Statistics HandWritten Notes

- Darshan R M

"Learn, Share, and Collaborate"

Statistics

Basic

① Introduction to Statistics.

② Types of Statistics.

(i) Descriptive stats.

(ii) Inforential 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:

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

② Sample mean $(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$ ← Sample size

(ii) Median

① N is odd: $\frac{N}{2}$

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

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

(iii) Mode:

Element with Max Frequency.

⑥ Variable v/s Random variable.

⑦ Set

(i) union
(ii) Intersection
(iii) Difference
(iv) Subset
(v) Superset

⑧ Covariance and Correlation

* To find relationship between 2 variables.

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

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

disadv: No limit to range

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

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

$\sigma \Rightarrow$ standard deviation
$Cov \Rightarrow$ Covariance

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

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

$R(x) \Rightarrow$ Rank of x

(9) Measure of dispersion

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

① Population variance

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

$N \rightarrow$ population size.
$\mu \rightarrow$ population mean.

② Sample variance

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

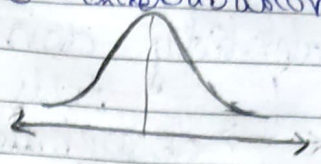
$\bar{x} \rightarrow$ sample mean
$n \rightarrow$ sample size
* $(n-1)$ because bessel correction

(ii) Standard deviation :

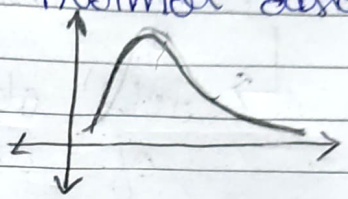
* standard deviation, $\sigma = \sqrt{\text{Variance}}$

(10) Skewness

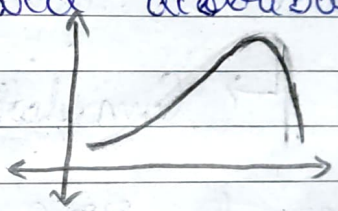
(i) Symmetric distribution.



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



(iii) left skewed distribution.



(11) Histogram.

* KDE
  - help to → Smoothen the histogram
  - we get → PDF

Advance - 1

① Probability density Function [PDF]
↳ tells about distribution of continuous data.

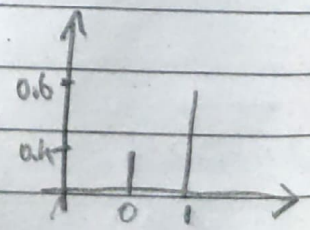
② Probability Mass Function [PMF]
↳ tells about distribution of discrete data.

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



$$q = 1 - p$$

PMF → $$p(x) = p^x q^{1-x} = p^x (1-p)^{1-x}$$

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

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

Variance, $\boxed{\sigma^2 = pq}$

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

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

⑤ Poission distribution.

- PMF
- Number of events occuring in a fixed time intervals

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

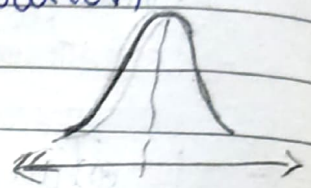
λ = Expected number of events to occur at every time interval

increases → shape of distribution become taller

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

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

⑥ Normal / Gaussian distribution



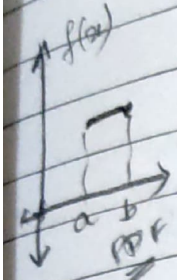
→ PDF
→ Empirical rule
68 - 95 - 99.7%
$(\mu + \sigma)$ $(\mu + 2\sigma)$ $(\mu + 3\sigma)$
→ Bell curve

⑦ Q-Q plot:
↳ to check whether data is normally distributed (or) not

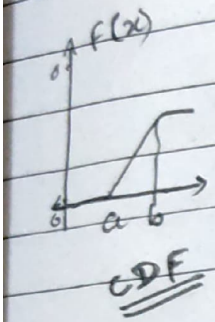
⑧ Uniform distribution

(i) Continuous uniform distribution
PDF ↳ describes an experiment where there is an arbitrary outcome in certain ranges

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



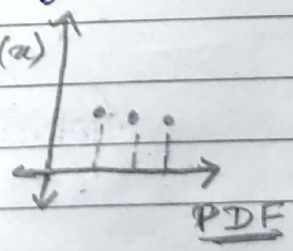
$$CDF = \begin{cases} 0, & \text{for } x < a \\ \frac{x-a}{b-a}, & \text{for } x \in [a, b] \\ 1, & \text{for } x > b \end{cases}$$



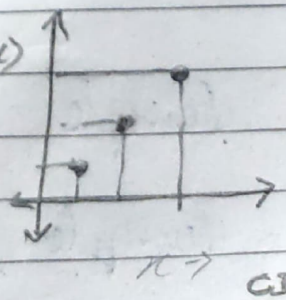
(ii) Discrete uniform distribution.

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

$$PMF = 1/n$$



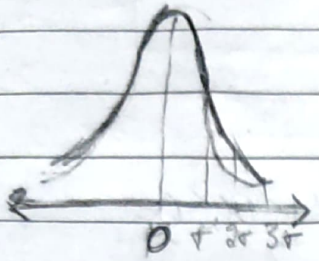
Mean & Median $= \frac{a+b}{2}$ $b \geq a$.



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

If 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) ⇒ [0, 1]

* Confidence interval, CI = 1 − α

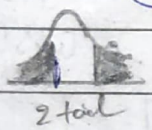
* Z test } Average
* t test }
chi-square → Categorical
ANOVA → ~~Variance~~ Mean/Average
F test → Variance.

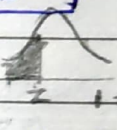
③ Z test

→ Condition
① Population std. [✓]
② $n \geq 30$

→ Steps to solve
define
① h_0 & h_1
② decision boundary → one tail / two tail

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





③ Find Z value
④ Find p (area)
→ from Z table with given 'α' & found 'Z'

⑤ output → $p < \alpha$

True [reject Null hypothesis]

False [fail to reject h_0]

④ Student t-test

→ Use when population std is not known [✓] (or)
$n < 30$

Z-test

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

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

$\bar{x}$ ⇒ sample mean

s ⇒ sample variance

↳ find using (α & dof)

degree of freedom = $n-1$

* Steps to Solve

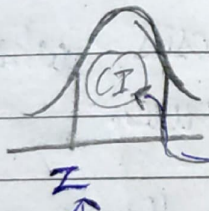
① define h_0 & h_1

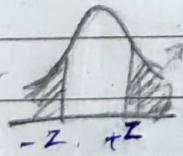
② α = ___

③ degree of freedom = ___

④ Decision boundary/rule

⑤ Finding Z score ∞ ↳ using (α and dof) in t table





* define decision rule [Condition]

ex: $t > +Z$ / $t < -Z$ ↳ reject

⑥ Calculate t using formula

⑦ check (Condition) decision rule

#

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

Mean $\mu = np$

$(p+q)^n = 1$ Variance $\sigma^2 = npq$

(5) Binomial distribution
- → extension of bernoulli's
- → No of sucess in fixed no independent trials

* CDF → discrete ⇒ Σ
→ Continous ⇒ $\int$

Advance - 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 ± margin of Error

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

$Z_{\alpha/2}$ → Z-score

③ Bayes theorem.

↳ independent & dependent variables.

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

$P(B/A)$ → Probability of B when A happened.

④ chisquare ~~test~~

↳ "chisquare test for goodness of fit"

↳ for - Categorical data

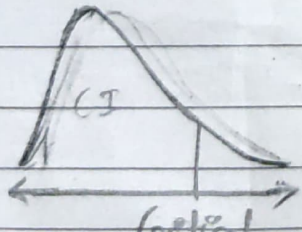
↳ steps to solve

① Define h_0 and h_1.

② $\alpha = $ ___ , $CI = 1 - \alpha$

③ degree of freedom $= n - 1$

④ Decision boundary



- Find critical value using (α and dof) using chisquare table.

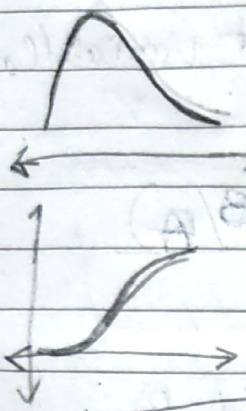
⑤ Calculate chisquare test value (define decision rule)

$\chi^2 \geq CV$ reject

$$\chi^2 = \frac{\sum (observed - Expect)^2}{Expected}$$

⑥ Conclusion

F - test



→ Continuous random variable
↳ probability distribution.

→ To analysis of variance b/w 2 groups.

→ Parameter: d_1 & d_2

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

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

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

steps to solve

① Define h_0 and h_1

② Calculate variance

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

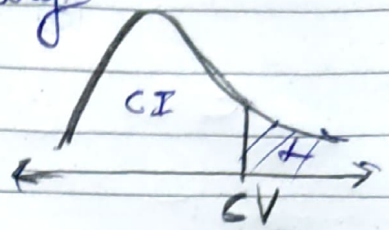
group 1 s_1^2 group 2 s_2^2

③ Calculation of variance ratio test (F test)

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

④ Decision boundary

dof 1 = ___
dof 2 = ___
α = ___



* Find CV. using (dof1, dof2 & α) in F-test.
* define decision rule

⑤ Conclu

IF < CV

True [reject H_0]
False [Fail to reject H_0]

⑥ ANOVA

→ To compare mean of 2 group
① Factors / variables
② levels.

* 4 Assumptions

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

* Steps to solve

$\mu = \mu = \mu$

① Define H_0 and H_1

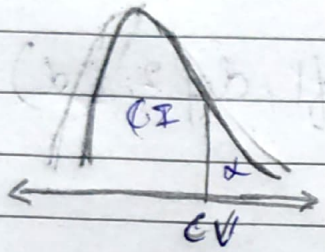
② α = ____ , C.I = 1 − α = ____

③ degree of freedom

$df_{between} = \underset{\text{no of level}}{a} - 1$ (df1)

$df_{within} = \underset{\text{Population size}}{N} - a$ (df2)

$df_{total} = N - 1$



④ Decision boundary

* Find CV using (df1, df2 & α) in F-Table

* define decision rule. $F > CV$ reject

⑤ Calculate F test stats

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

	Sum of Square [SS]	degree of freedom	Mean square [MS]
Between	$SS_{between}$	df1	SS/df
within	SS_{within}	df2	
total			

a = sum of value each level

Y = sum of square of value of any one level

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

$$SS_{within} = \sum Y^2 - \frac{\sum (\sum a_i)^2}{n}$$

⑥ Conclusion

10

Statistics

Week - 10

Basic Statistics

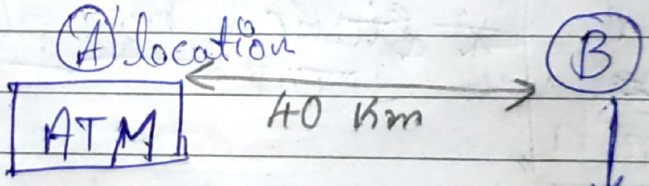
Defn: Statistics is the science of collecting, organizing and analyzing the data
→ decision making process

* **Data:** "facts or pieces of information."

ex: Heights.
{ 175 cm, 180 cm, ---}

IQ
{ 200, 300, ---}

ex:

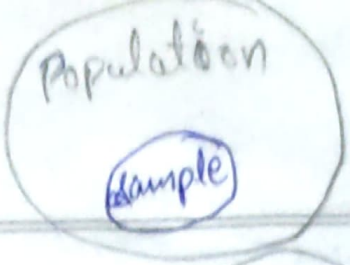


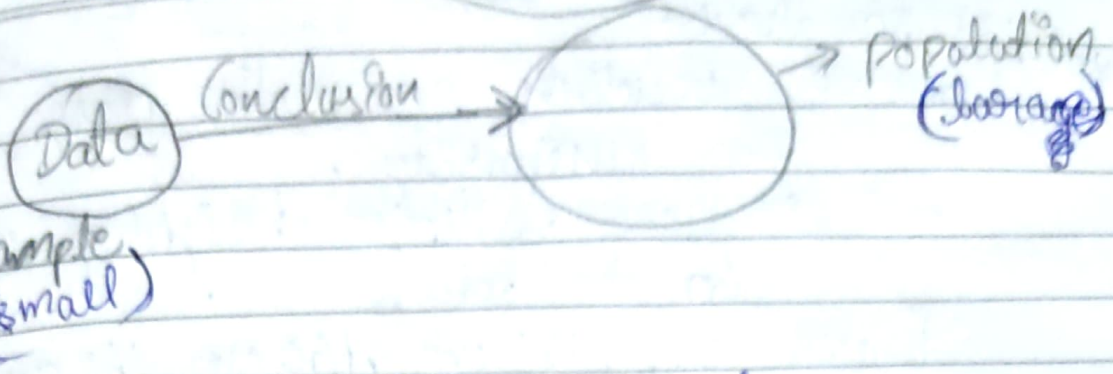
bank to be opened (or) not

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



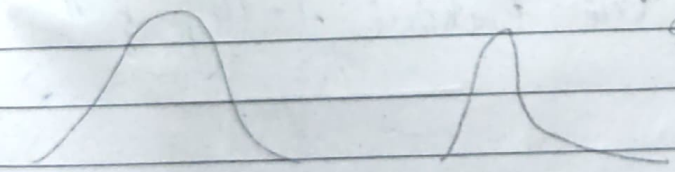


Sample (small)

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

② Measure of Dispersion
[variance, std] → used to understand how data is distributed.
↓
how our data is well spread

③ Histograms, Bar chart, Pie chart ↓
summarizes the data
– what is distribution/ relative dis – – –



different graph

④ For Conclusion, we need to do experiments. like →
① Z- test
② t test
Hypothesis testing, p value, significance,

Ex: Let say there are 50 students in maths class in university. (sample data)
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?" [mean]
* "what is common height of the students?" [mode]
(descriptive info)

⇒ Inferential Question
* "Are the average height of students in the classroom (sample) similar to what you expect in the entire college (population)?"

Ex: Sample data and population data
exit poll of election

Conclusions are drawn from sample data	Sample [few people voted]	State [Population]

Types of data

Col:	DMC	FWI	Regions
EDA step 1	3.4	1	0
	4.1	2	0
		0	0
		2.1	0

Numerical data → Data ← Categorical Data

Data
- Quantitative
  - Discrete
    - whole numbers [with specific range]
    - Ex: Number of bank accounts
      - 2 ✓
      - 2.5 ✗
  - Continous
    - Any value decimal
    - Ex: weight, speed, height
      - ✓ 75, 76.1, 26.9, 61.5
- Qualitative
  - nominal
    - Ex: Gender [M/F]
    - Blood group [A, B, AB]
    - * Cannot say M > F ✗ gender
    - Cannot rank
  - ordinal
    - Ex: Bank
      - 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?
⇒ Continous. [Quantitative]

③ what kind of variable movie duration?
⇒ Continous. [Quantitative]

Scale of measurement of data → helps to determine the types of statistical operation can be performed.

① Nominal Scale Data.
② ordinal Scale Data.
③ Interval Scale Data.
④ Ratio Scale Data.

① Nominal Scale Data:

* Qualitative / Categorical variable
* eg: gender, colors, labels.
* 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:

* Ranking and order matters.
* Difference cannot be measured.

Ex: Qualification → Rank

PHD	1st
BE	3rd
Masters	2nd
BCom	5th
BSc.	4th

Can be considered as ordinal scale of data.

Cannot find difference (Can't rank)

Normal data can be convert into ordinal scale data.

Can Calculate difference

Cannot Consider as ordinal scaled data

Works

100	1st
90	2nd
65	4th
70	3rd
30	5th

imp

(3) Interval scale data:

* Rank & order matters.
* ✓ Difference can be measured.
* ✗ Ratio cannot be measured.
* Donot have any '0' starting value.
  ↳ Can have any value may be -ve → -3F, +ve → 6F

Ex: Temperature

30F
60F
80F
90F

difference
60 - 30 → 30

Cannot measure here

ratio = 60/30 = 2

it will not tell the ratio correctly i.e the heat felling on 60F is double the heat felling on 30F. which is wrong

so ratio donot matter.

④ Ratio Scale data:

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

Ex: Simple Grades

Grades	order	difference	rank
100	100	10	1st
90	90	30	2nd
60	60	10	4th
50	50	5	5th
45	45	10	6th
35	35		

Ratio

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

↳ 100 : marks is double the 50 marks. which is Correct.

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

Ex:

① length of different rivers in the world ? ⇒ Ratio scale data

② Marital status ? ⇒ Nominal scale data

③ IQ measurement ? ⇒ Ratio scale data.

Measures of Central tendency

① Mean. ② Median. ③ Mode.

[imp for EDA and feature engineering]

* Population (N) | Sample (n)

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

① Population mean $(\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$ | Sample mean $(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$

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

$= \frac{32}{10}$

$= 3.2$

⑨ Median (.

* Sorted: 2, 2, 3, 4, 5

Q) 4, 5, 2, 3, 2, 1 — even

1, 2, [2, 3], 4, 5

→ average $= \frac{2+3}{2} = 2.5$

∴ median = 2.5

* {1, 2, 2, (3), 4, 5, 7} — odd

∴ median = 3

* Why median even after mean is present.

① ⇒ $\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

{1, 2, 3, 4, 5}

median = 3 ✓

1, 2, (3), 4, 5 ✓

② ⇒ but what if outlier is present in sample.

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

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

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

= 19.75 ≈ 19

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

median = $\frac{7}{2}$ = 3.6 ✓

In ① & ② → outlier present

∴ mean is very different in both.

but even in presence of outlier median is deviated a little.

③ Mode ⇒ (Frequency)

↳ Max Frequency

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

mode = 1 occured 3 times

where mode is used?

Nominal ⇒ Types of flower

Types of flower	Age
daisy	10
Rose	3
Rose	5
Sunflower	-
Rose	8

handle

① drop row but loss of data

② Replace with mode value i.e. Rose

→ if outlier replace with median else with mean.

*Practical

import numpy as np

numpy { ① np.mean (data)
② np.median (data)

scipy - ③ stats.mode (data) mode.mode[0]

from scipy import stats

Random variables

* Variables & RV are different

x + y = 7
x + y = 10
} → Here x & y are just variables.

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

ex: ① Tossing a coin.
↳ 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 = {Sum of the rolling of dice 7 times}

{4, 5, 6, 1, 2, 2} = 20 (sum)

= other value

A random variable can take any outcome based on case/process but variable take only one value & it is fixed.

Ex: $P(Y \geq 15)$ $P(\text{"Head"})$

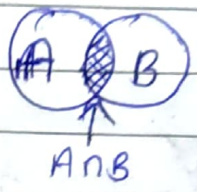
Sets

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

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

① **Intersection:** Common

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



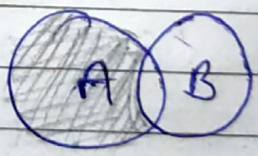
② **Union:** all

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

③ **Difference** present in set 1 not in set 2

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

$B = \{3, 5\}$



④ **Subset**

✗ $A \subset B$ [A is subset of B]

⇒ False, be 'B' donot have all the elements of 'A'

✓ $B \subset A$

⑤ Superset

✓ $A \supset B$ [A is a superset of B]

A Contain all elements of B.

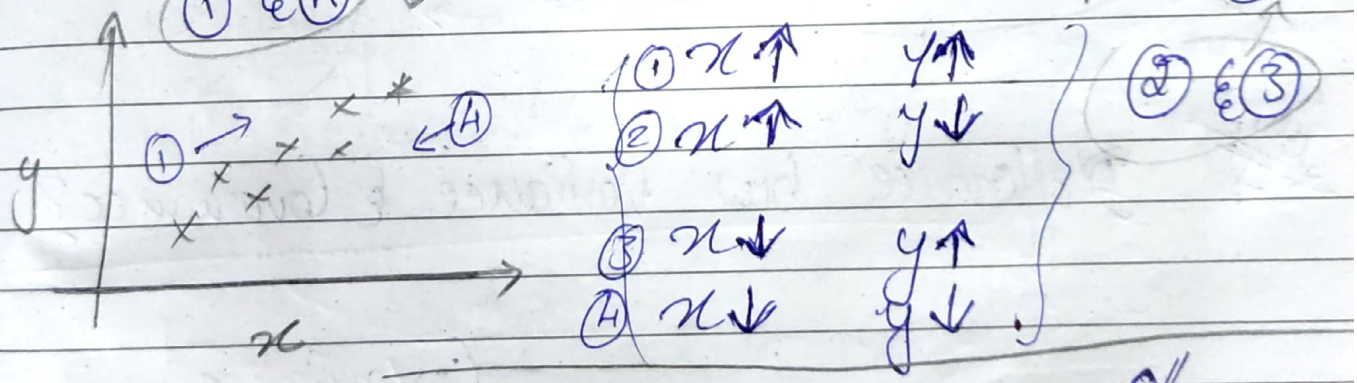
① Covariance and Correlation:

x	y
2	3
4	5
6	7
8	9

① & ④

Can we find out relationship b/w x & y

means



Ex: Size, location, Price o/p

2 Features are important

Determine price !...

like Ex: size↑ price↑

So, we need to apply Covariance.

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

(or)

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

spread of the data

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

checking Covariance of 'x' itself

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

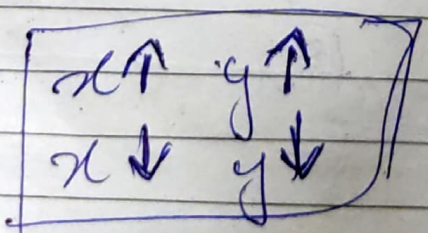
$$\text{Var}(x) = \text{Cov}(x, x)$$

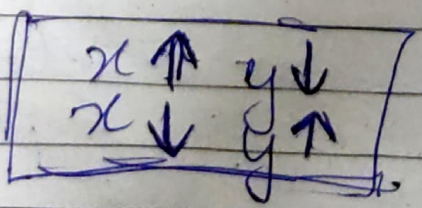
tells 'x' its own relationship.

Q) Difference b/w variance & Covariance?

Covariance means

+ve → positive relationship





-ve

x	y
2	3
4	5
6	7

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

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

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

$$= \frac{4+0+4}{2}$$

$$Cov(x,y) = 4$$

∴ Conclusion: x & y has +ve Covariance

Advantages

(i) Relationship b/w x & y → +ve / -ve

Disadvantage

(i) +100 Covariance of x y

+500 Covariance of x z

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

⇒ we need some restriction to this

① Covariance does not have specific limit value.

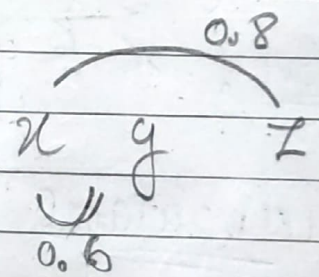
↓ To overcome this

we use other kind of Correlation

② Pearson Correlation Coefficient:

Can take any value in range b/w [-1 to +1]

As value is getting restricted.



Now, we can conclude same.

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

(Standard deviation)

① The more the value towards +1 the more +ve correlated it is

② The more it —"— -V

—"— -ve correlated it is.

ex: Dataset

1000 features

[For ML models to build we donot 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

③ Spearman Rank Correlation:

↗ work focus on x & y value
we focus of rank.

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

use

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

* we have to try both ② and ③

seaborn

* 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

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

where, x_i = data points

μ = population mean

N = population size

Sample Variance

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

[Bessels Correction]

why?

⇒ reason for we use $n-1$ rather than n is so that the sample variance will be what is called as unbaised estimator of the population variance.

⇒ x_i = data point

$\bar{x}$ = sample mean

n = sample size

Ex: Q) {1, 2, 3, 4, 5} ⇒ Sample

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

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

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

$$\bar{x} = 3.$$

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

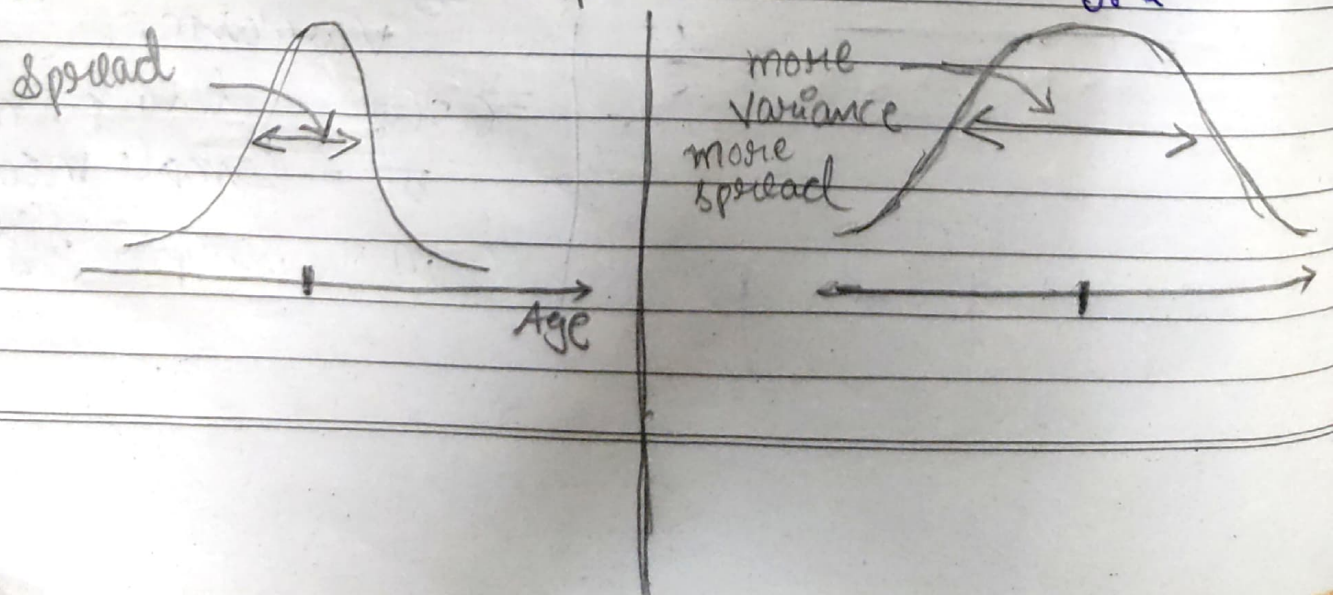
$$= 2.5$$

what it say?

⇒ Variance talks about the spread of data.

Ex: $\sigma^2 = 2.5$ ← less spread

$\sigma^2 = 6.5$ ← more spread



② Standard deviation

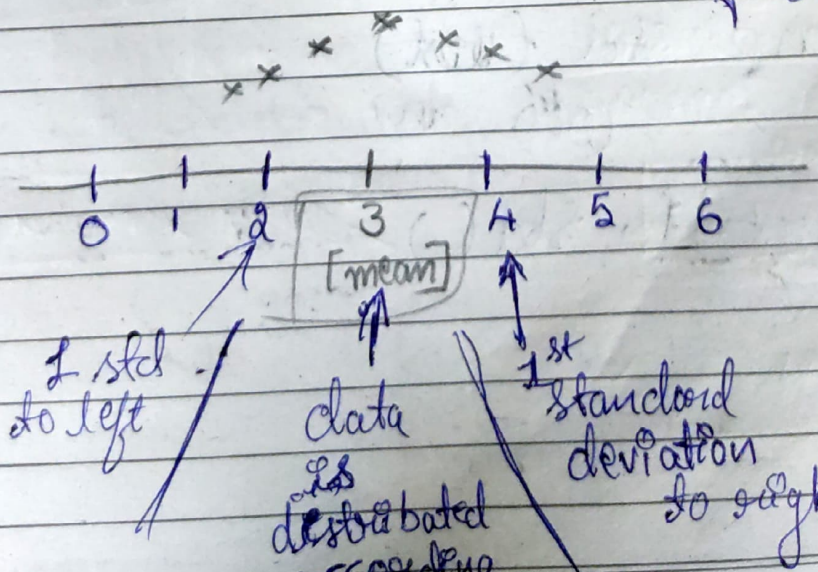
Population standard deviation	Sample standard deviation
$\sigma = \sqrt{\text{variance}}$	$s = \sqrt{\text{sample variance}}$
~~ex~~	ex: sample variance $s^2 = 2.5$
	sample stdl $= \sqrt{s^2}$

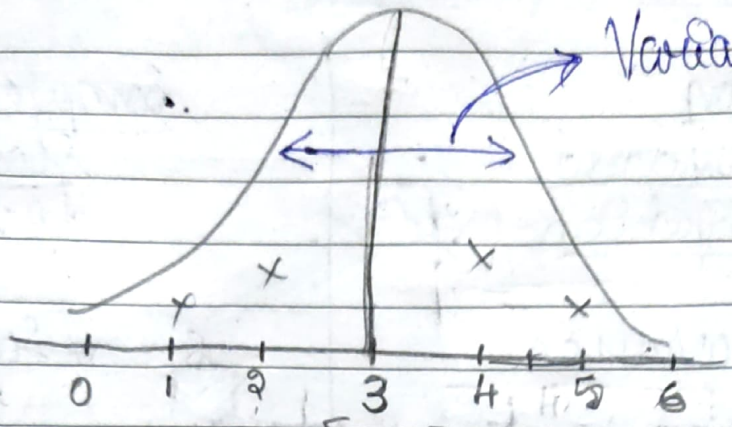
ex: Consider $\{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma = 1$

Most of the data will be falling in this 3 standard deviation





5 → is 2 std away from mean

numpy 1D

> array np.var(list)

(or)

Dataframe df.var() → default axis=0 → Column ||| axis=1 → row =

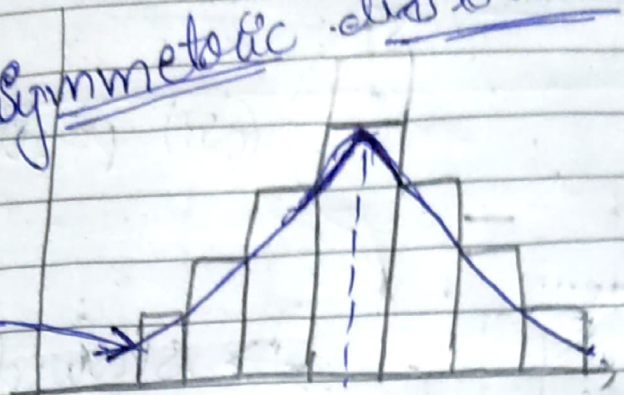
> np.std(list)

(or)

df.std()

Skewness

① Symmetric distribution:



when we smoothen this histogram we get

Symmetrical distribution → No skewness Here in SD.

area towards left = area towards right

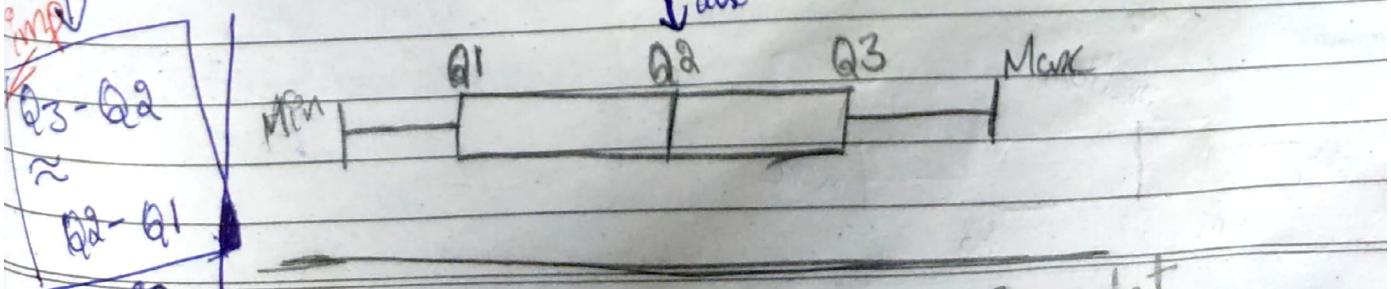
* The mean, median and mode all are perfectly at the center.

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

inter

mean = median = mode

(median) also



Box plot

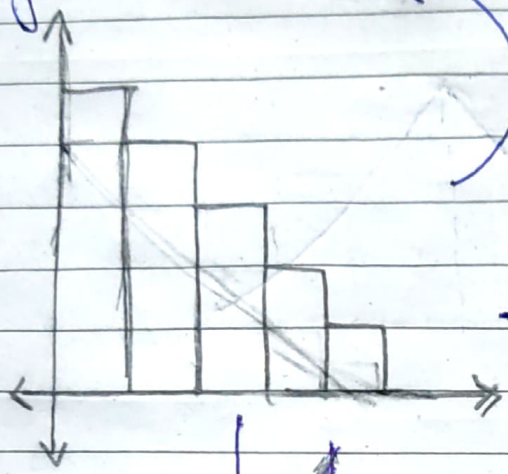
Most of data exist in our world follow this distribution

distribution bend towards right / +ve

Right [log Normal distribution]

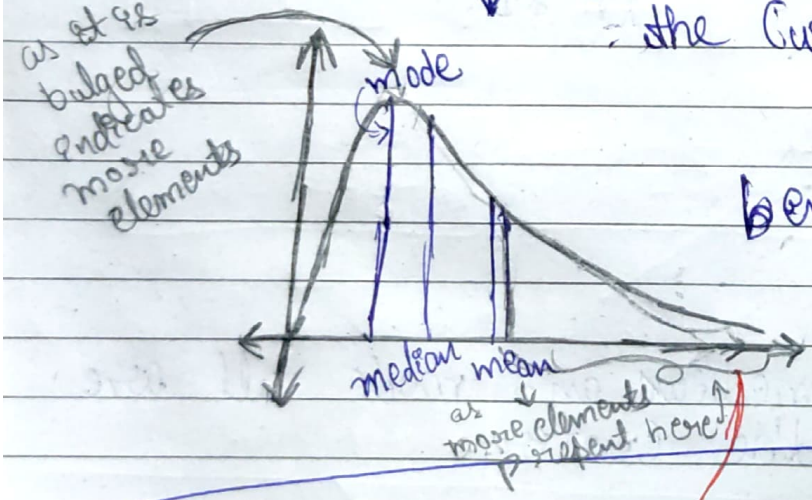
② Right skewed distribution:

(or) positive skewed



→ skewed towards right.

when smoother the curve

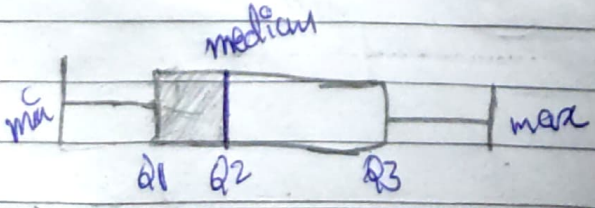


bent towards right so right / positive skewed.

when we draw box plot we get median somewhere towards left.

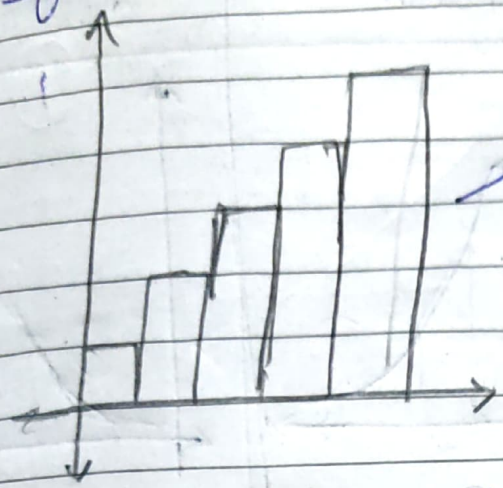
Q3-Q2 > Q2-Q1

mean > median > mode

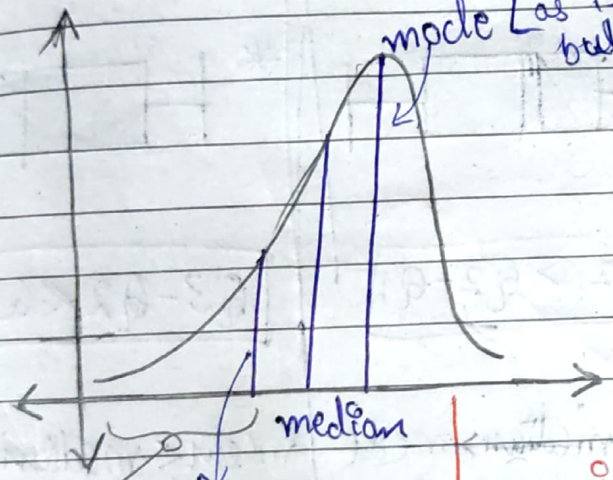


bend towards left /- ve.

③ left skewed distribution:



→ skewed towards left /- ve.



as more elements present

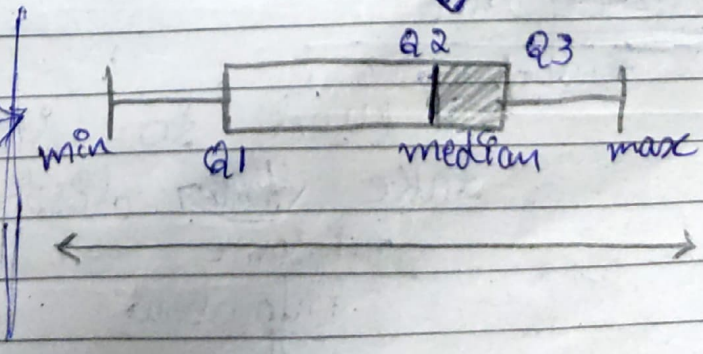
bend towards left so left /- ve skewed.

imp property

$Q3 - Q2 < Q2 - Q1$

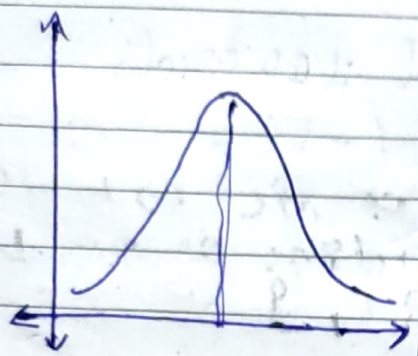
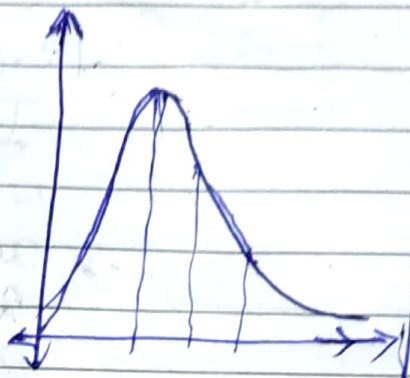
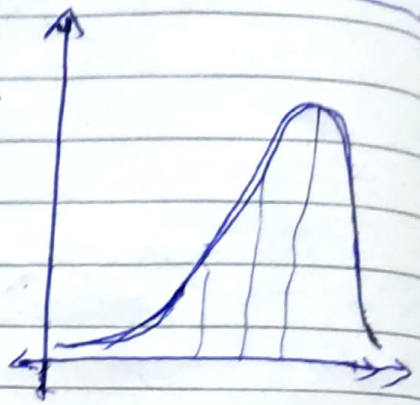
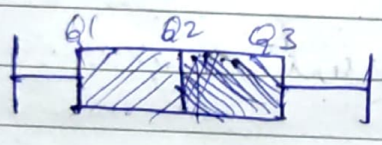
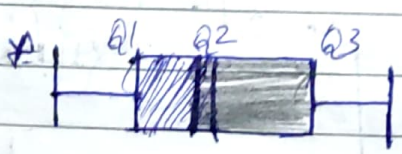
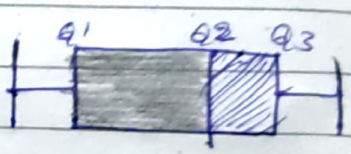
mean < median < mode

when we draw box plot we get median towards right.



interview

E log normal distribution

Symmetric distribution	Right skewed distribution	left skewed distribution
		
		
$Q3 - Q2 \approx Q2 - Q1$	$Q3 - Q2 > Q2 - Q1$	$Q3 - Q2 < Q2 - Q1$
mean = median = mode	mean > median > mode	mean < median < mode

Histograms

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

Histogram → basically deals with → Frequency of the data

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

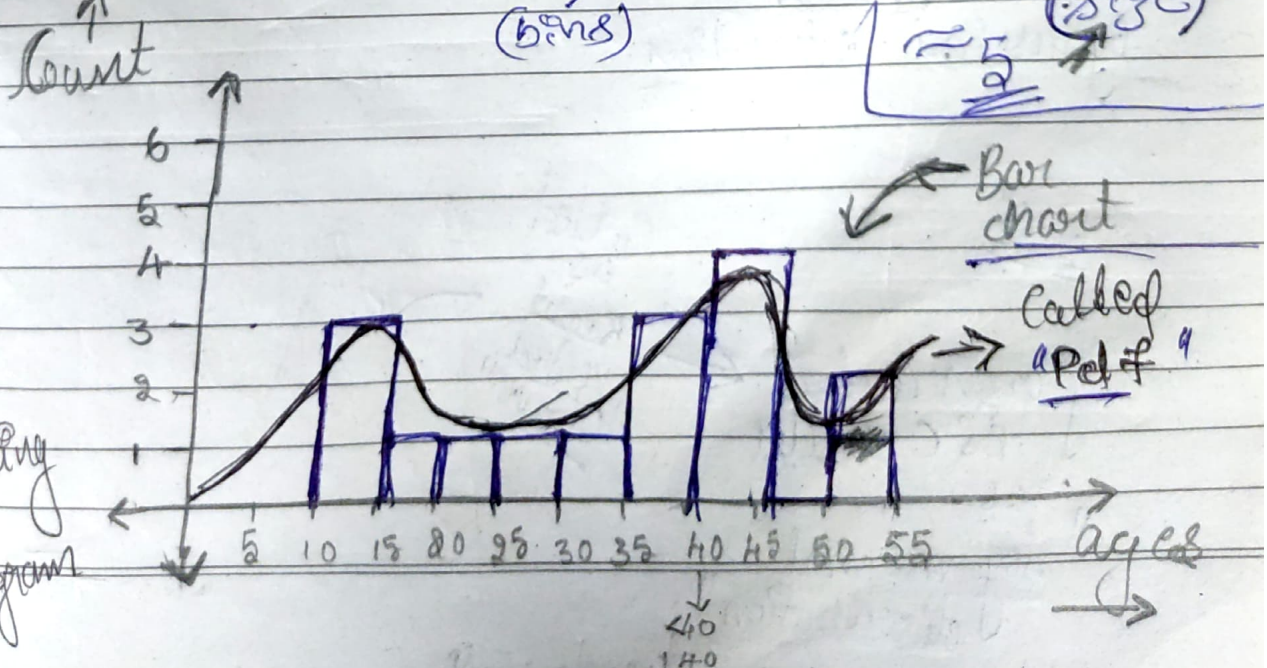
imp parameter Bins = 10 → we create 10 buckets according to our data and put data into that bucket ↓ and plot a graph

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

∴ bucket/size = $\frac{\text{highest} - \text{lowest}}{\text{bins}} = \frac{51 - 0}{10} = \frac{51}{10} = 5.1$ (size)

(bins)

≈ 5



How smoothing ⇒ KDE (Kernel density estimator)

By smoothing the histogram we get PDF.

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

* KDE → helps us to → smoothing of histogram → we get → PDF

Ex: Count / Frequency

Called as "Bell curve"

Graph contain this Bell curve called as Normal / Gaussian distribution

Smoothing of the histogram

Bar chart

Variable →

Ex:

Graph contain R&C called as "log normal distribution"

Called as "Right skewed curve"

* Why all this?

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

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

*

WEEK - 11

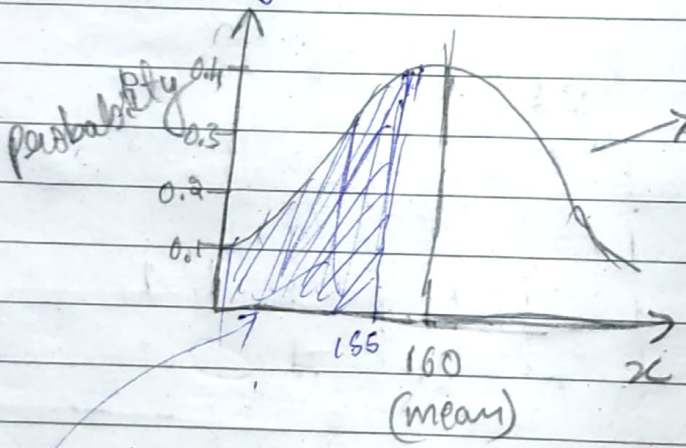
Statistics advance - 1

* (1) Probability distribution Function / Density Function:

[PDF] ↳ both are not same

① Probability density function → Tells about the distribution of continuous data.

ex: height of students.



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

using this distribution we can derive imp information.

ex: $P(x < 155)$ → Can be calculated like area under curve

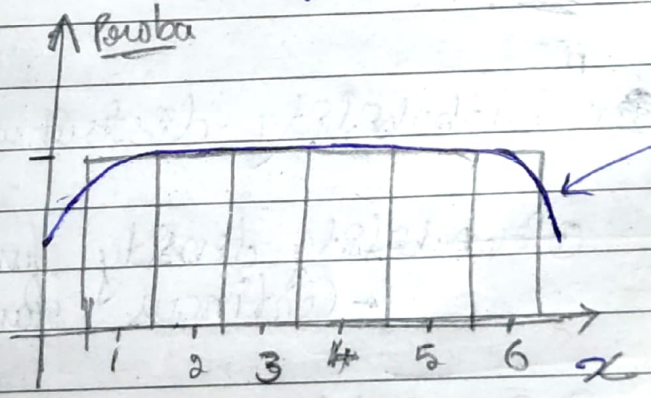
Probability distribution function contain various types.

① Prob density function. [PDF]
we try to create a curve for distribution of "Continous" data.

② Probability mass function [PMF]
we see the distribution of a "discreate" random variable.

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

For a fair/unbaised dice



Q) $P(x \leq 4)$

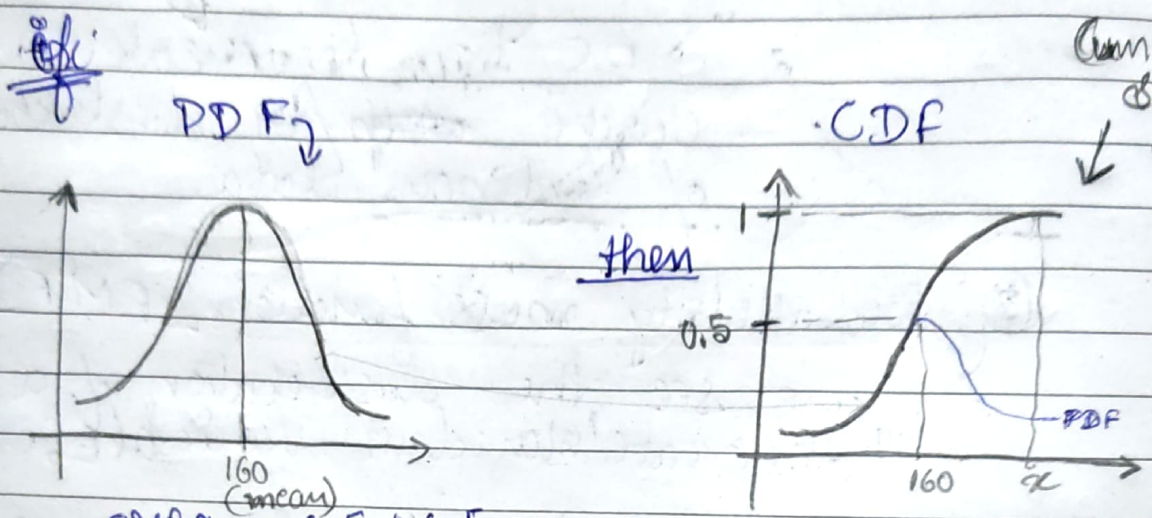
$= P(x=1) + P(x=2) + P(x=3) + P(x=4)$

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

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

③ Cumulative distribution Function (CDF):

To understand CDF



area = 0.5 + 0.5
= 1

- Probability distribution function

① Probability density function. [PDF]
- Continous random variable.

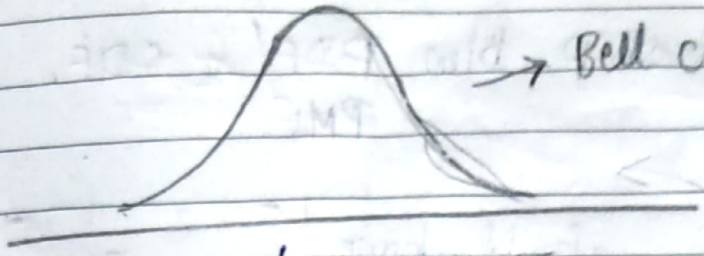
② Probability mass function [PMF]
- Discrete random variable.

③ Cumulative distribution function [CDF]
- just doing cumulative sum.

Imp

Types of probability distribution

① Normal / Gaussian distribution. [Part of PDF]



↓
we can do many assumptions

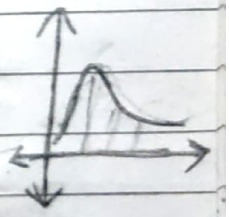
Most dataset follow this distribution

use discrete data

② Bernoulli distribution [Part of PMF]
↓
Outcome are only 2. → Success. / Failure.

③ Uniform distribution.

④ log normal distribution [Part of PDF]



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 follows gaussian Some distribution so convert it works good

* Probability density function [PDF] and Probability mass function [PMF]

learn about
* relationship blw PDF/PMF & CDF.

① PMF:
* should have Discrete random variable.

PMF

ex: rolling a dice
{1, 2, 3, 4, 5, 6}

$P(x=1) = 1/6$
$P(x=2) = 1/6$
1 }

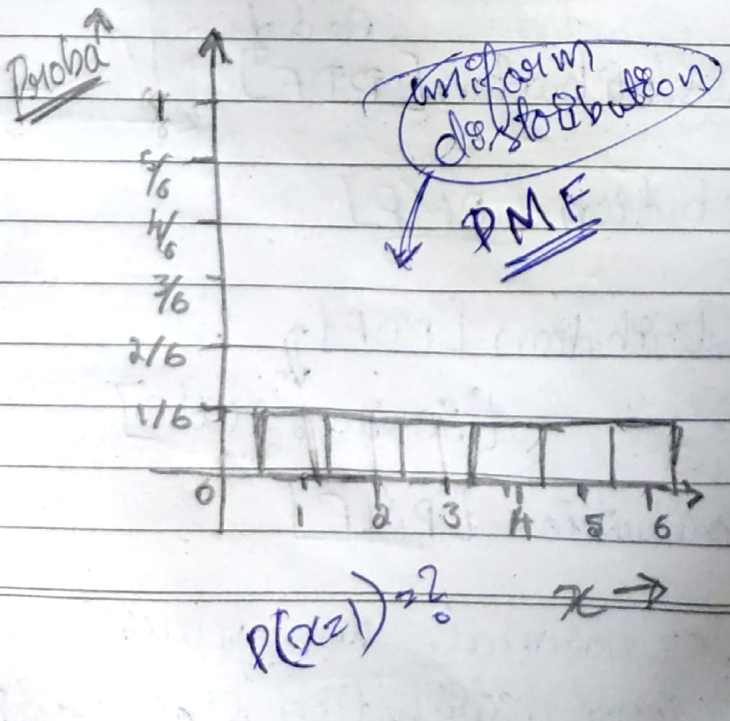
$$\sum_{i=1}^{n} P(x_i) = 1$$

CDF

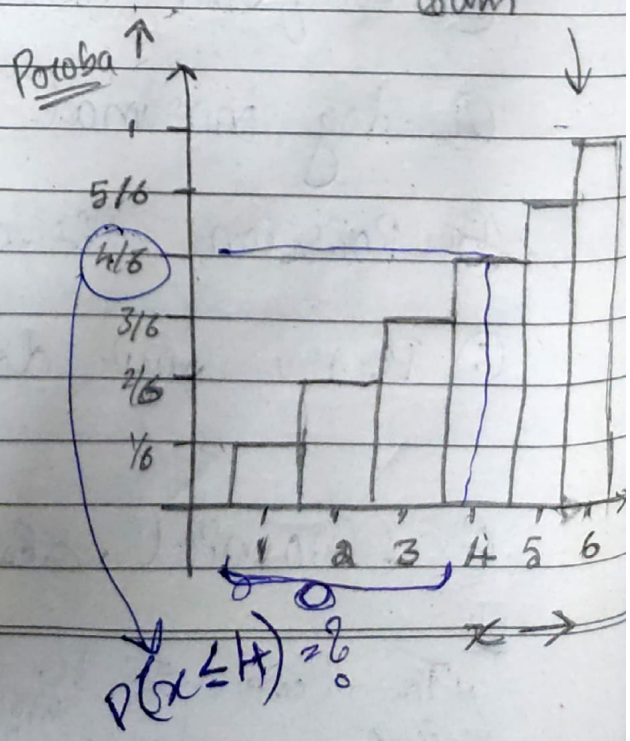
ex: rolling a dice

x	P	Cumulative
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6 = 1

$\sum P(x_i) = 1$



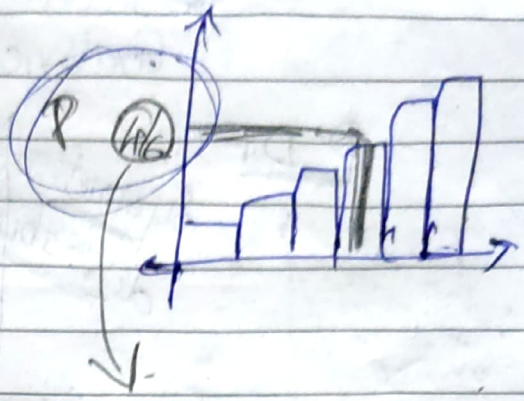
Cumulative sum



$P(x=1) = 1/6$
$P(x=2) = 1/6$ } easy can find.

but

$P(x \leq 4)$ → so we have to look at CDF



use CDF

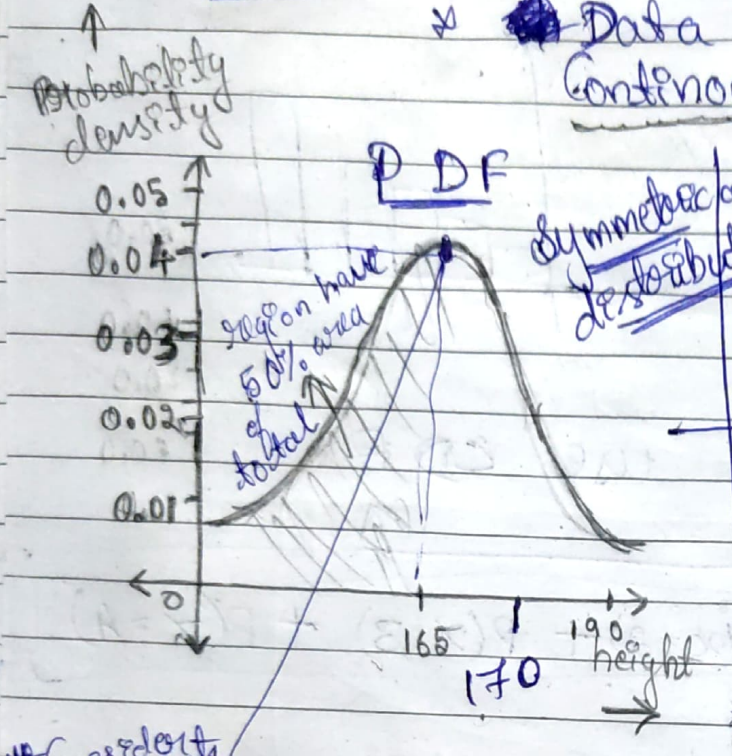
$$P(x \leq 4) = P(x=1) + P(x=2) + P(x=3) + P(x=4)$$

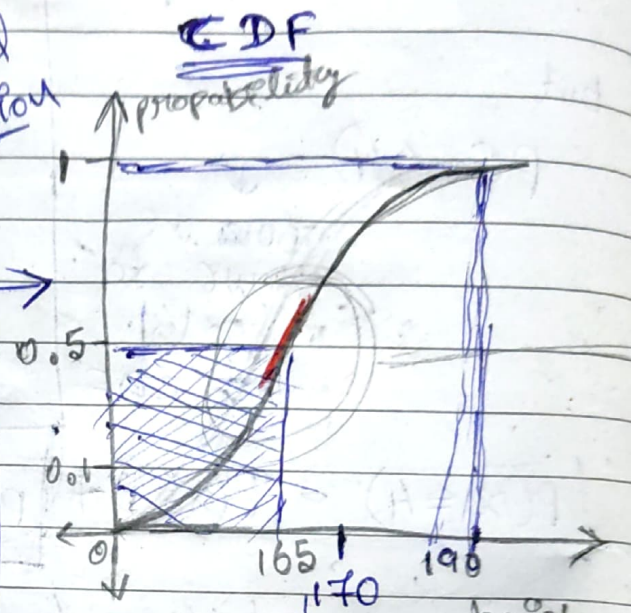
* For different different distribution we can have different formula.

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

② PDF:

* Data should be a continous random variable.





50% of distribution is there between this region.

* Consider this point. Does this point mean that probability density at 165cm is 0.04?

NO

does not make any sense

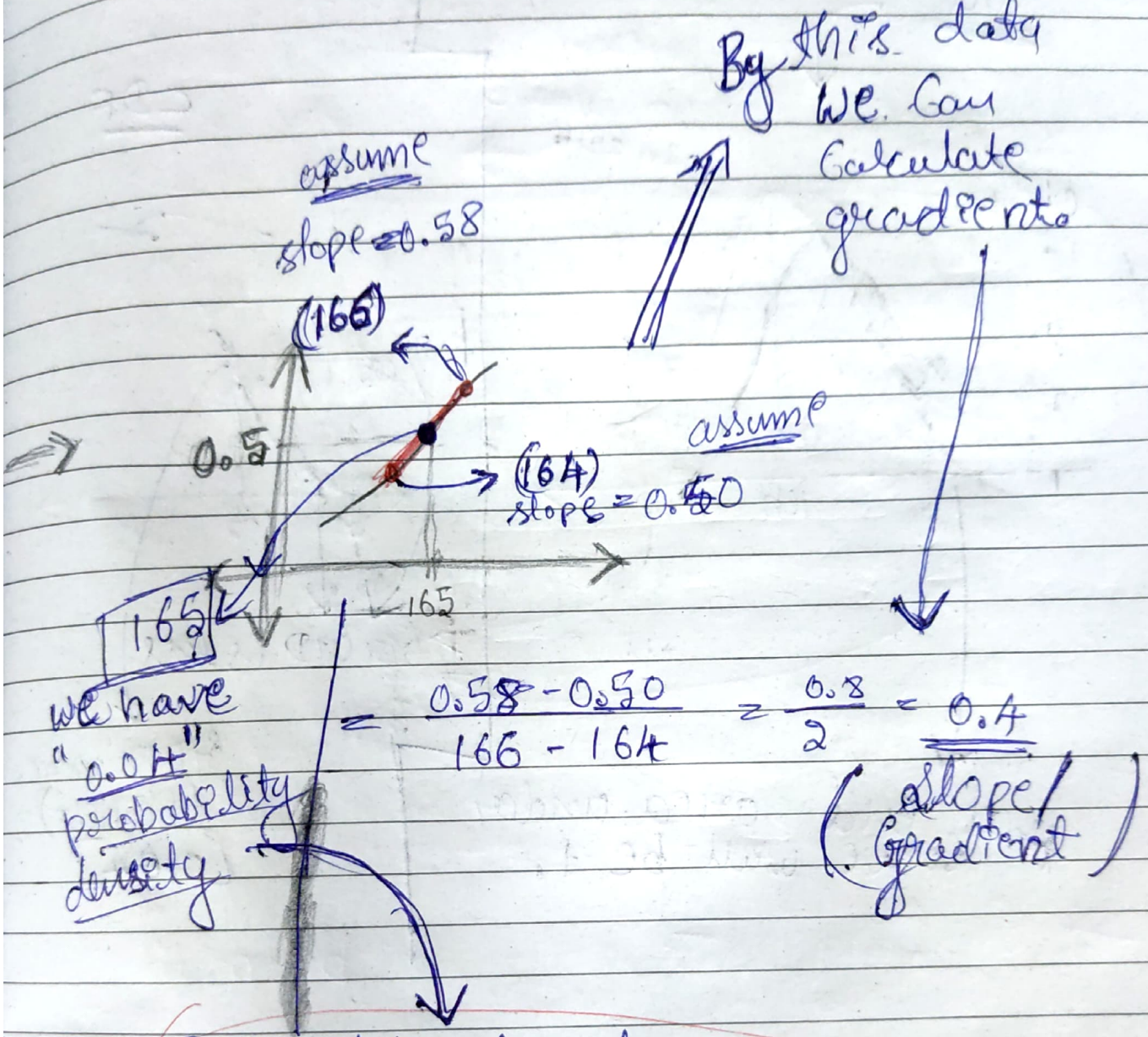
It's not the probability

* So to understand this probability density.

So, Further we try to Calculate the gradients

[slope] (or) [Derivative]

* then its the gradient descent of CDF



* Probability density is the gradient descent of CDF

*

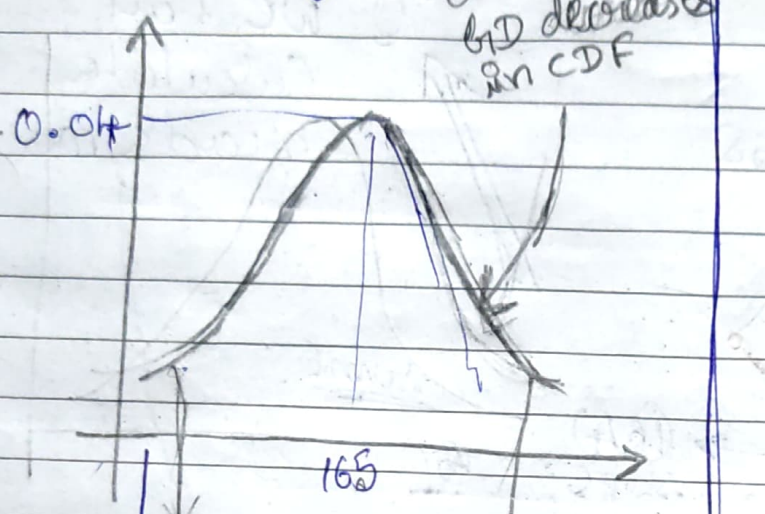
PDF : f(x) represents Rate of change of probability within the interval

PMF : P(X = x), represent a specific value.

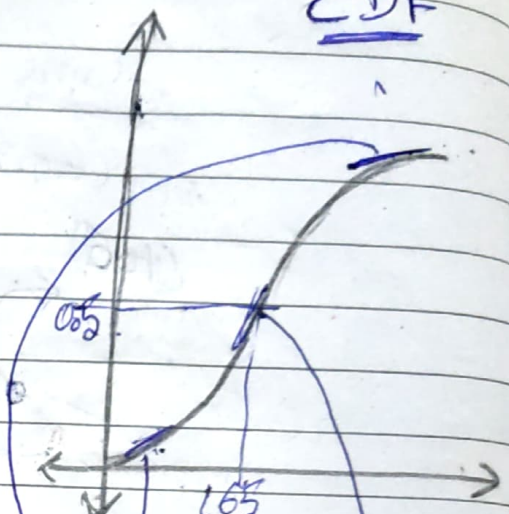
PDF

gradient descent [GD]

slope decrease because GD decreases in CDF



CDF



GD is less

GD 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 the probability p and value 0 with probability $q = 1 - p$.

$$q = 1 - p$$

* This
  ~~mate~~ models. Possible outcomes of an experiment → Yes / No

* outcomes may be boolean values.

Binary outcome

p → success / yes / true / one.

q → failure / no / false / zero.

Ex: ① Tossing a coin

⇒ outcomes → head [1] (May be)
→ Tail [0]

for Fair coin

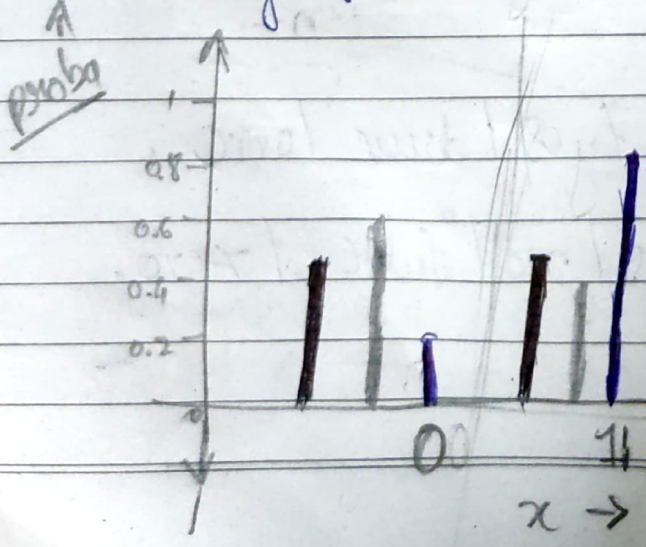
$P(H) = 0.5 = p$

Probability of getting a head

$q = 1 - p = 0.5$

Probability of not getting a head. (or) Here, we might also tell prob getting a tail.

* So any outcome in bernoulli's distribution wants to display as graph can be displayed using 'PMF'



will display the prob dist of the random variable.

ex:

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

② $P(x=0) = 0.6$
$P(x=1) = 0.4$

② Person pass / Fail exam.

* pmf v/s pdf
  - pmf → discrete random variable.
  - pdf → Continuous random variable

* we have specified range of values

* Specially, in Bernoulli's distribution we have 2 outcome only.

Proving using PMF

[Probability mass function]

'K' can be 0 (or) 1

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

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

① $P(x=1) = p^1 (1-p)^0$

$P(x=1) = p$ → in one equation we get 'p' value

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

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

$K \cdot p(x)$

$$E(K) = 1 * 0.6 + 0 * 0.4$$

$$= 0.6 = p$$

$K = 1(\text{or}) 0$

$p(K=1) = 0.6 = p$

$p(K=0) = 1 - 0.6 = q$

$= 0.4$

⇒ Median

we Consider

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

⇒ Variance & std

Variance $= p(1-p) = pq$

std $= \sqrt{pq}$

* Simplified way of PMF as proved

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

* Poission distribution [PMF]

① It is for discrete distribution

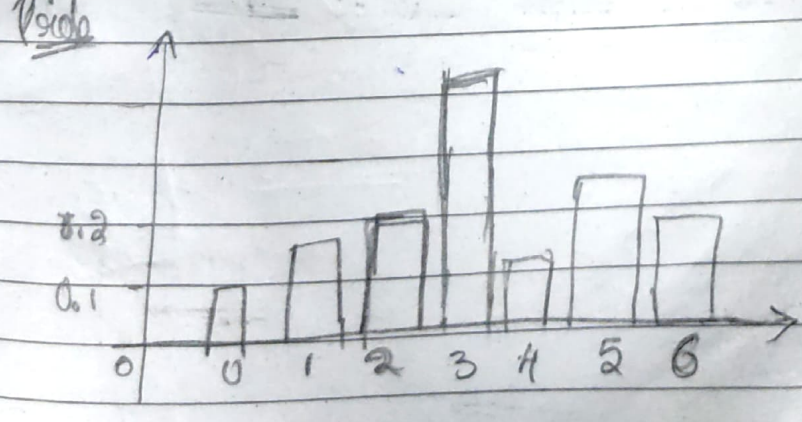
② It describes the number events occuring in a fixed time intervals.

ex: ① No of people visiting hospital every hour;

② — " — Bank — " —

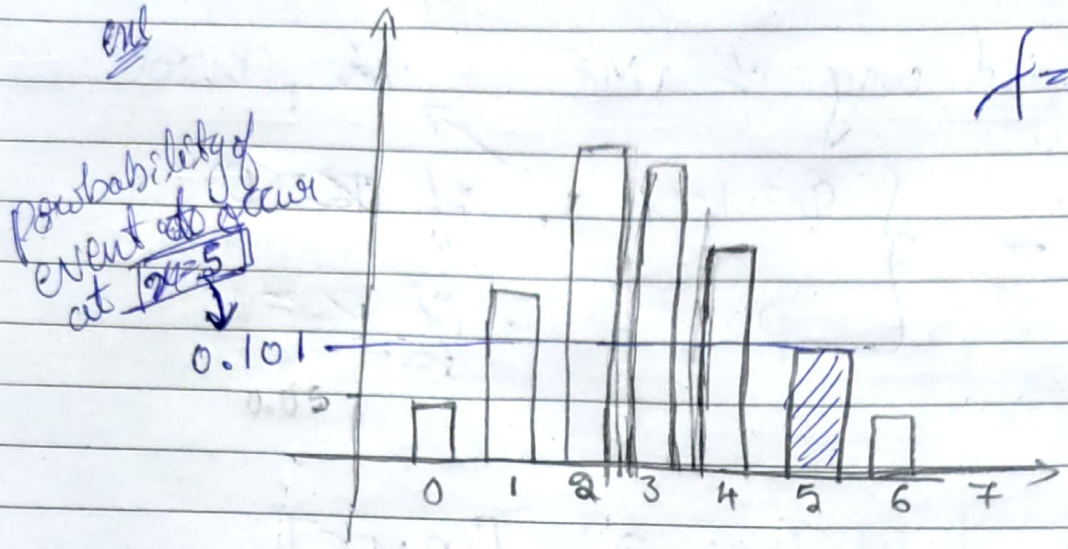
③ — " — airport — " —

is defined be "PMF"



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

also used in PMF



$\lambda = 3$

PMF: Probability of an event occur at 5th hour.

College

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

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

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

$$= 0.101 = 10.1\%$$

* Mean & Variance for poission distribution

→ Mean

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

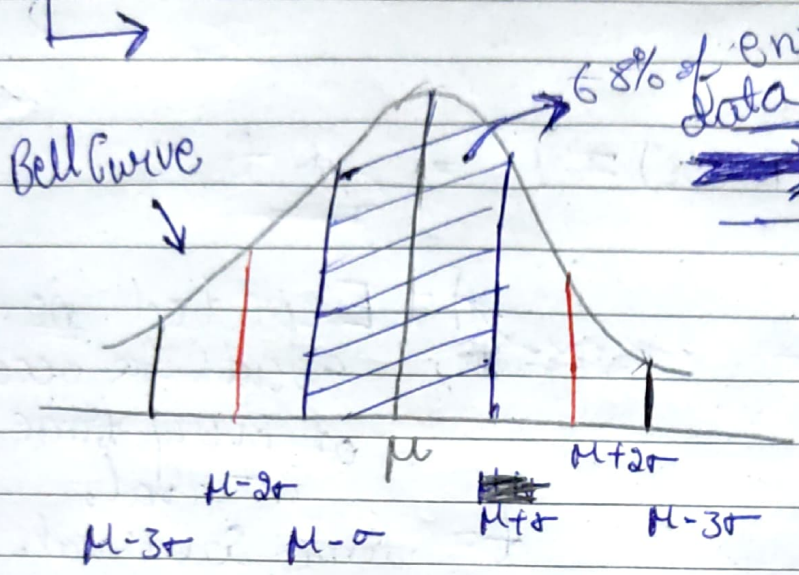
λ = Expected no of events to occur at every time interval.

t = time interval.

→ Variance

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

* Normal (or) gaussian distribution



Symmetric 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 then 68 data will fall in I std.

* within 2nd std 95% of data falls in this region.

* -11- 3rd std 99.7% of data falls.

* To check wheather the distribution is gaussion distribution (or) not ~~use plot~~

① Q-Q plot

enter

Pro.babi.lity

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

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

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

ex:

weight, Height, IBIS Dataset.

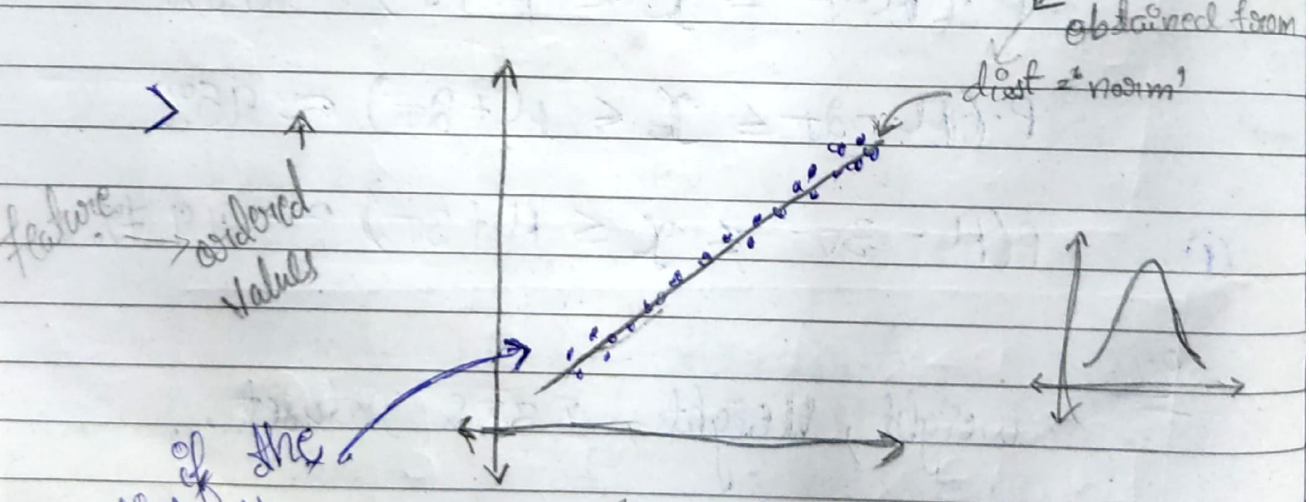
* QQ plot:

numpy
seaborn
matplotlib

pylab
scipy.stats

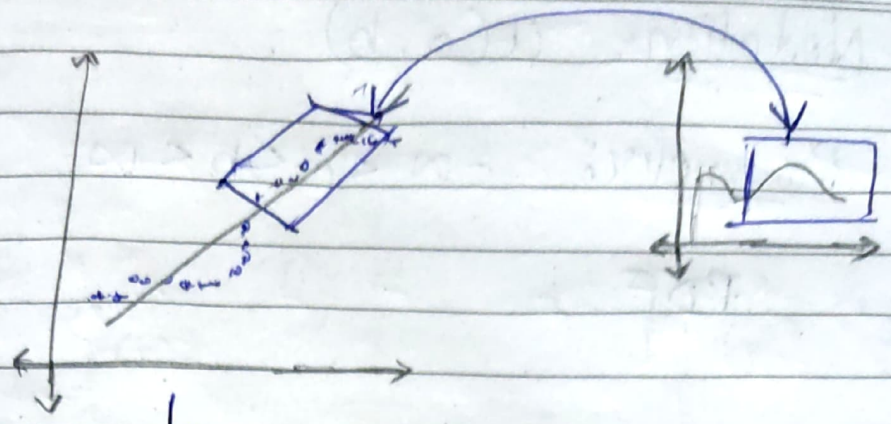
```
def plot_data(df, feature):
    plt.figure(fig=__)
    plt.subplot(1, 2, 1)
    sns.histplot(df[feat], kde=True)
    plt.subplot(1, 2, 2)
    stat.probplot(df[feat], dist='norm', plot=pylab)
    plt.show()
```

Calculate the quantiles for a probability plot and optionally show the plot.



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

QQ plot → tells the plot is normal distribution (or) not



Most of points donot lie on line so its not normal distribution.

* Uniform distribution

Constant probability density b/w 2 limits

① Continuous uniform distribution [PDF]

② Discrete uniform distribution [PMF]

① Continuous uniform distribution:

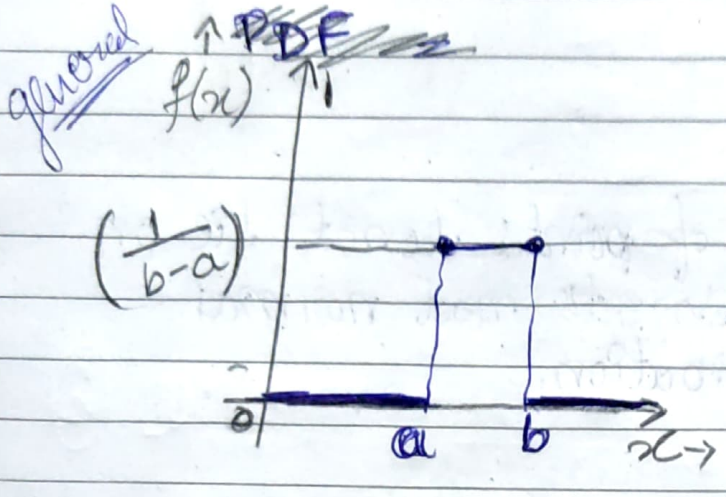
[Continuous random variable]

* 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)

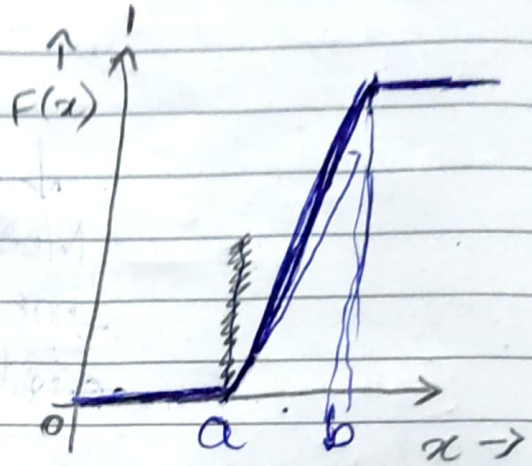
Notation: $U(a, b)$

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

general



PDF



CDF

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

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

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

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

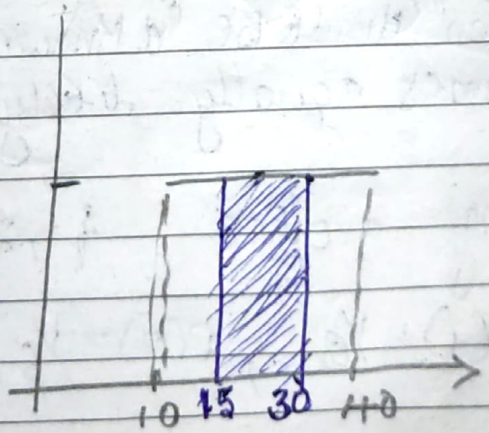
Variance $= \frac{1}{12}(b-a)^2$

Ex: The number of candies sold daily at a shop is uniformly distributed with a minimum of 10 and maximum of 40.

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

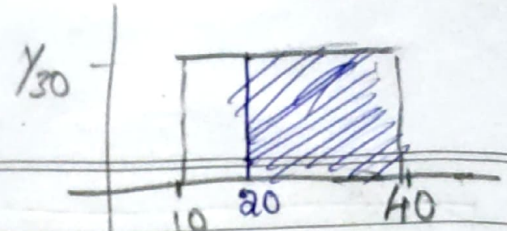
ans:

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



$P(15 \le x \le 30) = (30-15) * \frac{1}{30}$ (b * d)

$= \frac{15}{30} = 0.5$



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

$P(x \geq 20) = P(20 \leq x \leq 40) + P(x \geq 40)$

$= (40-20) * 1/30 + 0$

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

(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 $1/n$.
* Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen."

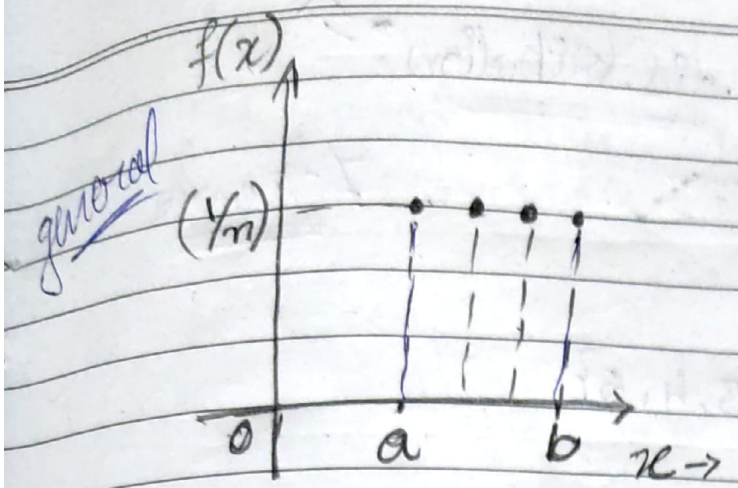
Ex: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$P(1) = 1/6$, $P(2) = 1/6$, ——

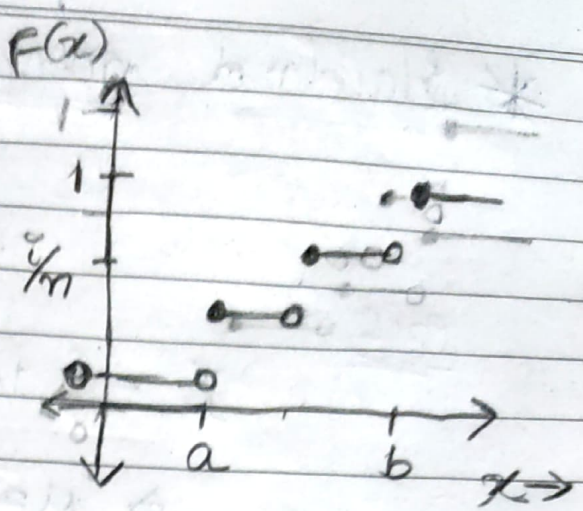
$\Rightarrow a = 1$, $b = 6$ → in dice

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

$= 6$



PDF



CDF

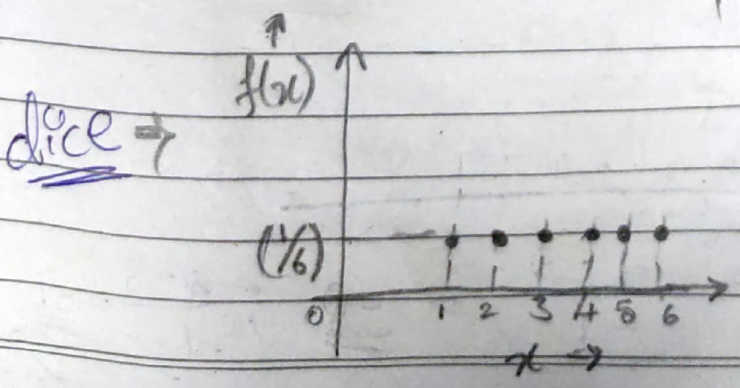
* All equal probability outcome.

Notation : $U(a, b)$

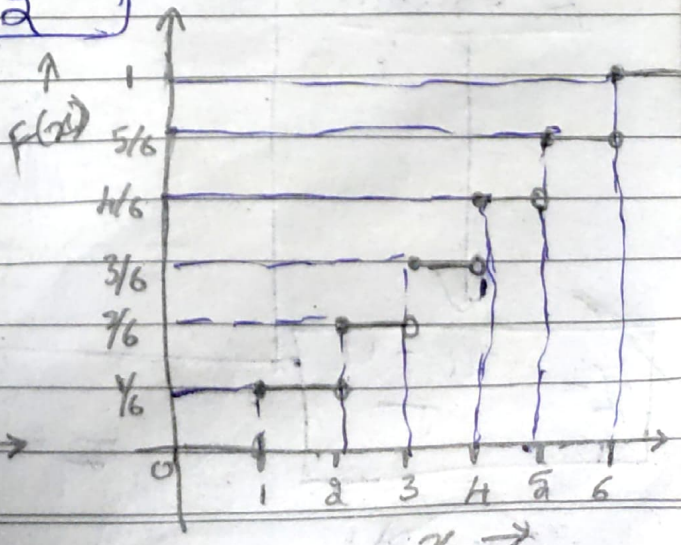
Parameters: a and b , $b \geq a$

PMF $\Rightarrow$ $\boxed{1/n}$

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



PDF



Use:
* To Find how many std away from mean.
* Standardization.

* Standard normal distribution:

Z-score

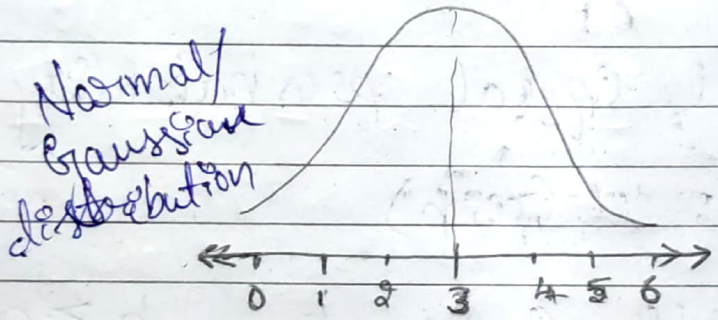
Eg:

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

$\Rightarrow \mu = 3$

Consider

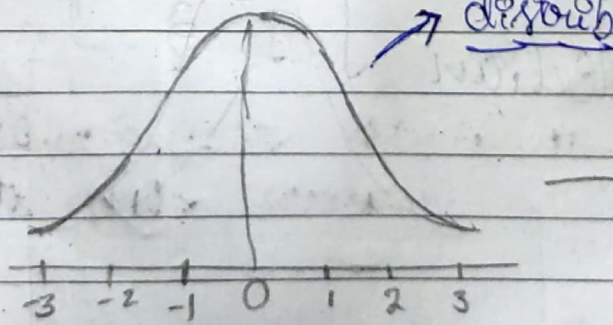
$\sigma = 1.414 \approx 1$



Convert

$\mu = 0$ (mean)
$\sigma = 1$ (std)
We want

Standard normal distribution



Convertion is done using Z-score

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

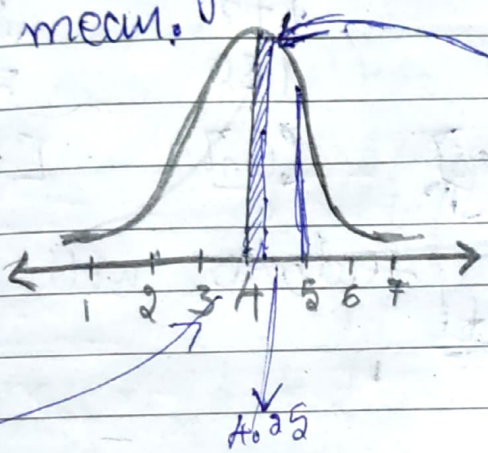
$$\therefore Z = \frac{1-3}{1} = -2 \quad ; \quad \frac{2-3}{1} = -1$$

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

mean (3) → mean converted to zero (0)

Normal Gaussian distribution → (using Z-score) → Standard normal deviation [$\mu = 0$, $\sigma = 1$]

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



area can be calculated using table.

(i) $x_i = 4.25$

$$Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1}$$

$$= 0.25$$

we get area

indicates that 4.25 is +0.25 std away from the mean (μ)

(ii) $x_i = 1.25$

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

$$= -2.75$$

tells that 1.25 is -2.75 std away from mean.

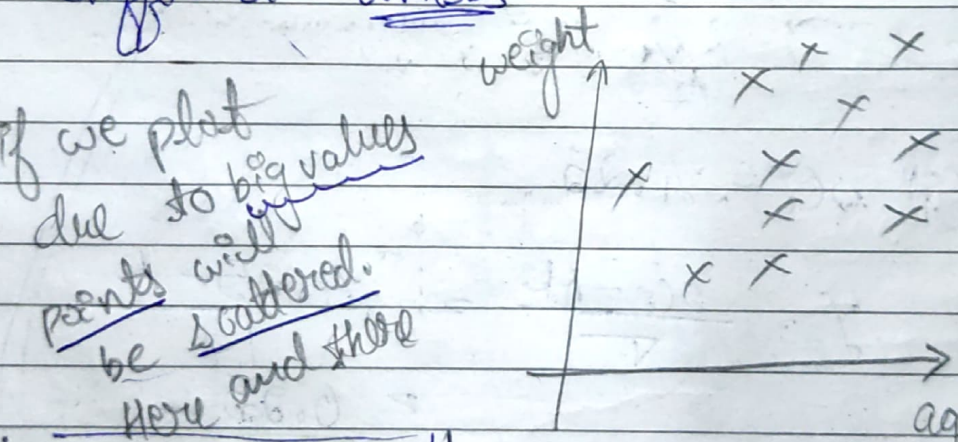
Standard normal deviation

Ex: 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
[years]	[Kgs]	[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 for few ML algo need features to be in same scale [clusting, KNN, oregress] They model gives performance increases

Convert using technique called "standardization" → we apply Z-score for all features.

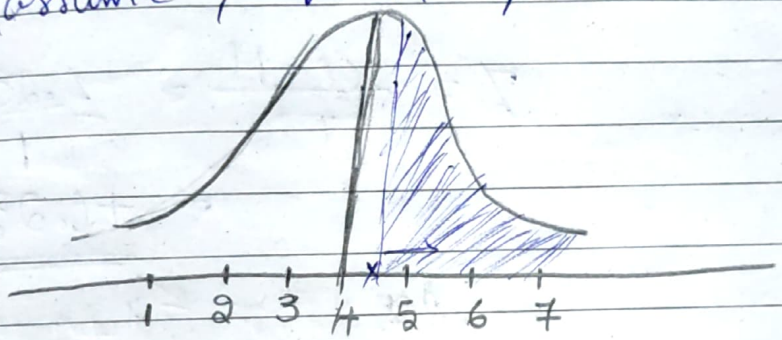
* Z score application and Z table.

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

⇒ problem statement

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

assume ; $\mu = 4.0$, $\sigma = 1$



a) what percentage of scores falls above 4.25 ?

⇒ i.e area under Curve should be found to Calculate percentage of scores.

we have to do standardization

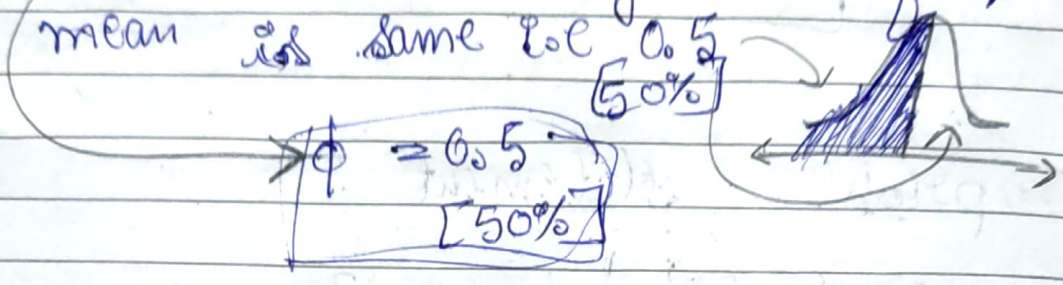
① clustering algo.
② linear regression
③ logistic regression, so many!

→ we get value in between (-3 to +3)

Curve

→ so, as bell curve it follows symmetric distribution.

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



→ $x_c = 4.25$, $\mu = 4$, $\sigma = 1$

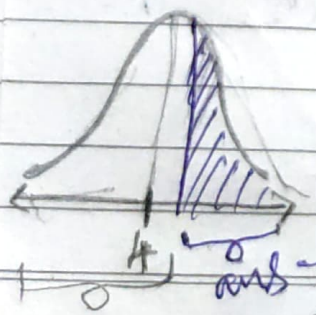
$$Z = \frac{x_c - \mu}{\sigma} = \frac{4.25 - 4}{1} = +0.25$$

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

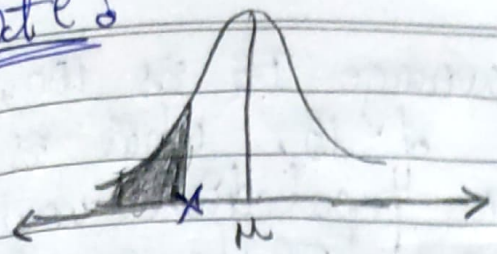


+Z table → 0.59871

Area = 1 − 0.5987 = 0.4013 ≈ 40%

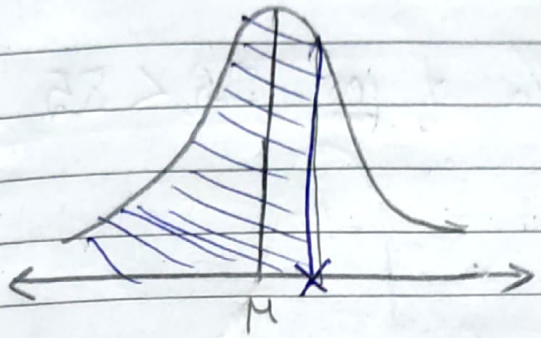
Area z = +0.25 → 40%

Notes

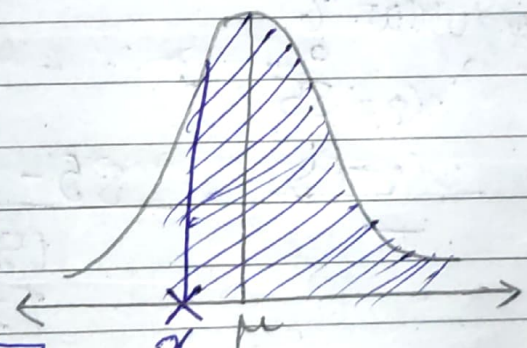


→ use negative Z table and Calculate area.

Basic



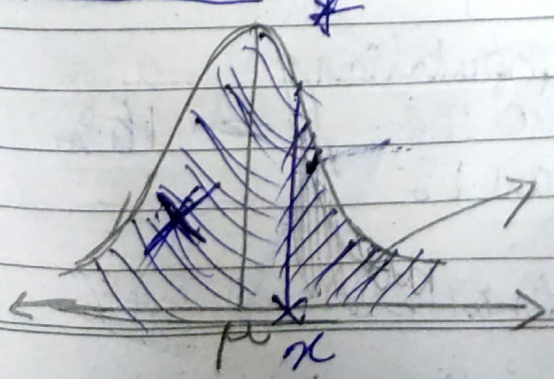
→ use positive Z table and Calculate area.



use -ve Z table

Area = 1 - (x) area

① Negative Z table.
② Positive Z table.



use +ve Z table

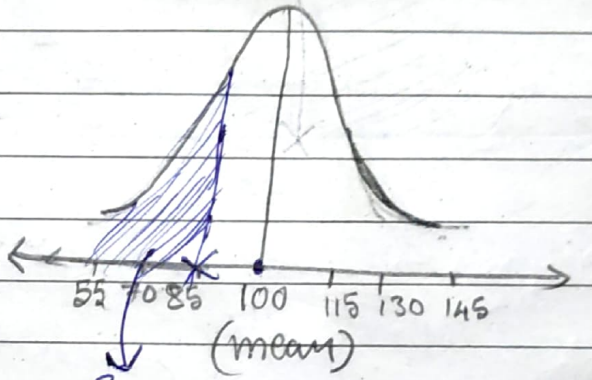
Area = 1 - (x) area.

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?

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

% of IQ < 85

initially



percentage/area = ?

$x_q = 85$

we convert into standard normal deviation

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

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

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

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

$\therefore$ Percentage of population who have IQ lower than 85 = 16%

$\therefore$ greater than 85 = 84%

$\therefore$ $75 < x < 100$ = 45%

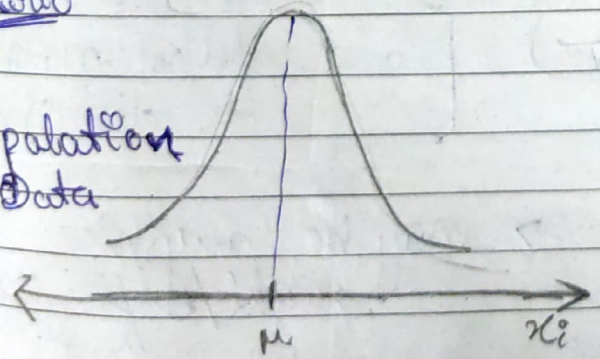
Central limit theorem [CLT]

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 wheather the population has a normal, poission, binomial, or any other distribution, the sampling distribution of the mean will be normal.

Scenario 1



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

→ Gaussian distribution.

(n) = Sample size ⇒ Can be any value

Sample		Mean
S_1	$= \{x_1, x_2, x_3, \ldots \rightarrow x_n\}$	$= \bar{x}_1$
S_2	$= \quad " \quad$	$= \bar{x}_2$
S_3	$= \quad " \quad$	
S_m	$= \quad " \quad$	$= \bar{x}_m$

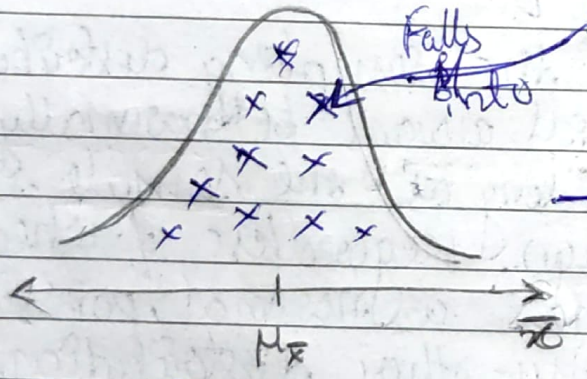
Mean of samples

Falls into

always

Follows Gaussian distribution

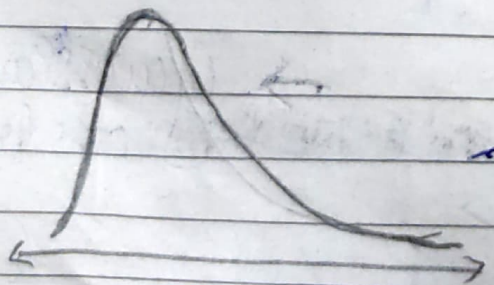
A/c CLT



Sampling distribution of the mean

Scenario

② $X \neq N(\mu, \sigma)$ [donot follow normal distribution]



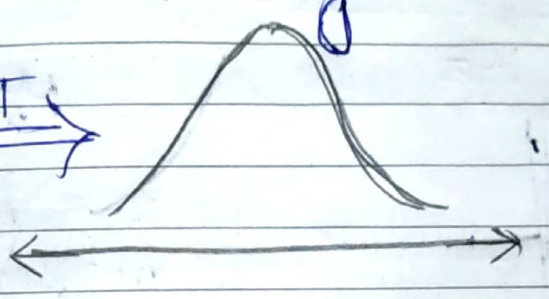
→ may be anything right / left skewed.

If $n \geq 30$ ⇒ Sample size

		mean
S_1 =	x_1 ... x_n	= $\bar{x}_1$
S_2 =		=
⋮		⋮
S_m =		= $\bar{x}_m$

then sample will follow normal / gaussian distribution

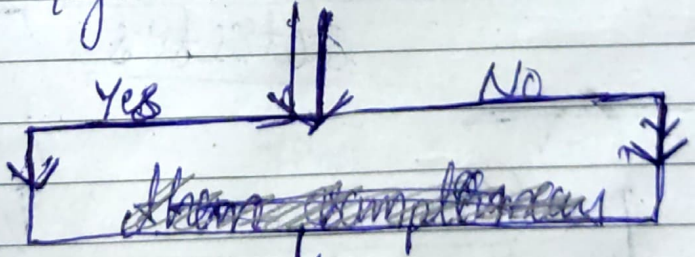
CLT ⟹



∴ Central limit theorem

↓ states

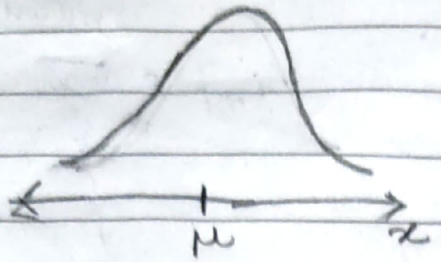
if population follows normal / gaussian distribution



then sampling distribution of mean will follow "normal distribution".

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

Normal distribution

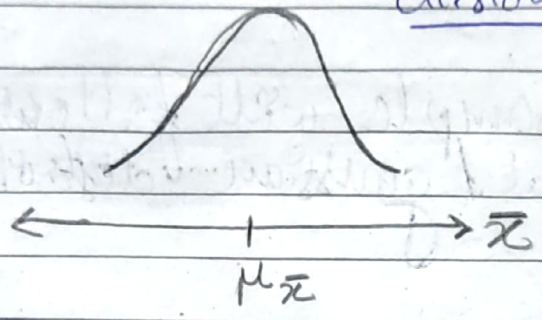


Population mean std

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

Population

↓ sampling distribution of means



$\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

$\mu_{\bar{x}}$ = mean of sample

σ = population std

n = sample size

Mean sample, $\mu_{\bar{x}} = \mu$

std of sample, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

↑ Until now descriptive statistics

From now inferential statistics ↓

* Estimate

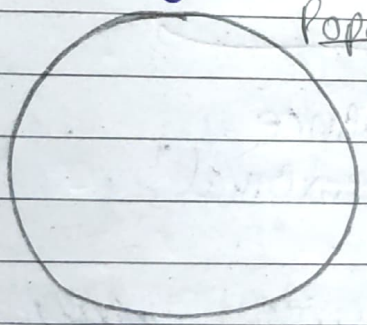
↳ It is a specific observed numerical used to estimate an unknown population parameter.

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)

$\bar{x}$ (Sample mean)

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 65 mean.

Disadvantage of point estimate

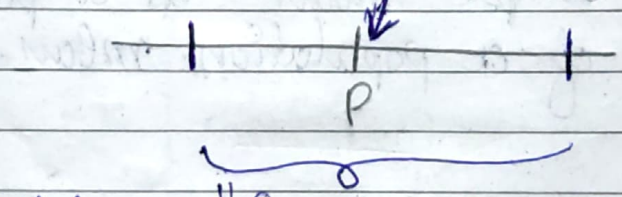
* There can have margin of error, may be + or −.

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"

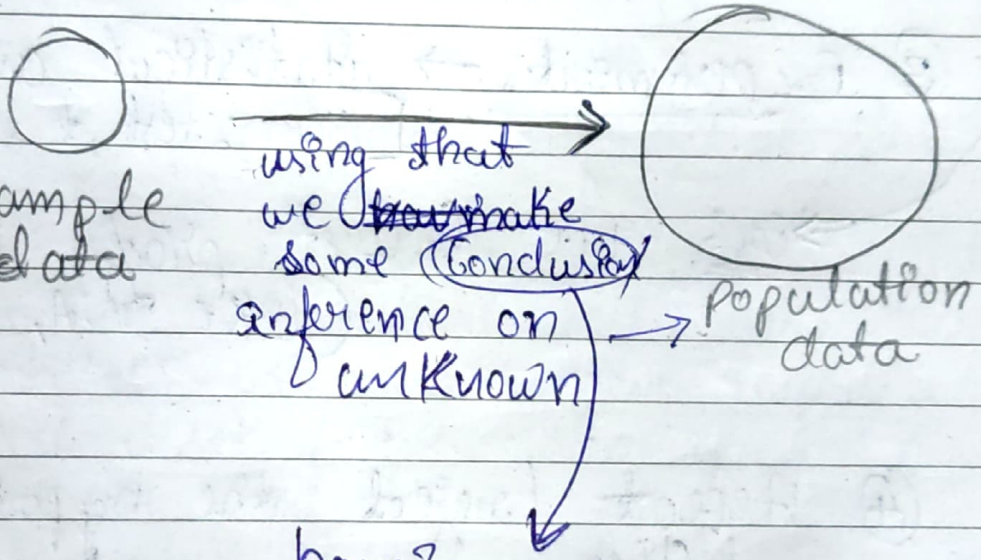
Ex: for previous example we can say that

population mean may be [55-65]

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

* A Hypothesis and Hypothesis testing mechanism

Inferential stats aim → Conclusion or inference from a sample data about a unknown population parameter.



how?

⇒ we use hypothesis ~~testing~~ testing.

⇒ Hypothesis Testing mechanism

① Null Hypothesis (H_0)
↳ The assumption that we are beginning with.

② Alternate hypothesis. (H_1)
↳ The opposite of null hypothesis.

Ex:

Person → Committed → Crime → went → Court

(h0) Initial : he do not fell guilty

(h1) He felt guilty (opposite)

(continue)

③ Experiments → statistical analysis
[like z test, t test, ANOVA, ...]

Ex: we collect proof
(like DNA, finger print)

④ Accept / reject the hypothesis testing.

Ex: Colleges at district a state's 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 90% with a standard deviation of 4%. college
Does this ~~serial~~ have a different passed percentage.

Null hypothesis

ans: $h_0 \rightarrow \mu = 85\%$

average passed percentage of student is 85%

alternate hypothesis $h_1 \rightarrow \mu \neq 85\%$

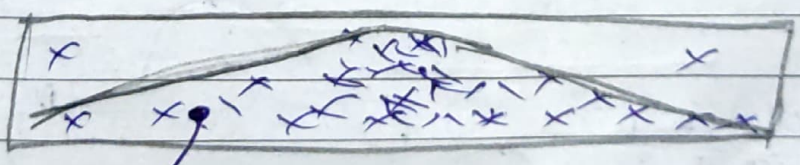
near College have different passed percentage.

---

* P value → 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.

Use: P value are used in hypothesis testing to help decide wheather to reject the null hypothesis.



Probability of a specific outcome in a specific region.

$P = 0.02$

Space Bar laptop

↳ Meaning: out of 100 touches, we touch around 20 times in this region.

*Hypothesis testing

Outcome of a statistical test is a Pvalue.

Eg: Coin is Fair (or) not
{100 times}

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

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

& 3 & $P(H) = 0.7$, $P(T) = 0.3$ → Coin not Fair

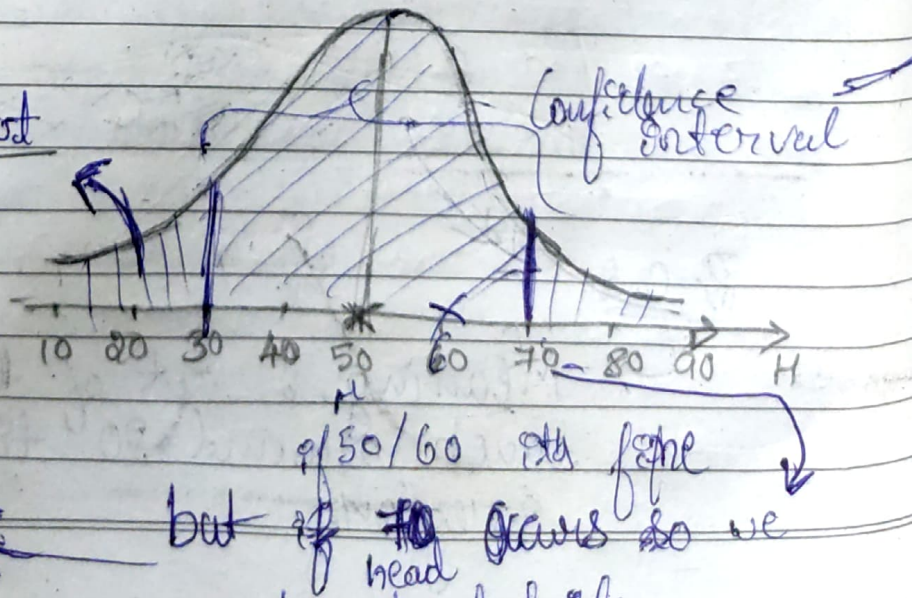
① Null hypothesis:

h_0 → Coin is Fair (based)

② Alternate hypothesis:

h_1 → Coin is not Fair

③ Experiment: 100 times



So, Now we well describe Confidence interval.

but if 70 occurs so we have to test it

determined by domain expertise

④ Significance value:

at 5% level of significance

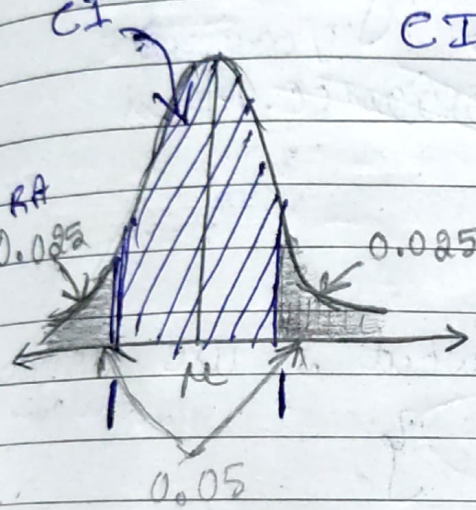
$\alpha = 0.05$

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

$CI = 1 - 0.05 = 0.95$

i.e 95% Confidence interval.

[except null hypothesis]



⑤ Conclusion:

$p <$ Significance value

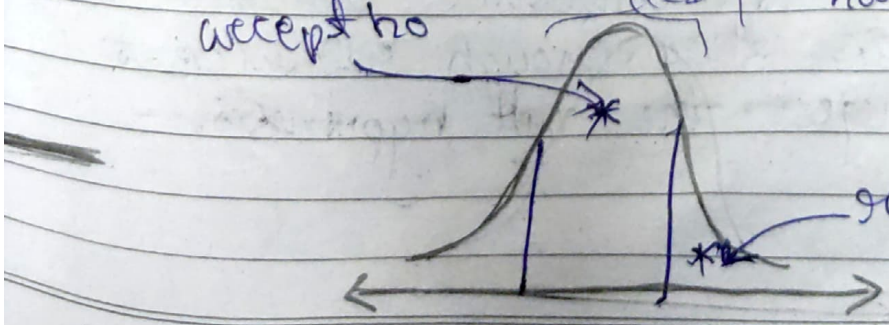
$p >$ Significance value:

we reject the null hypotheses

else Fail to reject null hypotheses

→ If Coin Comes in that region then "Coin is Fair."

accept h_0



imp

* Hypothesis testing and statistical analysis:

① Z test
② t test
} ⇒ Population Average ⇒ Z table / t table

③ Chi square ⇒ Categorical data
F test → Variance
④ ANNOVA ⇒ ~~Variance~~ Mean

① Z test:

Whenever we want to use ~~the~~ should follow
(i) Population std.
(ii) $n \geq 30$

Q) The average height of all residents in a city is 168 cm with a ~~mean~~ std of $\sigma = 3.9$ believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5 cm
(a) State null and alternate hypotheses.
(b) At 99% Confidence ~~level~~ level, is there enough evidence to reject the null hypothesis.

① Population mean should be there & $n >= 30 \Rightarrow$ Z test.

② Null & alternative hypothesis.

③ decision boundary.

④ Find Z value

⑤ Find area $= P$

⑥ Conclusion. $\Rightarrow P < \alpha \rightarrow$ reject
$P \not< \alpha \rightarrow$ not reject

we Find
→ Z score & P value

as $n \geq 30$
Z-test is used

population mean
$\Rightarrow \mu = 168$, $\sigma = 3.9$, $n = 36$.

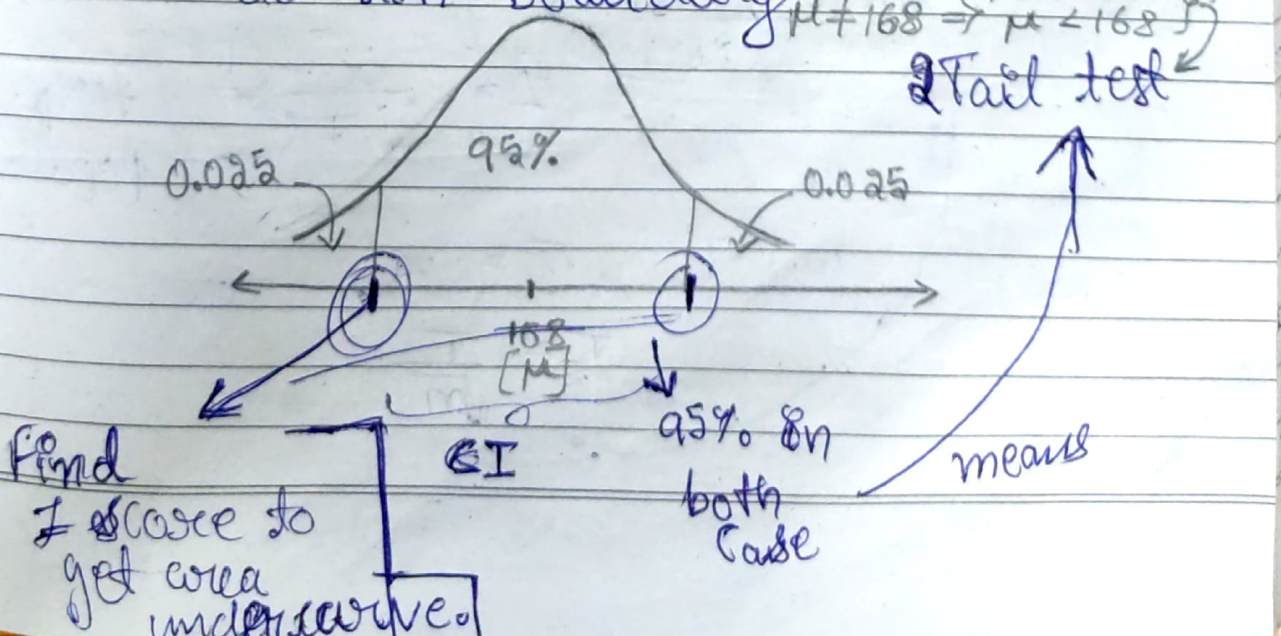
sample mean
$\bar{x} = 169.5$ cm

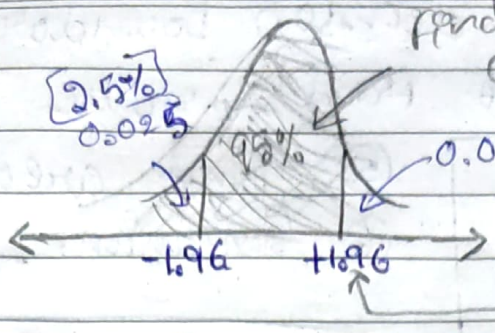
C.I $= 0.95$, $\alpha = 1 - 0.95$
$= 0.05$
→ significance value

① Null hypothesis.
$H_0 \Rightarrow \mu = 168$ cms

② Alternate hypothesis.
$H_1 \Rightarrow \mu \neq 168$ cm

③ Based on C.I we will draw decision boundary $\mu \neq 168 \Rightarrow \mu > 168$ & $\mu < 168$
2 Tail test





$1 - 0.025 = 0.9750 \Rightarrow$ Z-score

area → (converted using table) → +1.96

* if Z is less then −1.96 (or) greater then +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 is

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

w.k.t $\bar{x} = x$,

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

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

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

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

∴ 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.

$$2.31 > 1.96.$$
$$Z_t > Z_c.$$

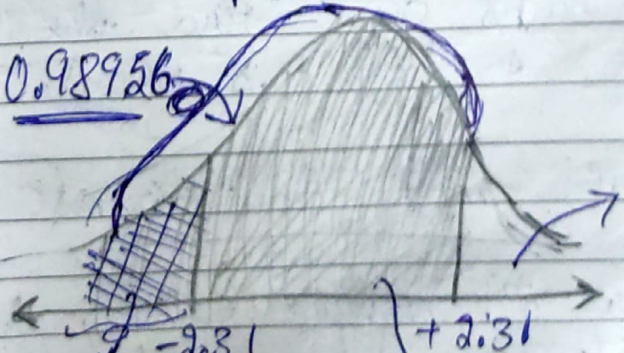
* Conclusion

$P < 0.05$

Area $_{Z=2.31}$ = 0.98956

Area = 1 − 0.98956 = 0.01044

Table



Area = 0.01044

Area = 1 − 2 * 0.01044 = 0.97912

p.value = 0.01044 + 0.01044
= 0.02088

5% level of significance

$p < 0.05$

$0.02088 < 0.05$ → True

⇓

reject the null hypothesis

Finall ↘

The average height ≠ 168 cm

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

---

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

only left than

left tail test

(a) State null and alternate hypothesis.

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

Null hypothesis

① $H_0 \Rightarrow$ Mean population $\mu = 5$ years

② $H_1 \Rightarrow \mu \neq 5$ years → 1 Tail test

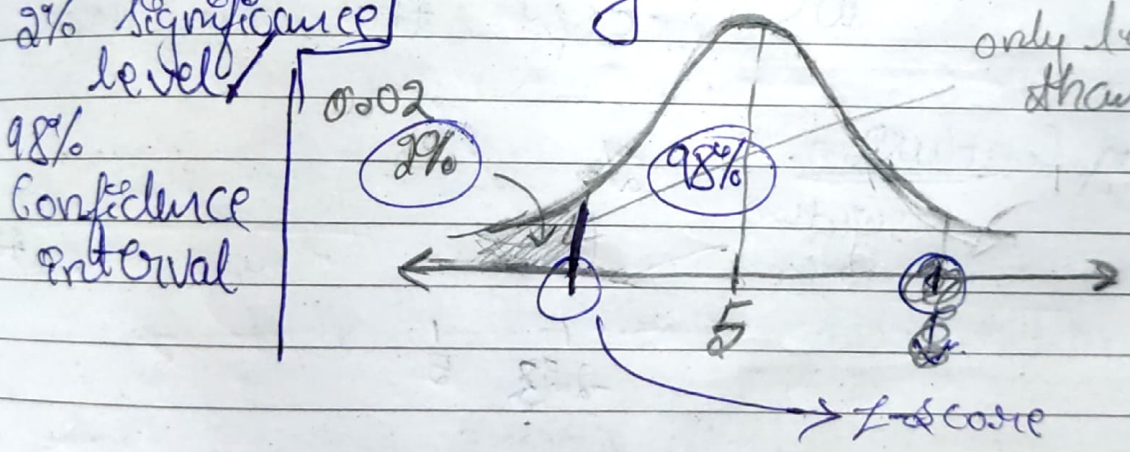
Mean population, $\mu = 5$

population std, $\sigma = 0.5$

sample size, $n = 40$ → Z test

sample mean, $\bar{x} = 4.8$

③ Based on significance level we draw decision boundary.



$1 - 0.02 = 0.98$ → 3.8

↓ area ↓ Z score

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

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

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

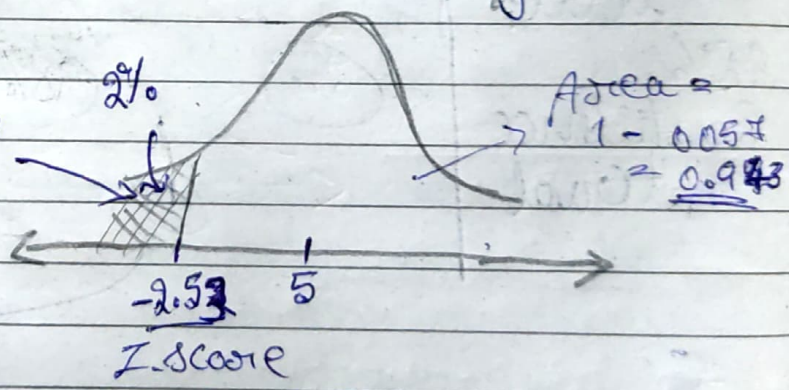
$$I_d = -2.53$$

As $I_d < I_c$

$-2.53 < -3.8$

We accept the hypothesis

* Conclusion.



$$\text{Area}_{z=-2.53} = 0.0057$$

P value = 0.057

Compare p-value with α

$0.057 < 0.02 \Rightarrow$ False

2% significance

my P value

which mean that point will not fall in 2% level of significance → as My P value < 0.05?

also fall in 95% CI

$\therefore$ 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) population
it is difficult to find σ as population data is not there.
σ is already known.

* 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

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

z ↓ table

s = sample standard deviation

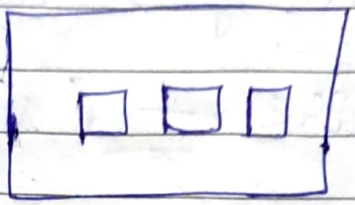
↓

we use t-table

we have parameter degree of freedom.

* Degree of freedom = n-1
  ↳ one

if we have 3 chairs in a Room



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

---

1) T-test — (One sample t-test)

In a population the average IQ is 100. A team researchers want to test a new medication to see if it 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 20. Did the medication affect intelligence? C.I = 95%

↓

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

⇒ population mean, $\mu = 100$
sample size, $n = 30$
sample mean, $\bar{x} = 140$
sample std d, $\sigma_{\bar{x}} = 20 \Rightarrow \boxed{\sigma = 20}$

population ~~std~~ std ⇒ not given

t-test

Confidence interval = 95%

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

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

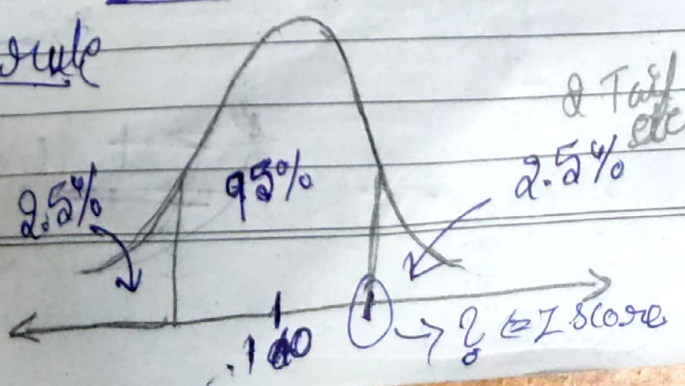
Alternate hypothesis, $H_1 \Rightarrow \mu \neq 100$
{2 Tail test}

For finding Z score

② $\alpha = 0.05$

③ Degree of freedom.
$dof = n - 1 = 30 - 1$
$dof = 29$

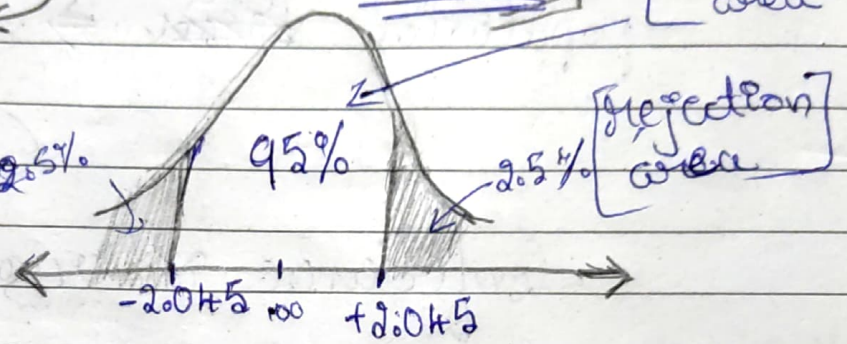
④ Decision rule



⑤ Finding T score using t-table for

dof = 29 & C.I = 95%

how much std away from mean ← T score = 2.045 [acceptance area]



⑥ ④ decision rule → if t-test is less than -2.045 (or) greater than +2.045 then we reject null hypothesis.

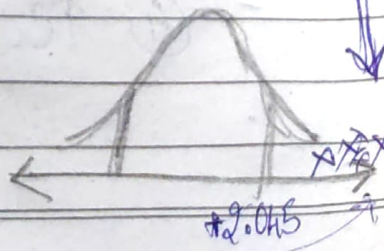
⑥ Calculating test statistics

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

$$t = \frac{140 - 100}{\frac{20}{\sqrt{30}}}$$

$$t = +10.96$$

falls in this region



Since, $|t = 10.96| > I = 2.045$

⇓

True

∴ We have to reject Null hypothesis

(7) Conclusion:

Medication used has affected the intelligence.

(may be more (+) (or) less (-))

⇒ But, specifically as it is "+ve" mean 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"

* Poission as mean increases → Shape of distribution become "Taller"

Code

* Calculating pdf using "scipy.stats"

→ scipy.stats.norm.pdf(point, loc = mean, scale = std)
p/p p/p

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

→ trials = 1

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

* poission distribution

np.random.poission(mean, sample size)

* CDF

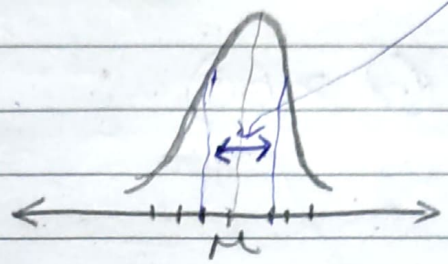
→ discrete ⇒ $F(x) = P(x \le x) = \sum [P(x = x_i) \text{ for } x_i \le x]$

→ Continuous ⇒ $P(x \le x) = \int [f(t)\,dt \text{ for } t \le x]$

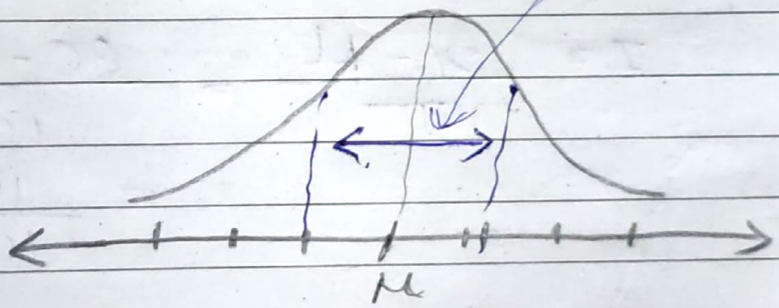
$\int_{t=0}^{x} f(t)\,dt$

* Normal distribution → μ & σ Controls the shape of curve.

→ Smaller σ (std)
⇒ Narrow curve



→ larger σ (std)
⇒ wider, more spread.



* Assignment

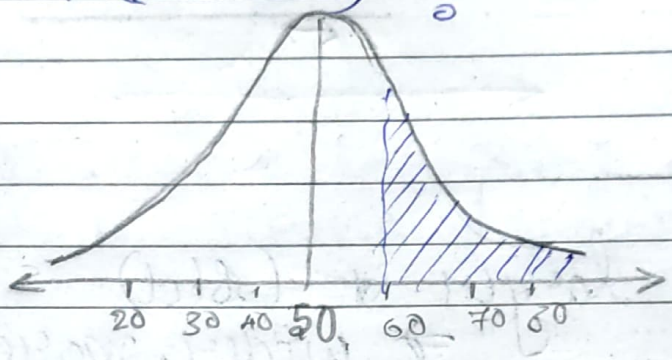
$\mu = 50$, $\sigma = 10$..

$P(x > 60)$

* Assigment Q6 Solution

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

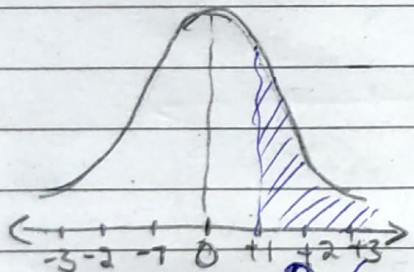
$P(x > 60) = ?$



$x = 60,$

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

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



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

(From Z Table)

$= 1 - 0.8413$

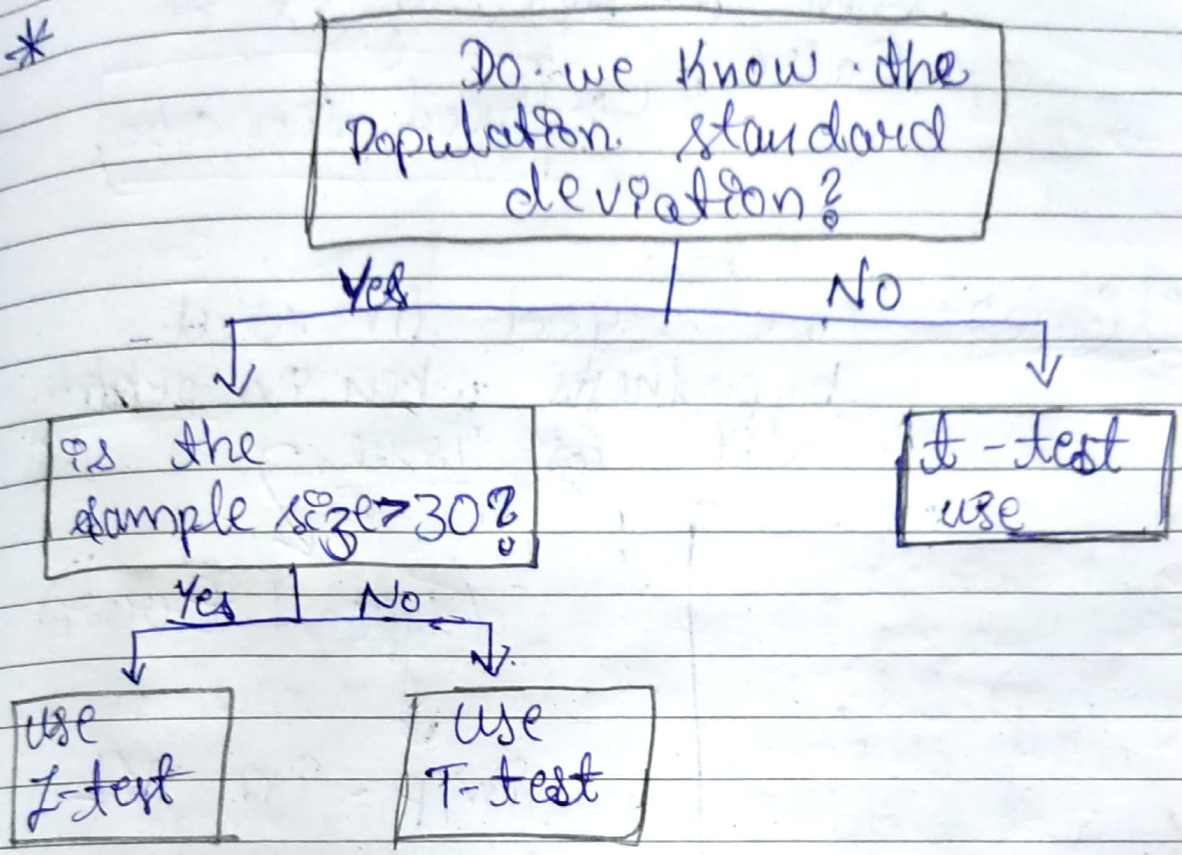
$P(x > 60) = 0.1587 = 15.87\% \simeq 16\%$

Answer: 16%

Week-12

Statistics advance-2

*



* Type 1 and Type 2 error

Tell NH can be true / can be false based on result of HT

Reality: Null hypothesis is True (or) Null Hypothesis is False.

Decision: —— | | ——

Outcome 1: we reject the Null hypothesis when in reality it is False. → Good Scenario

Reality → it is False

Type 1 Outcome 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 1 Error) → Imp - in ML i.e in Confusion matrix

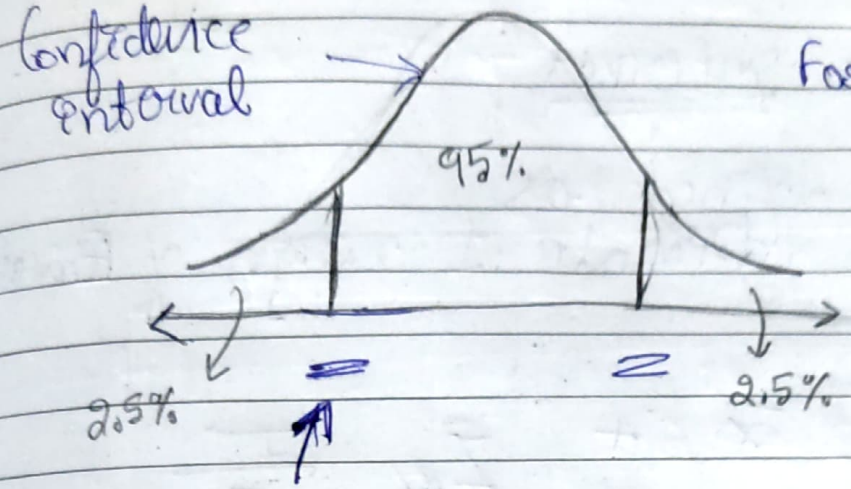
Type 2 Outcome 3: we accept the Null hypothesis when in reality it is False.

(Type 2 Error)

accept

outcome 4: we retain the Null hypothesis when in reality it is True. → Correct / Good Scenario

* Confidence Error and Margin of Error.



For 2-tail test with CI = 95%

what value can come in this interval.

Point Estimate

sample mean $\bar{x}$ may be PE → for → μ 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 ± margin of error

for
Z-test

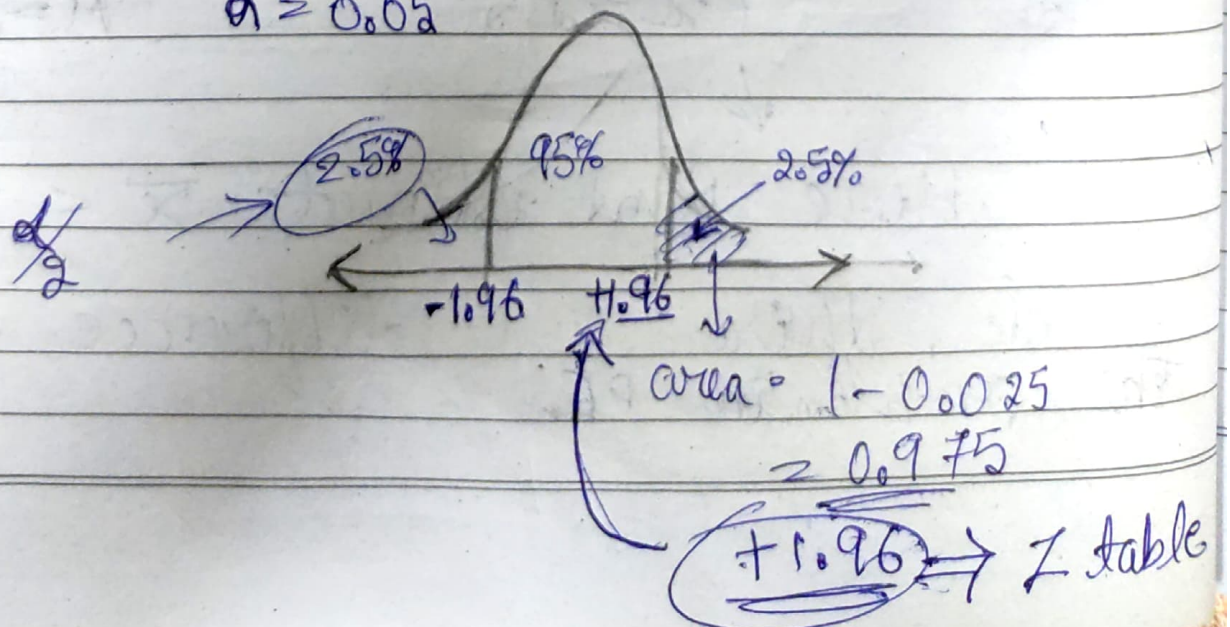
$$\Rightarrow \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Ex:- One the verbal section of CAT exams. the std is known to be 100, A sample 30 test takers has a mean of 520. Construct 95% CI about the mean.

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

$\alpha/2$ → 2 tail

$\alpha = 0.05$



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

$= 520 - 39.2$

Lower CI $= 480.8$

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

$= 559.2$

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 [V.imp]

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

→ V.imp For ML algo, Naive Bayes

→ Completely on Probability.

Probability → Independent Events
Dependent Events.

① Independent events:

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

P(1) = 1/6, P(2) = 1/6 ...

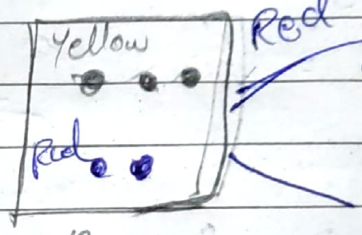
* Here, Happening of one event is not impacting other event. So, Independent

② Tossing a coin

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

② Dependent Event:

ex:



Bag

* Here happening 1 event affect the happening of other event. so dependent.

* P(R and Y) = P(R) * P(Y/R)

Probability of red and after red, yellow.

Probability of Yellow when red event has occured

Condition probability

$2/5 * 3/4 = 6/20$ == $3/5 * 2/4 = 6/20$

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

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

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

⇓

Bayes Theorem

(or)

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

I red already taken out

$P(Y) = 3/4$

college

$A, B \Rightarrow$ events

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

$P(B/A) \Rightarrow$ —11— B given A —11—

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

Dataset

[Independent] [dependent o/p]

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

Using this 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.

Chisquare test

we tell
The "chisquare 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: There is a population of male who likes different colors bikes

	Theory about Popu	Sample
Yellow Bike	1/3	22
Red Bike	1/3	17
orange Bike	1/3	20
		59

(*) goodness of fit test.

IMP
[The sample data which is collected does it support the theory present with us]

we check that which is present

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

$12\% \text{ of } 75 = \frac{12}{100} \times 75 = 9$

12% are left handed

9 → Expected value based on the theory

remaining: right handed $= 75 - 9 = 66$

Now, we have to test the 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

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

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

Sample data ⇒

<50	50-75	>75
140	160	200

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

⇒ (theory)

2010

Original data

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

⇓

Expected values

"given"

$n = 500$

Observed values/ data 2020 ⇒

<50 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 ⇒ H_0: The data meets the expectation

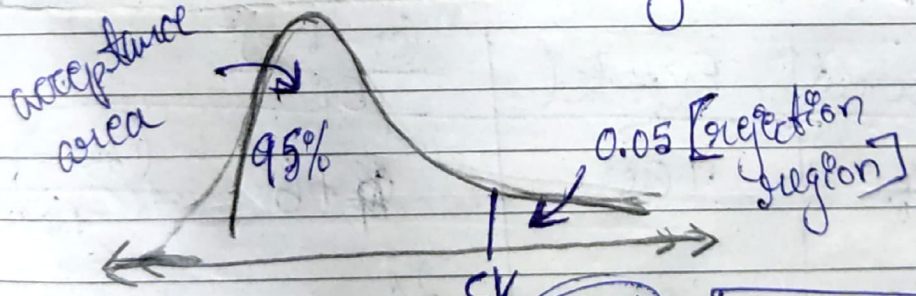
Alternate hypothesis ⇒ h_1: the data does not meet the expectation.

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

③ Degree of freedom

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

④ Decision Boundary



we get from "chisquare table" using α and df

but in chisquare test we have right skewed distribution.

In Z & t test we use symmetrical distribution.

for ($\alpha = 0.05$ & dof 2) chi square value $\Rightarrow$ 5.991

$\therefore$ Critical value

Decision rule: If x^2 is greater than 5.99, reject the H_0, else we fail to reject the Null Hypothesis

(5) Calculate Chi Square Test statistics

$$x^2 = \frac{\Sigma (\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

	Observed	Expected	$\frac{(O_i - E_i)^2}{E_i}$
<50	140	100	16 $\left(\frac{(140-100)^2}{100}\right)$
50-75	160	150	2/3
>75	200	250	10

$\Sigma =$ 26.66.

$x^2 =$ 26.66

6 Conclusion

As $x^2 = 26.66$ which is greater than 5.99. i.e

$x^2 > 5.99 \Rightarrow$ True

So, we reject the null hypothesis.

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

Chisquare in python

→ observed_data = [___]
→ expected_data = [___]

Find chisquare value → sum(obs) == sum(expe)

→ chisquare_test_stat, p_value = stat.chisquare(O, E)

Find critical value →
- significance = 0.05 # 5%
- CL = 1 − significance
- dof = len(obs) − 1
- crit_Cal_value = stat.chi2.ppf(CL, dof)

ppf = percent point function

output

→ if chisquare stat > CV:
    # reject
else:
    # accept

Conclusion

---

* F-distribution / Fration

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

↓

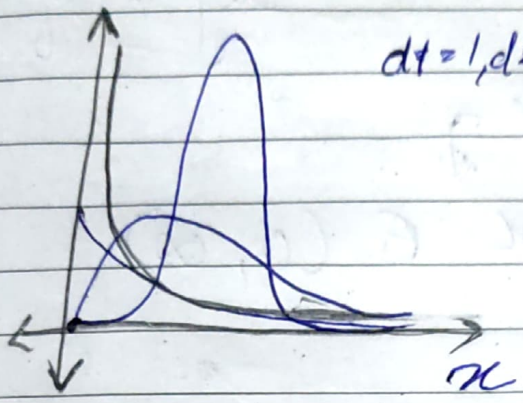
used for hypothesis testing to analysis the variance & mean b/w 2 different groups

we use Continous random variable

* From now we use F distribution in ANOVA, —

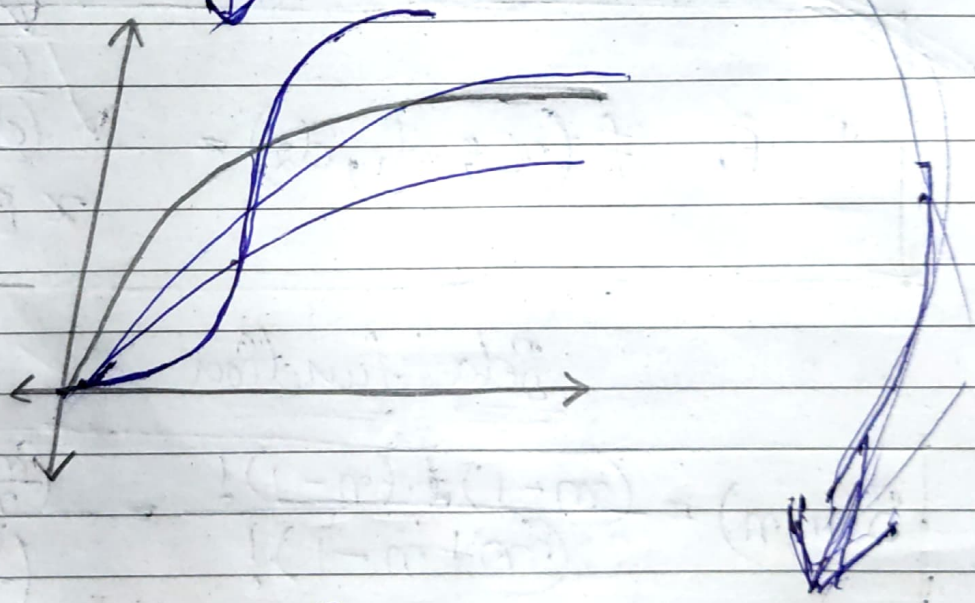
Fisher - Snedecor

Probability density function [PDF]



probably right skewed data

CDF



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

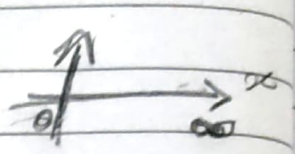
↓

because using F distribution we compare 2 groups of data.

* Parameters :- $d_1, d_2 > 0$
↓
degree of freedom

* Supporting variable ↓

$x \in (0, \infty)$



* Curve is created using PDF ↓

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

Beta function

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

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

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

d_1 & d_2 ⇒ degree of freedom

S_1 & S_2 ⇒ independent random variable with chi square distribution

* imp F-test

→ Called as "Variance ratio test".

→ Used to compare the variance 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).

$\alpha = 0.05$

Variance

Ans:

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

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

② Calculation of variance

~~worker A~~

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

worker A

x_1	$\bar{x}_1$	$(x_1 - \bar{x}_1)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
		$\Sigma = 80$

$\bar{x} = 37$

worker B

x_2	$\bar{x}_2$	$(x_2 - \bar{x}_2)^2$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
		$\Sigma = 84$

$\bar{x}_2 = 37$

$\sigma_1^2 = \frac{80}{6-1} = 13$

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

$\sigma_2^2 = 12$

③ Calculation of variance ratio test (F test)

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

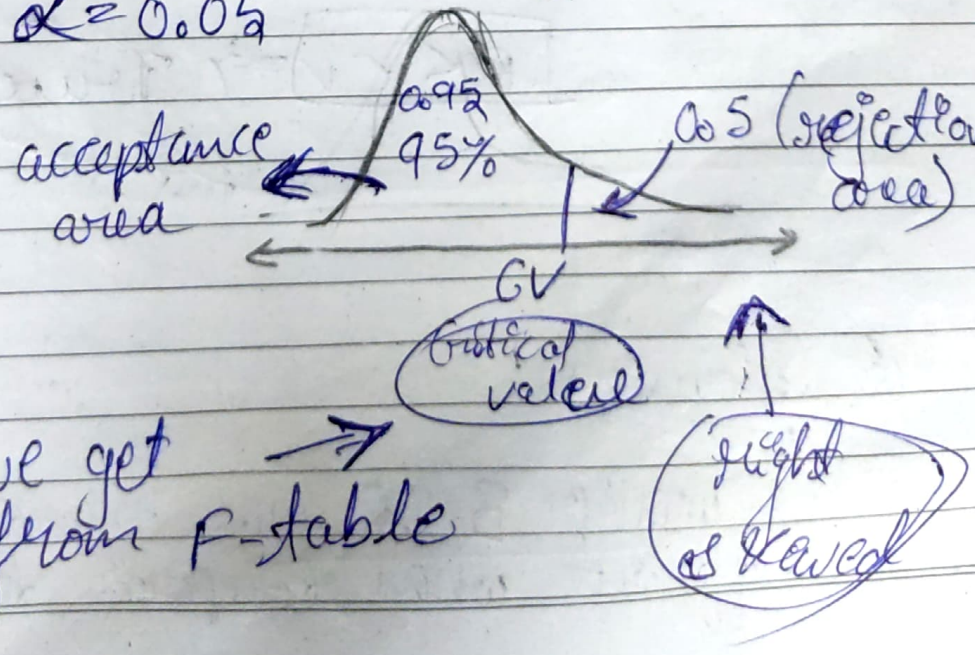
④ Decision rule

$dof\ 1 = 6-1 = 5$

$dof\ 2 = 8-1 = 7$

} use F table

$\alpha = 0.05$



CV ⇒ 3.9715

for $\alpha = 0.05$, $d_1 = 5$ and $d_2 = 7$

decision

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

Now,

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

$F = 1.33 < 3.9715$ (CV)

F < CV ⇒ True

we fail to reject null hypothesis

⑥ Conclusion:

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

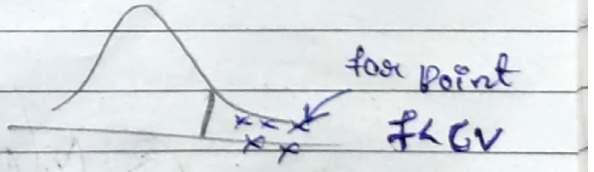
* F-test in python

ftest -> f_test = np.var(worker1) / np.var(worker2)

Critical value:
- dof1 = len(wor1) - 1
- dof2 = len(w2) - 1
- sign = 0.05
- CR = 1 - sig
- CV = stat.f.ppf(q = CR, dfn = dof1, dfd = dof2)

(Percent point function)

Conclusion:
- if f < CV: # reject
- else: # accept



imp

*ANOVA [Analysis of Variance]

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

ANOVA

① Factors (Variable) → Can be Independent & dependent

② levels

Eg: ① Medicine (Factor)

[Dosage] 5mg 10mg 15 mg → 3 levels

② Mode of payment (Factor)

levels GPAY Phonepe IMPS DD NETT

may be (interview)

Assumption in Anova

① Normality of sampling distribution of Mean
→ [C.L.T]
The distribution of sample mean is normal distributed.

② Absence of outliers.
→ Outling scores need to be removed from the dataset.

③ Homogenity of Variance
→ population variance in different levels of each independent variable are equal.

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

④ Sample are independent and random.

Hypothesis testing in ANOVA [Partitioning of Variance in the ANOVA]

Null hypothesis: $H0 \Rightarrow \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

Alternate hypothesis: $H1 \Rightarrow$ atleast 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 statistics

We use F test

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

Variance between samples ⇒ 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	2	8
3	1	4
Σx_1^2 3	$\bar{x}_2^2$ 19/5	$\bar{x}_3^2$ 4

Variance within sample ↓

how much far the point from the mean.

Here, in eg

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

H_1: atleast 1 sample mean is not equal

One Way ANOVA

① Doctors want to test a new medications which reduces headache. they splits the participant into 3 Condition [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 atleast 2 levels, levels are independent.

Std tells how far a point from mean.

Variance tells the spread of data.

Qns 6

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

① Define Null & alternate hypothesis?

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

H_1: ~~atleast~~ not all are equal
(or) atleast one of not

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

③ degree of freedom

Population size $N = 21$, no of sample $a = 3$, sample size $n = 7$

$df_{between} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

$18 + 2 = 20$

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

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

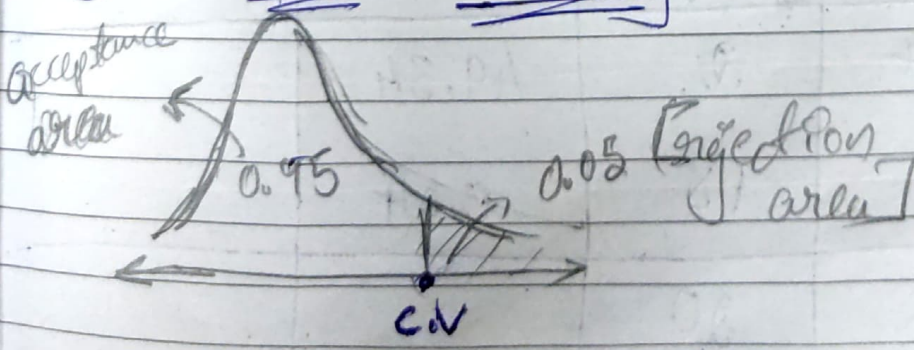
⇓

used on F table with $\alpha = 0.05$

⇓ Find

Critical Value

④ Decision boundary



Critical value
when $\alpha = 0.05$, $df_1 = 2$
$df_2 = 18$ ⇒ 3.155546

decision rule

If F is greater the 3.155546, reject the Null hypotheses

(5) Calculate F-test statistics

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

	(SS) Sum of squares	(df) degree freedo	(MS) Mean square	(F) F
Between	98.67	2	49.34	
within	10	18	0.54	
Tota	108 (calculate)	20 (before calculate)	(SS/df)	

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

Summation

15 mg $= 9+8+7+8+8+9+8 = 57$

30 mg $= 7+6+6+7+8+7+6 = 47$

45 mg $= 4+3+2+3+4+3+2 = 21$

$\therefore$ SS between $= \frac{57^2 + 47^2 + 21^2}{7} - \frac{[57+47+21]^2}{21}$

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

(ii) $SS_{within} = \Sigma y^2 - \frac{\Sigma [\Sigma a_i)^2}{n}$

$\Sigma y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + 8^2 = 853$

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

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

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

$$= \frac{19.34}{0.54} = 86.56$$

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

Here, $(F = 86.56) > 3.5546$ so we reject null hypothesis (h_0)

⑥ Conclusion

not all are equal

(or)

atleast one is not equal