# Statistiek I

## Questionnaires

John Nerbonne

CLCG, Rijks*universiteit* Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/

# Motivation: Questionnaires — a simple way to get data

- Do people find foreign accents attractive? (Martijn Wieling, Mona Timmermeister, Kaitlin Mignella)
- Do teenagers trust information from health organizations? (Ellen Hoogstraten)
- How effective are (health) campaigns appealing to fear? (Carel Jansen)

Just ask! ... noting obvious potential problems (honesty, "correctness", ...)

Analyse data as numerical (Likkert scale data) or categorical (as proportions), which have been treated earlier in the course.

# Problem: How to ask the right question?

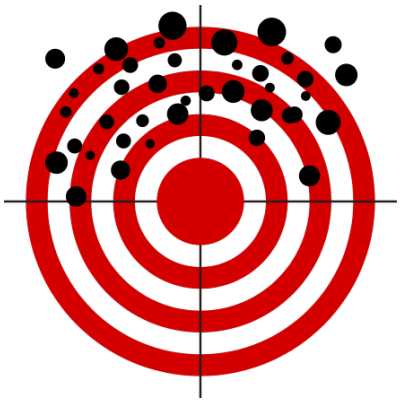
**Example:** Do teenagers trust information from the community health center?

- Is it important what sort of information is being sought?
  —Ask questions about different sorts of information?
- How many teenagers link the abstract question to concrete events, e.g. being worried about sexually transmitted diseases?
- How many know the community health center by name?
  —Or is it enough to show the sort of information offered (screenshot)?

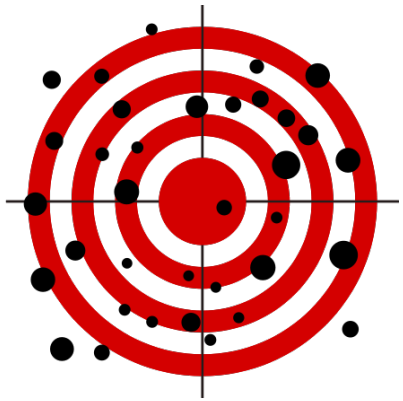
To solve some of these problems, researchers typically ask several questions, all aimed at acquiring similar information.

If the answers CORRELATE well, the results of the questionnaire are more RELIABLE.

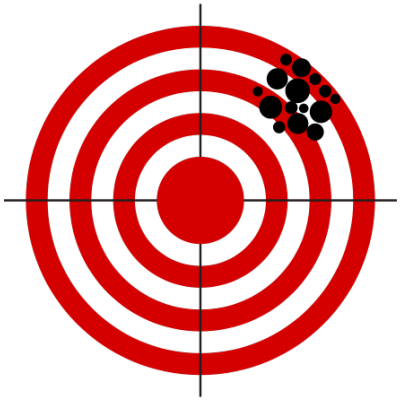
# Reliability and Validity in Tests



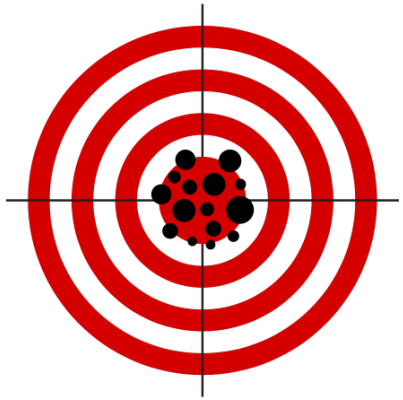
Unreliable & Unvalid



Unreliable, But Valid



Reliable, Not Valid



Both Reliable & Valid

Cronbach's $\alpha$ shows how to combine questionnaire items to improve reliability.

# Correlation

We often wish to compare two different variables

**Examples**: compare results on two distinct tests

- age and ability
- education (in years) and income
- speed and accuracy

**Methods** to compare two (or more) variables:

- Correlation coefficient

**Notice**:

- Correlation only for numeric variables!
  —Yes/No will be converted to $0/1$

# Correlation coefficient

How do you know if you are going to do well in a stats course?

Suppose you spend a lot of time on the material—more than your average class mate—then you'll have a high z-score in the distribution of study time.

You know that, generally, study time predicts grades.

So you know that you should have a high z-score in the distribution of grades.

If your final grade is not so good, you probably didn't spend much time studying. You would be below the mean in both distributions and have negative z-scores.

# Correlation coefficient

If $x = (x_1, \ldots, x_n)$ is study time, and $y = (y_1, \ldots, y_n)$ are grades, we can measure correlation between the two variables as

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

- compute everyone's z-score (study time and grades)
- multiply both z-scores and sum for everyone in class
- divide by the degrees of freedom (# students $-1$)

**Note**: positive sum results from multiplying two positive or negative z-scores for $x$ and $y$ (positive correlation)

Negative sum (correlation) results from multiplying positive and negative z-scores (and vice versa)

No correlation results from mixed-sign z-scores with sum close to zero.

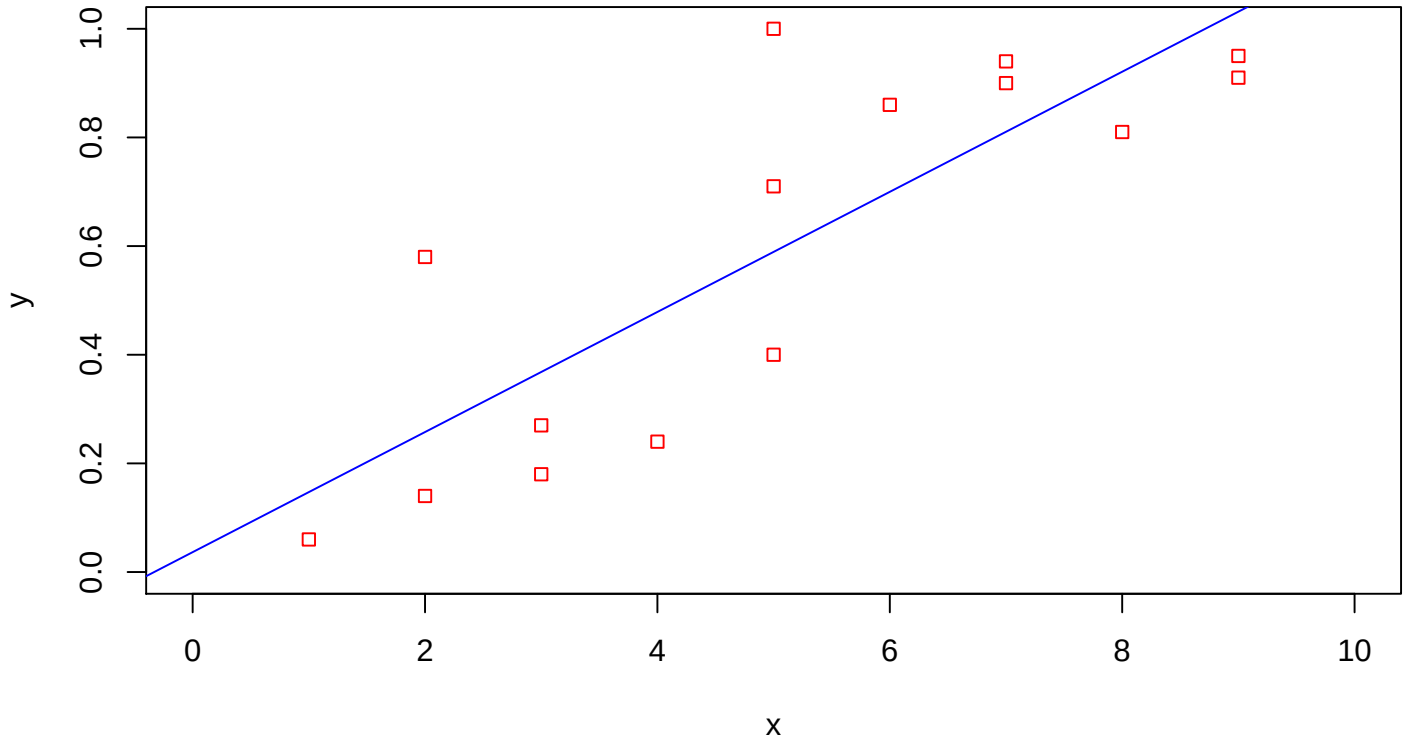
# Correlation coefficient

Correlation coefficient aka "Pearson's product-moment coefficient"

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

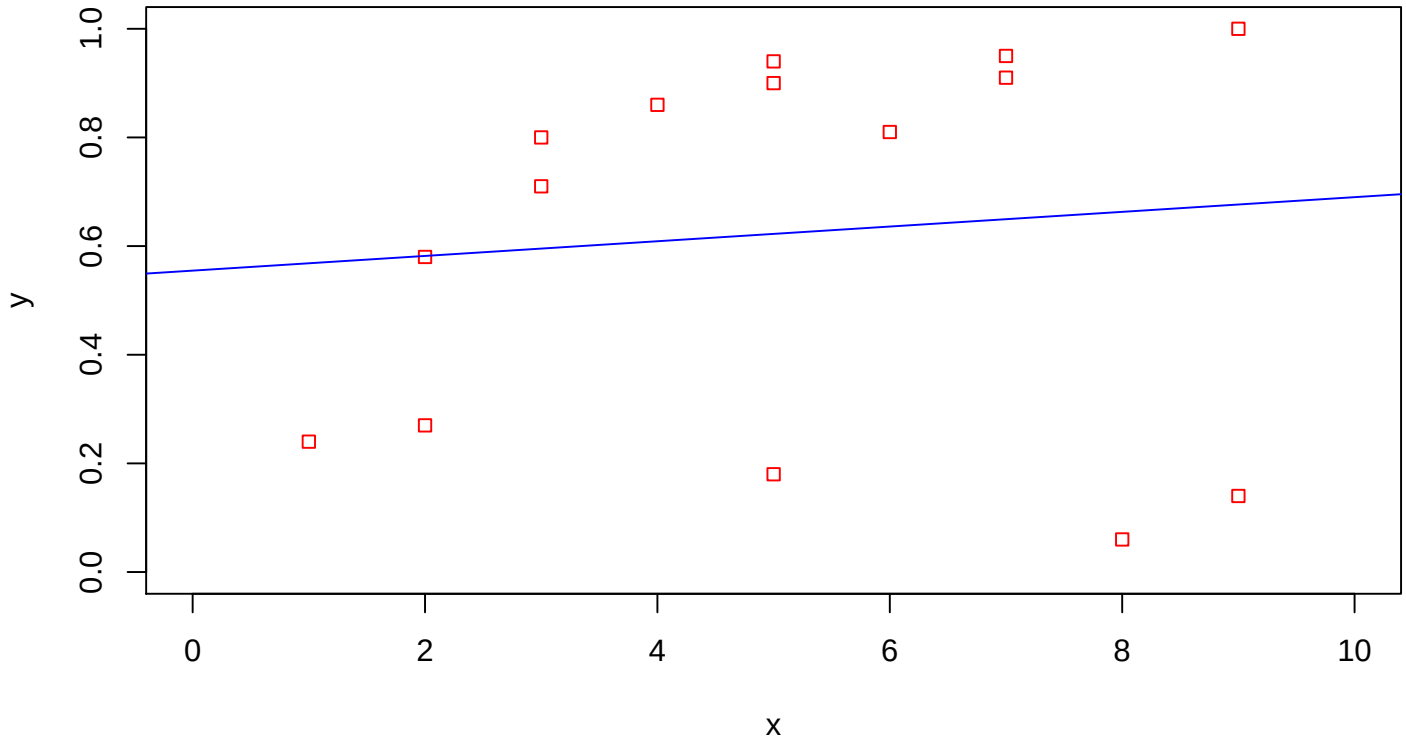
- $r_{xy}$ reflects the strength of the relation between $x$ and $y$
  - $r_{xy} = 0$ no correlation
  - $r_{xy} = 1$ perfect positive correlation (all data points on a straight line with positive slope)
  - $r_{xy} = -1$ perfect negative correlation
- no necessary dependence!
  - shoe size and reading ability correlate—both dependent on age

# Visualizing correlation



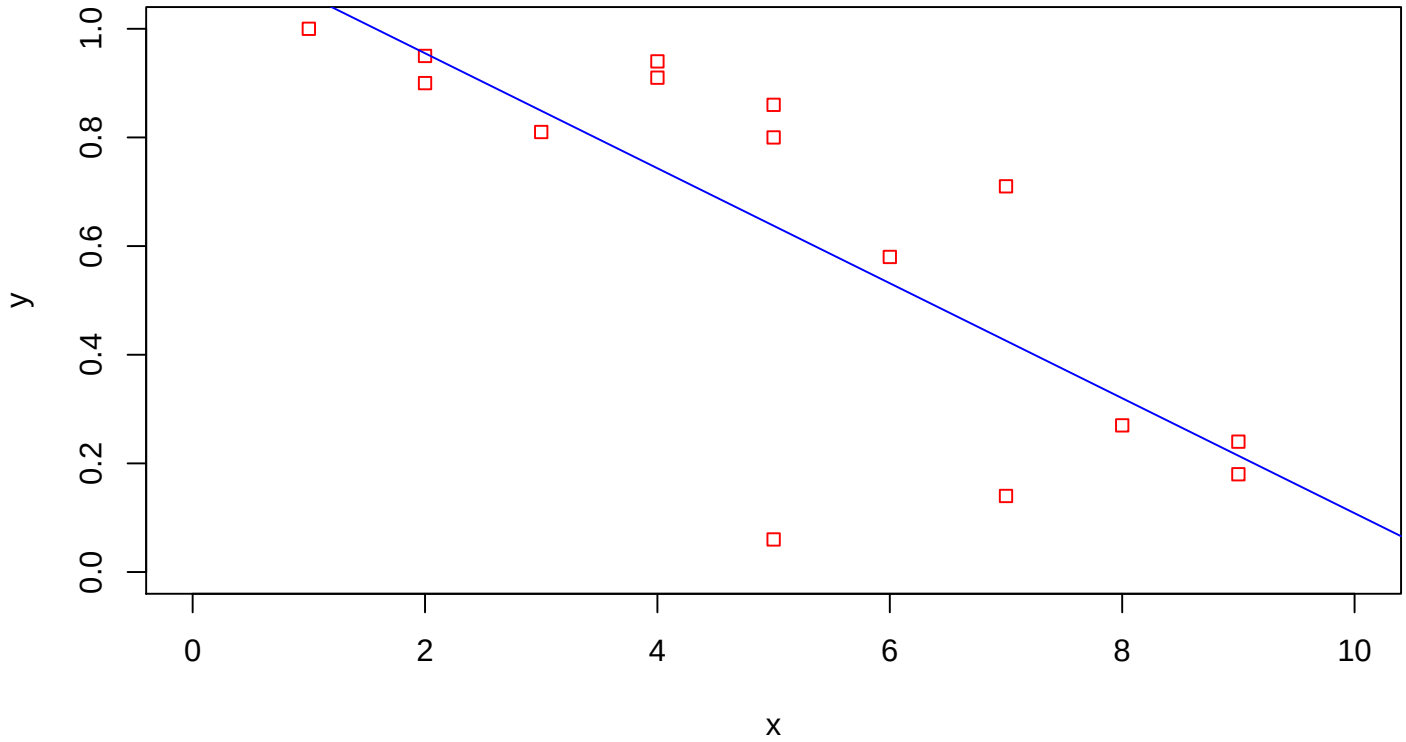

- data points lie close to the regression line
- correlation coefficient $r_{xy} = 0.83$
- strong positive correlation

# Visualizing correlation




- data points scatter in a cloud around regression line
- correlation coefficient $r_{xy} = 0.1$
- no correlation (there might be correlation in both subsets)

# Visualizing correlation



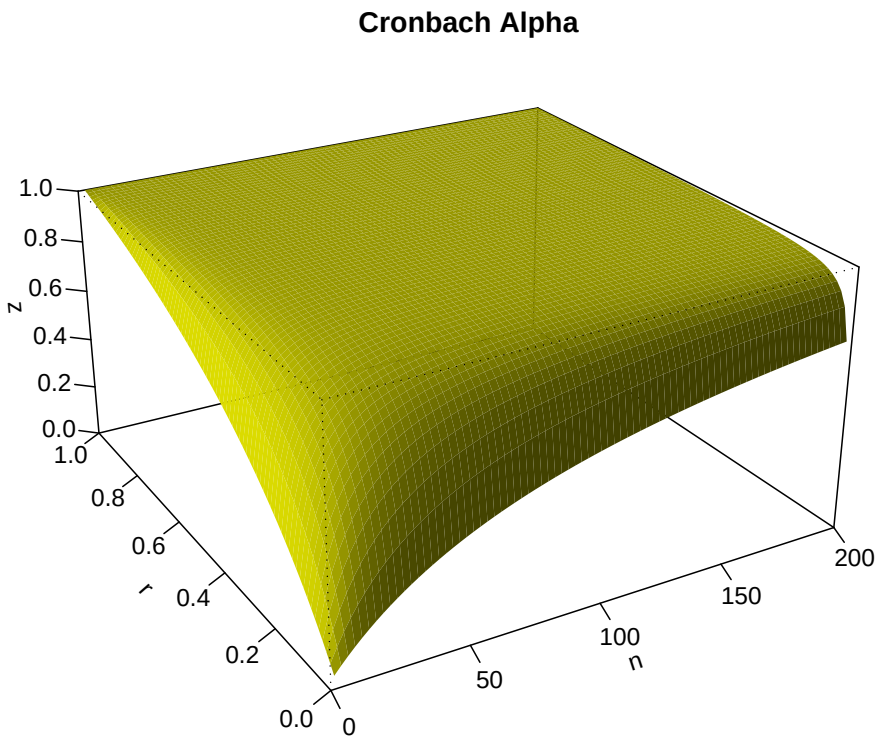

- data points close to regression line with negative slope
- correlation coefficient $r_{xy} = -0.77$
- correlation, but negative

# Cronbach's $\alpha$

Cronbach's $\alpha$: If we ask *n* questions, & answers correlate $\bar{r}$ (mean correlation coefficient), then we can derive a measure of reliability, Cronbach's $\alpha$.

Another way of looking at this: If we split the questions, how well would the two halves of the questionnaire agree? And if there are lots of questions, how well would they agree if we looked at all the ways of splitting?

# Cronbach's $\alpha$ depends on number of items and $\bar{r}$




The higher the inter-item correlation, the higher the reliability. The more items (with a high inter-item correlation), the higher the reliability.

# Cronbach's $\alpha$ can rise with fewer items if $\bar{r}$ rises

- Yfke Ongena (CIW) works on the *European Social Survey* (ESS 2010)
  - 600 Variables (!)
  - 38.000 respondenten
  - Four variables indicating interest in politics and trust in people

Reliability Statistics

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| ,632 | ,496 | 4 |

# Reliability example in ESS 2010

- Cronbach's $\alpha \approx 0.5$, even 0.63 too unreliable
- Examine inter-item correlations

Inter-Item Correlation Matrix

| | Most people can be trusted or you can't be too careful | Most people try to take advantage of you, or try to be fair | Most of the time people helpful or mostly looking out for themselves | How interested in politics |
|---|---|---|---|---|
| Most people can be trusted or you can't be too careful | 1,000 | ,569 | ,503 | -,146 |
| Most people try to take advantage of you, or try to be fair | ,569 | 1,000 | ,499 | -,139 |
| Most of the time people helpful or mostly looking out for themselves | ,503 | ,499 | 1,000 | -,103 |
| How interested in politics | -,146 | -,139 | -,103 | 1,000 |

- Problem with variable "interested politically"
- Inter-item correlation $\overline{5} = 0.2$
- Let's try eliminating the negatively correlating variable

# Reliability example in ESS 2010

- Cronbach's $\alpha \approx 0.5$, even 0.63 too unreliable
- Examine inter-item correlations

Inter-Item Correlation Matrix

| | Most people can be trusted or you can't be too careful | Most people try to take advantage of you, or try to be fair | Most of the time people helpful or mostly looking out for themselves | How interested in politics |
|---|---|---|---|---|
| Most people can be trusted or you can't be too careful | 1,000 | ,569 | ,503 | -,146 |
| Most people try to take advantage of you, or try to be fair | ,569 | 1,000 | ,499 | -,139 |
| Most of the time people helpful or mostly looking out for themselves | ,503 | ,499 | 1,000 | -,103 |
| How interested in politics | -,146 | -,139 | -,103 | 1,000 |

- Problem with variable "interested politically"
- Inter-item correlation $\overline{5} = 0.2$
- Let's try eliminating the negatively correlating variable

# Reliability example in ESS 2010

- Cronbach's $\alpha \approx 0.5$ too unreliable
- Examine inter-item correlations

Reliability Statistics

| Cronbachs Alpha | Cronbachs Alpha Based on Standardized Items | N of Items |
|---|---|---|
| ,767 | ,768 | 3 |

- We obtain an improved Cronbach's $\alpha$ by using fewer variables!
  —because these three correlate better ($\bar{r} = 0.52 \gg 0.2$)
- Cronbach's $\alpha > 0.7$ is acceptable, 0.8 is good, and 0.9 is very good.

# Reliability example in ESS 2010

- Cronbach's $\alpha \approx 0.5$ too unreliable
- Examine inter-item correlations

Reliability Statistics

| Cronbachs Alpha | Cronbachs Alpha Based on Standardized Items | N of Items |
|---|---|---|
| ,767 | ,768 | 3 |

- We obtain an improved Cronbach's $\alpha$ by using fewer variables!
  —because these three correlate better ($\bar{r} = 0.52 \gg 0.2$)
- Cronbach's $\alpha > 0.7$ is acceptable, 0.8 is good, and 0.9 is very good.

# Cronbach's $\alpha$ — Summary

- Cronbach's $\alpha$ measures RELIABILITY of (different) measurements, taken together
  - A measure is RELIABLE when it consistently yields the same result.
  - Reliability $\neq$ validity (i.e., whether a measure serves its purpose)
- Cronbach's $\alpha$ rises w. iter-item correlation and w. number of items
- $\alpha$ may rise when a questionnaire item is removed (when $\bar{r}$ rises)
- Once a good set is identified, mean values of the items may be used in the derived measure.
- Caution required
  - Reliability $\neq$ validity
  - Negatively correlating items will degrade $\alpha$