UNIVERSITAT DE BARCELONA

# Correlation tests

## Research Methods in Cyberspace, Behavior and e-Therapy

David Leiva Ureña

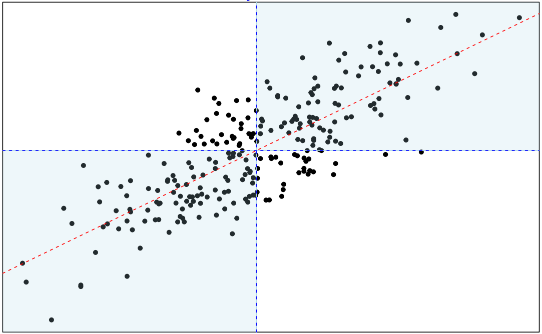
dleivaur@ub.edu

Quantitative Psychology Section. Faculty of Psychology

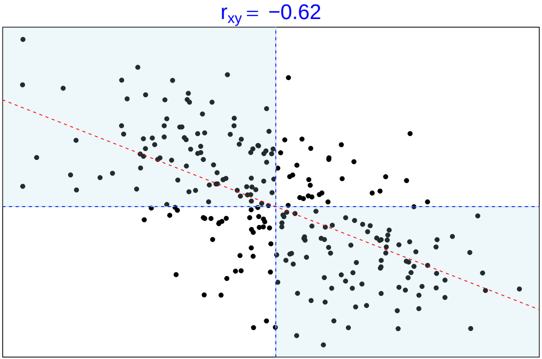
- In this unit we will introduce some inferential procedures related to correlations.
- Specifically, statistical tests for Pearson's (a.k.a. product-moment correlation coefficient), Spearman's (a.k.a. rank correlation coefficient), and Kendall's correlations are presented in the following slides.
- These procedures allow researchers to make decisions regarding the possible relationship between to quantitative variables.
- Pearson's correlation test is parametric and non distribution-free, whereas, Spearman's and Kendall's correlations tests are non-parametric and distribution free.
- Let's start by briefly reviewing these association indices.

- Pearson's correlation ($r_{xy}$) is adequate to quantify linear relationship between two quantitative variables.
- $-1 \leq r_{xy} \leq 1$. It is equal to 0 when the variables are linearly independent.
- It requires interval or ratio scales to be used.

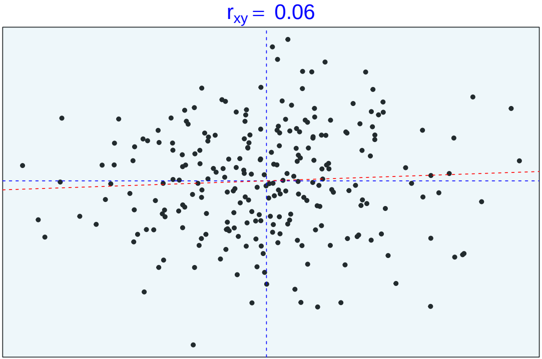
$r_{xy} = 0.8$



- Pearson's correlation ($r_{xy}$) is adequate to quantify linear relationship between two quantitative variables.
- $-1 \leq r_{xy} \leq 1$. It is equal to 0 when the variables are linearly independent.
- It requires interval or ratio scales to be used.



- Pearson's correlation ($r_{xy}$) is adequate to quantify linear relationship between two quantitative variables.
- $-1 \leq r_{xy} \leq 1$. It is equal to 0 when the variables are linearly independent.
- It requires interval or ratio scales to be used.



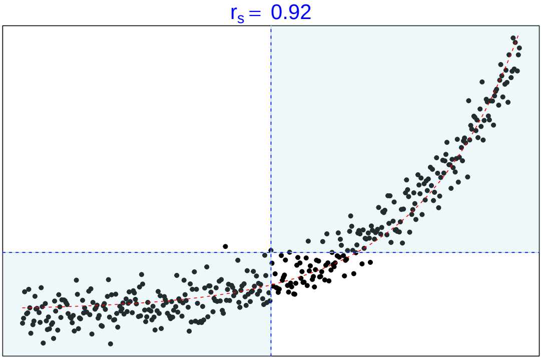
**Introduction**
Pearson's correlation test
Spearman's correlation test
Kendall's correlation test

- General expression of Pearson's correlation index is given by:

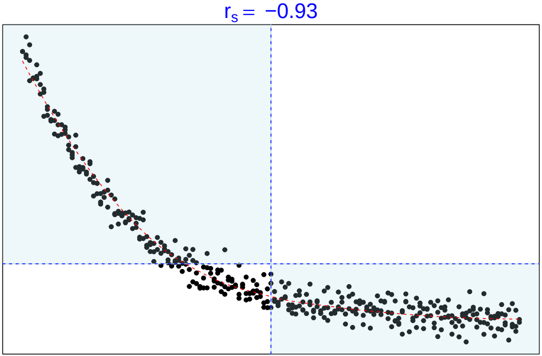
$$r_{XY} = \frac{\frac{\sum_{i=1}^{i=n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \overline{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{i=n}(y_i - \overline{y})^2}{n-1}}} = \frac{s_{XY}}{s_X s_Y}$$

- Some limitations of this correlation index:
  - Non-linear relationships cannot be quantified with this index.
  - Pearson's correlation is not resistent to the presence of outliers.
  - It can be inadequate when there is a range restriction.

- Spearman's correlation ($r_s$) allows researchers to quantify (non-)linear relationships between two quantitative variables.
- $-1 \leq r_s \leq 1$. $r_s = 0$ when the variables are not monotonically related.
- It can be used with ordinal scales as well as with ratio and interval ones.



- Spearman's correlation ($r_s$) allows researchers to quantify (non-)linear relationships between two quantitative variables.
- $-1 \leq r_s \leq 1$. $r_s = 0$ when the variables are not monotonically related.
- It can be used with ordinal scales as well as with ratio and interval ones.



- Original variables $X$ and $Y$ are transformed into ranks.
- When there are ties in the sample, average ranks are assigned to the tied observations.
- General expression of Spearman's correlation index is given by:

$$r_s = \frac{\frac{\sum_{i=1}^{i=n}(r_i-\bar{r})(s_i-\bar{s})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{i=n}(r_i-\bar{r})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{i=n}(s_i-\bar{s})^2}{n-1}}} = 1 - \frac{6\sum_{i=1}^{i=n}(R_i - S_i)^2}{n(n^2-1)}$$

- Some limitations of this correlation index:
  - It is not the best choice when there is a range restriction.
  - The previous expressions are biased when there are tied observations.
  - Non-monotonic relationships (e. g. quadratic relationships) cannot be adequately quantified with this index.

- Kendall's correlation index ($\tau$) allows us to quantify the relationship between ordinal scales.
- Originally thought for continuous variables, this estimator is also affected by the presence of tied observations.
- In general, it is based on the comparison of the frequencies of concordant and discordant pairs in the sample.
- $-1 \leq \tau \leq 1$. Basic interpretation guidelines:
  - $\tau = -1$, when all pairs are discordants and the relationship is thus inverse.
  - It is equal 0 when the two variables are non-associated.
  - $\tau = 1$, when all pairs are concordants and the relationship is thus direct.
- There are three $\tau$ indices:
  - $\tau_a = \frac{2(C-D)}{n(n-1)}$: Adequate when there are no ties in the sample.
  - $\tau_b = \frac{C-D}{\sqrt{\left(\frac{n(n-1)}{2} - T_Y\right)\left(\frac{n(n-1)}{2} - T_X\right)}}$: It is adequate for square tables (same number of rows and columns) and tied observations in the sample.
  - $\tau_c = \frac{2q(C-D)}{n^2(q-1)}$: The statistic of choice for rectangular tables.

- This statistical test allows us to make a decision regarding a linear relationship between two quantitative variables (measured by means interval or ratio scales) in a specific population.
- It is possible to approximate sampling distribution for Pearson's correlation by using normal distribution in case of relying with large samples.
- General statistical hypotheses in this case are the following:

$$H_0 : \rho \geq \rho_0 \qquad H_0 : \rho \leq \rho_0 \qquad H_0 : \rho = \rho_0$$

- In case of being interested in an independence test: $H_0 : \rho = 0$.
- Necessary condition to approximate sampling distribution of Pearson's correlation by means of a normal distribution is that the joint distribution of the two quantitative variables is normal bivariate.
- Thus, by using *Fisher's transformation*, the sampling distribution is approximately normal:

$$\rho \sim N\left(\frac{1}{2} ln \frac{1+\rho_0}{1-\rho_0}; \frac{1}{\sqrt{n-3}}\right)$$

- Note that in the previous expression we can use any real value (between $-1$ and 1) for the parameter correlation.

- In case of testing whether $\rho = 0$, and if having very large samples, the Pearson's correlation would approximately follow a $N\left(0; \frac{1}{\sqrt{n-3}}\right)$ distribution.
- When dealing with not such a large samples, Pearson's correlation is more suitably described by $t_{\nu=n-2}$ distribution.
- Thus, a less restrictive test for making a decision regarding the independence between two quantitative variables is:

$$t = \frac{\sqrt{(n-2)}r_{XY}}{\sqrt{1-r_{XY}^2}} \sim t_{\nu=n-2}.$$

- To make a decision regarding any other value of the parameter $\rho$, we can use the above-mentioned *Fisher's transformation* as long as we are dealing with a large sample:

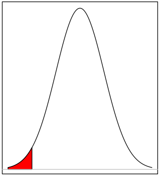
$$z = \frac{\sqrt{n-3}}{2} ln\left(\frac{(1+r_{XY})(1-\rho_0)}{(1-r_{XY})(1+\rho_0)}\right) \sim N(0,1).$$

# Pearson's correlation test

## Description

- A statistical decision can be made after obtaining the probability associated with the test statistic, that is, its *p*-value. Several expressions to get this probability depending on the statistical hypothesis:

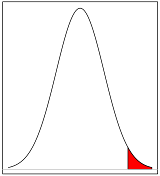
Left-tailed test



$p = P[Z \leq z_{\text{obs}} \mid H_0]$

$p = P[t_\nu \leq t_{\text{obs}} \mid H_0]$

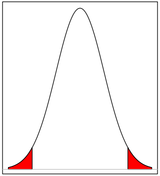
Right-tailed test



$p = P[Z \geq z_{\text{obs}} \mid H_0]$

$p = P[t_\nu \geq t_{\text{obs}} \mid H_0]$

Two-tailed test



$p = P[|Z| \geq |z_{\text{obs}}| \mid H_0]$

$p = P[|t_\nu| \geq |t_{\text{obs}}| \mid H_0]$

Note: $H_0 : \rho = \rho_0$.

- Given the *p*-value, we make a decision depending on the maximum risk (i.e. significance level) we are willing to take of committing a type I error.

# Pearson's correlation test

## Description

- Effect size can be reported by using the correlation index itself. If $r_{XY}$ is lower than 0.3, the effect is weak; if $0.3 < r_{XY} < 0.5$,then it is moderate; and finally, if $r_{XY} \geq 0.5$ the association intensity is large.
- Another possibility is using coefficient of determination: $R^2 = r^2_{XY}$. As for its interpretation, if $R^2 < 0.16$ the effect is weak; if $0.16 < R^2 < 0.26$ it is moderate; and if $R^2 \geq 0.26$ the effect size is large.
- As presented previously, the statistic $z = \frac{1}{2} ln \frac{1+r_{XY}}{1-r_{XY}}$ is normally distributed if conditions are met. Thus, we can obtain a confidence interval for the true correlation parameter:

$$z_l = z - z_{\alpha/2} \frac{1}{\sqrt{n-3}} \text{ and } z_u = z + z_{\alpha/2} \frac{1}{\sqrt{n-3}}.$$

- And to invert *Fisher's transformation* and obtaining thus boundaries in a correlation metric:

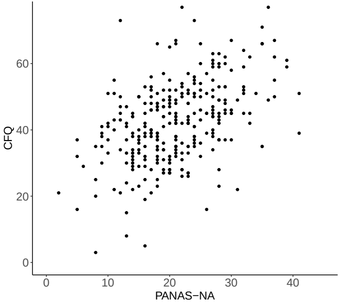
$$r_l = \frac{e^{2z_l} - 1}{e^{2z_l} + 1} \text{ and } r_u = \frac{e^{2z_u} - 1}{e^{2z_u} + 1}.$$

# Pearson's correlation test

## Example

- A researcher wants to demonstrate that negative affect (measured with PANAS-NA subscale) is correlated with cognitive failures during completion of everyday tasks (measured with Cognitive Failures Questionnaire, CFQ). To verify this hypothesis, she selected a random sample of 300 undergraduates. The following scatterplot shows the joint distribution of PANAS-NA and CFQ, as well as some descriptive statistics:

| | PANAS.NA | CFQ |
|---|---|---|
| Mean | 21.17 | 42.54 |
| SD | 7.18 | 12.43 |
| $S_{XY}$ | 41.35 | |

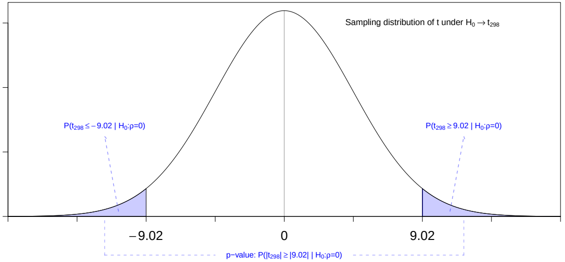


# Pearson's correlation test
## Example



- The null hypothesis in this case is:
$$H_0 : \rho_{XY} = 0.$$
- We assume bivariate normality in the case of the joint distribution of PANAS-NA and CFQ scores.
- The point estimate of the parameter corresponding to the linear correlation in the population is given by:
$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{41.35}{7.18 \times 12.43} = 0.463.$$
- Given that the test in this case is with respect to 0 we can use the *t* distribution:
$$t = \frac{\sqrt{(n-2)}r_{XY}}{\sqrt{1 - r_{XY}^2}} = \frac{\sqrt{298} \times 0.463}{\sqrt{1 - 0.463^2}} = 9.02.$$
- Under the null hypothesis $t \sim t_{298}$. Therefore, the *p*-value can be easily obtained by using a statistical software (e. g., R) as shown: $Prob(|t_{298}| \geq |9.02|) < 0.001$.

- Provided that the probability associated with the test statistic is very low, we reject the null hypothesis. Therefore, given the sample of 300 individuals, negative affect and cognitive failures are significantly related. The effect observed in the sample is moderate.




- NOTE: *p*-value was obtained by using the statistical program R:

```
2*pt(9.02,df=298,lower.tail=FALSE)
```

[1] 2.361091e-17

# Pearson's correlation test

## Example

- Let's estimate the true correlation by using *Fisher's transformation* and a confidence level of 0.95 ($z_{\alpha/2} = z_{0.025} = 1.96$):

$$z = \frac{1}{2} ln \frac{1 + r_{XY}}{1 - r_{XY}} = \frac{1}{2} ln \frac{1 + 0.463}{1 - 0.463} = 0.501.$$

$$z_l = z - z_{\alpha/2} \frac{1}{\sqrt{n-3}} = 0.501 - 1.96 \frac{1}{\sqrt{297}} = 0.387$$

$$z_u = z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} = 0.501 + 1.96 \frac{1}{\sqrt{297}} = 0.615.$$

- By using inverse transformation we obtain lower and upper boundaries of the confidence interval for the true correlation:

$$r_l = \frac{e^{2zl} - 1}{e^{2zl} + 1} = \frac{e^{2 \times 0.387} - 1}{e^{2 \times 0.387} + 1} = 0.369$$

$$r_u = \frac{e^{2zu} - 1}{e^{2zu} + 1} = \frac{e^{2 \times 0.615} - 1}{e^{2 \times 0.615} + 1} = 0.548.$$

- Thus, $CI_{0.95} = [0.369; 0.548]$.

# Pearson's correlation test

## Example

- Pearson's correlation test obtained with built-in function `cor.test()` in R:

```
with(datsim,cor.test(PANAS.NA,CFQ,
       alternative='two.sided'))
```

```
Pearson's product-moment correlation

data:  PANAS.NA and CFQ
t = 9.0148, df = 298, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3689993 0.5474421
sample estimates:
      cor
0.4628976
```

# Pearson's correlation test

## Example

- Pearson's correlation test obtained with jamovi:

Correlation Matrix

| | | PANAS.NA | CFQ |
|---|---|---|---|
| PANAS.NA | Pearson's r | — | |
| | p-value | — | |
| | 95% CI Upper | — | |
| | 95% CI Lower | — | |
| CFQ | Pearson's r | 0.46 | — |
| | p-value | < .001 | — |
| | 95% CI Upper | 0.55 | — |
| | 95% CI Lower | 0.37 | — |

# Spearman's correlation test

## Description

- As previously commented, Spearman's rank correlation coefficient, allows researchers to make a decision regarding the presence of a monotonic relationship between two random variables in the population.
- It corresponds to a non-parametric, distribution-free statistical test.
- As for the measurement scales required by this statistical test, it was originally thought for continuous variables measured either by means of an interval or a ratio scale.
- Given that it is based on rank orderings, it can also be used for ordinal scales.
- As it occurs with other non-parametric tests, when ties are present in the sample some sort of correction needs to applied in order to reduce the possible bias in the test.
- General hypothesis test is as follows:

$$H_0 : F(X, Y) = F(X)F(Y).$$

- The previous hypothesis establishes that both variables are independent.
- Sometimes you'll see the following hypothesis as $H_O : \rho_s = 0$ but it is unaccurate and favors misinterpretations regarding the test actually carried out.

# Spearman's correlation test

## Description

- Two possibilities in order to estimate $p$-value associated with the test statistic:
  - An exact test, based on a permutation procedure, is available in case that in the sample there are not observed ties.
  - An approximation, using $t$ distribution is also adequate for samples with $n \geq 10$:

$$t = \frac{\sqrt{(n-2)} r_s}{\sqrt{1 - r_s^2}} \sim t_{n-2}.$$

- NOTE: R also reports the $S$ statistic which is a transformation of the rank correlation coefficient. This statistic is used in those cases that the exact $p$-value can be obtained:

$$S = (n^3 - n)\frac{1 - r_s}{6}.$$

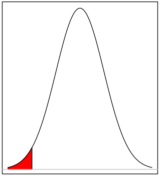
# Spearman's correlation test

## Description



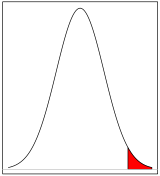
- A statistical decision can be made after obtaining the probability associated with the test statistic, that is, its $p$-value. Several expressions to get this probability depending on the statistical hypothesis:

Left-tailed test



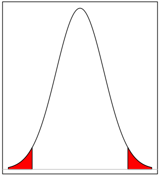
$p = P[t_\nu \leq t_{\text{obs}} \mid H_0]$

Right-tailed test



$p = P[t_\nu \geq t_{\text{obs}} \mid H_0]$

Two-tailed test

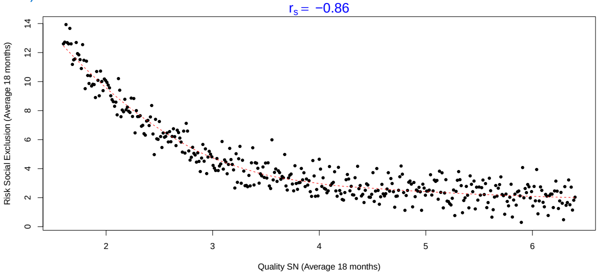


$p = P[|t_\nu| \geq |t_{\text{obs}}| \mid H_0]$

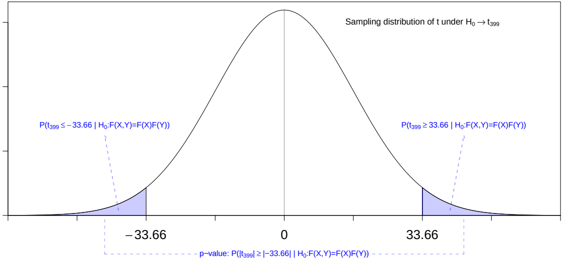
Note: $H_0 : F(X, Y) = F(X)F(Y)$.

- Given the $p$-value, we make a decision depending on the maximum risk of committing a type I error (i.e. significance level) we are willing to take.
- As for the effect size, Spearman's correlation can be used and intepreted as commented in the Pearson's correlation section.

- In a study on the effect of the quality of personal social networks (measured by an ordinal scale from 0 to 10, a higher score would indicate social networks of high quality) over the risk of social exclusion (measured with an interval scale from 0 to 30) during adolescence, 401 teenagers were assessed during a year and a half. The following scatterplot shows average monthly measures of the two mentioned variables. Given the pattern in the relationship between quality of social networks and risk of social exclusion, it was decided to use Spearman's correlation (which is shown in the plot title):



- The null hypothesis in this case establishes that quality of networks and social exclusion risk are independent in the population from which the sample was drawn:
$$H_0 : F(XY) = F(X)F(Y).$$
- Let's compute now the test statistic, $t$:
$$t = \frac{\sqrt{(n-2)}r_s}{\sqrt{1-r_s^2}} = \frac{\sqrt{399} \times (-0.86)}{\sqrt{1-0.86^2}} = -33.66.$$
- Under the null hypothesis $t \sim t_{\nu=399}$. Therefore, the $p$-value can be easily obtained by using a statistical software:
$Prob(|t_{399}| \geq |-33.66|) < 0.001.$
- Provided that the probability associated with the test statistic is very low, we reject the null hypothesis. Therefore, given the sample of 401 individuals, quality of social networks and risk of social exclussion are significantly related. Besides, the observed effect in the sample is strong.




- NOTE: *p*-value was obtained by using a built-in function in R:

```
2*pt(-33.66,df=399,lower.tail=TRUE)
```

[1] 1.263378e-118

# Spearman's correlation test

## Example



- Spearman's correlation test obtained with built-in function `cor.test()` of R:

```
with(SNdata,cor.test(quality,risk,alternative='two.sided',
                     method='spearman'))
Spearman's rank correlation rho

data:  quality and risk
S = 20030998, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8639035
```

# Spearman's correlation test

## Example

- Spearman's correlation test obtained with jamovi:

Correlation Matrix

| | | qualitySN | riskSE |
|---|---|---|---|
| qualitySN | Spearman's rho | — | |
| | p-value | — | |
| riskSE | Spearman's rho | -0.86 | — |
| | p-value | < .001 | — |

# Kendall's correlation test

## Description



- Kendall's correlation test is the statistical procedure of choice when having ordinal scales and, specially, when restriction of observed ranges is present in the sample.
- General null hypothesis establishes that the two variables are independent: $H_O : \tau = 0$. NOTE: Remember that $\tau$ is associated with the difference between number of concordances and discordances (C-D).
- There are different alternatives in order to get the *p*-value associated with observed point estimate of $\tau$:
  - Exact test, using permutations, can be used when no ties are present in the sample.
  - Normal approximation, when $n \geq 10$, can be used:

$$z = \frac{S - \mu_S}{\sigma_S} \sim N(0, 1), \text{ where } S = C - D;$$

$$\mu_S = 0 \text{ and } \sigma_S = \frac{1}{3}\sqrt{\frac{2(2n+5)}{n(n-1)}}.$$

- Previous formula for $\sigma_S$ is adequate when there are no ties in the sample.

# Kendall's correlation test

## Description

- In case of observing tied ranks, we need to correct standard error of the statistic as follows:

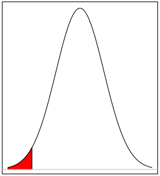
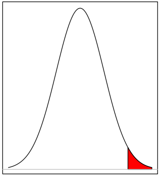
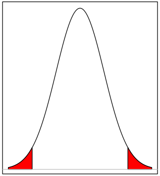
$$\sigma_S = \left( \frac{1}{18} \left[ n(n-1)(2n+5) - \sum_t t(t-1)(2t+5) - \sum_u u(u-1)(2u+5) \right] + \frac{1}{9n(n-1)(n-2)} \left[ \sum_t t(t-1)(t-2) \sum_u u(u-1)(u-2) \right] + \frac{1}{2n(n-1)} \left[ \sum_t t(t-1) \sum_u u(u-1) \right] \right)^{\frac{1}{2}}$$

  where $t$ are the tied observations in the first variable and $u$ are the ties in the second variable.
- Again, Kendall's correlation can also be used as an effect size indicator.

# Kendall's correlation test

## Description

- A statistical decision can be made after obtaining the probability associated with the test statistic, that is, its *p*-value. There are several expressions to get this probability depending on the statistical hypothesis:

| Left-tailed test | Right-tailed test | Two-tailed test |
|---|---|---|
|  |  |  |
| $p = P[Z \leq z_{\text{obs}} \mid H_0]$ | $p = P[Z \geq z_{\text{obs}} \mid H_0]$ | $p = P[\lvert Z \rvert \geq \lvert z_{\text{obs}} \rvert \mid H_0]$ |

Note: $H_0 : \tau = 0$.

- Given the *p*-value, we make a decision depending on the maximum risk (i.e. significance level) we are willing to take of committing a type I error.

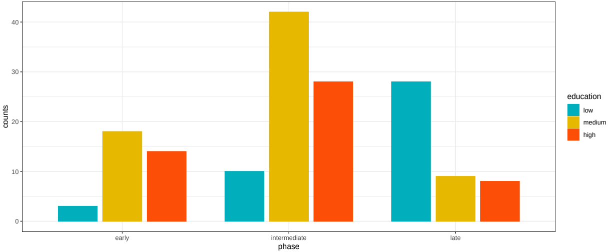
# Kendall's correlation test

## Example



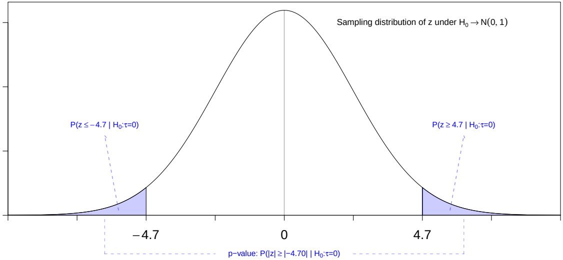
- A group of psychologists wanted to know whether education level (measured with an ordered factor: low, medium, high) is associated with the phase in which people diagnosed with a mental problem go to a health service to solve this problem for the very first time (ordered factor with levels early, intermediate, and late).

| | early | intermediate | late |
|---|---|---|---|
| low | 3 | 10 | 28 |
| medium | 18 | 42 | 9 |
| high | 14 | 28 | 8 |



- The null hypothesis in this case is:
$$H_0 : \tau = 0$$
- That is to say, education and phase are independent in the population.
- By using, a statistical software we know that in this sample of 160 individuals $\hat{\tau}_b = -0.33$, and $S = C - D = 1415 - 4142 = -2727$. Therefore, seems that there is a negative relationship between the two variables.
- In order to make a decision, and given that $n = 160$, we will use the normal approximation:
$$z = \frac{S - \mu_S}{\sigma_S} \sim N(0, 1), \text{ where}$$
$$\mu_S = 0 \text{ and } \sigma_S = 580.172.$$
- The standard error in this case was computed by using correction for ties (computation's details not shown here).
- Thus, $z = \frac{-2727}{580.172} = -4.7$.
- $p$-value is $Prob(|Z| \geq |-4.70|) < 0.001$. So, we can reject the null hypothesis.
- The effect size in this case is moderate.

- Provided that the probability associated with the test statistic is low, we can reject the null hypothesis. Therefore, there is empirical support in favour of the relationship between educational level and phase that a mental problem is detected by health professionals. The effect size observed in this sample is moderate.




- NOTE: *p*-value was obtained by using a built-in function in R:

```
2*pnorm(-4.70,mean=0,sd=1,lower.tail=TRUE)
```

```
[1] 2.601615e-06
```

# Kendall's correlation test
## Example



- Kendall's correlation test obtained with built-in function `cor.test()` in R:

```
with(datKendall,cor.test(as.numeric(education),
                         as.numeric(phase),
                         alternative='two.sided',
                         method='kendall'))
Kendall's rank correlation tau

data:  as.numeric(education) and as.numeric(phase)
z = -4.7003, p-value = 2.597e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
-0.334598
```

# Kendall's correlation test

## Example

- Kendall's correlation test obtained with jamovi:

Correlation Matrix

| | | phase | education |
|---|---|---|---|
| phase | Kendall's Tau B | — | |
| | p-value | — | |
| education | Kendall's Tau B | -0.33 | — |
| | p-value | < .001 | — |

Slides made with `knitr` and LaTeX