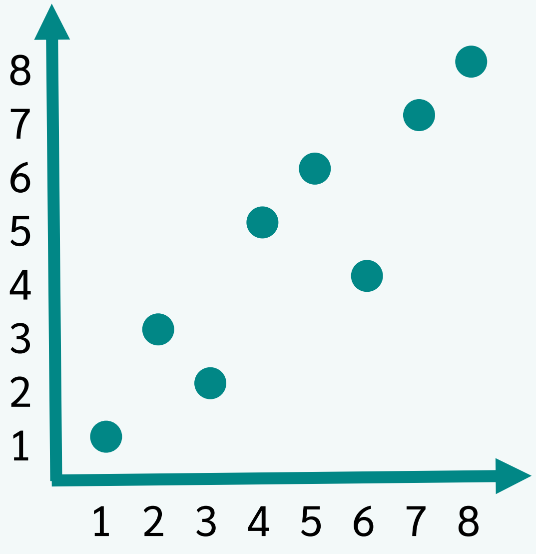
SPEARMAN CORRELATION

Playbook

Theory & Example

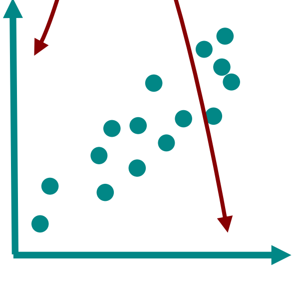


Author: Dr. Mathias Jesussek

DATAtab

What is a **Spearman correlation?**

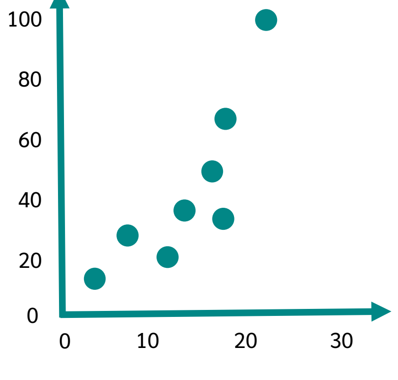
Spearman's rank correlation examines the **relationship** between two variables.

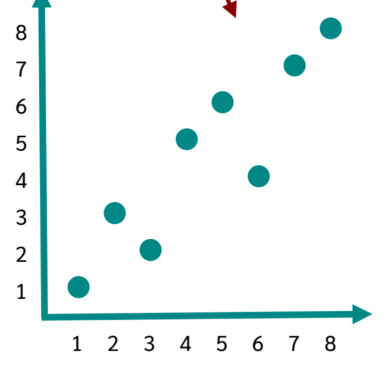


The Spearman rank correlation is the non-parametric counterpart of the Pearson correlation.

But there is an **important difference** between both correlation coefficients!

Spearman correlation does **not use the raw data,** but the **ranks** of the data.





DATAtab

Example:

We measured the **reaction time** of **8 computer players** and asked their **age**.

12	14
15	25
17	20
18	35
20	45
21	30
22	60
26	95

When we calculate a **Pearson correlation**, we simply take the two variables **reaction time** and **age** and calculate the **Pearson correlation coefficient**.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

DATAtab

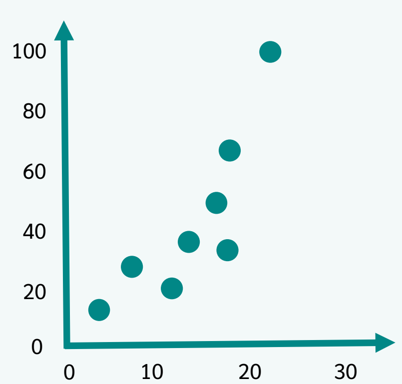
However, we now want to calculate the **Spearman rank correlation**, so first we assign a **rank** to each person for **reaction time** and **age**.

12	1	14	1
15	2	25	3
17	3	20	2
18	4	35	5
20	5	45	6
21	6	30	4
22	7	60	7
26	8	95	8

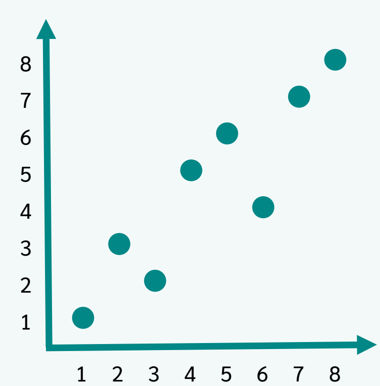
The **reaction time** is already sorted by size. **12** is the smallest value, so gets **rank 1**; **15** is the second smallest, so gets **rank 2** and so on and so forth.

We are now doing the same with **age**. **14** is the smallest value so gets **rank 1**; **25** is the third smallest and gets **rank 3**; **20** is the second smallest an gets **rank 2**, and so on and so forth

Let's take a look at this in a **scatter plot**. Here we see the raw data of **age** and **reaction time:**

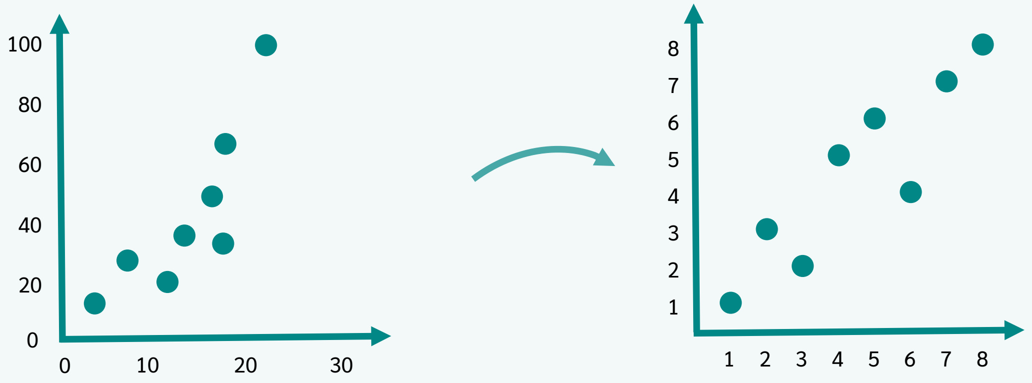


But now we would like to use the **rankings**. So, we **form ranks** from the variables **age** and **reaction time:**



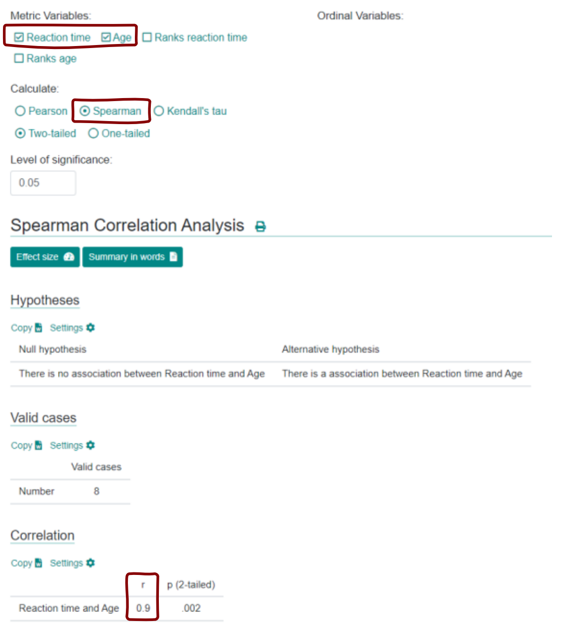
Through this **transformation**, we have now distributed the data more evenly.

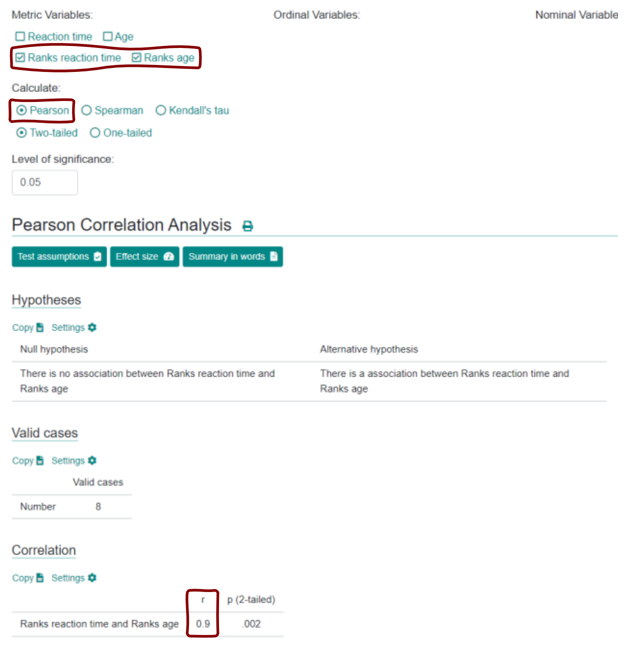
To calculate the **Spearman correlation**, we simply calculate the **Pearson correlation** from the **ranks**.



So, the **Spearman correlation** is equal to the **Pearson correlation**, only that the **ranks** are used instead of the **raw values.**

Have a quick look at this in **DATA***tab*





Spearman Correlation

Pearson Correlation

Now we can either calculate the **Pearson correlation** of the reaction time and the age where we get a **correlation of 0.9**.

Or we can calculate the **Spearman correlation from the ranks,** there we also get **0.9** .

So, exactly the same as before.

DATAtab

! If there are **no rank ties**, we can also use this **equation** to calculate the **Spearman correlation**.

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

d is the **difference in ranks** between the two variables.

n is the **number of cases**

Referring to our example:

we get the different **d's** with this:
1-1 = 0,
2-3 = -1,
3-2=1, and so on.

Now we **square the individual d's** and **add** them all up.

					d	d^2
	12	1	14	1	1-1 = 0	0
	15	2	25	3	2-3 = -1	1
	17	3	20	2	3-2 = 1	1
	18	4	35	5	4-5 = -1	1
	20	5	45	6	5-6 = -1	1
	21	6	30	4	6-4 = 2	4
	22	7	60	7	7-7 = 0	0
	26	8	95	8	8-8 = 0	0
						$\sum 8$

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

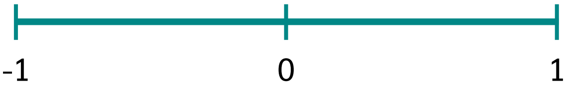
n, which is the **number of people**, is **8**.

So, the sum of d_i^2 is **8**.

If we put everything in, we get a **correlation coefficient of 0.9**.

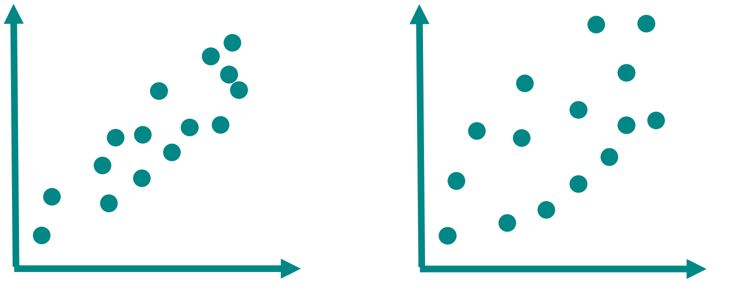
$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{48}{504} = 0.90$$

Just like the Pearson correlation coefficient r the **Spearman correlation coefficient r_s** also varies between **-1** and **1**.

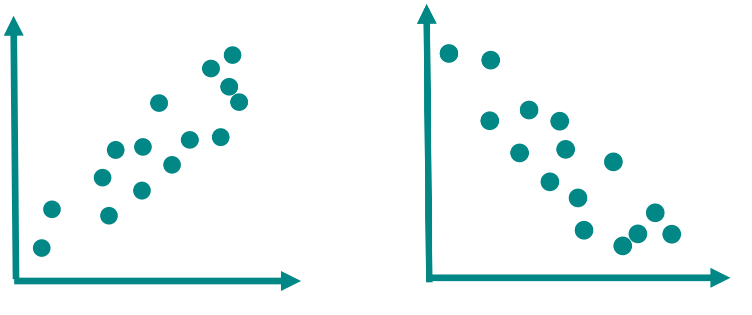


With the help of the **coefficient,** we can now **determine two things**.

1. How **strong** the **correlation** is



2. and in **which direction** the **correlation goes**.



DATAtab

!

The **strength of the correlation**, can be read in a **table**.

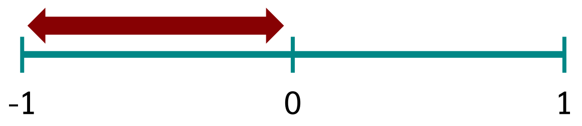
Amount of r	Strength of the correlation
0.0 < 0.1	no correlation
0.1 < 0.3	low correlation
0.3 < 0.5	medium correlation
0.5 < 0.7	high correlation
0.7 < 1	very high correlation

From Kuckartz et al.: Statistik, Eine verständliche Einführung, 2013, p. 213

If **r** is between **0 and 0.1**, we speak of **no correlation**.

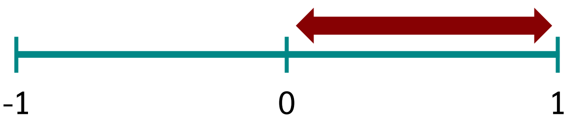
If **r** is between **0.7 and 1**, we speak of **very strong correlation**.

If we have a coefficient **between -1 and less than 0**, there is a **negative correlation**,



thus a **negative relationship** between the variables.

If we have a coefficient **between greater than 0 and 1**, there is a **positive correlation**,



thus a **positive relationship** between the variables.

! If the result is **0**, we have **no correlation**.

The **correlation coefficient** is usually calculated with data taken from a **sample**.

However, we often want to test a **hypothesis** about the **population**.

Population

Sample

In the case of **correlation analysis**, we then want to know if there is a **correlation** in the **population**.

For this we check, whether the **correlation coefficient** in the **sample** is statistically significantly **different from zero**.

DATAtab

Null hypothesis

H0 The correlation coefficient **does not differ significantly** from **zero**. → There **is no** linear relationship.

$$r = 0$$

Alternative hypothesis

H1 The correlation coefficient **differs significantly** from **zero**. → There **is a** linear relationship.

$$r \neq 0$$

Whether the correlation coefficient is significantly different from zero based on the **sample** collected

Sample

can be checked using a **t-test**.

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Where **r** is the **correlation coefficient**

and **n** is the **sample size.**

A **p-value** can then be calculated from the **test statistic t**.

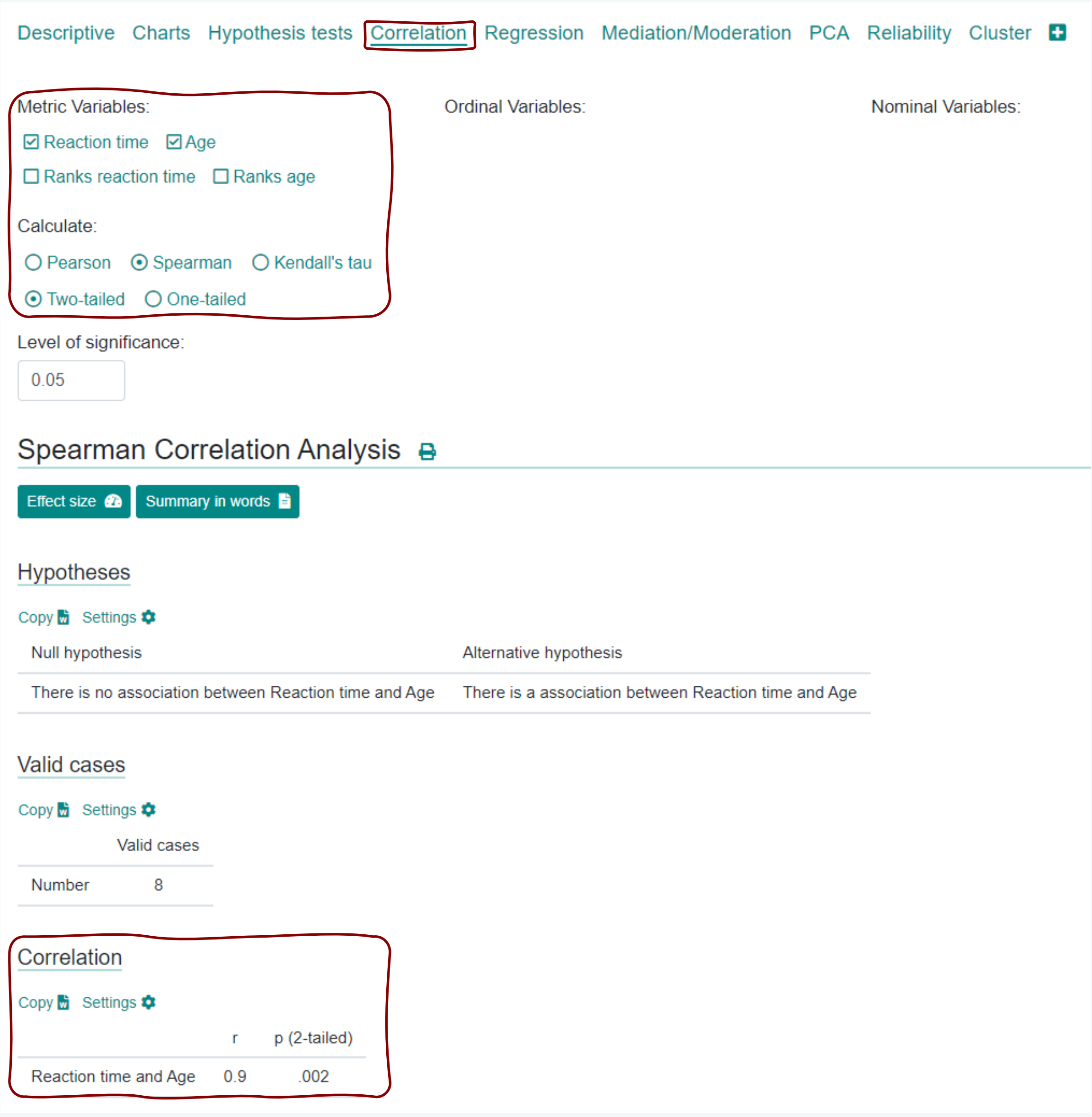
If the **p-value** is less than the specified **significance level**, which is **usually 5%**, then the **null hypothesis** is rejected, otherwise it is not.

DATA*tab*

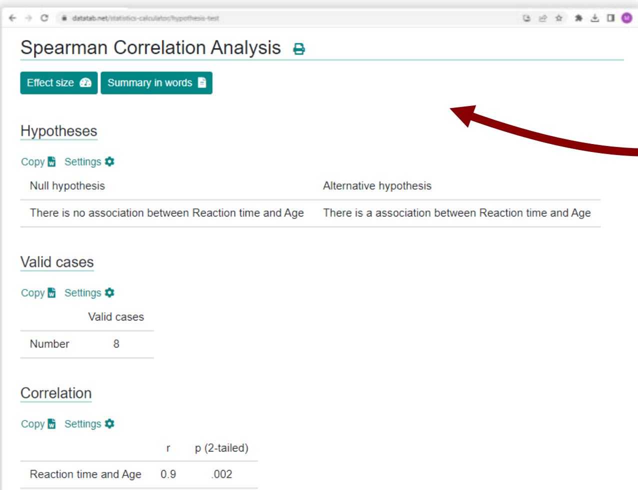
Referring to our example:

If we use **DATA*tab*** for the calculation of the example, we get a **p-value of 0.002**.

The **p-value** is therefore **smaller than 0.05** and we can therefore **reject the null hypothesis** that in the population the correlation coefficient is zero.



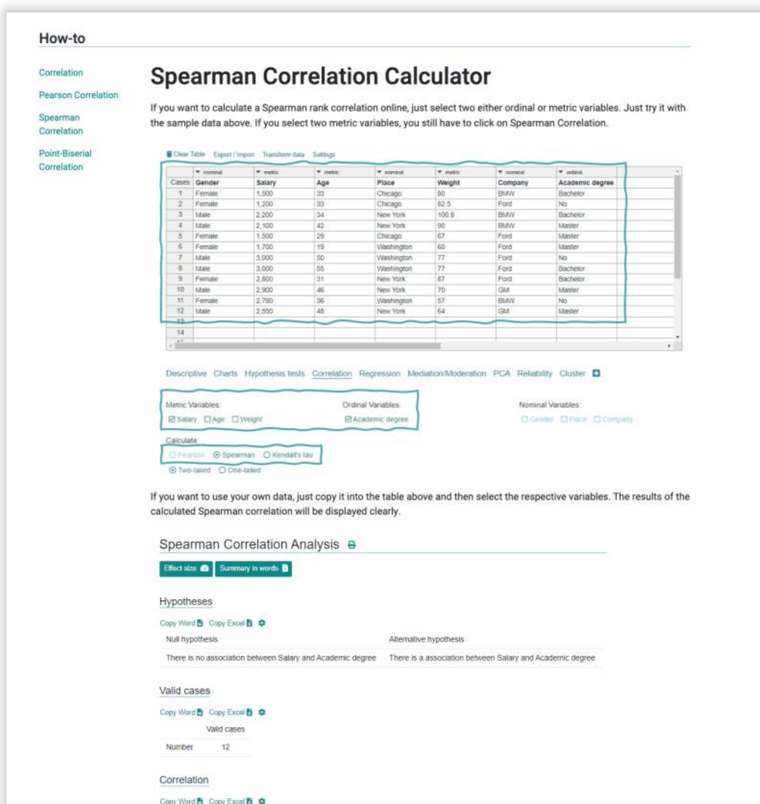
How do we calculate a **Spearman correlation** with **DATA***tab*?



If you like, you also can calculate the **Spearman correlation** with **DATA***tab*

GO TO DATA*tab*

To get a correlation analysis of our data, just copy your data into this **table** on **DATA***tab* and click on either the Hypotheses or Correlation tab.



More about

SPEARMAN CORRELATION

on our website **datatab.net**

Feel free to share!