

Statistics

Basic

① Introduction to Statistics.

② Types of Statistics.

(i) Descriptive stats.

(ii) Inferential stats.

③ Types of data.

(i) Quantitative.

① Discrete.

② Continuous.

(ii) Qualitative.

① Nominal.

② Ordinal.

④ Scale of measurement

(i) Nominal scale data.

(ii) Ordinal scale data.

(iii) Interval scale data.

(iv) Ratio scale data.

⑤ Measure of Central tendency.

(i) Mean:

$$\text{Population mean} = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Sample mean} = \frac{\sum_{i=1}^n x_i}{n}$$

↑
Population size
↓
Sample size

(ii) Median

① N is odd:

$$\boxed{\frac{N+1}{2}}$$

② N is even: ~~$\frac{N}{2} + \frac{(N+1)}{2}$~~

③ N is even:

$$\boxed{\frac{\left(\frac{N}{2}\right)^{th} + \left[\left(\frac{N}{2}\right) + 1\right]^{th}}{2}}$$

(iii) Mode:

Element with Max Frequency.

⑥ Variable vs Random Variable.

⑦ Set

- (i) Union
- (ii) Intersection \cap
- (iii) Difference
- (iv) Subset \subset
- (v) Superset \supset

⑧ Covariance and Correlation

To find relationship between 2 variables.

(i)

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

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

disadv: No limit to range

(ii) Pearson Correlation Coefficient:
range $\Rightarrow [-1 \text{ to } +1]$

$$\text{Correlation, } r_{x,y} = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y}$$

$\sigma \Rightarrow$ standard deviation

$\text{Cov} = \text{Covariance}$

(iii) Spearman Rank Correlation
Focus on Rank than Value.

$$\text{Correlation, } r_s = \frac{\text{Cov}(R(x), R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}}$$

$R(x) \Rightarrow$ Rank of x

⑨ Measure of dispersion

(i) Variance : gives idea about spread of data.

① Population variance

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

$N \rightarrow$ population size.

$\mu \rightarrow$ population mean.

② Sample variance

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

$\bar{x} \rightarrow$ sample mean

$n \rightarrow$ sample size

* $(n-1)$ because Bessel correction

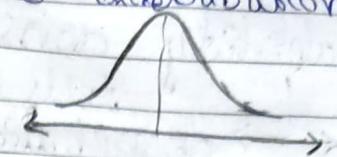
(iii) Standard deviation:

* standard deviation,

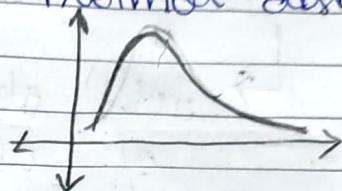
$$\sigma = \sqrt{\text{Variance}}$$

⑩ Skewness

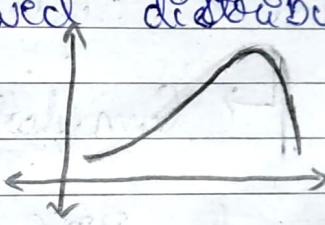
(i) Symmetric distribution.



(ii) Right skewed distribution.
[log-normal distribution]



(iii) Left skewed distribution.



⑪ Histograms.

* KDE
help to → & smoothen the histogram → we get PDF

Advance - I

① Probability density Function [PDF]

↳ tells about distribution of continuous data.

② Probability Mass Function [PMF]

↳ tells about distribution of discrete data.

③ Cumulative distribution Function [CDF]

↳ cumulative sum

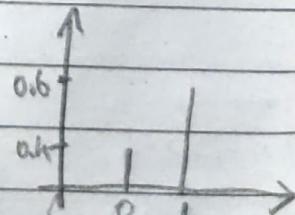
* ~~slope of PDF gives the g.~~

* Point in PDF indicates the gradient descent of point in CDF.

④ Bernoulli distribution

↳ PMF

↳ only 2 outcome



$$q = 1 - p$$

$$\text{PMF} \Rightarrow P(x) = p^x q^{1-x} = p^x (1-p)^{1-x}$$

Mean, $E(K) = \sum_{i=1}^K k_i \cdot p(x_i)$

Median, $\begin{cases} 0 & \text{if } p < \frac{1}{2} \\ 0 \text{ or } 1 & ; \\ 1 & \text{if } p = \frac{1}{2} \\ ; & \text{if } p > \frac{1}{2} \end{cases}$

Variance, $\sigma^2 = pq$

standard deviation $\Rightarrow \sigma = \sqrt{pq}$

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

⑤ Poisson distribution.

PMF

Number of events occurring in a fixed time intervals

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

λ = Expected number of events to occur at every time interval

increases \rightarrow shape of distribution
Mean, $E(x) = \mu = \sigma^2 t$ become taller

Variance,

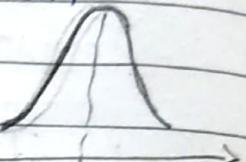
$$E(x) = \mu = \sigma^2 t$$

⑥ Normal / Gaussian distribution

→ PDF

→ Empirical rule

$$68 - 95 - 99.7\% \\ (\mu + \sigma) \quad (\mu + 2\sigma) \quad (\mu + 3\sigma)$$



→ Bell curve



⑦ Q-Q plots

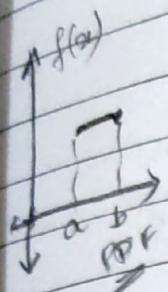
↳ to check whether data is normally distributed (or) not

⑧ Uniform distribution

(i) Continuous uniform distribution

PDF \rightarrow describes an experiment where there is an arbitrary outcome in certain range

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

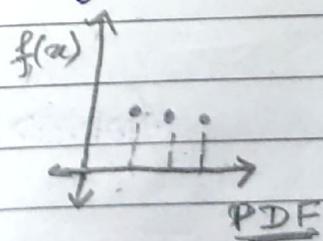


$$\text{CDF} = \begin{cases} 0, & \text{for } x < a \\ \frac{x-a}{b-a}, & \text{for } x \in [a, b] \\ 1, & \text{for } x > b \end{cases}$$

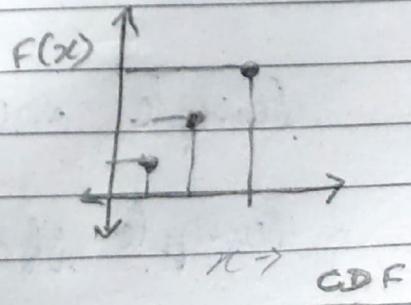
(ii) Discrete uniform distribution.

PMF \rightarrow symmetrical probability distribution where in finite no. of values are equally likely to occur.

$$\text{PMF} = \frac{1}{m}$$

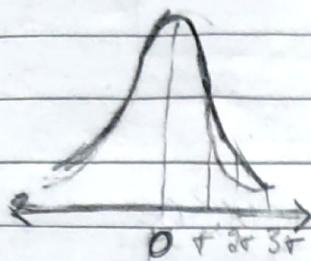


$$\text{Mean } \bar{x}, \text{ Median } \frac{a+b}{2}, \quad b \geq a$$



⑨ Standard normal distribution

→ Normal distribution with mean = 0 and std = 1



? Use

* Standardization

* To find how many std a point is away from mean

Z score,

$$Z = \frac{x - \mu}{\sigma}$$

⑩ Central limit theorem

→ Probability of sampling distribution of mean

of population follows gaussian / normal distribution

then

for $n > 30$, sampling distribution also follow normal distribution.

Sample mean, $\mu_{\bar{x}} = \mu$ μ = population mean

Sample std $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ σ = population std

Inferential statistics

① Estimate

↳ to estimate an unknown population parameter

(i) Point estimate.

(ii) Interval estimate.

② Hypothesis testing

(i) Null hypothesis [H_0] → ^{initial} assumption

(ii) Alternate hypothesis [H_1]

* significance value (α) $\Rightarrow [0, 1]$

* confidence interval, $C I = \bar{x} - d$

*	Z test	Average
	t test	
	chi-square	→ Categorical
	ANOVA	→ Parameter Mean/Varian
	F test	→ Variance.

③ Z-test

→ Assumptions Condition

① Population std. [+]

② $n \geq 30$

→ Steps to solve

defence

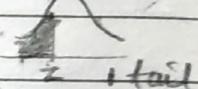
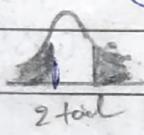
① H_0 & H_1

② decision boundary \Rightarrow One tail \rightarrow two tail



③ Find Z value $Z = \frac{x - \mu}{\sigma}$

④ Find p (area)



→ from Z-table

with given ' α ' & found ' Z '

⑤ Output \rightarrow

$$P < \alpha$$

True

[reject Null hypothesis]

False

[fail to reject H_0]

④ Student t-test

→ Use when population std. is

not known [+] (0.01)

$n < 30$

Z-test

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{\text{S.E.}/\sqrt{n}}$$

~~2~~ \Rightarrow sample mean

~~3~~ \Rightarrow sample variance

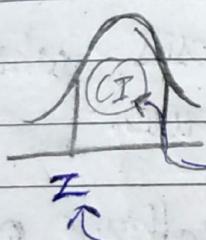
\hookrightarrow find using (d & dof)

$$\text{degree of freedom} = n - 1$$

* steps to solve

① Define hypothesis

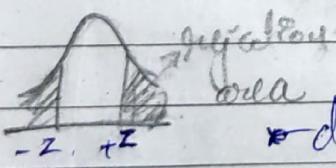
② d = _____



③ degree of freedom = _____

④ Decision boundary rule

⑤ Finding Z as core \Rightarrow
 \hookrightarrow using (d and dof) in
t table



* define decision rule (Condition)

ex: $t > +Z \quad / \quad t < -Z$

⑥ Calculate t using formula

⑦ check (Condition) decision rule

* -

$$P(x) = {}^n C_x p^x q^{n-x}$$

$$\text{Mean } \mu = np$$

$$E(X) = np \quad \text{Variance } \sigma^2 = npq$$

⑤ Binomial distribution

→ extension of binomial

→ No of success in fixed no independent trials

* CDF \rightarrow Discrete $\Rightarrow \Sigma$
 \rightarrow Continuous $\Rightarrow \int$

Advances - 2

① Types of errors

(i) Type 1: rejecting Null hypothesis when
in reality is true.

(ii) Type 2: accepting Null hypothesis when
in reality is false.

② Margin of Error.

* Confidence interval = point estimate \pm margin of error

$$C.I. \equiv \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Z-score

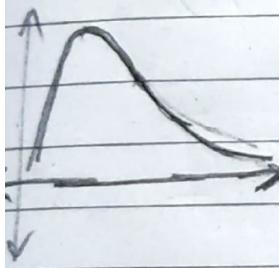
③ Bayes theorem.

→ independent & dependent variables.

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

$P(B/A) \rightarrow$ Probability of B when A has happened

④ Chi-square test



→ "chi-square test for goodness of fit"

for - Categorical data

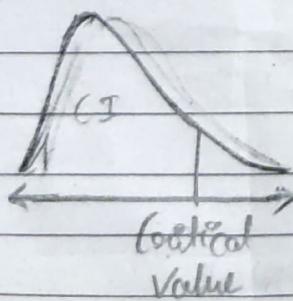
→ steps to solve

① Define H_0 and H_1 .

② $\alpha = \dots$, CI = $1 - \alpha$

③ degree of freedom = $n - 1$

④ Decision boundary



- Find critical value using (α) and (df) using Chi-square table.

$\chi^2 > CV$
reject

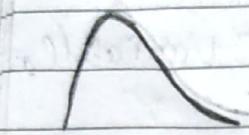
(define decision rule)

⑤ Calculate chi-square-test value

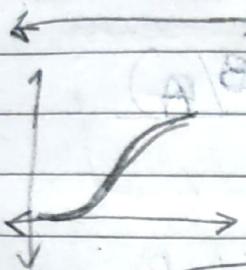
$$\chi^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

⑥ Conclusion

F-test



Continuous random variable
↳ Probability distribution.



To analysis of variance
b/w 2 groups

Parameters: d_1 & d_2

$$\text{PDF, } f(x; d_1, d_2) = \frac{(d_1 x)^{d_1} (d_2 x)^{d_2}}{(d_1 x + d_2)^{d_1+d_2}} x \beta(d_1/2, d_2/2)$$

Beta function, $\beta(m, n) = \frac{(m-1)! \cdot (n-1)!}{(m+n-1)!}$

$$x = \frac{\alpha_1/d_1}{\alpha_2/d_2}$$

* ~~steps~~ to solve

① Define H_0 and H_1

② Calculate variance

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

group 1 σ^2

group 2 σ^2

③

Calculation of Variance ratio test
(F-test)

$$F = \frac{s_1^2}{s_2^2}$$

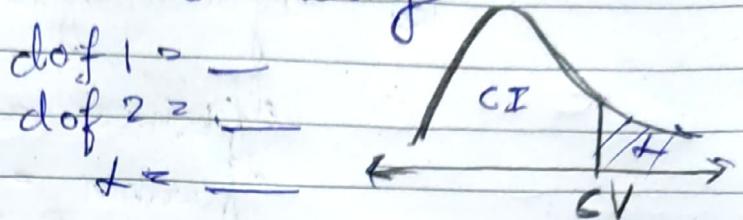
④

Decision ~~stat~~ boundary

dof 1 = _____

dof 2 = _____

F = _____



* Find CV using. (dof1, dof2, F)

In F-test.

* define decision rule

⑤

Case

$$F < CV$$

True
[reject H_0]

False
[Fail to
reject H_0]

⑥

~~ANOVA~~ ANOVA

→ To compare means of 2 groups

① Factors / variables

② levels.

* Assumptions

*

$$F = \frac{\text{Variance between sample}}{\text{Variance within sample}}$$

* steps to solve

$$\mu_1 = \mu_2 = \mu$$

① Define μ_1 and μ_2

② ~~σ^2~~ $\tau = \frac{\sigma^2}{N}$, C.I. $= 1 - \tau =$

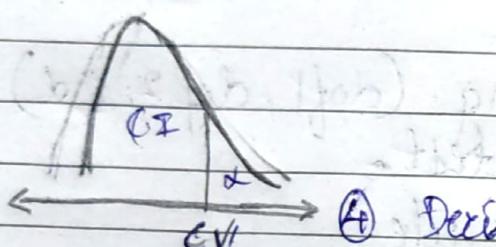
③ degree of freedom

$$df_{\text{between}} = a - 1$$

$$df_{\text{within}} = N - a$$

Population size

$$df_{\text{total}} = N - 1$$



④ Decision boundary

* Find CV using (df_1, df_2, α)
in F-table

* define decision rule. $F > CV$

⑤ Calculate F test stats

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

Between	Sum of square [SS]	degree of freedom	Mean square [MS]
Within	SS between	df1	$\frac{SS}{df1}$
	SS within	df2	
Total			

a = sum of value
each level

y = sum of square
of value of any
one level

$$SS_{\text{between}} = \sum \frac{(Ea_i)^2}{n} - \frac{T^2}{N}$$

$$SS_{\text{within}} = \sum y^2 - \frac{\sum (Ea_i)^2}{n}$$

⑥ Conclusion

~~10~~

Statistics

Week - 10

Basic Statistics

Defn:

Statistics is the science of collecting, organizing and analyzing the data
→ decision-making process

* Data: "facts or pieces of information."

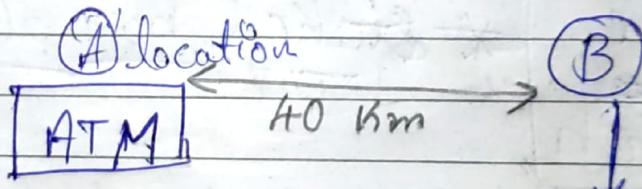
Ex: Heights:

{ 175 cm, 180 cm, --- }

IQ

{ 200, 300, --- }

Cas:



bank to be opened (or) not

Types of Stats

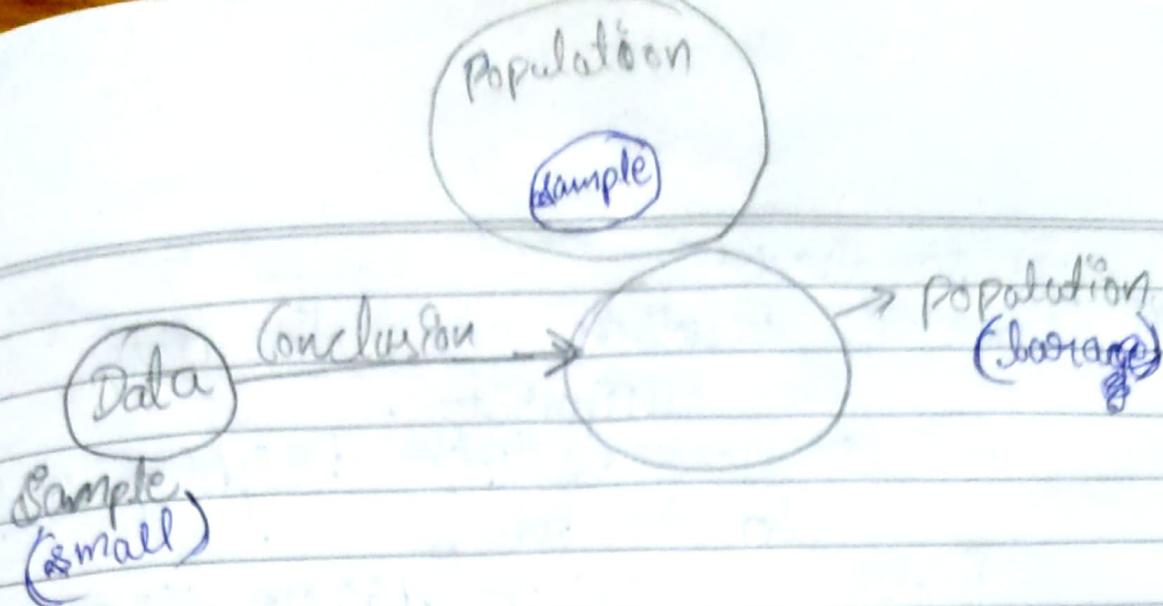
① Descriptive stats

def: organizing and summarizing data

② Inferential

stats

def: It consists of using data you have measured to form conclusion



① Measure of Central Tendency
[Mean, Median, Mode]

② Measure of Dispersion

[Variance, std]

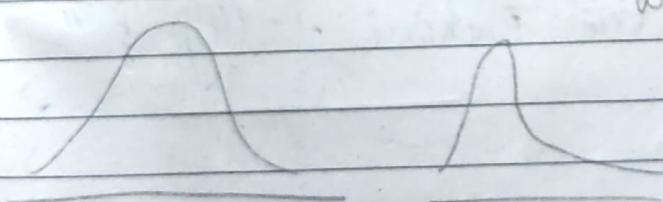
how our data
is well spread

used to
understand how
data is
distributed.

③ Histograms, Bar chart, Pie chart

summarizes the data

- what is distribution
relative obs - -



different graph

Ⓐ For Conclusion, we need to do
Experiments like →

- ① Z-test
- ② t-test

Hypothesis testing, p-value,
significance,

Ex: Let's say there are 50 students in maths class in university.

We have collected height of student in the class.

[175 cm, 180 cm, 160 cm, 140 cm, 130 cm,

⇒ Descriptive question

↳ "What is the average height of the students in the class?"
↳ "What is common height of the students?"

⇒ Inferrential Question

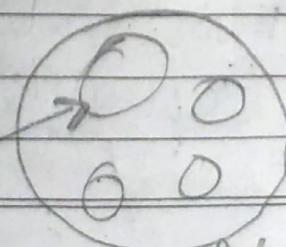
* "Are the height of students in the classroom similar to what you expect in the entire college?"
↑
population

Ex: Sample data and population data

Elect pool of election

Conditions are drawn from sample data

sample few people voted



State [Population]

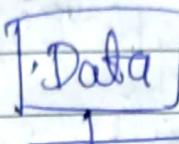
Types of data

EDA Step 1

col^o

DMC	R.W.F	Region
3.4	1	0
4.1	2	0
5	0	0
2.1		0

Numerical data



Categorical data

Quantitative

Qualitative

Discrete

Continuous

Nominal

Ordinal

whole number

[with specific range]

Number of book借 books

books

2.1 ✓

Any value decimal: -

Ex: weight, speed, height

175, 76.1

26.9, 61.5

Ex: Gender
[M/F]

Blood group
[A, B, AB]

* Cannot

Say

M > F ✗

gender
cannot
rank

Ex: Good ①

Bad ②

Average ③

[Review]

Which day holiday needed?

Assignment

→ ① what kind of variable marital status is?

→ Nominal. [Qualitative]

② what kind of variable Nile river length is?

→ Continuous. [Quantitative]

③ what kind of variable movie duration?

→ Continuous. [Quantitative]

Scale of measurement of data →

helps to determine the types

①	Nominal	Scale	Data.
②	ordinal	Scale	Data.
③	Interval	Scale	Data.
A	Ratio	Scale	Data.

of statistical operation
can be performed.

① Nominal Scale Data:

- * Qualitative / Categorical Variable
- * ex: gender, colors, labels.
- * ~~order~~ order does not matter.

~~ex:~~ Survey to select colors

Red	→ 5	→ 50%
Blue	→ 3	→ 30%
Yellow	→ 2	→ 20%

- * Focus more on count and distribution side.

② Ordinal Scale Data:

Can calculate difference

- * Ranking and order matters.

- * Difference cannot be measured.

Normal data
be convert
into
ordinal
scale data.

Qualification:

PHD	1 st
BE	2 nd
Masters	3 rd
B.Com	4 th
B.Sc.	5 th

Rank

100°	1 st
90	2 nd
85	4 th
70	3 rd
30	5 th

cannot consider
as ordinal
scale of
marks data

(not rank)
cannot find
difference

can be considered as
ordinal scale of data.

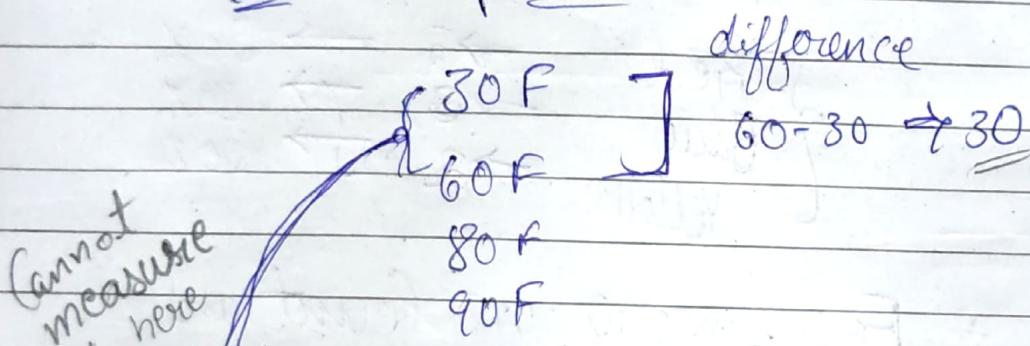
Ques
③

Interval scale data

- * Rank & order matter.
- * Difference can be measured.
- * Ratio cannot be measured.
- * Does not have any '0' starting value.
→ Can have any value

Ex: Temperature

may be $-10 \rightarrow -3F$
 $+10 \rightarrow 6F$



cannot measure here

$$\text{Ratio} = \frac{60}{30} = 2$$

It will not
be

so ratio does not
matter.

too correctly
feel heat falling on 60F
feel heat falling on 30F
double the heat
which is wrong

④ Ratio Scale data:

- * Order & rank matter.
- * Difference and ratios are measurable
- * It Does have '0' starting point.

Sample
Ques: Grades

100
90
60
45
50.
35

<u>order</u> ✓	<u>difference</u> ✓	<u>rank</u> ✓
100	10	1st
90	30	2nd
60	10	3rd
50	5	4th
45	10	5th
35		6th

✓
Ratio

$$\frac{100}{50} = \underline{\underline{2:1}}$$

↳ 100 marks is double the 50 marks. ✓
Thus is correct.

marks always starts with '0' with is lowest. ✓

Ques

① length of different rivers in the world? \Rightarrow Ratio Scale data

② Marital status \Rightarrow Nominal Scale data

③ IQ measurement? \Rightarrow Ratio Scale data.

Measures of Central tendency

① Mean . ② Median. ③ Mode.

[Imp for EDA and Feature Engineering]

* Population (N)

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

Sample (n)



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

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

$$\mu = \frac{1+2+2+2+3}{10} =$$

$$= \frac{12}{10}$$

$$= \underline{\underline{3.0}}$$

④ Median (.)

* ~~odd~~, 2, 2, 3, 4, 5
 even
 Q1 1, 5, 2, 3, 2, 1

~~even~~

1; 2, 2, 3, 4, 5
 average $= \frac{2+3}{2} = \underline{\underline{2.5}}$

$$\therefore \text{median} = \underline{\underline{2.5}}$$

* {1, 2, 2, 3, 4, 5, 7} odd

$\therefore \text{median} = \underline{\underline{3}}$

* why median? Even after mean is present.

① $\Rightarrow \text{Mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

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

median = 3 ✓

1, 2, 3, 4, 5

② \Rightarrow but what if outlier is present
in sample

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

a number that does not belong to distribution that doesn't look similar to sample

mean = $\frac{1+2+3+4+5+100}{6} = \frac{115}{6}$

≈ 19.17 ≈ 19

median = 1, 2, 3, 4, 5, 100

median = $\frac{7}{2} = \underline{\underline{3.5}}$ ✓

in ① & ②, ~~even~~
outlier present

∴ mean is very different in both.
but even on presence of outlier
median is deviated a little.

③ Mode \Rightarrow (Frequency)
 \rightarrow Max Frequency

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

mode = 1 occurred 3 times

where mode is used

Nominal \Rightarrow Types of flower | Age

daffy

10

Rose

3

Rose

5

Sunflower

—

Rose

8

handle

\rightarrow

X drop now

but loss of

data

④ Replace with

mode value i.e. Rose

if outlier replace
with median
else with mean.

* Practical

import numpy as np

① np.mean (~~list~~)

numpy ② np.median (~~list~~)

scipy - ③ stats.mode (.dist)

mode.mode[0]

from scipy import stats

Random variables

* Variables & RV are different

$$x+y=7$$

$x+y=10$ \rightarrow Here x & y are just variables.

* defn: Random variable is a process of mapping the output of a random process / experiment to a number.

Ex ① Tossing a coin.

\rightarrow This is a random process can get output anything

② Rolling a dice.

③ Measure the temperature for the next day

$X = \begin{cases} 0, & \text{if Head} \\ 1, & \text{if Tail} \end{cases}$

always Capital
output to a number

$Y = \{ \text{Sum of the rolling of dice } 7 \text{ times} \}$

$\{ A, 5, 6, 1, 2, 2 \} = \frac{\text{sum}}{20}$

————— = other value

{ } { }

* A random variable can take any outcome based on case / process but variable take only one value fixed

Ex^b
 $P(Y \geq 15) \quad P(\text{out} = \text{"Head"})$

Sets

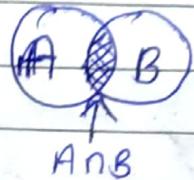
$$A = \{1, 2, 3, 4, 5, 6\}$$

$$B = \{3, 5, 8, 9\}$$

① Intersection:

Common

$$A \cap B = \{3, 5\}$$



② Unions: all



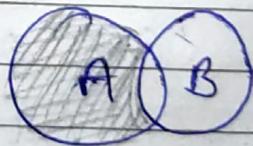
$$A \cup B = \{1, 2, 3, 4, 5, 6, 8, 9\}$$

③ Difference

~~B~~ present in set 1 not
in set 2

$$A - B = \{1, 2, 4, 6\}$$

$$B = \{3, 5\}$$



④ Subset

$$\times \quad A \subset B \quad [A \text{ is subset of } B]$$

→ False, be 'B' don't have all the elements of 'A'

$$\checkmark \quad B \subset A$$

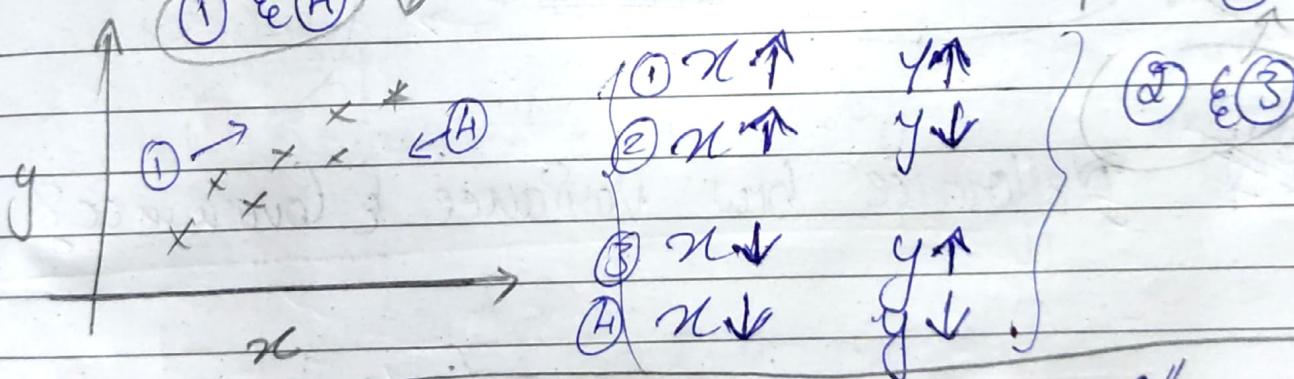
⑤ Superset

✓ $A \supseteq B$ [A is a superset of B]
A contain all elements of B.

① Covariance and Correlation:

X	Y
2	3
4	5
6	7
8	9

Can we find out relationship b/w X & Y?



Ex: ~~Size~~ ~~location~~ ~~price~~ ~~o/p~~

Determine
Features
are important
price !...
like ~~size~~ size↑ price↑

So, we need to apply Covariance.

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

(cont)

spread of the data

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

Checking Covariance

of 'x' itself

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

Ques

$$\text{Var}(x) = \underline{\text{Cov}(x, x)}$$

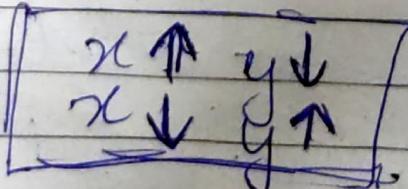
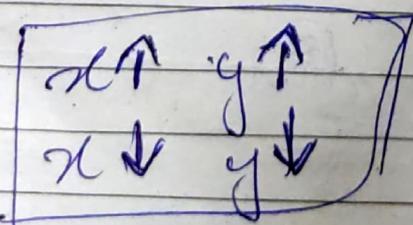
Helps 'x'
tell its own
relationship.

Ques

Difference b/w variance & covariance?

Covariance means

+ve \rightarrow positive relationship



-ve

x	y
2	3
4	5
6	7

$\bar{x} = 4$ $\bar{y} = 5$

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

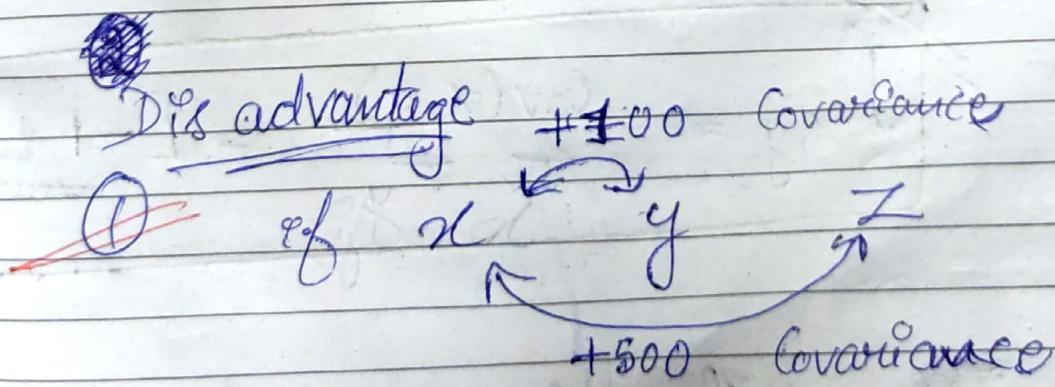
$$= \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{3-1} = \frac{4+0+4}{2}$$

$$\text{Cov}(x, y) = 4$$

Conclusion: x & y has +ve covariance

Advantages

① Relationship b/w x & y \rightarrow +ve



~~we cannot conclude that x is correlated to y than z . as dispersion / spread is less than z~~

\Rightarrow we need some restriction to this

① covariance does not have specific limit value.

To overcome this

we use other kind of correlation

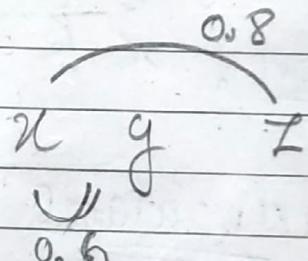


Pearson Correlation Coefficient:

Can take

any value $[-1 \text{ to } +1]$

orange blw



As value is getting restricted.

Now, we can conclude some.

$$\therefore \boxed{r_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}} = [-1 \text{ to } +1]$$

standard deviation.

② The more the value towards +1 the more correlated it is.

③ The more it is $-1 \rightarrow -V$
 $-1 \rightarrow -ve$ correlated.

Dataset

1000 features

[For ML models]

to build we don't need all features.

so, we can remove less correlated value.

price	size of the house	No of rooms	location	No of people staying

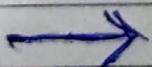
dependent features

independent features

Here, all features except ~~"No of people staying"~~ are required
∴ So, we do "feature selection."

↓ Correlation

~~No of people staying & price~~



near to 0

means not related so we will drop features

↑ won't focus on sig value
we focus of rank.

③ Spearman Rank Correlation

$$\rho_s = \frac{\text{Cov}(R(x), R(y))}{\sqrt{R(x)} \cdot \sqrt{R(y)}}$$

use

pl	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

* we have to try both ② and ③

* `import numpy as np`

`(covariance) > df.cov()`

`correlation > df. corr(method='__')`

① 'spearman'

② 'pearson'
(default)

Measure of dispersion

① Variance.

② Standard deviation.

① Variance

Population Variance

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

where, x_i = data points

μ = population mean

N = population size

Sample Variance

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

[Bessel's Correction]

why?

⇒ we also use $n-1$ rather than n so that the sample variance will be what is called an unbiased estimator of the population variance.

⇒ x_i = data point

\bar{x} = sample mean

n = sample size

Ex: Q1 {1, 2, 3, 4, 5} \Rightarrow Sample

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

x_i	$(x_i - \bar{x})^2$
1	4
2	1
3	0
4	1
5	4
$\Sigma = 10$	

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

$$\bar{x} = \underline{3.0}$$

$$\sigma^2 = \frac{10}{5-1} = \frac{10}{4}$$

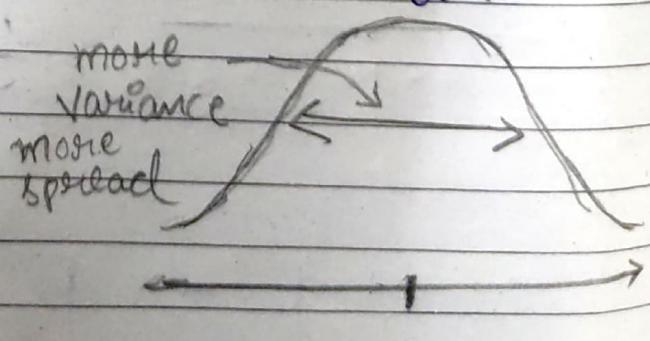
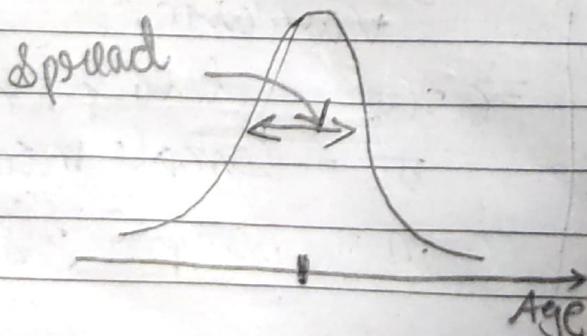
$$= \underline{2.5}$$

what let say?

\Rightarrow Variance talks about the spread of data.

Ex: $\sigma^2 = 2.5$ less spread

$\sigma^2 = 6.5$ more spread



② Standard deviation

population
standard
deviation

$$\sigma = \sqrt{\text{variance}}$$

~~E.D.~~

sample
standard
deviation

$$s = \sqrt{\text{sample variance}}$$

sample variance
ex $s^2 = 2.5$

$$\text{sample std} = \sqrt{s^2}$$

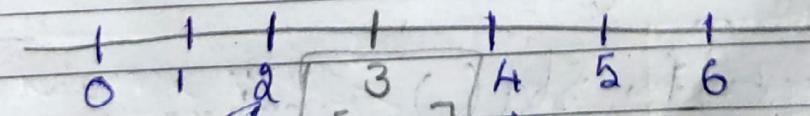
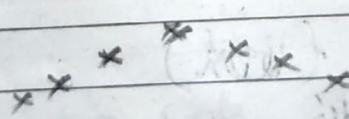
Consider
ex $\{1, 2, 3, 4, 5\}$

$$\mu = 3$$

$$\sigma = 1$$

Most of the data will
be falling in this 3

standard deviation

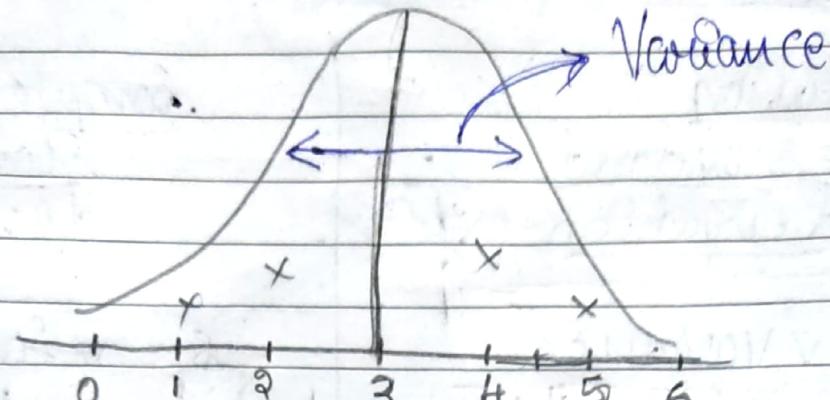


2 std
to left

[mean]

1st standard
deviation
to right

data
is
distributed
according
to this mean



$$\mu - 3\sigma \quad \mu - \sigma \quad \mu - \sigma \quad \mu \quad \mu + \sigma \quad \mu + 2\sigma \quad \mu + 3\sigma$$

$\frac{1}{2}$ std
to
left

5 → 2.5 std away
from mean

> array numpy 1D
> np.var (list)

Dataframe df. var() → default axis=0 → column
df. var (list) → axis=1 → row

> np.std (list)
(0.0)

df. std ()

Skewness

① Symmetrical distribution:

when we smoothen this histogram we get

Symmetrical distribution

→ No skewness here
in SD.

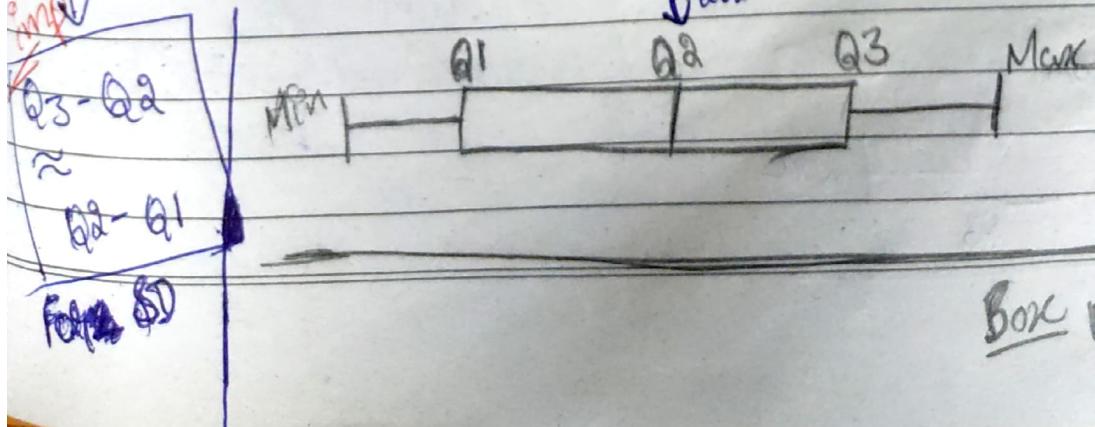
area towards left = area towards right

* The mean, median and mode all are perfectly at the center.

When we construct the box plot for SD we will not get outliers

(median)
↓ also

$\text{mean} = \text{median} = \text{mode}$



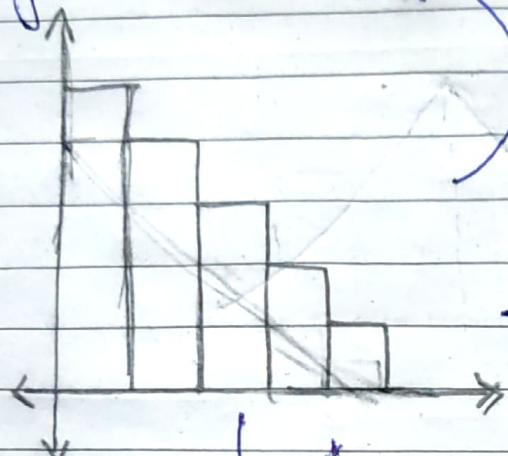
Box plot

Most of data exhibit unimodal distribution follow this distribution

bend towards right / +ve

Right [log Normal distribution]

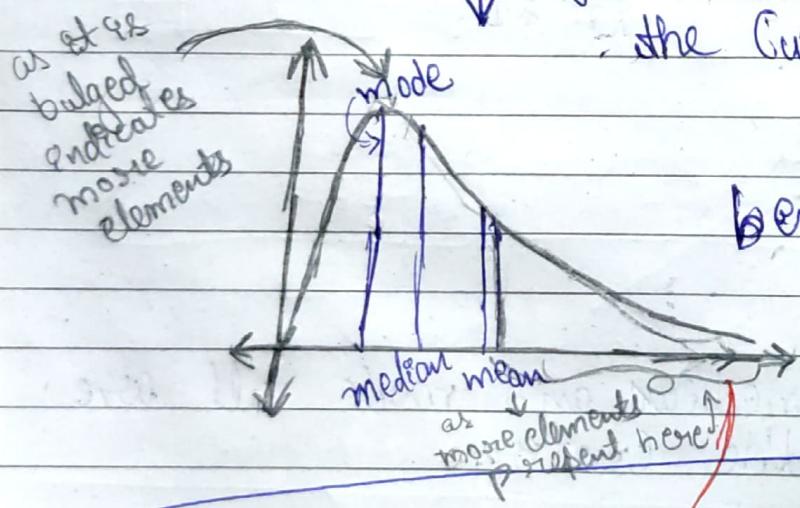
(2) Right skewed distribution:



(a) positive skewed

→ skewed towards right.

when ~~smoothen~~ smoothen the curve

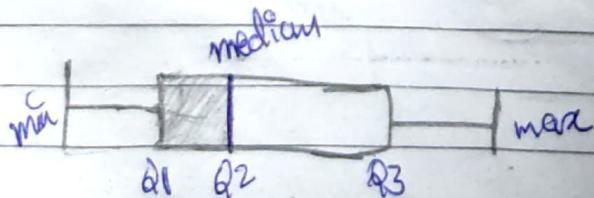


bend towards right so slight positive skewed.

when we draw box plot we get median somewhere towards left, ^{inter} mean > median > mode

$$Q_3 - Q_2 > Q_2 - Q_1$$

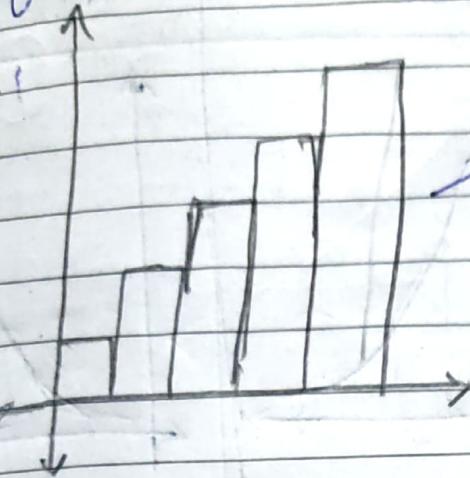
$$\text{mean} > \text{median} > \text{mode}$$



bend towards
left / -ve.

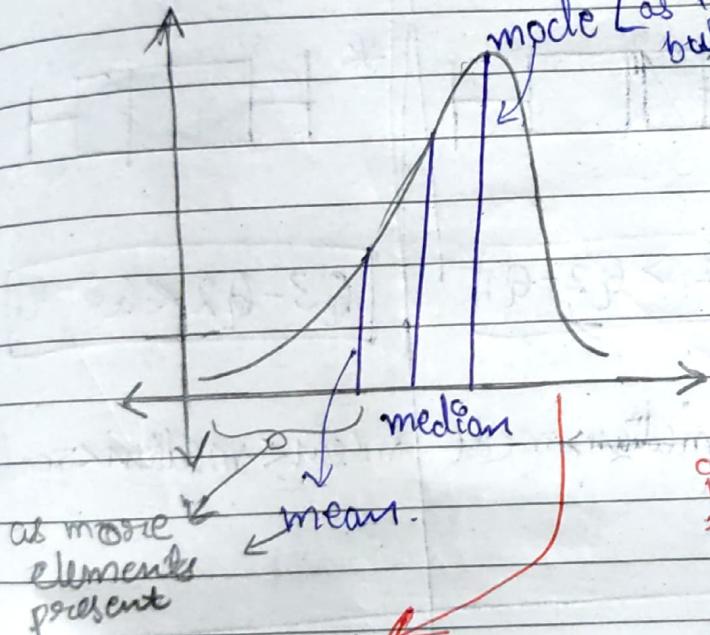
③ left

Skewed distribution:



skewed towards
left / -ve.

mode [as it
bulged]



bend towards
left so left / -ve
skewed.

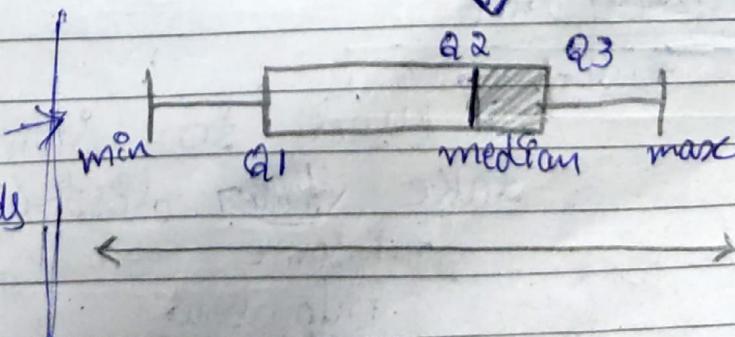
imp

$$Q_3 - Q_2 < Q_2 - Q_1$$

property

mean < median < mode

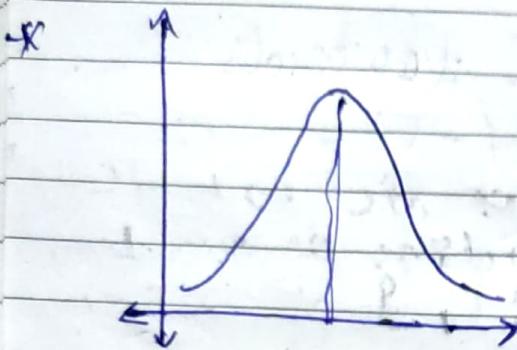
when we draw
box plot we
get median towards
right.



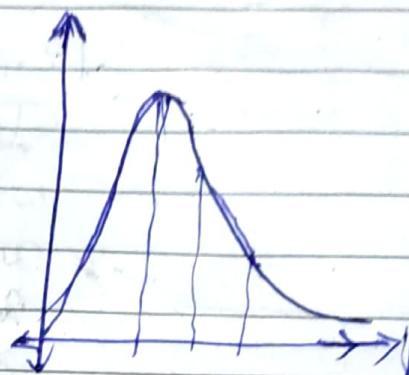
~~interview~~

E log normal distribution

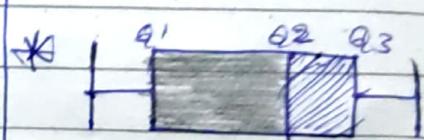
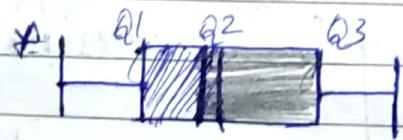
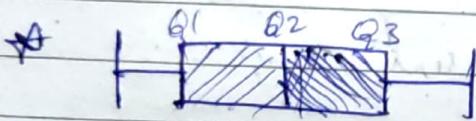
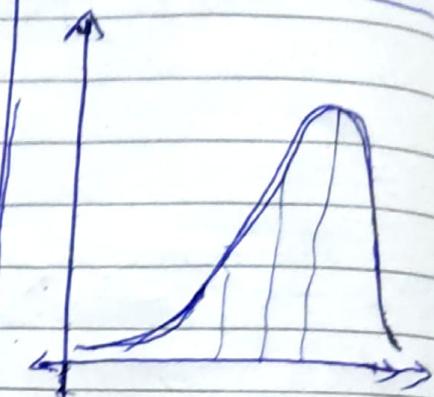
* Symmetric distribution



Right skewed distribution



Left skewed distribution



$$Q_3 - Q_2 \approx Q_2 - Q_1$$

$$Q_3 - Q_2 > Q_2 - Q_1$$

$$Q_3 - Q_2 < Q_2 - Q_1$$

* mean = median \Rightarrow mean $>$ median $>$ mode mean $<$ median $<$ mode

Histograms

→ used to summarize data and take ~~out~~ distribution of the data.
out some info about

Histogram \rightarrow Frequency of the data
 basically deals with

~~ex6~~ Consider.

- Ages = {10, 12, 14, 18, 21, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

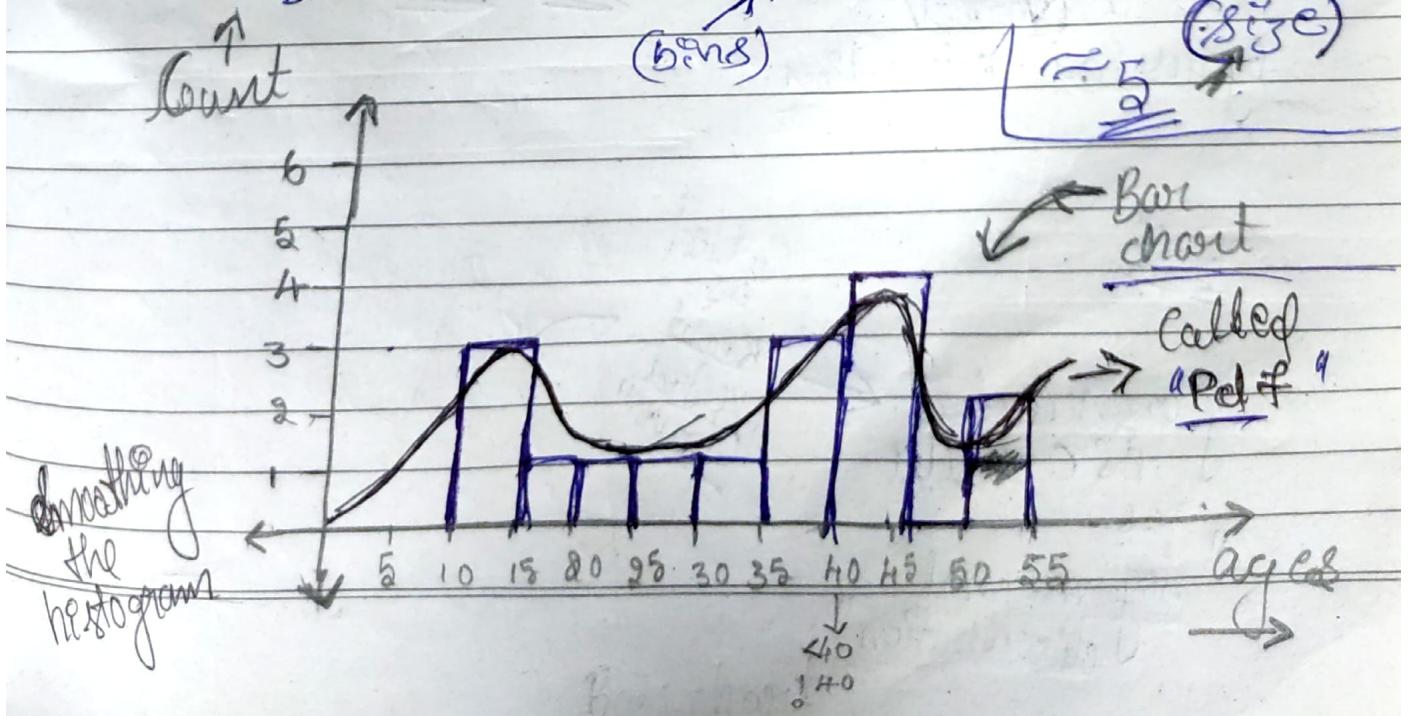
~~imp parameter~~ Bins = 10

we create 10 buckets according to our data and put data into that bucket

and plot a graph

~~Bucket size~~
 our data ranges from 10 to 51 and bins = 10 so we have to divide data into 10 buckets.

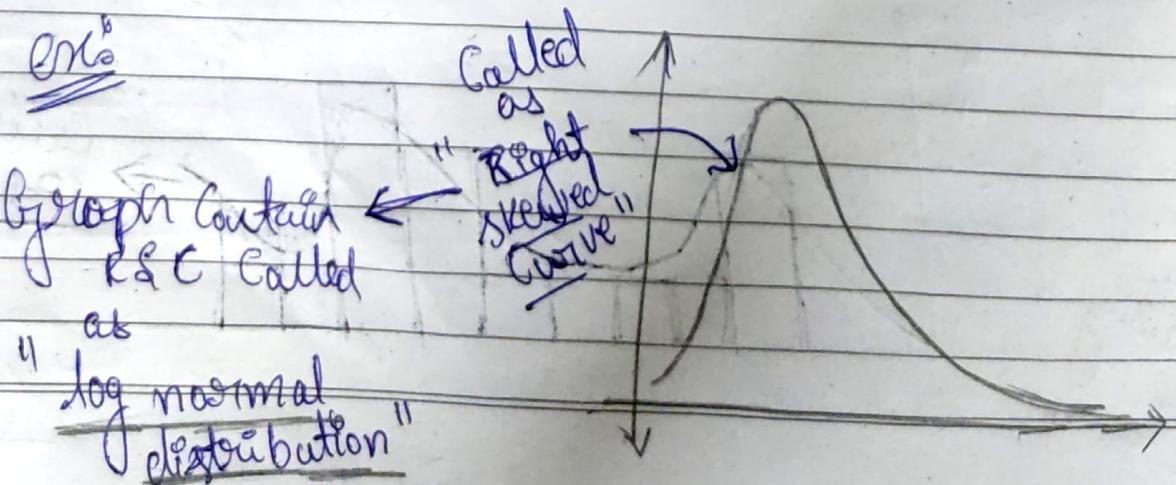
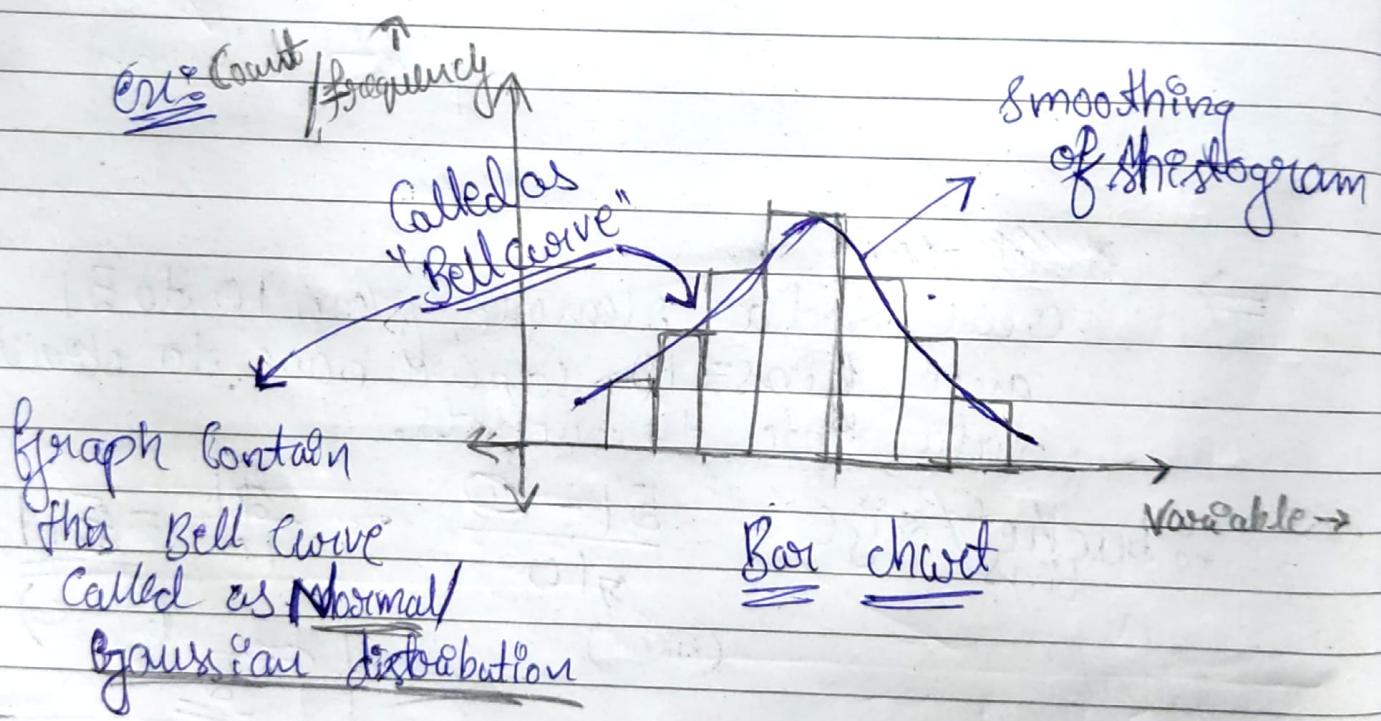
$$\therefore \text{bucket size} = \frac{\text{highest value}}{\text{bins}} = \frac{51}{10} = 5.1$$



How smoothing \rightarrow KDE (Kernel density estimator)
By smoothing the histogram we get PDF.

* PDF [Probability distribution function] tells how our data is distributed.

* KDE \rightarrow helps us do
smoothing of histogram $\xrightarrow{\text{we get}}$ PDF



* why call this?

* We have data to get information
assumptions we use these concepts.

* Variance → gives idea about spread of data
around median.

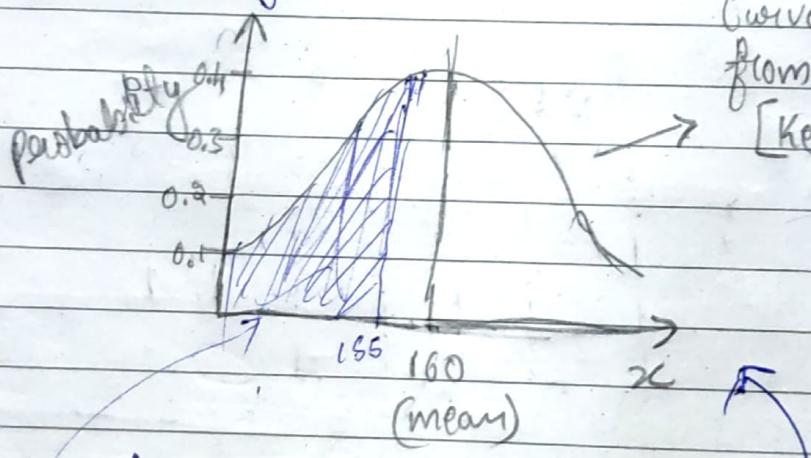
* Probability distribution Function / Density Function:

[PDF]

↳ both are not same

① Probability density function \rightarrow tells about the distribution of continuous data.

e.g. height of students.



Curve is obtained from 'KDE'
[Kernel density estimator]

using this distribution we
can derive imp information.

e.g. $P(x < 155)$,

↳ Can be calculated like area under curve

Probability distribution function contain various types.

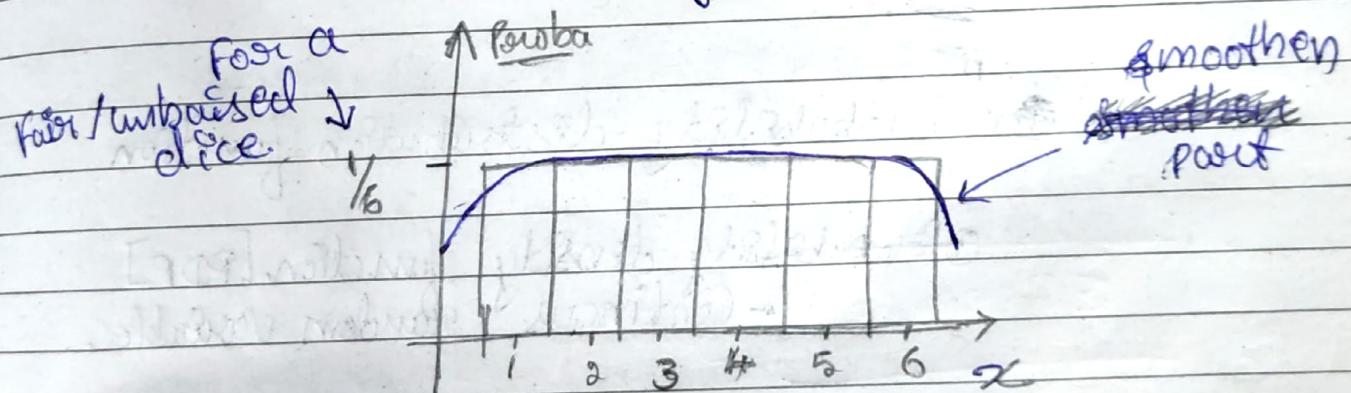
① Prob density function [PDF]

we try to create a curve ~~with~~ for distribution of "Continuous" data.

② Probability mass function [PMF]

"we see the distribution of a discrete random variable."

Eg: Rolling a die $\{1, 2, 3, 4, 5, 6\}$



$$\text{Q} \quad P(x \leq 4)$$

$$= P(x=1) + P(x=2) +$$

$$P(x=3) + P(x=4)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

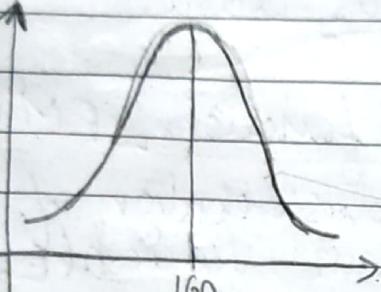
$$= \frac{4}{6} = \frac{2}{3}$$

③ Cumulative distribution function (CDF):

To understand CDF

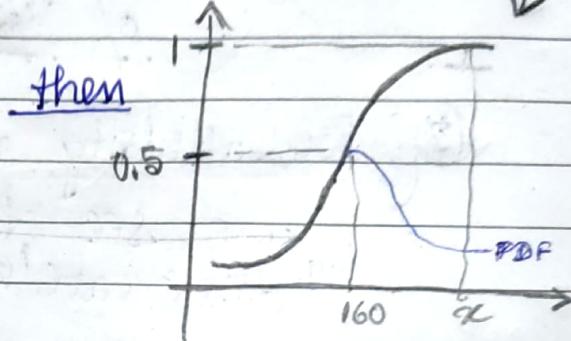
~~f(x)~~

PDF F_x



$$\text{area} = 0.5 + 0.5 \\ = 1$$

CDF



Cumulative sum
of all value

• Probability distribution function

① Probability density function [PDF]

- Continuous random variable.

② Probability mass function [PMF]

- Discrete random variable.

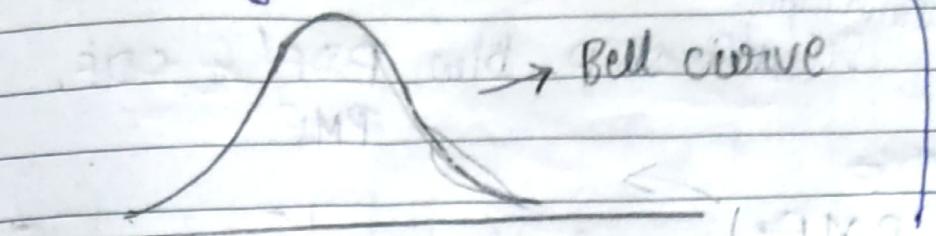
③ Cumulative distribution function [CDF]

- just doing cumulative sum.

~~Var~~ ~~MF~~

Types of probability distribution

① Normal (Gaussian) distribution. [PDF] Part of



we can do
many assumptions

Most datasets
follow this
distribution

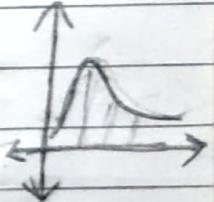
use
discrete
data

② Bernoulli distribution [PMF] Part of

↓
Outcome are only 2. → Success.
→ Failure.

③ Uniform distribution.

④ Log normal distribution [PDF] Part of



⑤ Poisson distribution [PMF]

⑥ Power law distribution [PDF]

[80-20% rule]

⑦ Binomial distribution [PMF]

use

* To convert to log normal to Gaussian.

* Some ml algo follow some distribution so convert

* Probability density function [PDF]

and

Probability mass function [PMF]

* Learn about

* Relationship b/w PDF & CDF.
PMF

~~PDF~~

① PMF:

* should have
Discrete random variable.

PMF

Ex: Rolling a dice

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

$$P(X=1) = \frac{1}{6}$$

$$P(X=2) = \frac{1}{6}$$

| |

$$\boxed{\sum_{i=1}^n P(X_i) = 1}$$

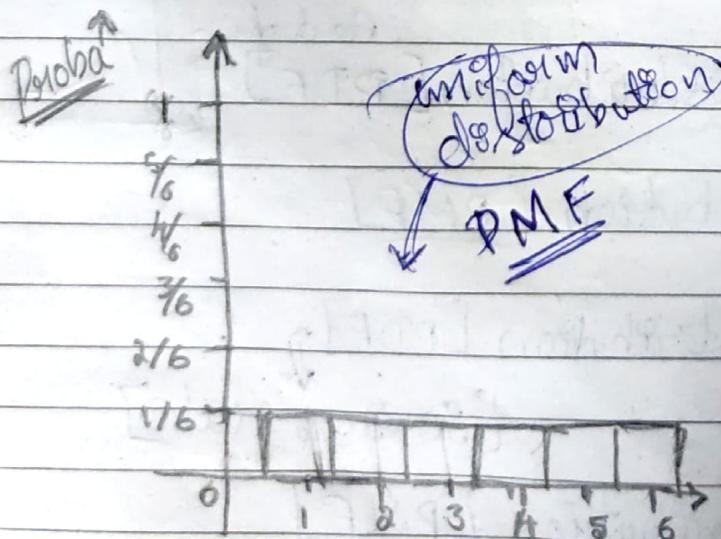
CDF

Ex: Rolling a dice

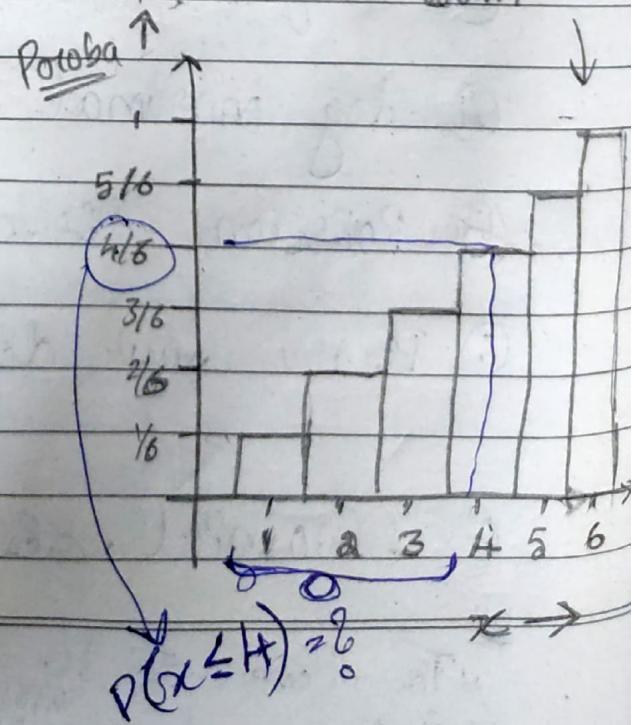
<u>X</u>	<u>P</u>	<u>Cumulative</u>
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$
6	$\frac{1}{6}$	$\frac{6}{6} = 1$

$$\boxed{\sum P(X_i) = 1}$$

Cumulative sum



$$P(X=2) = ? \quad x \rightarrow$$



$$P(X \leq 4) = ? \quad x \rightarrow$$

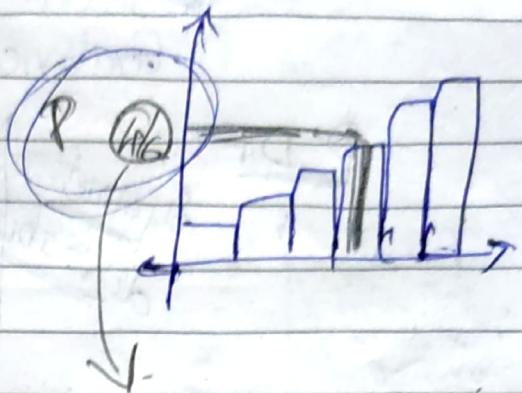
$$P(x=1) = \frac{1}{6}$$

easy
can
find.

but

$$P(x \leq 1) \rightarrow$$

so we
have to
look
at CDF



use CDF

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

* For different
different distribution we
can have different
formula.

* we took ex
of uniform
distribution
the prob of getting
a outcome is equal

(a) PDF:

↑
Probability
density

0.05 ↑

0.04 ↑

0.03 ↑

0.02 ↑

0.01 ↑

↓ 0

PDF

Region having
50% area
total

165 170 190 height

Symmetrical
distribution

CDF

Probability

0.5

0.1

0

165 170 190

height

* Consider point

Does ..

This point mean that ~~no~~
probability density ~~No~~
at 165 cm is 0.05.



↓
don't make
any sense.

* Then this
is the gradient
descent of
CDF

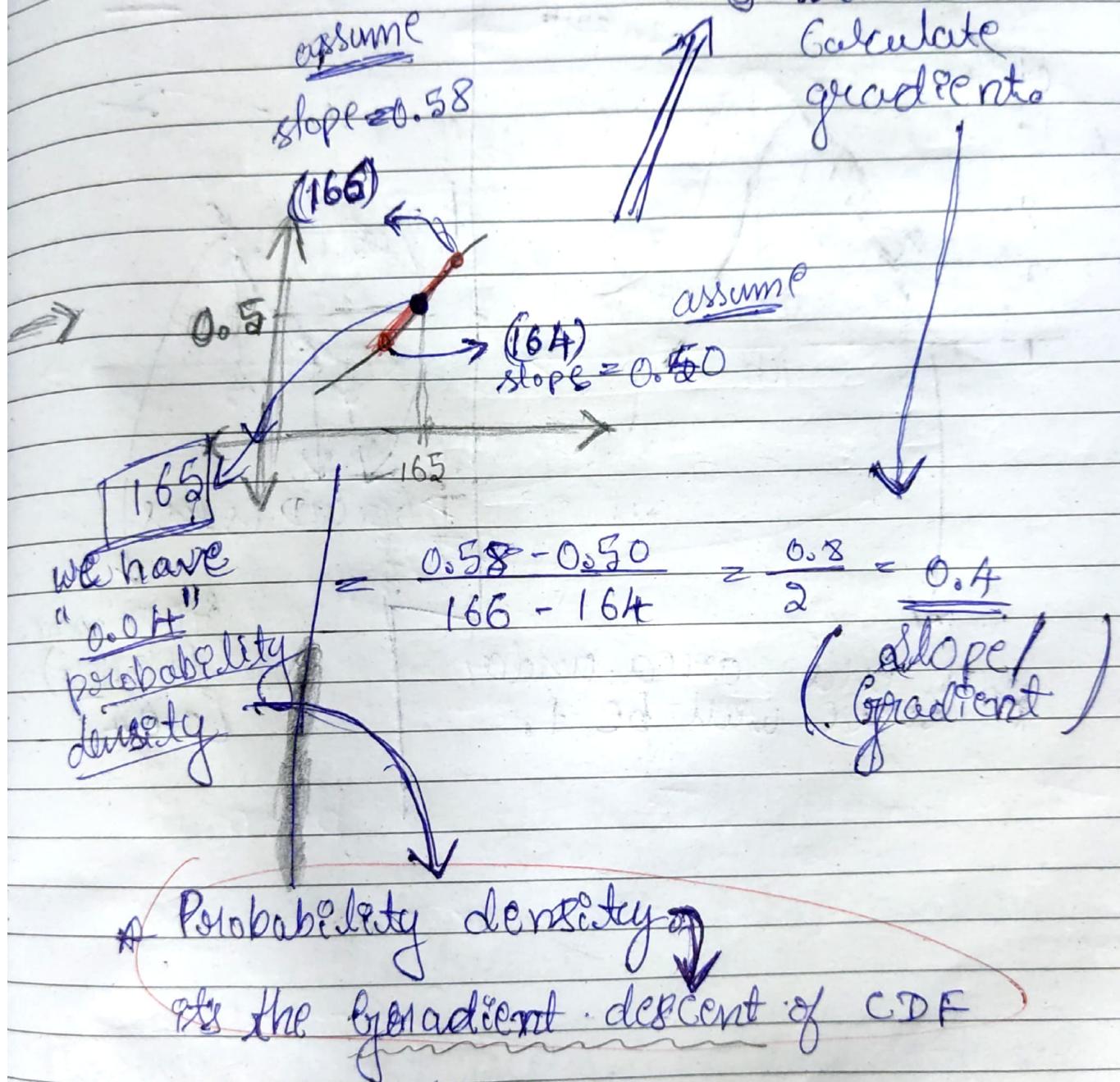
If it's not
the probability

* So to understand
this probability
density.

So further we
try to calculate
the gradient

[Slope] (or)
[Derivative]

By this data
we can
calculate
gradient.



* PDF : $f(x)$ represents state of change of probability within the interval

PMF : $P(X=x)$, not represent the a specific value.

PDF

* So

slope
decrease because
GD decreases
in CDF

gradient
descent [GD]

CDF

0.04

165

0.5

165

GD is less
GP is more.
(highest)
[0.04]

* at end area under
curve will be 1.

To remember
[Binary outcome]

* Bernoulli distribution [PMF]

* Named after "Swiss mathematician"
Jacob Bernoulli

* It is the discrete probability distribution of a random variable which takes the value 1 with probability p and value 0 with probability $q = 1 - p$.

$$\boxed{q = 1 - p}$$

* This ~~model~~ models Possible outcomes of an experiment \rightarrow Yes \rightarrow No

* Outcomes may be boolean values.

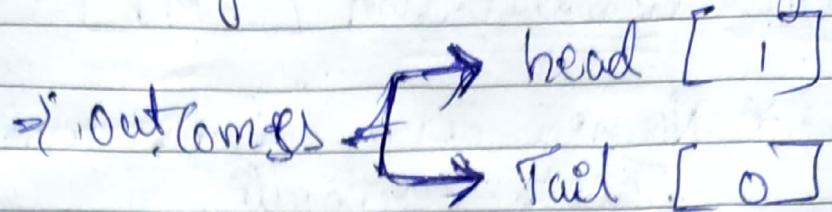
binary outcome

$p \rightarrow$ success / yes / true / one.

$q \rightarrow$ failure / no / false / zero.

Eg: ① Tossing a coin

May be



For Fair Coin

$$P(H) = 0.5 \Rightarrow P$$

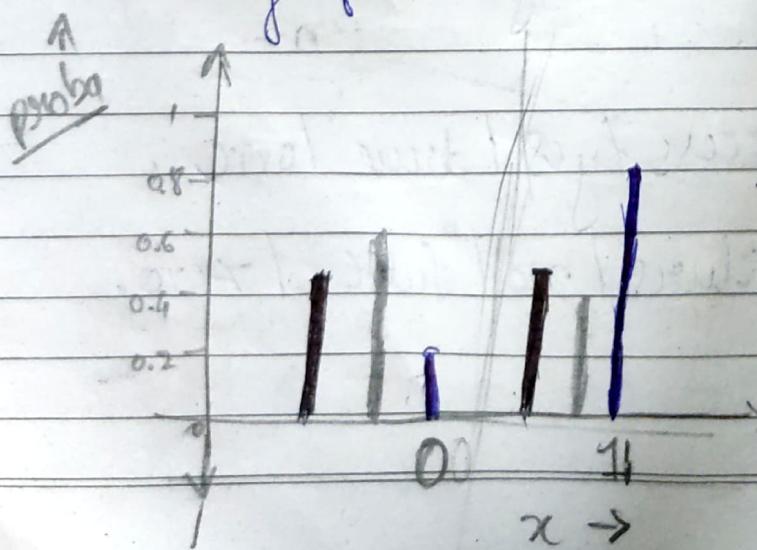
Probability
of getting
a head

$$q = 1 - p = 0.5$$

Probability of not
getting a head.

Here, we might also tell
prob getting a tail.

* So, any outcome in binomial distribution wants to display as graph can be displayed using PMF



We'll display the
prob of discrete
random variable.

ex:

$$\textcircled{1} \quad P(x=0) = 0.9 \\ P(x=1) = 1 - 0.9 = 0.1$$

$$\textcircled{2} \quad P(x=0) = 0.6 \\ P(x=1) = 1 - 0.6 = 0.4$$

② Person pass/fail exam.

* PMF VS pdf

\downarrow

discrete
random
variable

\rightarrow Continuous
random
variable

* ~~We have~~ we have
specified range of
values

* Specially, in Bernoulli's
distribution we have
1 outcome only.

~~Proving using [Probability mass function]~~
~~PMF~~ K can be 0 or 1

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

$$\text{① } P(x=1) = p^1 (1-p)^0$$

$\underline{p(x=1)} = p$

using this
we can get
'p' and 'q'
value.

\rightarrow in one scenario
we get 'p' value

$$\text{② } P(x=0) = p^0 (1-p)^{1-0}$$

$$= (1-p) = q$$

* Mean, variance & standard deviation
for Bernoulli distribution

⇒ Mean

Expected
value of
 K

$$E(K) = \sum_{x=1}^K K \cdot p(x)$$

$$E(K) = 1 * 0.6 + 0 * 0.4 \\ = \underline{0.6} = p$$

$$K = 1(0.6) 0$$

$$P(K=1) = 0.6 = p \\ P(K=0) = 1 - 0.6 = q \\ = \underline{\underline{0.4}}$$

⇒ Median

we consider

$$\text{Median} \left\{ \begin{array}{ll} 0, & \text{if } p < \frac{1}{2} \\ [0, 1], & \text{if } p = \frac{1}{2} \\ 1, & \text{if } p > \frac{1}{2} \end{array} \right.$$

⇒ Variance & std

$$\text{Variance} = p(1-p) = pq$$

$$\text{std} = \sqrt{pq}$$

* Simplified way of PMF ab proved

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

* Poisson distribution [PMF]

→ It is a discrete distribution

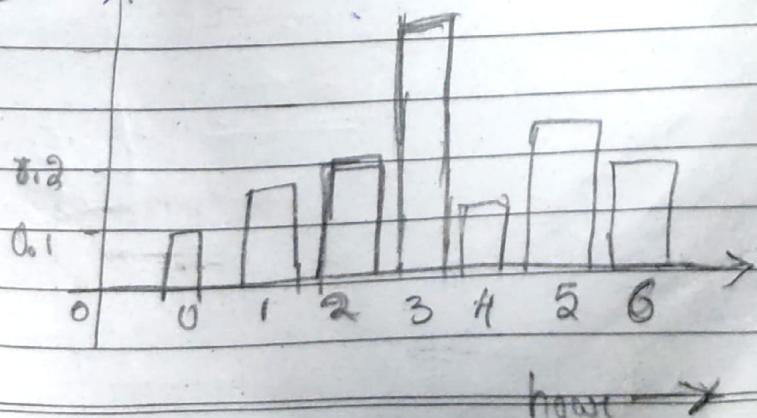
② It describes the number events occurring in a fixed time intervals.

ex: ① No of people visiting hospital every hour;

is defined
be "PMF"
 \downarrow
 \uparrow Pois

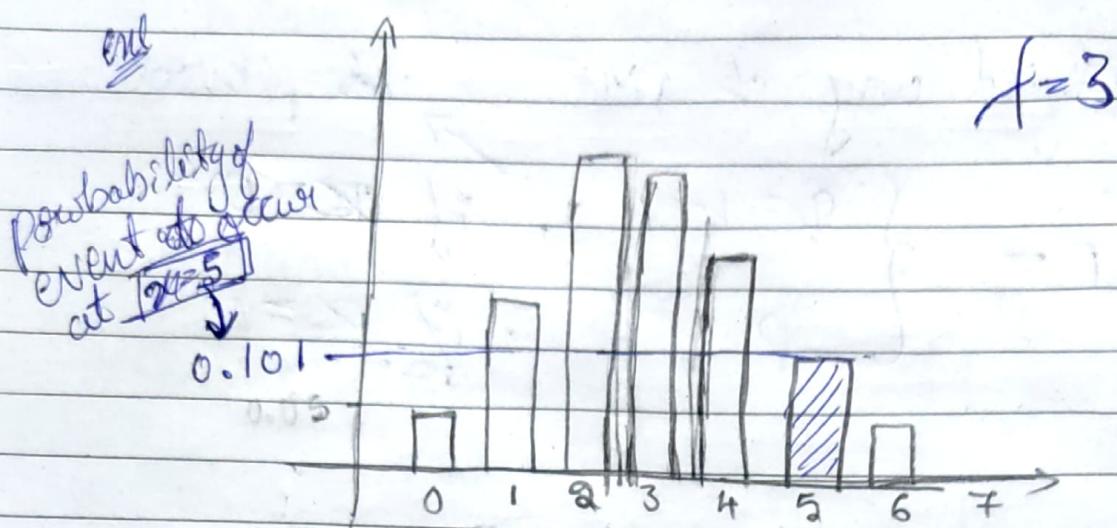
② — 11 — Bank — 11 —

③ — 11 — Airport — 11 —



$\lambda = 3$
→ Expected
No of event
to occur
at every
time interval

also used in
PMF



PME Probability of an event to occur
at 5th hour.

~~P(x=5)~~

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

College

$$e^{-m} m^x$$

$$x!$$

$$P(x=5) = \frac{e^{-3} 3^5}{5!}$$

$$= \underline{\underline{0.101}} = \underline{\underline{10.1\%}}$$

* Mean & Variance for poisson distribution

→ Mean

$$E(x) = \mu = \lambda * t$$

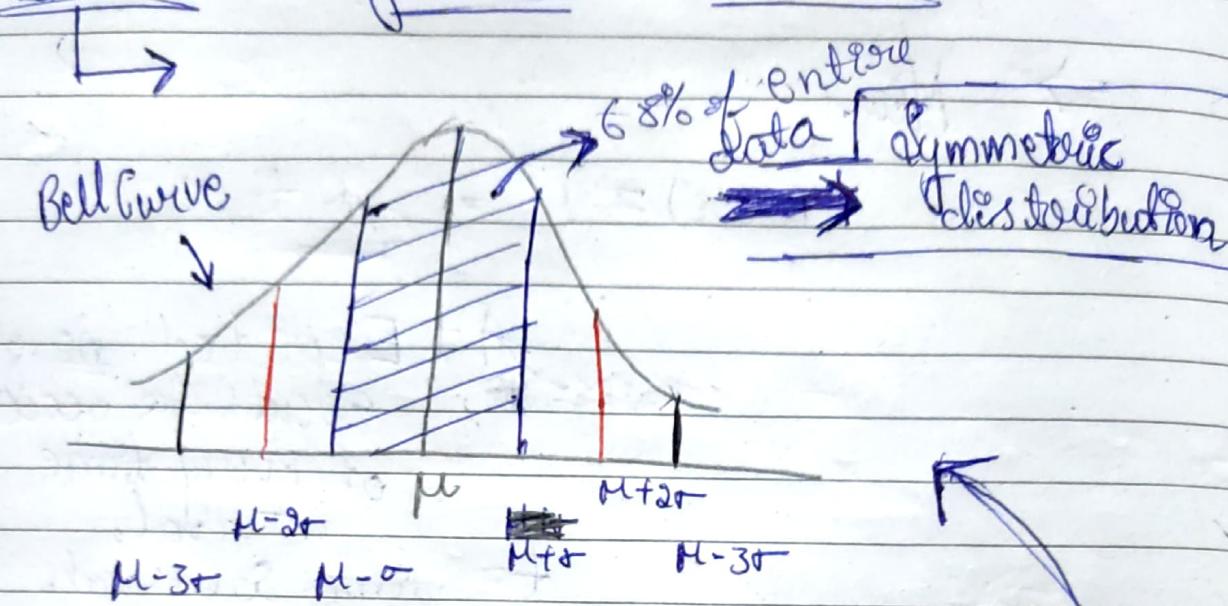
λ = Expected no of events to occur at every time interval.

t = time interval.

* Variance

$$E(x) = \mu = \lambda * t$$

* Normal (or) gaussian distribution.



Empirical rule [3 sigma rule]

$$68 - 95 - 99.7\%$$

Whenever we have a random variable looks exactly similar to (or) follow symmetric distribution

then within the first S.D from the mean 68% data will fall in that region

Ex: if we have a dataset contain 100 data ~~they~~ they 68 data will fall in I std.

- * Within 2nd std 95% of data falls in this region.
 - * - - - 3rd std 99.7% of data falls.
 - * To check whether the distribution is Gaussian distribution (or) not ~~not good~~
- ① Q-Q plot

Ans

Probability

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

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

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

Ex:

weight, height, IBIS Dataset.

* Q-Q plot:

mumpy
seaborn
matplotlib

pylab
scipy.stats

def plot_data(df, feature):
 plt.figure(figsize=)

plt.subplot(1, 2, 1)

~~df[feature].hist()~~

sns.histplot(df[feat], kde=True)

plt.subplot(1, 2, 2)

stat.probplot(df[feat], dist='norm')

Calculate the quantiles

for a probability plot

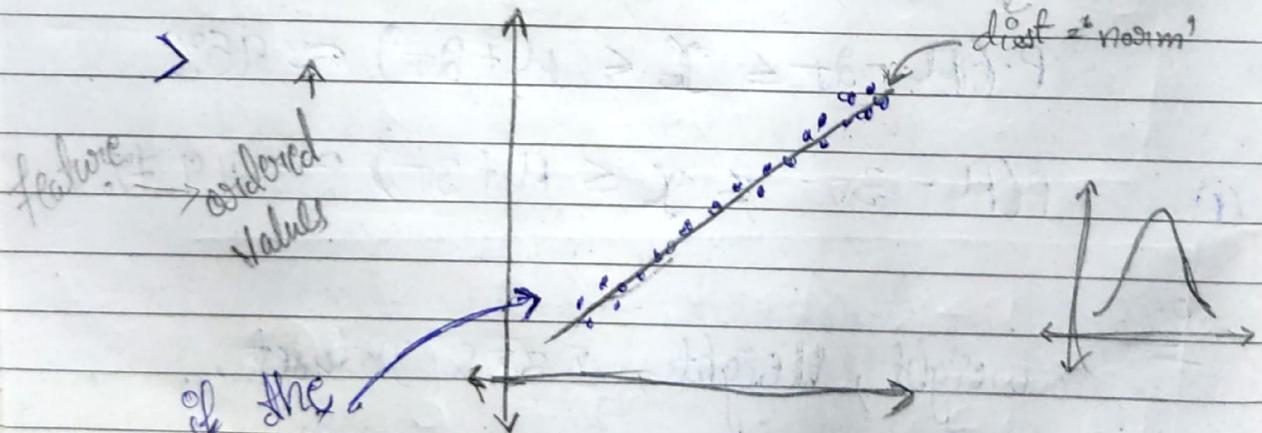
and optionally

show the plot: plt.show()

plot = pylab

obtained from

dist = 'norm'



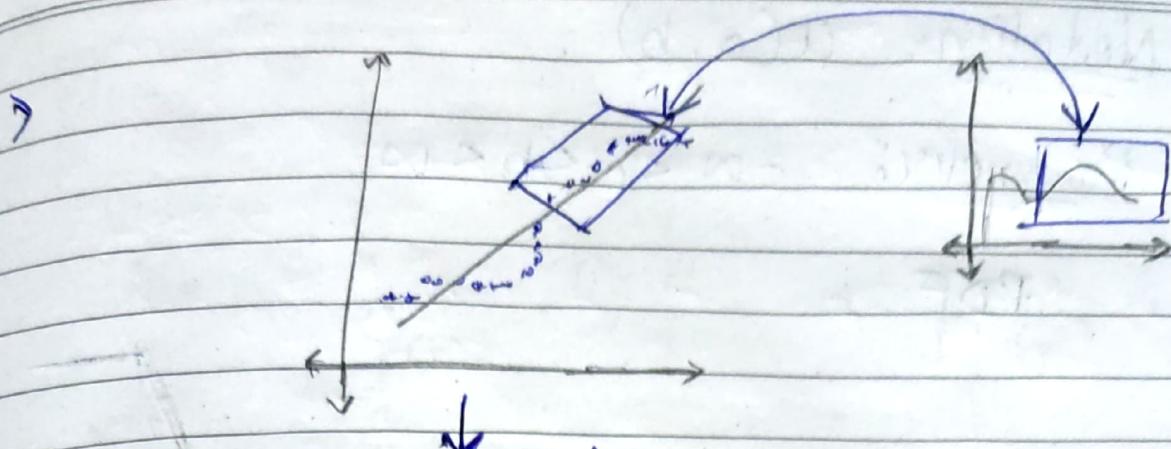
of the
distribution
normal
distribution
then points lie
in a single
line.

Theoretical
quantiles →

Q-Q plot ↴

tells the plot is

normal distribution
(or) not



Most of points don't lie on line \Rightarrow it's not normal distribution.

* Uniform distribution [constant probability density function]

① Continuous uniform distribution [PDF]

② Discrete uniform distribution [PMF]

① Continuous uniform distribution:

[continuous random variable]

* In probability theory and statistics, the continuous uniform distribution / rectangular distribution is a family of symmetric probability distributions.

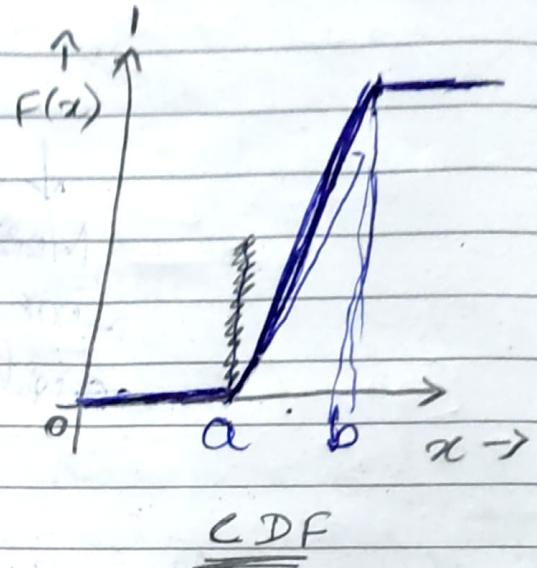
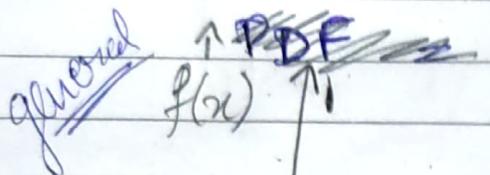
* The distribution describes an experiment where there is an arbitrary outcome that lies between "certain bounds".

* The bounds are defined by the parameters a and b , which are minimum & maximum values.

* interval: $[a, b]$ (or) (a, b) .

Notation: $u(a, b)$

Parameters: $-\infty < a < b < \infty$



PDF

CDF

~~PDF =~~

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

~~CDF =~~

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

$$\text{Mean} = \frac{1}{2} (a+b)$$

$$\text{Median} = \frac{1}{2} (a+b)$$

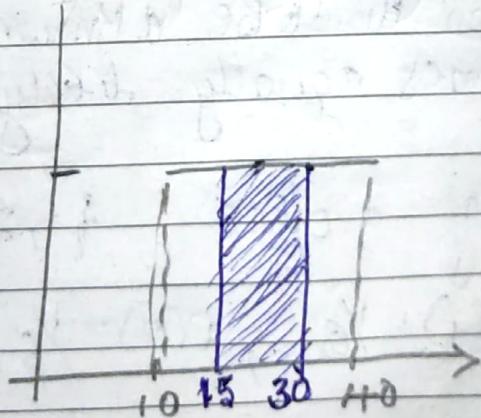
$$\text{Variance} = \frac{1}{12} (b-a)^2$$

Q6 The numbers of candies sold daily at a shop is uniformly distributed with a minimum of 10 and maximum of 40.

(i) What is the probability of daily sale to fall between 15 and 30?

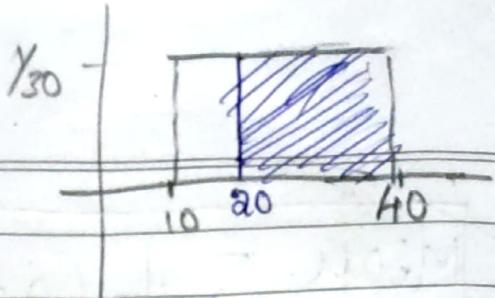
ans:

$$\frac{1}{30} = \frac{1}{b-a}$$



$$\begin{aligned} P(15 \leq x \leq 30) &= \frac{(b-a)}{30} \\ &= \frac{15}{30} = \underline{\underline{0.5}} \end{aligned}$$

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



$$\begin{aligned} P(x \geq 20) &= \cancel{P(20 \leq x \leq 40)} + P(x \geq 40) \\ &= (40 - 20) * \frac{1}{30} + 0 \end{aligned}$$

$$= \frac{20}{30} \rightarrow \frac{2}{3} \rightarrow 0.66 \approx 66.6\%$$

② Discrete uniform distribution: [PMF]

"Discrete random variable"

* In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein a finite number of values are equally likely to be observed, every one of n values has equal probability $\frac{1}{n}$.

* Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen."

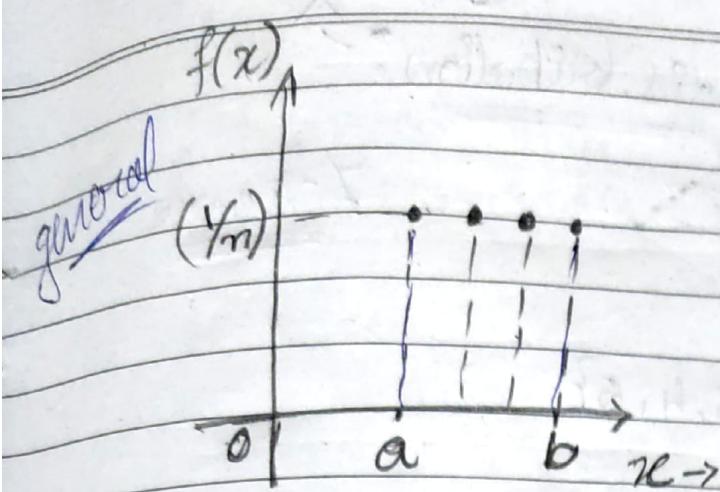
Ex: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$$P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, \dots$$

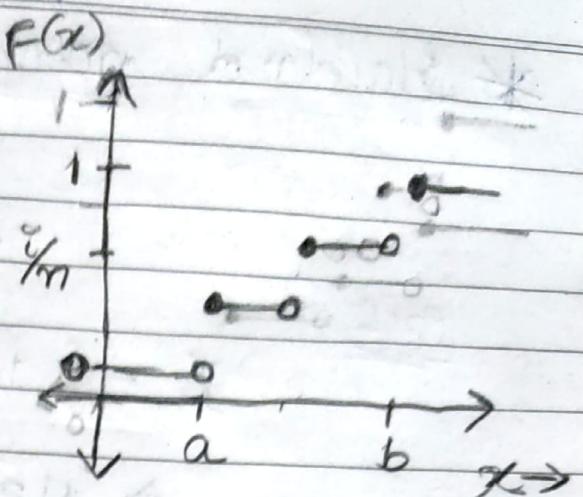
$$\Rightarrow a=1, b=6 \rightarrow \text{in dice.}$$

$$m = b - a + 1 = 6 - 1 + 1$$

= 6



PDF



CDF

* All equal probability outcome.

Notation : $U(a, b)$

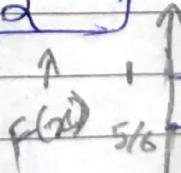
Parameters: a and b , $b \geq a$

$$\text{PMF} \Rightarrow \boxed{\frac{1}{m}}$$

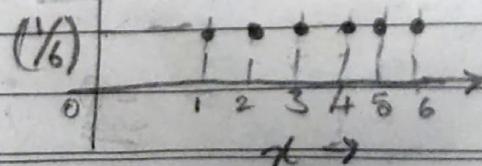
Mean

$$\frac{a+b}{2}$$

Median



dice \Rightarrow



PDF

Uses

- * To find how many std away from mean.
- * Standardization.

* Standard normal distribution:

Z-score

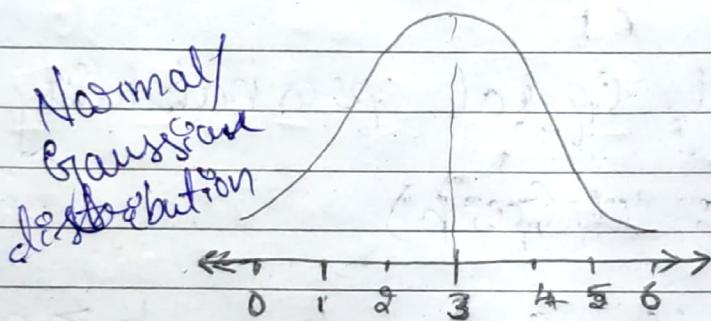
Ex:

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

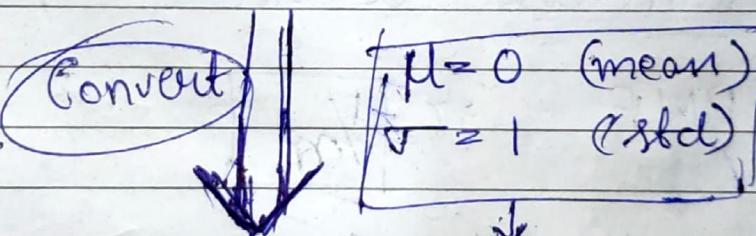
$$\Rightarrow \mu = 3$$

Consider

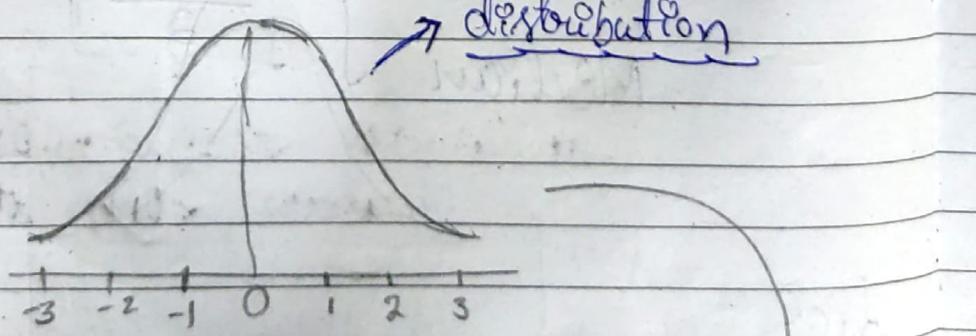
$$\sigma \approx 1.414 \approx 1$$



Conversion
is done
using
Z-score



Standard normal
distribution



$$Z = \frac{x - \mu}{\sigma}$$

$$z = \frac{1-3}{1} = -2 ; \frac{2-3}{1} = -1$$

Gaussian $X = \{1, 2, 3, 4, 5\}$ $\Rightarrow Z = \{-2, -1, 0, 1, 2\}$

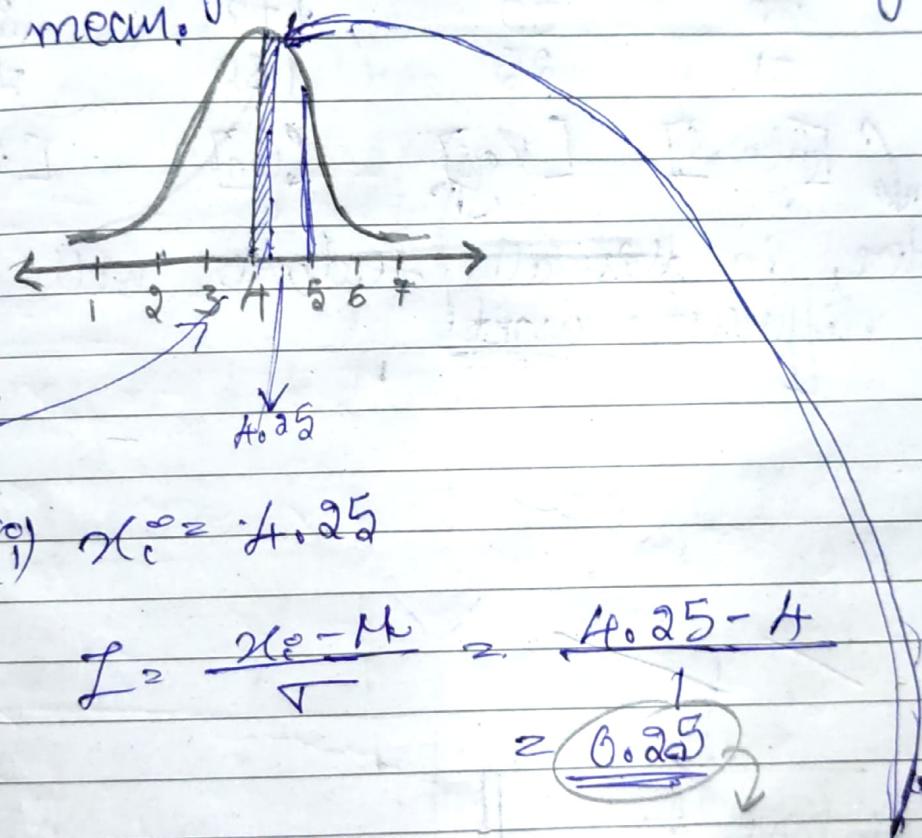
mean converted to zero

* Normal / Gaussian distribution

(using
Z-score)

Standard normal deviation
 $\mu = 0$,
 $\sigma = 1$

Q) How many std 4.25 is away from the mean.



indicates that 4.25 is $+0.25$ std away from the mean. (4)

(ii) $x_c = 1.25$

$$Z = \frac{x_c - \mu}{\sigma} = \frac{1.25 - 4}{1} = -2.75$$

Standard normal deviation

tells that 1.25 is -2.75 std away from mean.

~~CRIS~~ Dataset



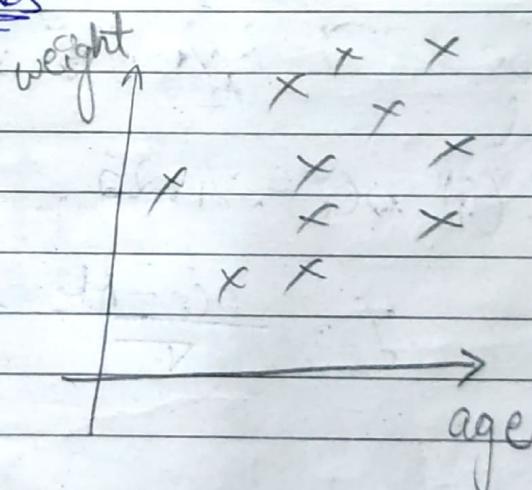
Age	weight	Height	Salary
-----	--------	--------	--------

24	70	175	40K
25	60	160	50K
26	55	180	60K
27	40	130	30K
30	30	175	20K
31	25	180	70K

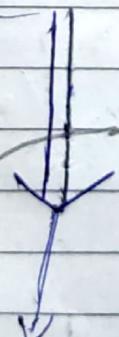
\downarrow [years] [kg] [cm] [INR]

Here, in this all feature will have different units

if we plot
due to big values
points will
be scattered.
Here and there



so we have to
bring feature
in same scale



because few ML
algo need features
to be in some scale
clustering, RNN, regression

They ~~get~~ model gives
more performance
increased

Convert using
technique called

"Standardization"

→ we apply Z-score
for all features

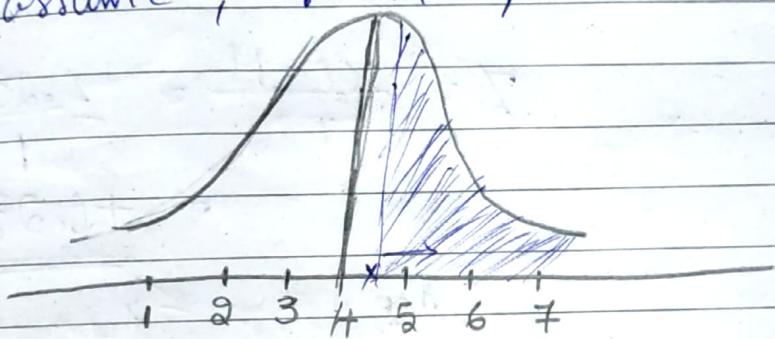
* Z score application and Z table.

$$Z = \frac{x_e - \mu}{\sigma}$$

→ problem statement

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

assume ; $\mu = 4.$, $\sigma = 1$



of what percentage of scores falls above

4.25

we have to do
standardization

- ① clustering algo.
- ② linear regression
- ③ logistic regression
so many

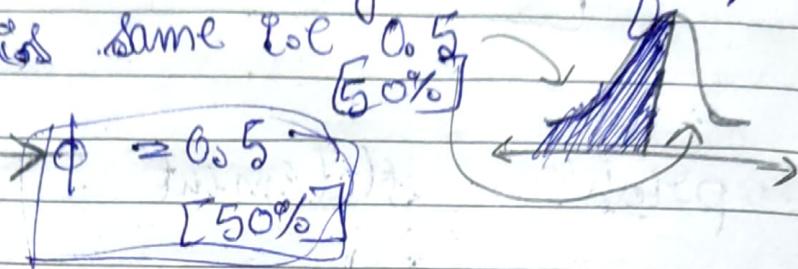
→ the area
under curve
should be
found to calculate
percentage of
scores.

→ we get value
in between
(-3 to +3)

Curve

\Rightarrow also, as bell Curve is follows symmetric distribution.

so area to right and left of mean is same i.e 0.5 [50%]



$$\rightarrow x_0 = 4.25, \mu = 4, \sigma = 1$$

$$Z = \frac{x - \mu}{\sigma} = \frac{4.25 - 4}{1} = +0.25$$

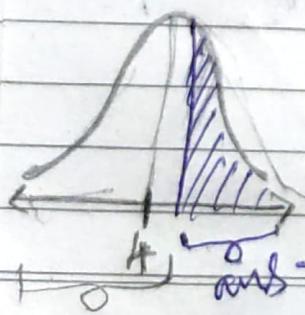
indicates 4.25 is $+0.25$ std away from mean.

Now, we have to find area to know the % of scores fall in the region.

we use

[Z-table] -

using (+ve) Z-table

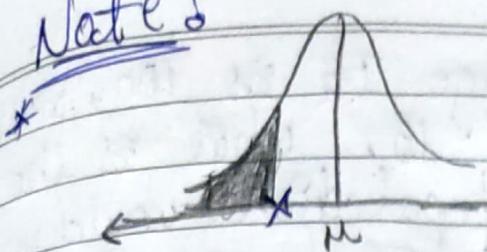


$+Z$ table $\rightarrow 0.5981$

$$\begin{aligned} \text{Area} &= 0.5981 \\ \text{ans} &= 1 - 0.5981 \\ &\approx 0.4019 \approx 40\% \end{aligned}$$

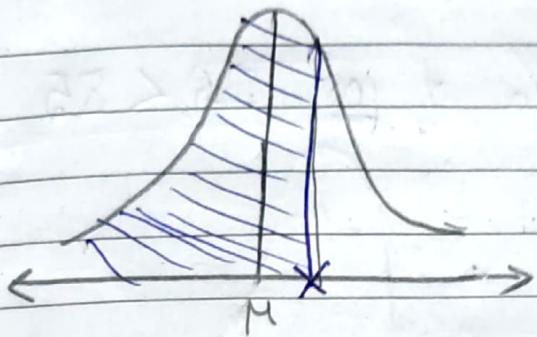
Area	$Z = +0.25$	40%
------	-------------	----------------

Note:



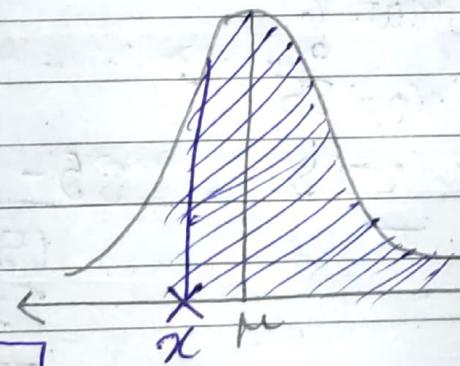
→ Use negative Z-table and calculate area.

BASIC



→ Use positive Z-table and calculate area.

*



↓

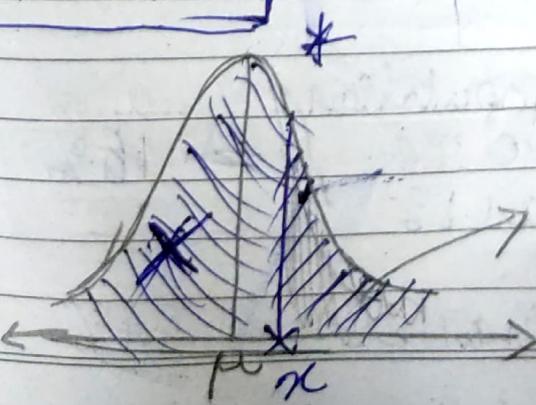
use -ve Z-table

$$\text{Area} = \underline{1 - (x) \text{area}}$$

- ① Negative Z-table.
- ② Positive Z-table.

use +ve Z-table

$$\text{Area} = \underline{1 - (x) \text{area.}}$$

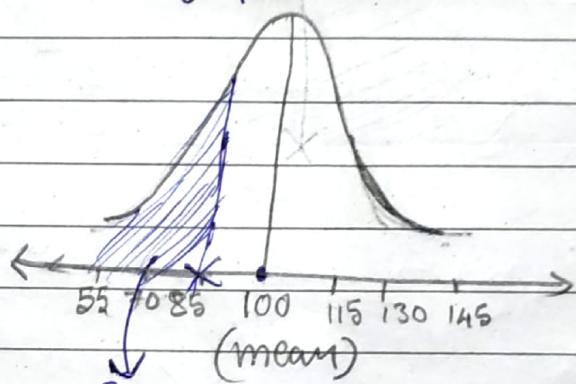


Q) In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have an IQ lower than 85?

$$\Rightarrow \mu = 100, \sigma = 15$$

~~Prob.~~ % of pop. IQ < 85

Initially



Percent Area = ?

$$x_0 = 85$$

We convert onto standard normal deviation

$$[\mu = 0, \sigma = 1]$$

$$Z = \frac{x_0 - \mu}{\sigma} = \frac{85 - 100}{15} = -1$$

$$\boxed{Z = -1}$$

$$\therefore \text{Area}_{Z=-1} = 0.15866 = 15.86\% \approx 16\%$$

∴ Percentage of population

~~that~~ who have IQ lower than 85 = 16%

∴ greater than 85 = 84%

∴ $75 < x < 100 = 45\%$

* Central limit theorem, [CLT]

Imp To understand sampling distribution & probability of sampling distribution.

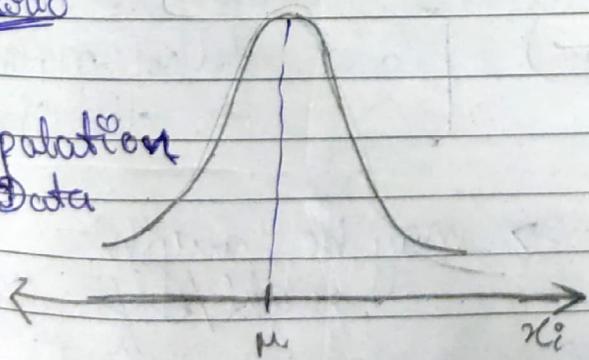
* The central limit theorem relies on the concept of a sampling distribution, which is probability distribution of a statistics for a large number of samples taken from a population.

* CLT says that the sampling distribution of the mean will always be normally distributed. as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

Scenario

0

Population Data



$$x \approx N(\mu, \sigma)$$

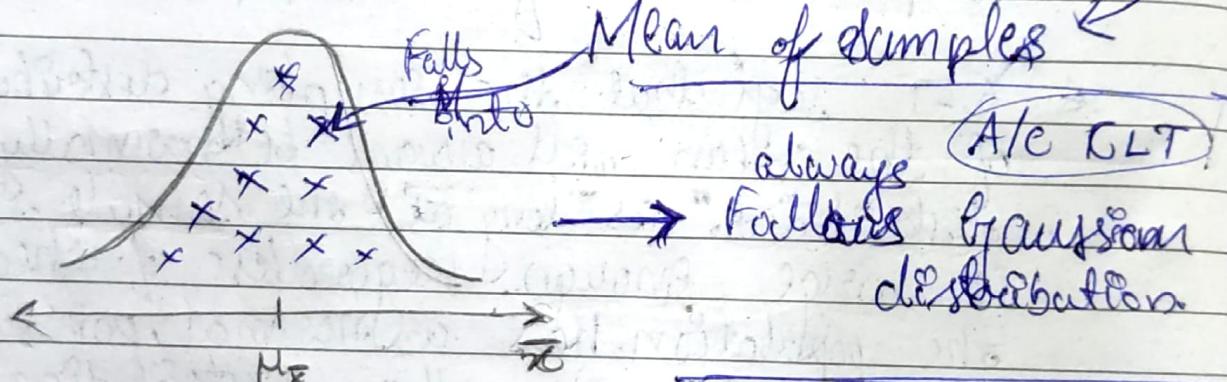
→ Gaussian distribution.

$n = \text{Sample Size} \Rightarrow n \text{ can be any value}$

sample
 $S_1 = \{x_1, x_2, x_3, \dots, x_n\} = \bar{x}_1$
 $S_2 = \dots = \bar{x}_2$
 $S_3 = \dots = \bar{x}_3$

$\{$

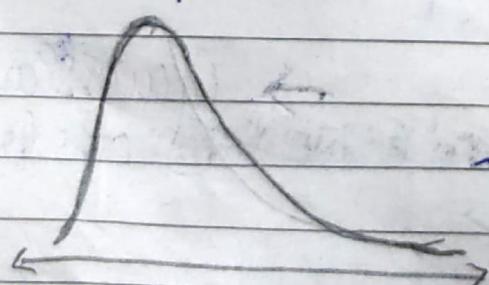
$S_m = \dots = \bar{x}_m$



Sampling distribution of the mean

Scenario

② $X \neq N(\mu, \sigma)$ [do not follow normal distribution]



→ may be anything right/left skewed.

$n \geq 30$ & Sample size

if

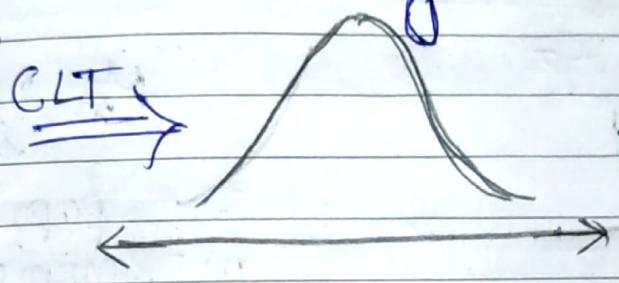
$$\begin{aligned} \delta_1 &= 2 \\ \delta_2 &= 2 \\ &\vdots \\ &\vdots \end{aligned}$$

$$\delta_m = \text{_____} = \overline{\delta_m}$$

mean

$$\sigma_{(n)} = \sqrt{\frac{1}{n}}$$

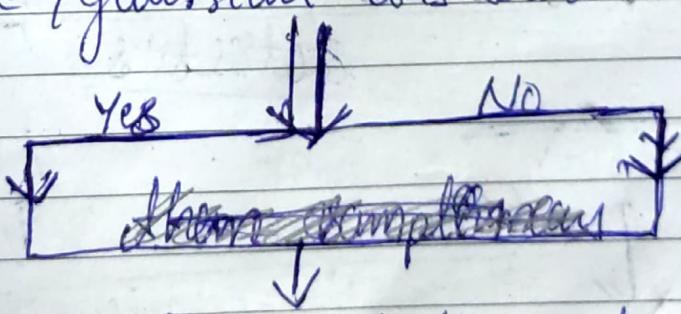
then sample will follow normal / gaussian distribution



∴ Central Limit theorem

states

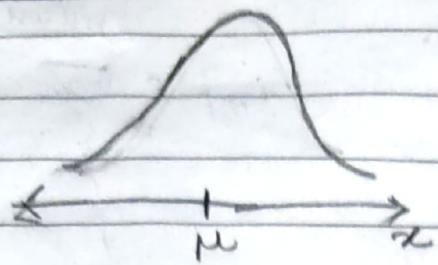
if population follows normal / gaussian distribution



then sampling distribution of mean
will follow "normal distribution"

* for sample size $n \geq 30$, if popu
donot follow nor. D.

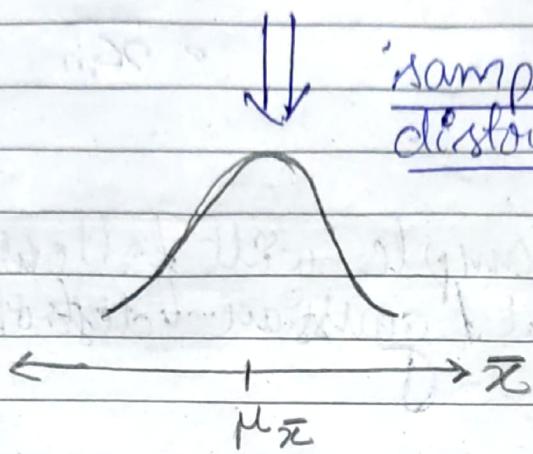
Normal distribution



Population
mean std

$$x \approx N(\mu, \sigma)$$

Population



sampling
distribution of mean

$$x \approx N(\mu_x, \frac{\sigma}{\sqrt{n}})$$

μ_x mean of
sample

σ = population std

n = sample size

Mean sample, $\mu_x = \mu$
std of sample, $\sigma_x = \frac{\sigma}{\sqrt{n}}$

↑ Under now descriptive
statistics

↓ From now inferential
statistics

* Estimate

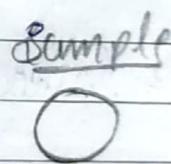
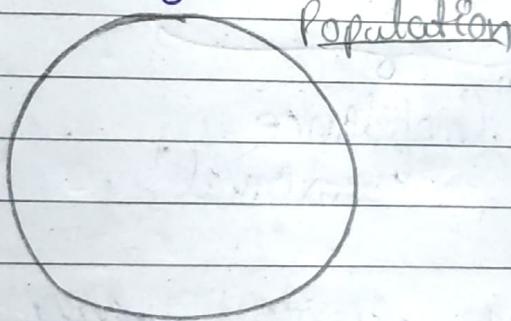
→ It is a specific observed numerical used to estimate an unknown population parameter.

Types of estimates

① Point estimate:

→ It is a single numerical value used to estimate an unknown population parameter.

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



μ
(Population mean)

\bar{x}

Here we find a point estimate and make some conclusion on the population.

Consider,

$$\mu = 65$$

$$\bar{x} = 60$$

↓
60 become a point estimate to assume about μ mean.

Disadvantage of point estimate



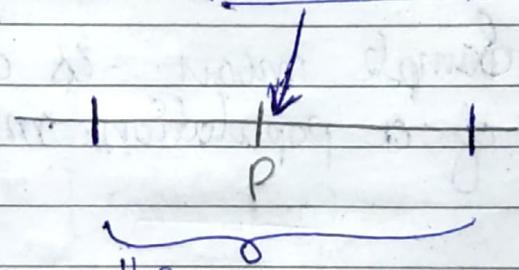
* There can ~~be~~ have margin of error, may be + or -.

So, better to create a range / interval ~~for estimation~~

② Interval estimate:

Range of values used to estimate the unknown population parameter

* along with point estimate we have ~~range~~ for estimation.



to avoid margin of error.

This is called as "Confidence interval"

Ex: for previous ~~example~~ example

we can say that

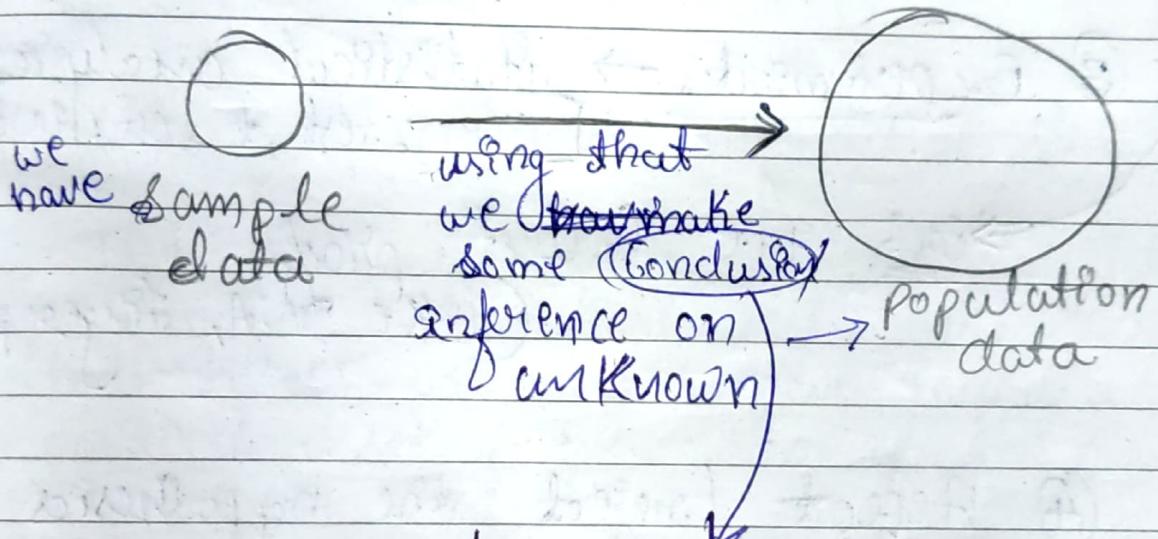
~~sample~~.

Population mean $\xrightarrow{\text{may be}} [55-65]$

↓
interval estimate
of a sample mean for
estimation of population mean.

* A Hypotheses and Hypotheses testing mechanism

Inferential stats \rightarrow Conclusion or inference from a sample data about a unknown population parameter.



how?
 \Rightarrow we use hypothesis testing.

\Rightarrow Hypotheses Testing mechanism

① Null Hypotheses (H_0)

\hookrightarrow The assumption that we are beginning with.

② Alternative hypotheses. (H_1)

\hookrightarrow The opposite of null hypothesis.

Ex:

Person $\xrightarrow{\text{Committed}} \text{Crime} \xrightarrow{\text{went}} \text{Court}$

(h₀) initial : he ~~do not~~ fell guilty

(h₁) He fell guilty (opposite)

③ Experiments \rightarrow Statistical analysis
(like z-test, t-test, ANOVA, ...)

Ans we collect proof
(like DNA, finger print)

④ Accept / reject the hypotheses testing.

Ex:

Colleges at district a states its average passed percentage of student are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 80% with a standard deviation of 4%.

Does this ~~college~~ have a different passed percentage.

Null hypothesis

Ans: $H_0 \rightarrow \mu = 85\%$

Average passed percentage
of student is 85%

Alternate
hypothesis

$H_1 \rightarrow \mu \neq 85\%$

near Godleg ~~do~~ have different
passed percentage.

* P value \rightarrow imp in inferential stats.

The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

uses: P value are used ~~to~~ in hypothesis testing to help decide whether to reject the null hypotheses.



Probability
of a $P=0.02$

Space Bar
laptop

specific \rightarrow Meaning: out of 100 touches, we
outcome in touch around 20 times in this
a specific region.

* Hypotheses testing

Outcome of a statistical test is a P-value.

Eg: Coin is Fair (or) not ~~Fair~~.
q 100 times

Scenario 1: $P(H) = 0.5$, $P(T) = 0.5$

Scenario 2: $P(H) = 0.6$, $P(T) = 0.4$

3rd: $P(H) = 0.7$, $P(T) = 0.3$
Coin not Fair

① Null Hypothesis:

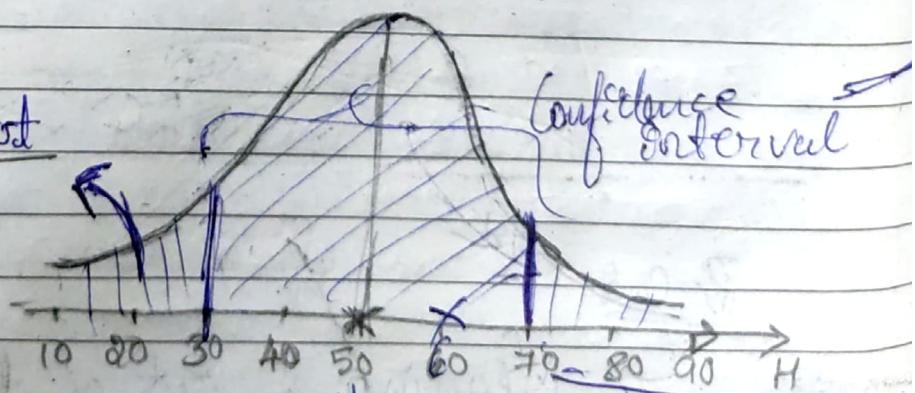
$H_0 \rightarrow$ Coin is Fair (biased)

② Alternate hypothesis:

$H_1 \rightarrow$ Coin is not Fair

③ Experiment: 100 times

"Coin is not fair"



So, Now we will
describe Confidence
interval.

but if ~~it~~ occurs so we
have to test it

determined by
domain expertise

A) Significance value

at 5% level of
significance

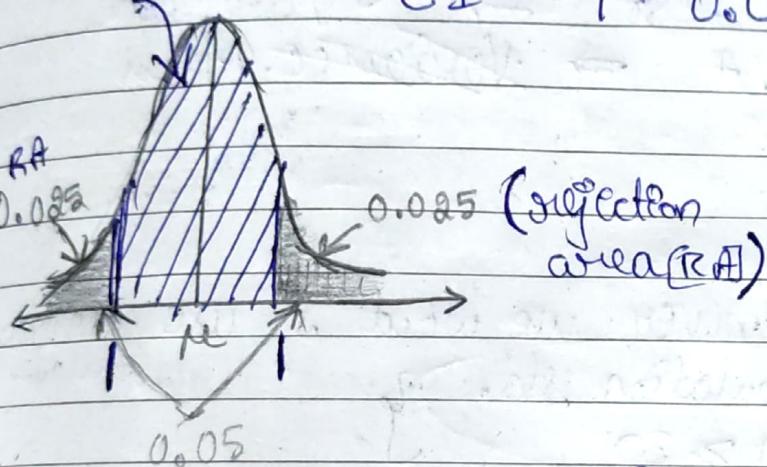
$$\alpha = 0.05$$

Except
null hypothesis
95%
CI

says that we have
to find Confidence Interval
by taking entire area 1 -

$$CI = 1 - 0.05 = 0.95$$

i.e 95% Confidence
interval.



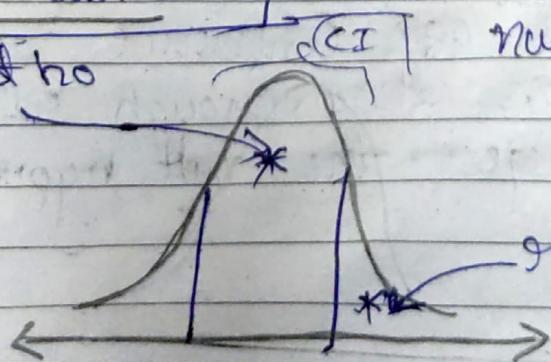
⑤ Conclusion:

$P < \text{Significance value} \Rightarrow$
 $P > \text{Significance value} \Rightarrow$

we reject the
null hypothesis

else - Fail to reject
null hypothesis

accept h_0



* Hypothesis testing and statistical analysis:

① Z test $\left\{ \begin{array}{l} \text{Population} \\ \text{Average} \end{array} \right\} \Rightarrow$ Z table

② t test \Rightarrow t table

③ Chi square \Rightarrow Categorical data

F test \rightarrow Variance

④ ANNOVA \rightarrow Variance. Mean

① Z test:

Whenever we want to use Z test we should follow

- (i) Population std.
- (ii) $n \geq 30$

a) The average height of all residents in a city is 168 cm with a ~~mean~~ std of $\sigma = 34$ believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5 cm

① State null and alternate hypotheses.

② At 95% confidence level, is there enough evidence to reject the null hypothesis.

- ① Population mean should be there & $n \geq 30 \Rightarrow Z$ test.
- ② Null hypothesis.
- ③ Alternative hypothesis.
- ④ Decision boundary.
- ⑤ Find Z value

We find
Z score & P value

population mean
 $\Rightarrow \mu = 168$, $t = 3.9$, $n = 36$.
 sample mean
 $\bar{x} = 169.5 \text{ cm}$.

$$C.I = 0.95, \alpha = 1 - 0.95 = 0.05$$

Significance value

① Null hypothesis.

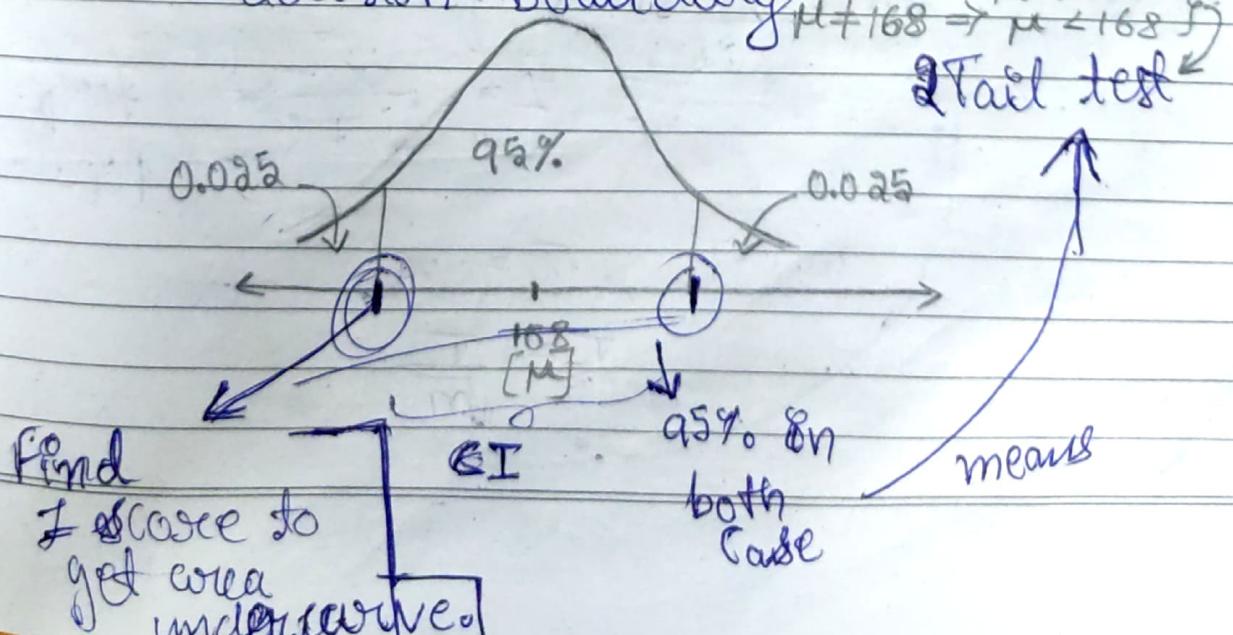
$$H_0 \Rightarrow \mu = 168 \text{ cms}$$

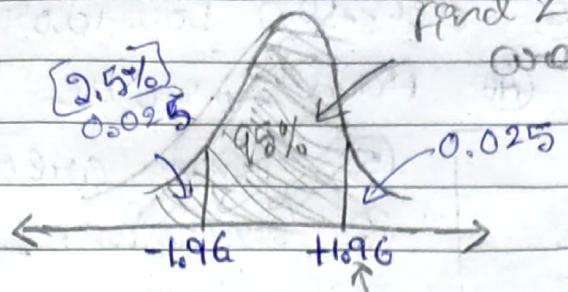
② Alternate hypothesis.

$$H_1 \Rightarrow \mu \neq 168 \text{ cm}$$

③ Based on C.I we will draw decision boundary $\mu > 168 \Rightarrow \mu < 168$

Fail test





$$1 - 0.025 = 0.9750 \Rightarrow \text{Z-score}$$

↓ ↓
 area converted using table → **+1.96**

* if Z is less than -1.96 (or) greater than +1.96 ; then we have to reject the Null hypothesis.

* Z-test

$$Z_d = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

for sample

but Z-score for normal distribution

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

for population

$$\text{W.K.T} \quad \bar{x} = \bar{x},$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{169.5 - 168}{3.0/\sqrt{36}}$$

$$Z_d = \frac{1.5}{0.65} = +2.31$$

As $Z_d = +2.31$ which is greater than $+1.96$.

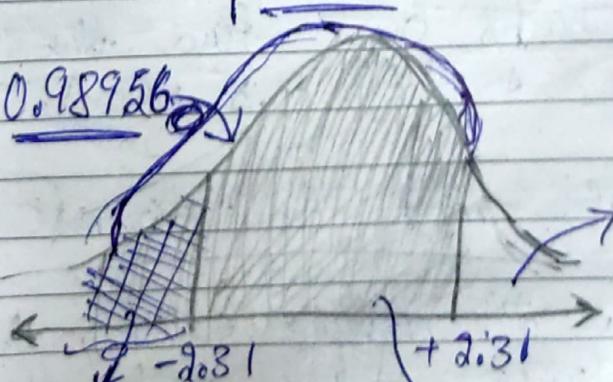
We reject the hypothesis at 5% level of significance.
i.e. 95% confidence interval.

$$\left. \begin{array}{l} 2.31 > 1.96 \\ Z_d > Z_{c_1} \end{array} \right\}$$

* Conclusion

$$P < 0.05$$

$$\text{Area } Z = 2.31 = 0.98956$$



↓ Table

$$\begin{aligned} \text{Area} &= 1 - 0.98956 \\ &= 0.01044 \end{aligned}$$

$$\text{Area} = 0.01044$$

$$\begin{aligned} \text{Area} &= 1 - \frac{2}{2} * 0.01044 \\ &= 0.97912 \end{aligned}$$

$$\begin{aligned} p\text{-value} &= 0.01044 + 0.01044 \\ &= \underline{\underline{0.02088}} \end{aligned}$$

3% level
of significance

$p < 0.05$
 $0.02088 < 0.05$ → True
 ↓
 reject the null hypothesis

Final

The average height ≠ 168 cm

→ The average height seems to increasing based on sample height.

a) A factory manufactures bulbs with a average warranty of 5 years with std of 0.50. A worker believes only less than that the bulb will malfunction in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years.

② State null and alternate hypothesis.

③ At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

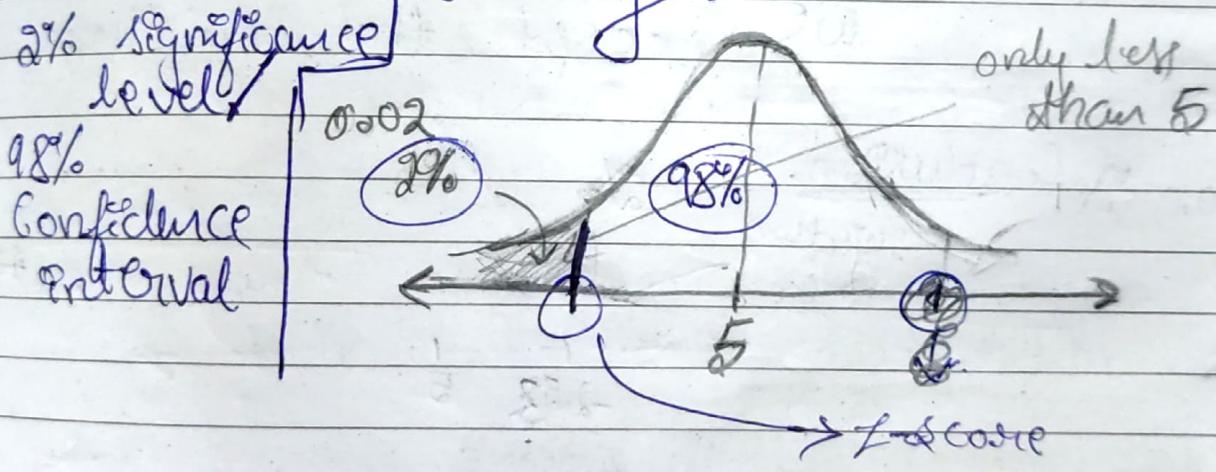
Null hypothesis
 \neq ① $H_0 \Rightarrow \mu = 5$ years
 Mean population
 ② $H_1 \Rightarrow \mu < 5$ years → 1 tail test

Mean population, $\mu = 5$

std. population std, $\sigma = 0.5$

sample size, $n = 40 \rightarrow z$ test
 sample mean, $\bar{x} = 4.8$

③ Based on confidence significance level - we draw decision boundary.



$$1 - 0.02 = 0.98 \rightarrow 3.8$$

area

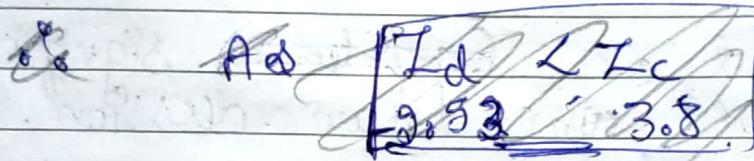
z score
Z score

* If $Z < -3.8$ or $Z > +3.8$ we reject null hypothesis.

$$\Rightarrow Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

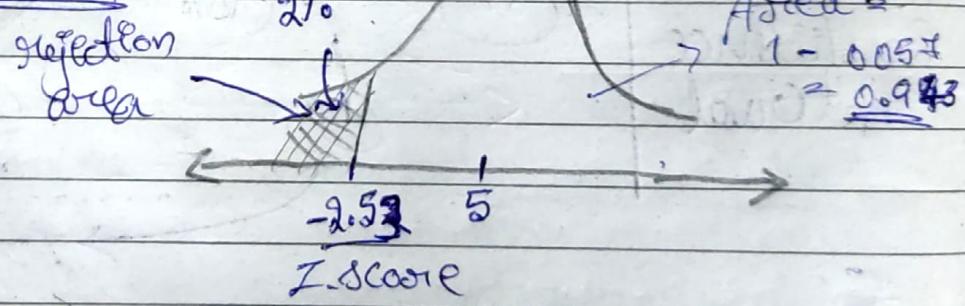
$$= \frac{4.8 - 5}{0.5/\sqrt{40}}$$

$$\underline{Z_d = -2.53}$$



~~we accept the hypothesis~~

* Conclusion.



$$\text{Area}_{Z = -2.53} = \cancel{0.005} \quad \underline{0.05}$$

$$P \text{ value} = \underline{0.05}$$

Compare P-value with α

$$0.05 > 0.02 \Rightarrow \text{Fals}$$

my P value

which mean that point
 also fall in $\approx 2\%$ level of significance
 will not fall in as My P value! < 0.05⁺
 we accept the Null Hypothesis.
 (fail to reject)

* Student t distribution [t-stats]

If Z stats when we perform any analysis using Z-score.

* we require σ (standard deviation) $\xrightarrow{\text{population}}$
 it is difficult $\xrightarrow{\text{to find}}$ $\xrightarrow{\text{is already known.}}$
 as population data is not there.

* How do we perform any analysis when we ~~don't~~ know the population std?

Students t distribution

Z-test

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

but

t-test

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Z table

t = sample standard deviation

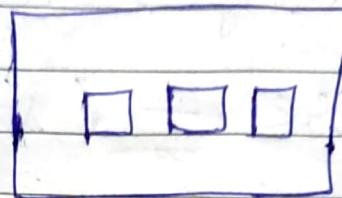
we use t-table

we have parameter degree of freedom.

* Degree of freedom = $n - 1$

→ ex:

if we have 3 chair in a Room



* if 3 person comes
only 2 person gets ~~one~~ option to seat
on chair.

*) T-test — (One sample t-test)

In a population the average is 100. A team of researchers want to test a new medication to see if it ~~has~~ has either a positive/negative 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 10. Did the medication affect intelligence? $C.I = 95\%$

* Here, we basically checking whether the intelligence increased / not.

\Rightarrow population mean, $\mu = 100$
sample size, $n = 30$
sample mean, $\bar{x} = 140$
sample std, $S_{\bar{x}} = 20 \Rightarrow S = 20$

population std ~~mean~~ \Rightarrow not given
 \Rightarrow t-test

Confidence interval = 95%

level of significance = 5% , $\alpha = 0.05$

① Null hypothesis, $H_0 \Rightarrow \mu = 100$

Alternate hypothesis, $H_1 \Rightarrow \mu \neq 100$
{ 2 tail test }

② $\alpha = 0.05$

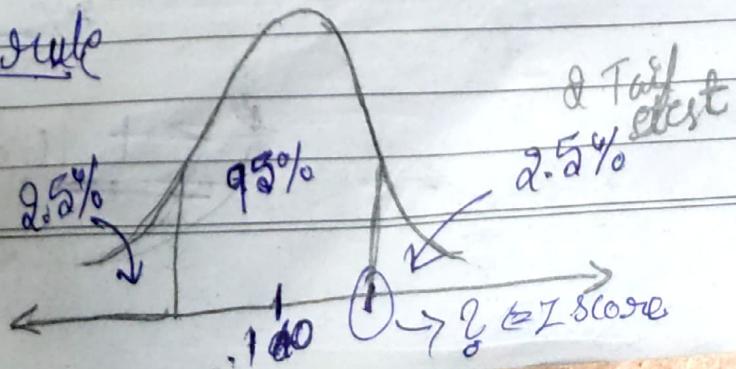
For finding

t score { ③ Degree of freedom:

$$dof = n - 1 = 30 - 1$$

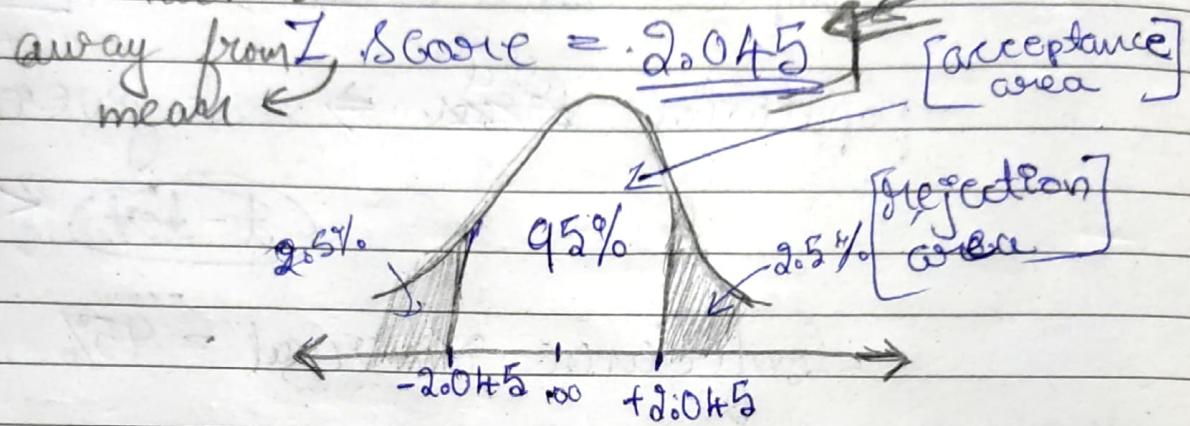
$$dof = 29$$

④ Decision rule



⑤ Finding t score using t -table
~~for~~
 $d.f = 29 \text{ & C.I} = 95\%$

how much std



⑥ if t -test is less than -2.045 (or) greater than $+2.045$ then we reject null hypothesis.

~~decision rule~~ →

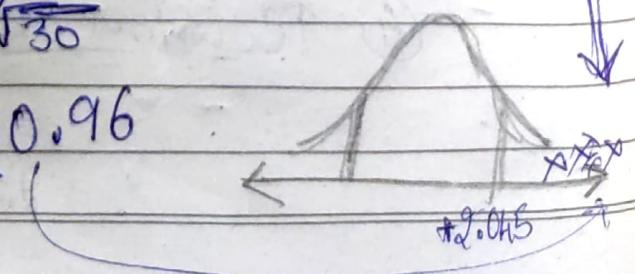
⑥ Calculating test statistics

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{140 - 100}{\frac{80}{\sqrt{30}}}$$

$$\underline{t = \pm 10.96}$$

falls in this region



Since, $t = 10.96 > t = 2.045$
↓
True

∴ We have to reject Null hypothesis

③ Conclusion:

Medication used has affected
the intelligence.

→ But, specifically as it is "t+ve"
means that medication
has increased intelligence.

* Binomial distribution ↴

no of successes in a fixed no of
independent trials.

* When no of trials increases

shape of distribution become
"Narrow"

* ~~Bin Poisson~~

as mean increases

shape of distribution become "taller"

Code

* Calculating pdf using "scipy.stats"

→ ~~Scipy.stats~~ norm.pdf(point,
loc=mean, scale=std)
 \uparrow \uparrow
P/P P/P

* taking 1000 binomial distribution sample with the 0.4 (p) probability
 (size)

→ trials = 1

random.sample = np.random.binomial(total, p,
 (size))

* poission distribution

np.random.poisson(mean, sample size)

* CDF

$F(x) =$

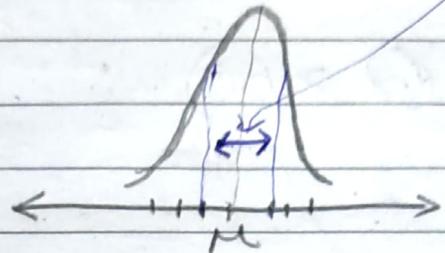
→ discrete $\Rightarrow P(X \leq x) = \sum [P(X=x_i) \text{ for } x_i \leq x]$

→ Continuous $\Rightarrow P(X \leq x) = \int_{-\infty}^x [f(t) dt] \text{ for } t \leq x$

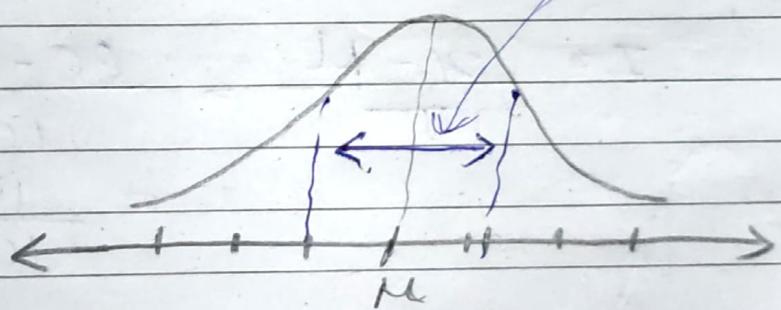
μ & σ Controls
the shape
of curve.

* Normal distribution

→ smaller σ (std)
 \Rightarrow Narrow Curve



→ larger σ (std)
 \Rightarrow wider, more spread.



* Assignment

$$\mu = 50, \sigma = 10..$$

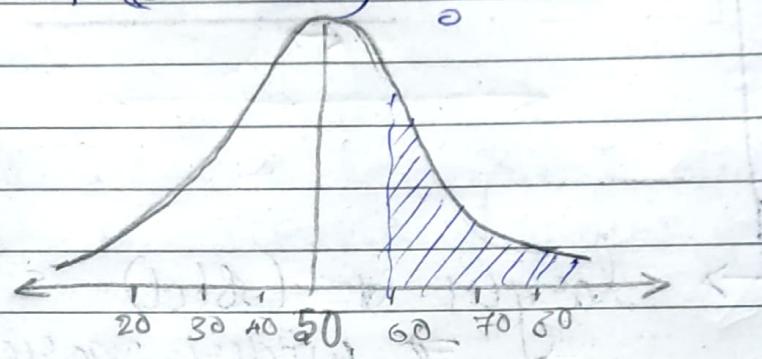
$$P(x > 60)$$

19

* Assignment Q6 Solution

mean, $\mu = 50$
 standard deviation, $\sigma = 10$

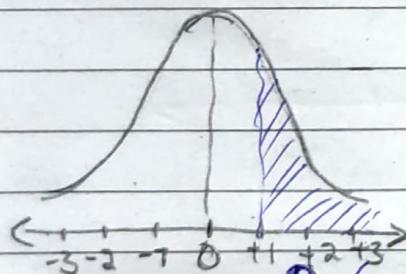
$$P(x > 60) = ?$$



$$x = 60,$$

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 50}{10}$$

$$= \frac{10}{10} = 1$$



$$P(x > 60) = P(x > 1)$$

(From Table)

$$= 1 - 0.8413$$

$$P(x > 60) = \underline{\underline{0.1587}} = 15.87\% \approx 16\%$$

Answer: 16%

Week-19

Statistics advance-2

*

Do we know the population standard deviation?

Yes

No

Is the sample size ≥ 30 ?

t-test
use

Yes | No

use Z-test

use T-test

* Type I and Type II error

H₀ can be true / can be false based on result of H₁

Actual Reality: Null hypothesis is True (H₀)
Null hypothesis is False.

Decision: ————— || —————

Out Come 1: we ~~reject the Null hypothesis~~ when in reality it is ~~True~~ ~~False~~. ↗ Good Scenario

Reality → it is False

Type Out Come 2: we ~~reject the Null hypothesis~~ when in reality it is True. ↗

(from hypothesis testing we are rejecting) But, (in reality that is true condition)



It's a Error

(Type I Error)

→ PMP - In ML
use in Confusion matrix

Type 2 Out Come 3: we ~~accept the Null hypothesis~~ when in reality it is False. ↗

(Type 2 Error)

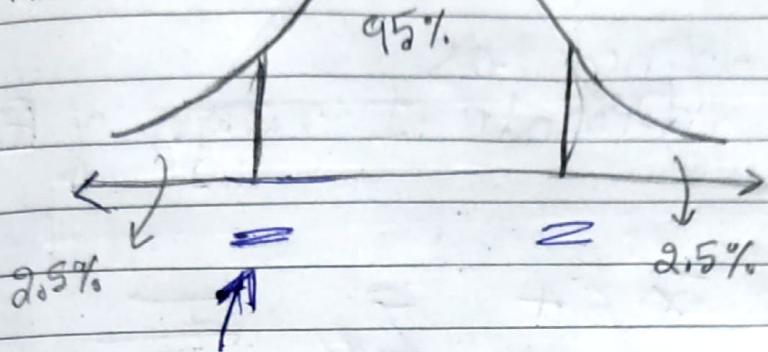
accept

Out Come 4: we ~~retain the Null hypothesis~~ when in reality it is True. ↗

Correct / Good Scenario

* Confidence Interval

Confidence Interval



For 2-tail test
with
 $CI = 95\%$

what value
can come in
this interval.

Point Estimate

sample
mean \bar{x} for μ
may be PE population
mean

$$\bar{x} = 2.5$$

$$\mu = 3$$

Here Point estimate $\bar{x} < \mu(3)$

So, there is a difference this is
the problem in PE

But if we say - our mean lies between 2 to 4 then it will be fine.

Confidence interval

CI defined as
∴ Point estimate \pm margin of error

For
Z-test

$$\Rightarrow \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

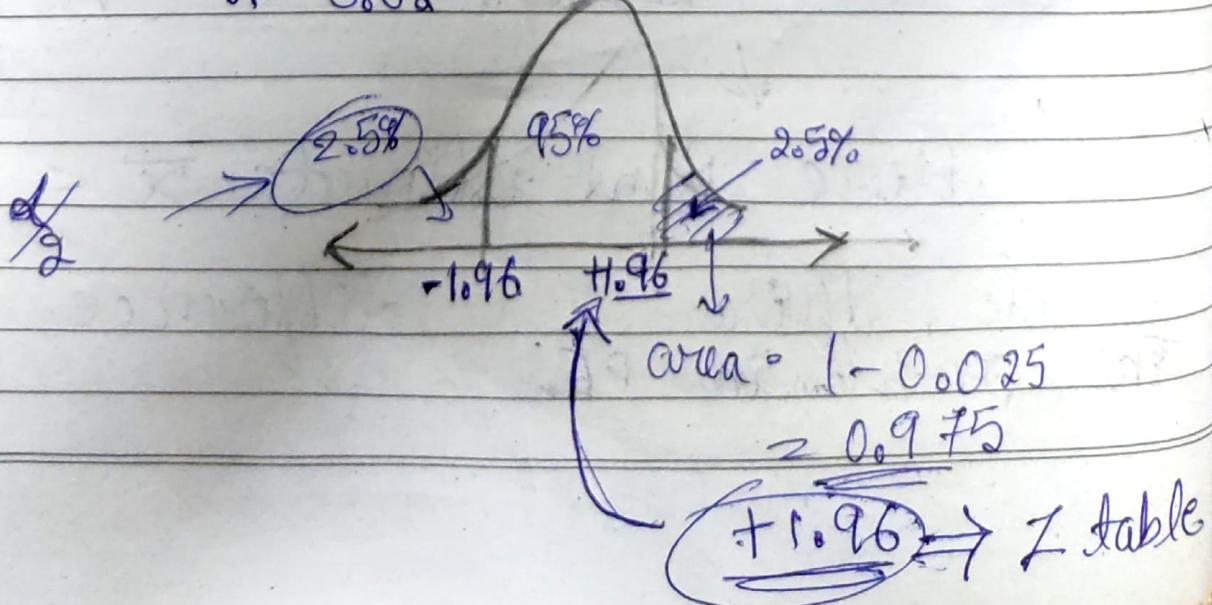
Ex: On the verbal section of SAT exam. The std is known to be 100. A sample of 30 test takers has a mean of 520. Construct 95% CI about the mean.

\Rightarrow

$$\bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

→ 2 tail

$$\alpha = 0.05$$



$$\text{lower CI} = 520 - (1.96) \frac{100}{\sqrt{30}}$$

$$= 520 - 39.2$$

$$\text{lower CI} = \underline{\underline{480.8}}$$

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

$$= \underline{\underline{559.2}}$$

Conclusion: I am 95% confident about the mean CAT score is between 480.8 and 559.2.

$$CI \Rightarrow [480.8, 559.2]$$

* Bayes Theorem [Vimp]

→ Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes theorem.

→ Vimp for ML algo, Naive Bayes
 → Completely on Probability.

Probability → Independent Events
Dependent Events.

① Independent events: ② Dependent Events:

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

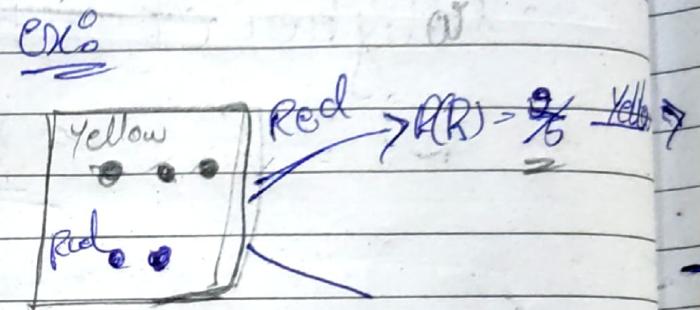
$$P(1) = \frac{1}{6}, P(2) = \frac{1}{6} -$$

* Here, Happening of one event is not impacting other event. So, Independent

② Tossing a coin

$$P(H) = 0.5, P(T) = 0.5$$

Probability of Red and after Red, yellow.



* Here, happening of one event affect the happening of other event.
So dependent.

* $P(R \text{ and } Y) = P(R) \times P(Y|R)$

Probability of Yellow when 1st event has occurred

Condition Probability

$$\frac{2}{5} \times \frac{3}{4} = \frac{6}{20} = \frac{3}{5} \times \frac{2}{4} = \frac{6}{20}$$

$$\therefore P(A \text{ and } B) = P(B \text{ and } A)$$

$$P(A) * P(B/A) = P(B) * P(A/B)$$

$$P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

1 red already
taken out

$$P(A) = \frac{3}{4}$$

Bayes' theorem

(Q1)

$$P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

~~college~~

$A, B \Rightarrow$ events

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

$P(B/A) \Rightarrow$ $\frac{-11}{A} - 11$ B given

$P(A), P(B) \Rightarrow$ independent probability of
A and B

Dataset

[Independent]

[Dependent]

size of house	No of rooms	location	Price
x_1	x_2	x_3	y

↓ ↓ ↓ ↑

Using these 3 features we can predict Price

$$P(y/x_1, x_2, x_3) = \frac{P(y) * P(x_1, x_2, x_3/y)}{P(x_1, x_2, x_3)}$$

From Bayes Theorem

We use this to calculate probability of output which here is price.

chi-square test

we tell ↴

The "chi-square test" for goodness of fit test claims about population proportions.

It is a non-parametric test that is performed on the categorical data.
[ordinal & nominal data].

~~ex~~ ~~test~~ There is a population of male who likes different colors bikes

Theory about sample

Yellow Bike

$\frac{1}{3}$ pop 22

Red Bike

$\frac{1}{3}$ 17

orange Bike

$\frac{1}{3}$ 59

* Goodness of fit test ↴

GMP

The sample data which is present collected does not support the theory present with us]

we check that

Called

"theory. Categorical distribution"

Called

"observed categorical distribution"

Ex: In a science class of 75 student 11 are left handed. Does this class fit the theory that 12% of people are left handed.

Answer:

	Observed value	(Derived) Expected value
left handed	11	9
right handed	64	66
Total	75	75

$$12\% \Rightarrow \frac{12}{100} \times 75 = \underline{\underline{9}}$$

are left handed

Expected value

based on the theory

remaining: right handed = 75 - 9 = 66

Now, we have to test ~~the goodness~~
that observed value is the goodness of
fit for the expected values.

CHI-SQUARE for goodness of Fit

In 2010 census of the city, the weight
of the individuals in a small city were
found to be the following

$<50\text{kg}$	$50-75$	>75
20%	30%	50%

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

sample data \Rightarrow

<50	$50-75$	>75
140	160	200

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

→ (theory)

2010
Original
data



$< 50\text{ Kg}$	$50 - 75$	> 75
20%	30%	50%

Expected
values

~~Actual~~
~~observed~~
Values
data
[2020]

$$n = 500$$

$< 50\text{ Kg}$	$50 - 75$	> 75
140	160	200

"
" expected
data
in [2020]

	< 50	$50 - 75$	> 75
	$\frac{20}{100} \times 500$	$\frac{30}{100} \times 500$	$\frac{50}{100} \times 500$
	= 100	= 150	= 250

Now, we will apply chisquare test

① Null hypothesis $\Rightarrow H_0$: The data meets the expectation

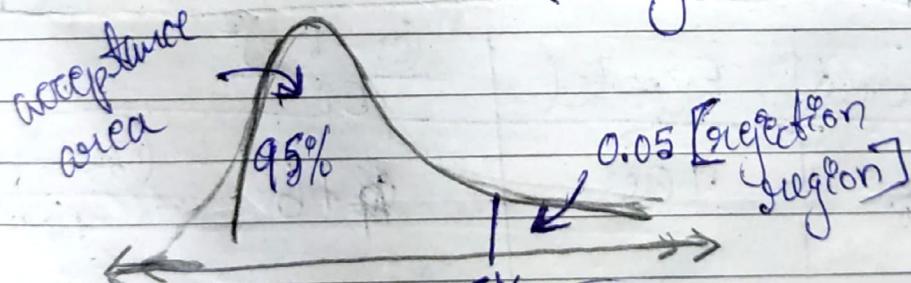
Alternate hypothesis $\Rightarrow H_1$: The data does not ~~not~~ meet the expectation.

② Significance level $\alpha = 0.05$, C.I = 95%

③ Degree of freedom

$$df = k - 1 = 3 - 1 = 2$$

④ Decision Boundary



we get
from "chisquare
table" using
 d and df

In Z & t test
we use symmetrical
distribution.
but in
chisquare test
we have
right skewed
distribution.

for $(\alpha = 0.05 \text{ &})$ ch.
 ch. square value $\Rightarrow 5.991$
 critical value

~~revision note~~
 If χ^2 is greater than 5.99, reject the H_0 , else we fail to reject the Null Hypothesis

⑤ Calculate Chi Square Test Statistics

$$\chi^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
≤ 50	140	100	16
50-75	160	150	2/3
> 75	200	250	10
			$\sum = 26.66$

$$\underline{\underline{\chi^2 = 26.66}}$$

B Conclusion

As $\chi^2 = 26.66$, which is greater than 5.99. i.e ~~reject~~

$$\boxed{\chi^2 > 5.99} \Rightarrow \text{True}$$

So, we reject the null hypothesis.

Answer

The weights of 2020 population are different than those expected in 2010.

chi-square in Python

$\rightarrow \text{observed_data} = [\quad]$

$\rightarrow \text{expected_data} = [\quad]$

Find χ^2 value $\rightarrow \text{sum(obs)} == \text{sum(expe)}$

value $\rightarrow \text{chi-square.test(stat, p.value = stat.chisquare(O, E)}$

And $\rightarrow \text{Significance} = 0.05$ # 5%

(critical) value $\rightarrow \text{dof} = \text{len(O)} - 1$

$\rightarrow \text{Crit Cal value} = \text{stat.chisq.ppf(0.05, dof)}$
percent point function

if chisquare stat \geq CV
reject
else: accept

Conclusion

* F-distribution / F-ratio

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

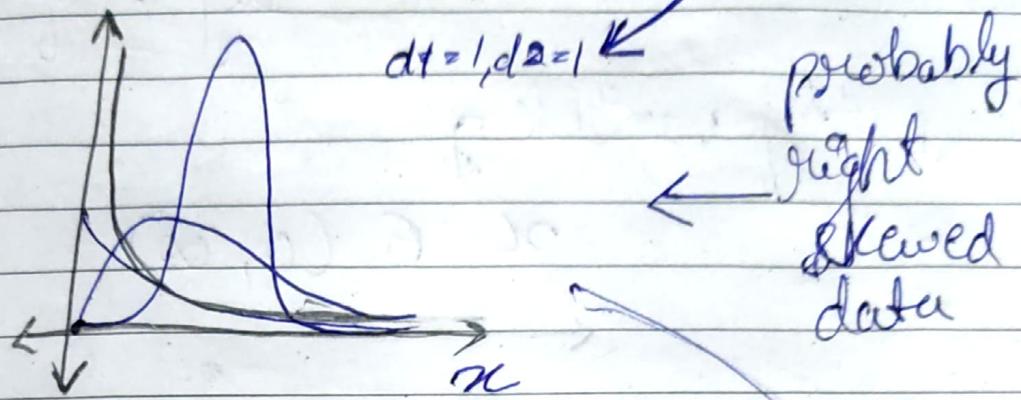
we use
continuous random
variable

used for hypothesis testing
to analyse the variance &
mean-blw 2 different groups

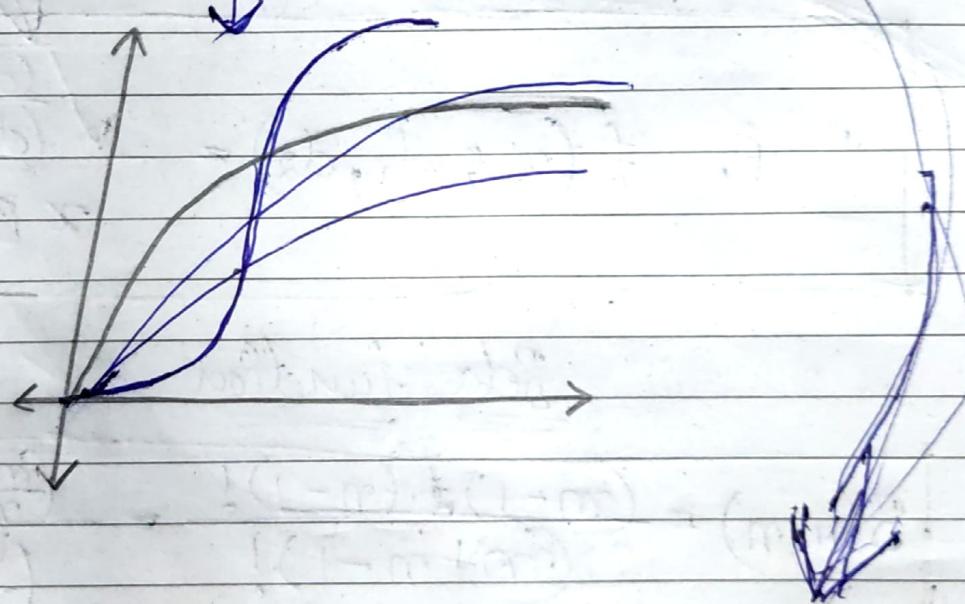
* From now we use F distributions
ANOVA, --

Fisher = Snedecor

Probability density function [PDF]



CDF



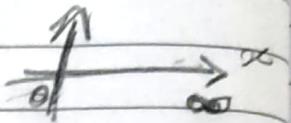
* In F distribution we also have 2 degrees of freedom [d₁ and d₂]

because using F distribution we compare 2 groups of data.

* Parameters $\equiv d_1, d_2 > 0$ ↴
degree of freedom

* Supporting variable ↴

$$x \in (0, \infty)$$



* Curve is created using PDF ↴

$$\text{PDF, } f(x; d_1, d_2) = \frac{(d_2)^{d_1} x^{d_2}}{(d_1 x + d_2)^{d_1+d_2}} \cdot x \cdot B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

Beta function

$$B(m, n) = \frac{(m-1)! \cdot (n-1)!}{(m+n-1)!} = \frac{\frac{(m+1)}{mn}}{\frac{(m+n)}{m}}$$

* F-distribution with d_1 and d_2 degree of freedom is the distribution of

$$F = \frac{s_1^2/d_1}{s_2^2/d_2}$$

$d_1, d_2 \neq$ degree of freedom

$d_1, d_2 \neq$
independent
random variable
with chi square distribution

* inv F-test

→ Called as "Variance ratio test".

→ Used to compare the variance ~~between~~ of 2 groups.

Ex:

- ① The following data shows the no of bulbs produced daily for some days by 2 workers A and B.

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

(Can we consider based on the data? Worker B is more stable and efficient).

$$\alpha = 0.05$$

Ans:

Variance

① Null hypothesis: $H_0 \Rightarrow \sigma_1^2 = \sigma_2^2$

alternative hypothesis: $H_1 \Rightarrow \sigma_1^2 \neq \sigma_2^2$

② Calculation of variance

~~Method~~

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

worker A

x_i	\bar{x}_1	$(x_i - \bar{x}_1)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
$\bar{x}_1 = 37$		$\Sigma = 80$

worker B

x_i	\bar{x}_2	$(x_i - \bar{x}_2)^2$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
$\bar{x}_2 = 37$		$\Sigma = 84$

$$\sigma_1^2 = \frac{80}{8-1} = \underline{\underline{13}}$$

$$\sigma_2^2 = \frac{84}{8-1}$$

$$\sigma_2^2 = \underline{\underline{12}}$$

③ Calculation of variance ratio test (F-test)

$$F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{13}{12} = \underline{\underline{1.33}}$$

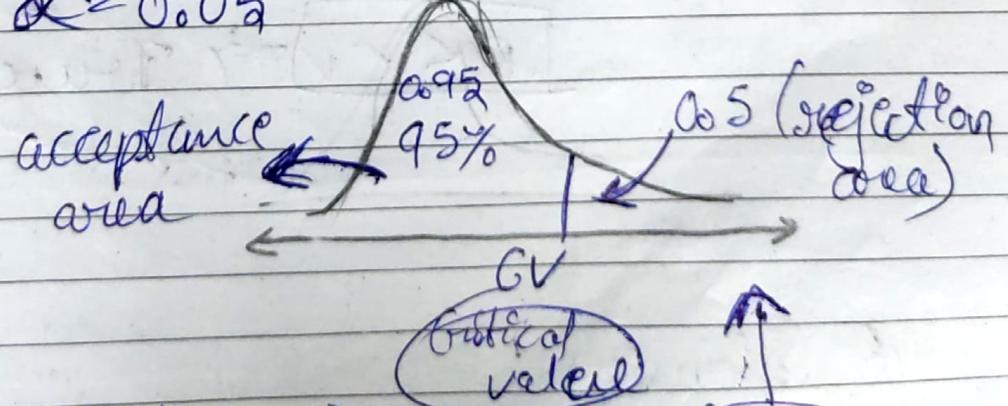
④ Decision rule

$$dof 1 = 6 - 1 = \underline{\underline{5}}$$

$$dof 2 = 8 - 1 = \underline{\underline{7}}$$

use
F-table

$$\alpha = 0.05$$



we get →
from F-table

↑
right
as skewed

$$\boxed{\begin{array}{l} CV \Rightarrow \\ \text{for} \\ \alpha = 0.05, \\ d_1 = 5 \\ \text{and } d_2 = 7 \end{array}}$$

3.9715

decision

if F test is greater than 3.9715,
we reject null hypothesis

~~Decision~~

Now,

F we found as 1.33
which is less than 3.9715
so, we accept hypothesis.

$$F = 1.33 < 3.9715$$

$$\boxed{F < CV} \Rightarrow \text{True}$$

we fail to reject null hypothesis

⑥ Conclusion:

worker B is not more stable / efficient when compared to worker A.

* F-test in python

fstat \rightarrow fstat = np.var(worker1) /
np.var(worker2)

$$\gamma \text{ dof1} = \text{len}(w1) - 1$$
$$\text{dof2} = \text{len}(w2) - 1$$

~~critical
value~~

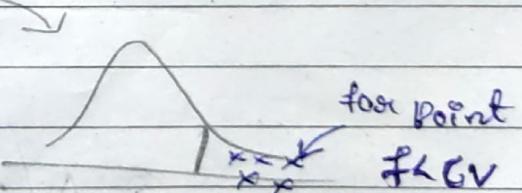
$$\text{Sign} \approx 0.05$$

$$COP \approx 1 - \text{Sign}$$

(Percent point function)

\rightarrow CV = stats.f.ppf(q=1 - sign, dfn=dof1,
dfd=dof2)

If $f < CV$
~~Conclusion~~ # reject
else: # accept



int

*ANOVA [Analysis of Variance]

Def : ANOVA is a statistical method used to compare the means of 2 or more groups.

ANOVA

① Factors (Variable) → can be independent & dependent

② levels

Medicine (Factor)

[Dosage] 5mg 10mg 15 mg ↴
levels

② Mode of payment (Factor)

Levels G.PAY Phonepe IMPIS DD NETT

may be
Gut feel (intuition)

Assumption in Anova

- ① Normality of sampling distribution of Mean

[CLT]

The distribution of sample mean is normal distributed.

- ② Absence of outliers.

Outling scores need to be removed from the dataset.

- ③ Homogeneity of Variance

→ population variance on different levels of each independent variable are equal.



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

- ④ Sample are independent and random.

Hypothesis testing in ANOVA

[Partitioning of Variance in the ANOVA]

Null hypothesis: $H_0 \Rightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

Alternate hypothesis: $H_1 \Rightarrow$ at least one of the sample mean is not equal

~~$\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_k$~~

~~wrong~~ because we are telling all are not equal to each other

Test Statistical

We use F test

$F = \frac{\text{Variance between samples}}{\text{Variance within sample}}$

Variance between samples \Rightarrow Here we calculate variance of each sample and then we find out difference

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	9	8
3	1	4
$\bar{x}_1^2 = 3$	$\bar{x}_2^2 = \frac{19}{6}$	$\bar{x}_3^2 = 4$

Variance within sample

how much far the point from the mean

Note, on 22

$$H_0: \bar{x}_1 = \bar{x}_2 = \bar{x}_3$$

H_1 : at least 1 sample mean is not equal

One Way ANOVA

① Doctors want to test a new medication which reduces headache. They splits the participant into 3 conditions [15 mg, 30mg, 45mg]. later on the doctor ask the patient to state the headache between [1-10]. Are there any differences between the 3 conditions using $\alpha=0.05$?
[Alpha]

~~means~~ that one factor with at least 2 levels, levels are independent.

Std tells - how far a point from mean.

Variance tells - the spread of data.

Ques 6

	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	5	2	

① Define Null & alternate hypothesis?

$$H_0: \mu_{15} = \mu_{30} = \mu_{45}$$

$H_1:$ ~~at least~~ not all are equal
(or) at least one is not

② Significance level $\alpha = 0.05$ $C.I = 0.95$

③ degree of freedom

$$\text{Population size} = 21, \alpha = 0.3, n = 7$$

$$df_{\text{between}} = a - 1 = 3 - 1 = \underline{\underline{2}} \quad \left. \right\}$$

$$df_{\text{within}} = N - a = 21 - 3 = \underline{\underline{18}} \quad \begin{matrix} \\ \\ 18 + 2 = 20 \end{matrix}$$

$$\therefore df_{\text{total}} = N - 1 = \underline{\underline{20}}$$

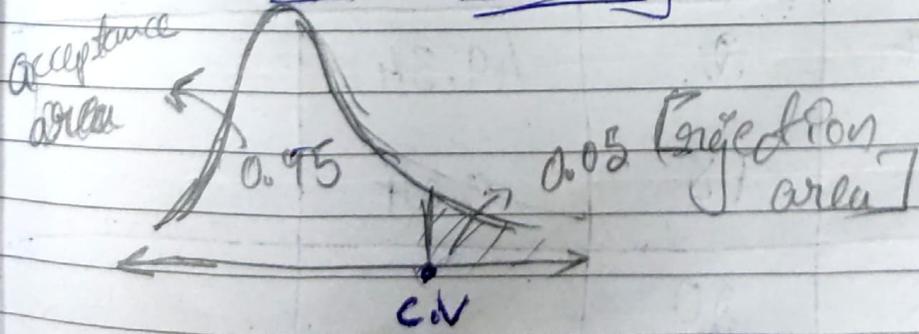
$$(2, 18) \Rightarrow (df_1, df_2)$$

↓, ↓

used on F-table with $\alpha = 0.05$

Find
critical
Value

① Decision boundary



Critical value
 when $\alpha = 0.05$, $df_1 = 2$ $\Rightarrow \underline{\underline{3.15546}}$
 $df_2 = 18$

~~Conclusion~~
decision rule

If F is greater than 3.15546 ,
 reject the Null hypothesis

⑤ Calculate F-test statistics.

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

	(SS)	(df)	(MS)	(F)
Between	98.67	2	49.34	
Within	10.51	18	0.54	
Total	109.18	20 <small>(before calculate)</small>	$\frac{(SS)}{(df)}$	

$$(i) \text{ SS between} = \frac{\sum (\sum a_i)^2 - \frac{T^2}{N}}{n}$$

Sample Total

$$15 \text{ mg} = \frac{9+8+7+8+8+9+8}{7} = \underline{\underline{57}}$$

$$30 \text{ mg} = 7+8+6+7+8+7+6 = \underline{\underline{47}}$$

$$16 \text{ mg} = 1+3+2+3+4+3+2 = \underline{\underline{21}}$$

$$\therefore \text{SS between} = \frac{57^2 + 47^2 + 21^2 - [57^2 + 47^2 + 21^2]}{21}$$

$$\boxed{SS_{\text{between}} = 98.67}$$

$$(ii) \text{ SS}_{\text{within}} = \sum y^2 = \frac{\sum (\sum a_i)^2}{n}$$

$$\sum y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + 8^2 = \underline{\underline{853}}$$

$$\therefore \text{SS}_{\text{within}} = 853 - \frac{57^2 + 47^2 + 21^2}{7}$$

$$\boxed{SS_{\text{within}} = 10.29}$$

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

$$\frac{19.34}{0.54} = \underline{\underline{36.56}}$$

If F is greater than 3.5546 then we reject the H_0 (null)

Here, ($F=36.56$) > 3.5546 so we reject null hypothesis (h_0)

⑥ Conclusion

not all are equal

($\alpha \neq 0$)
at least one is not equal