Al-Farabi Kazakh National University

Pearson’s Correlation

Spearman's Rank Correlation

Saktapov Akylbek

Correlation

Correlation is a measure of an association between two variables. A relationship between two variables is a one in which either as the value of one variable increases, so does the value of the other variable; or as the value of one variable increases, the other variable value decreases.

Correlation is measured by a statistic called the correlation coefficient, which represents the strength of the putative linear association between the variables in question.

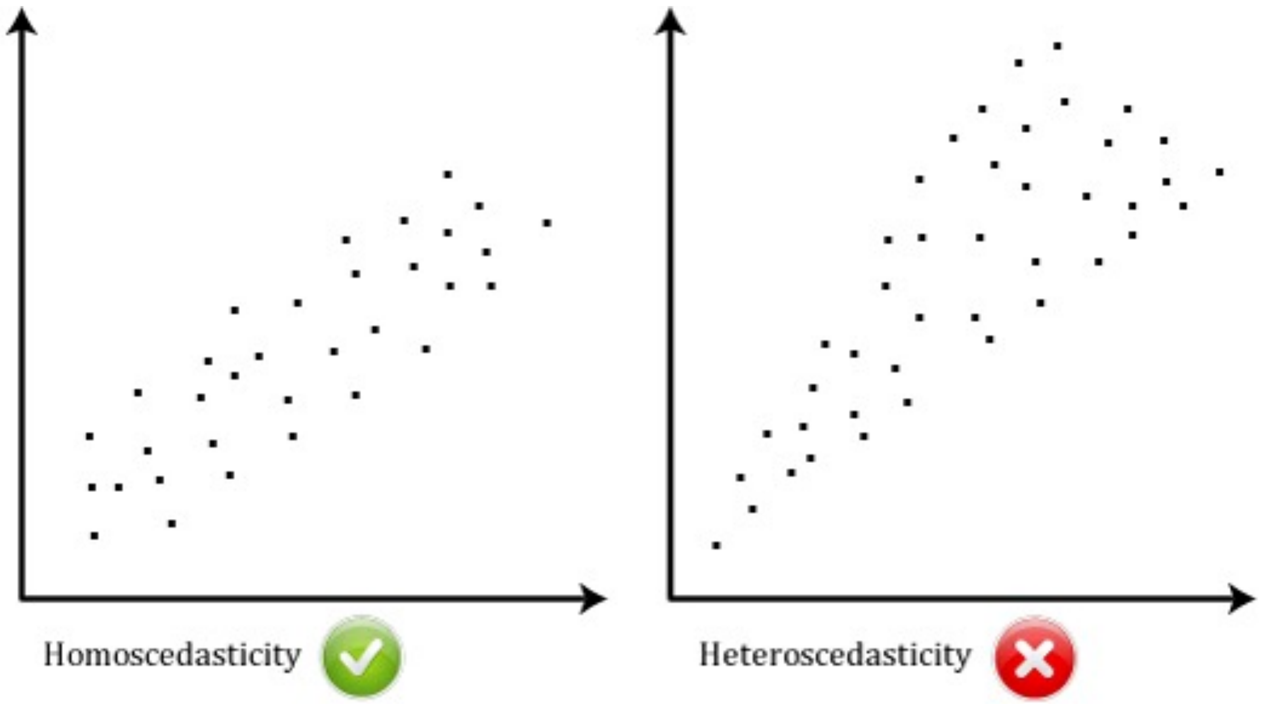
Pearson correlation coefficient (PCC)

In statistics, the **Pearson correlation coefficient** (PCC) is a statistic that measures linear correlation between two variables X and Y. Correlation analysis can only determine the strength and direction of the relationship between the variables.

Assumptions of PCC

- Both variables are quantitative, continuous; independence of research participants from each other;
- Sufficiently large sample size, at least 25 observations, the sample must be representative;
- The features have a normal distribution;
- Homoscedasticity and the relationship between variables must be linear.

Homoscedasticity vs Heteroscedasticity

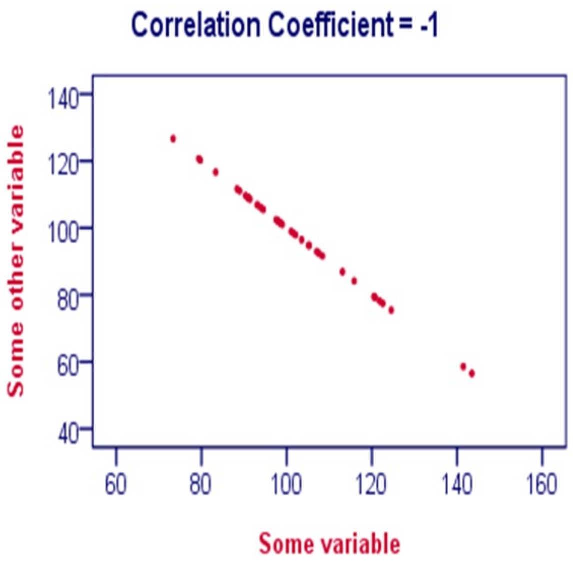


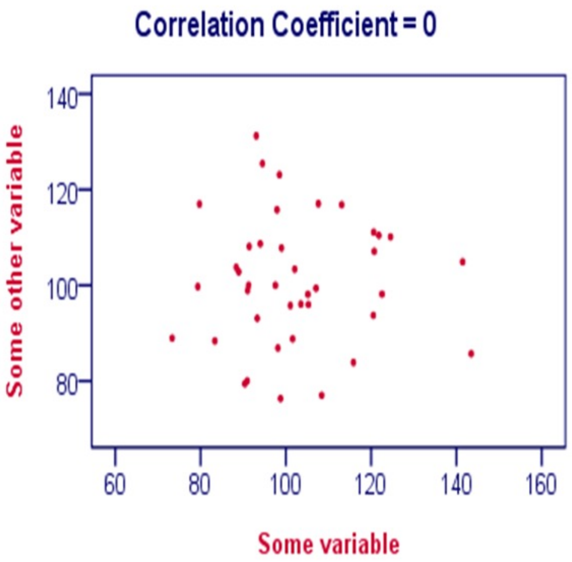
Calculation of PCC

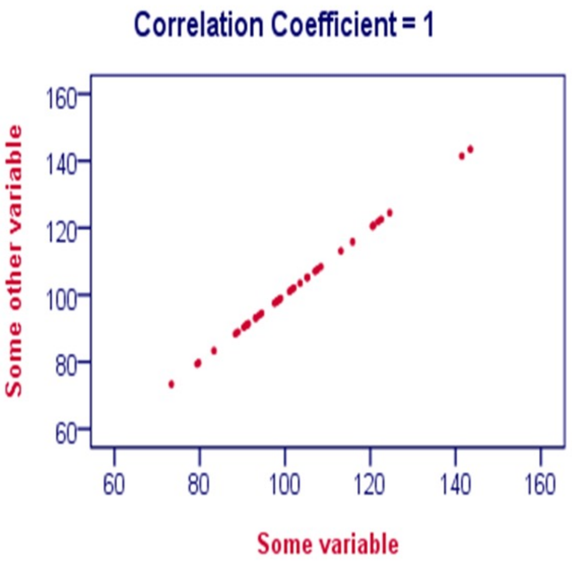
Pearson correlation coefficient	**95% Confidence interval:**

r = $\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{(n\sum x^2)-(\sum x)^2)(n\sum y^2-(\sum y)^2)}}$
where,
r – correlation coefficient;
x – first variable;
y – second variable;
n – sample size
	Z_L = 0.5 * ln$\left(\frac{1+r}{1-r}\right)$ – $\frac{1.96}{\sqrt{n-3}}$
	Z_U = 0.5 * ln$\left(\frac{1+r}{1-r}\right)$ + $\frac{1.96}{\sqrt{n-3}}$
	where, ln – logarithm, r – correlation coefficient, n – sample size
	after that you have to use the formula to find 95% CI
	from $\frac{e^2*(Z_L)-1}{e^2*(Z_L)+1}$ **to** $\frac{e^2*(Z_U)-1}{e^2*(Z_U)+1}$
	where, e (Euler's number) ≈ 2.7.

Interpretation of the Correlation Coefficient


Correlation Coefficient = -1


Correlation Coefficient = 0


Correlation Coefficient = 1

Size of Correlation	**Meaning**
0.0 – 0.10	Negligible correlation
0.1 – 0.39	Weak correlation
0.4 – 0.69	Moderate correlation
0.7 – 0.89	Strong correlation
0.9 – 1.00	Very strong correlation

- **Limit:** Coefficient values can range from -1 to +1, where, -1 indicates a perfect negative relationship, +1 indicates a perfect positive relationship, and 0 indicates no relationship exists..
- **Pure number:** It is independent of the unit of measurement. For example, if one variable’s unit of measurement is in inches and the second variable is in quintals, even then, Pearson’s correlation coefficient value does not change.
- **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Spearman Rank Correlation

- Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables.
- The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

Types of research questions a Spearman Correlation can examine:

- Is there a statistically significant relationship between participants' level of education (high school, bachelor's, or graduate degree) and their starting salary?
- Is there a statistically significant relationship between horse's finishing position a race and horse's age?

Assumptions

- Two variables should be measured on an ordinal, interval or ratio scale. Examples of ordinal variables include Likert scales (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-pont scale explaining how much a customer liked a product, ranging from "Not very much", to "It is OK", to "Yes, a lot").
- Two variables represent paired observations.
- There is a monotonic relationship between the two variables. A monotonic relationship exists when either the variables increase in value together, or as one variable value increases, the other variable value decreases.

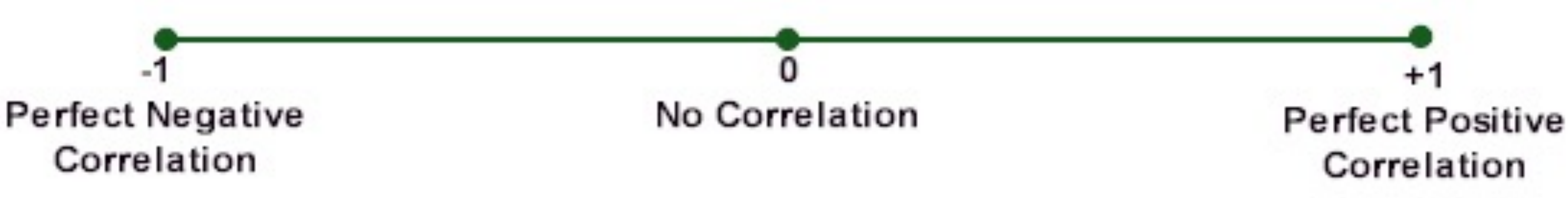
The coefficient (R_s) is calculated on this calculator using the common formula:

$$R_s = 1 - \left(\frac{6 \Sigma d^2}{n^3 - n} \right)$$

where d = difference in paired ranks and n = number of cases.

The answer will always be between 1.0 (a perfect positive correlation) and -1.0 (a perfect negative correlation).

An R_s of 0 indicates no association between ranks.



The strength of a correlation

Value of coefficient *Rs* (positive or negative)	Meaning
0.00 to 0.19	A very weak correlation
0.20 to 0.39	A weak correlation
0.40 to 0.69	A moderate correlation
0.70 to 0.89	A strong correlation
0.90 to 1.00	A very strong correlation

T-test for R

$$t = r \text{ x } \surd n - 2) / \surd(1 - r^\wedge 2)$$

$p < 0,05$ significant

$p > 0,05$ not significant

Thank you for your attention!