Statistics.

* It is a branch of applied Mathematics that involves the collection, description, analysis and inference of conclusions from quantitative data.

Qunantitative data - Something that can be Measured in Numbers.

Types of Statistics.

1. Descriptive Statistics

* It helps to describe the data

For ex: charts, Bar or graph.

2. Inferential Statistics.

* It allows you to make predictions ("inferences") from that data.

* With inferential statistics you take data from samples & make generalizations about a population.

Statistics for Data Science

Basic Statistics.

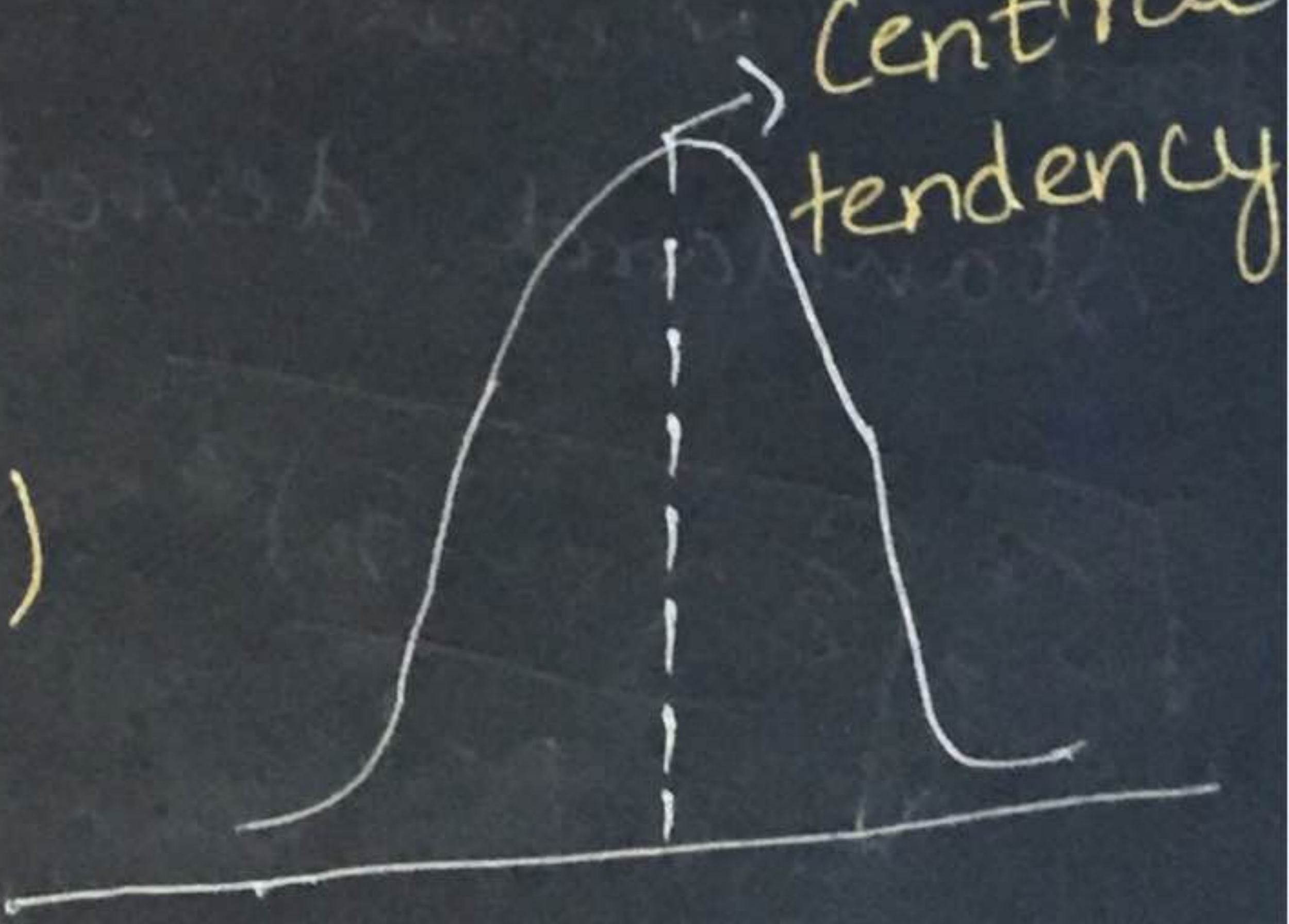
1. Central Tendency.

* It refers to the central position of the given data set

Mean - Average (Sum | total)

Median - Middle value (After Sorting)

Mode - Most repeated Value



2. Population

* It is the entire group that you want to draw conclusions group (total)

Sample

* It is the specific group from population (specific data). Sample size is always lesser than population.

Population Mean - Average of Population

Sample Mean - Averrage of Sample

3. Measure of Dispersion.

* It is used to describe variability in sample or population.

a. Range = Max val - Min val

* It is the spread of data from lowest to highest val. in the distribution.

b. Variance $= \sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$ → Population Variance

$S^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$ → Sample Variance

* It is Average of squarred distance from Mean.

c. Standard deviation $= \sigma = \sqrt{\text{Variance}}$

$\sqrt{S^2} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ $\sigma = \sqrt{\frac{\Sigma(X-\mu^2}{N}}$

* Average distance from Mean

4. Random Variable (features)

* It is something that stores some value in it ($x = 24$ & $x =$ "Hey")

Types:

i. Numerical

ii. Categorical

discrete R.V

Eg: no. of people in family

* It will be a whole num & can't be negative

* It is countable

Continuous R.V

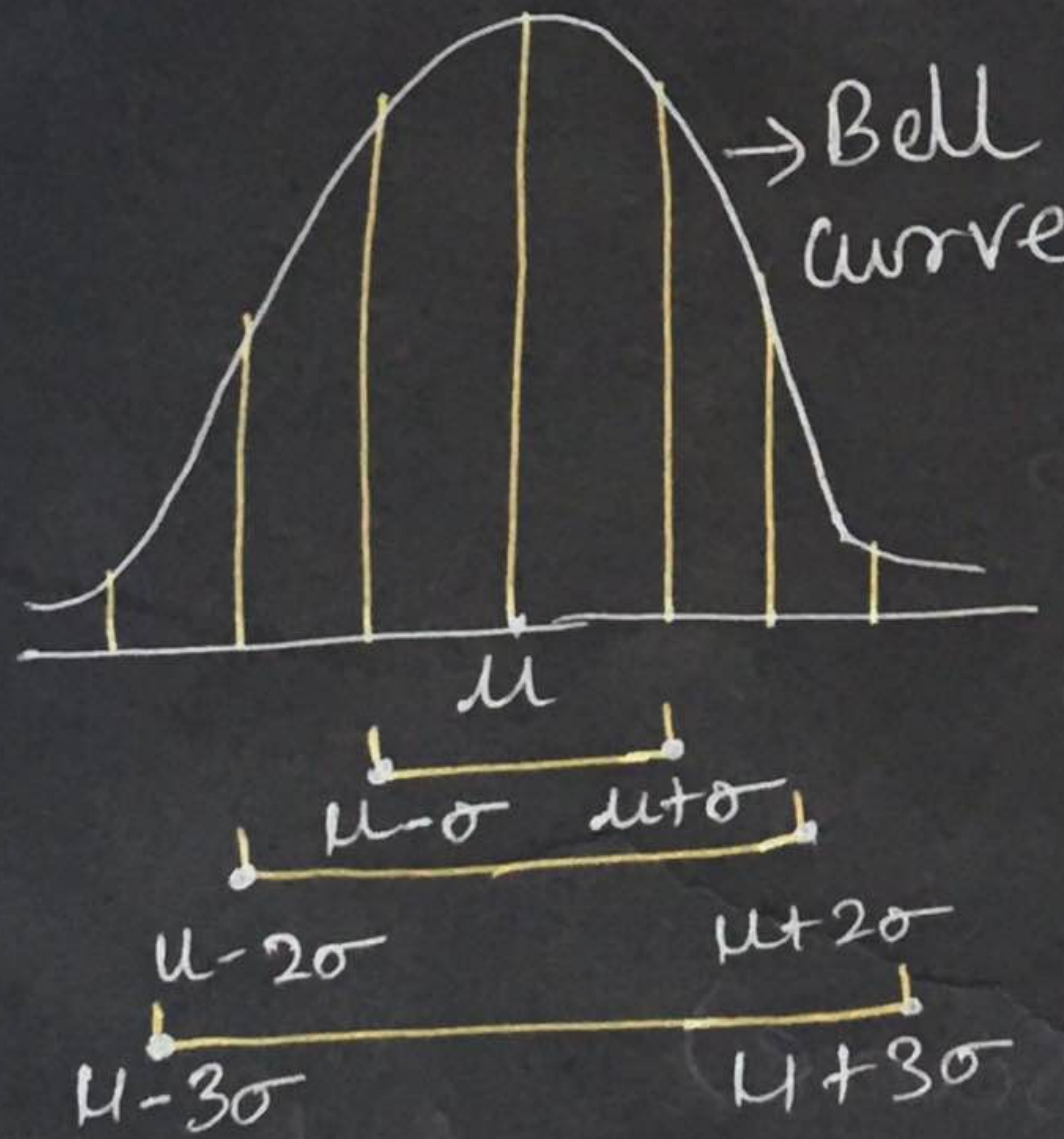
Eg: salary, age, loan

* It can be any number (decimal or integer)

* It is not countable

5. Normal | Gaussian Distribution

$X \approx G.D(\mu, \sigma)$



Emphevical Formula

1. $Pr[\mu-\sigma \leq X \leq \mu+\sigma] \approx$ 68% of data
2. $Pr[\mu-2\sigma \leq X \leq \mu+2\sigma] \approx$ 95% of data
3. $Pr[\mu-3\sigma \leq X \leq \mu+3\sigma] \approx$ 99.7% of data

* It is a probablity distribution that is symmetric about the mean, showing that the data near mean are more frequent in ccurrence than data from far the mean.

Probablity distribution

* It is a statistical function that describes all the possible values and likelihoods that a random Variable can take within a given Range

6. Percentage - Out of one hundred

Percentile - It is a measure in Stats. It shows the value below which a given percentage of observation falls.

Quantile - It is just a line that divide data into equally sized groups.