Data Analysis in Sociology

Lecture 5. Correlation

March 2023

This lecture

1. Correlation among other tests
2. How to read a correlation
3. Correlation coefficients
4. Assumptions
5. Common issues

Correlation among other bivariate tests

	Binary	Categorical (3+ cats)	Numeric
Binary	Chi-square test (Yates correction)	Chi-square test	T-test MW U-test
Categorical (3+ cats)	Chi-square test	Chi-square test	Anova K-W test
Numeric	T-test MW U-test	Anova K-W test	**Correlation**

Disclaimer: there are many other tests of association

1. Correlation among other bivariate tests

Correlation is a measure of the extent to which two variables are linearly related (Miles, Shevlin 2001).

1. Correlation among other bivariate tests

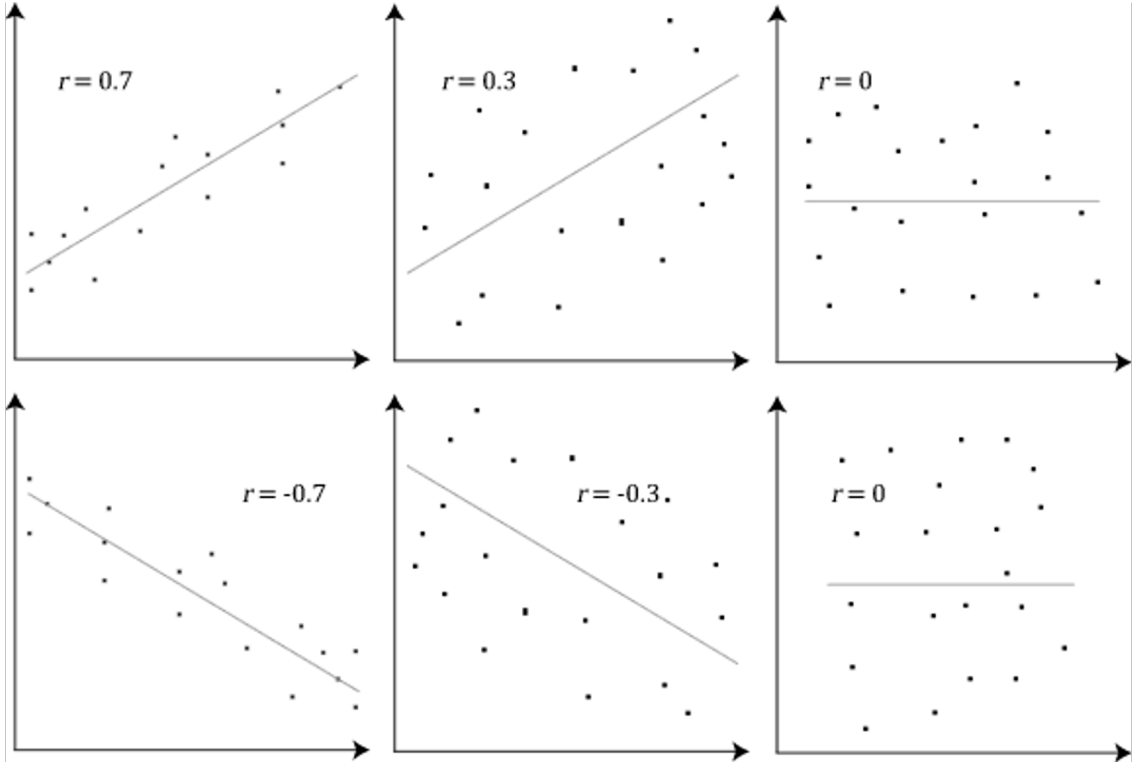
It answers three questions:

a) Is there an association between the variables?
b) If yes, is it positive or negative?
c) If yes, how strong is the association?

Correlation is a reciprocal, mutual association. The relationship remains linear regardless of the measurement scales.

1. Correlation among other bivariate tests

A correlation can be thought of as the extent to which the scattergraph of the relationship between two variables fits a straight line:



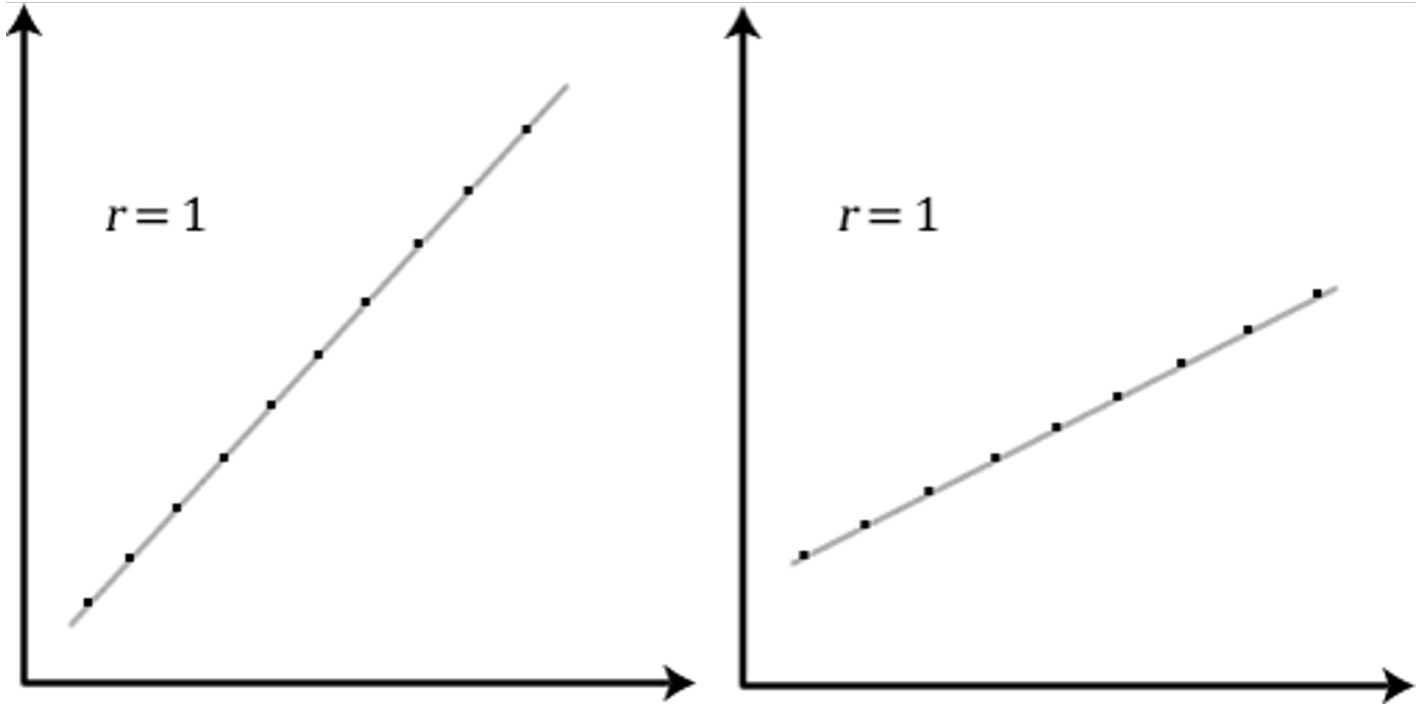
2. How to read a correlation

A **perfect linear association**: all points fall on a straight line going from bottom left to top right or from top left to bottom right. Correlation coefficients range from -1 to +1 where +-1 is the perfect association and 0 mean lack of it.

When the points lie **closer to the shape of a line** drawn in the diagonal the correlation is **higher**.

The **sign** of the correlation gives the **direction** of the diagonal line.

2. How to read a correlation



2. How to read a correlation

1. Is the correlation statistically significant?
   Look at the probability value: how likely the true correlation to be zero.
   Smaller sample produce more uncertainty: random errors have a greater impact, a large r occurs more often. Larger samples make even very small correlations significant.
2. Is the correlation positive or negative?
   Look at the sign. If there is no sign, it is positive.
3. How large is the correlation?
   Jacob Cohen (1988):
   small ~ 0.1 (absolute value)
   medium ~ 0.3 (absolute value)
   large ~ 0.5 (absolute value)
   (these are not strict rules)

3. Correlation coefficients: Pearson's product moment coefficient

- Pearson's r is used for bivariate relationships, i.e. to describe the relationship of two continuous variables
- **Covariance** (Cov) is a measure of joint variability of two variables: it is positive if variables go in the same direction (X↑Y↑ or X↓Y↓), negative if not. It is a non-normalised measure that depends on the original scales of variables.
- **Correlation** is normalised covariance: covariance of X and Y divided by the product of their standard deviations (Sx * Sy). Correlation can be significant or not (there is a test for that).

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

3. Correlation coefficients: Pearson's product moment coefficient

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

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

formula for the test of statistical significance of r_{xy}

df = n - 2

$$\mathrm{cov}\ (x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$$

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

$$S_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2}$$

Formulas to recall, See:
https://www.onlinemath4all.com/karl-pearson-product-moment-correlation-coefficient.html

3. Correlation: non-parametric coefficients

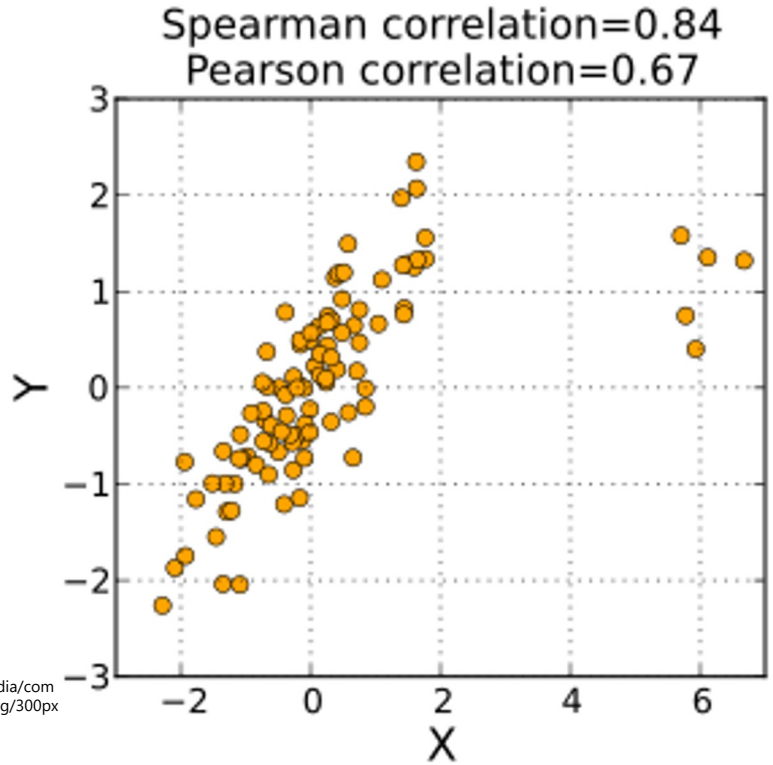
Spearman's rank correlation coefficient: fits ordinal data; does not imply linearity; assesses the monotony of joint variability, or similarity between rankings; ranges from -1 to 1

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}}$$

Kendall's tau rank correlation coefficient is also based on ranks but rests on whether two variable ranks are concordant/ discordant; tau is better for *smaller samples* and it is more robust than r_s. The tau coefficient is often smaller than r_s but most often they agree with each other.

3. Correlation: non-parametric coefficients

Compare:



Source:
https://upload.wikimedia.org/wikipedia/commons/thumb/6/67/Spearman_fig3.svg/300px-Spearman_fig3.svg.png

4. Assumptions for Pearson's correlation

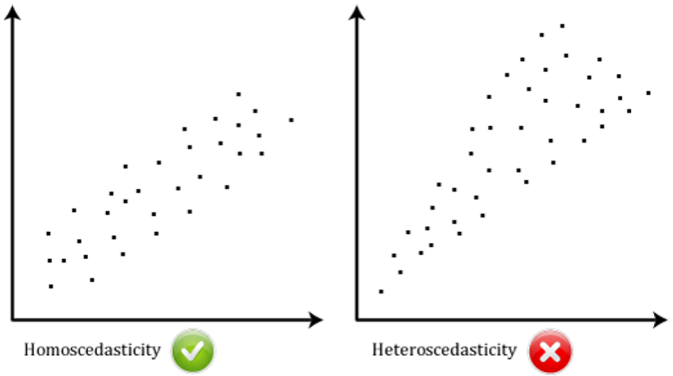
1. Measurement scale - continuous
2. Independent observations
3. Data points for each observation
4. Linearity (plot your data on a scatterplot)
5. Bivariate normality or univariate normality in both variables
6. Homoscedasticity (similar variance along the line)
7. No outliers

Recommended: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

4. Assumptions for Pearson's correlation

5. Bivariate normality or univariate normality in both variables (QQ plot, histogram)

6. Homoscedasticity
(similar variance along the line)



7. No outliers
(an outlier is an observation within your sample that does not follow a similar pattern to the rest of your data -- visualise to check). Outliers can affect linearity, normality and the value of coefficient.

Recommended: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

4. Assumptions for Pearson's correlation

Null hypothesis testing

The null hypothesis for a correlation is that no association exists between the two constructs and therefore the correlation coefficient in the population is zero:
H_0: r = 0
H_A: r != 0 (two-sided in most cases)

Spearman's: H_0: there is no monotonic relationship between the two variables in the population

Recommended: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

4.* Assumptions of these tests are similar as they belong to the same framework of linear model

The t-test assumptions:

- measurement level, normality, homogeneity of variances, no outliers

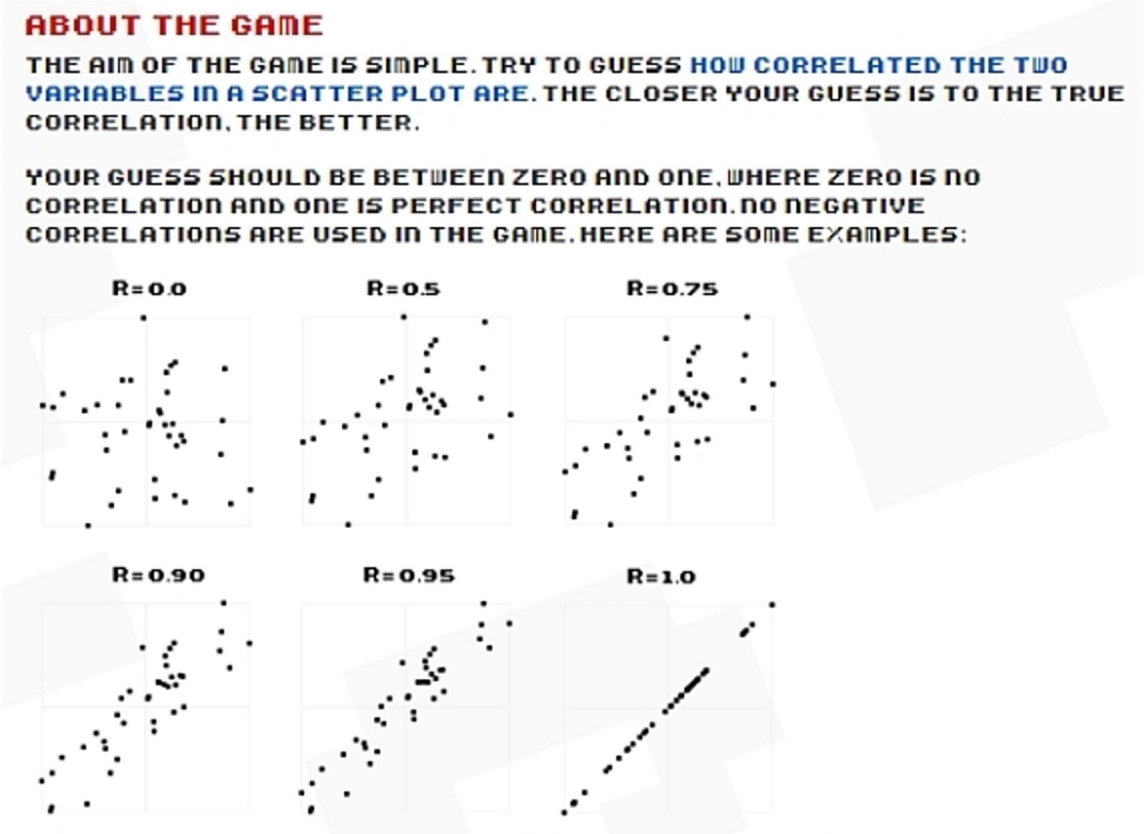
ANOVA assumptions:

- measurement level, normality, homogeneity of variances, no outliers

Correlation assumptions:

- measurement level, linearity, normality, homoscedasticity, no outliers

Play at home: guess the correlation to train your eye



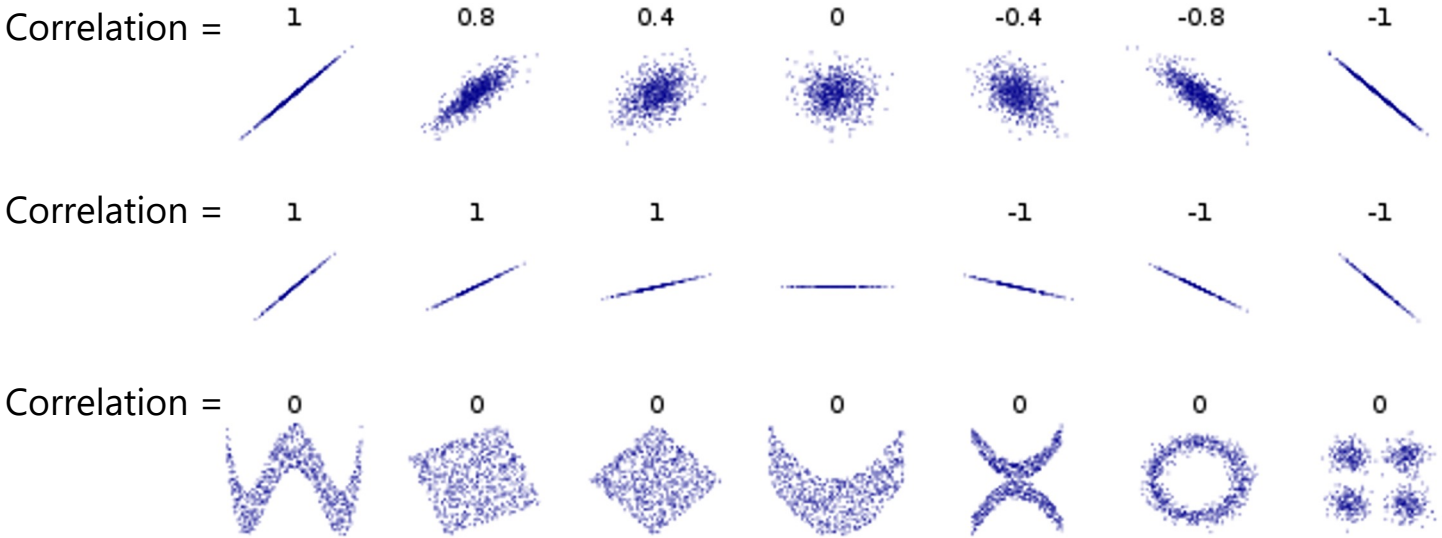
Where to practise:

basic
http://guessthecorrelation.com/

advanced
https://www.geogebra.org/m/KE6JfuF9

5. Common **pitfalls** with correlation

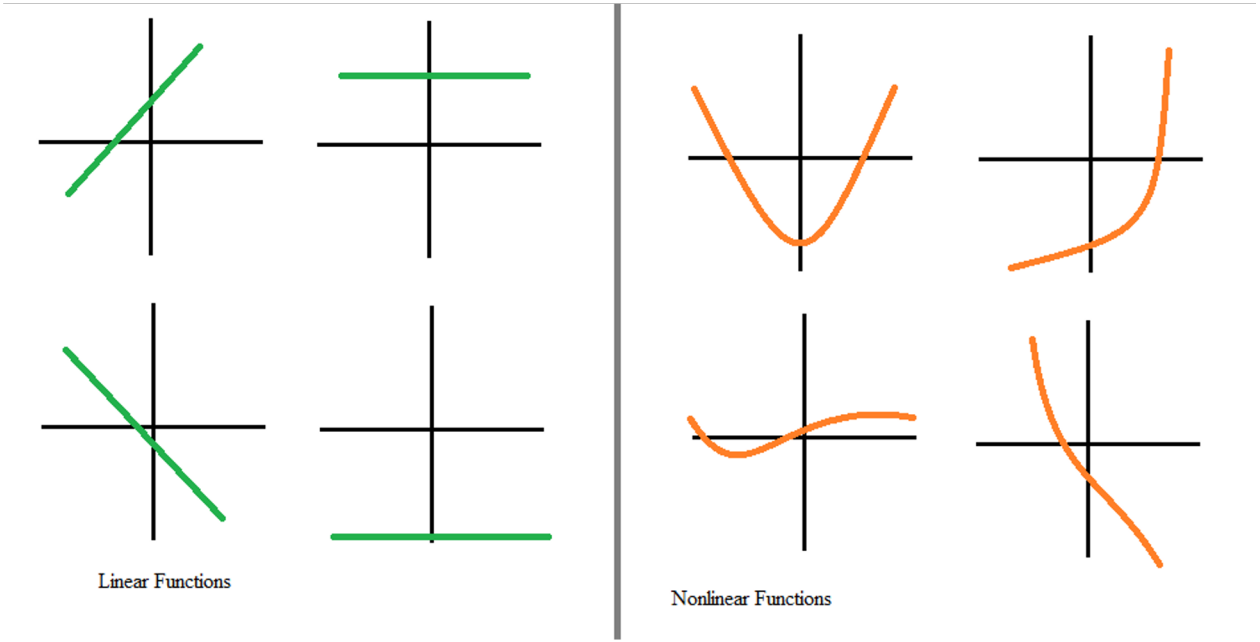
Despite simple appearance, correlation coefficients can be highly deceptive as they do not capture everything about the data:



https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/506px-Correlation_examples2.svg.png

5. Correlation describes a linear relationship

*linear (left) vs. nonlinear (right) relationships:

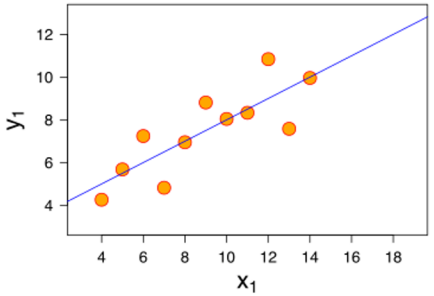


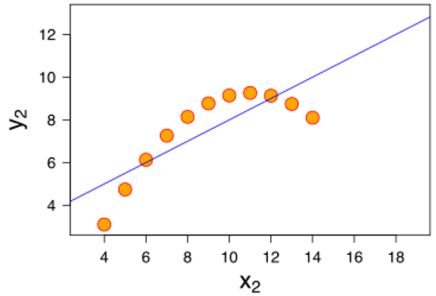
5. Common pitfalls with correlation

1. *Correlation is not causation*
    a. Ex.: smoking and lung cancer, industry and TB, etc.
2. *Zero correlation does not always mean there is no relationship between the variables*
    a. Not all relations are linear, see Anscombe's quartet
    b. Heterogeneous samples, see Simpson's paradox
3. *Non-zero correlation coefficient does not always mean a relationship between the variables*
    a. Spurious correlations, see http://www.tylervigen.com/spurious-correlations

Anscombe's quartet: four data sets with Y having the same mean and SD, the same correlation coefficient of .86, and different distributions

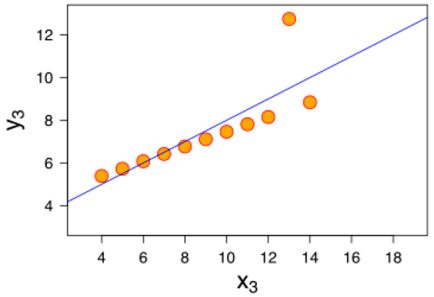
no outliers
linearity

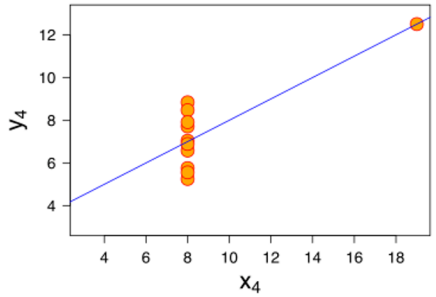




nonlinearity

outlier
linearity

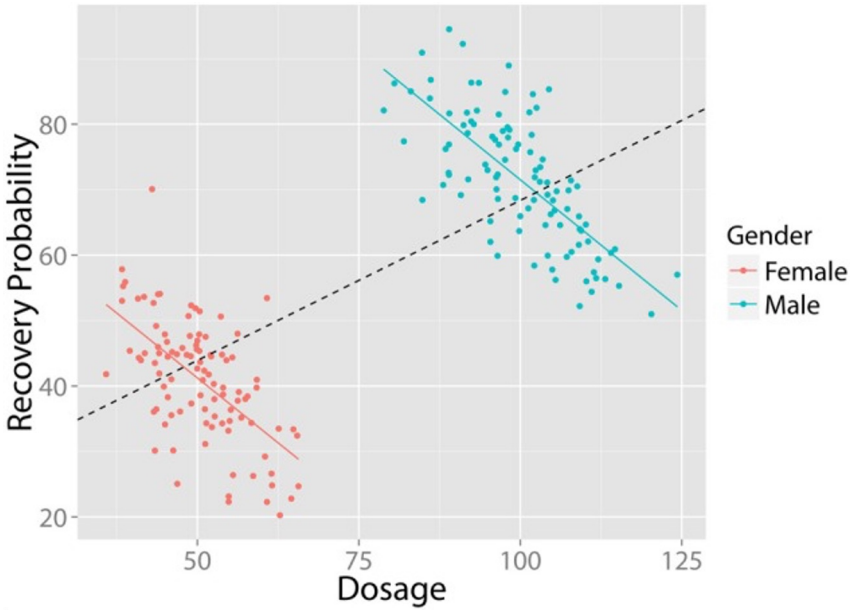




outlier
nonlinearity

Conclusion: visualise your data before interpreting the coefficients

Simpson's paradox: in a heterogeneous sample, relationships within groups and on the aggregate level can go in opposite directions



Read more here: https://www.researchgate.net/publication/256074671_Simpson's_Paradox_in_Psychological_Science_A_Practical_Guide

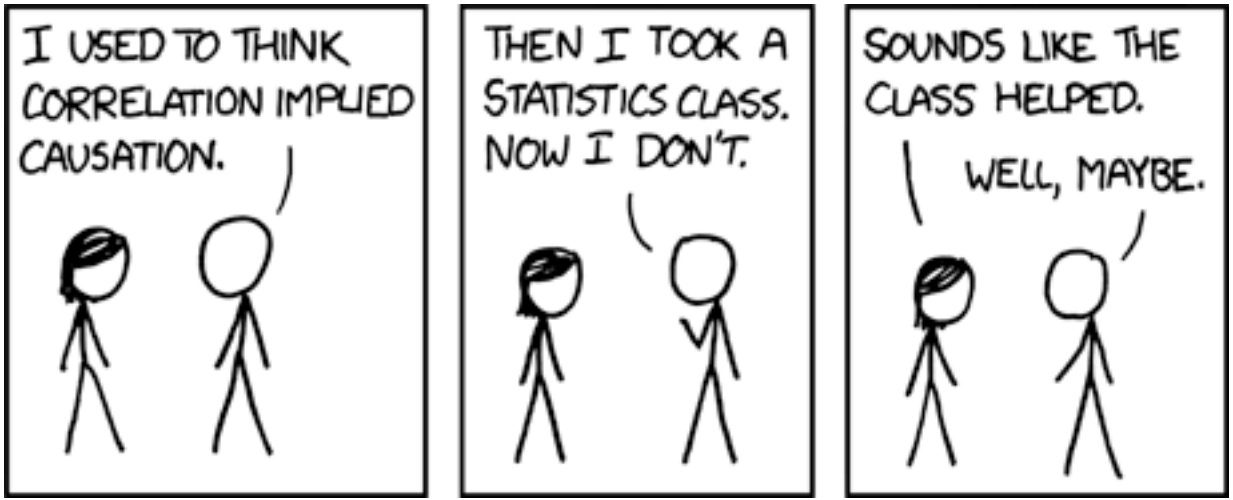
Conclusion: visualise data before interpreting, mind the groups, do not generalise individual-level conclusions to the group level without checking

Sources

Textbook https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
TUTORIAL https://www.sheffield.ac.uk/polopoly_fs/1.43991!/file/Tutorial-14-correlation.pdf

Short notes from Boston U https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression3.html

Measures of Association How to Choose?
https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.