Correlation

APPLIED STATISTICS - FCUP

Rita Gaio
argaio@fc.up.pt

2020

Covariance and Sample Covariance

The **covariance** between two continuous r.v. X and Y describes the degree to which those variables tend to deviate from their expected values:

$$Cov(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

or, equivalently,

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

For two random samples $x_1, \ldots, x_n$, of X, and $y_1, \ldots, y_n$, of Y, the **sample covariate** corresponds to the sample equivalent of the previous formulae, namely

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

or

$$\widehat{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\overline{x}\,\overline{y}, \quad \text{respectively.}$$

Covariance

Proposition: If X and Y are independent then $Cov(X, Y) = 0$; however, the inverse is not necessarily true.

Moreover,

- $Cov(X, X) = Var(X)$
- $Cov(X, Y)$ takes values in $\mathbb{R}$ (has no upper nor lower bound)
- units of covariance $=$ (units X)$\times$(units Y).

Pearson's Correlation Coefficient

A dimensionless quantity bounded below and above is the **correlation coefficient**:[1]

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Proposition: for any r.v. X and Y,

(a) $Cor(X,X) = 1$

(b) $Cor(X,Y) = Cor(Y,X)$

(c) $-1 \leq Cor(X,Y) \leq 1$

(d) X, Y independent $\implies Cor(X,Y) = 0$
(the inverse is not necessarily true)

(e) $|Cor(X,Y)| = 1$ if and only if X and Y have a linear relationship with a nonzero slope

[1] also denoted by $Cor(X,Y)$ or $Corr(X,Y)$

Pearson's Correlation Coefficient

Pearson[2] correlation coefficient ρ_{XY} measures the degree of the linear association between X and Y

- sign of ρ_{XY}: X and Y vary in the same way (positive correlation) or in opposite ways (negative correlation)
- absolute value ρ_{XY}: measures the strength of the linear association.

[2] Karl Pearson, 1857-1936

Pearson's Correlation Coefficient

Proposition: for any r.v. X and Y,

- **(f)** correlation is invariant under linear transformations of a single variable, up to the sign of the transformation:
  $Cor(aX + b, Y) = sign(a)Cor(X, Y)$, for all $a \in \mathbb{R}$
- **(d)** $Cor(aX + b, cY + d) = sign(ac)Cor(X, Y)$, for any $a, b, c, d \in \mathbb{R}$
- **(e)** $Cor\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right) = Cor(X, Y)$

Remark: high correlation does not necessarily imply causality. Indeed, two r.v. may be highly correlated for several reasons:

- X causes Y
- Y causes X
- a 3rd factor, directly or indirectly, causes X and Y
- an unlikely event has occured.

Sample Pearson's Correlation Coefficient

For two random samples $x_1, \ldots, x_n$, of X, and $y_1, \ldots, y_n$, of Y, the **sample (Pearson) correlation coefficient** corresponds to the sample equivalent of the previous formula, ie,

$$r_{xy} = \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}}$$

where

$$\begin{array}{rcl} SS_{xy} & = & \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ SS_{xx} & = & \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ SS_{yy} & = & \sum_{i=1}^{n}(y_i - \overline{y})^2. \end{array}$$

Sample Pearson's Correlation Coefficient

Proposition: for any random samples $x_1, \ldots, x_n$ of X and $y_1, \ldots, y_n$ of Y, and denoting the vectors of observations by $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the following holds:

(a) $r_{xx} = 1$

(b) $r_{xy} = r_{yx}$

(c) $r_{ax+b,y} = sign(a) r_{xy}$

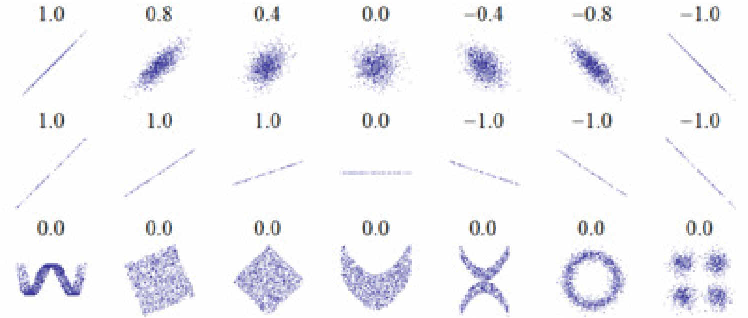
(d) if x and y are centered vectors (with zero mean), then $r_{xy} = \cos(\theta)$ where θ is the angle defined by x and y in $\mathbb{R}^n$

(e) $-1 \leq r_{xy} \leq 1$

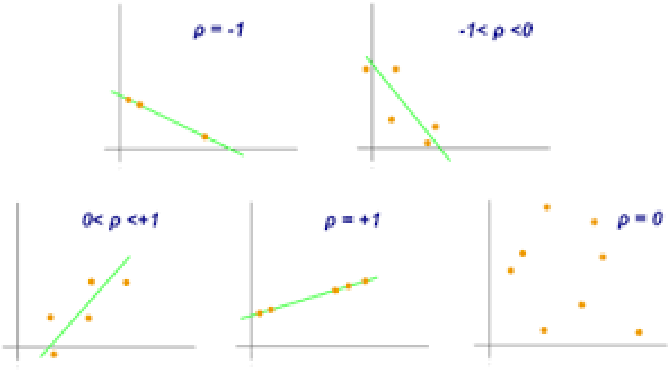
(f) $|r_{xy} = 1|$ if and only if $y = ax + b$, for some real constants a and b with $a \neq 0$.

(g) $r_{x,y} = 0$ if and only if x and y are orthogonal vectors in $\mathbb{R}^n$.

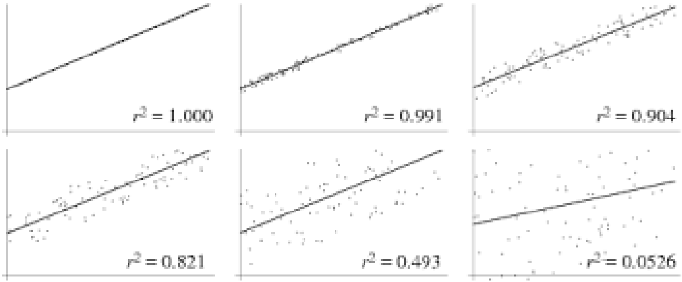
Sample Pearson's Correlation Coefficient



Sample Pearson's Correlation Coefficient



Sample Pearson's Correlation Coefficient



Pearson's Correlation Test

- **Data**: random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of a pair of continuous random variables (X, Y)
- H_0: $Cor(X, Y) = 0$, $\quad H_1 : Cor(X, Y) \neq 0$
  H_1 implies that X and Y are not independent
- **Requirements**: the pair (X, Y) follows a bivariate normal distribution[3]
- **Test Statistic**: assuming H_0, $T = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2)$
- **Decision**: to reject H_0 at an α level whenever $|t| > t_{1-\alpha/2}(n-2)$

If H_0 is $Cor(X, Y) = \rho_0 \neq 0$, then a different test statistic is required (Fisher's Z-statistic or Hotelling's statistic, depending on the sample size).

[3] the test remains true if at least one of the variables follows a normal distribution

Pearson's Correlation Test

Instructions in **R**:

```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95, ...)
```

where

- x and y are the vector of observations in the random sample
- *alternative* corresponds to the formulation of the alternative hypothesis
- *method* indicates the correlation coefficient to be used.

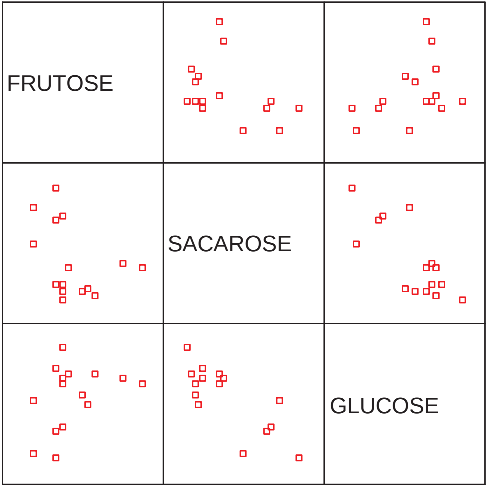
Pearson's Correlation Test - Example 1

The table below represents concentrations (g, l^{-1}) of frutose, sacarose and glucose present in 15 samples of apple juice.

Frutose	Sacarose	Glucose
40	20	6
49	27	11
47	26	10
47	34	5
40	29	16
49	6	26
47	10	22
51	14	21
49	10	20
49	8	19
55	8	17
59	7	21
68	15	20
74	14	19
57	9	15

Question: Assuming the normality requirements, is there enough statistical evidence to say that sacarose decreases linearly with glucose?

Pearson's Correlation Test - Example 1



Pearson's Correlation Test - Example 1

Sample Pearson's correlation is -0.775 for which we obtain p=0.001.
Hence:

- At a 5% significance level, we can reject H_0 and conclude that the variables are not independent, being negatively linearly associated.
- if sacarose and glucose were independent, sample values or a more extreme situation (further from zero correlation) only occurs 1% of the time, due to random sampling.

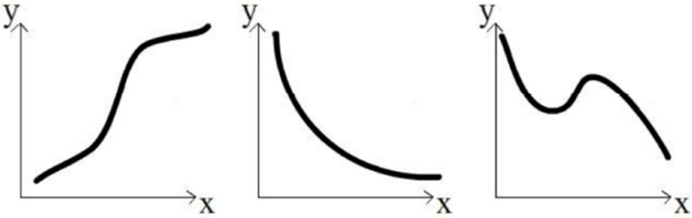
Spearman[5] Correlation Coefficient

- uses ranks instead of the original values of the variables, thus being possible to apply to ordinal variables
- corresponds to Pearson's correlation coefficient applied to the ranks of the observations within each sample
- $-1 \leq \rho_S \leq 1$ and $1 \leq r_S \leq 1$
- detects monotone associations (not simply linear)[4]
- $\rho_S > 0$ (resp < 0) corresponds an increasing (resp.decreasing) monotony
- $|\rho_S|$ gives the strength of the monotone association between X and Y:
  - $|\rho_S| \approx 1 \implies$ very strong monotone association
  - $|\rho_S| \approx 0 \implies$ very weak monotone association.

[4] a linear association is a monotone association but the inverse is not true

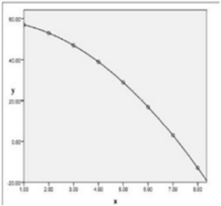
[5] Charles Spearman, 1863-1945

Spearman Correlation Coefficient

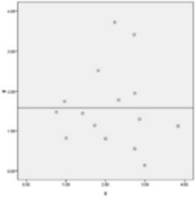


monótona crescente   monótona decrescente   não monótona

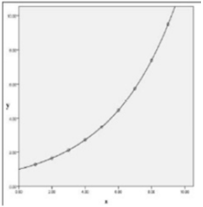
Spearman Correlation Coefficient



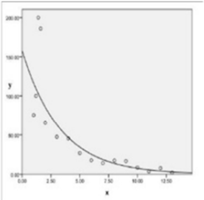
$r_S = -1$



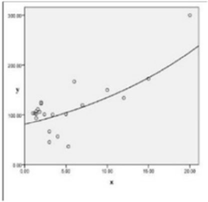
$r_S = 0$



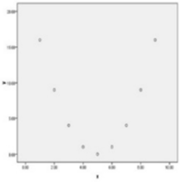
$r_S = 1$



$r_S = -0.941$



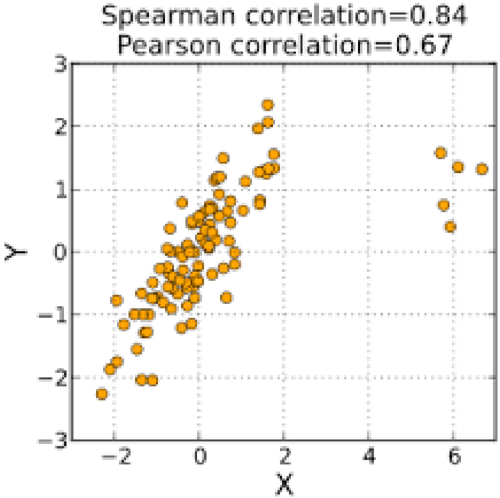
$r_S = 0.372$



$r_S = 0$

Spearman Correlation Coefficient

Spearman correlation coefficient is less sensitive to the presence of outliers than Pearson's correlation coefficient.



Spearman's Correlation Test

- **data**: random sample $(x_1, y_1), \ldots, (x_n, y_n)$ of a pair of continuous or ordinal random variables (X, Y)
- denote by $\xi_1, \ldots, \xi_n$ (resp. $\eta_1, \ldots, \eta_n$) the ranks of $x_1, \ldots, x_n$ (resp. $y_1, \ldots, y_n$) and define $d_i = \xi_i - \eta_i$; then

$$r_S = r_{(\xi_1,\ldots,\xi_n),(\eta_1,\ldots,\eta_n)} = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)}$$

  X, Y independent $\Longrightarrow$ $r_S \approx 0$

- H_0 : There is no (monotonic) association between the two variables (in the population), ie, $\rho_S = 0$
  H_1: There is a (monotonic) association between the two variables (in the population), ie, $\rho_S \neq 0$[6]
- **Test Statistic**: assuming H_0, $r_S \sim \mathcal{S}(n)$ where $\mathcal{S}(n)$ is a known distribution (implemented in softwares)
- **Decision**: to reject H_0 at an α level whenever $|r_S| \geq \mathcal{S}_{1-\alpha/2}(n)$

[6] $\rho_S \neq 0 \Longrightarrow X$ and Y not independent

Spearman's Correlation Test

Remarks:

- Spearman's test does not assume conditions on the distribution of (X, Y)
- statistical significance does not provide any information about the strength of the relationship between the two variables.
  For example, p=0.001 does not mean a stronger relationshio than the one found with p=0.04.

Spearman's Correlation Test

Instructions in **R**:

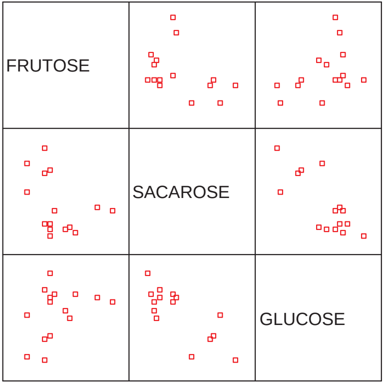
```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95, ...)
```

where

- *x* and *y* are the vector of observations in the random sample
- *alternative* corresponds to the formulation of the alternative hypothesis
- *method* indicates the correlation coefficient to be used.

Spearman's Correlation Test - Example 1

Consider again the previous data



Question: Is there a significant monotonic association between frutose and glucose?

Spearman's Correlation Test - Example 1

Sample Spearman's correlation is $r_S = 0.392$ and we obtain $p = 0.148$. Hence:

- assuming no monotonic association between frutose and glucose, a result at least as extreme as the one observed in the sample occurs 14.8% of the time, due to random sampling.
- if no association exists, there is more than a 5% chance that the strength of the relationship found (0.392) happened by chance.