

Parametric Statistics

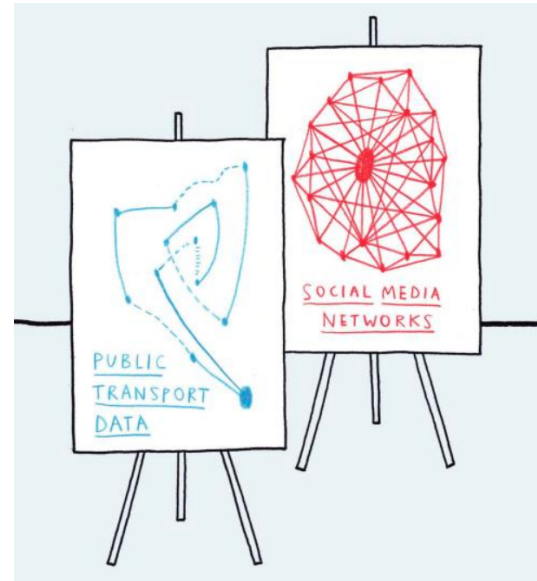
Week 5 - Correlation

Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 19. November 2019

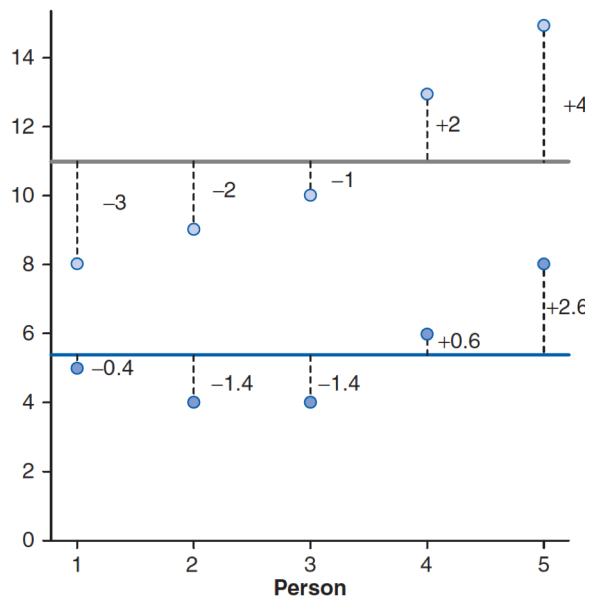
political
data
science
<https://politicaldatascience.blogspot.de>



Comparing Relationships

No more comparing the means between groups...for now... Now, we compare the *relationship* between variables:
How does the value of one variable changes when the value of another variable changes.

We expect that when one variable deviates from its mean, the other variable deviates from the mean in a similar way:



Covariance

We can check this by looking at the **covariance** of the two variables

Remember Variance:

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{N - 1} = \frac{\sum (x_i - \bar{X})(x_i - \bar{X})}{N - 1}$$

Now we define covariance:

$$\text{cov}(X, Y) = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N - 1}$$

Covariance can help us to see if two variables are related to each other: The higher covariance, the more similar variance between variables BUT...

The variables may have different units, they are NOT standardized.

Standardization

Converts variables into a standard set of units

Remember the Z-scores?

$$Z = \frac{x - \bar{X}}{s}$$

It turned any normal distribution into a standard normal distribution with mean 0 and standard deviation 1.
We can use the same standardization with any data, no matter how the distribution is.

In data analysis subtracting the mean and dividing by the standard deviation is called **normalization**. Very important data pre-processing step.

Pearson Correlation Coefficient

Time to standardize the covariance!

What is the covariance missing? Division by standard deviation!

$$r = \frac{\text{cov}(X, Y)}{s_X * s_Y} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{(N - 1) * s_X * s_Y}$$

This is the famous **Pearson correlation coefficient**.

The value lies between -1 and 1

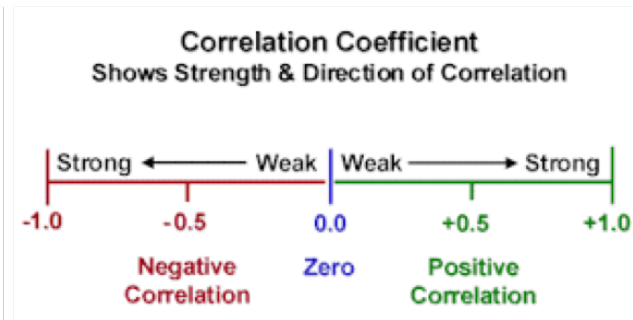
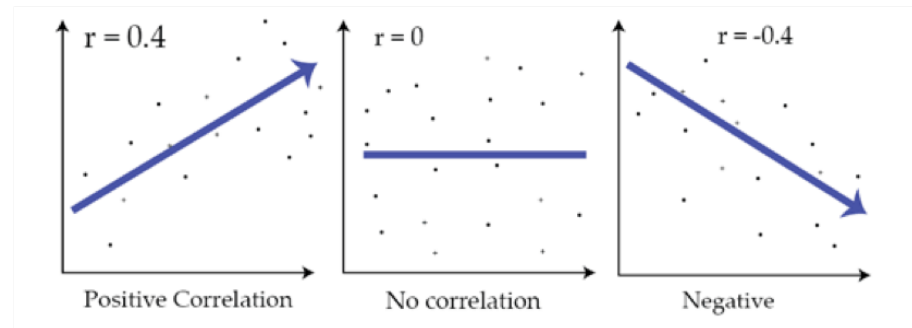
Pearson Correlation Coefficient

Positive correlation: Variables tend to move in the same direction when they change

Negative correlation: Variables tend to move in the opposite direction when they change

Negative does not mean that it is not correlated or something bad... it tells us that the effect is contrary between variables

Use **scatter plots** to visualize the correlation:



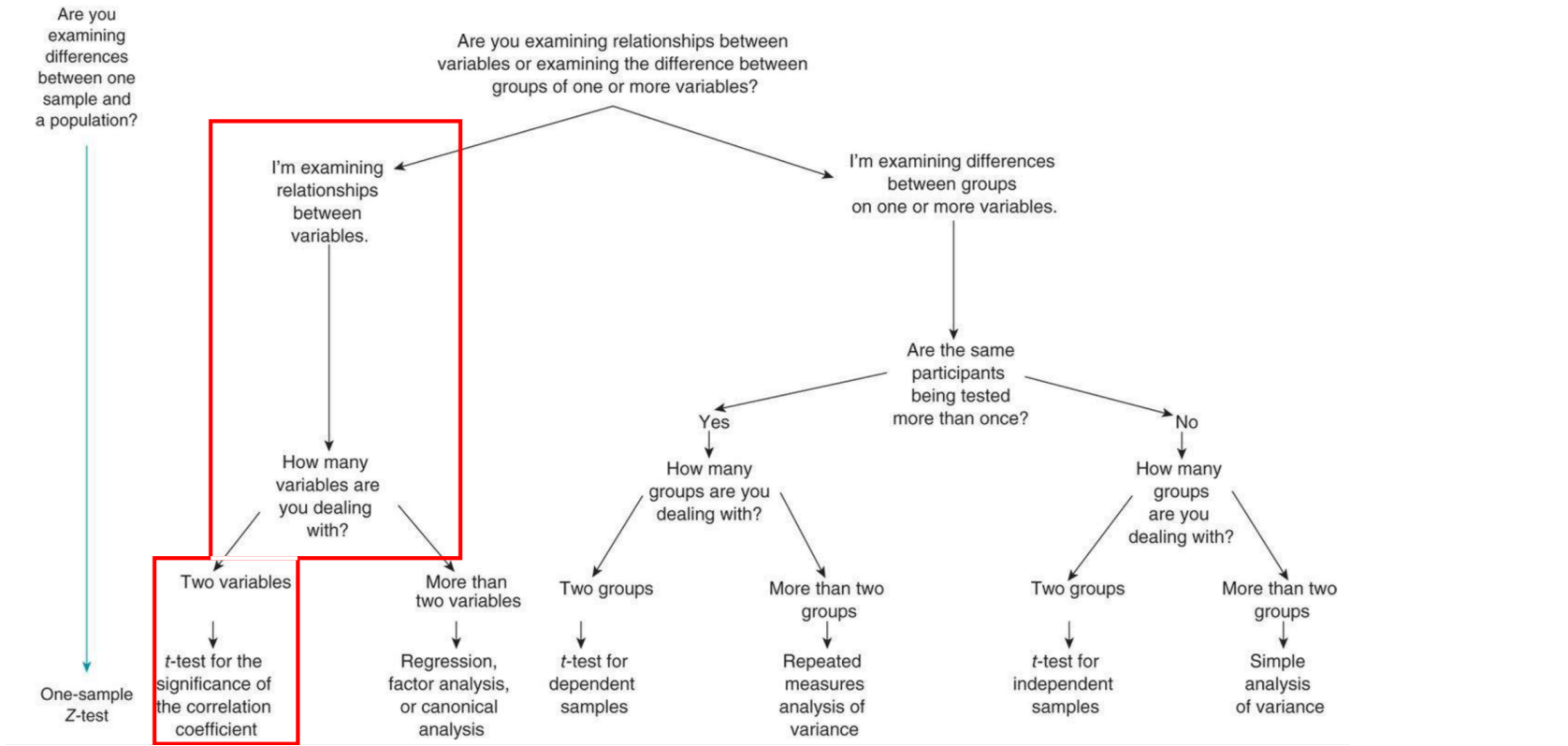
Question: Is the correlation coefficient statically significant?

Significance Tests

Types of significance tests:

- ▶ Comparing means
- ▶ **Asses correlation**
- ▶ Check assumptions

Significance Tests



t-tests for Correlation

Usage

Finding significance for the Pearson correlation coefficient

Assumptions

- Distribution of the variables is normally distributed.

t-tests for Correlation

Similar to a simple two-sided t-test, but with the following t statistic:

$$t = \frac{\text{effect}}{\text{error}} = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

Like a simple t-test, degrees of freedom is $N - 1$ and the hypotheses are:





$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

What does $1 - r^2$ in the t statistic formula mean?

Coefficient of determination

r^2 (also known as R^2) is the **coefficient of determination**: Percentage of variance in one variable that is accounted by the variance of the other variable. (shared variance)

Correlation	Coefficient of Determination	Variable X	Variable Y
$r_{xy} = 0$	$r^2_{xy} = 0$		0% shared 
$r_{xy} = .5$	$r^2_{xy} = .25$ or 25%		25% shared
$r_{xy} = .9$	$r^2_{xy} = .81$ or 81%		81% shared

$1 - r^2$ is the variance that is not shared or explained between the variables, and is part of the error in the t test statistic

Confidence Intervals

Every statistical test includes confidence intervals. Remember when we wanted to estimate the mean and we calculated the 95% confidence intervals?

When comparing the means, the confidence intervals are not used as much. However, for tests of relationship (any kind, not only Pearson's r) confidence intervals are used:

If the confidence interval for a given relationship coefficient(for example r) includes zero, then there is a high high chance that the coefficient IS zero. So the test fails to reject the null hypothesis.

This will be important again when we tackle linear regression

Effect Size

What about the effect size?

The correlation coefficients ARE effect sizes! No more calculations.

In week 3 we already used Pearson correlation coefficient as a standardized effect size to find out how many subjects are needed in an experiment. (Go check that slide again)

Correlation Matrix

What happens when we have more than one variable?

Use a **correlation matrix**

Example

	Income	Education	Attitude	Vote
Income	1.00	.574	-.08	-.291
Education	.574	1.00	-.149	-.199
Attitude	-.08	-.149	1.00	-.169
Vote	-.291	-.199	-.169	1.00

Partial Correlation

The relation of two variables is explored, but the impact of a third variable is removed from the relationship.

Statistical expression: *"We **control** for a confounding variable"*

A significant correlation may become not significant if the confounding variable explains most of the variance.

Next step would be to calculate the corrected correlation between variables and an appropriate significance t-test ...
However, the math here is advanced and we let R do the heavy lifting

Partial Correlation

Example In Munich, police found out that the consumption of ice is correlated to the crime rate! Should we ban ice cream in Bayern?

	Consumption of Ice Cream	Crime Rate
Consumption of Ice Cream	1.00	.743
Crime Rate		1.00

Partial Correlation

Example continued However, the smart police woman Judy Hopps also measured the correlation with the weather temperature and found out this:

	Consumption of Ice Cream	Crime Rate	Average Outside Temperature
Consumption of Ice Cream	1.00	.743	.704
Crime Rate		1.00	.655
Average Outside Temperature			1.00

After controlling for weather temperature, the correlation between ice cream consumption and crime rate reduces to 0.52 and the significance test shows a p-value of .147. Significance is gone!

Good job Hopps!

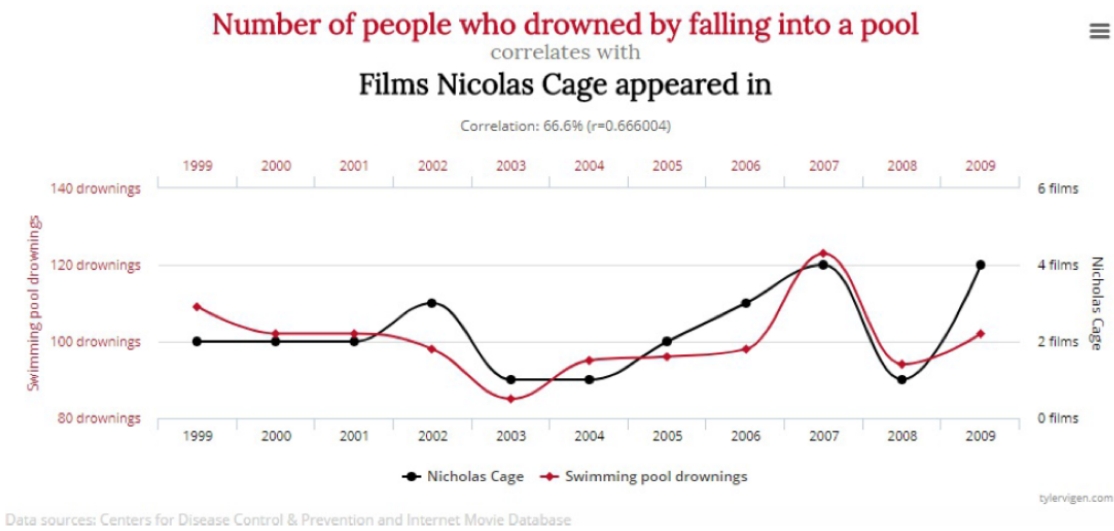
A Word of Caution

Correlation does not imply causation!

Even if there were no confounding variables at all, the correlation coefficients do not indicate in which direction the causality exists.

Important: Try to collect sufficiently diverse data. If we restrict the variables to a certain range, the correlation will be larger as when considering all the ranges the variables can have.

A Word of Caution



Literature

- ▶ Statistics for people who hate statistics: Chapters 5 and 15
- ▶ Discovering Statistics using R: Chapter 6