



DATA SCIENCE RETREAT®
SINCE 2014



Introduction to Probability & Statistics

Dr. Harald Stein
September 2025

1 day - Version 1



Content

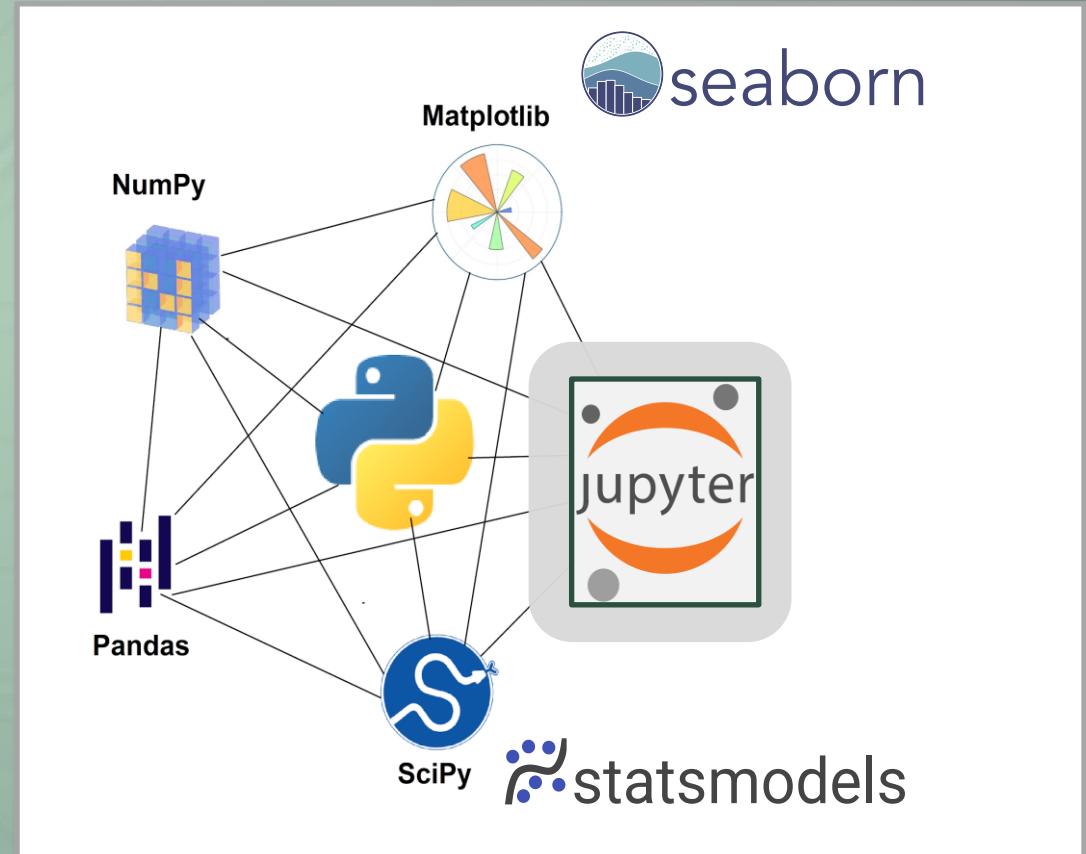
- **Technical Prerequisites**
- **Probability, Statistics and Information Theory**
- **Descriptive Statistics**
- **Combinatorics**
- **Probability Distributions**
- **Hypothesis Testing**
- **Model Estimation:**
 - Linear Regression
 - Binary Logistic Regression



Technical Prerequisites

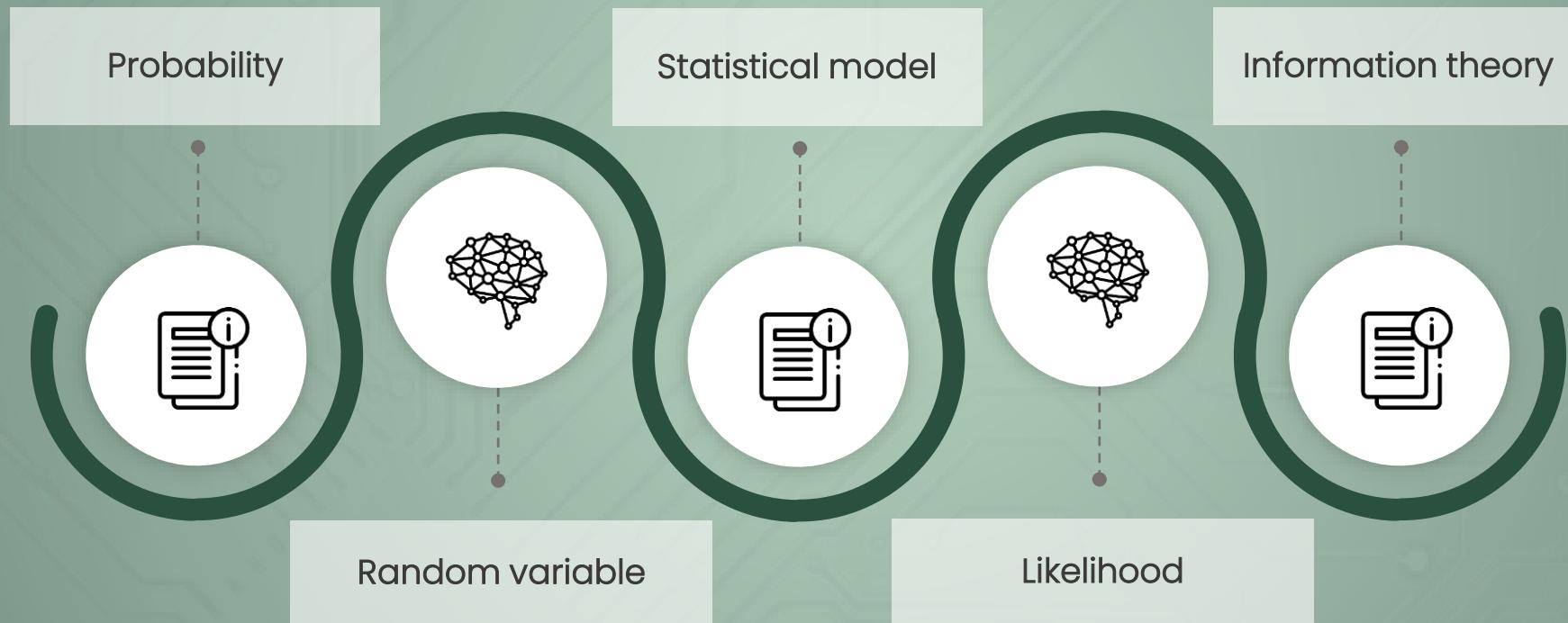
... means installation of Python environment with Jupyter

- GitHub Link:
https://github.com/MathforDataScience/DSR_Statistics1/
- List of packages:
requirements.txt
- How to install:
readme.txt
- Start Jupyter Notebook



Probability, Statistical Models and Information Theory

Probability provides foundation for statistical models, which use probability to analyze data, make predictions, while information theory quantifies uncertainty and information content within those models.



What is Probability?

Probability is measure quantifying the chance that random events will occur. It is expressed as number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

Example

- Tossing a fair (unbiased) coin:
 - Two possible outcomes: "heads" and "tails"
 - Probability of "heads" = Probability of "tails" = 0.5 or 50%
 - Random event

Application Areas

- Mathematics, Statistics, Finance, Gambling
- Science (including Physics)
- Artificial Intelligence/Machine Learning
- Computer Science, Game Theory, Philosophy

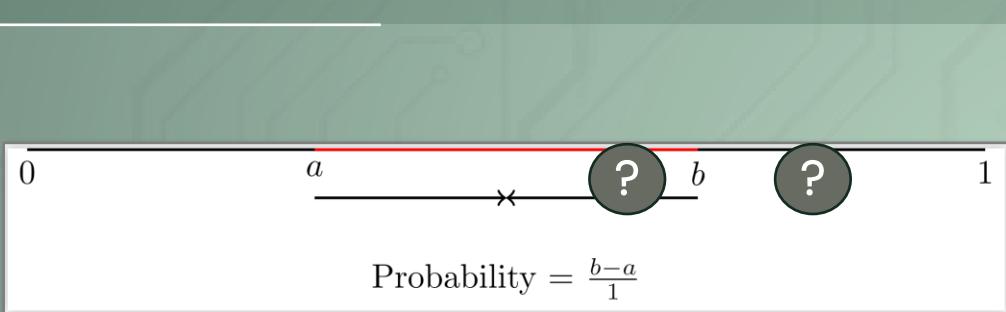
Purpose of Probability

- Used to infer the expected frequency of events.
- Describes the underlying mechanics and regularities of complex systems.

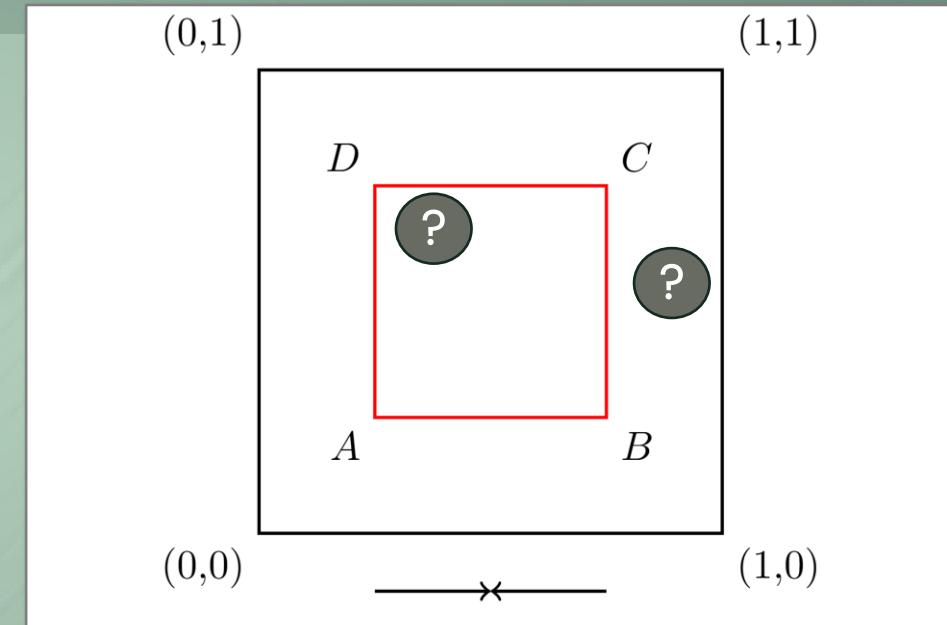


Geometric Interpretation of Probability

... visualizes probabilities as areas (or volumes) within geometric space. Probability is proportional to size of favorable region relative to total possible region.



- Consider a point chosen at random on a line segment of length 1
- The probability that the point falls within the sub-segment is $\frac{b-a}{1}$



$$\text{Probability} = \frac{\text{Area of } ABCD}{\text{Area of Unit Square}} = \frac{(b-a)^2}{1} = (b-a)^2$$

- Suppose a point is randomly selected within a unit square.
- The probability that the point lies within specific region (e.g., smaller square or circle) is equal to area of that region divided by area of the unit square (1)

Random Variable

... is variable that takes on numerical values determined by the outcome of random experiment. It maps outcomes of random process to numbers.

Aspect	Discrete Random Variable	Continuous Random Variable
Definition	Takes on a countable number of distinct values.	Takes on an infinite number of possible values within a given range.
Examples	Number of heads in 10 coin flips, number of students in a class.	Height of students, time to run a marathon.
Possible Values	Finite or countably infinite set (e.g., $\{0, 1, 2, \dots\}$).	Any value within a continuous range (e.g., $[0, \infty)$, $(-\infty, \infty)$).
Probability Distribution	Described by a probability mass function (PMF).	Described by a probability density function (PDF).
Cumulative Distribution	Step function with jumps at each possible value.	Smooth curve, the integral of the PDF.
Sum of Probabilities	Sum of probabilities for all possible values equals 1.	The area under the PDF curve equals 1.
Graphical Representation	Histogram or bar chart	Continuous curve (e.g., bell curve for normal distribution).
Real-World Analogy	Counting the number of occurrences (e.g., number of cars passing).	Measuring continuous quantities (e.g., measuring the exact weight).



Examples of Random Variables

Discrete vs. continuous

Type of Random Variable	Example	Description
Discrete Random Variable	Number of Students in a Classroom	Can only take specific integer values (e.g., 20, 21, 22).
	Number of Cars Passing Through a Toll Booth	Count of cars, taking on values like 0, 1, 2, etc.
	Number of Emails Received in a Day	Exact count of emails received (e.g., 5 emails).
	Number of Defective Items in a Batch	Number of defective items, counted in whole numbers (e.g., 0, 1, 2)
Continuous Random Variable	Height of Students in a Class	Can take any value within a range, measured in units like centimeters
	Time Taken to Run a Marathon	Measured in hours, minutes, and seconds with infinite possible values.
	Temperature Throughout the Day	Can be any value within a temperature range, measured in degrees (e.g., 25.3°C).
	Amount of Milk in a Bottle	Volume can be measured to any degree of precision (e.g., 1.5 liters).

Statistical Model

... is mathematical framework that represents relationships between variables in data and allows for analysis, inference, and prediction

Variables:

- Dependent Variable: The outcome or response being predicted or explained.
- Independent Variables: The predictors or inputs that influence the dependent variable.

Parameters

- Constants estimated from the data that define the model's specific form.
- Through estimated parameters the statistical model becomes generalization rule for the data sample referring to population

Probability Distributions

Assumptions about how data is distributed (e.g., normal distribution).

Example: Linear Regression $Y = \beta_0 + \beta_1 X + \varepsilon$

- Y : Outcome, dependent variable
- X : Predictor, independent variable
- β_0, β_1 : Parameters that are estimated
- ε : Random error, unpredictable part of model



Likelihood vs Probability

Likelihood measures how well a set of parameters explains observed data.

Probability measures the chance of a specific event occurring.

Aspect	Probability	Likelihood
Definition	The measure of the chance that a specific event will occur.	The measure of how likely a particular set of parameters is, given the observed data.
Context	Used to describe the probability of future events based on known conditions.	Used in statistical models to estimate parameters based on observed data.
Focus	Focuses on predicting the occurrence of outcomes.	Focuses on finding the best parameters that explain the observed data.
Example	Probability of rolling a 6 on a fair die is 1/6.	Likelihood of a die being biased towards 6 given repeated observations of rolling a 6.
Mathematical Form	Expressed as $P(A)$, where A is the event.	Expressed as $L(\theta X)$, where θ are parameters and X is data.
Use in Inference	Probability helps to infer about future events or outcomes.	Likelihood is used to estimate the parameters of a statistical model.

→ Statistical model is good predictor, i.e. generalization of data, if parameters are approximated in a way that likelihood is maximized.



Information Theory

... is branch of applied mathematics focused on quantifying information in signals.

Main idea: Learning about unlikely event is more informative than learning about likely event.

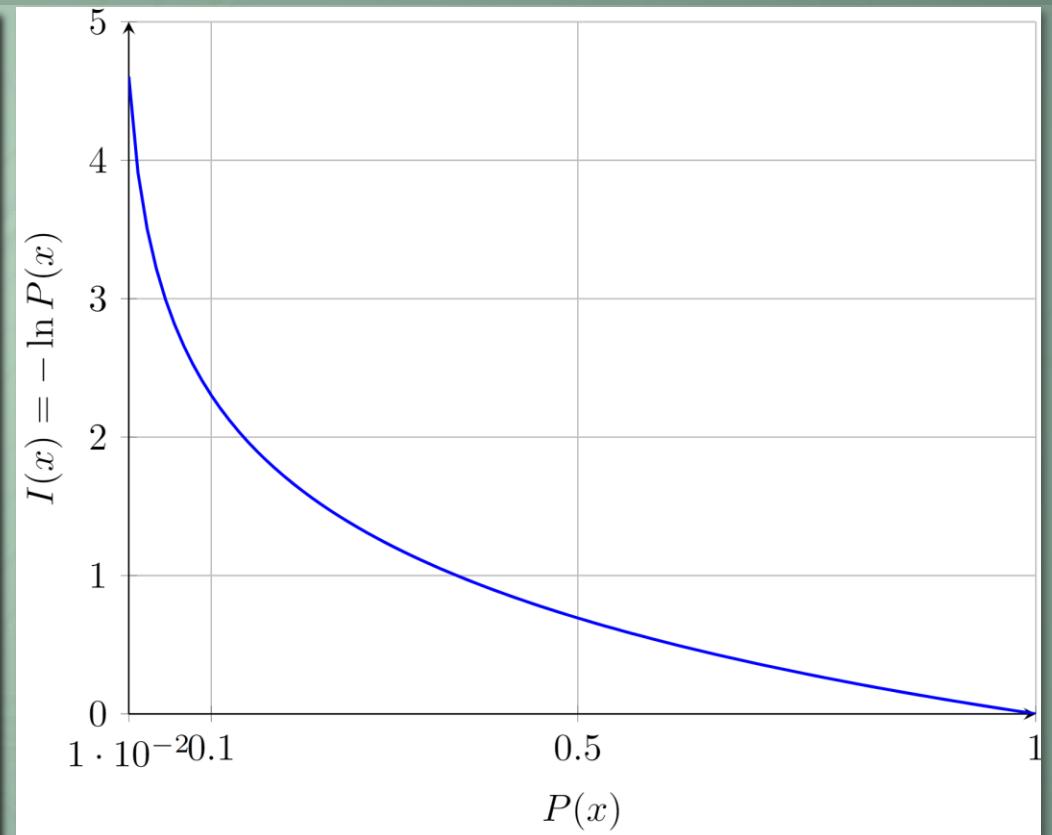
Key Properties of Information:

- Low frequency events: Have high information content.
- High frequency events: Have low information content.
- Independent events: Information should be additive.

Self-Information Formula

- mathematical way to quantify how much information is gained from observing outcome of random event.
- Rare events are more informative, surprising than common ones

$$I(x) = -\ln P(x)$$



Shannon Entropy

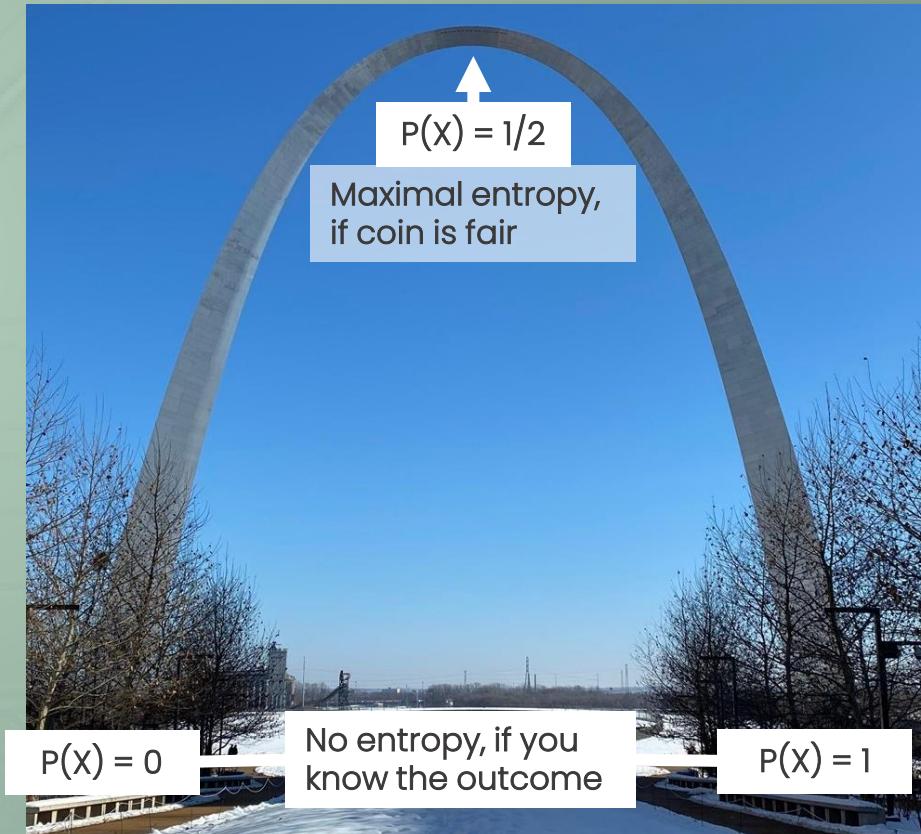
... is measure of randomness in set of possible outcomes. It quantifies expected amount of information/surprise, that event drawn from a probability distribution will provide.

- Weighted average of self-information, i.e. $-\ln P(x)$ of all possible outcomes.
- Represents level of uncertainty/surprise inherent in outcome, examples:
 - **High Entropy:**
Fair six-sided die has higher entropy than biased die where one number appears with very high probability.
 - **Low Entropy:**
If die is heavily biased towards one particular number, entropy is low.

Formula

$$H(X) = - \sum_{i=1}^n P(x_i) \ln P(x_i)$$

→ If any of the terms $P(x_i), \ln P(x_i)$ become 0, then $H(X)$ becomes 0



Gateway Arch, St. Louis, USA

Shannon Entropy and Outliers

Outliers can increase the entropy if they represent rare but possible outcome, adding to the distribution's unpredictability.



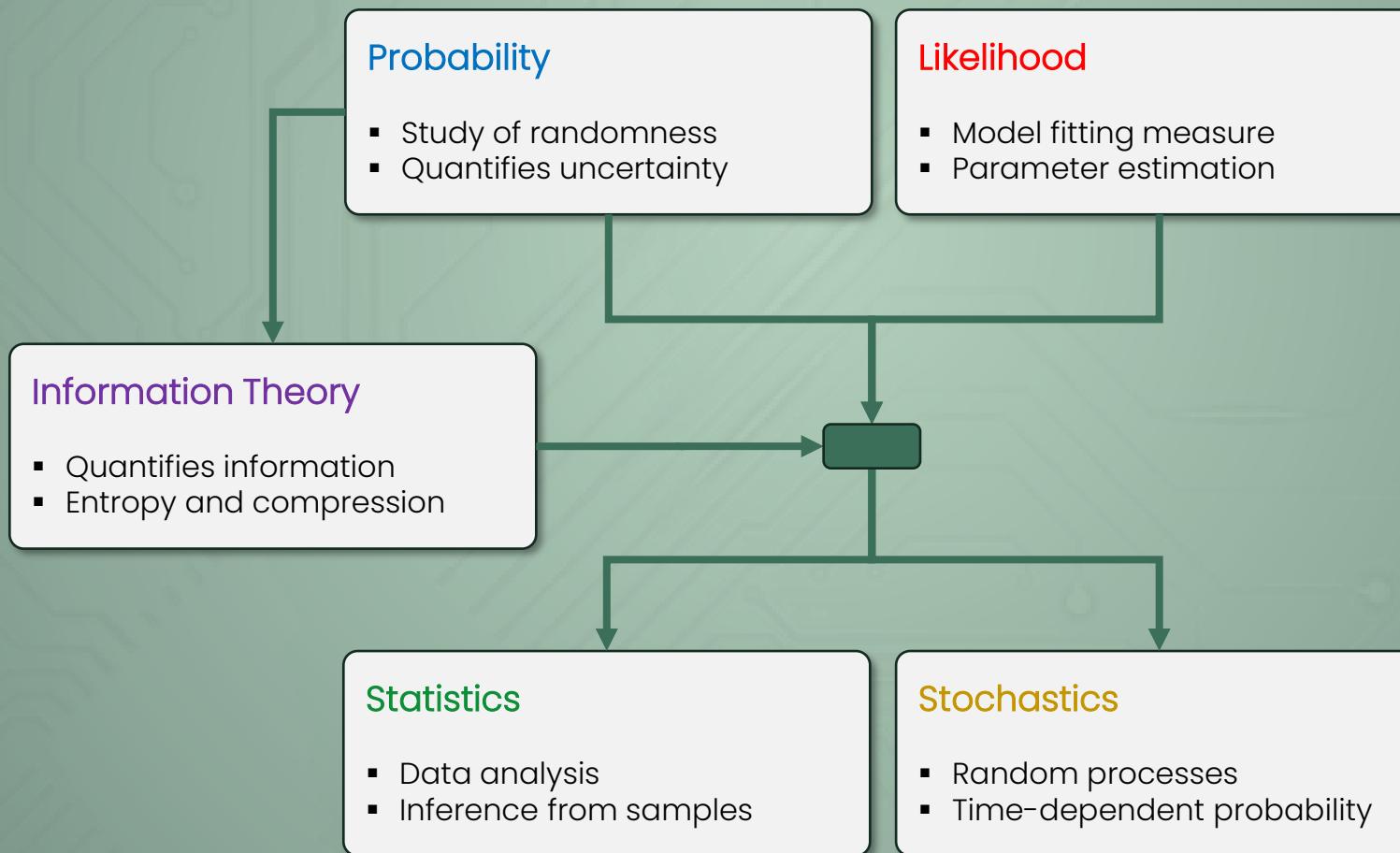
Black Swan: Symbol for rare event
Book: The Black Swan, Taleb, 2007



"I have seen horses vomiting in front of pharmacy"
Translation of German phrase

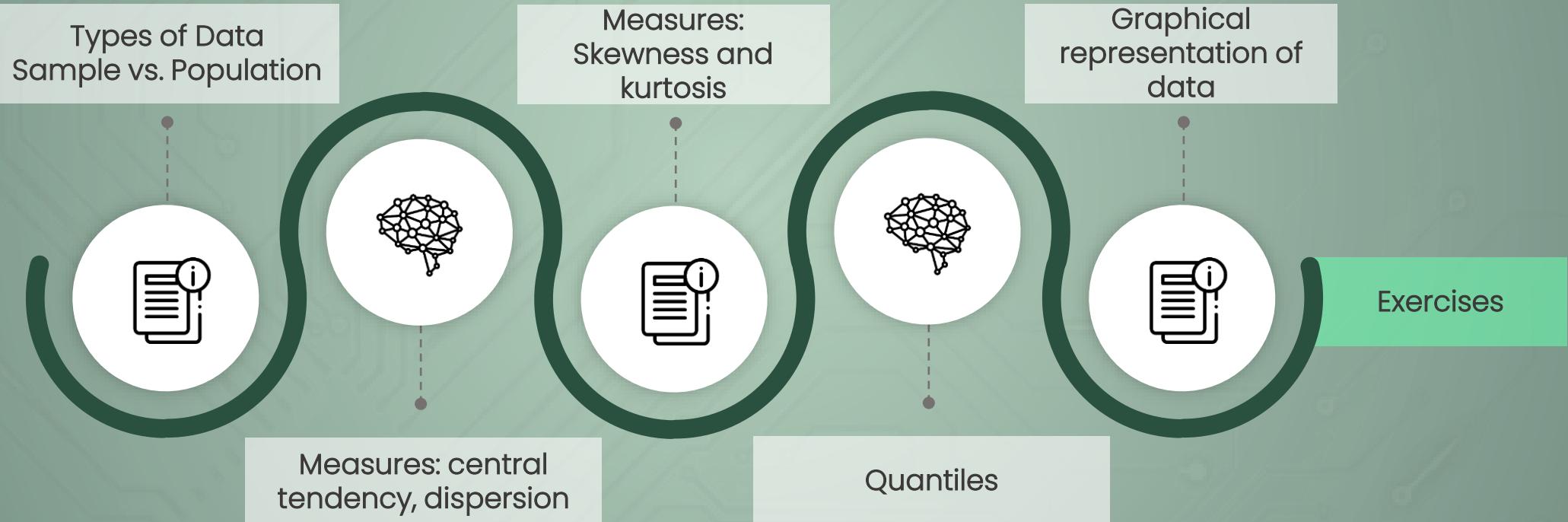
Probability, Likelihood, Information, Statistics, Stochastics

Probability quantifies uncertainty, Statistics analyzes data, Stochastics applies both to time-dependent processes, Likelihood assesses model fit, Information theory measures data content



Descriptive statistics

... involves the organization, summarization, and visualization of data. It provides simple summaries about the sample and the measures



Types of Data

Definition and examples of Qualitative and Quantitative data

Type of Data	Subtype of data	Example
Quantitative / Numerical: <ul style="list-style-type: none">• can be measured• or counted	Discrete: <ul style="list-style-type: none">• can only take certain values	<ul style="list-style-type: none">• number of students in a class
	Continuous: <ul style="list-style-type: none">• can take any value within certain range	<ul style="list-style-type: none">• Age, Height, Weight, Temperature, Income
	Interval	<ul style="list-style-type: none">• Temperature• IQ scores• Years
	Ratio	<ul style="list-style-type: none">• Probability
Qualitative / Categorical: <ul style="list-style-type: none">• non-numerical• often relates to subjective qualities, characteristics, or descriptions	Nominal: <ul style="list-style-type: none">• No order, hierarchy or sequence	<ul style="list-style-type: none">• Colors• Types of cuisine• Genders• Blood types
	Ordinal: <ul style="list-style-type: none">• can be arranged in a particular order	<ul style="list-style-type: none">• Survey responses• Educational level• Military rank

... other types of data: binary, time-series, geodata, textual, multimedia, etc.

Sample vs Population

A well-chosen sample should be representative of the population, allowing statisticians to draw accurate conclusions about the population based on sample data

Population

- Entire group of individuals or objects of interest
- Often too large to study completely
- Represented by N (size)

Sample

- Subset of the population
- Used to make inferences about the population
- Represented by n (size)

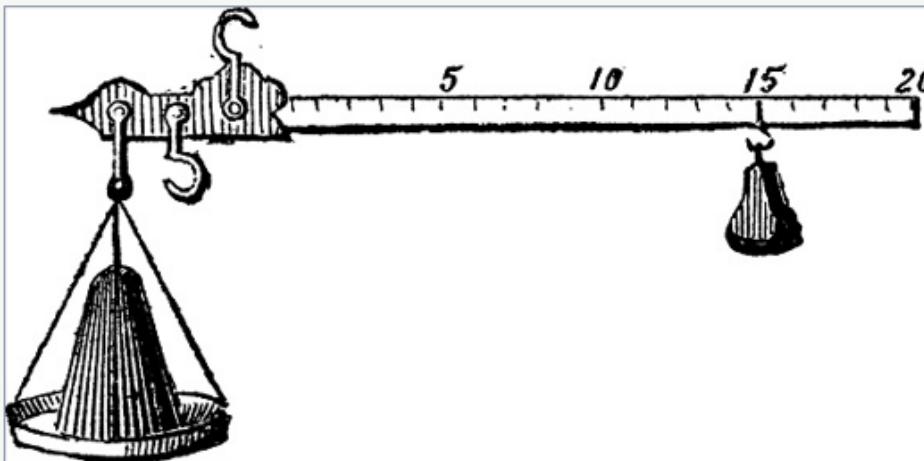


Measures of Central Tendency

... calculate typical central points that describe or represent dataset

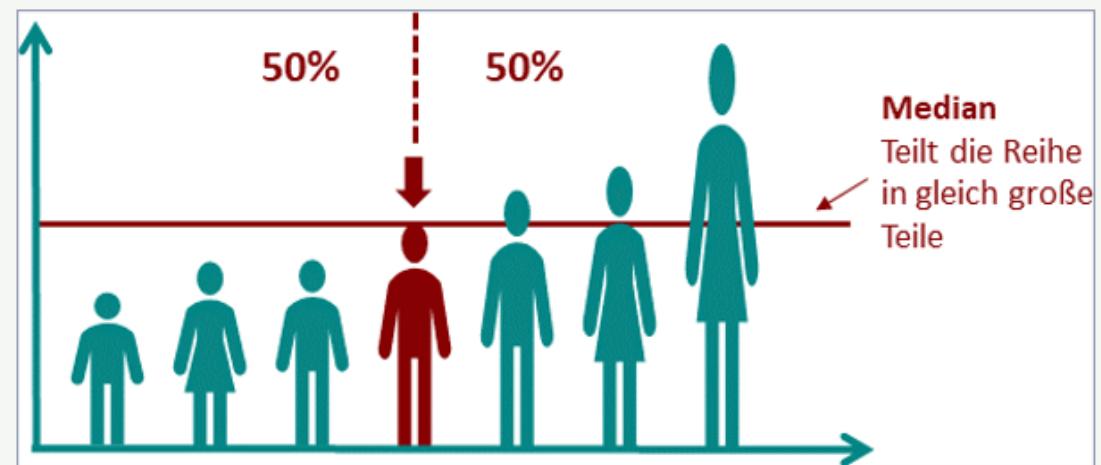
Arithmeric Mean = Average

- Sum of all data points divided by number of data points
- E.g.: $(1+2+3)/3 = 2$
- \bar{X} = Sample Mean; μ = Population Mean



Median

- middle value in an ordered dataset
- E.g.: For {1, 2, 3}, the median is 2
- **Interesting because of outlier resistance!**



... Other measures of centrality:

- Geometric mean: e.g. population growth, investment return
- Harmonic mean: e.g. F-score (Machine Learning, binary classification, evaluation of results)

Expectation or Expected Value

Random variable's expected value represents arithmetic mean / average of large number of independent realizations of the random variable

- X is a random variable
- with finite number of outcomes (x_1, x_2, \dots, x_k)
- occurring with probabilities (p_1, p_2, \dots, p_k).
- **For discrete variables:**
 - $E[X] = \sum_{i=1}^k x_i p_i = x_1 p_1 + x_2 p_2 + \dots + x_n p_n$
- **For continuous variables:**
 - we compute it with an integral.

Example: Rolling a Fair Six-Sided Die

Let X be the random variable representing the outcome of rolling a fair six-sided die.

Outcomes (x_i): 1, 2, 3, 4, 5, 6

Probabilities (p_i): 1/6 for each outcome (since it's a fair die)

To calculate the expected value $E[X]$, we'll use the formula:

$$E[X] = \sum_{i=1}^k x_i p_i = x_1 p_1 + x_2 p_2 + \dots + x_k p_k$$

$$\begin{aligned} E[X] &= 1 * (1/6) + 2 * (1/6) + 3 * (1/6) + 4 * (1/6) + 5 * (1/6) + 6 * (1/6) \\ &= (1 + 2 + 3 + 4 + 5 + 6) / 6 \\ &= 21 / 6 \\ &= 3.5 \end{aligned}$$

Therefore, the expected value of rolling a fair six-sided die is 3.5.

Measures of Dispersion

... provide information about how data values are distributed around central tendency

Measure of Dispersion	Definition	Example
Range	The difference between the highest and lowest values in a dataset	For $\{1, 2, 3, 4\}$, the range is $4 - 1 = 3$
Sum of Squares	$SS = \sum (x - \mu)^2$	For $\{1, 2, 3\}$, SS: $(1+0+1) = 2$
Variance (population)	$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$	For $\{1, 2, 3\}$, pop. variance: $(1+0+1)/3 = 0.67$
Variance (sample)	$s^2 = \frac{\sum(x-\mu)^2}{n-1}$	For $\{1, 2, 3\}$, sample variance: $(1+0+1)/2 = 1$
Standard Deviation	The square root of the variance	The square root of 0.67 is approximately 0.82

Variance and Standard Deviation

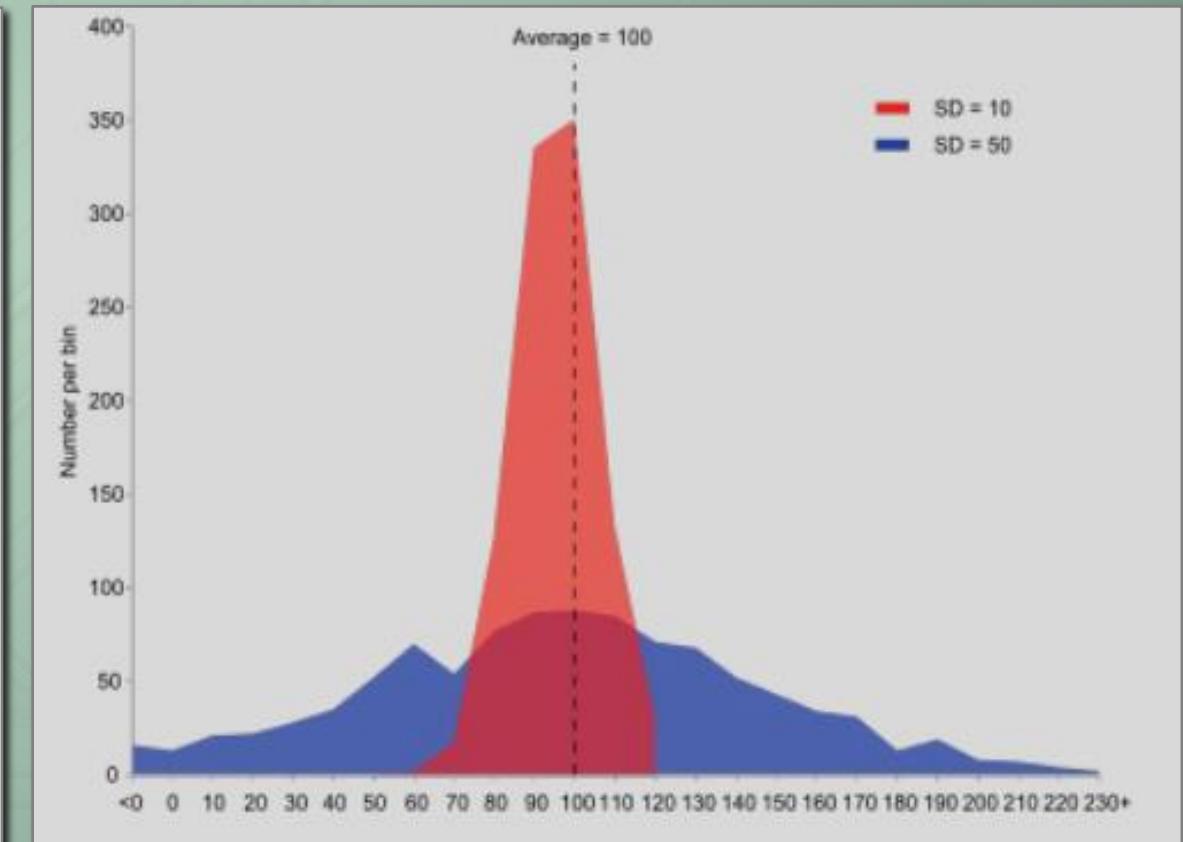
Variance $VAR(X) = E[(X - \mu)^2]$ of standard X (function) is expected value of squared deviation from mean X; $\mu = E[X]$

Example with Standard Deviation = 10

- Heights of certain species of plants
- normally distributed with mean height of 50 cm
- standard deviation of 10 cm.
- Heights are spread around mean
- Most plants' heights ranging 40 - 60 cm.

Example with Standard Deviation = 50:

- Annual income of group of people in certain profession
- normally distributed with mean income of \$200,000
- standard deviation of \$50,000.
- Incomes vary widely
- most individuals earn \$150,000 - \$250,000.



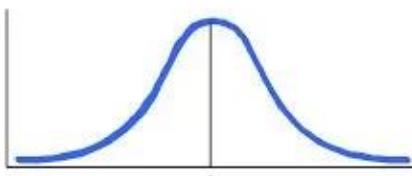
Skewness and Kurtosis

... describe asymmetry of distributions deviation from normal distribution

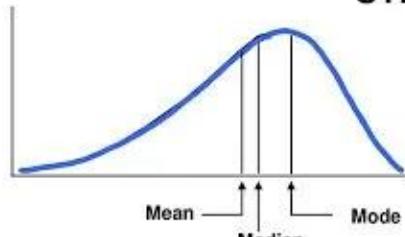
Skew

- Measure of asymmetry of the probability distribution of a real-valued random variable about its mean
- Positive (negative): Long tail on the right (left)

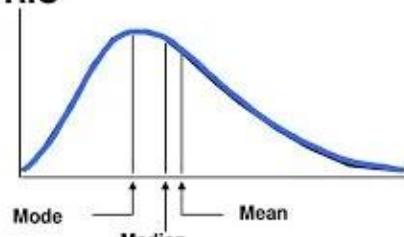
Describe the shape, center, and spread of a distribution... for shape, see below...



SYMMETRIC



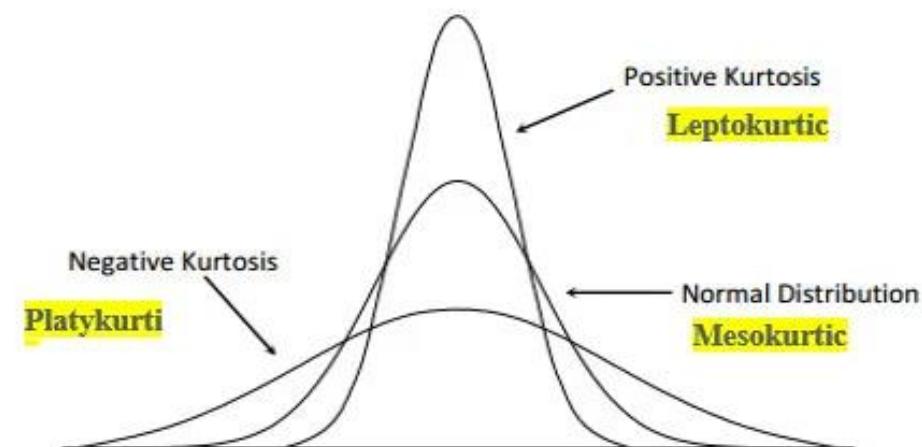
SKEWED LEFT
(negatively)



SKEWED RIGHT
(positively)

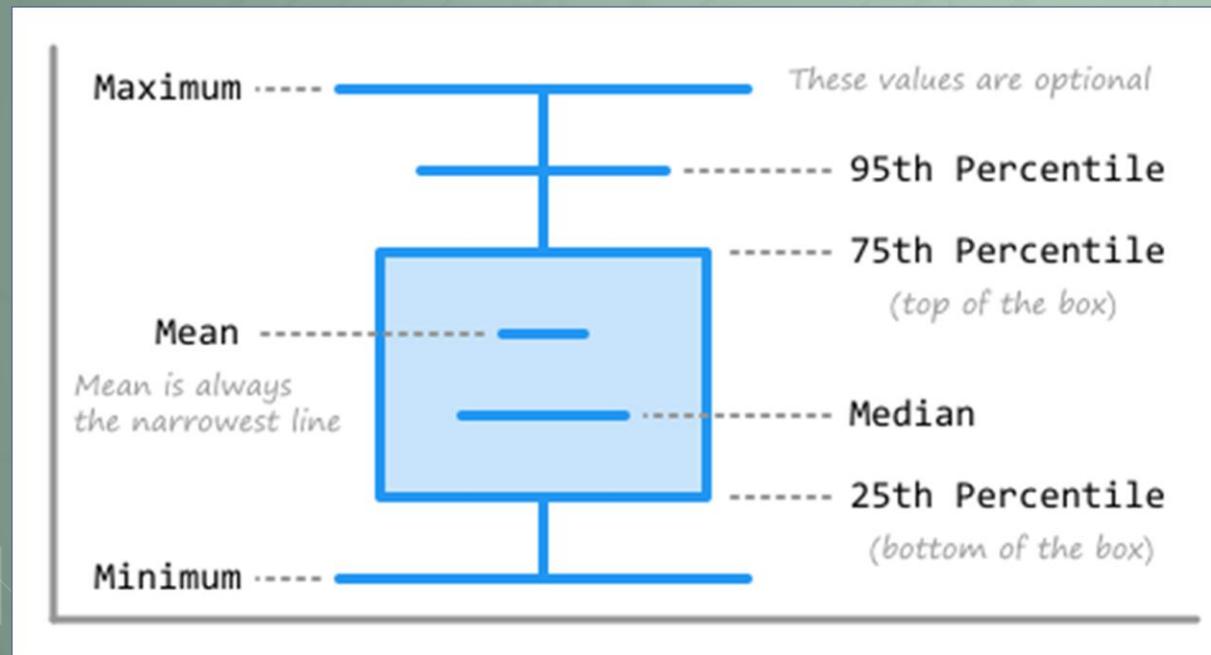
Kurtosis

- defines how heavily the tails of a distribution differ from the tails of a normal distribution
- Outliers may facilitate positive kurtosis



Quantiles

... statistical measures that divide a dataset into intervals of equal probability



Quantiles ...

- are used to understand data distribution and identify outliers
- partition data into several subsets containing equal number of observations.

Several types of quantiles:

- Quartiles: divide the data into four equal parts.
- Deciles: These divide the data into ten equal parts.
- Percentiles: These divide the data into 100 equal parts.
- Median: separates data into two halves, with 50% of the data falling below and 50% above the median.

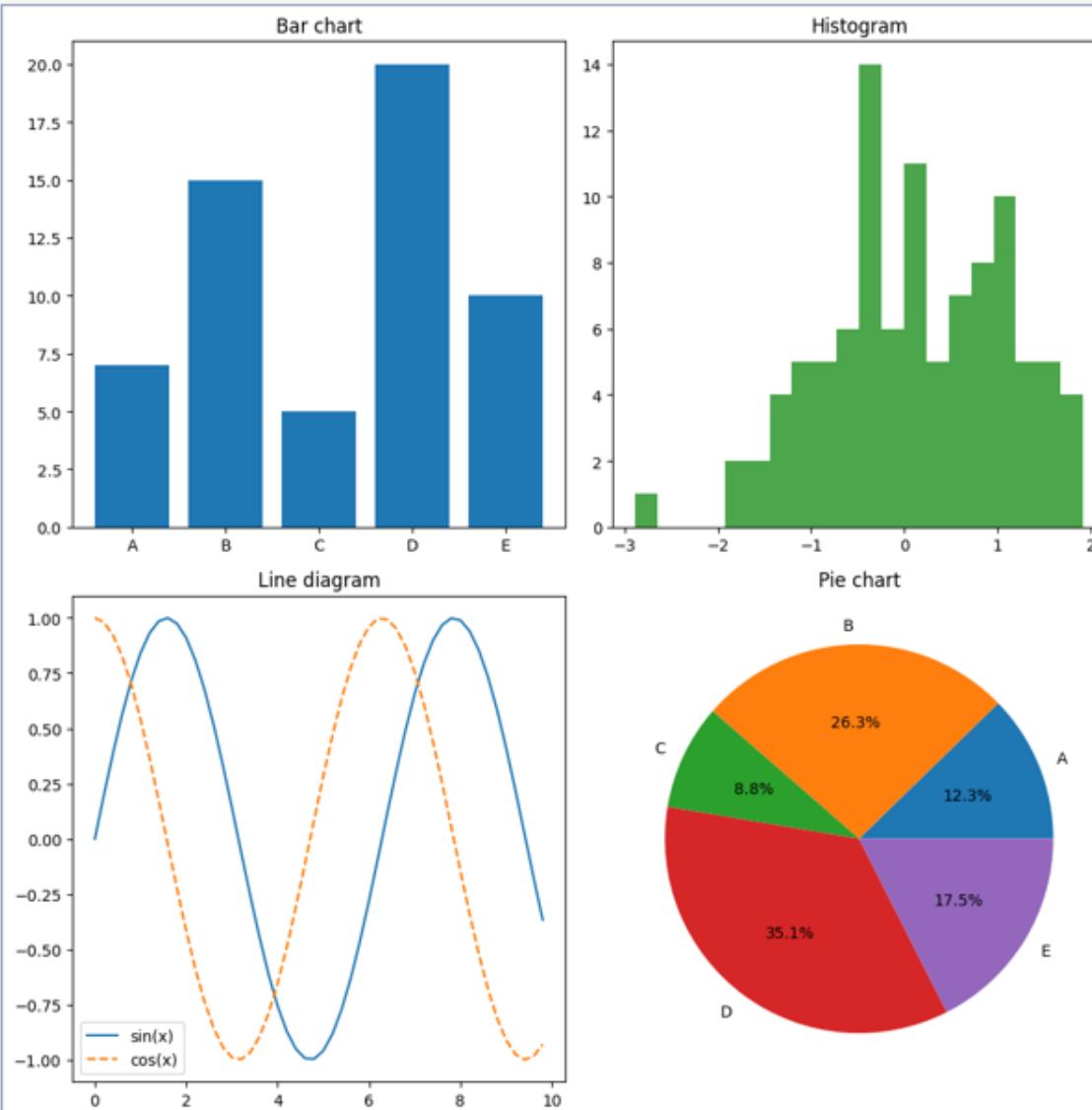
Graphical Representation of Data

Input:

- Categorical
- Nominal
- unordered

Output:

- Numerical
- Continuous
- Absolute



Input:

- Numerical
- continuous

Output:

- Numerical
- continuous

Input:

- Categorical
- Nominal
- ordered

Output:

- Numerical
- Continuous
- Absolute

Input:

- Categorical
- Nominal
- Unordered or ordered

Output:

- Numerical
- Continuous
- relative

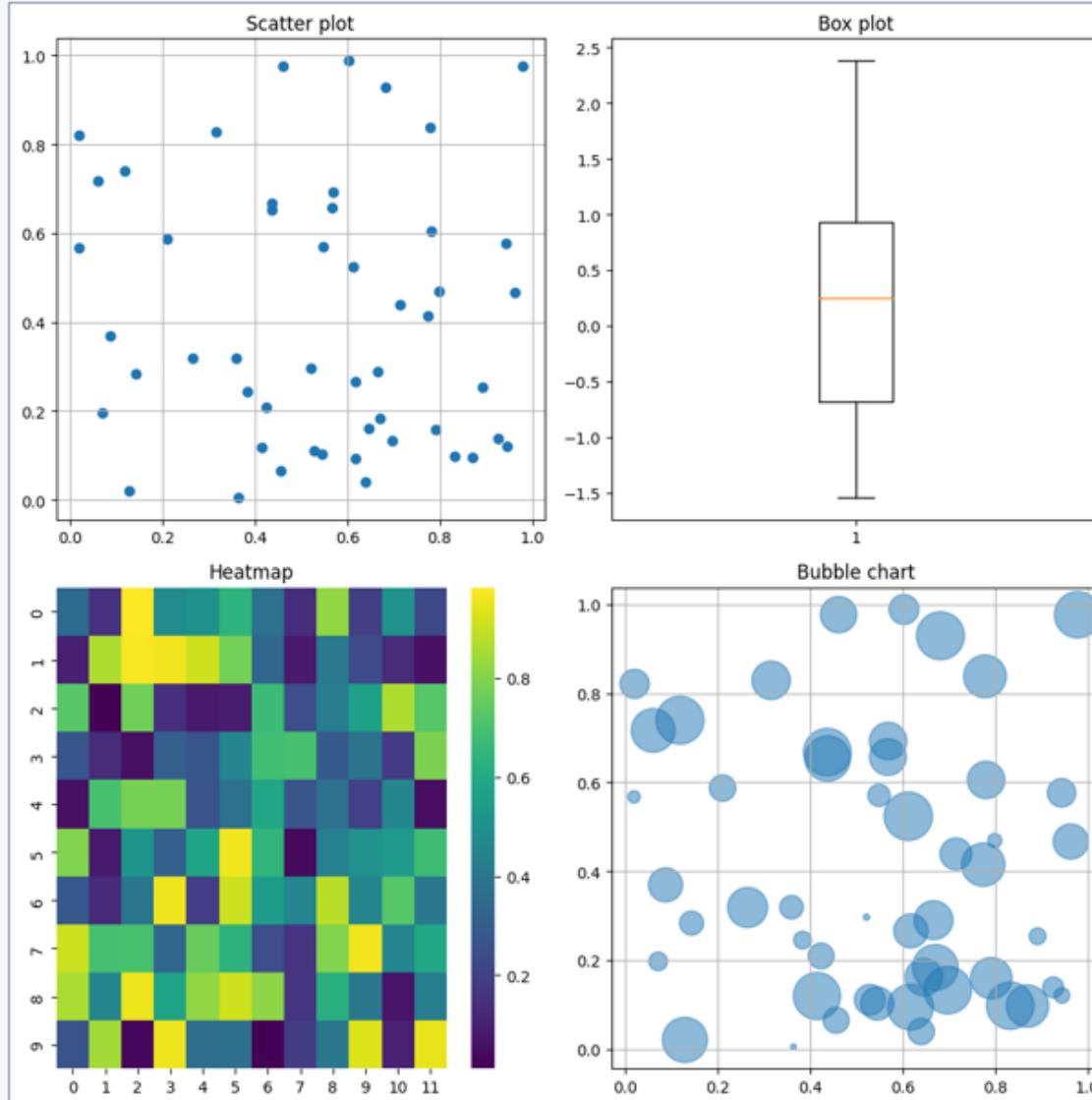
Graphical Representation of Data

Input:

- 2 Features
- Both numerical

Output:

- Boolean (1;0)



Input:

- 2 Features
- Both categorical

Output:

- Numerical

Input:

- 1 feature
- numerical

Output:

- At least ordered
- Or numerical

Input (like scatter plot):

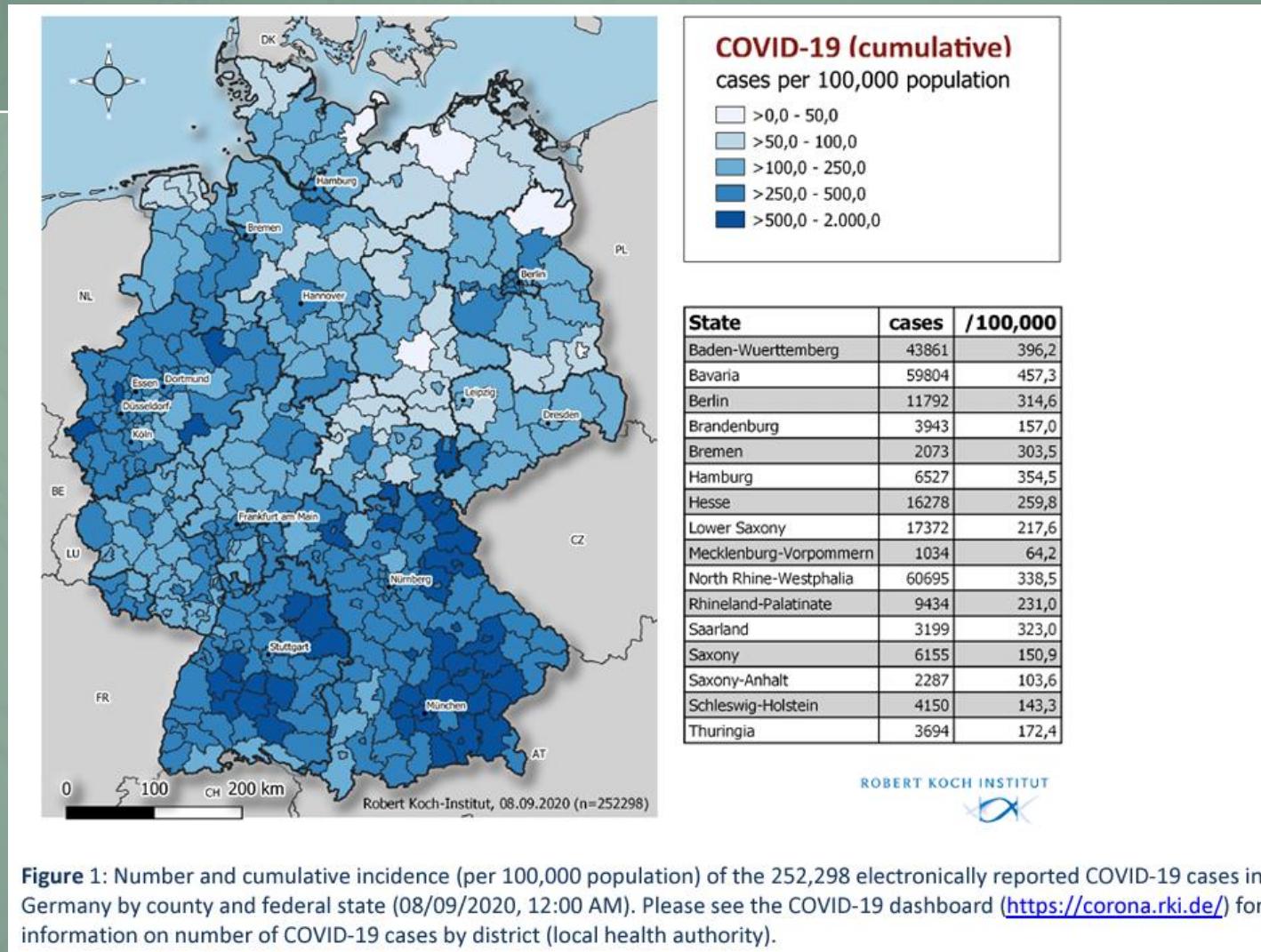
- 2 Features
- Both numerical

Output:

- Numerical

Graphical Representation of Data

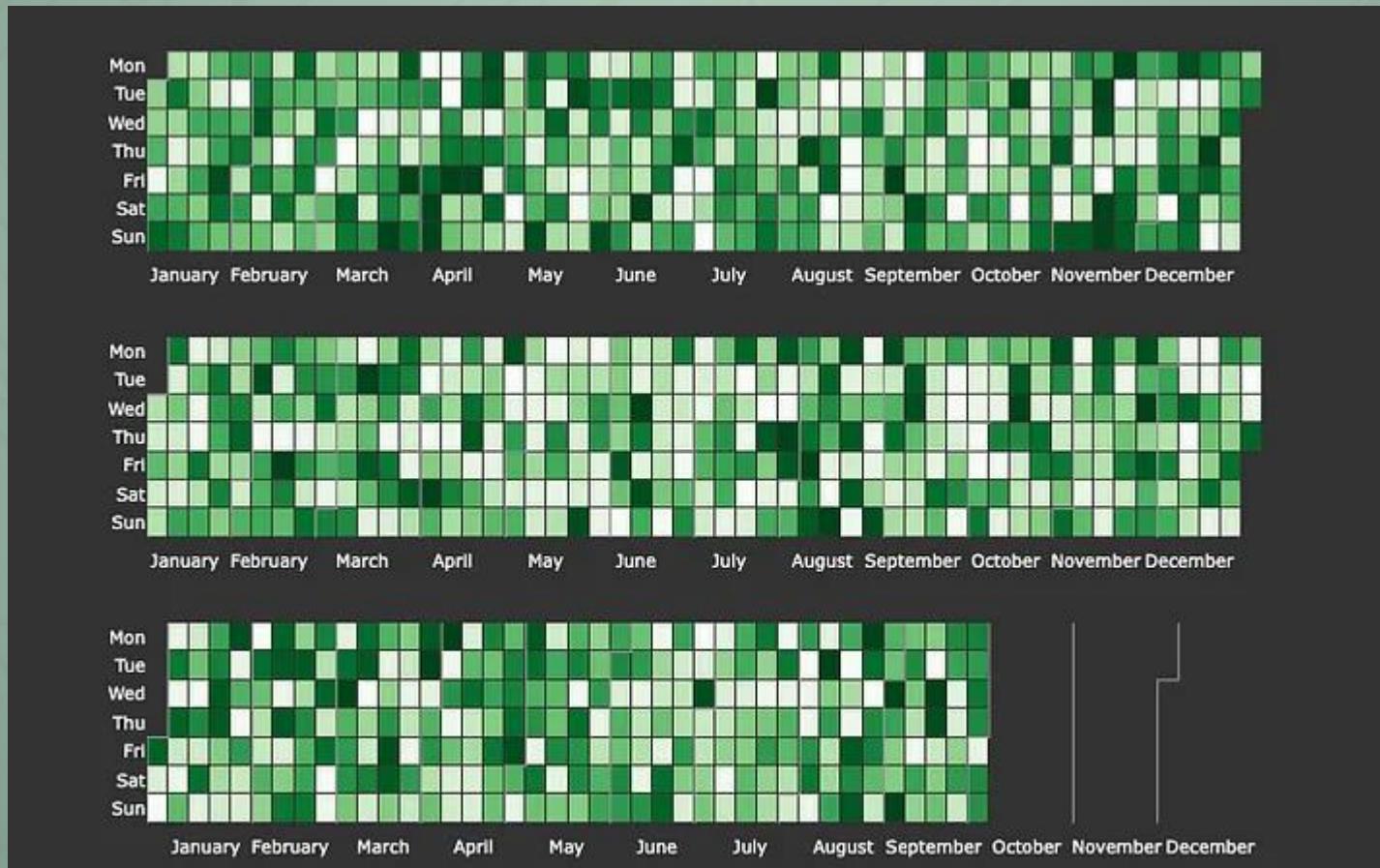
... Geovisualization with Heat Map. Example: COVID incidents Germany



Graphical Representation of Data

... Calendar with Heat Map

Example: Visualization of visitors' prediction over the year



Exercises

Exercise 1:

Question: Given the array numbers = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10]), compute the mean of the numbers.

Exercise 2:

Question: Given the array numbers from Exercise 1, compute the median of the numbers.

Exercise 3:

Question: Compute the variance of the numbers.

a) Use population variance b) Use sample variance

ages = [21, 23, 25, 27, 29, 31, 33]

b) Let's say we take a sample of 5 ages from the original group: sample_ages = [21, 23, 27, 31, 33]

Exercise 4:

Question: Given the array numbers from Exercise 1, compute the standard deviation of the numbers.

Exercise 5:

Question: Given the array numbers from Exercise 1, compute the first quartile (25th percentile), the second quartile (50th percentile, or median), and the third quartile (75th percentile) of the numbers.

Exercise 6:

show with python plots in a 2x2 subfigure matrix for ...

bar chart, histogram, line diagram, pie diagram

Scatter plot, Box plot (Box-and-whisker plot), Heatmap, bubble chart

Combinatorics

... is concerned with the study of countable, discrete structures and includes the arrangement, combination, and permutation of sets.

Permutation



Combination



... with replacement



k-permutation

Pascal's triangle

Exercises

Permutation

... is an arrangement of objects in a specific order.

Table 5-1

**Possible Rearrangements for
Four People Sitting in a Straight Line**

Rearrangement #	Listing	Rearrangement #	Listing
1	Tim, Syd, Elena, Mark	13	Elena, Tim, Syd, Mark
2	Tim, Syd, Mark, Elena	14	Elena, Tim, Mark, Syd
3	Tim, Elena, Syd, Mark	15	Elena, Syd, Tim, Mark
4	Tim, Elena, Mark, Syd	16	Elena, Syd, Mark, Tim
5	Tim, Mark, Syd, Elena	17	Elena, Mark, Tim, Syd
6	Tim, Mark, Elena, Syd	18	Elena, Mark, Syd, Tim
7	Syd, Tim, Elena, Mark	19	Mark, Tim, Syd, Elena
8	Syd, Tim, Mark, Elena	20	Mark, Tim, Elena, Syd
9	Syd, Elena, Tim, Mark	21	Mark, Syd, Elena, Tim
10	Syd, Elena, Mark, Tim	22	Mark, Syd, Tim, Elena
11	Syd, Mark, Tim, Elena	23	Mark, Elena, Tim, Syd
12	Syd, Mark, Elena, Tim	24	Mark, Elena, Syd, Tim

The number of permutations

- without repetition
- of a set of n objects is given by:

$$P(n) = n!$$

For $n = 4$ there are

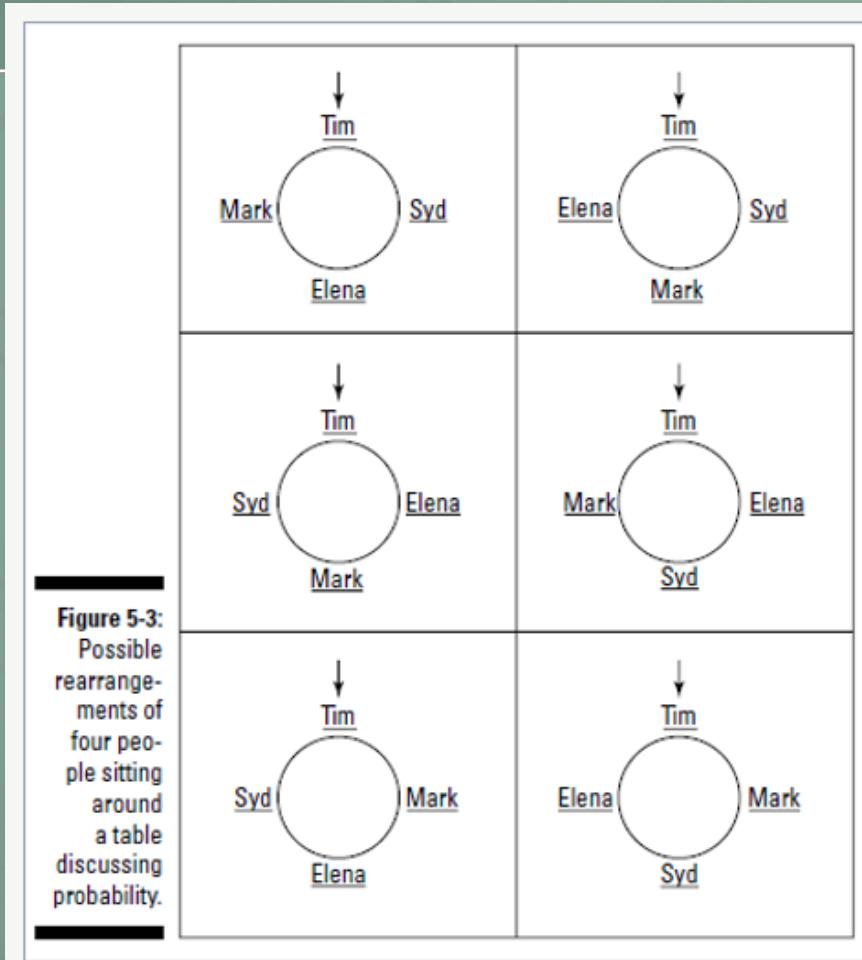
$$P(4) = 4! = 4 * 3 * 2 * 1 = 24$$

possibilities.

Here: No replacement

Permutation

... is an arrangement of objects in a specific order.



Example

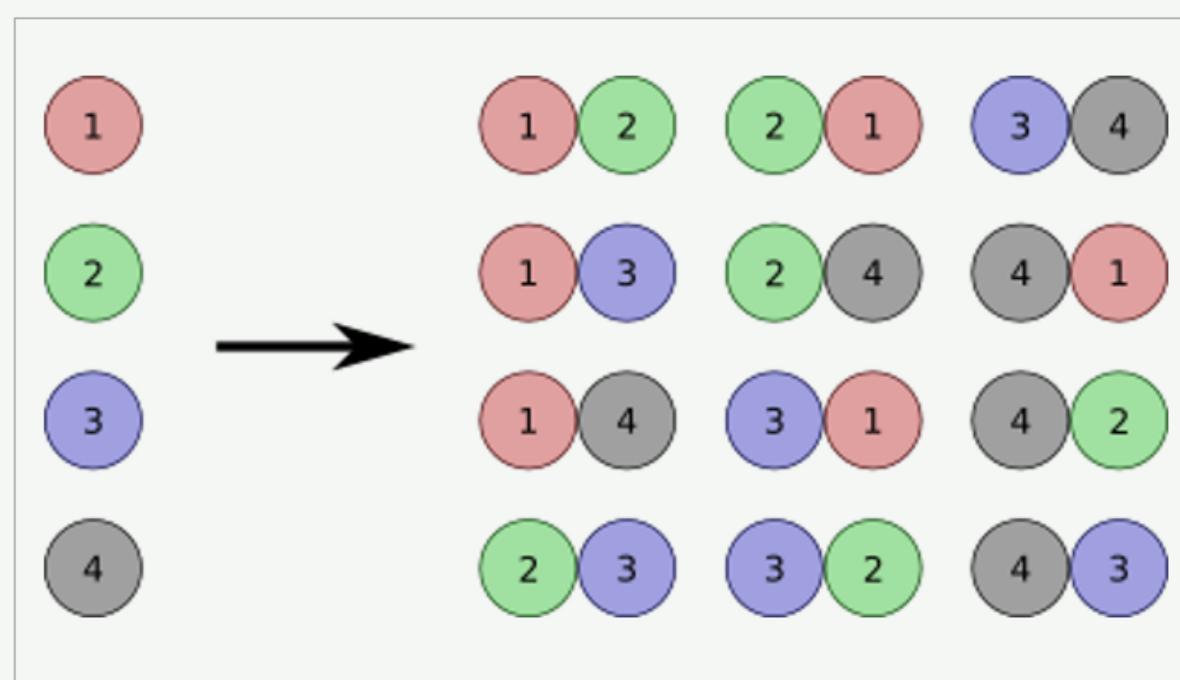
- 4 people sit at a table
- Restriction: The position of one of them is fixed.
- How many possible arrangements are there?

$$P(3) = 3! = 3 * 2 * 1 = 6$$

Here: No replacement

K-permutation

... is an ordered arrangement of k elements selected from that set. In German it is called "Variation"



[File:Kpermutation.png - Wikimedia Commons](#)

Here: No replacement

The number of k-permutations

- without repetition
- of a set of n objects and k selections is given by

$$P(n, k) = \frac{n!}{(n-k)!}$$

For $n = 4$ and $k = 2$ there are

$$P(4, 2) = \frac{4 \cdot 3 \cdot 2}{2} = \frac{24}{3 \cdot 2} = 12$$

possibilities.

Other k-permutations:

$$P(4, 1) = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = \frac{24}{3 \cdot 2} = 4$$

$$P(4, 3) = \frac{4 \cdot 3 \cdot 2}{1} = \frac{24}{1} = 24$$

$$P(4, 4) = \frac{24}{0!} = 24$$

Combination

... is a selection of items from a larger set where the order of selection does not matter.



Here: No replacement

The number of combinations

- without repetition
- of a set of n objects and k selections is given by

$$C(n, k) = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

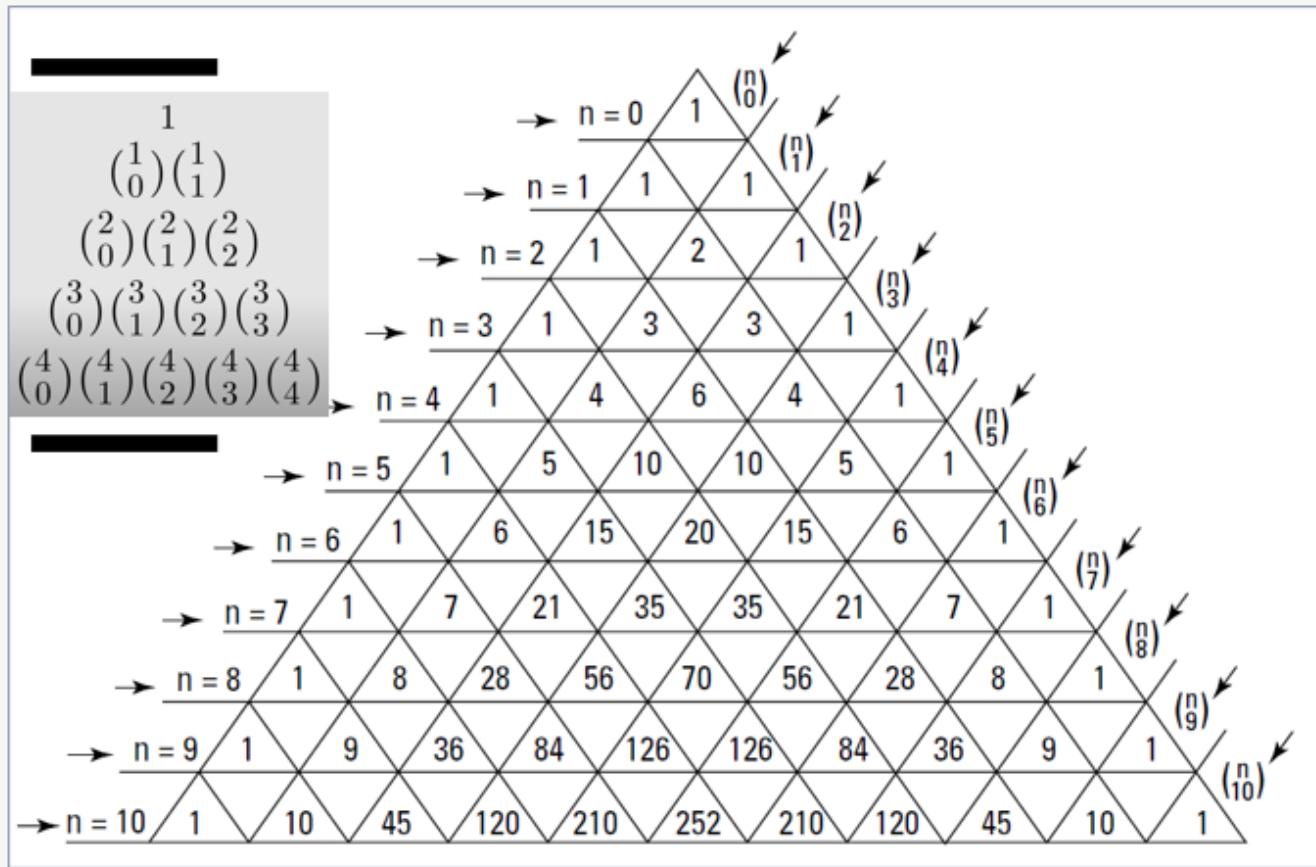
In the example of German lottery there are

- 49 balls
 - 6 selected balls
- i.e. there are

$$C(49, 6) = \binom{49}{6} = 13.983.816 \text{ possibilities.}$$

Pascal's triangle

... is a triangular array of numbers in which each number is the sum of the two directly above it. It is used for binomial expansions, combinatorics, probability calculations, etc.



Binomial theorem

- Coefficients from binomial expansion:

$$(a + b)^0 = 1$$

$$(a + b)^1 = 1a + 1b$$

$$(a + b)^2 = 1a^2 + 2ab + 1b^2$$

$$(a + b)^3 = 1a^3 + 3a^2b + 3ab^2 + 1b^3$$

- Sum of rows $\rightarrow 2^n$

Coming in next chapter:

- $\binom{n}{k}$: crucial for explaining distributions like binomial distribution

Exercises

Exercise 1: Count the permutations

Write a function that returns the number of all possible permutations of a list.

```
def count_permutations(lst): # Your code here  
    print(count_permutations(['A', 'B', 'C']))
```

Additionally: Try also lists with other lengths

Exercise 2: Generate all permutations of a list: `l1 = ['A', 'B', 'C']`

Exercise 3: Count the k-permutations

Write a function that returns the number of all possible k-permutations of a list.

Exercise 4: Generate k-permutations of a list

In this exercise, your task is to write a function that generates all possible k-permutations of a given list using `itertools`.

Exercise 5: Generate all combinations of size k

Write a function that generates all possible combinations of size k from a given list of size n=8.

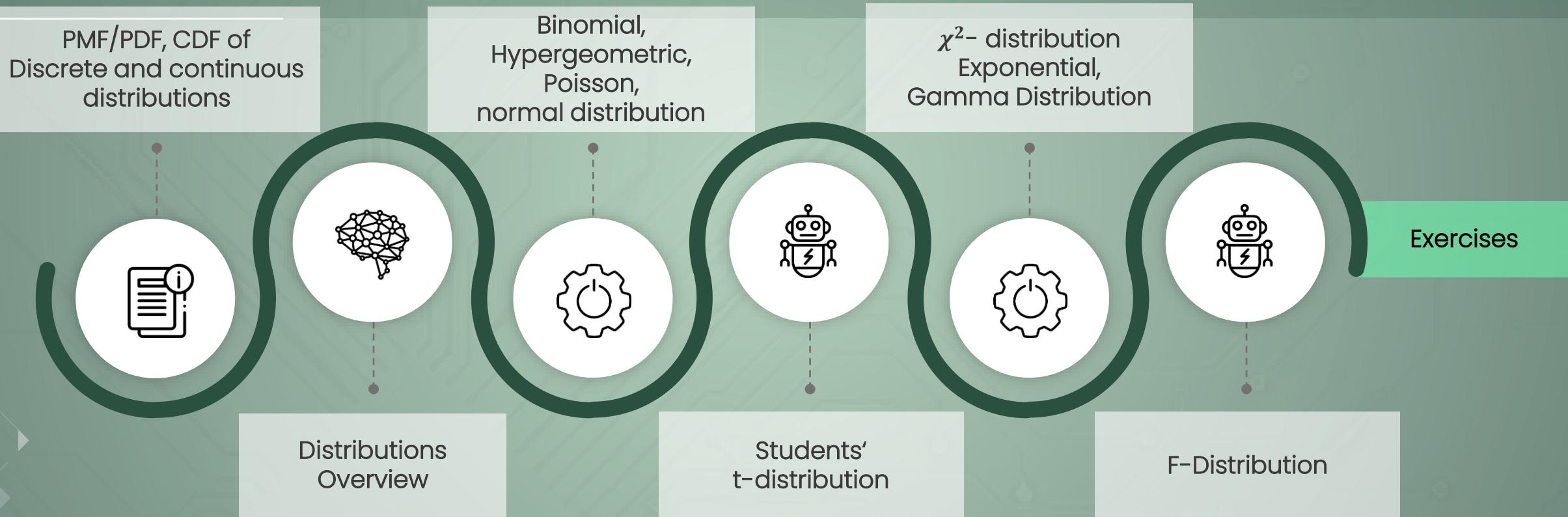
```
['apple', 'banana', 'pear', 'grapes', 'orange', 'mango', 'blueberry', 'strawberry']
```

Exercise 6: Calculate the number of combinations

$n = 8$ $k = 2$

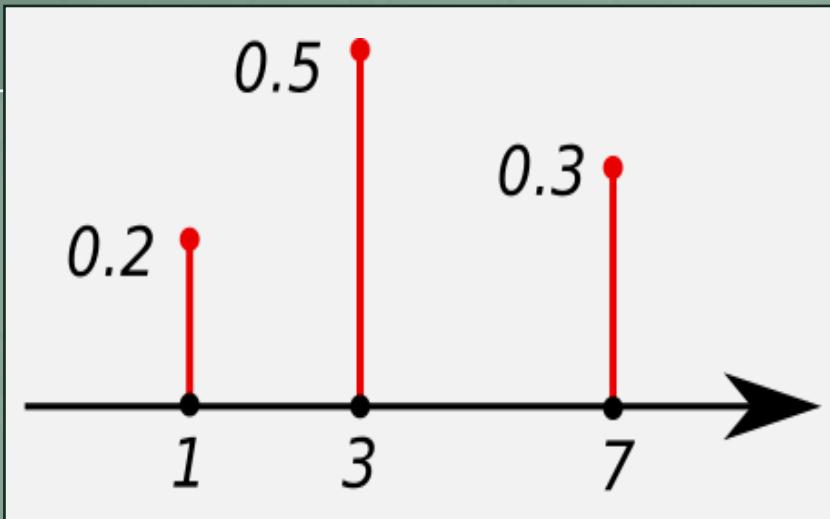
Probability Distributions

They describe the probabilities of occurrences of different possible outcomes in an experiment



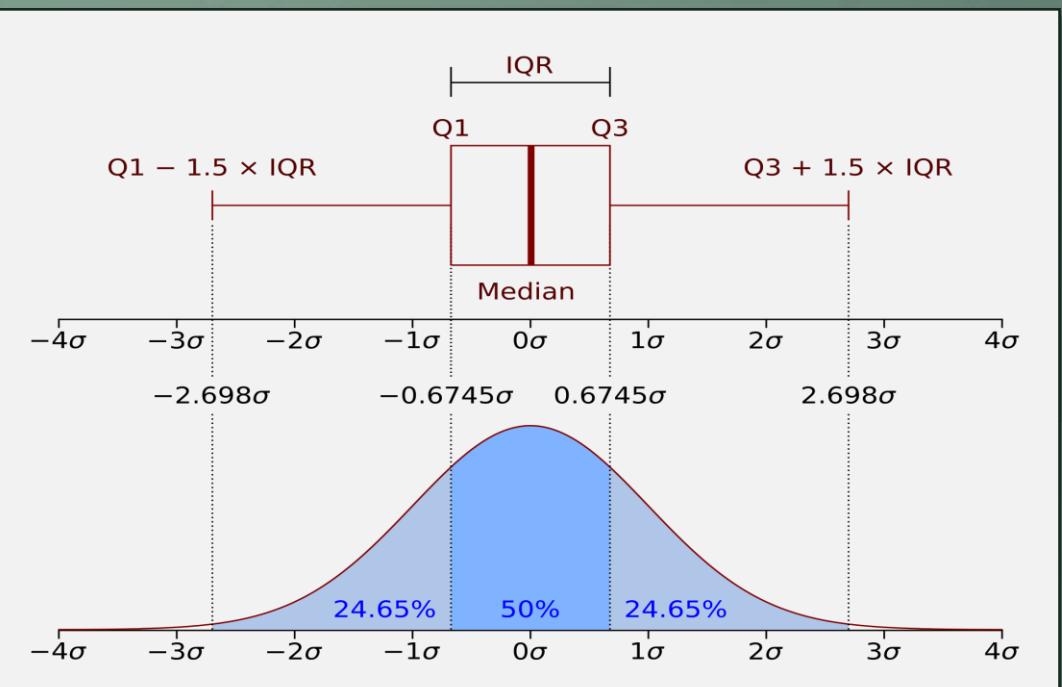
Probability Distributions

... are discrete or continuous



Discrete Probability Distribution

- Represented by vertical lines (lollipop chart)
- Each point represents a specific outcome
- Height of line indicates probability
- Example shown: {1: 0.2, 3: 0.5, 7: 0.3}
- Sum of all probabilities equals 1



Continuous Probability Distribution

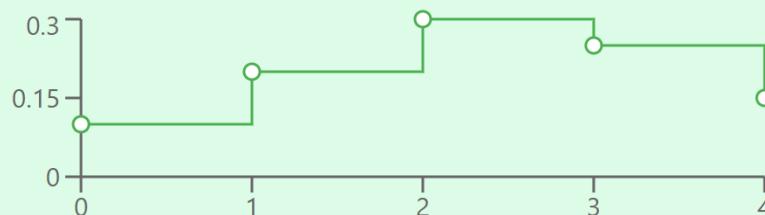
- Represented by a smooth curve (bell curve shown)
- Area under the curve represents probability
- Total area under curve equals 1
- Example: Normal (Gaussian) distribution
- Shows
 - median, quartiles, and standard deviations
 - Interquartile Range (IQR): Distance between Q1 and Q3
 - Standard Deviation: Measure of spread 1σ , 2σ , 3σ

Probability Distributions

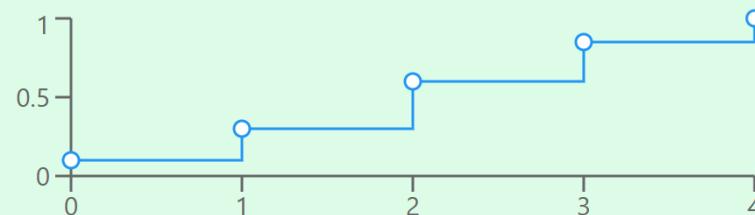
Comparing Probability Mass/Density and Cumulative Distribution Functions

Discrete Distributions

Probability Mass Function (PMF)

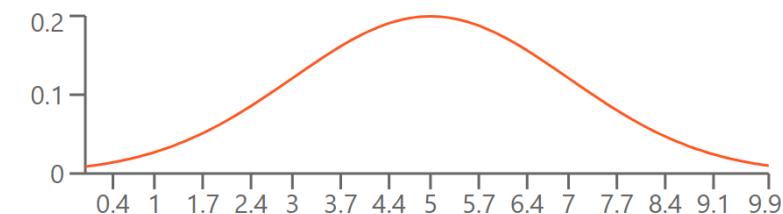


Cumulative Distribution Function (CDF)

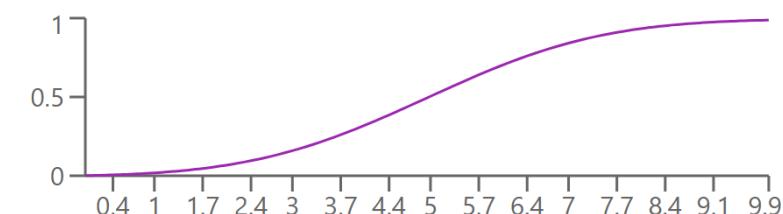


Continuous Distributions

Probability Density Function (PDF)



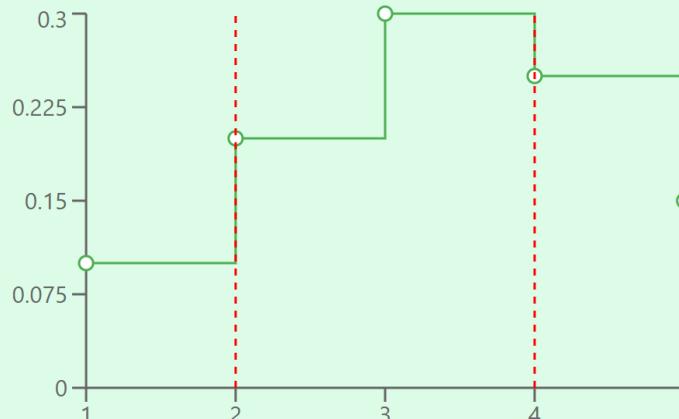
Cumulative Distribution Function (CDF)



Measuring Probability on Distribution Intervals

From Sums to Integrals: Calculating Probabilities Across Distribution Types

Discrete Distributions



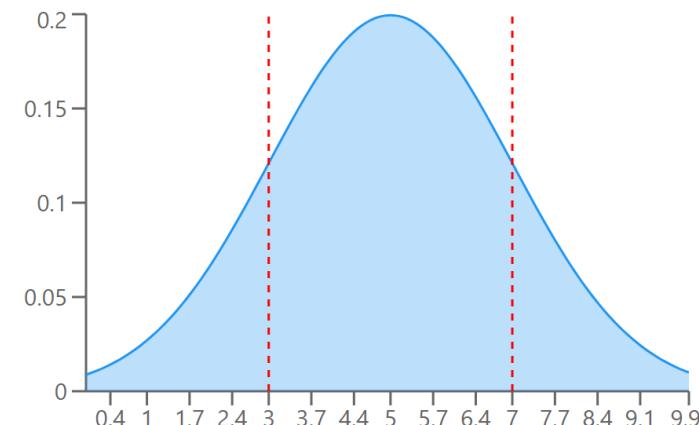
Probability = Sum of probabilities for each point in the interval

$$\text{Example: } P(2 \leq X \leq 4) = 0.2 + 0.3 + 0.25 = 0.75$$

Given the shown distribution ...

What is the probability that the guest has drunk 2,3 or 4 beer?

Continuous Distributions



Probability = Area under the curve between interval points

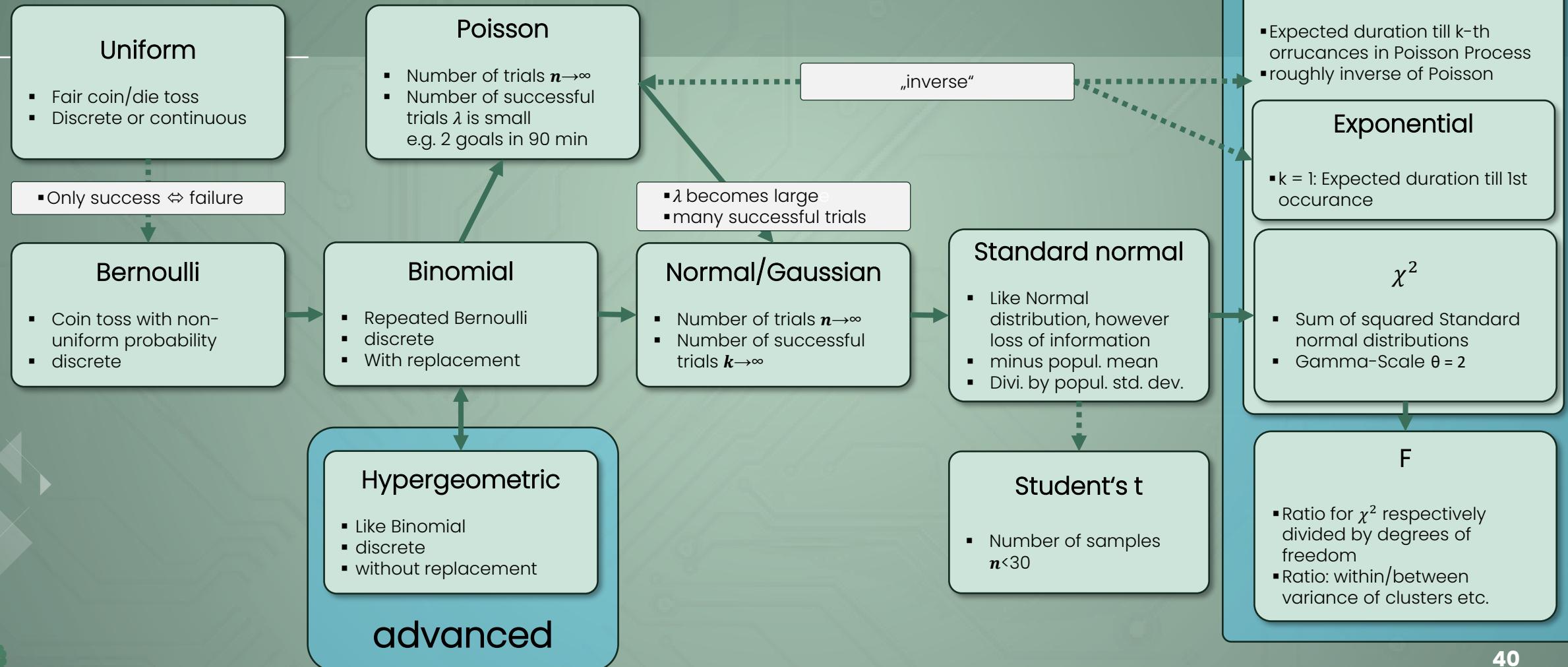
$$\text{Example: } P(3 \leq X \leq 7) = \int[3 \text{ to } 7] f(x) dx$$

Given the shown distribution ...

What is the probability that the person's income is between 3k and 7k € a month?

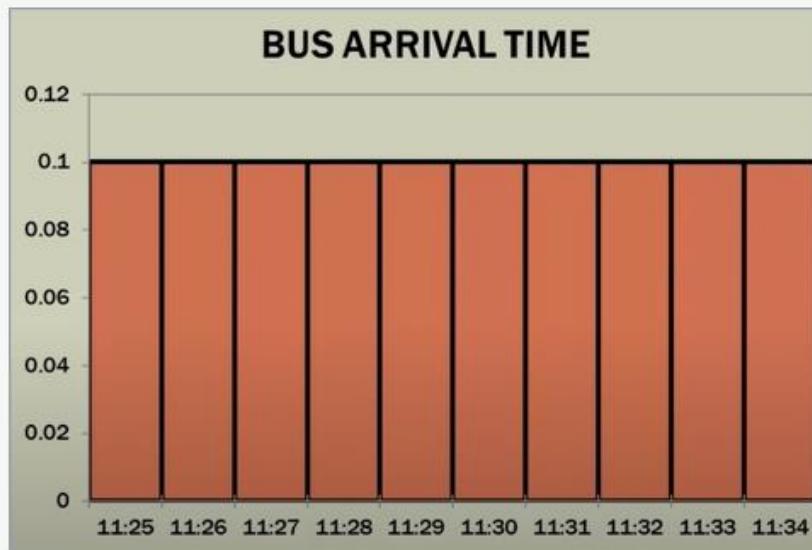
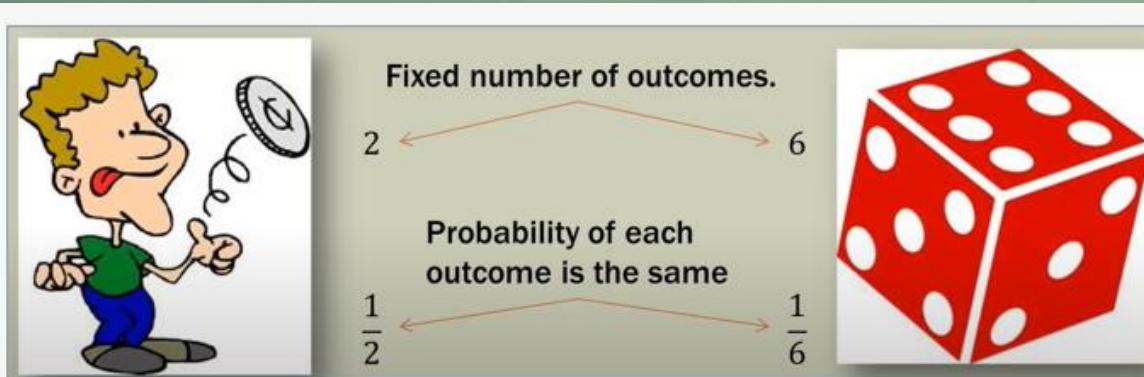
Overview of important distributions

... Most common distributions are based on the binomial distribution.



Uniform distribution

... is a probability distribution that describes the number of successes in a fixed number of independent binary experiments, each with the same probability of success.



Expected value:

- Arithmetic mean

Use Cases:

- Generating Random Numbers:
e.g.: fair dice, fair coin tosses
- Quality Control and Manufacturing:
e.g.: check the uniformity of products

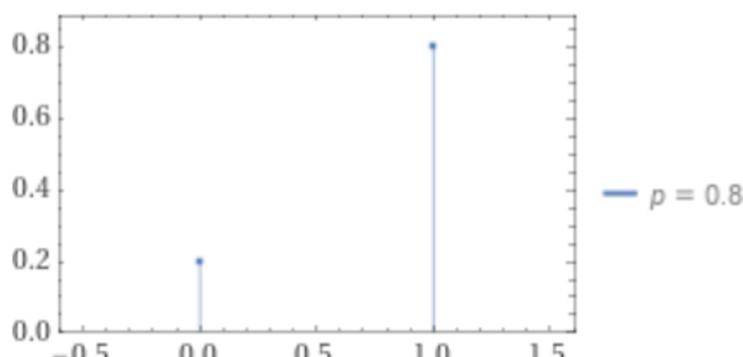
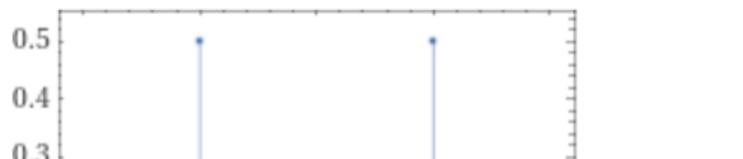
Be careful:

- Sum of 2 dice is not uniformly distributed!
(it is multinomial distribution, out of scope)

Bernoulli distribution

... is a discrete probability distribution for a random variable which can take a binary outcome: one (success) or zero (failure); example: toss of coin (that is not necessarily fair)

Plots of PDF for typical parameters



Formula:

$$P(x; p) = p^x * (1 - p)^{1-x} \text{ for } x \in [0; 1]$$

p: probability of success. E.g. if the coin is fair, the probability of getting a head (considered a success) would be 0.5

x: outcome of Bernoulli trial; it can take only two values:
1(success) or 0(failure)

Attention: $x \in [0; 1]$ is not an interval!

$P(x; p)$: probability of outcome x , given probability p of success.

Binomial distribution

Example for probability calculation:

Basketball player that is known for 60% prob of making free shots makes 7 of 10

Probability a 60% free throw shooter makes

1 of 1?

60%



Step 1:

- $P(\text{Make the shot}) = 60\%$
- $P(\text{Miss the shot}) = 40\%$
 $= 1 - 60\%$

Binomial distribution

Example for probability calculation:

Basketball player that is known for 60% prob of making free shots makes 7 of 10

Probability a 60% free throw shooter makes...

0 of 2?

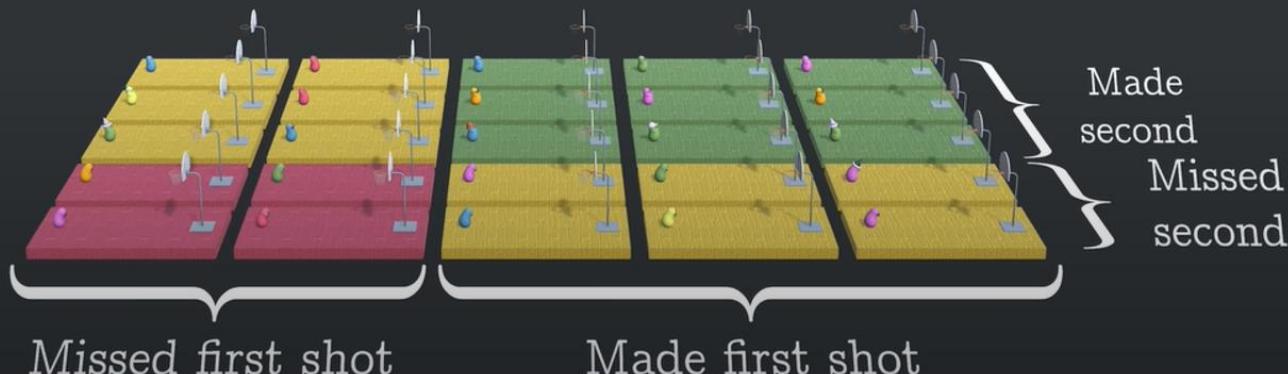
$$P(\text{miss, miss}) = \frac{4}{25} = 16\% \\ = 0.4 \times 0.4$$

1 of 2?

$$P(\text{make, miss}) = \frac{6}{25} = 24\% \\ = 0.6 \times 0.4 \\ P(\text{miss, make}) = \frac{6}{25} = 24\% \\ = 0.4 \times 0.6$$

2 of 2?

$$P(\text{make, make}) = \frac{9}{25} = 36\% \\ = 0.6 \times 0.6$$



Step 2:

- $P(\text{miss, miss}) = 16\%$
- $P(\text{miss, make}) = 48\%$
- $P(\text{make, make}) = 36\%$

Assumption:
Independence of shots

Binomial distribution

Example for probability calculation:

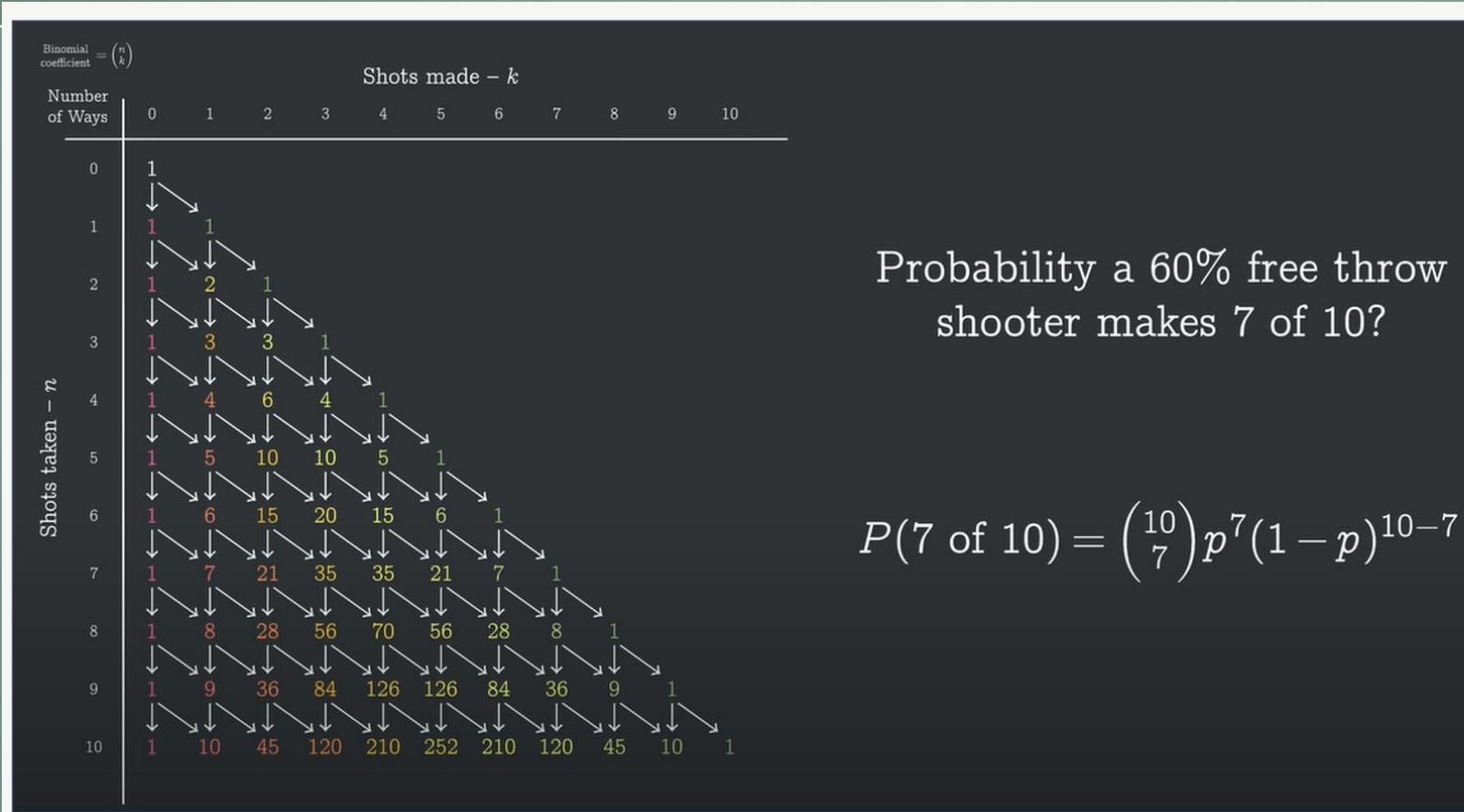
Basketball player that is known for 60% prob of making free shots makes 7 of 10



Binomial distribution

Example for probability calculation:

Basketball player that is known for 60% prob of making free shots makes 7 of 10



$$P(7 \text{ of } 10) = 120 \times 0.6^7 \times 0.4^3 \\ \approx 21.5\%$$

Probability a 60% free throw shooter makes 7 of 10?

$$P(7 \text{ of } 10) = \binom{10}{7} p^7 (1-p)^{10-7}$$

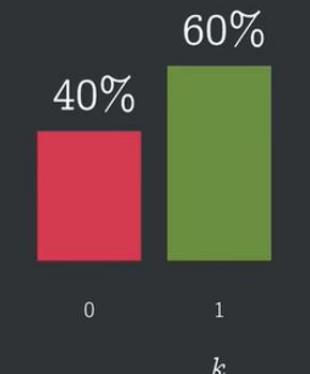
Binomial distribution

Example for probability calculation – Generalization:

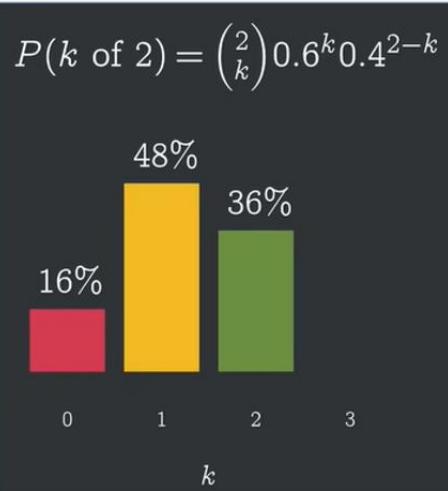
Basketball player that is known for 60% prob of making free shots makes k of n

$$P(k \text{ of } n) = \binom{n}{k} 0.6^k 0.4^{n-k}$$

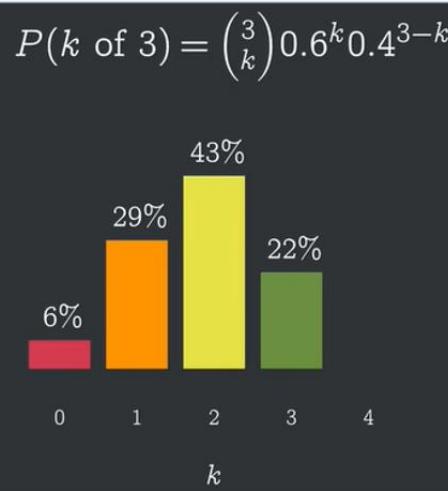
$$P(k \text{ of } 1) = \binom{1}{k} 0.6^k 0.4^{1-k}$$



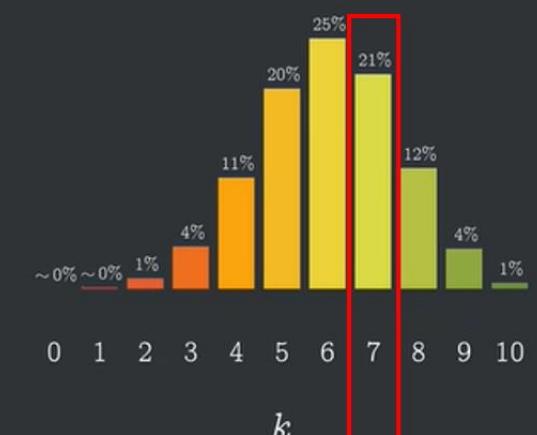
$$P(k \text{ of } 2) = \binom{2}{k} 0.6^k 0.4^{2-k}$$



$$P(k \text{ of } 3) = \binom{3}{k} 0.6^k 0.4^{3-k}$$



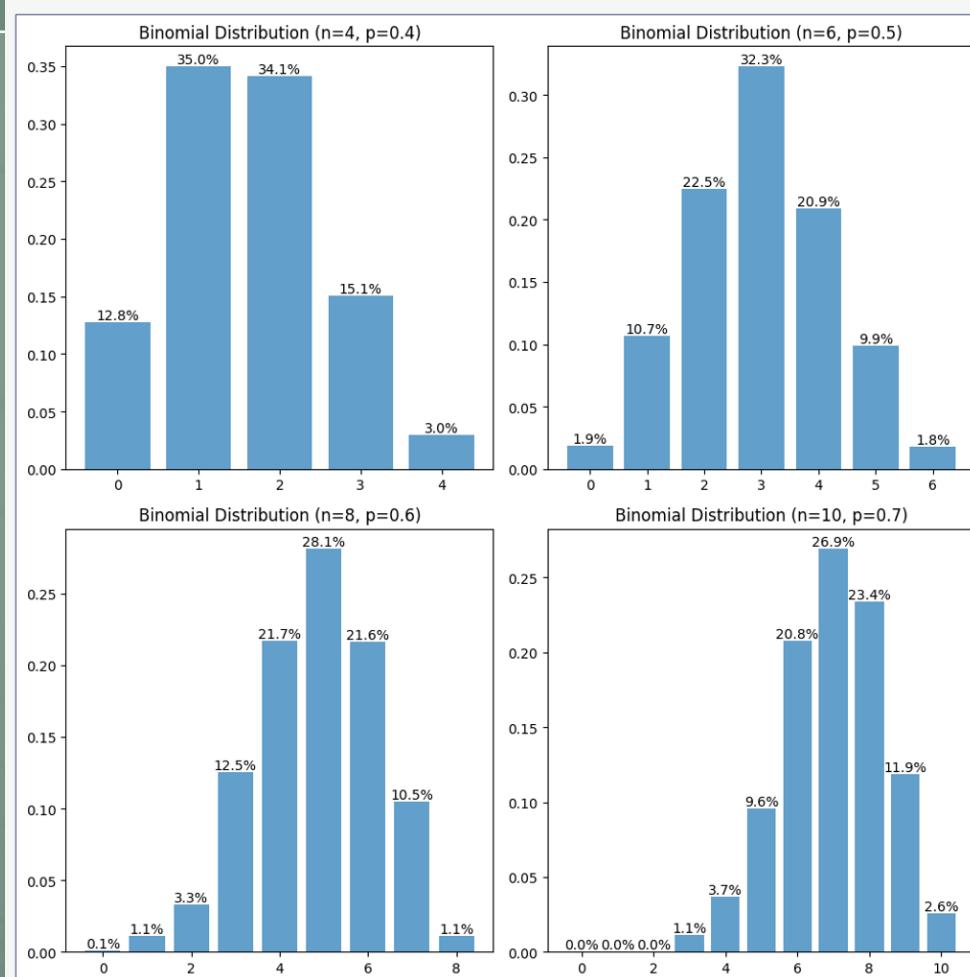
$$P(k \text{ of } 10) = \binom{10}{k} 0.6^k 0.4^{10-k}$$



This example

Binomial distribution

... is repeated Bernoulli distribution, i.e. probability distribution that describes number of successes in fixed number of independent binary experiments, each with same probability of success.



Formula:

$$P(k \text{ of } n) = \binom{n}{k} p^k (1-p)^{n-k}$$

Parameters:

- n: number of trials
- p: probability of success on each trial
- k (input factor on x-axis): number of actual positive outcomes

Use cases:

- Quality Control: In a manufacturing company the probability of producing a defective item is 0.05, and it manufactures 200 items per day
- Spam email classifier

Connection between Joint and Binomial Probability

Binomial probability is sum of joint probabilities for all possible ways to achieve k successes in n trials.

Joint Probability

For n independent trials:

$$P(X_1=x_1, X_2=x_2, \dots, X_n=x_n) = p^k * (1-p)^{n-k}$$

Where:

- p = probability of success on each trial
- k = number of successes
- n = total number of trials

Binomial Probability

Probability of exactly k successes in n trials:

$$P(X = k) = C(n,k) * p^k * (1-p)^{n-k}$$

Where:

- $C(n,k)$ = number of ways to choose k items from n
- p, k, and n are as defined for joint probability

Example: Coin Flips

For 3 coin flips (H = heads, T = tails):

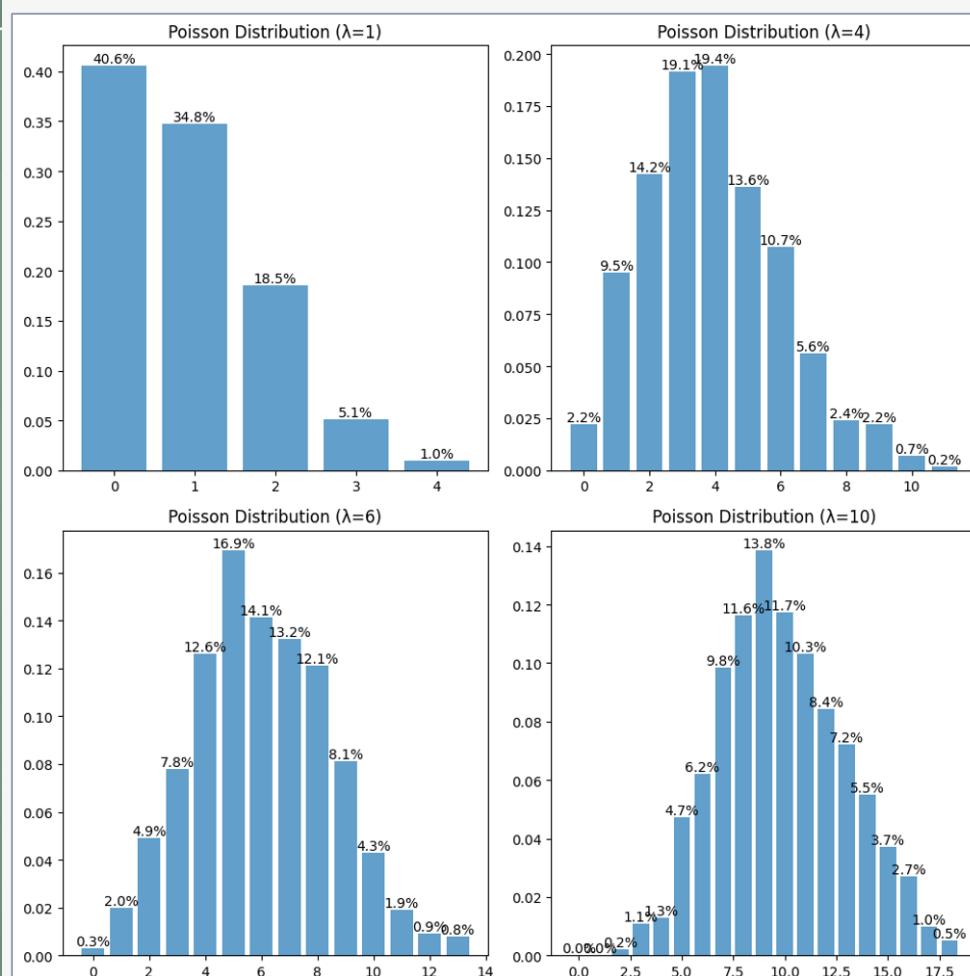
Joint Probability: $P(H,H,T) = (1/2)^2 * (1/2)^1 = 1/8$

Binomial Probability: $P(2 \text{ heads in } 3 \text{ flips}) = C(3,2) * (1/2)^2 * (1/2)^1 = 3 * 1/8 = 3/8$

The binomial probability includes all sequences with 2 heads: HHT, HTH, THH

Poisson distribution

... probability distribution of a given number of independent events occurring in a fixed interval of time or space, given fixed average rate of occurrence.



Events occur with a known constant mean rate and independently of the time since the last event.

Relation to binomial distribution:

1. The number of trials (n) is very large, i.e. continuous.
2. The probability of success (p) is small.
3. The average number of successes (np) is moderate.

Parameter:

$\lambda = np$ (the mean of the distribution)

Use cases:

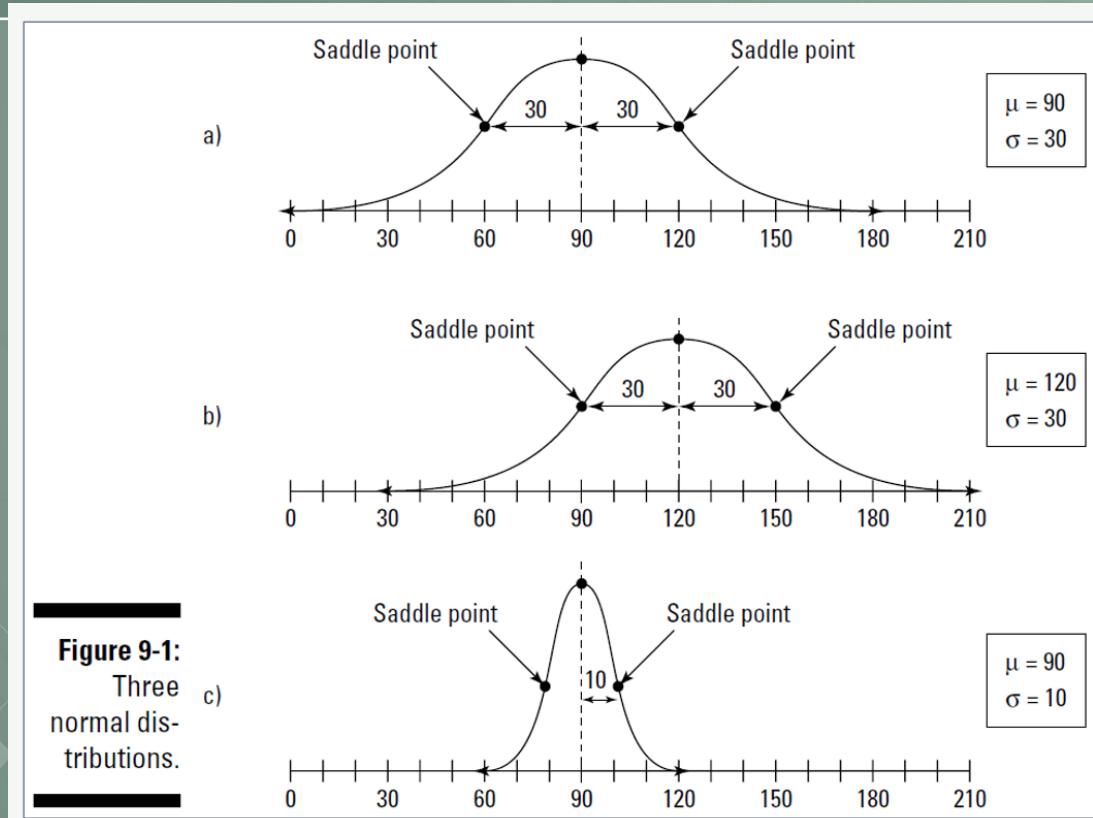
- Number of calls received in a given hour, sales per hour, etc.
- emails arriving in inbox per hour, given average arrival rate
- Very few defect items per hour

AI use cases:

- Sport bets: goals/baskets per match
(Prediction of football, basketball results)

Normal distribution

... is continuous probability distribution characterized by symmetric bell-shaped curve, defined by its mean (average) and standard deviation, also known as the Gaussian distribution



Parameters:

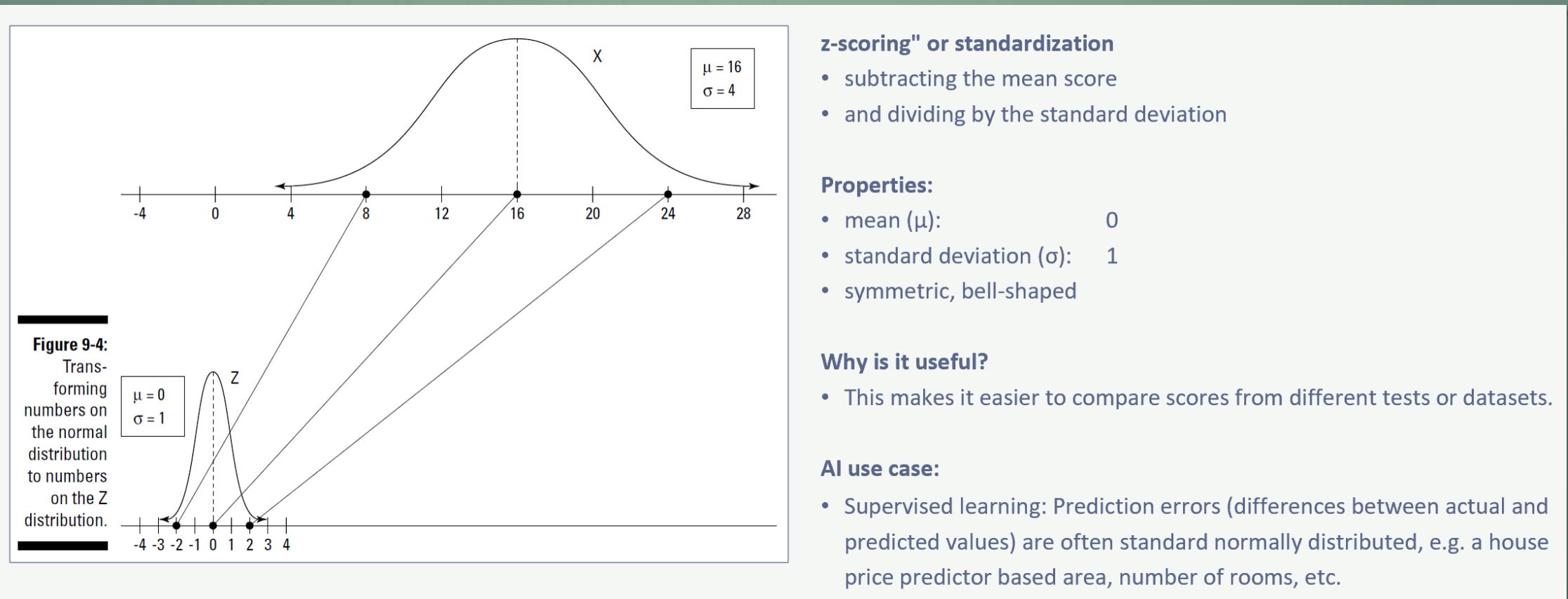
- Mean: where the peak of the bell occurs
- Standard deviation: determines the width of the bell

Use cases:

- Human Heights: most individuals have heights around the average, very tall or small people are much less frequent
- Test Scores: most students will score around average, fewer students have very high/low scores.

Standard normal distribution

... is a special case of the normal distribution with a mean of zero and standard deviation of one.



Students' t-distribution

... is used when sample size is small or when population standard deviation is unknown. It is similar to normal distribution, but with heavier tails, e.g. more flat.

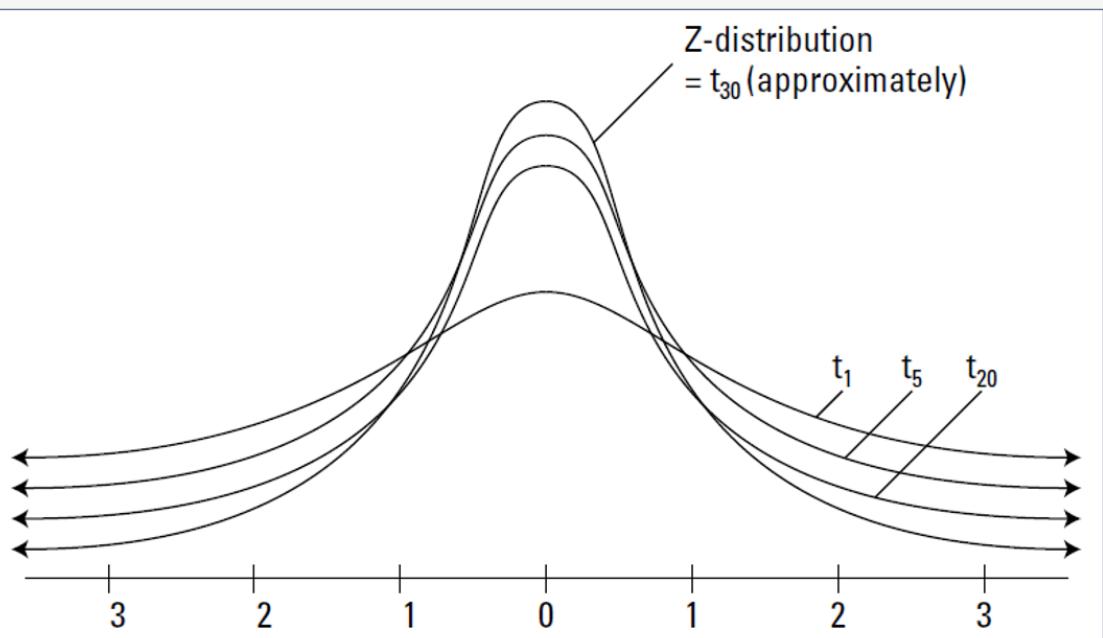


Figure 9-1: t -distributions for different sample sizes

If $n < 30$ (number of samples), t-distribution gives more accurate confidence intervals and hypothesis tests compared to normal distribution.

Use case:

- new drug is tested on small (say, 15 patients). Then, population standard deviation is unknown, and sample that is used for estimation is small.

AI use case:

- Any models with very small sample, e.g. for healthcare predictions.

Exercises

Exercise 1:

Problem: You are about to flip a coin five times. The probability of getting heads (success) is 0.5. What is the probability of getting 3 heads?

Exercise 2:

Problem: A manufacturer knows that 10% of his products are defective. He sells products in boxes of 20. What is the probability that a box will contain exactly 2 defective products?

Exercise 3:

Problem: A multiple-choice quiz contains 10 questions. Each question has four possible answers, of which one is correct. If a student guesses the answer to each question at random, what is the probability that the student will answer exactly 4 questions correctly?

Exercise 4:

Problem: A fast food restaurant serves an average of 10 customers every 15 minutes. What is the probability of serving exactly 7 customers in a 15 minute interval?

Exercise 5:

Problem: An IQ test is scored such that the mean score is 100 and the standard deviation is 15. What is the probability of a person scoring higher than 130?

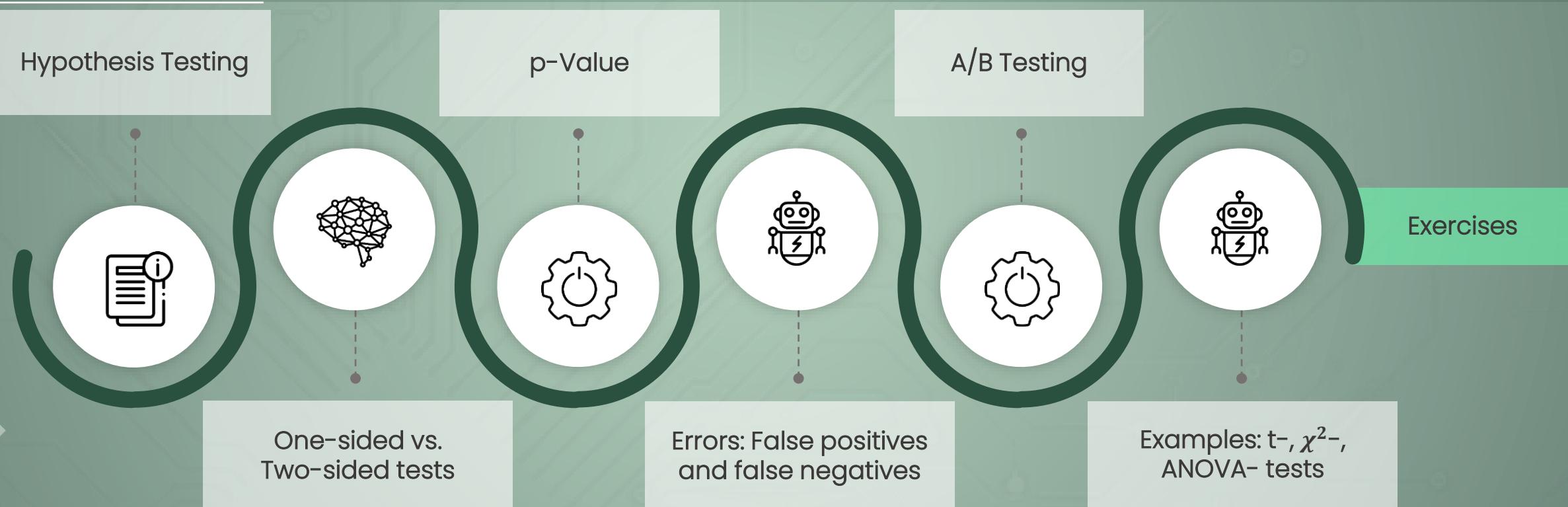
Exercise 6:

Problem: A sample of 20 students' test scores has a mean of 76 and a standard deviation of 10. What is the probability of a student scoring less than 70?



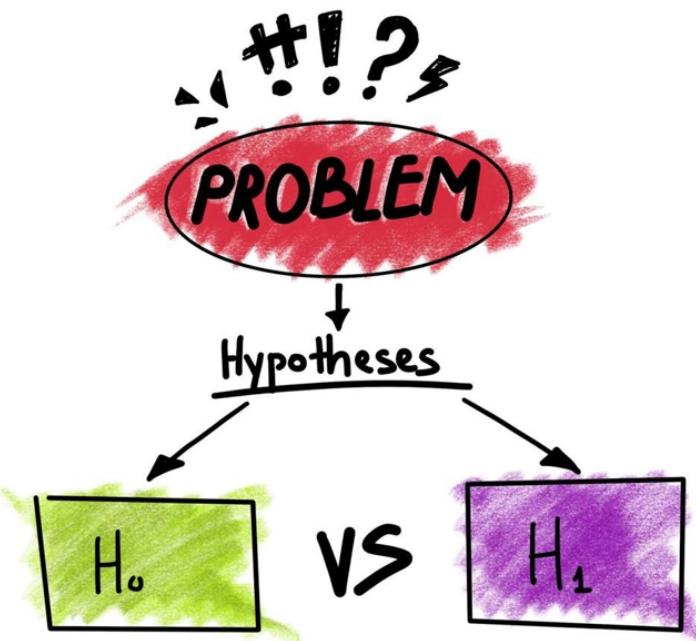
Hypothesis testing

... is statistical method that uses sample data to evaluate the plausibility of null hypothesis ("there is no effect"). Is there enough evidence in the data in order to reject the null?



Hypothesis testing

... is statistical method that uses sample data to evaluate the plausibility of null hypothesis ("there is no effect"). Is there enough evidence in the data in order to reject the null?

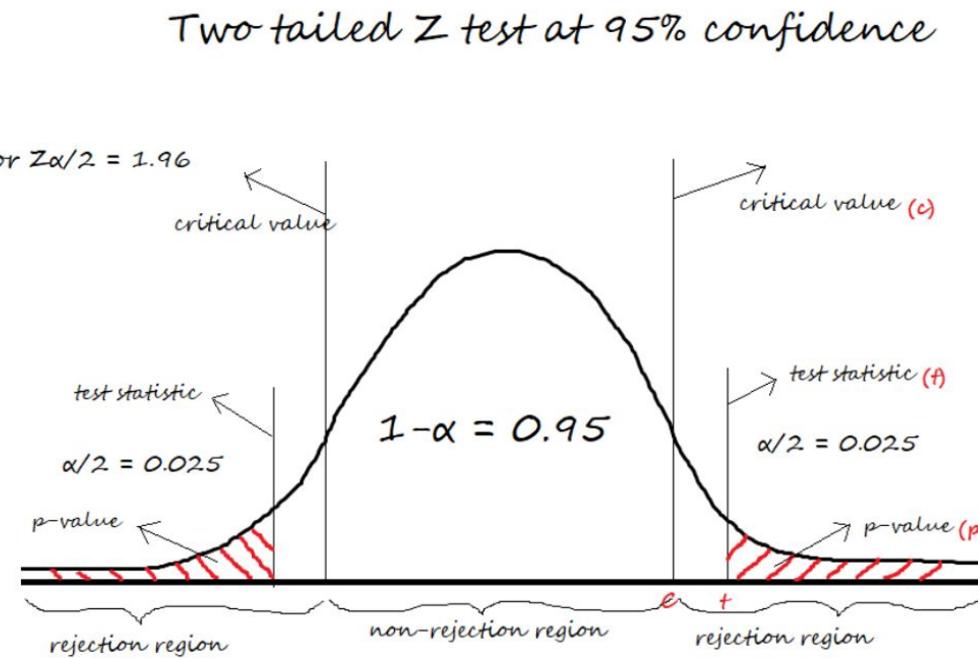


It involves formulating two competing hypotheses

- the null hypothesis (H_0): initial claim, status quo, which usually means: No effect/difference
- and the alternative hypothesis (H_1); we want evidence for the alternative hypothesis
- We assess evidence from the data, in order to determine whether to reject or fail to reject the null hypothesis.

Hypothesis testing

... involves specifying significance level, analyzing sample data using statistical test rejecting or not rejecting null hypothesis



Workflow:

1. State Hypotheses:

Null Hypothesis (H_0)
and Alternative Hypothesis (H_1)

2. Specify critical value α :

usually 0.05 for 2-sided tests

3. Calculate test statistic t with:

- t-test
- χ^2 -test
- z-test
- ANOVA-test, and many more

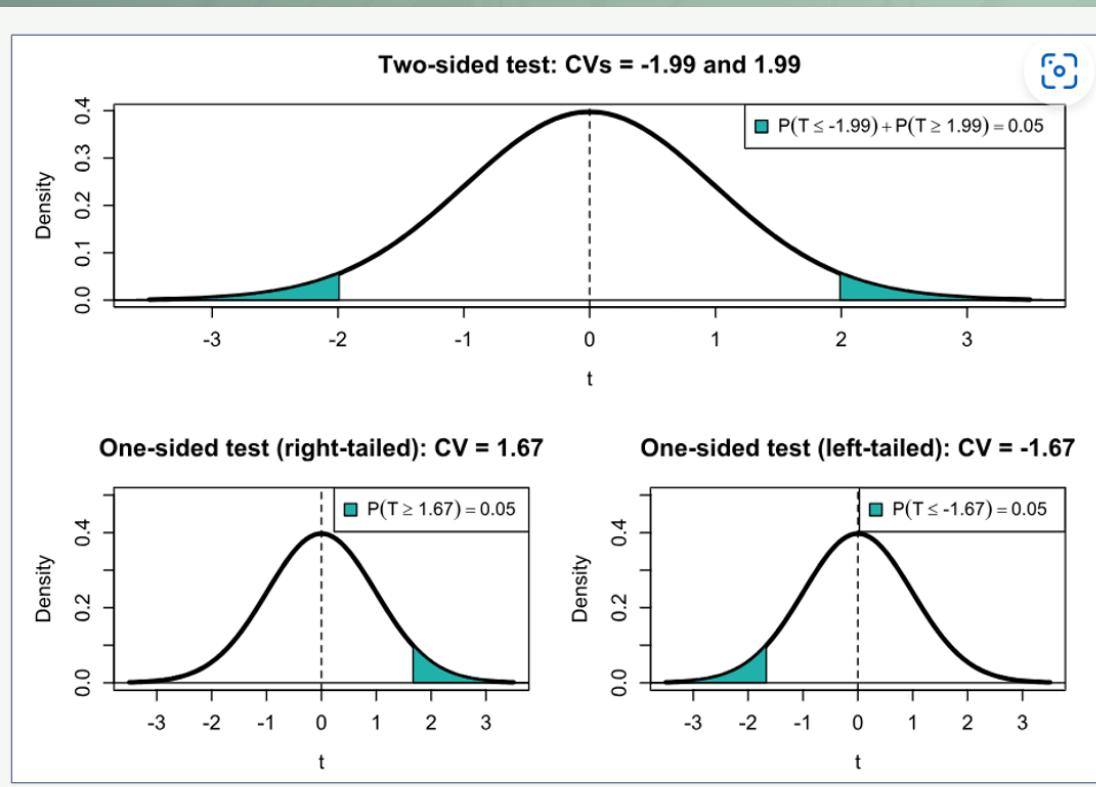
4. Interpret results comparing t and α :

- Reject null hypothesis (support H_1) or not (support H_0)?

One-sided vs. Two-sided tests

One-sided hypothesis test checks if parameter is greater/less than a certain value.

Two-sided test checks if parameter is not equal to a certain value, regardless of direction.



Two-sided (or Two-tailed) Tests

- ... assess if a parameter is either greater or less than the null hypothesis value.
- Example:**
Company is testing if a website redesign affects, i.e. either increases or decreases user time on site.
 - H_0 : Website redesign has no effect on user time on site.
 - H_1 : Website redesign affects user time on site.

One-sided (or One-tailed) Tests

- ... check if parameter is either greater or less than null hypothesis value, but not both.
- Example:**
Pharmaceutical company is testing a new drug and wants to know if it reduces disease recovery time (not interested if it increases).
 - H_0 : Drug does not reduce recovery time.
 - H_1 : Drug reduces recovery time.

Deciding to reject H_0 or not

... is standardized procedure like an “API”

- Initial assumption: “There is no effect”, i.e. we don’t reject H_0 yet
- Samples from population vary around the population mean μ
- Theoretical probability of samples
 - around μ (or further away, the integral) is high
 - far away from μ (or even further, the integral) is low
- Now you take the real sample and focus on distance from μ

If it is **very close**

sample is **not extreme**

theoretical probability of getting more extreme sample is **high**

we say: “There is no effect, there is **not enough evidence** to reject H_0 ”

simpler: “**We choose H_0** ”

If it is **very far**

sample is **extreme**

theoretical probability of getting more extreme sample is **low**

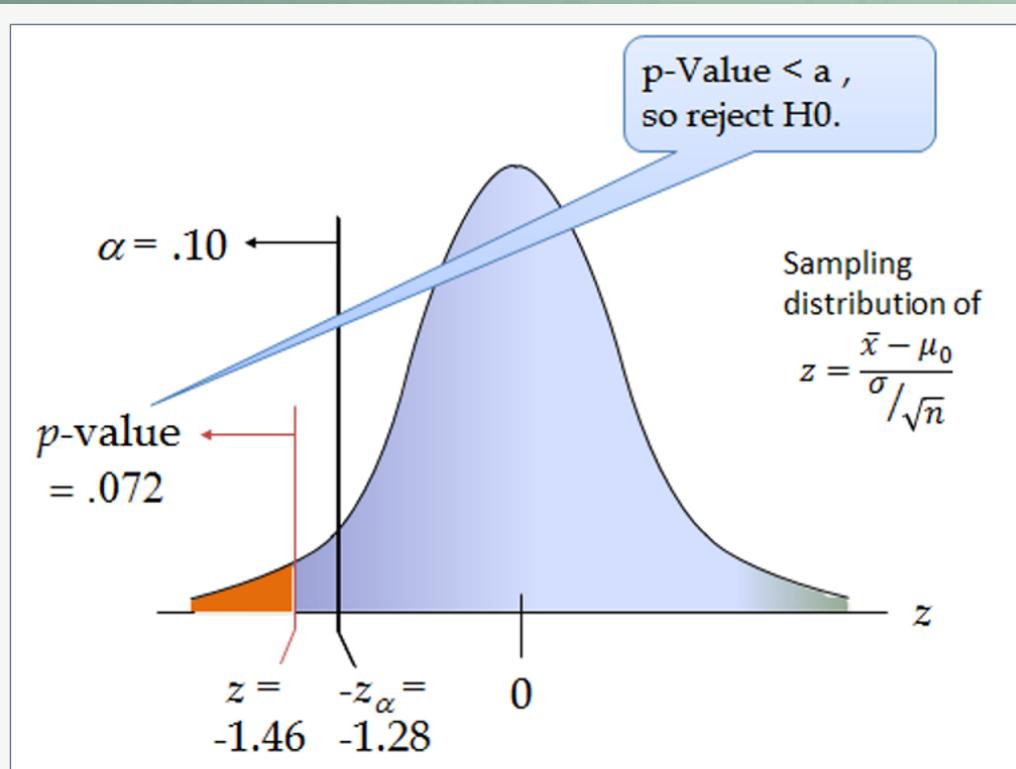
we say: “There is an effect, there is **enough evidence** to reject H_0 ”

simpler: “**We choose H_1** ”

We decide based on this integral. How is it called?

p-value

... is a tool for deciding whether to reject or not reject the null hypothesis



Example for rejecting H_0 , the 2 groups are different

Interpreting the P-value

- Small p-value (typically $\leq \alpha = 0.05$) suggests strong evidence against the null hypothesis, "we reject the null", "we support H_1 "
- Large p-value $\alpha = 0.05$) suggests weak evidence against the null hypothesis, "we fail to reject the null", "we support H_0 "
- If we reject the null, there is typically significant difference between the 2 groups

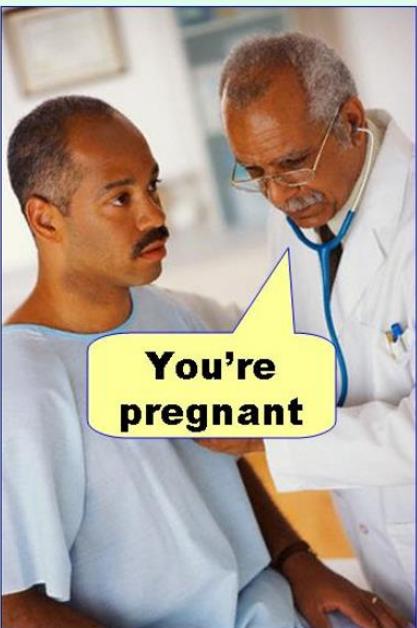
Properties

- p-values can be affected by the size of the sample.
- Figure: in 0.72 of 100 times you should have selected H_0 .

Errors: False positives and false negatives

... matter, as they represent incorrect conclusions that can impact decisions based on test results

Type I error
(false positive)



Type II error
(false negative)



- Selecting H_1 is the “positive” statement
- Selecting H_0 is the “negative” statement

Type I error means “false positive”:

- When H_0 is true and therefore “selecting H_1 is incorrect”
- precise formulation: H_0 is incorrectly rejected

Type II error means false negative:

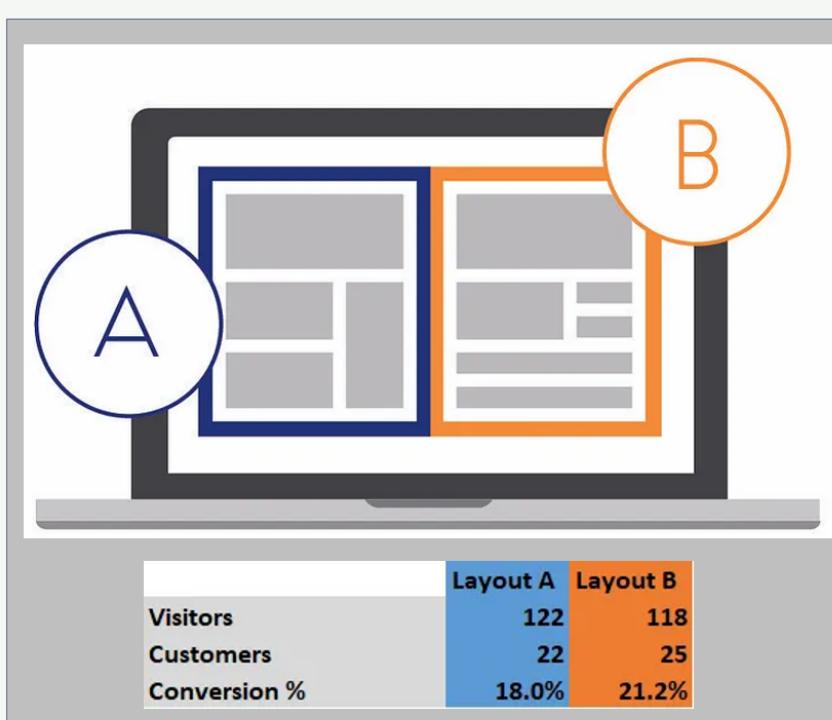
- When the H_0 is false, but “we select it”
- precise formulation: we fail to reject it

Connection to significance level α :

- The maximal probability of making Type I error is defined by α , such as 0.05 or 0.01.

What is A/B-Testing?

... also known as split testing, is method of comparing two versions of webpage, email, or other user experience to determine which one performs better.



<https://medium.com/towards-data-science/data-science-you-need-to-know-a-b-testing-f2f12aff619a>

- **Identify goal:**

Your goal might be to increase website conversions, improve email click-through rates, or boost the number of sign-ups.

- **Create Variants:**

Develop two versions of element you want to test - the control (A) and the variant (B). The versions should be identical except for one change.

- **Split audience:**

Randomly divide audience into two equal groups.
One sees version A, the other sees B.

- **Conduct test:**

Release both versions at the same time and collect data on how each version performs in relation to your goal.

- **Analyze the Results:**

Use statistical analysis to determine which version performed better.

- **Implement Changes:**

If B performs significantly better, consider replacing version A with B.

A/B Testing: Choosing the Right Test Type

Tests types are nuanced, slight differences in hypothesis determine selection

Test Type	Scipy method	Use Case, compare:	Examples
1. Two sample t	ttest_ind	<ul style="list-style-type: none">• means of• continuous data between• two groups of independent samples	<ul style="list-style-type: none">• Company tests new website B and compares it with old one A• Visitors are divided into 2 groups that only see one of them• Visit durations are measured <p>→ Does the subgroup stay longer on new website B?</p>
2. Paired t	ttest_rel	<ul style="list-style-type: none">• means of• continuous data for the same group at different times	<ul style="list-style-type: none">• Like 1• But: The same group is tested• Visit durations are measured, first with website A later with B <p>→ Does the same group stay longer on new website B?</p>
3. χ^2 for independence	chi2_contingency or chi2.sf	<ul style="list-style-type: none">• Variance of• categorical data between• two or more groups	<ul style="list-style-type: none">• A company has employees in 4 experience levels• On each level a part of them received promotion (yes/no) <p>→ Does experience level have impact on promotion?</p>
4. One factor ANOVA	f_oneway	<ul style="list-style-type: none">• means of• continuous data between• ≥ three groups of independent samples	<ul style="list-style-type: none">• Education program in company,• 3 learning styles: self-paced online courses, instructor led, mixed• Employees are divided in 3 subgroups <p>→ Does the learning style have impact on learning success?</p>

Two sample t-tests

... are statistical tests used to compare the means of continuous data between two groups and determine if they are significantly different from each other.

The formula for the t-statistic in a two-sample t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With:

- \bar{X}_1 and \bar{X}_2 are the sample means of the two groups
- s_1^2 and s_2^2 are the sample variances of the two groups.
- n_1 and n_2 are the sample sizes of the two groups
- Degrees of freedom: $df = n_1 + n_2 - 2$

Variant A: No significant difference

```
# Randomly generating test scores for Group A and Group B
np.random.seed(0) # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(80, 10, 30)
→ df = 58
```

Variant B: Significant difference

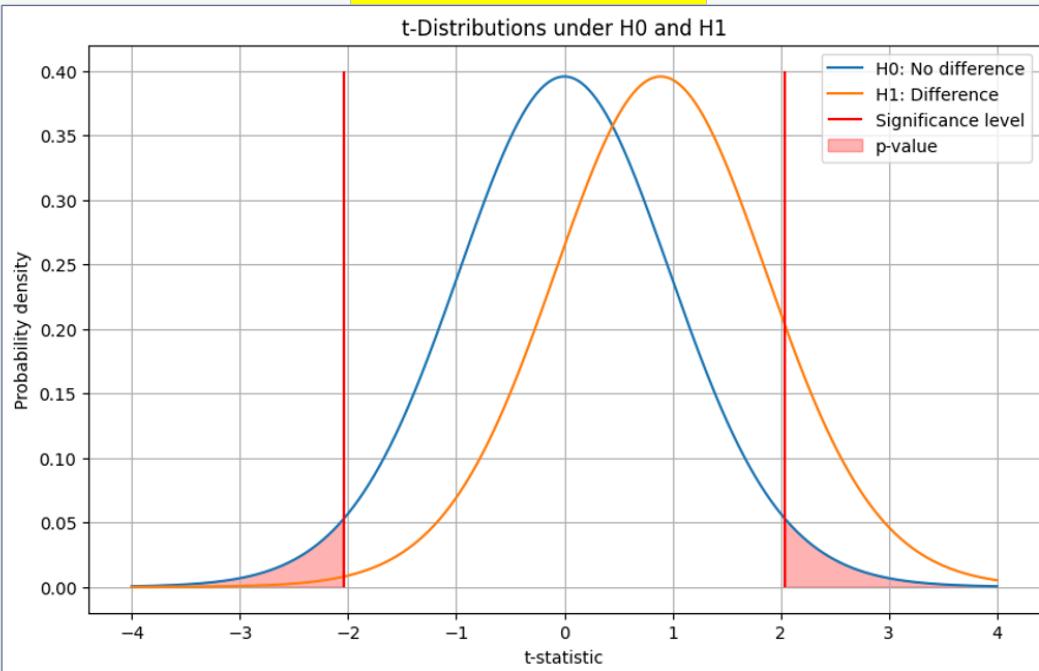
```
# Randomly generating test scores for Group A and Group B
np.random.seed(0) # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(90, 10, 30)
→ df = 58
```

Two sample t-tests

... are statistical tests used to compare the means of continuous data between two groups and determine if they are significantly different from each other.

ttest_stat = 0.8897

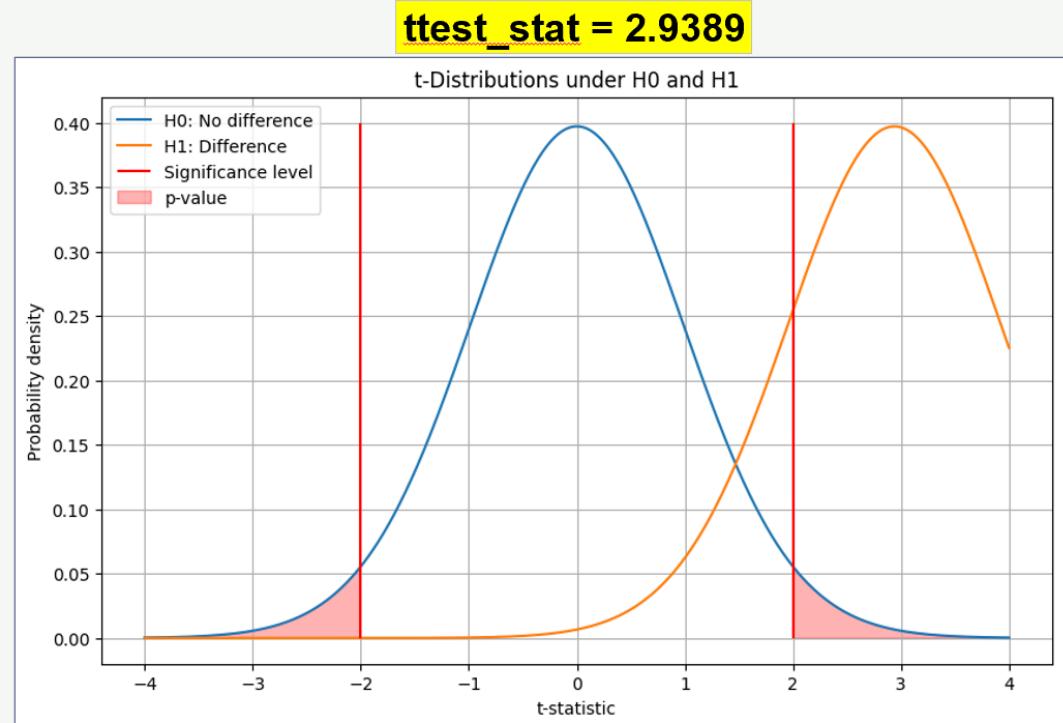
Critical value: 2.001 (2-sided!)



H0:

There is no evidence that subgroup B stays longer on the website ...

ttest_stat = 2.9389



H1:

There is evidence that subgroup B stays longer on the website ...

Paired t-tests

... are used when you want to compare the means of continuous data for the same group at two different times or under two different conditions

The formula for the t-statistic in a paired t-test is:

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

With:

- \bar{D} is the mean of the differences between the paired observations.
- s_D is the standard deviation of the differences between the paired observations.
- n is the number of pairs.
- **Degrees of freedom: $df = n - 1$**

Variant A: No significant difference

```
# Randomly generating test scores for Group A and Group B
np.random.seed(0) # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(80, 10, 30)
→ df = 29
```

Variant B: Significant difference

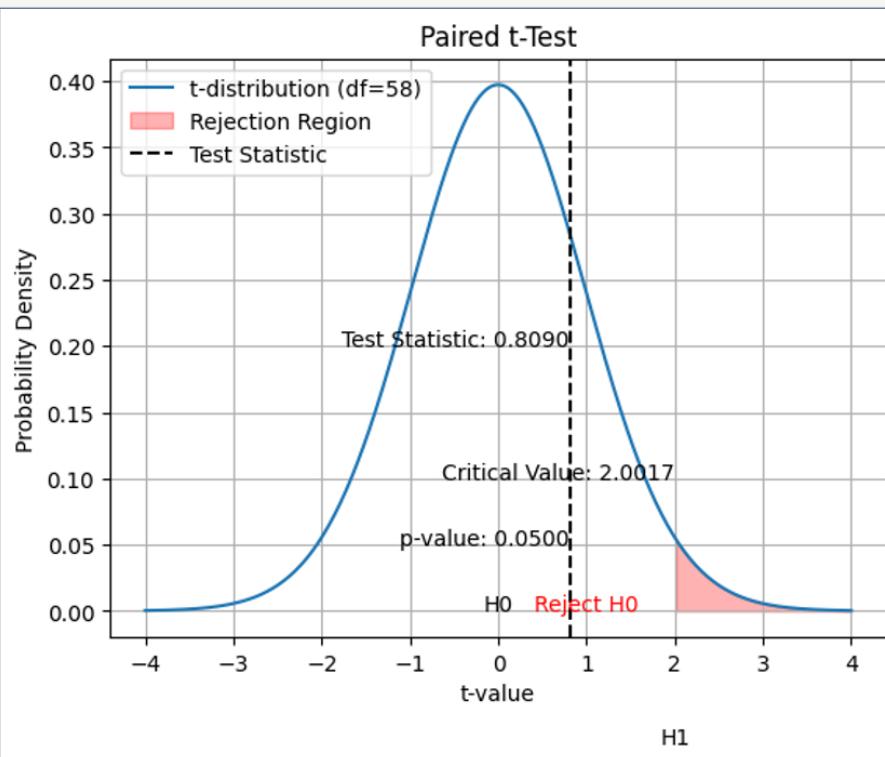
```
# Randomly generating test scores for Group A and Group B
np.random.seed(0) # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(100, 10, 30)
→ df = 29
```

Paired t-tests

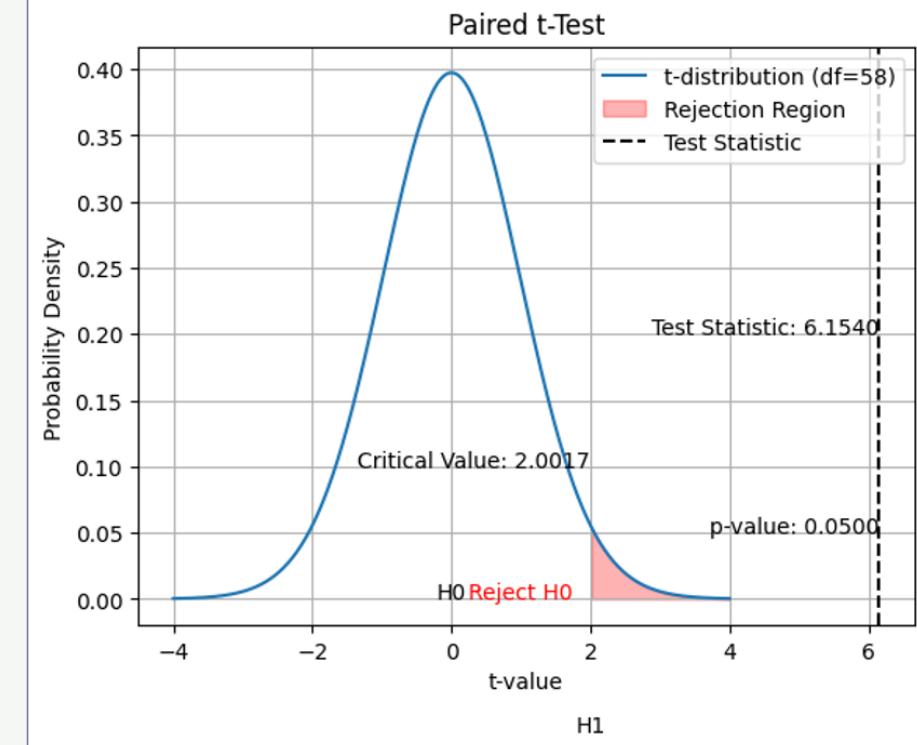
... are used when you want to compare the means of continuous data for the same group at two different times or under two different conditions

paired_t_stat = 0.8090

Critical value: 2.0452 (2-sided!)



paired_t_stat = 6.1540



H0:

There is no evidence that group stays longer on the new website ...

H1:

There is evidence that group stays longer on the new website ...

Exercises

Exercise for ttest_ind:

Based on the ttest_ind example, try new data points and find data point configurations, where H_0 is very close of being rejected or is just rejected.

Generate figure according to the given figures.

Exercise for ttest_rel:

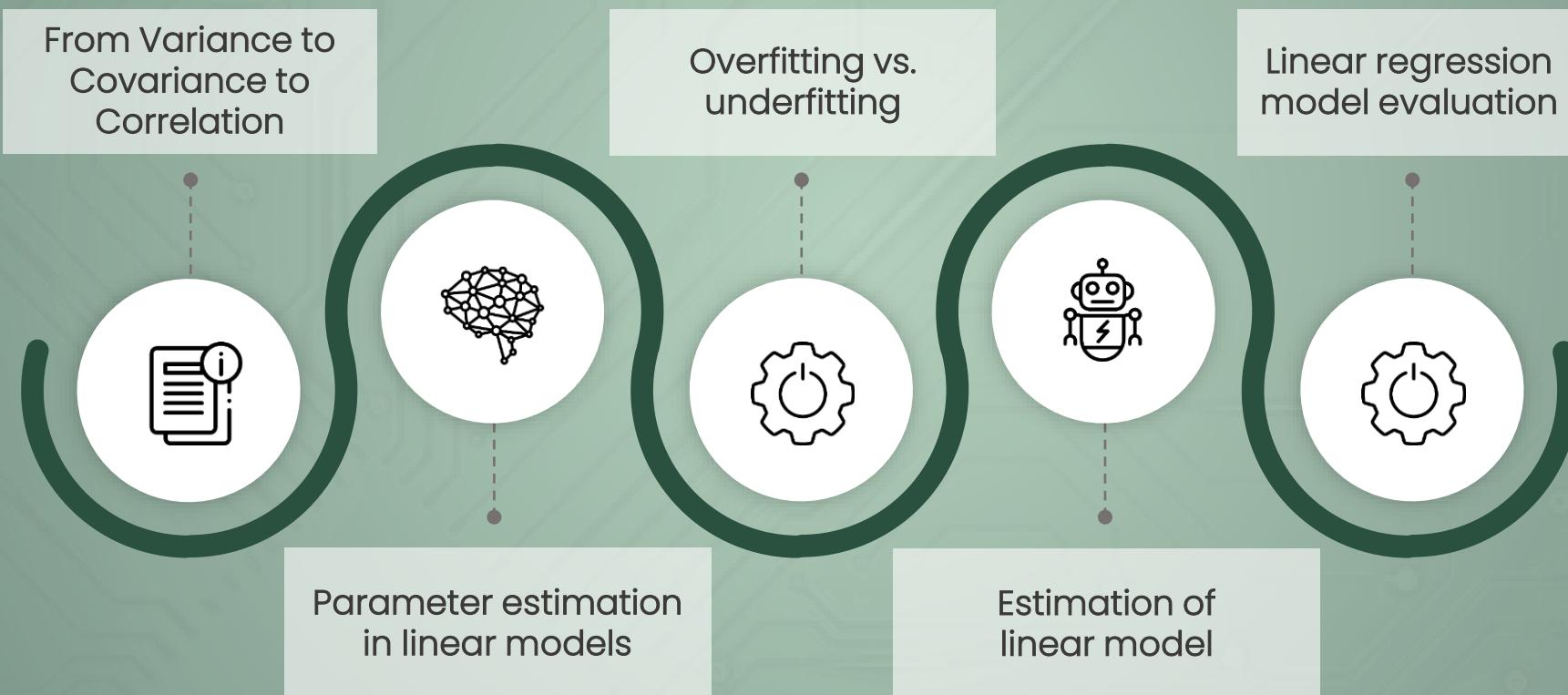
Based on the ttest_ind example, try new data points and find data point configurations, where H_0 is very close of being rejected or is just rejected.

Generate figure according to the given figures.



Model estimation: Linear Regression

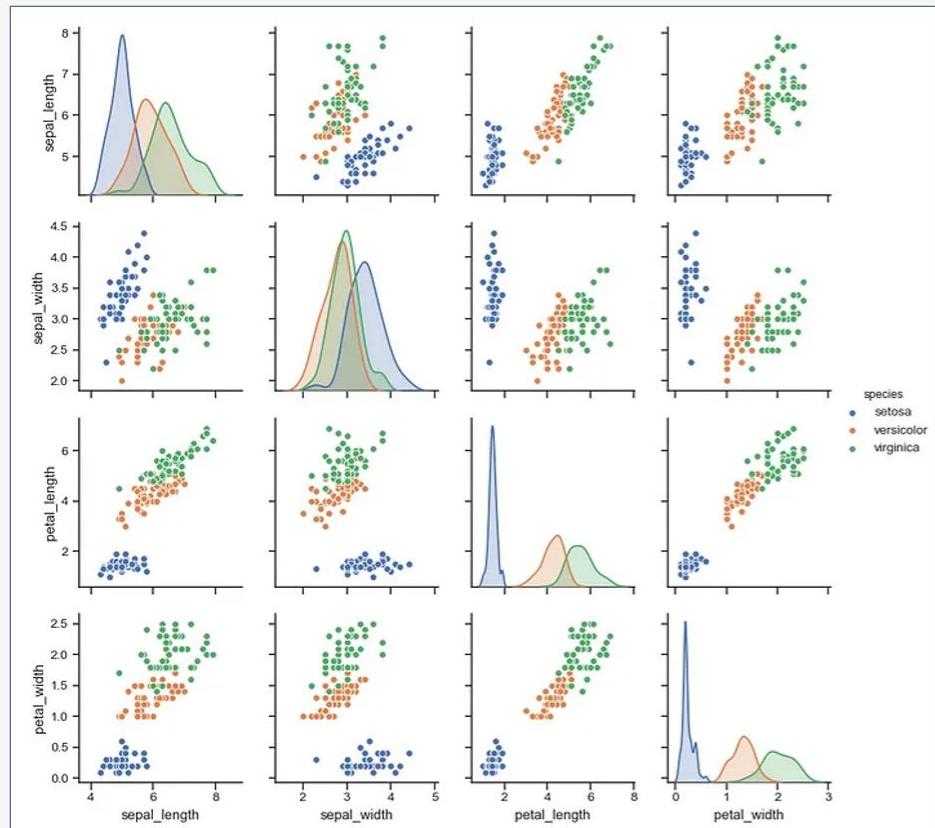
Model estimation is the process of determining the parameters of statistical model that best fit observed data. Linear regression models relationships using straight line.



From Variance to Covariance to Correlation

Covariance is a generalization of variance to multiple dimensions or variables.

Correlation is a normalized form of covariance to have values between -1 and +1.



Variance (σ^2):

- Measures how far a set of numbers spread out from mean.
- In the figure: The diagonal

Covariance:

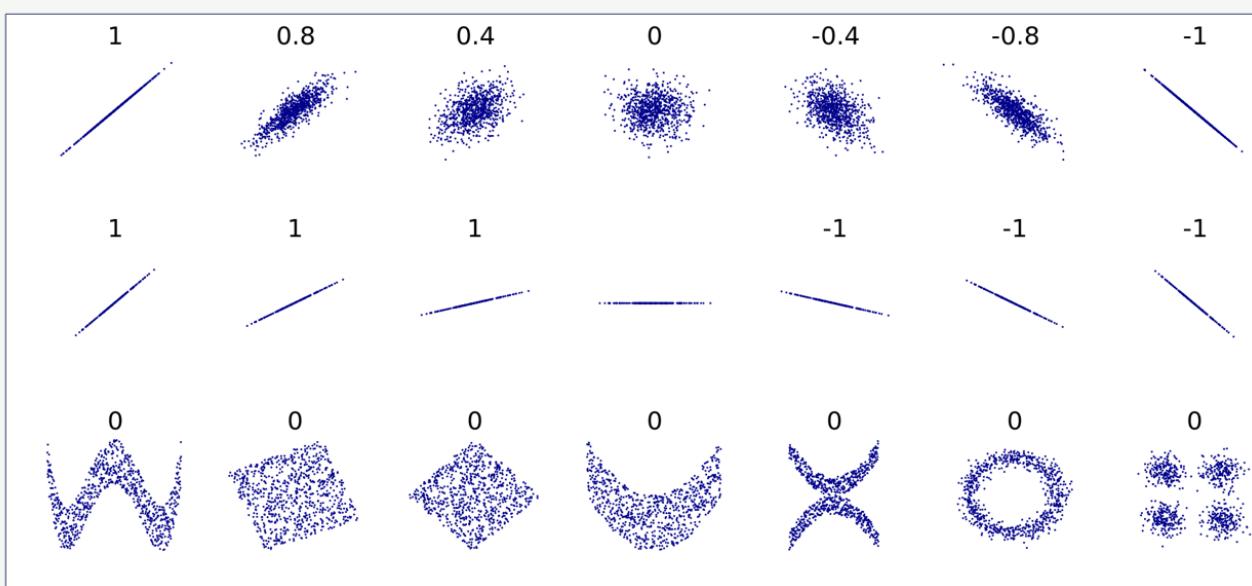
- Measures how much two random variables vary together.
- In the figure: The triangles above and below the diagonal

Correlation:

- Standardized measure of the relationship between two variables
- ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), e.g. it is the normalized covariance.

Correlation

... is the standardized measure of the relationship between two variables



Positive Correlation:

- When two variables increase or decrease together.
- For example, the more time spent studying (variable 1), the higher the grades (variable 2). Correlation value ranges from 0 to +1.

Negative Correlation:

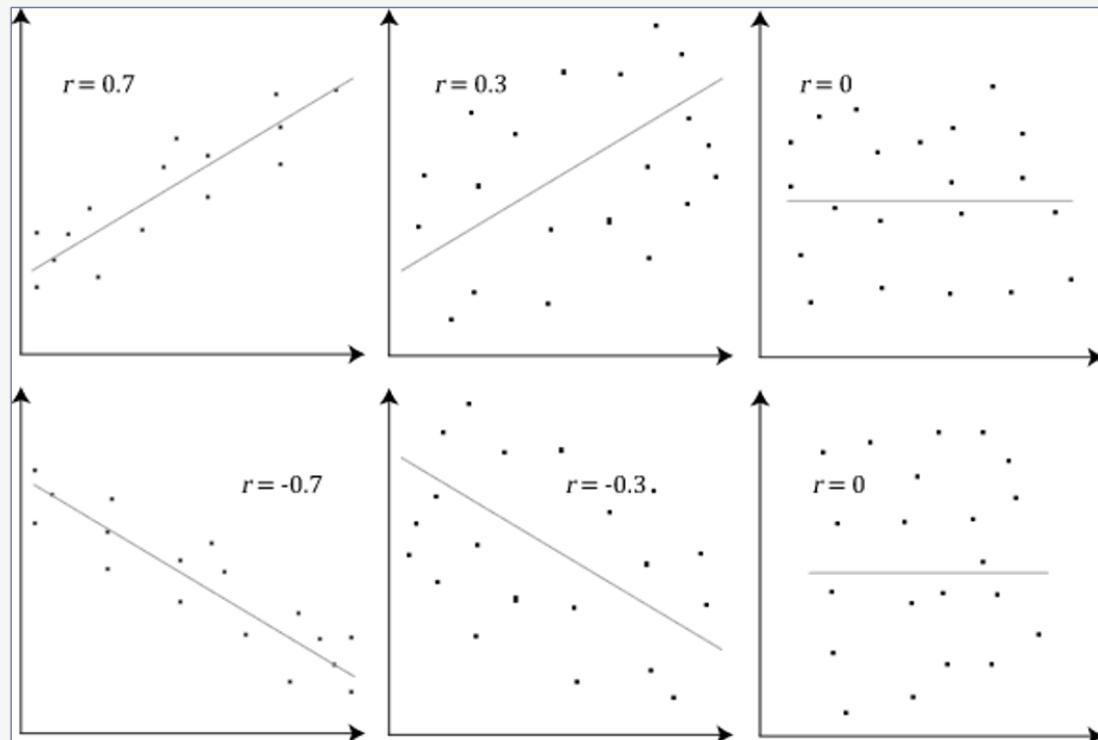
- When one variable increases as the other decreases.
- For example, the more time spent watching TV (variable 1), the lower the grades (variable 2). Correlation value ranges from 0 to -1.

Source:

<https://en.wikipedia.org/wiki/Correlation>

Pearson's Correlation Coefficient - Test

Tests whether two samples have a linear relationship.



Source:

<https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>

Assumptions

- Observations in each sample are independent and identically distributed (iid).
- Observations in each sample are normally distributed.
- Observations in each sample have the same variance.

Interpretation

- H₀: the two samples are independent.
- H₁: there is a dependency between the samples.

```
# Example of the Pearson's Correlation test
from scipy.stats import pearsonr
data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
data2 = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536, 3.350, -1.578, -3.537, -1.579]

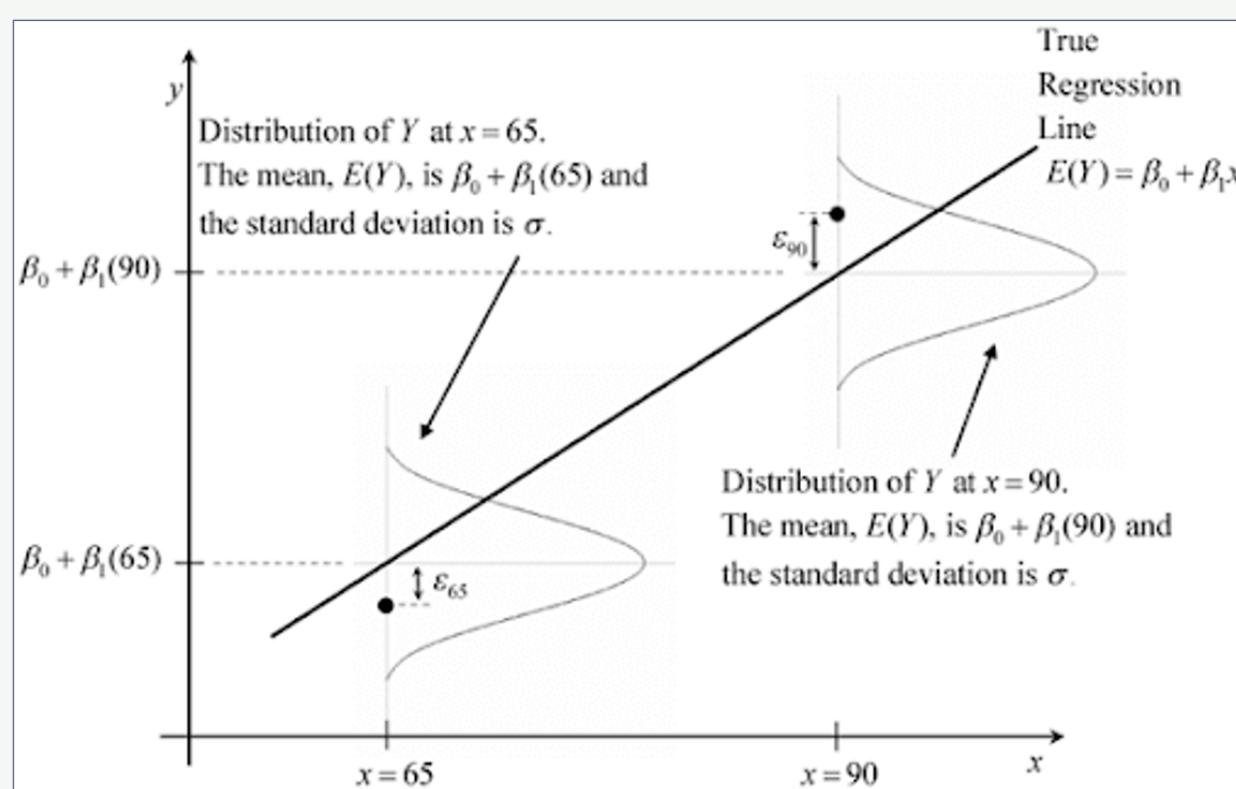
stat, p = pearsonr(data1, data2)
print(f'stat={stat:.3f}, p={p:.3f}')

if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')

stat=0.688, p=0.028
Probably dependent
```

Parameter estimation in linear models

... involves finding the coefficients that best fit the observed data, i.e. represent the relationship between two correlated variables (X and Y)



Simple linear regression model:

$$Y = \beta_0 + \beta_1 * X + \varepsilon$$

β_0 (intercept) and β_1 (slope) are the parameters to be estimated.

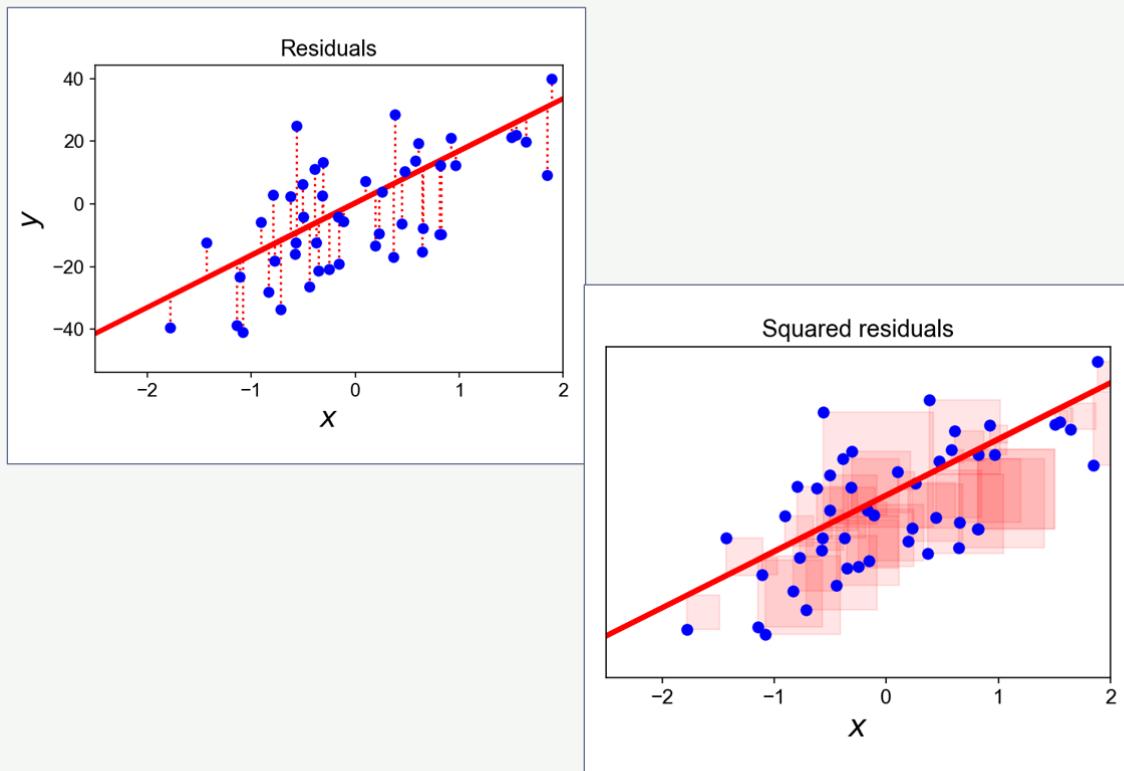
- β_0 : Value of Y when X = 0.
- β_1 : Change in Y for each one-unit change in X.

Use Cases:

- Generalization of sample data to overall population data
- Forecasting like prediction of house prices based on independent features

Ordinary Least Squares (OLS) estimation

... represents a relationship between two correlated variables (X and Y) and is used in linear regression analysis



Ordinary Least Squares (OLS) estimation:

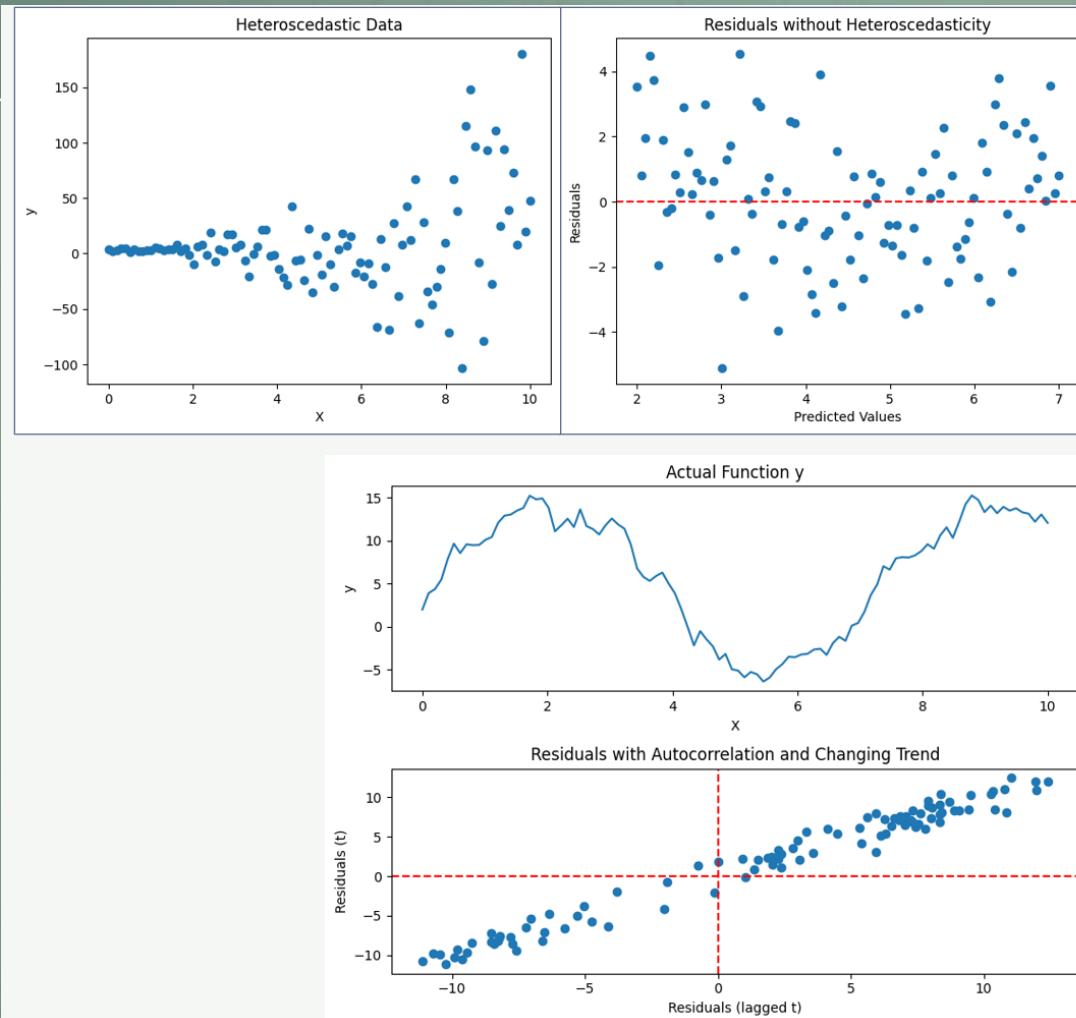
- Method used in linear regression to estimate the model's parameters
- by minimizing the sum of the squares of the differences between the observed and predicted values of the dependent variable.
- Separation systematic relations from white noise

IID & BLUE:

If the residuals are identically and independently distributed (iid) the OLS estimator is used, because it provides best linear unbiased (BLUE) estimation.

OLS - What if IID-ness of residuals is violated?

Here it gets interesting 😊



Heteroscedasticity

- Variance is not identical
- Breusch-Pagan-Test: H_1 means that there is enough evidence from residuals to assume heteroscedasticity

Autocorrelation (typical example: time series)

- Residuals depend on preceders
- Durbin-Watson-Test:

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

- $d \approx 2$: No autocorrelation
- $0 < d < 2$: Positive autocorrelation
- $2 < d < 4$: Negative autocorrelation

Transformations can be used in order to make OLS estimation usable...

Overfitting

... is a modeling error where a model is trained too well on the training data

Problems associated with Overfitting:

Poor Generalizability:

Overfit models perform well on training data but poorly on new, unseen data, which limits their predictive accuracy.

Complexity:

Overfitting often results in overly complex models that have too many parameters.

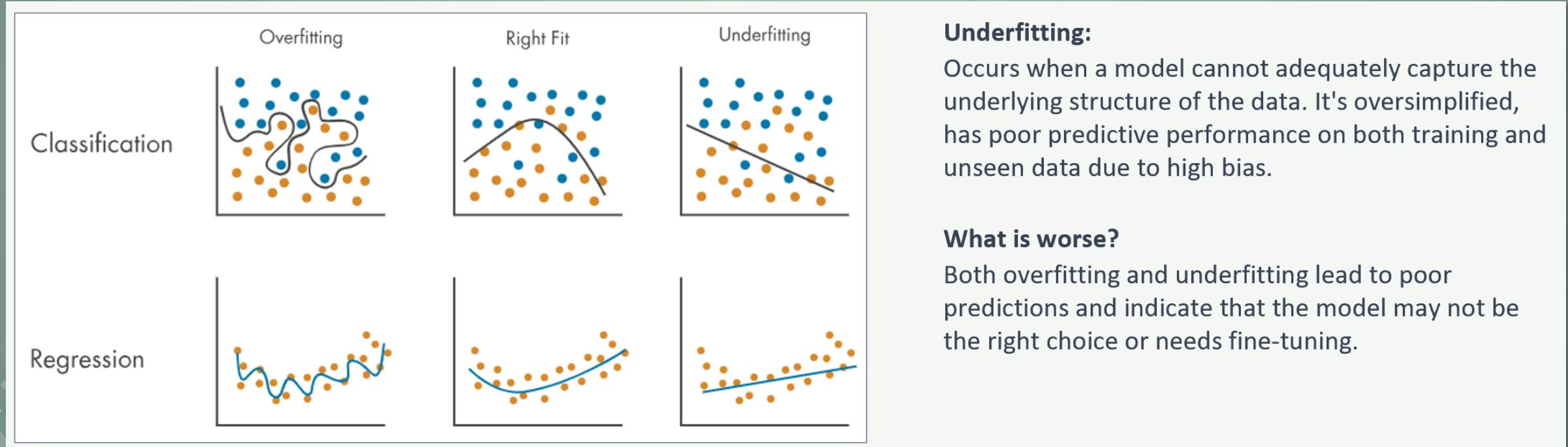
Unreliable Predictions:

Overfit models are sensitive to minor fluctuations in data, leading to unreliable, inconsistent predictions over time.



Overfitting vs. underfitting

In opposite to overfitting, underfitting is when the model fails to capture important trends in the training data, resulting in poor performance on both the training and unseen data.



Underfitting:

Occurs when a model cannot adequately capture the underlying structure of the data. It's oversimplified, has poor predictive performance on both training and unseen data due to high bias.

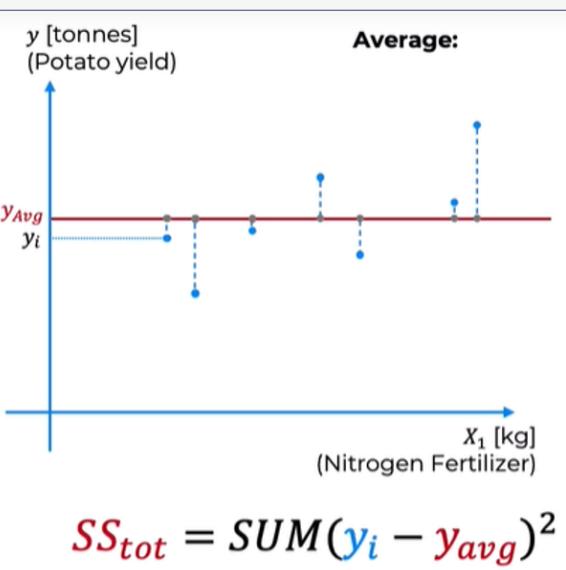
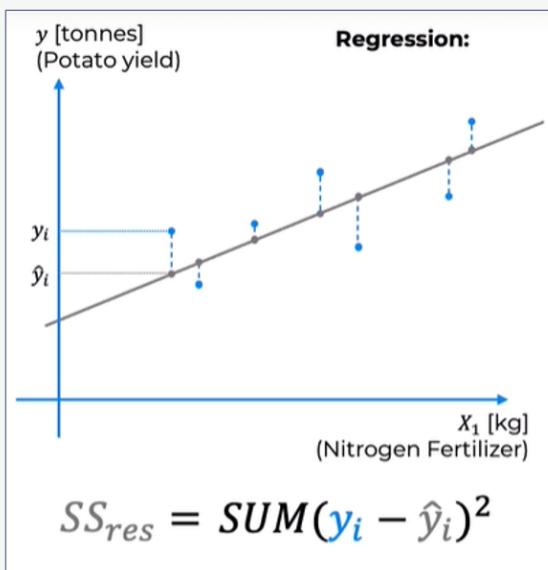
What is worse?

Both overfitting and underfitting lead to poor predictions and indicate that the model may not be the right choice or needs fine-tuning.



R-squared: Goodness of fit

... is a statistical measure in regression models that represents the proportion of the variance in the dependent variable that is predictable from the independent variables



Properties:

- Ranges from 0 to 1.
- Value of 1 → independent variable(s) perfectly predict dependent variable.
- Value of 0 → no predictive power.
- Higher R-Squared values → better fit, however very high R-Squared indicates overfitting.
- Only interpretable in the context of the data and domain. Low R-Squared could be acceptable in some fields where prediction is difficult.

Adjusted R-squared:

- Problem to be solved: Many input factors may improve fit, but lead to multicollinearity, overfitting etc.
- In order to incentivize leaner models, number of input factors is punished

$$R^2 = 1 - \frac{SSR}{SST}$$

Statsmodels: Summary for OLS

... provides a comprehensive summary of the results from fitted OLS model

OLS Regression Results																																																									
Model Info		<pre>Dep. Variable: y Model: OLS Method: Least Squares Date: Wed, 05 Jul 2023 Time: 22:35:07 No. Observations: 100 Df Residuals: 94 Df Model: 5 Covariance Type: nonrobust</pre>																																																							
Model Info		<pre>R-squared: 0.998 Adj. R-squared: 0.998 F-statistic: 8156. Prob (F-statistic): 2.50e-122 Log-Likelihood: -143.90 AIC: 299.8 BIC: 315.4</pre>																																																							
		<p>The closer to 1 the better</p>																																																							
Coefficients Table		<table><thead><tr><th></th><th>coef</th><th>std err</th><th>t</th><th>P> t </th><th>[0.025</th><th>0.975]</th></tr></thead><tbody><tr><td>const</td><td>2.1643</td><td>0.459</td><td>4.718</td><td>0.000</td><td>1.254</td><td>3.075</td></tr><tr><td>x1</td><td>3.0226</td><td>0.037</td><td>82.532</td><td>0.000</td><td>2.950</td><td>3.095</td></tr><tr><td>x2</td><td>1.9670</td><td>0.039</td><td>50.758</td><td>0.000</td><td>1.890</td><td>2.044</td></tr><tr><td>x3</td><td>4.0132</td><td>0.036</td><td>113.018</td><td>0.000</td><td>3.943</td><td>4.084</td></tr><tr><td>x4</td><td>0.9767</td><td>0.037</td><td>26.148</td><td>0.000</td><td>0.903</td><td>1.051</td></tr><tr><td>x5</td><td>4.9814</td><td>0.035</td><td>141.395</td><td>0.000</td><td>4.911</td><td>5.051</td></tr></tbody></table>								coef	std err	t	P> t	[0.025	0.975]	const	2.1643	0.459	4.718	0.000	1.254	3.075	x1	3.0226	0.037	82.532	0.000	2.950	3.095	x2	1.9670	0.039	50.758	0.000	1.890	2.044	x3	4.0132	0.036	113.018	0.000	3.943	4.084	x4	0.9767	0.037	26.148	0.000	0.903	1.051	x5	4.9814	0.035	141.395	0.000	4.911	5.051
	coef	std err	t	P> t	[0.025	0.975]																																																			
const	2.1643	0.459	4.718	0.000	1.254	3.075																																																			
x1	3.0226	0.037	82.532	0.000	2.950	3.095																																																			
x2	1.9670	0.039	50.758	0.000	1.890	2.044																																																			
x3	4.0132	0.036	113.018	0.000	3.943	4.084																																																			
x4	0.9767	0.037	26.148	0.000	0.903	1.051																																																			
x5	4.9814	0.035	141.395	0.000	4.911	5.051																																																			
		<p>The smaller the better</p>																																																							
Skewness and kurtosis		<pre>Omnibus: 2.625 Prob(Omnibus): 0.269 Skew: -0.361 Kurtosis: 3.101 Durbin-Watson: 1.810 Jarque-Bera (JB): 2.212 Prob(JB): 0.331 Cond. No. 50.2</pre>																																																							
		<p>The closer to 0 the better</p>																																																							
		<p>The smaller the better</p>																																																							
		<p>Autocorrelation, e.g. basis of time series</p>																																																							
		<p>Multicollinearity</p>																																																							
<p>Notes:</p> <p>[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.</p>																																																									

Statsmodels: Summary – Coefficients table

... provides detailed information about each parameter in the model

	coef	std err	t	P> t	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

coef:

- Estimates of the parameters

1.std err:

- Standard error of the estimate of the coefficient
- not standard deviation!
- Smaller values indicate a more precise estimate.

t:

- t-statistic value
- Ratio: coefficient / std. err
- measure of how statistically significant the coefficient is.

[0.025 0.975]

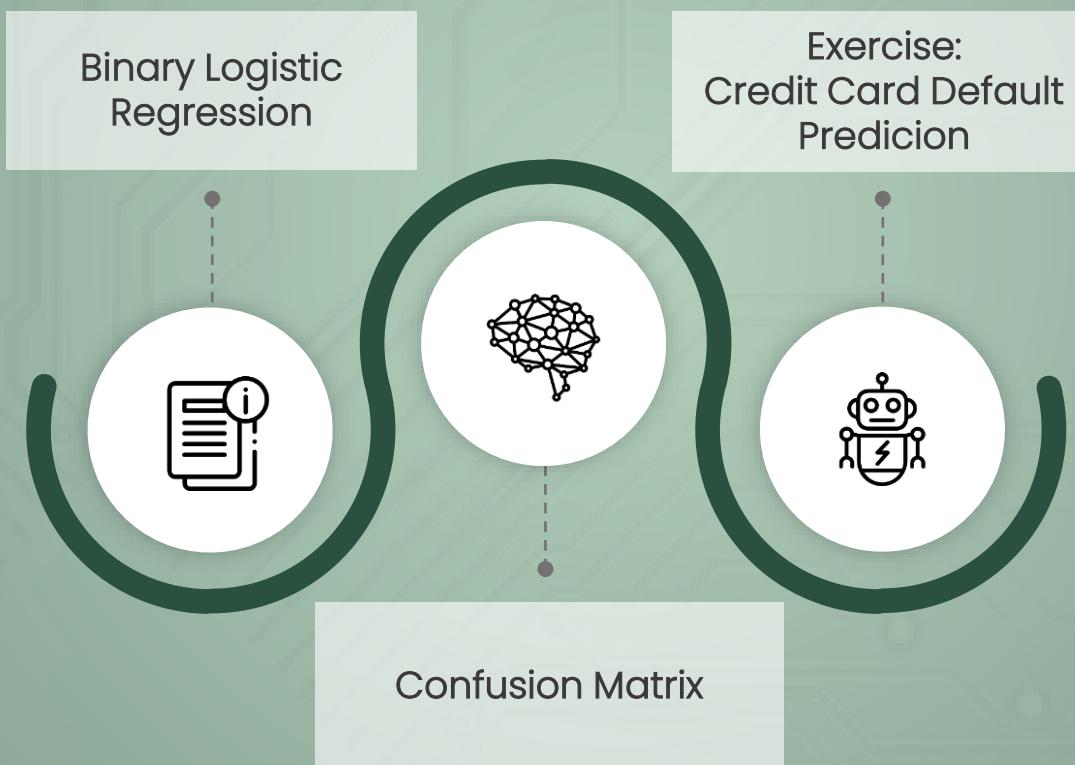
- 95% confidence intervals for the coefficient.
- If zero is not in this interval, that suggests the variable is a significant predictor at the 0.05 level.

P>|t|:

- p-value associated with the t-statistic.
- Smaller p-value (<0.05, traditionally) suggests that null hypothesis can be rejected, i.e. independent variable **is a significant predictor** of dep. variable.

Model estimation: Logistic Regression

... is statistical method for modeling binary or categorical outcomes using logistic function



Aim of binary logistic regression

... is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable) and a set of independent variables.

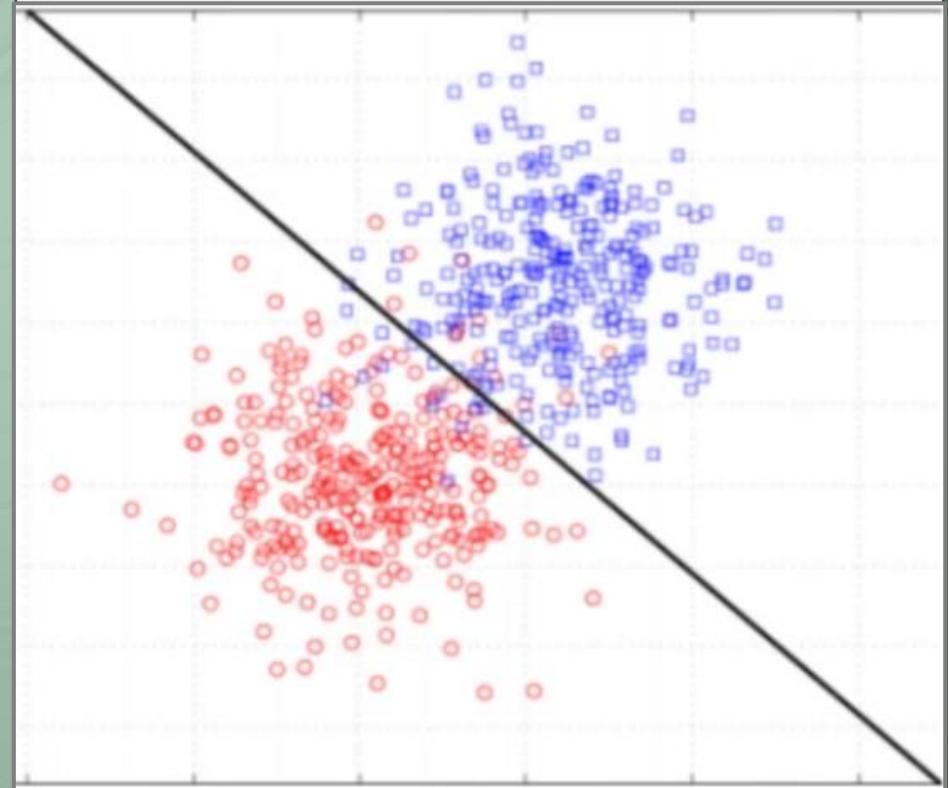
Probability Estimation

- Logistic Regression predicts the probability of occurrence of event by fitting data to logistic function.
- Trained model estimates probability that given instance falls into one of these two categories.
- Linear discriminant is trained in a way that errors, i.e. squared distances to each data point are minimized

Model Application

- Credit scoring
- measuring disease prevalence
- Any generalization of data that has binary target feature

Separation of data points with 2 features using linear discriminant



Centerpiece of logistic regression

... is linear combination of features (as in linear regression) projected on inverse logit curve

Linear combination of features

$$b_0 + b_1 x_1 + \dots + b_n x_n = \omega^T x$$

Logit (Logged-odds) equation

$$\ln\left(\frac{P}{1-P}\right) = \omega^T x$$

Inverse of logit equation by P

$$P(y = 1|x) = \frac{1}{1+exp(-\omega^T x)}$$

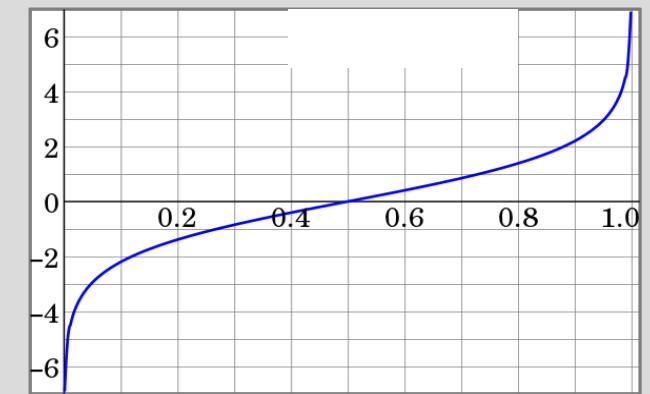
Abbreviated notation as sigmoid function

$$P(y = 1|x) = \sigma(\omega^T x)$$

Last step for binary classification: Rounding to 0 or 1

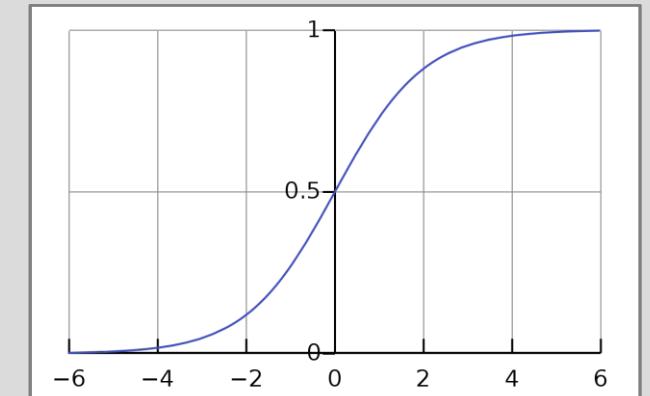
Logit equation

$$\ln\left(\frac{P}{1-P}\right) = \omega^T x$$



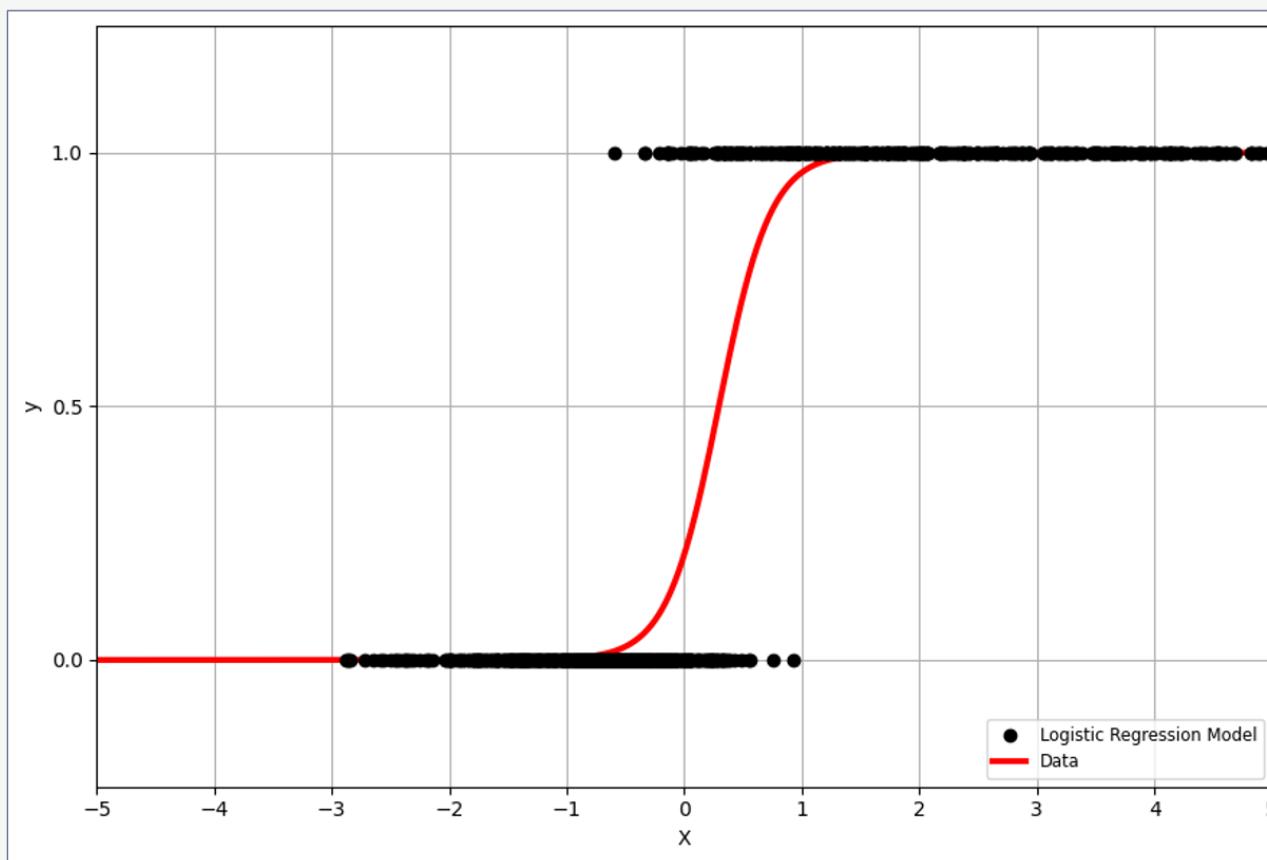
**Sigmoid function:
Inverse of logit equation**

$$P(y = 1|x) = \sigma(\omega^T x)$$



Logistic regression

... is a statistical model to predict a binary outcome (0 or 1, true or false) based on one or more predictor variables, using the logistic function to model the probability of the binary outcome.



Logistic function (= sigmoid function):

$$P(z) = \frac{1}{1+e^{-z}} \text{ with } z = \beta_0 + \beta_1 x$$

- It converts any input into a value between 0 and 1, which can be interpreted as probability.
- For given input 'x', output represents probability that 'x' belongs to 1
- Simple rounding can be used for predicting 0 or 1

It is based on ...

Logged odds:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

→ Linear model is projected on S-curve

Logistic regression

... is statistical model to predict a binary outcome (0 or 1, true or false) based on one or more predictor variables, using logistic function to model the probability of binary outcome.

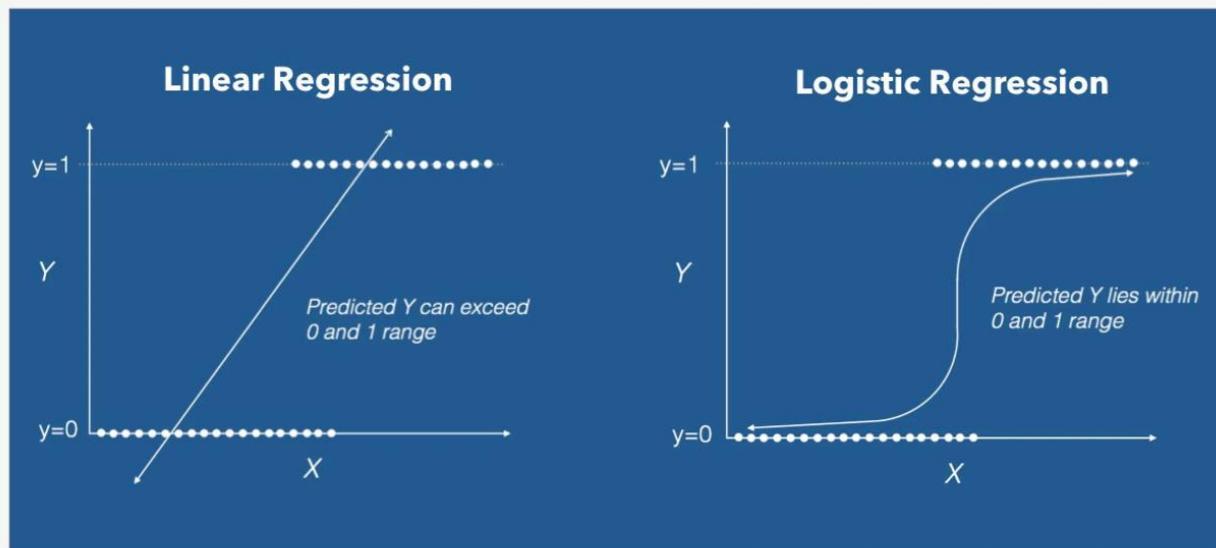


Fig 12: Typical logistic regression curve

Logistic function (= sigmoid function):

$$P(z) = \frac{1}{1+e^{-z}} \text{ with } z = \beta_0 + \beta_1 x$$

- It converts any input into a value between 0 and 1, which can be interpreted as a probability.
- For given input 'x', the output represents probability that 'x' belongs to class 1 (usually coded as positive outcome).
- Simple rounding can be used for predicting 0 or 1

It is based on ...

Logged odds:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

→ Linear model is projected on S-curve

<https://medium.com/@cmukesh8688/logistic-regression-sigmoid-function-and-threshold-b37b82a4cd79>

Backpropagation (of errors)

... is general optimization approach for minimizing error in supervised or reinforcement learning.
It calculates gradient of error function with respect to each weight by using chain rule.

Process

1. Feedforward Pass:

Input a dataset and compute the predicted output using current weights.

2. Compute Loss:

Calculate error, i.e. difference: predicted output \leftrightarrow actual target.

3. Backward Pass:

Propagate loss backward through network, calculating gradients for each weight.

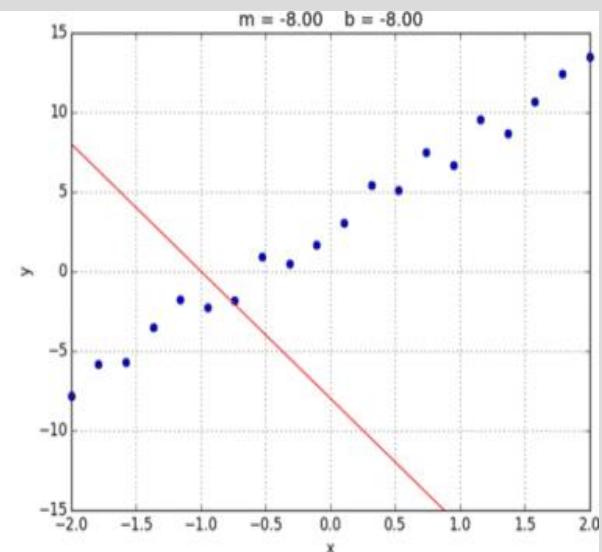
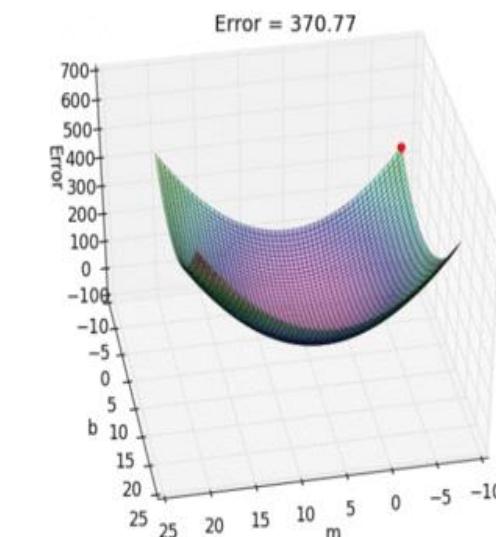
4. Weight Update (using Gradient Descent):

Once computed gradients are used to update weights of network.
Aim: to adjust weights in direction that minimizes error.

5. Iterate 1-4

Challenges

- Can get stuck in local minima
(mitigated by variants like stochastic gradient descent).
- Sensitive to hyperparameters and initial weight values.



Fit the model with L-BFGS in Sklearn

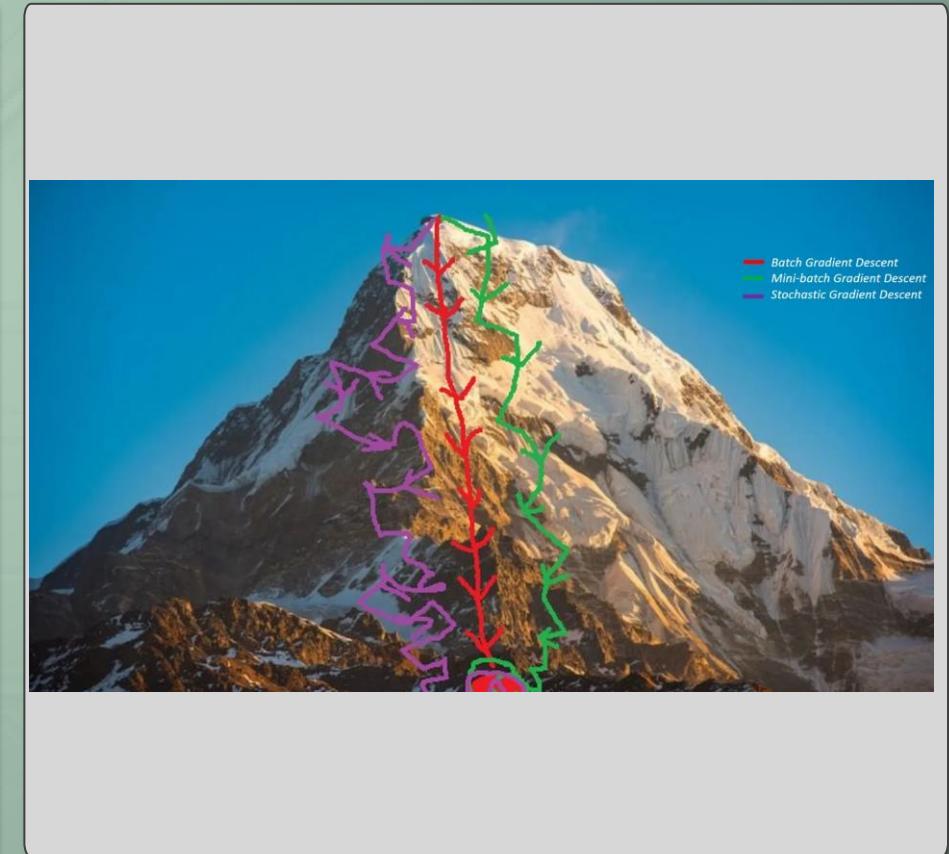
It looks for best parameters by navigating the cost function landscape, aiming to find the lowest point (global minimum)

Overview

- Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm.
- Iterative method for solving large-scale nonlinear optimization problems.
- Standard solver for logistic regression models in Python / sklearn

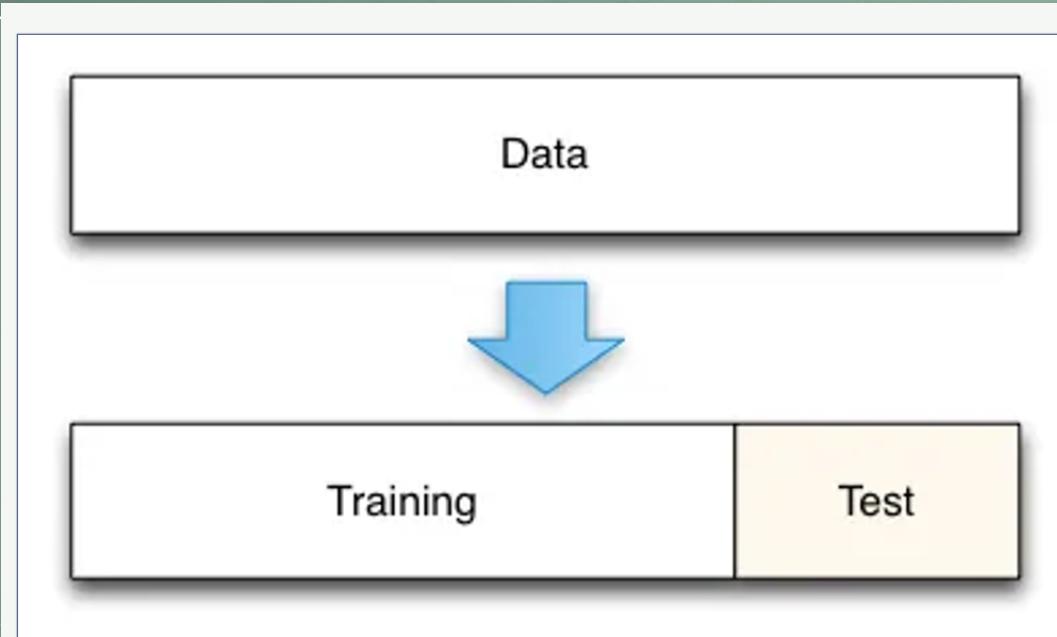
How L-BFGS Works

- **Approximation:**
It approximates Broyden-Fletcher-Goldfarb-Shanno (BFGS) Hessian matrix (particular second derivation matrix) using limited amount of memory.
- **Updates:**
Utilizes previous gradient evaluations to construct approximation, which is then used to determine direction of next step.
- **Iteration:**
Each iteration consists of line search followed by calculation of next approximation.



ML perspective: Train test split

... is a method where the dataset is divided into subsets: the training set, used to train the model, and the testing set, used to evaluate the model's performance on unseen data



Training models vs. analytical solution

- Logistic regression does not have analytical solution in opposite to linear regression.
- Therefore squared error minimizing parameters are approximated, e.g. with Newton's approximation method
- However: Danger of overfitting

Training Set:

- Portion of data used to create the model, i.e. to learn/approximate the parameters
- Usually 70-80% of the entire dataset

Testing Set:

- Subset of data that the model has not seen during learning phase
- Used to evaluate performance of the model on unseen data and check for overfitting. Typically 20-30% of entire dataset

<https://medium.com/@rathinavel.mph/how-the-train-and-test-samples-are-split-d7e46a8e2361>

Statistics perspective: Summary for Logistic regression

... provides a comprehensive summary of the results from fitted logistic regression model

# Print the summary statistics of the regression model print(result.summary())									
Logit Regression Results									
Dep. Variable:		default	No. Observations:	24000					
Model:		Logit	Df Residuals:	23975					
Method:		MLE	Df Model:	24					
Date:		Wed, 05 Jul 2023	Pseudo R-squ.:	0.1211					
Time:		21:36:53	Log-Likelihood:	-11162.					
converged:		True	LL-Null:	-12700.					
Covariance Type:		nonrobust	LLR p-value:	0.000					
		coef	std err	z	P> z	[0.025 0.975]			
		const	-1.4592	0.018	-79.195	0.000 -1.495 -1.423			
		x1	-0.0039	0.017	-0.233	0.815 -0.037 0.029			
		x2	-0.1064	0.023	-4.677	0.000 -0.151 -0.062			
		x3	-0.0561	0.017	-3.344	0.001 -0.089 -0.023			
		x4	-0.0774	0.018	-4.194	0.000 -0.114 -0.041			
		x5	-0.0791	0.018	-4.290	0.000 -0.115 -0.043			
		x6	0.0768	0.018	4.203	0.000 0.041 0.113			
		x7	0.6476	0.022	29.092	0.000 0.604 0.691			
		x8	0.1086	0.027	4.008	0.000 0.055 0.162			
		x9	0.0776	0.030	2.548	0.011 0.018 0.137			
		x10	0.0532	0.033	1.633	0.103 -0.011 0.117			
		x11	0.0205	0.034	0.605	0.546 -0.046 0.087			
		x12	0.0163	0.028	0.579	0.562 -0.039 0.072			
		x13	-0.3759	0.092	-4.087	0.000 -0.556 -0.196			
		x14	0.1462	0.118	1.240	0.215 -0.085 0.377			
		x15	0.1065	0.101	1.050	0.294 -0.092 0.305			

coef:

- Input factors that have impact and should be used

The closer to 1 the better
The closer to 0 the better

If LLR p-value is small (typically less than a pre-specified significance level, such as 0.05), full model provides significantly better fit to the data compared to reduced model.

How many zebras do you see?



10 out of 55 animals are zebras



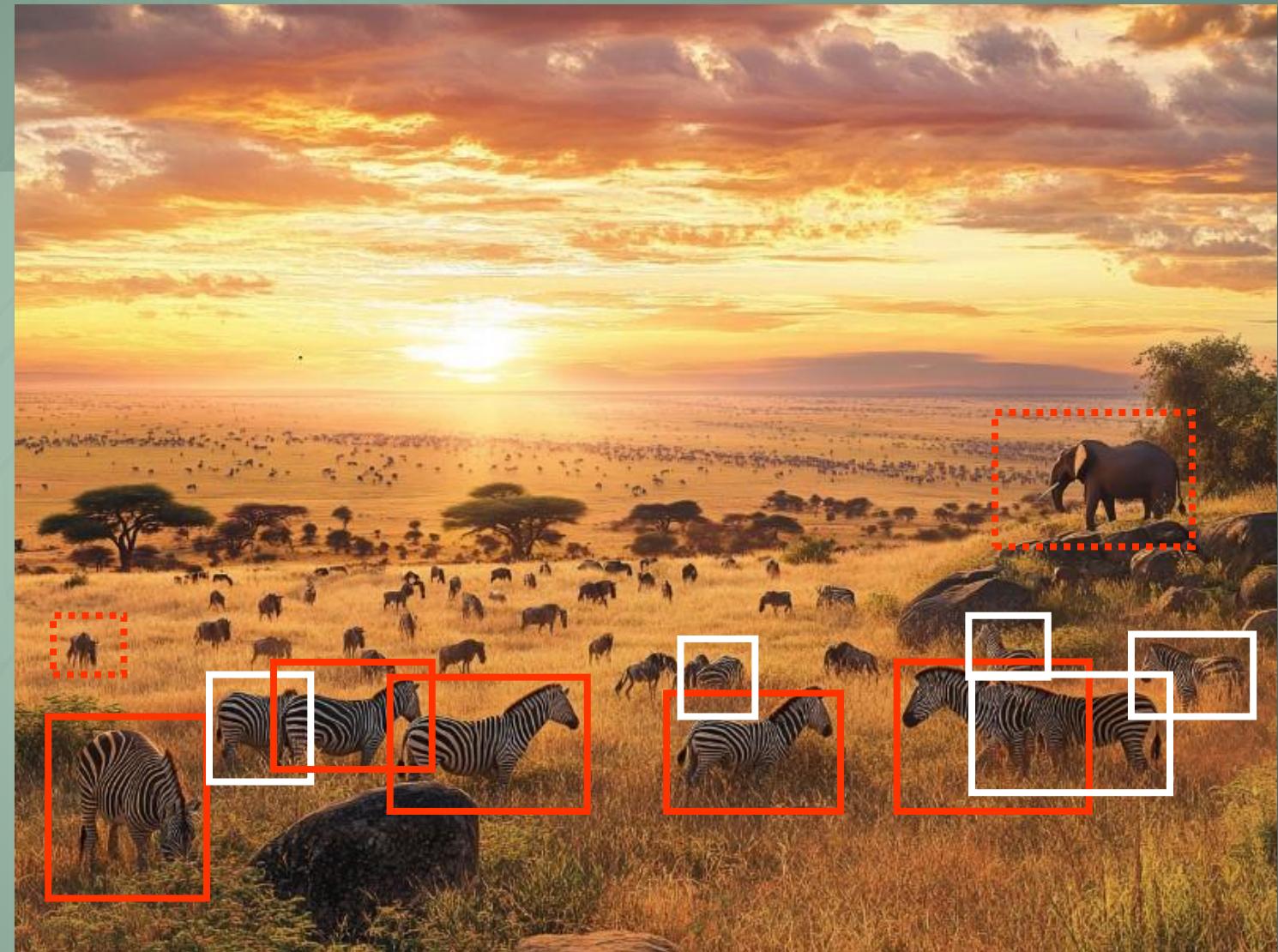
Example: Model marks 7 animals as zebras, but makes mistakes



Model marks 7 animals as zebras, but makes mistakes

Confusion Matrix:

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43



Confusion Matrix

.. Is a table used to evaluate classification model performance by comparing predicted outcomes with actual results.

		Actual		
		Total	Positive	Negative
Predicted	Total	Total number of items	Total of actual Positive	Total of actual Negative
	Positive	Total of predicted Positive	True Positive	False Positive (type I error)
	Negative	Total of negative Positive	False (Incorrect) Negative (type II error)	True Negative



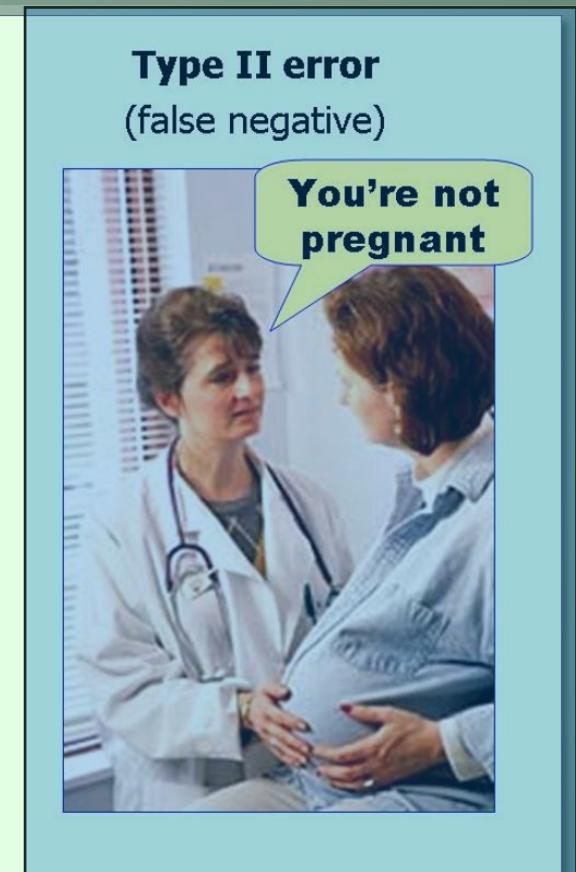
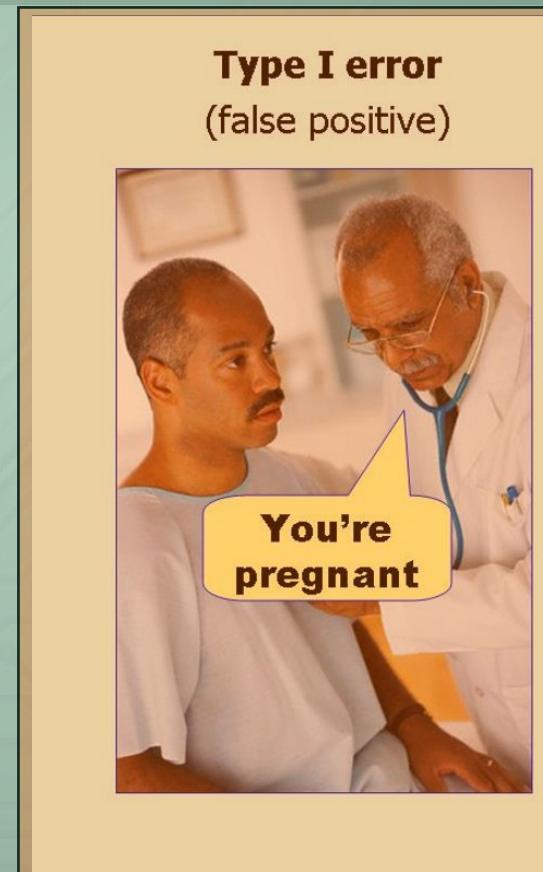
Mistakes or errors: False positives and false (incorrect) negatives

False positives and false negatives

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43

False positive

False negative



Evaluation metrics for classification

... ensure fair and thorough comparison of trained classification models

Key Metrics

Accuracy:

Overall correctness of predictions in %

Precision:

Correctly identified entities out of all predicted entities in %.

From all predicted items, how many are correct?

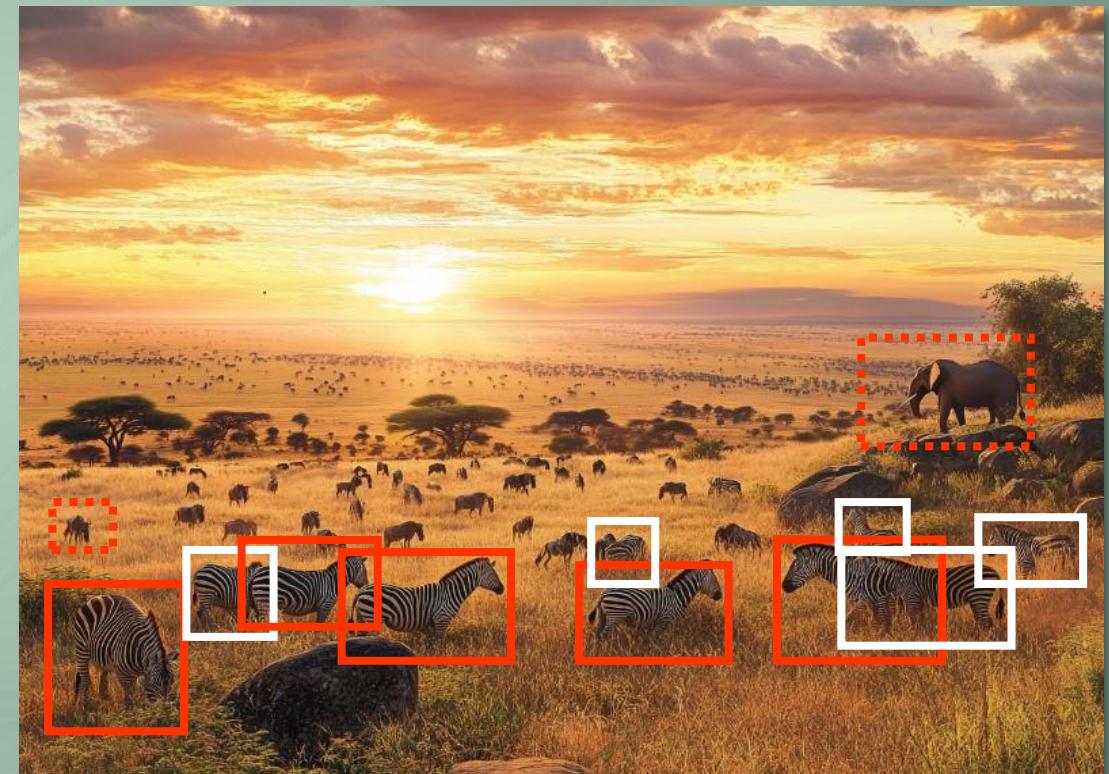
Recall Pass:

True positives identified out of all actual positives in %.

From all positive items, how many have been identified as positive?

F1-Score:

- Provides balance between precision and recall
- by calculating their harmonic mean



Fallacy of “accuracy” as evaluation metric of classification

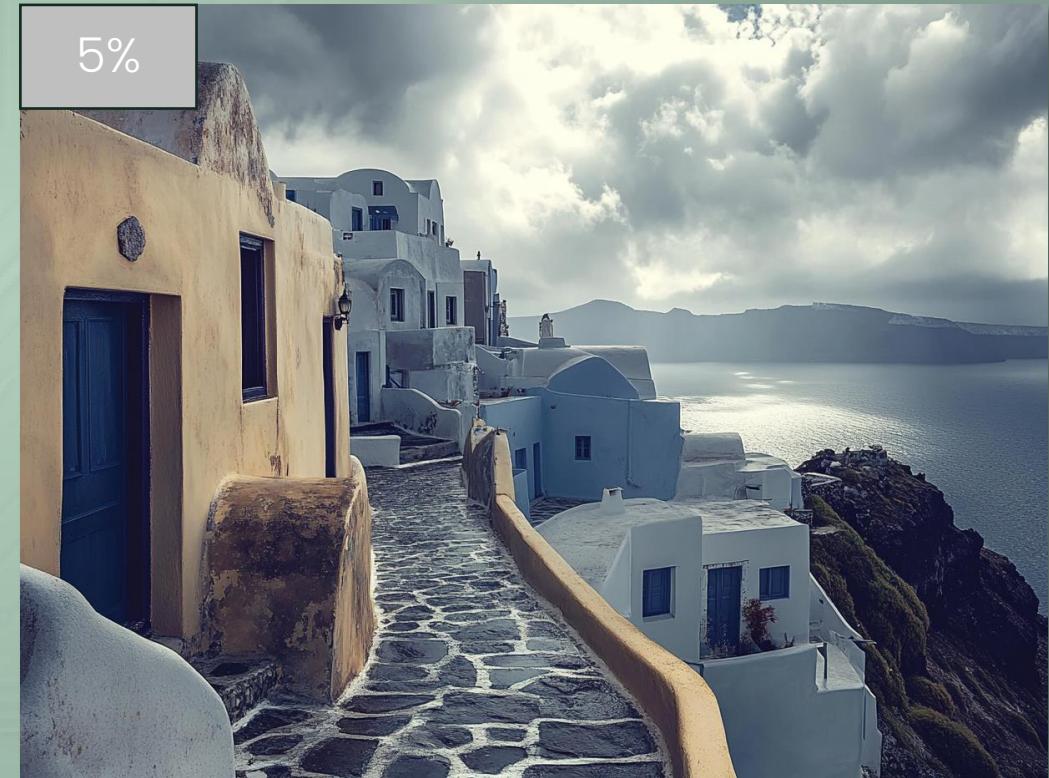
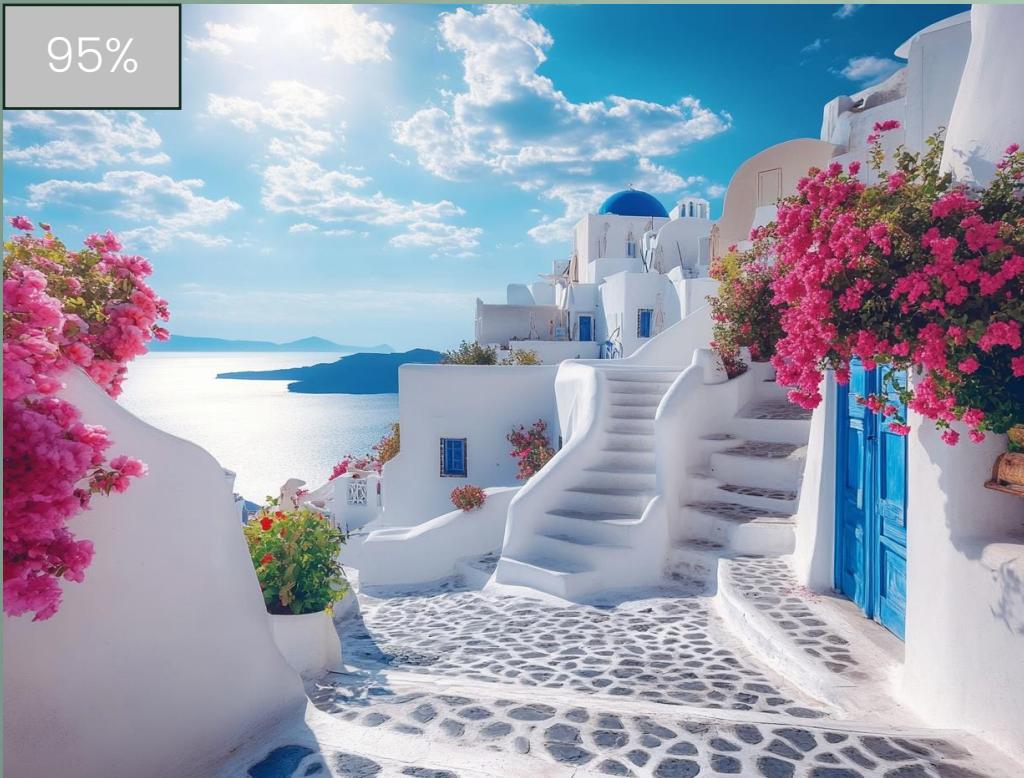
It's **July** in Greece, model predicts sun (often, 95%) or rain (rare, 5%) for next day.

It predicts sun (often). Next day this takes place.

It predicts rain (rare). Next day this takes place.

Was this prediction good or bad?

Was this prediction good or bad?



Model that always predicts sun has accuracy = 95%.
→ Therefore, baseline for accuracy must be at least 95%

Confusion matrix metric: Accuracy

How many items were classified correctly out of all items?

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43

$$\text{Accuracy} = \frac{\text{Correct Queries}}{\text{Total Queries}} \times 100$$

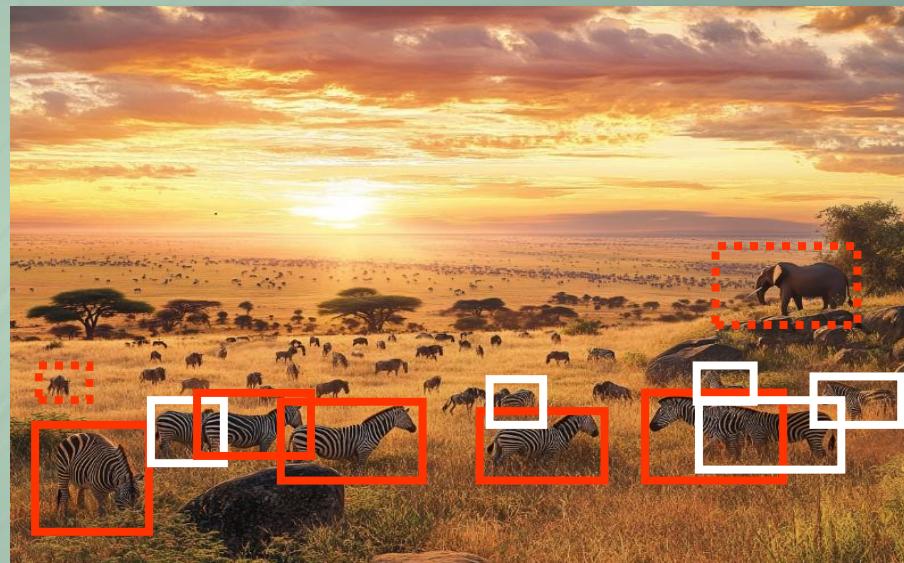


$$\text{Accuracy}(\%) = \frac{5 + 43}{55} = 87\%$$

Confusion matrix metric: Recall

Measures the proportion of actual positives correctly identified by the model, i.e. completeness of prediction. Out of all positive items, how many were predicted correctly (in %)?

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43



$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Precision}(\%) = \frac{5}{5+5} = 50\%$$

"Recall" is also called "Sensitivity", they are the same.

Confusion matrix metric: Precision

From all predicted items, how many are actually positive (in %)?

Or: From all as positive predicted items, how many are actually positive (in %)?

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43



$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Precision}(\%) = \frac{5}{5+2} = 71\%$$

Confusion matrix metric: F1-Score

... is the harmonic mean of precision and recall

Balances the trade-off between precision and recall

		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



$$F1 (\%) = \frac{2 \times 0.71 \times 0.5}{0.71 + 0.5} = \frac{0.71}{1.21} = 59\%$$

Confusion matrix metric: Interpreting F1-Score

... Let's finds balance between precision and recall,
especially useful for imbalanced datasets.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Metric	Definition	Focus	Strengths	Weaknesses
Precision	<ul style="list-style-type: none">▪ Proportion of correctly predicted positives▪ out of all predicted positives.	<ul style="list-style-type: none">▪ Accuracy of positive predictions	<ul style="list-style-type: none">▪ Minimizes false positives▪ useful in scenarios where false positives are costly▪ e.g., spam detection	<ul style="list-style-type: none">▪ May miss true positives,▪ leading to low recall.
Recall	<ul style="list-style-type: none">▪ Proportion of actual positives▪ correctly identified by the model.	<ul style="list-style-type: none">▪ Completeness of positive predictions.	<ul style="list-style-type: none">▪ Minimizes false negatives▪ critical in scenarios where missing true positives is costly▪ e.g., disease diagnosis	<ul style="list-style-type: none">▪ May increase false positives▪ leading to low precision.
F1-Score	<ul style="list-style-type: none">▪ Harmonic mean of precision and recall;▪ balances the two metrics.	Balance between precision and recall.	<ul style="list-style-type: none">▪ Useful for imbalanced datasets▪ or when both false positives and false negatives are important.	Does not distinguish between cases where precision or recall is more critical.

Exercise: Predicting Credit Card Default

Expected Outcome: Trained logistic regression model capable of predicting credit card defaults with interpretable results

Exercise Overview

- **Objective:** Build a logistic regression model to predict whether a client will default on their credit card payment (Yes = 1, No = 0)
- **Dataset:** Default of Credit Card Clients dataset (2005).

Key Variables

- Response Variable (Y): Default payment status (1 = Yes, 0 = No).
- Explanatory Variables:
 - X1: Credit limit (NT dollars)
 - X2-X5: Demographics (Gender, Education, Marital Status, Age)
 - X6-X11: Past repayment history (-1 = on time, 1-9 = months delayed)
 - X12-X17: Bill statement amounts (April–September 2005)
 - X18-X23: Previous payments made (April–September 2005)



Predicting Credit Card Default

Dataset:

Credit limit ID	Demographics		Past repayment history										Bill statement amounts						Previous payments made						Response Variable	
	X1 LIMIT_BAISEX	X2 EDUCATI	X3 MARRIA	X4 CAGE	X5	X6 PAY_0	X7 PAY_2	X8 PAY_3	X9 PAY_4	X10 PAY_5	X11 PAY_6	X12 BILL_AM	X13 BILL_AM	X14 BILL_AM	X15 BILL_AM	X16 BILL_AM	X17 BILL_AM	X18 BILL_AM	X19 PAY_AM	X20 PAY_AM	X21 PAY_AM	X22 PAY_AM	X23 PAY_AM	Y default payment next month		
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	0	1	
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	0	1	
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0	0	
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0	0	
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0	0	
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0	0	
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0	0	
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0	0	
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0	0	
10	20000	1	3	2	35	-2	-2	-2	-1	-1	0	0	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0	
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0	0	
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0	0	
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0	0	0	
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1	0	
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0	0	
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0	0	
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1	0	
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0	0	
19	360000	2	1	1	49	1	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	180000	2	1	2	29	1	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0	0	
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	0	632	316	316	0	1	0	
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1	0	
24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	0	1128	1	0
25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0	0	
26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0	0	
27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1	0	
28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0	0	
29	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0	0	
30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1600	0	0	0	0	
31	230000	2	1	2	27	-1	-1	-1	-1	-1	-1	16646	17265	13266	15339	14307	36923	17270	13281	15339	14307	37292	0	0	0	
32	50000	1	2	2	33	2	0	0	0	0	0	30518	29618	22102	22734	23217	23680	1718	1500	1000	1000	1000	716	1	0	
33	100000	1	1	2	32	0	0	0	0	0	0	93036	84071	82880	80958	78703	75589	3023	3511	3302	3204	3200	2504	0	0	
34	500000	2	2	1	54	-2	-2	-2	-2	-2	-2	10929	4152	22722	7521	71439	8981	4152	22827	7521	71439	981	51582	0	0	
35	500000	1	1	1	58	-2	-2	-2	-2	-2	-2	13709	5006	31130	3180	0	5293	5006	31178	3180	0	5293	768	0	0	
36	160000	1	1	2	30	-1	-1	-2	-2	-2	-1	30265	-131	-527	-923	-1488	-1884	131	396	396	565	792	0	0	0	

Predicting Credit Card Default

Expected Outcome: Trained logistic regression model capable of predicting credit card defaults with interpretable results

Steps to Complete the Exercise

1. Data Preprocessing:

Only if relevant:

Handle missing values, normalize numeric features, and encode categorical variables.

2. Model Training:

Use logistic regression to train model with explanatory variables.

3. Evaluation:

Assess model performance using metrics like

- Accuracy
- Precision & Recall
- F1-Score

4. Insights:

Identify key predictors of default and interpret coefficients for actionable insights.

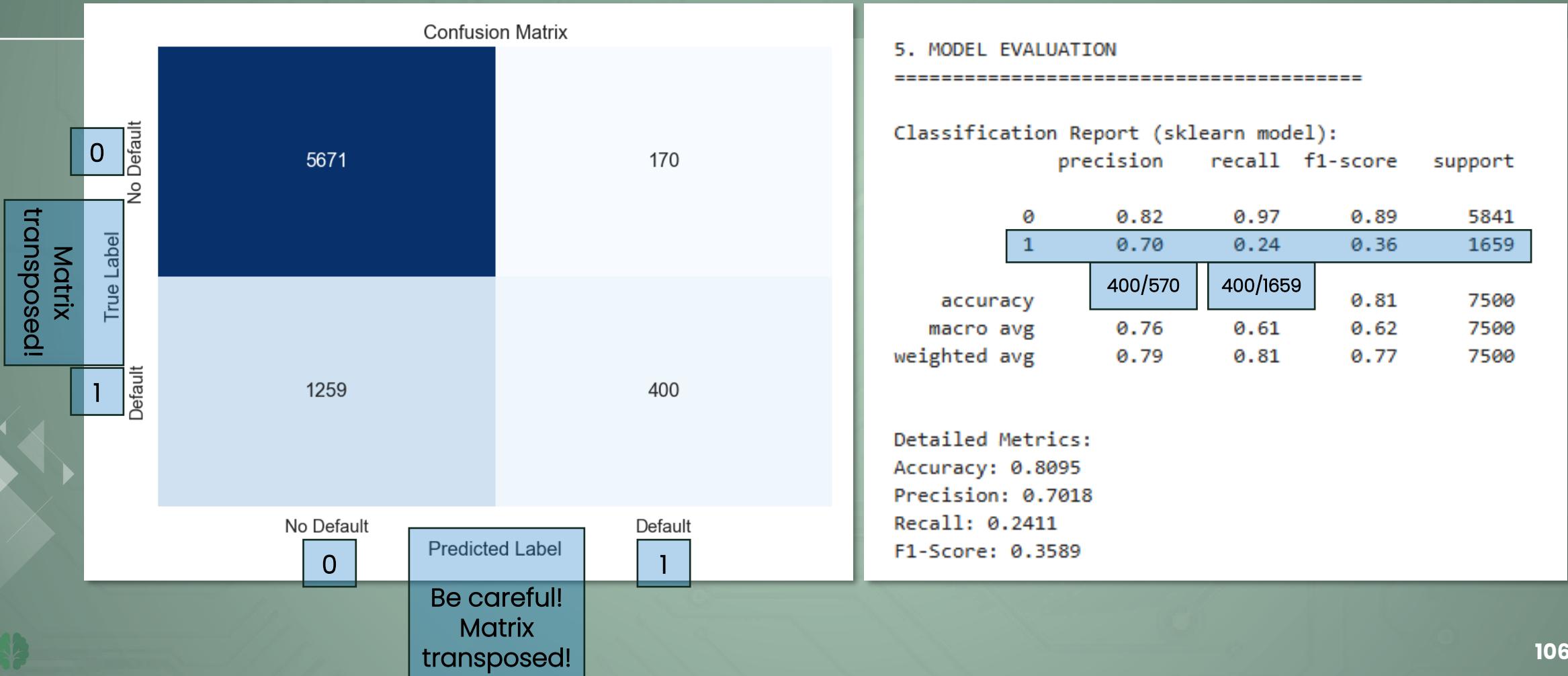
```
# Train our custom logistic regression model
print("\nTraining our custom LogisticRegression model...")
# Using a smaller number of iterations for time constraints
my_model = MyLogisticRegression(learning_rate=0.1, num_iterations=10000)
my_model.fit(X_train_scaled, y_train)
```

4. MODEL TRAINING: LOGISTIC REGRESSION

```
Training our custom LogisticRegression model...
Cost at iteration 500: 0.46402681019559955
Cost at iteration 1000: 0.46391625110907125
Cost at iteration 1500: 0.463881906608544
Cost at iteration 2000: 0.4638625289985032
Cost at iteration 2500: 0.46385059653172084
Cost at iteration 3000: 0.4638429819015393
Cost at iteration 3500: 0.4638379809509058
Cost at iteration 4000: 0.4638346091873176
Cost at iteration 4500: 0.46383228109929964
Cost at iteration 5000: 0.46383063939171343
Cost at iteration 5500: 0.463829460346282
Cost at iteration 6000: 0.4638286002434648
Cost at iteration 6500: 0.4638279644346782
Cost at iteration 7000: 0.4638274891188923
Cost at iteration 7500: 0.4638271303666043
Cost at iteration 8000: 0.4638268573547721
Cost at iteration 8500: 0.46382664809774654
Cost at iteration 9000: 0.463826486689604
Cost at iteration 9500: 0.4638263614823324
```

Predicting Credit Card Default: Evaluation

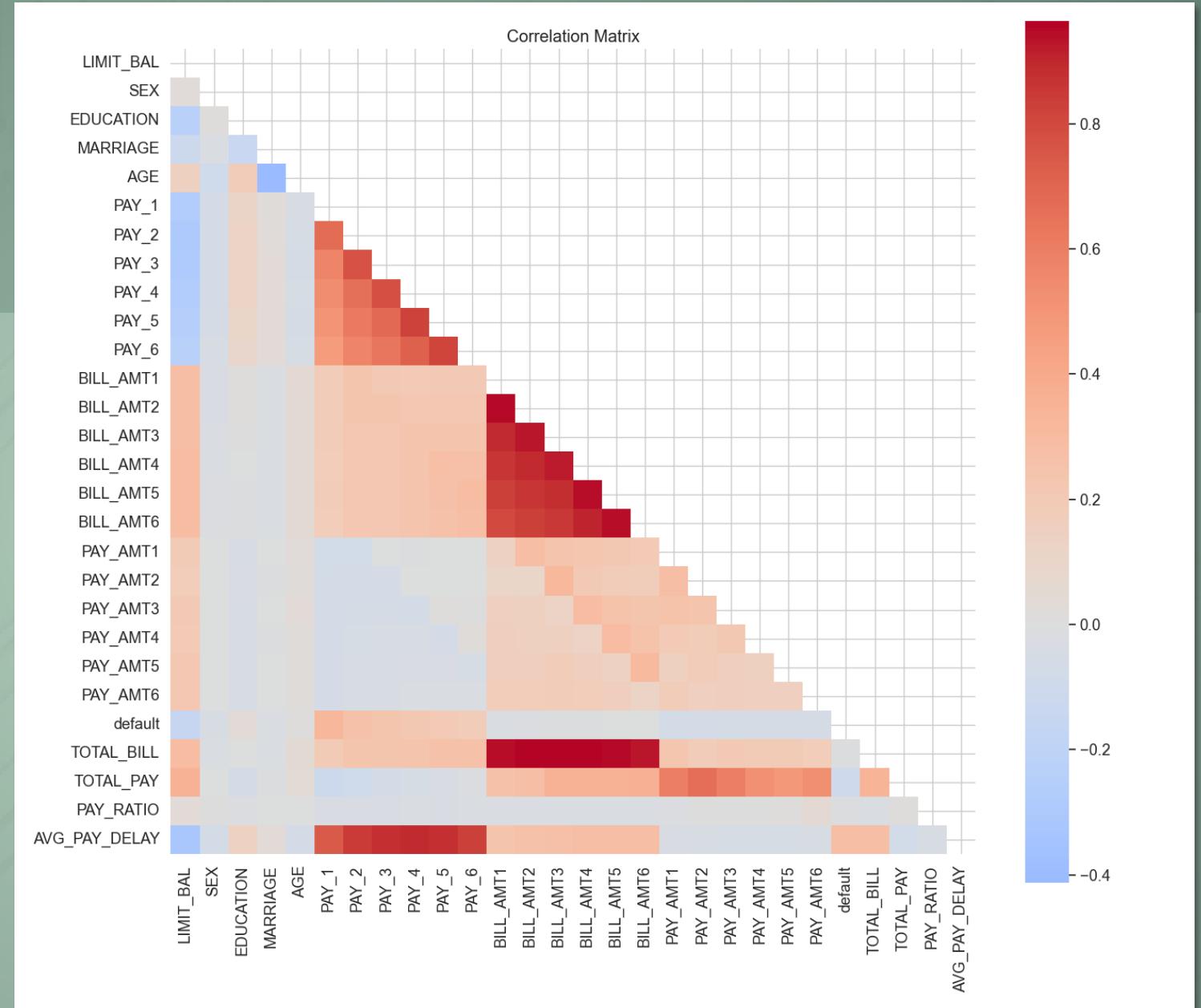
... with Confusion Matrix, accuracy, precision, recall, F1-Score
→ becomes relevant when comparing models



Correlation Matrix

Influence factors might also influence each other

- For unbiasedness of estimation high correlation between factors should be avoided.
- Therefore, part of factors should be removed from model.



Statistical tables

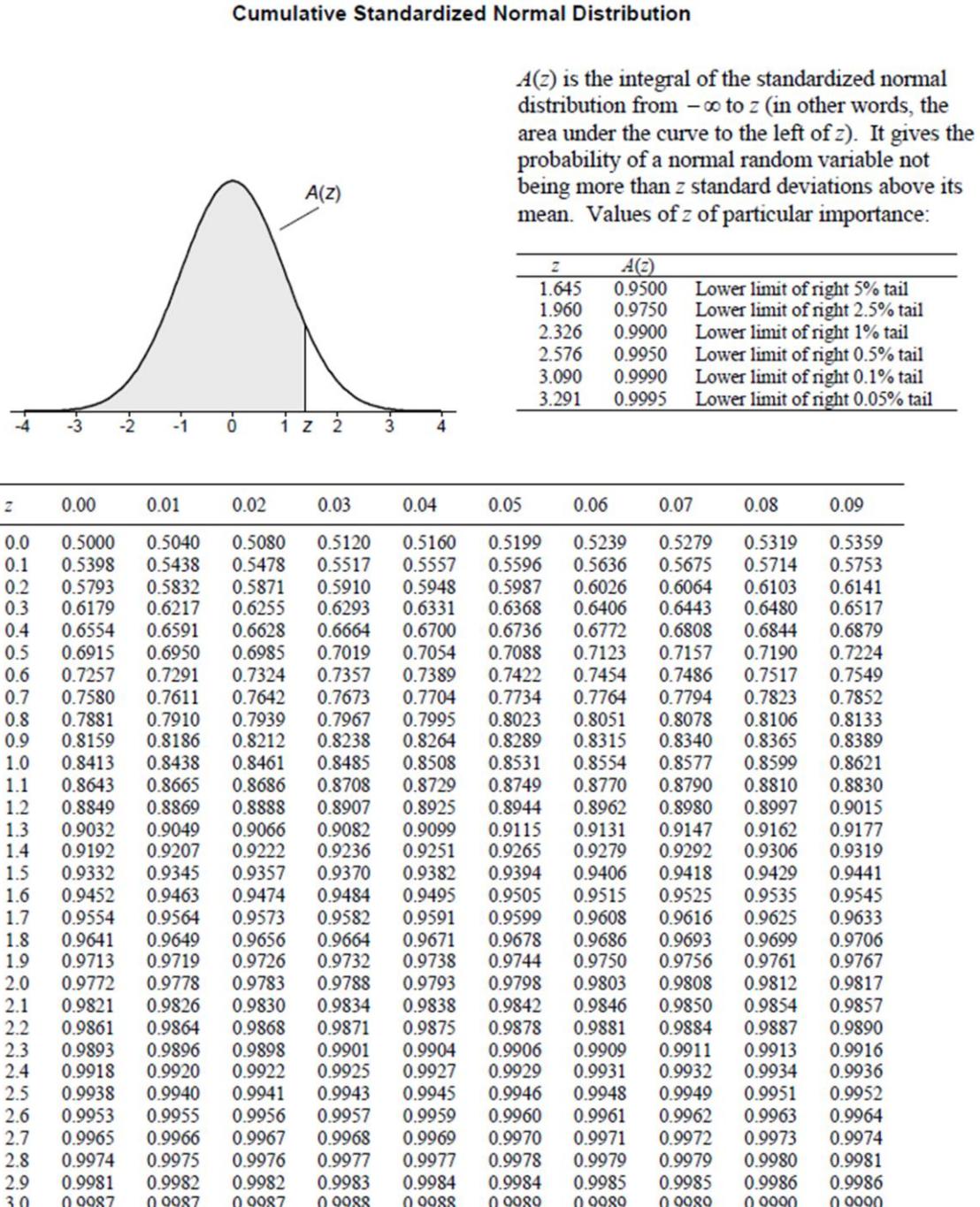
... let you decide whether to reject
 H_0 or not by hand

t Distribution: Critical Values of t

Degrees of freedom	Two-tailed test: One-tailed test:	Significance level						
		10%	5%	2.5%	2%	1%	0.2%	0.1%
1		6.314	12.706	31.821	63.657	318.309	636.619	
2		2.920	4.303	6.965	9.925	22.327	31.599	
3		2.353	3.182	4.541	5.841	10.215	12.924	
4		2.132	2.776	3.747	4.604	7.173	8.610	
5		2.015	2.571	3.365	4.032	5.893	6.869	
6		1.943	2.447	3.143	3.707	5.208	5.959	
7		1.894	2.365	2.998	3.499	4.785	5.408	
8		1.860	2.306	2.896	3.355	4.501	5.041	
9		1.833	2.262	2.821	3.250	4.297	4.781	
10		1.812	2.228	2.764	3.169	4.144	4.587	
11		1.796	2.201	2.718	3.106	4.025	4.437	
12		1.782	2.179	2.681	3.055	3.930	4.318	
13		1.771	2.160	2.650	3.012	3.852	4.221	
14		1.761	2.145	2.624	2.977	3.787	4.140	
15		1.753	2.131	2.602	2.947	3.733	4.073	
16		1.746	2.120	2.583	2.921	3.686	4.015	
17		1.740	2.110	2.567	2.898	3.646	3.965	
18		1.734	2.101	2.552	2.878	3.610	3.922	
19		1.729	2.093	2.539	2.861	3.579	3.883	
20		1.725	2.086	2.528	2.845	3.552	3.850	
21		1.721	2.080	2.518	2.831	3.527	3.819	
22		1.717	2.074	2.508	2.819	3.505	3.792	
23		1.714	2.069	2.500	2.807	3.485	3.768	
24		1.711	2.064	2.492	2.797	3.467	3.745	
25		1.708	2.060	2.485	2.787	3.450	3.725	
26		1.706	2.056	2.479	2.779	3.435	3.707	
27		1.703	2.052	2.473	2.771	3.421	3.690	
28		1.701	2.048	2.467	2.763	3.408	3.674	
29		1.699	2.045	2.462	2.756	3.396	3.659	
30		1.697	2.042	2.457	2.750	3.385	3.646	
32		1.694	2.037	2.449	2.738	3.365	3.622	
34		1.691	2.032	2.441	2.728	3.348	3.601	
36		1.688	2.028	2.434	2.719	3.333	3.582	
38		1.686	2.024	2.429	2.712	3.319	3.566	
40		1.684	2.021	2.423	2.704	3.307	3.551	

Statistical tables

... let you decide whether to reject H_0 or not by hand



Statistical tables

... let you decide whether to reject H_0 or not by hand

χ^2 (Chi-Squared) Distribution: Critical Values of χ^2

<i>Degrees of freedom</i>	<i>Significance level</i>		
	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

Statistical tables

... let you decide whether to reject
 H_0 or not by hand

TABLE A.3

F Distribution: Critical Values of F (5% significance level)

v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	
v_2	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03	
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01	
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97	
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96	
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93	
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84	
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75	
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72	
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70	
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69	
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68	
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66	
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64	

Links

i.e. sources for self-learning

	Title	Link
Descriptive statistics	Variance	https://www.youtube.com/watch?v=JgMFhKi6f6Y
	Quantiles	https://docs.eggplantsoftware.com/epp/9.4.0/analyzer/analyzer-understanding-charts-events.htm
	Python Data Visualization Tutorial	https://www.youtube.com/watch?v=Nt84_TzRkbo

	Title	Link
Combinatorics	Introduction to Permutations and Combinations	https://www.youtube.com/watch?v=gAnKvHmrJ0g
	Miracles of Pascal's triangle	https://www.youtube.com/watch?v=J0iINuxUcpQ

Links

i.e. sources for self-learning

	Title	Link
Distributions	Uniform probability distributions	https://www.youtube.com/watch?v=aCW8wm6nrRw
	Binomial distribution	https://www.youtube.com/watch?v=6YzrVUV09M0
	Chi^2	https://www.youtube.com/watch?v=JCOeytq7F0E

	Title	Link
Conditional probability	Bayes' Theorem of Probability With Tree Diagrams & Venn Diagrams	https://www.youtube.com/watch?v=OByl4RJxnKA
	What is the Bayes' Theorem?	https://medium.com/mlearning-ai/what-is-the-bayes-theorem-545a2ef0b91c
	Data Science. Bayes theorem	https://luminousmen.com/post/data-science-bayes-theorem
	Bayesian Machine Learning in Python: A/B Testing	https://www.udemy.com/course/bayesian-machine-learning-in-python-ab-testing

Links

i.e. sources for self-learning

	Title	Link
Sampling	Central limit theorem	https://www.geeksforgeeks.org/python-central-limit-theorem
	Confidence intervals	https://data.library.virginia.edu/the-intuition-behind-confidence-intervals/
	Resampling	https://www.linkedin.com/pulse/resampling-methods-pranshu-jaryal
	Bootstrapping	https://www.youtube.com/watch?v=9IFcRTyhd5Y
		https://dgarcia-eu.github.io/SocialDataScience/2_SocialDynamics/025_Bootstrapping/Bootstrapping.html
		https://janhove.github.io/teaching/2016/12/20/bootstrapping
	Title	Link
Hypothesis testing	Bayesian Machine Learning in Python: A/B Testing	https://www.udemy.com/course/bayesian-machine-learning-in-python-ab-testing
	Introduction to the Chi-square Test	https://www.youtube.com/watch?v=SvKv375sacA
	F-ratio Test for Two Equal Variances	https://www.youtube.com/watch?v=UWQO4gX7-IE
	17 Statistical Hypothesis Tests in Python (Cheat Sheet)	https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/
	Chi Squared Testing	https://www.youtube.com/watch?v=qYOM083ZIWU

Links

i.e. sources for self-learning

	Title	Link
Model estimation	Scatter matrix , Covariance and Correlation Explained	https://medium.com/@raghavan990/scatter-matrix-covariance-and-correlation-explained-14921741ca56
	Logistic regression sigmoid function and threshold	https://medium.com/@cmukesh8688/logistic-regression-sigmoid-function-and-threshold-b37b82a4cd79
	Introduction to logistic regression	https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148#:~:text=Logistic%20regression%20transforms%20its%20output,to%20return%20a%20probability%20value.
	Linear Regression Diagnostics	https://www.youtube.com/watch?v=HLAglyBfNk8&list=PLlbbWgBRF8EePgK40-i7aGU2_kylyujgL&index=11
	Machine Learning Classifier evaluation using ROC and CAP Curves	https://towardsdatascience.com/machine-learning-classifier-evaluation-using-roc-and-cap-curves-7db60fe6b716
	Classification Model Performance Evaluation using AUC-ROC and CAP Curves	https://medium.com/geekculture/classification-model-performance-evaluation-using-auc-roc-and-cap-curves-66a1b3fc0480
	Deep Learning Prerequisites: Logistic Regression in Python	https://www.udemy.com/course/data-science-logistic-regression-in-python/
		https://en.wikipedia.org/wiki/Correlation
	Ordinary Least Squares	https://gregorygundersen.com/blog/2020/01/04/ols/
	Scatter matrix , Covariance and Correlation Explained	https://medium.com/@raghavan990/scatter-matrix-covariance-and-correlation-explained-14921741ca56



About me

Dr. Harald Stein

- Data Scientist ~ 9 years experience
 - Algotrader ~ 4 years experience
 - Ph.D. in Economics, Game Theory
-
- LinkedIn: <https://www.linkedin.com/in/harald-stein-phd-1648b51a>
 - ResearchGate: <https://www.researchgate.net/profile/Harald-Stein>

