**Statistics & Machine Learning**

Statistics is the science of collecting, organizing, and analysing *data*.

Data is facts or pieces of information that can be measured. E.g.: The age of students in class, Height of students in class, etc.

**Types of Statistics**: -

**Descriptive Stats**: - It consists of organizing and summarizing data.

E.g.: We have 20 students in a Maths class. What are the average marks of students in the class?

**Inferential Stats: -** Technique where we used the data that we have measured to form conclusions.

E.g.: Are the marks of the students in Maths class similar to the marks of students in all the classes in the college?

**Population(N) and Sample(n):**

Let's take an example of the Election Exit Pole. Assume we have an election

in Maharashtra and a news channel to show the exit poll. In the image

below the bigger blue circle is the population(N) and the small multiple

circles are the sample(n) from the population. Not in these individual

samples, they find which candidate got the maximum number of votes,

and on the basis of that, we find which party is winning the election.

**Sample Techniques:**

1. **Simple Random Sampling**: In this, there is no set pattern, we can just randomly pick the objects in the sample. Every member of the population (N) has an equal chance of being selected for your sample(n).
2. **Stratified Sampling:** Where the population(N) is split into non-overlapping groups (Strata). Like Gender; Male and Female. Both categories can have different surveys
3. **Systematic Sampling**: In this, we have a specific Systematic pattern to select the sample. Like we from the population(N), we can select every nth object.
4. **Convenience Sampling**: The object expert in the field will be part of the sample(n).

**Variables**

**Type of Variables:**

Two kinds of variables are

1. **Quantitative variable**: Measured numerically, can perform operations like add, sub, multiply, divide, etc. E.g., Age, Weight, height.
   1. Discreet variable: Whole Number e.g., No of Children, No of Bank Account, etc
   2. Continuous Variable: Double (with decimal). E.g., Height {172.3,174.5,161.3, etc}, Weight {62.8, 71.5,82.8, etc}.
2. **Qualitative/categorial variable**: Based on some characteristics we can define some categories. E.g., Gender, Blood Group, etc.

**Variable Measurement Scales:**

1. Nominal data: Categorial data; e.g., Colour, Gender, etc.
2. Ordinal data: In this the order of the data matters but values don’t. E.g., Rank of student.
3. Interval data: Order matters, the value also matters, and natural zero is not present. E.g., measuring Temperature range in Fahrenheit, Distance Range.
4. Ratio Data

-----------------------------------------------------------------------------------------------------------------

**Measure of central Tendency:**

Used to determine the centre of the distribution of the population.

**Arithmetic Mean for Population(N) and Sample(n)**

The mean is similar to the Average. Population Mean is represented by: (μ = (Σ Xi) / N), whereas Sample Mean is represented by: (x̄ = (Σ xi) / n)

**Median** The central element after sorting the data.

**Mode** The most frequent element.

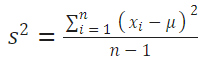
**Measure of dispersion**

Dispersion🡺 Spread

1. Variance:
   1. Population Variance

 Formula

* 1. Sample Variance



\*Degree of Freedom

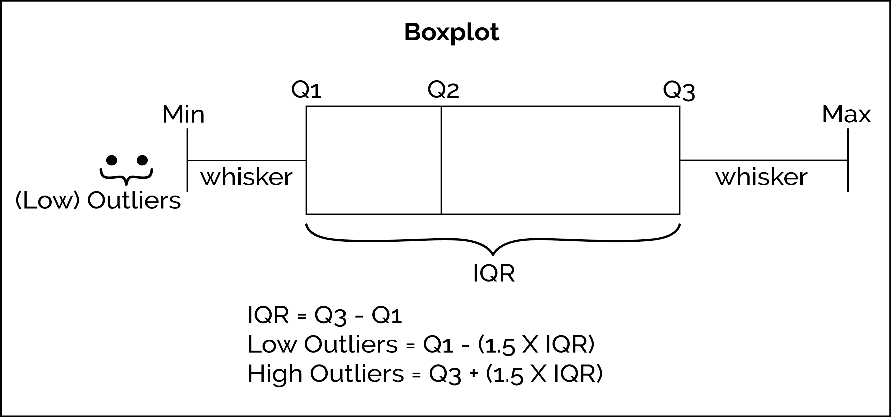
1. Standard deviation \sigma =\sqrt{\dfrac{\sum{(X - \mu)^2}}{N}}
2. Percentile: A percentile is a value below which a certain percentage of observation lies.
3. Five Number Summary:
   1. Minimum
   2. First Quartile(Q1) – 25 Percentile
   3. Medion

Removing the Outliers

* 1. Third Quartile(Q3) – 75 Percential
  2. Maximum
  3. Interquartile Range (IQR) = Q3-Q1
  4. Lower Fence/Higher Fence

Lower Fence = Q1-1.5(IQR)

Higher Fence = Q3+1.5(IQR)



**Distribution**

**Gaussian And Normal Distribution**: Whenever we have Gaussian/Normal distribution data, it will follow 66,95,99.7% rules.

To calculate, how much SD a value is away from the mean. We use Z-Score.

*z = ( x - μ ) / σ*

**Standard Normal Distribution**: where the Mean is 0, and the Standard distribution is 1. Mainly used for ML Algorithm.

**Standardization** is a process, where we convert a distribution into standard normal distribution where the mean is 0 and SD is 1.

**Normalization**: for this, we use MinMax Scaler.

**Probability**

Probability is a measure of the likelihood of an event to occur.

**Additional Rule (probability “or”):**

1. Mutual Exclusive Event: Two events can be mutually exclusive if they cannot occur at the same time.
2. Non-Mutual Exclusive Event: Multiple events can occur at the same time.

**Multiplication Rules**

1. Independent Events: Each and every event is independent.
2. Dependent Events:

-----------------------------------------------------------------------------------------------------------------

**P-Value (Probability Value)** Probability of the event/object occurring at every time of the activity. P=0.8 refers to every 100 times we do an activity then 80 times with the occurrence of the event.

**Hypothesis Testing and The Null Hypothesis: -**

An act in statistics whereby an analyst tests an assumption regarding a population parameter.

Hypothesis testing uses sample data to evaluate a hypothesis about a population. A hypothesis test assesses the unusual result, whether it is reasonable to accept the assumption or reject it.

Problem Statement: We need to test whether the coin is fair or not by performing 100 tests.

Step 1: Null Hypothesis (); Coin is fair

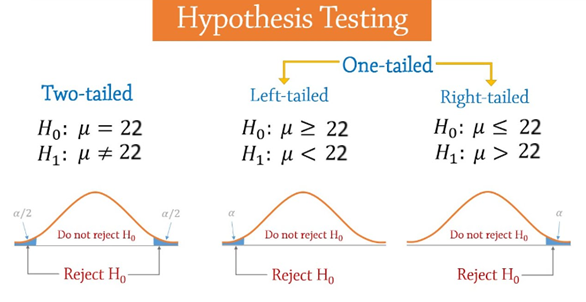
Step 2: Alternative Hypothesis (Ha or H1); Coin is unfair

Step 3: Experiment; Z-test, t-test, etc

Step 4: Reject or accept the null Hypothesis

**Significance Values (α)**

The significance value is compared to a predetermined cutoff (the significance level) to determine whether a test is statistically significant. If the significance value is less than the significance level (by default, 0.05), the test is judged to be statistically significant.



**Confidence Interval,**

**Z-test and T-test**:

If we have Population’s(N) Standard Deviation (SD) and the sample size is more than 30, then use the Z-test, else we use the t-test.

**Type I and Type II Error**

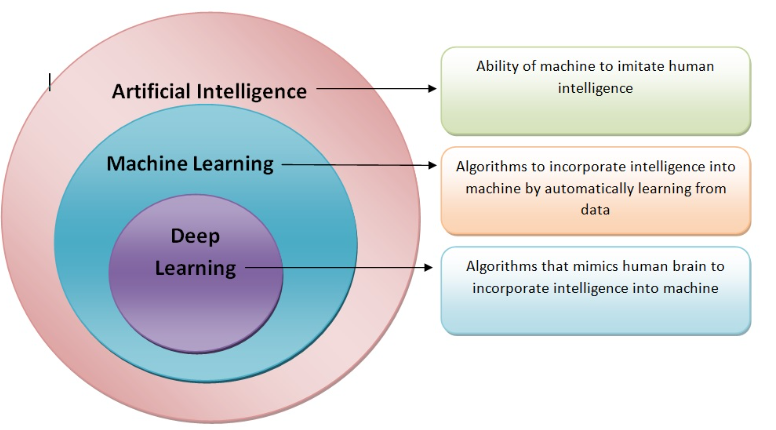
p-value = P (statistic | Null Hypothesis true)

if p-value < α 🡺 reject

if p-value≥ α 🡺 fail to reject

**Machine Learning:**

**AI vs ML vs DL vs DS**

****