# Quantitative Data Analysis: A Companion for Accounting and Information Systems Research

## Teaching Materials

*Created by Willem Mertens, Amedeo Pugliese & Jan Recker*

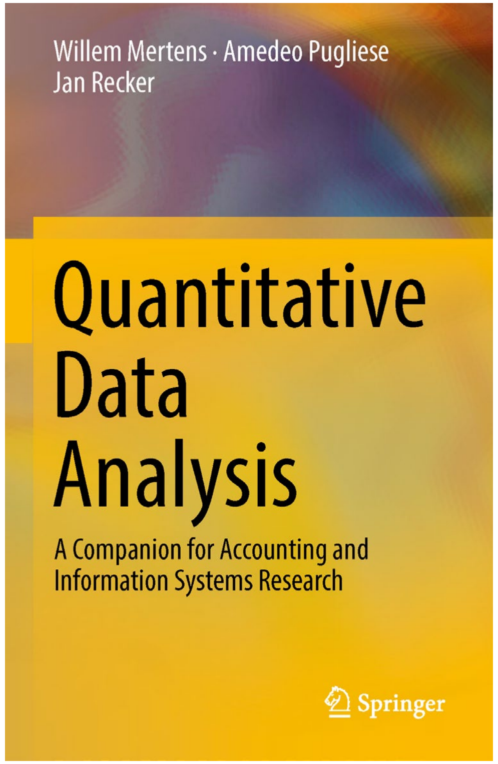
  

# Copyright Notice

# What these materials are about

Offering a guide through the essential steps required in quantitative data analysis

1. **Introduction**
2. Comparing Differences Across Groups
3. Assessing (Innocuous) Relationships
4. Models with Latent Concepts and Multiple Relationships: Structural Equation Modeling
5. Nested Data and Multilevel Models: Hierarchical Linear Modeling
6. Analyzing Longitudinal and Panel Data
7. Causality: Endogeneity Biases and Possible Remedies
8. How to Start Analyzing, Test Assumptions and Deal with that Pesky p-Value
9. Keeping Track and Staying Sane



# Part 1: **Exploring Data and Testing Assumptions**

# Warning

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

Benjamin Disraeli

*Statistics are no substitute for judgment.*

Henry Clay

# Agenda

1. Exploring Data
   - Structuring data
     - Basics
     - Variable types
   - Cleaning data and eliminating outliers
   - Visualising data

2. Understanding data
   - Distributions, means and standard deviations
   - Models and significance
   - Correlations and differences

3. Testing assumptions
   - Independence
   - Homoscedasticity
   - Normality
     - Skew and kurtosis
     - Transformations

4. Scales and factors
   - Basics
   - PCA/EFA vs. CFA

# Structuring data

## 1. Exploring data

- One row per case, one variable per column

| | Age | Gender | Role | … |
|---|---|---|---|---|
| Person 1 | 19 | F | Student | … |
| Person 2 | 53 | F | Professor | … |
| Person 3 | 27 | M | Admin | … |
| … | … | … | … | … |

- Depends on unit of analysis (e.g. person)

# Structuring data

## 1. Exploring data

- Nested data

| | Age | Gender | Role 1 | Role 2 | Role 3 |
|---|---|---|---|---|---|
| Person 1 | 19 | F | Student | Tutor | - |
| ~~Person 1~~ | ~~19~~ | ~~F~~ | ~~Tutor~~ | | |
| Person 2 | 53 | F | Professor | Head of School | Supervisor |
| ~~Person 2~~ | ~~53~~ | ~~F~~ | ~~Head of School~~ | | |
| ~~Person 2~~ | ~~53~~ | ~~F~~ | ~~Supervisor~~ | | |
| Person 3 | 27 | M | Admin | - | - |
| … | … | … | … | … | … |

# Structuring data

## 1. Exploring data

- Recoding data: variable types
  - Categorical variables
    - Nominal (e.g. role)
    - Dichotomous (e.g. gender)
    - Ordinal (e.g. hierarchical level)
  - Continuous variable
    - Interval (e.g. degrees): 5-10 = 15-20
    - Ratio (e.g. weight): O is nothing, 10 = 2*5

| | Age | Gender | Role 1 | Role 2 | Role 3 |
|---|---|---|---|---|---|
| Person 1 | 19 | 1 | Student | Tutor | - |
| Person 2 | 53 | 1 | Professor | Head of School | Supervisor |
| Person 3 | 27 | 2 | Admin | - | - |
| … | … | … | … | … | … |

# Cleaning data and eliminating outliers

## 1. Exploring data

- Cleaning data

  = Taking out *unreliable* (not inconvenient) cases

    - Missing data (or listwise/pairwise)
    - Extreme tendencies (e.g. all 6/all 1)
    - Improbable response time (e.g. outliers)
    - Inconsistent responses (e.g. age < tenure)

  ≠ Introducing bias

    - Consistent application of rules
    - Mindful of hypotheses and method (IV/DV)

  - Consider power and credibility

# Cleaning data and eliminating outliers

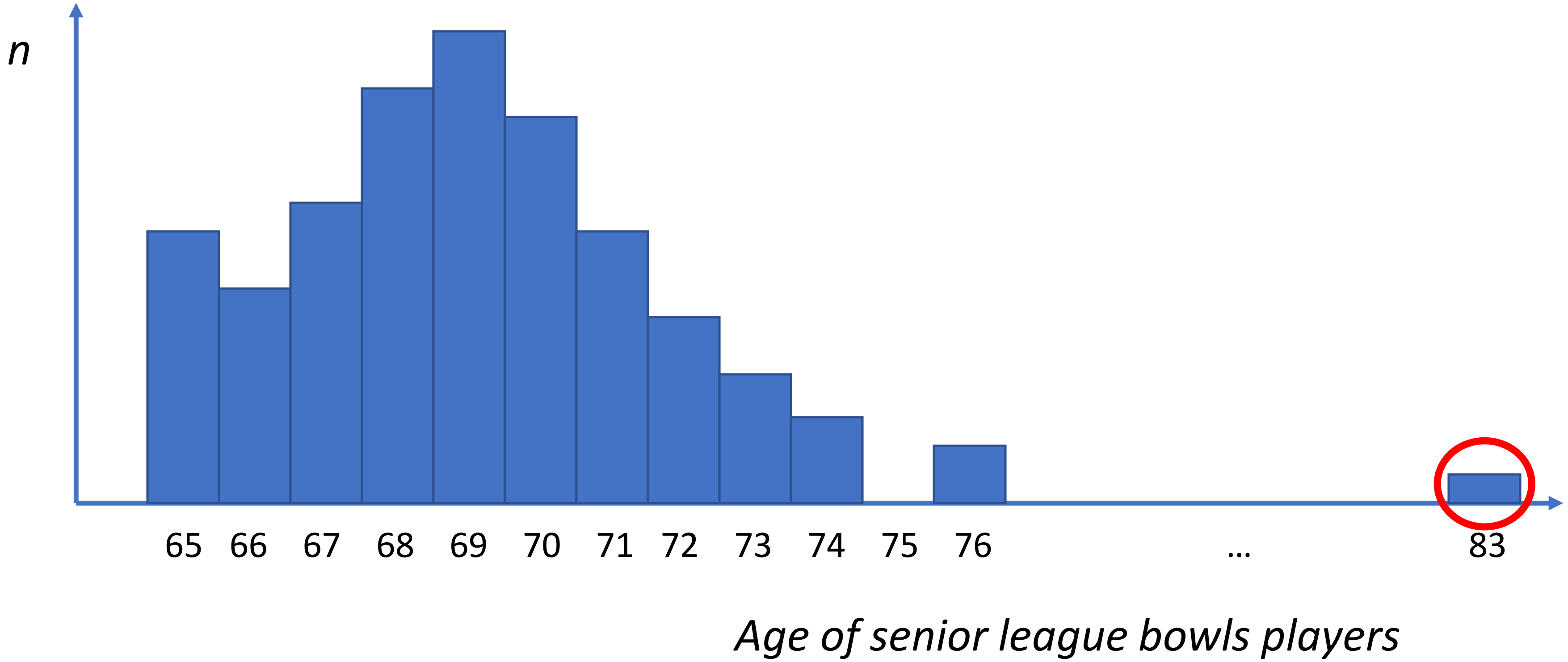
## 1. Exploring data

- Eliminating outliers
  - Outliers are highly improbable or erroneous values
    - They can influence statistics --> introduce bias
    - They affect generalizability
    - The decision to exclude depends on the RQs
  - How to find outliers
    - Box-plots
    - Histograms
    - Scatter plots
    - *z*-scores <-3.29 or >3.29 (see slide 16)

# Visualising data

## 1. Exploring data

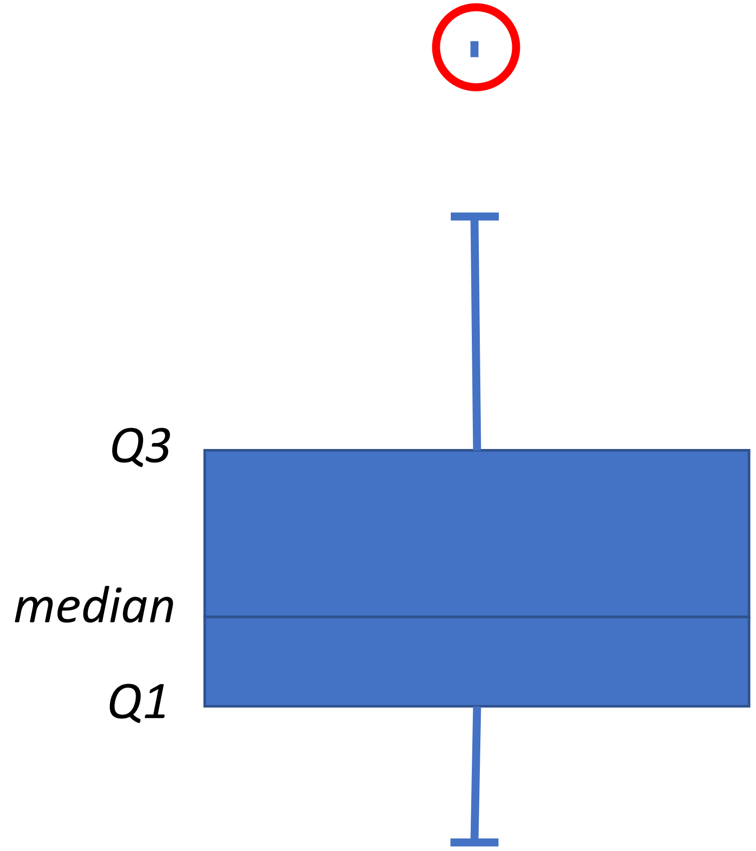
- Histograms



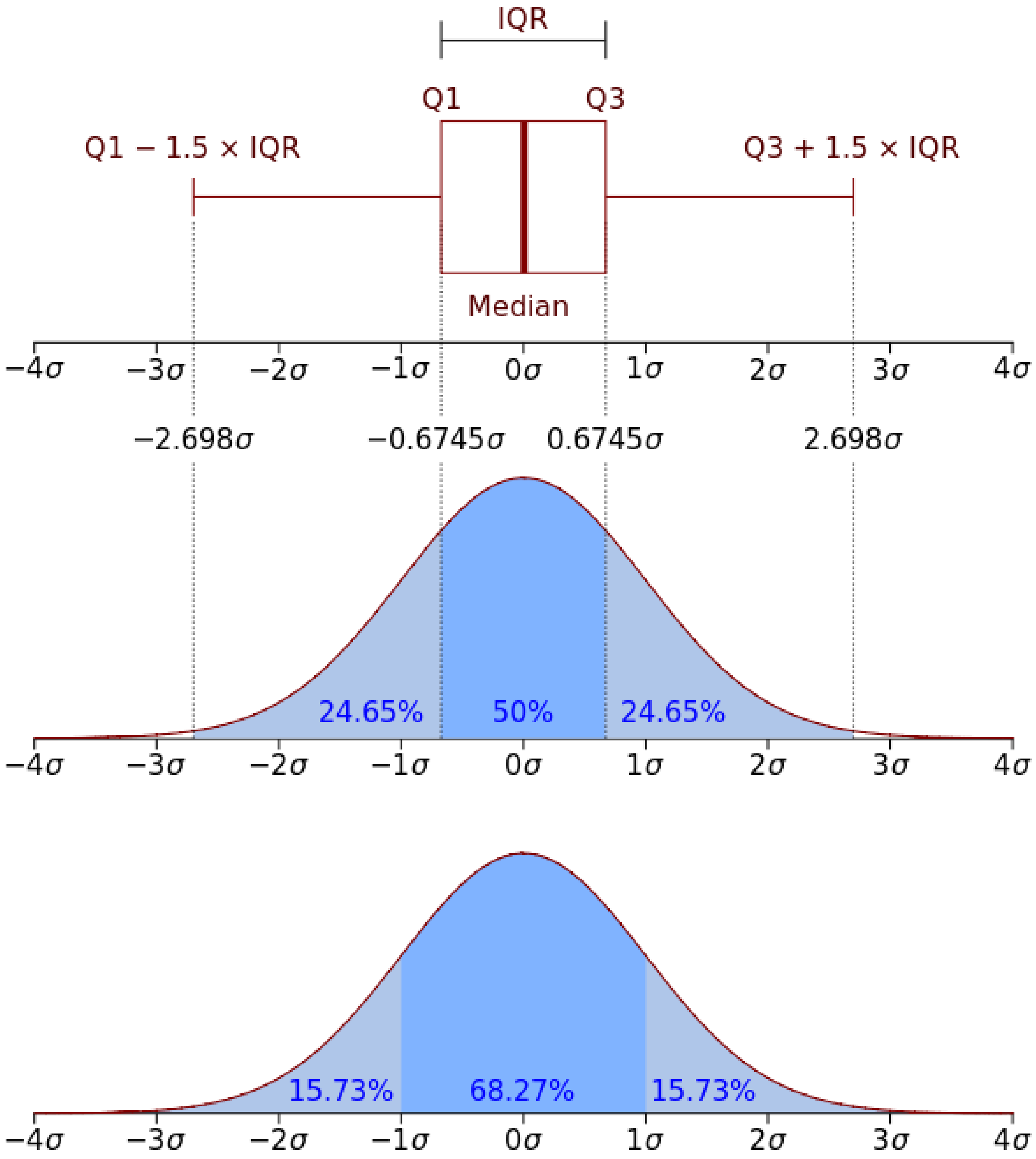
# Visualising data

## 1. Exploring data

- Box plots



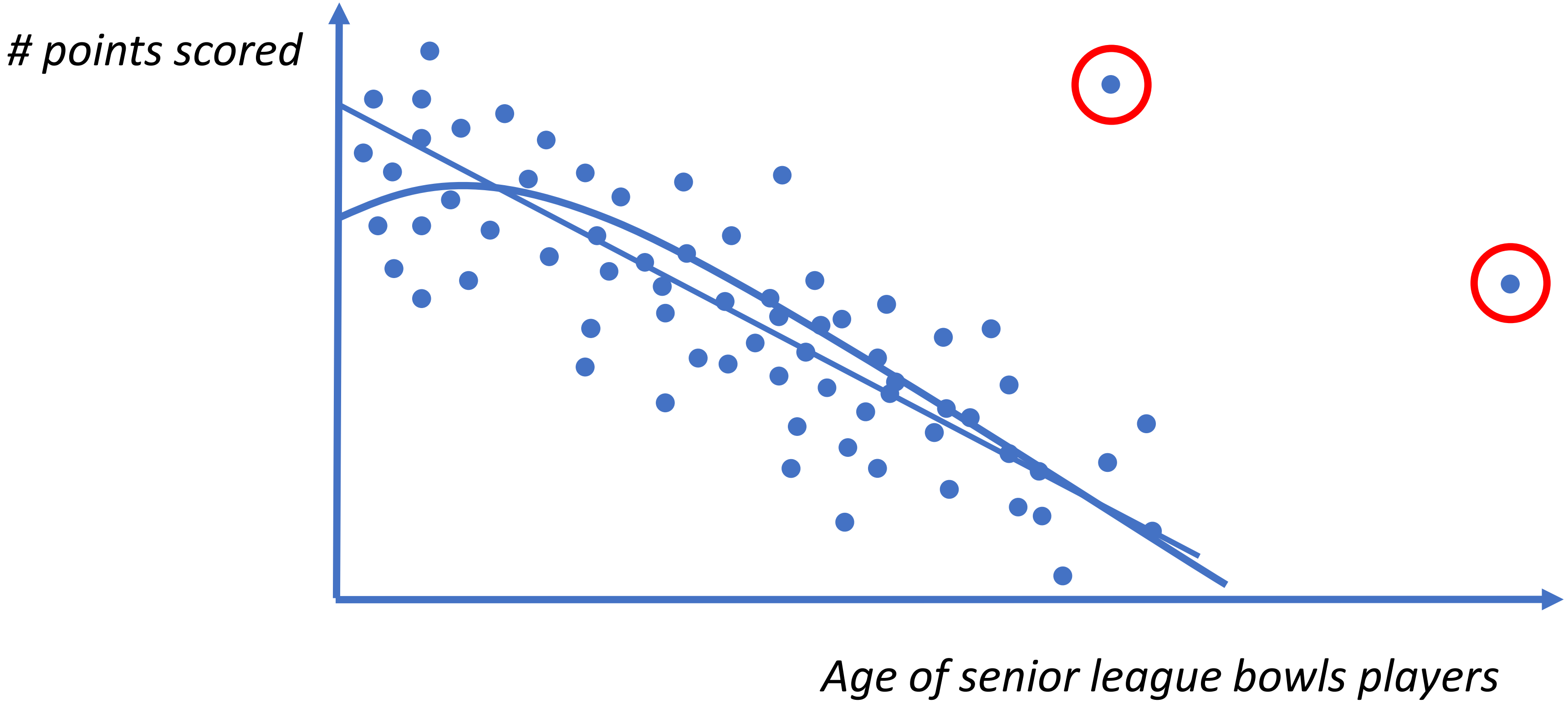

"Boxplot vs PDF" by Jhguch - Wikipedia



# Visualising data

## 1. Exploring data

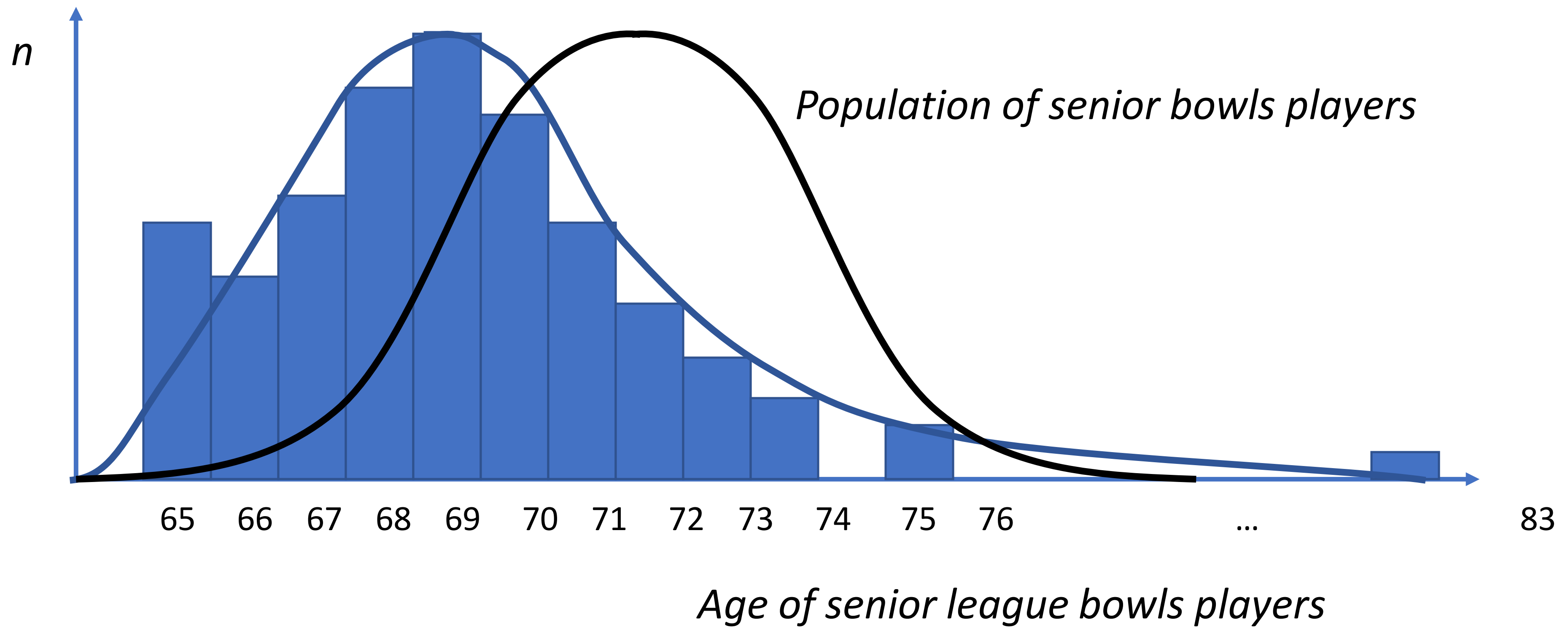
- Scatter plots



# Distributions, means and standard deviations

## 2. Understanding data

- Frequency distributions



# Distributions, means and standard deviations

## 2. Understanding data

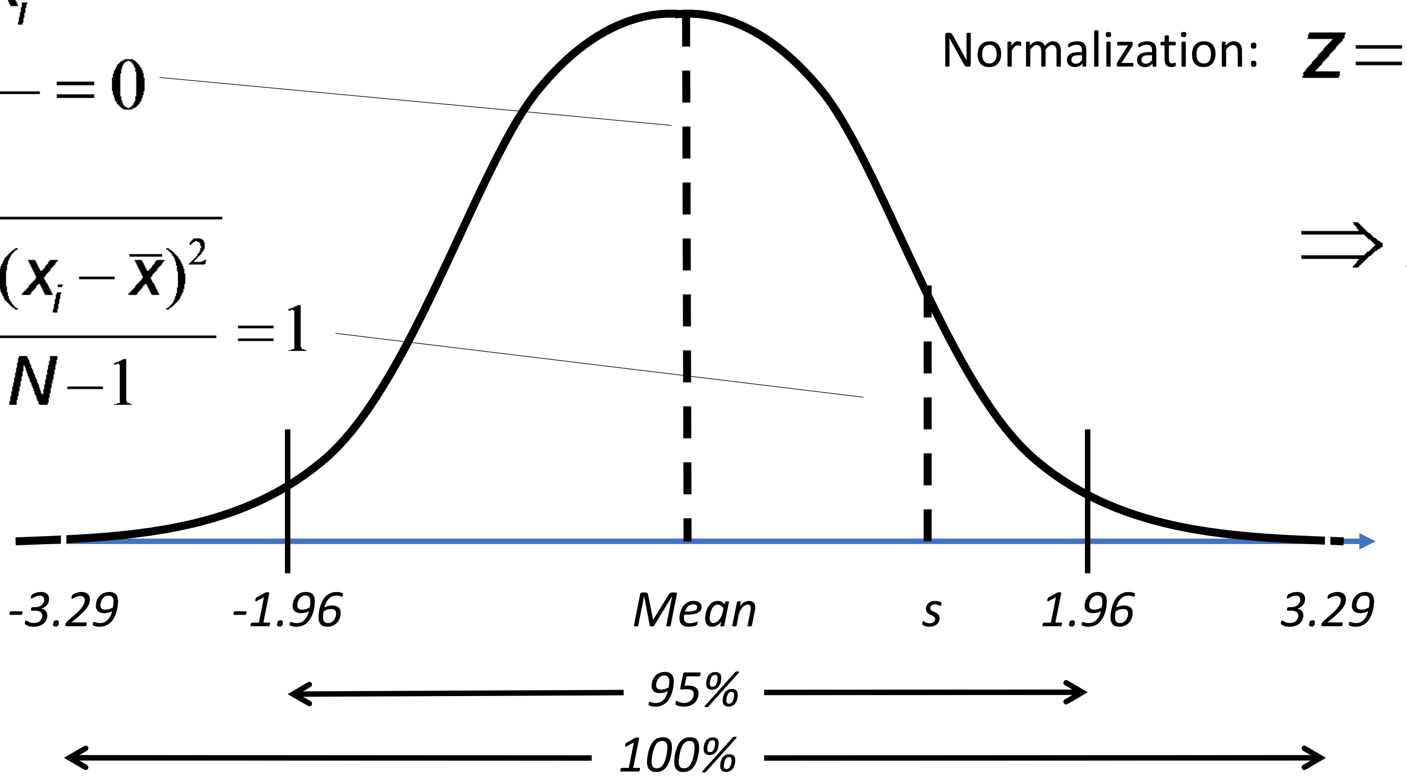
- Probability distributions - e.g.: normal distribution

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 0$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} = 1$$

Normalization: $$z = \frac{X - \bar{X}}{s}$$

$$\Rightarrow X = s * z + \bar{X}$$



# Models and significance

## 2. Understanding data

- Models
  - Attempt to explain/summarise data
  - Vary in how well they "fit" the data
    - E.g.: mean is a model; *s* illustrates fit
  - Fit
- Significance
  - Hypothesis testing involves comparing two models ($H_0$ vs. $H_1$)
  - Comparing models is done using test statistics:
    variance explained by the model/variance not explained by the model
  - If the probability of observing this test statistic, or anything more extreme, is smaller than .05/.01/.001, then we conclude statistical significance (i.e. $H_1$ explains the data better than $H_0$)

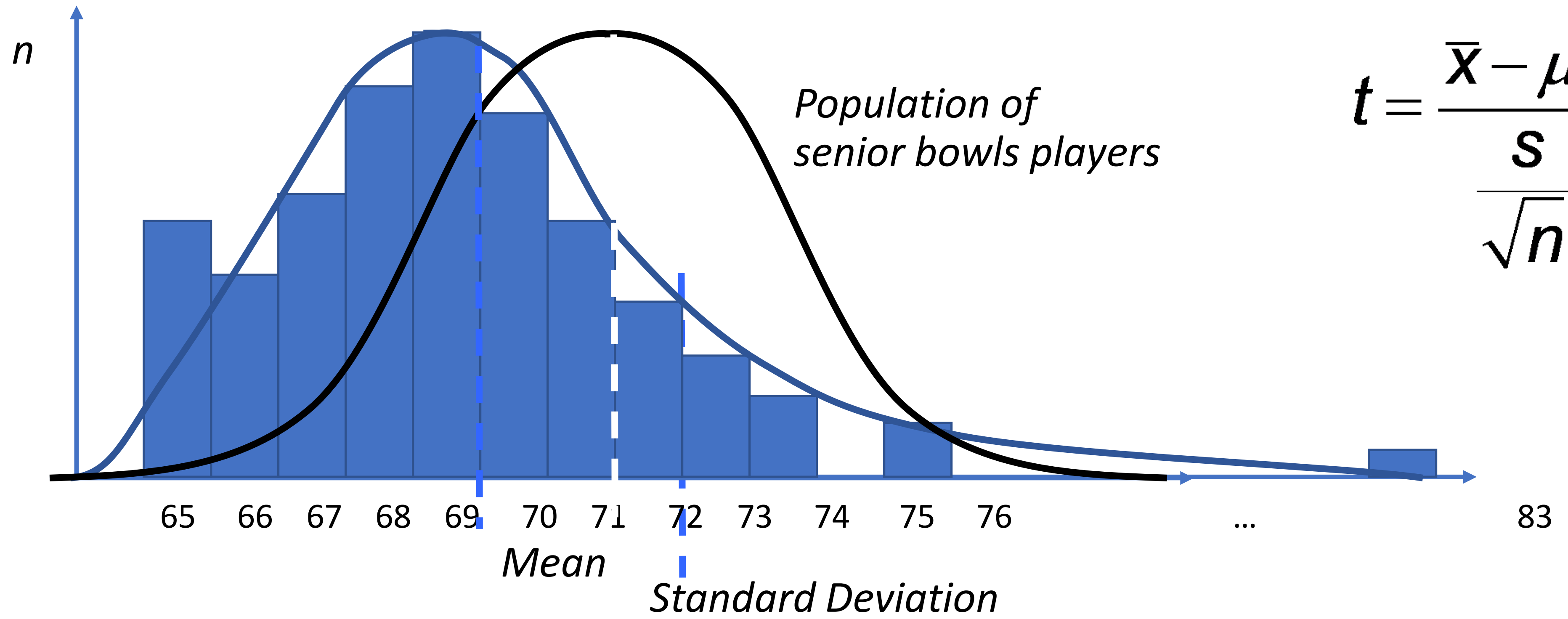
**Significance ≠ importance**

**Non-significance does not say anything about $H_0$**

# Correlations and differences

## 2. Understanding data

- Example of a model/hypothesis test: difference between means = *t*-test



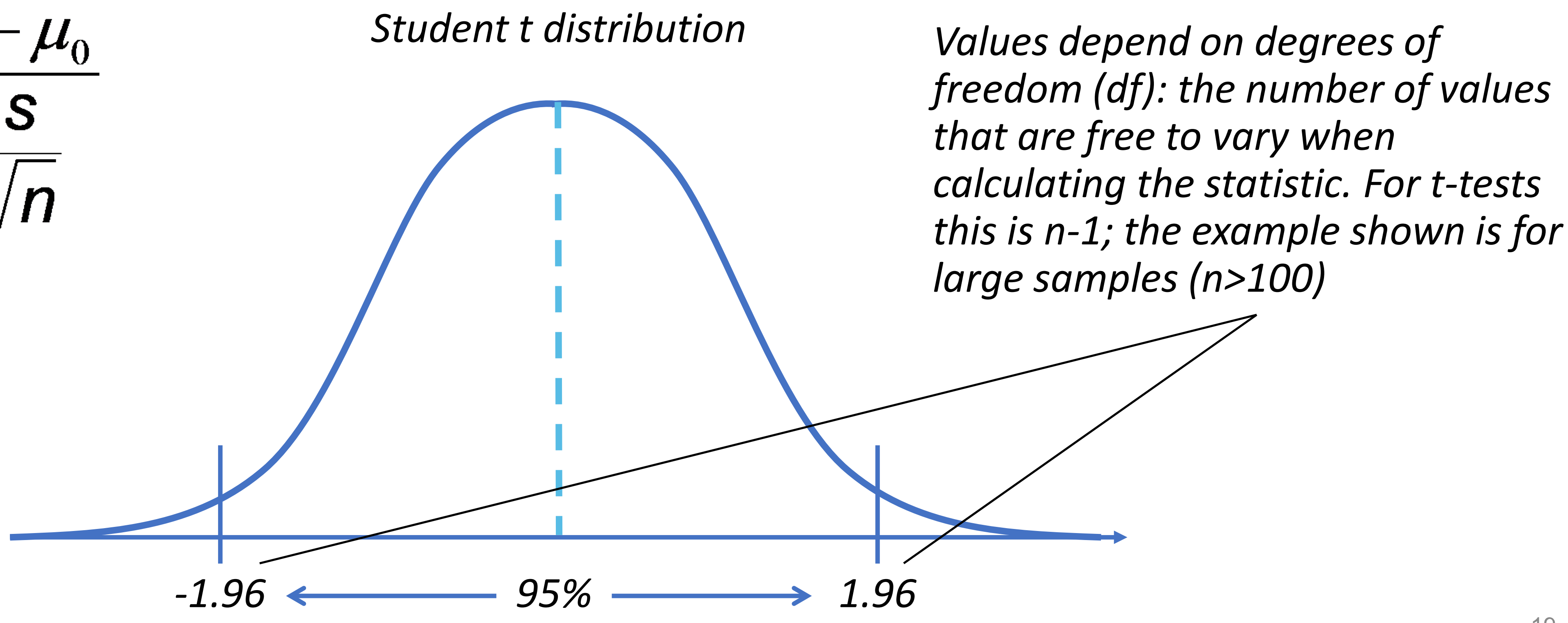

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

# Correlations and differences

## 2. Understanding data

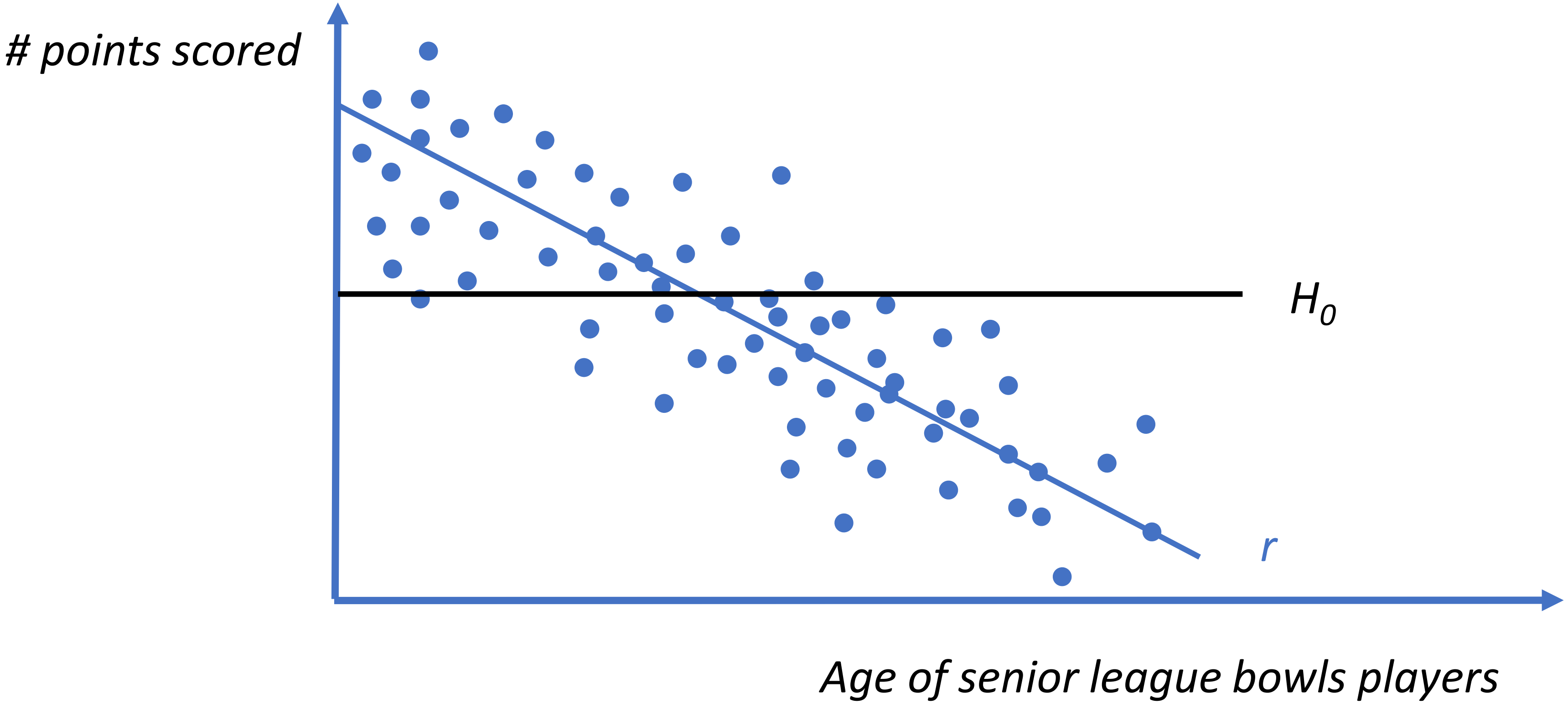
- Example of a model/hypothesis test: difference between means = *t*-test

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$



# Correlations and differences

## 2. Understanding data



# Most common assumptions for linear analyses

## 3. Testing assumptions

- Independence
  - Data was collected from independent sources
  - Variable measurements were independent (e.g. regression)
- Homoscedasticity/homogeneity of variance
  - Variance is equal in different (sub-)samples
- Normality
  - Sampling distribution/errors/data follow a normal distribution --> have limited skew and kurtosis

# Independence

## 3. Testing assumptions

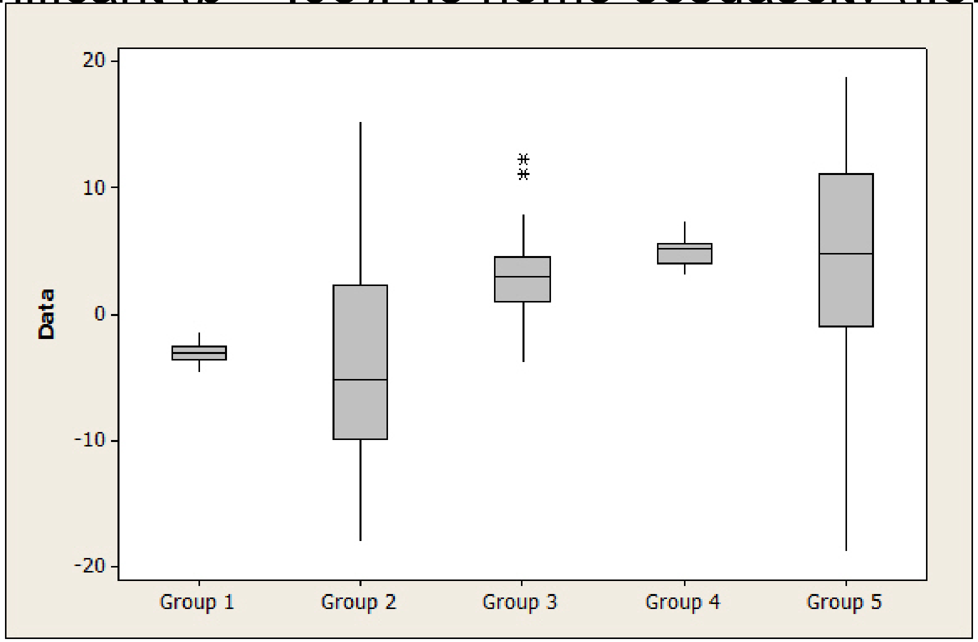
- Data was collected from independent sources
  - No repeated measures
  - No mutual influence between participants
  - No nested structures (see HLM module)

- Variable measurements were independent
  - No priming, framing, context or other question order effects
  - In regression-based models:
    - Variables are unrelated to external (exogenous) variables
    - Errors are independent

# Homoscedasticity/homogeneity of variance



- One variable, multiple groups (e.g. *t*-test): spread of values is equal across different groups
  - Visual test: scatter- or boxplot
  - Statistical test: Levene's test for equality of variance
    - When significant ($p < .05$): no homo-scedascity (i.e. heteroscedascity)

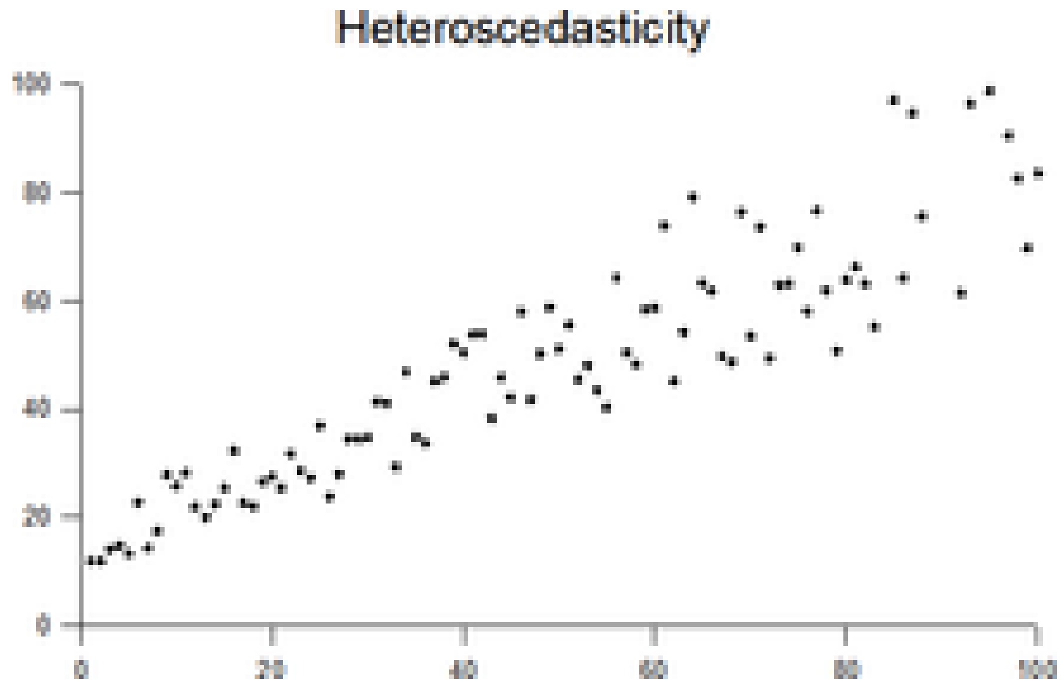


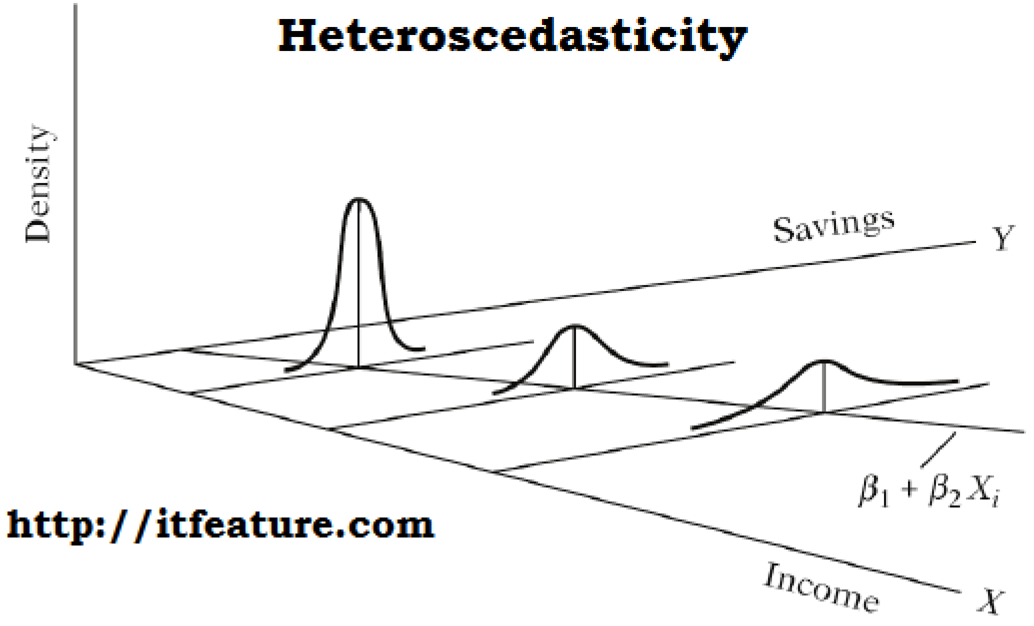

**Levene's test will usually be significant in large samples; use other tests (e.g. Hartley's $F_{max}$)**

# Homoscedasticity/homogeneity of variance

## 3. Testing assumptions

- Two variables (e.g. regression): spread of errors/residuals is equal across different values of *x*

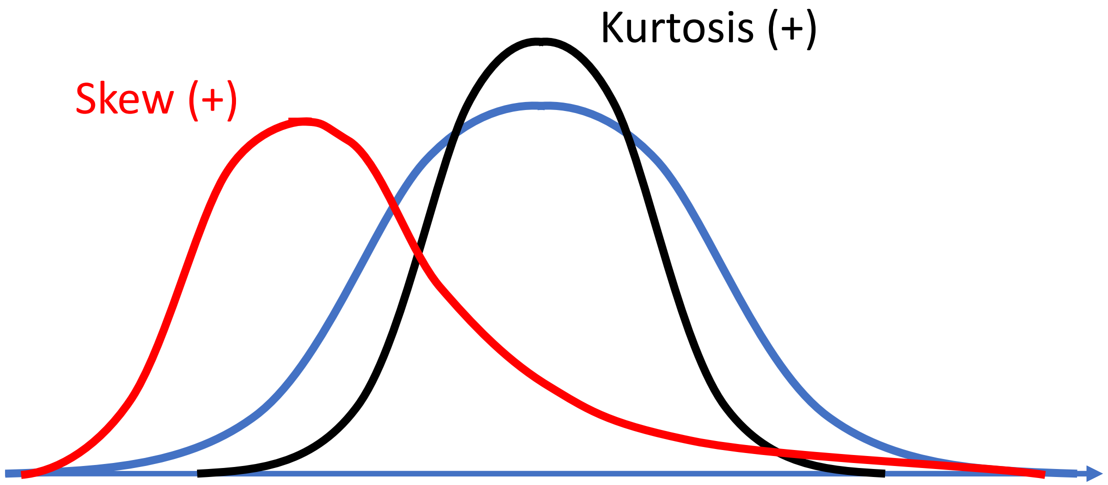


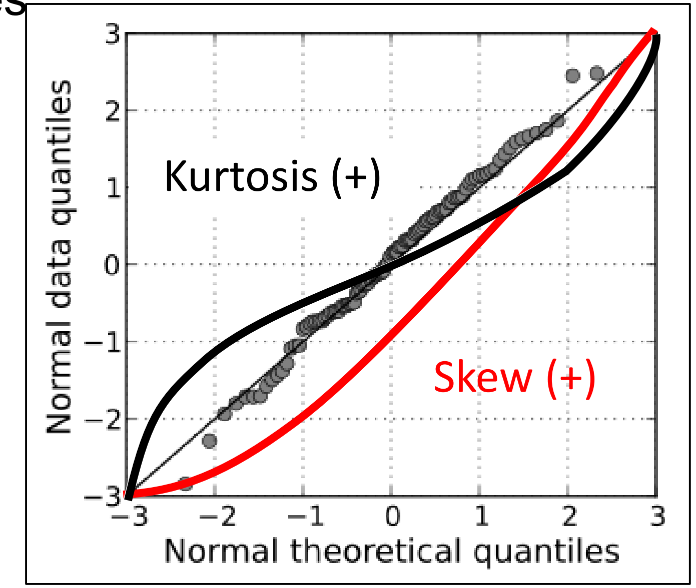



# Normality

## 3. Testing assumptions

- In many statistical tests
  - Sampling distribution is normally distributed
    --> test normality of sample
    - Visually testing normality of (sub-)sample data
      - Histograms (see slide 10)
      - Q-Q plots: theoretical vs. actual quantiles




"Normal normal qq" by Skbkekas - Wikipedia

# Normality

## 3. Testing assumptions

- Statistical tests for normality of (sub-)sample data
  - Compute descriptives including skew and kurtosis
  - Convert skew and kurtosis to *z*-scores, e.g.:

$$z_{skewness} = \frac{skewness - 0}{SE_{skewness}} \Rightarrow \frac{|\,skewness\,|}{SE_{skewness}} \text{ must be} \leq 1.96$$

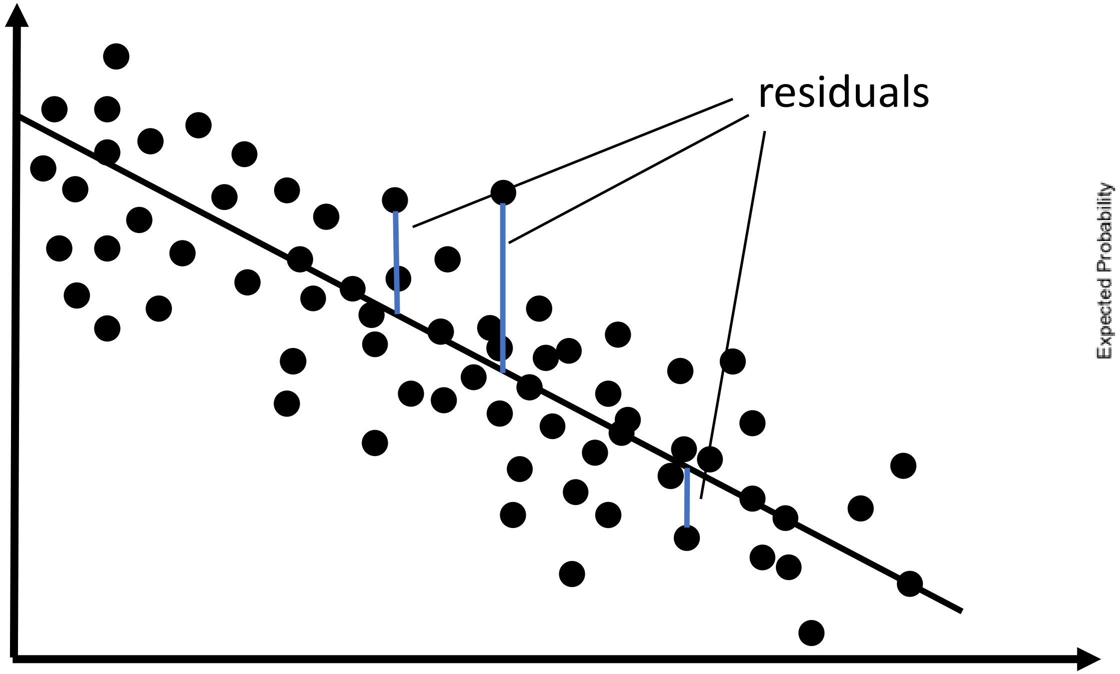
**Increase to 2.58 in larger samples and do not use in very large samples (*n* > 200)**

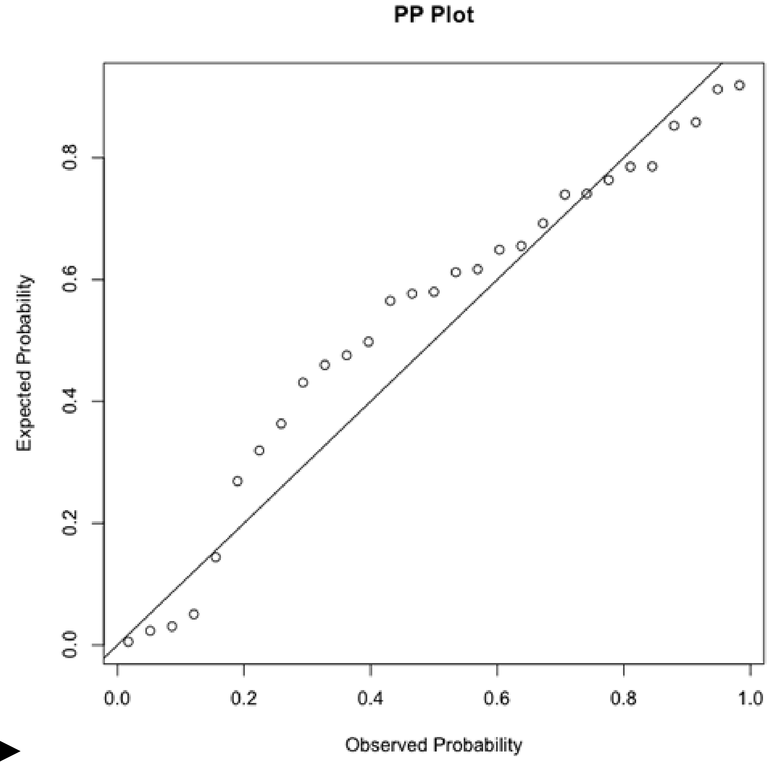
  - Shapiro-Wilk test: significant ($p < .05$) when NOT normal

# Normality

## 3. Testing assumptions

- In regression-based models
  - Errors/residuals, not indicators need to be normally distributed
  - Same visual principles as Q-Q plot apply




Please note: in this case, both graphs do not represent the same data

# What if assumptions are violated?

## 3. Testing assumptions

- Correct data
  - Exclude outliers
  - Transform data, e.g.:
    - Log-, square root and reciprocal (1/x) transformations shorten the right tale (i.e. correct positive skew)
    - The same transformations applied to the reverse score (score – highest score + 1) correct for negative skew

⚠ **The same transformation has to be applied to variables that are compared directly**

- Turn to tests that are robust against violations or to non-parametric tests, e.g.
  - Mann–Whitney U for group comparisons
  - Kendall's tau for dependence between two variables

# Scales and factors - basics



- Scales are sets of indicators that measure the same latent variable / factor

  ≠ response scales!

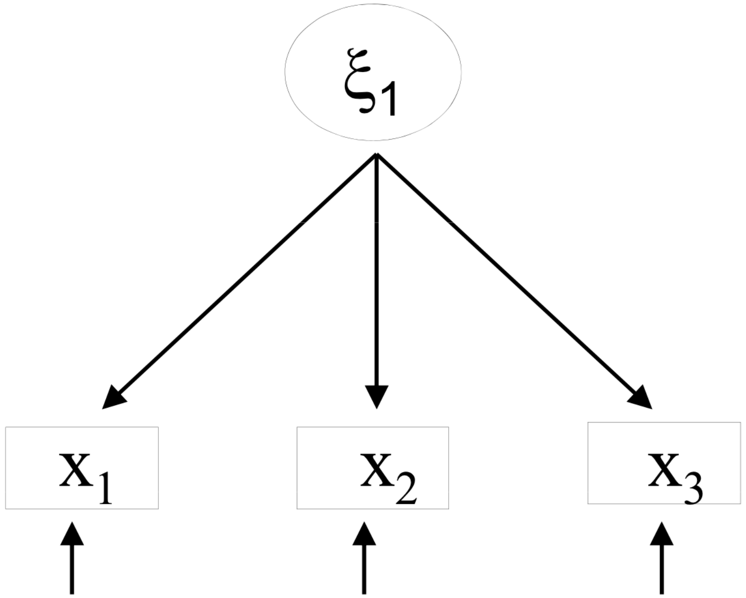
- E.g. To aid me in my teaching, overall, I feel Powerpoint … is:
  - Easy to Learn
  - Easy to manipulate
  - Clear to interact with
  - Flexible to interact with
  - Difficult to master (reverse scored)
  - Very cumbersome (reverse scored)

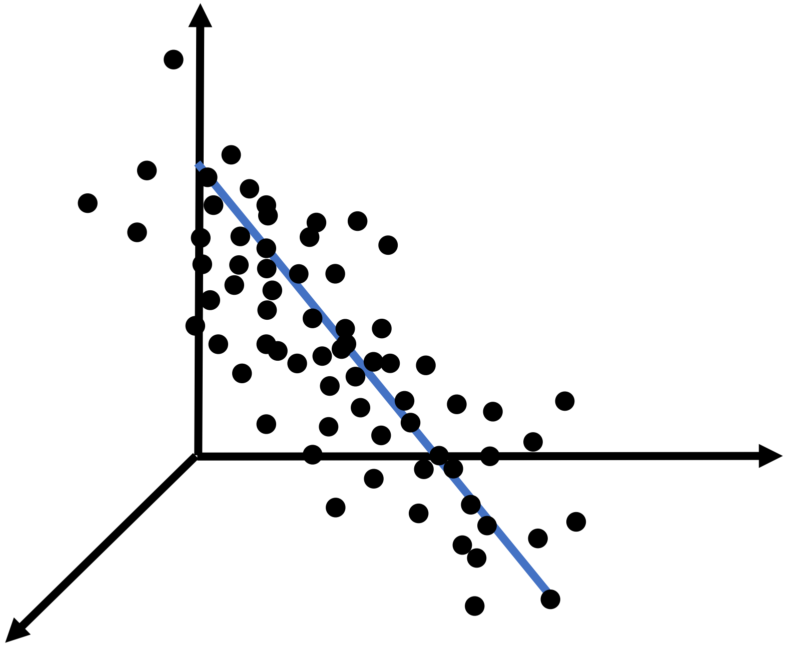
  } Ease of use

# Scales and factors - basics

## 4. Scales and factors

- Visualisation of scale with three indicators measuring one latent variable / factor:



# Principal component analysis

## 4. Scales and factors

- Run PCA with no restriction on the number of factors and with a scree plot
- Decide how many factors to retain based on eigenvalues, scree plot and $R^2$
  - Separate mountain from scree
  - Eigenvalue > 1
    - Eigenvalue: proportion of variance explained by factor (sum = # variables)
  - Cumulative $R^2$ > .6

# Principal component analysis

## 4. Scales and factors

- Run PCA again
  - Restrict the number of extracted factors
  - Rotate factors orthogonally or oblique based on theory (or trial and error/inspection of the component correlation matrix)
  - Study the component matrix (orthogonal) or pattern matrix (oblique) to interpret factors and exclude indicators when
    - Loading is small (< .4/.7) on all factors
    - Loadings are high for multiple factors (> .4/.7)
    - Difference between loadings on different factors < .2
  - Run PCA again after each exclusion

# Principal component analysis

## 4. Scales and factors

- Once a stable solution has been reached, evaluate reliability and uni-dimensionality of scales
  - Inter-item correlation when # indicators for factor is 2
    - Should be significant
  - Chronbach’s Alpha when # indicators for factor is > 2
    - Should be higher than .7
    - “Alpha if item deleted” should be lower than Alpha
      - If not: exclude item and run PCA again

# End of Part 1