

# A Practical Introduction to Data Science

## Part 2

# Business and Data Understanding



Gergely Zsombor Haász



[haasz.zsombi@gmail.com](mailto:haasz.zsombi@gmail.com)

# Course Agenda

I. Introduction to Data Science

II. Business and Data Understanding

III. Introduction to Supervised Learning

IV. Advanced Supervised Learning

V. Unsupervised Learning

VI. Time Series Analysis

VII. Deep Learning

VIII. Machine Learning Operations

# Business and Data Understanding

The Importance of B&D Understanding

Data Collection

Exploratory Data Analysis

Statistical Inference

Confidence Intervals

Hypothesis Testing


Common Mistakes

Case Studies

# The Importance of B&D Understanding

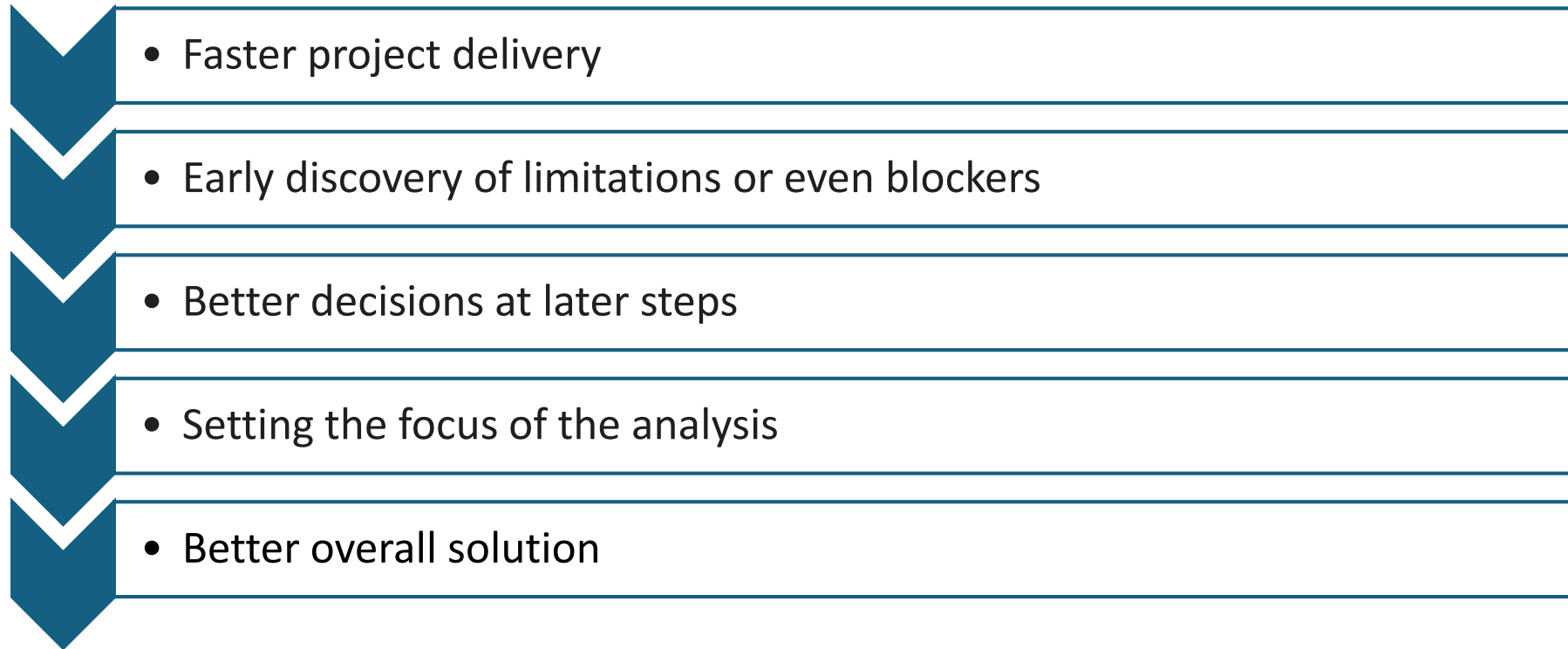
# The Importance of B&D Understanding

A lack of business/data understanding can lead to:

- 
- Wasted time and resources
  - Poor adoption of the model
  - Useless model (bad generalization)
  - Unsatisfied stakeholders (e.g. lack of explainability)

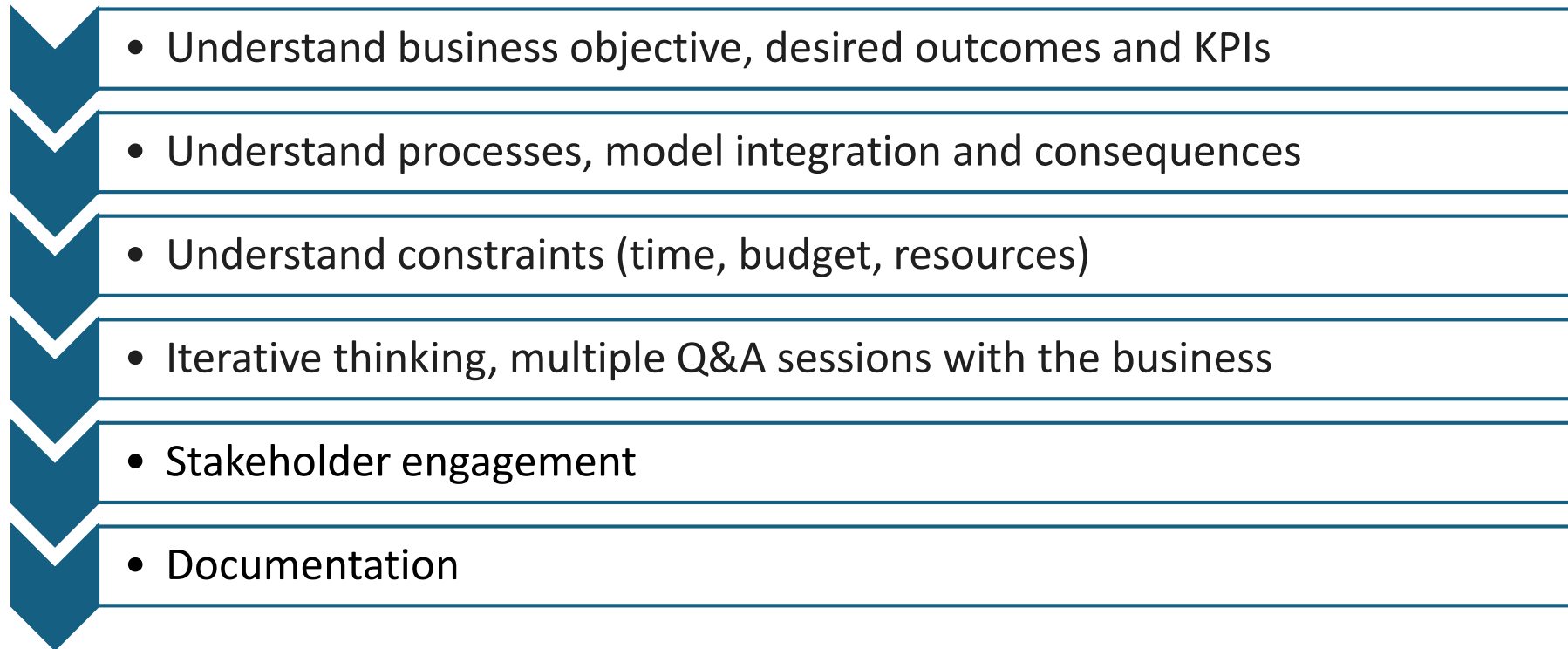
# The Importance of B&D Understanding

Good business/data understanding leads to:

- 
- Faster project delivery
  - Early discovery of limitations or even blockers
  - Better decisions at later steps
  - Setting the focus of the analysis
  - Better overall solution

# The Importance of B&D Understanding

## Requirements:

- 
- Understand business objective, desired outcomes and KPIs
  - Understand processes, model integration and consequences
  - Understand constraints (time, budget, resources)
  - Iterative thinking, multiple Q&A sessions with the business
  - Stakeholder engagement
  - Documentation

# Data Collection



# Data Collection

## Common Data Sources

- On-prem / Cloud Data Warehouse
  - Customer data
  - Product data
  - Transactional data
  - Financial data
- Sensor data (IoT)
- Third-Party data
- Web scraping

## Challenges

- Quality
- Volume
- Security
- Privacy
- Integration
- Processing
- Governance

# Data Collection

Extract > Transform > Load (ETL)

Tools for data ingestion, processing and storing:


- Apache: Kafka, Hive, Spark, Hadoop
- GCP: Data Fusion, Dataproc, Dataflow, BigQuery

*Data Collection itself is usually the responsibility of the **Data Engineer**. However, the **Data Scientist** is responsible for*

- *specifying data needs and the required format*
- *monitoring data quality*
- *collaborating with Data Engineers to solve data issues*

# Data Collection

## Requirements:

- 
- Data source reliability
  - Proper data collection methodology
  - Availability and consistency over time
  - Data ingestion automation
  - Data completeness, bias and limitations
  - Process Documentation and Data Dictionary

# Exploratory Data Analysis

# Exploratory Data Analysis (EDA)

## 1. Data Structure Assessment

- Table Relations & Joining
- Number of rows & columns (observations & features)
- Data Types
- Column Descriptions
- Data Granularity & IDs
- Long and Wide Format
- Aggregation

## 2. Data Quality Assessment

- Data Errors
- Duplications
- Missing Values
- Outliers
- Label Correctness
- Data Inconsistencies

# Exploratory Data Analysis (EDA)

## 3. Looking for Patterns

- Distributions
- Relationships
- Data Bias
- Trend and Cointegration
- Data Drift
- Data Segments

## 4. Methods

- Data Visualization
  - Gain insights: understand patterns, detect anomalies
  - Communicate results to non-technical stakeholders: enhanced storytelling
- Statistical Analysis
  - Descriptive
  - Inferential
- Ask business/data owner

# Exploratory Data Analysis (EDA)

## Univariate Analysis

- Feature distributions
  - Histogram
  - Bar plot
  - Box plot
- Descriptive statistics:
  - mean, std
  - median, IQR
  - min, max

## Multivariate Analysis

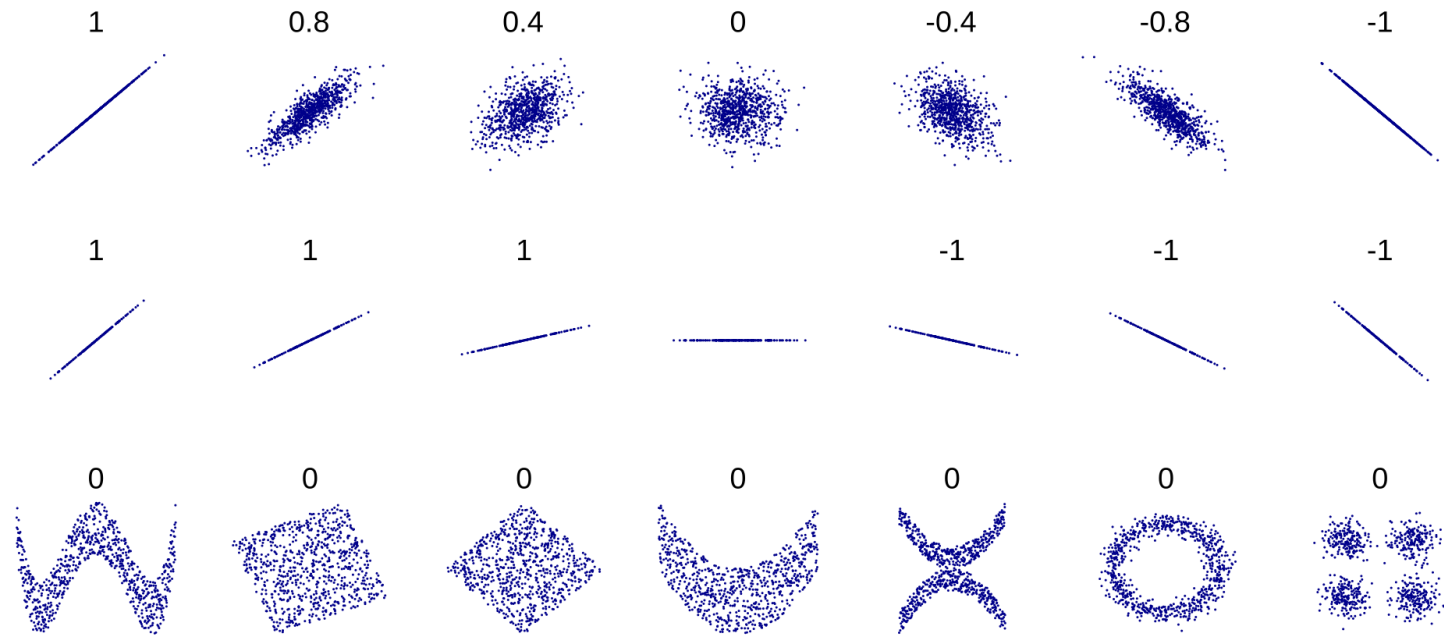
- Correlations (Pearson, Spearman)
- Contingency tables (crosstabs)
- Joint distributions
  - scatter plot
  - box plot
  - heatmap

# Correlation

## Pearson correlation

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$



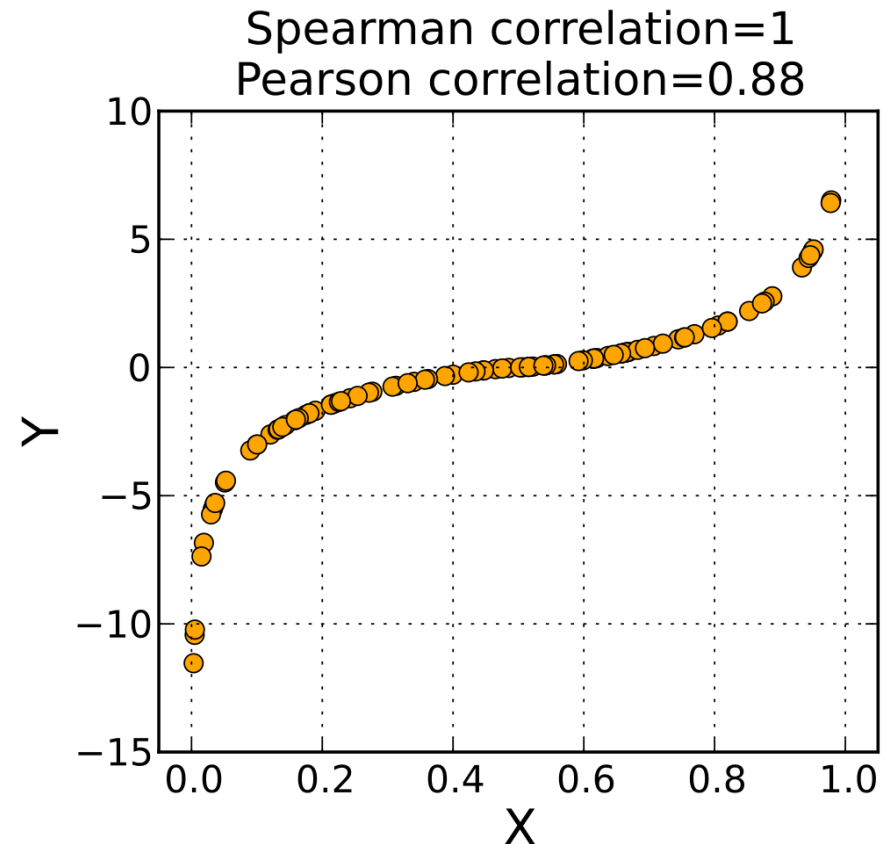
- The correlation reflects the strength and direction of a linear relationship (top row),
- but not the slope of that relationship (middle),
- nor many aspects of nonlinear relationships (bottom)



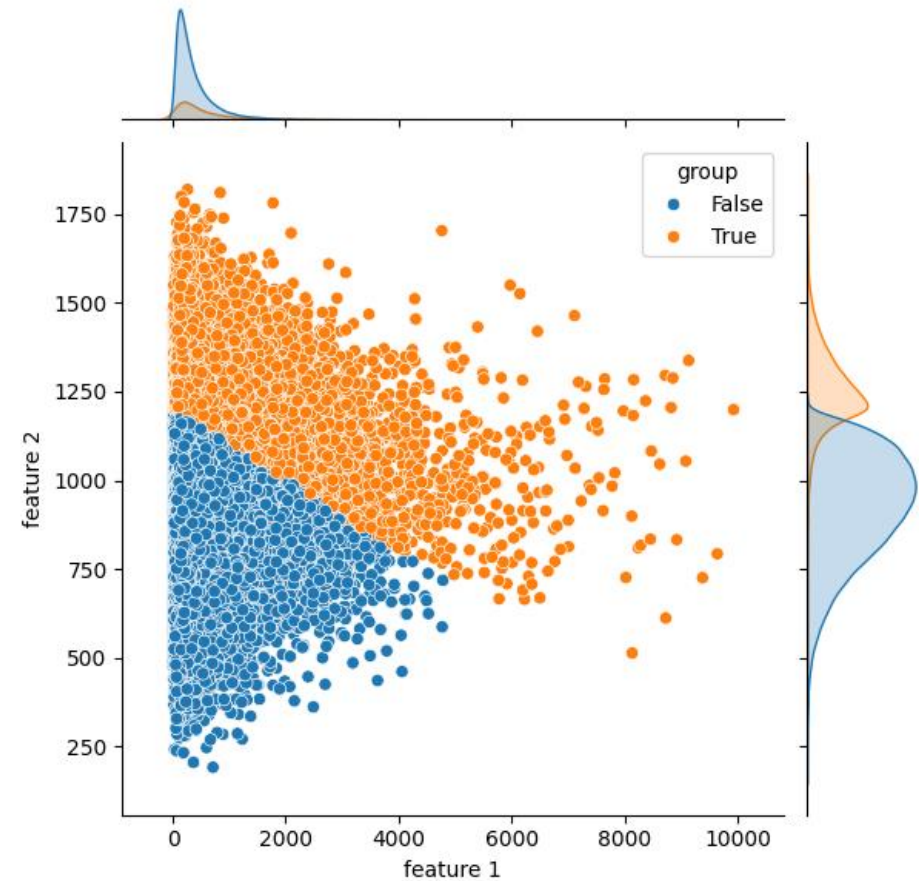
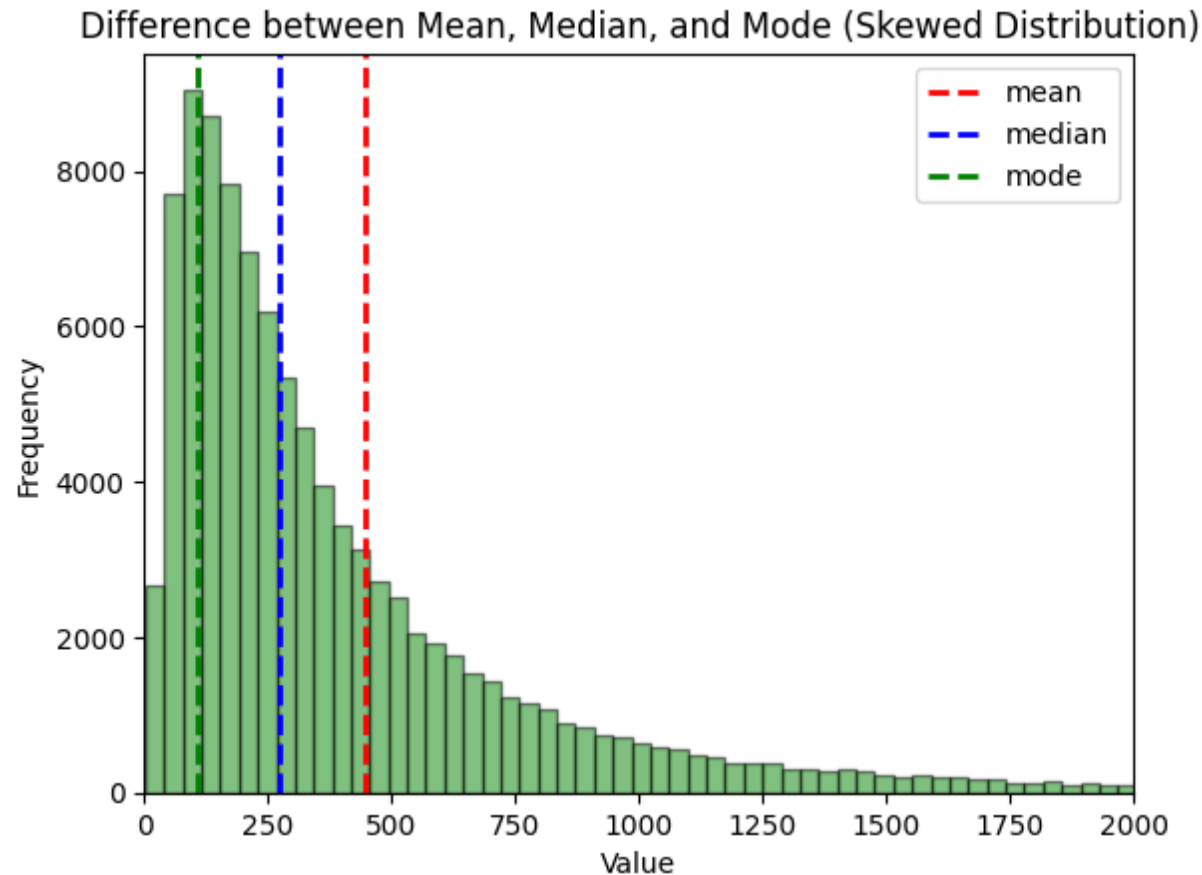
# Correlation

## Spearman's rank correlation

- Spearman's correlation assesses monotonic relationships (whether linear or not)
- The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables
- $r_S = \rho(R[X], R[Y])$



# Exploratory Data Analysis (EDA)

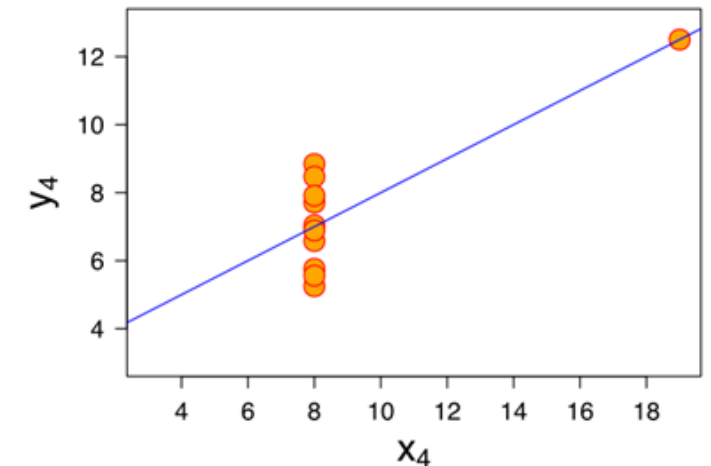
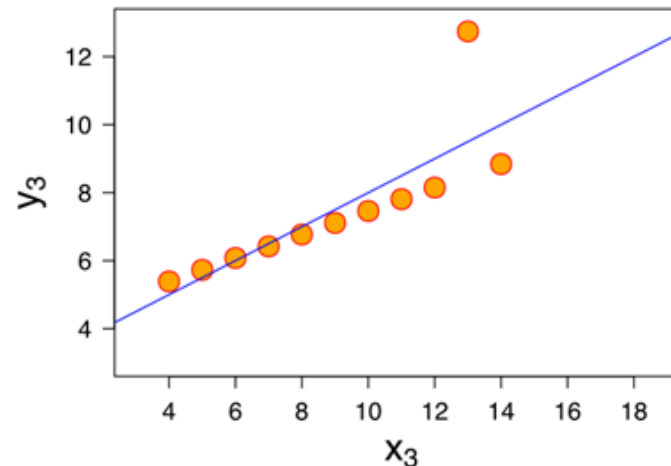
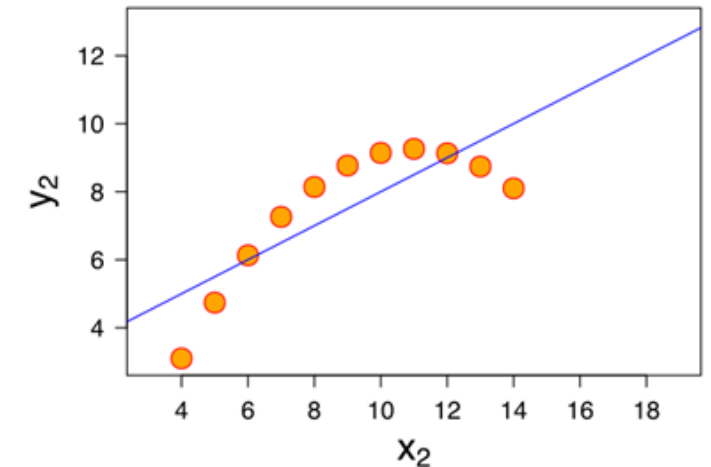
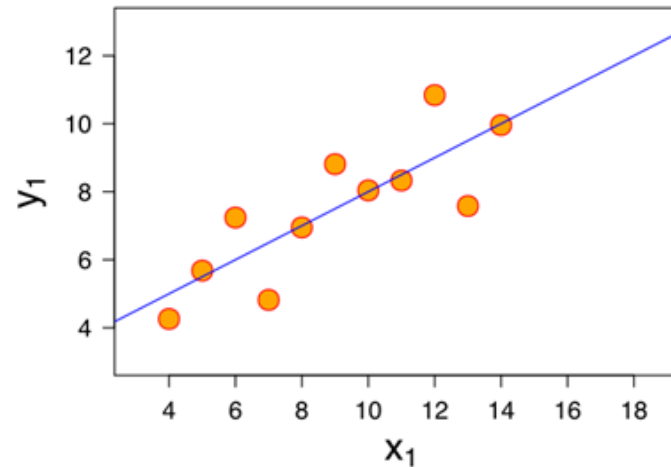


# Exploratory Data Analysis (EDA)

## Anscombe's quartet

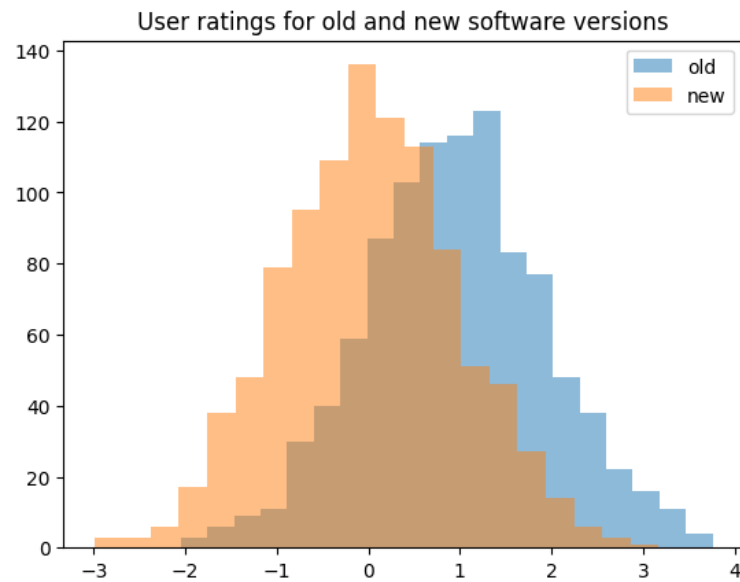
*Same mean, variance and correlation*

- The importance of visualization
- Pearson vs Spearman correlation
- OLS vs Robust regression
- Outliers and influential points



# Hand's paradox

	Old version	New version
Distribution	Normal	Normal
Mean	1	0
Std	1	1

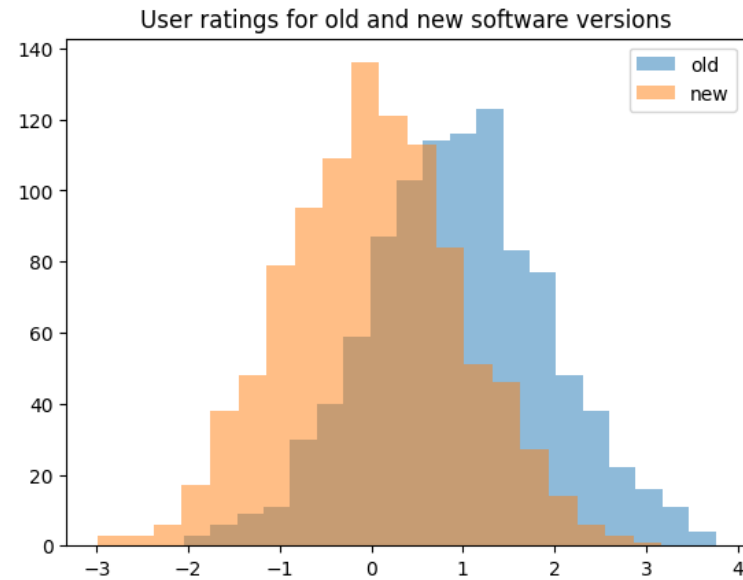


At first glance:

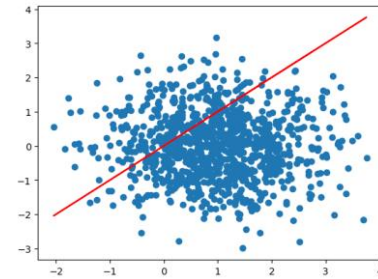
- People liked the old version better on average
- In healthcare: patients responded better to the old treatment

# Hand's paradox

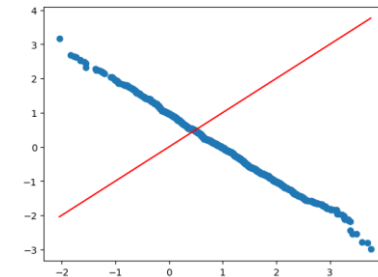
	Old version	New version
Distribution	Normal	Normal
Mean	1	0
Std	1	1



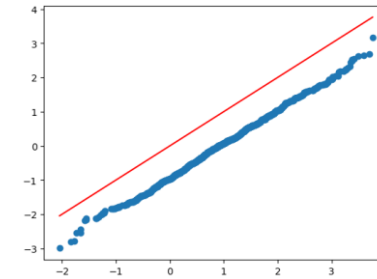
Correlation 0  
Positive change rate 24%



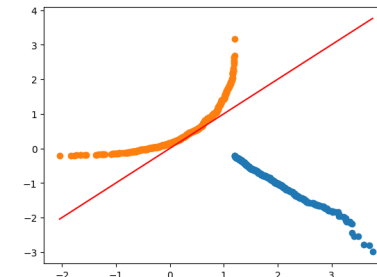
Correlation -1  
Positive change rate 30%



Correlation 1  
Positive change rate 0%

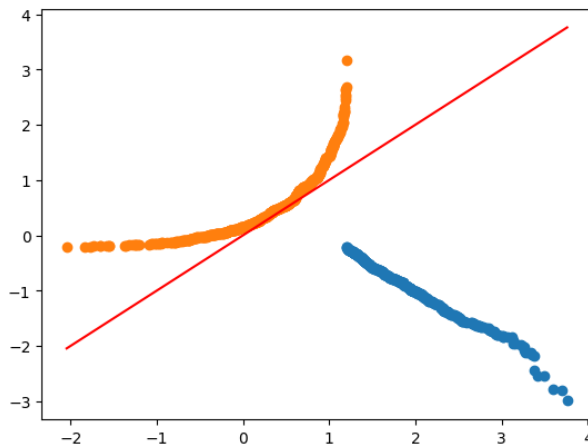


Correlation -0.5  
Positive change rate 60%

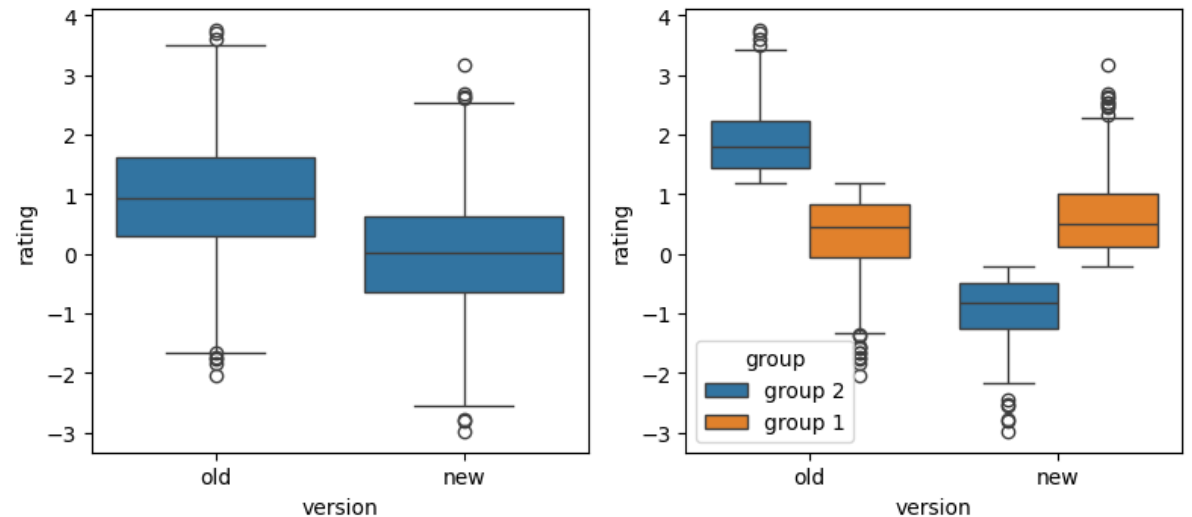


# Hand's paradox

	Old version	New version
Overall mean	1	0
Group 1 mean	0.3	0.6
Group 2 mean	1.9	-0.9
Correlation		-0.5
Positive change rate		60%



User ratings for old and new software versions




- Old version has a higher average rating
- However, more users prefer the new version
- Moreover, results are different in business segments
- What is the business goal?

A comparison between two randomly chosen patients, one from each group, and a comparison of treatment effects on a randomly chosen patient, can lead to different conclusions. ([Hand's paradox](#))

# Exploratory Data Analysis (EDA)

## Conclusions:

- 
- Identify issues and limitations
  - Find insights
  - Collaborate with business stakeholders and data owners
  - Document findings and communicate to stakeholders
  - Guide data cleaning
  - Guide feature selection and engineering
  - Guide model selection

# Statistical Inference



# Statistical Inference

*Statistical inference helps to determine if observed differences or relationships are **statistically significant or due to random chance**. Useful for validating research hypotheses, making data-driven decisions, and avoiding false conclusions.*

- Prerequisite: Probability Theory
- Methods:
  - Estimation (Point and Interval)
  - Hypothesis Testing
  - Regression Analysis
- Python packages: [scipy](#), [statsmodels](#)
- There are many use cases that don't need ML, only EDA and statistical testing

# Conditional Probability and Bayes' Theorem

## Conditional Probability

The probability of an event occurring given that another event has already occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

## Law of total probability

The total probability of an outcome which can be realized via several distinct events (scenarios).

$$\begin{aligned} P(B) &= \sum_n P(B \cap A_n) = \\ &= \sum_n P(B|A_n)P(A_n) \end{aligned}$$

## Bayes' Theorem

A mathematical rule for inverting conditional probabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_n P(B|A_n)P(A_n)}$$

# Three prisoners problem

- Three prisoners are sentenced to death. The governor has selected one of them at random to be pardoned. The warden knows which one is pardoned. Prisoner A begs the warden to let him know the identity of one of the two who are going to be executed.
- "If B is to be pardoned, give me C's name. If C is to be pardoned, give me B's name. And if I'm to be pardoned, secretly flip a coin to decide whether to give me name B or C."
- The warden gives him B's name.
- What is the probability that A / C will be pardoned?

# Three prisoners problem

- $A_n$  is the event that the corresponding prisoner will be pardoned
- $b$  is the event that the warden tells A that prisoner B is to be executed

$$P(A|b) = \frac{P(A \cap b)}{P(b)}$$

$$P(A_1|b) = \frac{P(b|A_1)P(A_1)}{P(b|A_1)P(A_1) + P(b|A_2)P(A_2) + P(b|A_3)P(A_3)} = \frac{\left(\frac{1}{2} * \frac{1}{3}\right)}{\left(\frac{1}{2} * \frac{1}{3}\right) + \left(0 * \frac{1}{3}\right) + \left(1 * \frac{1}{3}\right)} = \frac{1}{3}$$

$$P(A_3|b) = \frac{P(b|A_3)P(A_3)}{P(b|A_1)P(A_1) + P(b|A_2)P(A_2) + P(b|A_3)P(A_3)} = \frac{\left(1 * \frac{1}{3}\right)}{\left(\frac{1}{2} * \frac{1}{3}\right) + \left(0 * \frac{1}{3}\right) + \left(1 * \frac{1}{3}\right)} = \frac{2}{3}$$

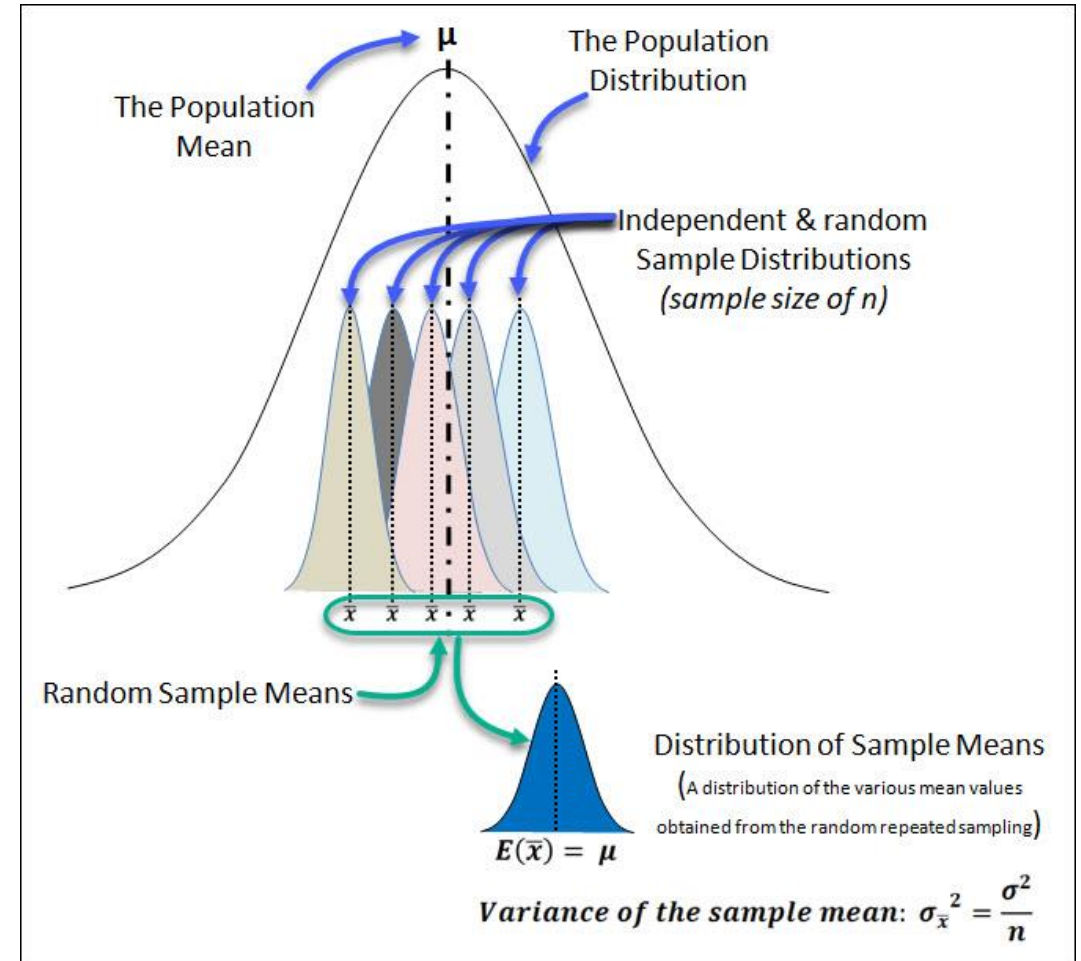
# Confidence Intervals

# Interval Estimation

## Central Limit Theorem (CLT)

- The sample mean is approx. normally distributed around the population mean (regardless of the original distribution)
- As the sample size increases, the variance of the sample mean decreases

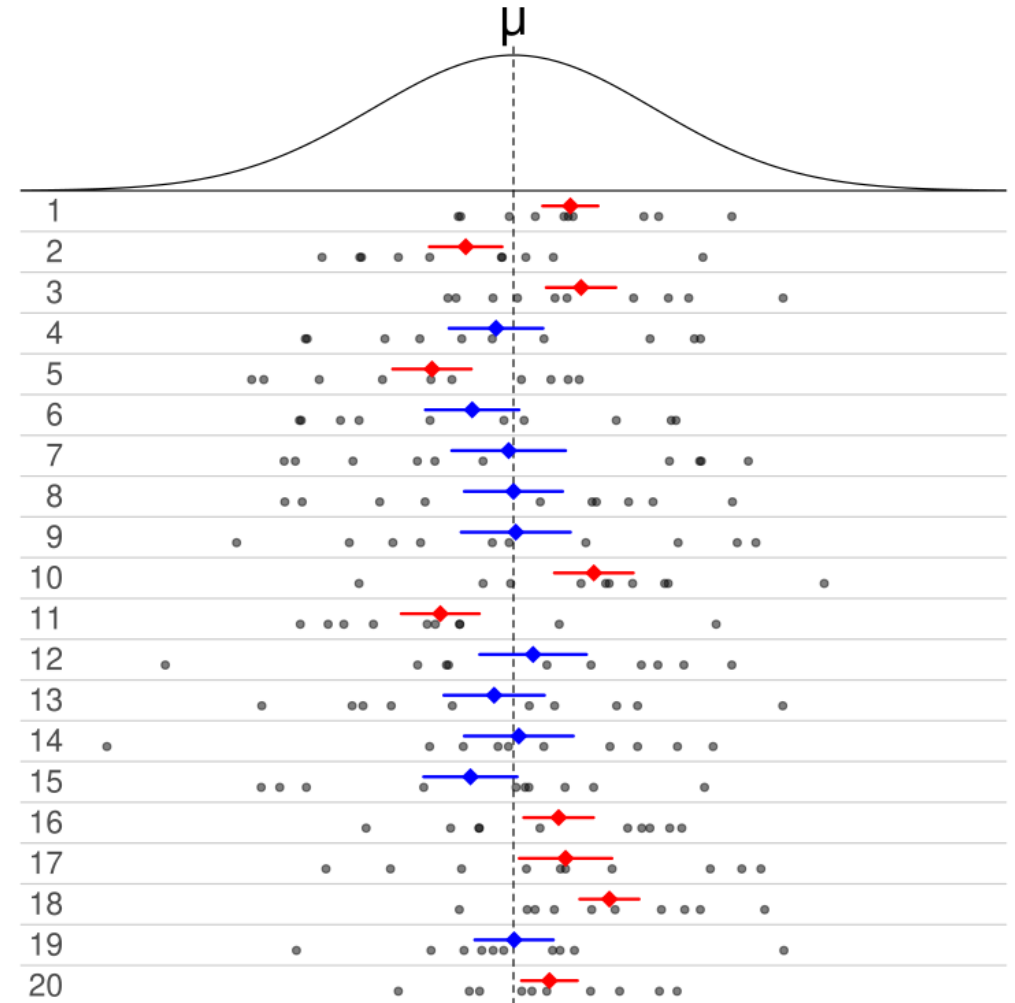
$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



# Interval Estimation

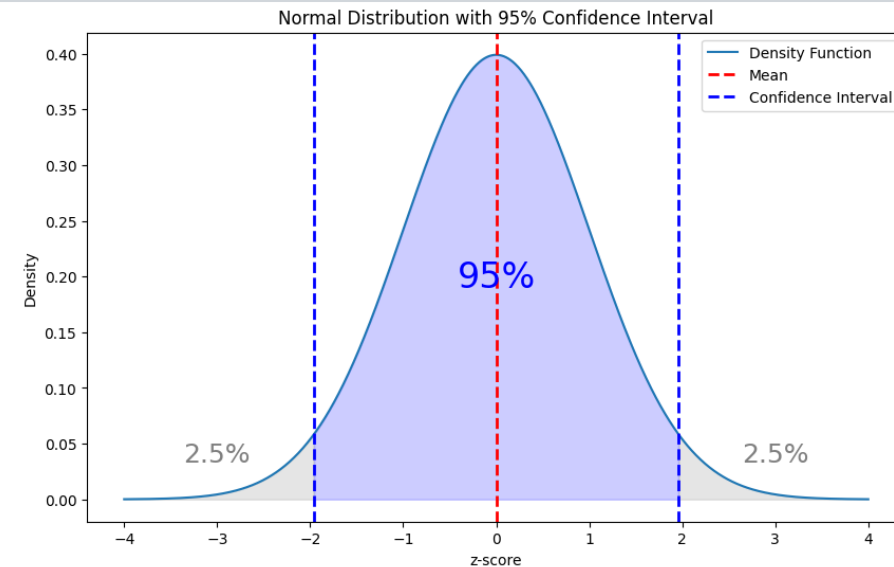
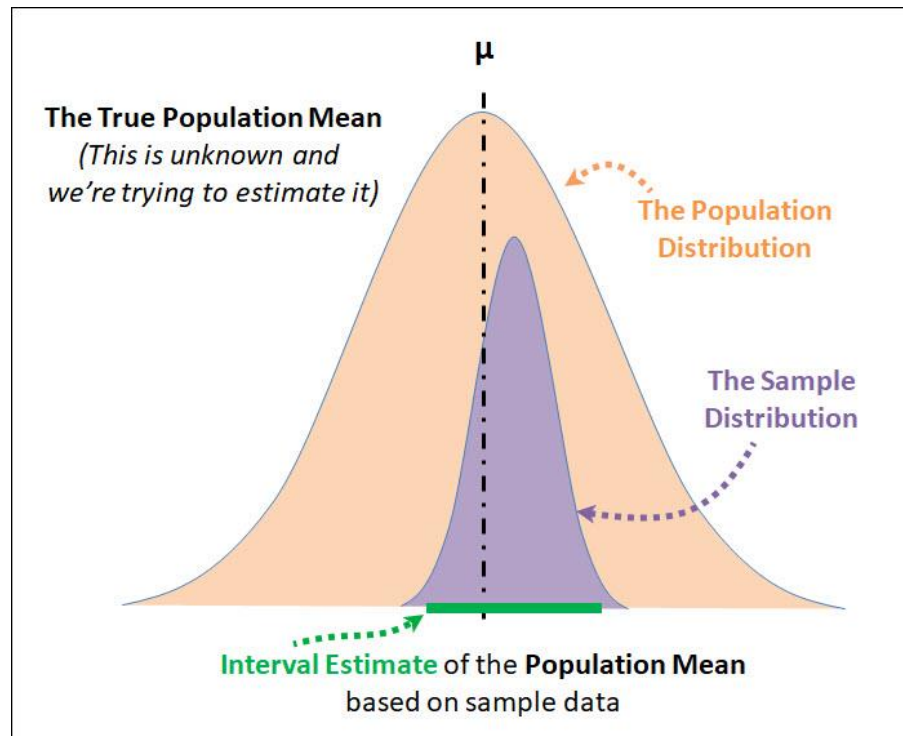
## 95% Confidence Interval:

- If we repeated the sampling process many times, 95% of the confidence intervals would include the true population mean.
- The higher the confidence level, the wider the confidence interval



# Interval Estimation

The **z-score** is the **quantile function** (or inverse cumulative distribution function) of the **standard normal distribution**. For a normal distribution, the z-score measures how far a given quantile is from the **mean**, in units of the **standard deviation**.



For a 95% **confidence level** we calculate  $z(0.975) = 1.96$ . This means that 95% of the possible **sample means** fall within 1.96 **standard error** from the true **population mean**. So, the **interval estimate** equals the **point estimate** plus/minus 1.96 times the standard error:

$$CI = \left( x - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, x + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$



# Hypothesis Testing

# Hypothesis Testing

## Steps:

1. Formulate **Null and Alternative Hypotheses**

- *Null Hypothesis ( $H_0$ ): There is no effect or difference.*
- *Alternative Hypothesis ( $H_1$ ): There is an effect or a difference.*

2. Choose a **significance level** ( $\alpha$ )

3. Calculate the test statistic and the p-value

*The **test statistic** is a value derived from the sample data, that follows a specific distribution under the null hypothesis. It helps determine the **p-value**, which is: the probability of observing the sample data, or something more extreme, given that the null hypothesis is true*

4. Based on the p-value, **accept or reject  $H_0$**

*Reject  $H_0$  if  $p < \alpha$*

# Student's t-distribution

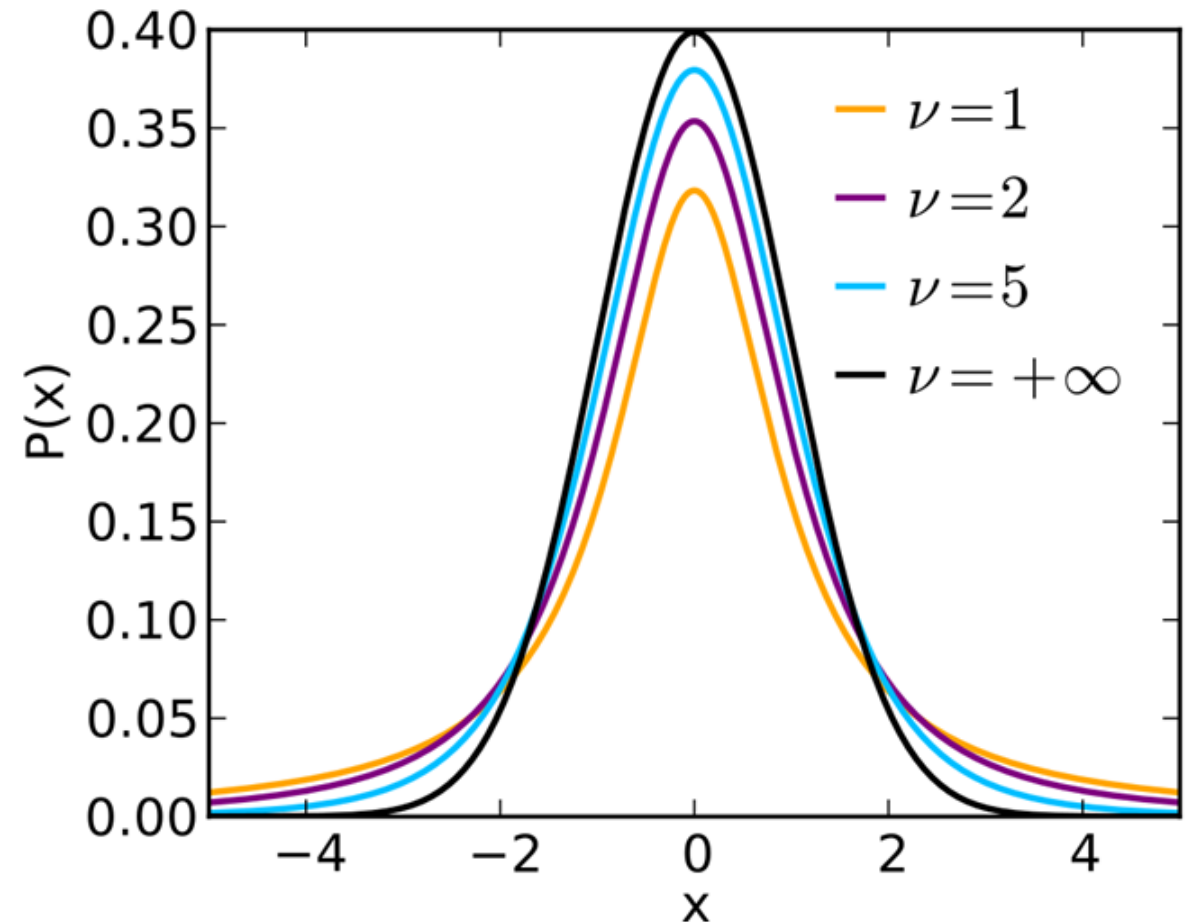
- The t-distribution is the generalization of the standard normal distribution (with heavier tails)
- It is used when the sample size is small, and the population variance is unknown

## Student's t-test

- The test statistic follows the t-distribution under the null hypothesis:

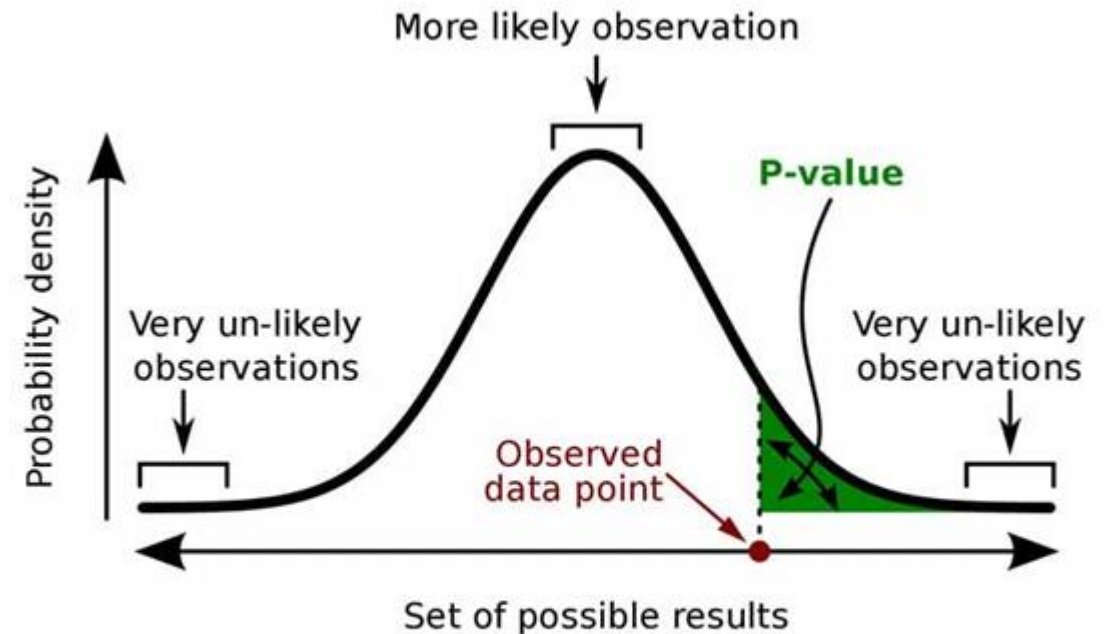
$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

- It measures the distance between the sample mean and the assumed population mean in units of the standard error.



# Understanding the p-value

- If the p-value is very low, then we can conclude that a difference exists (we reject  $H_0$ ) because it would be unlikely to get this sample if  $H_0$  was true.
- If the p-value is high, it means that our sample would not be extreme if  $H_0$  was true. Hence, we cannot reject  $H_0$
- We determine alpha before testing, as a threshold for the p-value
- Note: the p-value is not the probability of  $H_0$  being true! A high p-value does not mean that  $H_0$  is proven, only that we can't reject it based on the data.

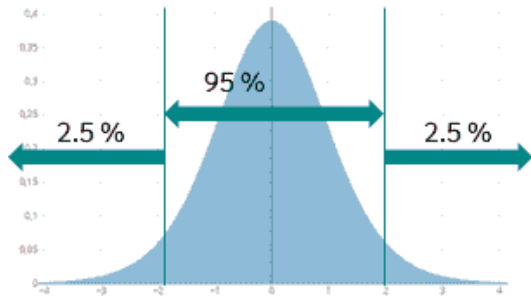


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

# One-sample t-test

$$\begin{aligned}H_0 : \mu &= 28 \\ H_1 : \mu &\neq 28 \\ \alpha &= 0.05\end{aligned}$$

Two tailed (non-directional)



Student	Score
1	28
2	29
3	35
4	37
5	32
6	26
7	37
8	39
9	22
10	29
11	36
12	38

Significance level

$$\alpha = 0.05$$

Number of sample values

$$n = 12$$

Mean value

$$\bar{x} = 32.33$$

Standard deviation

$$s = 5.47$$

Standard error of the mean

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{5.47}{\sqrt{12}} = 1.58$$

t-value

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{32.33 - 28}{1.58} = 2.75$$

Degrees of freedom

$$df = n - 1 = 11$$

We calculate the p-value from the test statistics (t-value) based on the Student's t-distribution.

**p-value = 0.02** which means that the probability of a sample like this (or more extreme), if the null hypothesis is true, is 2%. **We reject the null hypothesis.**

# Common Tests

## **t-tests:**

1. One-Sample t-test: Compares the sample mean to a known or hypothesized population mean.
2. Two-Sample t-test: Compares the means of two independent samples
3. Paired Sample t-test: Compares the means from the same group at different times or under different conditions.

## **F-tests:**

1. F-test of equality of variances: Compares the variances of two samples
2. ANOVA (Analysis of Variance): Compares the means of three or more independent samples

## **Chi-squared tests:**

1. Goodness of Fit Test: Determines if a sample data matches a population with a specific distribution.
2. Test of Independence: Assesses whether two categorical variables are independent of each other.

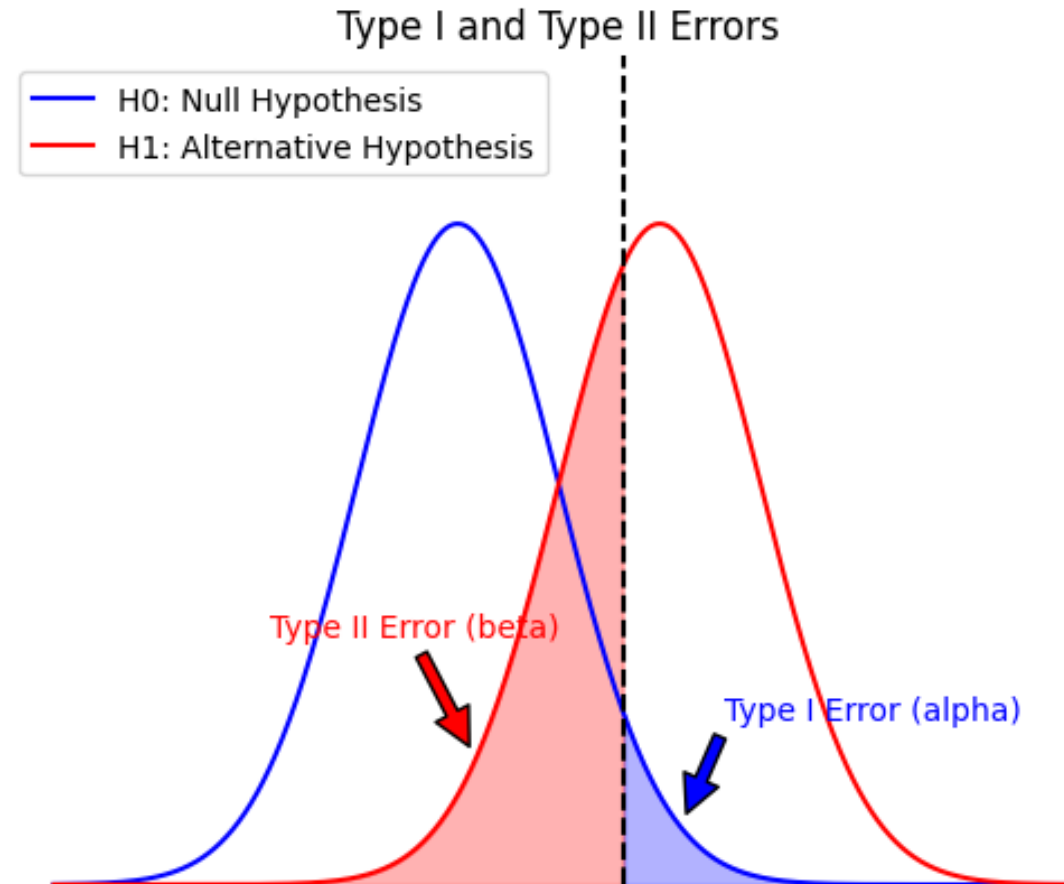
*Statistical tests are widely used to test hypotheses and validate results in medical studies and surveys, as well as for EDA, feature selection, evaluation and time series analysis.*

# Common Mistakes

# Common Mistakes in Statistics

## Misinterpreting the p-value

- p-value is *not* the probability of the null hypothesis being true
- p-value is *not* the probability of making a mistake
- p-value does *not* indicate the size of the observed effect
- alpha and beta are conditional probabilities
- p-hacking or data dredging
  - Multiple comparisons (look-elsewhere effect) and selective reporting

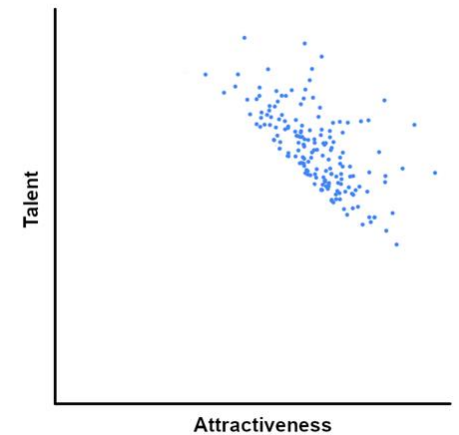
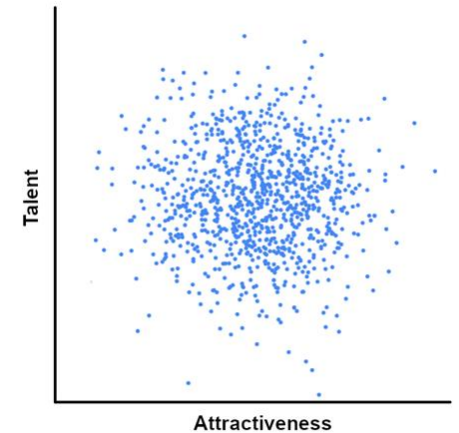
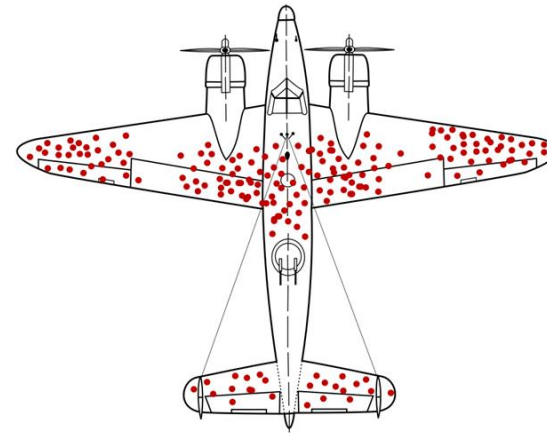




# Common Mistakes in Statistics

## Incorrect conclusions

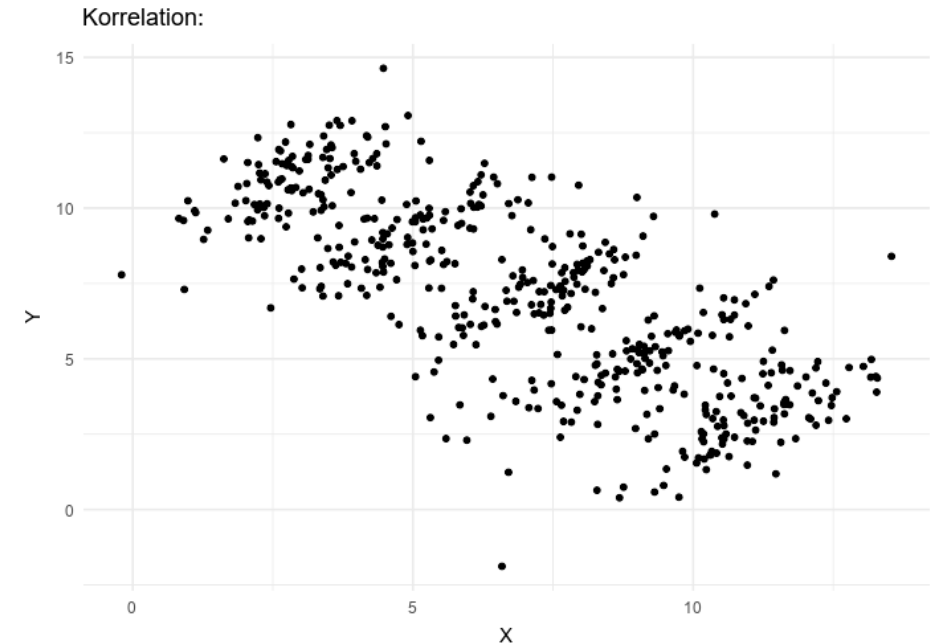
- Small sample size, low statistical power
- Ignoring the Assumptions of Statistical Tests
  - e.g. normality, independence, equal variances
- Overgeneralization due to Selection/Sampling bias
  - Participation bias
  - Survivorship bias
  - Berkson's paradox



# Common Mistakes in Statistics

## Incorrect conclusions

- Small sample size, low statistical power
- Ignoring the Assumptions of Statistical Tests
  - e.g. normality, independence, equal variances
- Overgeneralization due to Selection/Sampling bias
  - Participation bias
  - Survivorship bias
  - Berkson's paradox
- Correlation does not imply causation
  - False correlation
  - Simpson's paradox
  - Confounder variables
- Confirmation bias
- Post hoc analysis (Texas sharpshooter fallacy)



Treatment Stone size	Treatment A	Treatment B
Small stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

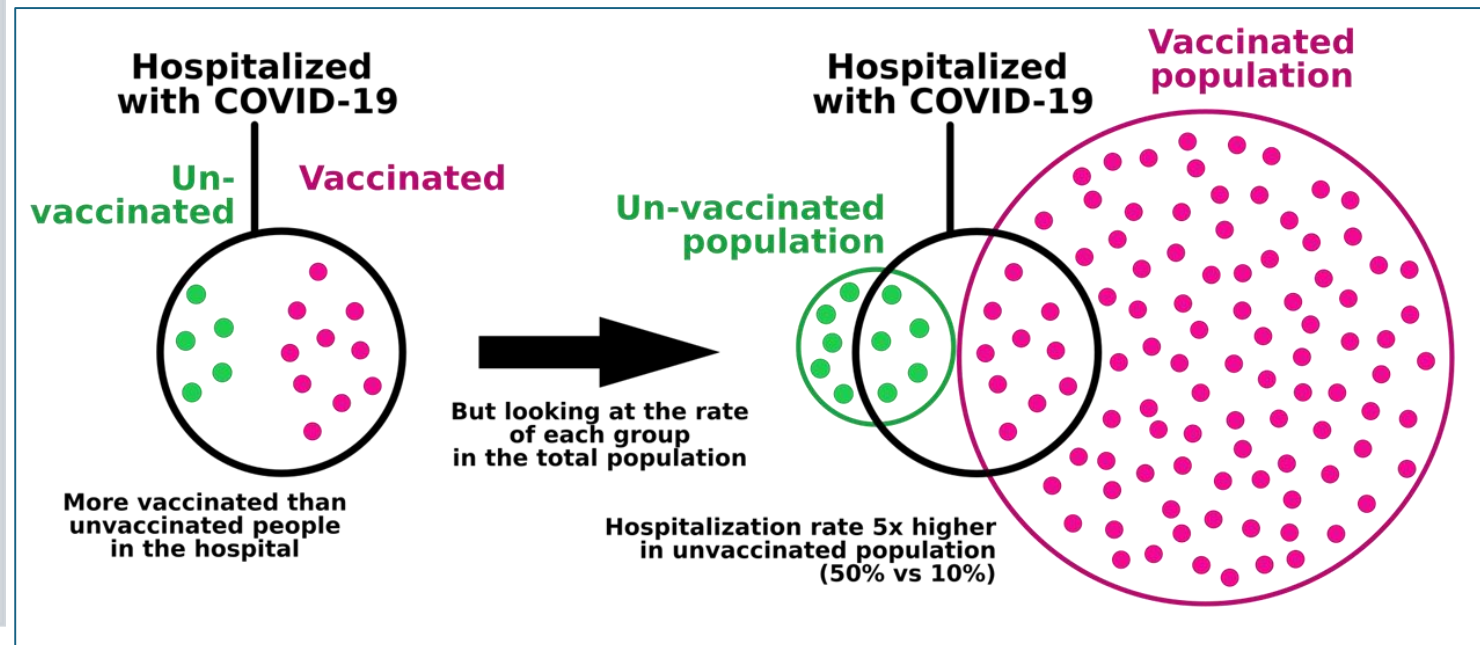
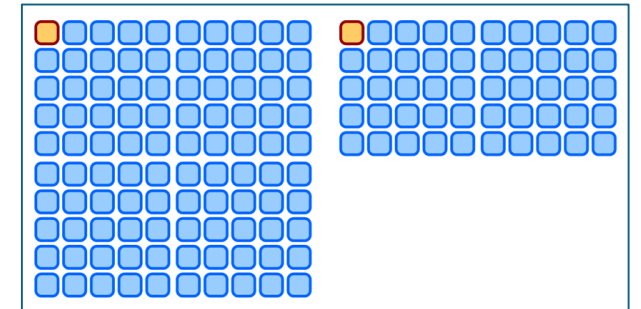
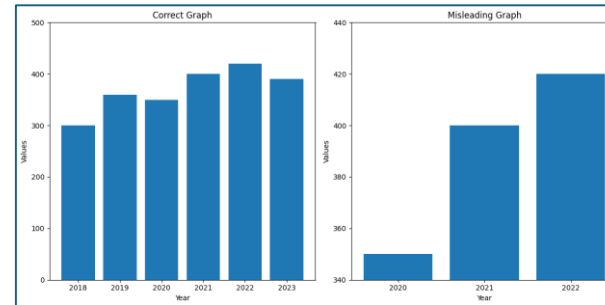


# Common Mistakes in Statistics

## Misinterpretations

- Data (e.g. outliers, skewness)
- Statistical Significance (p-values) versus Practical Significance (effect size)
- Loaded questions
- Misleading graph
- Potato paradox
- Preparedness paradox
- Base Rate Fallacy
  - Prevention paradox
  - False positive paradox

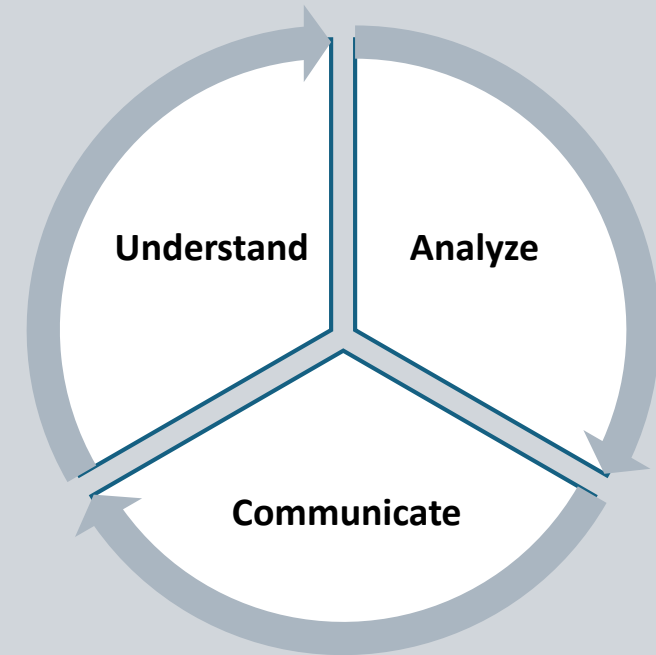
	Positive	Negative
Drunk	2	0
Sober	100	1900



# Case Studies

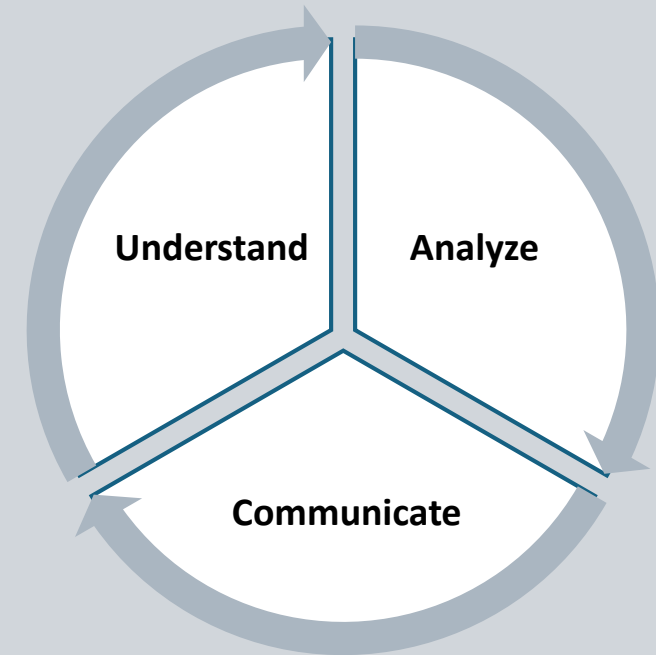
# Case Study – Customer churn

1. Business target
  - Churn in the given month
  - Churn in the next 6 months
2. Data availability at inference time
  - When was the data collected?
  - Usage data of the last month
3. Historical data
  - What to collect, what to calculate?
  - Storing data
  - New customers
4. Actions
  - No improvements?



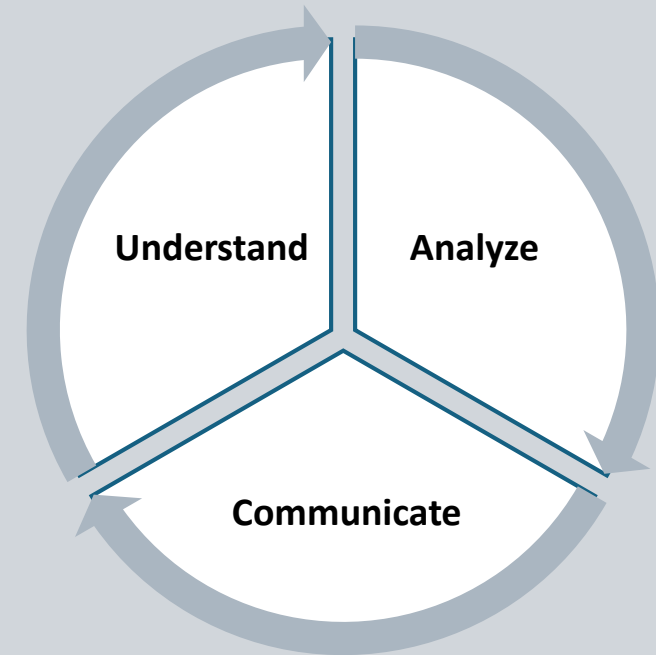
# Case Study – Movie popularity prediction

1. Business target
  - Views or ratings
  - Base drift
  - Introduce – Where?
  - Introduce – Which?
2. Data availability at inference time
  - External ratings, popularity
3. Historical data
  - Representativeness
4. Actions
  - Marketing: self-fulfilling predictions



# Case Study – Anomaly Detection

1. Business target
  - Types: outlier, shift, variance, interactions
  - Special events
  - Seasonality
2. Link between data sources



# Business and Data Understanding

The Importance of B&D Understanding

Data Collection

Exploratory Data Analysis

Statistical Inference

Confidence Intervals

Hypothesis Testing

Common Mistakes

Case Studies



# Thank you for your attention!

Your feedback would be much appreciated:



## Any Questions?



Gergely Zsombor Haász



[haasz.zsombi@gmail.com](mailto:haasz.zsombi@gmail.com)