# Pearson Correlation and Spearman Correlation test

## Pearson Correlation test

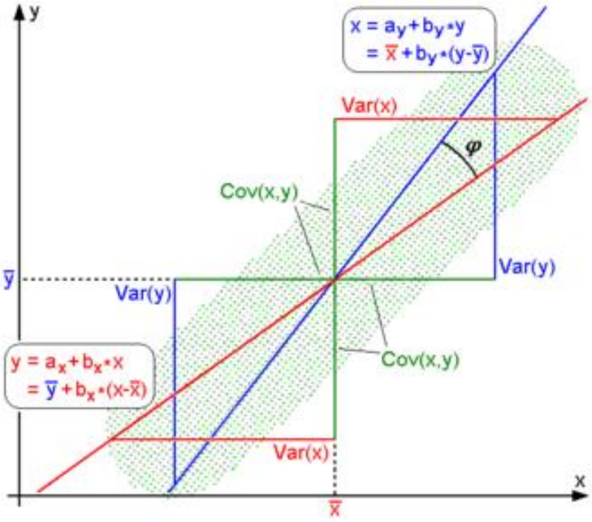
### Definition and purpose

Measures the strength and direction of a linear association between two variables. It defines as

$$p_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

### Geometric Interpretation

Assume we have two regression line, $y = g_x(x)\ and\ x = g_y(y)$, with angle $\varphi$



## Spearman Correlation test

### Definition and purpose

Measures the strength and direction of a **monotonic association** (whether linear or not) between **two ranked variables (or ordinal data)**

For a sample of size n, the n raw scores $X_i, Y_i$ are converted to ranks $R(X_i)$, $R(Y_i)$, and $r_s$ is computed as

$$r_s = \frac{cov(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$ (Only if in both variables **all ranks are distinct**)

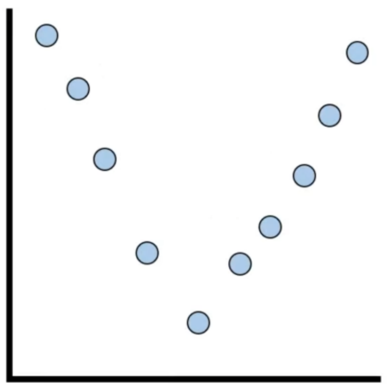
$$d_i = R(X_i) - R(Y_i), n\ is\ the\ number\ of\ observations$$

**If the trend line between two datasets goes up and down, the spearman correlation is not suitable for this situation.**

## Assumption

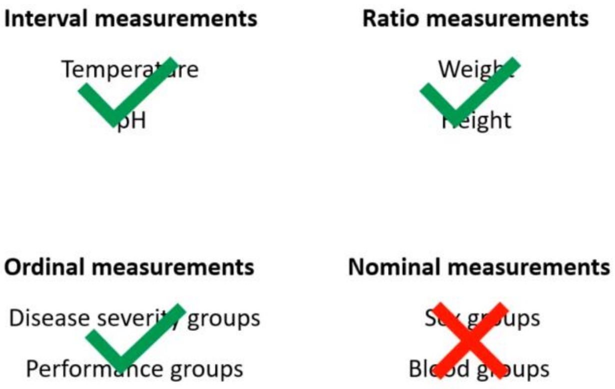
1. Random Sample
2. **A monotonic association exists**

**The variables must exhibit a monotonic correlation <u>before</u> you run the Spearman correlation test**



3. Variables are at least ordinal

**Both variables should be measured on a <u>continuous (interval or ratio) or ordinal scale</u>**



**Interval measurements**

Temperature

pH

**Ratio measurements**

Weight

Height

**Ordinal measurements**

Disease severity groups

Performance groups

**Nominal measurements**

Sex groups

Blood groups

4. Data contains paired samples
5. Independence of observations

# Determining significance

## Permutation Test

Use the premutation test (rely on resampling (bootstrapping) the original data assuming the null hypothesis, based on the resampled data to conclude how likely the original data to occur under the null hypothesis)

## Fisher Transformation (Fisher Z-transformation)

The fisher transformation can be used to test hypotheses about the value of the population correlation coefficient $p$ between X and Y.

### Definition

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

And standard error equals to

$$\frac{1}{\sqrt{N-3}}(N\ is\ the\ sample\ size)$$

Or directly use this equation

$$z_{std} = \sqrt{\frac{n-3}{1.06}} * \frac{1}{2}\ln(\frac{1+r}{1-r})$$

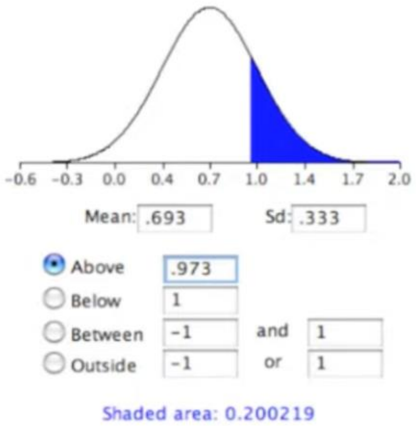
To get the standard z score (std error = 1)

### Examples

Suppose we have the null hypothesis assuming $r = 0.6$, our sample size is 12, what is the p-value (risk of $\alpha$ error) for $r > 0.75$

After the Fisher Z-transformation, we get a normal distribution with $\mu = \frac{1}{2} * \ln\frac{1+0.6}{1-0.6} = 0.693$ with standard error $e = \frac{1}{\sqrt{12-3}} = 0.333$.

For the alternative hypothesis, we get the Z score after the transforming $z = \frac{1}{2} * \ln\frac{1+0.75}{1-0.75} = 0.973$



Mean: .693 Sd: .333

● Above .973

○ Below 1

○ Between -1 and 1

○ Outside -1 or 1

Shaded area: 0.200219

# Kendall rank correlation coefficient

## Definition

Kendall’s rank correlation is a statistic used to measure the ordinal association (**degree of concordance**) between two measured quantities. Both **Kendall’s $\tau$** and **Spearman’s $\rho$** can be formulated as special cases of **General correlation coefficient.**

The greater the number of “**inversions**”, the smaller the coefficient will be. Range [-1,1]

**Here we talk about $\tau - a$, not $\tau - b$ (can handle the tied correlations)**

$$\tau = \frac{C - D}{C + D}$$

$C$ is the number of concordant pairs, $D$ is the number of discordance pairs.

**Concordant pairs:** The number of observed ranks below a particular rank which are larger than that particular rank.

**Discordant pairs:** The number of observed ranks below a particular rank which are smaller in value than that particular rank.

## Determining significance

$$Z = \frac{3 * \tau * \sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

## Example

| **Master** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Student** | 2 | 1 | 4 | 3 | 6 | 5 | 8 | 7 | 10 | 9 | 12 | 11 |
| **C** | 10 | 10 | 8 | 8 | 6 | 6 | 4 | 4 | 2 | 2 | 0 | |
| **D** | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |

$$Kendall's\,\tau = \frac{60-6}{60+6} = .818$$

$$Z = \frac{3*.818*\sqrt{12(12-1)}}{\sqrt{2(2*12+5)}} = 3.7019 > 1.96$$

So, it is significance association here.

Another example is

| **Master** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Student2** | 12 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 1 |
| **C** | 0 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | |
| **D** | 11 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |

$$Kendall's\,\tau = \frac{45-21}{45+21} = .364$$

## Kendall's $\tau$ vs Spearman's $\rho$

| | **Master** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Student | 2 | 1 | 4 | 3 | 6 | 5 | 8 | 7 | 10 | 9 | 12 | 11 |
| $\boldsymbol{\tau}$ | C | 10 | 10 | 8 | 8 | 6 | 6 | 4 | 4 | 2 | 2 | 0 | |
| | D | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | |
| $\boldsymbol{\rho}$ | $d$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | $d^2$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

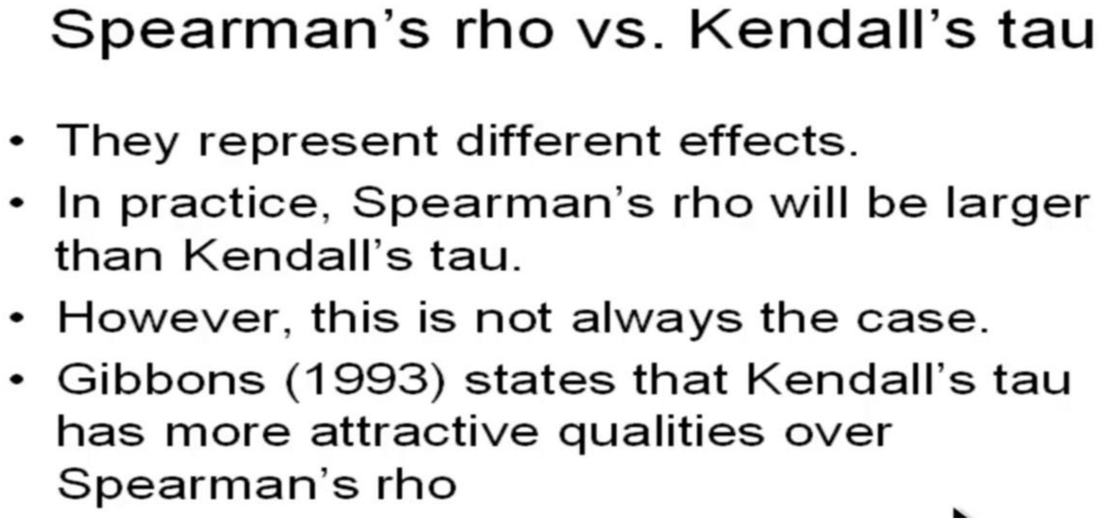
$$Kendall's\,\tau = \frac{60-6}{60+6} = .818$$

$$Spearman's\,\rho = 1 - \frac{6\sum\left(12_i^2\right)}{12(12^2-1)} = .958$$

| | **Master** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Student | 12 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 1 |
| $\boldsymbol{\tau}$ | C | 0 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | |
| | D | 11 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| $\boldsymbol{\rho}$ | $d$ | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| | $d^2$ | 121 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 121 |

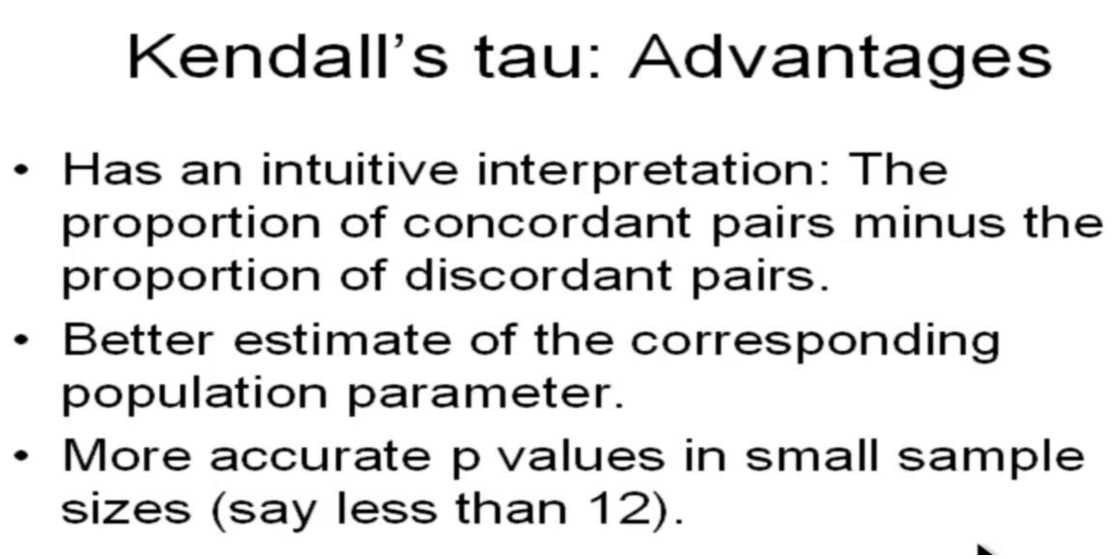
$$Kendall's\,\tau = \frac{45-21}{45+21} = .364$$

$$Spearman's\ \rho = 1 - \frac{6\sum(242_i^2)}{12(12^2-1)} = .154$$



## Spearman’s rho vs. Kendall’s tau

- They represent different effects.
- In practice, Spearman’s rho will be larger than Kendall’s tau.
- However, this is not always the case.
- Gibbons (1993) states that Kendall’s tau has more attractive qualities over Spearman’s rho



## Kendall’s tau: Advantages

- Has an intuitive interpretation: The proportion of concordant pairs minus the proportion of discordant pairs.
- Better estimate of the corresponding population parameter.
- More accurate p values in small sample sizes (say less than 12).