# Summarizing Data through Numbers.

* Measure of Central Tendency — Mean of the distribution
* Dispersion → Variance
* Skewness & kurtosis.
* Minimum Value
~~Measure~~
* Maximum Value.

↓

Mean is at exactly 50% area in the histogram.

# Measure of Central tendency.

* Mean
* Median
* Mode.

Ex $X = [3, 4, 3, 1, 2, 3, 9, 5, 6, 7, 4, 8]$

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = 4.583$$

## Median

arrange the data in accending order.

Ex

$1, 2, 3, 3, 3, 4, 4, 5, 6, 7, 8, 9.$

here $n = \cancel{4}\,12$ Median $= 4.$

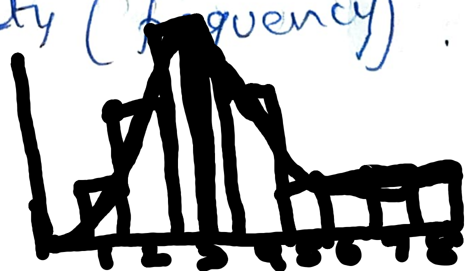
hence Median is mean of the central pair

$$\text{Median} = \frac{x_{n/2} + x_{n/2+1}}{2} \quad \text{if } n \text{ is even}$$

$$= \cancel{x_{n/2}}\; x_{\lceil n/2 \rceil} \quad \text{if } n \text{ is odd}$$

Mode : which particular outcome have highest probability (frequency).

Mode $[x] = 3$



Where do we want to use Mean, Median, Mode

* Choosing between mean & Median.

    * If the data have a Outlier and it is not contributing in the problem then Median is more robust. (error in the data). Ex Salary data/Income

    * If the outlier is the part of the story. Then mean is more appropriate as outlier is contributing in the inference (transient behavior of the system)

      ex ~~Level of water on the Dam~~ ~~reservoir~~.

Ex: Participate in Lottery/Gambling. hitting a Jackpot is rare.

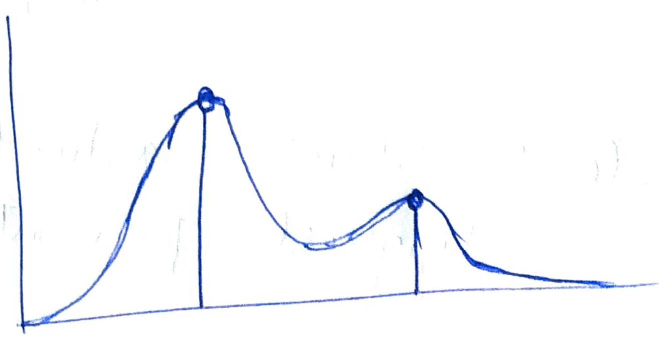
[redacted]

Mode

Multi Modal distribution: Level of water in the Reservoir (DAM)

decision for height & Strength of the dam (Reservoir)

Mean & Median will not help in taking the decision



**Skewness :** is a measure of the asymmetry of the probability distribution of a random variable about its mean.

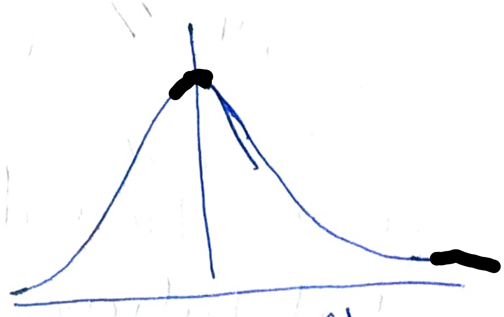
Skewness can have values positive, negative zero & undefined.

for unimodal distribution

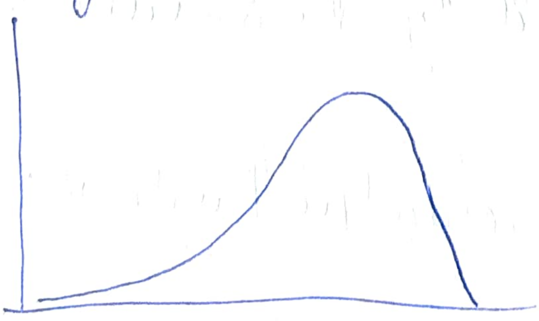
-ive Skewness means the tail is on the left side of the distribution.

+ive Skewness means the tail is on the ~~p~~ right side of the distribution.

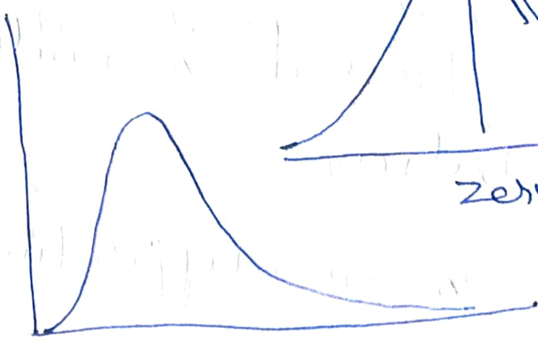
Zero Skewness means tails on the both side balance each other. it is also called Symmetric distribution.



Zero Skew.



-ive Skew



+ive Skew

Skewness of a random variable is defined as third Standardized Moment $\hat{\mu}_3$

$$\hat{\mu}_3 = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E[(X-\mu)^3]}{\sigma^3}$$

$\mu_3$ : **third Central Moment**.

Scaled version of the (Proportional)

Kurtosis: fourth central moment of the distribution.
(by Karl Pearson)

* Kurtosis of univariate Normal distribution is 3.

* Distribution with kurtosis less than 3 are said to be platykurtic, it means that the distribution produces very less extreme outliers than the normal.

* ~~Distribution with kurtosis great~~

Example of Platykurtic distribution is ~~Nor~~ Uniform distribution.

Distribution with kurtosis greater than 3 are said to be <u>leptokurtic</u> distribution

Ex Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian. & hence produce more outliers than the Normal distribution

$$\text{kurtosis}[X] = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

$$= \frac{E[(X-\mu)^4]}{\left(\underbrace{E[(X-\mu)^2]}_{\sigma^2 = \text{Var}}\right)^2} = \frac{\mu_4}{\sigma^4}$$

where $\mu_4$ is fourth central moment.

$\sigma \to 0$

$K[X] \to \infty$

# Relation between Skewness & kurtosis

Kurtosis is bounded below by Skewness

$$\frac{\mu_4}{\sigma^4} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1$$

Lower bound is realized by <u>Bernoulli distribution</u>

there is no upper limit to the kurtosis

delta $f^n$

$$\delta(x) = \begin{cases} \int \delta(x) = 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$