Probability and Statistics

Introduction to Statistics

(decision making Process)

def: Statistics is the science of collecting, organising and analysing data.

Data: "facts & pieces of information"

eg: Height of students in classroom.
IQ of students in classroom.

Types of Statistics

Used to predict something

Descriptive Statistics

It consist of organising and summary data.

(1) Measure of Central tendency.
   (i) Mean
   (ii) Median
   (iii) Mode

(2) Measure of dispersion
   (i) Variance
   (ii) Standard deviation

(3) Diff type of distribution of data.
   eg: Histogram
   Probability distribution function.
   Probability Mass function.

Inferential Statistics

It consist of using data you have measured to form conclusion.

from Sample data → Population data (concluding)

(i) Z-test
(ii) t-test
} Hypothesis Testing H_0, H_1
Pvalue
Significa value

(iii) Chi square test

* for creating sample, be careful
   ⇒ Sample size
   ⇒ Random (No Bias)
   ⇒ Enough representative.

eg: There are 20 Statistics Classes at your University and you have collected the heights of student in the class

Heights are recorded as

{175, 180, 140, 140, 125, 160, 135, 190}

Descriptive Question

"What is the ~~common~~ avg ~~age~~ height of the entire classroom"

Inferential Question → Sample data

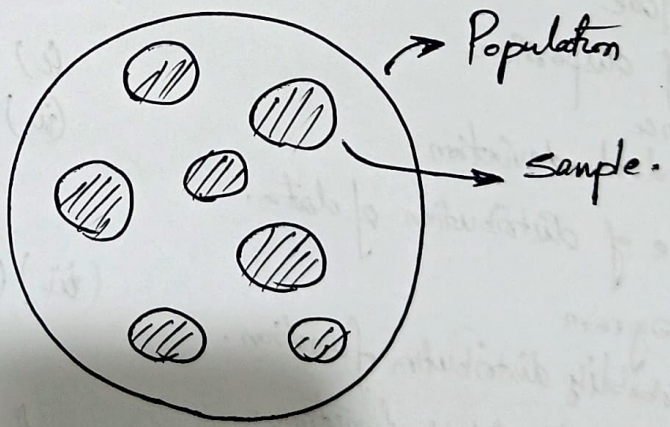
"Are the height of the students in the classroom similar to what you expect in the entire University?" → population data

Population data & Sample data.

Population : The group you are interested in studying.

Sample : A subset of population.

eg: Exit Poll



Types of data

- Data
  - Quantitative (Numerical)
    - Discrete
      - Whole No
      - always have fixed range.
      - eg. No of bank accounts
      - No of children in family
    - Continous
      - Any value
      - eg: Weight, Speed, Height.
  - Qualitative (Categorical)
    - Nominal
      - No Ranks
      - eg: Gender, Blood Group, Pin Code.
    - Ordinal
      - Ranks
      - eg: Customer Feedback
        - (i) Good
        - (ii) Better
        - (iii) Bad
        - (iv) Worst.

Scale of Measurement of Data

(1) Nominal Scale Data
(2) Ordinal Scale Data
(3) Interval Scale Data
(4) Ratio Scale Data

(1) Nominal Scale Data

(i) Qualitative
(ii) Categorical Data
(iii) Order/Rank doesnot matter

eg: Survey on favorite color.
(Here order doesnot matter on what color people like.)

(2) Ordinal Scale Data:

(i) Ranking is Important
(ii) Order Matters
(iii) Difference cannot be measured
↳ Just on the basis of ranks we cannot calculate difference → like in this case we donot know the reason for this rank

eg: An feedback form
(i) Best
(ii) Good
(iii) Bad
(iv) Worst

(3) Interval Scale of Data

(i) Order Matters
(ii) Diff can be Measured
(iii) Ratio cannot be measured
(iv) No true zero starting point.

eg: Temperature Variation in a place.

30F 60F 90F 120F

Diff: 60-30 = 30F

Ratio cannot be measured because it's just can't conclude that if at 30F we feel certain warmth then at 60F we will feel double the warmth there can be no of other factors affecting.

(4) Ratio Scale Data

(1) The Order matters
(2) Diff are measurable (including ratio)
(3) Common '0' starting point.

eg Grades of students

0, 90, 60, 30, 75, 40, 50

Measure of Central Tendency

(i) Mean or Average
(ii) Median
(iii) Mode

Mean

Population (N)

- Weighted Mean

	Score	wt
LR →	0.2	10
RF →	0.3	15
Xgboost →	0.5	12

$$\Rightarrow \frac{0.2\times10+0.3\times15+0.5\times12}{0.2+0.3+0.5}$$

- Trimmed Mean
  There is removal of certain value from bottom and top to remove the impact of outliers.

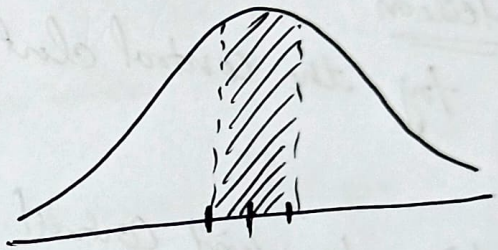
$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\text{Population Mean}(\mu) = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{[1+1+2+2+3+3+4+5+5+6]}{10}$$

$$= 3.2$$

Sample Mean (n)

$$\text{Sample Mean } (\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$



Median

$X = \{4, 5, 3, 2, 1\}$

Step
1. Sort the random variable X. : $\{1, 2, 3, 4, 5\}$
2. No of elements count : 5
3. if (count % 2 == 0)
   find central element
   if count = 6 $\{1, 2, \boxed{2, 3}, 4, 5\}$

$$\frac{2+3}{2} = 2.5 \text{ Median}$$

or if (count % 2 != 0)
find central element
2.5 → (3) Median

Why Median?

$x = \{1, 2, 3, 4, 5\}$

$\bar{x} = \frac{1+2+3+4+5}{5} = 3$

Shift

if now introduce another elent.

$x = \{1, 2, 3, 4, 5, 100\}$ → (outlier) → because it does not belong the distribution

$\bar{x}$ (mean) $= \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.1$

Now Median

for the central elent in (1) case → 3

(2) Case $= \frac{3+4}{2} = 3.5$

* Median to find central tendency when outliers present.

Mode

Frequency → Maxm frequency.

$\{2, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10\}$

Maxm presency of an elent : 1 so, Mode = 1

Application of Mean, Median, Mode in Feature Engineering

Age	Weight	Salary	Gender	Degree
24	70	40K	M	BE
25	80	70K	F	—
27	95	45K	M	—
24	—	80K	—	PhD
32	—	60K	—	Masters
—	60	—	—	Bsc
—	65	55K	M	—
40	72	—	F	—

Numerical Value → We can find mean and fill null values.
↳ If outlier → Median value.

Categorical Value: We can fill it with mode values.

Measure of Dispersion → (It describes the spread or variability of dataset.)
(Spread of the data)

(i) Variance → (proportional to spread of the data tells us how much the values are far from spread.)
(ii) Standard deviation.

Variance

Population Variance (σ^2)

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

x_i = Data points → x_i
μ = population mean → $\bar{x}$ = Sample mean
N = population size → n = sample size.

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

↓
Imp

Q Why we divide sample variance by $(n-1)$?

In order to create unbiased estimator } Bessel's corr^n

Q What is unbiased and biased estimator?

Unbiased: Like expected value is equal to the true value of the parameter

Biased: expected value is not equal to the true value due to bia

eg: If we pick 20 random students from a school and find avg

Unbiased } there is a good chance our estimator will give avg height true to the actual height

Biased } If we pick up student only from the basketball team where students are typically taller our avg value will not be equal to true value.

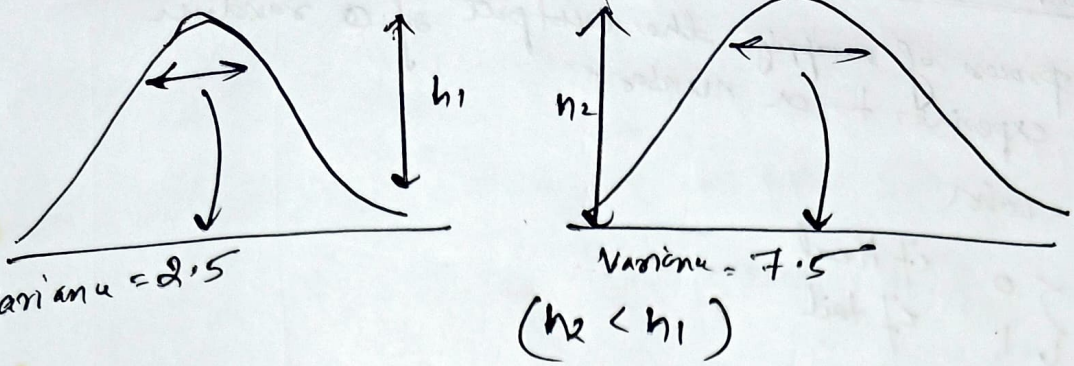
In Variance formula, we could have used mod, then square of it, but we don't because inference. conclusion is difficult for a population.

MAD (Mean Absolute Deviation)

$$MAD = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{n}$$

less prone to outliers like variance still can't use to find result for a population.

Dispersion or Spread (variance)



$(h_2 < h_1)$

Standard Deviation

Population S.D

$$\sigma = \sqrt{\text{variance}} \quad (\sigma^2)$$

Sample S.D

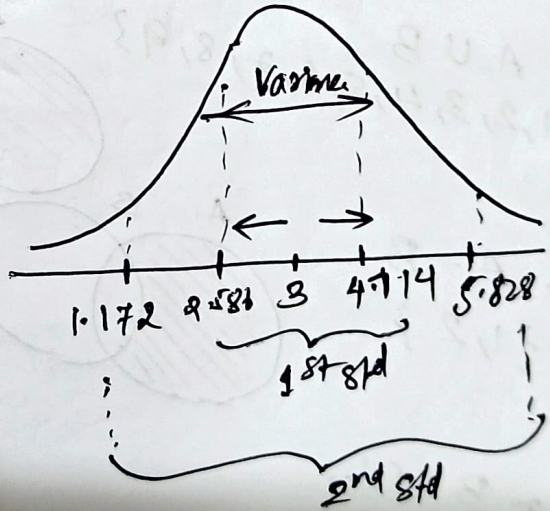
$$\text{Std} = \sqrt{S^2}$$

S^2 → Sample variance

eg: $x = \{1, 2, 3, 4, 5\}$

$\bar{x} = 3$

$\sigma = 1.414$



Coeff of Variation

$$CV = \frac{\text{Stand deviation}}{\text{mean}} \times 100$$

It is basically volatility in a dataset, a measure of variability b/w columns which do not have same units

(High CV, high variability
Low CV, low variability)

Random Variable

It is the process of mapping the output of a random process or experiment to a number.

eg:- Tossing a coin

$$x \begin{cases} 0 & \text{if Head} \\ 1 & \text{if tail} \end{cases}$$

(2) Rolling a dice

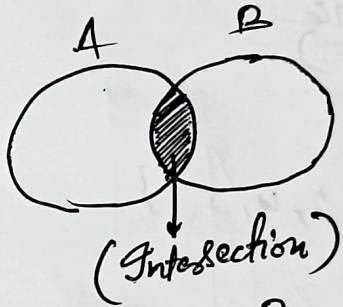
Sets

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7, 8, 9\}$

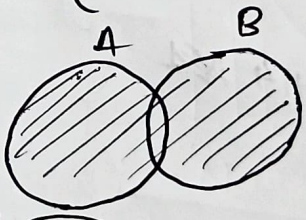
(1) Intersection: $A \cap B$

$\{3, 4, 5, 6, 7, 8\}$



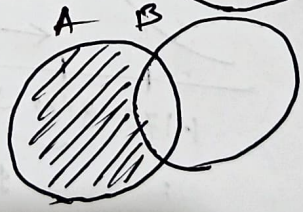
(2) Union: $A \cup B$

$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$



(3) difference: $A - B$

$\{1, 2\}$



(4) Subset:

$A \rightarrow B$ = False

$B \rightarrow A$ = False

Histogram & Skewness → (Frequency)

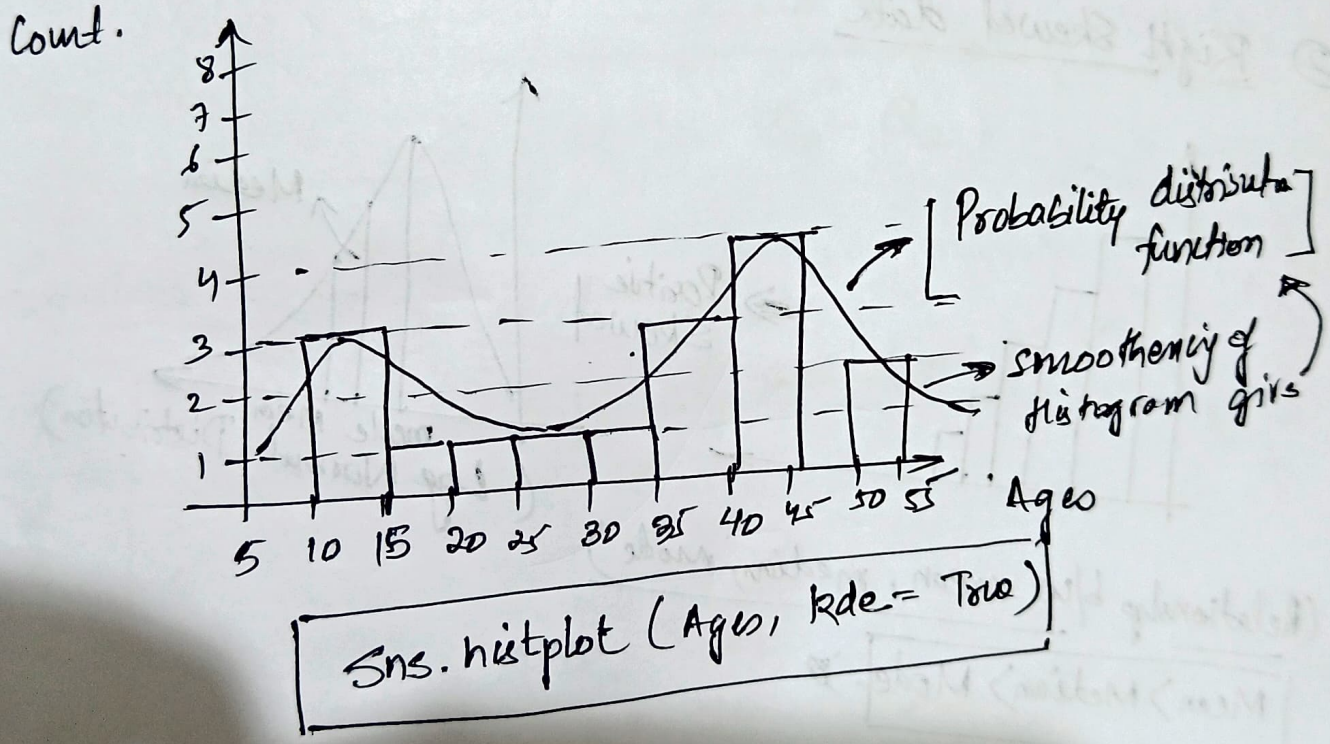
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Consider, we have 60 values and we want bin of 10 bins.

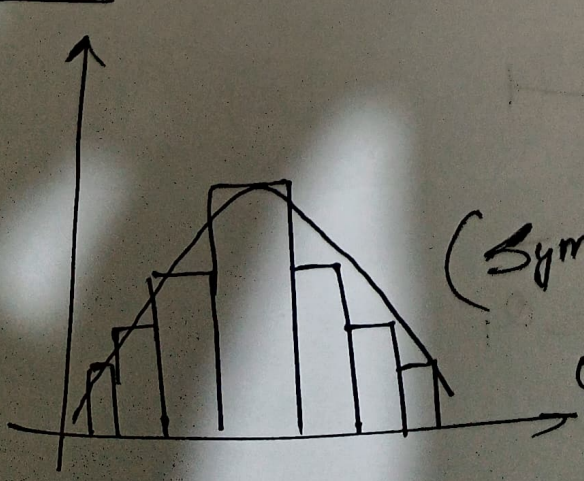
$\frac{60}{10} = \boxed{6}$ → bin size (No of bins = 10)

$\frac{60}{20} = \boxed{3}$ bin size (No of bins = 20)

Histogram



Symmetrical



(Symmetrical Distribution)

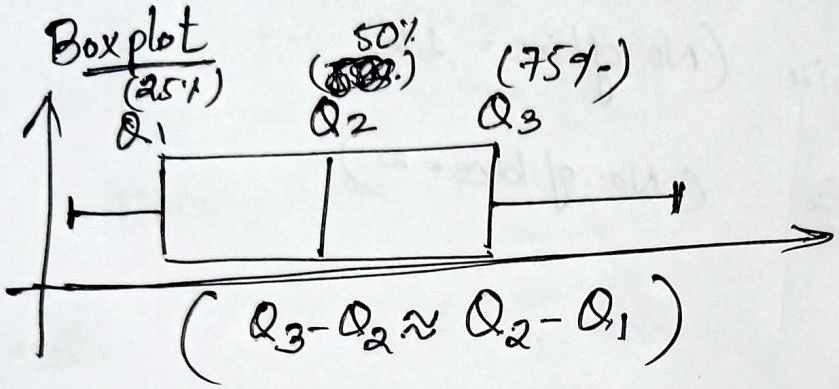
↓

(No skewness)

No skewness

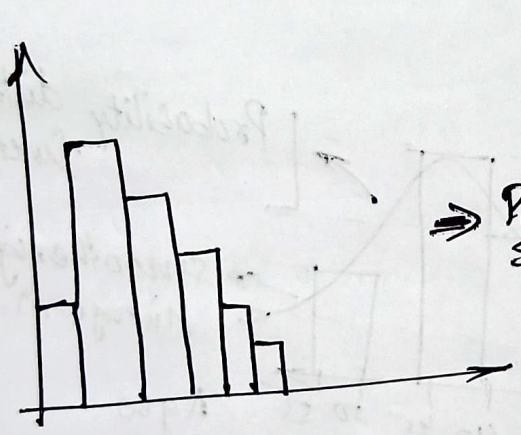
① The mean, median, mode all are at perfect centre
⇒ (Mean = Median = Mode)

Boxplot

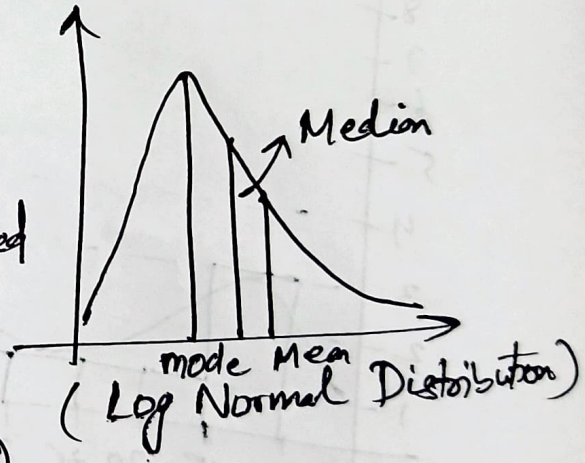


$(Q_3 - Q_2 \approx Q_2 - Q_1)$

② Right skewed data



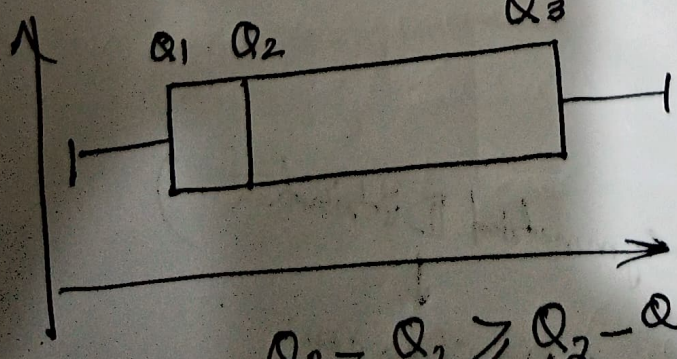
⇒ Positive Skewed



(Log Normal Distribution)

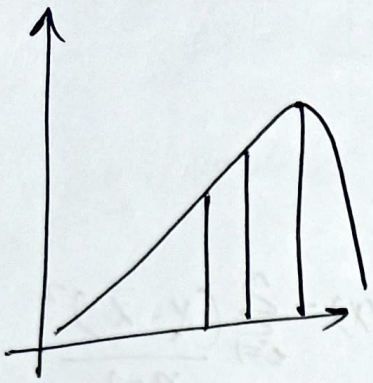
(Relationship b/w mean, median, mode)

Mean > Median > Mode



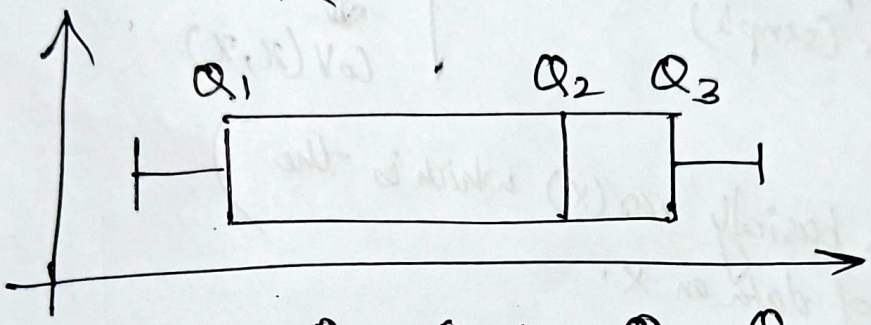
$Q_3 - Q_2 \geq Q_2 - Q_1$

Left skewed distribution



⇒ Negative skewed

Mean ≤ Median ≤ Mode



$Q_2 - Q_1 \geq Q_3 - Q_2$

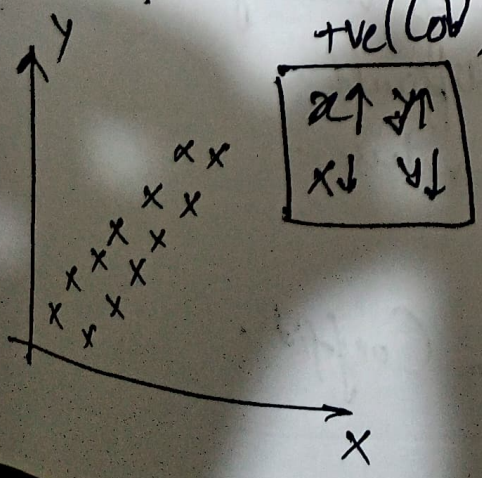
Covariance & Correlation

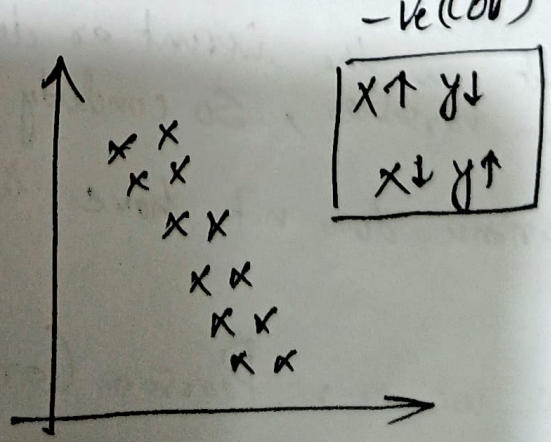
X	Y
2	3
4	5
6	7
8	9

[Relationship b/w X and Y]

can be

X↑ Y↑
X↓ Y↑
X↓ Y↓
X↑ Y↓





Covariance

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{(n-1)}$$

x_i = data points
$\bar{x}$ = mean (sample)
y_i = data points
$\bar{y}$ = mean (sample)

$$Var(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$\Downarrow$

$Cov(x,x)$

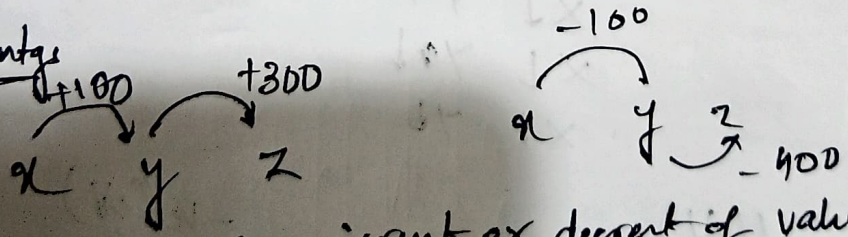
** (Cov(X, X) is basically var(x) which is the spread of data on x.)

Advantages of Covariance

(1) find out the relationship b/w x and y.

(2)

Disadvantage



* No limit on the increase or decrease of value of random variables, so corelation is difficult.

* Covariance does not have limit value.

To fix this we use; Pearson Correlation Coefficient.

Pearson Correlation Coeffi [-1 to 1]

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

σ_x : Std of x
σ_y : Std of y

(for linear relationships preffred) (or approx Normal distribution) → because of this CoV is limited

① The more the value towards +1, the more +vely correlated it is.

② The more the value toward -1, the more -vely correlated it is.

Spearman Rank Correlation [-1 to 1]

$$r_s = \frac{cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

R(x) = Rank of x
R(y) = Rank of y.

(preffred)

eg

x	y	R(x)	R(y)
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

for data Not
① Normally distbd
② has Outliers
③ Not strictly linear

Implementation (Choose imp features)

① Feature Selection

- Size of flat (+ve) ↑
- No of Rooms (+ve) ↑
- Location (+ve) ↑
- No of people stay ≈ 0 (No relation)
- Leakage (-ve)

→ Price ↑

(we can delete feature which do not affect the output)

✱ Correlation and Covariance relationship

Covariance : tells you how two random variables are connected to each other. but does not give you a measure of strength.

Correlation : makes the relationship a standardised relationship which is a simple scale from -1 to 1. It tells you the strength of a relationship.

Kendall Correlation Coeffit

It also tells the strength and direction, but it maps its value on basis of its agreement or disagreement.

C - Concordant : if most of all pairs agree with each other.

D - Discordant, if most of all the pairs dont agree with each other.

formula

$$T = \frac{C - D}{\frac{1}{2} n(n-1)}$$

If most pairs are in same order = 1
If most pairs are in opposite order = -1
If most pairs are in random = 0

* Use "Pearson" = if you care about the actual values and assume a straight line relationship.
* Use "Spearman" if you care about the ranks of the values
* Use "Kendall" if you want to measure the agreement or disagreement of the order of data points

Syntax Covariance

df.cov()

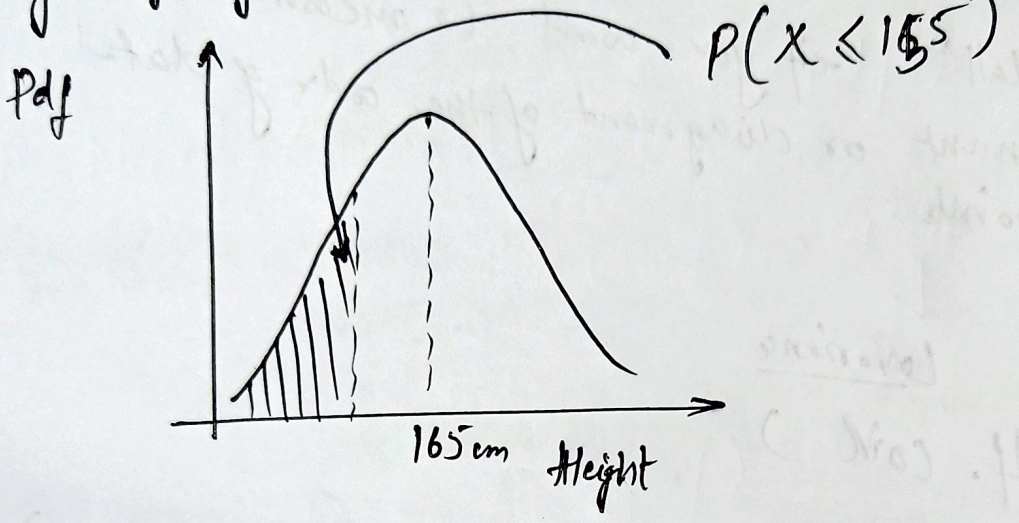
Corr

① df.corr(method="pearson", numeric_only=True)
② df.corr(method="spearman", numeric_only=True)
③ df.corr(method="kendall", numeric_only=True)

Probability Distribution function

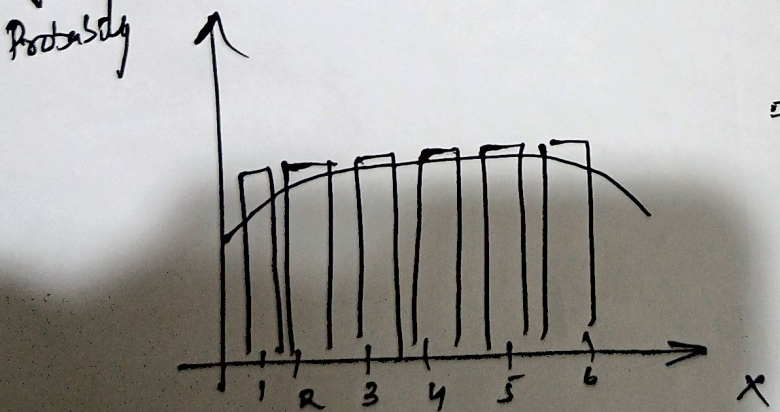
(1) Continous Random Variable — Probabilty Density function

Eg: Height of student in Classroom.



(2) Probability Mass function for discrete Random Variable

Eg: Rolling a Dice {1, 2, 3, 4, 5, 6}

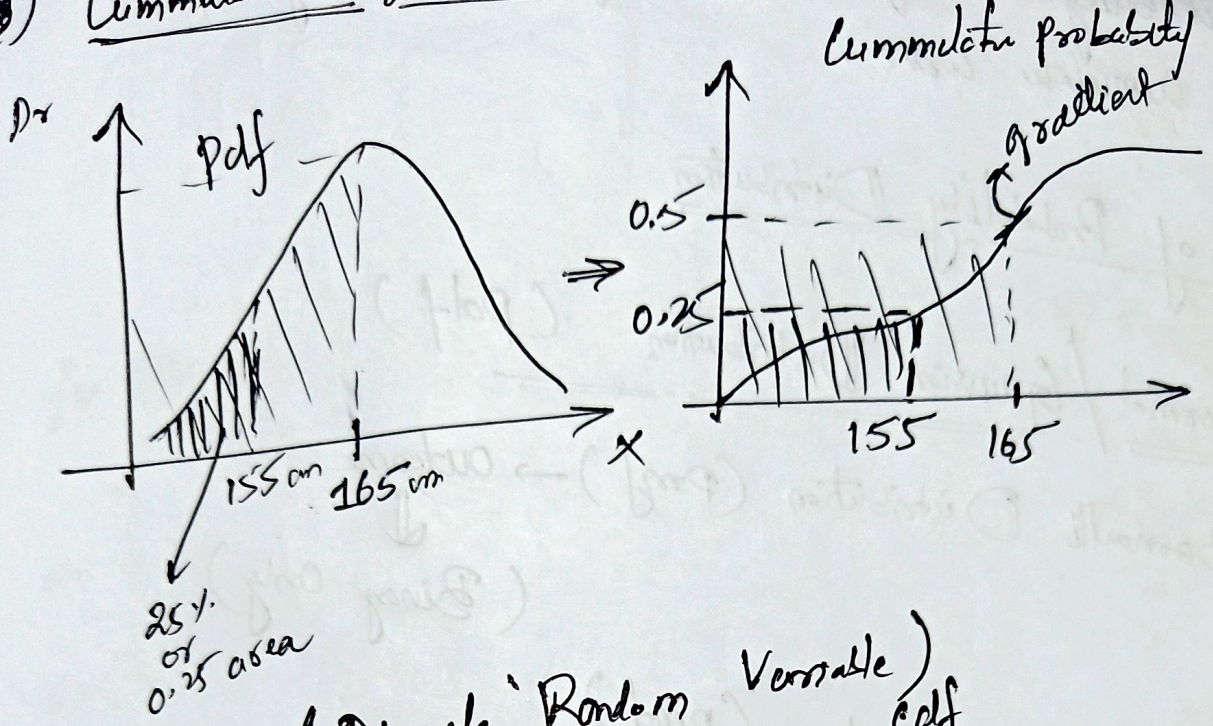


$P(x \leq 4)$

$= P(X=1) + P(X=2) + P(X=3) + P(X=4)$

$= 4/6 = 2/3$

(6) Cummulative Distribution function (cdf)

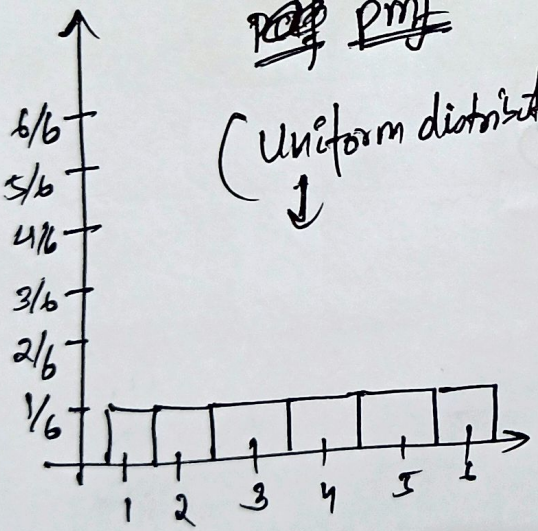


(Discrete Random Variable)

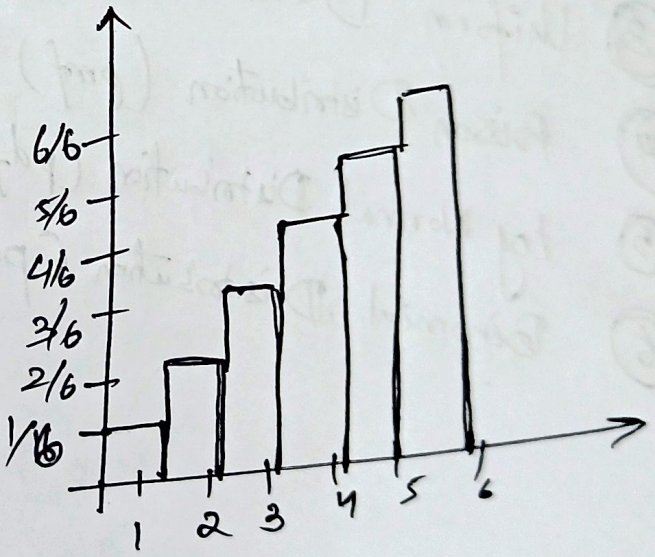
for die

~~pdf~~ pmf

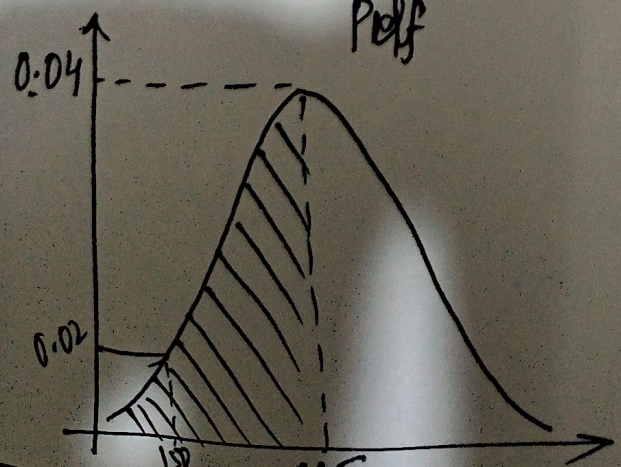
(Uniform distribution)



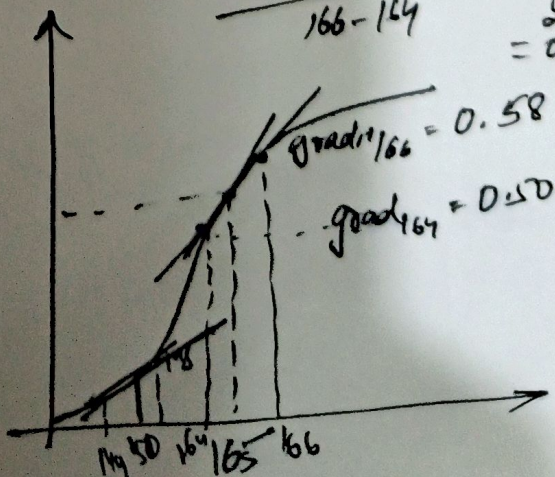
cdf



Distribution of Continous Random Variable



$$\frac{grad_{166} - grad_{164}}{166 - 164} = \frac{0.08}{2} = 0.04$$



Probability Mass function is the gradient value of cumulative curve.

Types of Probability Distribution

1. Normal / Gaussian Distribution (pdf)
2. Bernoulli Distribution (pmf) → outcome ⇓ (Binary Only)
3. Uniform Distribution (pmf)
4. Poisson Distribution (pmf)
5. Log Normal Distribution (pdf)
6. Binomial Distribution (pmf)

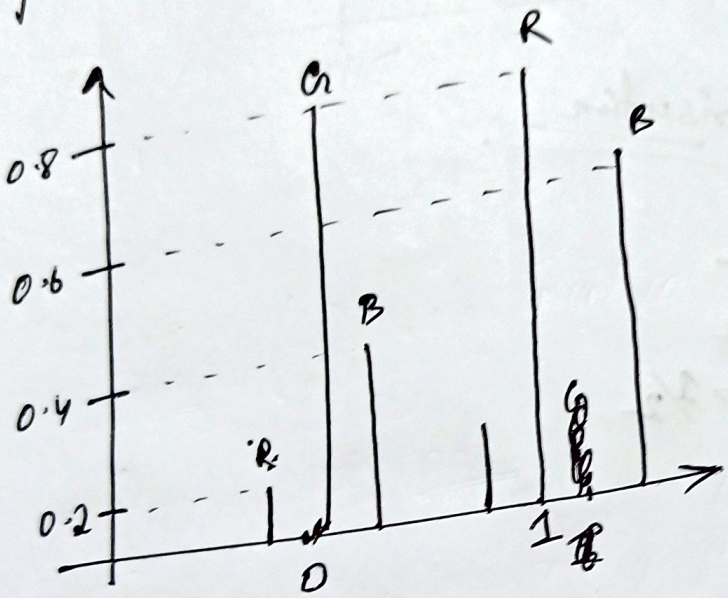
Bernoulli Distribution

↓

two outcomes ; i.e outcomes are binary

↙ success (p) ↘ failure (q = 1-p)

for discrete random variable, it follows pmf (probability mass function)



Red $P(x=0) = 0.2$ $P(x=1) = 0.8$

Blue $P(x=0) = 0.4$ $P(x=1) = 0.6$

Green $P(x=0) = 0.8$ $P(x=1) = 0.2$

Parameters:

$0 \leq p \leq 1$

$q = 1 - p$

$K = \{0, 1\}$

pmf

$$\boxed{p^k (1-p)^{1-k}}, \quad k \in \{0, 1\}$$

Mean of Bernoulli Distribution

$$E(K) = \sum_{e=0}^{K} K \cdot p(K)$$

assume

$P(K=1) = 0.6 \rightarrow p$

$P(K=0) = 0.4 \rightarrow 1-p = q$

if,

$$\sum_{e=0}^{K} K \cdot p(K)$$

$$\Rightarrow [0 * 0.4 + 1 * 0.6]$$

$$= 0.6$$

Median of Bernoulli Distribution

$$\text{Median} \begin{cases} 0 & \text{if } P < 1/2 \\ [0,1] & \text{if } P = 1/2 \\ 1 & \text{if } P > 1/2 \end{cases}$$

Variance

Var $p * (1-p)$

Standard Deviation

Std $= \sqrt{p}$

Binomial Distribution

for discrete random variables.

- ★ Binomial Distribution (n, p)
- ★ Every exp outcome is binary; this experiment is performed for n trials.

★ Bernoulli is special case of Binomial where $n=1$.

Binomial $(B(n,p))$

Params : $n \in \{0, 1, 2, \ldots\}$
$p \in \{0, 1\}$ → Success probability for each trial.

eg: Tossing a coin Ten times.

$K \rightarrow$ no of success.

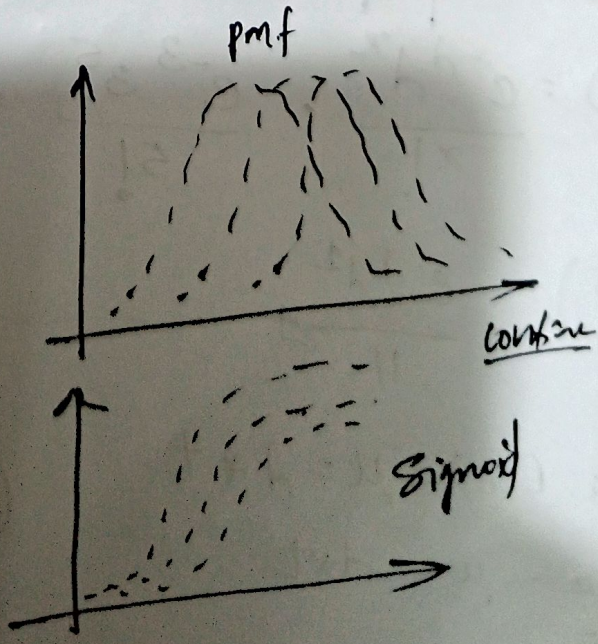
$$\text{Pmf} = P(K, n, p) = {}^{n}C_{K}\, p^{K} (1-p)^{n-K}$$

for $K = 0, 1, 2, \ldots, n$, where $\quad {}^{n}C_{K} = \frac{n!}{K!(n-K)!}$

Mean $= np$

Variance $= npq$

Std deviat $= \sqrt{npq}$



Poisson Distribution

① for Discrete Random Variable

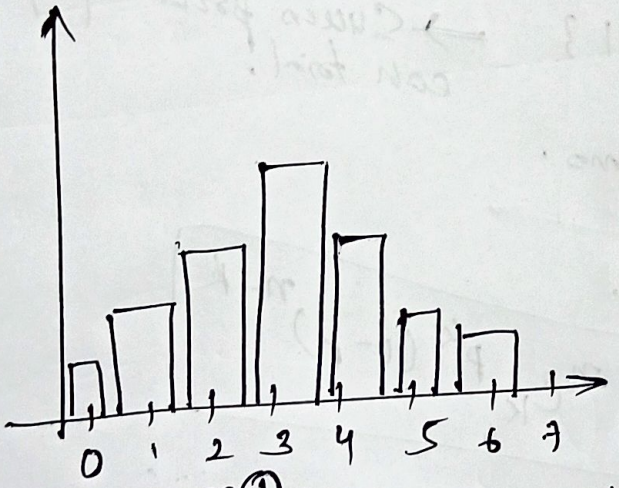
② Describe the no of events occur in a fixed time interval.

eg: ① No. of people visiting hospital every hour.

② No. of people visiting banks every hour.

pmf

λ = (expected no of events to occur at every time interval)



In context to eg ①.

When we say $\lambda = 3$, we mean that we expect atleast 3 people to visit every hour.

$$P(X=5) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-3}3^{5}}{5!} = 0.101 = 10.1\%$$

$$P(X \leq 5) = \frac{e^{-\lambda}\lambda^{1}}{1!} + \cdots + \frac{e^{-\lambda}\lambda^{5}}{5!}$$

Mean : $E(X) = \mu = \lambda * t$ (t = time interval)

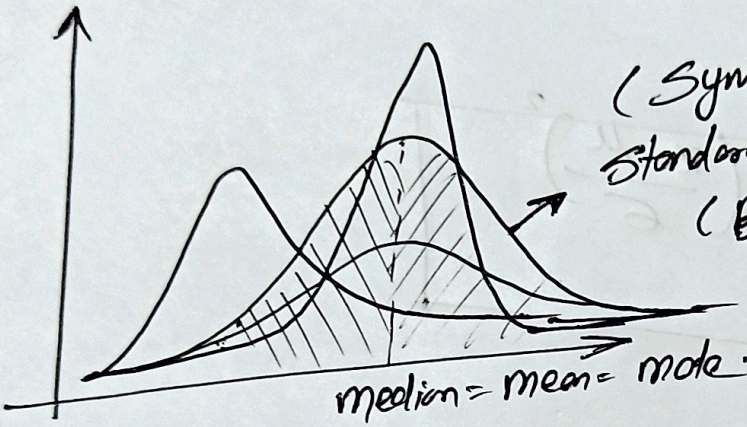
Variance = $\mu = \lambda * t$

Std =

Normal or Gaussian Distribution

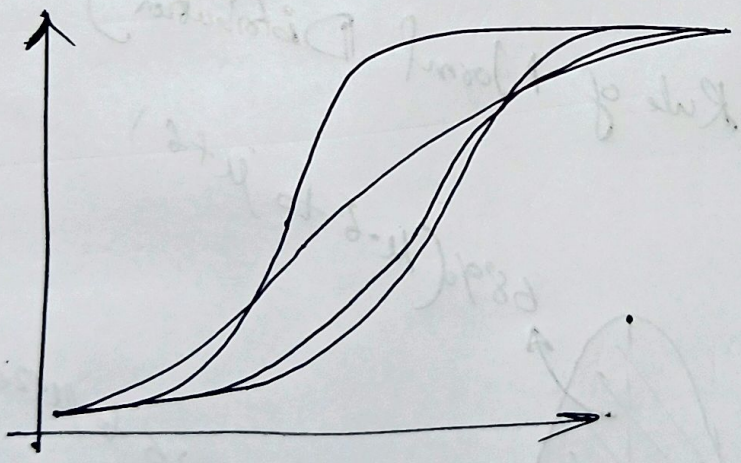
* It is for continous random values.
* It is for a real valued random variable.

Pdf (Probability Density function)



(Symmetrical)
Standard normal deviation.
(Bell curve)

Commulative Density function



Varience ↑ grad ↑

Most of the datasets have this kind of curve

eg: Ages
IQ
[weight
[weight
[length
[width

Notation

$N(\mu, \sigma^2)$

Parameters
1. Mean (μ) $\in R$
2. Varina (σ^2) $\in R$
3. $x \in R$

↓
data points

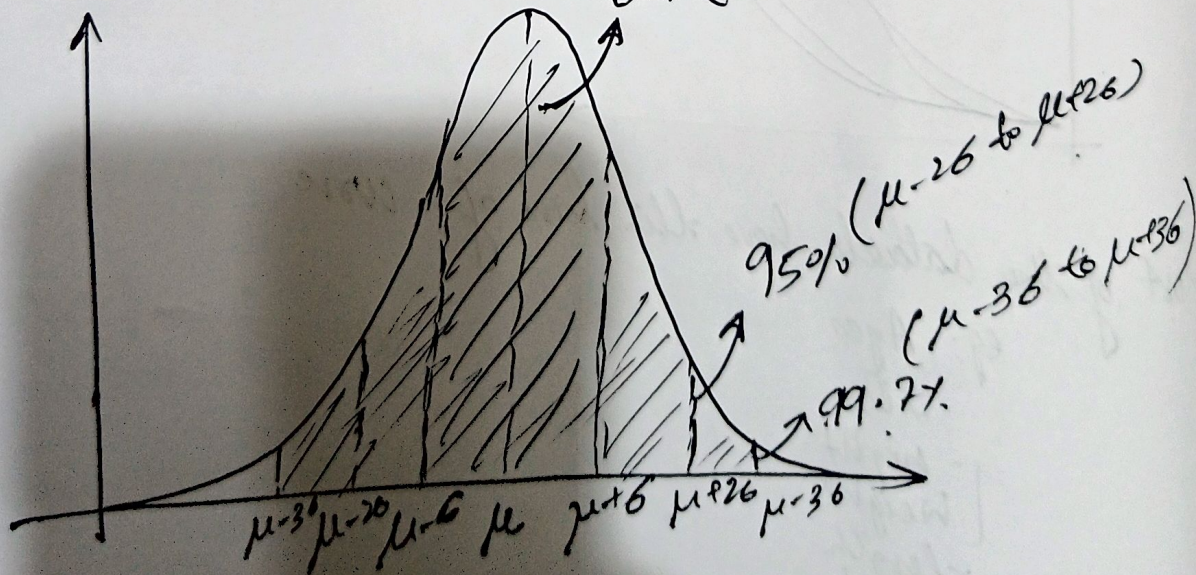
$$\text{Pdf} = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

Mean (μ) = Avg

Varina = σ^2

Std = $\sqrt{\sigma^2} = \sigma$

Assumption (Empirical Rule of Normal Distribution)



$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$

$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$

$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$

Example

Dataset which contains
(i) Weights of students in the class.
(ii) Height of " " "
(iii) IRIS Dataset (sepal width)