UNIVERSITÄT BIELEFELD
Faculty of Technology

Introduction to Data Mining
03 - Statistical Testing

Benjamin Paaßen

WS 2023/2024, Bielefeld University

Exercise Prep

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)

Exercise Prep

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)
- ChatGPT/external help policy: language models are permitted IF you tell us where you used it (indicate in your solution) and you could still present the solution yourself; ideally: link the chatlog

Exercise Prep

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)
- ChatGPT/external help policy: language models are permitted IF you tell us where you used it (indicate in your solution) and you could still present the solution yourself; ideally: link the chatlog Why?

Exercise Prep

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)
- ChatGPT/external help policy: language models are permitted IF you tell us where you used it (indicate in your solution) and you could still present the solution yourself; ideally: link the chatlog Why?
    - sometimes, language models get it wrong and you need to be able to check the result

Exercise Prep

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)
- ChatGPT/external help policy: language models are permitted IF you tell us where you used it (indicate in your solution) and you could still present the solution yourself; ideally: link the chatlog Why?
  - sometimes, language models get it wrong and you need to be able to check the result
  - you need to take full responsibility for correctness of the result

Exercise Prep

- First sheet is out TODAY with deadline friday in 11 days (Nov 03, noon, i.e. 12:00)
- ChatGPT/external help policy: language models are permitted IF you tell us where you used it (indicate in your solution) and you could still present the solution yourself; ideally: link the chatlog Why?
  - sometimes, language models get it wrong and you need to be able to check the result
  - you need to take full responsibility for correctness of the result
  - we want to understand how students use the available tools

Example

B

Write python code to automatically identify outliers, which are defined as any students with a pre-test score higher than 3 standard deviations above the mean. Write python code that removes these outliers from the data set.

Example

UNIVERSITÄT BIELEFELD

Faculty of Technology

B

Write python code to automatically identify outliers, which are defined as any students with a pre-test score higher than 3 standard deviations above the mean. Write python code that removes these outliers from the data set.

You can identify and remove outliers from a dataset using Python with the help of libraries like NumPy and pandas. Here's an example code to do this:

python Copy code

```python
import numpy as np
import pandas as pd

# Generate a sample dataset (you can replace this with your own data)
data = {
    'StudentID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'PreTestScore': [85, 92, 88, 110, 60, 105, 95, 102, 80, 120]
}

# Create a DataFrame from the data
df = pd.DataFrame(data)

# Calculate the mean and standard deviation of the PreTestScore
mean_score = df['PreTestScore'].mean()
std_dev = df['PreTestScore'].std()

# Define a threshold for identifying outliers
threshold = mean_score + 3 * std_dev

# Identify outliers
outliers = df[df['PreTestScore'] > threshold]

# Remove outliers from the dataset
df = df[df['PreTestScore'] <= threshold]
```

Python introduction

- First tutorial sessions: Python introduction with our intelligent tutoring system (bring your own device!)

Python introduction

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First tutorial sessions: Python introduction with our intelligent tutoring system (bring your own device!)
- Who wants to take part today, right after the lecture?

Python introduction

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First tutorial sessions: Python introduction with our intelligent tutoring system (bring your own device!)
- Who wants to take part today, right after the lecture?
- Who wants to take part Thursday, 16-18

Python introduction

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First tutorial sessions: Python introduction with our intelligent tutoring system (bring your own device!)
- Who wants to take part today, right after the lecture?
- Who wants to take part Thursday, 16-18
- Who wants to take part next monday?

Python introduction

UNIVERSITÄT BIELEFELD
Faculty of Technology

- First tutorial sessions: Python introduction with our intelligent tutoring system (bring your own device!)
- Who wants to take part today, right after the lecture?
- Who wants to take part Thursday, 16-18
- Who wants to take part next monday?
- Who wants to take part next thursday?

Recap: Random variables

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We define random variables via a domain of possible values $\mathcal{X}$ and a probability mass/density function p assigning probability mass to each possible value

Recap: Random variables

- We define random variables via a domain of possible values $\mathcal{X}$ and a probability mass/density function p assigning probability mass to each possible value
- discrete random variables for finite/countably infinite sets; continuous random variables for (compact subsets of) the real numbers

Recap: Random variables

- We define random variables via a domain of possible values $\mathcal{X}$ and a probability mass/density function p assigning probability mass to each possible value
- discrete random variables for finite/countably infinite sets; continuous random variables for (compact subsets of) the real numbers
- marginal probabilities for single variables, joint/conditional probabilities for combinations of variables

Recap: Random variables

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We define random variables via a domain of possible values $\mathcal{X}$ and a probability mass/density function p assigning probability mass to each possible value
- discrete random variables for finite/countably infinite sets; continuous random variables for (compact subsets of) the real numbers
- marginal probabilities for single variables, joint/conditional probabilities for combinations of variables
- most important density functions for continuous variables (in this lecture): Gaussian (!), t-distribution

Recap: Random variables

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We define random variables via a domain of possible values $\mathcal{X}$ and a probability mass/density function p assigning probability mass to each possible value
- discrete random variables for finite/countably infinite sets; continuous random variables for (compact subsets of) the real numbers
- marginal probabilities for single variables, joint/conditional probabilities for combinations of variables
- most important density functions for continuous variables (in this lecture): Gaussian (!), t-distribution
- means of random variables tend to be Gaussian (central limit theorem)

Preamble: Study Design

Setup

Imagine you have invented a cool pedagogical intervention

Child with VR glasses by Julia M Cameron (Link); Usage according to pexels license.

Setup

Imagine you have invented a cool pedagogical intervention

Child with VR glasses by Julia M Cameron (Link); Usage according to pexels license.

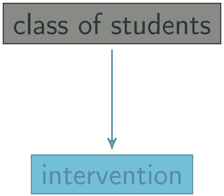
... but how do you know if it is effective?

Attempt 1: Single Class Study

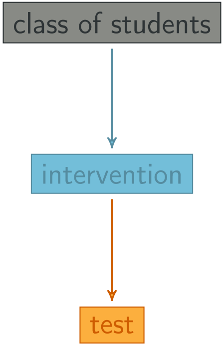
UNIVERSITÄT BIELEFELD
Faculty of Technology

class of students

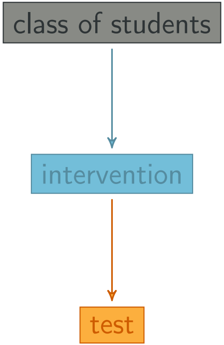
Attempt 1: Single Class Study



Attempt 1: Single Class Study

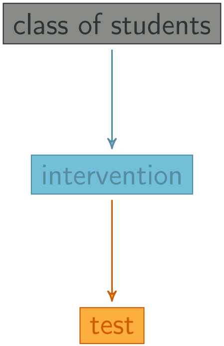


Attempt 1: Single Class Study



What are the problems here?

Attempt 1: Single Class Study



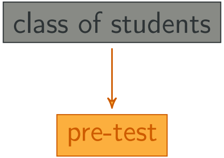
What are the problems here?

- Maybe the students already knew everything before the intervention

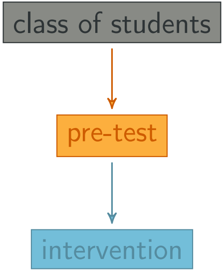
Attempt 2: Pre- and post-test

class of students

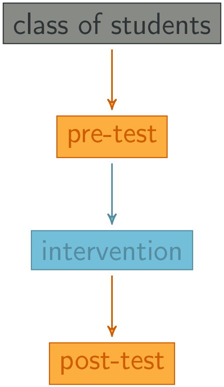
Attempt 2: Pre- and post-test



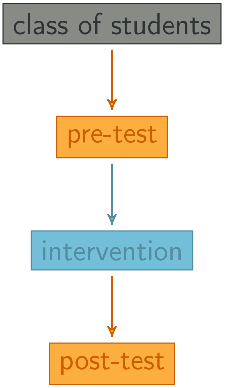
Attempt 2: Pre- and post-test



Attempt 2: Pre- and post-test



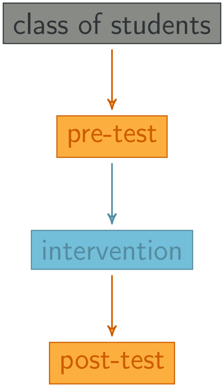
Attempt 2: Pre- and post-test



What are the problems here?

Attempt 2: Pre- and post-test

UNIVERSITÄT BIELEFELD
Faculty of Technology



What are the problems here?

- Maybe the learning is not due to the intervention but something else

Attempt 3: Controlled study

intervention group

Attempt 3: Controlled study

UNIVERSITÄT BIELEFELD

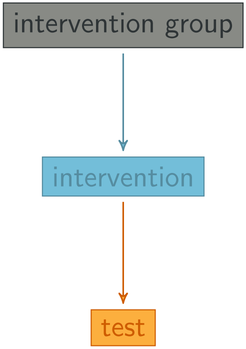
Faculty of Technology

intervention group

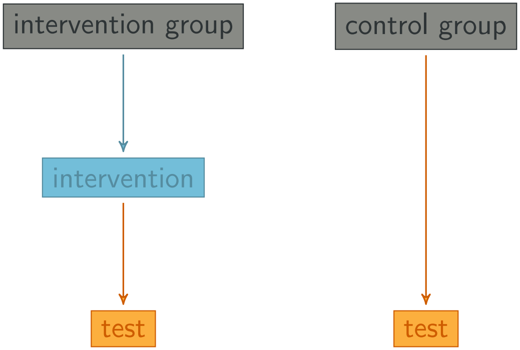
↓

intervention

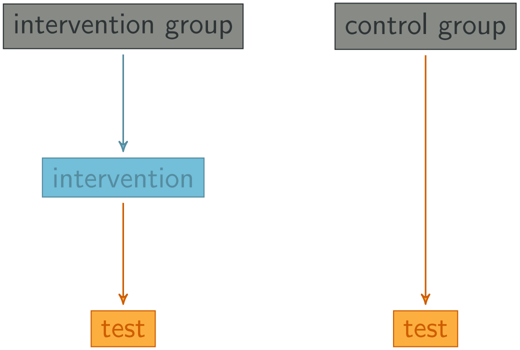
Attempt 3: Controlled study



Attempt 3: Controlled study



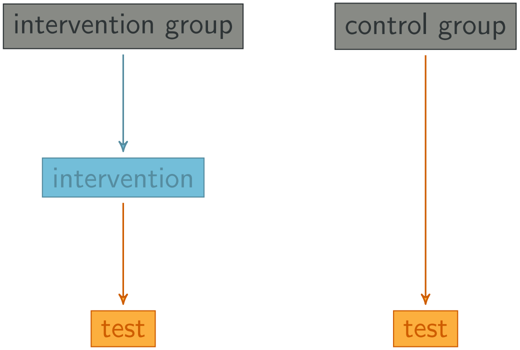
Attempt 3: Controlled study



What are the problems here?

Attempt 3: Controlled study

UNIVERSITÄT BIELEFELD
Faculty of Technology

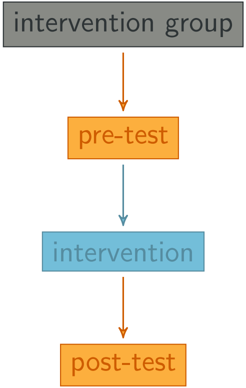


What are the problems here?

- Maybe students in the intervention condition are just better

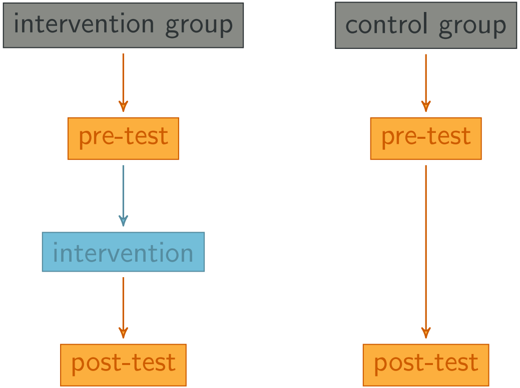
Attempt 4: Controlled pre-/post-test study

UNIVERSITÄT BIELEFELD
Faculty of Technology



Attempt 4: Controlled pre-/post-test study

UNIVERSITÄT BIELEFELD
Faculty of Technology



Further aspects

- Regression to the mean if intervention group are the (seemingly) weaker students

Further aspects

- Regression to the mean if intervention group are the (seemingly) weaker students
- Are the study participants representative of the underlying population?

Further aspects

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Regression to the mean if intervention group are the (seemingly) weaker students
- Are the study participants representative of the underlying population?
- Are there additional confounders? $\Rightarrow$ randomize groups and blind participants & teachers, if possible

Further aspects

- Regression to the mean if intervention group are the (seemingly) weaker students
- Are the study participants representative of the underlying population?
- Are there additional confounders? $\Rightarrow$ randomize groups and blind participants & teachers, if possible
- Are the tests representative of what you want to test?

Further aspects

- Regression to the mean if intervention group are the (seemingly) weaker students
- Are the study participants representative of the underlying population?
- Are there additional confounders? $\Rightarrow$ randomize groups and blind participants & teachers, if possible
- Are the tests representative of what you want to test?
- ... many more (Kulik and Fletcher 2016)

Example data

- Let's say we have the following data from a study

Example data

- Let's say we have the following data from a study

	control condition				
student	1	2	3	4	5
pre-test score	24	17	29	85	31
post-test score	67	60	-	80	75

Example data

- Let's say we have the following data from a study

control condition

student	1	2	3	4	5
pre-test score	24	17	29	85	31
post-test score	67	60	-	80	75

intervention condition

student	6	7	8	9
pre-test score	20	27	23	16
post-test score	71	75	73	68

Example data

- Let's say we have the following data from a study

control condition

student	1	2	3	4	5
pre-test score	24	17	29	85	31
post-test score	67	60	-	80	75

intervention condition

student	6	7	8	9
pre-test score	20	27	23	16
post-test score	71	75	73	68

attention! This fictional data set is too small for actual statistics! This is only for illustration purposes!

Statistical questions

Assume we have recorded pre- and post-test scores for control and intervention group

Statistical questions

UNIVERSITÄT BIELEFELD
Faculty of Technology

Assume we have recorded pre- and post-test scores for control and intervention group

- Are there students with implausibly high pre-test scores? (outlier detection)

Statistical questions

UNIVERSITÄT BIELEFELD
Faculty of Technology

Assume we have recorded pre- and post-test scores for control and intervention group

- Are there students with implausibly high pre-test scores? (outlier detection)
- If students have missing post-test values, what can we most plausibly fill in? (imputation)

Statistical questions

Assume we have recorded pre- and post-test scores for control and intervention group

- Are there students with implausibly high pre-test scores? (outlier detection)
- If students have missing post-test values, what can we most plausibly fill in? (imputation)
- Do students with high pre-test scores also tend to have high post-test scores? (correlation)

Statistical questions

UNIVERSITÄT BIELEFELD
Faculty of Technology

Assume we have recorded pre- and post-test scores for control and intervention group

- Are there students with implausibly high pre-test scores? (outlier detection)
- If students have missing post-test values, what can we most plausibly fill in? (imputation)
- Do students with high pre-test scores also tend to have high post-test scores? (correlation)
- Do students in the intervention group improve their scores more than in the control group? (statistical testing)

Statistical questions

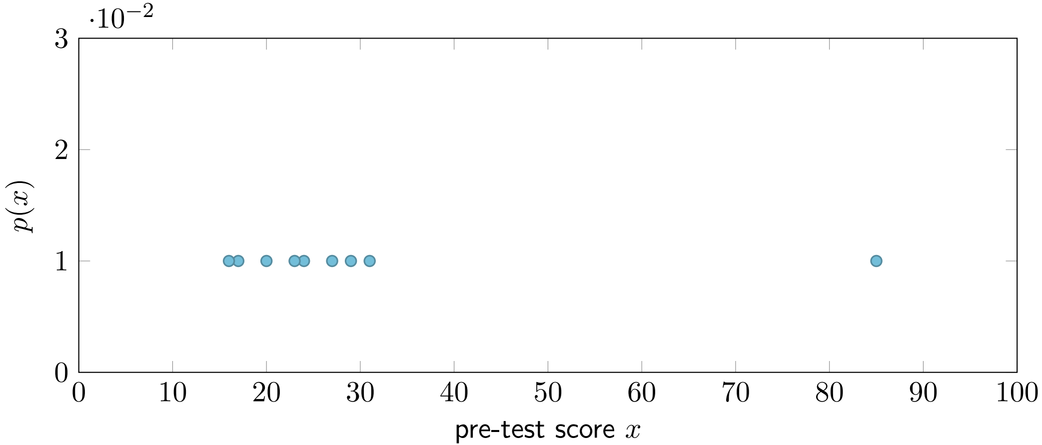
UNIVERSITÄT BIELEFELD
Faculty of Technology

Assume we have recorded pre- and post-test scores for control and intervention group

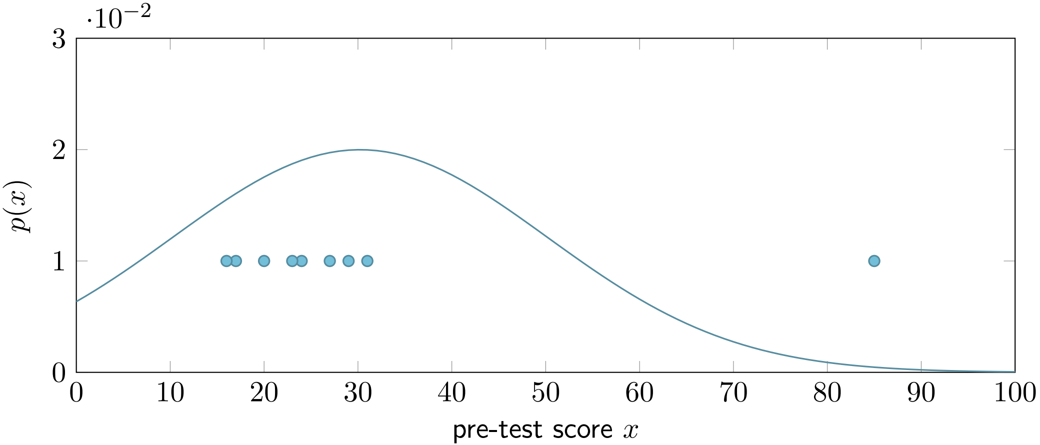
- Are there students with implausibly high pre-test scores? (outlier detection)
- If students have missing post-test values, what can we most plausibly fill in? (imputation)
- Do students with high pre-test scores also tend to have high post-test scores? (correlation)
- Do students in the intervention group improve their scores more than in the control group? (statistical testing)
- Do students improve between pre- and post-test? (paired statistical testing)

Outlier detection

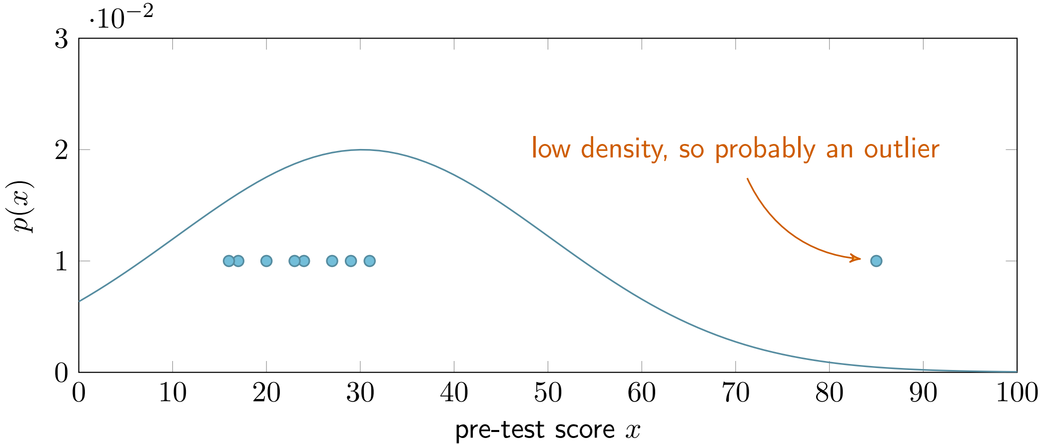
Example: Outlier detection



Example: Outlier detection



Example: Outlier detection



Outlier Detection

Assume example data $x_1, \ldots, x_m \in \mathcal{X}$.

Outlier Detection

Assume example data $x_1, \dots, x_m \in \mathcal{X}$.
Assume a probability density p fitted to that data.

Outlier Detection

Assume example data $x_1, \dots, x_m \in \mathcal{X}$.
Assume a probability density p fitted to that data.
We call points x_i **outliers** for threshold $\epsilon > 0$ if $p(x_i) < \epsilon$.

Gaussian outlier detection

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data

Gaussian outlier detection

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data
- Compute z-score for each data point $x_1, \ldots, x_m$:

$$z = \frac{|x_i - \mu|}{\sigma}$$

Gaussian outlier detection

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data
- Compute z-score for each data point $x_1, \ldots, x_m$:

$$z = \frac{|x_i - \mu|}{\sigma}$$

- Remove data point if $z > k$ for some threshold k

Gaussian outlier detection

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data
- Compute z-score for each data point $x_1, \ldots, x_m$:

$$z = \frac{|x_i - \mu|}{\sigma}$$

- Remove data point if $z > k$ for some threshold k
- $\epsilon = \frac{1}{\sqrt{2\pi \cdot \sigma}} \cdot \exp(-\frac{1}{2}k^2)$

Multi-dimensional Gaussian outlier detection

- Approach 1: perform outlier detection separately for each dimension

Multi-dimensional Gaussian outlier detection

UNIVERSITÄT BIELEFELD
Faculty of Technology

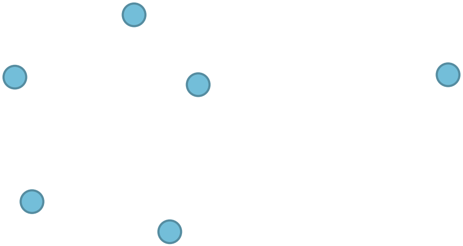
- Approach 1: perform outlier detection separately for each dimension
- Approach 2: Compute covariance matrix $\mathbf{\Sigma}$; z-score becomes Mahalonobis distance:

$$z = \sqrt{(x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu)}$$

Multi-dimensional Gaussian outlier detection

- Approach 1: perform outlier detection separately for each dimension
- Approach 2: Compute covariance matrix $\mathbf{\Sigma}$; z-score becomes Mahalonobis distance:

$$z = \sqrt{(x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu)}$$

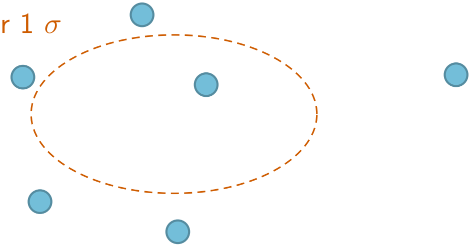


Multi-dimensional Gaussian outlier detection

- Approach 1: perform outlier detection separately for each dimension
- Approach 2: Compute covariance matrix $\mathbf{\Sigma}$; z-score becomes Mahalonobis distance:

$$z = \sqrt{(x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu)}$$

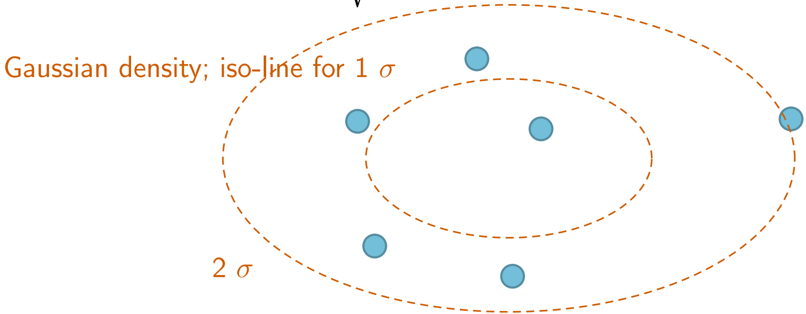
Gaussian density; iso-line for 1 σ



Multi-dimensional Gaussian outlier detection

- Approach 1: perform outlier detection separately for each dimension
- Approach 2: Compute covariance matrix $\mathbf{\Sigma}$; z-score becomes Mahalonobis distance:

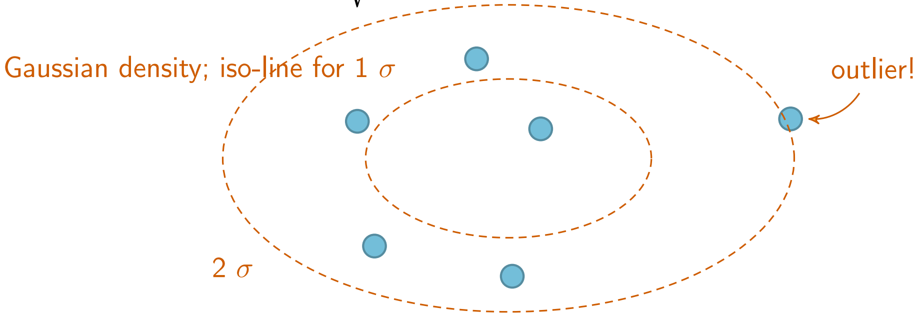
$$z = \sqrt{(x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu)}$$



Multi-dimensional Gaussian outlier detection

- Approach 1: perform outlier detection separately for each dimension
- Approach 2: Compute covariance matrix $\mathbf{\Sigma}$; z-score becomes Mahalonobis distance:

$$z = \sqrt{(x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu)}$$



Challenges in outlier detection

- p needs to be an accurate model of the data – how to ensure that?

Challenges in outlier detection

- p needs to be an accurate model of the data – how to ensure that?

$\Rightarrow$ e.g. generative models, density estimation

Challenges in outlier detection

UNIVERSITÄT BIELEFELD
Faculty of Technology

- p needs to be an accurate model of the data – how to ensure that?

$\Rightarrow$ e.g. generative models, density estimation

- We fit p to the data **including** outliers; chicken-and-egg problem

Challenges in outlier detection

- p needs to be an accurate model of the data – how to ensure that?

$\Rightarrow$ e.g. generative models, density estimation

- We fit p to the data **including** outliers; chicken-and-egg problem

$\Rightarrow$ iterative models (remove an outlier, fit again, etc.)

Challenges in outlier detection

- p needs to be an accurate model of the data – how to ensure that?

$\Rightarrow$ e.g. generative models, density estimation

- We fit p to the data **including** outliers; chicken-and-egg problem

$\Rightarrow$ iterative models (remove an outlier, fit again, etc.)

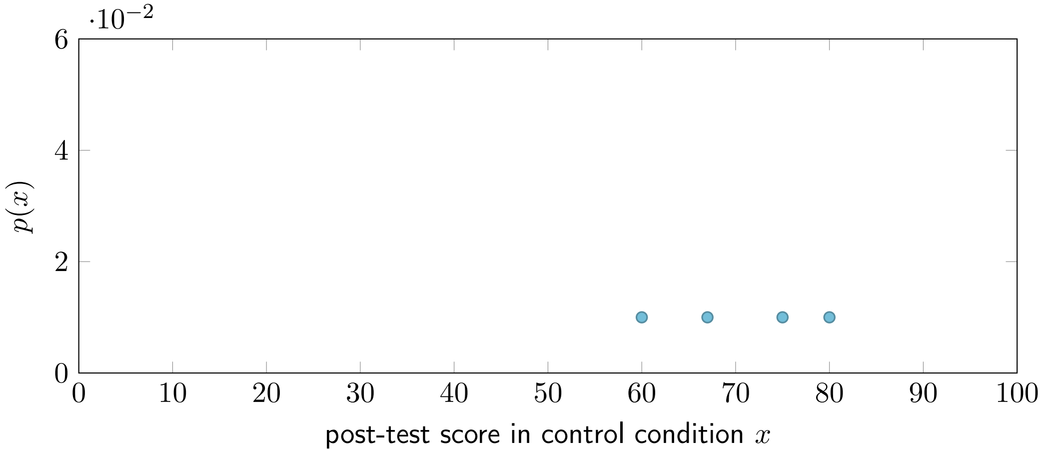
$\Rightarrow$ density estimation & outlier exclusion in one go, e.g. one-class SVM

Imputation

Example: Imputation

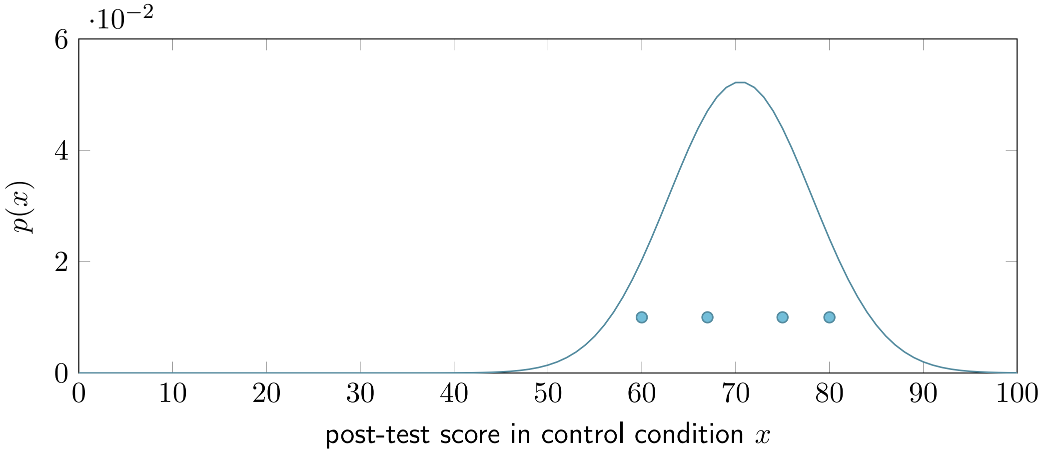
UNIVERSITÄT BIELEFELD

Faculty of Technology

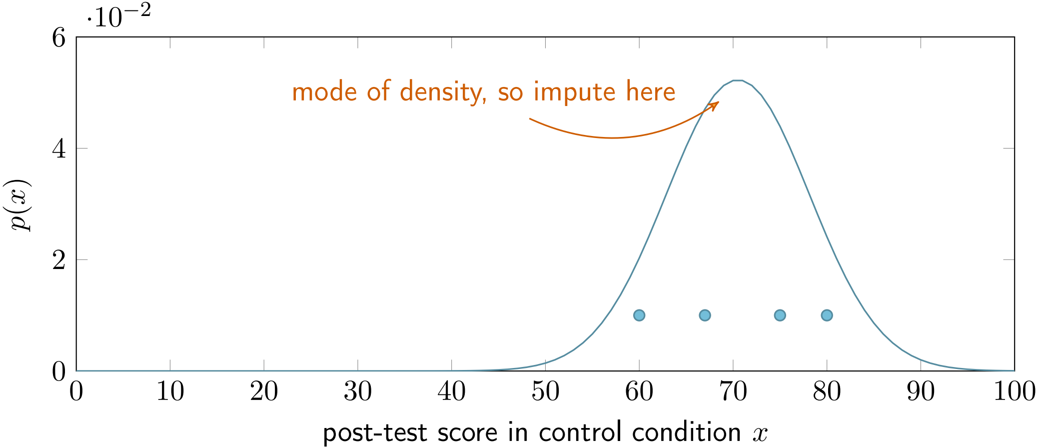


Example: Imputation

UNIVERSITÄT BIELEFELD
Faculty of Technology



Example: Imputation



Universität Bielefeld
Faculty of Technology

Outlier Detection

Assume example data $x_1, \ldots, x_m \in (\mathbb{R} \cup \{\mathrm{NaN}\})^n$.

Outlier Detection

Assume example data $x_1, \ldots, x_m \in (\mathbb{R} \cup \{\mathrm{NaN}\})^n$.
We call a value $x_{i,j} = \mathrm{NaN}$ a **missing value**.

Outlier Detection

Assume example data $x_1, \ldots, x_m \in (\mathbb{R} \cup \{\text{NaN}\})^n$.
We call a value $x_{i,j} = \text{NaN}$ a **missing value**.
Assume a probability density p fitted to the data.

Outlier Detection

Assume example data $x_1, \ldots, x_m \in (\mathbb{R} \cup \{\text{NaN}\})^n$.
We call a value $x_{i,j} = \text{NaN}$ a **missing value**.
Assume a probability density p fitted to the data.
Imputation means to replace a missing $x_{i,j}$ with one that maximizes p.

Gaussian imputation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values

Gaussian imputation

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values
- Replace NaNs with $\mu \Rightarrow$ maximizes $p(x_{i,j})$

Gaussian imputation

UNIVERSITÄT BIELEFELD
Faculty of Technology

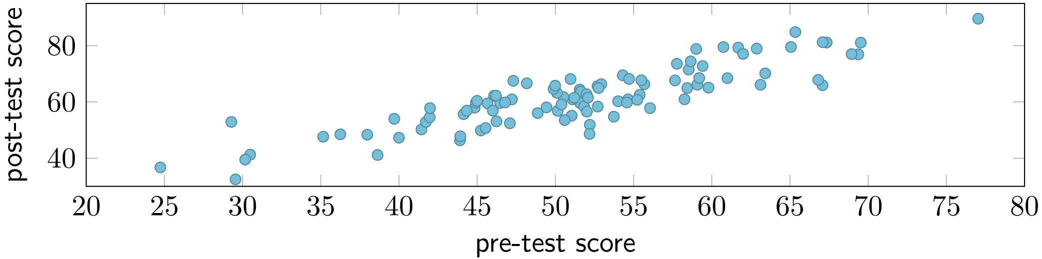
- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values
- Replace NaNs with $\mu \Rightarrow$ maximizes $p(x_{i,j})$
- Approach 1 for multi-dimensional data: treat each dimension separately

Gaussian imputation

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values
- Replace NaNs with $\mu \Rightarrow$ maximizes $p(x_{i,j})$
- Approach 1 for multi-dimensional data: treat each dimension separately
- Approach 2: maximize conditional $p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,n})$

Gaussian imputation

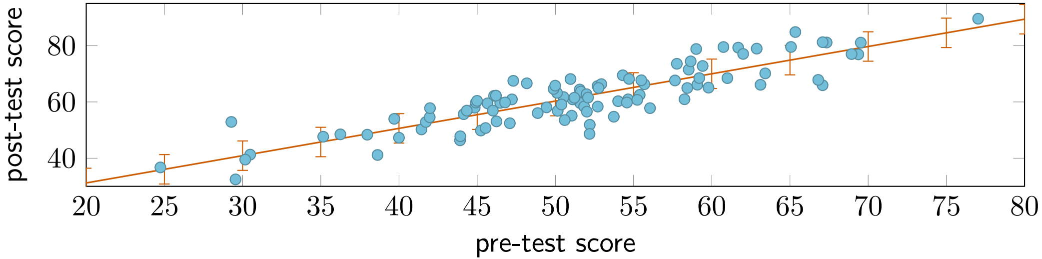
- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values
- Replace NaNs with $\mu \Rightarrow$ maximizes $p(x_{i,j})$
- Approach 1 for multi-dimensional data: treat each dimension separately
- Approach 2: maximize conditional $p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,n})$



Gaussian imputation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Fit Gaussian, i.e. compute empiric mean μ and standard deviation σ of data, excluding missing values
- Replace NaNs with $\mu \Rightarrow$ maximizes $p(x_{i,j})$
- Approach 1 for multi-dimensional data: treat each dimension separately
- Approach 2: maximize conditional $p(x_{i,j}|x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,n})$



Challenges in imputation

- Constant value imputation artificially decreases variance of data

Challenges in imputation

- Constant value imputation artificially decreases variance of data

$\Rightarrow$ Sample from p instead of always taking argmax

Challenges in imputation

- Constant value imputation artificially decreases variance of data

$\Rightarrow$ Sample from p instead of always taking argmax

- better models of conditionals?

Challenges in imputation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Constant value imputation artificially decreases variance of data

$\Rightarrow$ Sample from p instead of always taking argmax

- better models of conditionals?

$\Rightarrow$ Regression approaches

Correlations

Correlation: Example

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Intuition: How strongly are two random variables associated?

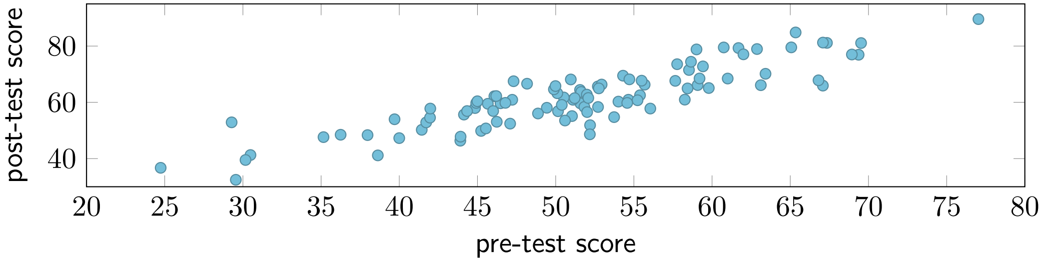
Correlation: Example

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Intuition: How strongly are two random variables associated?
- Attempt 1: Slope of linear regression line

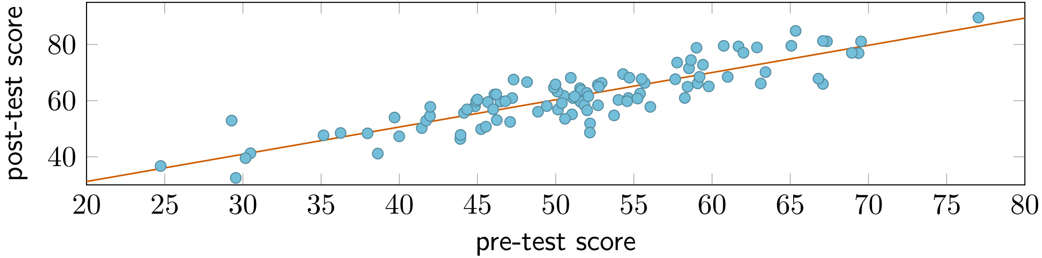
Correlation: Example

- Intuition: How strongly are two random variables associated?
- Attempt 1: Slope of linear regression line



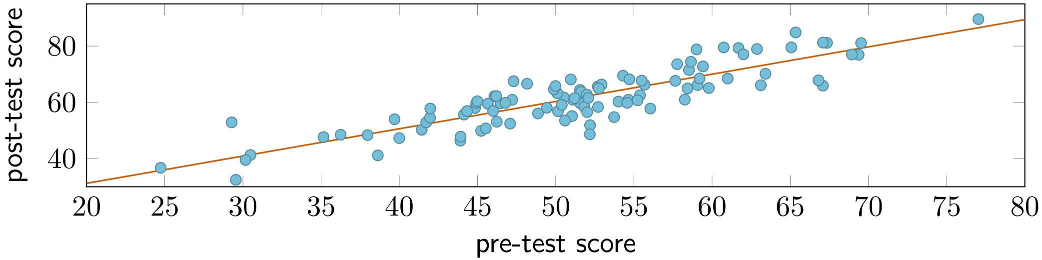
Correlation: Example

- Intuition: How strongly are two random variables associated?
- Attempt 1: Slope of linear regression line



Correlation: Example

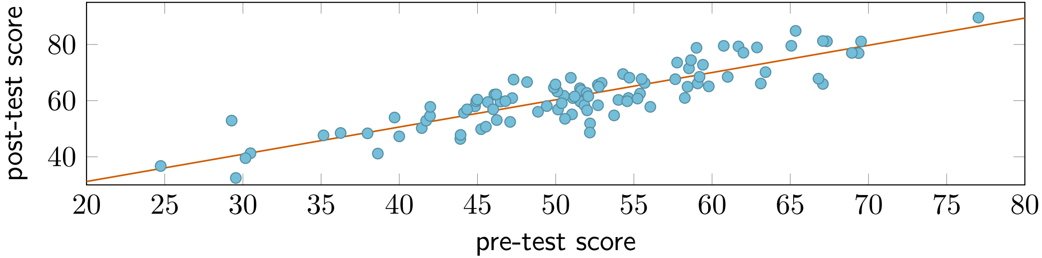
- Intuition: How strongly are two random variables associated?
- Attempt 1: Slope of linear regression line



- Problem: What about the **scaling** of the data?

Correlation: Example

- Intuition: How strongly are two random variables associated?
- Attempt 1: Slope of linear regression line



- Problem: What about the **scaling** of the data?

$\Rightarrow$ normalize data beforehand: $\tilde{x} = (x - \mu_x)/\sigma_x$ and $\tilde{y} = (y - \mu_y)/\sigma_y$.

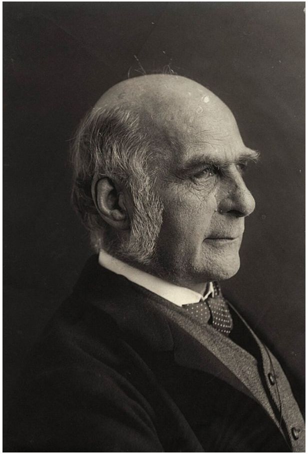
Linear/Pearson correlation

UNIVERSITÄT BIELEFELD
Faculty of Technology

Sir Francis Galton FRS FRAI (/ˈɡɔːltən/; 16 February 1822 – 17 January 1911) was a British polymath and the originator of the eugenics movement during the Victorian era.[1][2]

Galton produced over 340 papers and books. He also developed the statistical concept of correlation and widely promoted regression toward the mean.

In recent years, he has received significant criticism for being a proponent of social Darwinism, eugenics, and scientific racism; he was a pioneer of eugenics, coining the term itself in 1883.



Linear/Pearson correlation

Karl Pearson FRS FRSE[1] (/ˈpɪərsən/; born **Carl Pearson**; 27 March 1857 – 27 April 1936[2]) was an English mathematician and biostatistician. He has been credited with establishing the discipline of mathematical statistics.[3][4] He founded the world's first university statistics department at University College London in 1911, and contributed significantly to the field of biometrics and meteorology. Pearson was also a proponent of social Darwinism and eugenics, and his thought is an example of what is today described as scientific racism.

Linear/Pearson correlation (Galton 1889)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Linear correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.

Linear/Pearson correlation (Galton 1889)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Linear correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
the linear correlation coefficient is defined as:

$$r = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y},$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the data.

Linear/Pearson correlation (Galton 1889)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Linear correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
the linear correlation coefficient is defined as:

$$r = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y},$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the data.

- range: $[-1, +1]$; symmetric for x and y

Linear/Pearson correlation (Galton 1889)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Linear correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
the linear correlation coefficient is defined as:

$$r = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y},$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the data.

- range: $[-1, +1]$; symmetric for x and y
- interpretation: slope of the linear regression line **after** normalization of x and y

Linear/Pearson correlation (Galton 1889)

UNIVERSITÄT
BIELEFELD
Faculty of Technology

Linear correlation coefficient

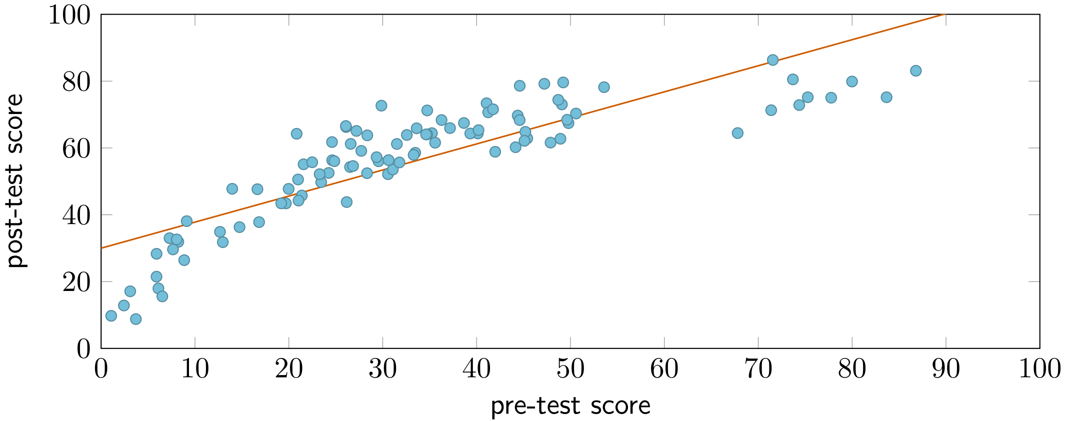
Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
the linear correlation coefficient is defined as:

$$r = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y},$$

where μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of the data.

- range: $[-1, +1]$; symmetric for x and y
- interpretation: slope of the linear regression line **after** normalization of x and y
- Rules of thumb: $r \leq 0.3$ is very small, $r \in (0.3, 0.5]$ is small, $r \in (0.5, 0.7]$ is moderate, $r \in (0.7, 0.9]$ is high, $r \in (0.9, 1.0]$ is very high (Mukaka 2012)

Problems: Nonlinear correlations



Rank/Spearman correlation

UNIVERSITÄT BIELEFELD
Faculty of Technology

Charles Edward Spearman, FRS[1][3] (10 September 1863 – 17 September 1945) was an English psychologist known for work in statistics, as a pioneer of factor analysis, and for Spearman's rank correlation coefficient.

The Eugenics Review

Eugen Rev. 1914 Oct; 6(3): 219–237.

PMCID: PMC2987066
PMID: 21259592

The heredity of abilities

C. Spearman

Rank/Spearman correlation (Spearman 1904)

Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.

Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
Replace x_i and y_i by their **rank**.

Rank/Spearman correlation (Spearman 1904)

Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
Replace x_i and y_i by their **rank**.
The rank correlation coefficient is then defined as r on the ranks.

Rank/Spearman correlation (Spearman 1904)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
Replace x_i and y_i by their **rank**.
The rank correlation coefficient is then defined as r on the ranks.

- similar interpretation to linear correlation coefficient r – but in rank space

Rank/Spearman correlation (Spearman 1904)

Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
Replace x_i and y_i by their **rank**.
The rank correlation coefficient is then defined as r on the ranks.

- similar interpretation to linear correlation coefficient r – but in rank space
- no direct relation to regression line on the original data; no "drawing" of fit $\Rightarrow$ "non-parametric statistic"

Rank/Spearman correlation (Spearman 1904)

UNIVERSITÄT BIELEFELD
Faculty of Technology

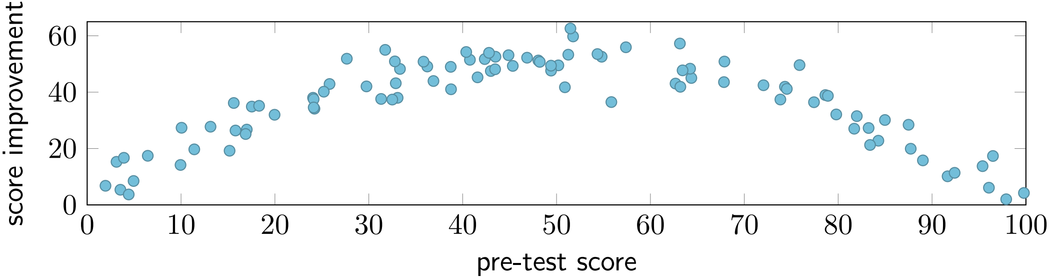
Rank correlation coefficient

Let $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$.
Replace x_i and y_i by their **rank**.
The rank correlation coefficient is then defined as r on the ranks.

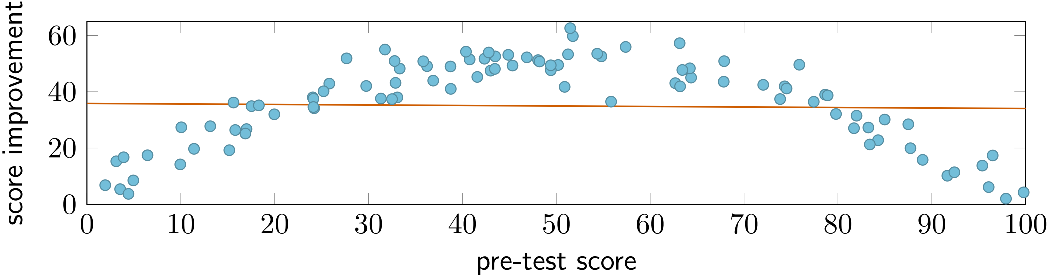
- similar interpretation to linear correlation coefficient r – but in rank space
- no direct relation to regression line on the original data; no “drawing” of fit $\Rightarrow$ “non-parametric statistic”
- alternatives: Cohen’s κ, Krippendorff’s α, ...

Non-monotonic relations

UNIVERSITÄT BIELEFELD
Faculty of Technology



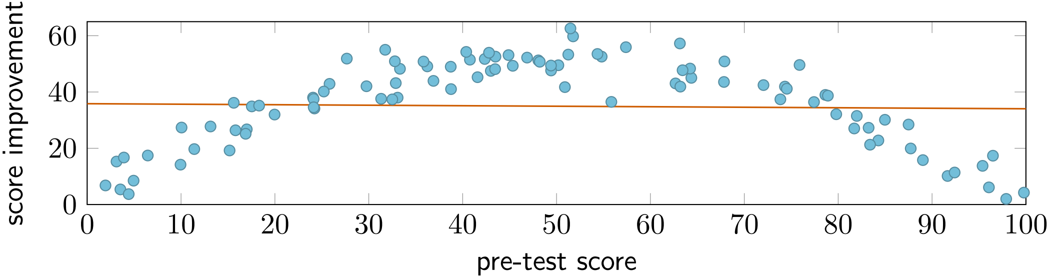
Non-monotonic relations



Non-monotonic relations

UNIVERSITÄT BIELEFELD

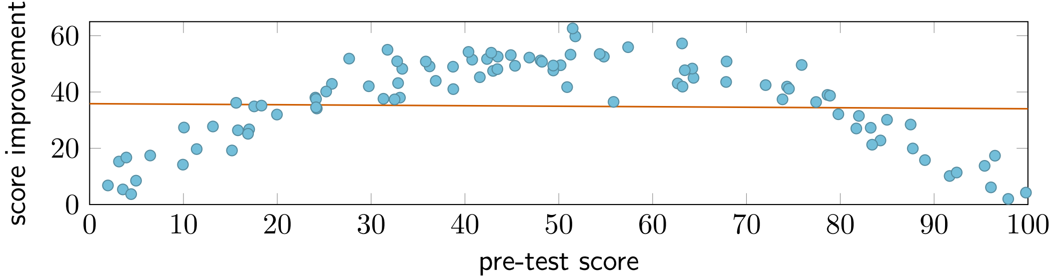
Faculty of Technology



- correlation measures capture only **monotonic** relationships

Non-monotonic relations

UNIVERSITÄT BIELEFELD
Faculty of Technology



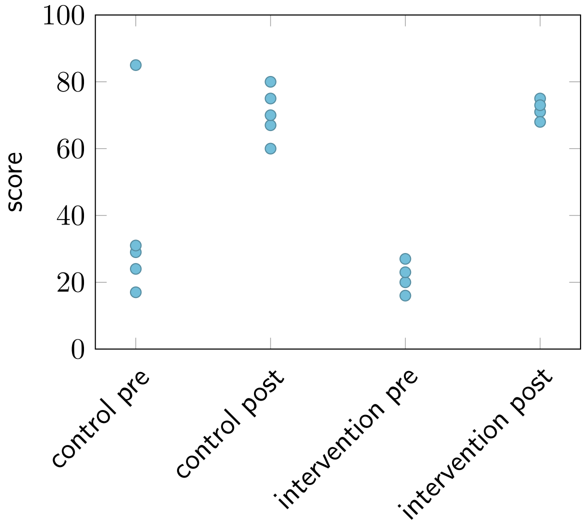
- correlation measures capture only **monotonic** relationships

$\Rightarrow$ independence $\Rightarrow$ no correlation, but not vice versa

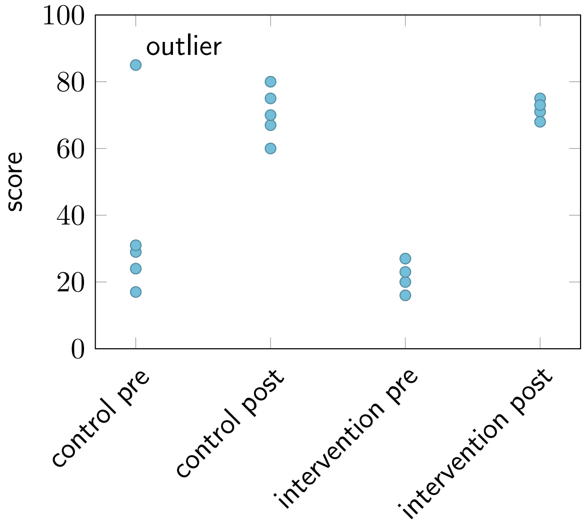
Statistical Tests

Example: Bar plot

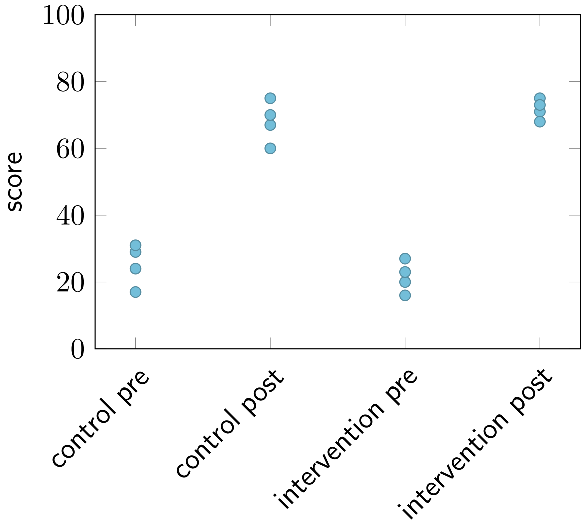
UNIVERSITÄT BIELEFELD
Faculty of Technology



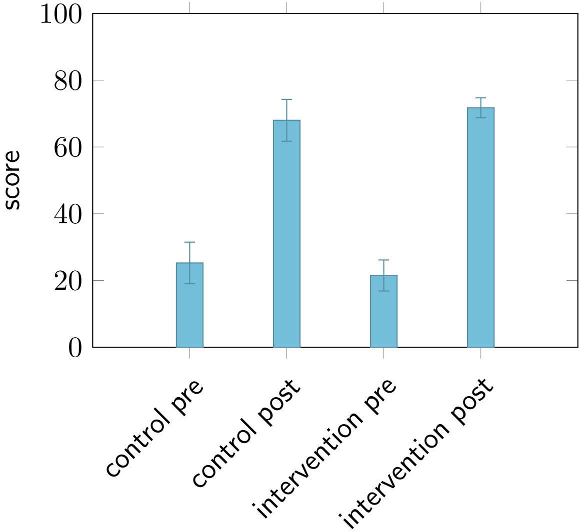
Example: Bar plot



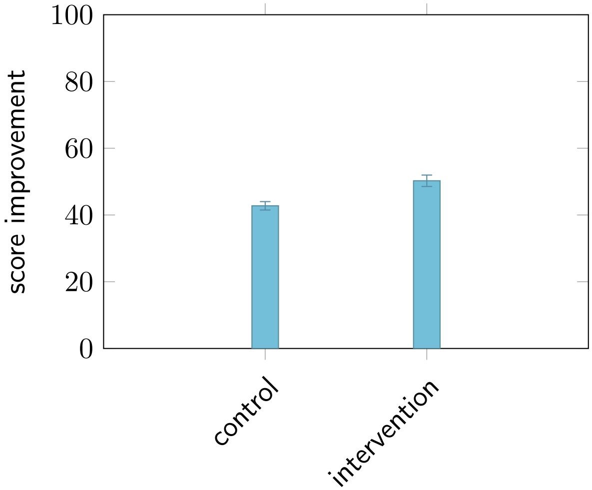
Example: Bar plot



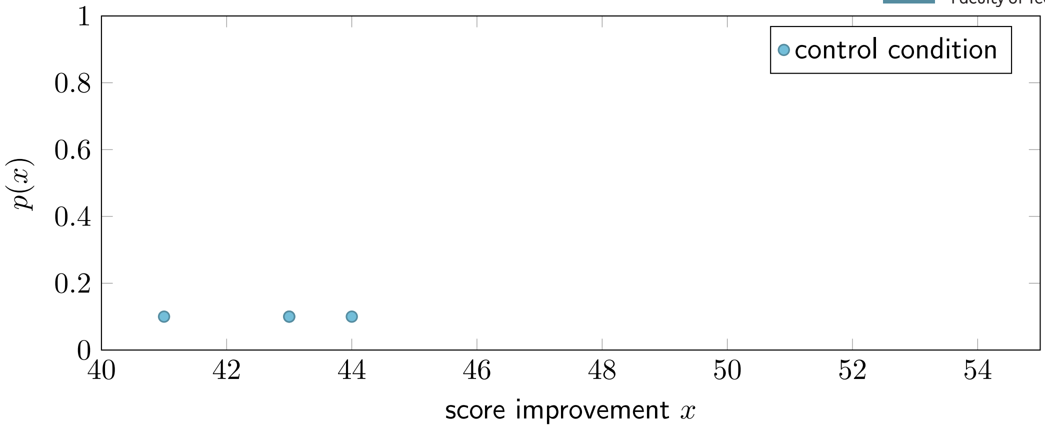
Example: Bar plot



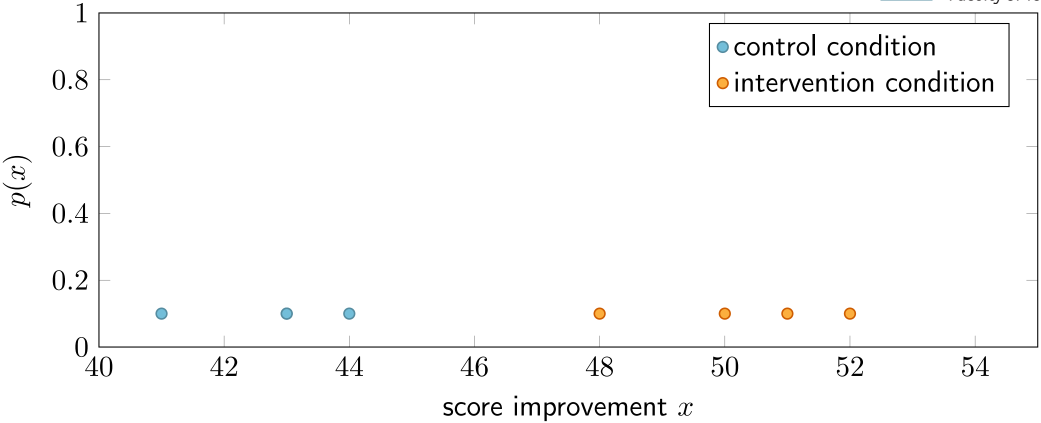
Example: Bar plot



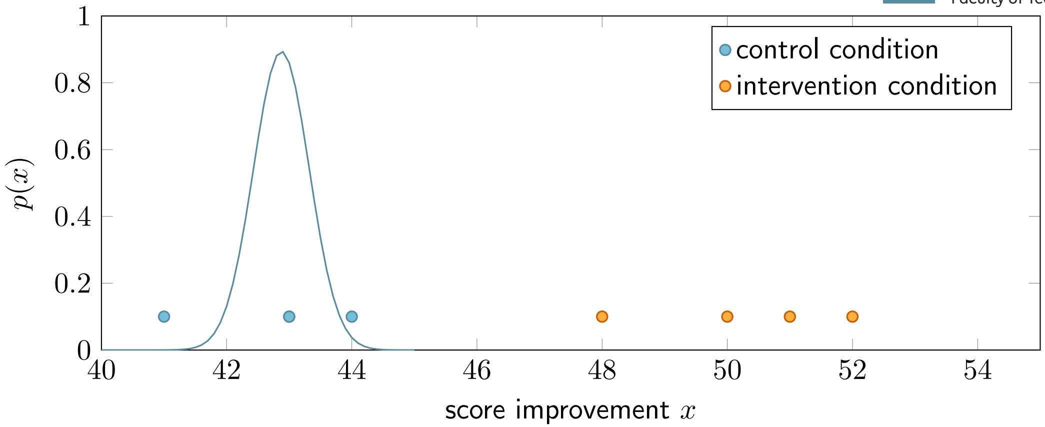
Example (continued)



Example (continued)



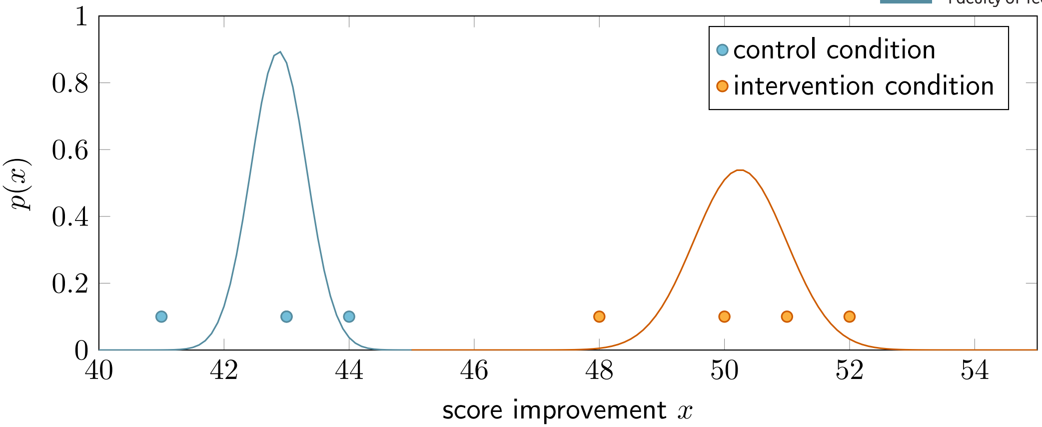
Example (continued)



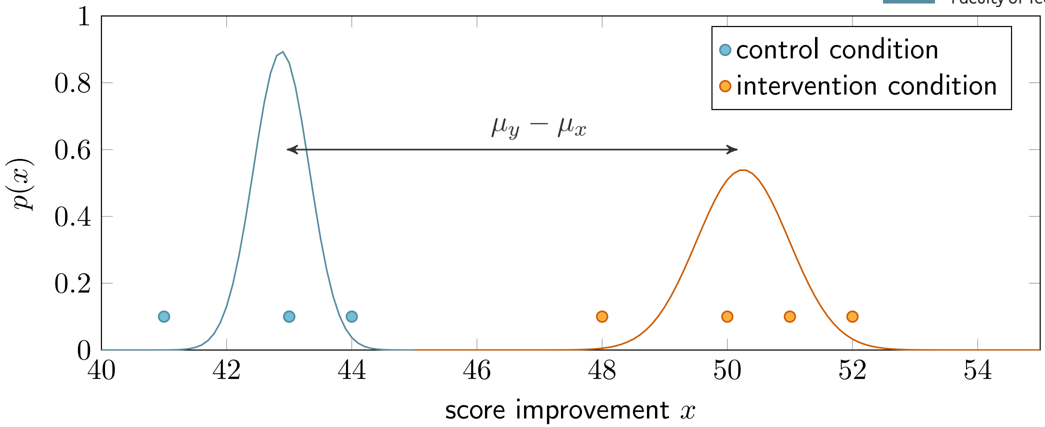
Example (continued)

UNIVERSITÄT BIELEFELD

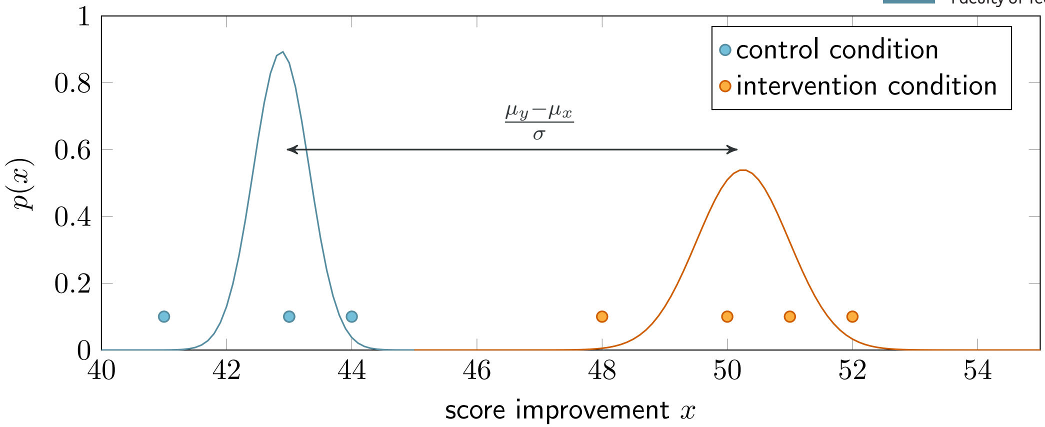
Faculty of Technology



Example (continued)

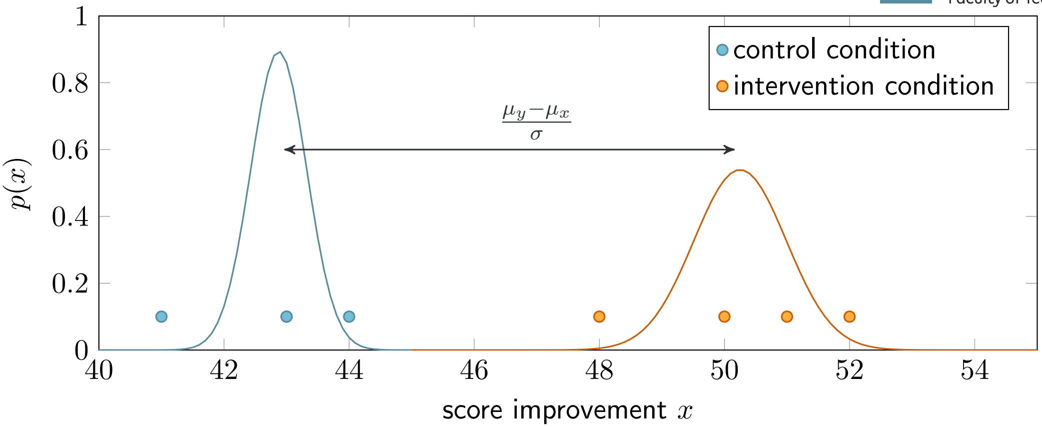


Example (continued)



Example (continued)

UNIVERSITÄT BIELEFELD
Faculty of Technology



- What is the probability of $\frac{\mu_y - \mu_x}{\sigma}$ being that large just by random chance?

Null hypothesis testing: Motivation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We want to be certain that an observed effect could not also have been generated by mere chance

Null hypothesis testing: Motivation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We want to be certain that an observed effect could not also have been generated by mere chance – more precisely: by a random baseline model (the **null hypothesis**)

Null hypothesis testing: Motivation

UNIVERSITÄT BIELEFELD
Faculty of Technology

- We want to be certain that an observed effect could not also have been generated by mere chance – more precisely: by a random baseline model (the **null hypothesis**)
- If we are certain that the null hypothesis could not have generated the observed effect, we **reject the null hypothesis**

Null hypothesis testing: Motivation

- We want to be certain that an observed effect could not also have been generated by mere chance – more precisely: by a random baseline model (the **null hypothesis**)
- If we are certain that the null hypothesis could not have generated the observed effect, we **reject the null hypothesis**
- Type I error: Rejecting the null hypothesis even though it did generate the observed effect (p value)

Null hypothesis testing: Motivation

- We want to be certain that an observed effect could not also have been generated by mere chance – more precisely: by a random baseline model (the **null hypothesis**)
- If we are certain that the null hypothesis could not have generated the observed effect, we **reject the null hypothesis**
- Type I error: Rejecting the null hypothesis even though it did generate the observed effect (p value)
- Type II error: Failing to reject the null hypothesis even though it did not generate the observed effect (statistical power)

Null hypothesis testing framework

UNIVERSITÄT BIELEFELD
Faculty of Technology

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.

Null hypothesis testing framework

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.
The null hypothesis is that $\hat{s}$ is generated just due to random chance.

Null hypothesis testing framework

UNIVERSITÄT BIELEFELD

Faculty of Technology

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.
The null hypothesis is that $\hat{s}$ is generated just due to random chance.
We model the probability $p = P(s \geq \hat{s})$ under the null hypothesis.

Null hypothesis testing framework

UNIVERSITÄT BIELEFELD
Faculty of Technology

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.
The null hypothesis is that $\hat{s}$ is generated just due to random chance.
We model the probability $p = P(s \geq \hat{s})$ under the null hypothesis.
We reject the null hypothesis if p is sufficiently small (e.g. $p < 0.05$). We say the effect is significant (at 0.05-level).

Null hypothesis testing framework

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.
The null hypothesis is that $\hat{s}$ is generated just due to random chance.
We model the probability $p = P(s \geq \hat{s})$ under the null hypothesis.
We reject the null hypothesis if p is sufficiently small (e.g. $p < 0.05$). We say the effect is **significant** (at 0.05-level).

- Key advantage: We only need a model of the statistic under the null hypothesis

Null hypothesis testing framework

UNIVERSITÄT BIELEFELD
Faculty of Technology

Null hypothesis testing

To quantify an effect, we compute a statistic $\hat{s}$.
The null hypothesis is that $\hat{s}$ is generated just due to random chance.
We model the probability $p = P(s \geq \hat{s})$ under the null hypothesis.
We reject the null hypothesis if p is sufficiently small (e.g. $p < 0.05$). We say the effect is **significant** (at 0.05-level).

- Key advantage: We only need a model of the statistic under the null hypothesis
- Tests differ in statistic, assumptions, and model

t-Test (Welch version)

UNIVERSITÄT BIELEFELD
Faculty of Technology

William Sealy Gosset (13 June 1876 – 16 October 1937) was an English statistician, chemist and brewer who served as Head Brewer of Guinness and Head Experimental Brewer of Guinness and was a pioneer of modern statistics. He pioneered small sample experimental design and analysis with an economic approach to the logic of uncertainty. Gosset published under the pen name **Student** and developed most famously Student's t-distribution – originally called Student's "z" – and "Student's test of statistical significance".[1]

t-Test (Welch version)

(Student 1908; Welch 1947)

UNIVERSITÄT BIELEFELD
Faculty of Technology

t-Test (Welch version)

(Student 1908; Welch 1947)

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Assume two samples with n_x and n_y data points, means μ_x and μ_y, and standard deviations σ_x and σ_y

t-Test (Welch version)

(Student 1908; Welch 1947)

- Assume two samples with n_x and n_y data points, means μ_x and μ_y, and standard deviations σ_x and σ_y
- Test statistic: mean difference scaled by standard error: $t = \frac{\mu_x - \mu_y}{\tilde{\sigma}}$

t-Test (Welch version)

(Student 1908; Welch 1947)

- Assume two samples with n_x and n_y data points, means μ_x and μ_y, and standard deviations σ_x and σ_y
- Test statistic: mean difference scaled by standard error: $t = \frac{\mu_x - \mu_y}{\tilde{\sigma}}$
- Problem: Which $\tilde{\sigma}$ do we take?

$$\tilde{\sigma} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

t-Test (Welch version)

(Student 1908; Welch 1947)

- Assume two samples with n_x and n_y data points, means μ_x and μ_y, and standard deviations σ_x and σ_y
- Test statistic: mean difference scaled by standard error: $t = \frac{\mu_x - \mu_y}{\tilde{\sigma}}$
- Problem: Which $\tilde{\sigma}$ do we take? $\Rightarrow$ compromise between standard error of sample one and sample two:

$$\tilde{\sigma} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

t-Test (Welch version)

(Student 1908; Welch 1947)

- Assume two samples with n_x and n_y data points, means μ_x and μ_y, and standard deviations σ_x and σ_y
- Test statistic: mean difference scaled by standard error: $t = \frac{\mu_x - \mu_y}{\tilde{\sigma}}$
- Problem: Which $\tilde{\sigma}$ do we take? $\Rightarrow$ compromise between standard error of sample one and sample two:

$$\tilde{\sigma} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

- Under the null hypothesis, μ_x and μ_y stem from the same Gaussian with $\tilde{\sigma}$ and t is t-distributed with parameter

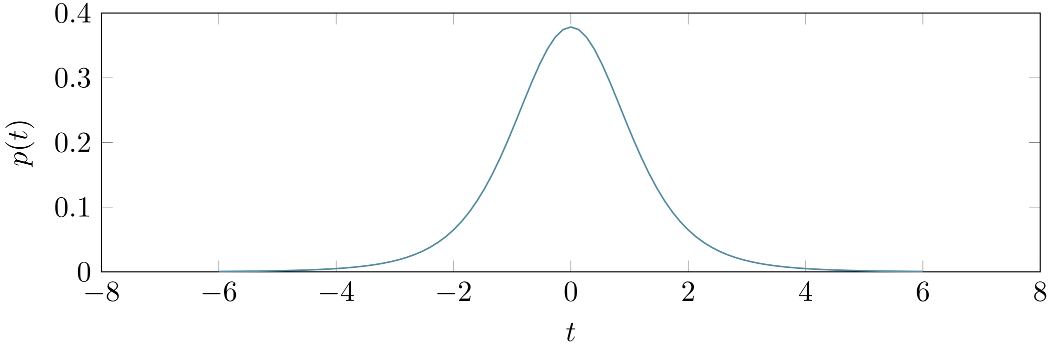
$$\nu = \frac{\tilde{\sigma}^4}{\frac{\sigma_x^4}{n_x^2 \cdot (n_x - 1)} + \frac{\sigma_y^4}{n_y^2 \cdot (n_y - 1)}}$$

t-distribution illustration

- For our toy data set: $\tilde{\sigma} \approx 1$, $t = 7.39$, $\nu = 4.93$, $p = 0.0007$

t-distribution illustration

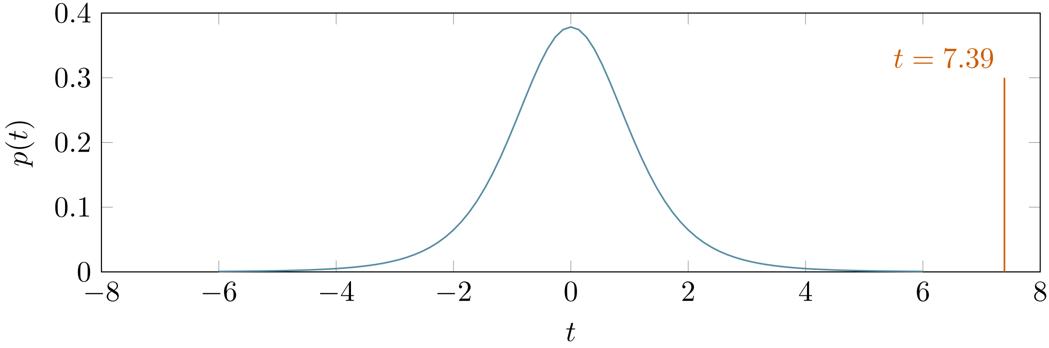
- For our toy data set: $\tilde{\sigma} \approx 1$, $t = 7.39$, $\nu = 4.93$, $p = 0.0007$



t-distribution illustration

UNIVERSITÄT BIELEFELD
Faculty of Technology

- For our toy data set: $\tilde{\sigma} \approx 1$, $t = 7.39$, $\nu = 4.93$, $p = 0.0007$



Assumptions of the t-test

- Samples are independent

Assumptions of the t-test

- Samples are independent
- Sample means are Gaussian distributed (Central limit theorem)

Assumptions of the t-test

- Samples are independent
- Sample means are Gaussian distributed (Central limit theorem)

$\Rightarrow$ one way of verification: check normality of original data distribution via Shapiro-Wilk-test (or other tests) $\Rightarrow$ mean of Gaussian variables is Gaussian

Effect size

(Cohen 1988)

- Problem: significance testing only tells us whether an effect could be generated by the null hypothesis – not how strong it is

Effect size

(Cohen 1988)

- Problem: significance testing only tells us whether an effect could be generated by the null hypothesis – not how strong it is
- Effect size: Cohen's d, defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma}$

Effect size

(Cohen 1988)

- Problem: significance testing only tells us whether an effect could be generated by the null hypothesis – not how strong it is
- Effect size: Cohen's d, defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma}$
- Again: σ needs to be a compromise between σ_1 and σ_2, e.g.

$$\sigma = \sqrt{\frac{(n_1 - 1) \cdot \sigma_1^2 + (n_2 - 1) \cdot \sigma_2^2}{n_1 + n_2 - 2}}$$

Effect size

(Cohen 1988)

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Problem: significance testing only tells us whether an effect could be generated by the null hypothesis – not how strong it is
- Effect size: Cohen's d, defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma}$
- Again: σ needs to be a compromise between σ_1 and σ_2, e.g.

$$\sigma = \sqrt{\frac{(n_1 - 1) \cdot \sigma_1^2 + (n_2 - 1) \cdot \sigma_2^2}{n_1 + n_2 - 2}}$$

- Rules of thumb: $d < 0.2$ is very small, $d \in [0.2, 0.5)$ is small, $d \in [0.5, 0.8)$ is moderate, $d \in [0.8, 1.2)$ is large, $d \in [1.2, 2.0)$ is very large, $d \geq 2.0$ is huge (Sawilowsky 2009)

Effect size

(Cohen 1988)

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Problem: significance testing only tells us whether an effect could be generated by the null hypothesis – not how strong it is
- Effect size: Cohen's d, defined as $d = \frac{|\mu_1 - \mu_2|}{\sigma}$
- Again: σ needs to be a compromise between σ_1 and σ_2, e.g.

$$\sigma = \sqrt{\frac{(n_1 - 1) \cdot \sigma_1^2 + (n_2 - 1) \cdot \sigma_2^2}{n_1 + n_2 - 2}}$$

- Rules of thumb: $d < 0.2$ is very small, $d \in [0.2, 0.5)$ is small, $d \in [0.5, 0.8)$ is moderate, $d \in [0.8, 1.2)$ is large, $d \in [1.2, 2.0)$ is very large, $d \geq 2.0$ is huge (Sawilowsky 2009)
- Side note: largest measured effect sizes in psychology for educational studies

Example: signed rank test

- Question: Are post-test scores significantly higher than pre-test scores?

Example: signed rank test

- Question: Are post-test scores significantly higher than pre-test scores?

$\Rightarrow$ **paired** data

Example: signed rank test

- Question: Are post-test scores significantly higher than pre-test scores?

⇒ **paired** data

student	1	2	3	4	5	6	7	8	9
pre-test score	24	17	29	85	31	20	27	23	16
post-test score	67	60	70	80	75	71	75	73	68

Example: signed rank test

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Question: Are post-test scores significantly higher than pre-test scores?

$\Rightarrow$ **paired** data

student	1	2	3	4	5	6	7	8	9
pre-test score	24	17	29	85	31	20	27	23	16
post-test score	67	60	70	80	75	71	75	73	68
improvement	43	43	41	-5	44	51	48	50	52

Example: signed rank test

- Question: Are post-test scores significantly higher than pre-test scores?

$\Rightarrow$ **paired** data

student	1	2	3	4	5	6	7	8	9
pre-test score	24	17	29	85	31	20	27	23	16
post-test score	67	60	70	80	75	71	75	73	68
improvement	43	43	41	-5	44	51	48	50	52

- Idea 1: If there are more positive than negative numbers, probably yes

Example: signed rank test

- Question: Are post-test scores significantly higher than pre-test scores?

⇒ **paired** data

student	1	2	3	4	5	6	7	8	9
pre-test score	24	17	29	85	31	20	27	23	16
post-test score	67	60	70	80	75	71	75	73	68
improvement	43	43	41	-5	44	51	48	50	52

- Idea 1: If there are more positive than negative numbers, probably yes
- Idea 2: Put higher weights on bigger numbers

Wilcoxon signed rank test (Wilcoxon 1945)

- Sort improvements acc. to absolute value

Wilcoxon signed rank test (Wilcoxon 1945)

- Sort improvements acc. to absolute value
- test statistic T: sum of ranks, with signs

Wilcoxon signed rank test (Wilcoxon 1945)

- Sort improvements acc. to absolute value
- test statistic T: sum of ranks, with signs

improvement	43	43	41	-5	44	51	48	50	52

Wilcoxon signed rank test (Wilcoxon 1945)

- Sort improvements acc. to absolute value
- test statistic T: sum of ranks, with signs

improvement	43	43	41	-5	44	51	48	50	52
signed rank	3.5	3.5	2	-1	5	8	6	7	9

Wilcoxon signed rank test (Wilcoxon 1945)

- Sort improvements acc. to absolute value
- test statistic T: sum of ranks, with signs

improvement	43	43	41	-5	44	51	48	50	52
signed rank	3.5	3.5	2	-1	5	8	6	7	9

$$\Rightarrow T = 43$$

Wilcoxon signed rank test (Wilcoxon 1945)

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Sort improvements acc. to absolute value
- test statistic T: sum of ranks, with signs

improvement	43	43	41	-5	44	51	48	50	52
signed rank	3.5	3.5	2	-1	5	8	6	7	9

$$\Rightarrow T = 43$$

- Under null hypothesis, p_T can be computed exactly (for small n) – or via Gaussian approximation (for large n)

Exact Wilcoxon distribution $n = 3$

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Under the null hypothesis, any combination of signs amongst the ranks is equally likely

rank	sign combination							
1	-	+	-	-	+	+	-	+
2	-	-	+	-	+	-	+	+
3	-	-	-	+	-	+	+	+
T	-6	-4	-2	0	0	+2	+4	+6

Exact Wilcoxon distribution $n = 3$

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Under the null hypothesis, any combination of signs amongst the ranks is equally likely

rank	sign combination							
1	-	+	-	-	+	+	-	+
2	-	-	+	-	+	-	+	+
3	-	-	-	+	-	+	+	+
T	-6	-4	-2	0	0	+2	+4	+6

- $p(-6) = p(-4) = p(-2) = p(+2) = p(+4) = p(+6) = \frac{1}{8}$, $p(0) = \frac{1}{4}$

Assumptions of Wilcoxon signed rank test

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Samples are independent

Assumptions of Wilcoxon signed rank test

- Samples are independent
- Differences are symmetrically distributed

Summary

- This visualization summarizes all key concepts from null hypothesis testing neatly:

  `https://rpsychologist.com/d3/nhst/`

Literature I

UNIVERSITÄT BIELEFELD
Faculty of Technology

Cohen, Jacob (1988). **Statistical Power Analysis for the Behavioral Sciences**. New York, NY, USA: Routledge. DOI: 10.4324/9780203771587.

Galton, Francis (1889). "I. Co-relations and their measurement, chiefly from anthropometric data". In: **Proceedings of the Royal Society of London** 45.273-279, pp. 135–145. DOI: 10.1098/rspl.1888.0082.

Kulik, James A and JD Fletcher (2016). "Effectiveness of intelligent tutoring systems: a meta-analytic review". In: **Review of educational research** 86.1, pp. 42–78. DOI: 10.3102/0034654315581420. URL: https://www.researchgate.net/profile/J-D-Fletcher/publication/277636218_Effectiveness_of_Intelligent_Tutoring_Systems_A_Meta-Analytic_Review/links/5707a6fe08ae8883a1f7e55f/Effectiveness-of-Intelligent-Tutoring-Systems-A-Meta-Analytic-Review.pdf.

Literature II

UNIVERSITÄT BIELEFELD
Faculty of Technology

Mukaka, Mavuto M (2012). "A guide to appropriate use of correlation coefficient in medical research". In: **Malawi medical journal** 24.3, pp. 69–71. URL: https://www.ajol.info/index.php/mmj/article/view/81576.

Sawilowsky, Shlomo S. (2009). "New effect size rules of thumb". In: **Journal of modern applied statistical methods** 8.2, p. 26. DOI: 10.22237/jmasm/1257035100.

Spearman, Charles (1904). "The Proof and Measurement of Association between Two Things". In: **The American Journal of Psychology** 15.1, pp. 72–101. DOI: 10.2307/1412159.

Student (1908). "The Probable Error of a Mean". In: **Biometrika** 6.1, pp. 1–25. DOI: 10.2307/2331554.

Welch, Bernard Lewis (1947). "The Genralization of Student's Problem when several different population variances are involved". In: **Biometrika** 34.1-2, pp. 28–35. DOI: 10.1093/biomet/34.1-2.28.

Literature III

UNIVERSITÄT BIELEFELD
Faculty of Technology

Wilcoxon, Frank (1945). “Individual Comparisons by Ranking Methods”. In: **Biometrics Bulletin** 1.6, pp. 80–83. DOI: 10.2307/3001968. URL: http://www.jstor.org/stable/3001968.