Statistics Refresher II

19th JRC Annual training on Composite Indicators and Scoreboard

Jaime Lagüera González

European Commission

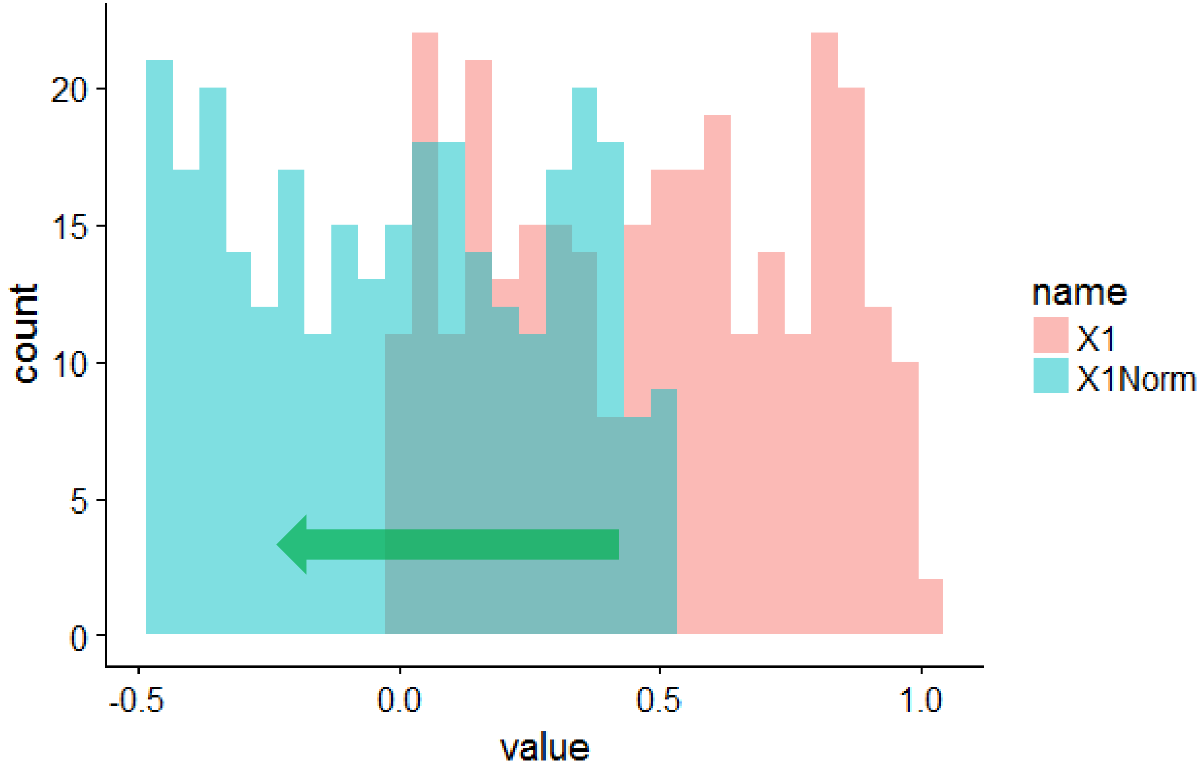
Measures of dependence

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

Covariance

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

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



European Commission

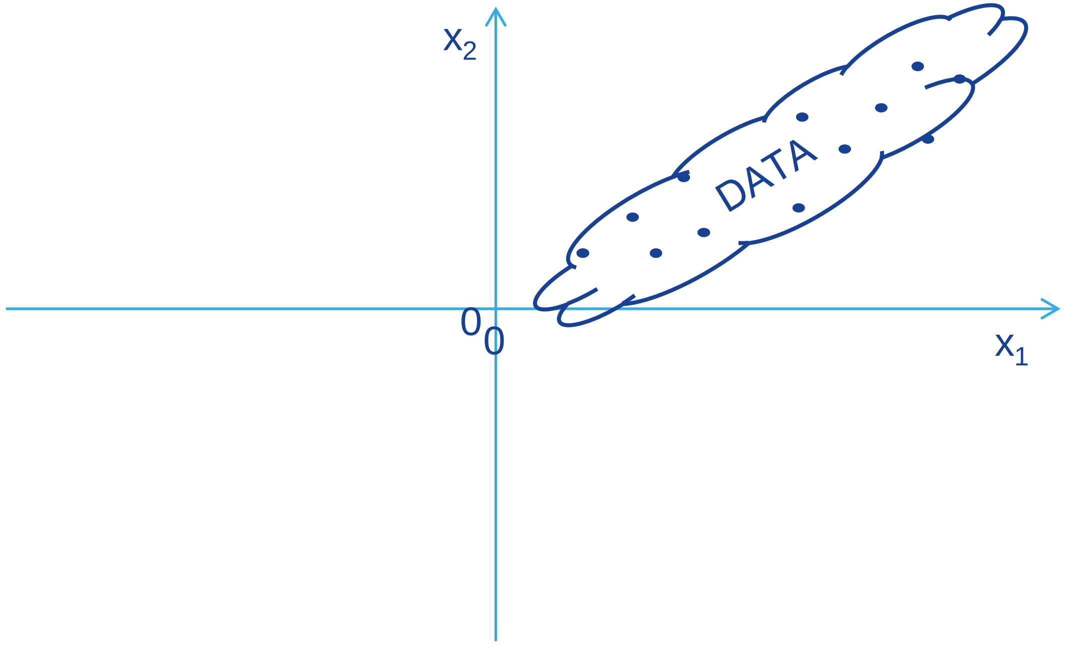
Measures of dependence

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

Covariance

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

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



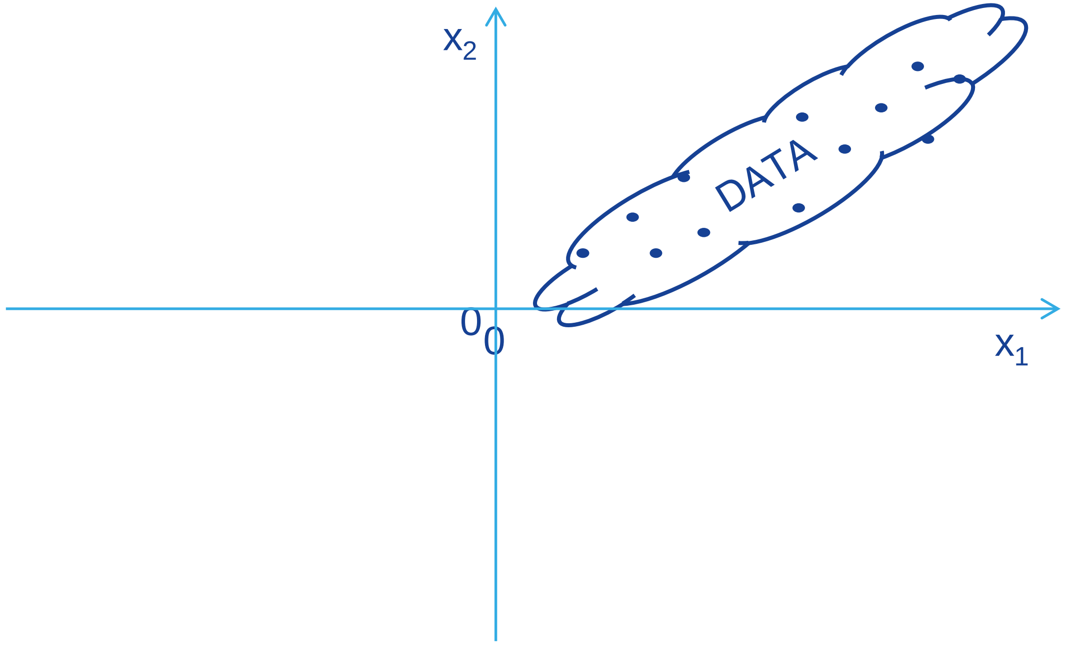
Measures of dependence

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

Covariance

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

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



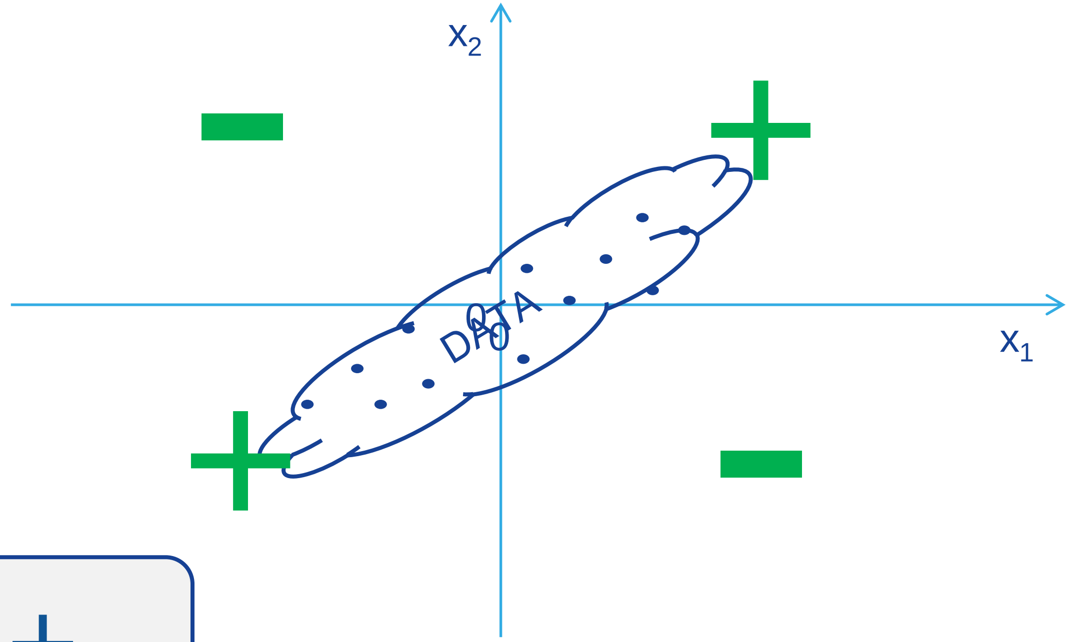
Measures of dependence

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+

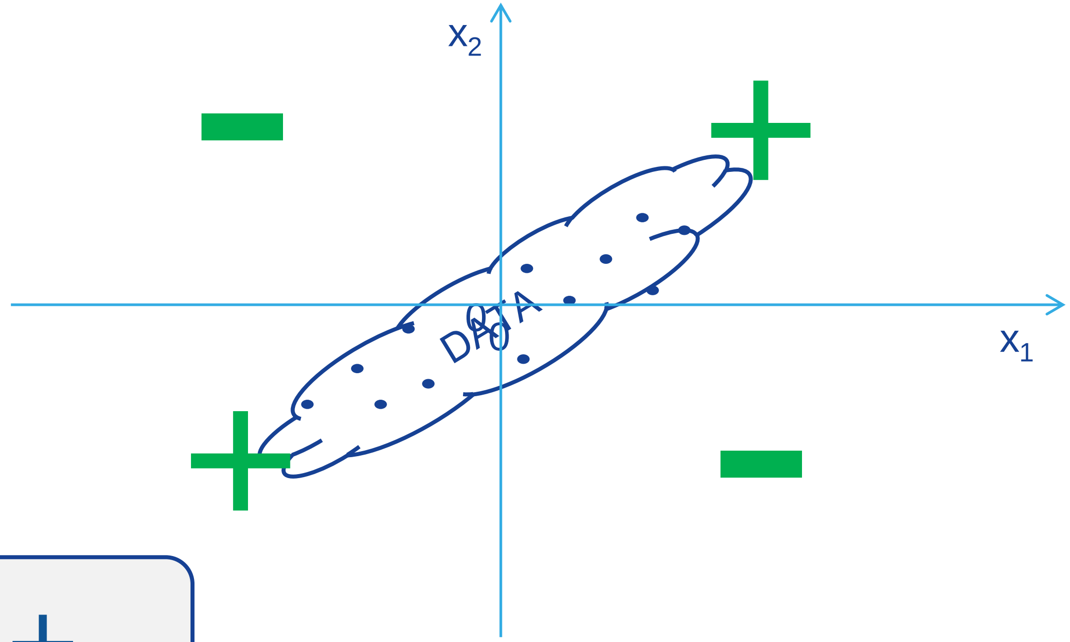
Measures of dependence

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



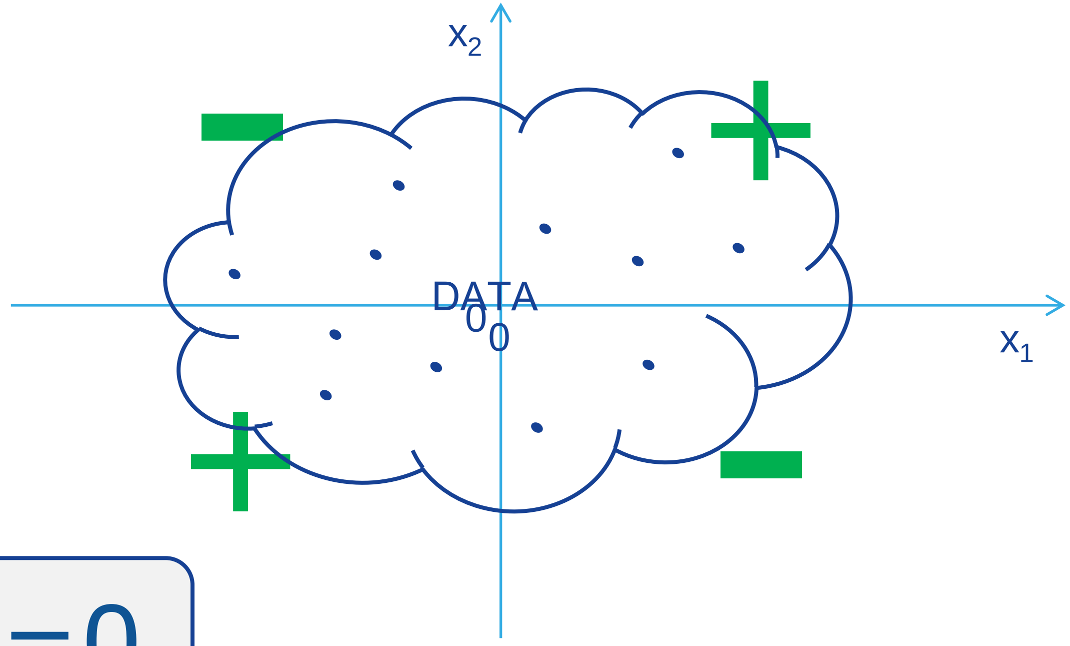
2+

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

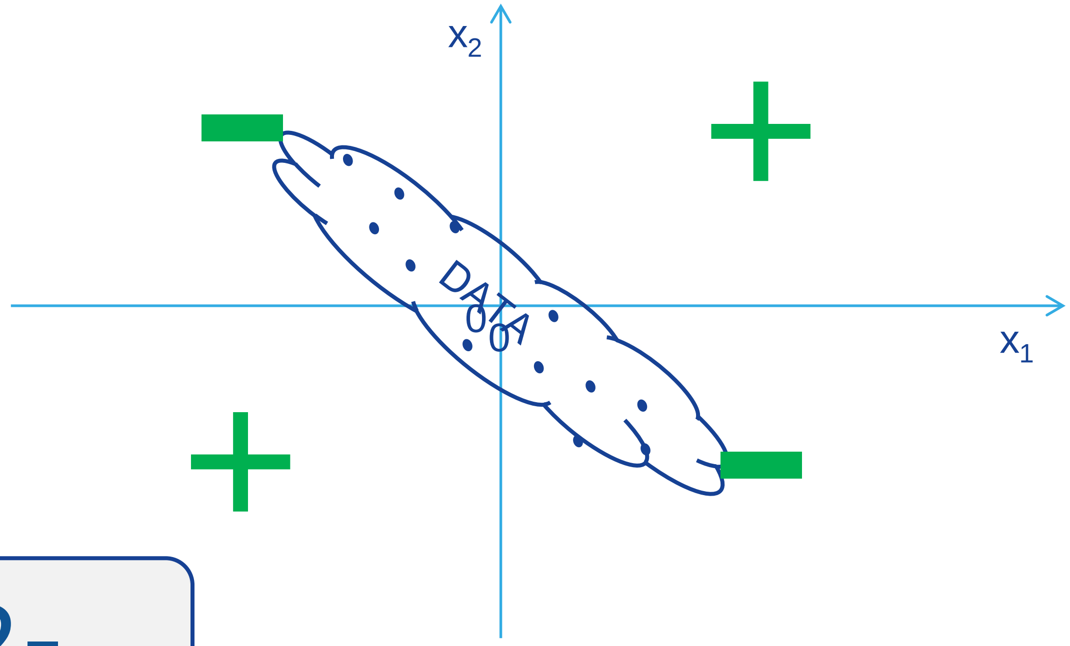
Measures of dependence

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

Covariance

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

$(x_2 - \mu_2)$ 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

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

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:

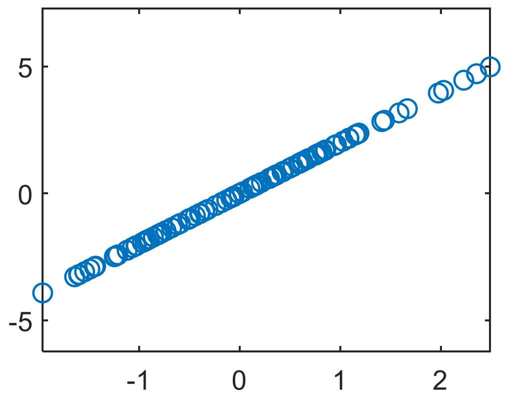
$$\frac{\text{variance explained by regression}}{\text{total variance}}$$

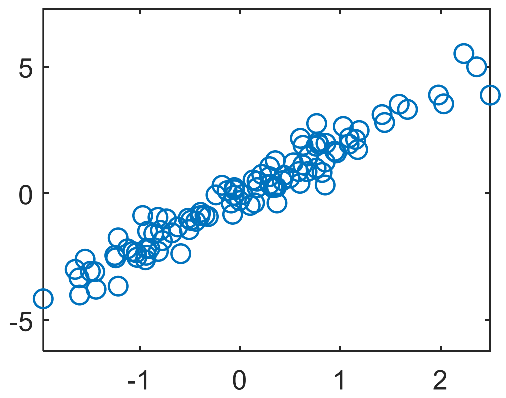
R^2 is a measure of linear dependence.

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

Measures of dependence

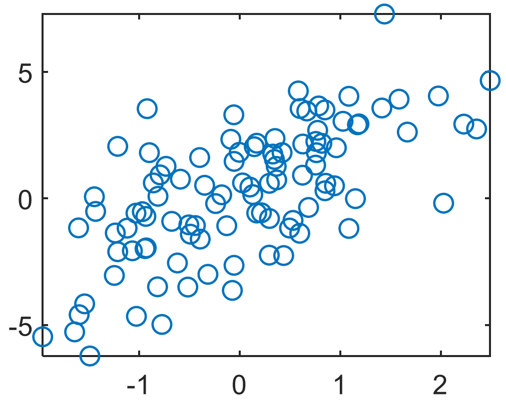
Perfect positive correlation
$R = 1$

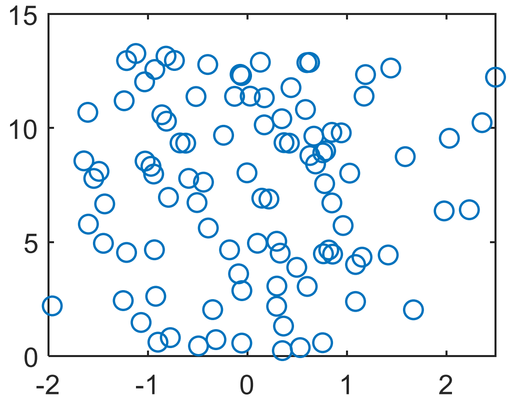




Very strong positive correlation
$R = 0.97$

Moderate positive correlation
$R = 0.66$





No positive correlation
$R = 0.02$

Measures of dependence

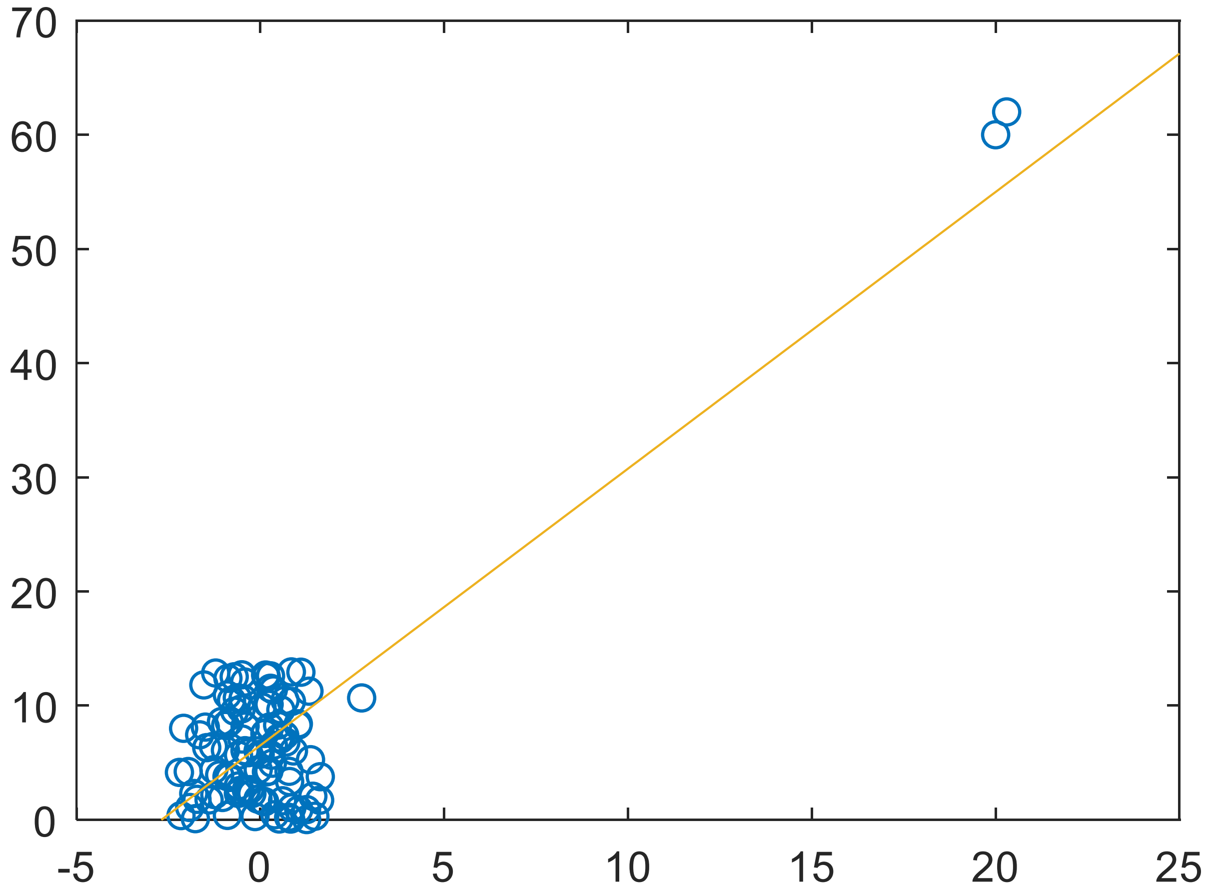
Size of correlation	Interpretation
0.9 to 1 (-0.9 to -1)	Very high positive (negative) correlation
0.7 to 0.9 (-0.7 to -0.9)	High positive (negative) correlation
0.5 to 0.7 (-0.5 to -0.7)	Moderate positive (negative) correlation
0.3 to 0.5 (-0.3 to -0.5)	Low positive (negative) correlation
0 to 0.3 (0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

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?

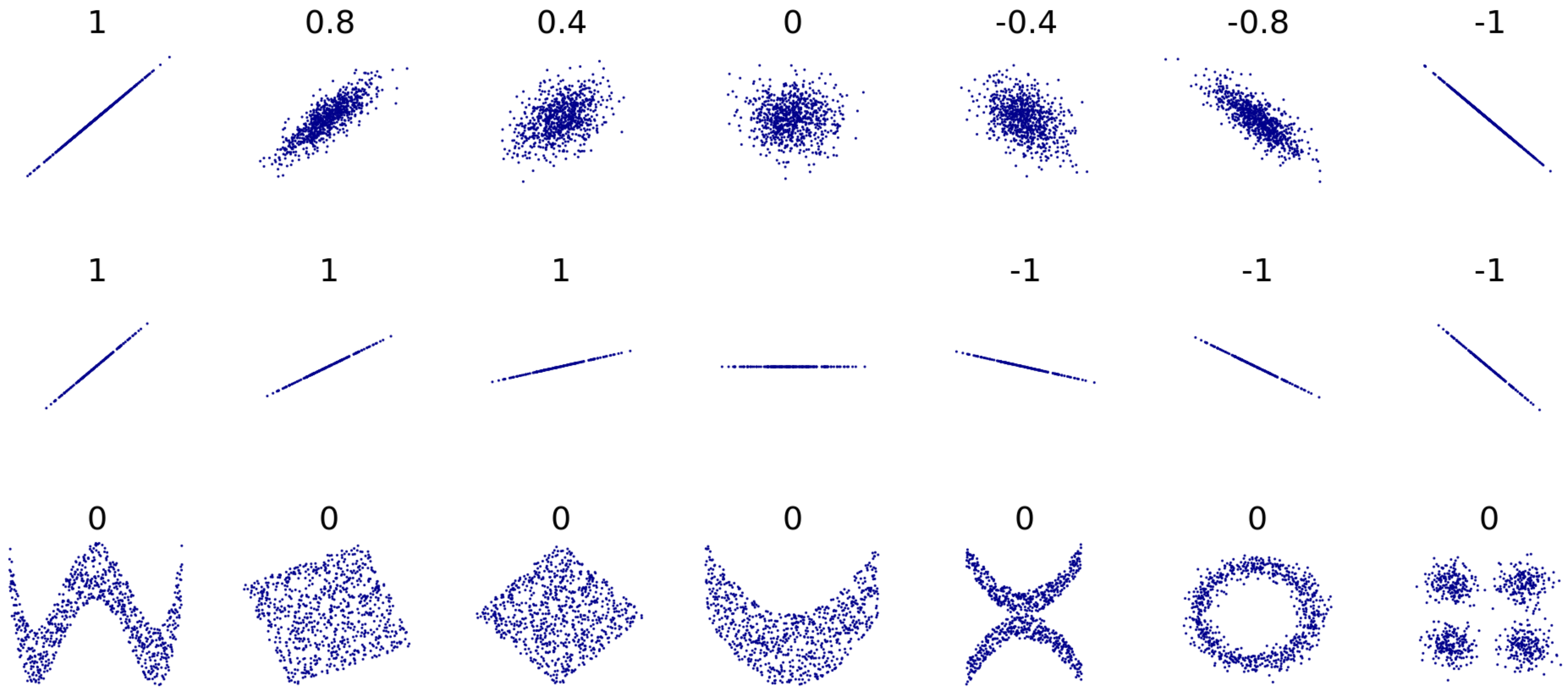
correlation:

0.02 → 0.84

Be careful with outliers.

Always plot your data!

Measures of dependence



Finally

Correlation is *linear* dependence. Extensions or variations of correlation include:

- Spearman rank correlation (correlation coefficient between ranks)
- Kendall rank correlation (slightly different than Spearman)
- Nonlinear dependence measures: *correlation ratio* (first order sensitivity index) and friends

But

Linear correlation is usually a pretty good approximation. Using both is a safe bet.

Thank you

jaime.laguera-gonzalez@ec.europa.eu | jrc-coin@ec.europa.eu

composite-indicators.jrc.ec.europa.eu

European Commission