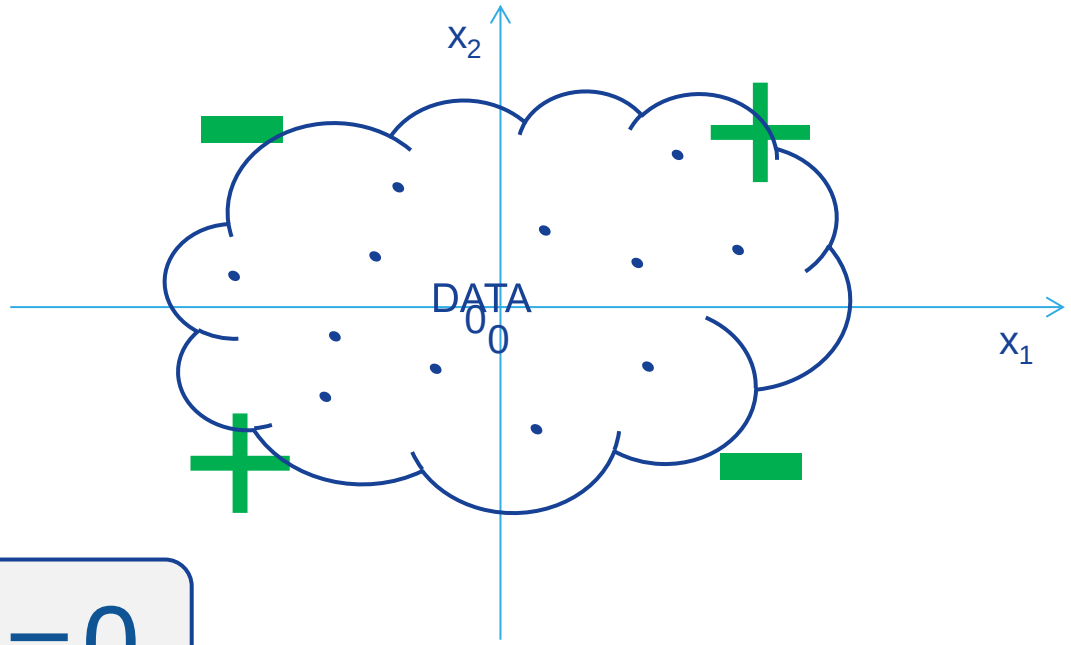
Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)]$$

$(x_2 - \mu_2)$ is shifting x_2 so that it has a mean of zero



(2+)+(2-)=0

Measures of dependence

Fundamentally, we are interested in how the random variable x_1 depends on the random variable x_2.

Covariance

$$\mathrm{cov}(x_1, x_2) = E[(x_1 - \mu_1)(x_2 - \mu_2)$$

Correlation

$$R(x_1, x_2) = \mathrm{corr}(x_1, x_2) = \frac{\mathrm{cov}(x_1, x_2)}{\sigma_1 \sigma_2}$$

Standardises covariance so that $R \in [-1,1]$: 1 or -1 is perfect correlation, 0 is no correlation. Allows comparability.

Measures of dependence

Correlation

$$R(x_1, x_2) = \mathrm{corr}(x_1, x_2) = \frac{\mathrm{cov}(x_1, x_2)}{\sigma_1 \sigma_2}$$

Coefficient of determination: R^2

$$R^2(x_1, x_2) = [\mathrm{corr}(x_1, x_2)]^2 = \left[\frac{\mathrm{cov}(x_1, x_2)}{\sigma_1 \sigma_2}\right]^2$$

More generally, R_i^2 can be defined as:

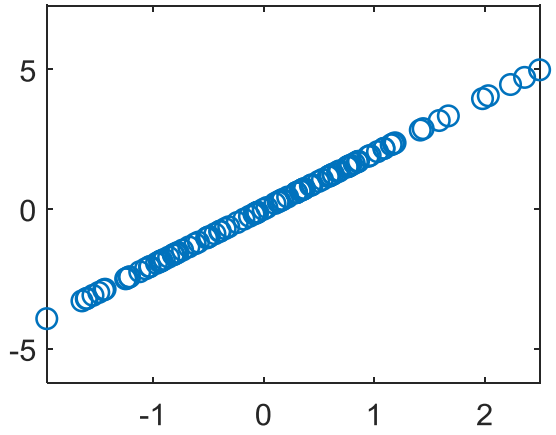
$$\frac{\text{variance explained by regression}}{\text{total variance}}$$

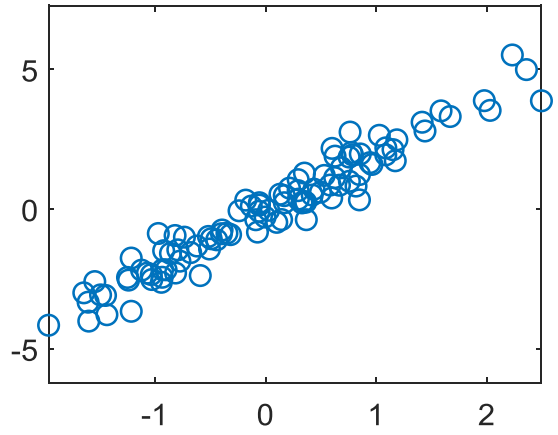
R^2 is a measure of linear dependence.

$R^2 \in [0,1]$: higher values indicate stronger dependence.

Measures of dependence

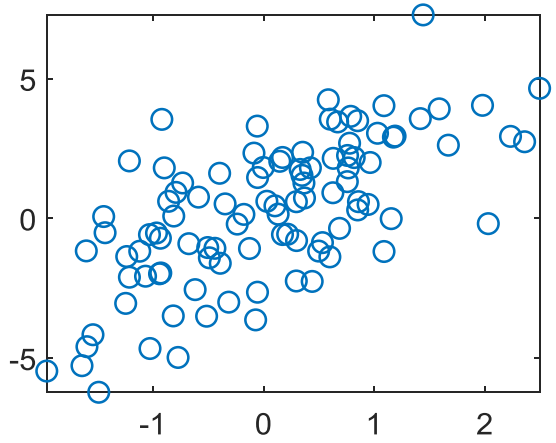
Perfect positive correlation
$R = 1$

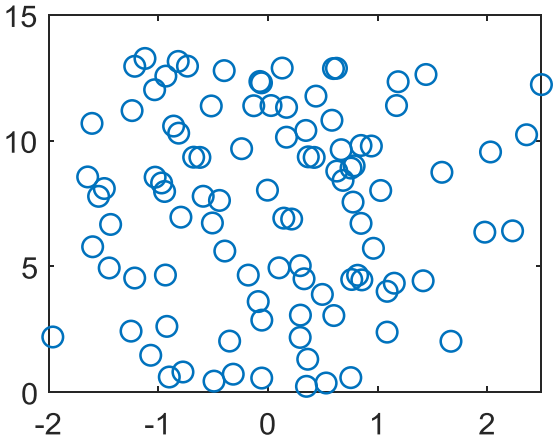




Very strong positive correlation
$R = 0.97$

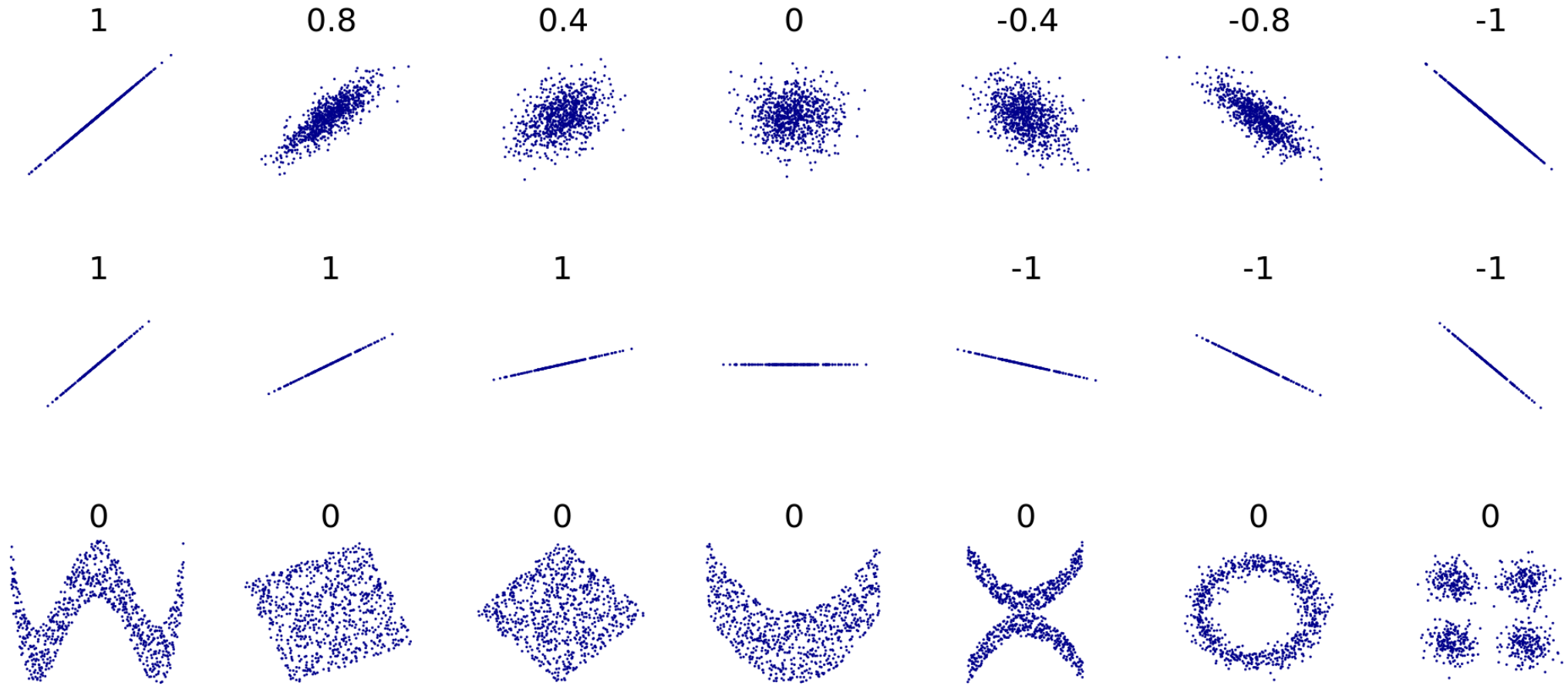
Moderate positive correlation
$R = 0.66$





No positive correlation
$R = 0.02$

Measures of dependence



European Commission

Measures of dependence

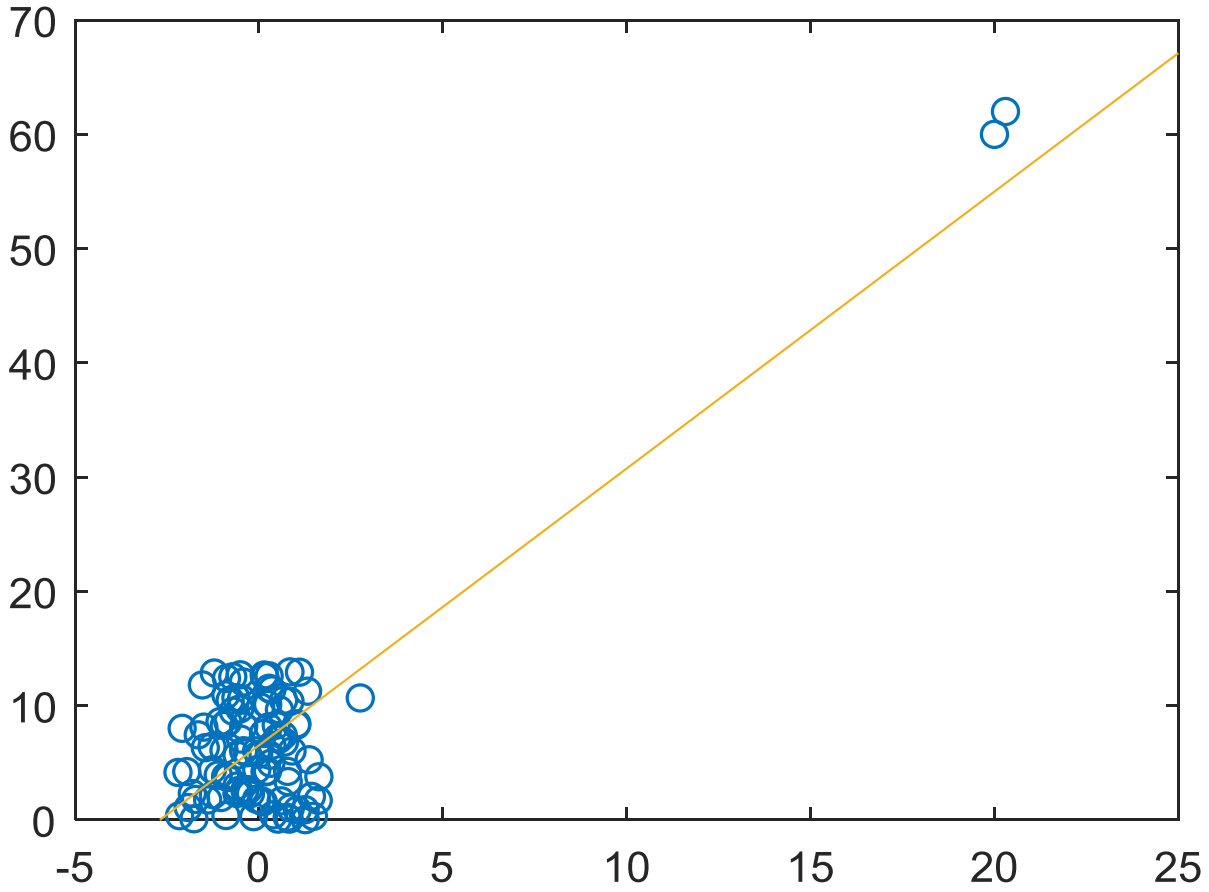
Size of correlation	Interpretation
0.9 to 1 (-0.9 to -1)	Very high positive (negative) correlation
0.7 to 0.9 (-0.7 to -0.9)	High positive (negative) correlation
0.5 to 0.7 (-0.5 to -0.7)	Moderate positive (negative) correlation
0.3 to 0.5 (-0.3 to -0.5)	Low positive (negative) correlation
0 to 0.3 (0 to -0.3)	Negligible correlation

BUT: check for statistical significance!

"Given the sample size (number of points, countries, …), is the correlation significantly different from zero?"

A quick test for significance is $|R| \geq \frac{2}{\sqrt{N}}$

Measures of dependence



What happens to correlation when there are outliers?

correlation:

0.02 ⟹ 0.84

Be careful with outliers.

Always plot your data!

THANK YOU

Welcome to email us at: jrc-coin@ec.europa.eu

COIN in the EU Science Hub
https://ec.europa.eu/jrc/en/coin

COIN tools are available at:
https://composite-indicators.jrc.ec.europa.eu/

European Commission