Introduzione alla data science e al pensiero computazionale
Lezione 10: Correlazione

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Content

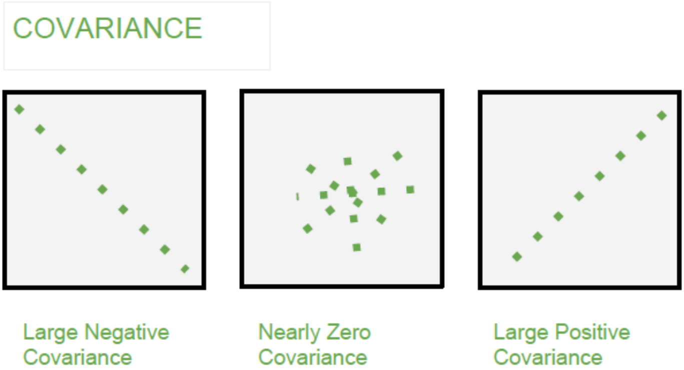
- Covariance
- Correlation (aka Pearson product-moment correlation coefficient)
- Relationship between Pearson correlation and linear regression
- Non parametric correlations

Covariance

- To proceed further with our analysis we will use the concept of **covariance**, which we have already seen
- It expresses the degree in which the variation of a random variable is connected to the variation of another random variable
- It is defined as follows:
  - Given two random variables X and Y

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Covariance – graphically



Source :
https://www.geeksforgeeks.org/mathematics-covariance-and-correlation

About the covariance - 1

- We notice that:
  - The covariance of a random variable with itself is the variance:

$$Cov(X,X) = Var(X)$$

  - There is a similar property as for the variance

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

since:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] =$$

$$= E(XY) - E(XE(Y)) - E(E(X)Y) + E(E(X)E(Y)) =$$

$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) =$$

$$= E(XY) - E(X)E(Y)$$

QED.

About the covariance - 2

- The covariance is symmetric:

$$Cov(X, Y) = Cov(Y, X)$$

- The covariance is linear with respect to multiplications by constants:

$$(\forall a, b \in \mathbb{R}) \;\; Cov(aX, bY) = abCov(X, Y)$$

- If $e \sim N(0, \sigma)$, $Cov(X, e) = 0$

$$Cov(X, e) = E(Xe) - E(X)E(e)$$

Since X and e are independent

$$E(Xe) = E(X)E(e)$$

Moreover, $e \sim N(0, \sigma)$

$$E(e) = 0$$

QED

Pearson Correlation Coefficient

- AKA Pearson product-moment correlation coefficient or just correlation coefficient
- It expresses the linear correlation between two random variables
- It is defined as follows:
  - Given two random variables X and Y

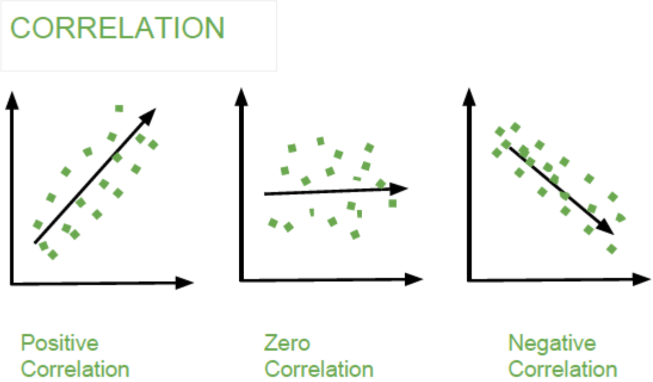
$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

  - Where:

$$\sigma_Z = \sqrt{Var(Z)}$$

For the time being we intentionally ignore the difference between sample and population.

Pearson Correlation Coefficient – graphically



Source :
https://www.geeksforgeeks.org/mathematics-covariance-and-correlation

About the Pearson Correlation Coefficient

- The Pearson correlation coefficient is also often expressed as:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- It is symmetric: $r_{X,Y} = r_{Y,X}$
- It is invariant with respect to multiplications by, and additions of constants:
  $(\forall a, b, c, d \in \mathbb{R}, b \neq 0, d \neq 0) \;\; r_{X,Y} = r_{(a+bX),(c+dY)}$
- It ranges from -1 to 1: $-1 \leq r_{X,Y} \leq 1$ $r_{X,Y} = 1$ means perfect linear relationship (all points lie on a monotonically increasing line)
  $r_{X,Y} = -1$ means perfect opposite linear relationship (all points lie on a monotonically decreasing line)
  $r_{X,Y} = 0$ means no linear relationship between X and Y

Back to Linear Regression (1/2)

- We now focus our attention to the case of the case of the linear regression
- Suppose we have two phenomena that we want to measure, X and Y
- Let us assume
  - that there is a linear relationship between them
  - that I can express the data I collect as:

$$\boldsymbol{y} = \boldsymbol{\theta}_0 + X\boldsymbol{\theta}_1 + \boldsymbol{\epsilon}$$

  - where ϵ is a stationary gaussian process $N(0, \sigma^2)$
- We know the solution that minimizes the square error

Back to Linear Regression (2/2)

- From this solution we have extracted the coefficient of determination

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- Where:
  - $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$
  - SS_{res} is the distance between the reality and the 1-degree best approximation, that is, the OLS model
- and
  - $SS_{tot} = \sum_i (y_i - \overline{y})^2$
  - SS_{tot} is the distance between the reality and the 0-degree best approximation, that is, the mean
- I want to know the relationship between R^2 and the correlation coefficient between X and Y, $r_{X,Y}$

Our goal – understanding R and r

- We focus on 1D
- We are now going to prove a fundamental point.
- Under the assumption that the noise is gaussian and centered in 0, in a linear regression:

$$R^2 = r_{X,Y}^2$$

$R^2 = r^2_{X,Y}$ (1/4)

- Since

$$\hat{y} = \theta_0 + \theta_1 x$$

- we have from above (see page 7) that:

$$r_{X,Y} = r_{\hat{Y},Y}$$

- We define now the explained sum of squares (ESS)
  - $ESS = \sum_i (\hat{y}_i - \overline{y})^2$
  - ESS is the additional knowledge we get on the random variable using a polynomial of degree 1 vs. using a polynomial of degree 0
- We will now prove that **under our hypotheses**:

$$ESS + SS_{res} = SS_{tot}$$

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (1/6)

- We start from:

$$(y_i - \overline{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

- which we square:

$$(y_i - \overline{y})^2 = (y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + (\hat{y}_i - \overline{y})^2$$

- and then we sum:

$$\sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \sum_i (\hat{y}_i - \overline{y})^2$$

Source with modifications:
https://en.wikipedia.org/wiki/Explained_sum_of_squares

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (2/6)

- Now we focus on:

$$\sum_i 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

- and we want to prove that it is 0, that is $\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$; considering:

$$y_i = \hat{y}_i + \epsilon_i$$

$$E(y_i) = E(\hat{y}_i + \epsilon_i) = E(\hat{y}_i) + E(\epsilon_i) = E(\hat{y}_i)$$

because ϵ is a stationary gaussian process $N(0, \sigma^2)$

Source with modifications:
https://en.wikipedia.org/wiki/Explained_sum_of_squares

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (3/6)

- We can build a system:

$$*\hat{y}_i = \theta_0 + \theta_1 x_i, \overline{y} = \theta_0 + \theta_1 \overline{x}$$

- from which we deduce by subtraction:

$$\hat{y}_i - \overline{y} = \theta_1 (x_i - \overline{x})$$

- remembering that:

$$\theta_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Source with modifications:
https://en.wikipedia.org/wiki/Explained_sum_of_squares

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (4/6)

- So:

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_i (y_i - \hat{y}_i)(\theta_1(x_i - \overline{x})) =$$

$$= \theta_1 \sum_i (y_i - \hat{y}_i)(x_i - \overline{x})$$

- Now, let's consider that:

$$(y_i - \hat{y}_i) = y_i - \hat{y}_i + \overline{y} - \overline{y} = (y_i - \overline{y}) - (\hat{y}_i - \overline{y}) =$$

$$= (y_i - \overline{y}) - \theta_1(x_i - \overline{x})$$

- Substituting $(y_i - \hat{y}_i)$ above we get:

$$\theta_1 \sum_i (y_i - \hat{y}_i)(x_i - \overline{x}) = \theta_1 \sum_i [(y_i - \overline{y}) - \theta_1(x_i - \overline{x})](x_i - \overline{x})$$

Source with modifications:

https://en.wikipedia.org/wiki/Explained_sum_of_squares

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (5/6)

- We can conclude:

$$\theta_1 \sum_i [(y_i - \overline{y}) - \theta_1(x_i - \overline{x})](x_i - \overline{x}) =$$

$$= \theta_1 [\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \theta_1 (x_i - \overline{x})(x_i - \overline{x})] =$$

$$= \theta_1 [\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \frac{\sum_j (x_j - \overline{x})(y_j - \overline{y})}{\sum_j (x_j - \overline{x})^2} (x_i - \overline{x})^2] =$$

Source with modifications:
https://en.wikipedia.org/wiki/Explained_sum_of_squares

$[R^2 = r^2_{X,Y}] - ESS + SS_{res} = SS_{tot}$ (6/6)

- And simplifying what is in [•]:

$$\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \frac{\sum_j (x_j - \overline{x})(y_j - \overline{y})}{\sum_j (x_j - \overline{x})^2}(x_i - \overline{x})^2 =$$

$$= \sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) \sum_i \frac{(x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} =$$

$$= \sum_i (x_i - \overline{x})(y_i - \overline{y}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) \frac{\sum_i (x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} =$$

$$= \sum_i (x_i - \overline{x})(y_i - \overline{y}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) = 0$$

QED.

Source with modifications:

https://en.wikipedia.org/wiki/Explained_sum_of_squares

$R^2 = r^2_{X,Y}$ (2/4)

- Now we know that, under the assumption to deal with a Gaussian noise centered in 0 we have:

$$ESS + SS_{res} = SS_{tot}$$

- Under this hypothesis we have:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = \frac{ESS}{SS_{tot}}$$

$R^2 = r^2_{X,Y}$ (3/4)

- We now consider the square of $r_{X,Y} = r_{\hat{Y},Y}$

$$r^2_{\hat{Y},Y} = \left(\frac{Cov(\hat{Y},Y)}{\sqrt{Var(Y)Var(\hat{Y})}} \right)^2 = \frac{Cov(\hat{Y},Y)Cov(\hat{Y},Y)}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{Cov(\hat{Y},\hat{Y}+\epsilon)Cov(\hat{Y},\hat{Y}+\epsilon)}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{(Cov(\hat{Y},\hat{Y}) + Cov(\hat{Y},\epsilon))(Cov(\hat{Y},\hat{Y}) + Cov(\hat{Y},\epsilon))}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{Cov(\hat{Y},\hat{Y})Cov(\hat{Y},\hat{Y})}{Var(Y)Var(\hat{Y})}$$

Source with modifications:

https://economictheoryblog.com/2014/11/05/proof/

$R^2 = r^2_{X,Y}$ (4/4)

- But we know that $Cov(\hat{Y}, \hat{Y}) = Var(\hat{Y})$, therefore we get that

$$r^2_{X,Y} = \frac{Var(\hat{Y})Var(\hat{Y})}{Var(Y)Var(\hat{Y})} = \frac{Var(\hat{Y})}{Var(Y)} =$$

$$= \frac{\frac{\sum_i (\hat{y}_i - \overline{\hat{y}})^2}{n}}{\frac{\sum_i (y_i - \overline{y})^2}{n}} = \frac{\sum_i (\hat{y}_i - \overline{\hat{y}})^2}{\sum_i (y_i - \overline{y})^2} = \frac{ESS}{SS_{tot}}$$

since we have already proven that $\overline{y} = \overline{\hat{y}}$

QED

Source with modifications:

https://economictheoryblog.com/2014/11/05/proof/

Comment on $R^2 = r^2_{X,Y}$

- This is a major result
- It is the center of our subsequent investigation, in the case of normality of error we can model, interconnect, and understand relationships in an easy way
- The next question is on how the slope of the regression line (θ_1) relates to the correlation coefficient $r_{X,Y}$

$r_{X,Y}$ and θ_1

- We know that:

$$\theta_1 = \frac{Cov(X,Y)}{Var(X)}$$

- And that:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Therefore:

$$\theta_1 Var(X) = r_{X,Y} \sigma_X \sigma_Y$$

- We can then conclude that:

$$\theta_1 = \frac{\sigma_X \sigma_Y}{Var(X)} r_{X,Y}$$

$$r_{X,Y} = \frac{Var(X)}{\sigma_X \sigma_Y} \theta_1$$

Comment on $r_{X,Y} \sim \theta_1$

- $r_{X,Y}$ and θ_1 are therefore directly and monotonically proportional
- It means that a positive relationships implies a positive slope and viceversa

General remark

- Right now we work with samples of larger populations of data
- We measure properties of samples, like mean, standard deviation, covariance, correlation coefficient
- All these properties are also random variable and have a distribution
- Our question is therefore, what kind of distribution is the one of the correlation coefficient
- Knowing its distribution allows us to understand the relationships existing between the variables it connect

Non parametric correlations

- Spearman's rank correlation coefficient
- Kendall's τ
- $\ldots$

Spearman's Rank Correlation Coeff. (1/3)

- What can we do when the data is not normally distributed?
- Or even if the data is not on a ratio scale, just on an ordinal scale?

- *If the data is on a nominal scale, the concept of correlation looses interest; at most we can consider clustering.*

Spearman's Rank Correlation Coeff. (2/3)

Idea:

- Transform the data into ranks
- Apply the Pearson correlation coefficient to ranks
- Indeed, the values can be different, and also the significance and the mutual relationship
- Remember that:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- And also that:

$$\theta_1 = \frac{\sigma_X \sigma_Y}{Var(X)} r_{X,Y}$$

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman's Rank Correlation Coeff. (3/3)

Definition:

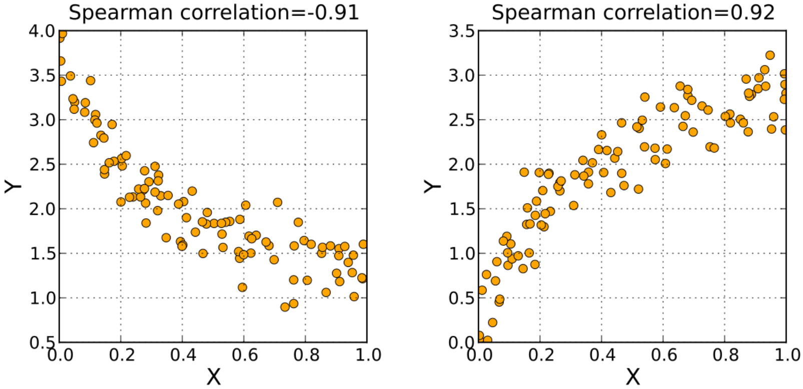
- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's define
  - $(\forall X_i \in X)\ rk_{X\,i} = \text{rank}(X_i, X)$, $Rk_X = \{rk_{X_i}\}$
  - $(\forall Y_i \in Y)\ rk_{Y_i} = \text{rank}(Y_i, Y)$, $Rk_Y = \{rk_{Y_i}\}$
- We define the Spearman's Rank Correlation Coefficient between X and Y, $r_S(X, Y)$ as:

$$r_S(X,Y) = r(Rk_X, Rk_Y) = \frac{Cov(Rk_X, Rk_Y)}{\sigma_{Rk_X} \sigma_{Rk_Y}}$$

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Visualization of r_S

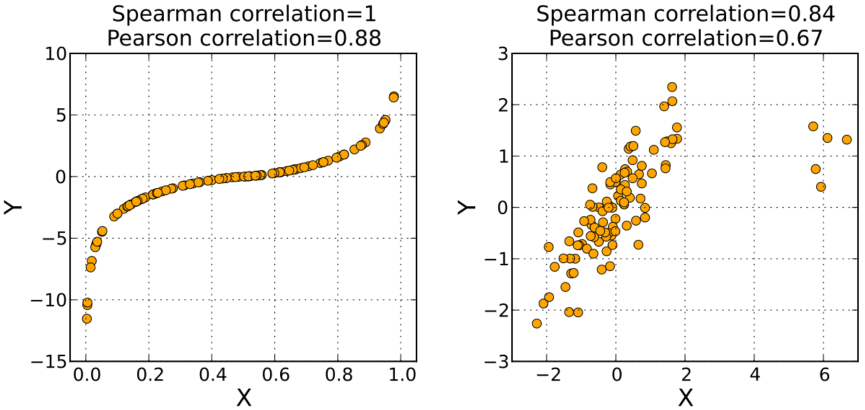
Spearman's Rank Correlation Coefficient is based on monotonicity:



Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

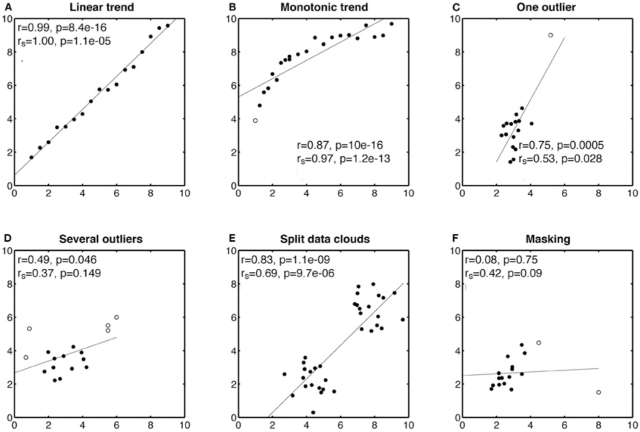
r and r_S (1/2)

Indeed, the values of r and r_S can be different:



Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

r and r_S (2/2)



Source with modifications: *https://www.researchgate.net/figure/Examples-of-Pearson-and-Spearman-correlations-In-each-subplot-r-is-Pearson-correlation_fig7_224915794*

About r_S

Note that:

- Two identical values are assigned their fractional rank
  - So if we have as values 20, 20, 30, 35, 36, then their ranks should be 1.5 (the average between 1 and 2), 1.5, 3, 4, 5 respectively
- Taking into account that we are dealing with integer ranks, we can simplify the formula as follows if all values are different:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  - where each d_i is equal to the difference in rank between X_i and Y_i:

$$d_i = Rk_{X_i} - Rk_{Y_i}$$

  - and n is the number of observations

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Significance of r_S (1/2)

- Being based on ordinals and non assuming anything on the distribution of the underlying populations, the computation of the significance of r_S is based on permutations
- This belong to the family of permutation tests
    - A permutation test (or exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points
- In our case, since I have sequence of ordinals, we can consider all possible pairs of mutual relationships and, based on this, determine if the monotonic relationship that we have obtained is significantly different from a random order

Source with modifications: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

Significance of r_S (2/2)

- Consider as an example the dataset $\{(X_i, Y_i)\} = \{(10, 2), (15, 0), (20, 4), (21, 50)\}$
- Does it have a significant positive correlation?
- We need to assign ranks the elements, leading to $\{(Rk_{X_i}, Rk_{Y_i})\} = \{(1, 2), (2, 1), (3, 3), (4, 4)\}$
- This leads to $r_S = 0.8$
- To compute the significance, I determine the number of times the comparison $Rk_{Y_i} \leq Rk_{Y_j}$ are true when $i < j$
- These are sequences of Bernoulli trials ...

Source with modifications: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

Kendall's τ (1/2)

An alternative non parametric correlation coefficient is the Kendall's τ

- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's assume that the two sets X and Y do not contain duplicates
- Let's define
  - a concordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i > Y_j)$ or $(X_i < X_j$ and $Y_i < Y_j)$
  - a discordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i < Y_j)$ or $(X_i > X_j$ and $Y_i < Y_j)$

Source with modifications: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

Kendall's τ (2/2)

- We can define the Kendall's τ as:

$$\tau = \frac{(\#\text{ concordant pairs}) - (\#\text{ discordant pairs})}{n(n-1)/2}$$

Source with modifications: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

Domande?

Fine della lezione dieci.