Great Learning
POWER AHEAD

Welcome! We will begin shortly

Live Virtual Class

The Must-Know Mathematics & Statistics Behind AI

Learning Outcomes

- Understand the key mathematical and statistical concepts in the World of Data Science
- Understand how some of different mathematical and statistical concepts are used in different scenarios
- Get questions & concerns on mathematical and statistical concepts resolved

Guidelines

Listen only mode

Ask questions at the interest of the larger audience

Questions in the Q&A Box

Thank you

Kindly utilize the chat box for subject-relevant questions only to maximize your learnings from the session.

Probability

“Measure of the likelihood or chance of an event occurring”

$$\text{Probability of an event occurring} = \frac{\text{\# outcomes associated with the event}}{\text{Total \# of outcomes}}$$

← 100% Chance (Certainty)

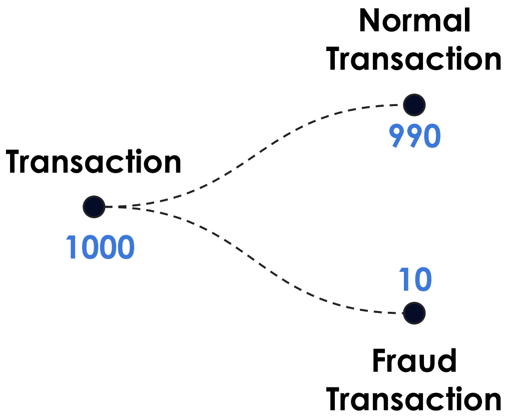
← 50% Chance (Equally Likely)

← 0% Chance (Impossibility)

Probability

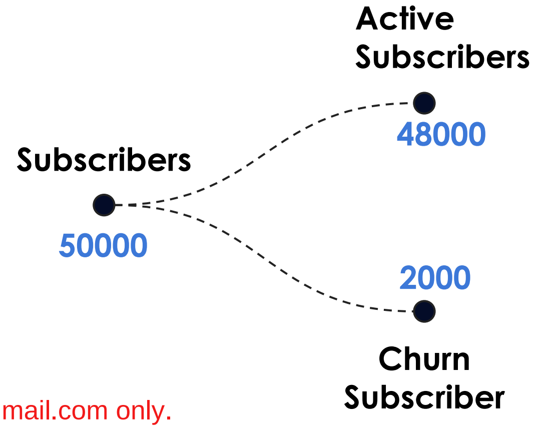
Example 1: 1000 transactions, 10 are fraudulent and 990 are normal. The probability of randomly selecting a fraudulent transaction is 0.01 or 1%.

- Fraudulent cases are rare but important to detect.



Example 2: In a telecom company with 50,000 subscribers, 2,000 churned last month. The probability of a randomly chosen subscriber churning is 4%.

- Understanding this helps in devising retention strategies for business sustainability.



Conditional Probability

“The probability of an event happening, given that another event has already happened”

$$\textbf{Probability of event A happening given B} = \frac{\textbf{Probability of both events A and B happening}}{\textbf{Probability of event B happening}}$$

Example: Out of 1000 transactions made in an hour. 50 transactions were fraudulent. Out of those 50, 35 transactions are related to phishing scams, and 15 are related to Robo call scams.

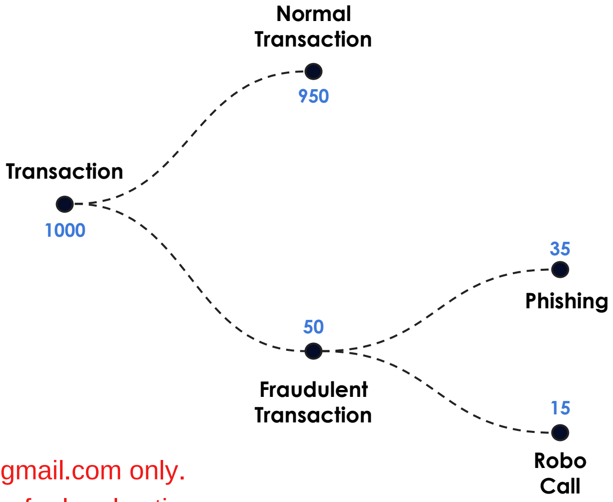
The probability of a transaction being a phishing scam given it is a fraudulent transaction can be computed as follows:

Probability = # phishing-related frauds / # fraudulent transactions

= 35/50

= 0.7 (or 70%)

So, if we know that a transaction is fraudulent, there is a 70% chance that it is a phishing scam.



Bayes Rule

“Determine probability of a hypothesis based on prior knowledge and new evidence”

$$\text{Probability of A happening given B} = \frac{\text{Probability of B happening given A * Probability of A happening}}{\text{Probability of event B happening}}$$

Business Applications

- **Spam Email Detection**: Calculate the probability that an email is spam given certain words in the subject/body
- **Medical Diagnoses**: Estimate the probability of a patient having a particular disease given their symptoms and medical history
- **Fraud Detection**: Detect fraudulent activities, such as credit card fraud, by analyzing transaction patterns

Bayes Rule

Example: In a township, 1% people have COVID. A new test for detecting COVID has been devised and it correctly detects COVID 80% of the time, but flags non-COVID cases as COVID 10% of the time. Given that a random citizen's test result was Yes, the chances that they have COVID would be computed as follows:

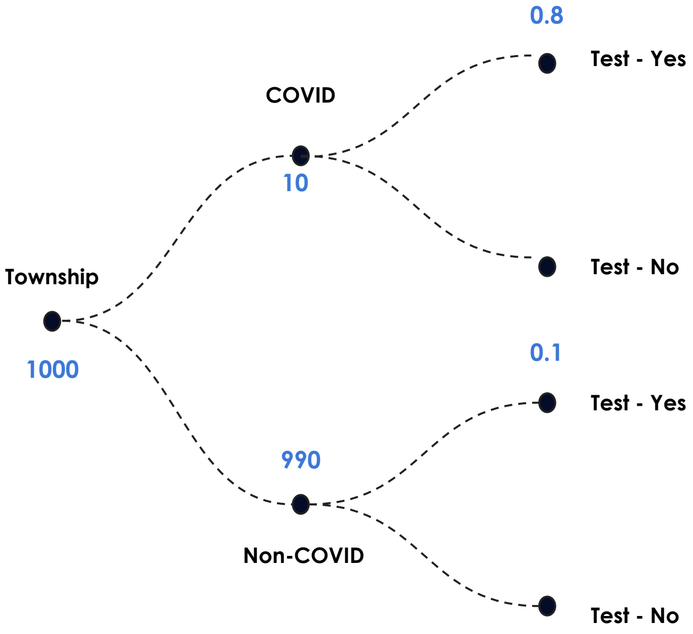
P(COVID | Test - Yes)

$$= \frac{P(\text{Test - Yes} \mid \text{COVID}) * P(\text{COVID})}{P(\text{Test - Yes})}$$

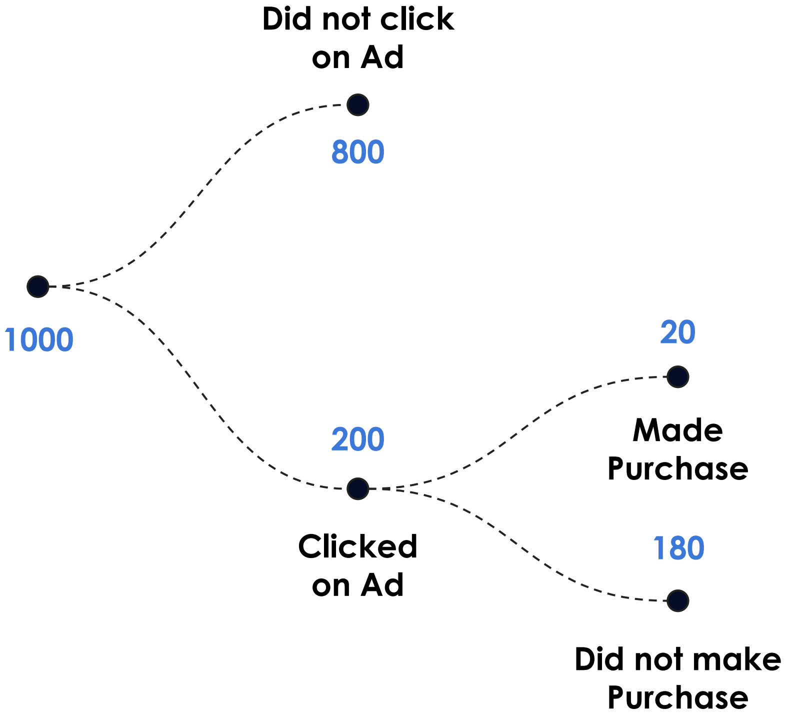
$$= \frac{P(\text{Test - Yes} \mid \text{COVID}) * P(\text{COVID})}{P(\text{Test - Yes} \mid \text{COVID}) * P(\text{COVID}) + P(\text{Test - Yes} \mid \text{Non-COVID}) * P(\text{Non-COVID})}$$

$$= \frac{0.8 * 0.01}{(0.8 * 0.01) + (0.1 * 0.99)} = \mathbf{0.0748 \sim 7.5\%}$$

So, if a random citizen's test result is Yes, then there is a 7.5% chance that they have COVID.



Probability - Quiz



A company is running a marketing campaign to promote its new product. From the adjacent flowchart, compute the probability that customer who clicks on the ad will make a purchase?

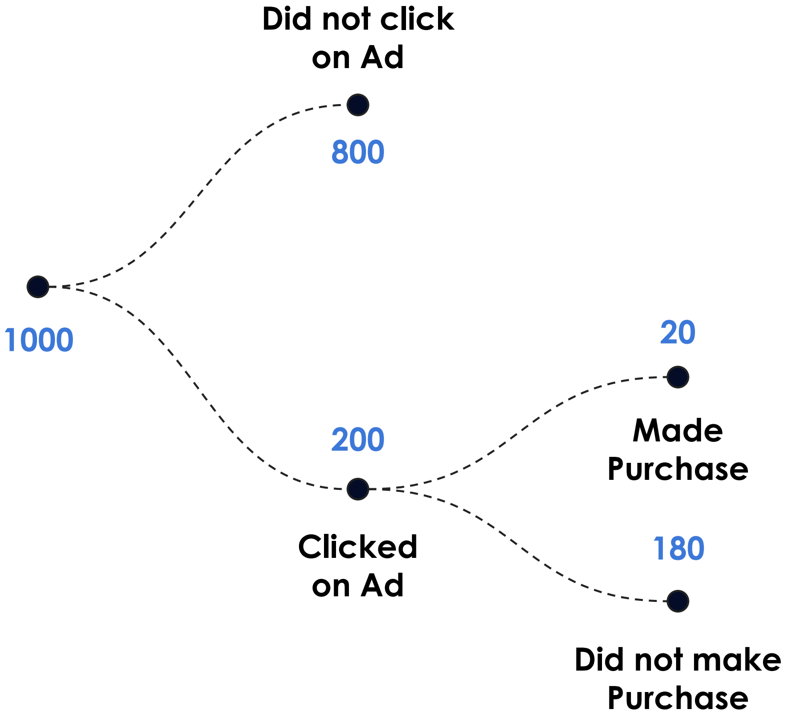
A 0.2

B 0.1

C 0.32

D 0.12

Probability - Quiz



Did not click on Ad

800

1000

200

Clicked on Ad

20

Made Purchase

180

Did not make Purchase

A company is running a marketing campaign to promote its new product. From the adjacent flowchart, compute the probability that customer who clicks on the ad will make a purchase?

- A 0.2
- B [✓ 0.1]
- C 0.32
- D 0.12

Probability - Quiz

0.2

0.1

0.32

0.12

The question aims to emphasize that the purchase was made after clicking on the ad. Once the user clicks on the ad, there are only two possible outcomes - either they make the purchase (buy the product) or they do not make the purchase.

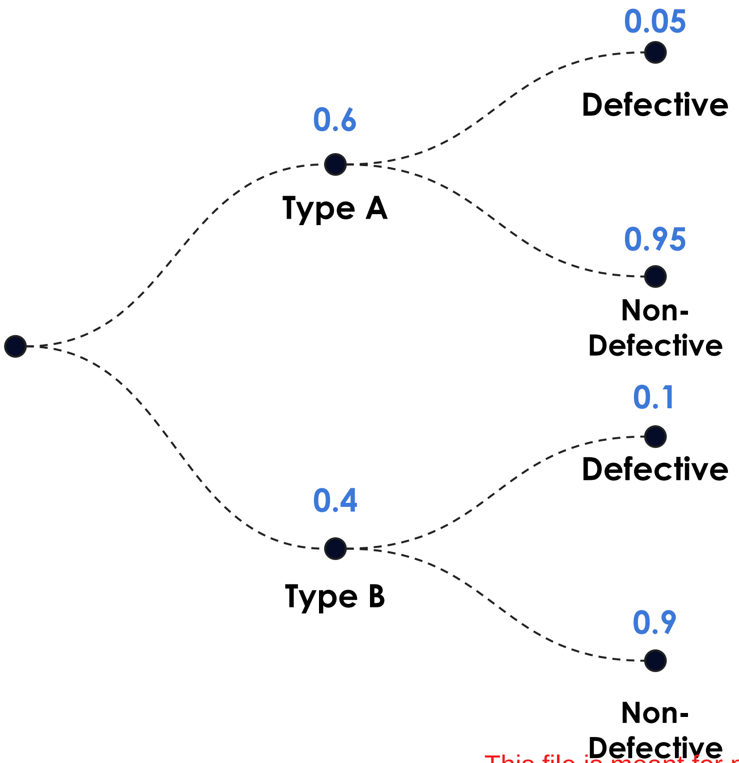
Out of the 200 ad clicks, 20 of them resulted in a purchase. So, we have

P(click on ad) = 200/1000 = 0.2
P(click on ad **and** make a purchase) = 20/1000 = 0.02

P(make a purchase given that they clicked on the ad)
= P(click on ad **and** make a purchase) / P(click on ad)
= 0.02 / 0.2
= 0.1

Therefore, the probability of a customer clicking on the ad and then making a purchase is 0.1.

Bayes Rule - Quiz



A factory produces two types of electronic components: Type A and Type B. Type A and Type B components make up 60% and 40% of the production, with defect rates of 5% and 10%, respectively. If a randomly selected component is found to be defective, what is the probability that it is Type B?

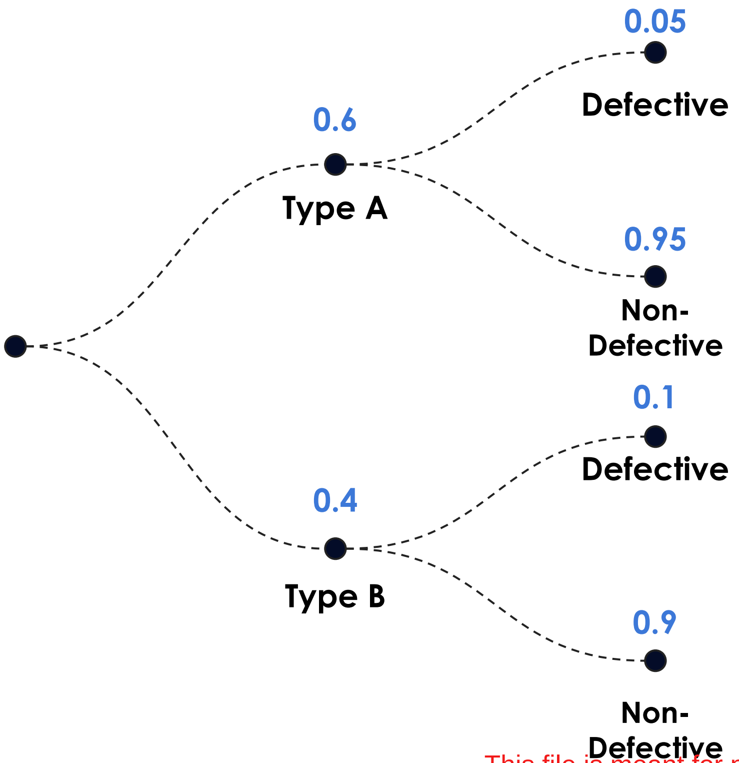
A 25%

B 30%

C 60%

D 40%

Bayes Rule - Quiz



A factory produces two types of electronic components: Type A and Type B. Type A and Type B components make up 60% and 40% of the production, with defect rates of 5% and 10%, respectively. If a randomly selected component is found to be defective, what is the probability that it is Type B?

- A 25%
- B 30%
- C [✓ 60%]
- D 40%

Bayes Rule - Quiz

25%

30%

[✓ 60%]

40%

We are asked to find the conditional probability of component being Type B given that it is defective. We can use Bayes' Rule to calculate this probability.

P(Type B | Defective)

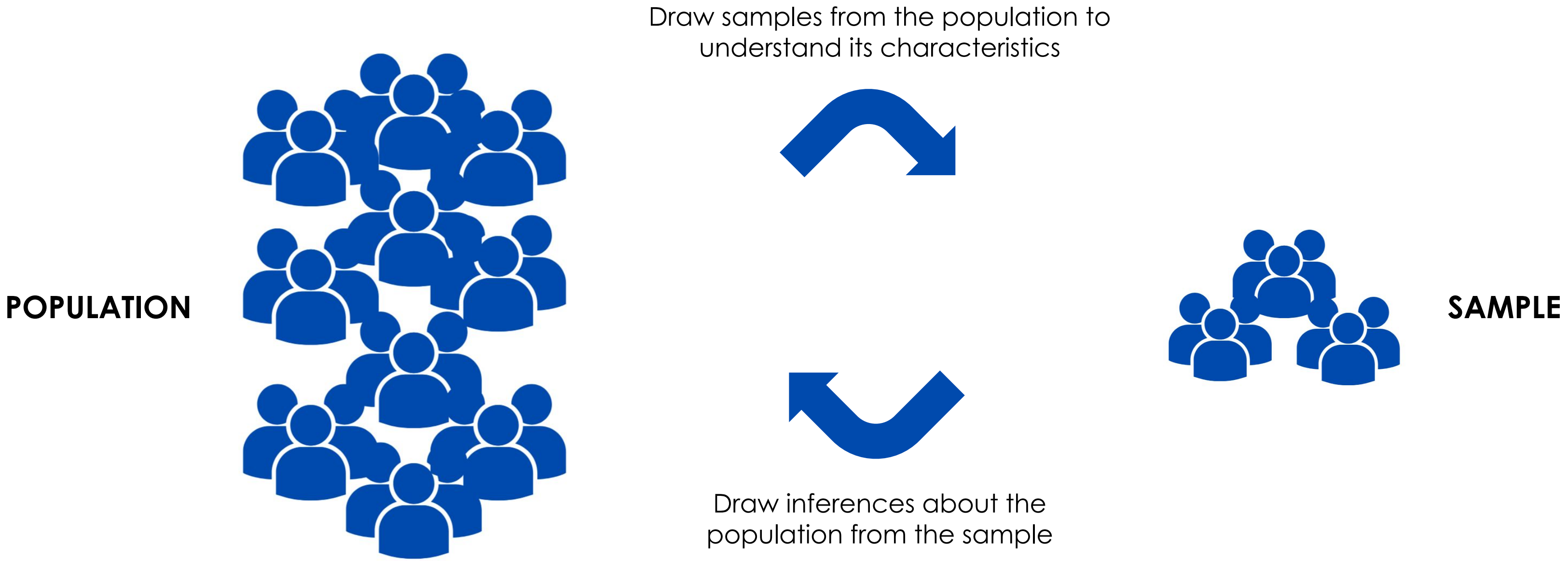
$$= \frac{P(\text{Defective} \mid \text{Type B}) * P(\text{Type B})}{P(\text{Defective})} = \frac{P(\text{Defective} \mid \text{Type B}) * P(\text{Type B})}{P(\text{Defective} \mid \text{Type A}) * P(\text{Type A}) + P(\text{Defective} \mid \text{Type B}) * P(\text{Type B})}$$

We know that
P(Type A) = 0.6
P(Type B) = 0.4
P(Defective | Type A) = 0.05
P(Defective | Type B) = 0.1

Putting these values in the formula, we get
P(Type B | Defective)
= (0.1 * 0.4) / [(0.05 * 0.6) + (0.1 * 0.4)]
= 0.5715 ~ 0.6

Therefore, if a randomly selected component is found to be defective, the probability that it is Type B is ~0.6, i.e., ~60%

Types of Statistics



Draw samples from the population to understand its characteristics

POPULATION

SAMPLE

Draw inferences about the population from the sample

Inferential Statistics

Confidence Intervals - The size of the transistor on a processor chip lies in the 95% confidence interval (4.95, 5.05) nm

Hypothesis Testing - Does the conversion rate of a marketing campaign vary with the font style of the infographic?

Descriptive Statistics

Central Tendency - Mean, Median, Mode

Dispersion - Variance, Range, Standard Deviation

Descriptive Statistics

Mean

A measure of the centre of the data, and is computed as the sum of all data points divided by the total number of data points

Example:

The annual salaries of 6 employees in an organization (in thousands dollars) are as follows:

40, 42, 39, 45, 48, 50

Mean = Sum of Salaries of Employees / Total Number of Employees

= 264 / 6

= 44

The Mean(average) salary of employees in organizations is $44K

Standard Deviation

A measure of how dispersed the data is in relation to the mean

- Low standard deviation - data is clustered around the mean
- High standard deviation - data is more spread out

Example:

In the above scenario, the standard deviation of employee salaries comes out to be ~$4.5K.

This is a comparatively low value, indicating that the data is clustered around the mean

Descriptive Statistics

- In the previous scenario, a new employee joined the organization at an annual salary of $150K
- New mean salary of employees - ~$60K
- $60K **not reflective of the centre of the data** - an impact of the one extremely high value
- Need a ‘better’ measure of the centre of the data

Descriptive Statistics

- In the previous scenario, a new employee joined the organization at an annual salary of $150K
- New mean salary of employees - ~$60K
- $60K **not reflective of the centre of the data** - an impact of the one extremely high value
- Need a 'better' measure of the centre of the data

Median

The middle value of the data when arranged in an order

- Odd number of data points - the actual middle number
- Even number of data points - the average of the two middle numbers

Example:

The annual salaries of the seven employees (in thousands dollars), arranged in an order, are as follows:

39, 40, 42, 45, 48, 50, 150

Since we have an odd number (7) of data points, the median is the middle value. So, we say that the median annual salary of employees in the organization is 45

Descriptive Statistics

- The employees whose salary we discussed previously are from the following department:

  Sales, Sales, Marketing, Sales, Marketing, HR, Finance

- What is the 'centre' of the data now?
- Can't use mean and median - **no numbers!**
- Need a 'better' measure of the centre of non-numerical data

Descriptive Statistics

- The employees whose salary we discussed previously are from the following department:

  Sales, Sales, Marketing, Sales, Marketing, HR, Finance
- What is the 'centre' of the data now?
- Can't use mean and median - **no numbers!**
- Need a 'better' measure of the centre of non-numerical data

Mode

The value that occurs the most often in the data

- Data can have multiple modes
- Best used when you want to indicate the most common response or item in a data set

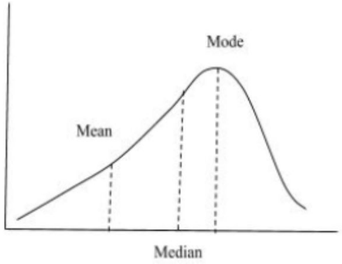
Example:

In the current scenario, the most frequently occurring value is Sales

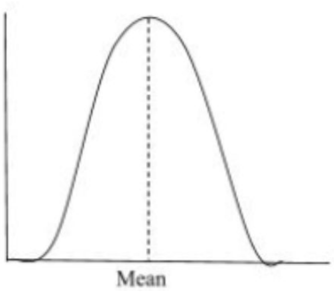
So, the mode of the data is Sales

This information helps business leaders understand which department forms the largest part of the workforce and they can allocate resources accordingly to ensure it functions efficiently and meets business objectives..

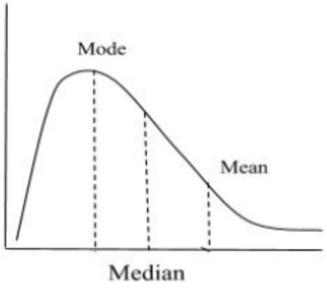
Mean vs Median vs Mode - When to use



Skewed Data
mean < median < mode



Symmetrical Data
mean = median = mode



Skewed Data
mode < median < mean

←

Negative Direction

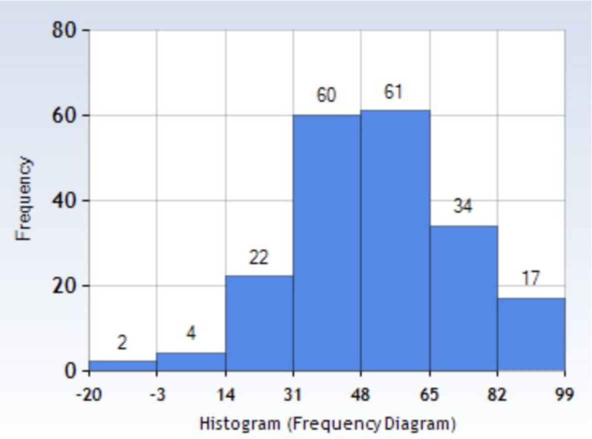
→

Positive Direction

Statistics - Quiz

Assume that the below distribution was created using 200 data points. The mean of this distribution is 50. The median of this dataset will also be roughly the same.

If you add another 100 data points with values between 60 and 75, what will the relation between the mean & the median look like?



A Mean > Median

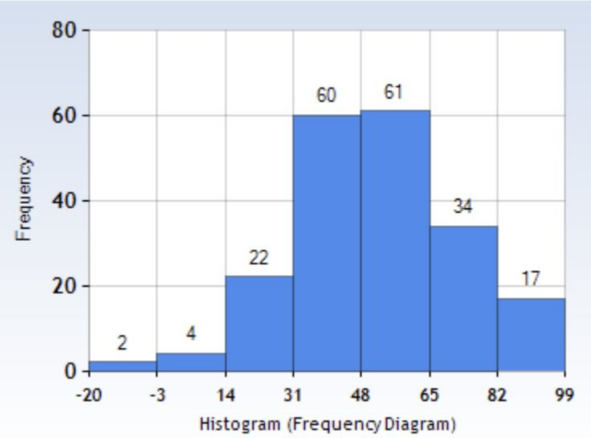
B Mean < Median

C Mean = Median

Statistics - Quiz

Assume that the below distribution was created using 200 data points. The mean of this distribution is 50. The median of this dataset will also be roughly the same.

If you add another 100 data points with values between 60 and 75, what will the relation between the mean & the median look like?



A Mean > Median

[✓ B Mean < Median]

C Mean = Median

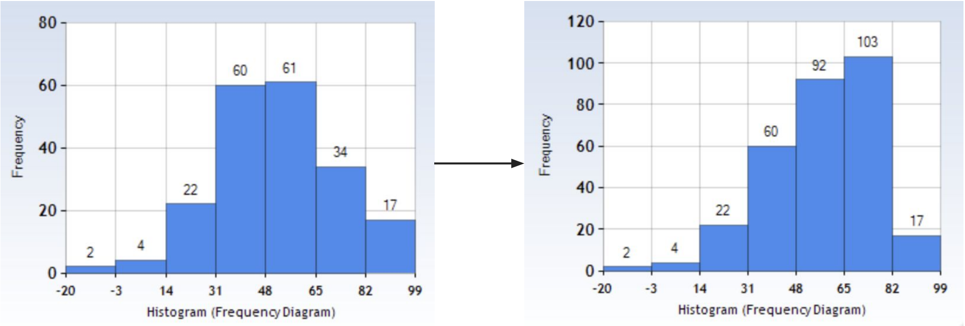
Statistics - Quiz

Mean > Median

Mean < Median

Mean = Median

While the exact numbers of mean & median will depend on the actual data that was inserted, it can be inferred that with the addition of higher-valued data into the dataset, the data will lose its symmetry and be positively skewed. This will result in the median of the data being higher than the mean.



Statistics - Quiz

An organisation pays $30,000 a year to 99 of its employees, while the salary of the CEO (not included in the 99 considered before) is $600,000 a year.

Assume that the organisation is hiring for a job position, which metric do you think will the organisation advertise, in order to make itself appear as a high-paying organisation?

A Mean

B Median

Statistics - Quiz

An organisation pays $30,000 a year to 99 of its employees, while the salary of the CEO (not included in the 99 considered before) is $600,000 a year.

Assume that the organisation is hiring for a job position, which metric do you think will the organisation advertise, in order to make itself appear as a high-paying organisation?

A Mean

B Median

Statistics - Quiz

Mean

Median

The median salary of the organisation will be roughly $30,000. However, if you consider the mean salary of the organisation, that number will be roughly ~$35,000.

While the figures may appear impressive at first glance, they do not provide a complete picture as one of the employees (who happens to be the boss) is earning considerably more than the other workers. As a result, such outcomes can be classified as **deceptive statistics**. This is an example of why it is usually better to check if the 'average' metric being used is median or mean, in order to be able to form a more informed opinion

Matrices

“Rectangular array of numbers, symbols, or expressions arranged in rows and columns”

5	5	5	2	1	4	1
4	4	4	3	1	3	3
3	3	3	3	2	4	4
2	2	2	1	2	2	5
5	5	5	3	2	4	3

Rows (m) = 5
Columns (n) = 7

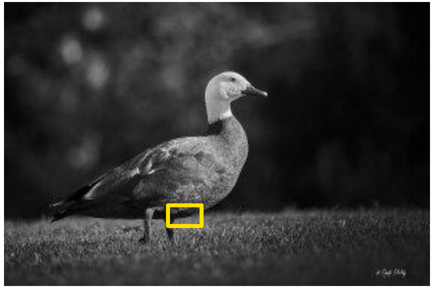
Tabular Data Representation using Matrices

**"Each row is a record (customer, object, etc.),
each column is an attribute (age, dimensions, etc.)"**

	Cylinders	**Weight**	**Displacement**	**Horsepower**	**Model Year**	**Country**	**MPG**
Car 1	8	3504	307	130	1970	USA	18
Car 2	4	2372	113	95	1970	Japan	24
Car 3	4	2130	97	88	1970	USA	27
Car 4	3	2320	70	97	1970	Japan	19
Car 5	3	2130	70	90	1970	Germany	18

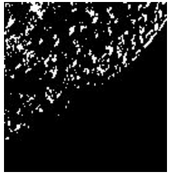
Image Representation using Matrices

“Convert an image into an array of numbers and each pixel has an intensity number”



128 x 128 matrix

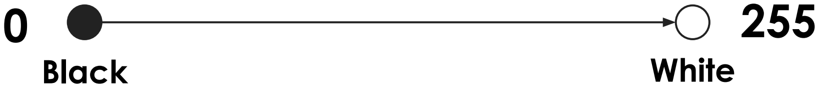
→



→

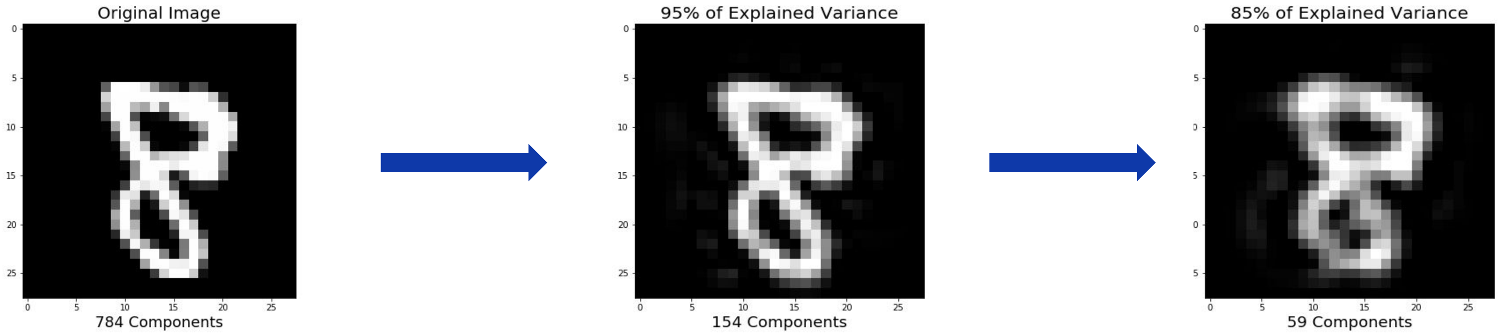
010	146	025
057	071	082
012	098	099

9 x 9 matrix



Dimensionality Reduction

"Increasing the interpretability of data while preserving the maximum amount of information"



Original Image

784 Components

95% of Explained Variance

154 Components

85% of Explained Variance

59 Components

Dimensionality Reduction - Quiz

Which of the following are the benefits of using Dimensionality Reduction?

A Increase quality of the data

B Reduce storage space consumed by the data

C Reduces the computational complexity of machine learning algorithms

D Help improve accuracy of your machine learning algorithms

Dimensionality Reduction - Quiz

Which of the following are the benefits of using Dimensionality Reduction?

A Increase quality of the data

B Reduce storage space consumed by the data

C Reduces the computational complexity of machine learning algorithms

D Help improve accuracy of your machine learning algorithms

Dimensionality Reduction - Quiz

Increase number of features in the data

Increasing the number of features in a dataset is the opposite of what dimensionality reduction aims to do. Adding more features can actually make it more difficult to analyze and understand the data, as well as increase the computational complexity of machine learning algorithms

Reduce storage space consumed by the data

It help with data compression and storage by reducing the size of the dataset and by helping remove redundant or irrelevant information from the dataset

Reduces the computational complexity of machine learning algorithms

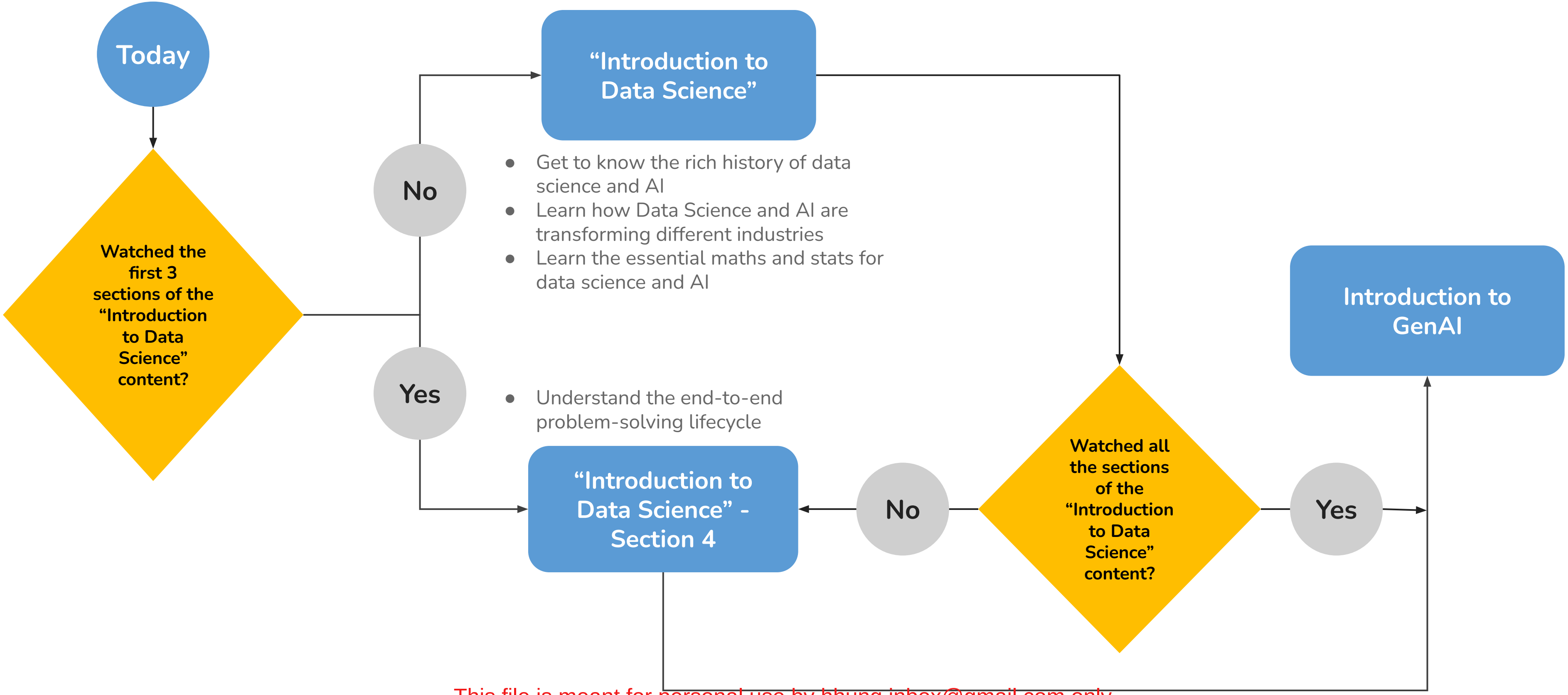
It helps reduce the number of features in a dataset, which can lead to faster and more efficient machine learning algorithms.

Improve accuracy of your machine learning algorithms

While the effects of dimensionality reduction on the data, might, in some cases, help improve the accuracy of the machine learning model as a byproduct. However, this is not a given and could only happen in some scenarios.

QnA

Next Steps


Today
Watched the first 3 sections of the "Introduction to Data Science" content?
No
Yes
"Introduction to Data Science"
"Introduction to Data Science" - Section 4
No
Watched all the sections of the "Introduction to Data Science" content?
Yes
Introduction to GenAI

- Get to know the rich history of data science and AI
- Learn how Data Science and AI are transforming different industries
- Learn the essential maths and stats for data science and AI

- Understand the end-to-end problem-solving lifecycle

Thank you!

We'd love to hear your feedback!
Please share your feedback for the session

Wish you all the very best!

Please feel free to raise a Support Request through Olympus in case of any queries