STATISTICS

for Data Science

Ronny Fahrudin | Data Science Fellowship at IYKRA

OUTLINE

Understanding about:

1. Definition Statistics
2. Why data science need statistics?
3. Data
   - What is data?
   - Type of variables
4. Descriptive Statistics
5. Inferential Statistics
   - Hypothesis Testing

Definition Statistics


? ? ?
What is Statistics?

Statistics is a mathematical science pertaining to data collection, analysis, interpretation, and presentation

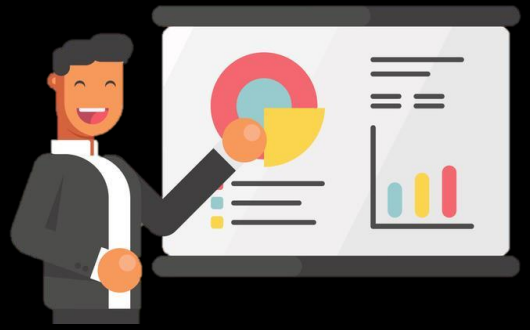

Why data science need statistics?

Collect

Organize

Analyze

Interpret



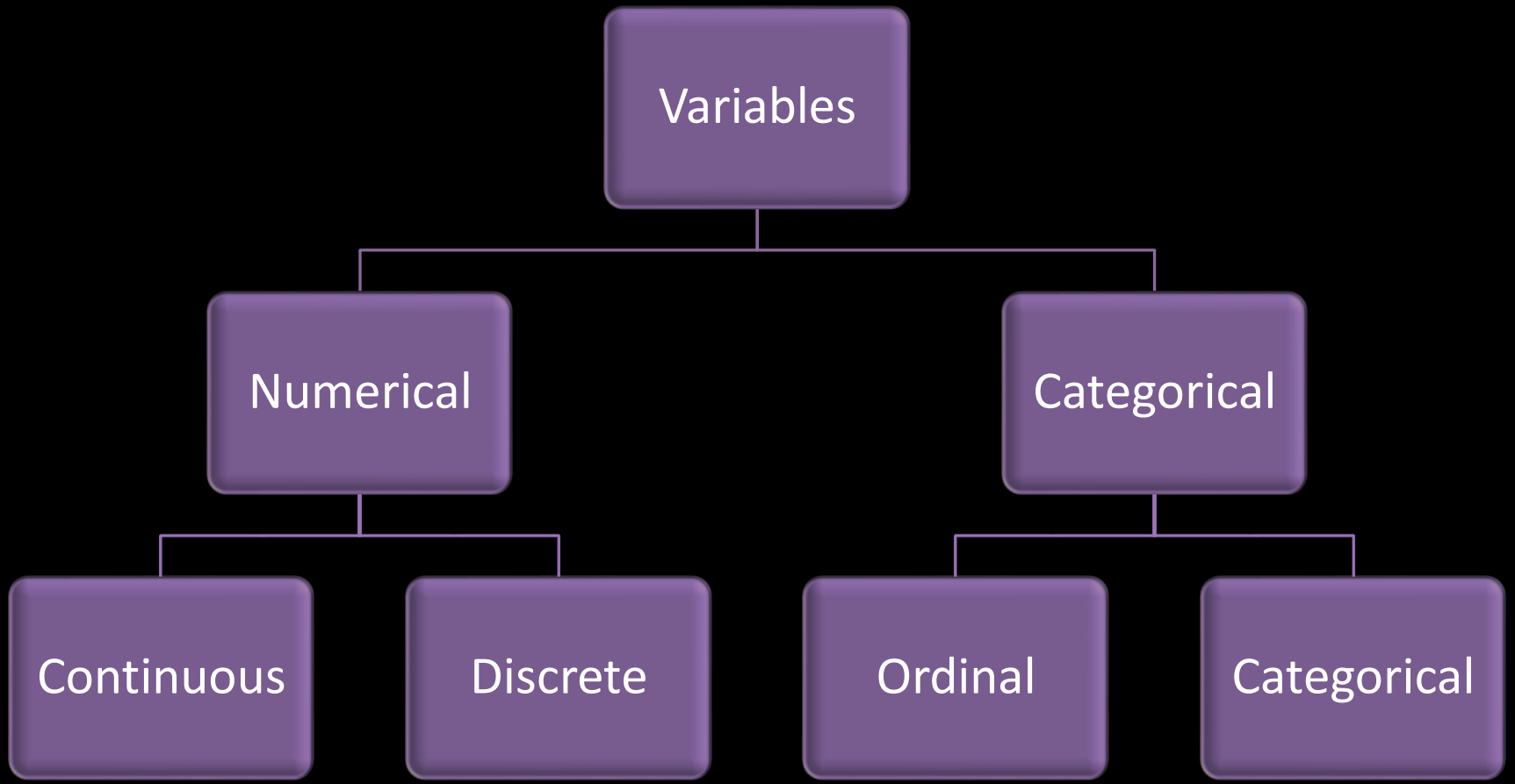
Present

What is Data?

Data in Statistics contexts

Data is information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used be a computer.

Variance Variables



- Continuous: height, temperature, length
- Discrete: Num of children, num of cars, num people.
- Ordinal: rank, education level, levels satisfaction.
- Categorical: binary, nominal, gender, Boolean, language

Types of Statistics

Applied Statistics

Inferential Statistics

Drawing conclusion a Population based on Data observed in sample

Descriptive Statistics

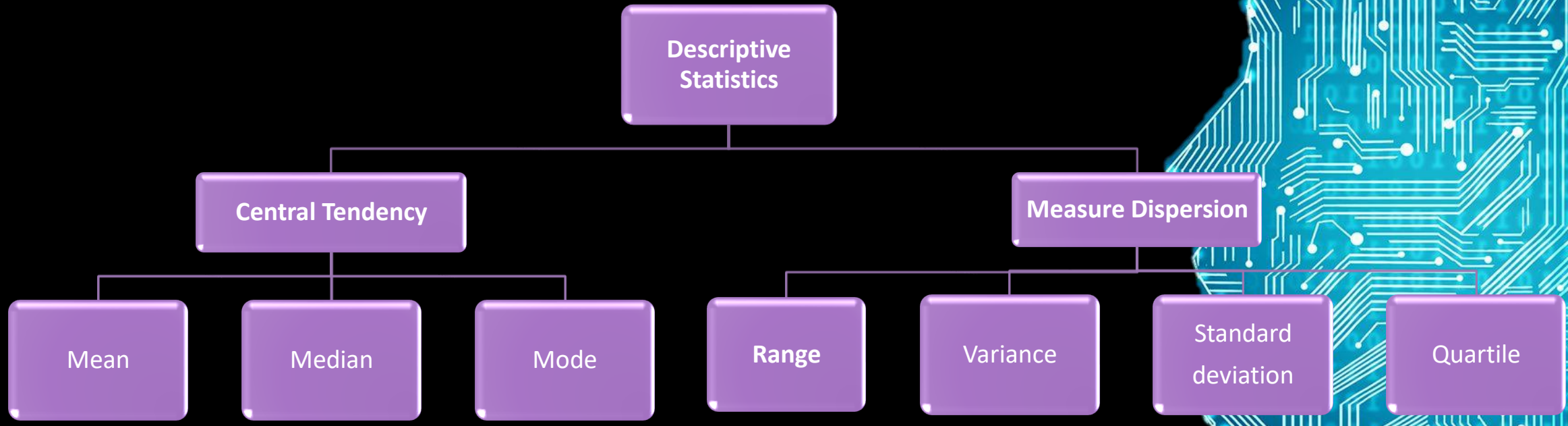
Organizing, Summarizing, and presenting data

Theoretical Statistics

Mathematical statistics philosophycally

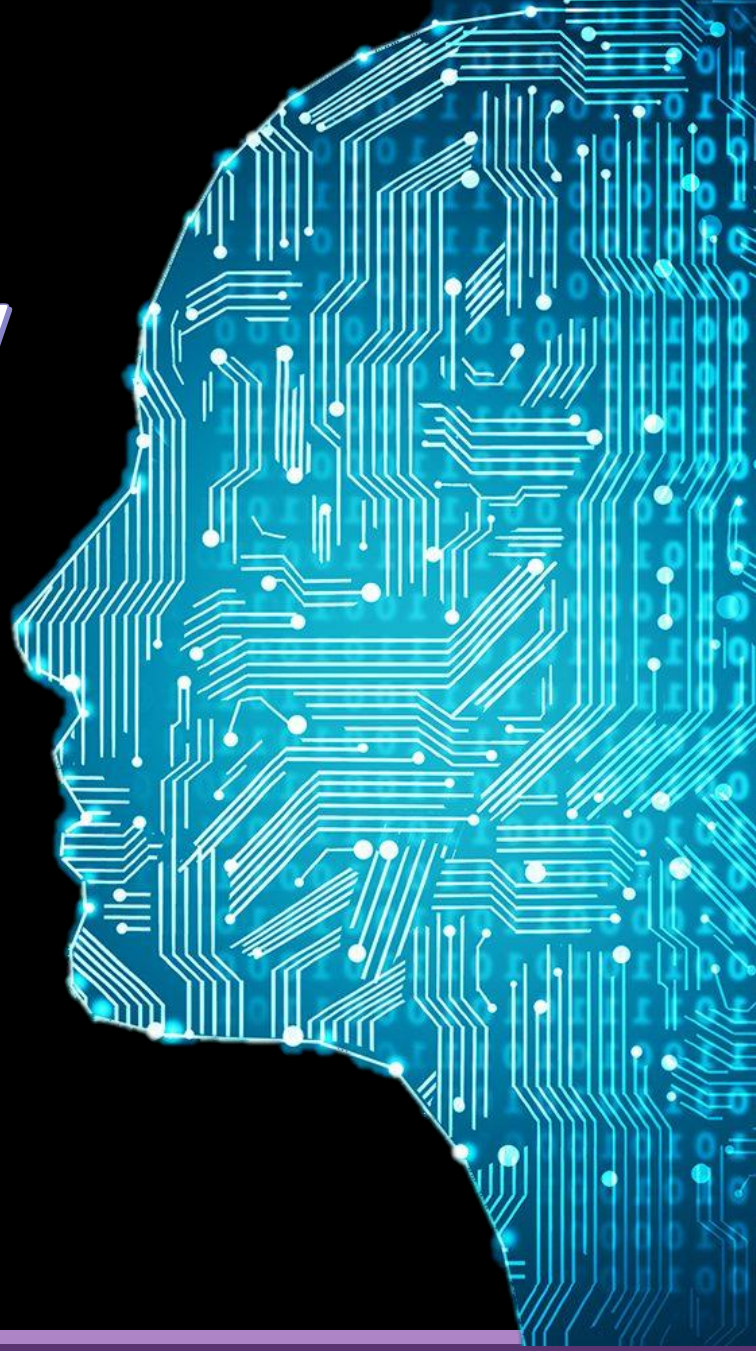
Descriptive Statistics

To Understanding



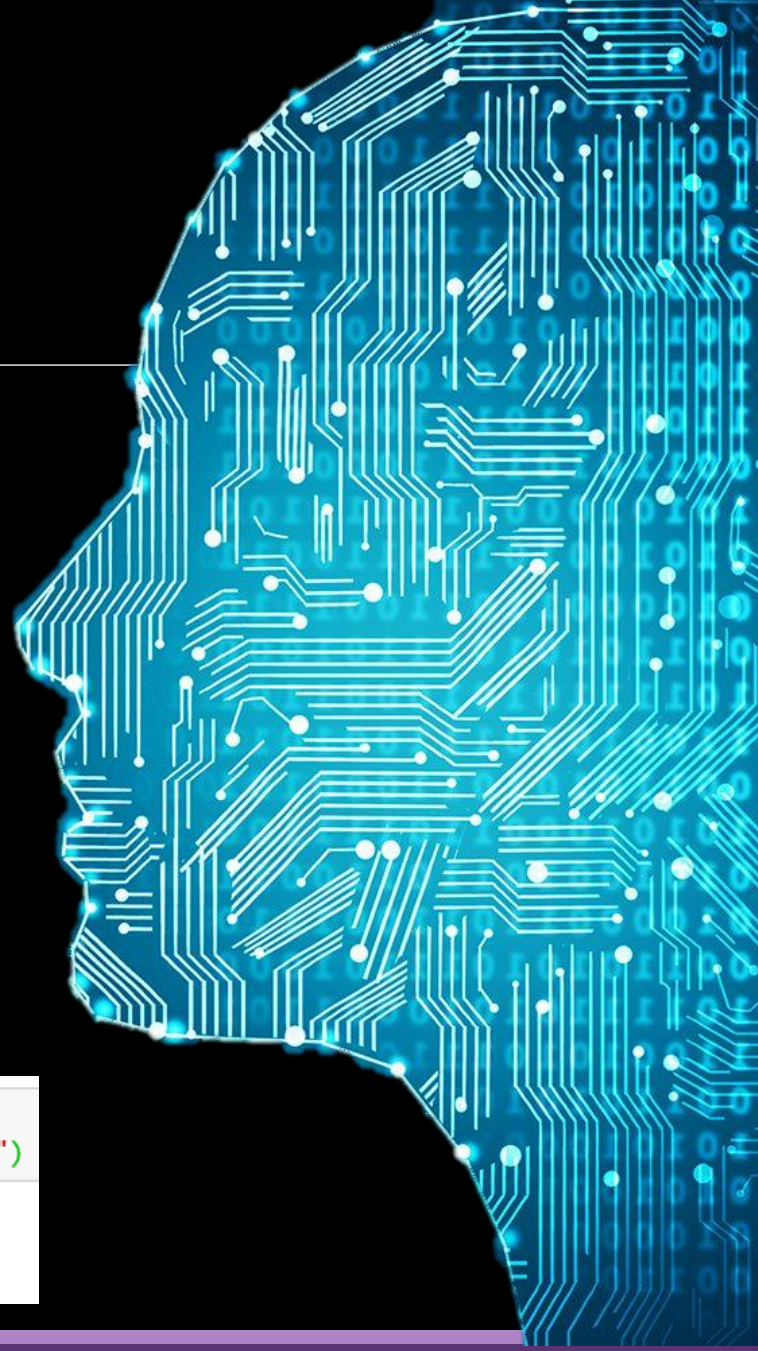
What is Descriptive Statistics?

Descriptive Statistics is summary Statistic that quantitatively describe or summary feature from of collection information.



CENTRAL TENDENCY

Central tendency is a single values that attempts to describe a set of data by identifying the central position within that set of data.



Mean

Arithmetic Mean Formula

$$\text{Arithmetic Mean} = \Sigma \frac{x_i}{n}$$

$$\text{Arithmetic Mean} = \Sigma \frac{(f_i * x_i)}{f_i}$$

Median

$$\text{Median Formula} = \frac{(n+1)}{2}$$

Mode

$$\text{Mode Formula} = L + \frac{(f_m - f_1) \times h}{(f_m - f_1) + (f_m - f_2)}$$

```
data = [1,2,3,4,5,5,5,5,6,7,3,11,11,4,15,20]
print(f"mean of data: {np.mean(data)},\nmedian of data: {np.median(data)},\nmode of data: {stats.mode(data)[0][0]}")
```

```
mean of data: 6.6875,
median of data: 5.0,
mode of data: 5
```

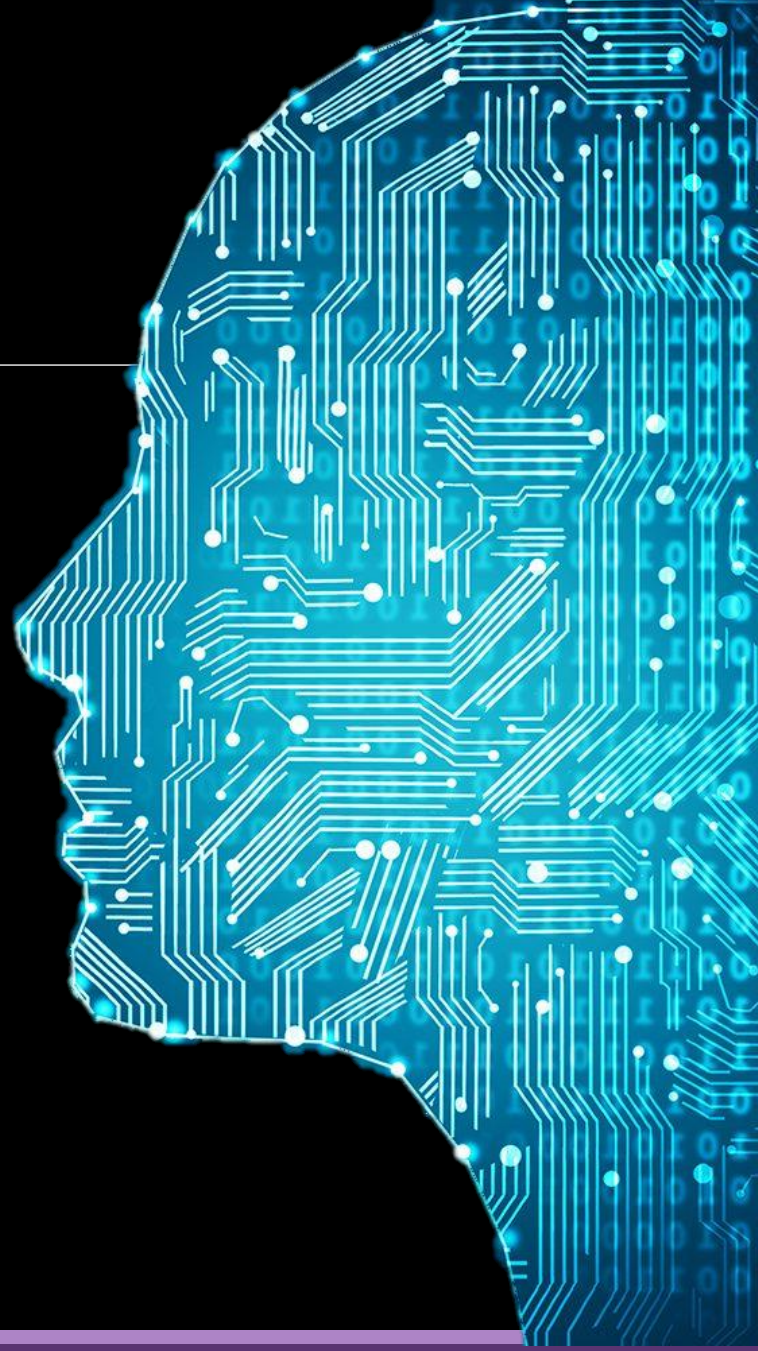
Range

Range = Max. Value – Min. Value

= 43 - 12

= 31

Name	Age
Dini	12
Dono	32
Doni	43
Dani	32
Dadang	22
Didik	24
Dudun	26
Deni	31



Variance & Standard Deviation

Used to measure the dispersion and variation of data

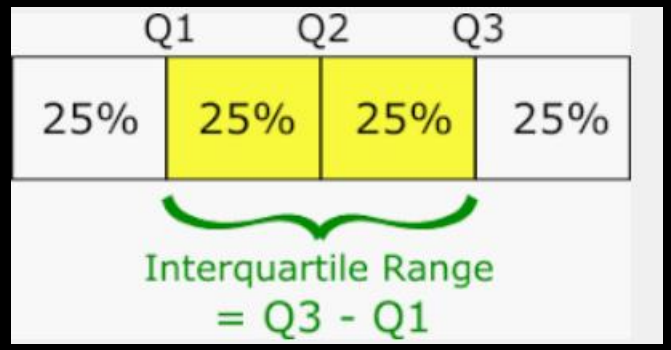
The **variance** of a population of n is $$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \mu^2.$$

The **standard deviation** of a population of n is $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

	age	bmi	charges
0	19	27.9	16884.924
1	18	33.77	1725.5523
2	28	33	449.462
3	33	22.705	21984.47061
4	32	28.88	3866.8552
5	31	25.74	3756.6216
6	46	33.44	8240.5896
7	37	27.74	7281.5056
8	37	29.83	6406.4107
9	60	25.84	289223.1369
variance	152.1	13.564	83016214.86
std	12.33	3.683	9111.3234

Quartile & Outlier

Values that divide your data into quarters provide data is sorted in an ascending order.



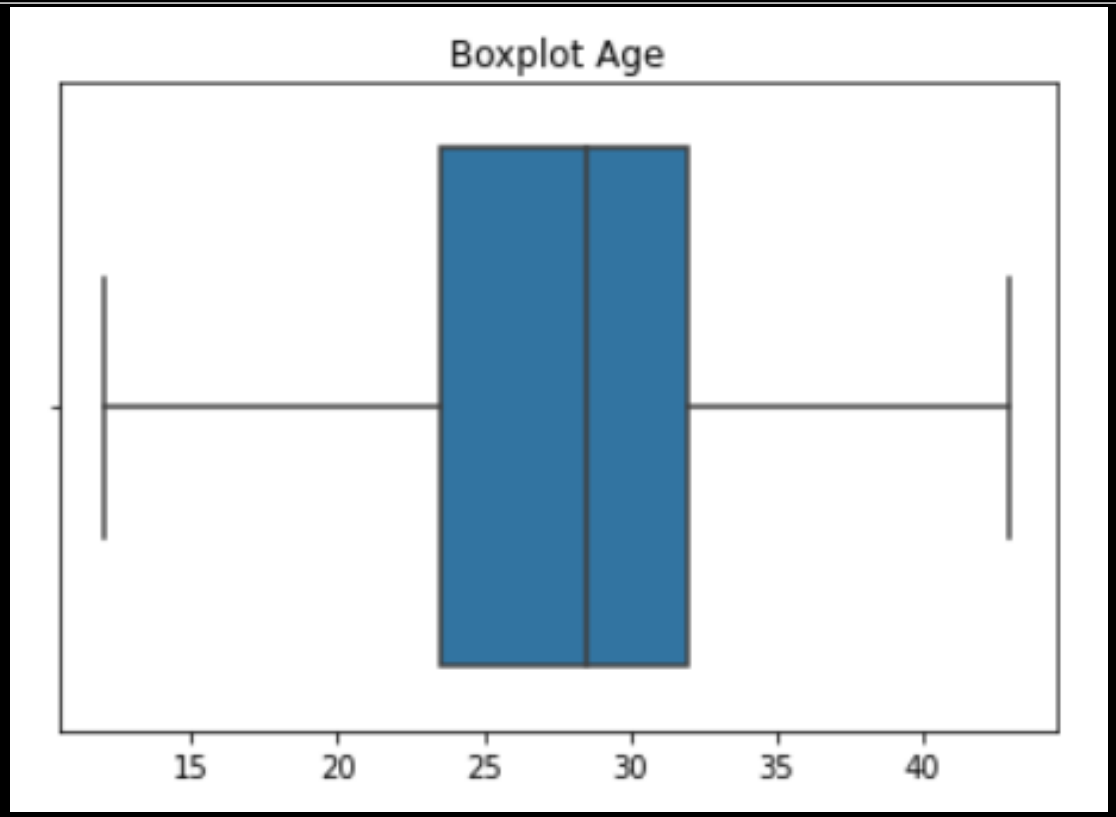
Lower outlier = Q1 – (1.5 x IQR)

Higher Outlier = Q3 + (1.5 x IQR)

Name	Age
Dini	12
Dono	32
Doni	43
Dani	32
Dadang	22
Didik	24
Dudun	26
Deni	31

```
Q1 :23.5
Q2 :28.5
Q3 :32.0
Lower outlier: 10.75
Higher Outlier: 44.75
```

Quartile & Outlier

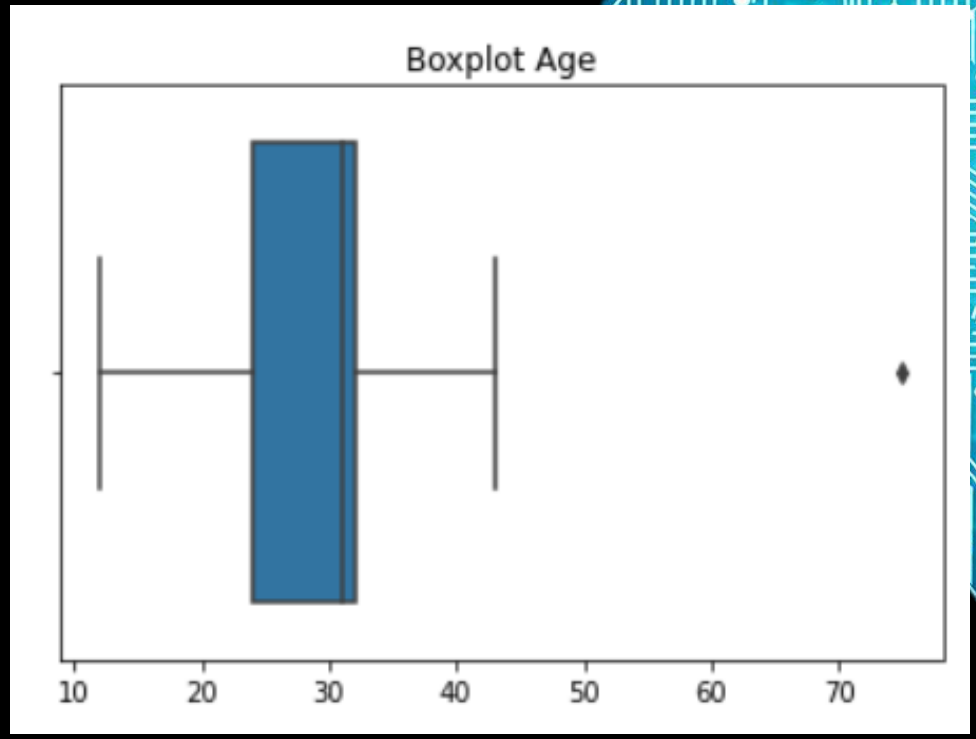


Name	Age
Dini	12
Dono	32
Doni	43
Dani	32
Dadang	22
Didik	24
Dudun	26
Deni	31

There is no Outlier

Quartile & Outlier

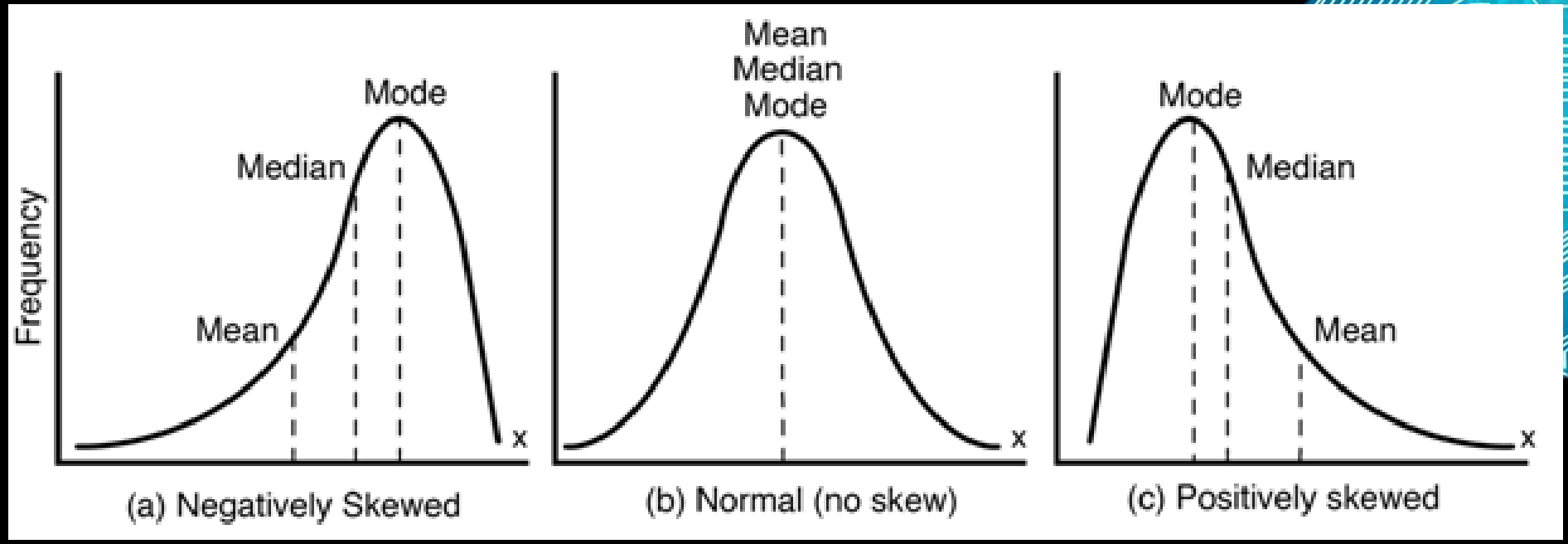
Name	Age
Dini	12
Dono	32
Doni	43
Dani	32
Dadang	22
Didik	24
Dudun	26
Deni	31
Bejo	75




```
Q1 :24.0
Q2 :31.0
Q3 :32.0
IQR: 8.0
Lower outlier: 12.0
Higher Outlier: 44.0
There is an outlier: 75
```

Skewness

Skewness is a measure of the asymmetry of the data

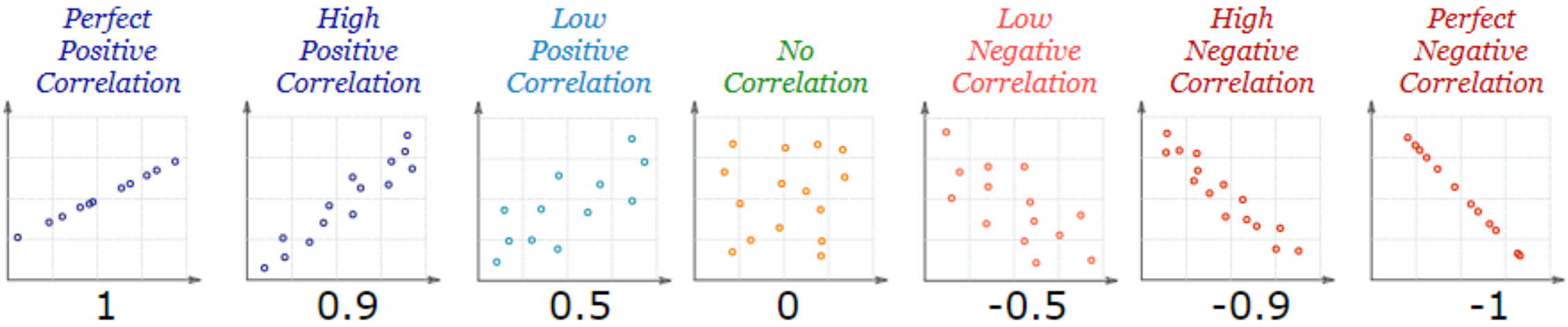


Correlation

Measure correlation between 2 variables or more than two variables.

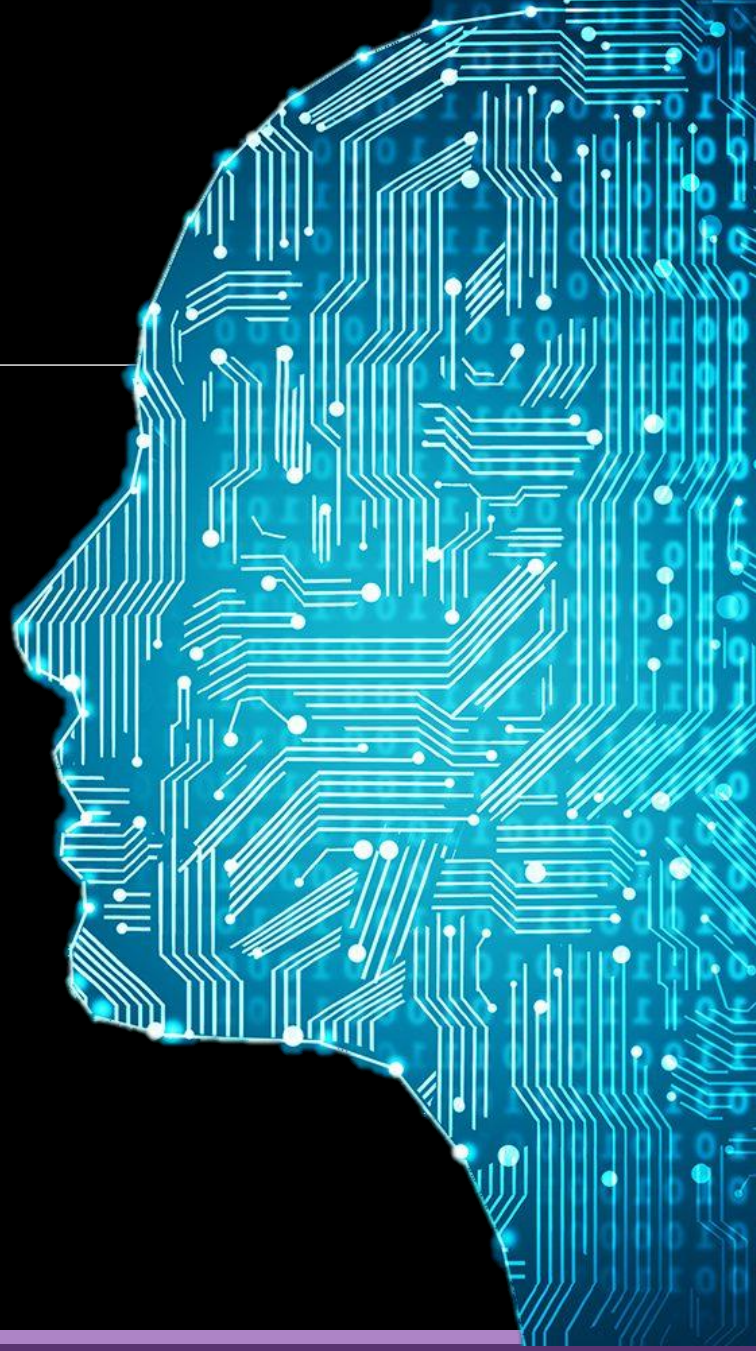
Correlation Coefficient Formula =

$$\frac{\Sigma[(X - X_m) * (Y - Y_m)]}{\sqrt{[\Sigma(X - X_m)^2 * \Sigma(Y - Y_m)^2]}}$$



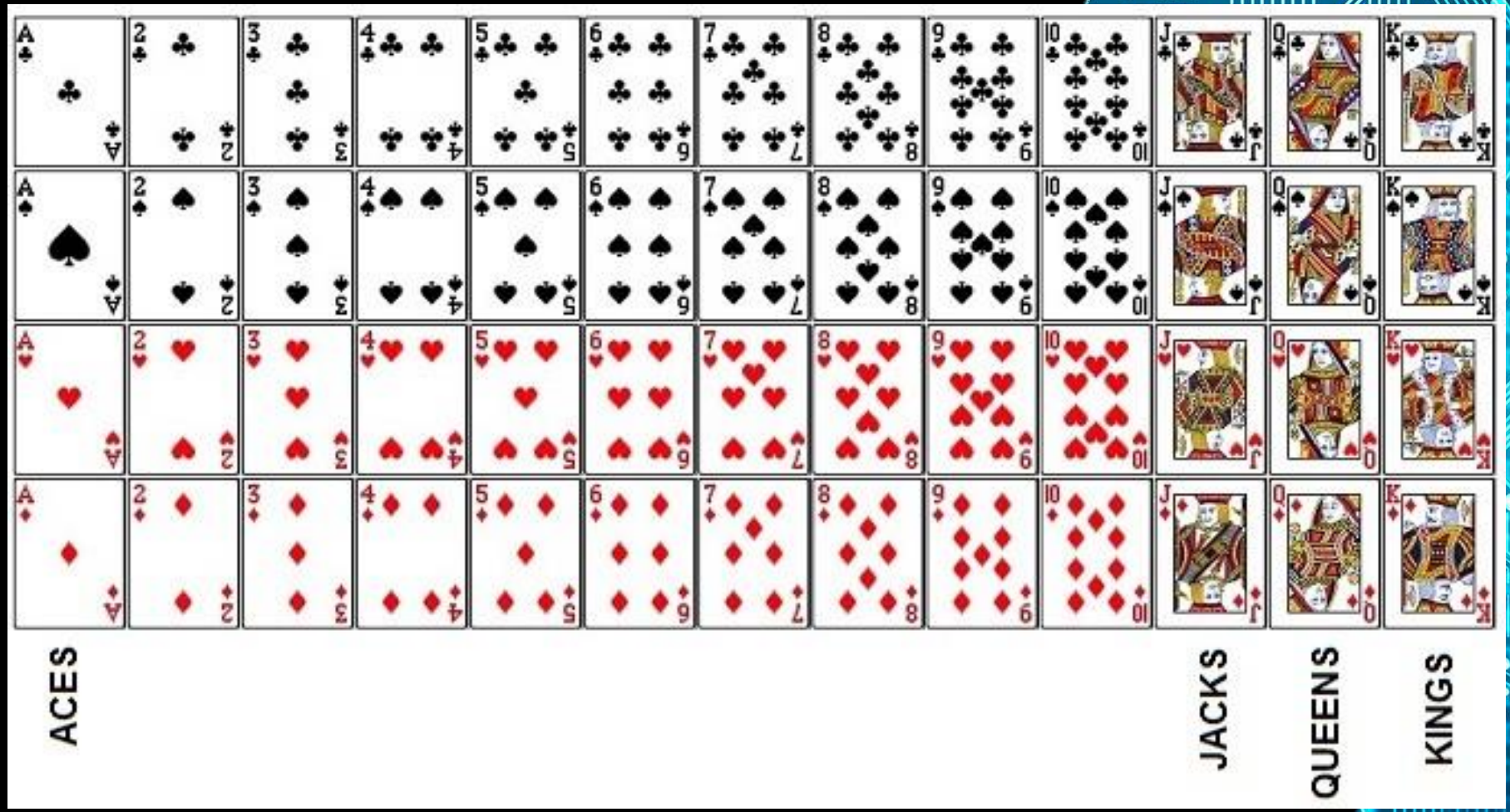
Probability

Probability is studies about randomness and finding a likelihood.



How's the probability of card?

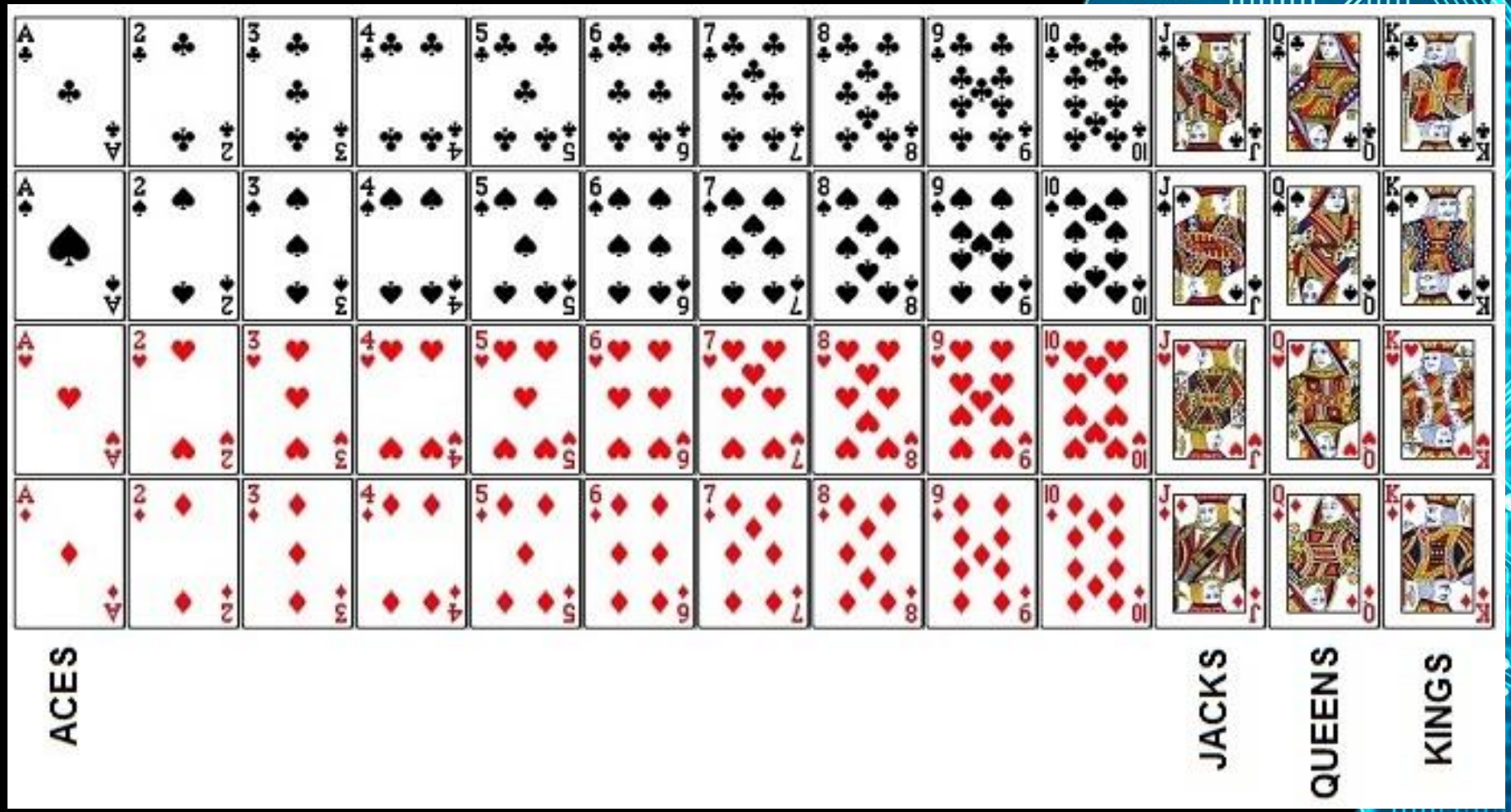
P(Queen) = 4/52
P(Black) = 26/52
P(5) = 4/52



Disjoint – Test

P(Queen or 5) =
4/52 + 4/52 = 8/52

P(Queen or Black) =
4/52 + 26/52-2/52 =
8/52



How's the probability of coin?

Two-tosses	(1000,Angklung)	(1000,1000)	(angklung,angklung)	(angklung,1000)
probability	0.25	0.25	0.25	0.25

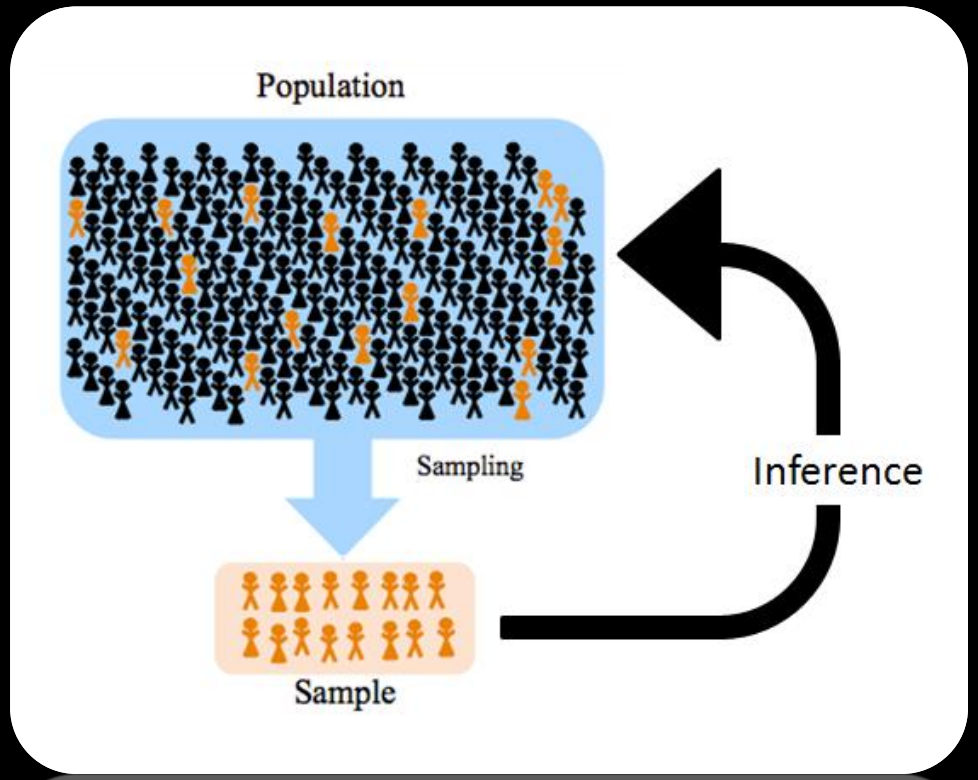
Inferential Statistics

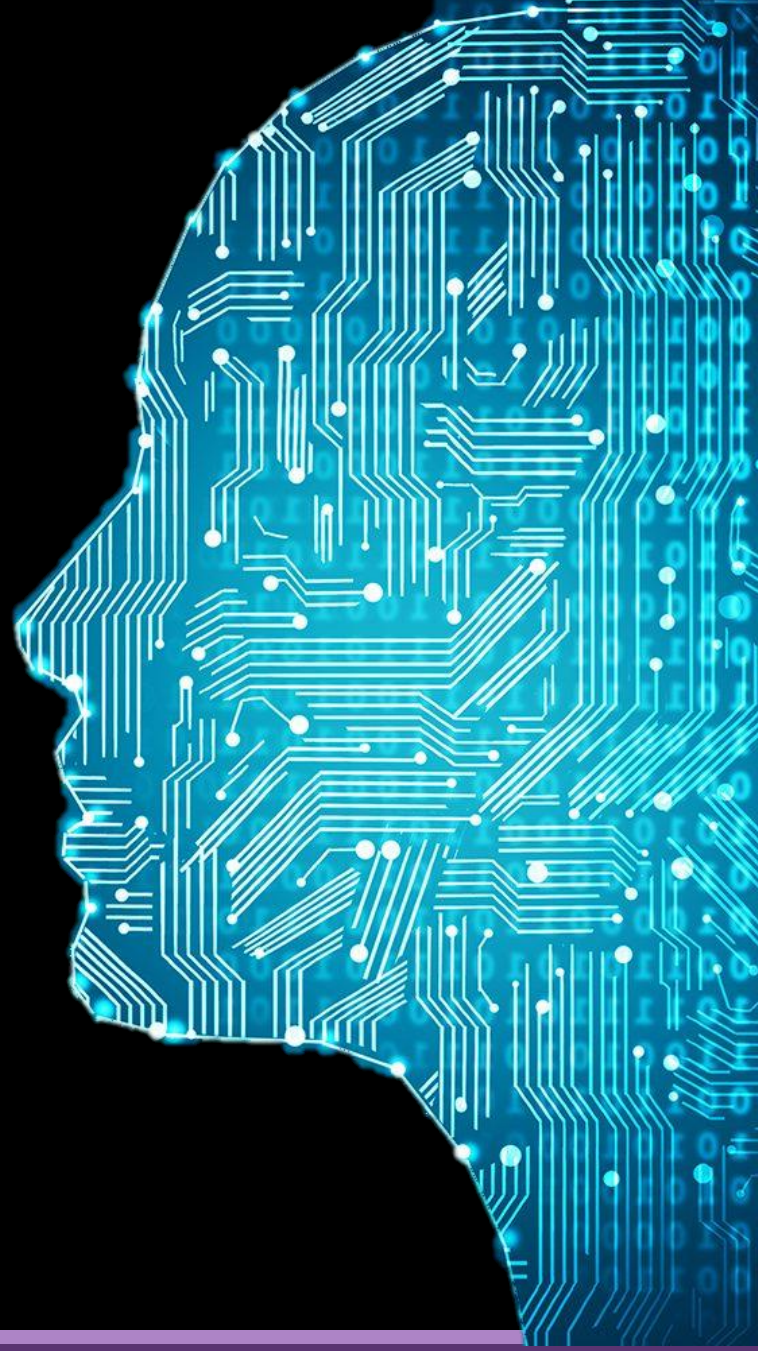
Kinds Of Inferential Statistics

- Inferential Statistics
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis

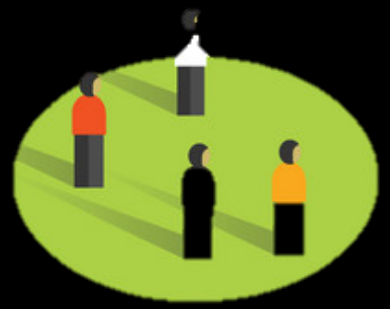
What is Inferential Statistics?

Inferential statistics is a Technique for inferring something About an entire population.

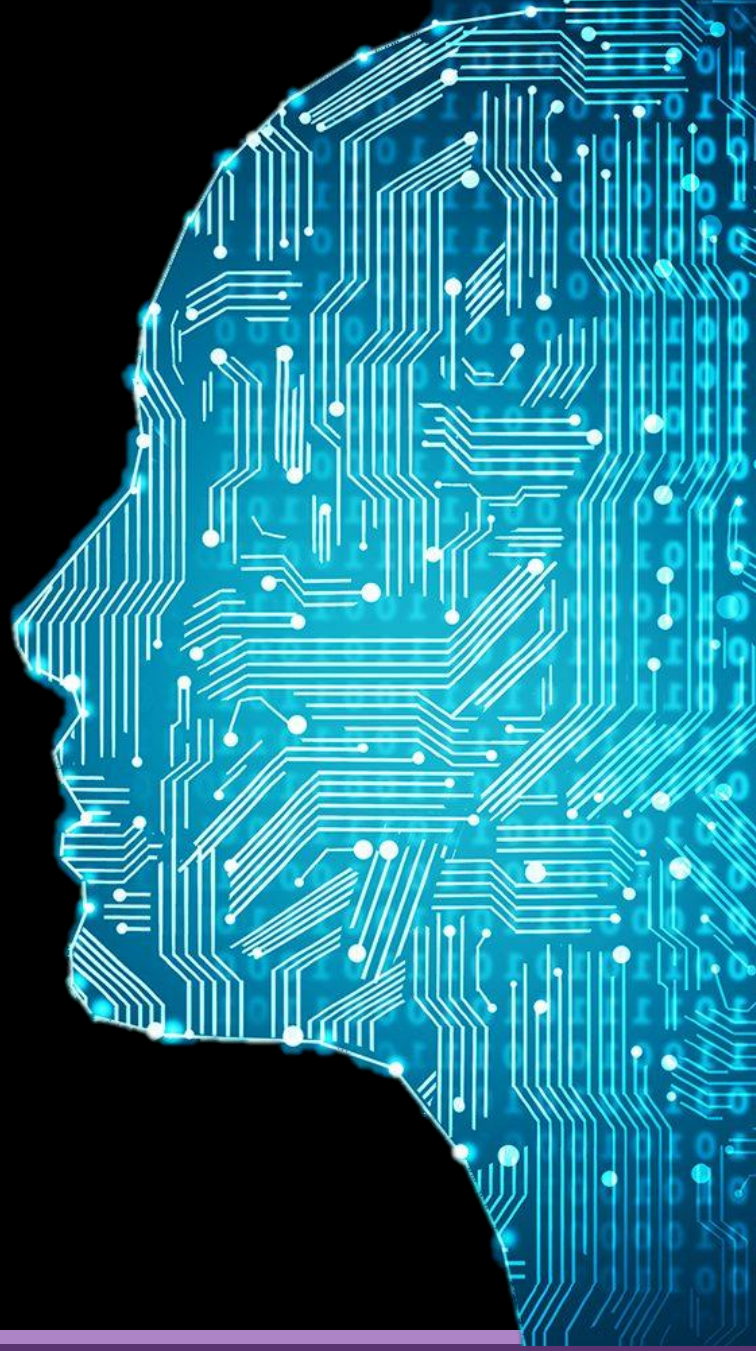


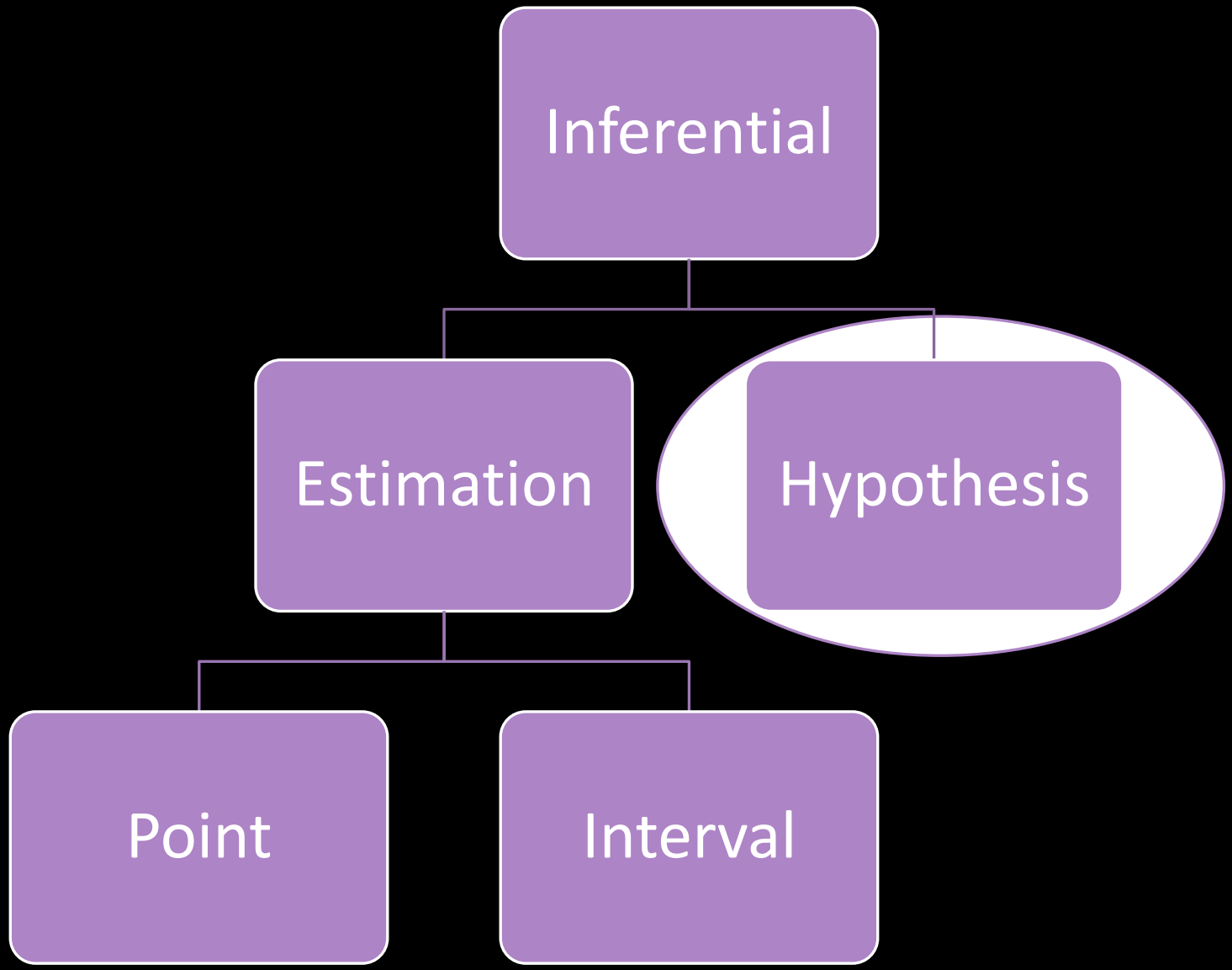


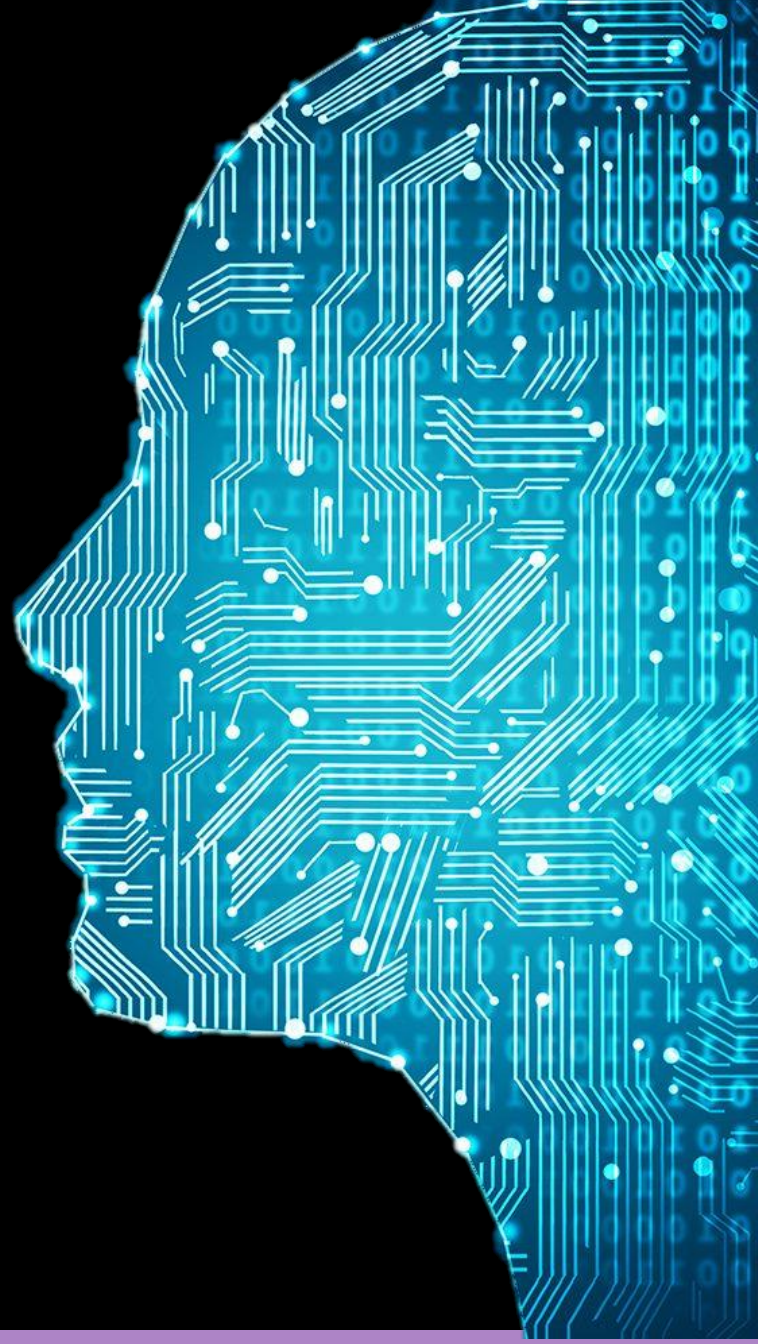
Population is the entire pool



Sample is a representative of a population.



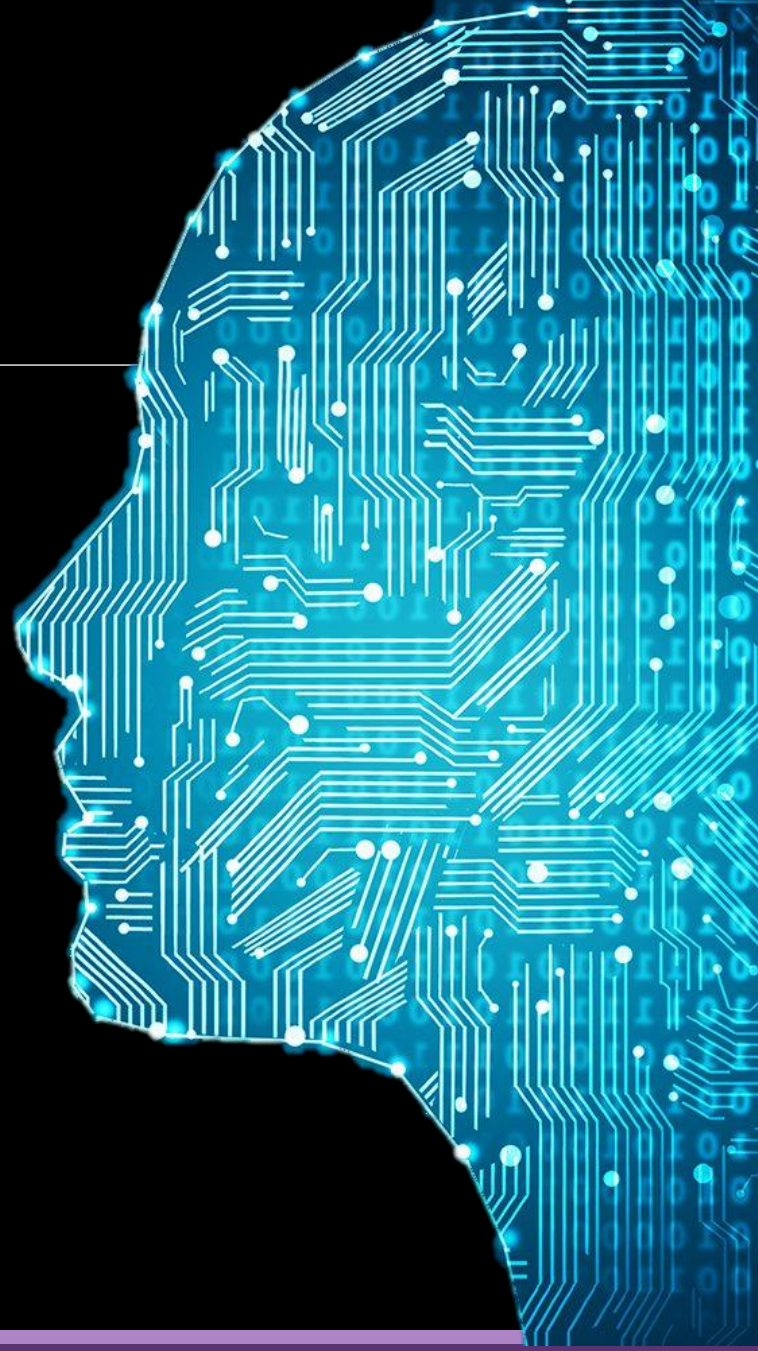

Inferential
Estimation
Hypothesis
Point
Interval

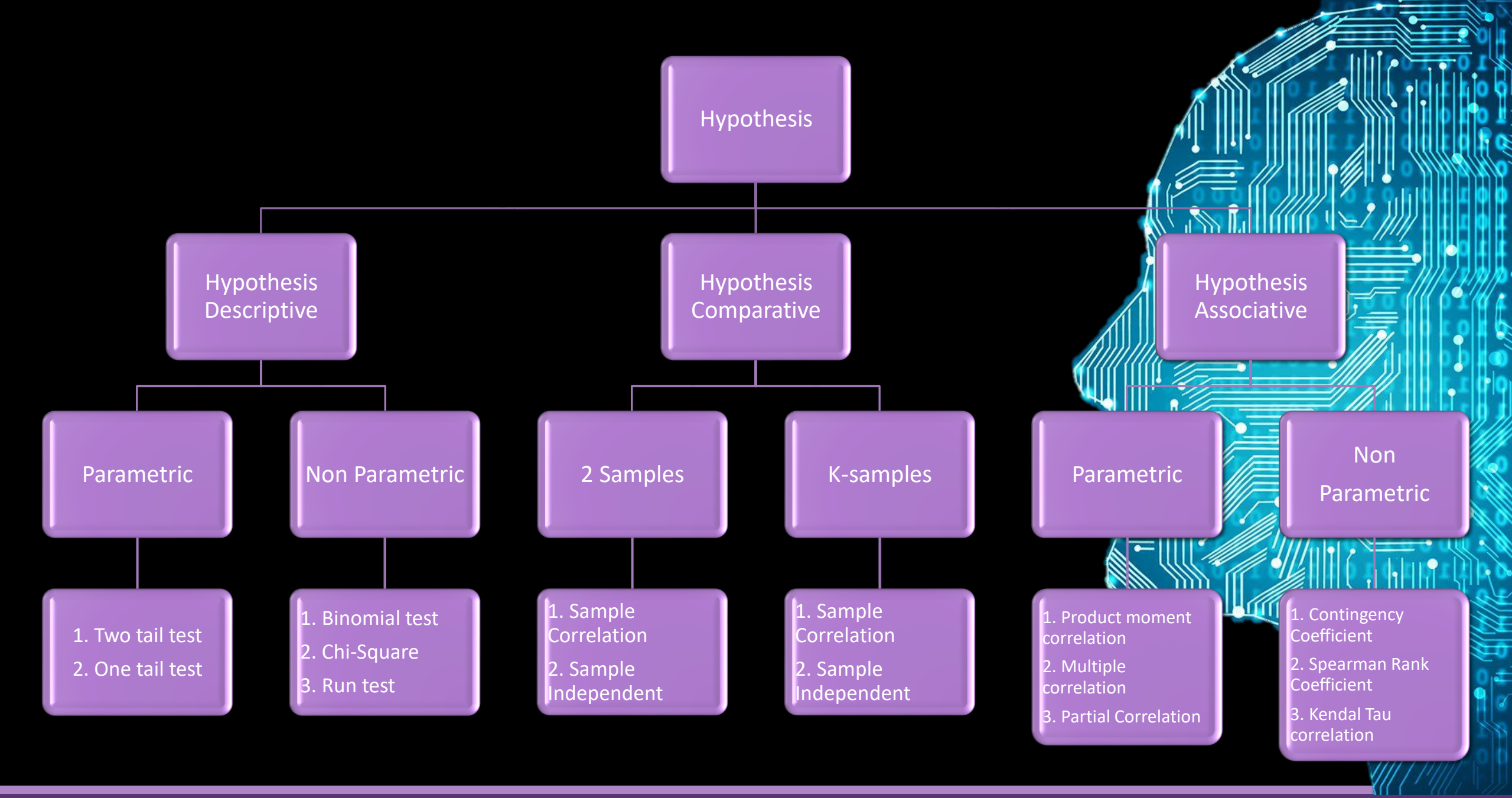


What is Hypothesis

is an estimation of

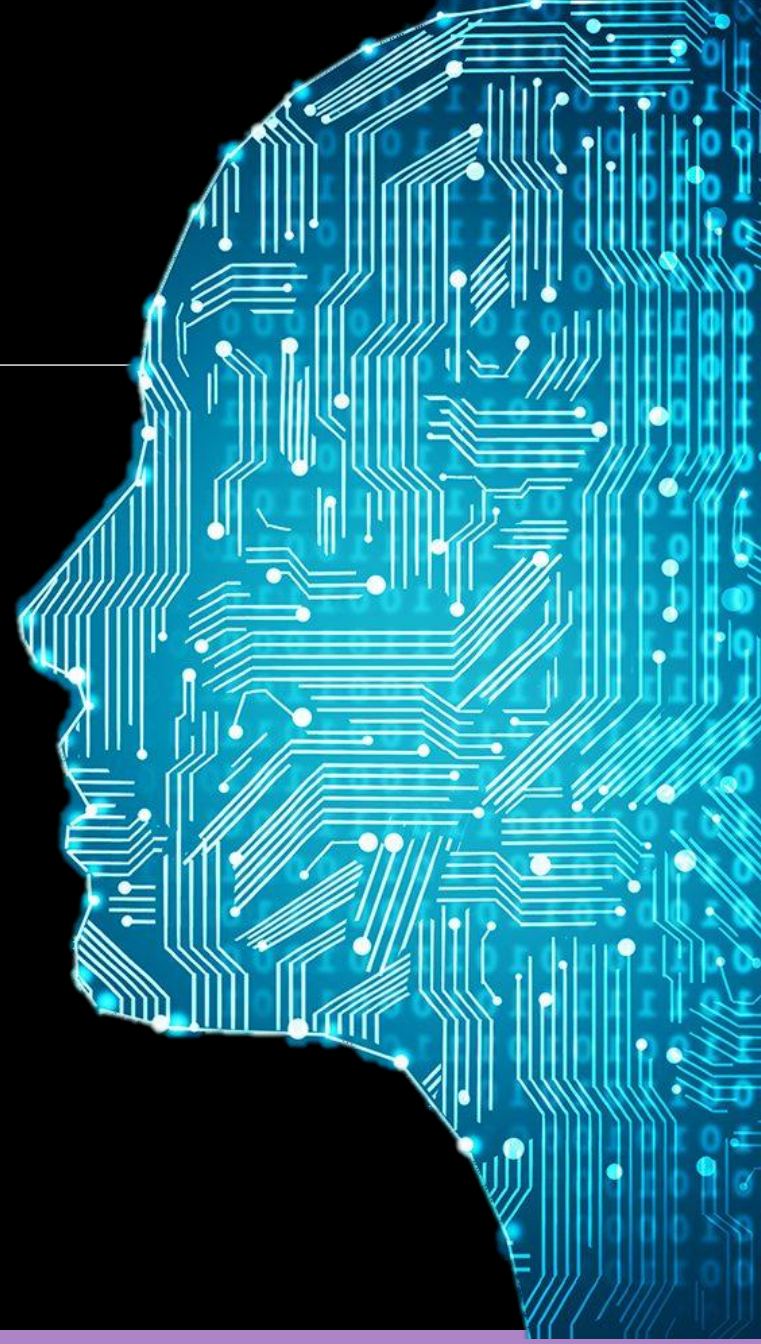
population parameters via samples data



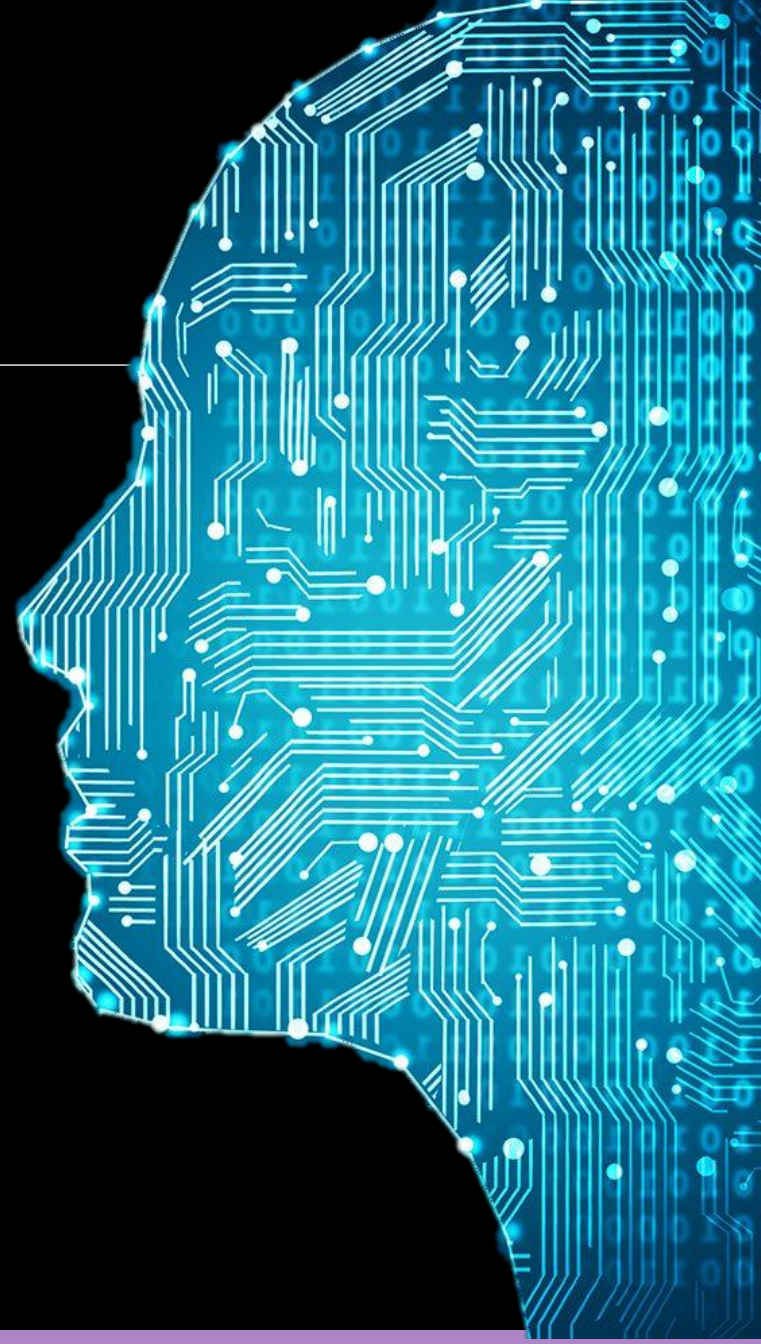

Hypothesis
Hypothesis Descriptive
Hypothesis Comparative
Hypothesis Associative
Parametric
Non Parametric
2 Samples
K-samples
Parametric
Non Parametric
1. Two tail test
2. One tail test
1. Binomial test
2. Chi-Square
3. Run test
1. Sample Correlation
2. Sample Independent
1. Sample Correlation
2. Sample Independent
1. Product moment correlation
2. Multiple correlation
3. Partial Correlation
1. Contingency Coefficient
2. Spearman Rank Coefficient
3. Kendal Tau correlation

Hypothesis Descriptive (1 sample) Technic

Data types	Technic
Nominal	Binomial Test
	Chi-Square (1 sample)
Ordinal	Run test
interval / Ratio	t-test (1 sample)



Hypothesis Comparative Technic

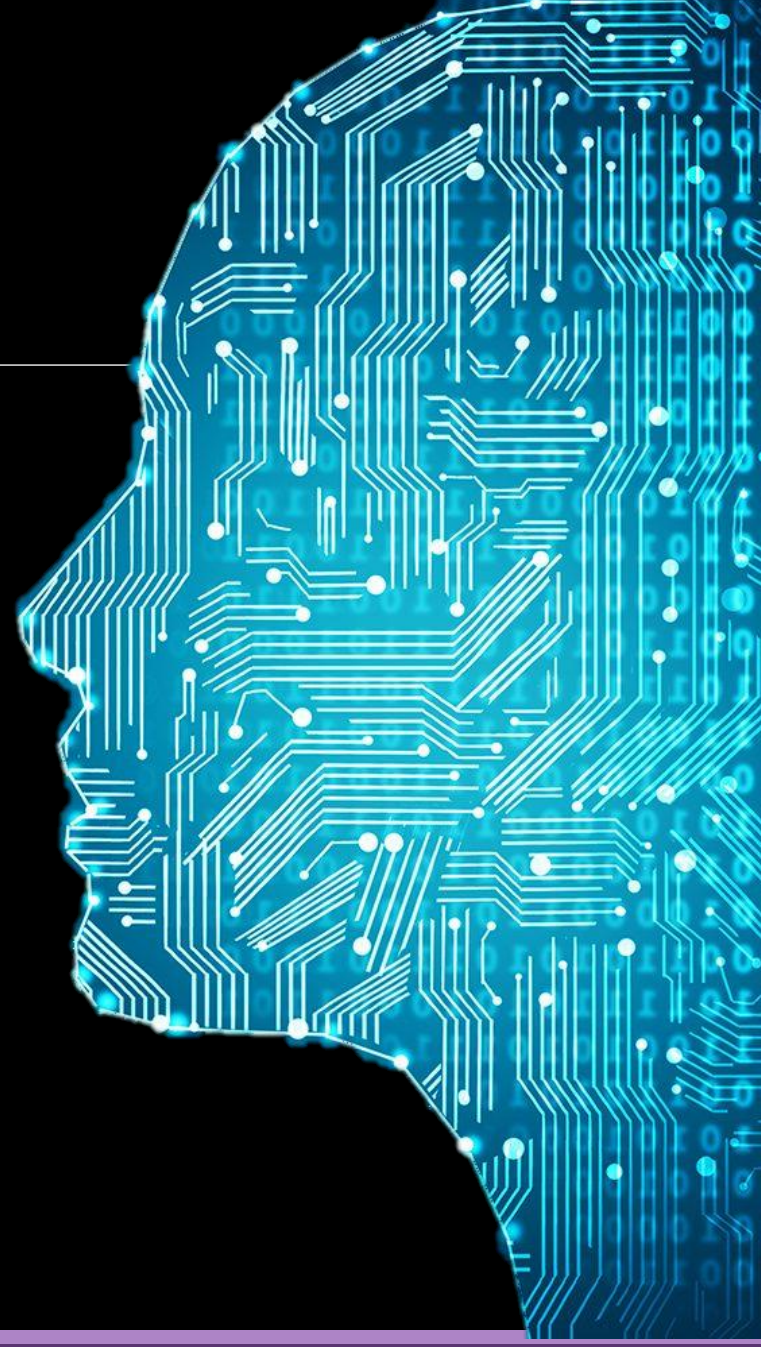


Data Types	Comparative Type			
	2 Samples		K-Sample	
	Correlation	Independent	Correlation	Independent
Interval/ Ration	t-test * two samples	t-test * two samples	One Way Anova*	One Way Anova*
			Two Way Anova	Two Way Anova
Nominal	Mc Nemar	Fisher Exact	Chi-Square k samples	Chi-Square k samples
		Chi-Square two samples	Cochran Q	
Ordinal	Sign test	Median Test	Friedman	Median Extension
		Mann- Whitney U test		
	Wilcoxon Matched Pairs	Kolomogorov Smirnov	Two Way Anova	Kruskal-Walls one Way Anova
		Wald_Wolfowitz		

*Parametric Statistics

Hypothesis Associative Technic

Data types	Technic
Nominal	Contingency Coefficient
Ordinal	Spearman Rank
	Kendals Tau
Interval and Ratio	Person product Moment
	Multiple Correlation
	Partial Correlation



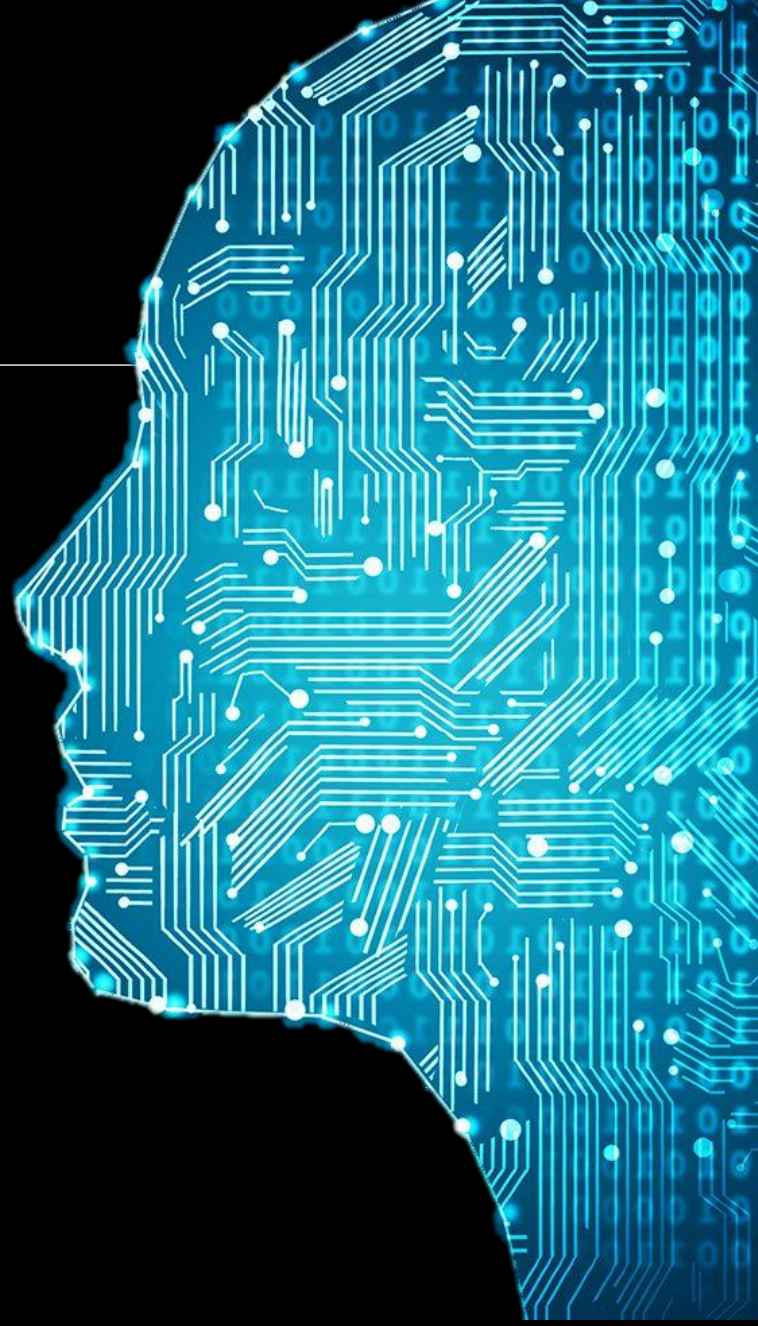
Hypothesis Testing with p-value

Use p-value to analyze how's significant of between two variables.

Hypothesis Null Ho	Hypothesis Alternative H_1
Assumed to be true, unless there is strong evidence to prove it.	Assumed to be true, if Ho rejected.

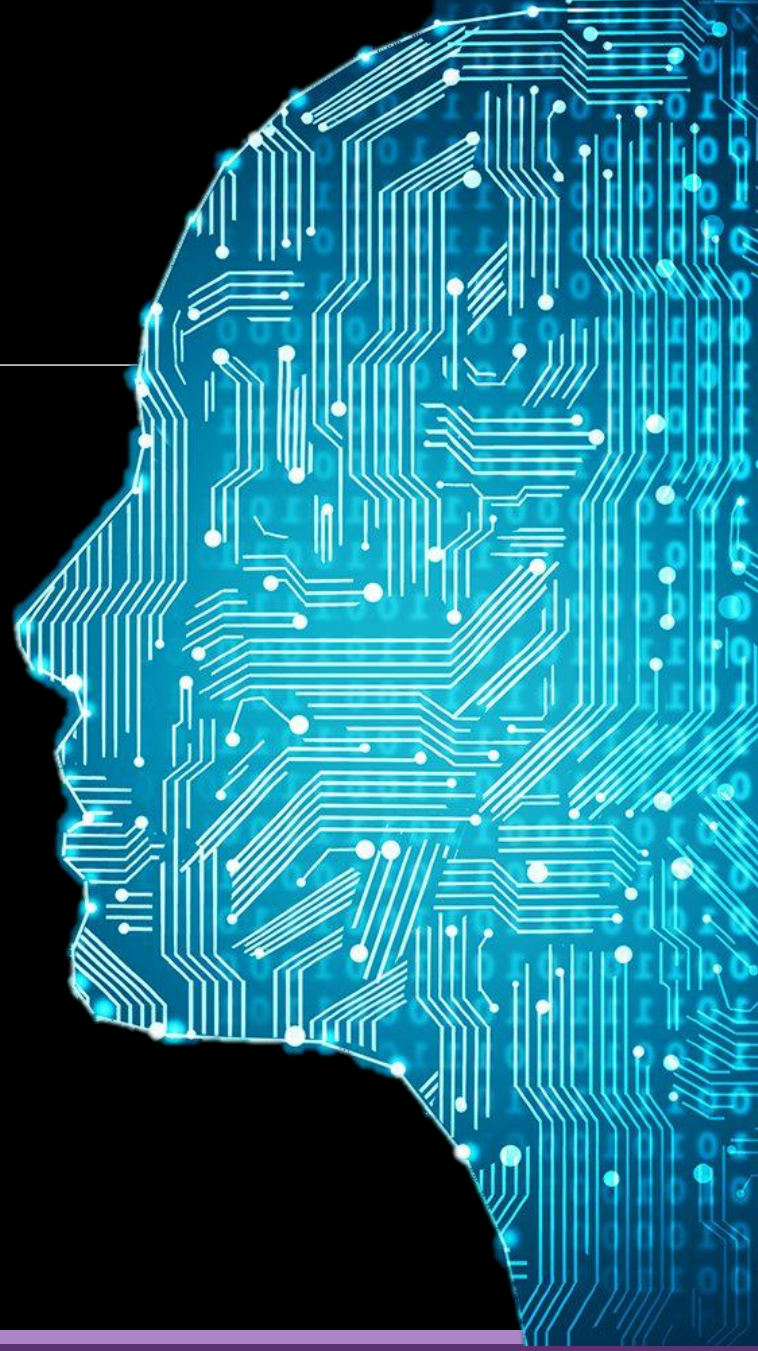
Standard significance

If p-value < 0.05, Ho rejected, H_1 accepted
If p-value > 0.05, Ho accepted, H_1 rejected



Resource

- https://dzone.com/articles/a-complete-guide-to-math-and-statistics-for-data-s#:~:text=Statistics%20is%20a%20Mathematical%20Science,trends%20and%20changes%20in%20Data.
- https://sixsigmastudyguide.com/kinds-of-statistics/
- https://www.xmind.net/m/GZ47/
- https://www.math-only-math.com/probability-of-tossing-two-coins.html
- https://www.math-only-math.com/playing-cards-probability.html#:~:text=Cards%20of%20hearts%20and%20diamonds,deck%20of%2052%20playing%20cards.
- https://dictionary.cambridge.org/dictionary/english/data
- https://www.dicoding.com/academies/177/tutorials/7956
- “Statistics untuk Penelitian” Books by Prof. Dr. Sugiyono



Thank You

Ronny Fahrudin | Data Science Fellowship at IYKRA