Data Analytics in R
Session 9

Maria Kunevich

Paper summary and work on your project

- Feedback on paper selection and project work: https://miro.com/welcomeonboard/SUw3RHpXWjBpUDF2V0dwTWhkOFVVUnlUTnZ4Qm1oVGtSTVN0SmNCUmJyODhydU9hUzA3VUpNZVZHRnNBenhqVHwzMDc0NDU3MzYzNjI1MDIzMjY2fDI=?share_link_id=164034942325

Homework on DataCamp

- Please send me a screenshot of your completed assignments **by the end of the day today** (I'll assign points based on the screenshots I get today)
- Thank you!

Assignments deadlines

Assignment	Date of assignment	Deadline (midnight 23:59)
HW1	22 Sept 2022	28 Sept 2022
HW2	29 Sept 2022	5 Oct 2022
HW3	6 Oct 2022	12 Oct 2022
HW4	13 Oct 2022	19 Oct 2022
HW5	20 Oct 2022	2 Nov 2022
Paper summary	20 Oct 2022	20 Nov 2022
HW6	3 Nov 2022	9 Nov 2022
HW7	10 Nov 2022	16 Nov 2022
HW8	17 Nov 2022	23 Nov 2022
HW9	24 Nov 2022	30 Nov 2022
HW10	1 Dec 2022	7 Dec 2022
Project	TBA	14 Dec 2022
Final Presentations		15 Dec 2022

Nov 17 - interim report

Quick revision

- What is the main difference between descriptive and inferential statistics?
- What is the difference between a population and a sample?
- What is a good sample?
- What is a statistic and what is a parameter? What is their relationship?
- What is sampling error?
- What are the different data types and why this difference is important?
- What are the main categories of measures for descriptive statistics?
- What are parametric and nonparametric statistics?

Covariance

- **Variance** - measure of dispersion, i.e. the degree of spread in the data set, measuring how far values are spread from the mean. Variance tells you how a single variable varies.
- **Covariance** measures how the two variables change in relation to each other. The higher this value, the more dependent the relationship is.
    - A **positive** number refers to positive covariance with a direct connection, which means that an increase in one variable would also lead to a corresponding increase in the other variable
    - A **negative** number refers to negative covariance with an inverse relationship between the two variables.

Covariance and correlation

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

```
> cov(salary, expenses)
[1] 95000
>
```

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Correlation: 0.8969985 -> 0.9

names	salary	expenses
Mark	2000	1800
Julia	1650	1500
David	2300	1800
Rose	1700	1700
Rick	2100	1900
Camilla	3100	2100

Correlation

- Correlation means there is a statistical association between variables, i.e. when one variable changes, there is a similar change in the other variable.
- The variables change together: they covary
- Correlation is described in two terms that have a statistical significance: the **strength** and the **direction** of the connection
    - **Strength**: signifies the relationship correlation between two variables, i.e how consistently one variable will change due to the change in the other
    - **Direction**: a positive linear or negative linear relationship between variables

Correlation coefficient

A correlation coefficient is **a number between -1 and 1** that tells you the **strength** and **direction** of a relationship between variables

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.

Interpreting a correlation coefficient

General indicator of the **strength** of the relationship:

- the value of the correlation coefficient ranges between **1** and **-1**
- the correlation coefficient closer to **1** and **-1** indicate a strong relationship, this happens when the data points fall on or are very close to the line of best fit
- When data points are further away from the line, the strength of linear relationship becomes weaker
- When we can’t draw a line because the data points are scattered, the strength of the linear relationship is the weakest

Interpreting a correlation coefficient

General indicator of the **direction** of the relationship:

- the sign of the coefficient reflects **the direction of change**: a positive value means the variables change together in the same direction, while a negative value means they change together in opposite directions.
- the line of best fit has **an upward slope** -> a positive linear relationship
    - This means an increase in the value of one variable will lead to an increase in the value of the other variable
- the line of best fit has a downward slope -> a negative linear relationship
    - This means an increase in the amount of one variable leads to a decrease in the value of another variable

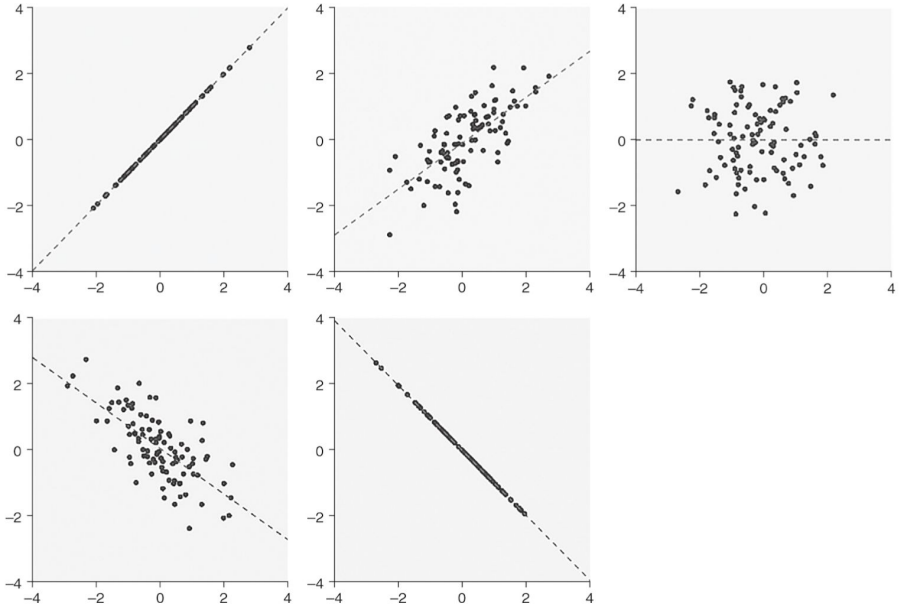
Interpreting a correlation coefficient

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Correlation visualisation

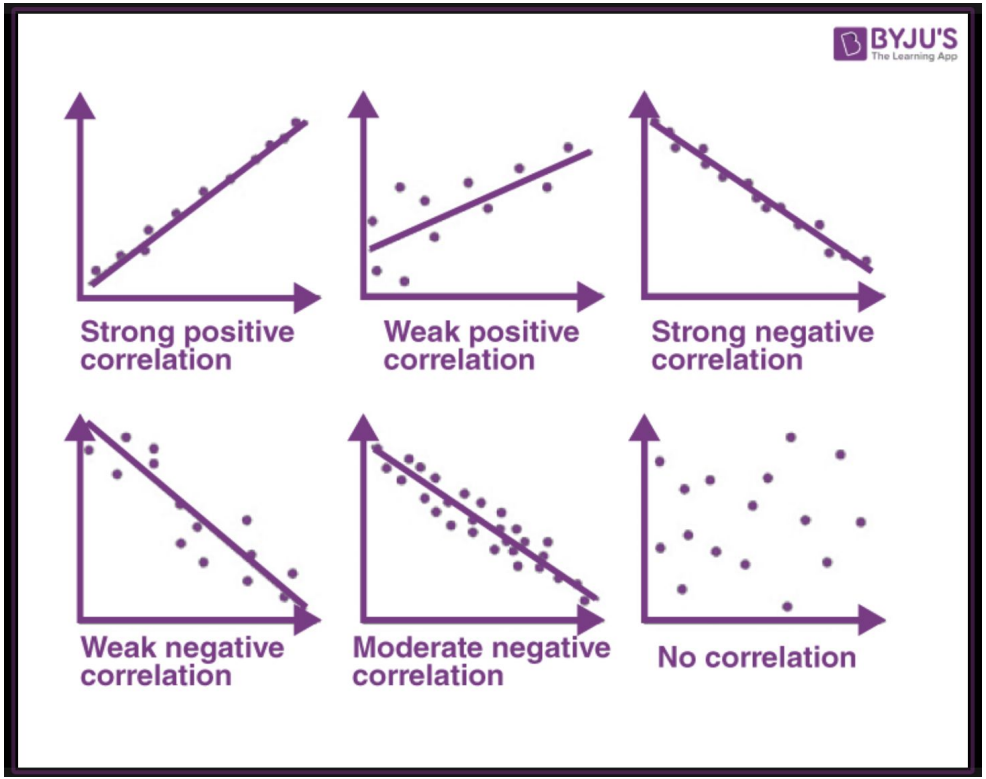
A scatter plot

- to understand relationship between two variables
- to identify trends
- to identify outliers



http://methods.sagepub.com/Reference//the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i15659.xml

Types of Correlation



Guess the correlation

A nice website to play with guessing the correlation for up to two decimal points

http://guessthecorrelation.com

What does a correlation tell you?

Summarising data

- A correlation coefficient is a **descriptive statistic** that summarises sample data but does not allow you to infer anything about the population.
- To generalise your results to the population, you will need a statistical test (an *F* test or a *t* test to calculate a test statistic that will tell you the statistical significance of your finding

Comparing studies

- A correlation coefficient also allows you to measure an effect size, which tells you the practical significance of a result
- Correlation coefficients can be compared directly between studies

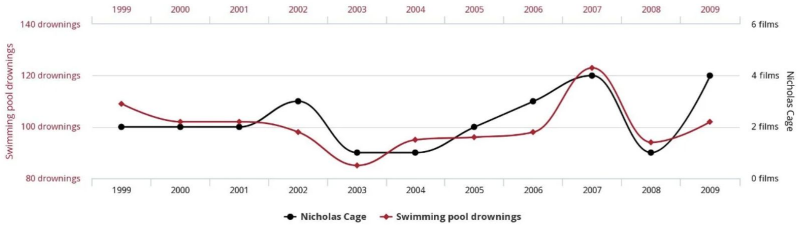
Source: https://www.scribbr.com/statistics/correlation-coefficient/

Correlation does NOT imply causation

- a lot of memes on the internet
- spurious correlations between random facts

A **spurious correlation** is when two variables appear to be related through hidden third variables or simply by coincidence

- https://tylervigen.com/spurious-correlations



Types of correlation coefficients

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Source: https://www.scribbr.com/statistics/correlation-coefficient/

Pearson’s r

The **assumptions** for the data that must be met in order to use Pearson’s *r*:

- Both variables are on an interval or ratio level of measurement, i.e. **continuous**
- Data from both variables follow **normal distributions**
- Your data have **no outliers**
- Your data is from **a random** or representative **sample**
- You expect **a linear relationship** between the two variables (the scatterplot to check)

The Pearson’s r is **a parametric test**, so it has high power. But it’s not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure

Source: https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Spearman’s rho

Spearman’s rank correlation coefficient - the most common alternative to Pearson’s *r* and it uses **the rankings of data** (from lowest to highest)

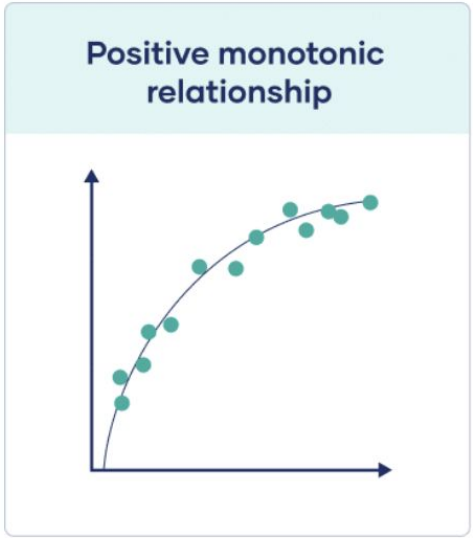
When the assumptions for Pearson’s *r* statistic are not met, you should use Spearman’s *rho* (ordinal variables)

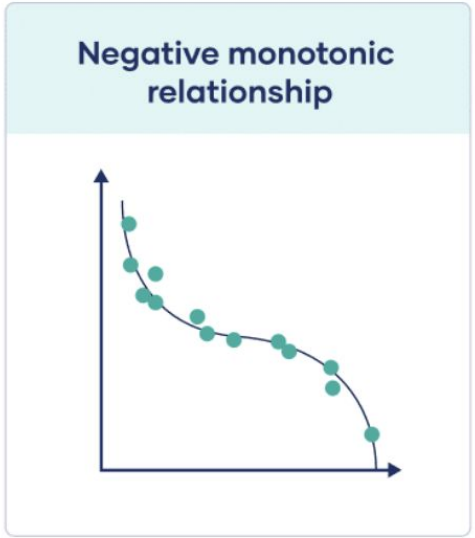
The Spearman correlation coefficient measures the **monotonicity** of relationships, i.e. each variable always changes in only one direction but not necessarily at the same rate.

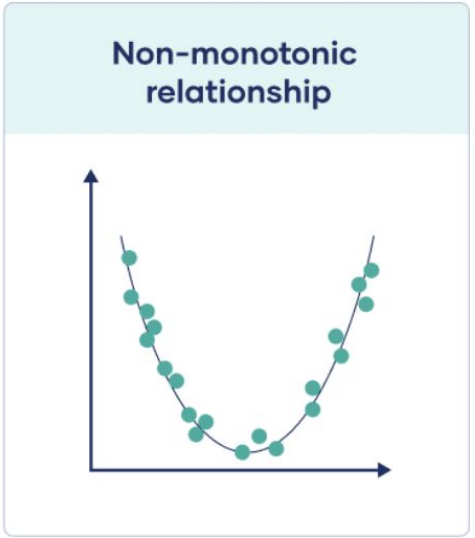
- Positive monotonic: when one variable increases, the other also increases.
- Negative monotonic: when one variable increases, the other decreases.

Source: https://www.scribbr.com/statistics/correlation-coefficient/

Spearman’s rho







Source: https://www.scribbr.com/statistics/correlation-coefficient/

Correlations in R

Syntax:

cor(x, y, method = “pearson”)

cor.test(x, y, method = “pearson”)

Parameters: x, y: numeric vectors **with the same length**

method: correlation method, the default method is “pearson”

Let’s practise in RStudio!