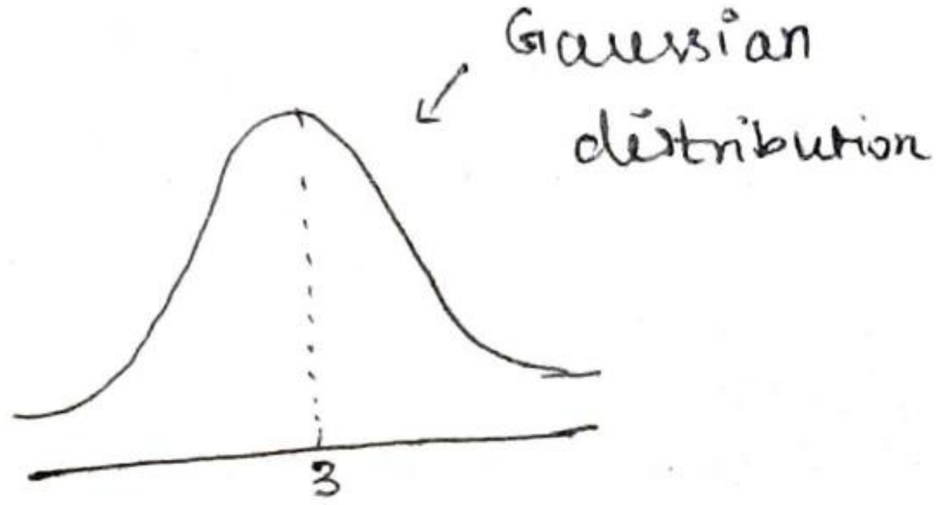
Statistics for data Science

Mean:-

i) sample = {1, 2, 3, 4, 5} ⇒ no outliers

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{15}{5} = 3$$

M = 3



ii) sample = {1, 2, 3, 4, 5, 50}

* Here in this sample we can find the value 50 as a outlier.

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{65}{6} = 10.4$$

M = 10.4

∴ Because of the presence of an outlier, the mean gets drastically changed.

* So in features like Age, if the value is missing in the feature mean is not used instead we can use median or mode.

Median:-

Sample = {1, 2, 3, 4, 5} no outliers

* Taking central values

If even no. of. records = $\frac{x_i + x_{i+1}}{2}$

If odd no of. records = $\frac{x_i}{2}$

where x_i is central values.

Here median is 3

Sample = {1, 2, 3, 4, 5, 50} with outlier

median = $\frac{3+4}{2}$ = 3.5

Here median is 3.5

So there is only slight difference blw with and without outliers

& so for features like Age, median is considered or mode

mode:-

sample = {1, 2, 3, 3, 4, 5} no outliers

mode = 3

sample = {1, 2, 3, 3, 4, 5, 50} with outliers

mode = 3

∴ so no change in both cases

∴ so it will be considered for columns like

Age.

Population and sample:-

∴ suppose if we want to calculate height of 1million people, it is impossible to collect details of 1million so that we can collect samples

1million ⇒ Population count (N)

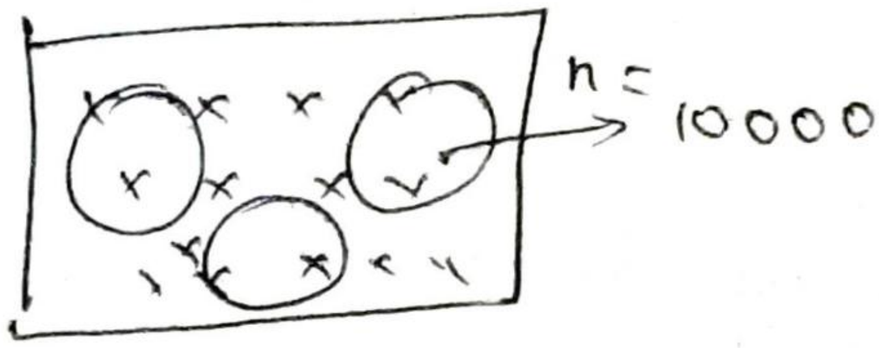
n ⇒ Sample count (n)

$$\text{Population mean} = \mu = \frac{1}{1\text{million}} * \sum_{i=1}^{1\text{million}} (x_i)$$

$$\text{Sample mean} = \bar{x} = \sum_{i=1}^{10000} x_i * \frac{1}{10000}$$

Re

If we want results from exit poll it is impo
ask 1million people so we will choose some
random sample and with that answer we
will predict who will win.



The dataset which we come across are the observed data which is a sample.