CS-E5250 Data-Driven Concept Design
5 ECTS

Quantitative Data Analysis
Assignment 2

Mika P. Nieminen   CS-E5250@aalto.fi

Markus Kirjonen

Agenda

- Today's lecture and its Learning Outcomes
- Theoretical introduction
- Assignment details and DLs

Quantitative Data Analysis

Introduction to quantitative data analysis using SPSS.

Learning Outcomes

- Become familiar with basic quantitative concepts and methodology.
- Learn basic quantitative research practices and experiment design.

After this Assignment:

- You know how to prepare and analyse real-world quantitative user research data to produce relevant user knowledge and insights.

Aalto University
School of Science

Quantitative Data Analysis

There are three kinds of **lies**:
lies, damned **lies**, and **statistics**.

Mark Twain

What is quantitative analysis?

"Quantitative research, in contrast to qualitative research, deals with data that are numerical or that can be converted into numbers."

Analysis of numerical data is commonly called 'statistics'.

The goal of quantitative analysis is to describe distributions and relationship between variables, and to test them using parametric and nonparametric tests.

When statistical analysis has become more accessible "..there is a danger that analysis may be performed on data without an understanding of the appropriate statistical tests to use and how they should be applied"

Examples of quantitative data

- Questionnaires and Surveys
- Interviews and Observations
- Usage logs
- …

How do we analyze qualitative data?

1. *Data preparation*
   – *Cleaning*
   – *Transformation*
2. *Statistical Analysis*
   – *Descriptive statistics*
   – *Inferential statistics*

Data cleaning

- Remove or replace erroneous or outlying data
  - Impossible, out-of-range, high deviations
- Remove unnecessary data
- Replace or substitute missing data

Missing data

"Because the values are subjective ratings of other people's competencies, we may consider them as statistically random inside each variable. In this case, based on Roth (1994) and Tsikriktsis (2005), the missing data should be imputed using the Hot-Deck missing data technique (MDT). Others have suggested more complex methods such as maximum likelihood (ML) and Bayesian multiple imputation (MI) (Schafer and Graham, 2002). Hot-Deck method imputes the missing data with an actual score from a similar case in the same data set (Roth, 1994). In this analysis the median value of each competency data set for End users role was used."

Nieminen, M.P. User-Centered Design Competencies - Construction of a Competency Model. Aalto University publication series DOCTORAL DISSERTATIONS, 36/2015

Data cleaning, transformation

- Prepare data from your data collection instrument to your analysis tool
  - Enumerate nominal scale values
  - Combined several data sources
  - SPSS format: Cases horizontally, variables vertically

Analysis - Descriptive statistics

- Central tendency

- Dispersion

- Shape

- Correlation

Analysis - Descriptive statistics

- Central tendency (Central limit theorem)
  – Mean
  – Median
  – Mode

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

280, 152, 318, 88, 6, 145, 123

280, 152, 318, 88, 6, 145, 123

Analysis - Descriptive statistics

- Central tendency (Central limit theorem)
    - Mean
    - Median
    - Mode

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

280, 152, 318, 88, 6, 145, 123 $\overline{X}_7 = 158,8571$

6, 88, 123, 145, 152, 280, 318

88, 123, 145, 152, 280, 318 $\overline{X}_6 = 184,3333$

148,5

Aalto University
School of Science

Analysis - Descriptive statistics

- Dispersion
  - Range (or Interquartile range)
  - Variance s^2
  - Standard deviation s

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

s^2 = sample variance
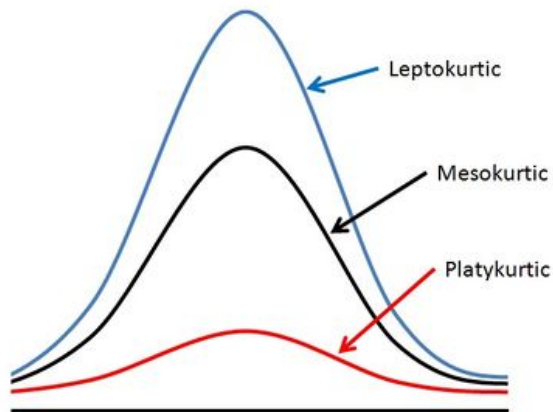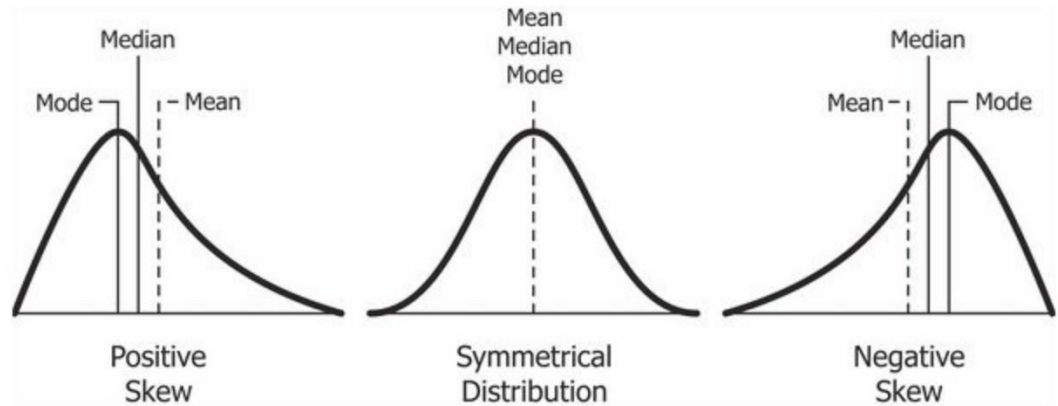
x_i = value of ith element

$\bar{x}$ = sample mean

n = sample size

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

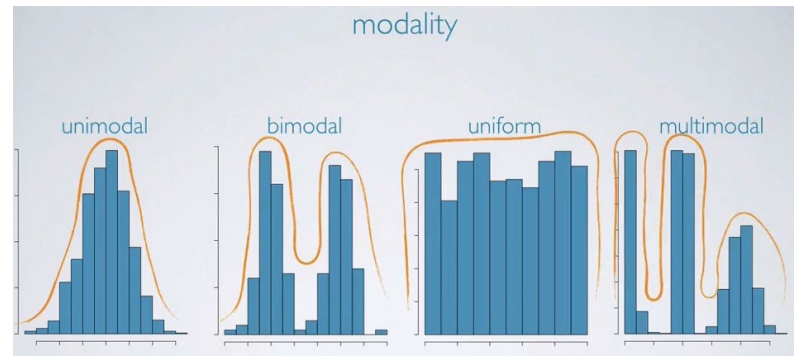# Analysis - Descriptive statistics

- Shape
  - Skewness
  - Kurtosis
  - Modality

Positive Skew · Symmetrical Distribution · Negative Skew

https://www.bogleheads.org/wiki/Excess_kurtosis

https://mathematica.stackexchange.com/questions/173275/a-simple-fast-way-to-estimate-distribution-modality

Aalto University
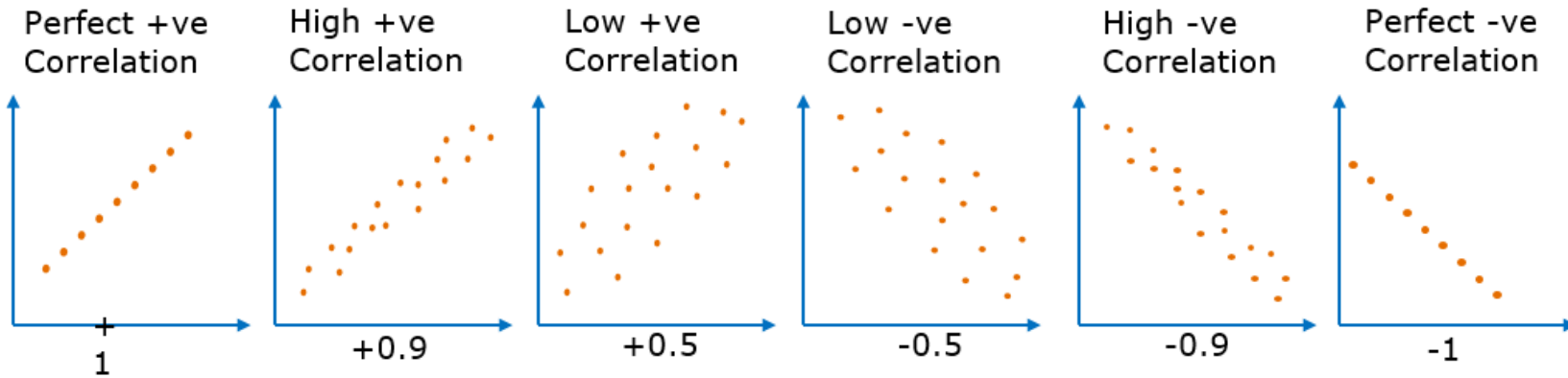School of Science

Analysis - Descriptive statistics

- r = Covariance standardized to std deviation

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by
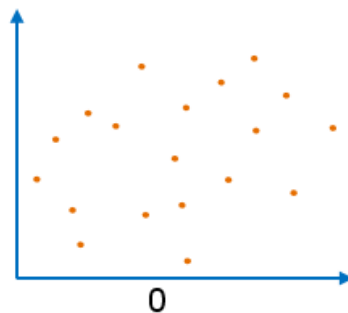
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.

Perfect +ve Correlation	High +ve Correlation	Low +ve Correlation	Low -ve Correlation	High -ve Correlation	Perfect -ve Correlation

+1   +0.9   +0.5   -0.5   -0.9   -1

0

1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

A!

Analysis -
Inferential statistics

- Conclude or infer population characteristics from samples
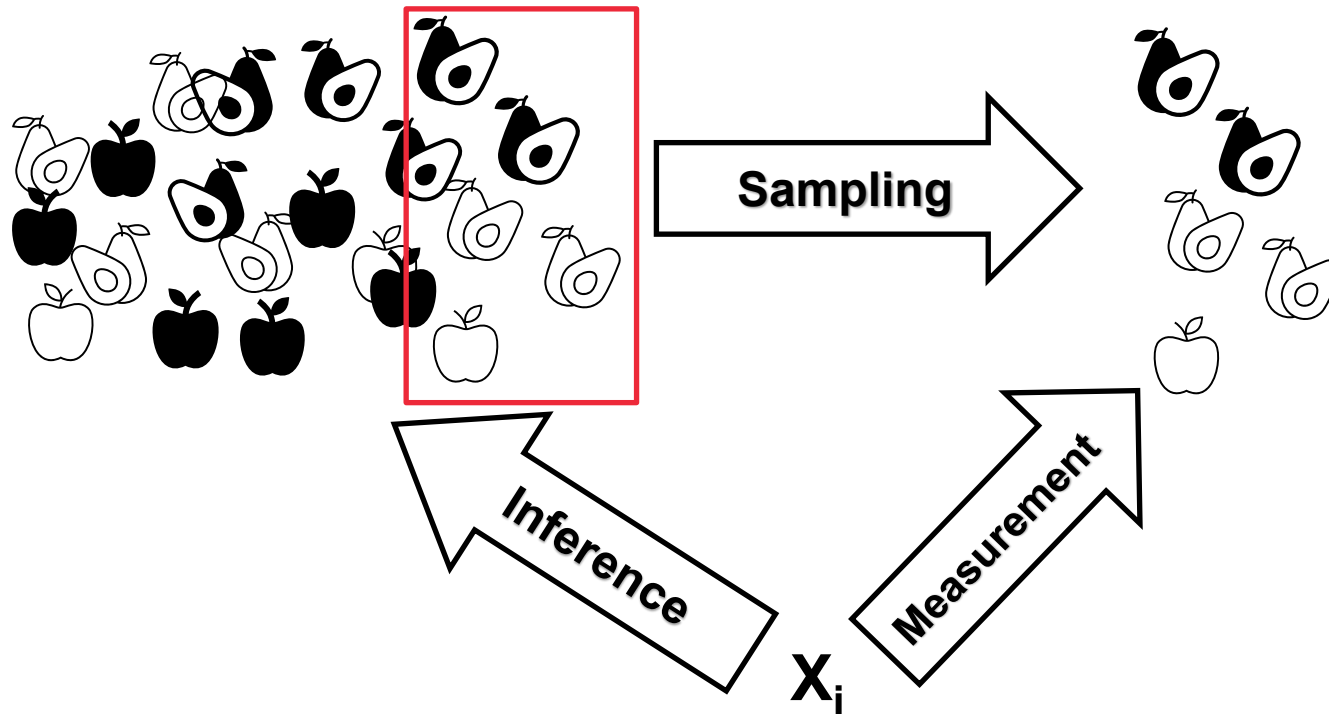
Analysis - Inferential statistics

- Concepts
- Parametric Tests
- Non-Parametric Tests
- Correlation analyses

Analysis - Inferential statistics

- Concepts
  - Null hypothesis, hypothesis of no relationship
  - Level of significance, can this happen by chance alone
  - *p value, commonly p < 0.05 used as a threshold*
    - *Less than 5% chance that difference is due to chance*
      -> reject null hypothesis
  - Confidence Interval
    - CI 95%
      5% chance that the population mean lies
      outside of the upper and lower confidence interval
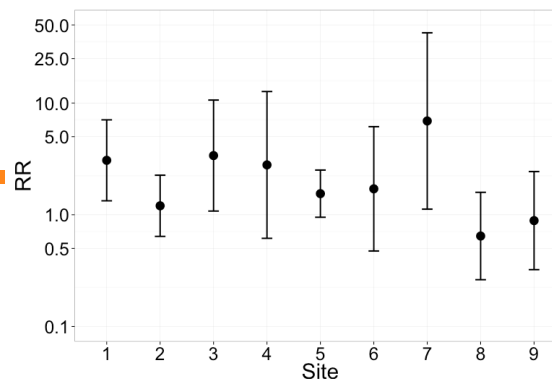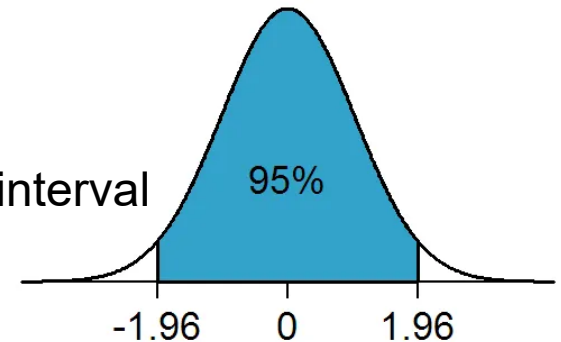
Analysis - Inferential statistics

- Parametric Tests
  - *2 or more data variables on the interval scale*
  - *Values are independent*
  - *Normal distribution*
  - *Homogenious variance (see Levene's test)*

- Tests the differences between the means of these distributions, calculates *p* value
  - *T-test (no difference in the means of two distributions)*
  - *ANOVA (no difference in the means of values of two or more distributions)*

A! Aalto University
School of Science

Analysis - Inferential statistics

- Non-Parametric Tests
  - *Used for **nominal** and **ordinal** data, or for interval scale data if the parametric test conditions are not met*

- Tests the differences in relative proportions of values between nominal values
  - Chi-squared (χ^2): Two variables are independent
  - Mann-Whitney: No difference in median values from two distr.
  - Wilcoxon signed rank: No difference in median values for matched pairs of subjects or repeated sapling
  - Kruskal-Wallis: No difference in median values from three or more distributions

Analysis - Testing for relationships

- Correlation analyses
  - Pearson:
    - Both distributions on the interval scale
    - Normal distributed population
    - Linear relationship
    - Homoscedasticity - variability in values is similar to all values
  - Spearman rho:
    - Both distributions on the ordinal scale
    - Relationship between the two variables is monotonic
  - Contingency coefficient:
    - Both distributions on the nominal scale

Recap: the quantitative analysis process

1. Prepare data
2. Install SPSS
3. Perform descriptive analysis
4. Perform inferential analysis
5. Report conclusions

Recommended reading

Sheard, J. (2018). Quantitative data analysis. In *Research Methods* (pp. 429–452). Elsevier. https://ebookcentral.proquest.com/lib/aalto-ebooks/reader.action?docID=5161869&ppg=464

Pelham, B. W. (2013). Intermediate statistics: A conceptual course. SAGE. https://www.sagepub.com/sites/default/files/upm-binaries/49259_ch_1.pdf

SPSS for Quantitative Data Analysis

Get started by downloading SPSS from (download.aalto.fi)

Suggested viewing for the TikTok generation:
https://edge.sagepub.com/field5e/spss-video-tutorials

SPSS workshop on Wednesday