Introduction to Probability & Statistics

Dr. Harald Stein

April 2025

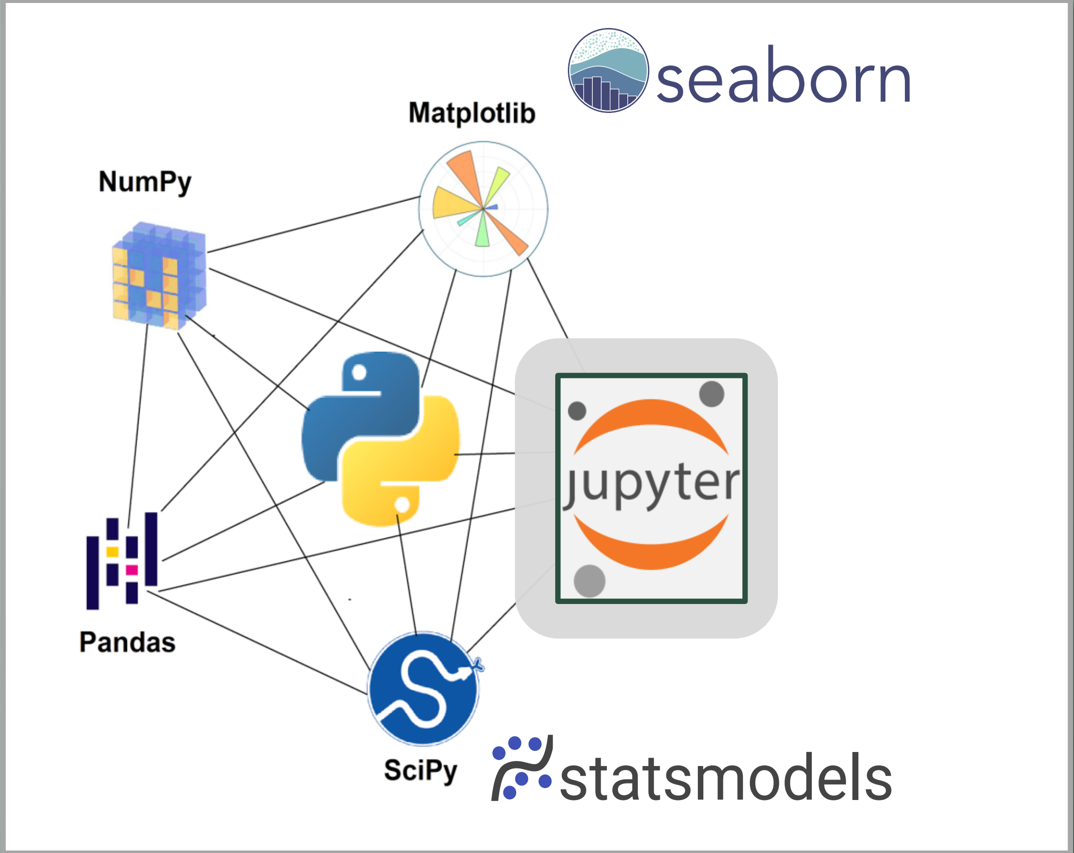
Content

- **Technical Prerequisites**
- **Probability, Statistics and Information Theory**
- **Descriptive Statistics**
- **Combinatorics**
- **Joint and Conditional Probability**
- **Probability Distributions**
- **Structured Probabilistic (Graphical) Models**
- **Sampling, Monte Carlo**
- **Hypothesis Testing**
- **Model Estimation:**
  - Linear Regression
  - Binary Logistic Regression

Technical Prerequisites

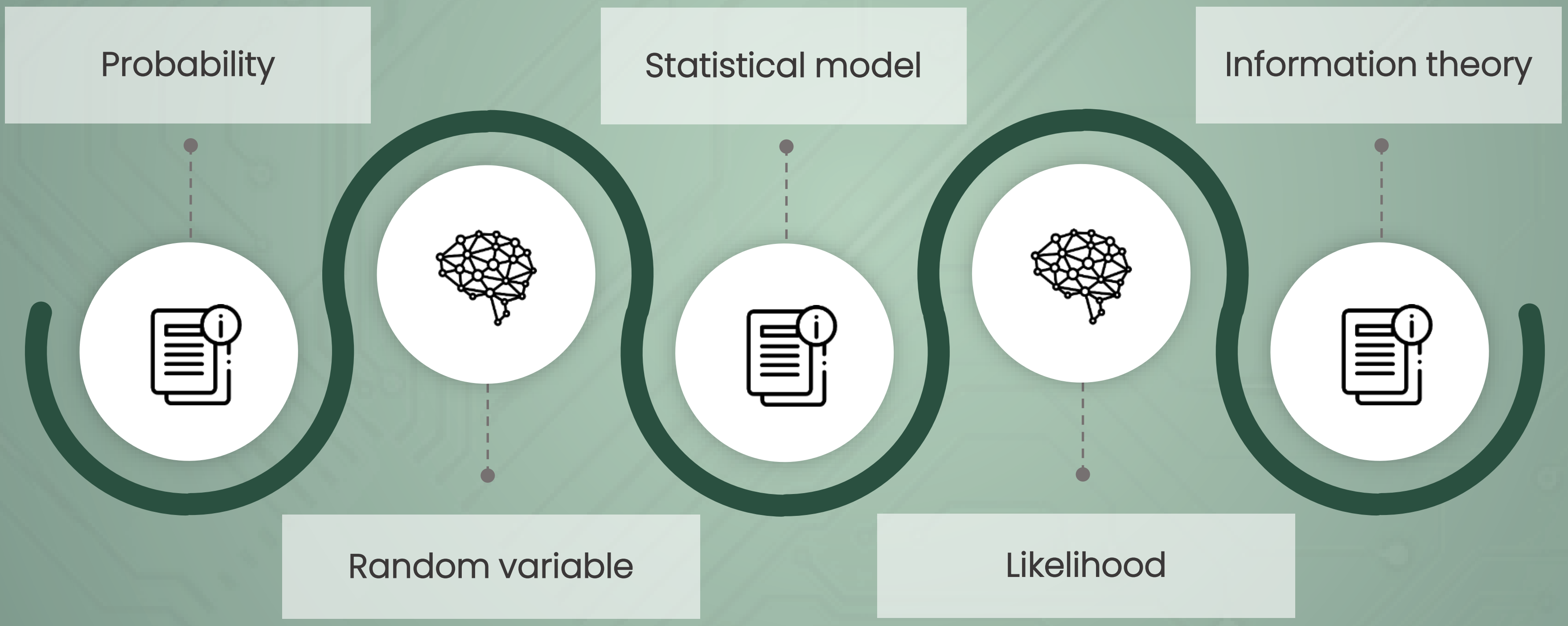
… means installation of Python environment with Jupyter

- 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

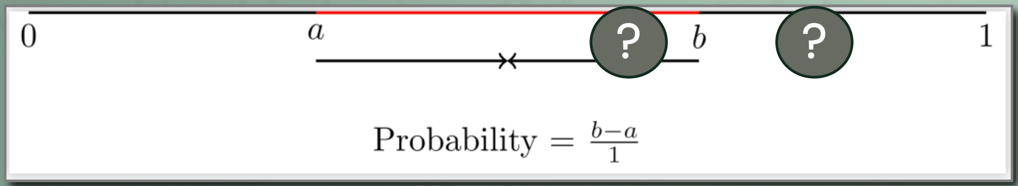
- Mathematics, Statistics, Finance, Gambling
- Science (including Physics)
- Artificial Intelligence/Machine Learning
- Computer Science, Game Theory, Philosophy

Purpose of Probability

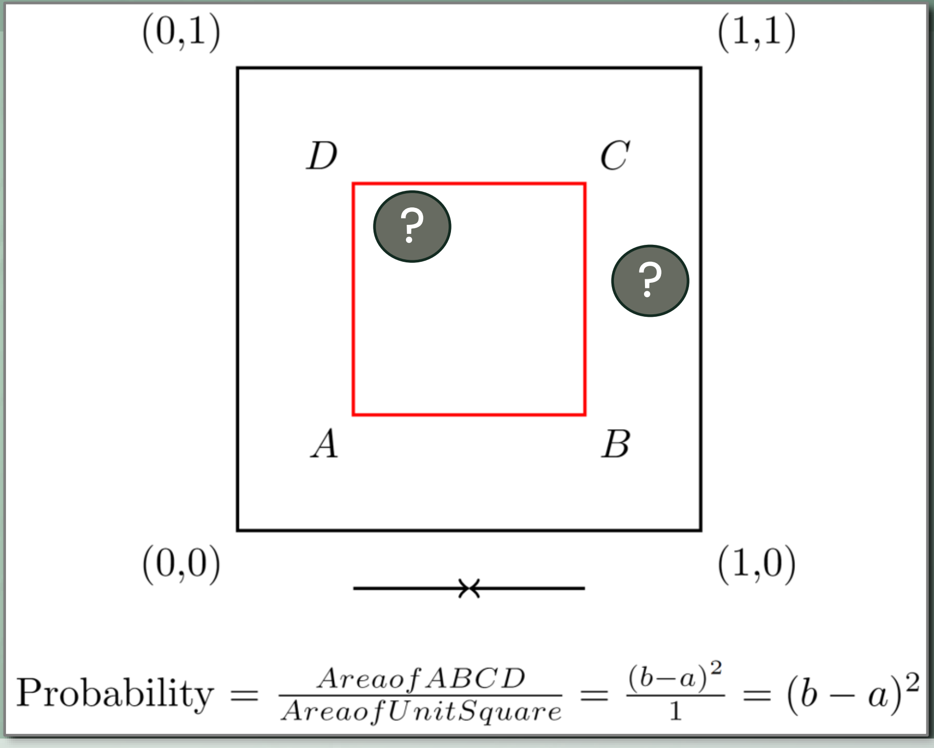
- Used to infer the expected frequency of events.
- Describes the underlying mechanics and regularities of complex systems.

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 $\frac{b-a}{1}$



- 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, ...}).	Any value within a continuous range (e.g., [0, ∞), (-∞, ∞)).
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).

https://towardsdatascience.com/understanding-random-variable-a618a2e99b93

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 $L(\theta \mid X)$, 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.

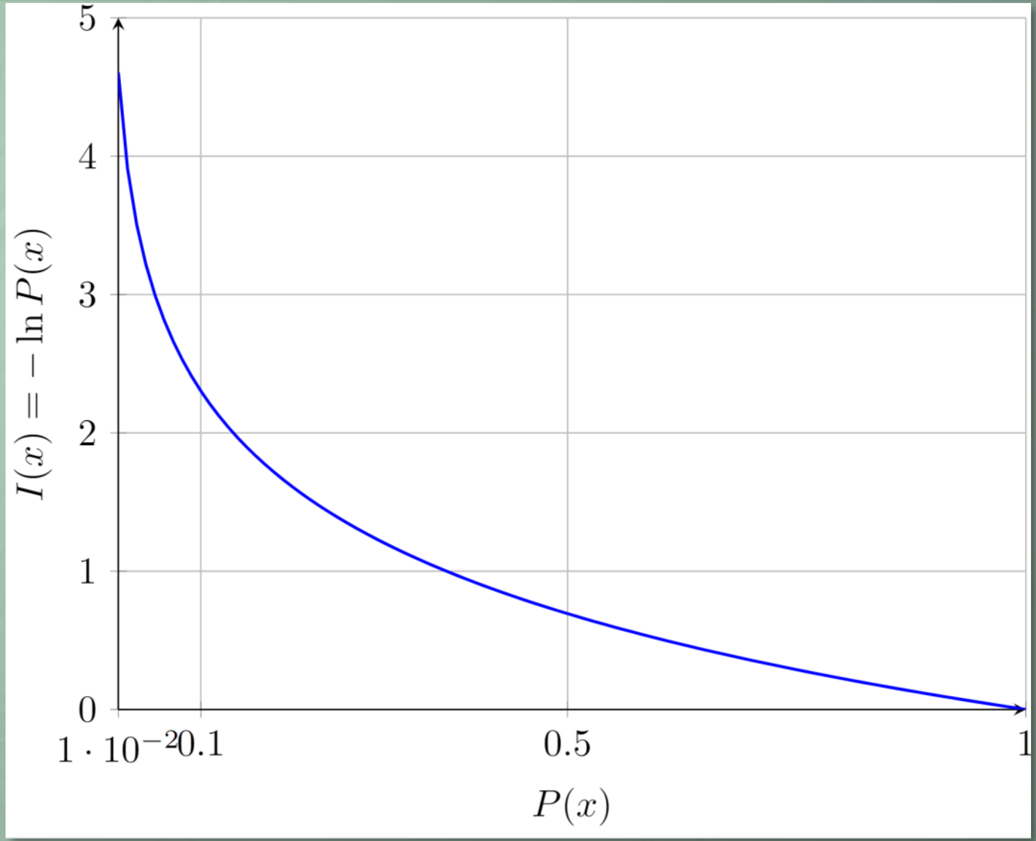
Key Properties of Information:

- 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

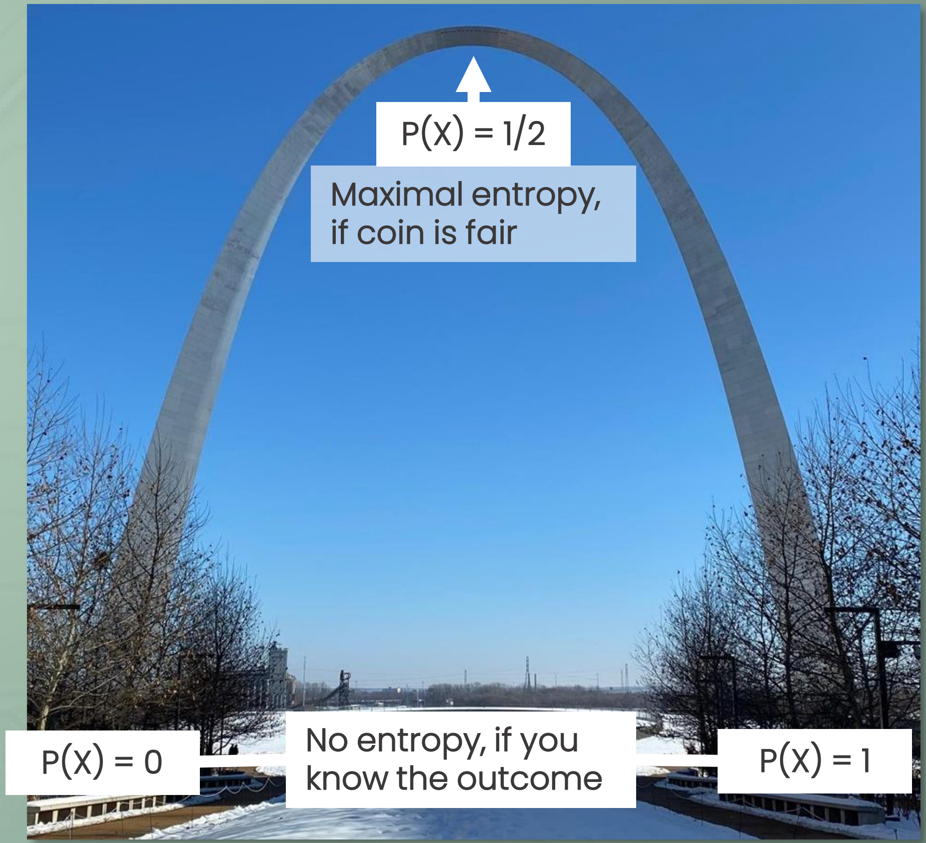
... is measure of randomness in set of possible outcomes. It quantifies expected amount of information/surprise, that event drawn from a probability distribution will provide.

- 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

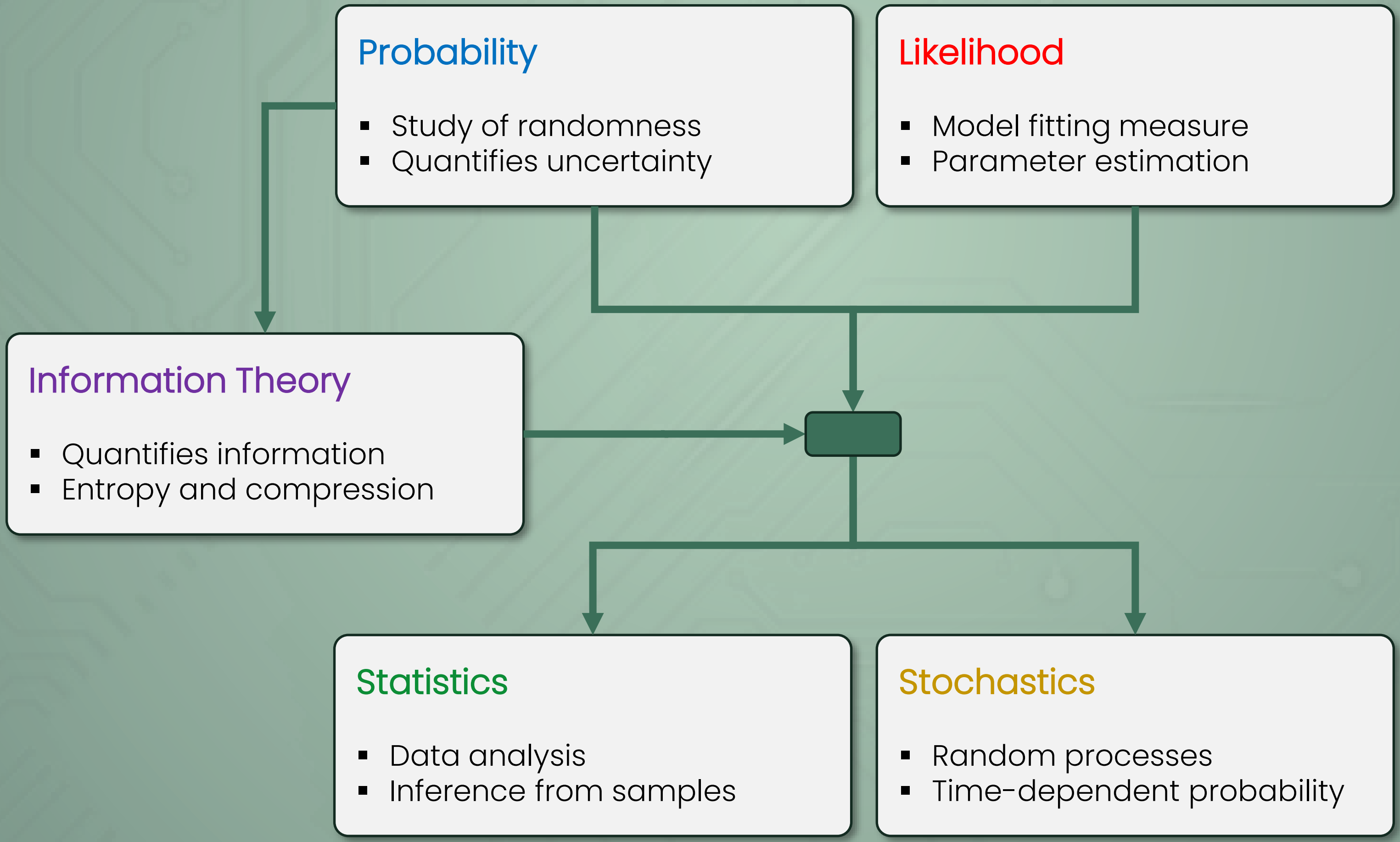
Outliers can increase the entropy if they represent rare but possible outcome, adding to the distribution’s unpredictability.

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
Sample vs. Population

Measures: central tendency, dispersion

Measures: Skewness and kurtosis

Quantiles

Graphical representation of data

Exercises

Types of Data

Definition and examples of Qualitative and Quantitative data

Type of Data	Subtype of data	Example
Quantitative / Numerical: • can be measured • or counted	**Discrete:** • can only take certain values	• number of students in a class
	Continuous: • can take any value within certain range	• Age, Height, Weight, Temperature, Income
	Interval	• Temperature • IQ scores • Years
	Ratio	• Probability
Qualitative / Categorical: • non-numerical • often relates to subjective qualities, characteristics, or descriptions	**Nominal**: • No order, hierarchy or sequence	• Colors • Types of cuisine • Genders • Blood types
	Ordinal: • can be arranged in a particular order	• 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)

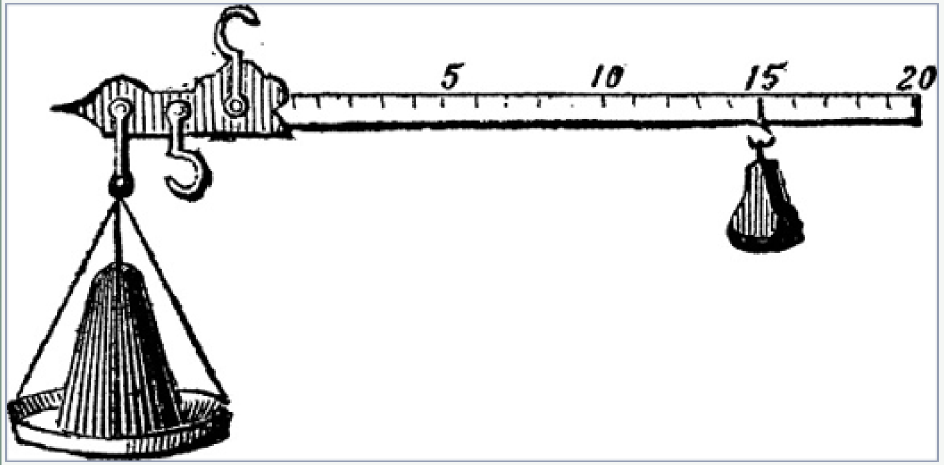
Measures of Central Tendency

… calculate typical central points that describe or represent dataset

Arithmeric Mean = Average

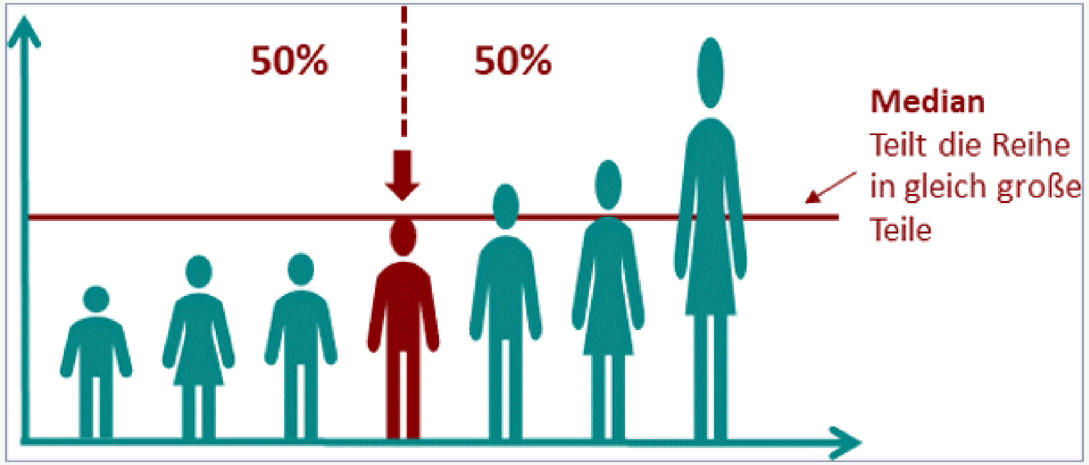
- Sum of all data points divided by number of data points
- E.g.: (1+2+3)/3 = 2

▪ $\overline{X}$ = 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 ($x_1, x_2, \ldots, x_k$)
- occurring with probabilities ($p_1, p_2, \ldots, p_k$).
- For discrete variables:
  - $E[X] = \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$
- For continues 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:

$E[X] = \Sigma_i^k{}_{=1}\, x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$

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

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

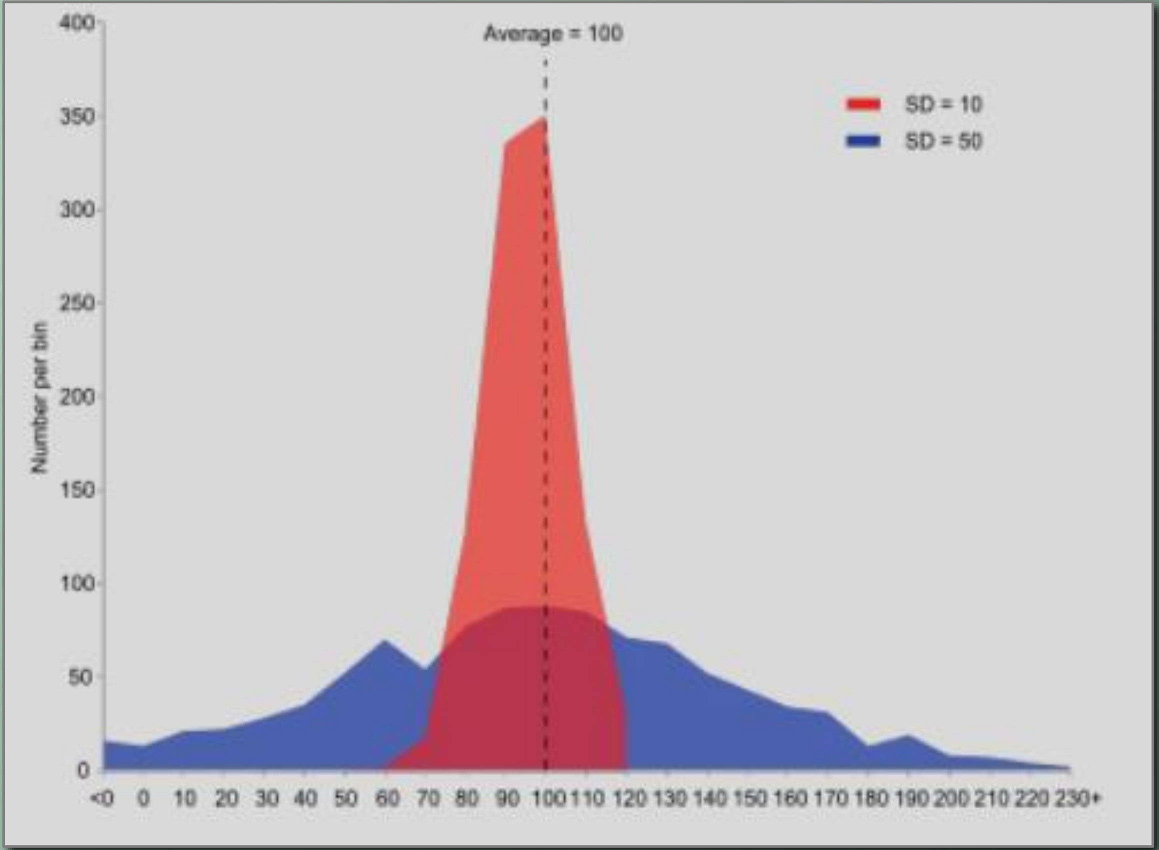
Variance $VAR(X) = E[(X - \mu)^2]$ of standard X (function) is expected value of squared deviation from mean X; $\mu = E[X]$

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.

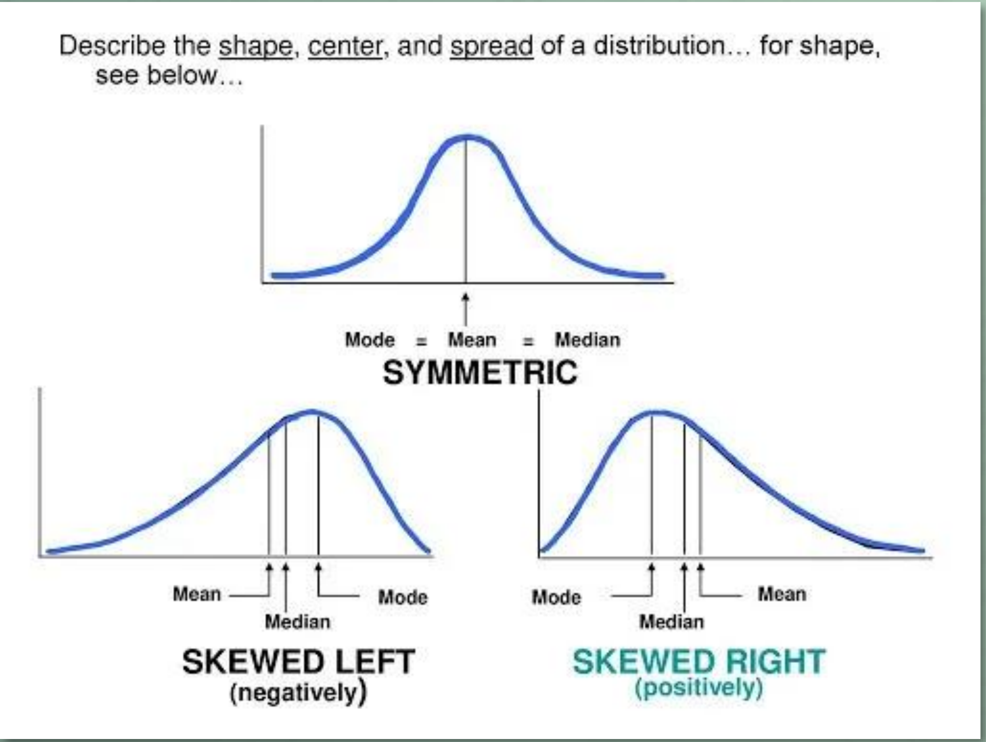


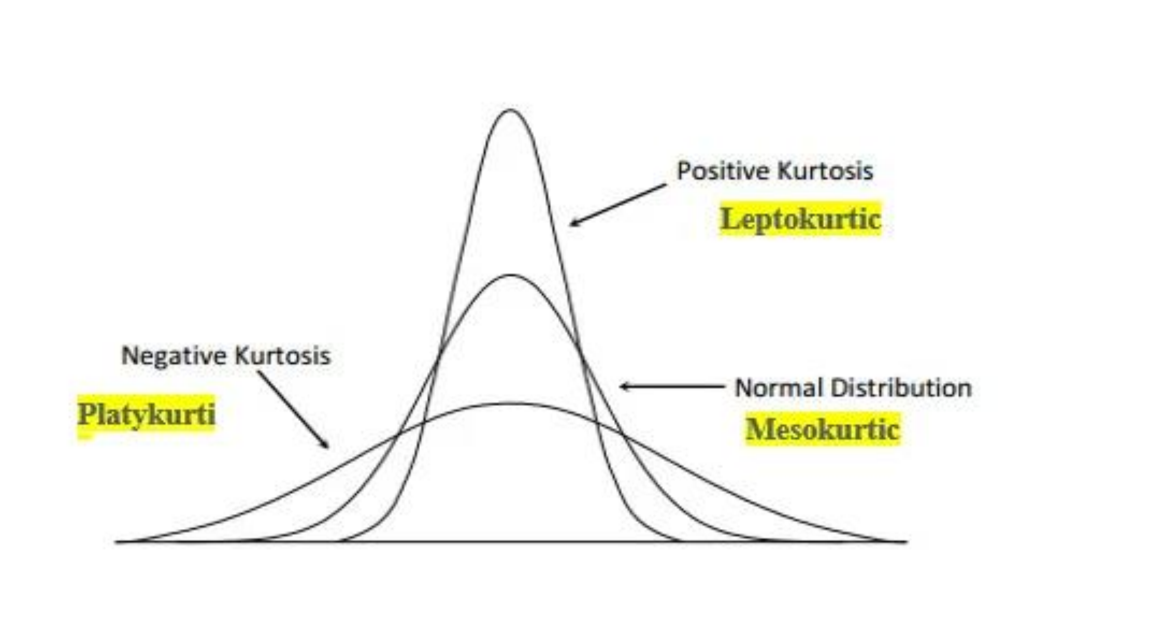
Skewness and Kurtosis

… describe asymmetry of distributions diviation 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)

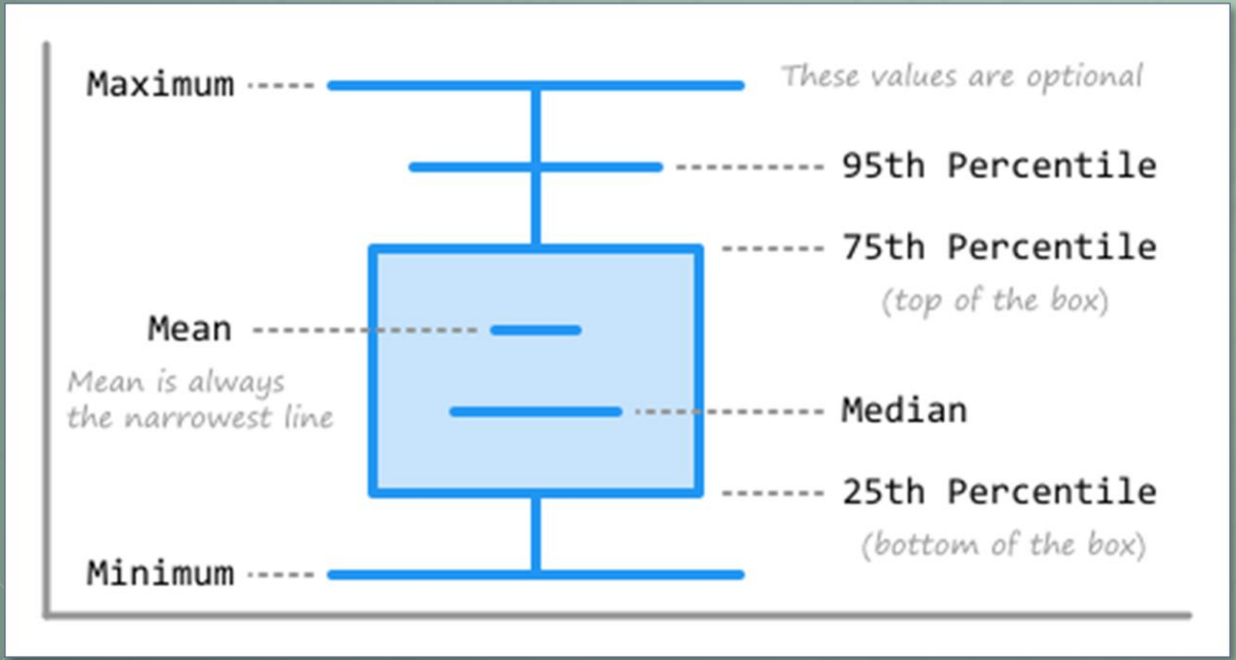
Kurtosis
- defines how heavily the tails of a distribution differ from the tails of a normal distribution
- Outliers may facilitate positive kurtosis





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.

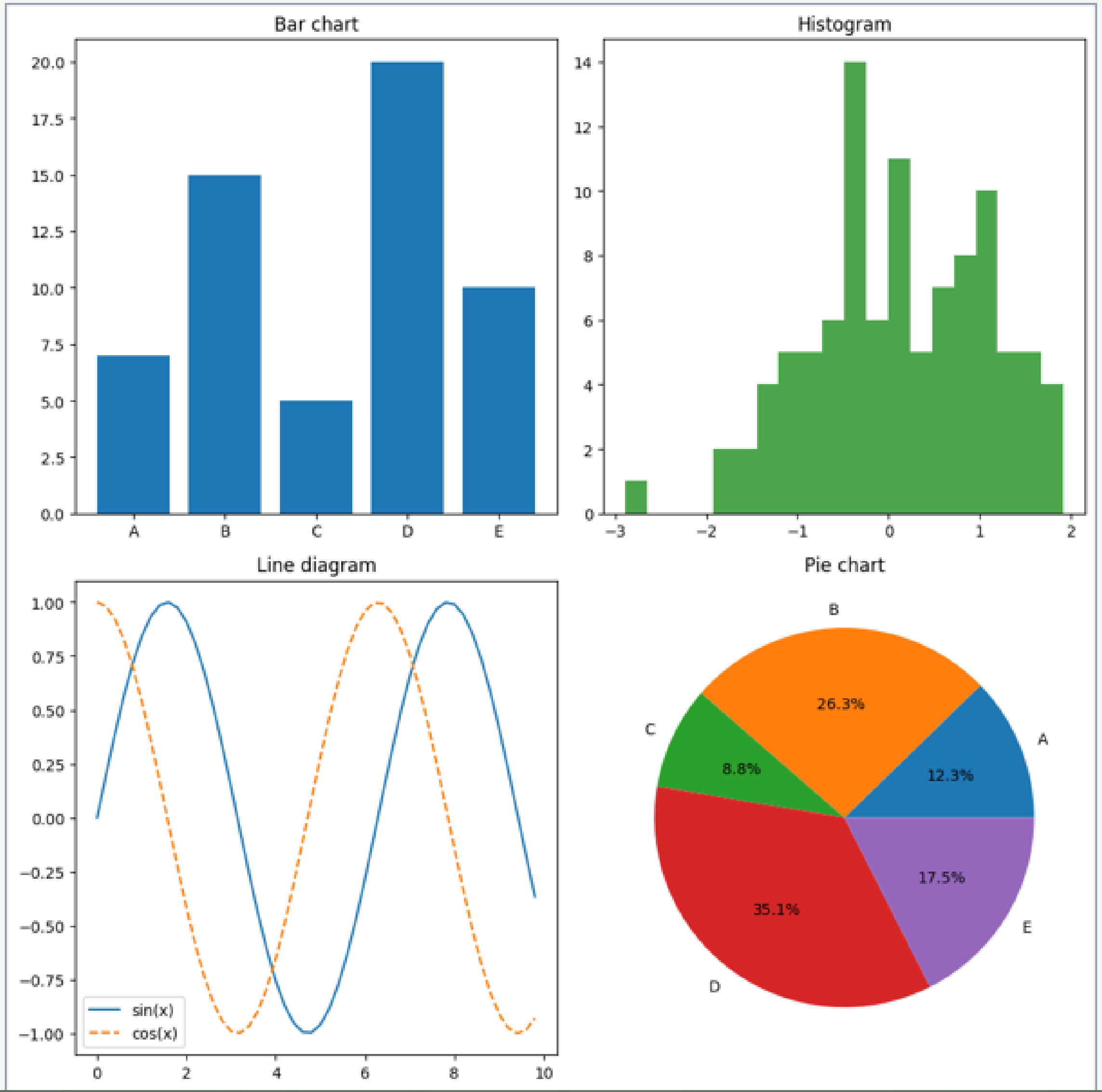
Graphical Representation of Data

Input:
- Categorical
- Nominal
- unordered

Output:
- Numerical
- Continuous
- Absolute

Input:
- Numerical
- continuous

Output:
- Numerical
- continuous



Input:
- Categorical
- Nominal
- ordered

Output:
- Numerical
- Continuous
- Absolute

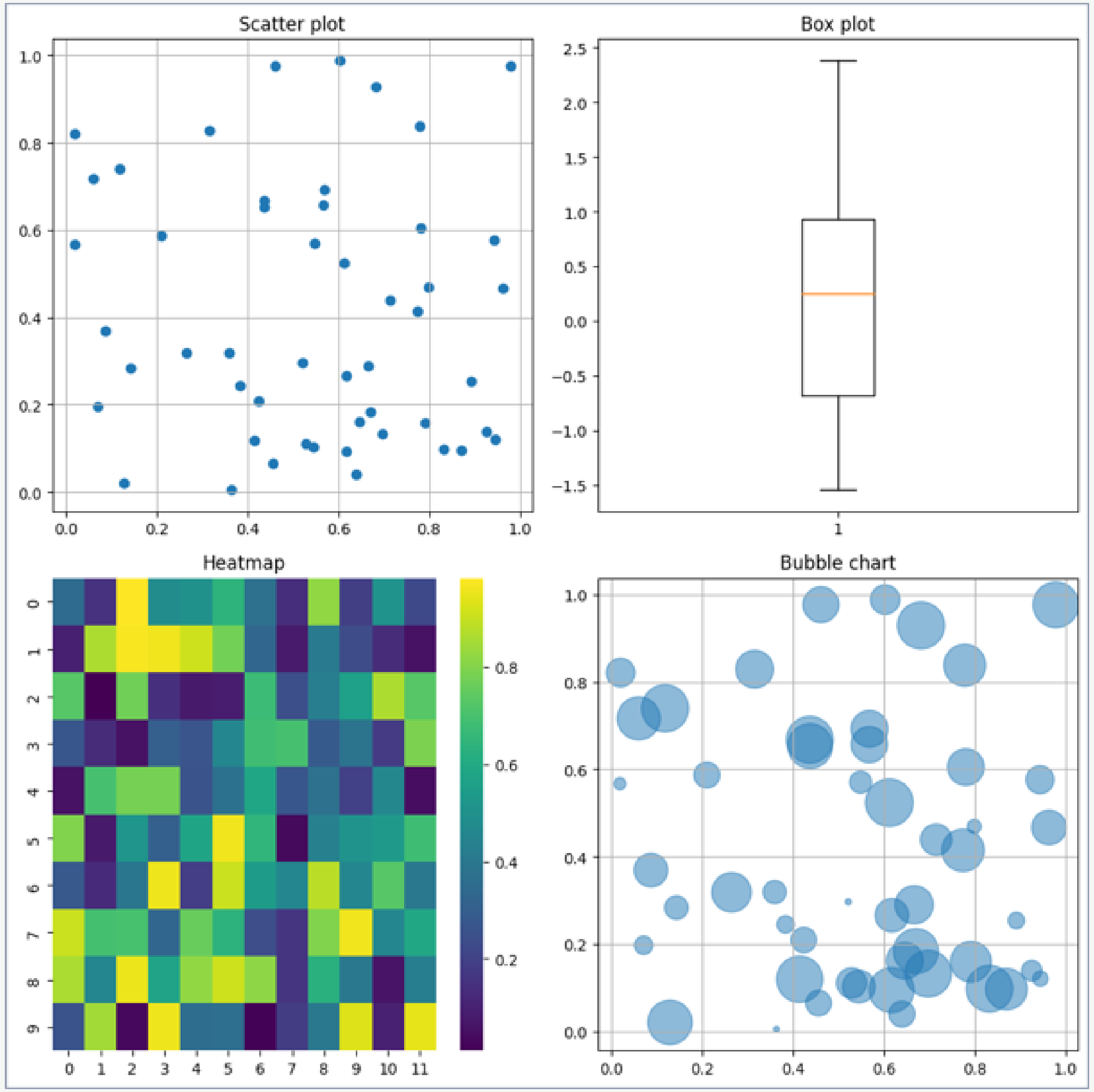
Input:
- Categorical
- Nominal
- Unordered or ordered

Output:
- Numerical
- Continuous
- relative

Graphical Representation of Data

Input:
- 2 Features
- Both numerical

Output:
- Boolean (1;0)



Input:
- 1 feature
- numerical

Output:
- At least ordered
- Or numerical

Input:
- 2 Features
- Both categorical

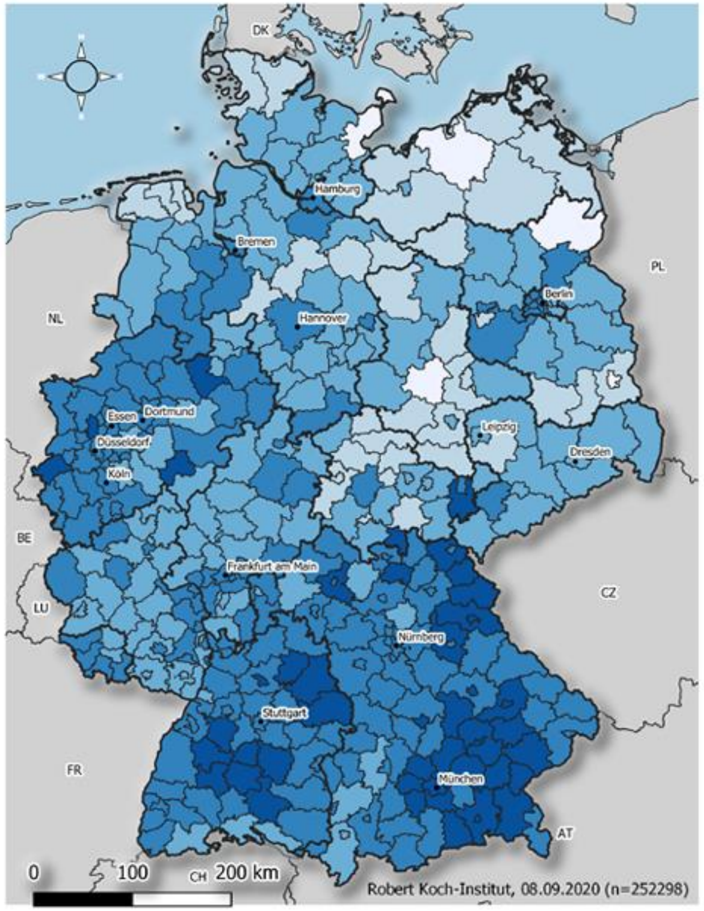
Output:
- Numerical

Input (like scatter plot):
- 2 Features
- Both numerical

Output:
- Numerical

Graphical Representation of Data

… Geovisualization with Heat Map. Example: COVID incidents Germany



COVID-19 (cumulative)
cases per 100,000 population

- >0,0 - 50,0
- >50,0 - 100,0
- >100,0 - 250,0
- >250,0 - 500,0
- >500,0 - 2.000,0

State	cases	/100,000
Baden-Wuerttemberg	43861	396,2
Bavaria	59804	457,3
Berlin	11792	314,6
Brandenburg	3943	157,0
Bremen	2073	303,5
Hamburg	6527	354,5
Hesse	16278	259,8
Lower Saxony	17372	217,6
Mecklenburg-Vorpommern	1034	64,2
North Rhine-Westphalia	60695	338,5
Rhineland-Palatinate	9434	231,0
Saarland	3199	323,0
Saxony	6155	150,9
Saxony-Anhalt	2287	103,6
Schleswig-Holstein	4150	143,3
Thuringia	3694	172,4

ROBERT KOCH INSTITUT

Figure 1: Number and cumulative incidence (per 100,000 population) of the 252,298 electronically reported COVID-19 cases in Germany by county and federal state (08/09/2020, 12:00 AM). Please see the COVID-19 dashboard (https://corona.rki.de/) for information on number of COVID-19 cases by district (local health authority).

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
- k-permutation
- Combination
- Pascal's triangle
- … with replacement
- 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:

$$\boldsymbol{P(n) = n!}$$

For $n = 4$ there are

$$\boldsymbol{P(4) = 4! = 4 * 3 * 2 * 1 = 24}$$

possibilities.

Here: No replacement

Permutation

… is an arrangement of objects in a specific order.

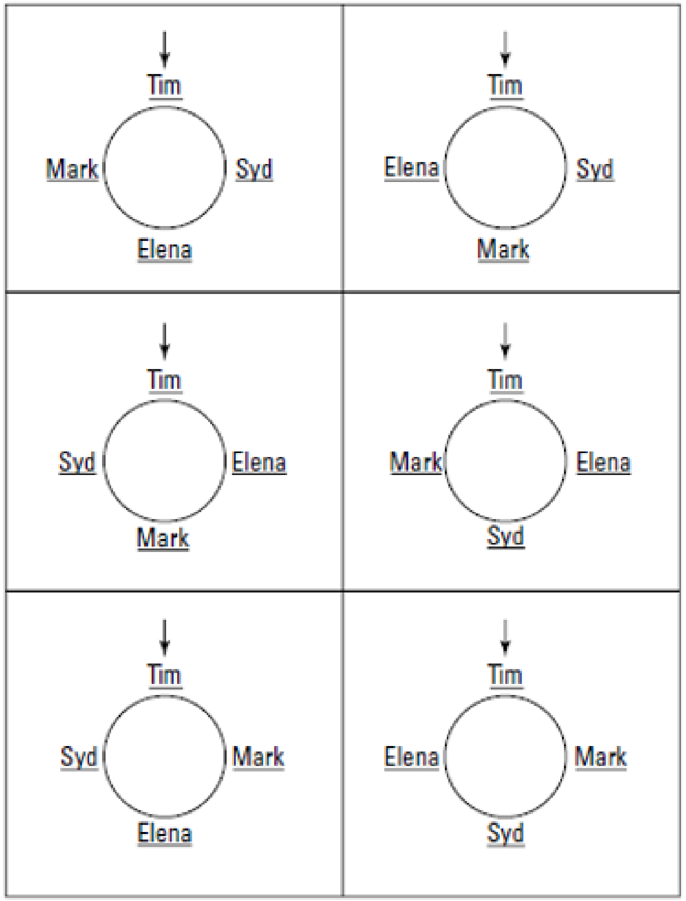


Figure 5-3: Possible rearrangements of four people sitting around a table discussing probability.

Example

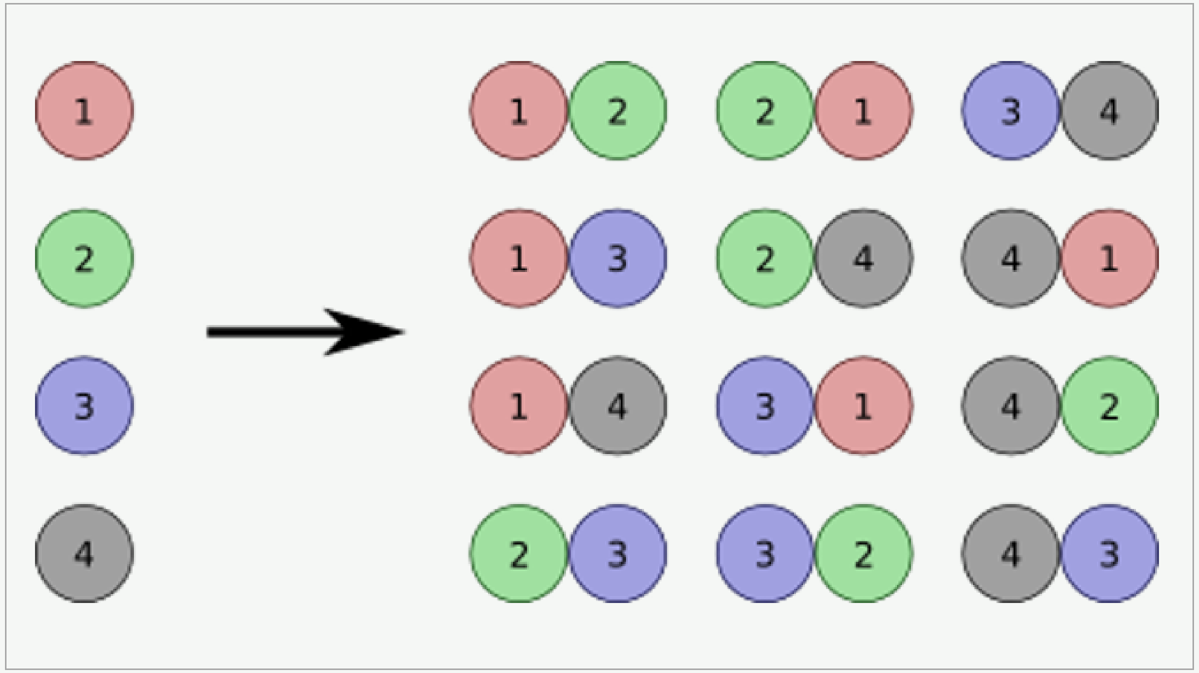
- 4 people sit at a table
- Restriction: The position of one of them is fixed.
- How many possible arrangements are there?

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

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]

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

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

possibilities.

Other k-permutations:

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

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

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

Here: No replacement

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

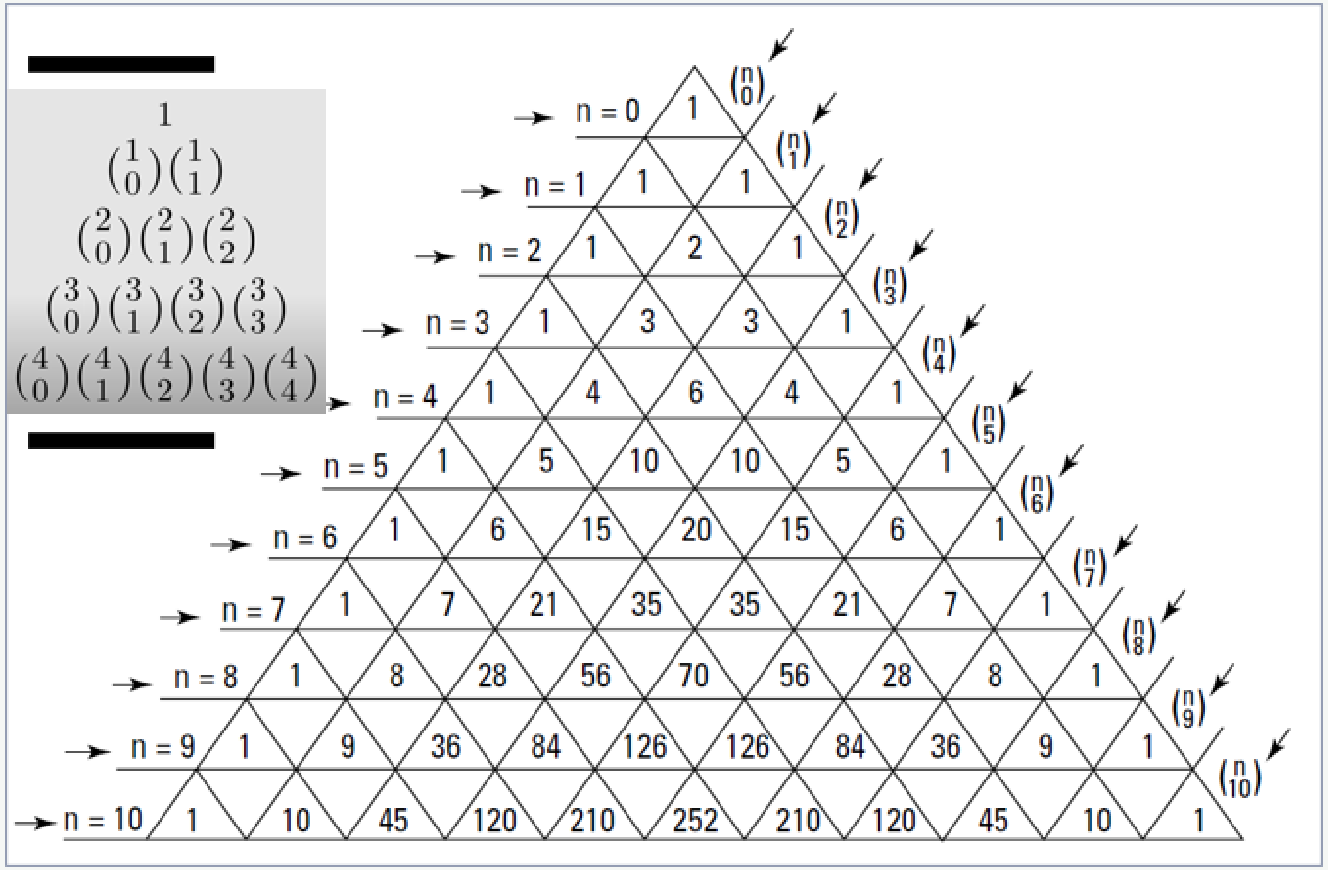
- 49 balls
- 6 selected balls

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:

$$(a+b)^0 = 1$$
$$(a+b)^1 = 1\mathrm{a} + 1\mathrm{b}$$
$$(a+b)^2 = 1a^2 + 2\mathrm{ab} + 1b^2$$
$$(a+b)^3 = 1a^3 + 3a^2\mathrm{b} + 3\mathrm{a}b^2 + 1b^3$$

- Sum of rows → 2^n

Coming in next chapter:

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

Combinatorics with Replacement

The number of ways to choose k items from n distinct items, where each item can be chosen more than once and the order doesn't matter.

Concept …with Replacement	Definition	Formula	Example
Permutation	▪ Arrangement of items where order matters ▪ each item can be selected more than once.	$\frac{n!}{k_1! * \cdots * k_s!}$	▪ Choosing 3-digit PIN from 10 digits (0-9). ▪ Possible permutations: 10^3 = 1000 combinations.
k-Permutation	▪ Subset of k items chosen ▪ arranged from n items ▪ allowing repetition.	n^k	▪ Selecting, arranging 2 letters from the alphabet with repetition. ▪ Example: 26^2 = 676 possibilities.
Combination	▪ Selection of items where order does not matter ▪ each item can be selected more than once.	$\binom{n+k-1}{k}$	▪ Choosing 3 fruits from basket of apples, oranges, and bananas ▪ where repetition is allowed ▪ 10 combinations.

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.

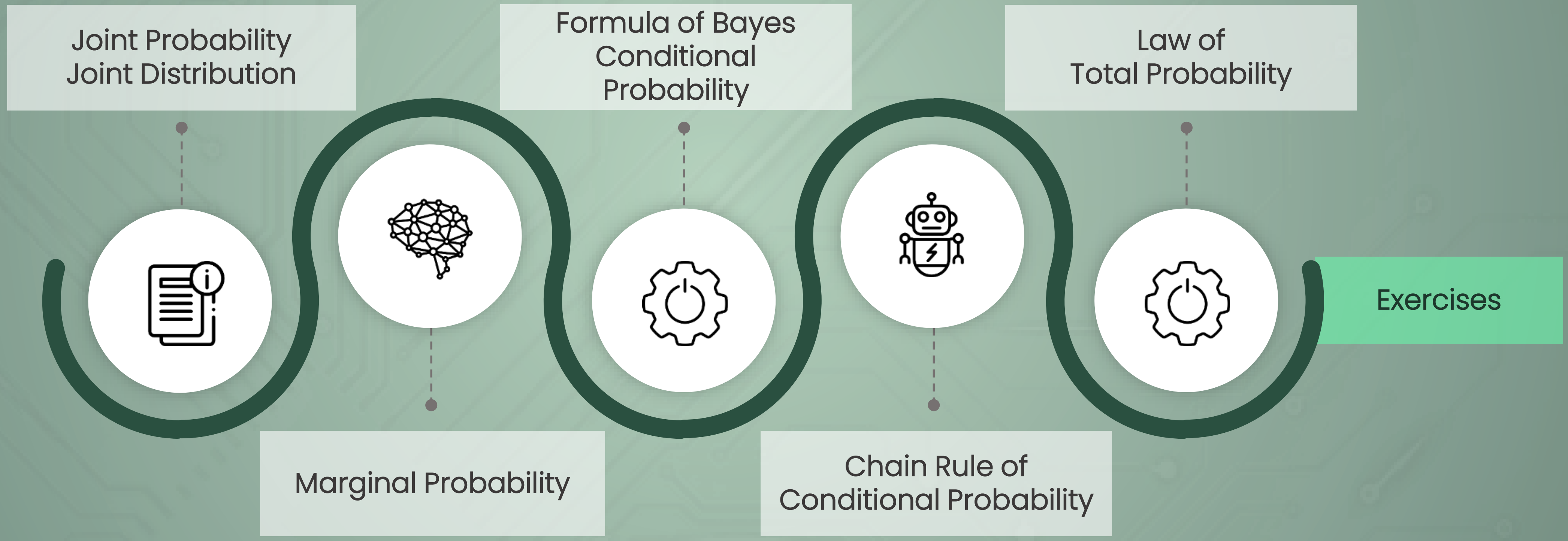
Exercise 4: Generate k-permutations of a list
In this exercise, your task is to write a function that generates all possible k-permutations of a given list using itertools.

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: Calcucate the number of combinations
n = 8 k = 2

Joint and Conditional probability

Bridging Events: How Joint Occurrences and Informed Probabilities interrelate



Joint Probability

is probability of two or more events occurring simultaneously.
It is denoted as P(A and B) or P(A ∩ B) for events A and B.

Example: Two rolling dice (independent events)

```
P(First die = 6 and Second die = 6)
= P(First die = 6) * P(Second die = 6)
= (1/6) * (1/6)
= 1/36
```

Formula

*P(A and B) = P(A) * P(B|A)*

Where:

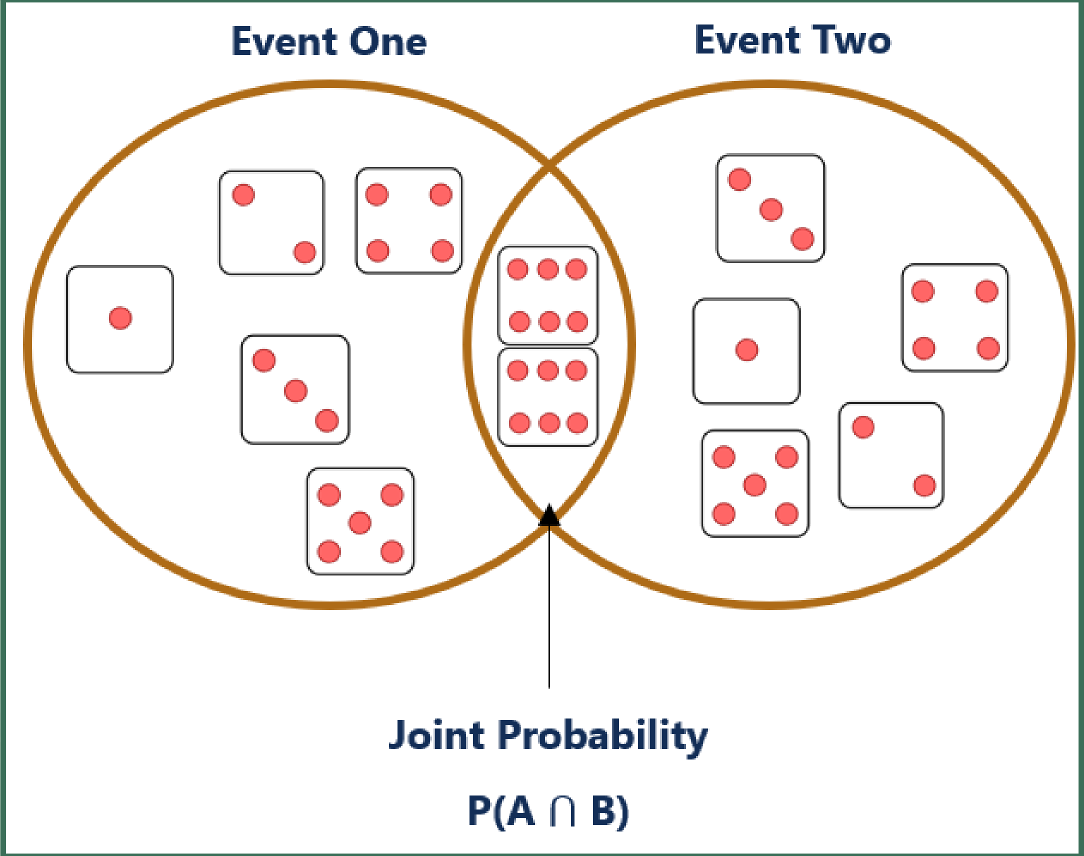
- P(A) is the probability of event A
- P(B|A) is the conditional probability of B given A

(In-) dependence

- If A and B are independent, P(A and B) = P(A) * P(B)
- Else: P(A and B) ≠ P(A) * P(B)
- Mutually Exclusive: If A and B cannot occur together, P(A and B) = 0

Joint Probability Distribution

- Describes probabilities of all possible combinations



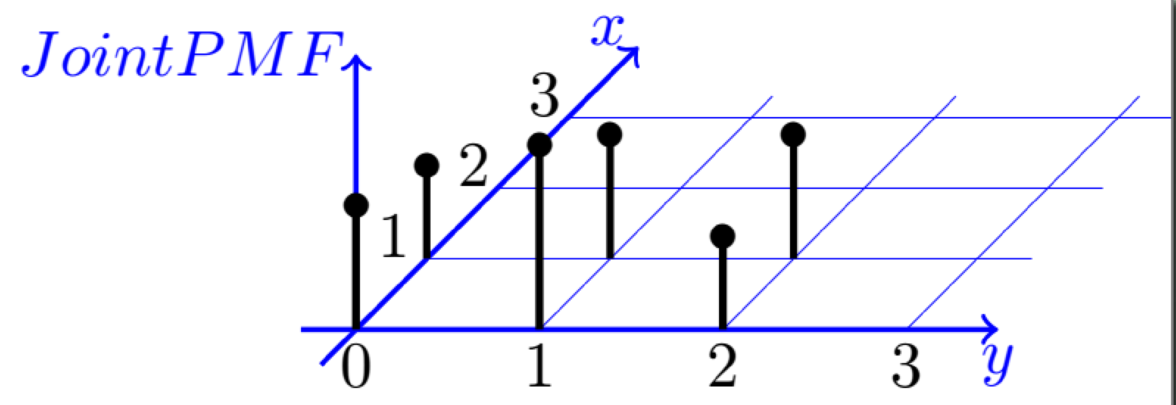
https://corporatefinanceinstitute.com/resources/data-science/joint-probability/

Joint Probability Distribution

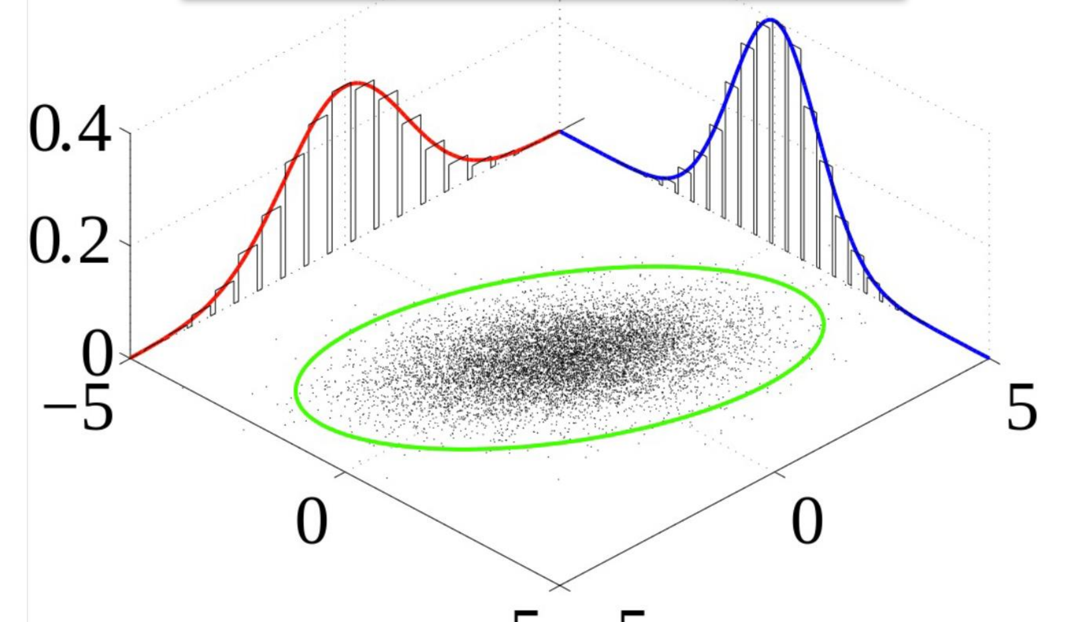
… is probability distribution that gives probability of two or more random variables occurring together. Results of random events are not added or averaged.

Discrete random variables

	Y=0	Y=1	Y=2
X=0	$^1/_6$	$^1/_4$	$^1/_8$
X=1	$^1/_8$	$^1/_6$	$^1/_6$



Continuous random variables



f(x, y) – joint probability density function

Marginal Probability

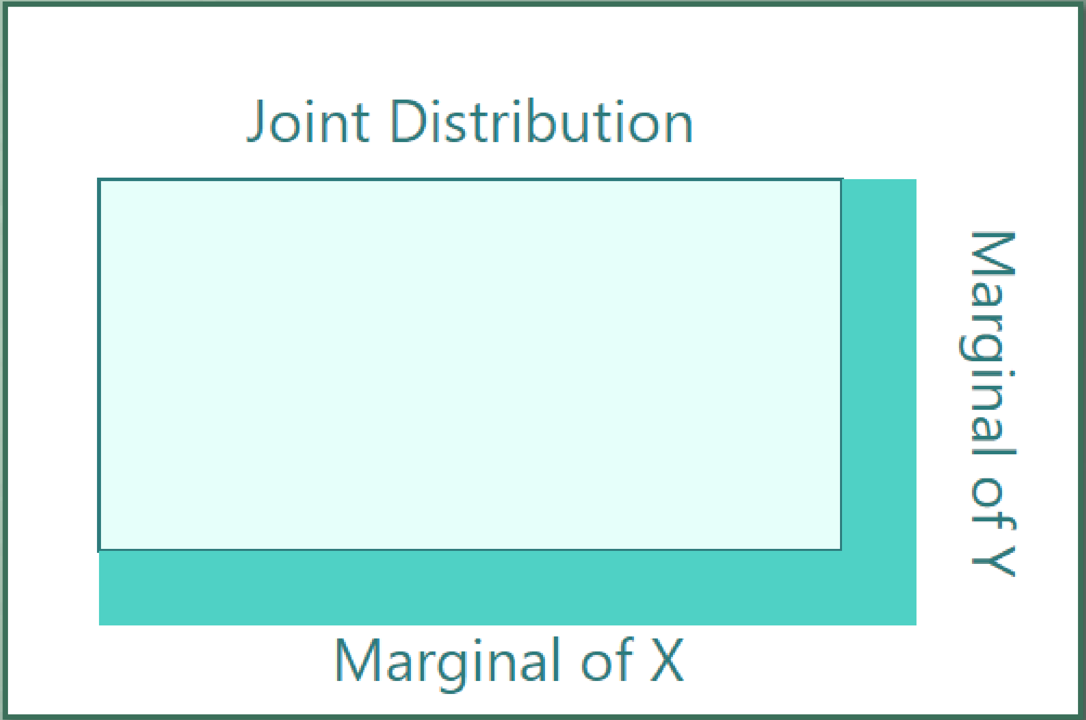
… is probability distribution of subset of variables from larger set of variables, ignoring the other variables.

Key Points

- Derived from joint distribution
- Represents distribution of (summarized) single variable, regardless of others
- Useful for analyzing individual variables in multivariate scenarios
- Can be used to check for independence between variables

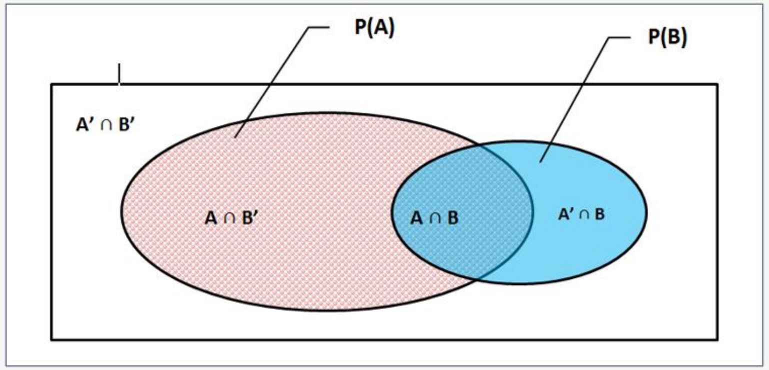
Calculation

- For discrete variables X and Y:
  - Sum over all possible values of Y
  - $P(X = x) = \Sigma P(X = x, Y = y)$
- For continuous variables:
  - Integrate over all possible values of Y
  - $f_X(x) = \int f(x, y)\, dy$



Bayes‘ Theorem

… describes how to update probabilities of hypotheses when given evidence. It’s is used to calculate the conditional probability, i.e., the probability of event based on prior knowledge.



From conditional probabilities to Bayesian theorem:

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

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

$$P(A \text{ and } B) = P(B \text{ and } A)$$

Thus: $P(A|B)P(B) = P(B|A)P(A)$

Or:

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

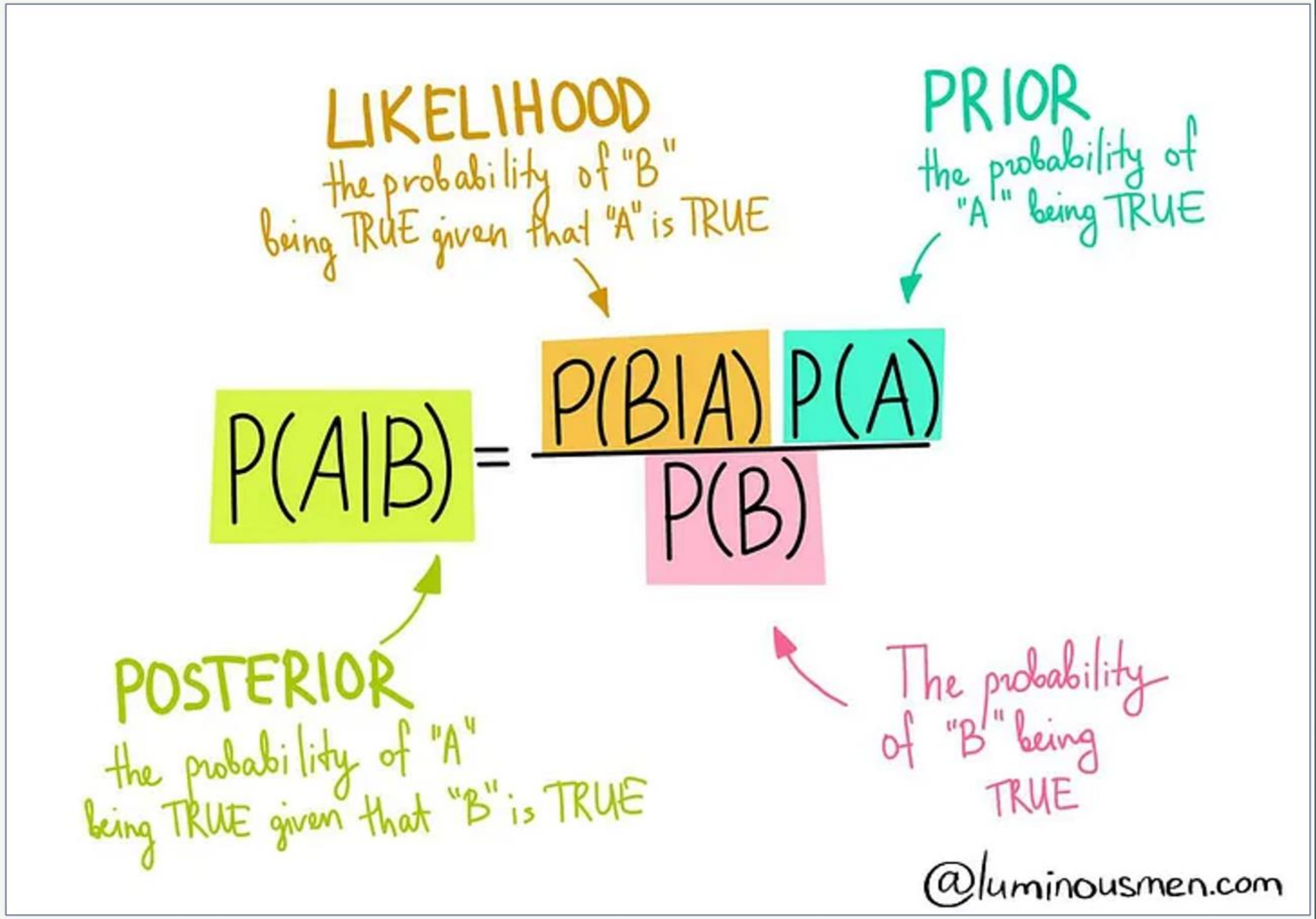


Table with conditional probabilities

… Use Case: How many visitors make a purchase depending on whether they visit website A or B

	Purchase	No Purchase	Total (MP)
Website A	80 (A1) JP: 4 % **CP: 8 %**	920 (A2) JP: 46 % CP: 92 %	$Total\_A$ **1000** $P(A)$ =50%
Website B	90 (B1) JP: 4.5 % CP: 9 %	910 (B2) JP: 45.5 % CP: 91 %	$Total\_B$ **1000** $P(B)$ =50%
Total (MP)	Purchase: 170 (T1) P(Purchase) = 8.5%	Non-purchase: 1830 (T2) P(Non-Purchase) = → 91.5 %	$Total = Total\_A + Total\_B$ 2000 $P(Total) = 100\%$

$$P(Purchase\ given\ Website\ A) = \frac{P(Website\ A\ given\ Purchase) * P(Purchase)}{P(Website\ A)}$$

$$0.08 = \frac{80/170 * 0.085}{0.5}$$

Joint probabilities (JP):

- P(A and Purchase) = A1; T = 80 / 2000 = 0.04
- P(B and Purchase) = B1; T = 90 / 2000 = 0.045

Conditional probabilities (CP):

- P(Purchase | A) = A1; TA = 80 / **1000** = 0.08
  (probability of a purchase given the user saw website A)
- P(Purchase | B) = B1; TB = 90 / **1000** = 0.09
  (probability of a purchase given the user saw website B)

Marginal probabilities (MP):

- P(A) = TA; T = 1000 / 2000 = 0.5
  (probability a visitor sees website A)
- P(B) = TB; T = 1000 / 2000 = 0.5
  (probability a visitor sees website B)
- P(Purchase) = T1; T = 170 / 2000 = 0.085
  (probability a visitor makes a purchase)

Conditional vs. Joint Probability

Bridging Events: Conditional Probability is based on Joint Probability, however considers effect of new information like "Event B has taken place"

Conditional Probability

$P(A|B)$
Probability of A given B has occurred

- Represents updated probability based on new information
- Calculates likelihood of one event considering another event's occurrence
- Formula: $P(A|B) = P(A \cap B) / P(B)$

Joint Probability

$P(A \cap B)$
Probability of both A and B occurring together

- Represents simultaneous occurrence of multiple events
- Calculates likelihood of all specified events happening
- Formula: $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$

Key Points

- Conditional probability can be derived from joint probability: $P(A|B) = P(A \cap B) / P(B)$
- For independent events: $P(A \cap B) = P(A) * P(B)$, and $P(A|B) = P(A)$
- Bayes' Theorem connects these concepts: $P(A|B) = P(B|A) * P(A) / P(B)$

Reference to Combinatorics

- Sampling without Replacement: Connected to Conditional Probability
- Sampling with Replacement: Connected to Joint Probability

Chain Rule of Conditional Probability

… allows to express joint probability of multiple events as product of conditional probabilities.

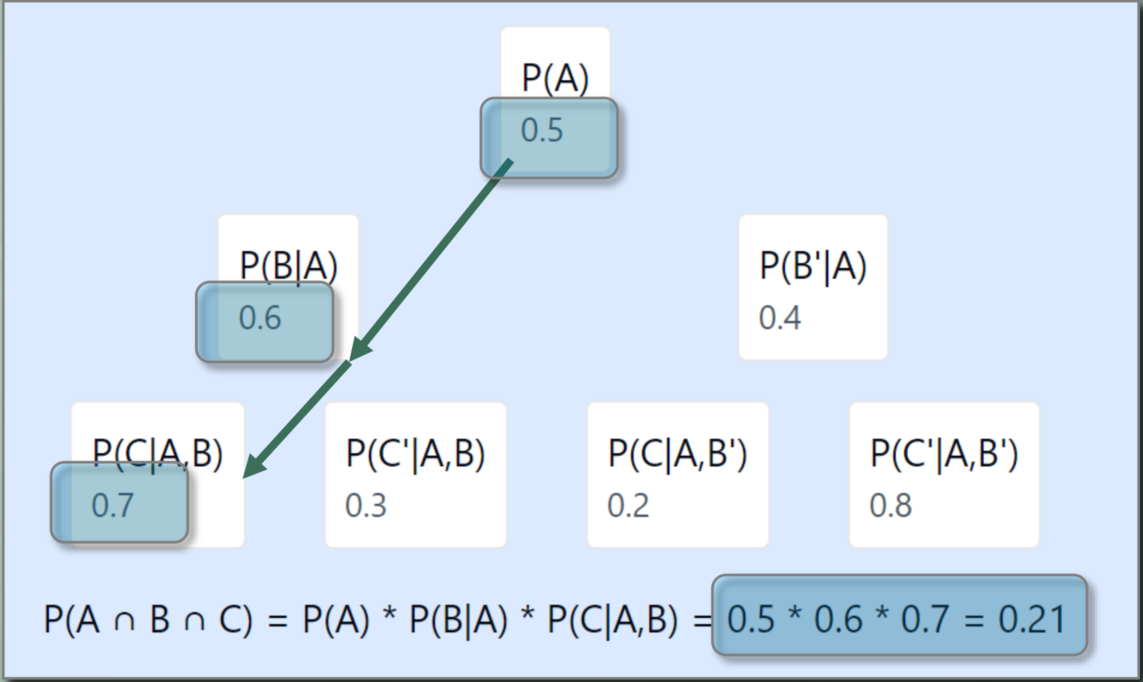
Key Points

- Breaks down complex joint probabilities
- Useful when direct calculation of joint probability is difficult
- Order of events can affect calculation complexity

Calculation

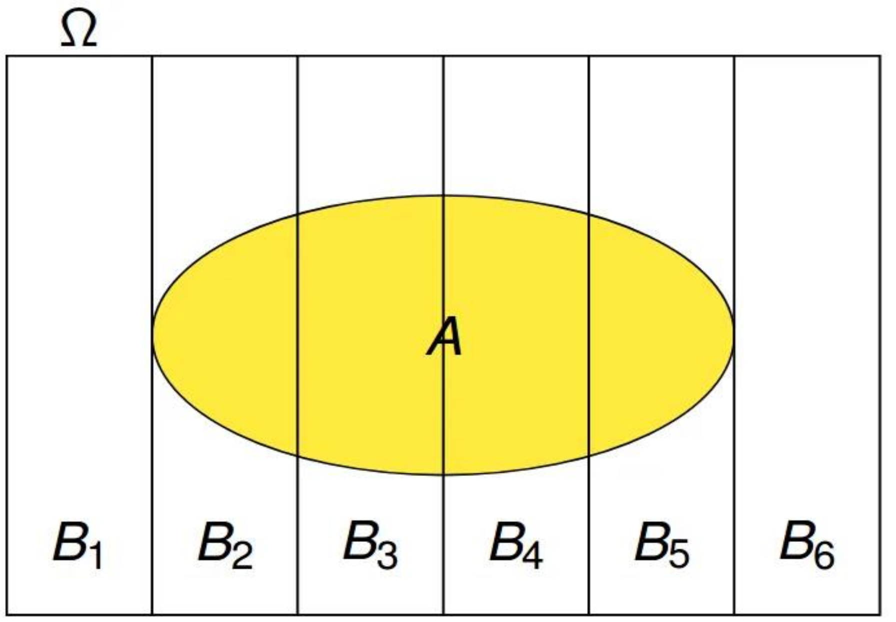
- For two events: A & B
  - $P(A \cap B) = P(A|B) * P(B)$
- Find an example?
- More than two events $A_1, \ldots, A_n$:
  - $P(A_n \cap \ldots \cap A_1) = P(A_n | A_{n-1} \cap \ldots \cap A_1) * P(A_{n-1} \cap \ldots \cap A_1)$
  - Also:

$$P(A_n \cap \ldots \cap A_1) = \prod_{k=1}^{n} P\left(A_k \middle| \cap_{j=1}^{k-1} A_j\right)$$



Law of Total Probability

… calculates the probability of an event by summing its conditional probabilities across all possible scenarios



$P(A) = \sum P(A \cap B_x) = 0 + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4) + P(A \cap B_5) + 0$

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

This can be derived using the fact that $P(A \cap B) = \boxed{P(A|B)P(B)}$

[Law of Total Probability] Let $\{B_1, B_2, \ldots, B_n\}$ be a partition of the sample space S, and let A be any event. Then:

$$P(A) = \sum_{i=1}^{n} \boxed{P(A|B_i)P(B_i)}$$

where $P(A|B_i)$ is the conditional probability of A given B_i, and $P(B_i)$ is the probability of B_i.

Exercises

Exercise 1:

Animal Joint Probability Problem
A wildlife researcher is studying a forest habitat. They are particularly interested in the presence of two animal species:

Red foxes (F)
Barn owls (O)
Based on previous studies:

The probability of observing a red fox in a given area is 0.4
The probability of observing a barn owl in a given area is 0.3
The probability of observing both a red fox and a barn owl in the same area is 0.15
Calculate the following probabilities:

a) $P(F \cap O)$: Probability of observing both a red fox and a barn owl in the same area b) $P(F \cup O)$: Probability of observing either a red fox or a barn owl (or both) in an area c) $P(F \mid O)$: Probability of observing a red fox, given that a barn owl has been observed d) $P(O \mid F)$: Probability of observing a barn owl, given that a red fox has been observed e) $P(F' \cap O')$: Probability of observing neither a red fox nor a barn owl in an area

Exercises

Exercise 2:
Consider the following frequency table for students' preferences for Programming Language and whether they prefer Dark Theme or not

Programming Language	Dark Theme	Count
Python	Yes	90
Python	No	30
Java	Yes	50
Java	No	40
JavaScript	Yes	60
JavaScript	No	30

Represent this table in Python using a list of dictionaries and calculate the conditional probability of a student preferring a Dark Theme given that they prefer Python, Java, JavaScript.

Probability Distributions

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

PMF/PDF, CDF of Discrete and continuous distributions

Distributions Overview

Binomial, Hypergeometric, Poisson, normal distribution

Students' t-distribution

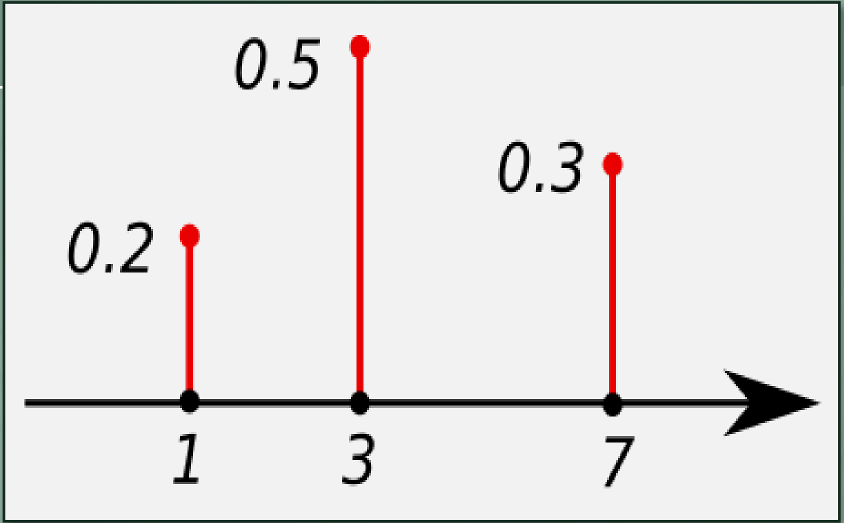
χ^2- distribution Exponential, Gamma Distribution

F-Distribution

Exercises

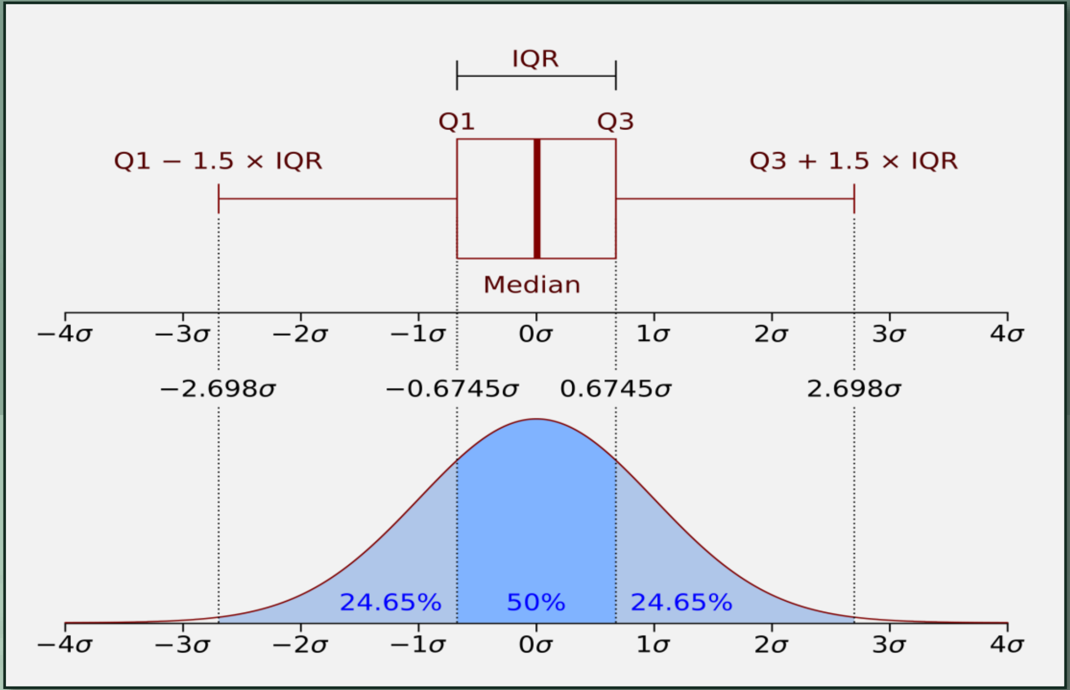
Probability Distributions

… 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

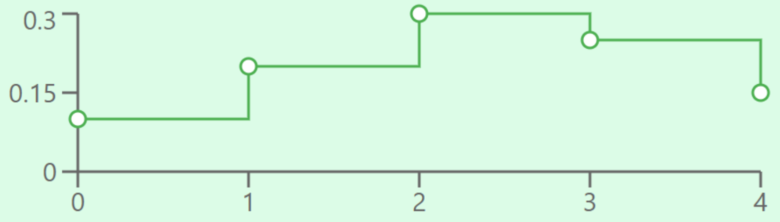
- Represented by a smooth curve (bell curve shown)
- Area under the curve represents probability
- Total area under curve equals 1
- Example: Normal (Gaussian) distribution
- Shows
  - median, quartiles, and standard deviations
  - Interquartile Range (IQR): Distance between Q1 and Q3
  - Standard Deviation: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

Probability Distributions

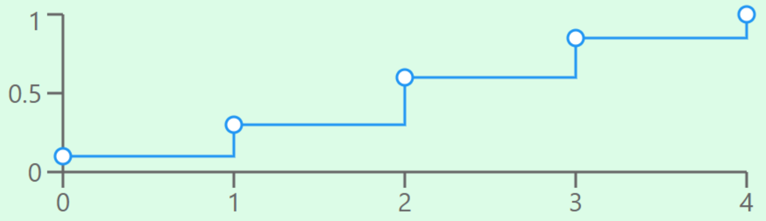
Comparing Probability Mass/Density and Cumulative Distribution Functions

Discrete Distributions

Probability Mass Function (PMF)

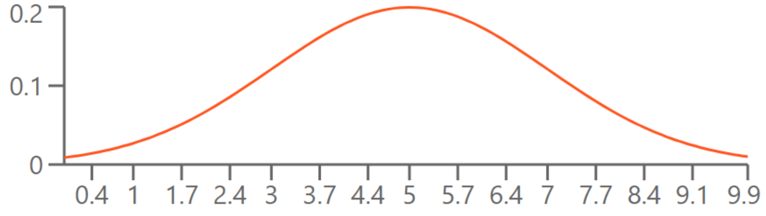


Cumulative Distribution Function (CDF)

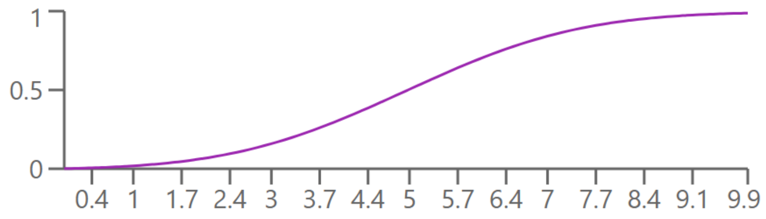


Continuous Distributions

Probability Density Function (PDF)



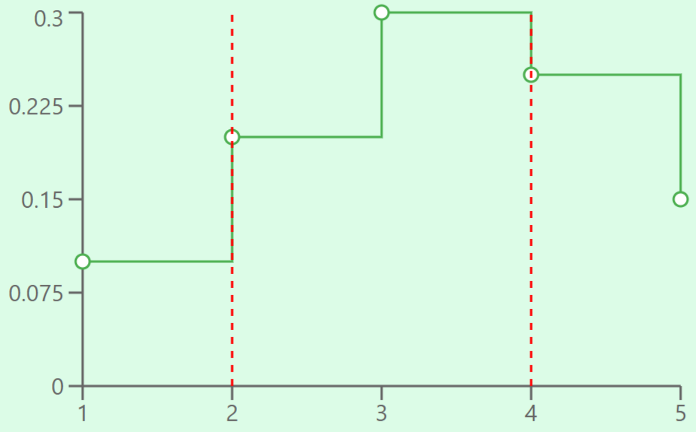
Cumulative Distribution Function (CDF)



Measuring Probability on Distribution Intervals

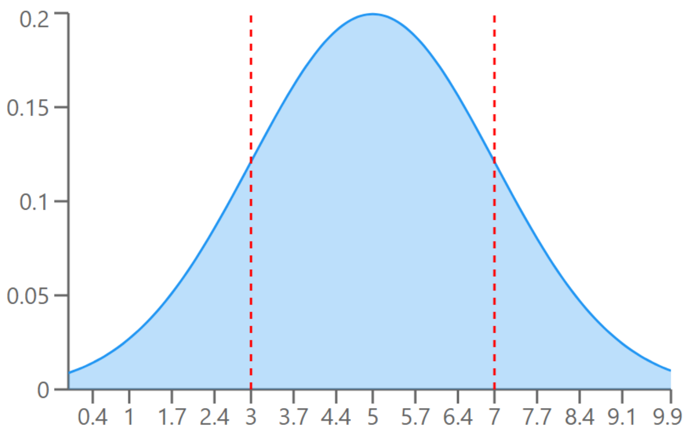
From Sums to Integrals: Calculating Probabilities Across Distribution Types

Discrete Distributions



Probability = Sum of probabilities for each point in the interval
Example: P(2 ≤ X ≤ 4) = 0.2 + 0.3 + 0.25 = 0.75

Continuous Distributions



Probability = Area under the curve between interval points
Example: P(3 ≤ X ≤ 7) = ∫[3 to 7] f(x) dx

Given the shown distribution ...
What is the probability that the guest has drunk 2,3 or 4 beer?

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

- Fair coin/die toss
- Discrete or continuous

- Only success ⇔ failure

Bernoulli

- Coin toss with non-uniform probability
- discrete

Binomial

- Repeated Bernoulli
- discrete
- With replacement

Hypergeometric

- Like Binomial
- discrete
- without replacement

Poisson

- Number of trials $n \to \infty$
- Number of successful trials λ is small e.g. 2 goals in 90 min

„inverse“

- λ becomes large
- many successful trials

Normal/Gaussian

- Number of trials $n \to \infty$
- Number of successful trials $k \to \infty$

Standard normal

- Like Normal distribution, however loss of information
- minus popul. mean
- Divi. by popul. std. dev.

Student's t

- Number of samples n<30

Gamma

- Expected duration till k-th orrucances in Poisson Process
- roughly inverse of Poisson

Exponential

- k = 1: Expected duration till 1st occurance

χ^2

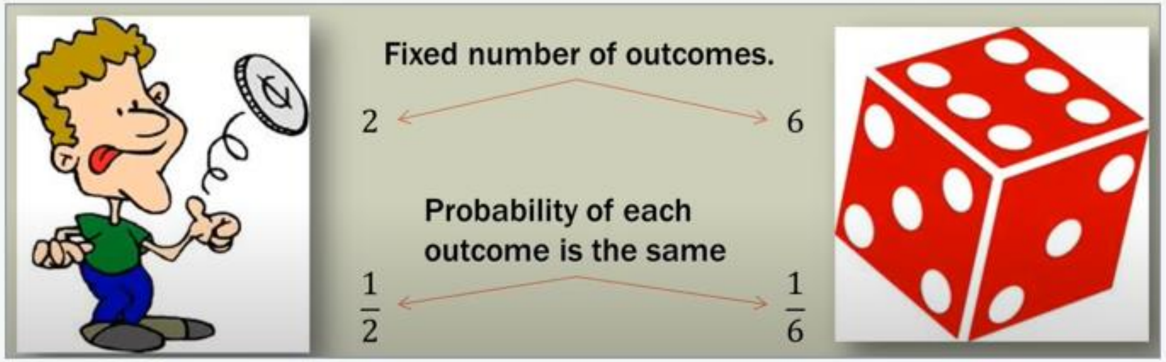
- Sum of squared Standard normal distributions
- Gamma-Scale θ = 2

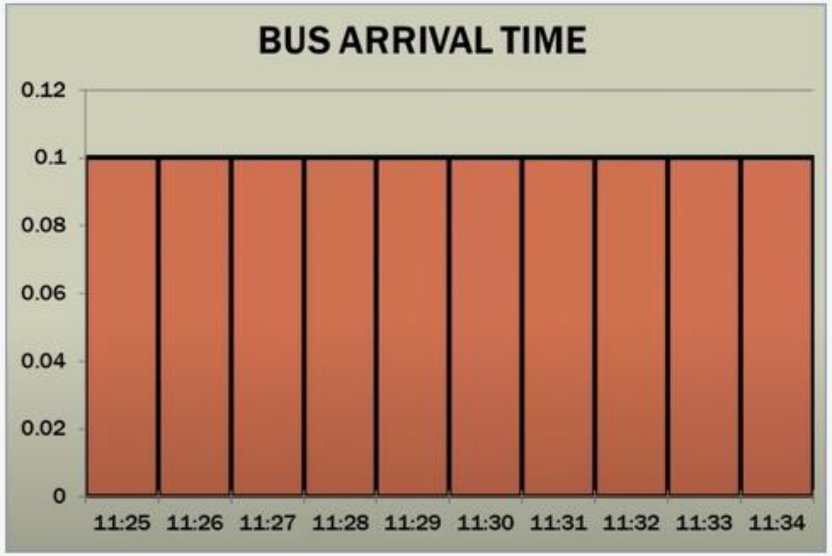
F

- Ratio for χ^2 respectively divided by degrees of freedom
- Ratio: within/between variance of clusters etc.

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:

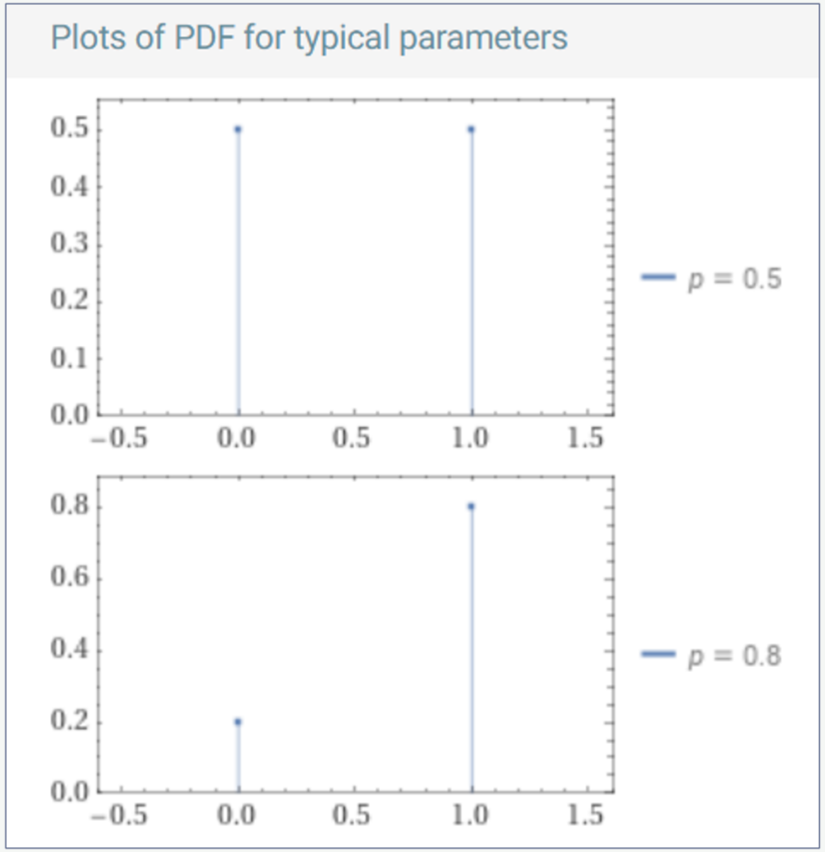
- Generating Random Numbers:
  e.g.: fair dice, fair coin tosses
- Quality Control and Manufacturing:
  e.g.: check the uniformity of products

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)



Formula:

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

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)

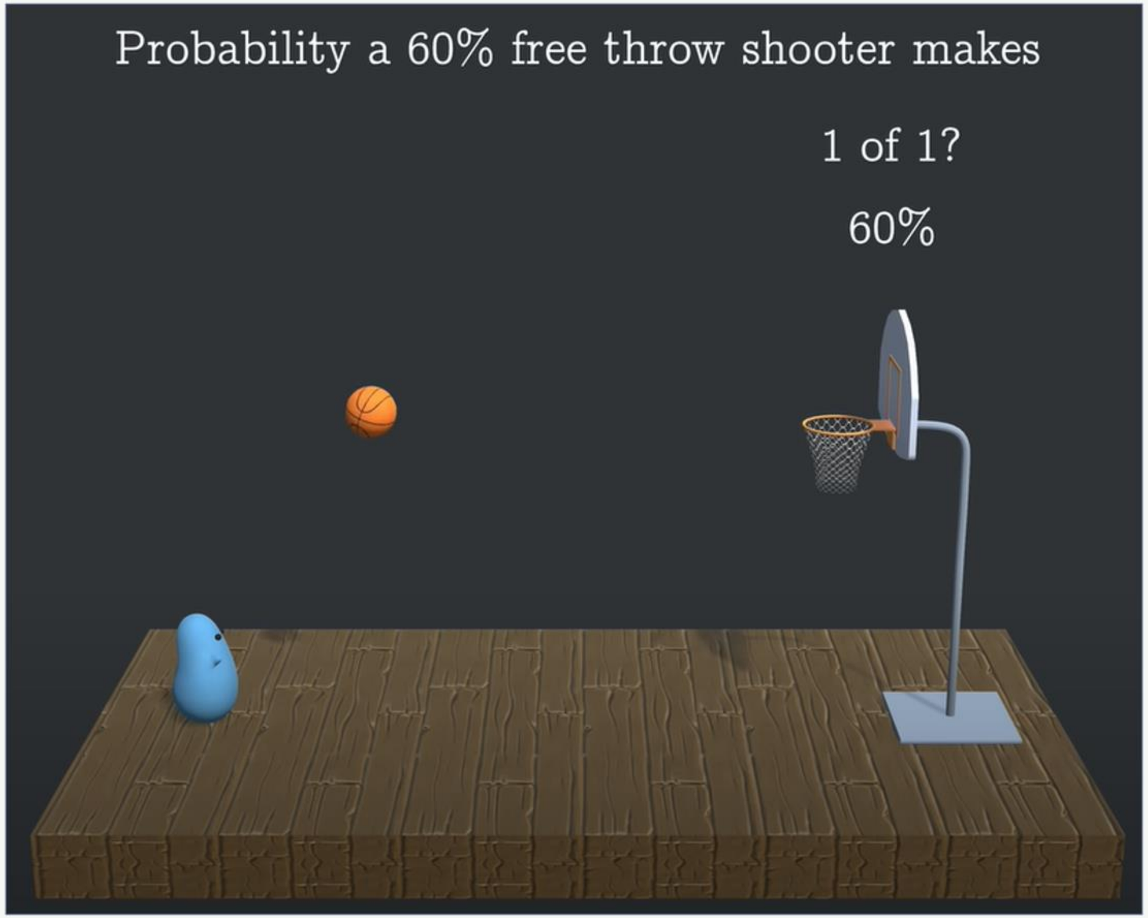
Attention: $x\epsilon[0;1]$ is not an interval!

$P(x;p)$: 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



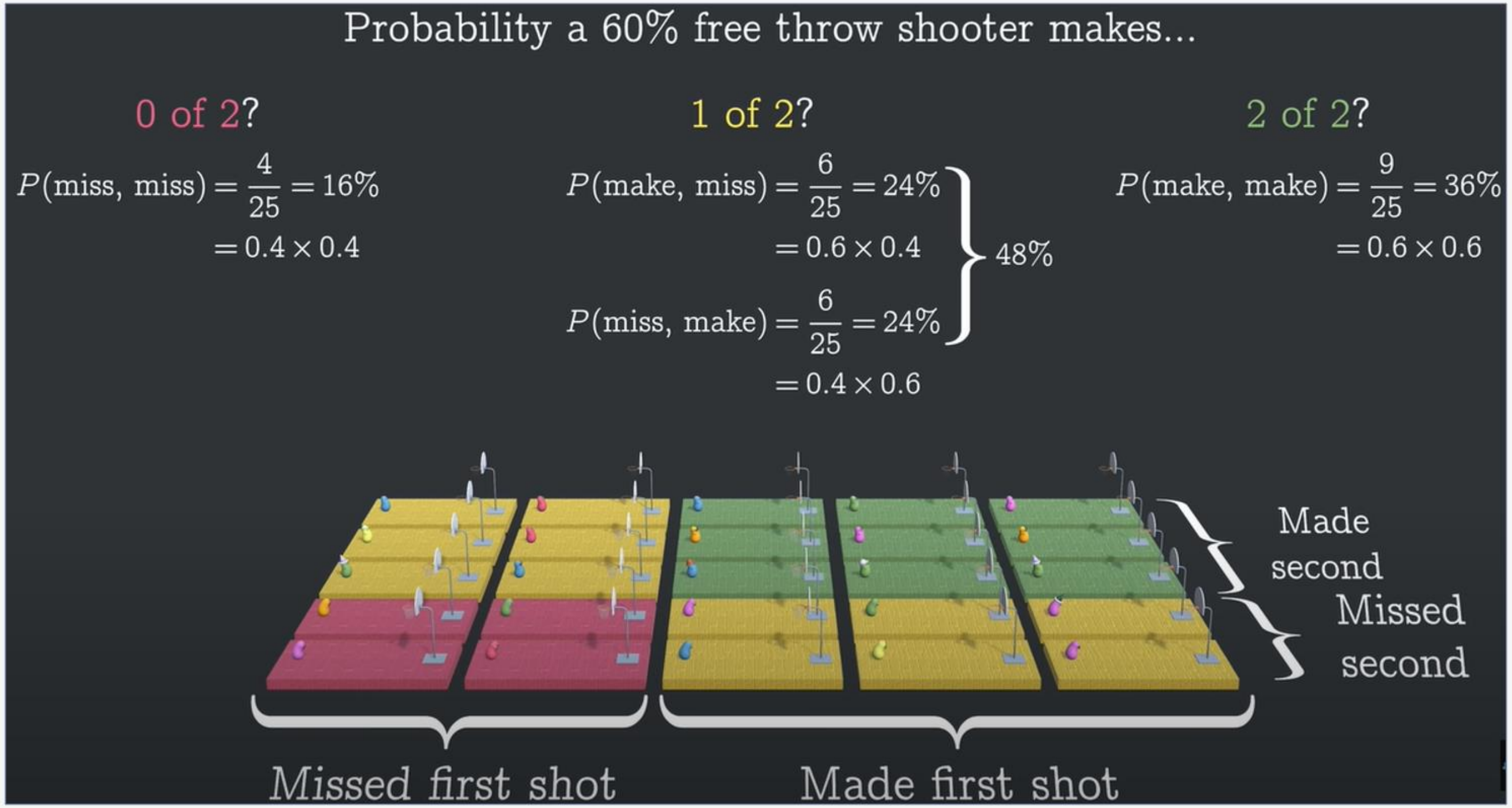
Step 1:

- P(Make the shot) = 60 %
- P(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



Step 2:

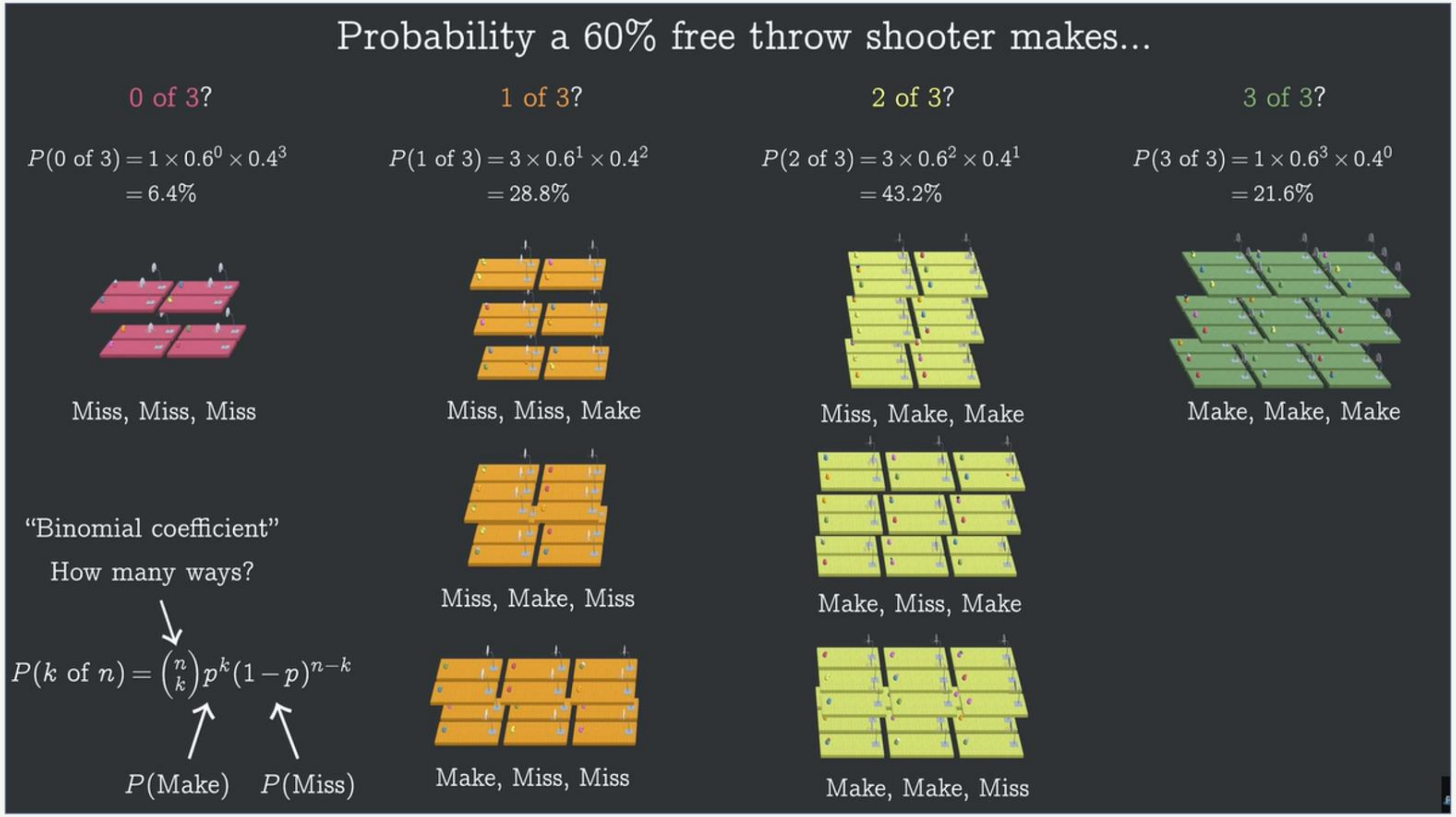
- P(miss, miss) = 16 %
- P(miss, make) = 48 %
- P(make, make) = 36 %

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



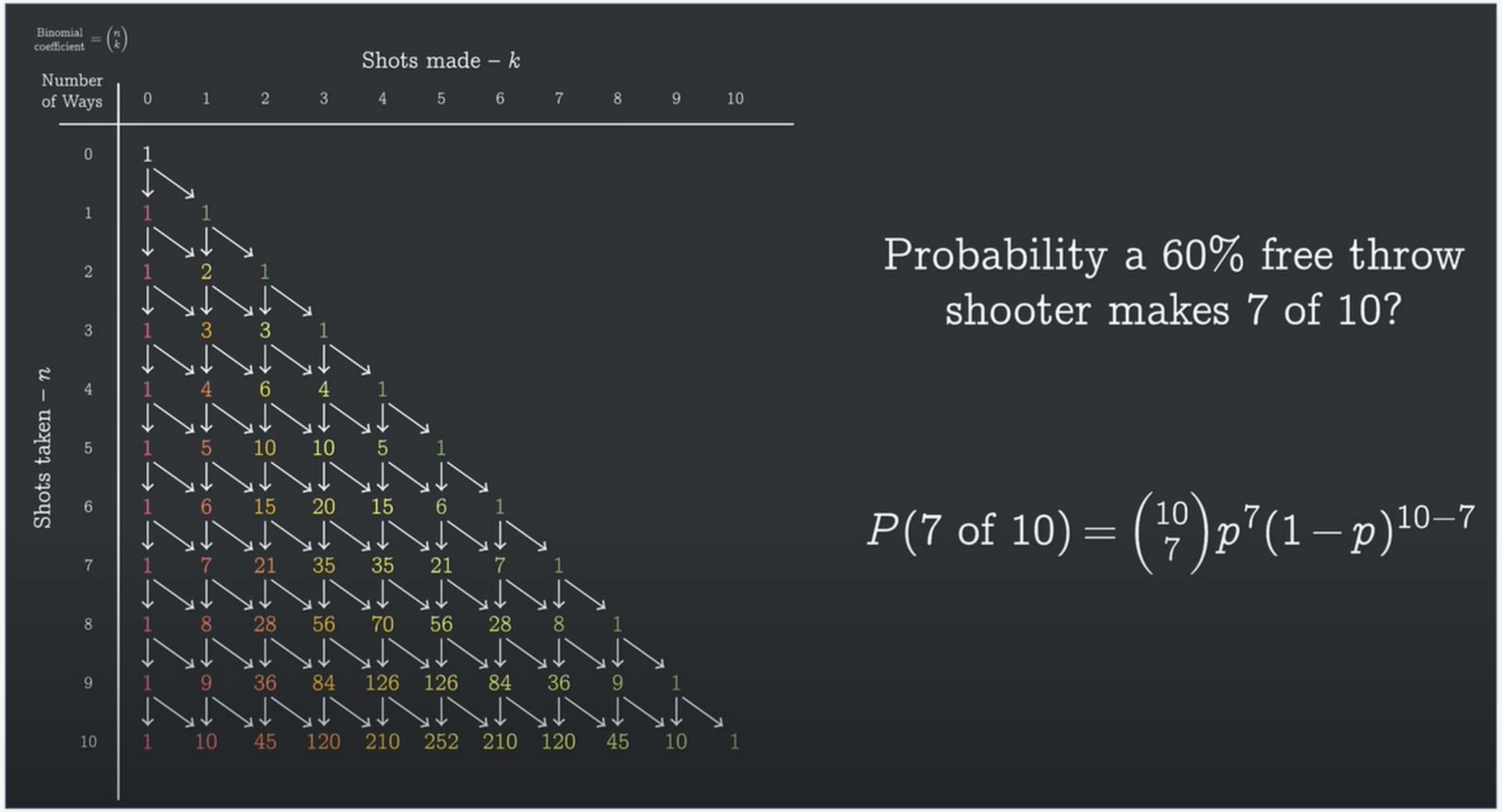
Step 3:

- P(miss, miss, miss) = 8 %
- P(miss, miss, make) = 28.8 %
- P(miss, make, make) = 43.2 %
- P(make, make, make) = 21.6 %

Binomial distribution

Example for probability calculation:

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

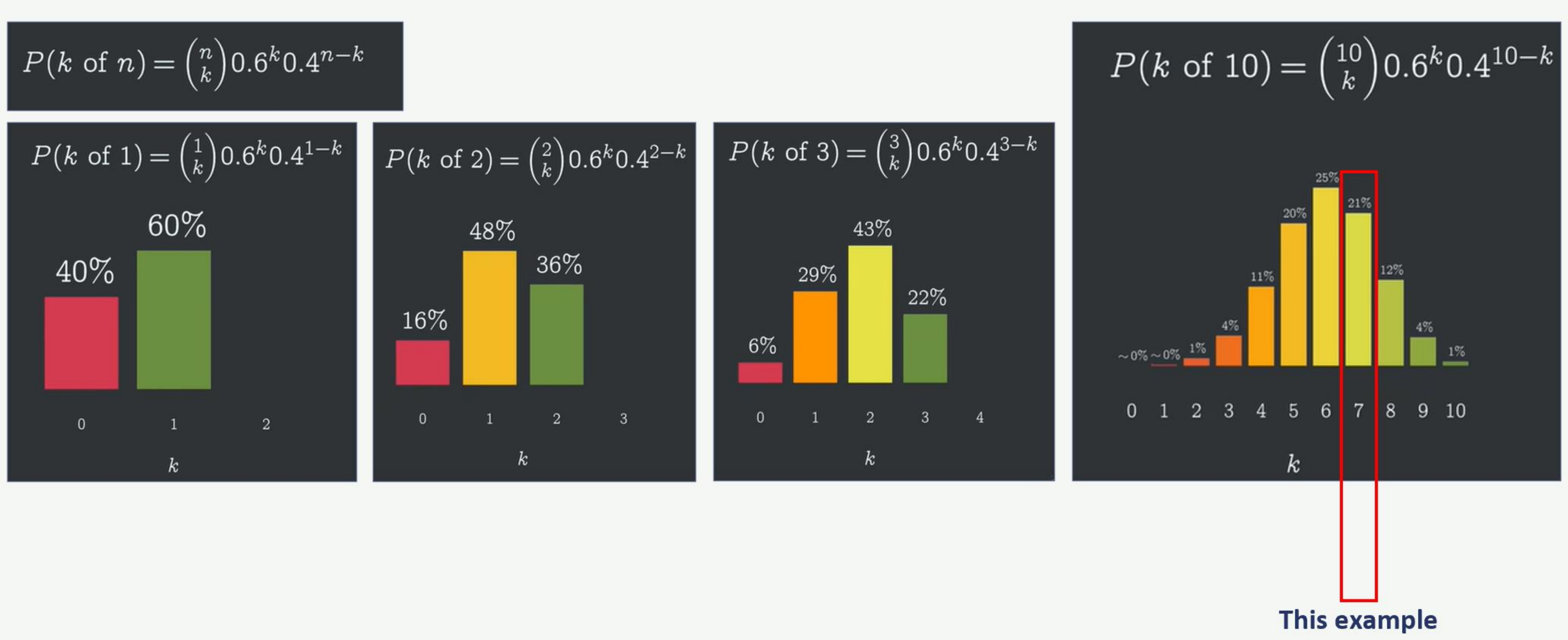


$$\boldsymbol{P(7\ of\ 10) = 120 \times 0.6^7 \times 0.4^3 \approx 21.5\%}$$

Binomial distribution

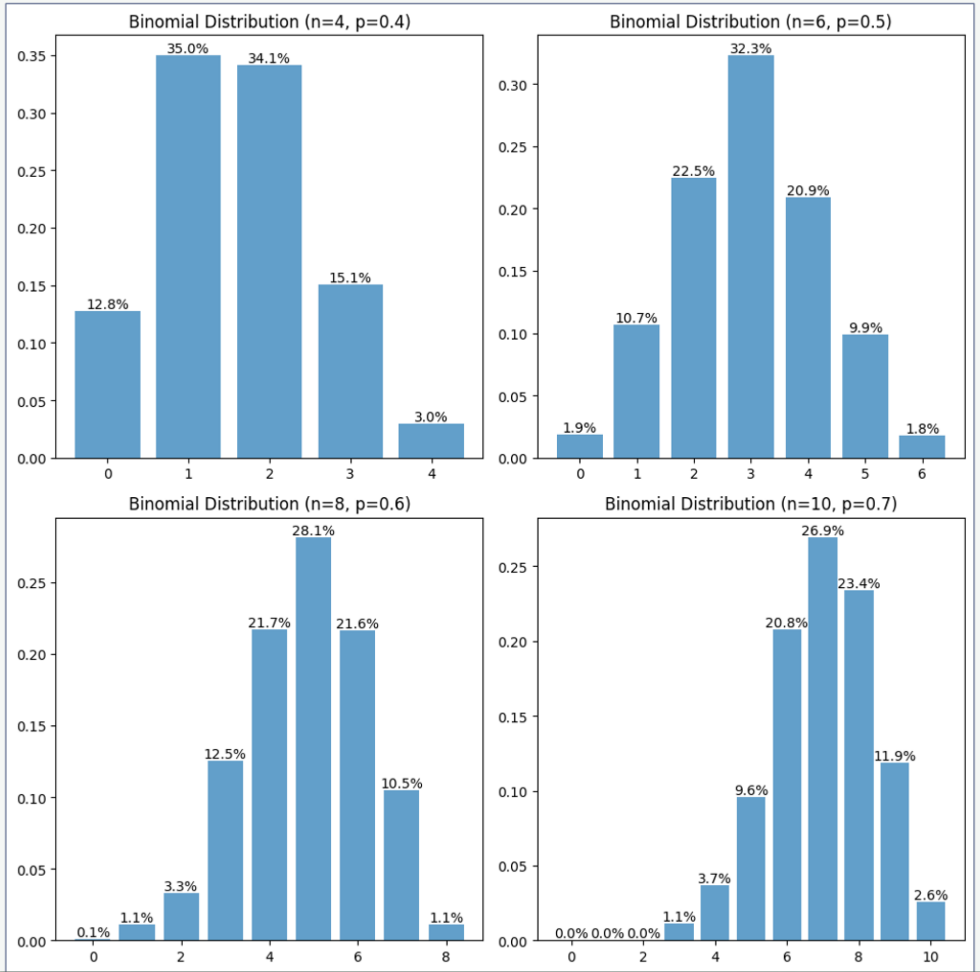
Example for probability calculation - Generalization:

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



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:

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

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:

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

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:

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

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) = $(1/2)^2 * (1/2)^1 = 1/8$

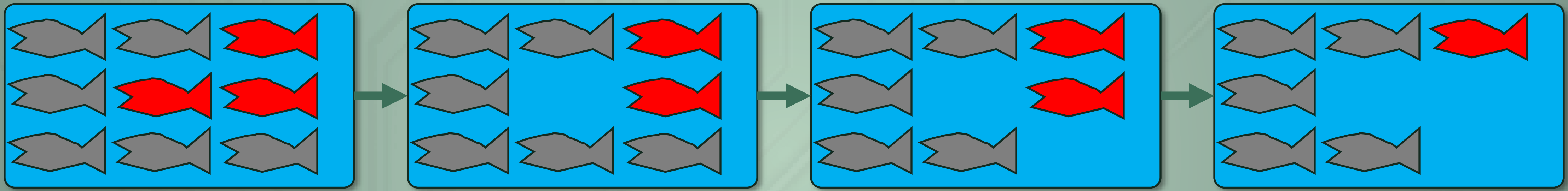
Binomial Probability: P(2 heads in 3 flips) = $C(3,2) * (1/2)^2 * (1/2)^1 = 3 * 1/8 = 3/8$

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

Hypergeometric distribution

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Example: In a lake there are 6 gray fishes and 3 red fishes.
What is the probability of selecting k red fish?



- Total number of fish (N):
  There are 9 fish in total
  → population size N = 9
- Number of red fish (K):
  3 red fish in the pool
  → "success" cases K = 3
- Number of grey fish:
  6 grey fish, no parameter,
  just N - K.
- Number of draws (n):
  Depends on how many fish are selected
- Number of successful draws (k)

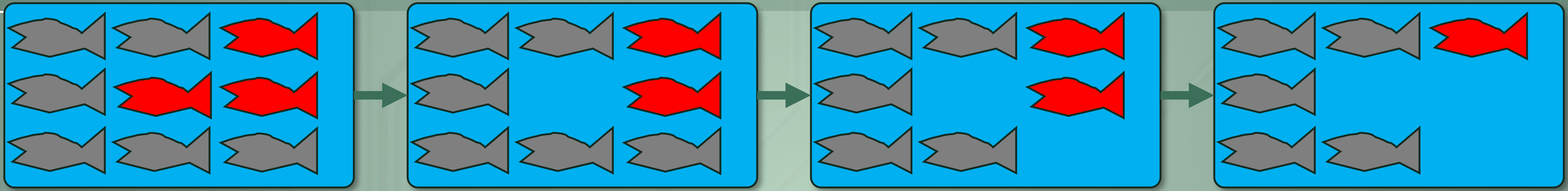
- Total number of fish (N):
  N = 8
- Number of red fish (K):
  K = 2
- Number of grey fish:
  N − K = 6

- Total number of fish (N):
  N = 7
- Number of red fish (K):
  K = 2
- Number of grey fish:
  N − K = 5

- Total number of fish (N):
  N = 6
- Number of red fish (K):
  K = 1
- Number of grey fish:
  N − K = 5

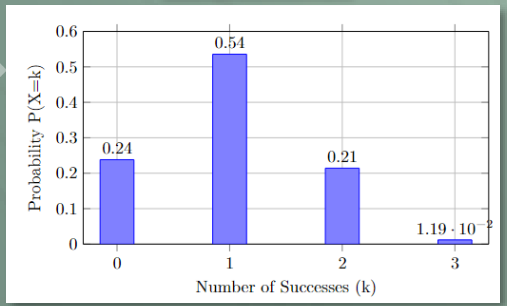
Hypergeometric distribution

Example: In a lake there are 6 gray fishes and 3 red fishes. Depending on the order of catching gray or red fish, probability distribution for next rounds changes



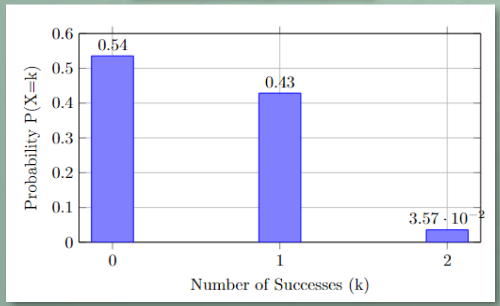
```
N = 9  # Total number of fish
K = 3  # Number of red fish
n = 3  # Number of draws (assumption)
```

k	P(X = k)
0	0.2381
1	0.5357
2	0.2143
3	0.0119



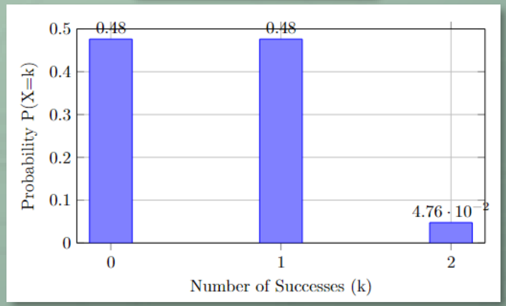

```
N = 8  # Total number of fish
K = 2  # Number of red fish
n = 2  # Number of draws (assumption)
```

k	P(X = k)
0	0.5357
1	0.4286
2	0.0357



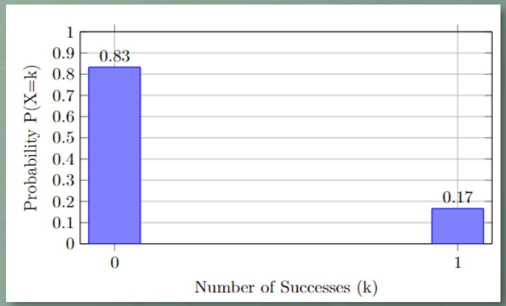

```
N = 7  # Total number of fish
K = 2  # Number of red fish
n = 2  # Number of draws (assumption)
```

k	P(X = k)
0	0.4762
1	0.4762
2	0.0476




```
N = 6  # Total number of fish
K = 1  # Number of red fish
n = 1  # Number of draws (assumption)
```

k	P(X = k)
0	0.8333
1	0.1667



Hypergeometric distribution

… models number of successes in sample drawn without replacement from finite population.

Probability mass function

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Where:

- N is the population size
- K is the number of success states in the population
- n is the number of draws
- k is the number of observed successes
- $\binom{a}{b}$ represents the binomial coefficient (combinations)

Connection to Conditional Probability

- Each draw changes the probability for subsequent draws:
- P(success on 2nd draw | success on 1st draw) = (K-1)/(N-1)
- This sequence of conditional probabilities forms the hypergeometric distribution.

Sampling Without Replacement

- Probability changes after each draw
- As N → ∞, approaches the binomial distribution

Use Cases

- Card game:
  - Consider deck of 52 cards (without jokers)
  - which includes 4 aces.
  - If you draw 5 cards without replacement, what is probability that exactly 2 of them are aces?
- Environmental Studies:
  - Ecologist is studying forest of 1,000 trees
  - 300 of which are of an endangered species.
  - Ecologist randomly samples 50 trees.
  - What is the probability that 15 of sampled trees are from endangered species?
  - What is the probability if new sample is collected that 20 more trees are from endangered species?

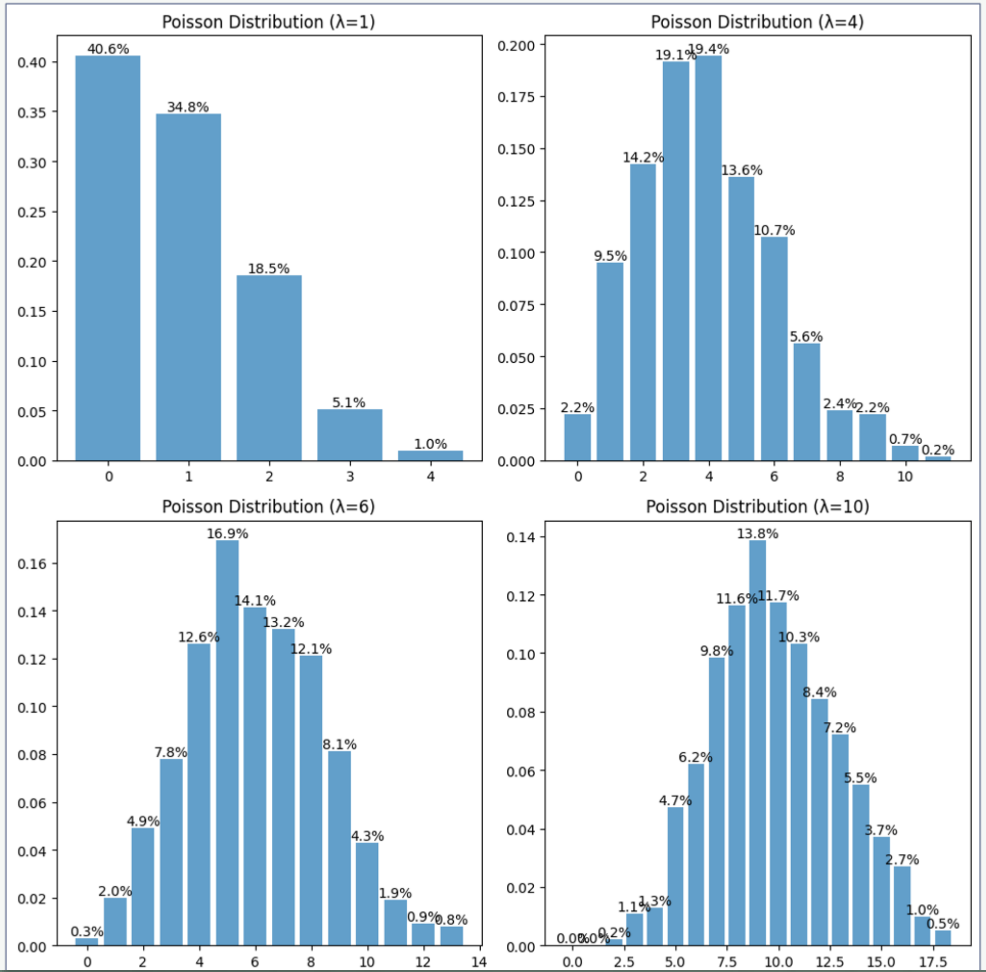
Hypergeometric vs. Binomial Distribution

As population size increases, hypergeometric distribution approaches binomial distribution. For large populations with small samples, binomial can approximate hypergeometric

Characteristic	Hypergeometric Distribution	Binomial Distribution
Sampling Method	Without replacement	With replacement
Population Size	Finite	Infinite or very large
Probability of Success	Changes after each draw	Constant for each trial
Probability Mass Function	P(X=k) = [C(K,k) * C(N-K,n-k)] / C(N,n)	P(X=k) = C(n,k) * p^k * (1-p)^(n-k)
Mean	n * (K/N)	n * p
Variance	n * (K/N) * ((N-K)/N) * ((N-n)/(N-1))	n * p * (1-p)
Trial Independence	Dependent	Independent

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)

Poisson Process

… is stochastic process that models occurrence of random events over time or space, where events occur continuously and independently at constant average rate

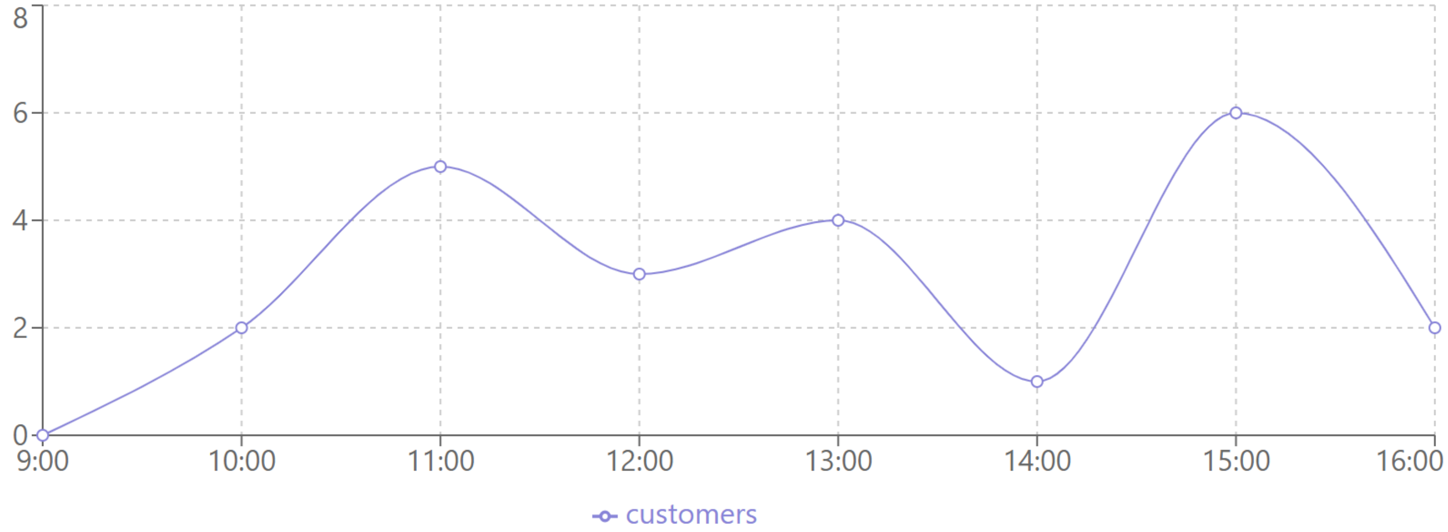
Poisson Process Characteristics:

- Random arrivals over time
- Customers arrive independently
- Average arrival rate (λ) is constant
- No simultaneous arrivals

Car Wash Application:

- Predict customer flow
- Optimize staff scheduling
- Manage queue lengths
- Plan for peak hours

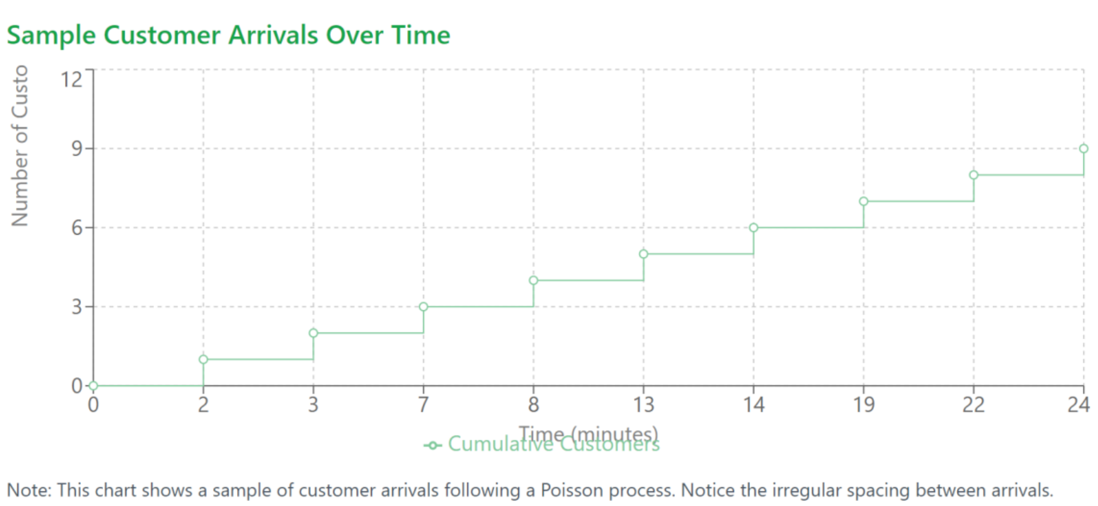
Customer Arrivals Throughout the Day

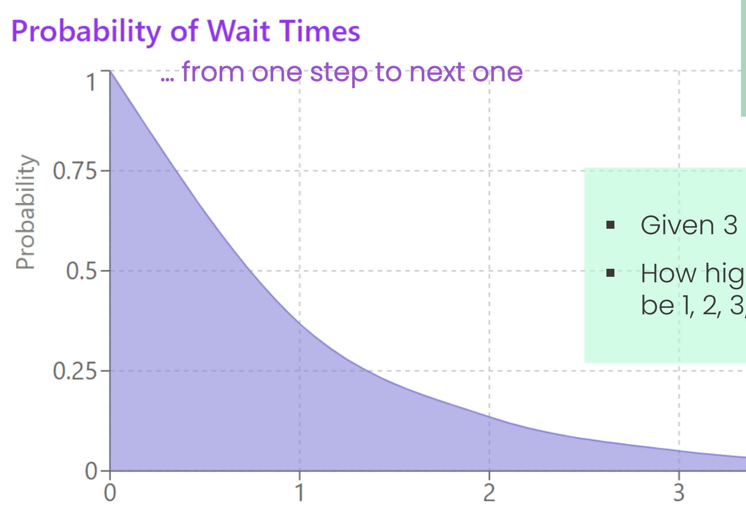


Note: This chart shows a sample of customer arrivals at a car wash following a Poisson process.

Poisson Processes: Inter-arrival Times and Variance

… is stochastic process that models occurrence of random events over time or space, where events occur continuously and independently at constant average rate





Later:
Compare this distribution with Gamma, Exponential, χ^2

- Given 3 events in average per time unit
- How high is probability that waiting time will be 1, 2, 3, 4, … time units (e.g. hours) or longer

- **Varying Wait Times:**
  Can range from very short to very long
- **Short vs. Long Waits:**
  Short waits more common, long waits possible, but rate
- **Average Wait Time:**
  1 / (average arrival rate)
- **Variance:**
  Equal to square of average wait time

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

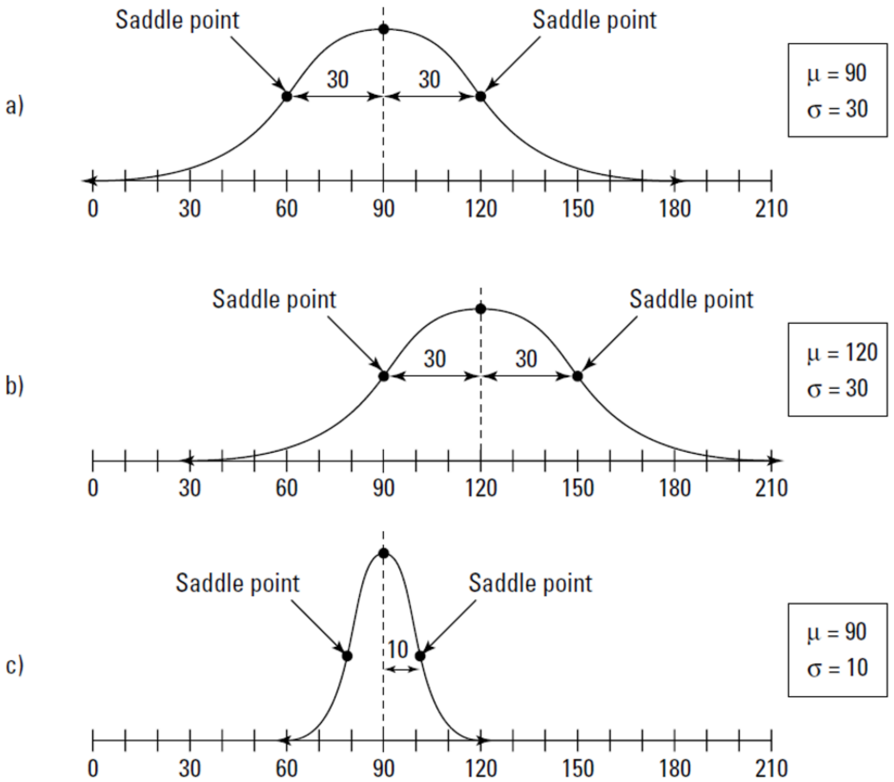


Figure 9-1: Three normal distributions.

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.

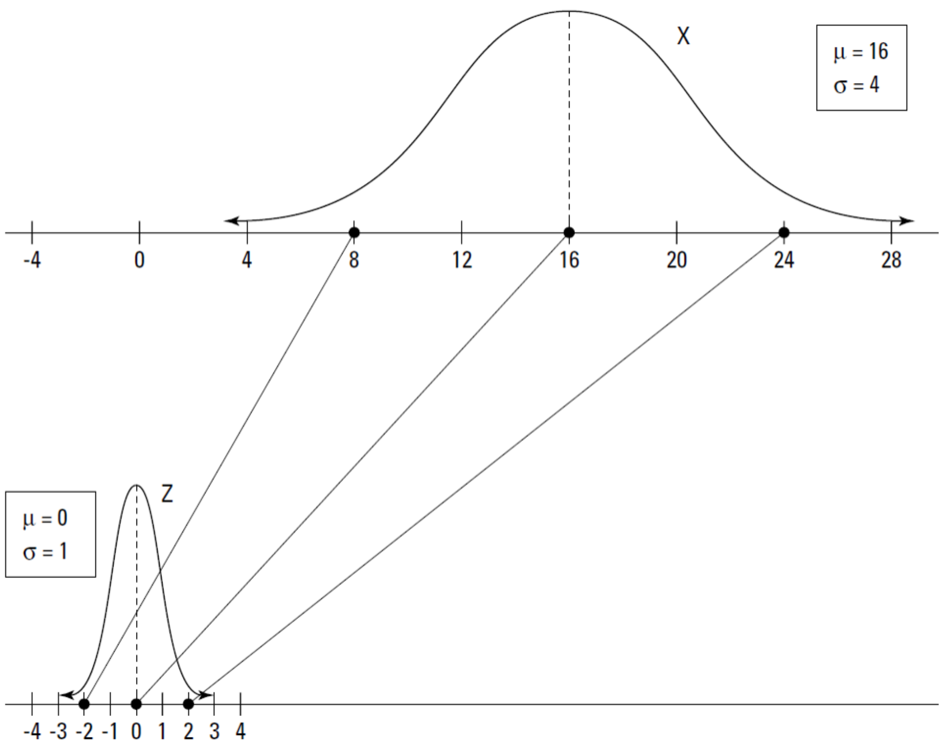


Figure 9-4: Transforming numbers on the normal distribution to numbers on the Z distribution.

z-scoring" or standardization

- subtracting the mean score
- and dividing by the standard deviation

Properties:

- mean (μ): 0
- standard deviation (σ): 1
- symmetric, bell-shaped

Why is it useful?

- This makes it easier to compare scores from different tests or datasets.

AI use case:

- Supervised learning: Prediction errors (differences between actual and predicted values) are often standard normally distributed, e.g. a house price predictor based area, number of rooms, etc.

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.

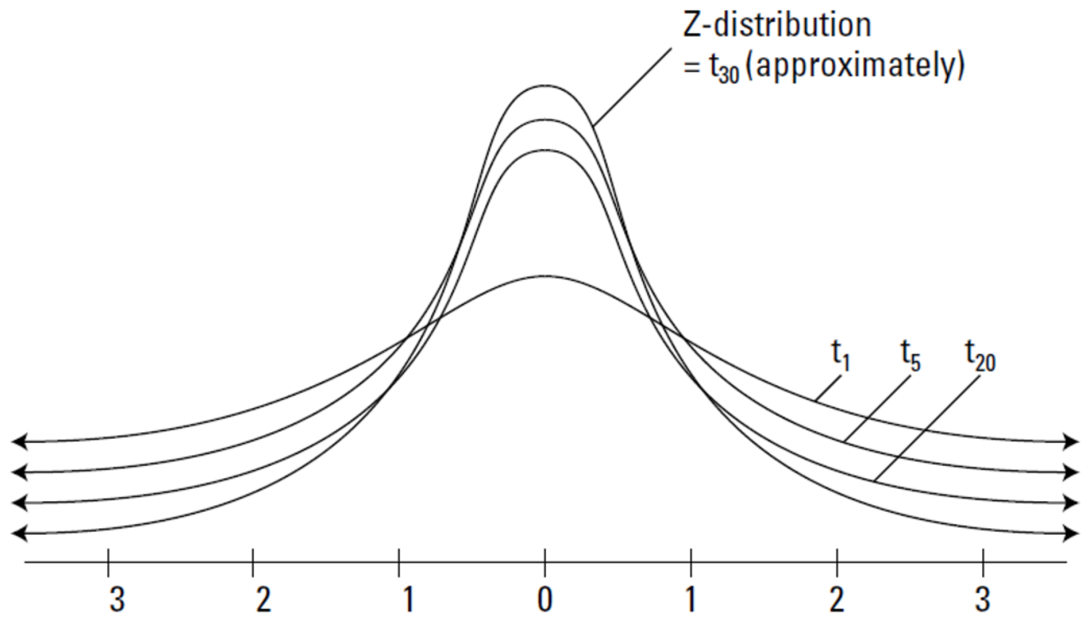


Figure 9-1: *t*-distributions for different sample sizes

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.

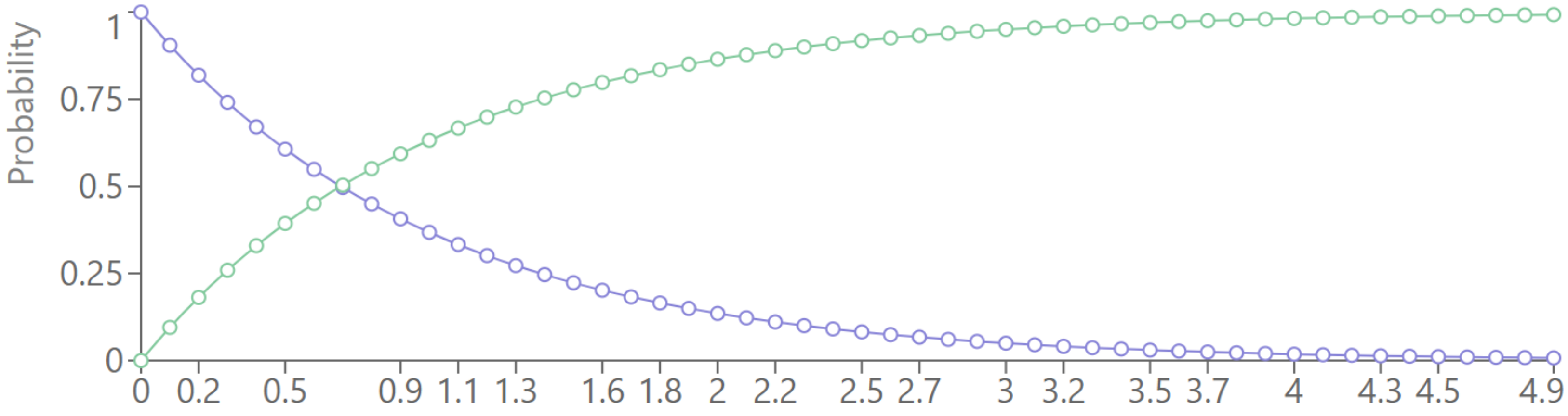
Exponential Distribution

Continuous probability distribution that describes time between events in Poisson process.

Use Cases

- Memoryless property: $P(X > s + t \mid X > s) = P(X > t)$
- Only parameter, constant hazard rate: λ
- Relationship with Poisson distribution

Probability Density Function (PDF) and Cumulative Distribution Function (CDF)



Example: $\lambda = 1$

Exponential vs Poisson Distribution

Exponential distribution models time between events in Poisson process, while Poisson distribution models number of events in fixed interval.

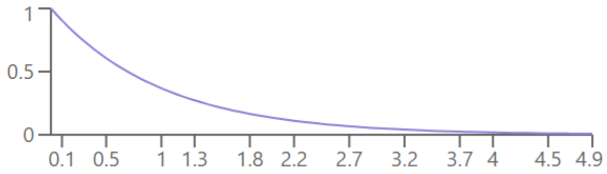
Example: Customer arrivals at a store

- Poisson: Number of customers arriving in one hour (λ = 3 customers/hour)
- Exponential: Time between customer arrivals (λ = 1/20 customers/minute or 3 customers/hour)
- If customer arrivals follow a Poisson process with rate λ, then the time between arrivals follows an Exponential distribution with rate λ

Exponential Distribution

Models time between events in a Poisson process

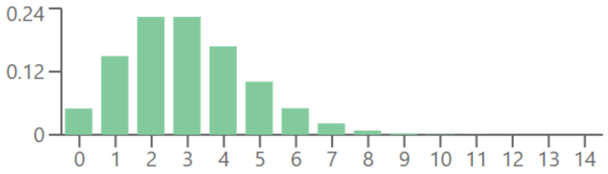
- PDF: f(x) = λe^(-λx) for x ≥ 0
- Mean: 1/λ
- Continuous distribution



Poisson Distribution

Models number of events in a fixed interval

- PMF: P(X = k) = (λ^k * e^(-λ)) / k!
- Mean: λ
- Discrete distribution



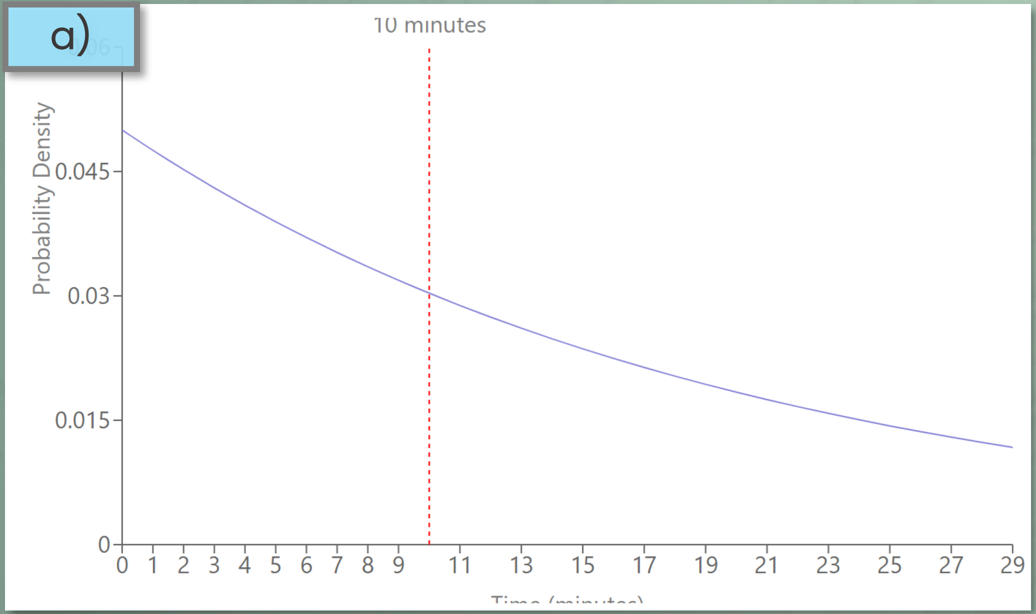
Exponential Distribution: Probability Calculation

λ = 1/20 customers/minute or 3 customers/hour

Case a) Probability of customer arrival within first 10 minutes

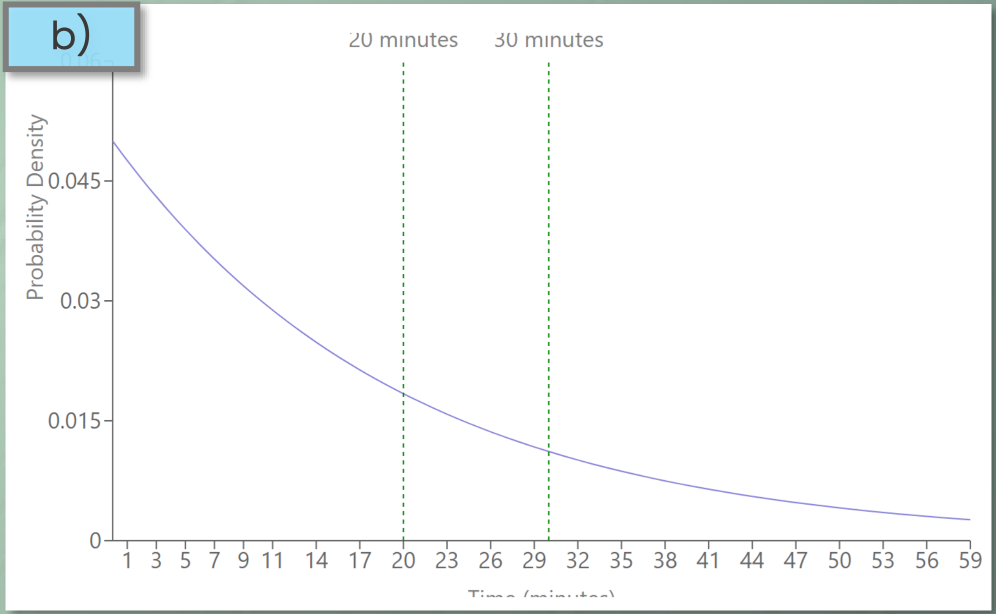
Case b) Probability of customer arrival between 20 and 30 minutes

a)



- Shaded area under curve represents probability of arrival within 10 minutes.
- Exponential curve shows how probability density decreases over time.
- Approximately 39.35% chance of a customer arriving within the first 10 minutes.

b)



- Shaded area between 20 and 30 minutes represents the probability of arrival in this interval
- Calculation uses difference of CDF: F(30) – F(20)
- Approximately 14.47% chance of customer arriving between 20 and 30 minutes.

Gamma Distribution

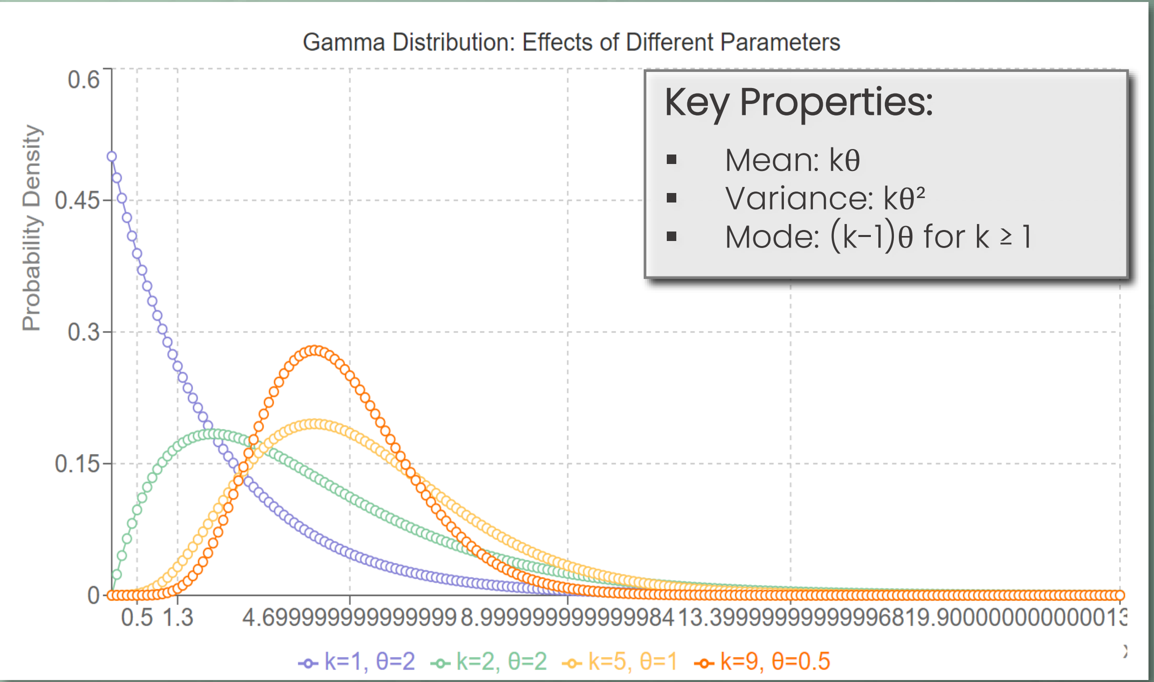
… is continuous, characterized by two parameters: shape (k) and scale (θ),modeling waiting times for k independent events in Poisson process, generalizing exponential distribution.

Parameters

- k (shape): Determines the basic shape of the distribution
- θ (scale): Stretches or shrinks the distribution along the x-axis

Examples and Interpretations

- k=1, θ=2 (Blue):
  Equivalent to an exponential distribution. Could model time between events in a Poisson process.
- k=2, θ=2 (Green):
  More bell-shaped. Might represent time until the second event in a process.
- k=5, θ=1 (Yellow):
  More symmetrical. Could model total processing time for 5 independent tasks.
- k=9, θ=0.5 (Orange):
  Nearly symmetrical, approaching normal distribution. Might represent a sum of many small, independent random variables.



Key Properties:

- Mean: kθ
- Variance: $k\theta^2$
- Mode: (k−1)θ for k ≥ 1

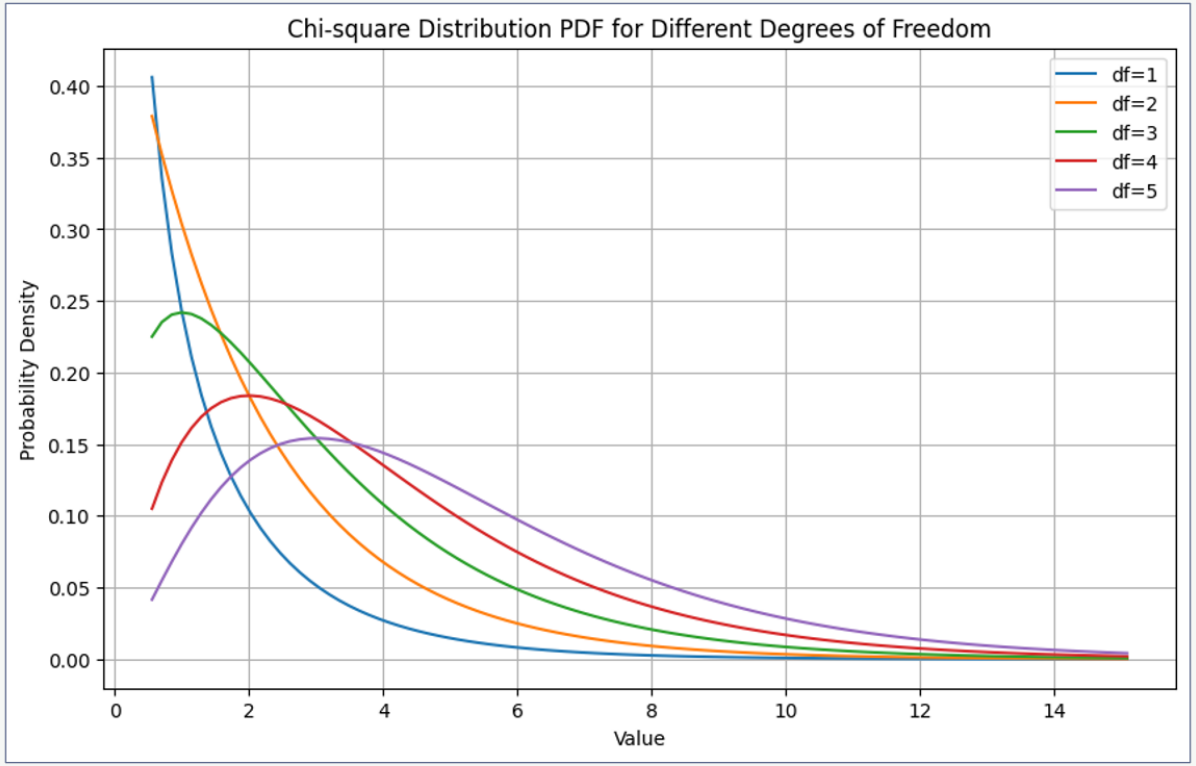
Connection of Poisson, Exponential, Gamma Distributions

Poisson counts events, Exponential measures time between events, and Gamma extends this to time until the k-th event, collectively modeling random processes.

Characteristic	Poisson	Exponential	Gamma
What it models	Number of events in a fixed interval	▪ Time between events ▪ Special case of Gamma: k=1	Time until k-th event
Parameters	λ (rate)	λ (rate)	k (shape), θ (scale) or λ (rate = $1/\theta$)
Support	Non-negative integers	Positive real numbers	Positive real numbers
Mean	λ	$1/\lambda$	$k\theta$
Variance	λ	$1/\lambda^2$	$k\theta^2$
Relationship to others	Inter-event times are Exponential(λ)	Special case of Gamma where k=1	Sum of k Exponential(λ) ~ Gamma($k,1/\lambda$)
Key property	Sum of independent Poisson is Poisson	Memoryless property	Reproductive property
Example application	Number of calls received in an hour	Time until next call	Time until 5th call

χ^2 distribution

… is a probability distribution of the sum of squares of independent standard normal random variables and is important particularly in the context of variance



https://en.wikipedia.org/wiki/Chi-squared_distribution

It's the distribution of a sum of the squares of k independent standard normal random variables:

$Q = \sum_{i=1}^{k} Z_i^2$

Parameter:

k (mean): independent standard normal random variables, square each and sum them up

Use cases:

- χ^2-tests of independence: They use χ^2 -distribution.
  Is preference for having cat or dog is independent from city?
- Several other hypothesis tests, particularly connected with variance

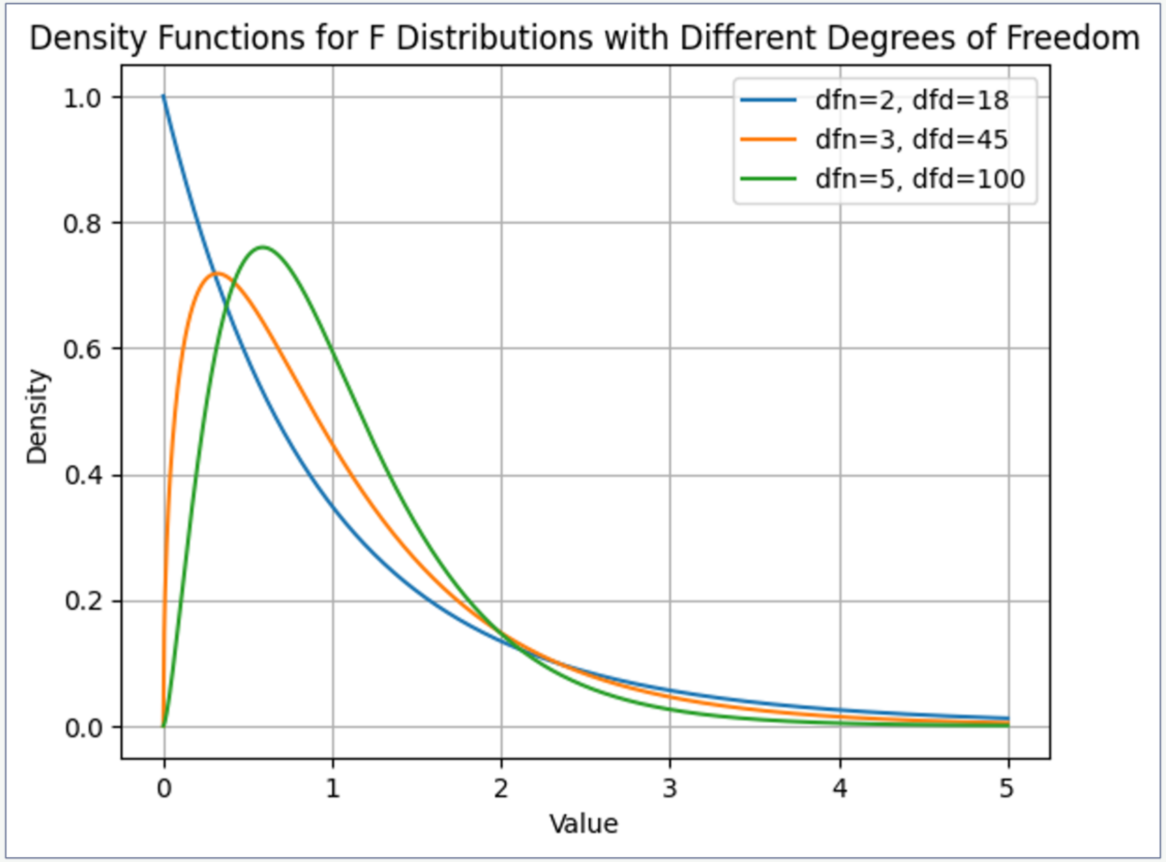
Connection of χ^2 and Gamma Distributions

$\chi^2(k)$ is equivalent to Gamma(k/2, 2), where k is degrees of freedom, i.e. number of occurences

Feature	Chi-Squared (χ^2)	Gamma
Parameters	k (degrees of freedom)	α (shape), θ (scale)
Relationship	▪ $\chi^2(k)$ = Gamma(k/2, 2) ▪ Special case of Gamma with scale = 2	
Support	[0, ∞)	[0, ∞)
Mean	k	$\alpha\theta$
Variance	2k	$\alpha\theta^2$
Skewness	$\sqrt{(8/k)}$	$2/\sqrt{\alpha}$
Common Applications	Hypothesis testing, goodness-of-fit tests	Modeling waiting times, rainfall amounts
Flexibility	Less flexible, shape determined by k	More flexible, can model various shapes

F-distribution

… is ratio of 2 independent χ^2-distributions divided by their respective degrees of freedom



$$X = \frac{S_1/df_1}{S_2/df_2}$$

Degrees of freedom (df):

- dfn (df of numerator):
  number of groups being compared minus 1
  between group comparison
- dfd (df of denominator):
  total number of observations minus the number of groups
  within group comparison

Use Cases

- is used in Analysis of Variance (ANOVA)
- comparing variances of $\geq$ 3 samples
- Evaluation of regression results, overall model significance

https://en.wikipedia.org/wiki/F-distribution

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?

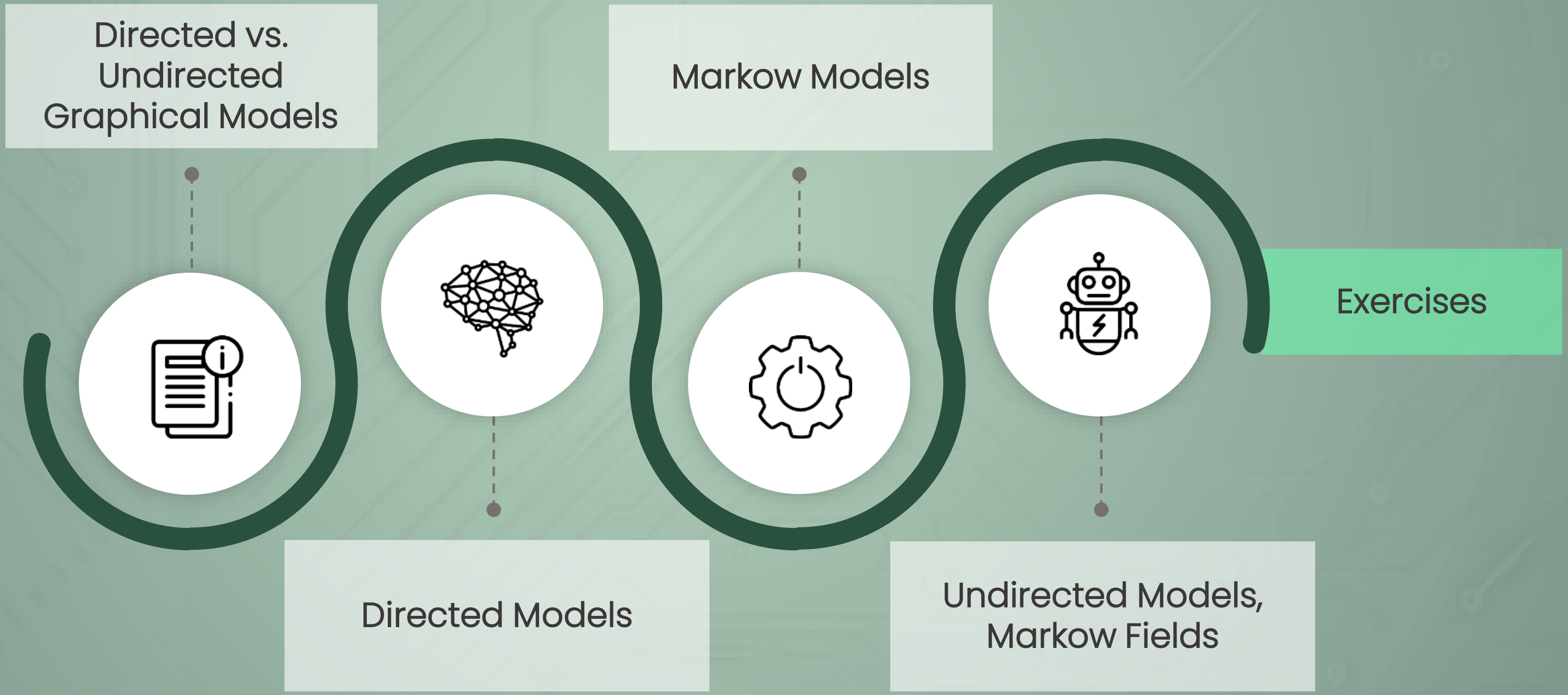
Exercise 5:
Problem: An IQ test is scored such that the mean score is 100 and the standard deviation is 15. What is the probability of a person scoring higher than 130?

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?

Extra-exercises: Relevant use cases are in context of hypotheses and model estimation
Exercise 7:
Suppose the variance of a sample of 20 observations is 5. What is the probability that the sample variance is greater than 6?
Exercise 8:
If two groups of data are sampled from normal distributions with the same variance, the ratio of their sample variances will follow an F-distribution. Suppose you have two samples of sizes 15 and 20 with variances 4 and 2, respectively. What is the probability that the ratio of these sample variances is less than 1?

Structured Probabilistic Models (Graphical Models)

… are powerful mathematical frameworks that use graphs to represent, visualize probabilistic relationships among set of variables, enabling efficient reasoning, inference in complex systems



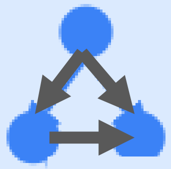
Structured Probabilistic Models

… use graph theory to intuitively and efficiently represent and reason about uncertainty, capturing complex variable dependencies better than traditional probabilistic methods.

Directed Models

- Also known as: Bayesian Networks
- Represent asymmetric relationships
- Use arrows to show direction of influence
- Model causal and temporal relationships
- Example: Disease → Symptom

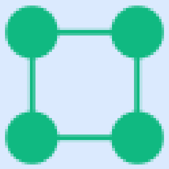
Capture causal or hierarchical relationships



Undirected Models

- Also known as: Markov Random Fields
- Represent symmetric relationships
- Use lines without arrows
- Model mutual dependencies
- Example: Protein interactions

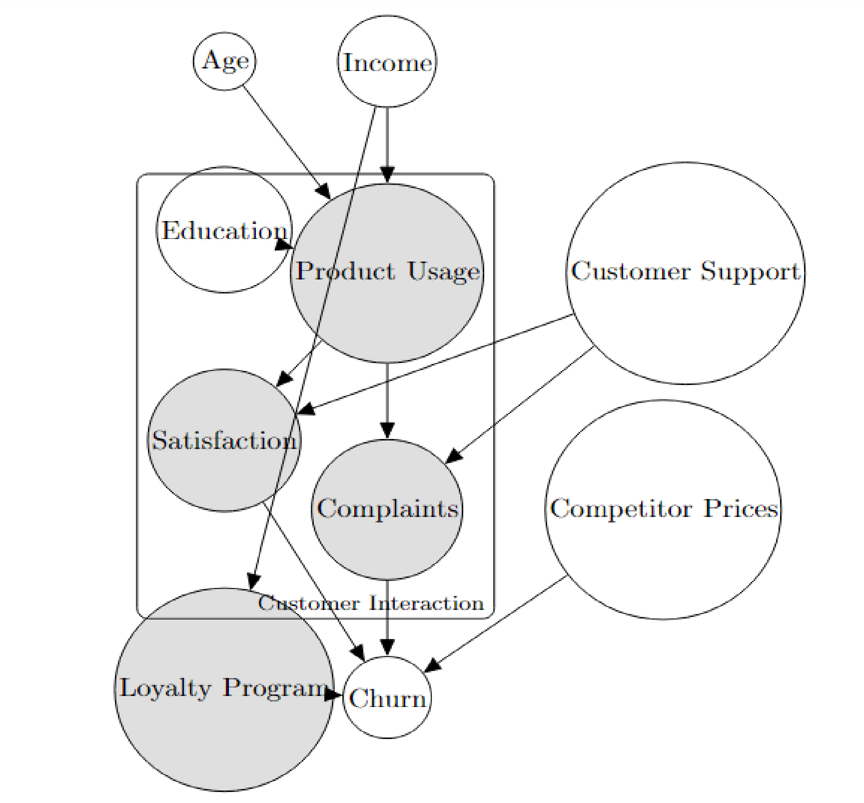
Represent mutual or symmetric interactions



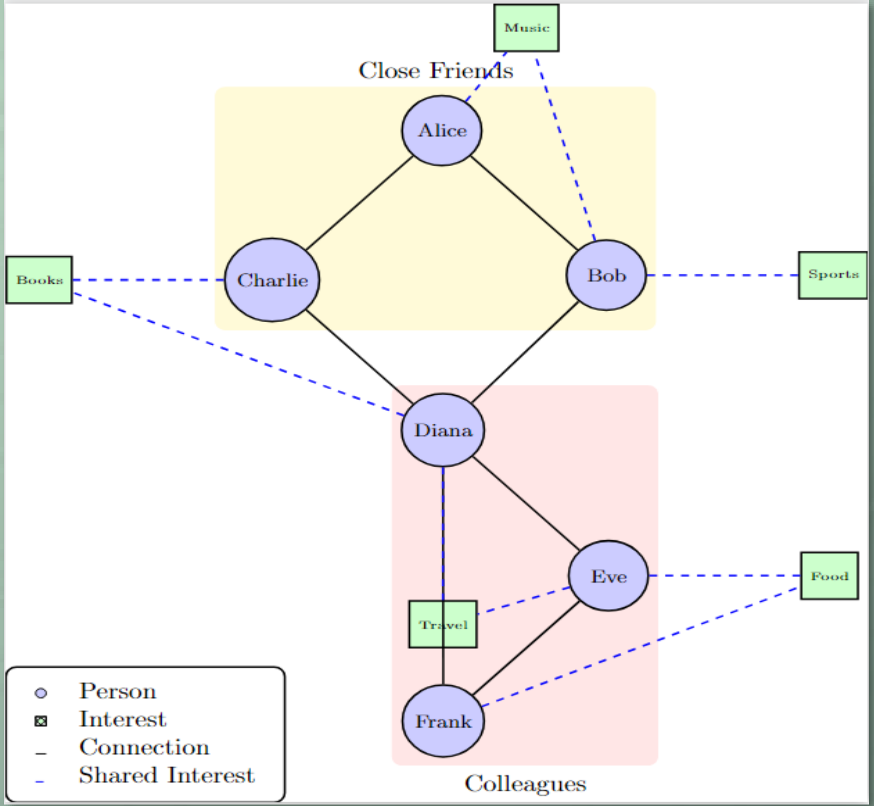
Directed vs. Undirected Graphical Models

Directed models excel in causal inference and decision-making, while undirected models are preferred for symmetric relationships like in image processing or social networks

Directed Graphs:
- Bayesian Networks
- Use Case: Customer Churn Prediction



Undirected Graphs:
- Markov Random Fields
- Use Case: Social Network



Directed Graphical Model: Home security alarm

… Use Case: Home security alarm can be triggered by burglaries or false alarms. What's the probability of an actual burglary when the alarm is triggered?

What is P(B/A)? 0.1919

Conditional Probabilities

Event	Probability
P(B) - Prior probability of burglary	0.001
P(F) - Probability of factors causing false alarm	0.02
P(A/B) - Probability of alarm given burglary	0.95
P(A/F) - Probability of alarm given false trigger	0.20

Answer

P(A/B)*P(B) P(A/F)*P(F)

P (A) = (0.95∗0.001)+(0.20∗0.02)
= 0.00095+0.004
= 0.00495

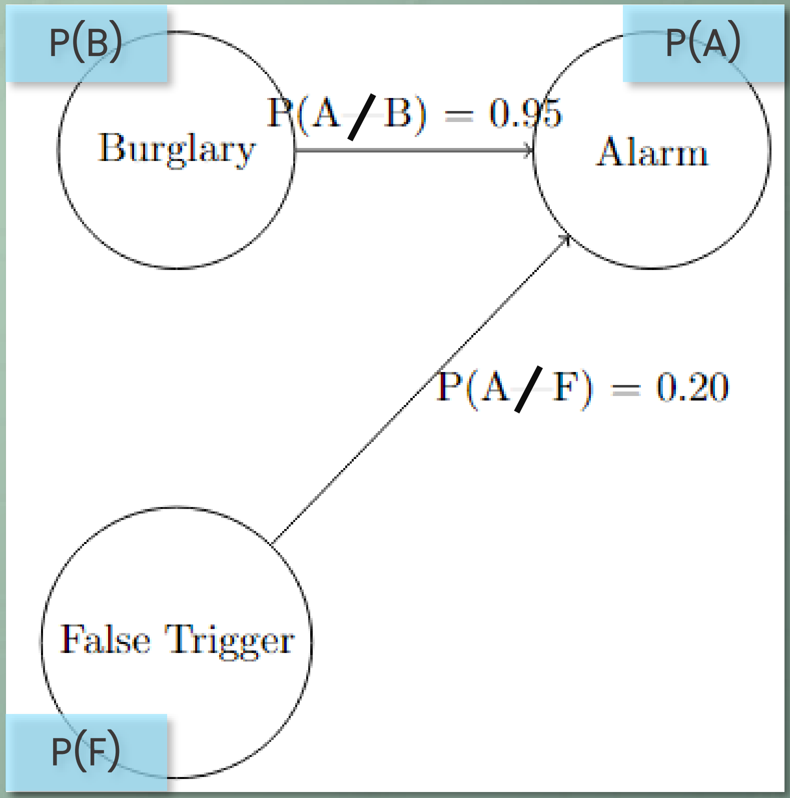
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

P(A/B)*P(B)

$$P(B|A) = \frac{0.95 * 0.001}{0.00495} \approx 0.1919$$

P(A)

Directed Graphical Model



Directed Graph Model: Use Cases

They are commonly used in scenarios where causality and conditional dependencies between variables are essential

Web Page Ranking

Model hyperlinks between web pages for search engine algorithms

Dependency Analysis

Model software package dependencies or project task dependencies

Recommendation Systems

Model user preferences and item relationships for personalized recommendations

Flow Networks

Represent material flow in supply chains or traffic flow in transportation

Citation Networks

Represent citations between academic papers or patents

Gene Regulatory Networks

Represent interactions and regulatory relationships between genes

Causal Inference

Represent cause-effect relationships in various domains

Social Influence Analysis

Model information propagation and influence in social networks

Markov Models

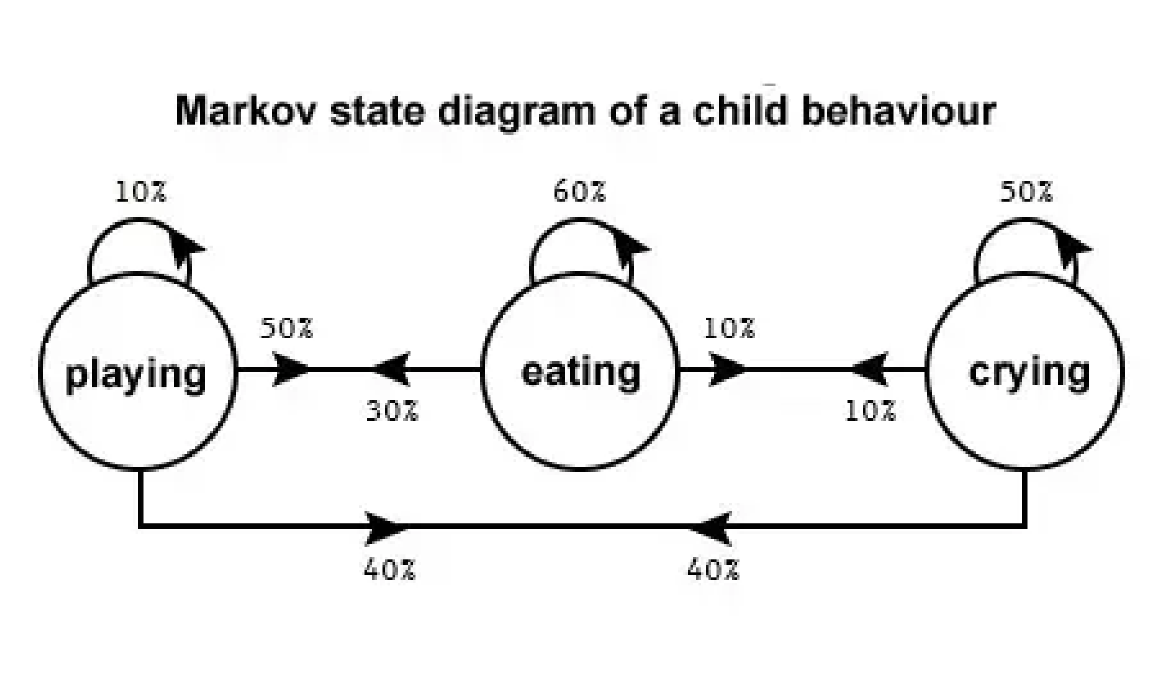
… are mathematical system that transitions between states with probabilities that depend only on current state

Key Characteristics

- Models random processes
- Exhibits the Markov property (memoryless)
- Named after Andrey Markov (1856 – 1922, Russia)
- Consists of states and transition probabilities

Markov Property Explained

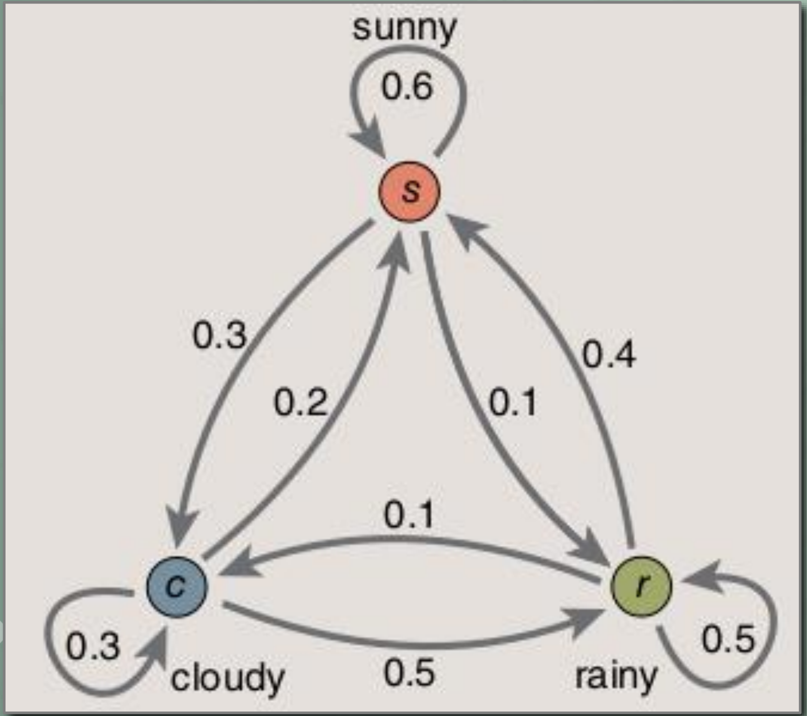
The probability of transitioning to any particular state depends solely on the current state and time elapsed, not on the sequence of states that preceded it.

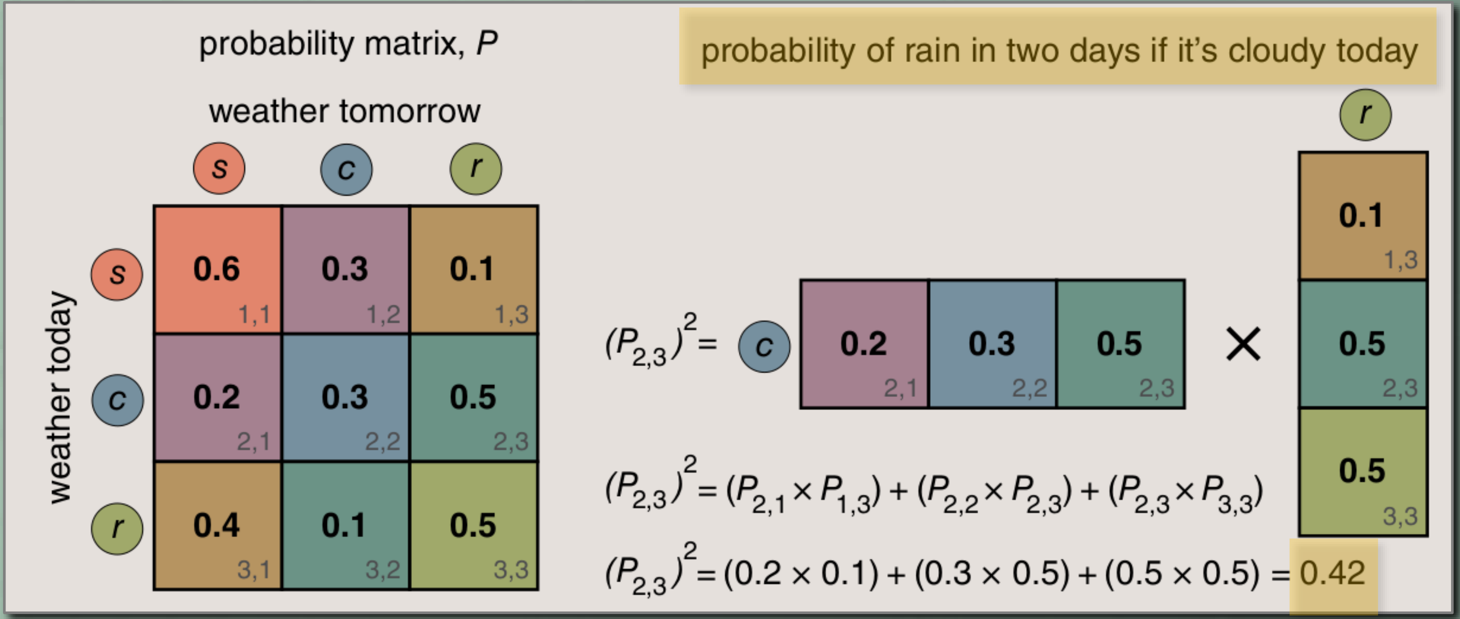


Markov Models

… Use Case: Weather prediction

How probable is rain in two days after cloudy day? Attention: Not 2 days of rain in a row!

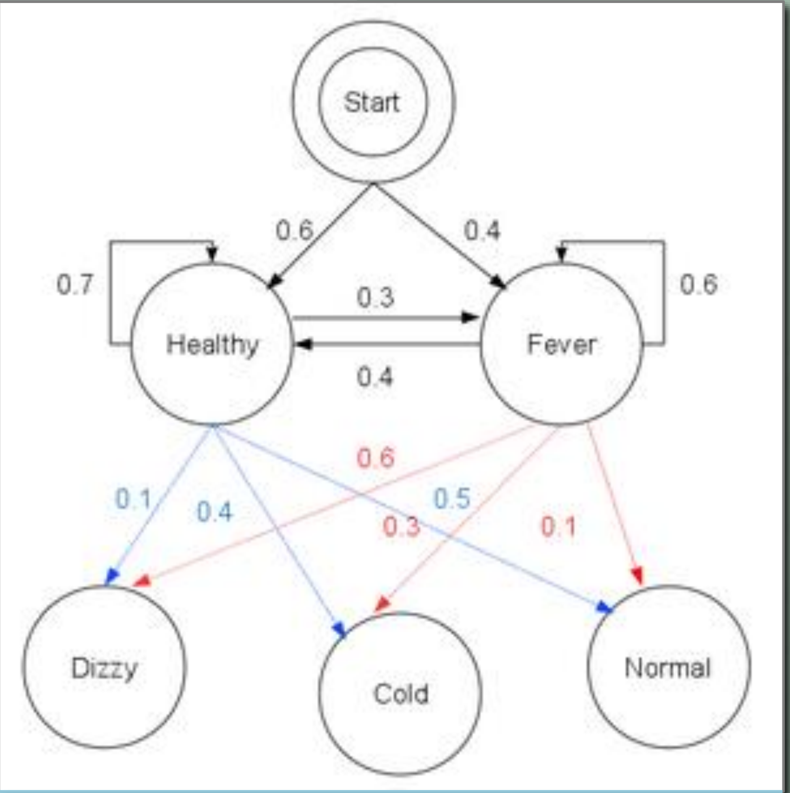




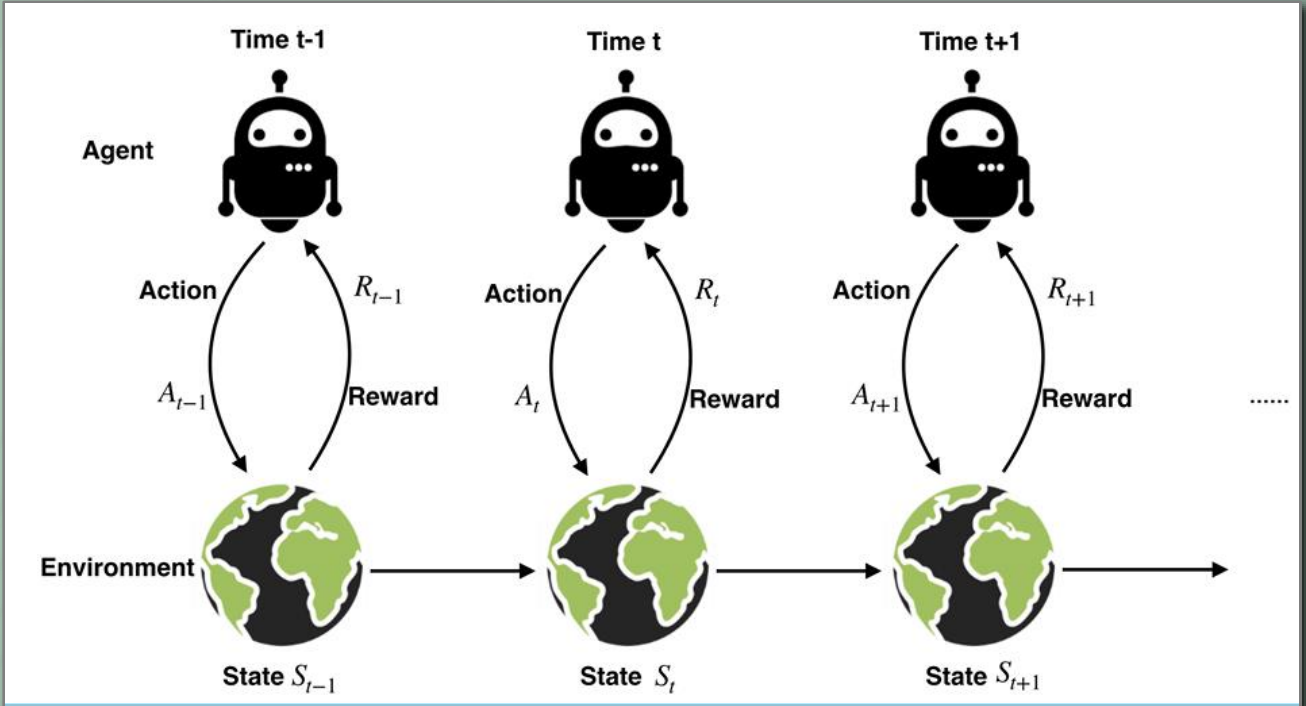
https://www.linkedin.com/pulse/weather-forecast-markov-chain-carlos-serra-traynor/

Advanced Markov Model Approaches

Hidden State Markow Model: have unobservable state. Only some outputs can be observed.
Markow Decision Processes: Optimal policy that maximized rewards needs to be found



Hidden State Markow Model
It is unknown whether patient is healthy or ill



Markow Decision Processes (for Reinforcement Learning)
Robot interacts with environment and explores optimal policy through reward and punishment

(Hidden) Markov Models, Markov Decision Processes

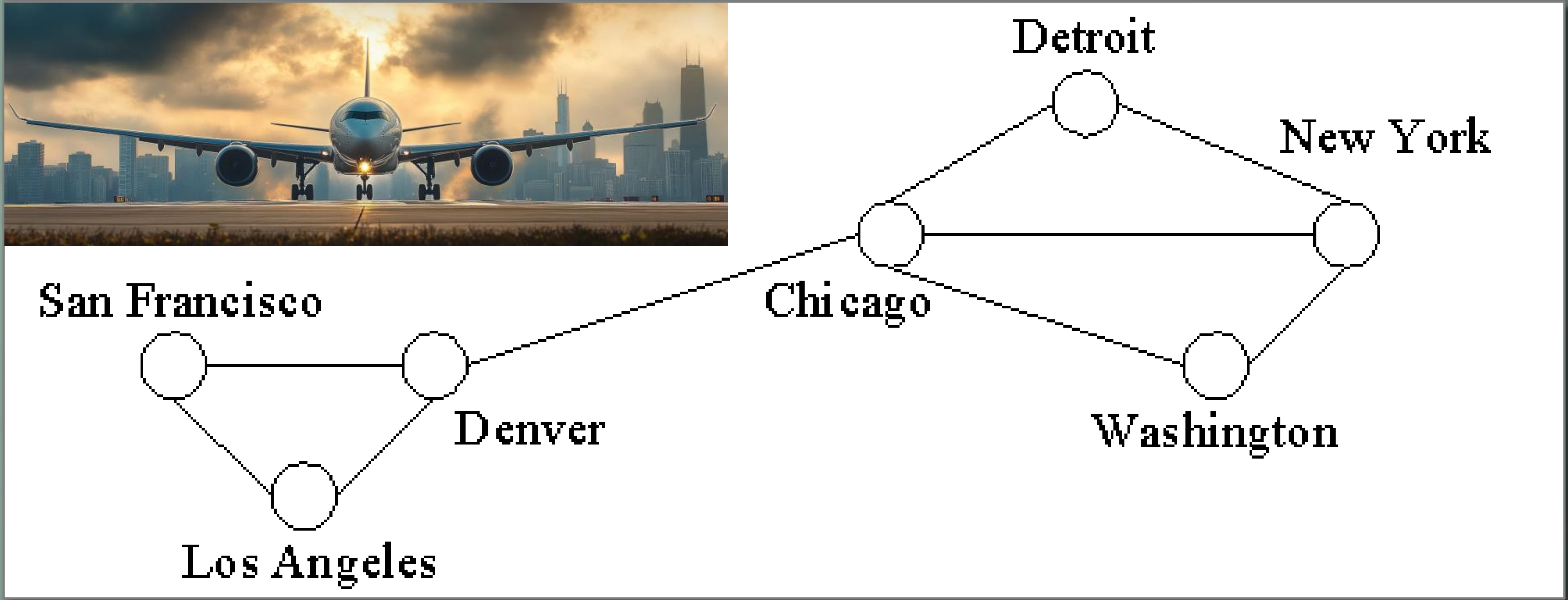
Comparison

Characteristic	MarkovModel (MM)	Hidden MarkovModel (HMM)	Markov Decision Process (MDP)
States	Observable	Hidden	Observable
Observations	Same as states	Emissions from states	Same as states
Transitions	Probabilistic	Probabilistic	▪ Probabilistic ▪ action-dependent
Decision Making	No	No	Yes
Key Components	▪ State space ▪ Transition probabilities	▪ Hidden state space ▪ Observable spac ▪ Transition probabilitie ▪ Emission probabilities	▪ State space ▪ Action space ▪ Transition probabilities ▪ Reward function
Main Purpose	State prediction	State inference	Decision optimization
Example	Weather forecasting	Speech recognition	Robot navigation

Undirected Graphical Model

… are mathematical structures used to represent relationships between objects where connections have no inherent direction. Causality is mutual

Example: Travel connections of airline



- Edges are bidirectional, allowing travel in both directions.
- Represented as G = (V, E), where V is the set of vertices and E is the set of edges.
- Properties include connectivity, degree (number of edges connected to one vertice), cycles.

Undirected Graph Model: Use Cases

… are used when modeling relationships or connections that are inherently symmetric or bidirectional, where the direction of the relationship is not important or meaningful

Computer Networks

Model network topology and connectivity between devices

Image Segmentation

Partition images into meaningful regions for computer vision tasks

Collaborative Filtering

Power recommendation systems based on user-item interactions

Financial Markets

Model correlations between different financial instruments

Power Grids

Model electricity distribution networks

Social Network Analysis

Model friendships, collaborations, and information flow in social groups

Ecological Networks

Represent species interactions in ecosystems

Transportation Networks

Model road networks, flight connections, or public transit systems

Collaboration Graphs

Represent joint authorship in academic publications

Molecular Structures

Represent chemical compounds and protein interactions

Typical Calculations in Undirected Graphs

Shortest Path

Find the path with minimum total edge weight between two nodes

Connectivity

Determine if the graph is connected or find connected components

Centrality Measures

Identify important nodes (e.g., degree, betweenness, closeness centrality)

Minimum Spanning Tree

Find a tree that connects all nodes with minimum total edge weight

Graph Clustering

Identify groups of densely connected nodes

Maximum Flow

Calculate the maximum flow between source and sink nodes

Exercises

Exercise 1:

Maritime Logistics Probabilistic Model Consider a directed graph model for predicting successful cargo delivery in a maritime logistics network:

W: Weather conditions (1: Favorable, 2: Moderate, 3: Severe)
P: Port congestion (1: Low, 2: Medium, 3: High)
S: Ship maintenance status (1: Excellent, 2: Average, 3: Poor)
D: Delivery outcome (1: On-time, 2: Delayed, 3: Significantly delayed)\

The directed graph structure is as follows:
W -> P
W -> D
P -> D
S -> D
Given the following probabilities:
$P(W = 1) = 0.5$, $P(W = 2) = 0.3$, $P(W = 3) = 0.2$
$P(S = 1) = 0.6$, $P(S = 2) = 0.3$, $P(S = 3) = 0.1$

$P(P = 1 \mid W = 1) = 0.7$, $P(P = 2 \mid W = 1) = 0.2$, $P(P = 3 \mid W = 1) = 0.1$
$P(P = 1 \mid W = 2) = 0.3$, $P(P = 2 \mid W = 2) = 0.5$, $P(P = 3 \mid W = 2) = 0.2$
$P(P = 1 \mid W = 3) = 0.1$, $P(P = 2 \mid W = 3) = 0.3$, $P(P = 3 \mid W = 3) = 0.6$

$P(D = 1 \mid W = 1, P = 1, S = 1) = 0.9$, $P(D = 2 \mid W = 1, P = 1, S = 1) = 0.1$, $P(D = 3 \mid W = 1, P = 1, S = 1) = 0.0$
$P(D = 1 \mid W = 2, P = 2, S = 2) = 0.5$, $P(D = 2 \mid W = 2, P = 2, S = 2) = 0.4$, $P(D = 3 \mid W = 2, P = 2, S = 2) = 0.1$
$P(D = 1 \mid W = 3, P = 3, S = 3) = 0.1$, $P(D = 2 \mid W = 3, P = 3, S = 3) = 0.3$, $P(D = 3 \mid W = 3, P = 3, S = 3) = 0.6$

Tasks:\

Visualize the directed graph for this maritime logistics model.
Calculate $P(D = 1 \mid W = 1, S = 1)$, the probability of on-time delivery given favorable weather and excellent ship maintenance.
Discuss how this model could be used to improve maritime logistics planning.

Exercises

Exercise 2:

Consider a simple weather model for a particular location with three possible states:

1. Sunny (S)\
2. Cloudy (C)\
3. Rainy (R)

The weather follows a Markov chain model with the following transition matrix:

```
          To: S    C    R\
From: S [0.7, 0.2, 0.1]\
      C [0.3, 0.4, 0.3]\
      R [0.2, 0.3, 0.5]
```

For example, if it's Sunny today, there's a 70% chance it will be Sunny tomorrow, 20% chance it will be Cloudy, and 10% chance it will be Rainy.

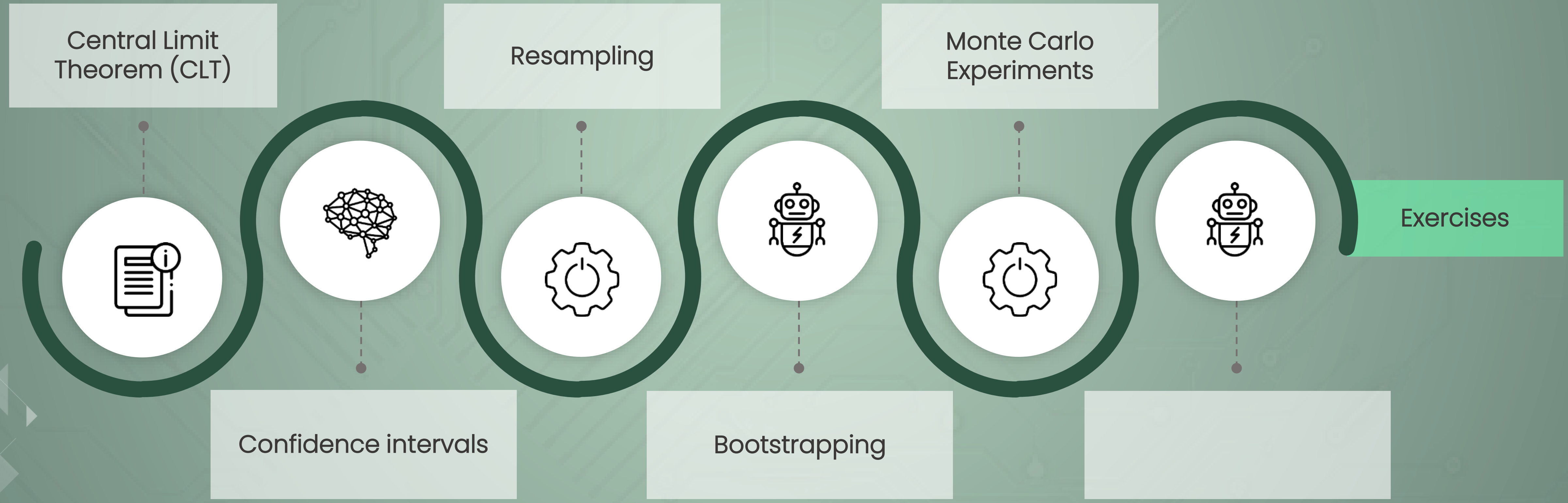
Tasks:\

1. Visualize the Markov chain as a directed graph.\
2. If it's Sunny today, what's the probability it will be Rainy two days from now?\
3. Calculate the steady-state probabilities for each weather state.\
4. Simulate the weather for the next 30 days, starting from a Sunny day.

Not 2 days of rain in a row!

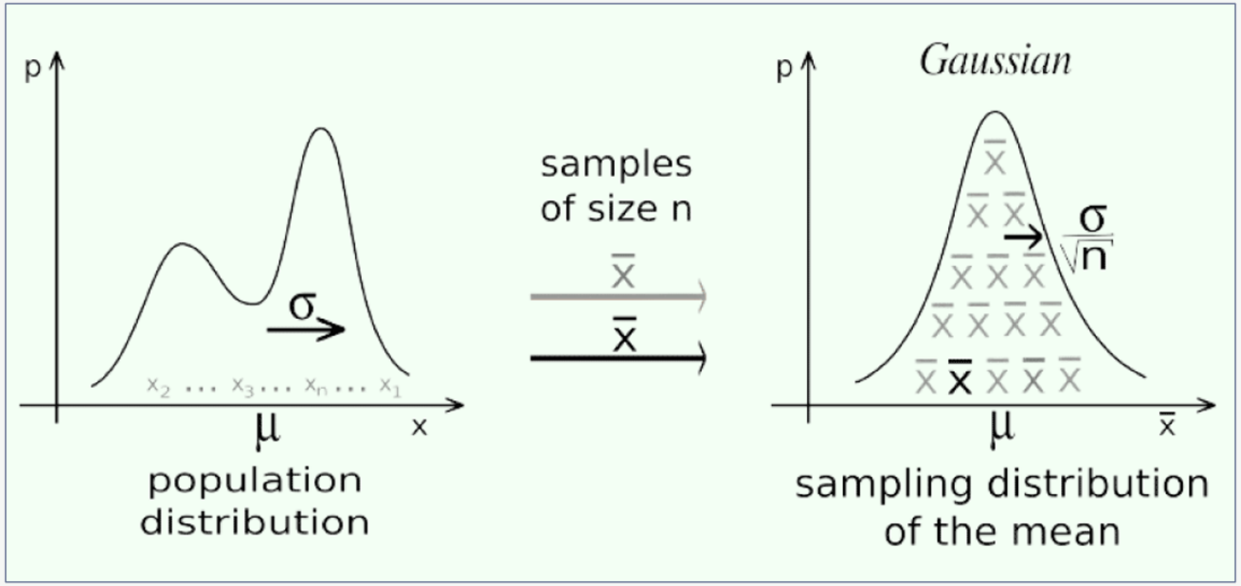
Sampling, Monte Carlo Experiments

… is the process of selecting a subset of individuals from a statistical population to estimate characteristics of the whole population



Central limit theorem

The distribution of sample means approximates a normal distribution as the sample size becomes larger, regardless of the population's distribution.



Definition:

- For population with mean μ and standard deviation σ
- take sufficiently large random samples from the population
- with replacement (independent)
- Independent and identically distributed (i.i.d) random variables
- Sample size is large enough (typically $n > 30$)

→ distribution of sample means is appr. normally distributed, regardless of population distribution

Why is it important?

- This principle enables us to make statistical inferences about populations based on sample data
- foundation for hypothesis testing, confidence intervals, etc.

Confidence intervals

… is range of values that is likely to contain an unknown population parameter with certain level of confidence, widely used e.g. for hypothesis testing

For large sample size

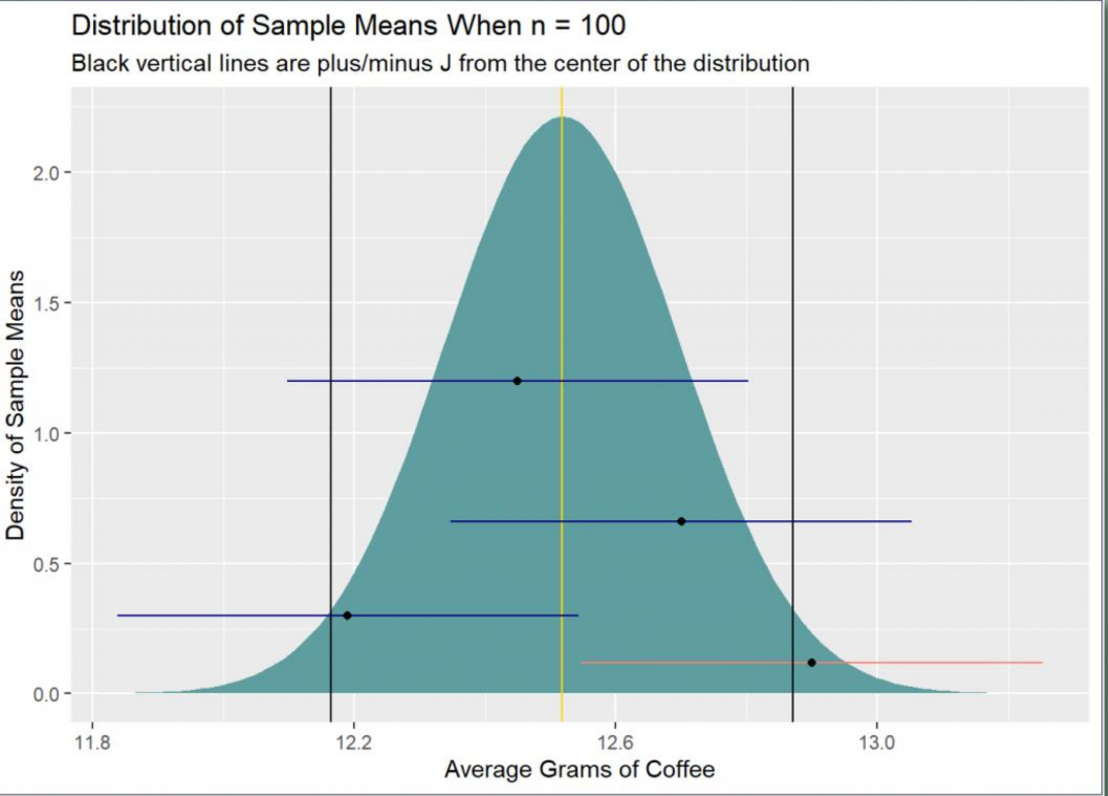
$$\mathrm{CI} = \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Where:

- $\bar{x}$ is the sample mean
- $z_{\alpha/2}$ is the z-score for the chosen confidence level
- σ is the population standard deviation
- n is the sample size

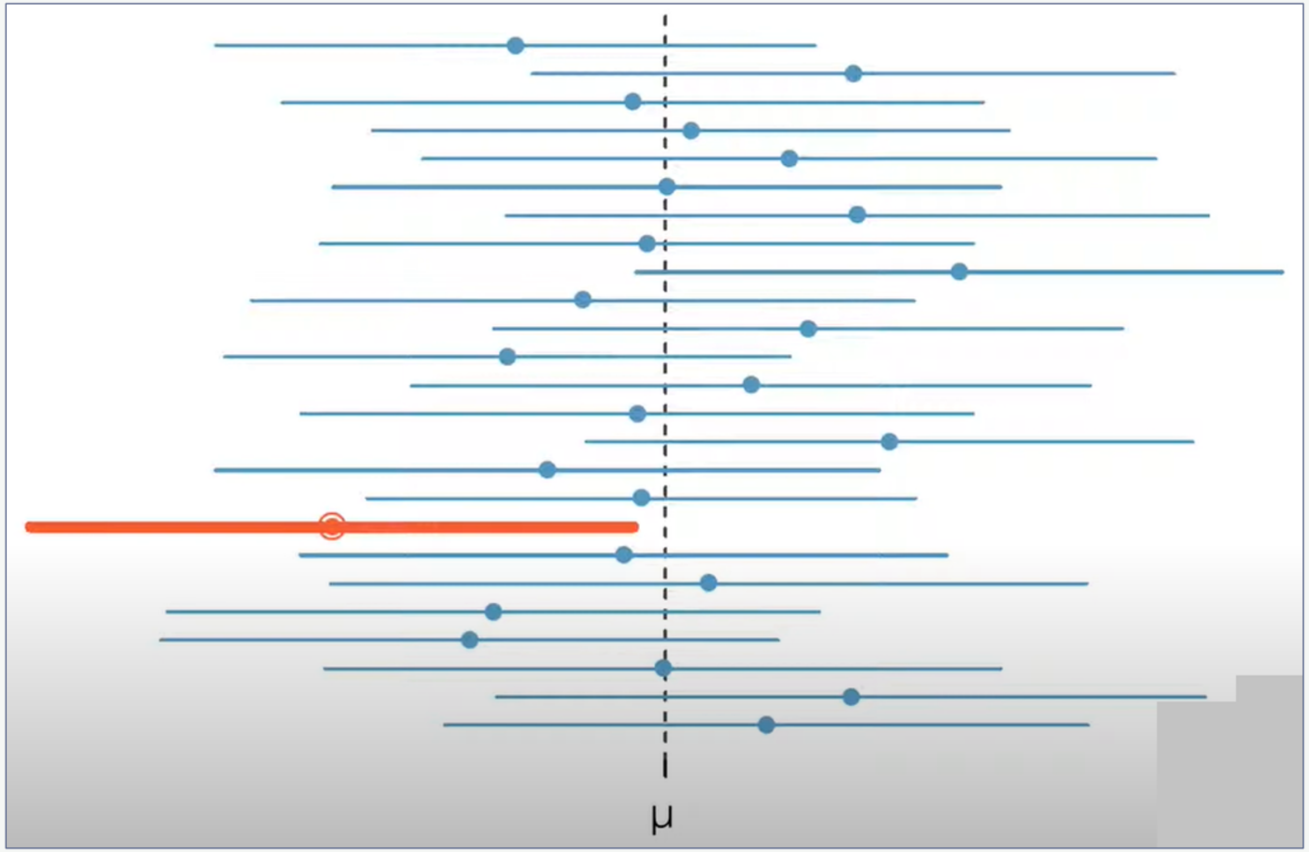
For small sample size

- Similar formula is used
- however not based on standard normal distribution but t-distribution



Confidence intervals

… need careful interpretation. Assumption: α=5%



Assumption:
population parameter μ (average of population) is fixed

Wrong interpretation:

- With a chance of 95% the confidence interval contains μ

Correct interpretation:

- If sampling process is repeated many times or infinitely
- confidence intervals are generated each time
- 95% of these confidence intervals contain μ

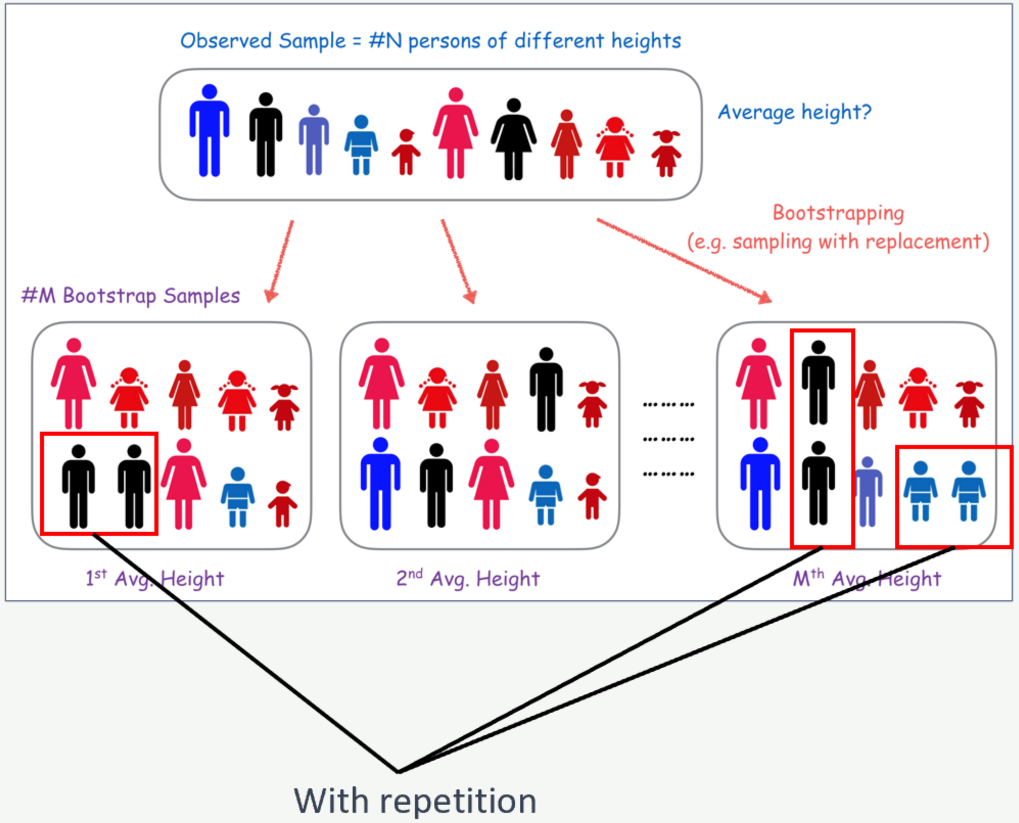
Resampling

… Is set of methods of statistical inference that involve drawing repeated samples from original data samples

What is it good for?	• Versatile method to make statistical inferences about a dataset • It helps to validate models by using random subsets (bootstrap) or sequences (cross-validation) dataset • It extends possibilities for hypothesis testing and estimation
Example: Bootstrap	• Many resamples (with replacement) of the observed dataset are generated • Arbitrary distribution is transformed into normal distribution • convenient for estimation and hypothesis testing.
Example: Cross-validation	• a form of resampling used to prevent overfitting in machine learning models • dataset is divided into 'k' subsets; used for training the model • respective residual data is used as test set

Bootstrapping

… is a resampling method that involves drawing repeated samples from the original data set, with replacement. Each sample drawn is the same size as the original dataset.



Procedure:

- Draw sample from the original dataset with replacement
- Often of same size like original dataset
- Compute desired statistic from this resample
- Repeat the process multiple times (often thousands or tens of thousands of times), each time storing the statistic.
- Use distribution of these bootstrapped statistics to make inferences about the population.

Why does it make sense?

- It provides method to estimate variability or confidence intervals of statistic when underlying distribution is unknown or complex.
- It's a simple and powerful method requiring very few assumptions

AI Use case:

- Random forests (like XGBoost):
  Classification or numeric prediction
  based on many resampled pruned decision trees

Monte Carlo Simulations

… are statistical technique using repeated random sampling to obtain numerical results

Process
1. Define input domain 2. Generate random inputs 3. Perform computations 4. Aggregate results

Applications
Risk analysis, optimization, integration, physics simulations, finance

Advantages
Can handle complex systems with many variables and uncertainties

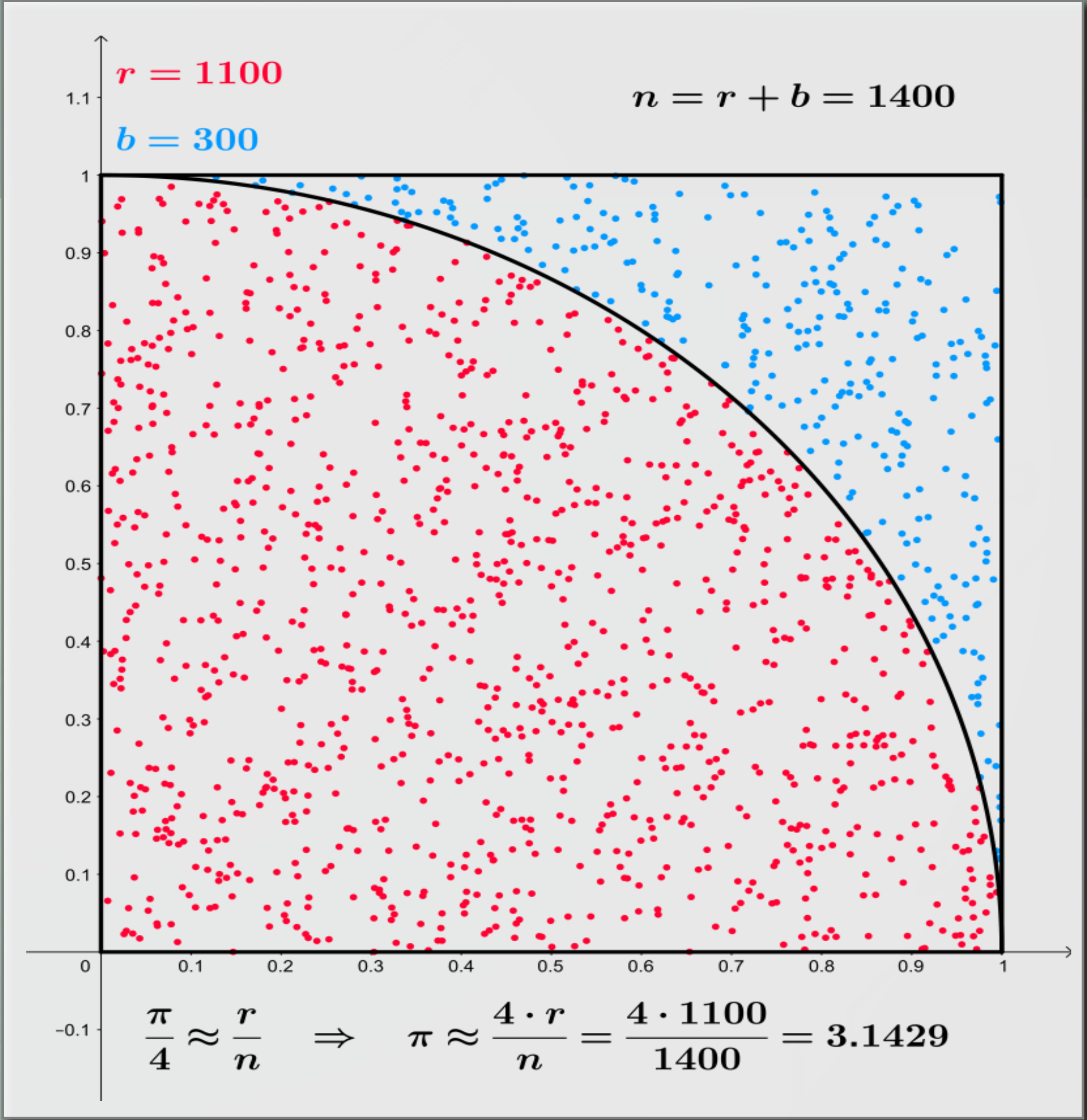
Limitations
Computationally intensive; accuracy depends on number of simulations

100s of dogs do random walk in Monte Carlo

Monte Carlo Simulations: Example

Estimate π using random points in square with an inscribed quarter circle

- **Concept:** Use random sampling to estimate π
- **Setup:** 1x1 square with inscribed quarter circle
- **Process:**
  - Generate random points in square
  - Count points inside quarter circle
  - Calculate: π ≈ 4 × (points in circle ÷ total points)
- **Key Points:**
  - More points increase accuracy
  - Simple but computationally intensive for high precision



Monte Carlo vs. Bootstrapping

Monte Carlo simulations generate new data from assumed distributions to model future scenarios, bootstrapping resamples existing data to estimate statistical properties of current sample

Aspect	Monte Carlo	Bootstrapping
Core Concept	Generate new data from known or assumed distribution	Resample with replacement from existing dataset
Data Source	Theoretical model or distribution	Observed data
Sample Size	Can generate large samples	Limited by original dataset size
Primary Use	▪ Simulating future scenarios ▪ risk assessment	▪ Estimating sampling distributions ▪ confidence intervals
Assumptions	Requires specifying underlying distribution	Assumes sample is representative of population
Computational Intensity	Can be very high for complex models	Generally less intensive
Flexibility	Highly flexible for various scenarios	Limited to inference about observed data

Exercises

Exercise 1
Implement a python function called central_limit_theorem that accepts three arguments: population_data, sample_size, and num_of_samples. Your function should simulate the central limit theorem. The function should draw the specified number of samples of the given size from the population data, calculate the means of these samples, and return a list of these means.

Test this function using a randomly generated population data. Plot the distribution of sample means and the mean of the population.

Exercise 2
Implement a python function called confidence_interval that accepts two arguments: data and confidence. Your function should calculate and return the confidence interval of population mean for the given data at the specified confidence level.
Test your function on a randomly generated data.

Exercise 3
How does the confidence interval of variance look like?

Exercise 4
Implement a python function called bootstrap that accepts three arguments: data, num_samples, and statistic. Your function should perform bootstrap sampling on the data (resampling with replacement) the specified number of times, applying the provided statistical function to each sample, and return an array of the calculated statistic for each sample.

Test this function using a randomly generated data and numpy's mean as the statistic. Calculate and print the mean and standard deviation of the bootstrap samples' statistic.

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

One-sided vs.
Two-sided tests

p-Value

Errors: False positives
and false negatives

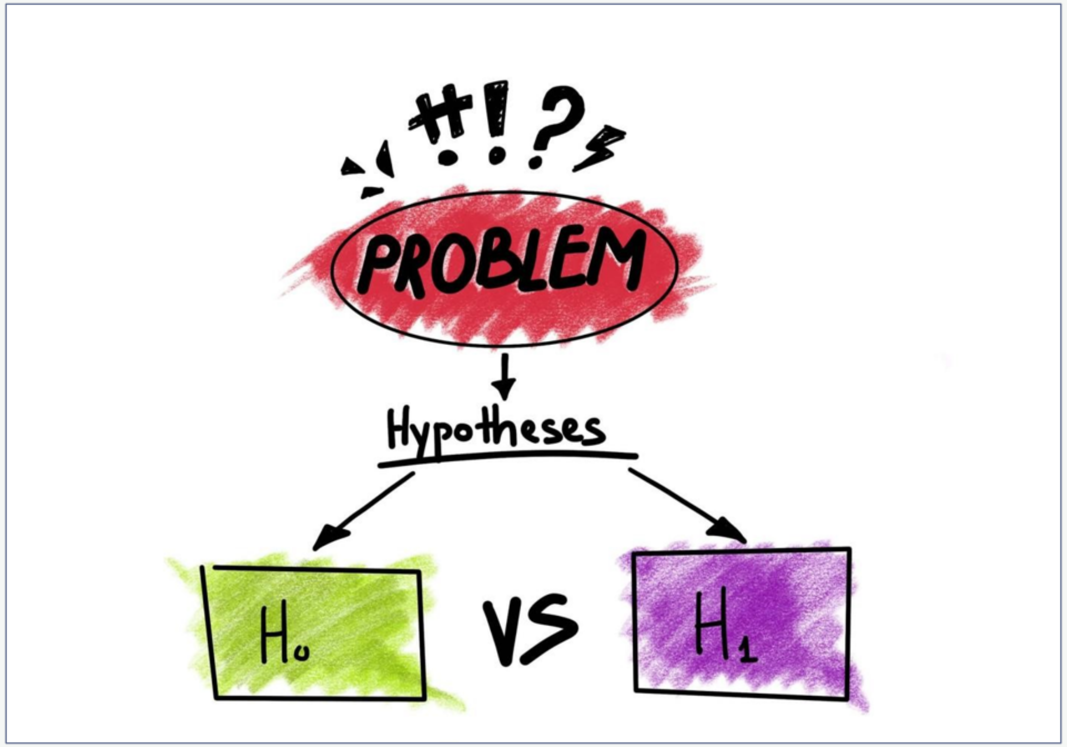
A/B Testing

Examples: t-, χ^2-,
ANOVA- tests

Exercises

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?

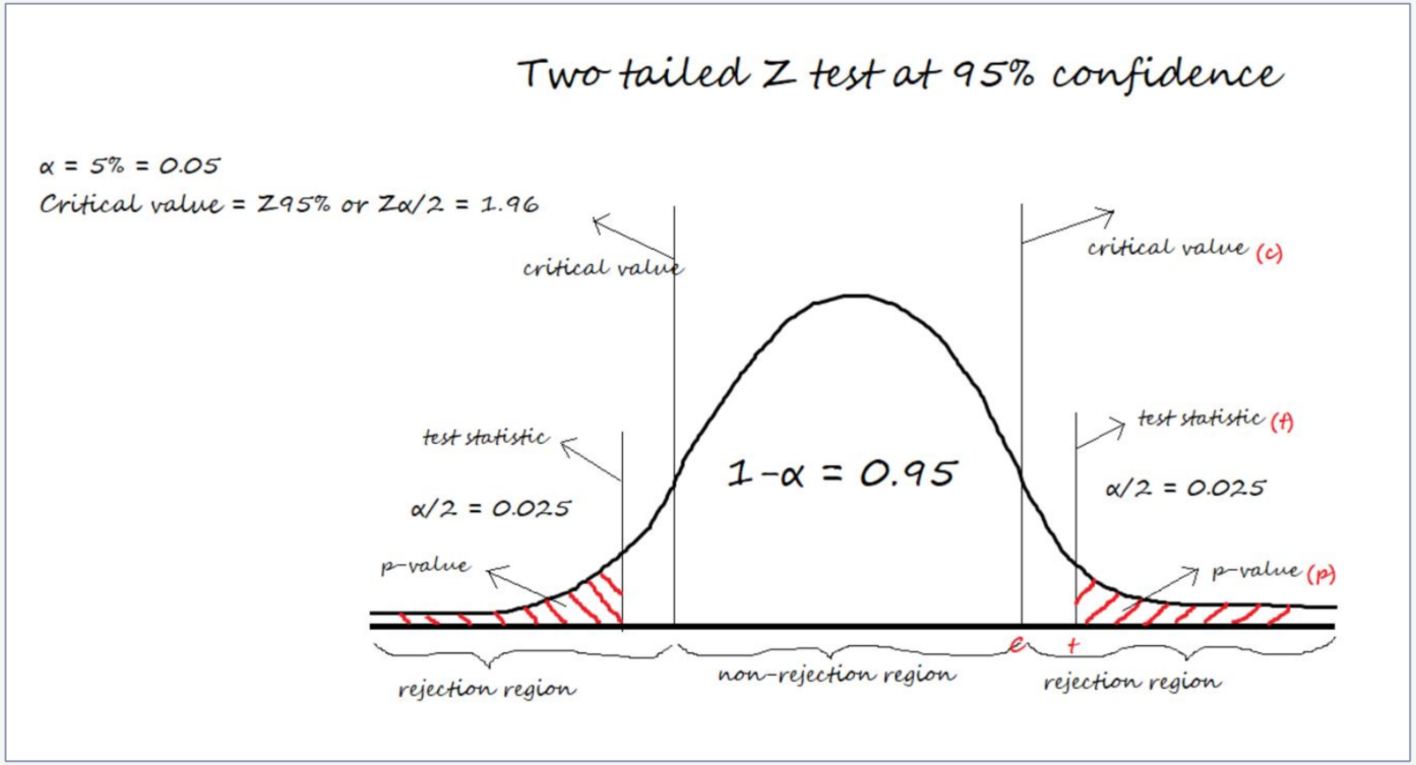


It involves formulating two competing hypotheses

- 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:**

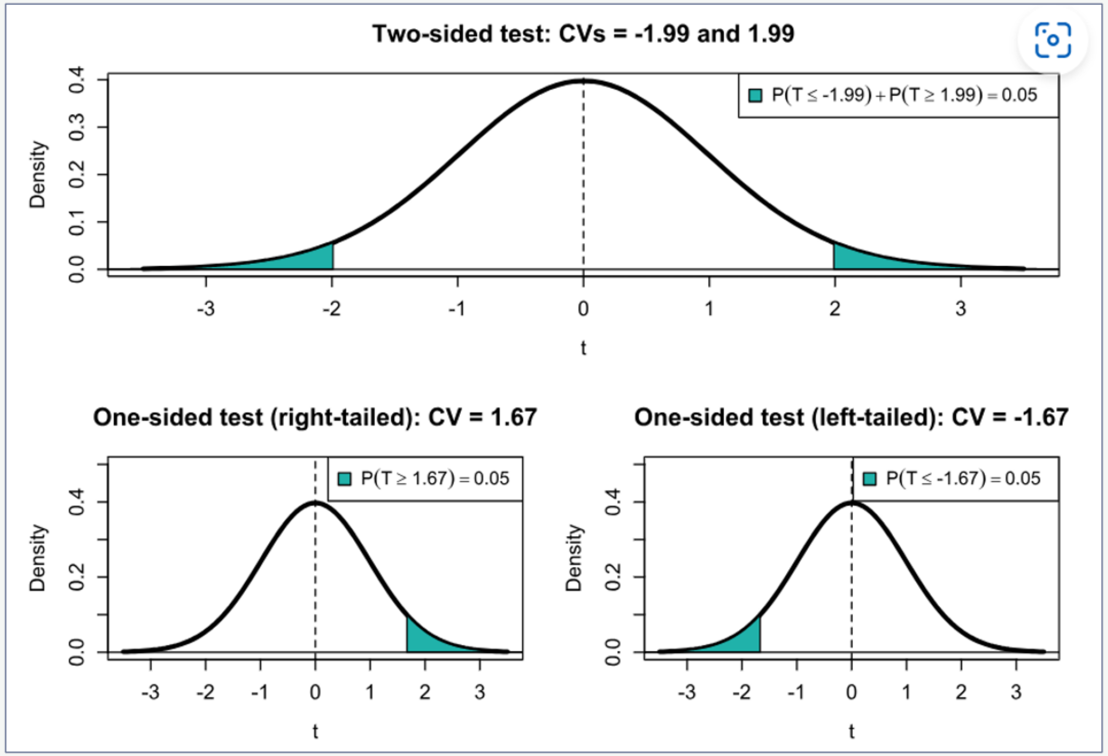
- t-test
- χ^2-test
- z-test
- ANOVA-test, and many more

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 <u>either</u> greater or less than null hypothesis value, but <u>not both</u>.
- **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 $\boldsymbol{H_0}$**"

If it is **very far**

sample is **extreme**

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

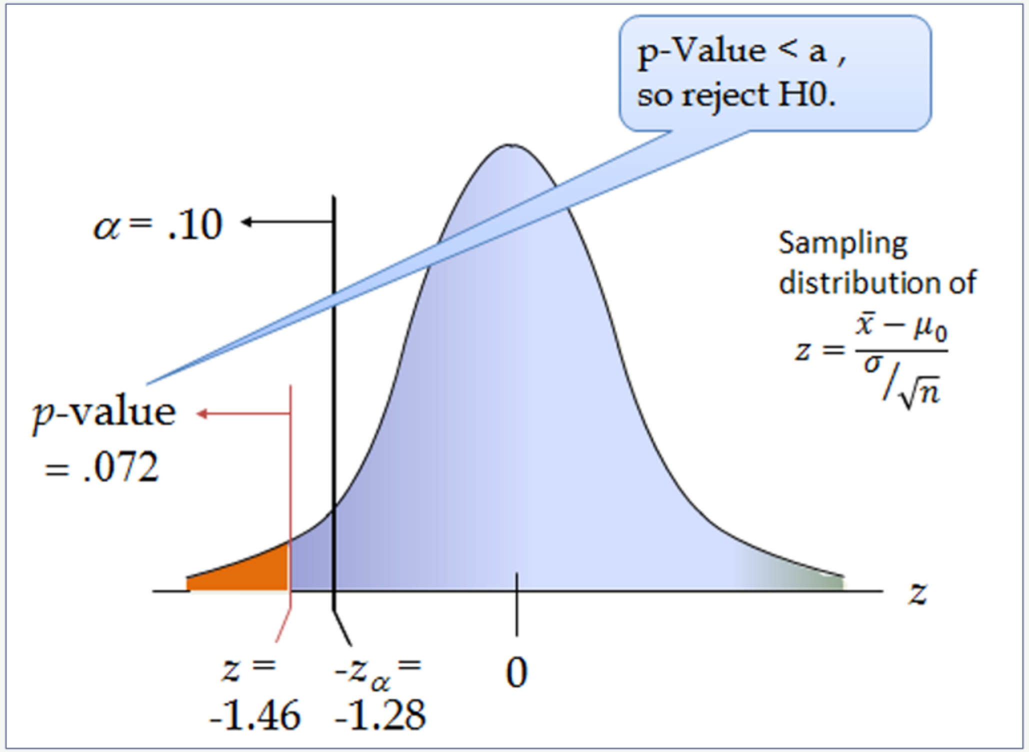
we say: "There is an effect, there is **enough evidence** to reject H_0"

simpler: "**We choose $\boldsymbol{H_1}$**"

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 ≤ $\alpha = 0.05$)
  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

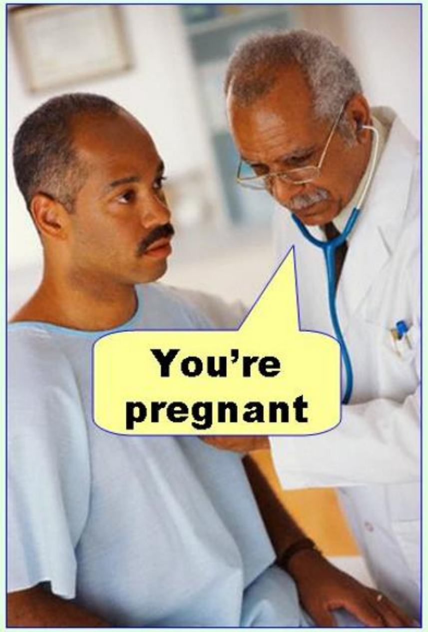
Properties

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

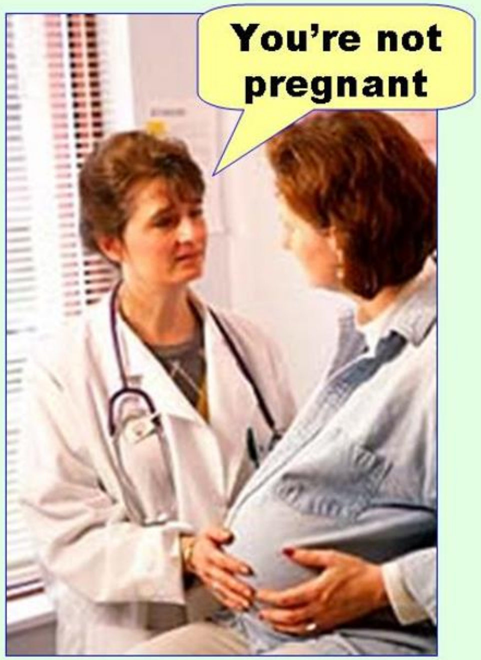
Errors: False positives and false negatives

… 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 <u>true</u> 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 <u>false</u>, 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.

	Layout A	Layout B
Visitors	122	118
Customers	22	25
Conversion %	18.0%	21.2%

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	• means of • continuous data between • two groups of independent samples	• 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 → Does the subgroup stay longer on new website B?
2. Paired t	ttest_rel	• means of • continuous data for the • same group at different times	• Like 1 • But: The same group is tested • Visit durations are measured, first with website A later with B → Does the same group stay longer on new website B?
3. χ^2 for independence	chi2_contingency or chi2.sf	• Variance of • categorical data between • two or more groups	• A company has employees in 4 experience levels • On each level a part of them received promotion (yes/no) → Does experience level have impact on promotion?
4. One factor ANOVA	f_oneway	• means of • continuous data between • ≥ three groups of independent samples	• Education program in company, • 3 learning styles: self-paced online courses, instructor led, mixed • Employees are divided in 3 subgroups → Does the learning style have impact on learning success?

Two sample t-tests

… are statistical tests used to compare the <u>means</u> of <u>continuous data</u> between <u>two groups</u> 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:** $\mathbf{df = n_1 + n_2 - 2}$

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

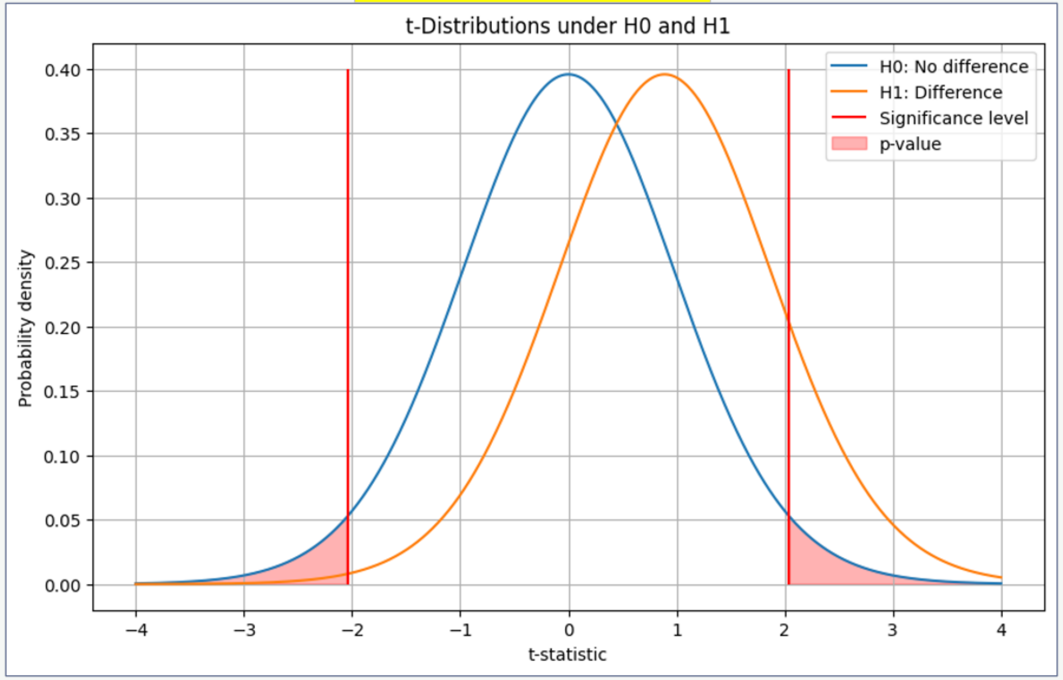
```
# Randomly generating test scores for Group A and Group B
np.random.seed(0)  # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(90, 10, 30)
→ df = 58
```

Two sample t-tests

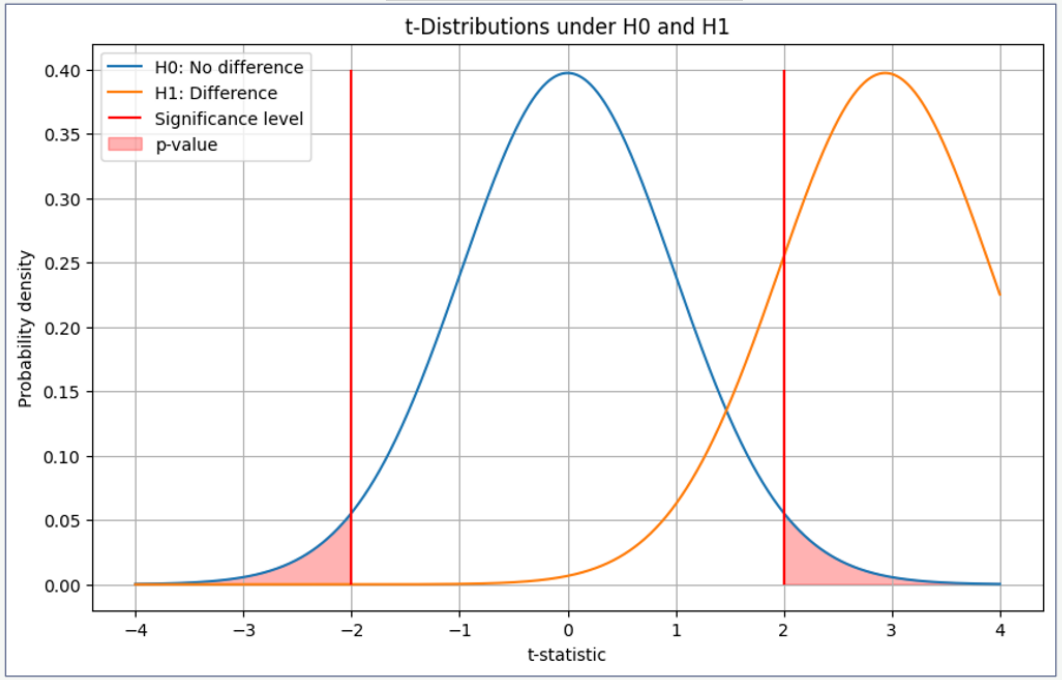
… are statistical tests used to compare the means of continuous data between two groups and determine if they are significantly different from each other.

Critical value: 2.001 (2-sided!)

ttest_stat = 0.8897



ttest_stat = 2.9389



H0:
There is no evidence that subgroup B stays longer on the website …

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: $\mathbf{df = n - 1}$**

Variant A: No significant difference

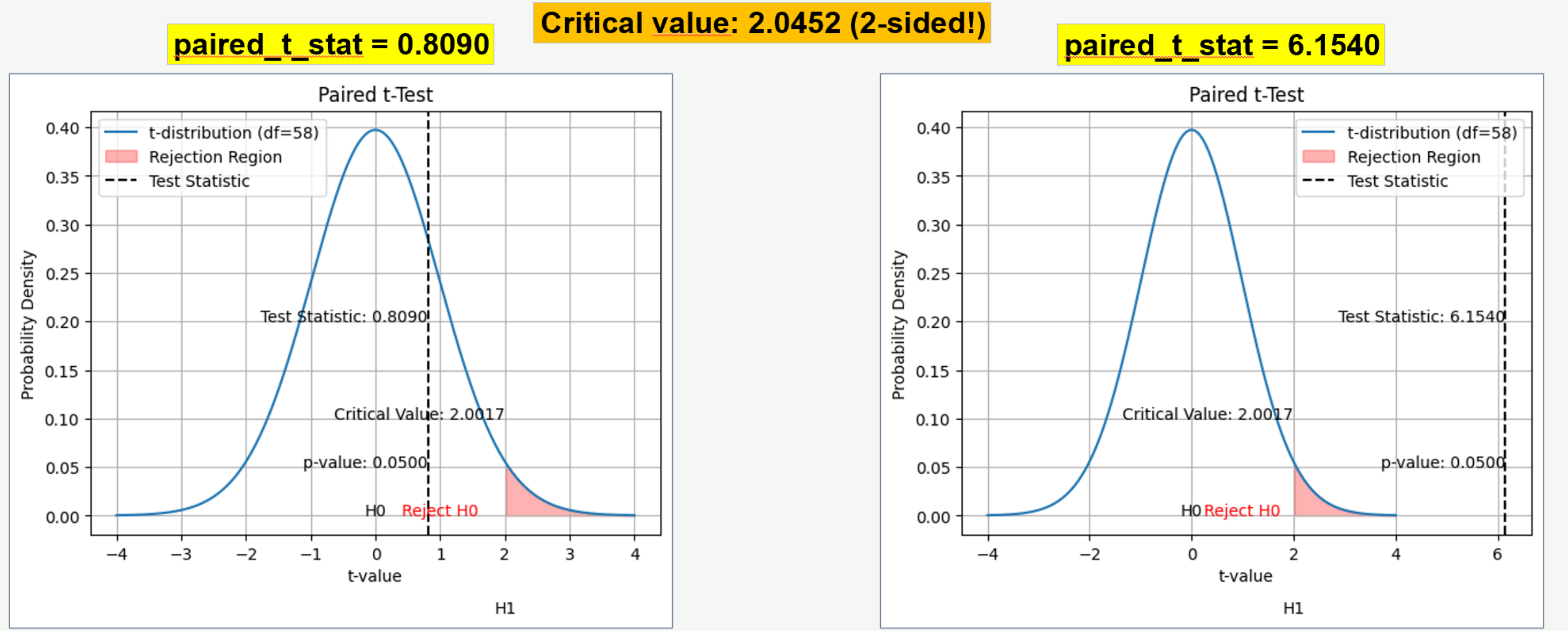
```
# Randomly generating test scores for Group A and Group B
np.random.seed(0)  # for reproducibility
group_A_scores = np.random.normal(75, 10, 30)
group_B_scores = np.random.normal(80, 10, 30)
→ df = 29
```

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



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 …

χ^2 independence test

… statistical tests used to determine if there's a significant association between two categorical variables in a sample.

Observed	2007	2008	2009	2010	2011	Total	
Freshman	560	495	553	547	512	2667	
Sophomore	369	385	358	361	393	1866	
Junior	209	226	248	268	285	1236	
Senior	267	277	304	328	340	1516	
Unclassified	64	70	93	77	126	430	
Total	1469	1453	1556	1581	1656	7715	7715

Observed	2007	2008	2009	2010	2011	Total	
Freshman	507,818924	502,287881	537,893973	546,536228	572,462994	2667	
Sophomore	355,301879	351,432016	376,344264	382,390927	400,530914	1866	
Junior	235,344653	232,781335	249,282696	253,287881	265,303435	1236	
Senior	288,658976	285,514971	305,754504	310,667012	325,404537	1516	
Unclassified	81,8755671	80,9837978	86,7245625	88,117952	92,2981205	430	
Total	1469	1453	1556	1581	1656	7715	7715

alpha	0,05
df	16
Chi^2 Critical value	26,2962276

Ch^2-Computation

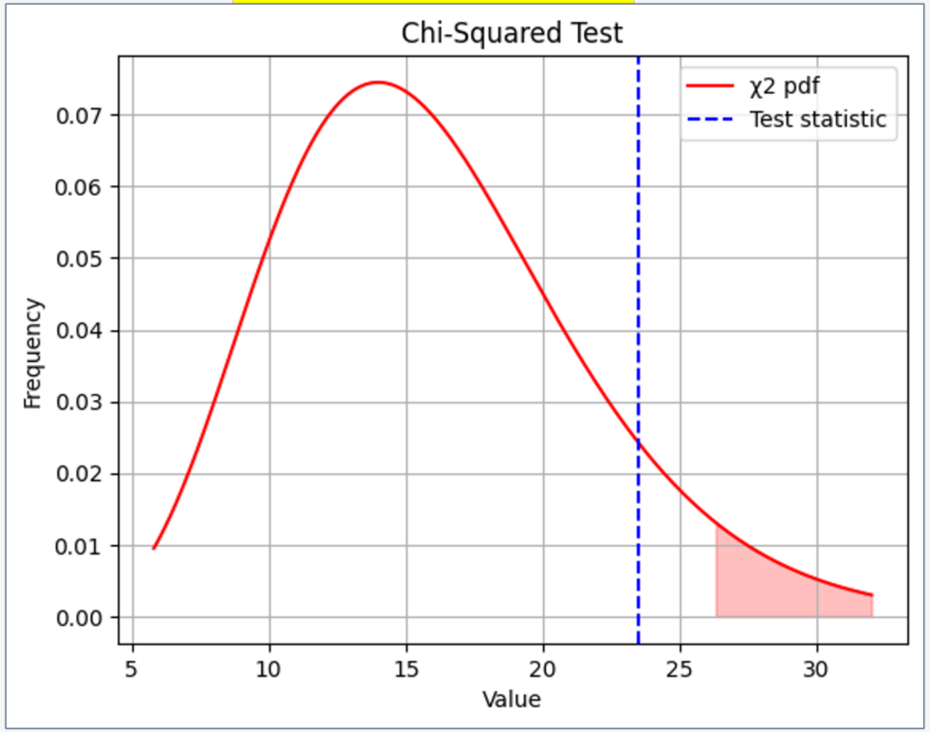
Observed	Expected	O - E	(O - E)^2	(O - E)^2/E
560	507,82	52,18	2.722,86	5,36188107
369	355,30	13,70	187,64	0,52811009
209	235,34	-26,34	694,04	2,94903983
267	288,66	-21,66	469,11	1,62513998
64	81,88	-17,88	319,54	3,9027015
495	502,29	-7,29	53,11	0,10574256
385	351,43	33,57	1.126,81	3,2063373
226	232,78	-6,78	45,99	0,19755237
277	285,51	-8,51	72,50	0,2539437
70	80,98	-10,98	120,64	1,48972779
553	537,89	15,11	228,19	0,42423241
358	376,34	-18,34	336,51	0,89416013
248	249,28	-1,28	1,65	0,00660017
304	305,75	-1,75	3,08	0,01006783
93	86,72	6,28	39,38	0,45409414
547	546,54	0,46	0,22	0,00039354
361	382,39	-21,39	457,57	1,19660723
268	253,29	14,71	216,45	0,85454721
328	310,67	17,33	300,43	0,96705621
77	88,12	-11,12	123,61	1,40276589
512	572,46	-60,46	3.655,77	6,38604364
393	400,53	-7,53	56,71	0,14159872
285	265,30	19,70	387,95	1,46230552
340	325,40	14,60	213,03	0,65465452
126	92,30	33,70	1.135,82	12,3059567
			Chi^2-val	46,7812601

χ^2 independence test

… statistical tests used to determine if there's a significant association between two categorical variables in a sample.

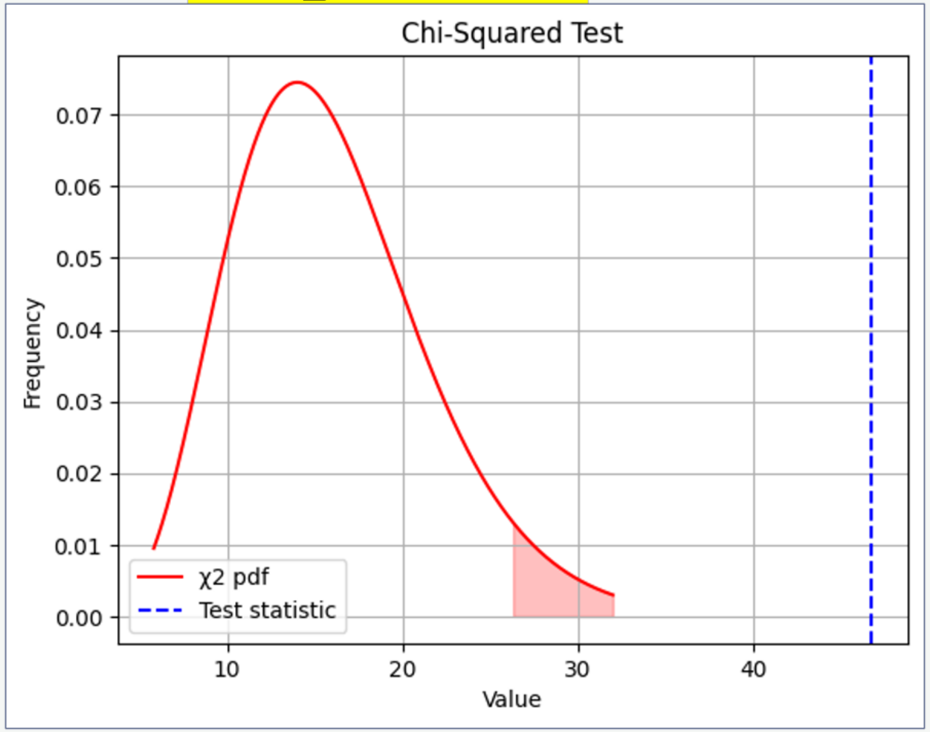
Critical value: 26.2962

Chi^2_stat = 23.5172



H0:
Experience level does not have significant impact on promotion…

Chi^2_stat = 46.7812



H1:
Experience level has significant impact on promotion…

Oneway – analysis of variance test

… statistical tests used to compare the means of three or more groups and determine if they are significantly different from each other.

Group 1	Group 2	Group 3	
82	71	64	
93	62	73	
61	85	87	
74	94	91	
69	78	56	
70	66	78	
53	71	87	
Mean	Mean	Mean	Overall mean
71,71	75,29	76,57	74,52

N_Total	21
Num_Groups	3
N_per_Group	7
N_per_Group - 1	6
alpha	5%

	df	SS	MS	F_stat
Between (SSC)	2	88,66667	44,33333333	**0,283726128**
Within (SSE)	18	2812,571	156,2539683	
Total (SST)	20	2901,238		

- The larger the differences between the group samples the larger F_stat
- Exceeding F_crit, we "reject the null" (H1)

F_crit(5%, 2, 18)	3,554557146
F_stat	0,283726128

- **F_stat < F_crit**

→ H0 is not rejected ("H0")

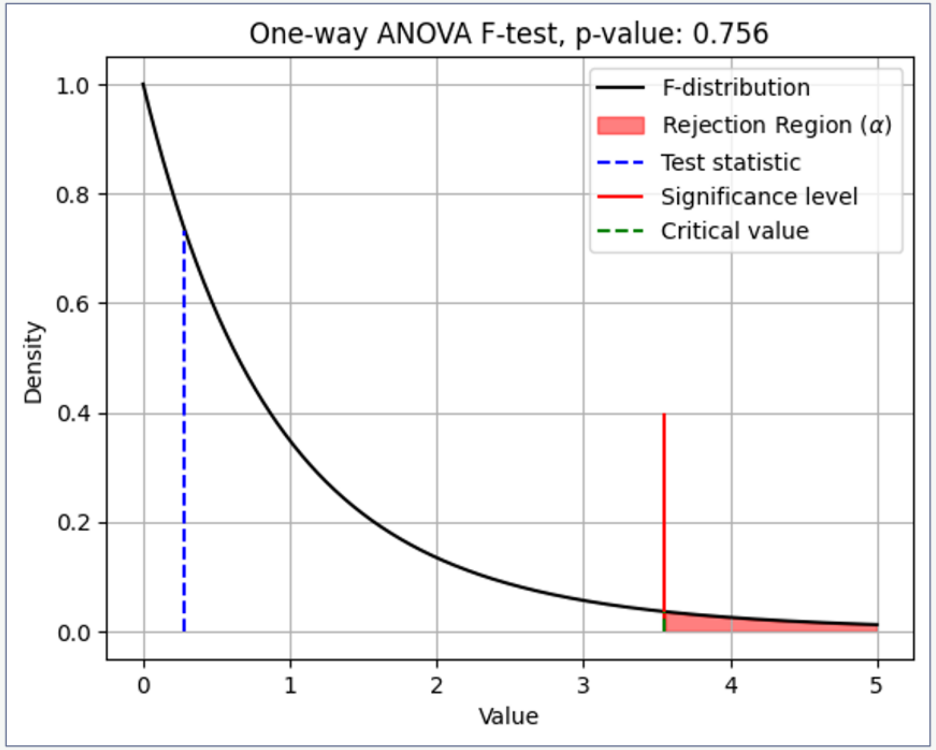
Oneway – analysis of variance test

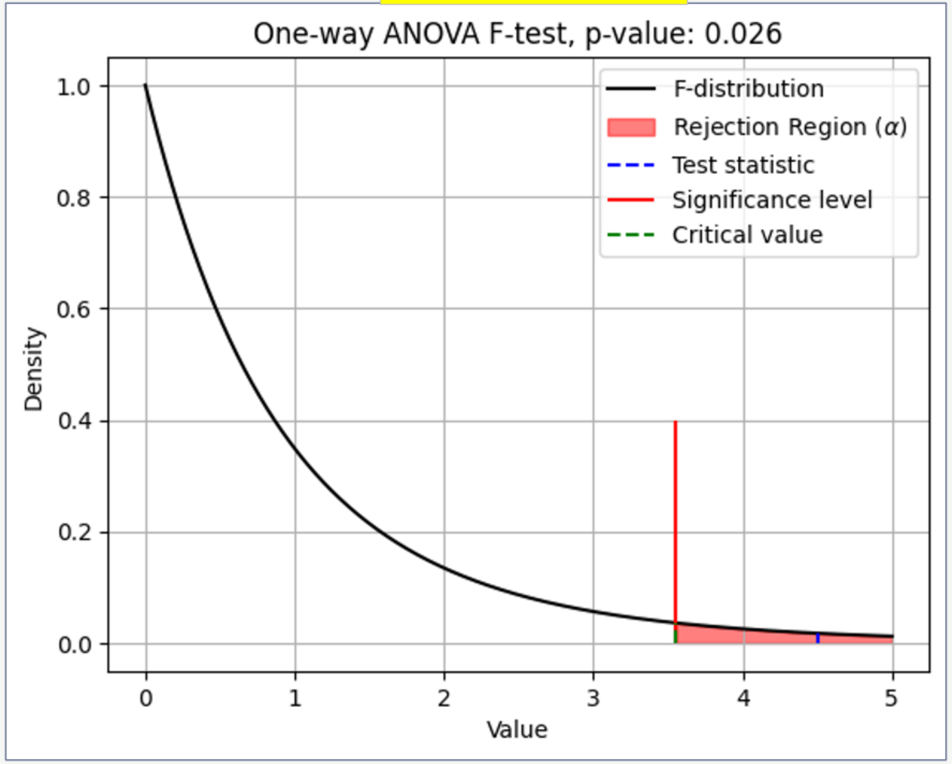
… statistical tests used to compare the means of three or more groups and determine if they are significantly different from each other.

F_stat = 0.2837

Critical value: 3.5545

F_stat = 4.4822





H0:
The learning style has no significant impact on learning success …

H1:
The learning style has significant impact on learning success …

Exercises

Exercise for ttest_ind:
Based on the ttest_ind example, try new data points and find data point configurations, where H0 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 H0 is very close of being rejected or is just rejected.
Generate figure according to the given figures.

Exercise for chi^2:
→ Next slide

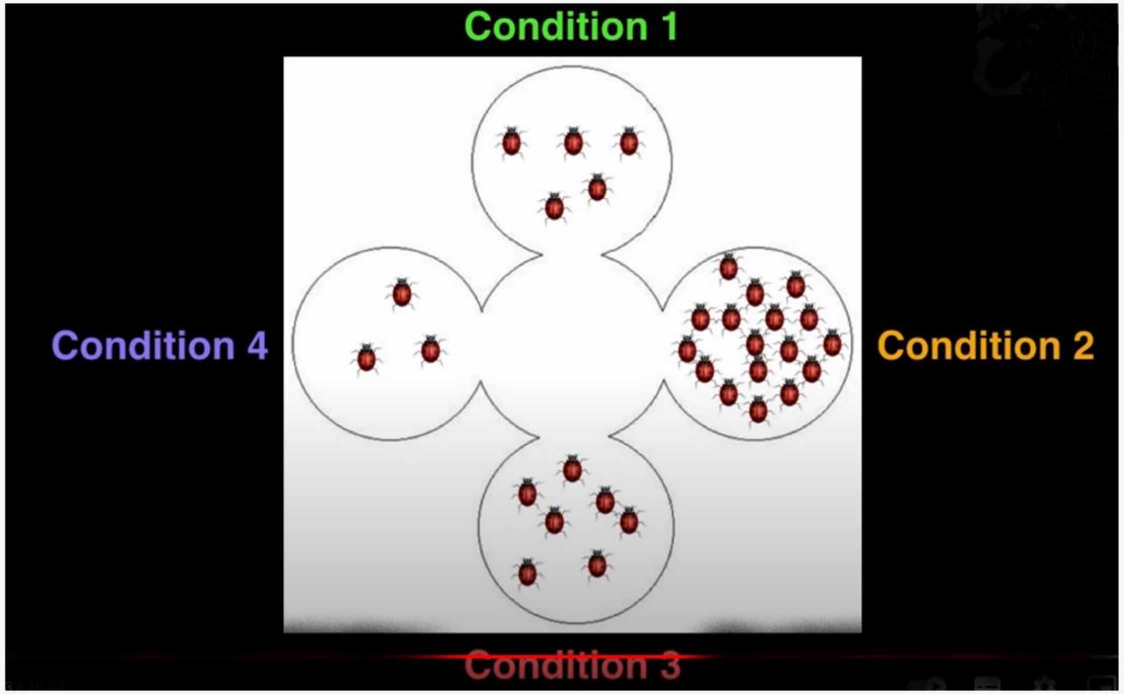
Exercise for F:
Based on the F example, try new data points and find data point configurations, where H0 is very close of being rejected or is just rejected.
Generate figure according to the given figures.

Additional exercise for chi^2:
Based on the chi^2 example, try new data points and find data point configurations, where H0 is very close of being rejected or is just rejected.
Generate figure according to the given figures.

Exercise – Chi squared

H_0: Bugs have no preference of moving to one or the other bins

H_1 : Or do we have enough evidence to reject H_0?



Nr	Obs
1	5
2	17
3	7
4	3

Source:
https://www.youtube.com/watch?v=qYOMO83Z1WU&t=552s

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

Parameter estimation in linear models

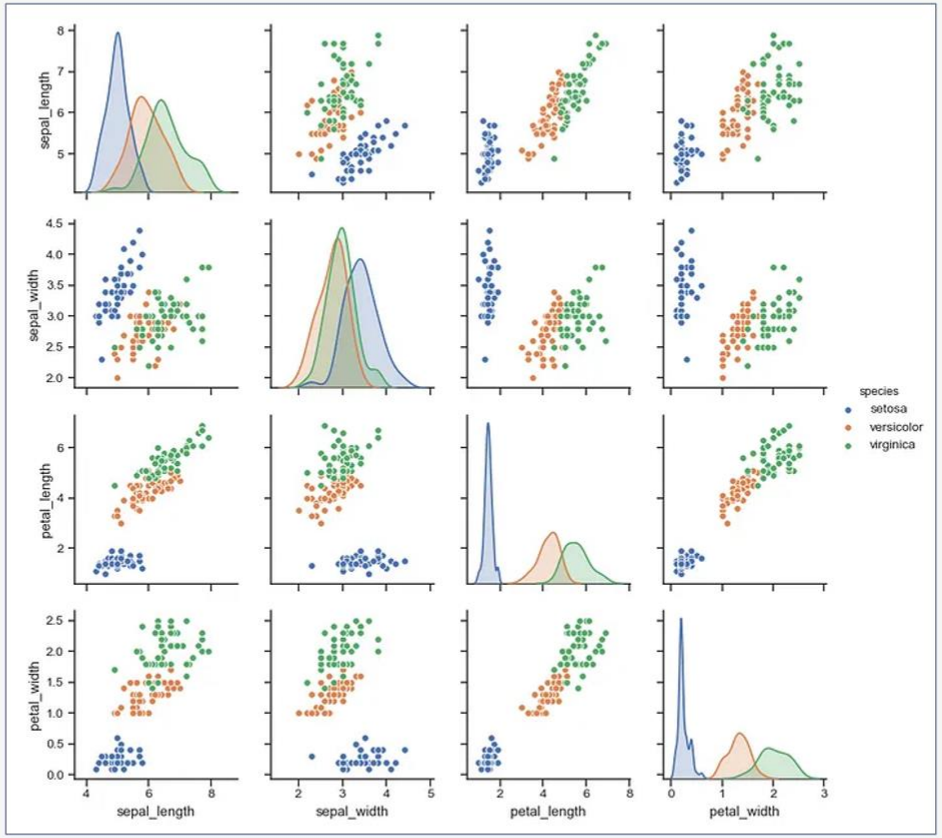
Overfitting vs. underfitting

Estimation of linear model

Linear regression model evaluation

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

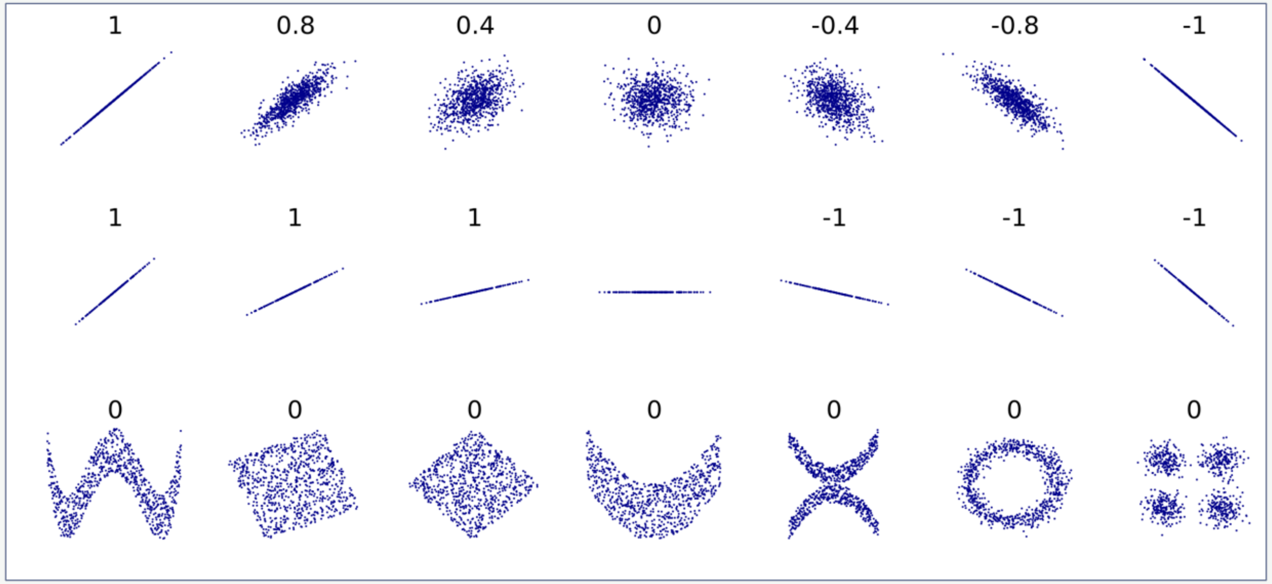
Correlation:

- 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.

https://medium.com/@raghavan99o/scatter-matrix-covariance-and-correlation-explained-14921741ca56

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.

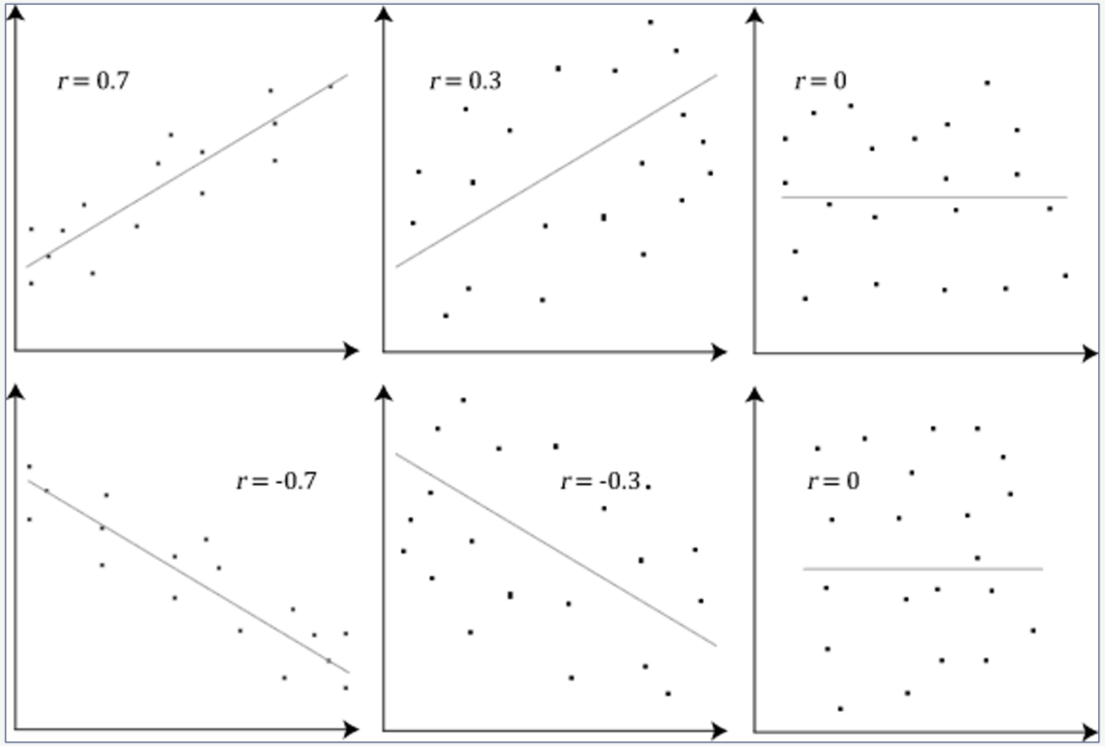
Negative Correlation:

- 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

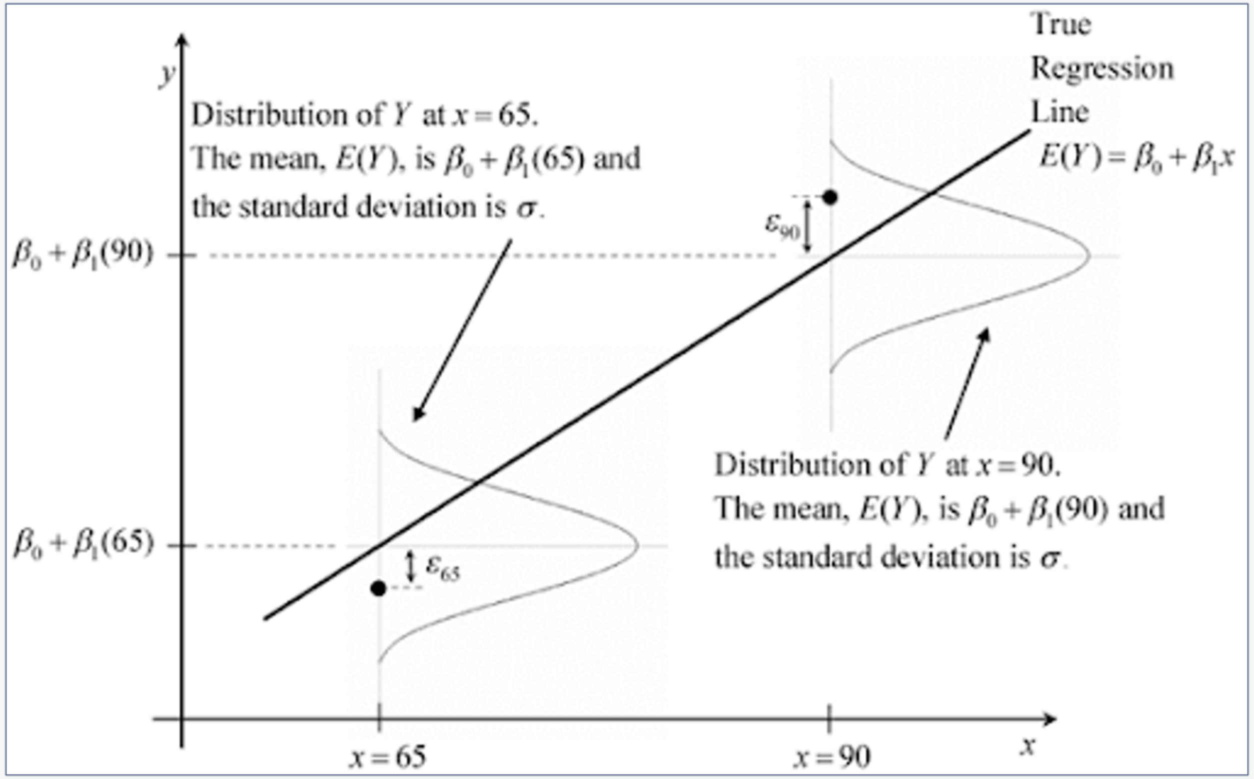
- H0: the two samples are independent.
- H1: there is a dependency between the samples.

```
# 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)



https://vitalflux.com/generalized-linear-models-explained-with-examples/

Simple linear regression model:

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

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

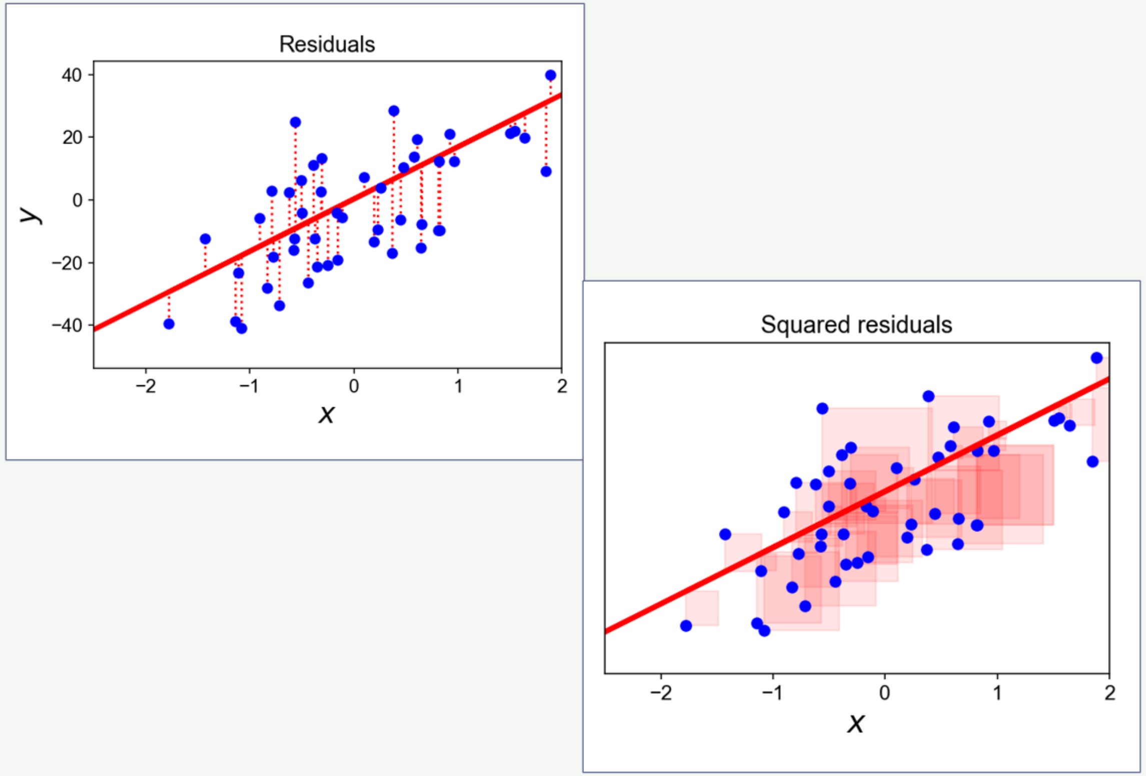
- β_0 : Value of Y when X = 0.
- β_1 : Change in Y for each one-unit change in X.

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



https://gregorygundersen.com/blog/2020/01/04/ols/

Ordinary Least Squares (OLS) estimation:

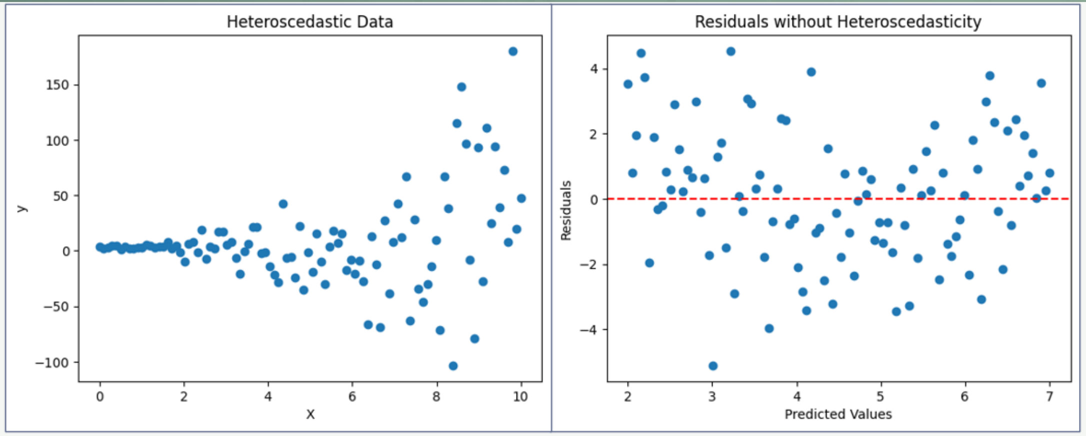
- Method used in linear regression to estimate the model's parameters
- by minimizing the sum of the squares of the differences between the observed and predicted values of the dependent variable.
- Separation systematic relations from white noise

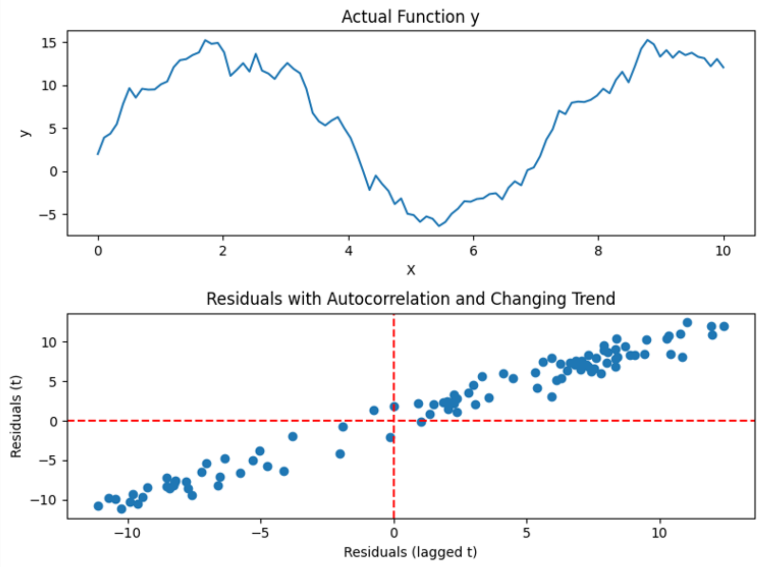
IID & BLUE:

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 precedessors
- 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:

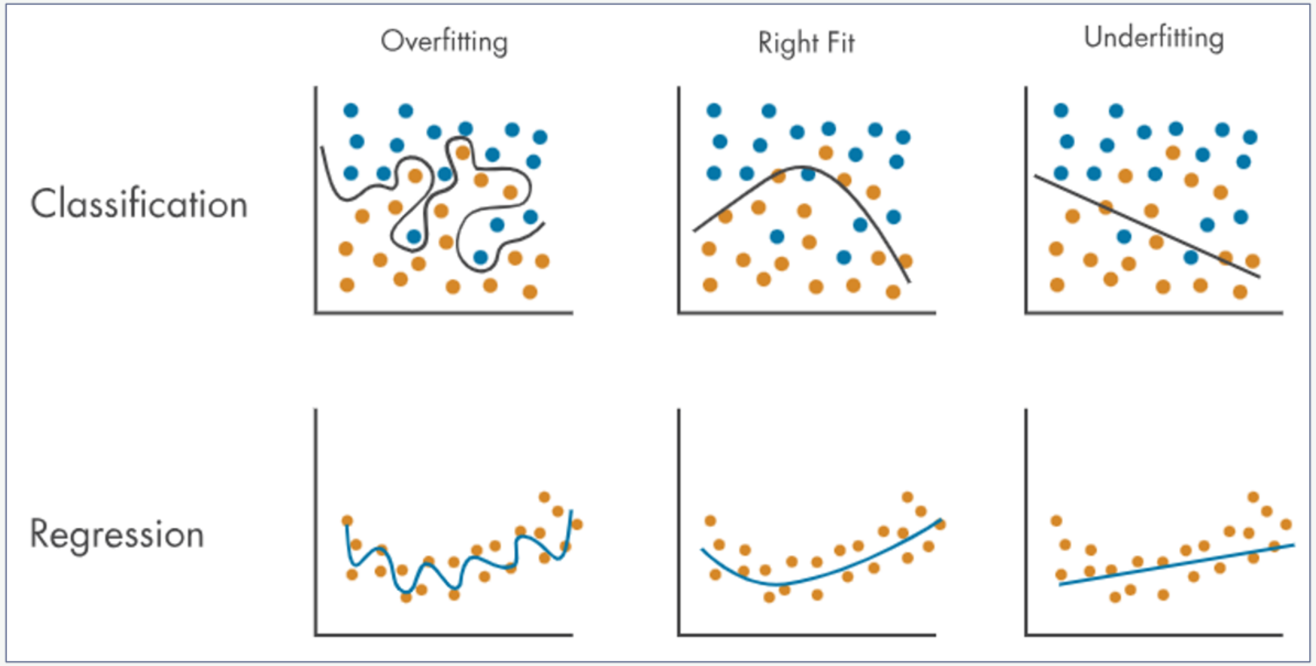
Overfitting often results in overly complex models that have too many parameters.

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:

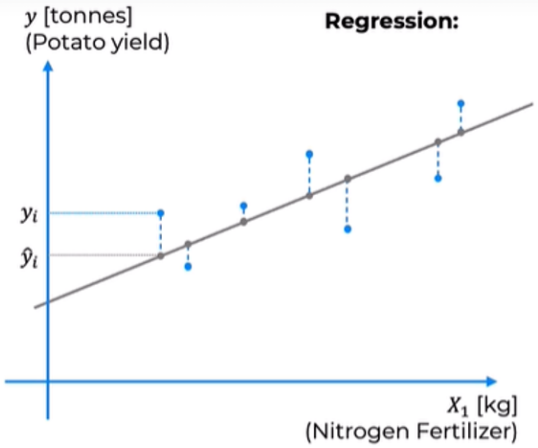
Occurs when a model cannot adequately capture the underlying structure of the data. It's oversimplified, has poor predictive performance on both training and unseen data due to high bias.

What is worse?

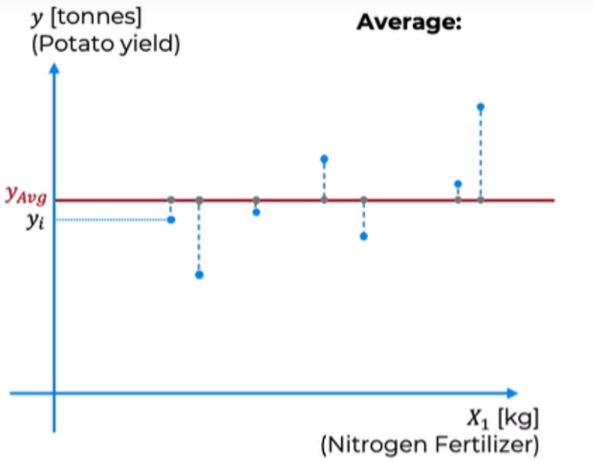
Both overfitting and underfitting lead to poor predictions and indicate that the model may not be the right choice or needs fine-tuning.

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



$$SS_{res} = SUM(y_i - \hat{y}_i)^2$$



$$SS_{tot} = SUM(y_i - y_{avg})^2$$

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

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 multicoliniarity, overfitting etc.
- In order to incentivate leaner models, number of input factors is punished

Statsmodels: Summary for OLS

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                     8156.
Date:                Wed, 05 Jul 2023   Prob (F-statistic):          2.50e-122
Time:                        22:35:07   Log-Likelihood:                -143.90
No. Observations:                 100   AIC:                             299.8
Df Residuals:                      94   BIC:                             315.4
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 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
------------------------------------------------------------------------------
Omnibus:                        2.625   Durbin-Watson:                   1.810
Prob(Omnibus):                  0.269   Jarque-Bera (JB):                2.212
Skew:                          -0.361   Prob(JB):                        0.331
Kurtosis:                       3.101   Cond. No.                         50.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model Info

Model Info

The closer to 1 the better

The smaller the better

The closer to 0 the better

The smaller the better

Coefficients Table

Skewness and kurtosis

Autocorrelation, e.g. basis of time series

Multicollinearity

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.

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.

[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.

Model estimation: Logistic Regression

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

Binary Logistic Regression

Exercise:
Credit Card Default Predicion

Confusion Matrix

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.

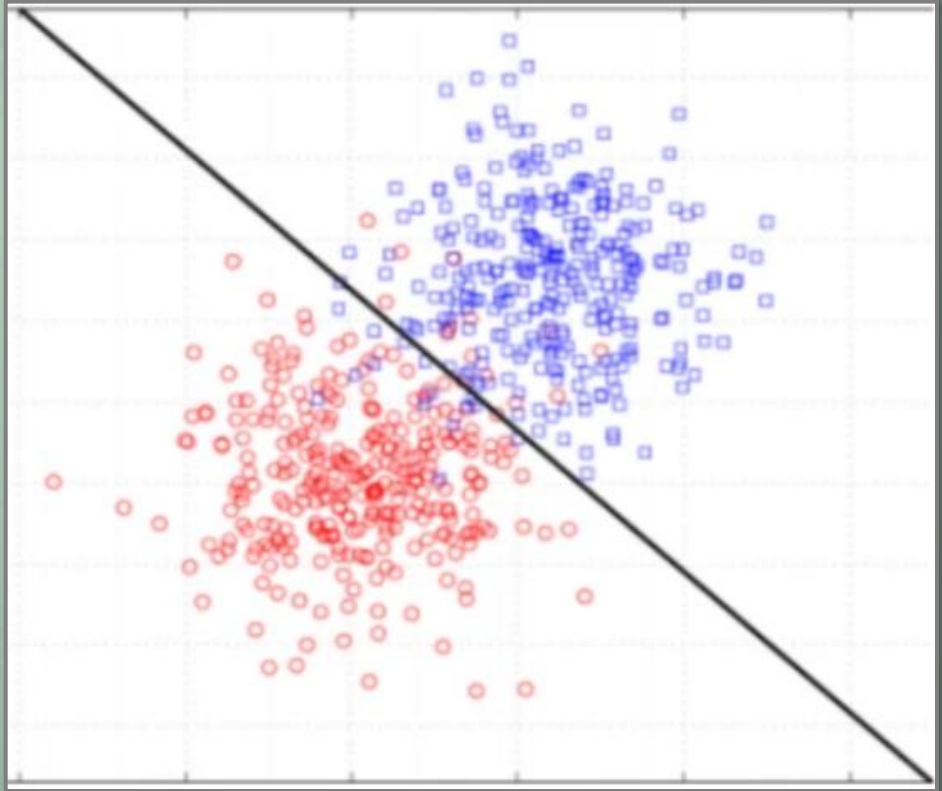
Probability Estimation

- 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

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

Logit (Logged-odds) equation

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

Inverse of logit equation by P

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

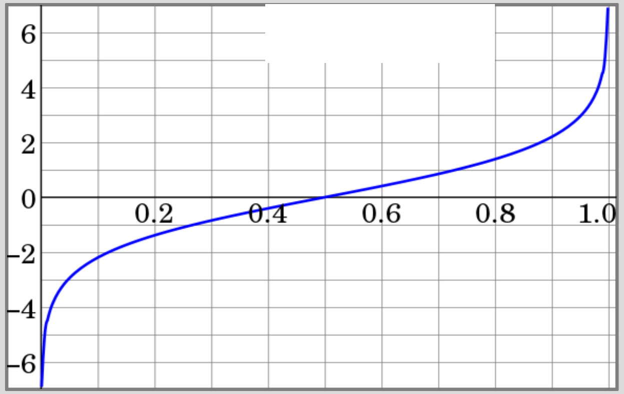
Abbreviated notation as sigmoid function

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

Last step for binary classification: Rounding to 0 or 1

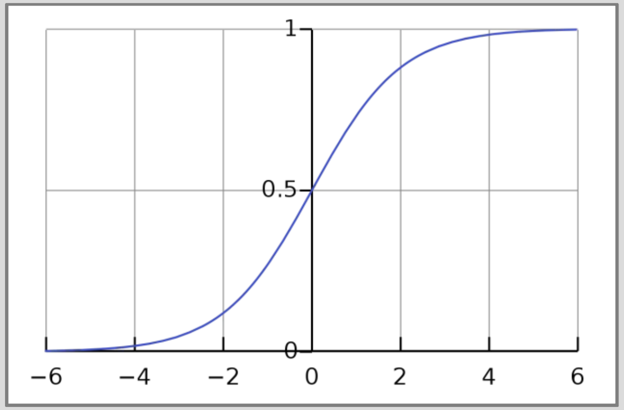
Logit equation

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



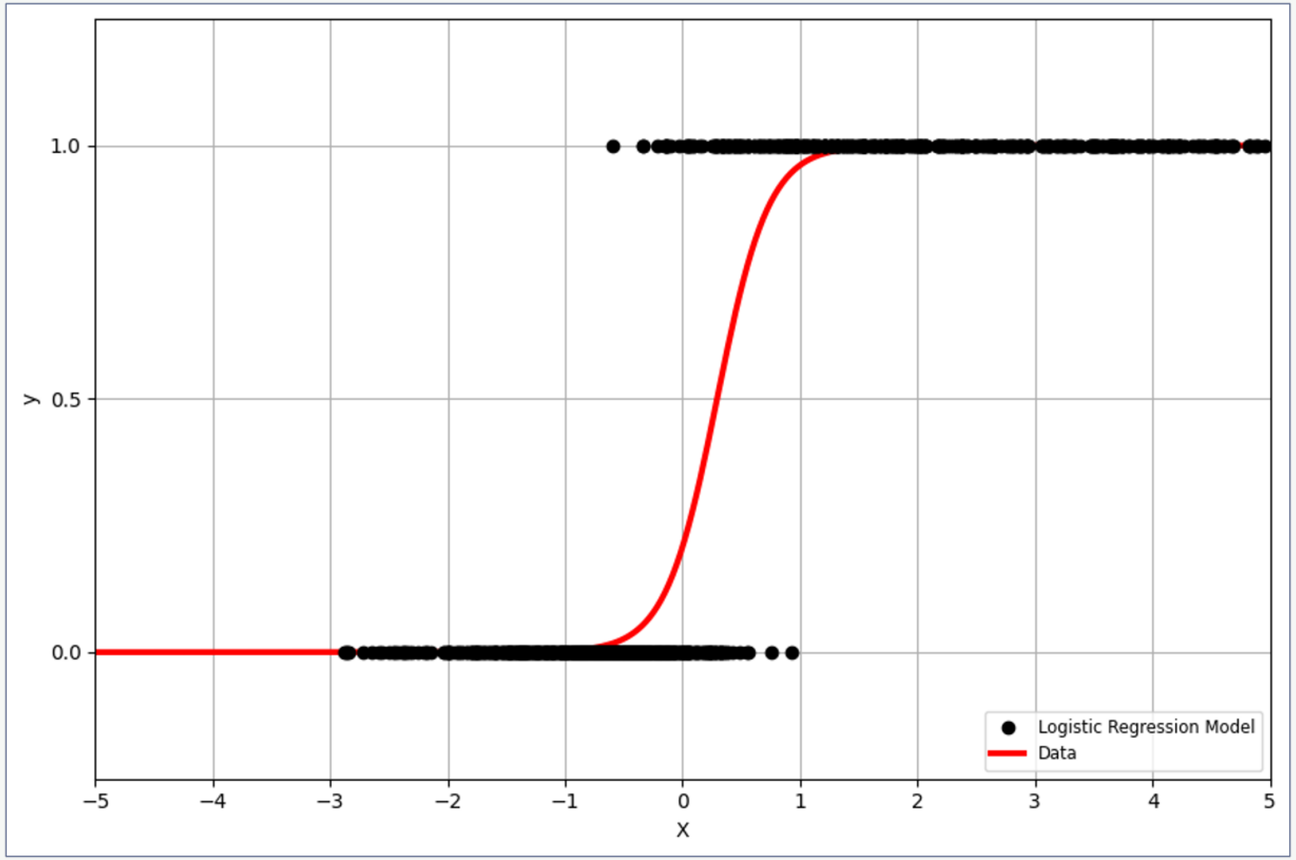
Sigmoid function:
Inverse of logit equation

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



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):

$P(z) = \frac{1}{1+e^{-z}}$ 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.

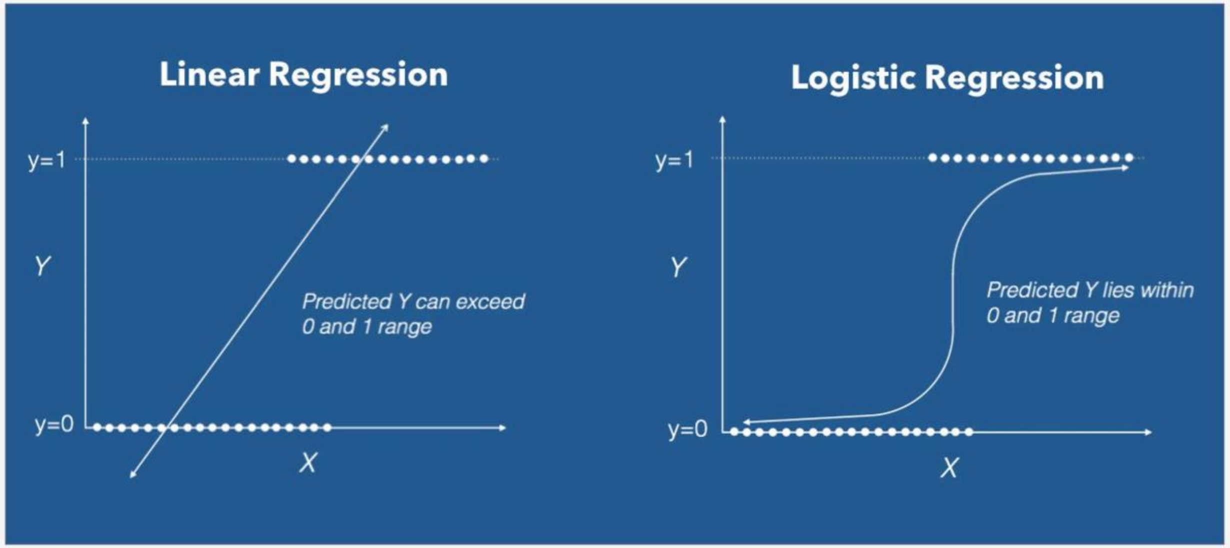


Fig 12: Typical logistic regression curve

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

Logistic function (= sigmoid function):

$P(z) = \frac{1}{1+e^{-z}}$ 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

Backpropagation (of errors)

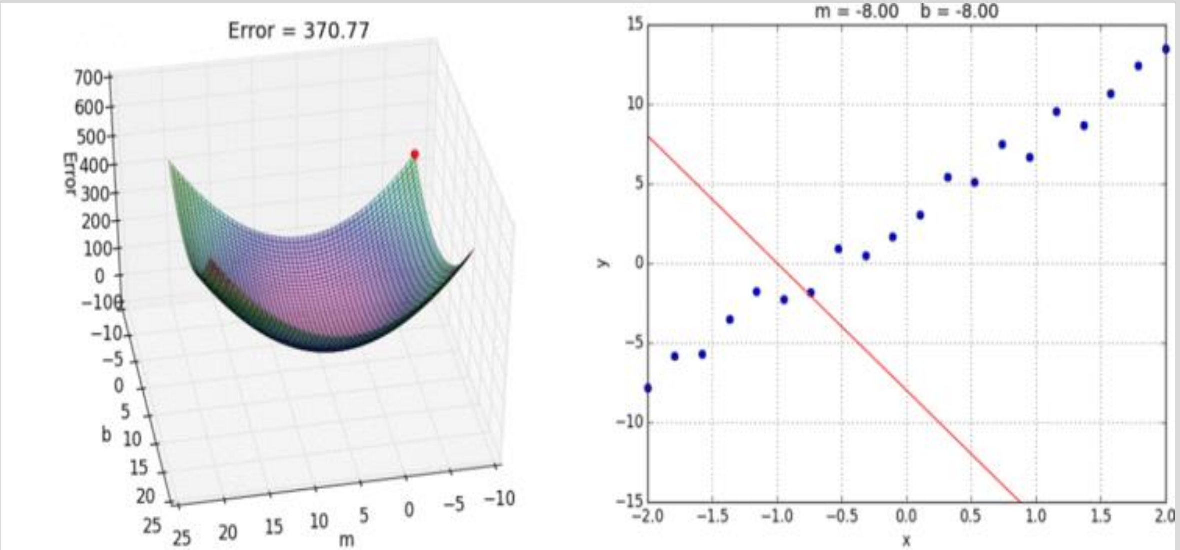
… is general optimization approach for minimizing error in supervised or reinforcement learning.
It calculates gradient of error function with respect to each weight by using chain rule.

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 ⇔ 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

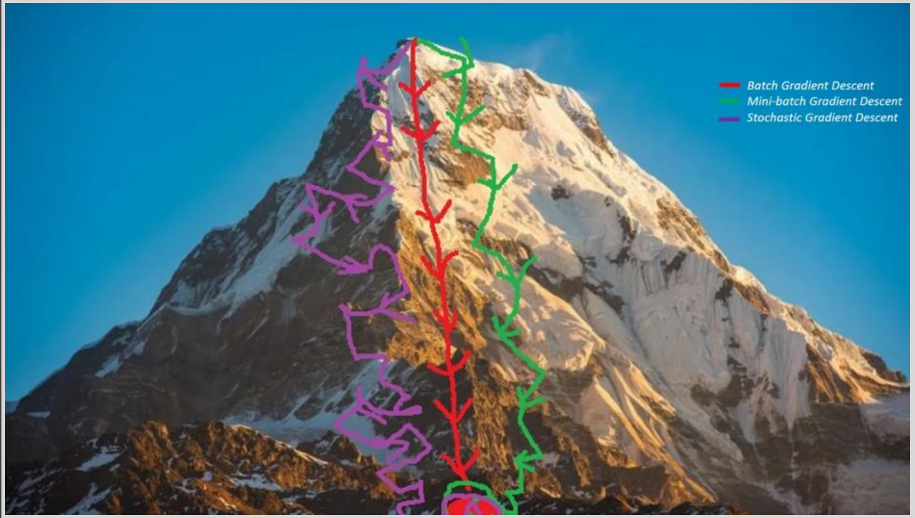
It looks for best parameters by navigating the cost function landscape, aiming to find the lowest point (global minimum)

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

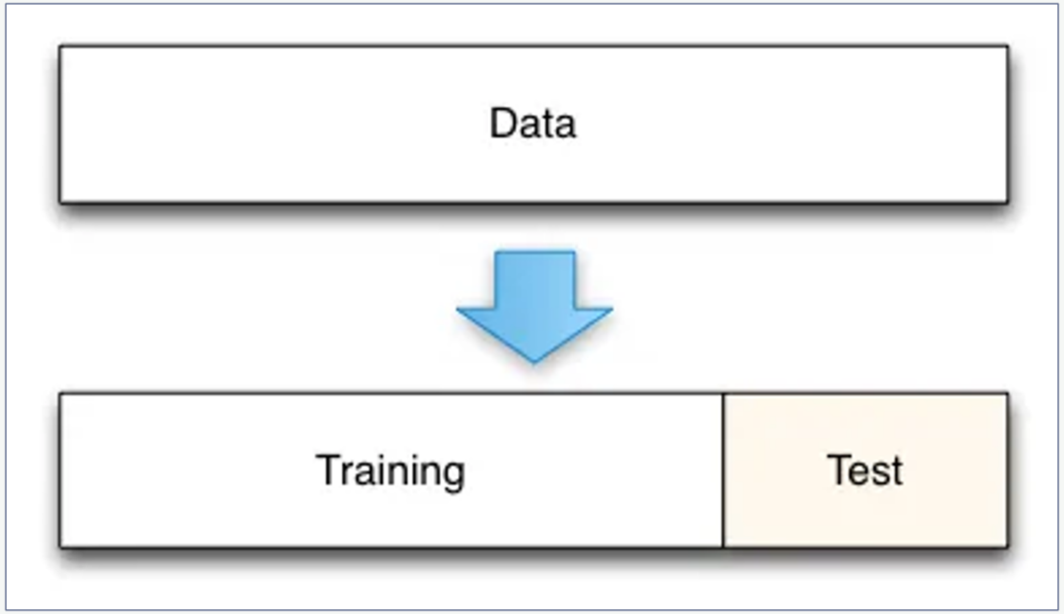
How L-BFGS Works

- 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



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

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

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
```

The closer to 1 the better

The closer to 0 the better

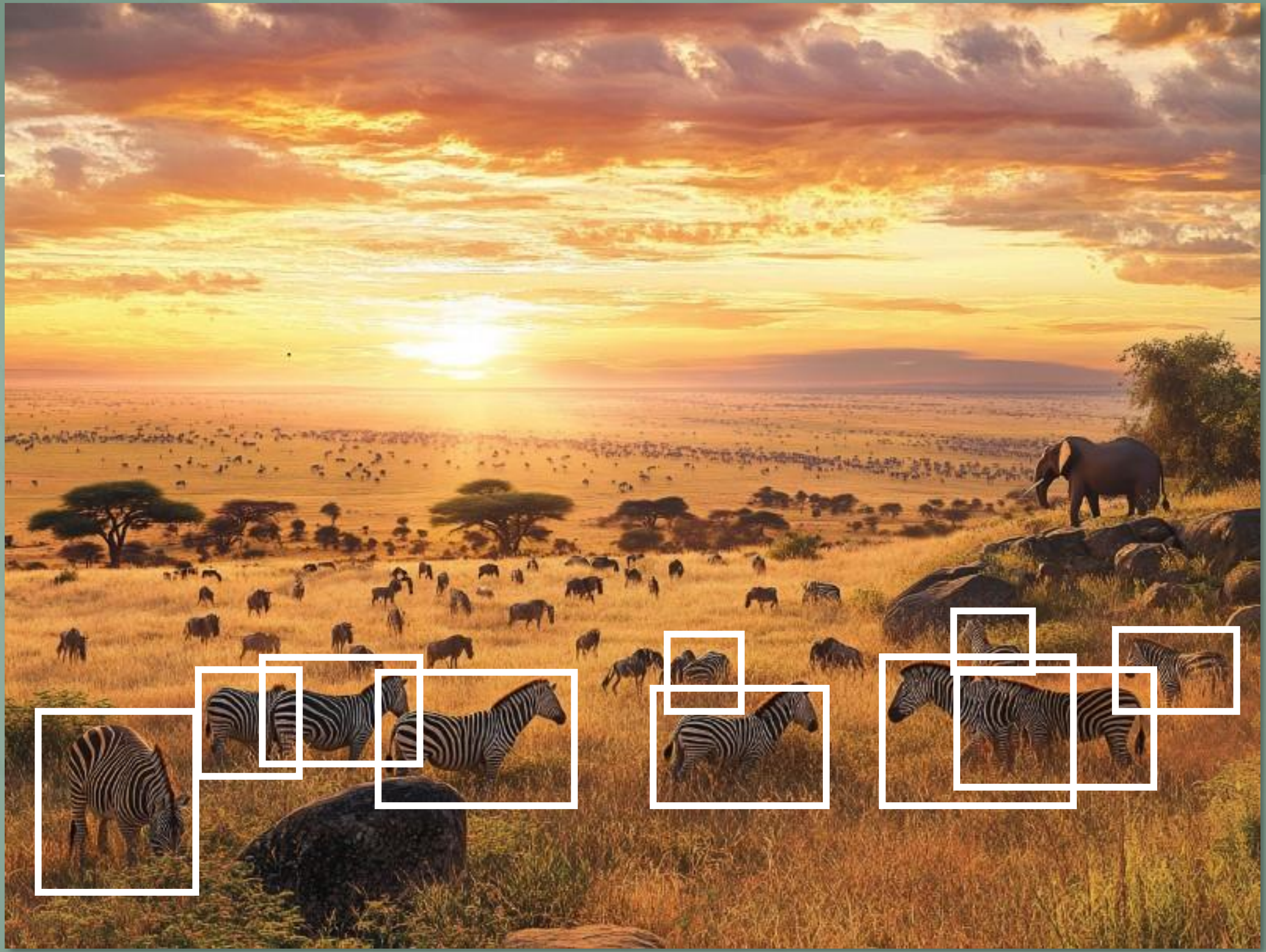
If LLR p-value is small (typically less than a pre-specified significance level, such as 0.05), full model provides significantly better fit to the data compared to reduced model.

coef:

- Input factors that have impact and should be used

How many zebras do you see?

10 out of 55 animals are zebras



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

Correctly recognized

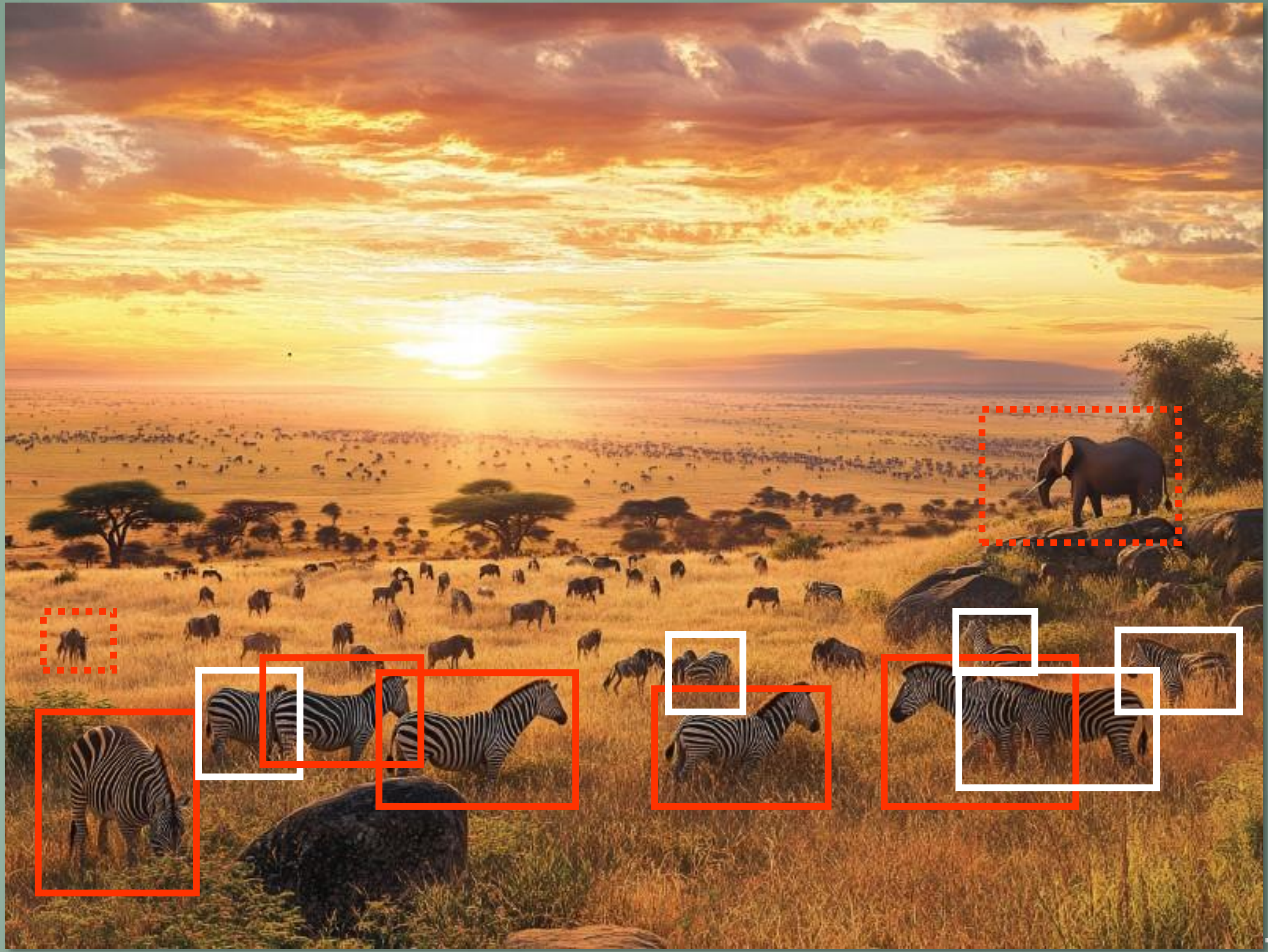
Mistakenly recognized

Mistakenly not recognized

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.

Correct Predictions

Incorrect Predictions

		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 Negative (type II error)	True Negative

Mistakes or errors: False positives and false negatives

False positives and false negatives

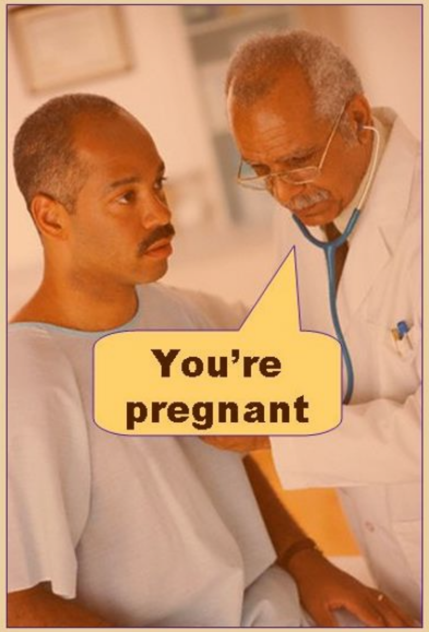
		Actual		
		Total	Pos	Neg
Predicted	Total	55	10	45
	Pos	7	5	2
	Neg	48	5	43

False positive

False negative

Type I error

(false positive)



Type II error

(false negative)

Evaluation metrics for classification

… ensure fair and thorough comparison of trained classification models

Key Metrics

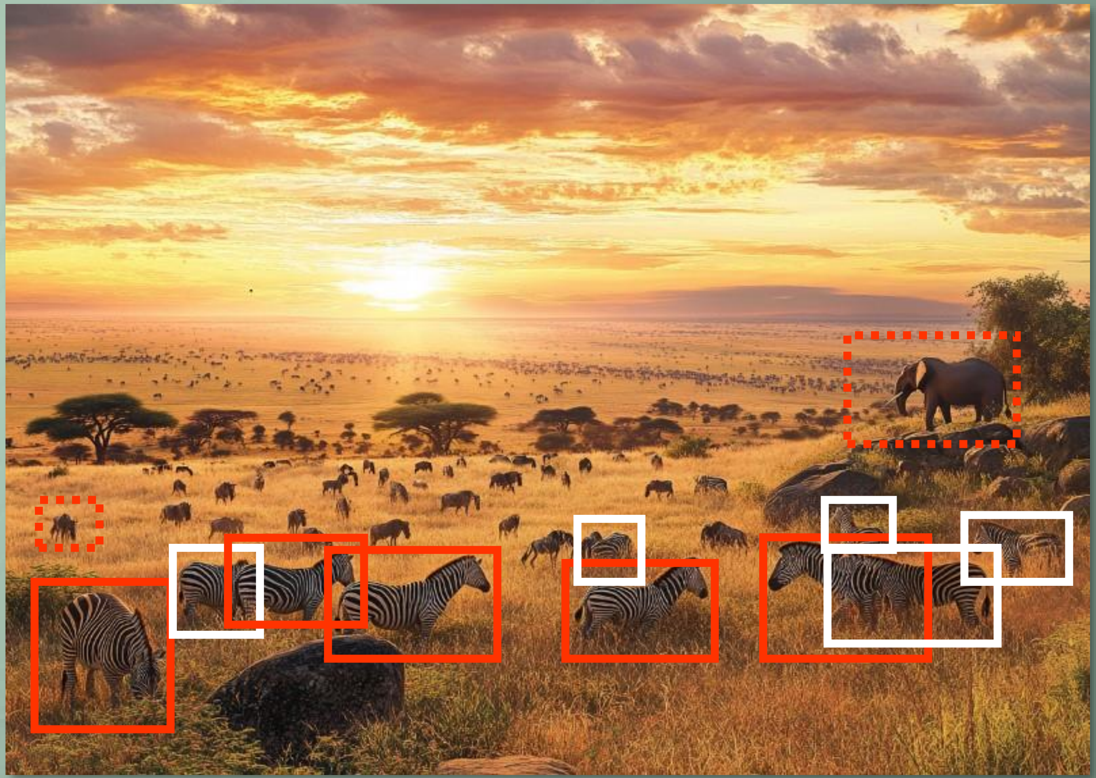
Accuracy:
Overall correctness of predictions in %

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

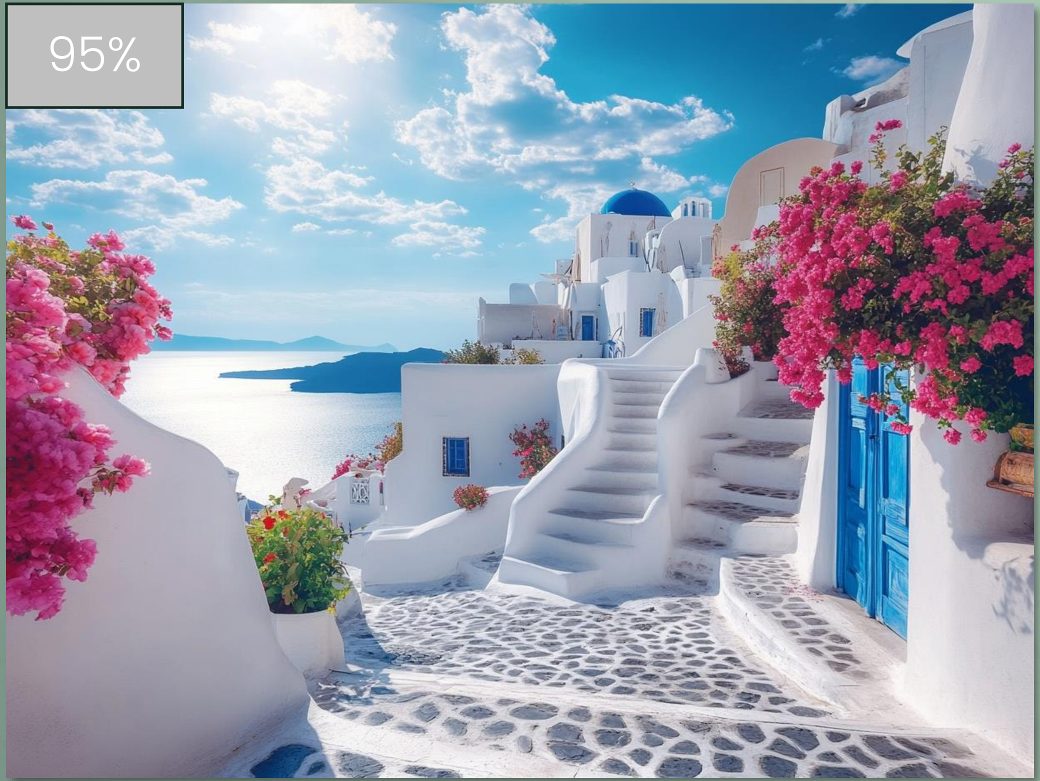


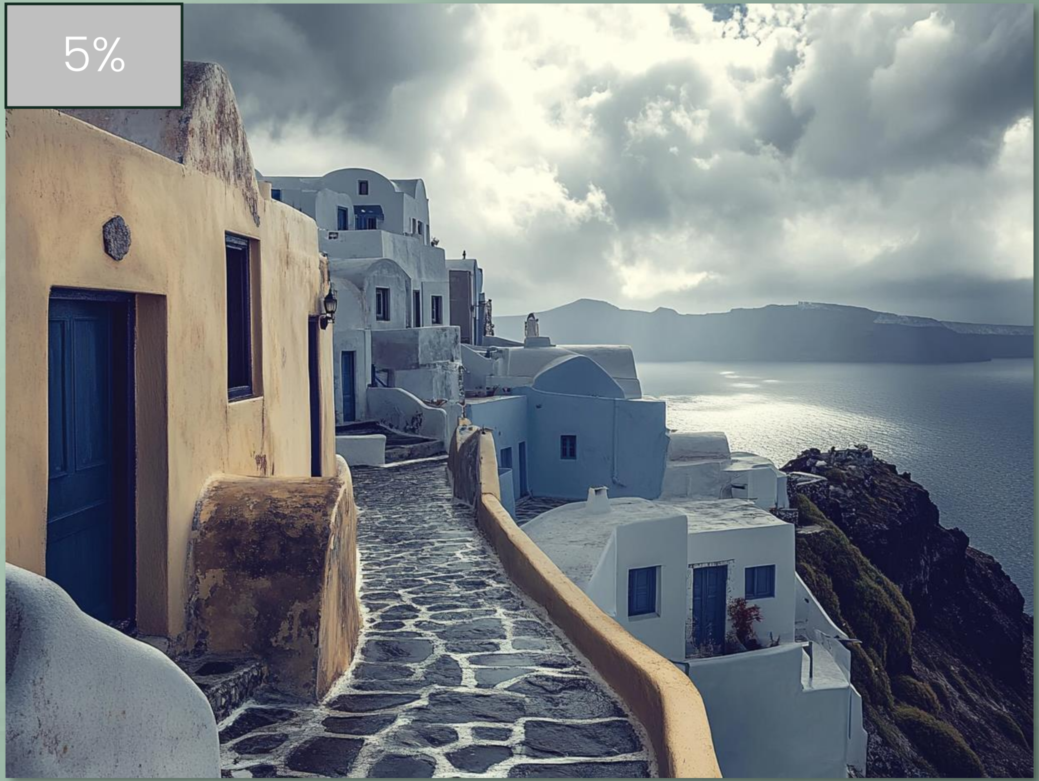
Fallacy of “accuracy” as evaluation metric of classification

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. Was this prediction good or bad?

It predicts rain (rare). Next day this takes place. Was this prediction good or bad?





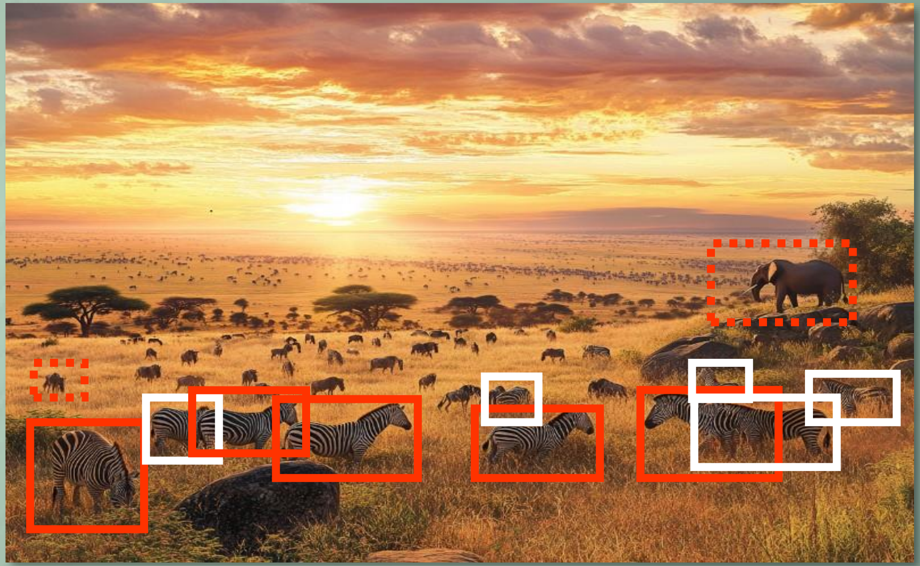
Model that always predicts sun has accuracy = 95%.
→ Therefore, baseline for accuracy must be at least 95%

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

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



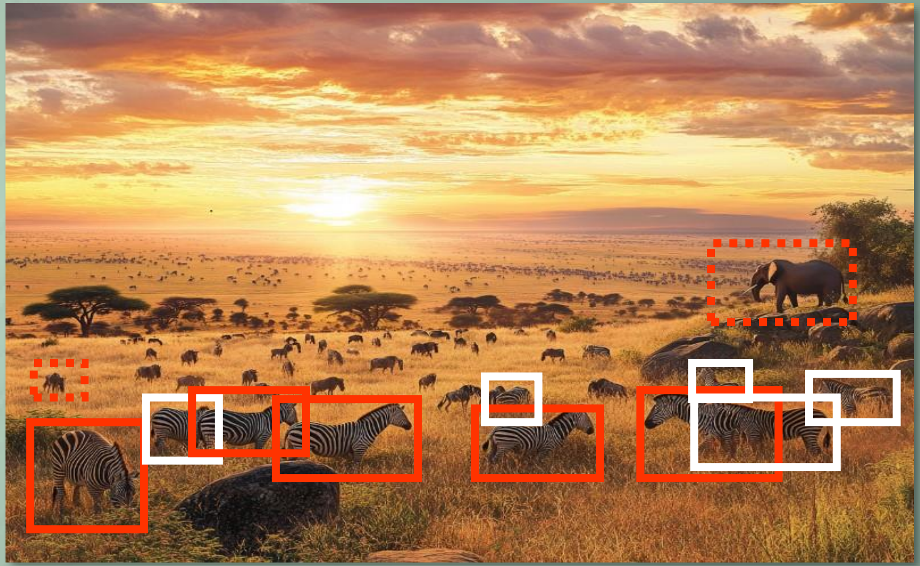
$$Accuracy(\%) = \frac{5 + 43}{55} = \mathbf{87\%}$$

Confusion matrix metric: Precision

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

Or: From all as positive predicted items, how many are actually positive?

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



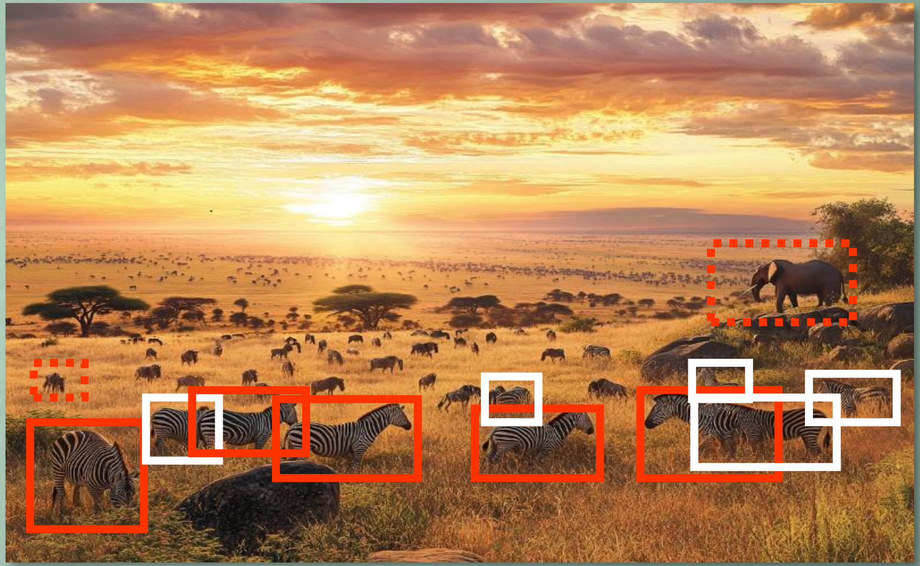
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

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

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



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

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

“Recall” is also called “Sensitivity”, they are the same.

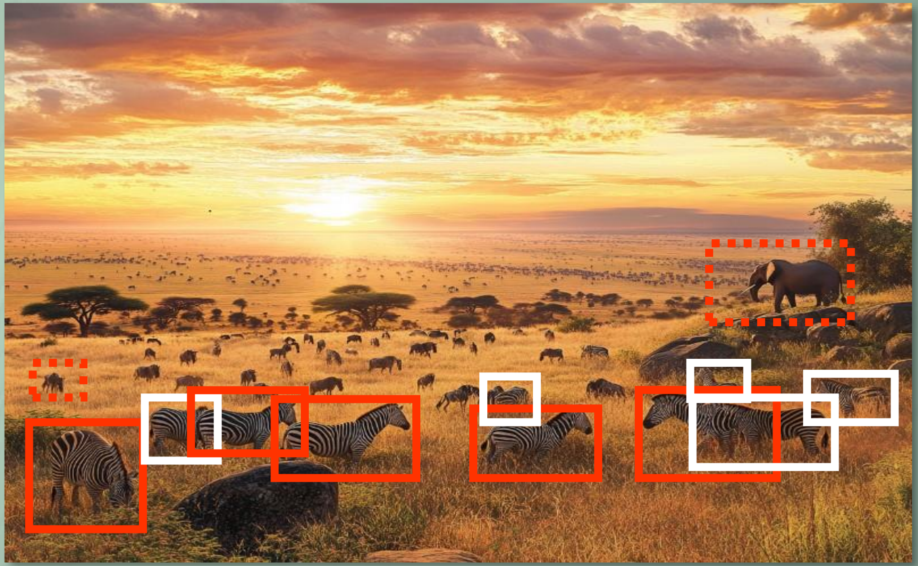
Confusion matrix metric: F1-Score

… 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

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



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

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	▪ Proportion of correctly predicted positives ▪ out of all predicted positives.	▪ **Accuracy** ▪ of positive predictions	▪ Minimizes false positives ▪ useful in scenarios where false positives are costly ▪ e.g., spam detection	▪ May miss true positives, ▪ leading to low recall.
Recall	▪ Proportion of actual positives ▪ correctly identified by the model.	▪ **Completeness** ▪ of positive predictions.	▪ Minimizes false negatives ▪ critical in scenarios where missing true positives is costly ▪ e.g., disease diagnosis	▪ May increase false positives ▪ leading to low precision.
F1-Score	▪ Harmonic mean of precision and recall; ▪ balances the two metrics.	Balance between precision and recall.	▪ 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: Default of credit card clients

Develop model for binary classification by generalizing the dataset

Number of zebras on image

Confusion matrix

Accuracy

Precision vs. Recall

F1-Score

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	Demographics				Past repayment history						Bill statement amounts						Previous payments made						Response Variable
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
ID	LIMIT_BAI	SEX	EDUCATI	MARRIAC	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	BILL_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	PAY_AM	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	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	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
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
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
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
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
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
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
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	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
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
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	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
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
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
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
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
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	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
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1
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
24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1
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
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
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
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
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
30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	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
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
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
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
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
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

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.

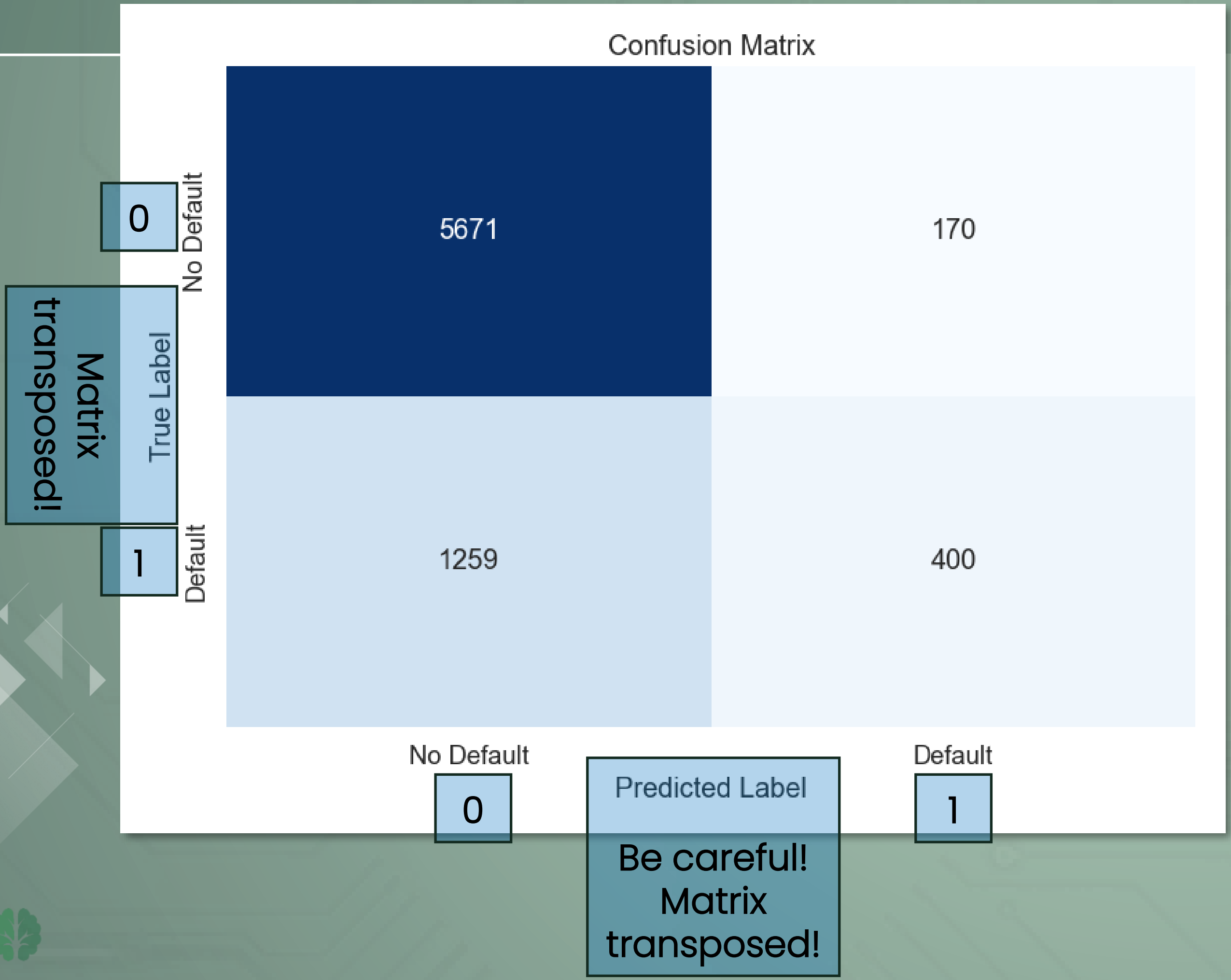
```python
# Train our custom logistic regression model
print("\nTraining our custom LogisticRegression model...")
# Using a smaller number of iterations for time constraints
my_model = MyLogisticRegression(learning_rate=0.1, num_iterations=10000)
my_model.fit(X_train_scaled, y_train)
```

```
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




```
5. MODEL EVALUATION
========================================

Classification Report (sklearn model):
              precision    recall  f1-score   support

           0       0.82      0.97      0.89      5841
           1       0.70      0.24      0.36      1659

    accuracy                           0.81      7500
   macro avg       0.76      0.61      0.62      7500
weighted avg       0.79      0.81      0.77      7500


Detailed Metrics:
Accuracy: 0.8095
Precision: 0.7018
Recall: 0.2411
F1-Score: 0.3589
```

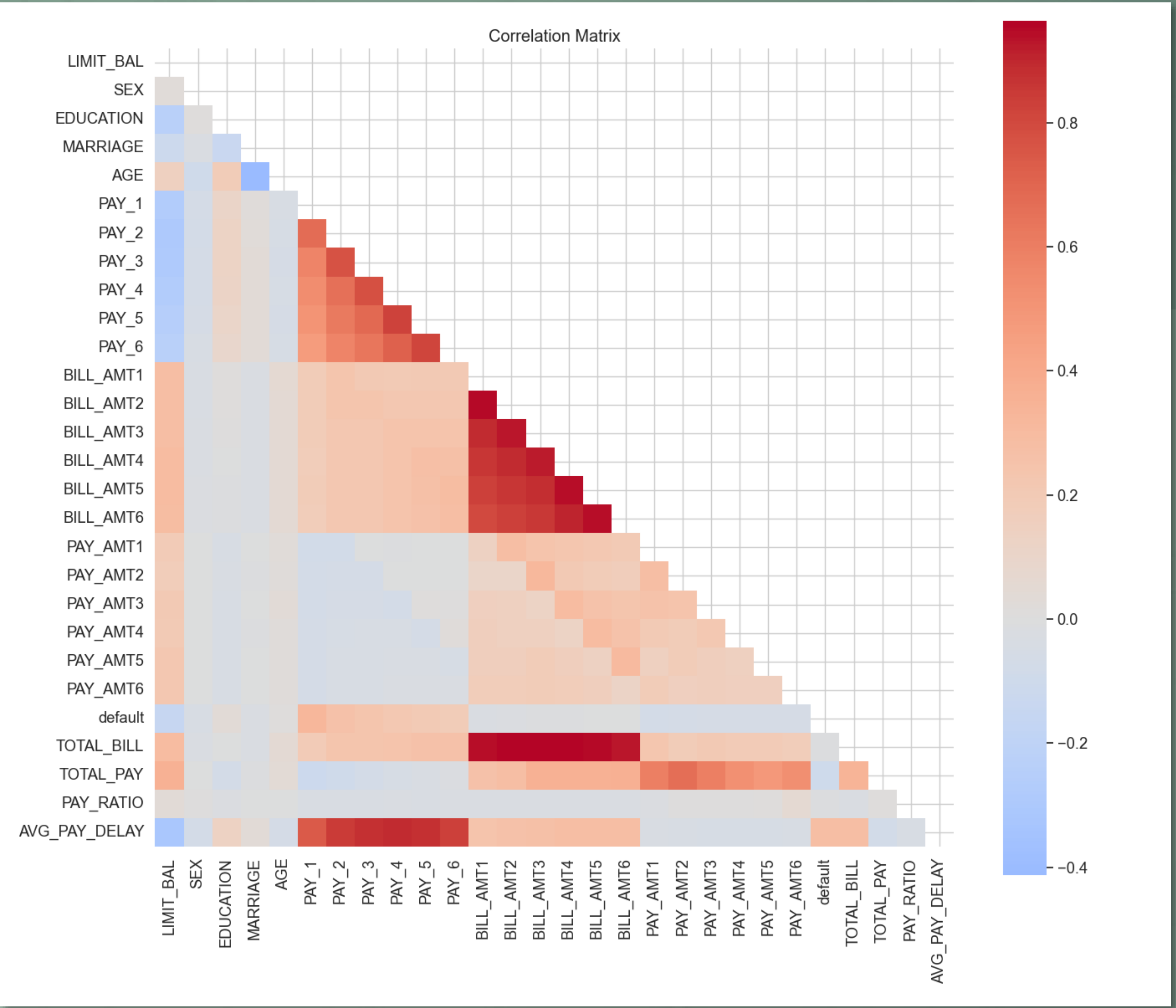
400/570

400/1651

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

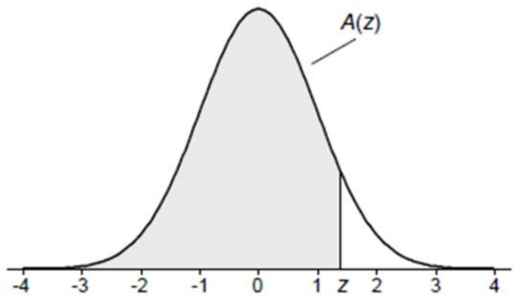
t Distribution: Critical Values of t

		Significance level					
Degrees of freedom	*Two-tailed test:*	10%	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
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

Cumulative Standardized Normal Distribution



$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Statistical tables

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

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

Degrees of freedom	*Significance level* 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_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
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=J0I1NuxUcpQ

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=6YzrVUVO9M0
	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=9lFcRTyhd5Y
		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-lE
	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=qYOMO83Z1WU

Links

i.e. sources for self-learning

	Title	Link
Model estimation	Scatter matrix , Covariance and Correlation Explained	https://medium.com/@raghavan99o/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=HLAg1yBfNk8&list=PLIbbWgBRF8EePgK40-i7aGU2_ky1yujgL&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/@raghavan99o/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=HLAg1yBfNk8&list=PLIbbWgBRF8EePgK40-i7aGU2_ky1yujgL&index=11

DATA SCIENCE RETREAT®
SINCE 2014

About me

Dr. Harald Stein

- Data Scientist ~ 8 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