Correlation

Data files for today

- CourseEvaluation2.sav
- pontokPrediktorok.sav
- Happiness.sav
- Catterplot.sav

Defining Correlation

- Co-variation or co-relation between two variables
- These variables change together
- Scale (interval or ratio) or ordinal variables

Does income vary as a function of school grades?

Does life expectancy increase with health expenditure?

Is speed of reaction inversely related to age?

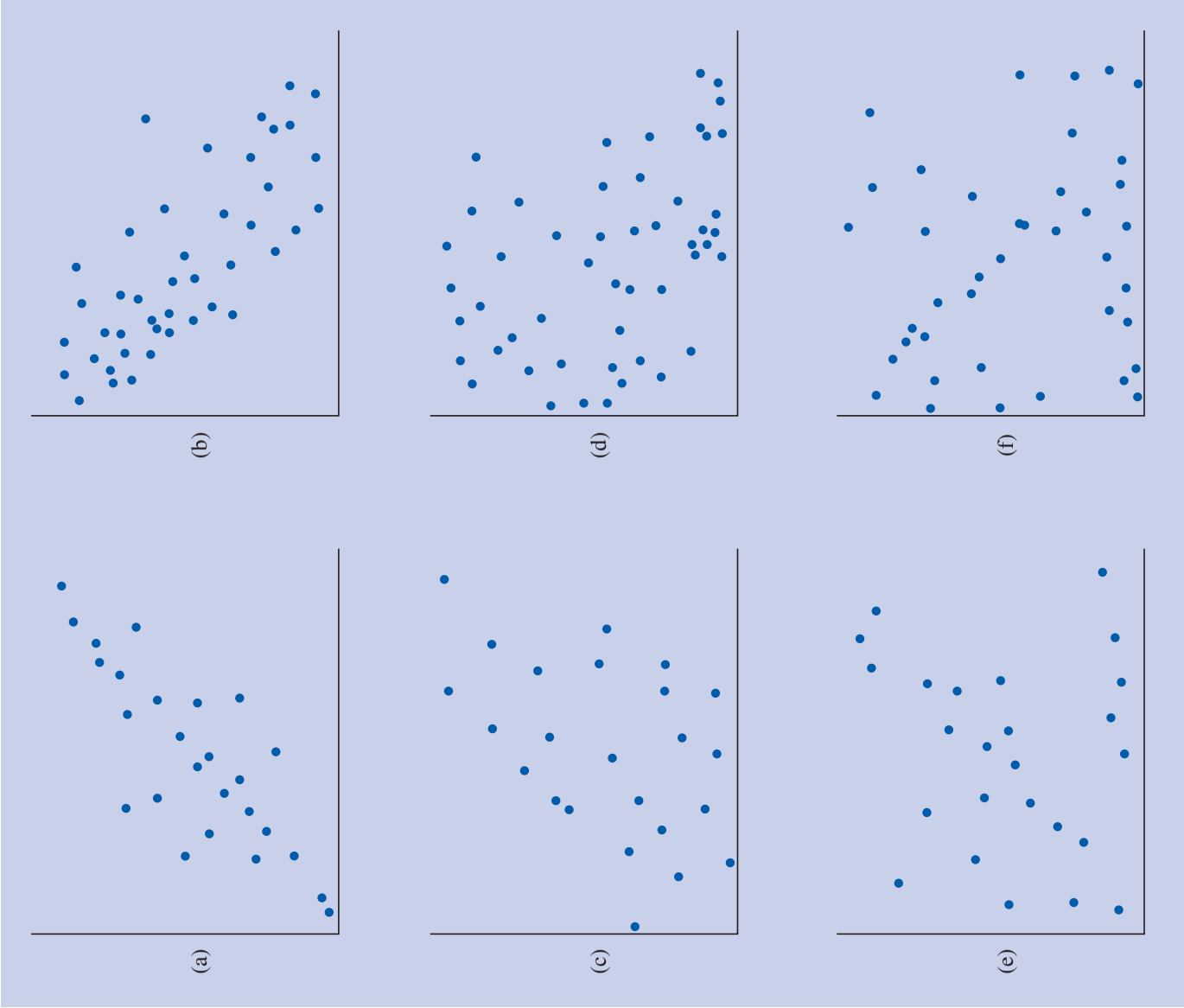
Requirements

- Plenty of data points
- Several different values of each variable

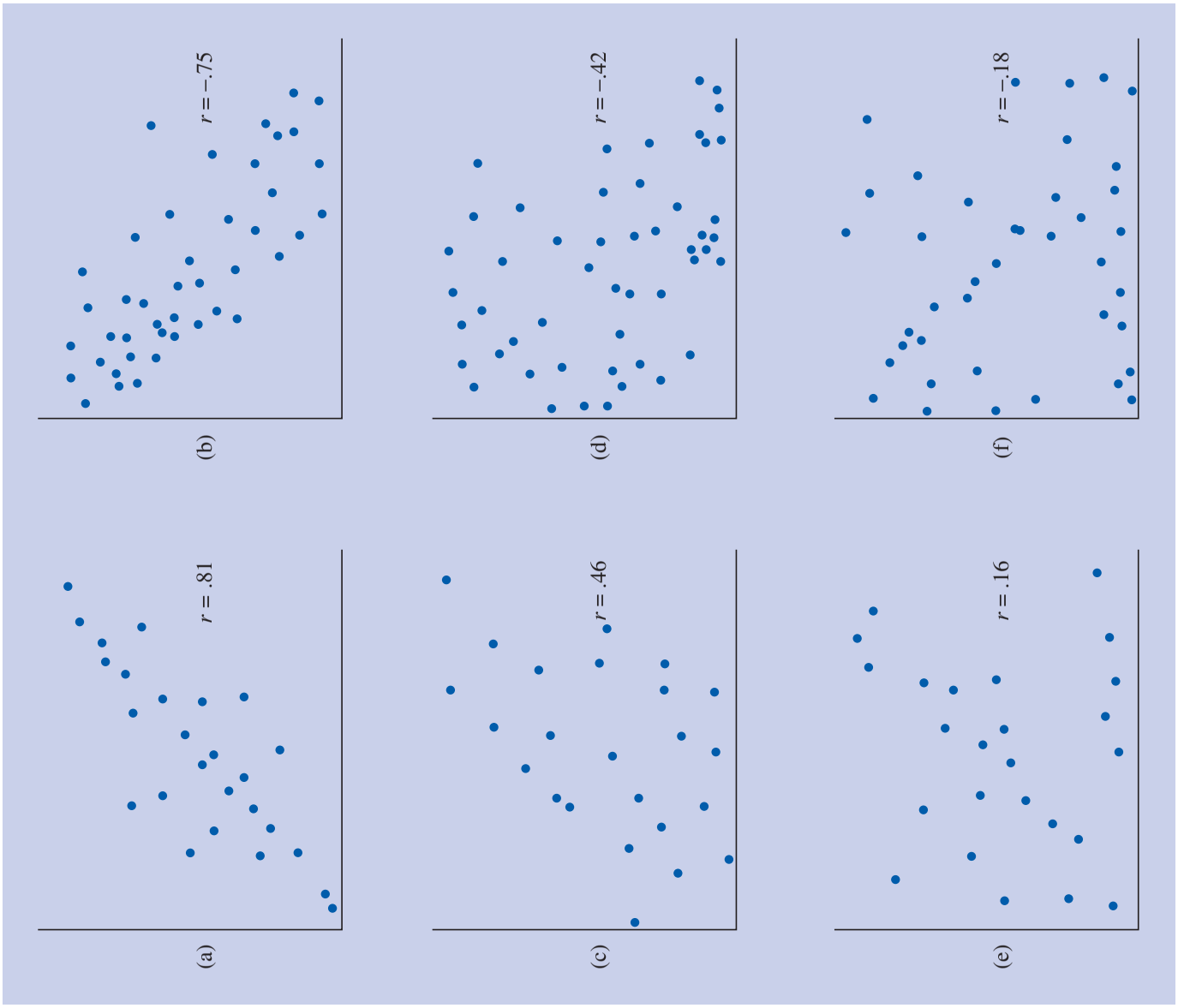
Linear vs. curvilinear correlation

- Scatterplots: one variable on the x axis, the other on the y axis, each point in the graph represents one observation
  - X axis: the presumed predictor
  - Y axis: the presumed outcome
- Linear: the relationship can be described by a straight line (height and shoe size)
  - Positive relationship
  - Negative relationship
- Curvilinear: the relationship can be described by a curved line or more than one straight line (processing speed and age)

Scatterplot: Strength of correlation



The correlation coefficient



Analyzing the Data

- Correlations range from -1.00 to +1.00
  - Size indicates strength of the relationship
  - Sign indicates direction of the relationship
- Most common types of correlations
  - Pearson product-moment correlation
  - Spearman rank-order correlation

Pearson vs. Spearman

- Pearson r for interval data
  - normally distributed scale data
  - linear relationships
  - no outliers
- Spearman's r_s for ordinal data
  - ordinal data or non-normally distributed data
  - converting scores to rank order closes the gap between outliers and the rest

Interpreting the Data

- Note size and sign of correlation coefficient
  - Indicates strength and direction of relationship
  - **Size is really and effect size**

- Is the correlation likely to be due to chance?
  - Is the *p* value > alpha?

Intercorrelation matrices

- If you test the correlations between more than two variables, each pair of variables will be tested
- These are shown in an intercorrelation matrix

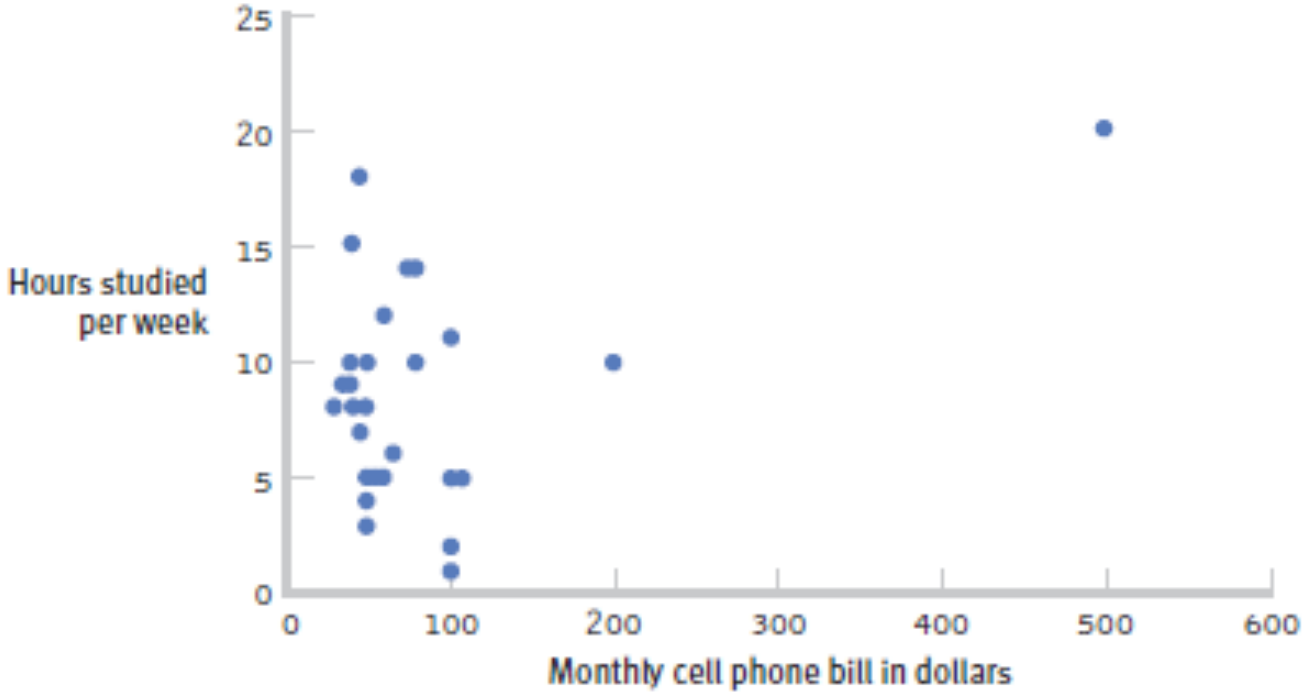
Limitations of these Methods

- Problems in determining causation
  - A correlation does not imply causality
    - If ***A*** and ***B*** are correlated, then
      - ***A*** could cause ***B***
      - ***B*** could cause ***A***
      - Some other variable could cause both

The Limitations of Correlation, cont.

The effect of an outlier:

In a study on the relationship between cellphone use and studying, one individual who both studies and uses her cell phone more than any other individual in the sample changed the correlation from -0.14, a negative correlation, to 0.39, a much stronger and positive correlation!



IN MORE TECHNICAL TERMS

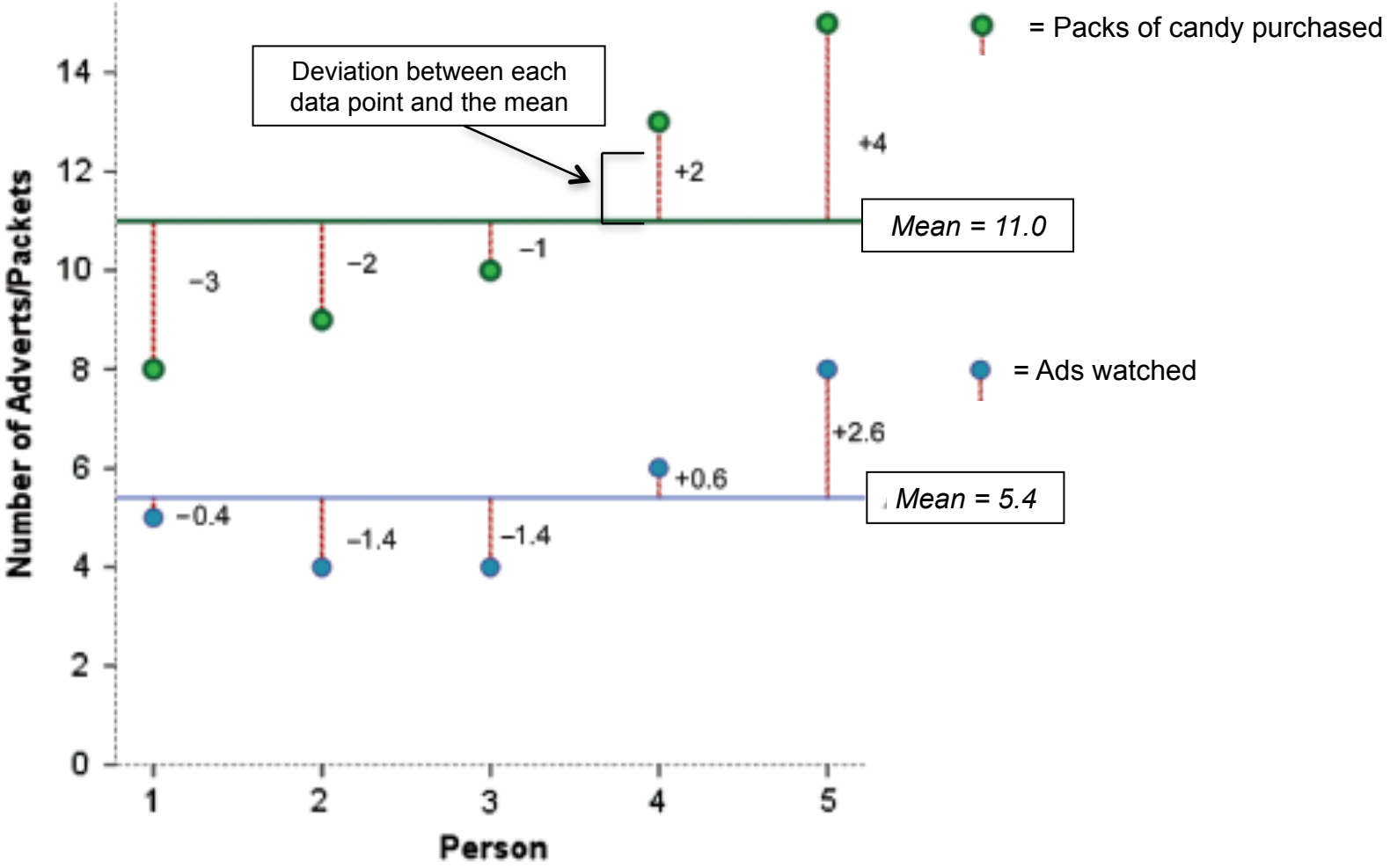
Covariance:

Assessing association between two variables

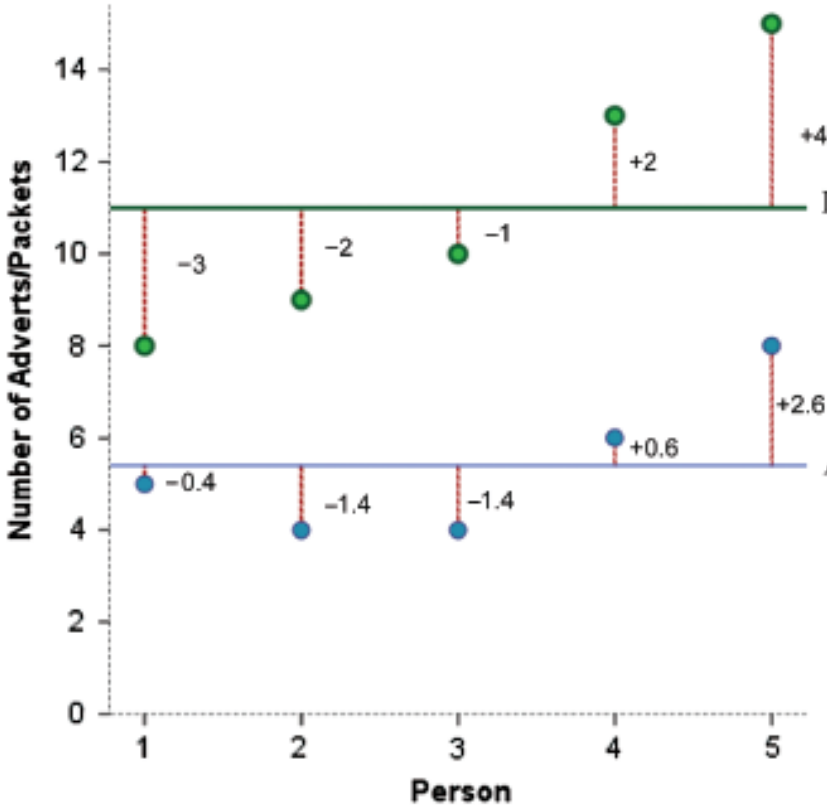
- Measures extent to which corresponding elements from two sets of ordered data move in the same direction

- Its value is influenced by:
  - the strength of the linear relationship between X and Y
  - the size of the standard deviations of X and Y (i.e., SD_x and SD_y)

Example: Imagine we expose five people to a specific number of advertisements promoting a particular type of candy and then measure how many packages of the candy each person purchases the following week



Cross product deviations and covariance



We quantify the level of similarity between the two variances:

We multiply the individual deviations for one variable by the corresponding deviations for the second variable to obtain the **cross-product deviations**, add them up and then divide by N-1 to compute the **covariance.**

Covariance: Equation Form

$$\text{cov}(x,y) = \frac{\Sigma(X_i - M_x)(Y_i - M_y)}{N - 1}$$

$$= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (-0.6)(2) + (2.6)(4)}{4}$$

$$= \frac{(1.2) + (2.8) + (1.4) + (1.2) + (10.4)}{4} = \frac{17}{4} = 4.25$$

Positive covariance = as one variable deviates from the mean, the other variable deviates in the same direction.

Negative covariance = as one variable deviates from the mean, the other variable deviates from the mean in the opposite direction.

The Pearson Product-Moment Correlation Coefficient (r)

To standardize the covariance:

We divide the covariance by the product of the standard deviations. (This essentially means that we use z scores (number of standard deviations from the mean).)

The standardized covariance is known as the correlation coefficient:

$$r = \frac{cov_{xy}}{SD_x SD_y} = \frac{\Sigma\left(x_i - Mx\right)\left(y_i - M_y\right)}{(N-1)SD_x SD_y}$$

Hypothesis testing

- r is converted into a t statistic:

$$t_r = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}} \qquad t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- t^r value is checked against the t distribution for N-2 degrees of freedom

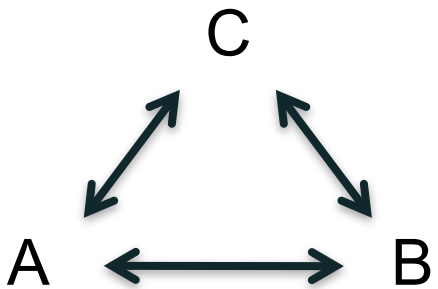
Types of Correlation

Bivariate correlation:

Used to assess the relationship between two variables.

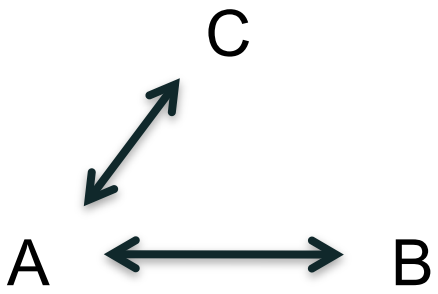
Partial correlation:

When we do a partial correlation between two variables, we control for the effects of a third variable. Specifically, the effect that the third variable has on *both* variables in the correlation is controlled.



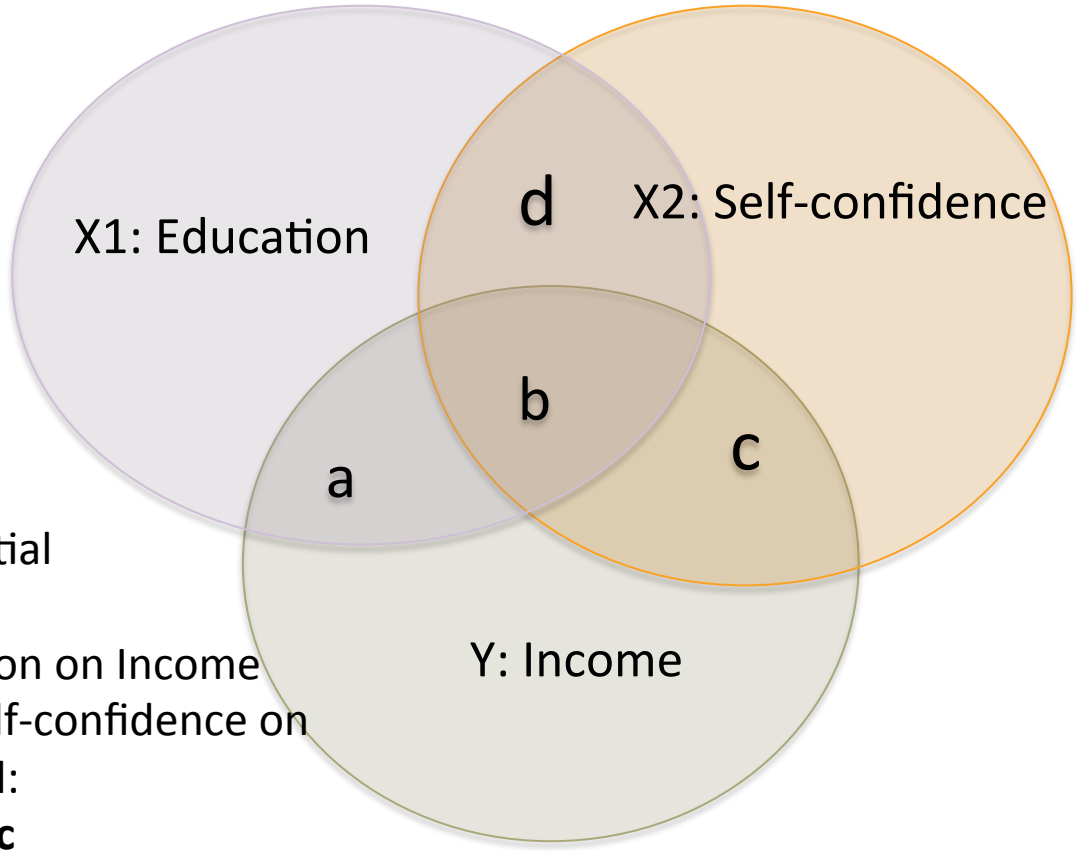
Semi-partial (or part) correlation:

When we do a semi-partial correlation, we control for the effect that the third variable has on only one of the variables in the correlation.



Partial Correlation

- A partial correlation is the relationship between two variables after removing the overlap with a third variable completely from both variables: Effect of Education on Income with all effects of Self-confidence removed: **a** without **b, c** and **d**



Part or semi-partial correlation:
Effect of Education on Income with effect of Self-confidence on Income removed:
a without **b** and **c**

Correlation: Non-Parametric Alternatives

Spearman's Correlation Coefficient (r_s or rho or ρ):

Used when the data violate parametric assumptions (e.g., non-normally distributed data; ordinal data).

Works by first ranking the data and then applying Pearson's equation to these ranks.

Kendall's tau (τ):

Used instead of Spearman's correlation coefficient when data set is small and has a large number of tied (equal) ranks.

Regression lines and Confidence intervals

- We can fit a regression line: line with minimum error (least squares: minimum amount of deviation)
  - R^2 (Coefficient of determination) = r^2: proportion of shared variance
- SPSS calculates bootstrapped confidence intervals for r (simulating repeated sampling)

Exercise 1: relationship between expected grade and rated course quality (courseEvaluation2.sav)

Grade	Course Evaluation
4.38	4.25
4.00	3.25
3.50	3.65
4.13	3.75
4.00	4.75
4.00	4.28
4.50	3.13
5.00	4.88
3.75	4.15
3.88	4.25
2.25	2.50
4.13	3.63
4.00	4.13
4.25	4.88
4.63	3.38

1. Look at distribution. Outliers?
2. Draw graph.
3. Run analysis.
4. Get bootstrapped confidence intervals.
5. Anything suspicious?
6. If yes, remove outliers.
7. Redraw graph.
8. Rerun analysis.

Writing up the results

It is often thought by course instructors that the way in which students evaluate a course will be related, in part, to the grades that are given in that course. In an attempt to test this hypothesis we collected data on 15 courses in a large university, asking students to rate the overall quality of the course (on a five-point scale) and report their anticipated grade in that course.

A Pearson correlation between mean rating and mean anticipated grade produced a fairly strong positive correlation (r = .52[-.42, .87], p = .047). Two outliers were, however, identified: One of the courses had an unusually high average grade and one an unusually low average grade. With these two courses removed from the analysis, the remaining 13 courses showed a slight (non-significant) negative correlation between expected grade and course evaluation (r = -.218[-.713, .417], p = .47)

We therefore have no conclusive evidence for the hypothesis that students' course ratings are affected by their grades.

Exercise 2: Homework and exam points (pontokPrediktorok.sav)

- Average homework grades and exam points
  - Distribution (normality, outliers)
  - Scatterplot
  - Analysis (Analyze –> Correlate -> Bivariate)
  - Bootstrapped confidence intervals
- Educational background variable

Exercise 3: happiness.sav

- Does happiness correlate with income?
- Does happiness correlate with health?
- Does happiness correlate with income WITH THE CONTRIBUTION OF HEALTH CONTROLLED? (Analyze -> Correlate -> Partial)
- Now let's look at men and women separately.

For fun

- Draw a scatterplot for: catterplot.sav

time since last meal x amount of purring

Homework

happiness.sav

- Is there a correlation between
  - Income and Size of town?
  - Number of siblings and Happiness?
  - Watching TV and Happiness?
- Draw scatterplots, run the analyses, calculate confidence intervals
  - Are you going to use Spearman or Pearson? Why?
- You can explore other variables (e.g. number of siblings and size of town; watching TV and age, separating men and women etc.)
- Choose ONE and write it up. Include graph.