Types of Data:

- **Data**
  - **Quantative** — They are basically numerical data
    - **Discrete** — whole no. with a specific range. Ex: No of Bank acc. No. of Children in a family
    - **Continous** — Can have any value. Ex: Weights, heights, Temperature
  - **Qualitative** — Also c/a Categorical Data
    - **Normal** — fixed categorical data. Eg: Gender, Blood grp. # Can't give ranks like ordinal
    - **Ordinal** — Same as Nominal. Difference: Rank. Eg:- Good, Bad, Satisfactory. So, we can provide good as rank 1 & so on

Assignment:

Q1) What kind of variable "Marital status" is? – Catagorical Nominal

Q2) What kind of variable "Nile River length" is? Quantative discrete.

Q3) What kind of variable "Movie Duration" is? Quantitative Continous

Q4) Define the type of datas for the table colums given below.

DMC	DC	ISI	BUI	FWI	Classes	Region
3.4	7.6	1.3	3.4	0.5	not fire	0
4.1	7.6	1	3.9	0.4	not fire	0
2.5	7.1	0.3	2.7	0.1	not fire	0
1.3	6.9	0	1.7	0	not fire	0
3	14.2	1.2	3.9	0.5	not fire	0
5.8	22.2	3.1	7	2.5	fire	0
9.9	30.5	6.4	10.9	7.2	fire	0
12.1	38.3	5.6	13.5	7.1	fire	0

Quantitative Continuos Variable

Catigorical Nominal Variable

Scale of Measurement of Data

In data collection stage we collect varities of data, so to measure that data through some scale is very imp.

There are 4 types of scales of measurement.

1. Nominal Scale Data
2. Ordinal Scale Data
3. Internal Scale Data
4. Ratio Scale Data

we must know in which collection measurement scale the data falls while collecting a data.

1) Nominal Scale Data :-

- type of data involved :- Qualitative / Catagorical variable
- Eg: gender, colour, Labels
- Order does not matter

ex:- 10 ppl asked to choose the Best Colour

"Red" → 5 ppl
"Black" → 3 ppl
"Blue" → 2 ppl

conclusion
50% ppl like red
30% Black
20% Blue

2) Ordinal Scale Data:-

Ranking and Order matters
Difference can'not be measured

Ex: Qualification Dataset

Phd	1st Rank
Masters	2nd
B.Sc	3rd
B.com	4th

Based on rank difference can't be calculated

3) Interval Scale Data :-

Rank and Order matters

Differences can be measured

Ratios can't be measured

Doesn't have '0' as starting point

Ex: Temperature

-30°F
0°F
30°F
60 F
80 F
90 F

- Difference can be measured
  $d = 0 - (-30) = 30°F$

- Ratio has no significance as;
  $\frac{30°F}{60°F} = \frac{1}{2}$, it does not mean that you feel 2 twice as hot in 60°F than in 30°F

4) Ratio Scale Data :-

Order and Rank matters

Differences and Ratios are mesurable

It does have a starting point '0'.

Ex:- Marks

100
90
80
70
60
50
⋮

$\frac{100}{50} = \frac{2}{1}$ Person with marks 100 has double (2x) the marks of a person with marks 50.

I) Measure of Centeral Tendencies :-

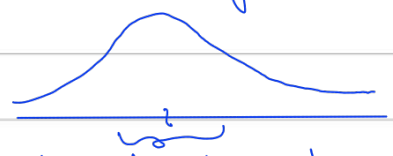
Here we discuss about mainly 3 main topics :-

1. Mean
2. Median
3. Mode

} # EDA (Explorotory Data Analysis) and Feature Engineering will be using these concepts

Q) What does measuring of Centeral Tendency mean?

Soln → Suppose below is the given data distribution :-



So, measure of centeral tendency talks about this centeral region where maximum amount of data is present

1.1) Mean :-

Suppose we have a population of 10 ppl and sample is taken from 5 ppl the calculating both population and sample mean :-

Population (N = 10)	Sample (n = 5)

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

$\therefore$ Population mean $(\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$

$\therefore$ Sample mean $(\mu_n) = \frac{\sum_{i=1}^{n} x_i}{n}$

$\Rightarrow \mu = \frac{32}{10} = \boxed{3.2} = \mu_N$

1.2) Median :-

It is the centeral Element of the sorted list of data

eg: data = 4, 5, 2, 3, 2, 1

$\Downarrow$

sorting data :- 1, 2, 2, 3, 4, 5

$\Downarrow$

Median: Count the the no. of Elements :-

if

a) **Even Count** :- then median is the mean of the two middle nos.

ex: 1, 2, $\boxed{2, 3}$, 4, 5.

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

b) **Odd Count** :- median is $(n+1)^{th}$ element

ex: - for 1, 2, $\boxed{2}$, 3, 4

median = 2

☆5

Q) Why do we need to calculate median?

Soln:- Suppose we have a sample data {1, 2, 3, 4, 5}

$$\text{mean} = \frac{1+2+3+4+5}{5} = 3 \quad , \quad \text{meadian} = 3$$

Now,

Suppose an outlayer (100) is added to the above data set such that the dataset becomes

{1, 2, 3, 4, 5, 100}

an outlyer is a value that does not belong/fit into the dataset

∴,

$$\text{median} = \frac{1+2+3+4+5+100}{6} = 19.16 \quad , \quad \text{mode} = \frac{3+4}{2} = 3.5$$

Observation :- Even after adding a large outlayer the value of median have not shifter that much but there is a significant shif in the value of mean

So, whenever we have an outlayer it is a better strategy to use meadian coz change in median value is not significant as compar-ed to the median.

1.34 Mode :-

Element with the maximum frequency will have the mode value

Ex: → for {2, 1, 4, 5, 1, 7, 8, 1, 9, 1, 10}

$\boxed{\text{mode} = 1}$

Q) Where mode is used?

⇒ suppose dataset :-

Type of Flowers	Age
Lilly	10
Rose	3
Rose	5
Sunflower	Median or Mean
Rose	8

Replaced by 'mode'

If outlyer is present the replace with Median or Mean

EDA on Feature Engineering

IIy Measure of Dispersion

two imp. topics to understand in measure of dispersion :-

1. Varience
2. Standard Deviation.

2.14 Variance :-

Population Varience

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

, x_i = Data points
μ = Population mean
N = Population size

Sample Varience

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

WHY DOES THE SAMPLE VARIANCE HAVE N-1 IN THE DENOMINATOR? The reason we use n-1 rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance

Proved from :- "Bessel Correction"

, x_i -> Data points
$\bar{x}$ -> Sample mean
n -> Sample size

Ex 1: Sample :- $\{1, 2, 3, 4, 5\}$

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

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4

$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 10$ => $s^2 = \frac{10}{4} = \frac{5}{2} = 2.5$

Q7 What does Varience Value Signifies?

Soln - $s^2 = 2.5$ $s^2 = 6.5$

So; Varience specifies the spread of the distribution as varience value ↑ed spread also ↑es.

2.27 Standard Deviation:-

Population Std

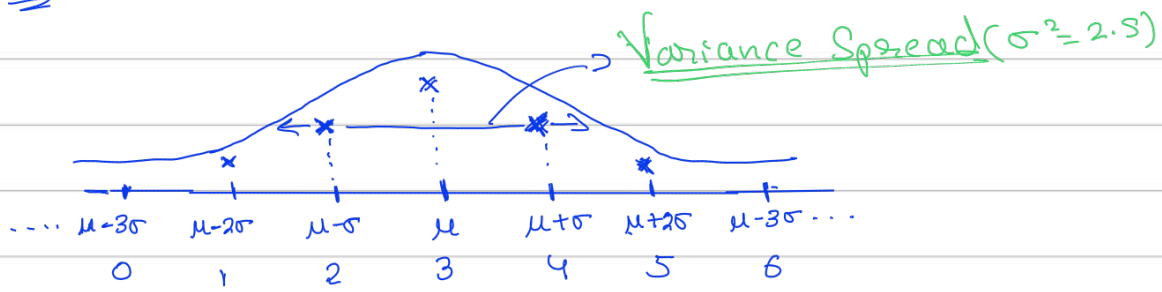
$\boxed{\sigma = \sqrt{\text{varience}}}$

Sample Std.

$\boxed{s = \sqrt{\text{Sample Varience}}}$

Ex: Suppose for a given data set: $\{1, 2, 3, 4, 5\}$, $\mu = 3$ and $\sigma = 1$ are the mean & std. deviation. Then Explain the std. deviation.

Sol^n -



Most of the element in a dataset are always present b/w 1st 3 std. deviations ($\mu-\sigma, \mu, \mu+\sigma$).

III Random Variables

Random variable is a process of mapping o/p of a random process or experiment to a no.

ex: tossing coin
rolling a dice
measure of temperature for next day
} - Random Processes

Q) How is random variable diff from algebric variables?

Random Variable (X)	Algebric Variable (x)
•) Can take any value based on the process outcome	•) Algebric Variable values is particularly fixed
•) Denoted by : Capital Letter	•) Denoted by : small letter
•) Ex: Sum (dice rolled 7 times) i.e $\sum_{i=1}^{7}$ (dice roll)$_i$ ⇒ o/p 1 :- {1, 2, 4, 6, 3, 4, 2} ⇒ o/p 2 :- {2, 3, 3, 6, 5, 5} ⋮	•) Ex: $x+5=7$; & $y+x=10$ ⇒ $x=2$ & $y=8$
•) Useful in Probability problems $X = \begin{cases} 0 & \text{if Heads} \\ 1 & \text{if tails} \end{cases}$	

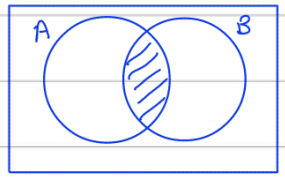
IV) Sets

Suppose two sets :- $A = \{1, 2, 3, 4, 5, 6, 7 8\}$
$B = \{3, 4, 5, 6, 7\}$

Following are some diff. set operations.

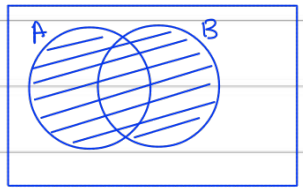
a) Intersection : $(\cap)$

$(A \cap B) = \{3, 4, 5, 6, 7\}$



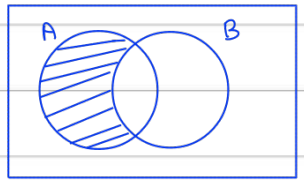
b) Union $(\cup)$:-

$(A \cup B) = \{1, 2, 3, 4, 5, 6, 7, 8\}$



c) Difference

$(A - B) = \{1, 2, 8\}$



d) Subset : $(\subset)$

is $A \subset B \rightarrow$ False

is $B \subset A \rightarrow$ True

is $A \not\subset B \rightarrow$ True

$8 \subset A \rightarrow (\times)$; $\{8\} \subset A$ $(\checkmark)$

e) Superset :- $(\supset)$

is $(A \supset B) \rightarrow$ True

is $(B \supset A) \rightarrow$ False

f) Belong to $(\in)$:

$8 \in A \rightarrow$ True

$8 \in B \rightarrow$ False

g) Not belongs to $(\notin)$:

$8 \notin A \rightarrow$ False

$8 \notin B \rightarrow$ True

IV Covarience & Correlation :-

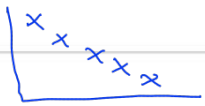
Suppose we have 2 features on 2 Random Variables (X, Y):-

X	Y
2	3
4	5
6	7
8	9

So, can we find a relationship b/w X and Y. (i.e like if X↑es then Y↑es etc.)

Plotting : $\boxed{\begin{matrix} X\uparrow, & Y\uparrow \\ X\downarrow, & Y\downarrow \end{matrix}}$ ≡

Plotting: $\boxed{\begin{matrix} X\downarrow, & Y\uparrow \\ X\uparrow, & Y\downarrow \end{matrix}}$ ≡



5.1 Covarience (X, Y) or Cov (X, Y) :-

now, covarience w.r.t random variables X and Y is given by formulae :-

$$\boxed{Cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}}$$

Note ⟹ what the difference b/w Varience and Covarience

Varience $(X^2) = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{n-1}$ & Covarience $(X,Y) = \sum_{i=1}^{n} \frac{(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$

(signifies spread)

So; Varience is nothing but Covarience of random variable X with itself i.e Cov(X, X)

i.e

$$Cov(X,X) = \sum_{i=1}^{n} \frac{(X_i-X)(X_i-X)}{n-1} = \sum_{i=1}^{n} \frac{(X_i-X)^2}{n-1} = Var(X^2)$$

∘) For (X↑, Y↑ / X↓, Y↓) Relationship : Covarience = +ve

X	Y
2	3
4	5
6	7
$\bar{X}=4$	$\bar{Y}=5$

$$Cov(X,Y) = \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{3-1}$$

$$= \frac{4+0+4}{2} = \boxed{4} \text{ (+ve)}$$

∴ X and Y are having a +ve covarience

Q4 Example where Covarience can be used?

Sol^n -> Suppose we have a dataset where Price and size of different houses is given, and In such case we have to predict the price of the house

So, Covarience can be used to find the relationship b/w the size and prices of the house.

Advantages and Disadvantages of Covarience:-

- **Advantage:-** Using covarience we can find a relation b/w X and Y. It can be +ve or -ve

- **Disadvantage:-** Covarience doesnot have a specific limit value

  Ex: Suppose we have X, Y, Z

  if Cov (X, Y) = +100 & Cov (X, Z) = +500

  it does not mean X is heavily corelated to Z as compared to Y cuz there is no specific limit to Cov value to compare high or low

So;

in order to overcome this advantage of covarience we specifically use another correlation technique c/a "Pearson Correlation Coefficient"

5.24 Pearson Correlation Coefficient (ρ) [range: -1 to +1]

as its value is restricted b/w -1 and +1 ∴, now we can differentiate which variables are highly correlated and which are not.

$$\boxed{\text{Correlation}(X,Y)\,(\rho_{X,Y}) = \frac{Cov(X,Y)}{\sigma_X\ \sigma_Y} = [-1 \text{ to } 1]}$$

- more the value is towards +1, the more +vely correlated X & Y are.
- more the value of ρ_{XY} is towards -1, the more -vely correlated X & Y are

Qy Where can this correlation concept is applied in data science project?

Soln Suppose a dataset with 1000 features,
now, for an ML project it is not feasible to take all of the features

Ex: dataset features: Size of house, No of Rooms, Location, No. of PPL staying, Price

Independent features

O/p or dependent feature

no. using Feature Selection :-
if Cov (No. of ppl, Price) ≈ or near 0 then it can be removed from our ML model.

5.34 Spearman Rank Correlation :- (γ_s)

sometime it is consider to be better than Pearson Correlation.
here instead of considering the value of X & Y we consider the rank of both X and Y i.e

$$\boxed{\gamma_s = \frac{Cov[R(X), R(Y)]}{\sigma_{R(X)}\ \sigma_{R(Y)}}}$$

Ex: if

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

VII Histogram and Skewness :-

Remember to learn : 5 no. summary, Box Plot

6.17 Histogram :-

Histogram are very useful in discriptive statistics to summarize the data and take out information regarding distribution of data.

Bins in Histogram Bin refers to no. of intervals or buckets for a perticular no. of datasets

ex:-

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

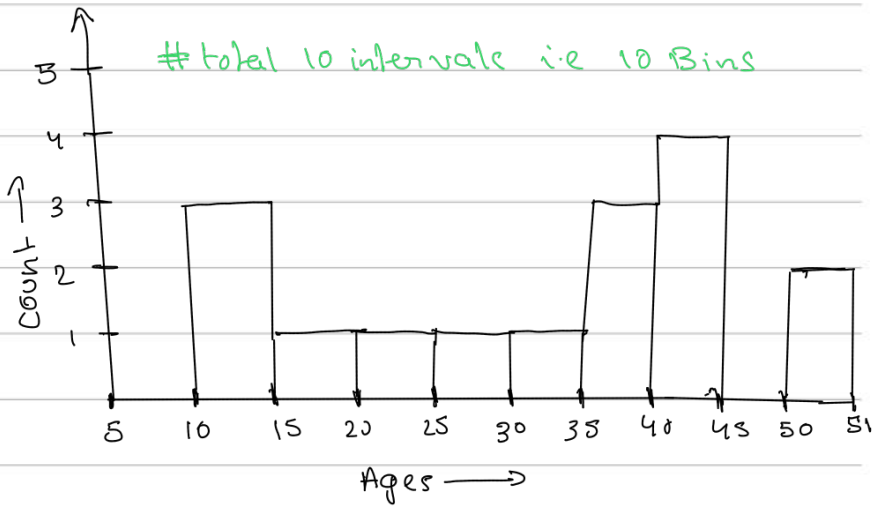
∴,

Calculating Ages difference for bin size = 10

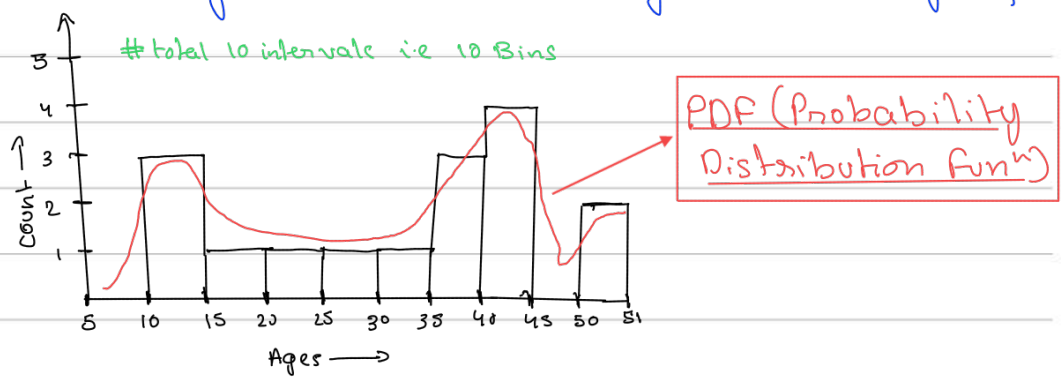
$$\boxed{\text{min data range} = \frac{\text{max} - \text{min}}{\text{Bin Size}} = \frac{51-0}{10} = 5.1 \approx 5}$$

So,

Histogram :-

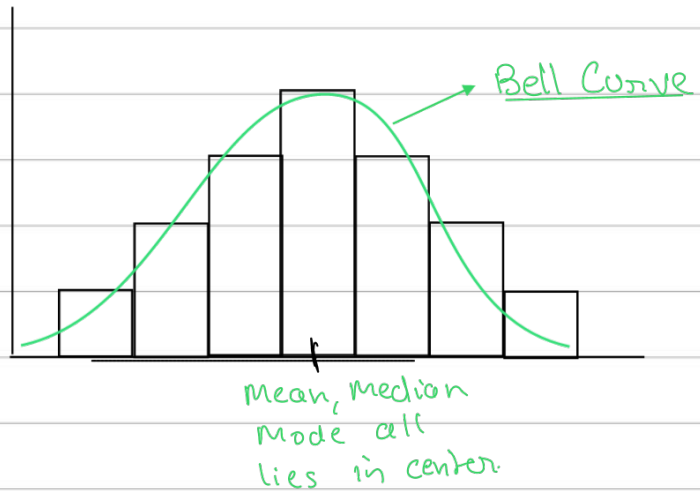


Now, smoothening the above Histogram we get,



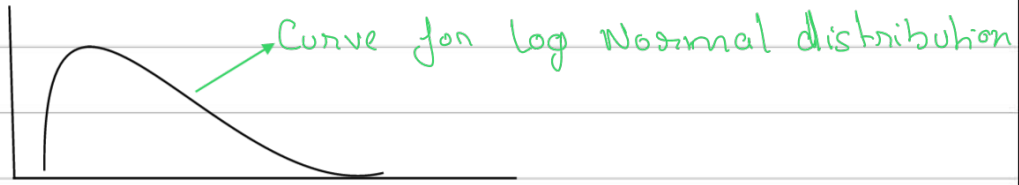
Probability Distribution Function (PDF):- This basically talks about how my data is getting distributed. PDF is achieved by smoothening the Histogram by a concept called as Kernal Density Estimator (KDE)

6.1.2Y Bell Curve



Bell Curve is a type of Probability distribution function for which all centeral tendencies (like mean, median & mode lies at the center of the distribution and such type of distribution is cfa "Normal / Gaussian Distribution".

6.1.3Y Log Normal Distribution Curve

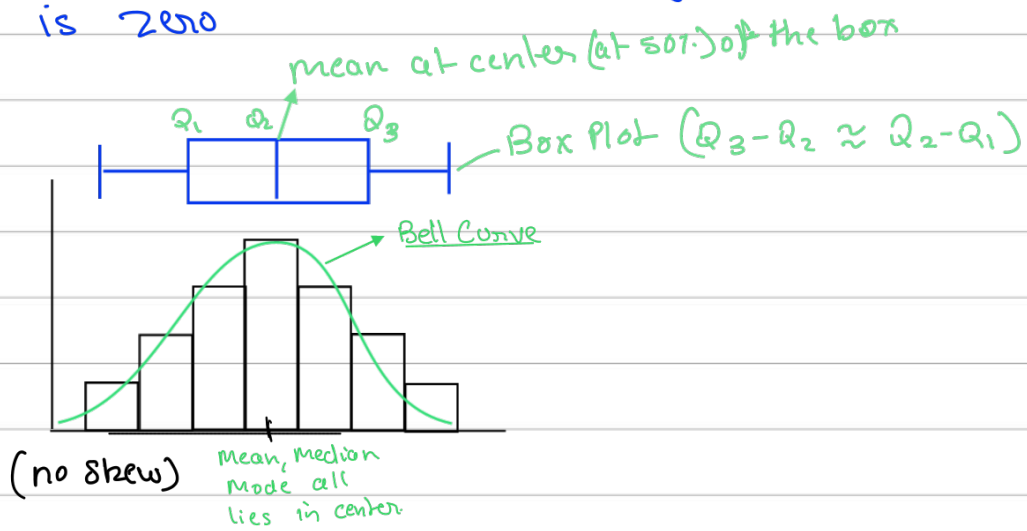


6.2Y Skewness :-

6.2.1Y No Skewness :-

Normal Distribution also known as Gaussian Distribution is a Bell shaped symmetrical distribution in which the mean, median and mode all are perfectly at the center (i.e all are

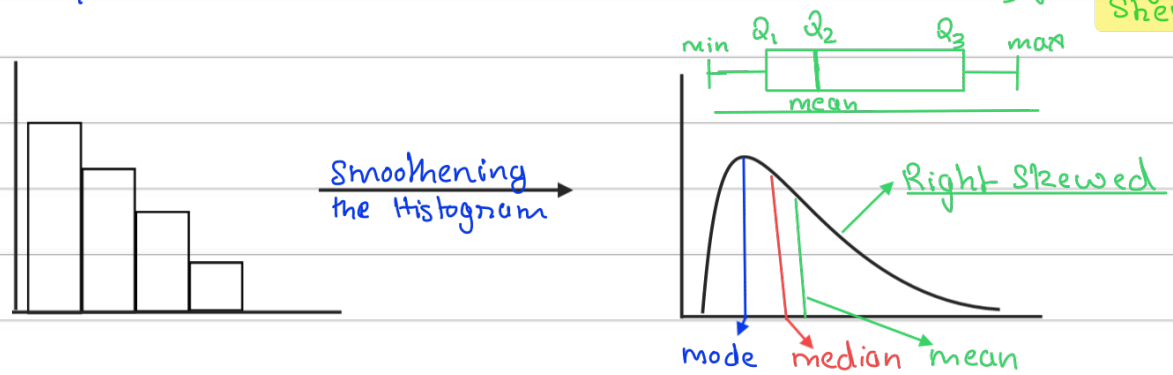
equal). For such kind of distribution skewness is zero



6.2.2) Right Skewed Distribution :- Log Normal Distribution

also c/a Positive Skewed

($Q_3-Q_2 > Q_2-Q_1$) for Right Skew



∴ For Right Skew distribution mean > median > mode

Q) Why mean > median > mode for Right Skew Distribution?

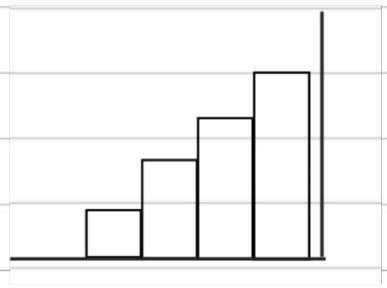
Soln → **Mean**: as for Right Skewed type of distribution the maximum no. of distribution or data is on the right so mean is shifted most to the right

Median :- As we know in case of an overlay median shift is alway less than the mean ∴ Median is shift is less that mean

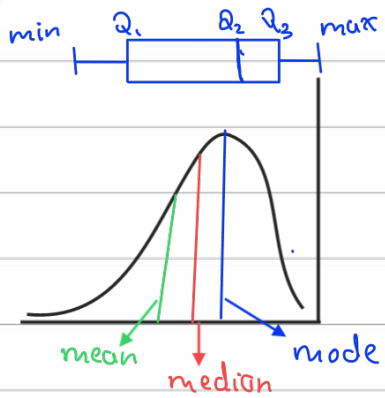
Mode :- As mode is always at the highest freq ∴ it will found at the maxima.

6.2.34 Left Skewed Distribution :-

$Q_2 - Q_1 > Q_3 - Q_2$ for Left Skew



Smothening →



∴,

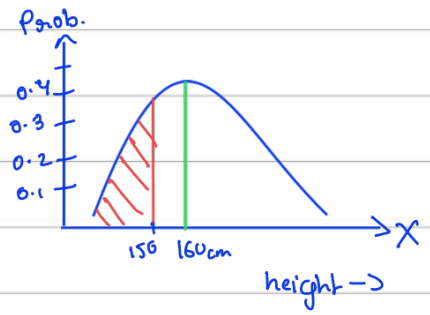
1) Probability Distribution and Density function

all PDF, PMF, CDF are part of Probability Distribution function

1.1) Probability Density Function :- (PDF)

It shows distribution of continous data.

Ex: Height of student



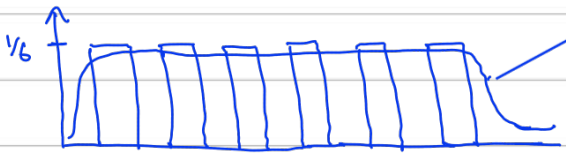
What is the probability for height less than 150?

Probability of having height < 150 is given by area of graph under 150

i.e $\boxed{P(x<150) = \text{Area under curve}}$

1.2) Probability Mass Function (PMF):- Distribution of a discrete Random variable

ex: Rolling of a dice $\{1, 2, 3, 4, 5, 6\}$



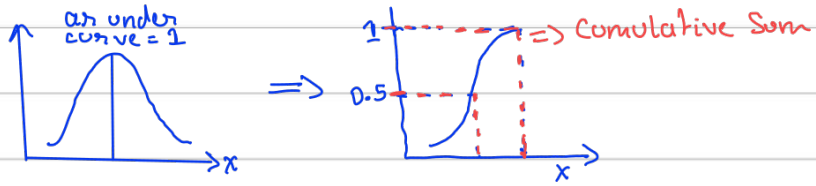
PMF: created by smoothening the histogram

Q) What is the the probability of $X < 4$?

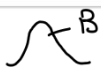
Sol^n -> $P(x<4) = P(x=1) + P(x=2) + P(x=3) + P(x=4)$

$= (1/6) \times 4 = \boxed{2/3}$

1.3) Cumulative Distribution Function (CDF):- CDF is cumulative sum of every value wrt area under the curve of PDF.



II) Types of Probability Distribution :-

① Normal / Gaussian Distribution :-  Bell curve

② Bernoulli Distribution (PMF) : 2 outcomes success or failure

③ Uniform Distribution

④ Log Normal Distribution (PDF)

⑤ Poisson Distribution Rule (PMF)

⑥ Power Law Distribution (PdF) : c/a 80 - 20% Rule

⑦ Binomial Distribution (PMF)

III) PDF vs CDF vs PMF

PDF → Probability density function

PMF → Probability mass function

CDF → Cumulative density function

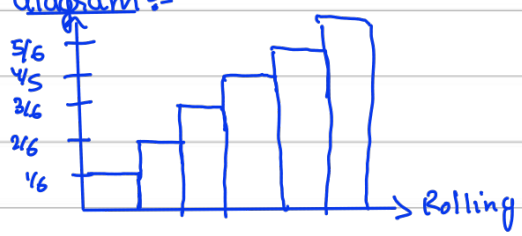
① PMF :-

Requirement : Discrete random variable

PMF	CDF
Ex:- Rolling a dice	"
$\{1,2,3,4,5,6\}$ $P(X)(1) = 1/6$	"
$P(X)(2) = 1/6$	"

diagram :-

diagram :-



in CDF probabilities are cumulative-ly added

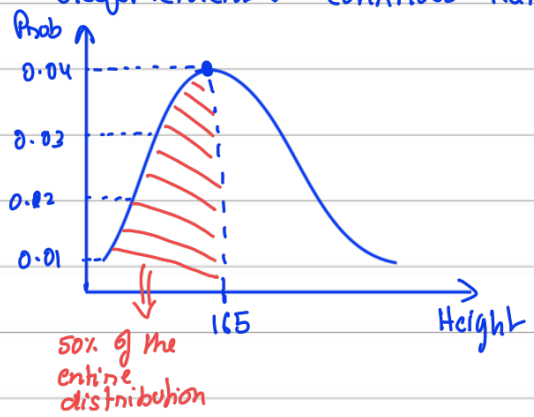
with CDF we can find the probability of lets say $P(x \leq 4)$

$P(x \leq 4) = P(1) + P(2) + P(3) + P(4)$

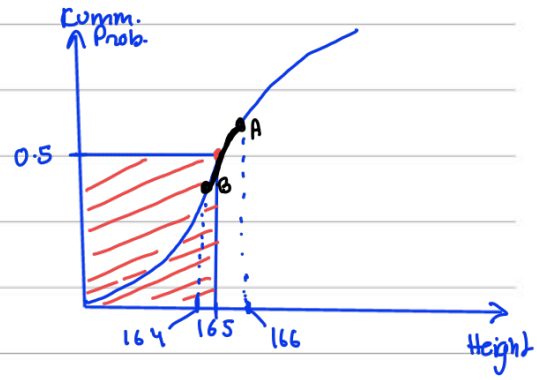
$= 4/5$

② **PDF**: Probability Density Function.

Requirement: Continous Random Variable.



CDF →



#finding gradient:-

Slope at pt A = 0.58

Slope at pt B = 0.50

$$\text{gradient} = \frac{0.58 - 0.50}{166 - 164} = \frac{0.08}{2} = 0.04 \quad (1)$$

therefore from equation (1) we conclude that Probability density is gradient descent of CDF

IV Bernauli Distribution: [Binary Outcomes]

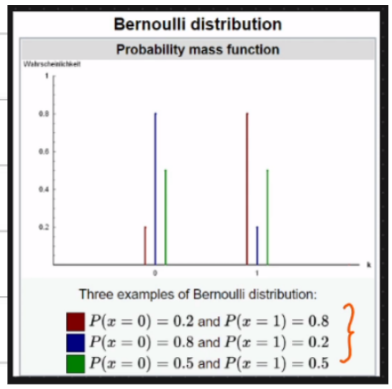
Bernauli Distrib^n and log normal distributions are not normal or gaussian distributions

In probability theory and Stastics, Bernaulli distribution, named after Swiss mathematician Jacob Bernaulli, is the discrette probability distribution of random variables which takes value 1 with probability p and value 0 with probability $q = 1 - p$. Less formally it can be thought as model for the set of possible outcomes of any single experiment that asks yes-no questions. Such questions leads to outcomes that are boolean-valued: a single bit whoose value is success/yes/true/one with probability p. and failure/no/false/zero with probability q.

Ex: ① tossing a fair coin. $\{\overset{H}{0}, \overset{T}{1}\}$:

$$P(T) = 0.5 = p$$

$$P(H) = 1 - 0.5 = 1 - p = q$$



② Weather a person will fail or pass?

Deriving probability mass function for Bernauli Distribution

for Bernaulli dist. R value is either 0 or 1

i.e, $k = 0$ or $k = 1$

so,

Pmf: probability of getting outcome $k = 0$ is give by:

$$P(x=k) = p^k (1-p)^{1-k}$$

→ for $k = 1$: $P(x=1) = p^1 (1-p)^0 = p$

→ for $k = 0$: $P(x=0) = p^0 (1-p)^{1-0} = (1-p) = q$

Hence its a Bernaulli distribution's pmf cuz we are getting p or q values for diff. scenarios.

PMF formula simplified

$$PMF = \begin{cases} q = 1-p & \text{if } k=0 \\ p & \text{if } k=1 \end{cases}$$

Calculating mean variance and std distribution for Bernauli

a). Mean:- mean of a Bernaulli distribution is always p.

mean given by :- $\boxed{E(k) = \sum_{i=1}^{k} k \cdot P(k)}$

ex: Suppose prob. of $P(k=1) = 0.6 = p$, then

$P(k=0) = 1 - 0.6 = 0.4 = q$

now using above eqn to find mean using above eqn

$$E(k=1) = \sum^{k} k \; P(k)$$

$$= (1 \times 0.6) + (0 \times 0.4)$$

$$= 0.6 = p$$

Hence, the mean of a Bernaulli Distribution is always p.

b) Median:

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

c) Variance & std

$$\boxed{\begin{aligned} \text{variance} &= p(1-q) = pq \\ \text{std} &= \sqrt{pq} \end{aligned}}$$

IV Bionomial Distribution

for a process whoose outcomes are binary but the process is repeated n no. of times such type of distribution is called as Bionomial Distribution.

In probability theory and statistics, the bionomial distribution with parameter n and p ——— discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking yes-no questions, and each with its own Boolean-valued outcome: (success with probability p) or failure (with prob. $q = 1-p$). A single Success/failure experiment is also called as Bernaulli trial or Bernaulli experiment, and a sequence of outcomes is called as Bernaulli process, for a single trial, i.e, $n=1$, the bionomial distribution is a Bernaulli distribution.

ex: Tossing a coin
{Bernaulli distribution}

$P(H) = 0.5 = 1-P = q \rightarrow 0$

$P(T) = 0.5 = P \longrightarrow 1$

Tossing a coin 10 times
{Bionomial distribution}

	1st time	2nd time	3rd time	...
$P(T) = P$		P	P	...
$P(H) = 1-P$		$1-P$	$1-P$	...

#So Bionomial Distribution can combine various Bernaulli distributions.

Parameters used:-

$n \in \{0,1,2,3 \ldots\}$: no. of Trials or Experiments

$p \in \{0 \text{ to } 1\}$: Success or probability of each trial

$q = 1-p$

PMF of Bionomial dist.:-

$$\boxed{P(k) = {}^nC_k \, p^k (1-p)^k}$$

, where $k\{0,1,2,\ldots\}$: no. of successes

Mean of Bionomial dist :- $\boxed{nP}$

Variance of Bionomial dist :- $\boxed{nPq}$

Std of Bionomial dist :- $\boxed{\sqrt{nPq}}$

VII Poisson Distribution :-

for : ① Discrete Distribution

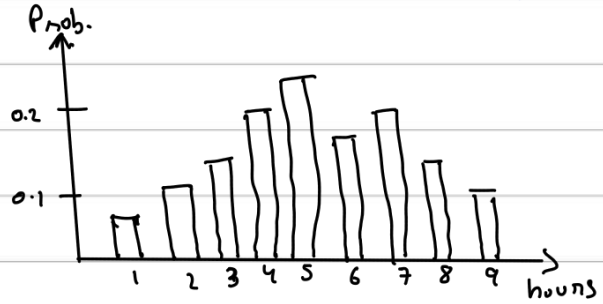
② Describes no. of events occuring in a fixed time interval.

ex: No. of people visiting hospital every hour

No. of people visiting banks " "

" " " " banks " "

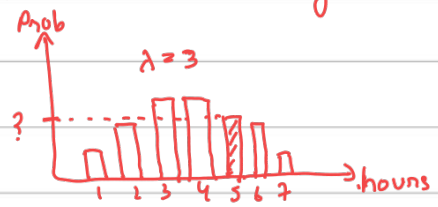
Distribution:



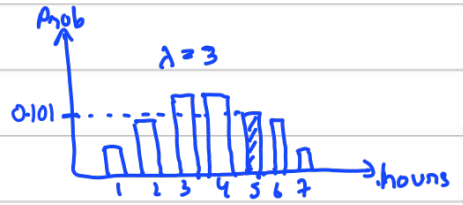
λ parameter: signifies expected no. of events to occur every hour

PMF for Poisson dist :- $\boxed{P(x) = \frac{e^{-\lambda} \lambda^{x}}{x!}}$

calculate the PMF for the following distribution with $\lambda = 3$.



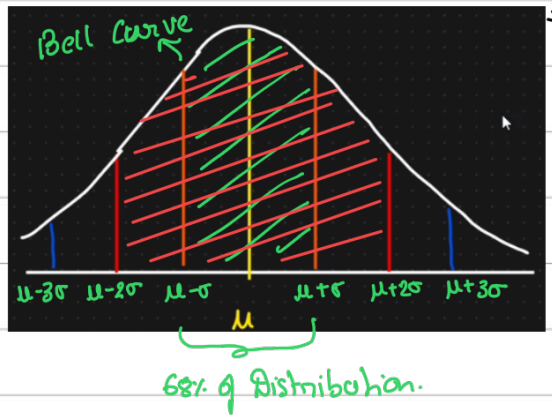
Sol^n: $P(x=5) = \frac{e^{-\lambda} \lambda^{x}}{x!} = \frac{e^{-3} 3^{5}}{5!} = 0.101$ ≈ 10.1%



Mean and Variance of Poisson dist :- $\boxed{E(x) = \mu = \lambda \times t}$

mean and variance are same in case of poisson distribution

VIII) Normal or Gaussian Distribution



⇒ Symmetric Distribution

also c/a "3 Sigma Rule"

Emperical Rule : within the 1st standard deviation from μ (i.e μ-σ to μ+σ) around 68% of the data falls on this region and

95% of the distribution will be falling in the 2nd std. deviation and

99.7% of the distribution will be falling in the 3rd std deviation

QQ Plot: used to determine wether a distribution is Guassian/Normal distribution or not.

Q) What is the probability of a random variable to fall b/w

① $P(\mu-\sigma \le X \le \mu+\sigma) \longrightarrow \approx 68\%$

② $P(\mu-2\sigma \le X \le \mu+2\sigma) \longrightarrow \approx 95\%$

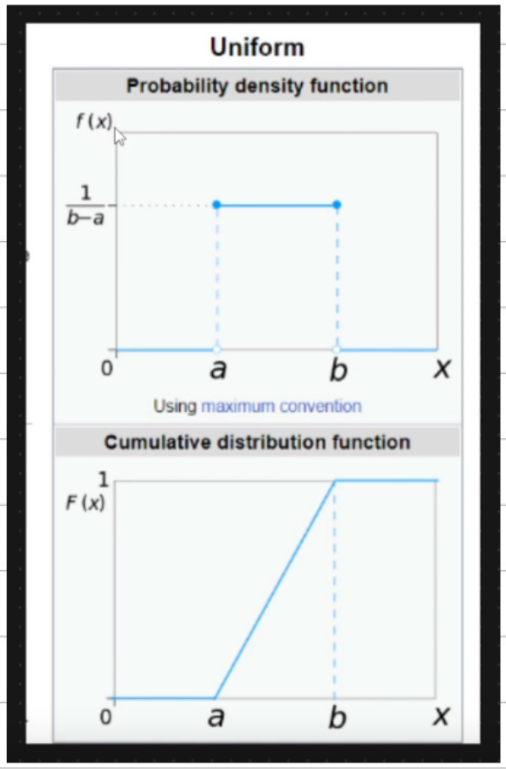
③ $P(\mu-3\sigma \le X \le \mu+3\sigma) \longrightarrow \approx 99.7\%$

IV {Uniform Distribution}

there are 2 types of Uniform distribution ① Continous
② Discrete

① Continous Normal Distribution :- (for Continous Random Variables)

In probability theory and stastistics, the continous uniform distribution or rectangular distribution is a family of symmetric probability distributions. The distributions describes an experiment where there is an arbitrary outcome that lies b/w a certain bounds. The bounds are defined by parameters a and b, which are the minimum and maximum values. The interval can either be closed (eg. $[a,b]$) or open (eg. (a,b)).



notation: $\boxed{U(a,b)}$

parameters: $-\infty < a \leq b < \infty$

$$\boxed{PDF = \begin{cases} 1/(b-a) & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}}$$

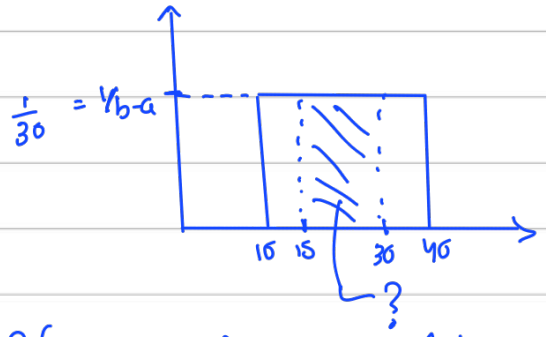
$$\boxed{CDF = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}}$$

$$\boxed{\text{mean} = \frac{1}{2}(a+b), \quad \text{median} = \frac{1}{2}(a+b), \quad \text{variance} = \frac{1}{12}(b-a)^2}$$

Ex:- No of candies sold daily at a shop is distributed with a min. of 10 and max. of 40.

i) what is the probability of daily sales to fall b/w 15 & 30?

Sol^n → as candies sold is uniformly distributed within 10 to 40



$P(15 \le x \le 30)$ = area of the shaded region

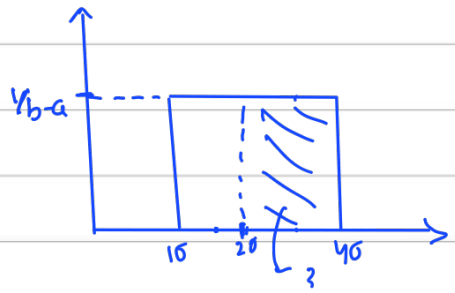
$= (30-15) \times \frac{1}{b-a}$

$= 15 \times \left(\frac{1}{30}\right)$

$= \underline{0.5}$

∴ 50% of daily sales will probably fall b/w this area

ii) $P(x \geq 20) = ?$



$P(x \geq 20) = (40-20) \times 1/30$

$= 20/30$

$= 0.\bar{6}$

$= 66.66\%$ of daily sales.

② Discrete Uniform Distribution: (for discrete random variables)

In probability theory and stastistics, the discrete uniform dist. is a symmetric probability distribution wherein a finite no. of values are equally likely to be observed; every one of n values has equal probability 1/n. Another way of saying "discrete uniform distribution" would be "a known, finite no. of outcomes equally likely to happen".

Ex:- Rolling a dice {1, 2, 3, 4, 5, 6}

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/n = 1/6$

where $a = 1$ & $b = 6$

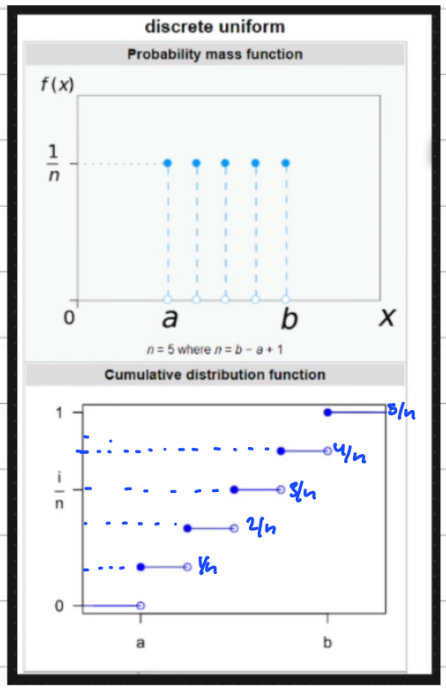
& $\boxed{n = b-a+1}$

Notations used: $U(a,b)$

parameters: a & b, $b > a$

$\boxed{PMF = 1/n}$

$\boxed{\text{Mean \& Median} = \frac{a+b}{2}}$



III) STANDARD NORMAL DISTRIBUTION :-

Let consider a random variable X such that

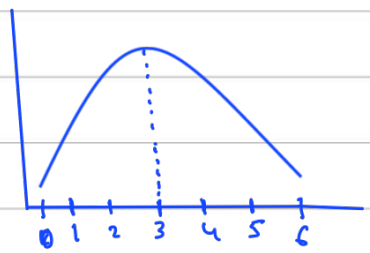
$$X = \{1, 2, 3, 4, 5\}$$

and suppose X is following a std. normal distribution such that

$$\text{mean} = \frac{\sum X}{n} = \frac{15}{5} = 3$$

$$\text{std deviation } (\sigma) = 1.414 \ (\approx 1 \Rightarrow \text{to make calculation easy})$$

So, the gaussian distribution is :-



$\Rightarrow \mu = 3$ & $\sigma = 1.414$

So if we convert the above distribution to a distribution where mean $= 0$ and std deviation $(\sigma) = 1$ then such type of distribution is c/a "<u>Standard Normal Distribution</u>".

Z-Score Method: it is used to convert a gaussian distribution to a normal distribution.

$$\boxed{Z \text{ score} = \frac{x_i - \mu}{\sigma}}$$

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

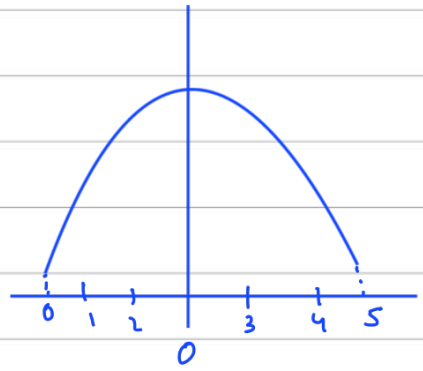
$Z(1) = \frac{1-3}{1} = -2$

$Z(2) = \quad = -1$

$Z(3) = \quad = 0$

$Z(4) = \quad = 1$

$Z(5) = \quad = 2$



Any distribution belongs to a standard normal distribution if its mean (μ) = 0 and $\sigma = 1$

i.e $\boxed{\text{dist}^n \in \text{Std. Normal dist} \ (\text{if } \mu = 0 \ \& \ \sigma = 1)}$

Q. What is the significance of Z-score? Explain with example

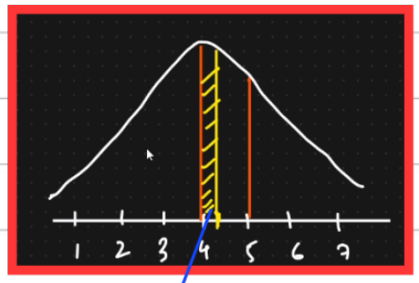
Sol^n → ① Z-score signifiece how much standard deviation is a value away from the mean.

Ex:

Suppose there is a deviation such that its $\mu = 4$ & $\sigma = 1$ then How many std. deviation 4.25 is away from the mean

Sol^n → calculating z-score

$$\text{z-score} = \frac{4.25 - 4}{1} = 0.25$$



$\therefore$ 4.25 is 0.25 std. deviation away from the mean

② Z-score can also be used to find the area under the curve b/w to deviations.

Q) Explain with an example that why we are using z-score and std. normal distribution?

Solⁿ →

Suppose we have a dataset with features: Ag, weight, height, salary

Eg: Dataset

Years Age	kg Weight	Cms Height	INR Salary
24	70	175	40K
25	60	160	50K
26	55	180	60K
27	40	130	30K
30	30	175	20K
31	25	180	70K

all the features have diff units.

But,

To solve a machine learning problem or create a ML model, 1 rule is to bring all the features to a same unit scale

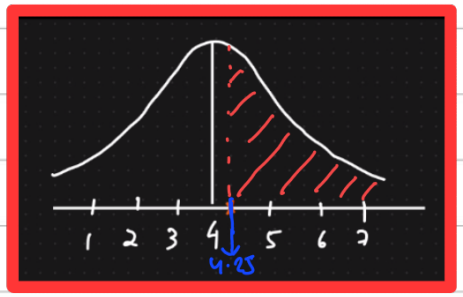
∴ we convert all these features by a technique c/a "Standardization (applying z-score method to all data-set)"

$$z(age) = \frac{x_i - \mu_{age}}{\sigma}$$

Clustering Algo, Linear Regression, Logistic Regression are some example of algorithms that use standardization technique.

III) Z-SCORE APPLICATION & Z-TABLE :-

Q) for the following guassian distribution what %age of of data falls over 4.25?



1-Step - computing z-score

$$z\ score = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

∴ 4.25 stands around 0.25 std. deviation away from mean

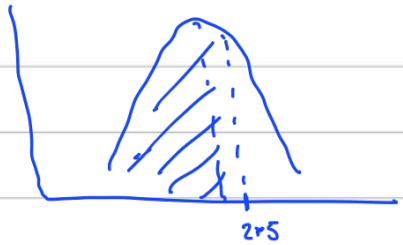
2nd Step: Using z-table to find to find are under the curve

two type of z-tables

① +ve z-table ② -ve z-table

z-table calculates area a left of the z score ∴

∴,



at 4.25 Left side area = 0.59871

to calculate for distribution above 4.25 :-

4.25 Right side area = 1 − 4.25 Left-side area

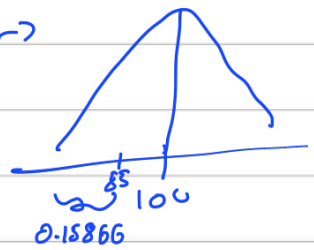
= 1 − 0.59871

= .40129 ≈ .4013

= 40.13% of data falls above 4.25

Qy In India avg IQ is 100, with a std. deviation of 15, What is the % of the population would you expect to have an IQ lower than 85?

Soln →



$\mu = 100$

$\sigma = 15$

$z \text{ score} = \frac{x_i - \mu}{\sigma} = \frac{85-100}{15} = -1$

according to z-table : 0.15866 ≈ 15.86% of IQs in india is below 85

∴ %age of people having IQ above 85 : 100 − 15.86 ≈ 84.14

★ find %age of ppl with b/w 75 ≤ IQ ≤ 100?

z score (75) = $\frac{75-100}{15} = \frac{-25}{15} = -1.66$ & ar = 0.4846 ≈ 48.46

z score (100) = $\frac{100-100}{15} = 0$ & ar = 50%

z score (100) − z score (75) = 1.54%

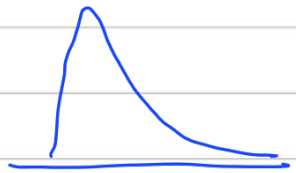
∴, 1.54% of ppl have IQ b/w 75 & 100.

IV) CENTERAL LIMIT THEOREM:-

Centeral limit theorem relies on concept of sampeling distribution, which is the probability distribution of a statistic for a large no. of samples taken from a population.

The Centeral Limit theorem says that the sampeling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of weather the population has a normal, Possion, bionomial, or any other distribution, the sampling distribution of mean will always be std. normal

ex1: Suppose $X \not\approx N(\mu, \sigma)$ not a normal distribution

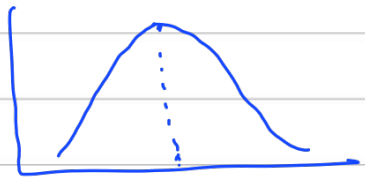


∴, creating $n \geq 30$ samples for non gaussion distibutions such that

$$S_1 \{x_{11}, x_{12}, x_{13} \ldots x_{1n}\} = \bar{x}_1$$
$$S_2 \{x_{21}, x_{22}, x_{23} \ldots x_{2n}\} = \bar{x}_2$$
$$\vdots \qquad\qquad \vdots$$
$$S_m \{x_{m1}, x_{m2} \ldots x_{mn}\} = \bar{x}_m$$

Now,

according to CLT the distribution of above sample means is always a normal i.e



★ mean of sample distribution remains same as population mean but the std deviation changes from σ to $\sigma/\sqrt{n}$, where $n \rightarrow$ Sample Size.

i.e $X \approx N(\mu, \sigma)$, $\mu_{new} = \mu$ & $\sigma_{new} = \frac{\sigma_{old}}{\sqrt{n}}$

previous part were based on discriptive statistics and now we are moving to infrential statistics

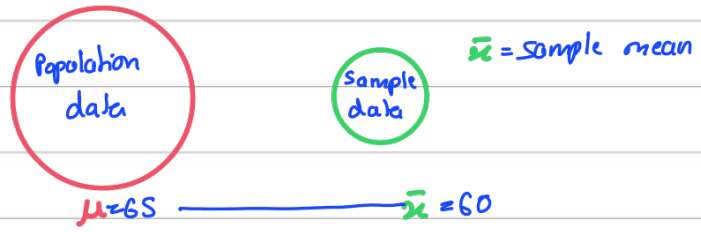
Infrential Statistics: In this type of statistics we make diff infrences/conclusions from a sample data wrt finding any population parameter.

Ty {Estimate}: It is a specified observed numerical value used to estimate an unknown population parameter.

Types of Estimates:-

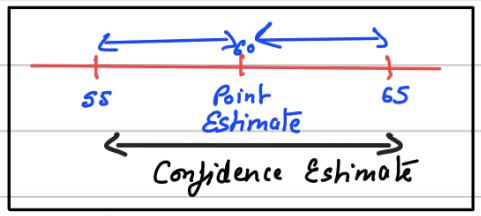
① **Point Estimate:** single numerical value used to estimate an unknown population parameter.

Ex:- Sample mean is a point estimate of a population data.



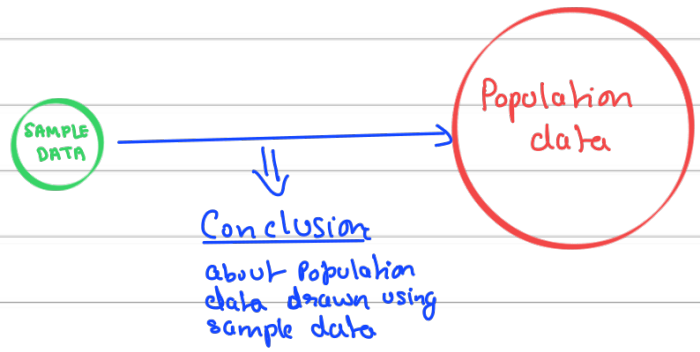
② **Interval Estimate:** Range of values used to estimate the unknown population parameter.

when we have a pt estimate along with have a range which we get by adding & substracting margin of error and this is c/a "Confidence Interval"



an this kind of interval is know as an **Interval Estimation** made for a sample mean

III {Hypothesis Testing & Mechanism}



these conclusion or inferences are drawn using "Hypothesis Testing"

Hypothesis Testing Mechanism:- # suppose person committed crime

Step ① Null Hypothesis (H_0) :

- it contains the assumption you are begining with

acc^n to H_0, court say "Incent until proven guilty"

Step ② Alternate Hypothesis (H_1)

- it is nothing but opposite of null hypothesis

acc^n to H_1, he is guilty

Step ③ Experiments : Statistical Analysis

at this stage various (DNA/Fringerprint test are conducted)

Step ④ Accept or Reject the Null hypothesis

Q) A College at district A states its average pass %age of students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass %age of 90% with a std deviation of 4%. Does this college have a diff. passed %age.

Sol^n → 1st step defining null hypothesis (H_0):

$H_0 = \mu$ (avg pass %age) = 85%

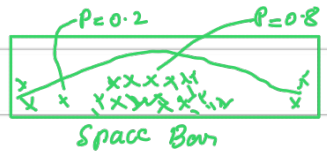
2nd step defining alternate Hypothesis (H_1):-

$H_1 = \mu \neq 85\%$

III) P-VALUE

It is a number, calculated from a stastical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide weather to reject the null hypothesis.

Simple Example: below is given a space bar where the x represents the place where the space bar has been touched :-



at region where P=0.2, it means at that region space bar has been touched 20 times out of 100 touches

region where P=0.8, that means keyboard's been touched 80 times out of 100.

Q) Do Hypothesis testing on a coin toss to determine weather the coin is fair or not

P(H) = 0.5 , P(T) = 0.5 ⇒ fair coin

P(H) = 0.6 , P(T) = 0.4 ⇒ somewhat fair

P(H) = 0.7 , P(T) = 0.3 ⇒ Toss biased towards Heads

Soln — ① H_1 : Coin is fair

② H_0 : Not a fair coin

③ Experiment : tossing for 100 times

IV Hypothesis Testing and Statistical Analysis

There are various types of Statistical Analysis.

① Z-test } - deal with avg. data } < Z-table, t table
② t-test }

③ CHI-Square test } - deals with Categorical data

④ ANNOVA test } - deals with variance.

1) Z-test :-

Requirements for using Z-test :-

① must know population std

② Sample size (n) ≥ 30

Q1) Avg. height of all the residents in a city is 168cm with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individuals and found the avg. height to be 169.5 cm

a) State null & alternate hypothesis

b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

Sol^n - population mean → 168 cm with std. (σ) = 3.9

Sample size (n) = 36 & sample mean ($\bar{x}$) = 169.5 cm

As,

population mean is given & $n > 30$ ∴ we will be using Z-test

S1) a) Null Hypothesis (H_0) - $\mu = 168$ cm

(H_1) = $\mu \neq 168$ cm — $\boxed{\text{2 tailed test}}$

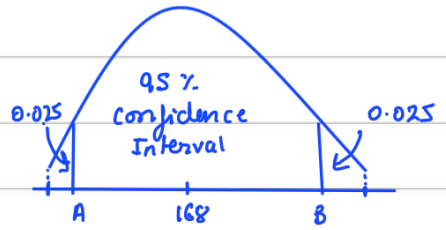
coz μ can be > 168 or $\mu < 168$ cm

S2) b) given confidence interval : 95%

∴, $\boxed{\text{Significance value : } 1 - CI}$ $= 1 - \frac{95}{100} = 1 - 0.95 = \underline{\underline{0.05}}$

Now,

S3) Based of C.I we will draw Decision Boundry



finding Z-score for B

$1-0.025 = 0.9750$ (+ve)

referring z-table we get

$\boxed{Z(B) = +1.96}$

finding z-score for A

"

"

$\boxed{Z(A) = -1.96}$

5V
$\boxed{0.025}$

Now

If new z-score is < -1.96 or $> +1.96$ then we must reject the Null Hypothesis (H_0):

S4) Applying Z-test

2.31

0.1044

0.01044

- for population data z-score $= \bar{x} - \mu / \sigma$
- for sample data z-score (Z_d) $= \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

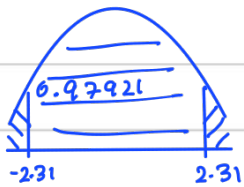
$$\boxed{Z_d = 2.31}$$

S5) Conclusion: based on above Z_d score we can conclude that as Z_d (2.31) > 1.96 ∴, we can reject the Null Hypothesis

$\Rightarrow \mu \neq 168$ and since $Z_d = 2.31 = +ve$ ∴ avg height $\neq 168$ seems to be ting i.e > 168.

Now,

New decision boundry is :-



using z-table area left of $-2.31 = \underline{.1044}$

∴ area on right of $2.31 = .1044$

new P value $= .1044 + .1044 = \underline{0.02088}$

as P value (new) < 0.05 ∴ we reject Null Hypothesis

Finding the confidence interval:

↳ $1 - 0.02088 = 0.97921 \approx 97.921\%$

Qy A factory Manufactures bulbs with a avg warranty of 5 yrs with std of 0.50. A worker believes that the bulb will malfunction in less than 5 yrs. He tests a sample of 40 bulbs and find the avg. time to be 4.8 yrs

a) State Null & Alternate Hypothesis

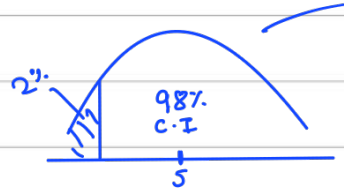
b) at 2% significance level, is there enough evidance to support the idea that warranty should be revised?

Soln. → S-1y – $H_0 \Rightarrow \mu = 5$ yrs

S2y → $H_1 \Rightarrow \mu < 5$ yrs # checking only <5yr & not ≠5 ∴ it is a $\boxed{\text{1 tailed Test}}$

S3y Decision Boundary

only 2% on left side cuz its a 1 tailed test



S4y Z-test:-

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = -2.53$$

for $Z_d = -2.53$ area under the curve is: 0.00570 P-value

So;

Now as P-value (0.00570) < Significance Value (0.02)

∴ reject the Null Hypothesis

In Z-Statistics when we perform any analysis using Z-score we require population standard deviation (σ)

But,

What happen if population std(σ) is not given? How can we analyse the sample data then?

In such cases we use <u>Student's t distribution</u>

wrt this distribution

$$\boxed{t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad , \; s \rightarrow \text{sample standard deviation}}$$

t-table is used for Student's t distribution analysis which has degree of freedom (df)

$\boxed{df = n-1}$, n = no. of sample data

Why df = n-1

Suppose 3 chairs are empty inside a room & 3 ppl looking for chair, 1st person to enter the room will have 3 choices to sit, 2nd person will have 2 choices & no choice will be available of 3rd person and he has to sit on a chair which is left empty

So, with 3 ppl only 2 choices are availabe

$\therefore$ DF = n-1

2) t-stats : t-test :- 1 sample t-test

Q1) In a population the avg IQ is 100. A team of researchers want to test a new medication to see if it has either a +ve or -ve effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence?

Sol^n → $\mu = 100$

Sample, $n = 30$, $\bar{x} = 140$, $s = 20$ – sample std

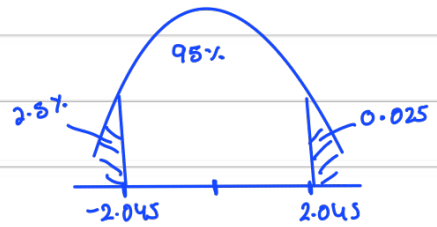
☆ Let's initialize our Confidence Interval (C.I) = 95% and Significance value (α) = 0.05 ≈ 5%

S1) $H_0 \rightarrow \mu = 100$

S2) $H_1 \rightarrow \mu \neq 100$ (∴, 2 tailed test)

S3) Computing Degree of Freedom for t-test: Dof = 30-1 = 29

S4) Decision Rule:-



Using t-table at Dof=29 & α(two tail) = 0.05

$\boxed{\text{t-value} = 2.045}$

☆ if → (t test < -2.045 & t test > 2.045) → reject Null Hypothesis.

S5) t-test:-

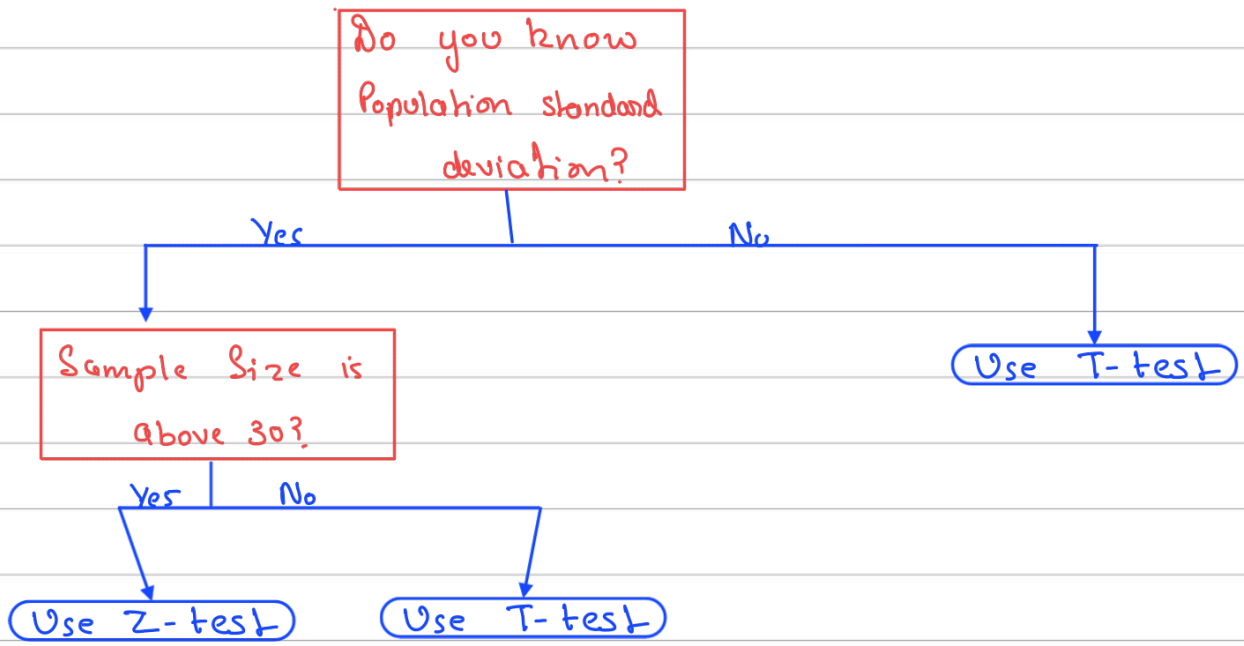
$$t = \frac{\bar{x} - \mu}{s\sqrt{n}} = \frac{140 - 100}{20\sqrt{30}} = \underline{\underline{10.96}}$$

Since,

t-test = 10.96 > 2.045 ∴, Rejecting the null Hypothesis

S6) Conclusion:- → Medication used has affected the intellegence
→ also as t-test value = 10.96 > 2.045 & +ve ∴ intellegence has increased.

Q4 When to use T-test Vs Z-test



11/03/23

Type-1 and Type-2 Error

Reality: Null Hypothesis is True or Null Hypothesis is False
Decision: Null Hypothesis is True or Null Hypothesis is False

Outcome 1:- We reject the Null Hypothesis when in reality it is false → Good Scenario

Outcome 2:- We reject the Null Hypothesis when in reality it is True → Type 1 Error

Outcome 3:- We retain the Null Hypothesis when in reality it is false → Type 2 Error

Outcome 4:- We retain the Null Hypothesis when in reality it is True → Good Scenario

{Confidence Interval & Margin of Error}

Till now we had point estimate i.e sample mean($\bar{x}$) and we were trying to make some assumptions about the population mean μ.

Q1 What is the problem with using point estimate?

Sol^n → Suppose now pt. estimate $\bar{x}$ is used to estimate the population mean (μ)

$$\boxed{\bar{x}} \longrightarrow \boxed{\mu}$$

if $\bar{x}=2.5$ $\mu=3$

But depending on diff. samples the $\bar{x}$ value can be greater than or less tha or equals to $\mu=3$.

So; in order to avoid this,

We define a range of Confidence interval which is used to correctly define the population μ.

Confidence Interval:-

$$\boxed{\begin{aligned} C.I &= \text{Point Estimate} \pm \text{Margin of Error} \\ &= \bar{x} \pm \underbrace{\boxed{Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ (for z-test)}}}_{\text{Margin Error (for z-test)}} \\ &= \bar{x} + \underbrace{\boxed{t_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ (for t-test)}}}_{\text{margin error (for t test)}} \end{aligned}}$$

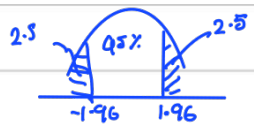
, α → Significance value

Q1 In the verbal section of CAT exam, the std is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean?

Sol^n → we know $CI = \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

we know for 95% confidence Index $\alpha = 5\% \approx \underline{0.05}$

∴,



$Z_{\alpha/2} = 1.96$ & -1.96

Lower CI $= 520 - (1.96)\frac{100}{\sqrt{25}} = 480.8$

Higher CI $= 520 + (1.96)\frac{100}{\sqrt{25}} = 559.2$

So, with 95% assurance we can say that our final mean CAT score will be lying b/w 480.8 and 559.2.

Bayes Statistics (Bayes Theorem)

Bayes statistics is an approach to data analysis and parameter estimation based on bayes' theorem.

Baye's Theorem:- # completely based on probabilities

Probability —— Independent Events / Dependent Events

Independent Event

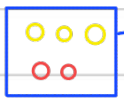
ex: Rolling a dice
$\{1, 2, 3, 4, 5, 6\}$
$P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}$ .....

ex: Tossing a coin
$P(H) = 0.5$ & $P(T) = 0.5$

Probabilities does not dependent of other outcomes.

Dependent Events:

ex: A Box of marbel with 3 yellow marbles and 2 red marbles



Red → $P(R) = \frac{2}{5}$ $\xrightarrow[\text{this event}]{\text{after}}$ $P(Y) = \frac{3}{4}$

$\therefore$

$$\boxed{P(R \,\&\, Y) = P(R) \times \boxed{P(Y/R)}} \quad \text{—(1)}$$

Conditional Probability

$= \frac{2}{5} \times \frac{3}{4}$

$= 6/20$

Deriving Baye's Theorem:-

from above eq ① we know $P(A \& B) = P(A) \times \boxed{P(B/A)}$

↳ Probability of B after event A already happened

as, $P(A \& B) = P(B \& A)$

$\Rightarrow P(A) \times P(B/A) = P(B) \times P(A/B)$

5★

$$\Rightarrow \boxed{P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}}$$

Baye's Theorem

also $\boxed{P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}}$

where, $A, B \rightarrow$ Events

$P(A/B) \rightarrow$ Probability of A given B is true

$P(B/A) \rightarrow$ Probability of B given A is true

$P(A)$ & $P(B) \rightarrow$ Independent probabilities of A & B.

Qy Use Baye's Theorem to predict the output price of the house from the bellow data set

Data set:-

Size of House x_1	No of Rooms x_2	Location x_3	Price y

features — outcome

Sol^n → using Baye's theorem to predict price:-

$$\boxed{P(y/x_1 x_2 x_3) = \frac{P(y) \times P(x_1 x_2 x_3 / y)}{P(x_1 x_2 x_3)}}$$

12/03/23

CHI SQUARE TEST

- The "Chi Square Tesh for Goodness of fit" test claims about population proportions.
- It is a non-parametric test that is performed on the categorical data (Nominal or Ordinal)

Ex:→ There is a population of male who likes different colour bikes. and there is a theory that 1/3 ppl like yellow, 1/3 like Red & 1/3 like orange.

	Theory	Sample
Yellow bike	1/3	22
Red bike	1/3	17
Orange bike	1/3	59

Theory Categorical distribution → observed categorical distribution

Now,
using Goodness of fit Chi^2 test we will confirm that weather the sample supports the theory or not

Goodness of fit test:-

Qy In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of the ppl are left handed.

Soln →

	O(observed)	E(expected: 11% are lefty)
left handed	11	9 (11% of 75)
Right handed	64	66
	75	75

Now Chi^2 test is performed to make sure that weather this observed value is Goodness of fit for the expected values.

In Chi^2 test we compare 2 categorical variable i.e Theoritical /expected and Observed Categorical distribution (Sample).

Performing Chi^2 Test:-

Q1) In 2010 census of the city, the weight of the individual in a smal city were found to be the following:

<50kg	50-75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are the results

<50kg	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population differences of weights changed in the last 10 years?

Soln

Expected data (2010)

<50kg	50-75	>75
20%	30%	50%

Sample / Observed data (2020)

<50kg	50-75	>75
140	160	200

Expected data (2020)

<50kg	50-75	>75
20% of 500 = 100	30% of 500 = 150	50% of 500 = 250

S1) Null Hypothesis (H_0): The data meets the expectation

S2) Alternate Hypothesis (H_1): The data does not meet the expectation

S3) as $\alpha = 0.05$ ∴, Confidence Interval (CI) = 95%

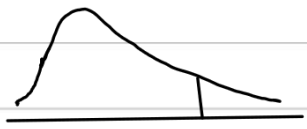
S4) Degree of Freedom:-

$$df = K-1 = 3-1 = 2$$

S5) Decision Boundary:

☆ In case of Chi^2 test we don't have a symmetrical distribution.

distribution looks like:



•) using CHI^2 table find critical value for $\alpha = 0.05$

$\therefore$ CV = 5.991 (using χ table)

If χ^2 is greater than 5.991, Reject H_0

If χ^2 is less than 5.991, fail to reject H_0

Sol Calculating χ^2 test statistics :-

Expected data (2020)

<50kg	50-75	>75
20% of 500 = 100	30% of 500 = 150	50% of 500 = 250

Sample / Observed data (2020)

<50kg	50-75	>75
140	160	200

$$\boxed{\chi^2 = \sum \frac{(O-E)^2}{E}}$$

$$\Rightarrow \chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250} =$$

$$= 26 + \frac{10}{15}$$

$$= \frac{390+10}{15} = \frac{400}{15} = \frac{80}{3} = 26.66$$

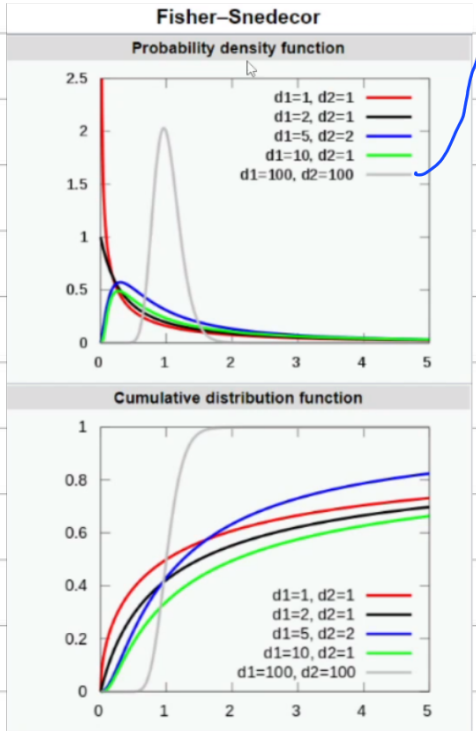
as $26.66 > 5.91$ $\therefore$ we reject H_0 (null hypothesis)

Sol Conclusion:- The weights of 2020 population are different than those expected in the 2010 population.

CHI SQUARE TEST for PYTHON

F- DISTRIBUTION

In probability theory and statistics F-distribution or f-ratio, also known as Sendecor's F-distribution or the Fisher-Sendecor distribution (after Ronald Fisher and George W. Sendecor) is a continous probability distribution that arises frequently as the null distribution of a statistic, most notably in the analysis of variance (ANOVA) and other F-Tests.



☆ as d1 & d2 increases the distribution becomes on starts to look like normal distribution

$$f(x;d_1,d_2) = \frac{\sqrt{\frac{(d_1x)^{d_1}\, d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

Beta function

$$\mathrm{B}(m,n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!} = \frac{m+n}{mn} \Big/ \binom{m+n}{m}.$$

☆ Parameter used : $d_1, d_2 > 0$: degree of freedom

☆ Supports : $x \in (0, +\infty)$

☆ Pdf = $f(x, d_1, d_2) = \sqrt{\dfrac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}} \Big/ x\,B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$ — Beta function

☆ F-distribution with d_1 and d_2 degree of freedom is the distribution of Random variable such that:

$$X = \frac{S_1/d_1}{S_2/d_2}$$

, d_1 & d_2 → Respective dof for S_1 & S_2

, S_1 & S_2 → Independent random variables with Chi Square distribution

Chi Sq distribution is a Right skewed distribution

F-test [Variance Ratio Test] :-

Q) The following data shows no. of bulbs produced daily for some days by 2 workers A and B

A	B
40	34
30	38
38	41
41	33
38	32
35	34
	40
	34

Can we consider based on the data that worker B is more sizable and efficient ($\alpha = 0.05$)

Soln S1) $H_0 = \sigma_1^2 = \sigma_2^2$

S2) $H_1 = \sigma_1^2 \neq \sigma_2^2$

S3) Calculation of varience

$$s_1^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} =$$

Worker A

x_1	$\bar{x}_1$	$(x_1-\bar{x}_1)^2$
40	37	9
30	"	44
38	"	1
41	"	16
38	"	1
35	"	4
$\bar{x}_1 = 37$		$\Sigma = 80$

$$s_1^2 = \frac{80}{n-1} = \frac{80}{8} = \boxed{13}$$

Worker B

x_2	$\bar{x}_2$	$(x_2-\bar{x}_2)^2$
34	37	9
38	"	1
41	"	16
33	"	16
32	"	25
34	"	4
40	"	9
34	"	9
$\bar{x}_2 = 37$		$\Sigma = 84$

$$s_2^2 = \frac{84}{7} = \boxed{12}$$

S4) F-test Calculation (i.e variance Ratio test)

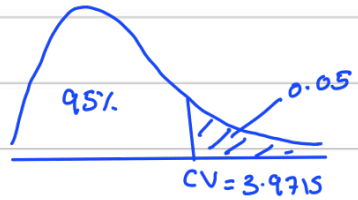
$$\boxed{F = \frac{S_1^2}{S_2^2}} = \frac{13}{12} = 1.33$$

S5) Decision Boundry

$$\left[\begin{array}{l} df_1 = 6-1 = 5 \quad \& \quad df_2 = 8-1 = 7 \\ \alpha = 0.05 \end{array}\right]$$ — # Using this to refer f-table

$\Downarrow$

$CV = 3.9715$



$\therefore$, If f-tes is $> 3.9715 \Rightarrow$ reject H_0.

f-test is $< 3.9715 \Rightarrow$ accept H_0

So,

as $1.33 < 3.9715$ $\therefore$ accepting Null Hypothesis

S6) Conclusion :- Worker B is not more stable or efficient when compared to worker A.

14/03

ANALYSIS OF VARIANCE (ANNOVA)

Annova is a statistical method used to compare the means of 2 or more groups

Components of Annova:-

① Factors (variable) → Independant / → dependent

② Levels

ex:- medicine (factors)

(dosage) {5 mg 10 mg 15 mg} → Levels

ex:→ Mode of Payment (factor)

G-Pay, Phone pe, Imps, NEFT } (Levels)

Assumptions of Annova:-

① **Normality of Sampling Distribution of Mean:-**

The distribution of sample mean is normally distributed.

② **Absence of Outliers:-**

Outlying score needs to be removed from the dataset

ex: using boxplot etc.

③ **Homogenity of Varience:-**

Population varience in different levels of each independent variable are equal.

$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$

④ **Samples are independent and random:**

Types of Annova:-

① **One Way Annova** :- One factor with atleast 2 levels, these levels are independent

ex: Doctor want to test a new medication to decrease headache. They split the participants in 3 conditions/groups based on no. of dosage of medication given to them [10mg, 20mg, 30mg]. Doctor ask participants to rate the headache from [1-10]

here,

factor → medication

level → 10 mg, 20 mg, 30 mg (independent levels)

10 mg	20 mg	30 mg
5	7	2
3	4	6
⋮	⋮	⋮

② **Repeated Measures Annova** :- One factor annova with atleast 2 levels, levels are dependent.

ex: factor → running

levels → Day 1, Day 2, Day 3 (dependent levels)

Day 1	Day 2	Day 3
8	5	4
7	4	9
⋮	⋮	⋮

③ **Factorial Annova** :- 2 or more factors each of which with at least 2 levels, levels can be independent and dependent.

ex: factors: Running, Gender

levels: Day 1, Day 2, Day 3 | Male, Female

Day 1	Day 2	Day 3
8	5	4
9	4	3
2	4	2
7	8	9

Hypothesis Testing in Annova (Partitioning of Variance in Annova)

Step 1) Define Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 \ldots\ldots$ (compaining mean of 2 or more groups)

Step 2) Alternate Hypothesis H_1: atleast one of the mean is not equal

★ To perform this we use F-Test:

$$\boxed{F = \frac{\text{Variance b/w samples}}{\text{Variance within Samples}}}$$

Q1) Explain Variance b/w & Variance within Sample?

Solⁿ →

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\bar{x}_1 = 3$	$\bar{x}_2 = 19/5$	$\bar{x}_3 = 4$

Variance b/w Sample:- is the difference b/w the variances of x_1, x_2, x_3 (i.e $\sigma_1, \sigma_2, \sigma_3$)

Varience within Sample: How much the variace of a sample differs from its mean.

One Way Annova

one factor with atleast 2 independent levels

Qy Doctors want to test a new medication which reduces headaches. They splits the participants into 3 condition [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headaches b/w [1-10]. Are there any differences b/w the 3 conditions using alpha = 0.05?

15mg	30mg	45mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

Soln → Step1) $H_0 = \mu_{15} = \mu_{30} = \mu_{45}$

H_1 = not all μ are equal

Step2) Significance value 0.05 ⇒ CI = 95%

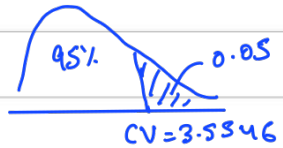
Step3) Calculate dof: total samples $N = 7 \times 3 = 21$

dof (b/w samples) $= a - 1 = 3 - 1 = \boxed{2}$ df1

dof (within sample) $= N - a = 21 - 3 = \boxed{18}$ df2

dof (total) $= N - 1 = 20$

Step4) Construct Decision Boundry (F test ⇒ Righ skewed)



if F. is > 3.5546 → reject H_0

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

- $F = \dfrac{\text{Variance b/w Sample}}{\text{Variance within Sample}}$
- SS – Sum of Squares
- MS → mean square

	SS	dof	MS	F
b/w				
within				
Total				

① $$\boxed{SS(b/w) = \frac{\Sigma(\Sigma a_i)^2}{n} - \frac{T^2}{N}}$$

Σ 15 mg : $9+8+7+8+4+8 = 57$

Σ 30 mg : . . . . . $= 47$

Σ 45 mg : . . . . . $= 21$

$\Sigma(\Sigma a_i)^2 = (57)^2 + (47)^2 + (21)^2 =$

$T^2 = (57)^2 + (47)^2 + (21)^2 =$

$$\Rightarrow SS(b/w) = \frac{57^2 + (47)^2 + (21)^2}{7} - \frac{[(57)^2 + (47)^2 + (21)^2]}{21}$$

$= 98.67$

② $$\boxed{SS(within) = \Sigma Y^2 - \frac{\Sigma(\Sigma a_i)^2}{n}}$$

$\Sigma Y^2 =$ sum of sq of every value of all the samples

$= 853$

$$SS(within) = 853 - \frac{[(57)^2 + (47)^2 + (21)^2]}{7} = 10.29$$

	SS	dof	$MS = \frac{SS}{dof}$	F
b/w	98.67	3-1 = 2	49.34	
within	10.29	21-3=18	0.54	
Total	108.96	20		

Now,

$$\boxed{F\text{-test} = \frac{MS(b/w)}{MS(within)}} = \frac{49.34}{0.54} = 86.56$$

Since, $F > 3.5546$ $\therefore$ we reject H_0