Aalto University
School of Science

CS-E5250 Data-Driven Concept Design

5 ECTS

Quantitative Data Analysis
Assignment 2

Mika P. Nieminen
Markus Kirjonen
Mathias Hörlesberger
Frans Welin

CS-E5250@aalto.fi

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.

Quantitative Data Analysis

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

Mark Twain

This slide deck based on
Sheard, J. (2018). Quantitative data analysis.
In *Research Methods* (pp. 429–452). Elsevier.
https://doi.org/10.1016/B978-0-08-102220-7.00018-2

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

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 quantitative 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

148,5

$\overline{X}_6$= 184,3333

Analysis - Descriptive statistics

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

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

s^2 = sample variance

x_i = value of i^{th} element

$\overline{x}$ = sample mean

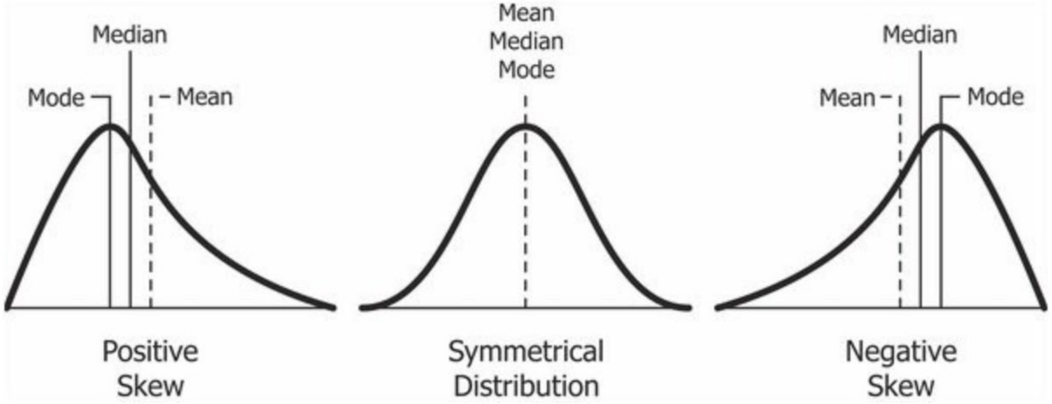
n = sample size

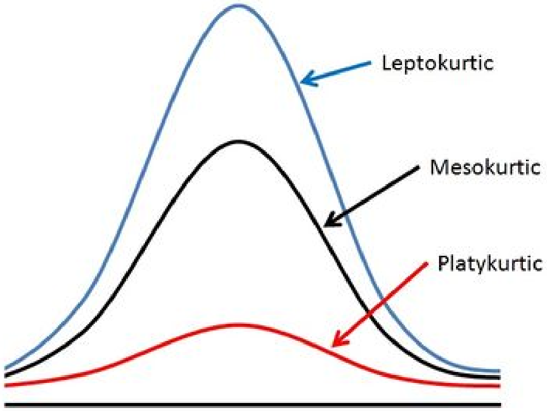
$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Analysis - Descriptive statistics

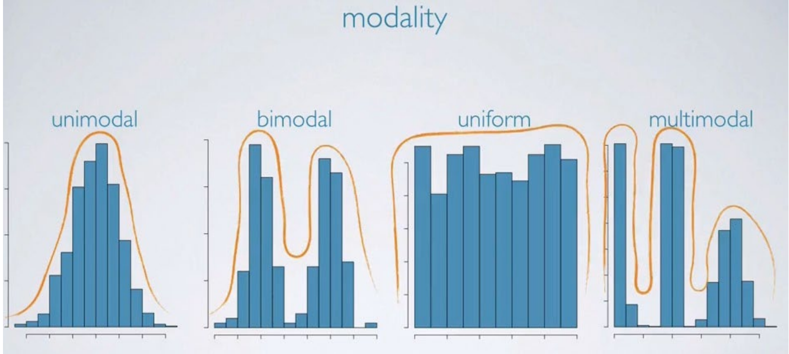
- Shape
  - Skewness
  - Kurtosis
  - Modality

By Diva Jain - https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=84219892





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



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

Analysis - Descriptive statistics

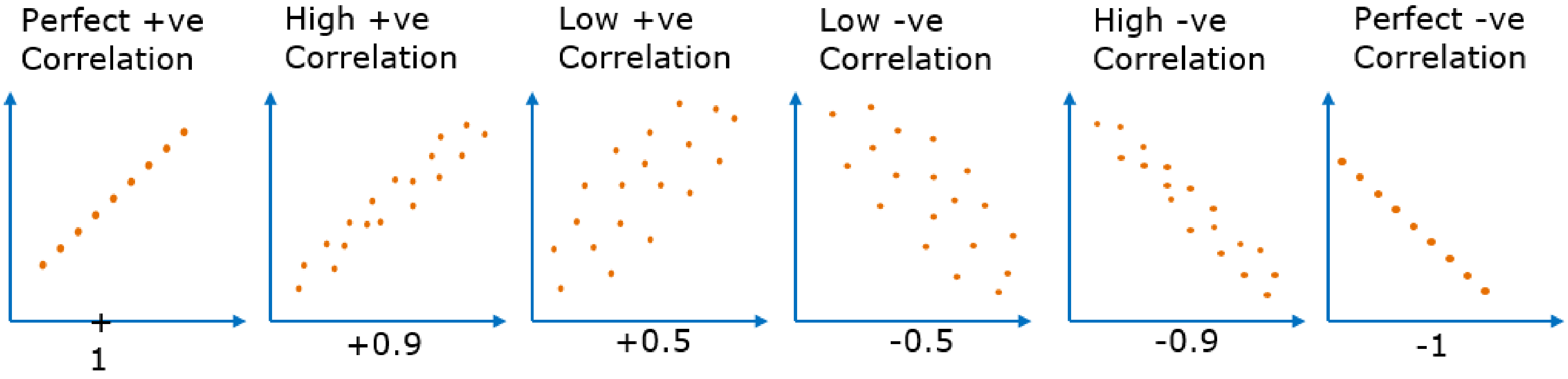
- r = Covariance standardized to std deviation

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

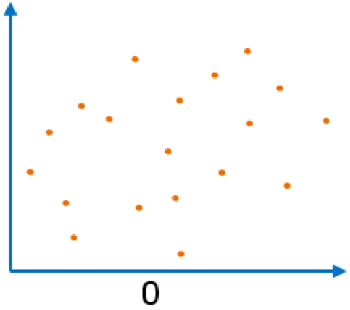
High +ve Correlation

Low +ve Correlation

Low -ve Correlation

High -ve Correlation

Perfect -ve Correlation

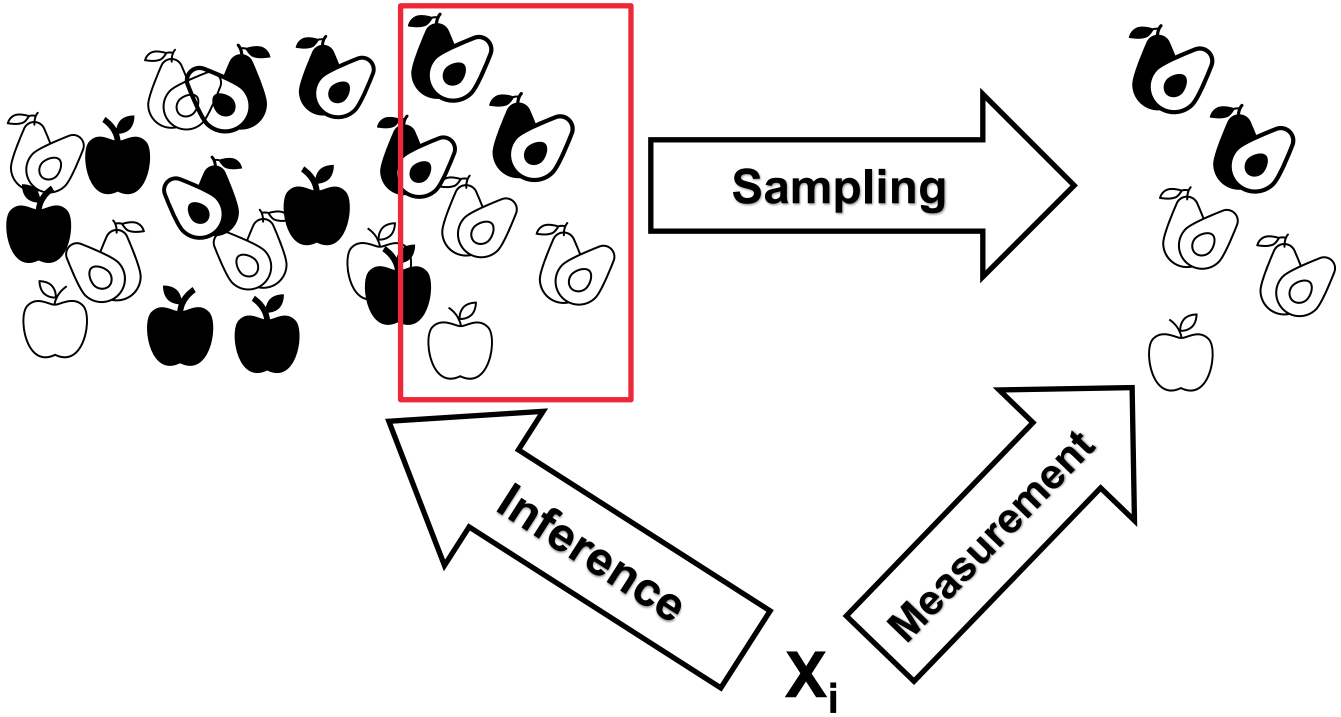


1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

www.ashutoshtripathi.com

Analysis - Inferential statistics

- Conclude or infer population characteristics from samples

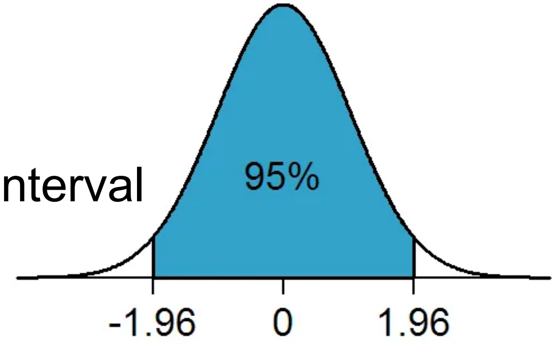


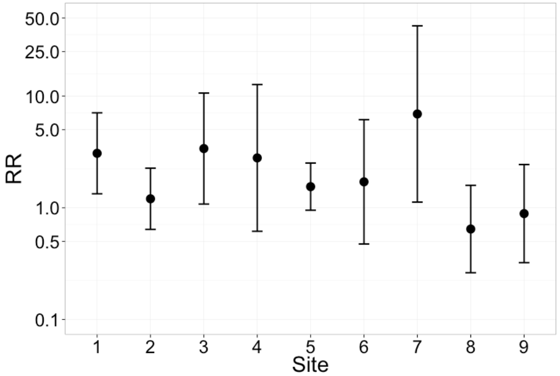
Analysis - Inferential statistics

- 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

Data measures:
Scale
Nominal
Ordinal

- 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
  - *Independent Samples T-test (no difference in the means of two distributions)*
  - *Paired Samples T-test (no difference in the means of matched pairs values from two occasions)*
  - *ANOVA (no difference in the means of values of two or more distributions)*

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

Mann-Whitney U Test is a nonparametric version of the independent samples t-test
Wilcoxon Signed Rank Test is a nonparametric counterpart of the paired samples t-test
Kruskal-Wallis Test is a nonparametric alternative to the one-way ANOVA

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
- Correlation shows a relationship that is associative, but tells nothing about causality

Analysis -
Testing for relatio

- Correlation analyses
  - Pearson:
    - Both distributions on t
    - Normal distributed po
    - Linear relationship
    - Homoscedasticity - va
  - Spearman rho:
    - Both distributions on t
    - Relationship between

- Correlation shows a relationship that is associative, but tells nothing about causality

Recap: the quantitative analysis process

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

Recommended reading

Chapter 18

Quantitative data analysis

Judithe Sheard
Monash University, Australia

This chapter provides a brief introduction to quantitative data analysis, explaining processes for the preparation of quantitative data and common techniques for data description and analysis. The context for the examples used in the chapter is a project management website, used by information technology (IT) students in the third year of their undergraduate degree. The description of data preparation includes data cleaning, as well as data transformation and integration. Statistics, used to investigate numerical data, include descriptive statistics for describing distributions and relationships between variables, and inferential statistics. The latter cover testing for differences between distributions using parametric and nonparametric tests (according to the type of data involved). Examples of different types of tests are provided, using the project management website dataset which is the focus of the chapter.

Research Methods: Information, Systems, and Contexts. **DOI:** http://dx.doi.org/10.1016/B978-0-08-102220-7.00018-2
Copyright © 2018 Judithe Sheard. Published by Elsevier Ltd. All rights reserved.

Sheard, J. (2018). Quantitative data analysis. In *Research Methods* (pp. 429–452). Elsevier.

A Review of Basic Statistical Concepts 1

The record of a month's roulette playing at Monte Carlo can afford us material for discussing the foundations of knowledge.

—Karl Pearson

I know too well that these arguments from probabilities are imposters, and unless great caution is observed in the use of them, they are apt to be deceptive.

—Plato (in *Phaedo*)

Introduction

It is hard to find two quotations from famous thinkers that reflect more divergent views of probability and statistics. The eminent statistician Karl Pearson (the guy who invented the correlation coefficient) was so enthralled with probability and statistics that he seems to have believed that understanding probability and statistics is a cornerstone of human understanding. Pearson argued that statistical methods can offer us deep insights into the nature of reality. The famous Greek philosopher Plato also had quite a bit to say about the nature of reality. In contrast to Pearson, though, Plato was skeptical of the "fuzzy logic" of probabilities and central tendencies. From Plato's viewpoint, we should only trust what we can know with absolute certainty. Plato probably preferred deduction (e.g., If *B* then *C*) to induction (In my experience, bees seem to like flowers).

Even Plato seemed to agree, though, that if we observe "great caution," arguments from probabilities may be pretty useful. In contrast, some modern nonstatisticians might agree with what the first author's father, Bill Pelham, used to say about statistics and probability theory: "Figures

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 tomorrow at 10:15