Introduction to Statistics

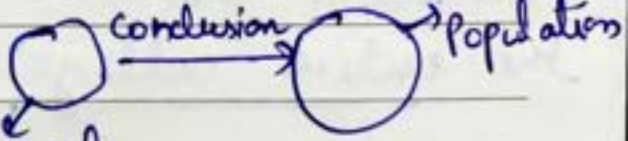
Date 06 / 03 / 23

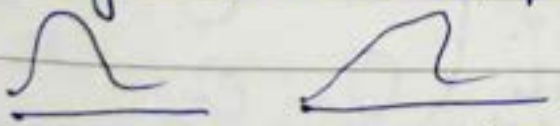
① Defⁿ :- Statistics is the science of collecting, organizing and analyzing the data. (Decision making Process)

Data → facts or pieces of information.

eg → Height of students in the class
{175cm, 160cm, 165 cm, 190 cm, 185cm - - }
IQ of students {85, 100, 95 - - - }

② Types of Statistics :-

① Descriptive Stats	② Inferential Stats
→ It consists of organising and summarising the data.	→ It consists of using data you have measured to form conclusion/output
→ Measure of central tendency [Mean, Median, Mode]	
→ Measure of Dispersion [Variance, Std]	→ z-test, t-test
→ Histogram, Bar chart, pie chart	→ Hypothesis testing, p value



Diff. charts

Population → The total collection of all the elements that we are interested in is called population.

Eg → Taking data of students in Maths class
- Total stud = 50
- Height of stud = [175cm, 180, 185 - - 145]

Descriptive Question

Q What is the avg height of students.

Sol^n Mean (Avg) = $\frac{\text{Summation of height}}{\text{Total students}}$

Q What is the common height of the students.

Sol^n Let us say 140 cm is the most frequent height. = 140 cm (Mode)

Inferential Question – (Sample data)

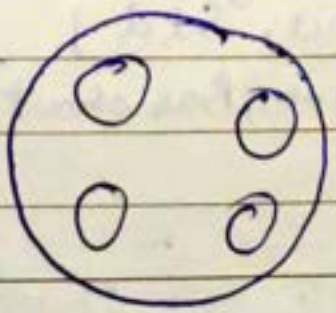
Q Are the avg height of the students in the classroom similar to what you expect in the entire college'. → (population)

→ Sample Data and Population Data:-

Eg → Exit poll

Party A gets 56% Vote

Party B get 46% Vote



Sample → A subgroup of the population that will be studied in detail is called sample. Date / /

→ In this eg. Sample data is collected from many parts of the country to know some rough data that who will win the upcoming election result.

Than conclusion will be party A will win because they get more no. of votes acc. to sample data so it is population data.

③ Types of Data :-

- ① Quantitative (Numerical)
  - Discrete — whole No
  - Continuous — Any num. value

→ Quantitative data are numerical form data in which all the arithmetic formula applies eg +, -, *, /;

→ Discrete are whole no's.

eg Eg. No. of Bank a/c,

- ② Qualitative (categorical data)
  - Nominal
    - eg Grender, (M/F)
    - Blood Gr.,
    - Color.
  - Ordinal
    - eg Ranking data.
    - Like review system
      - Good ①
      - Satisfactory ②
      - Bad ③

Mo Tu We Th Fr Sa Su

Date / /

- No. of children in a family. eg 2,3,4,5 not like 1.5, 2.5, 3.5
→ Continuous data contain any values.
eg Weight, Height, temp., speed.

→ In Nominal data we just categorise data.
→ In Ordinal data we assign the rank to the data acc. to preferences.

Ques What type of variable Marital Status is?
Ans Nominal Data because it is just categorical data like Married, Unmarried, Single,

Ques What kind of variable Nile River length?
Ans Continuous

Ques What type of variable is movie duration?
Ans Continuous

Date / /

④ Scale of Measurements of Data :—

i) Nominal Scale Data →

→ Qualitative / Categorical Variable
→ Eg Gender, Colors, Labels
→ Order does not matter

Ex → Survey b/w 10 people about colour preferences -

- Red → 5 person like it → 50% opted
- Blue → 3 " " " → 30% opted
- Yellow → 2 " " " → 20% opted

→ In this order/raking doesn't matter but only data matters in numerical form.

ii) Ordinal Scale Data →

→ Ranking and order matter
→ Difference can't be measured

Eg → Qualification analysis

PHD	1st Rank	Difference can't be measured based on rank.
BE	3rd Rank	
Masters	2nd Rank	
B.com	5th "	
BSC	4th "	

Date / /

eg

Marks	Rank	
100	1st	→ It is not a purely ordinal data because diff. b/w the marks can be measured.
90	2nd	ordinal data
65	3rd	
70	4th	
30	5th	

Diff. can be measured

iii) Interval Scale Data

→ The Rank & order matters
→ Difference can be measured (excluding ratio).
→ It doesn't have '0' (zero) starting value.

Eg → Temperature

It is not a starting value	30 F	→ 60 − 30 = 30	$\frac{30}{60} = \frac{1}{2}$ #
0f	60 F	→ 80 − 60 = 20	$\frac{60}{80} = \frac{3}{4}$
−1of	90 F		$\frac{80}{90} \to \frac{8}{9}$
−2of → Ratio can't be relatable.			$\frac{1}{2} \neq \frac{3}{4} \neq \frac{8}{9}$
−3of			

→ Temp. can have any negative value.

Date / /

iv) Ratio Scale Data →

→ Order and Rank matter

→ Diff. and ratio are measurable

→ It does have a '0' (zero) starting

eg Grades → 100, 90, 60, 55, 45, 50, 75

0, 100, 90, 75, 60, 50, 45

10 15 15

$\frac{100}{50} \Rightarrow 2:1$ → So 50 is 2 times less than 100 or we can say that marks obtained by 100 grade is too greater than 50.

→ So Ratio can be relate according the mark distribution.

→ It can be start from 0 because there will be no negative grade.

Ques → length of diff. rivers in the world?

Ans Ratio Scale Data

use in EDA & feature engineering

Date / /

⇒ Measure of Central Tendency :-

① Mean - It is the avg value of the data.

Population (N) | Sample (n)

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

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

Sample mean $\boxed{(s) = \sum_{i=1}^{n} \frac{X_i}{n}}$

$\mu = \frac{(1+1+2+2+3+\_ \_ \_ \_)}{10}$ $(N=10)$

$= \frac{32}{10} = 3.2.$

$s = \frac{(1+1+2+1 \_ \_ \_)}{10}$ $(n=10)$

$= \frac{32}{10} = 3.2$

→ In some condition N, n are different and values will also be different.

Date / /

(2) **Median** → Middle most value of the series.

4, 5, 2, 3, 2, 1

Sort → 1, 2, 2, 3, 4, 5.

If n is even,

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

$\frac{2+3}{2} \Rightarrow \frac{5}{2} = 2.5.$

If n is odd,

1, 2, 2, 3, 4, 5, 7

↓

Median → 3

→ **Why median is used ?**

{ 1, 2, 3, 4, 5 }

$S = \frac{1+2+3+4+5}{5} \Rightarrow \frac{15}{5} = 3.$

Median ⇒ {1, 2, 3, 4, 5}

↓

Median = 3.

→ If we add an outlier to this distribution

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

100 is not a member of this distribution or doesn't look like it belong to this distribution.

Date / /

Mean (S) = $\frac{1+2+3+4+5+100}{6} \Rightarrow \frac{115}{6} \approx 18$

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

$\frac{3+4}{2} \Rightarrow \frac{7}{2} \Rightarrow 3.5$

∴ In this median is more accurate as there is no influence of an outlier.

→ Therefore when we have an outlier to our distribution pt is always advicable to get Median it is more accurate.

(3) Mode = Most frequent number

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

Mode = 1

Type of Flower	Age
Lilly	10
Rose	3
— (Rose)	5
Sunflower	— (Mean or Median) in case of outlier.
Rose	8

→ Blank space will be replaced by Mode.

Date / /

⇒ Measure of Dispersion :—

① Variance → Spread of the data.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
x_i → Data points	$\bar{x}$ → sample mean
μ → population mean	n → population size
N → Population size	

Eg → {1, 2, 3, 4, 5} ⇒ sample

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \quad \bar{x} = \frac{1+2+3+4+5}{5}$$

$$\bar{x} = \frac{15}{5} = 3$$

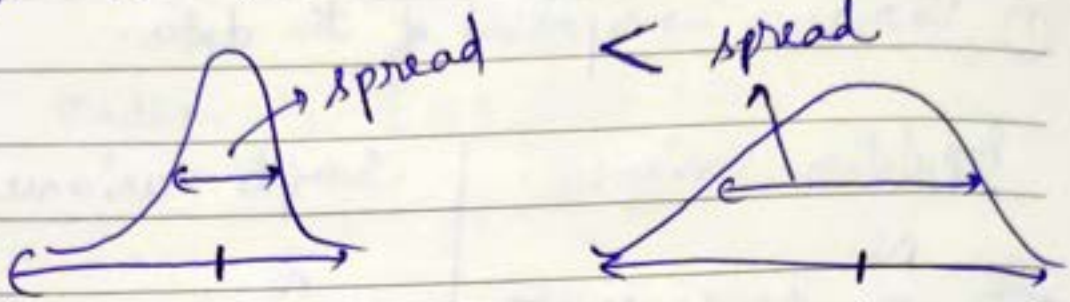
x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
		10

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

$$s^2 = 2.5.$$

Date / /

eg $S^2 = 2.5$ $S^2 = 6.5$

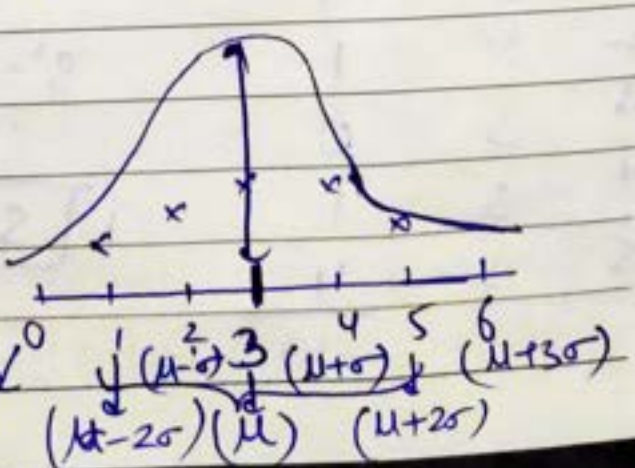


(2) Standard Deviation :-

Population Std	Sample Std
$\sigma = \sqrt{\text{variance}}$	$S = \sqrt{\text{sample variance}}$
	$S^2 = 2.5$
	$\sqrt{S^2}$ = sample std.

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

$\mu = 3$
$\sigma = 1$



Date / /

③ Random Variable :- It is a process of mapping the output of a random process or experiment to a number.

eg Tossing a coin
Rolling a dice
Measure the temp. for next day

$X = \begin{cases} 0 & , \text{if head} \\ 1 & , \text{if tail} \end{cases}$ Quantifying a Random process

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

$\{3,4,5,6,1,2,2\} = 23$

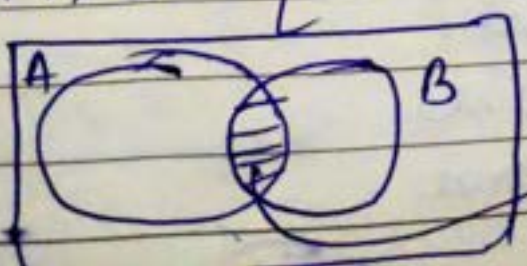
④ Sets

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

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

i) Intersection ($\cap$) → most common element

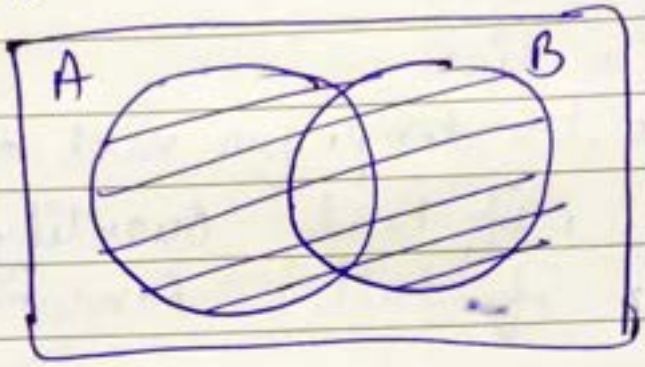
$A \cap B = \{3,4,5,6,7\}$



Date / /

ii) Union (U) → Combination of all the element present in the both sets

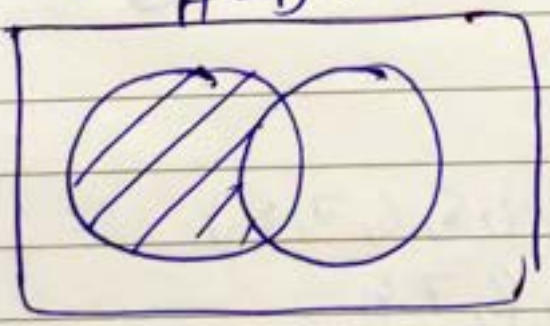
$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$



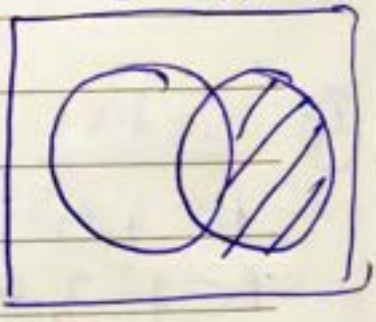
iii) Difference (-) → Removing the element

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

A - B



B - A.



iv) Subsets → All the elements in one set has to be present in other set to be called as subset.

$A \longrightarrow B \Rightarrow$ False

$B \longrightarrow A \Rightarrow$ True

Date / /

v) Supersets → One set has all the elements which another set have.

A ⟶ B ⇒ True

B ⟶ A ⇒ False

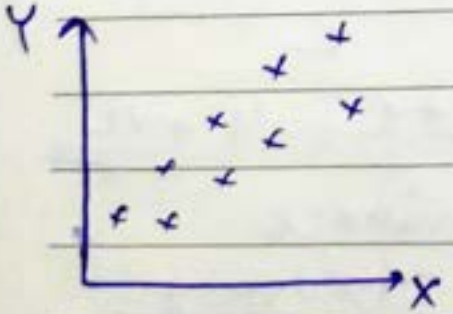
⇒ Covariance and Correlation :—

eg.

X	Y
2	3
4	5
6	7
8	9

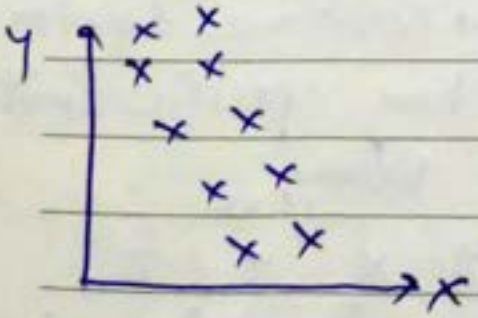
x ↑	y ↑
x ↓	y ↓
x ↑	y ↓
x ↓	y ↑

→ Possible relation



x ↑	y ↑
x ↓	y ↓

→ +ve cov



x ↑	y ↓
x ↓	y ↑

→ -ve cov

Date / /

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

Cov. (X, Y)

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

Spread of the data ⇐ $Cov.(X,X) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$

X	Y
2	3
4	5
6	7

$\bar{x} = \frac{12}{3}$ $\quad \bar{y} = \frac{15}{3}$

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

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

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

$$= \frac{-2 \times -2 + 2 \times 2}{2}$$

$$= \frac{4+4}{2} = 4 \text{ +ve}$$

X & Y having a positive variance.

Advantage	Disadvantage
→ Relationship b/w X & Y +ve or -ve value	→ Covariance doesn't have specific limit value

eg: X (+100) Y (+500) Z

Date / /

→ Pearson Correlation Coefficient :-

Range [-1 to +1]

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

- The more the value towards +1 the more the correlated (+ve) it is (x, y)
- The more the value towards -1 the more -ve correlated it is (x, y)

eg Dataset :- 1000 features (ML Models)

o/p or dependent feature ↑

Size of house	No. of Rooms	Location	No. of people staying	Price

Independent feature

- In above example there is a +ve correlation b/w size of house, no. of rooms, location and price as one increases other will also be increases. but in case of no. of people staying there will be no effect on the price value. i.e. no relation. and if the house is haunted therefore it is -ve correlated with price.

Date / /

- Feature Selection → near to 0 → Drop no. of people staying

→ Spearman Rank Correlation → It is better than pearson correlation.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \times \sigma_{R(y)}}$$

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

⇒ Histograms [Frequency] ⇒ Count

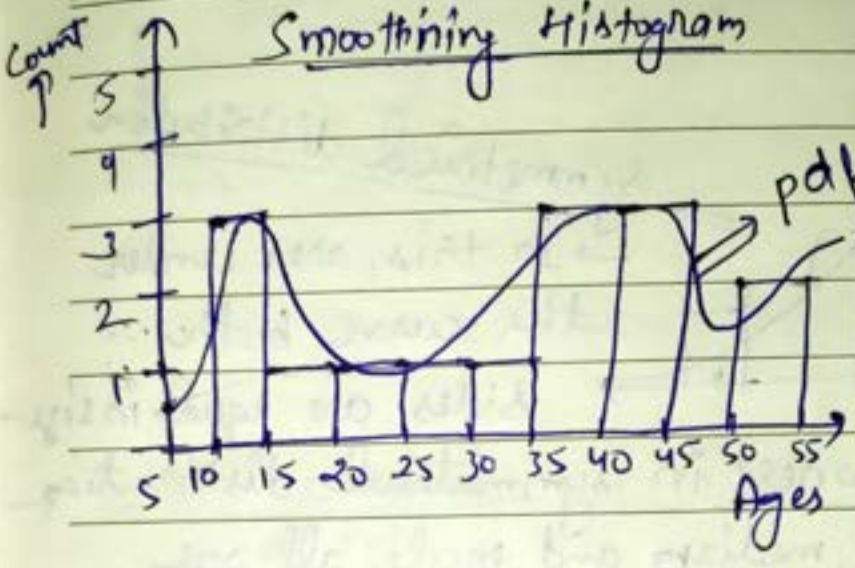
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 43, 50, 51}

Bins = 10 (buckets)

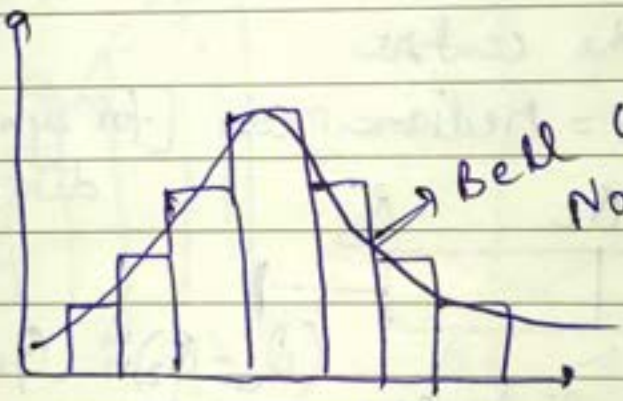
Date

bin size = $\frac{50}{10}$ = 5

Smoothing Histogram

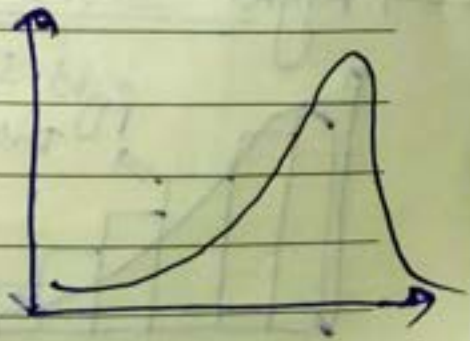


pdf [Probability Distribution function]
⇓
Kernel Density estimator



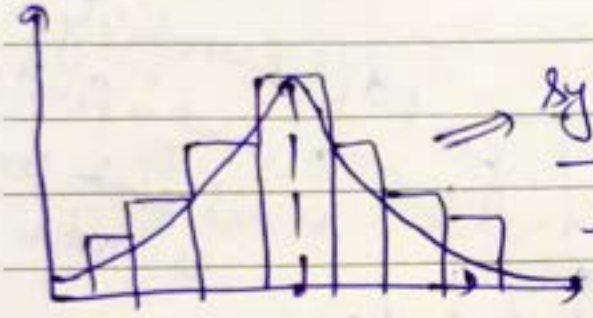
Bell curve
Normal/Gaussian Distribution

→ log Normal Distribution



Date / /

⇒ Skewness



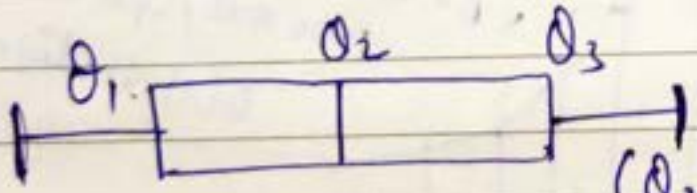
⇒ Symmetrical distribution

→ In this, area under the curve both sides are equal in size.

→ No skewness in symmetrical distribution

→ The mean, median and mode all are perfectly at the centre

i.e. Mean = Median = Mode (for symmetrical dist.)

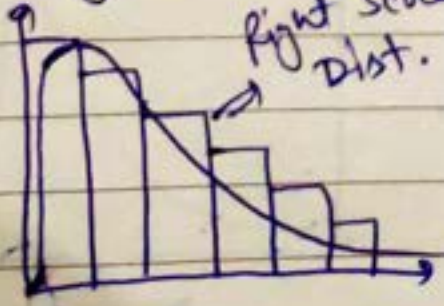


$(Q_3 - Q_2) \approx (Q_2 - Q_1)$

Box Plot (for symm. dist.)

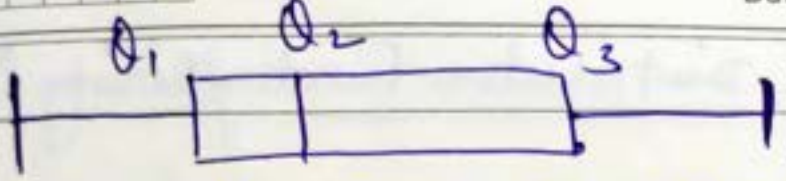
→ Normal / Gaussian Distribution.

→ Right Screwed



⇒ Positive Screwed (Log Normal Dist.)

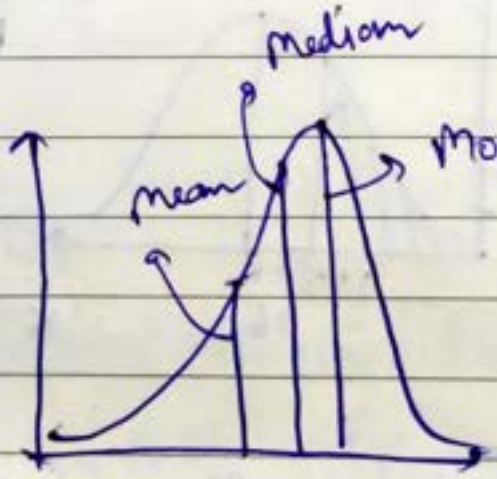
Date / /

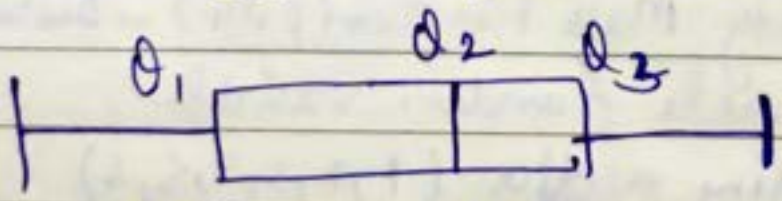


i.e. $(Q_3 - Q_2) > (Q_2 - Q_1)$ → for right screwed

i.e. Mean > Median > Mode.

→ Left Screwed





$(Q_2 - Q_1) > (Q_3 - Q_2)$ → for left screwed

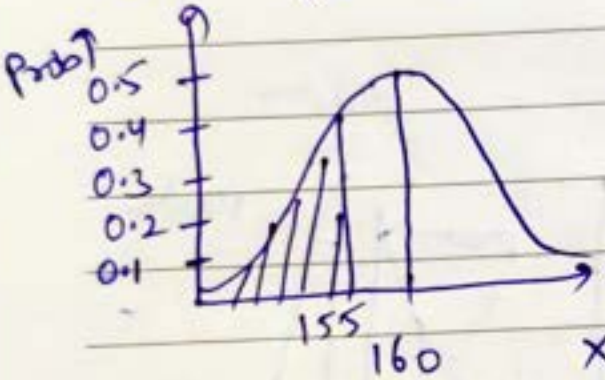
i.e. Mean < Median < Mode

Date / /

→ Probability Distribution Function/Density Function

① Probability Density Func^n (Pdf) → Distribution of the continuous data.

eg Height of students



[0 to 1]

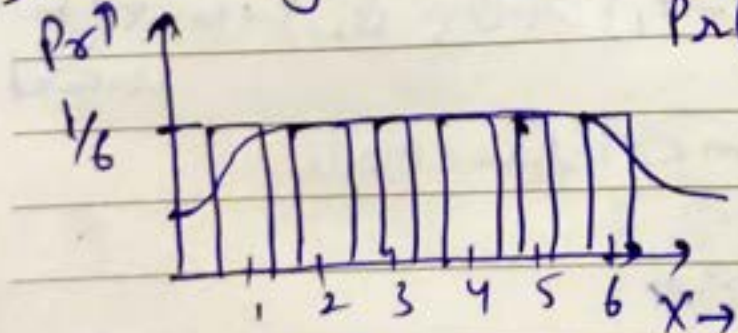
$Pr(X < 155)$ = Area under the curve.

0.3, 0.4, 0.5

→ Kernel Density Estimator

② Probability Mass Function (PMF) → Distribution of a discrete random variable

eg Rolling a die (1, 2, 3, 4, 5, 6)



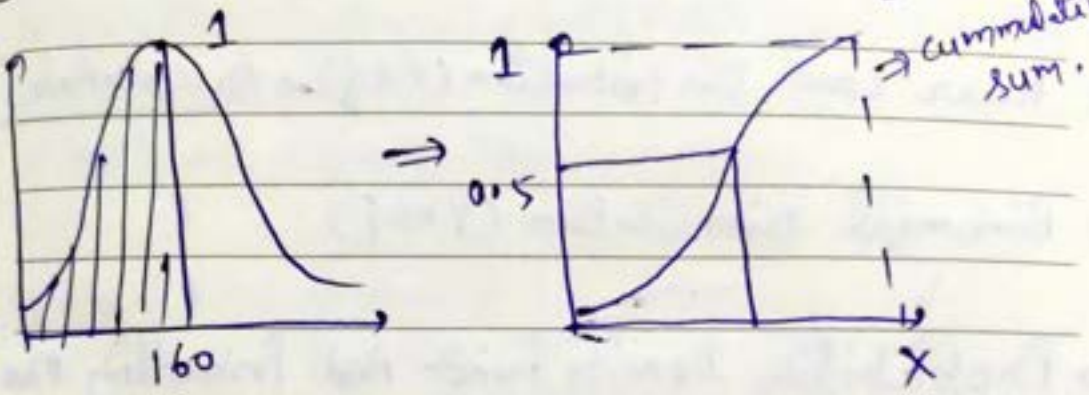
$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$

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

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

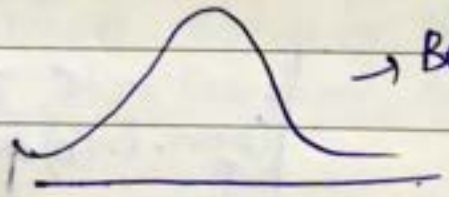
Date

③ Cummulative Distribution func^{n} (cdf) –



⇒ Types of Probability Distribution :–

① Normal/Gaussian Distribution → (in Pdf)



② Bernoulli Distribution → (in Pmf) → outcome are of success or failure

③ Uniform Distribution

④ Log Normal Distribution (Pdf)

Mo Tu We Th Fr Sa Su Date / /

(5) Poisson Distribution (Pmf)

(6) Power Law Distribution (Pdf) → 80-20% Rule

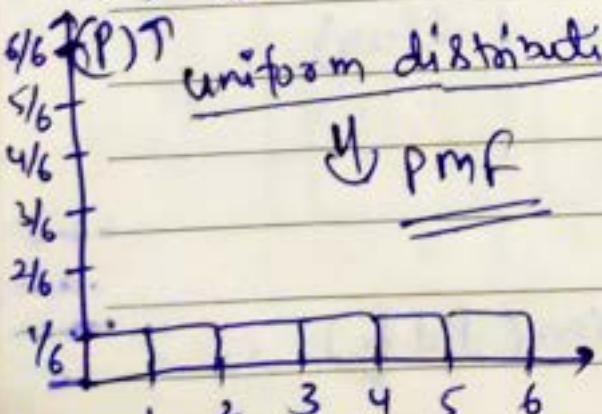
(7) Binomial Distribution (Pmf)

⇒ Probability Density Func^n And Probability Mass Fun.
Cummulative Distributive Func^n (CDF)-

(1) PMF :-
→ Used for discrete random variable

PMF

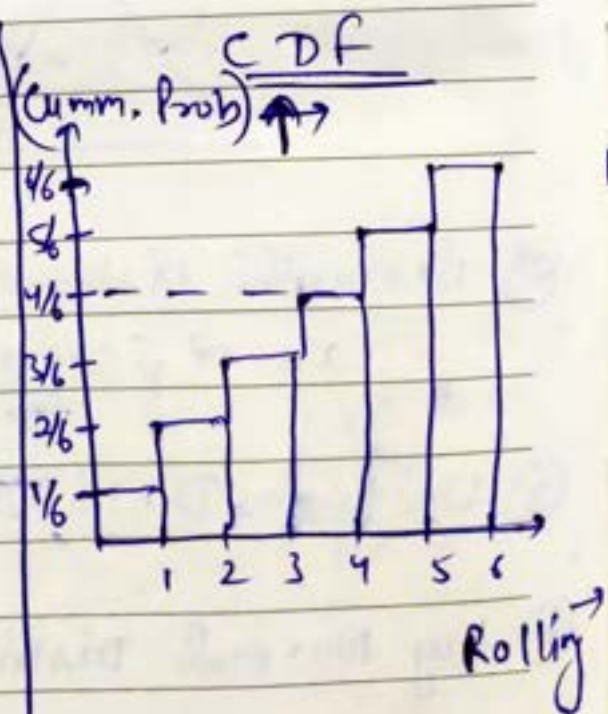
→ Rolling a dice
{1, 2, 3, 4, 5, 6}, P(1) = 1/6,
P(2) = 1/6 - - - - P(6) = 1/6.



uniform distribution

↓ PMF

CDF



Date / /

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

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

$= 1/6 + 1/6 + 1/6 + 1/6 = 4/6 = 2/3.$

$P(x \leq 6) = P(x=1) + P(x=2) + \dots + P(x=6)$

$= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6$

$= 6/6 = 1$

But in CDF,

$P(x \leq 4) = 4/6 = 2/3.$

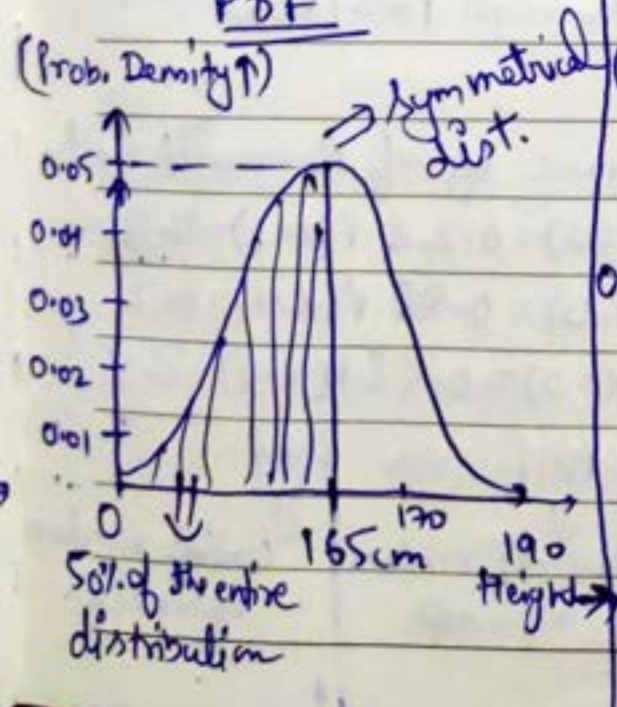
$P(x \leq 6) = 6/6 = 1.$

(2) Probability Density func^n (PDF) :-

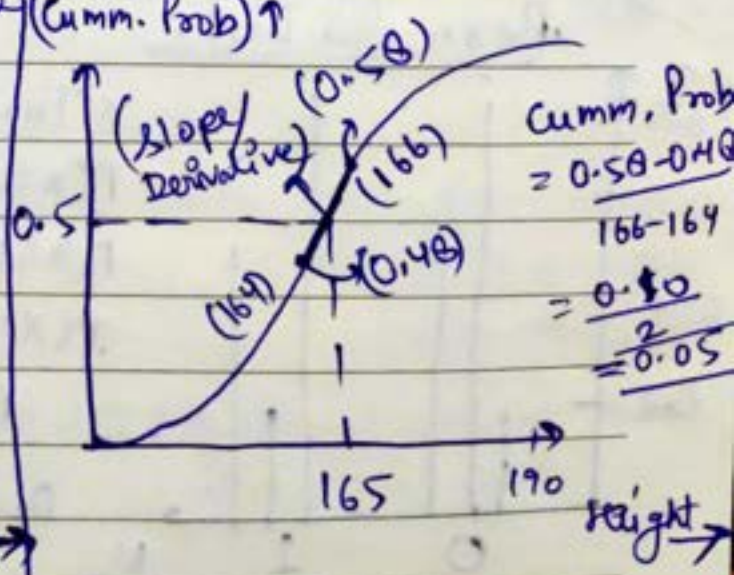
→ Continuous random variables

Prob. Density → Gradient descent of CDF

PDF



CDF



Date / /

→ Bernoulli Distribution :- [Binary o/c]

In probability theory and statistics, the Bernoulli distribution, named after Swiss mathematician Jacob Bernoulli, is the discrete probability distribution of a random variable which takes the value 1 with prob. p and the value 0 with prob. q = 1-p.

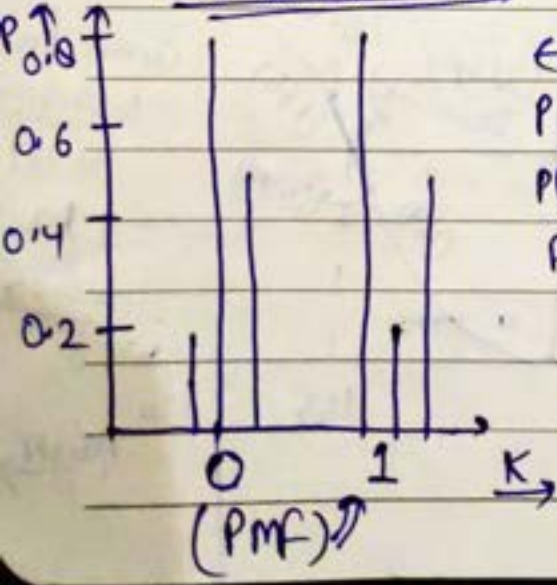
i.e. Outcomes are Binary {0, 1}

eg Tossing a coin. {0, 1}
(H, T)

$P(T) = 1/2 = 0.5 = p$

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

eg whether the person will Pass/fail → Bernoulli Dist.



← Three eg of Bernoulli dist.

$P(X=0) = 0.2$ & $P(X=1) = 0.8$

$P(X=0) = 0.8$ & $P(X=1) = 0.2$

$P(X=0) = 0.5$ & $P(X=1) = 0.5$

PMF vs Pdf

↓ Discrete random variable

↓ Contin: random variable.

Date / /

· PMf → K=0 or 1

$$\boxed{P(X=K) = p^K (1-p)^{1-K}}$$

① $P(X=1) = p^1 (1-p)^{1-1}$
$= p^1 (1-p)^0 = p$ //

② $P(X=0) = p^0 (1-p)^{1-0}$
$= (1-p)^1 = q$ //

Simplified way of showing PMf →

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

→ Mean, Variance & Std. Deviation of Bernoulli Dist :-

Mean $E(K) = \sum_{i=1}^{K} K \cdot p(K)$ → $1 \times 0.6 + 0 \times 0.4$
$= 0.6 = p$

K = 1 or 0

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

$P(K=0) = 1-p = 1-0.6 = 0.4 = q$ //

Date / /

Median of Bernoulli Dist.

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

Variance & Std Deviation →

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

$$\text{Std} = \sqrt{p \cdot q}$$

→ **Binomial Distribution →** The binomial dist. with parameters n and p is the discrete probability distribution of the no. of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued o/c: success (with prob. p) or failure (with prob. $q = 1-p$).

→ A single success/failure exp. is also called a Bernoulli trial or Bernoulli exp. ($n = 1$).

Date / /

eg Tossing a coin {Bernoulli Distribution}

$P(H) = 0.5 = 1-p$ ↓ 0 $\qquad$ $P(T) = 0.5 = p$ ↓ 1

eg Tossing a coin for 10 times.

	1st time	2nd	3rd	4th
$P(T) = p$	$\cdot p$	p	p	
$P(H) = 1-p$	$1-p$	$1-p$	$1-p$	

↓ Binomial Distribution

① Parameters → $n \in \{0, 1, 2, 3\}$ → no. of trials or exp.

$p \in [0, 1]$ → success or prob. for each trial

$q = 1-p$.

② PMF $\boxed{P(K) = {}^{n}C_{K}\, p^{K} \cdot (1-p)^{n-K}}$, K → no. of successes.

③ Mean of Binomial Dist.

$\text{mean} = n \cdot p$.

④ Variance and Std.

$\text{Variance} = npq$

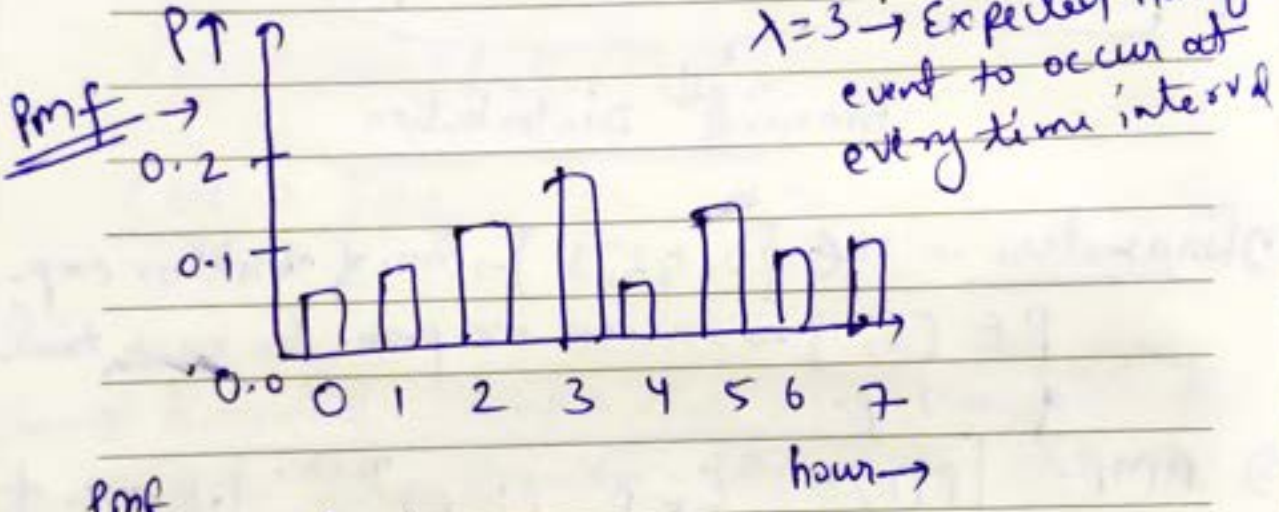
$\text{Std} = \sqrt{npq}$

Date / /

→ Poisson Distribution :-

- Discrete Distribution (Pmf)
- Describes the no. of events occuring in a fixed time interval.

Eg No. of people visiting hospital every hour.
bank ____
airport ____.



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

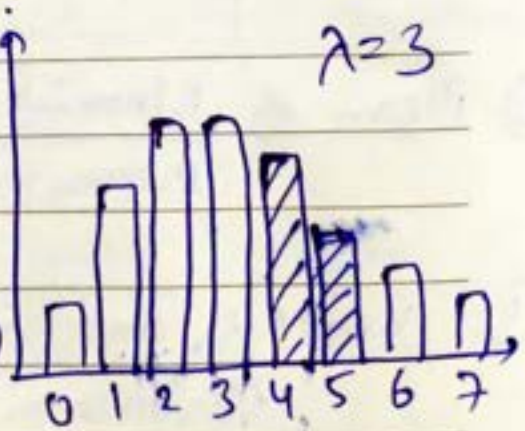
Pmf

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

$$= \frac{e^{-3} 3^5}{5!} = 0.101 = 10.1\%.$$

$$P(x=4) + P(x=5) = \frac{e^{-3} \times 3^4}{4!} + 0.101$$

$$= \text{____}$$



Date / /

Mean & Variance of Poisson Distr.

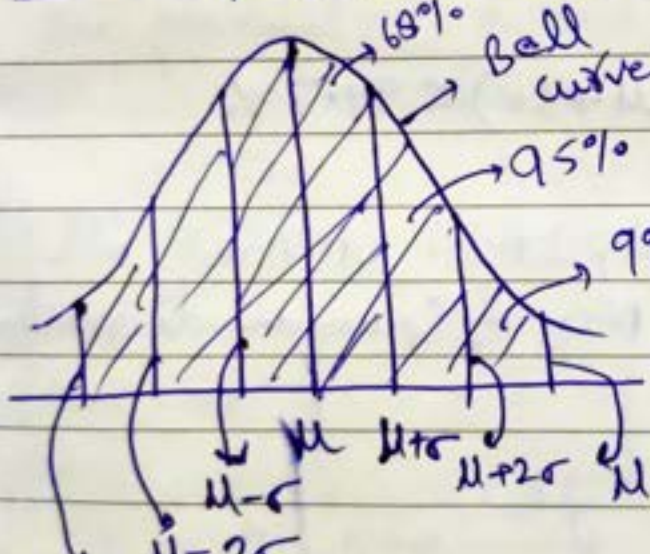
Mean = $E(X) = \mu = \lambda \times t$

Variance = $E(X) = \mu = \lambda \times t$

$\lambda \rightarrow$ Expected no. of events to occur at every time interval

$t \rightarrow$ time interval

→ Normal/Gaussian Distribution :—



⇒ Symmetric distribution.

$X = \{ \underbrace{- - -}_{68\%} \underbrace{- - - -}_{95\%} \underbrace{- - - - -}_{99.7\%} \}$

Date / /

- Emperical Rule [3-Sigma Rule]

68% - 95% - 99.7%

Q-Q Plot → To check whether a distribution is Gaussian/Normal Distribution.

Probability

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

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

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

eg Weight, height, IRIS DATASET all comes in the Normal/Gaussian distribution.

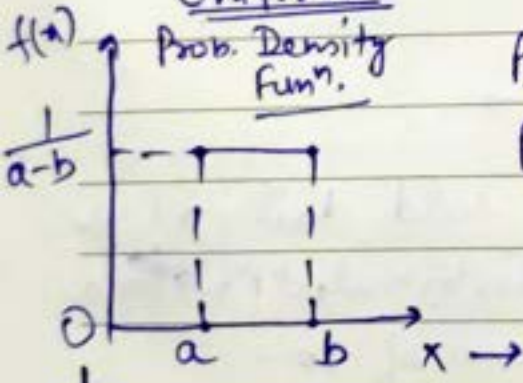
Date / /

⇒ Uniform Distribution :-

1. Continuous Uniform Dist. (pdf)
2. Discrete " " (pmf)

1. Continuous Uniform Dist. (Cont. Random Variable)

Cont. Unif. Dist or rectangular dist. is a family of symmetric prob. distributions. The distribution describes an exp. where there is an arbitrary outcome that lies b/w certain bounds. The bounds are defined by the parameters a and b which are the min. and max. values.

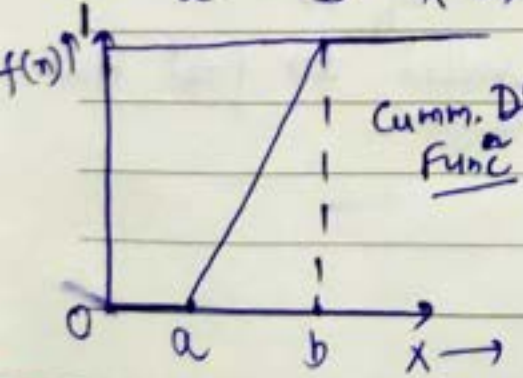


Notation :- $U(a, b)$

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

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

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



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

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

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

Date / /

Uniform Distribution →

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

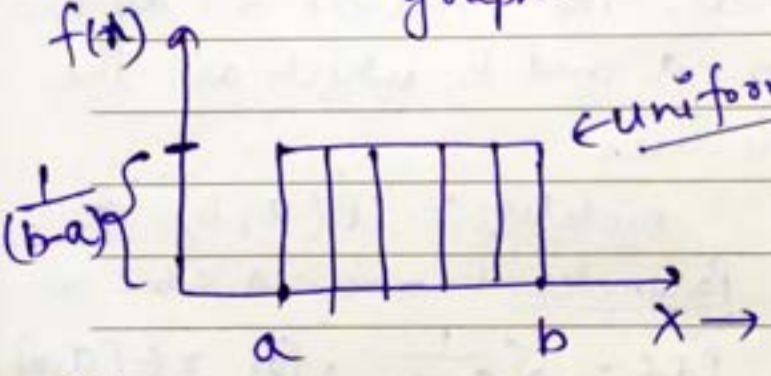
$P(1) = 1/6$, $P(2) = 1/6$, $P(3) = 1/6$

In uniform dist. the probability of getting the o/c is equal.

A variable X is said to be uniformly distributed.

$$f(X) = \frac{1}{b-a} \quad \text{for } -\infty \le a \le b < \infty$$

⇓ graph



(X) The no. of candies sold daily at a shop is uniformly distributed with a max. of 40 and minimum of 10.

i) Probability of daily sales to fall b/w 15 and 30 ?

Date / /

$x_1 = 15$, $a = 10$

$x_2 = 30$, $b = 40$

$\frac{1}{b-a} = \frac{1}{40-10} = \frac{1}{30}$

$$P(15 \leq x \leq 30) = (x_2 - x_1) * \frac{1}{(b-a)}$$

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

$$= 15 \times \frac{1}{30} \Rightarrow 0.5 // = 50\%$$

$$P(x \geq 20) = (40-20) \times \frac{1}{30}$$

$$= 20 \times \frac{1}{30} = 2/3 = 0.666$$

$$= 66.6\%.$$

→ Mean and Variance of Uniform Distribution

Mean → $E(X) = [a+b]/2 = (40+10)/2 = \frac{50}{2}$

$= 25.$

Variance → $V(X) = (b-a)^2/12 = \frac{(40-10)^2}{12} \Rightarrow \frac{30^2}{12}$

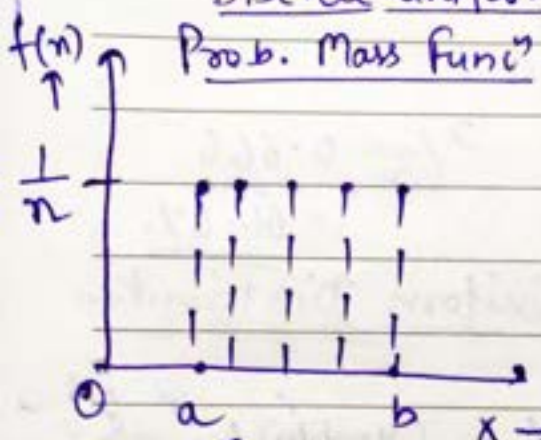
$\Rightarrow \frac{900}{12} = 75$

Date / /

② Discrete Uniform Distribution (Discrete Random Variable)

→ It is a symmetric prob. distr. wherein a finite no. of values are equally like to be observed; every one of n values has equal prob. 1/n. Another way of saying 'discrete uniform dist' would be 'a known, finite no. of outcomes equally likely to happen'.

Discrete uniform Prob. Mass Func"



n = 5, where $n = b - a + 1$

Eg Rolling a dice

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

$P(1) = \frac{1}{6}, P(2) = 1/6, P(3) = 1/6$

$a = 1, b = 6$

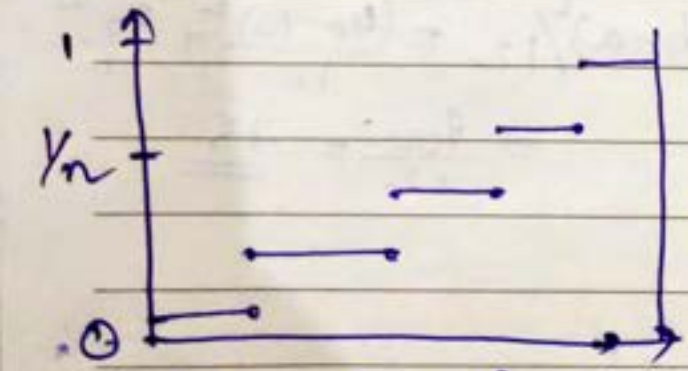
$n = b - a + 1 = 6$ //

Notation → U(a, b)

Parameters → a & b, $b \geq a$

PMF = 1/n.

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



Cumm. Dist. Func"

(SND) Date / /

→ Standard Normal Distribution → When we convert Normal/Gaussian distribution mean (μ) and variance (σ) as $\mu = 0$ and $\sigma = 1$ called as Standard Normal Distribution. Conversion of this possible by Z-score.

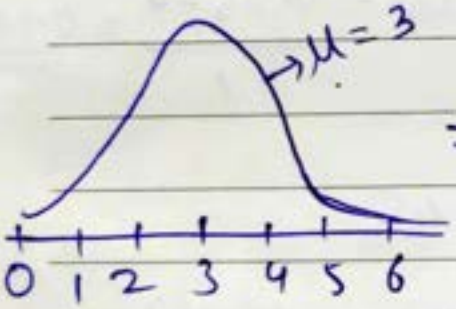
Notation

$X \sim SND(\mu = 0, \sigma = 1)$

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

$\mu = 3$

$\sigma = 1.414 \approx 1$



$\mu = 0$

$\Rightarrow \sigma = 1$

$X = \{1, 2, 3, 4, 5\}$, $\text{Z-score} = \frac{x_i - \mu}{\sigma}$, y

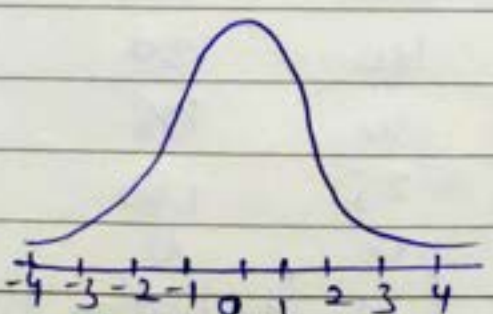
$1 \rightarrow \frac{1-3}{1} = -2$ $\quad y = \{-2, -1, 0, 1, 2\}$

$2 \rightarrow \frac{2-3}{1} = -1$

$3 \rightarrow \frac{3-3}{1} = 0$

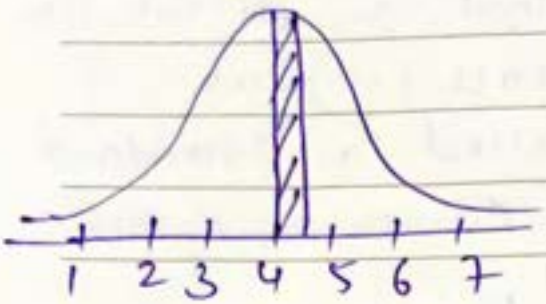
$4 \rightarrow \frac{4-3}{1} = 1$

$5 \rightarrow \frac{5-3}{1} = 2$



Date / /

eg $\mu = 4, \sigma = 1$



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

$1 \to \frac{1-4}{1} = -3$

$2 \to \frac{2-4}{1} = -2$

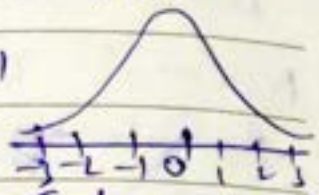
$3 \to \frac{3-4}{1} = -1$

$4 \to 0$

$5 \to 1$

$6 \to 2$

$7 \to 3$



Q How many Std. deviation 4.25 is away from the mean?

Ans $x_i = 4.25$

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

Q $x_i = 2.5$

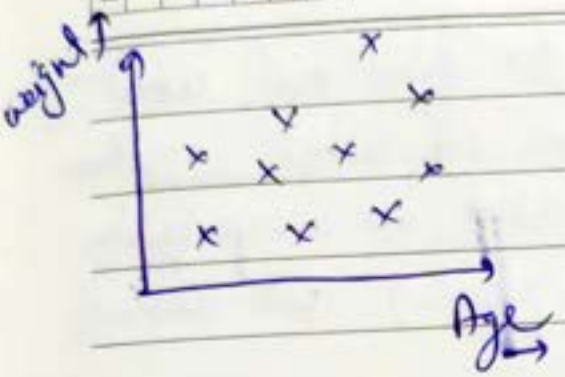
Ans $z\text{-score} \Rightarrow \frac{2.5-4}{1} = -1.5$

Eg → Dataset

Age (yr)	Weight (kg)	Height (cm)	Salary (INR)
24	70	175	40K
25	60	160	50K
26	55	150	60K
27	40	130	30K
30	30	175	20K
31	25	100	70K
↓	↓	↓	↓

(Standardization) using z-score

Date / /



$$Z\text{-score} = \frac{x_i - \mu_{age}}{\sigma} \quad , \quad \frac{x_i - \mu_{weight}}{\sigma}$$

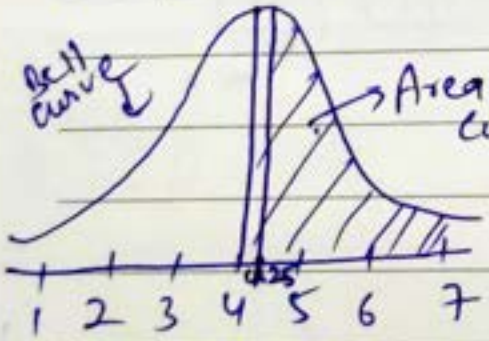
→ Problems statement on Z-score :-

Z-table

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

$\mu = 4, \sigma = 1$

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



⇒ Symmetric Distribution

Q] What % of scores fall above 4.25 ?

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

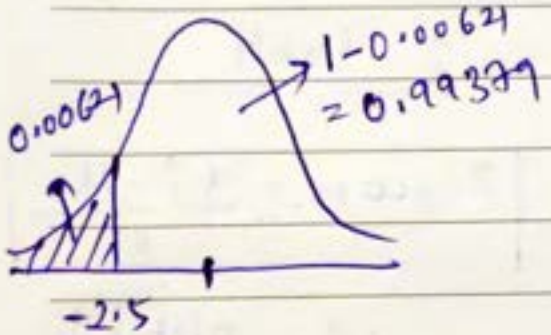
$= 0.25$ → Std. Deviation away from mean.

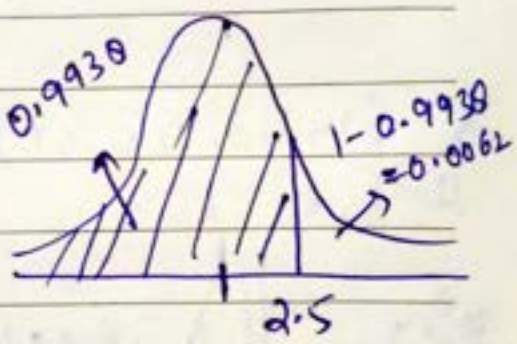
Date / /

Note → When we need to find out what is the % of distribution or what is the area under the curve based on a specific std. deviation (derive from z-score) we use z-table.

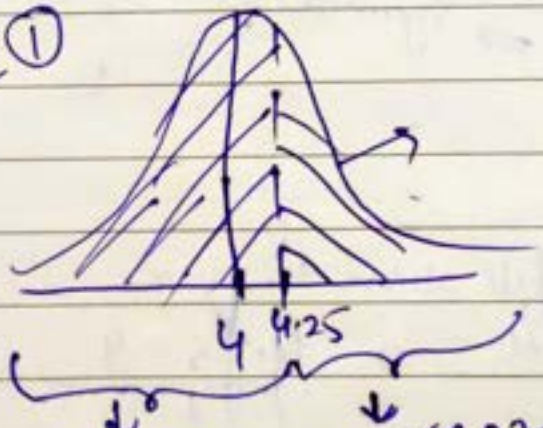
→ There are two types of z-table →

① Negative z table ② Positive z table





Soln ①



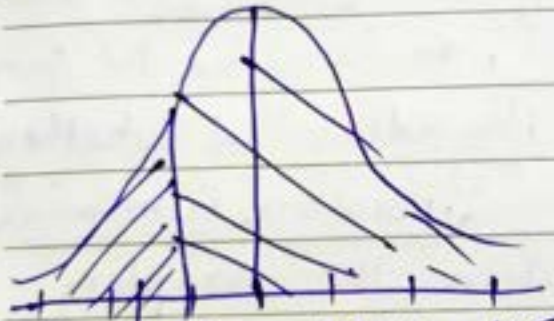
0.59871
= 59.871%

1-0.59871
= 0.4013
= 40.13%

Date / /

Q) In India the avg IQ is 100, with a std. deviation of 15. What is the % of the pop^n would you expect to have an IQ lower than 85?

Sol^n $\mu = 100, \sigma = 15$



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

$$= \frac{85 - 100}{15} \Rightarrow \frac{-15}{15}$$

$$= -1$$

on z-table using z-score $= -1.0$

we get $= 0.15866 = 15.866\%$

Q) $IQ \geq 85$

$\Rightarrow 1 - 0.15866 = 0.84134 \Rightarrow 84.134\%$

Q) $75 \geq IQ \leq 100$

$$z\text{-score} = \frac{75 - 100}{15} = \frac{-25}{15} \Rightarrow -1.666, 0.0485$$

$= 4.85\%$

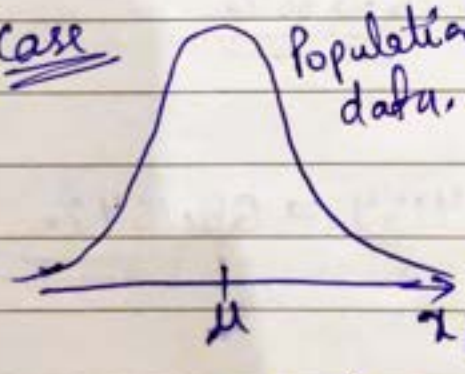
$50 - 4.85\% \Rightarrow 45.15\%$

Date / /

→ Central limit Theorem → It relies on the concept of a sampling distribution, which is the probability of a statistic for a large no. of samples taken from a population. The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the pop^n has a normal, Poisson, binomial or any other distribution, the sampling distribution of the mean will be normal.

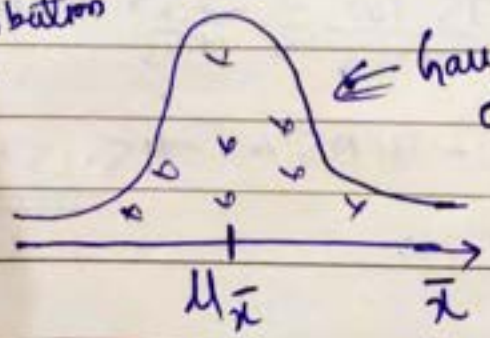
$X \overset{\sim}{} N(\mu, \sigma)$ $n \rightarrow$ sample size → any value

① Case



$S_1 = \{x_1, x_2, x_3 \dots x_n\} = \bar{x}_1$
$S_2 = \{x_{n+1}, x_3, x_4 \dots x_n\} = \bar{x}_2$
S_3 $\bar{x}_3$
S_4 $\bar{x}_4$
! !
S_m $\bar{x}_m$

Sampling distribution of the mean.



← Gaussian/Normal distribution.

Date / /

② Case $X \not\sim N(\mu, \sigma)$ $\boxed{n \geq 30}$ → sample size

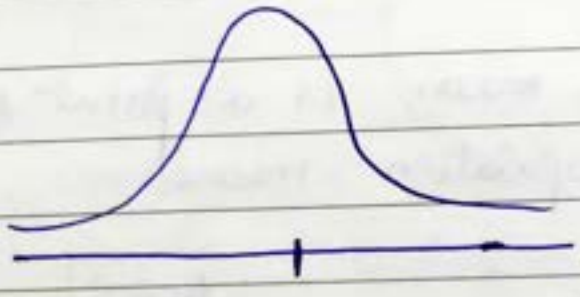
$s_1 \rightarrow \bar{x}_1$

$s_2 \rightarrow \bar{x}_2$

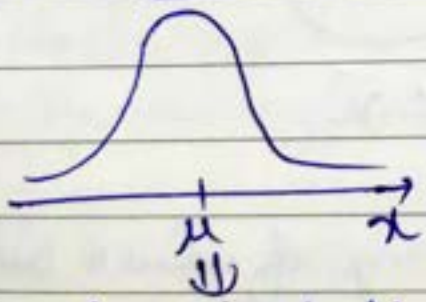
$\vdots$

$\bar{x}_m$

⇓ (CLT)



→ Normal Distribution



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

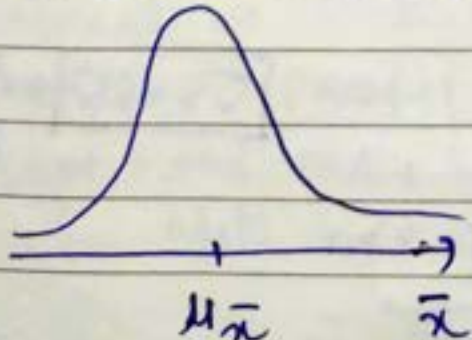
$\sigma \rightarrow$ population std. deviation

⇓

Sampling Distribution of mean

$\mu \rightarrow$ pop^n mean

$n \rightarrow$ sample size



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

Inferential Statistics

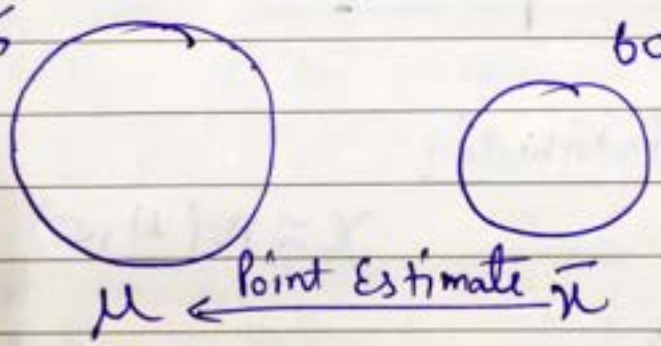
Date / /

⇒ **Estimates** → It is a specified observed numerical value used to estimate an unknown population parameter.

Types :—

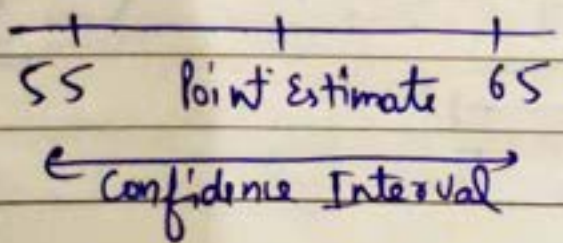
① **Point Estimate** → Single numerical value used to estimate an unknown population parameter.

eg Sample mean is a point estimate of a population mean



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

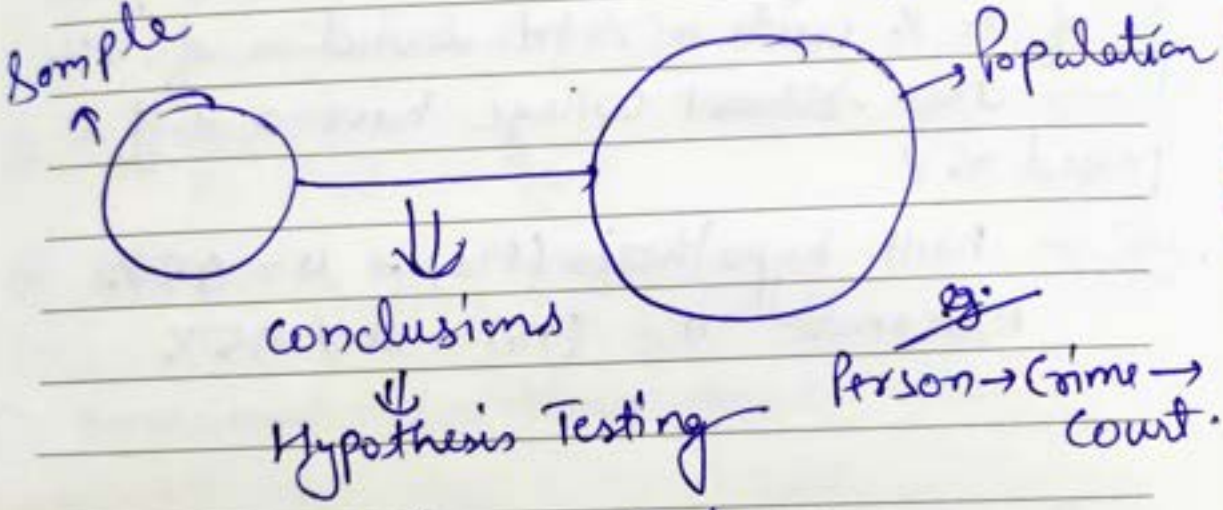
[55-65] → Sample mean



Date / /

→ Hypothesis And Hypothesis Testing Mechanism

Inferential Statistics → Conclusion or Inference.



Hypothesis Testing Mechanism :-

① Null Hypothesis (H_0) - Person is not guilt until proved
- The assumption you are beginning with

② Alternate Hypothesis (H_1) - The person is guilt
- Opposite of null hypothesis

③ Experiments → Statistical Analysis.
→ Collect proofs (DNA test, Fingerprint test)

④ Accept the Null hypothesis or Reject the Null hypothesis.

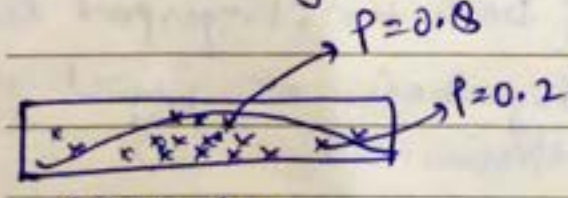
Date / /

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

Soln Null hypothesis (H_0) = $\mu = 85\%$

Alternate " (H_1) = $\mu \neq 85\%$

→ P Value – It is a no., calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis was true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.



Out of 100 touches, we touch around 20 times in side region and 80 times in middle region.

Date / /

Hypothesis Testing

Eg Coin is fair or Not [100 times tossed]

$P(H) = 0.5 \quad P(T) = 0.5$

$P(H) = 0.6 \quad P(T) = 0.4$

$P(H) = 0.7 \quad P(T) = 0.3$

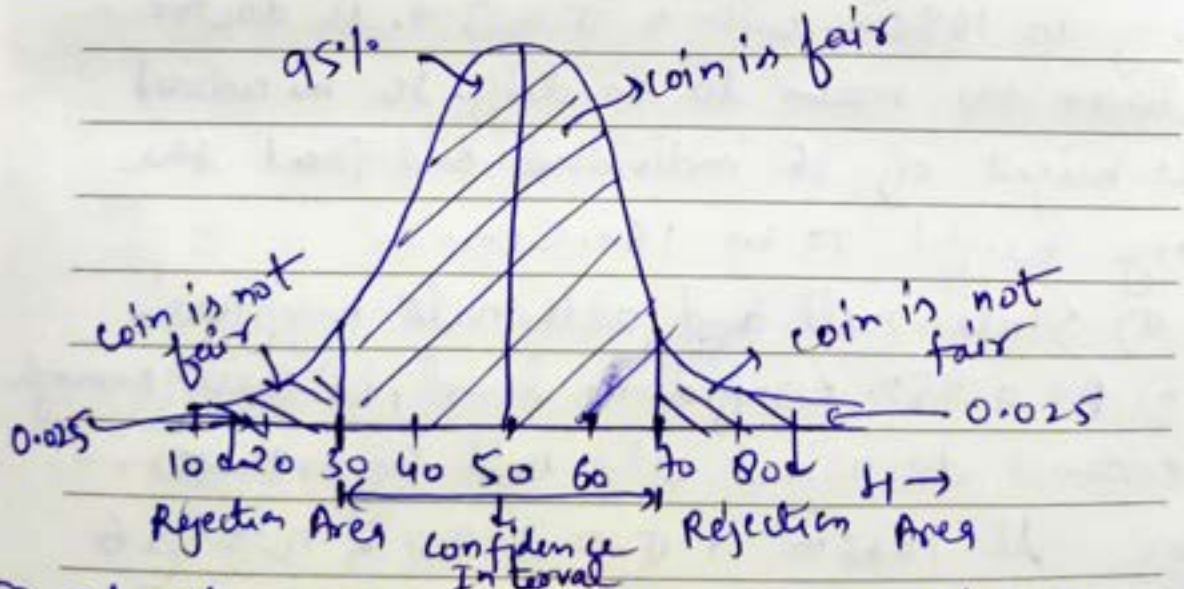
① Null hypothesis

$H_0 \rightarrow$ coin is fair

② Alternate hypothesis

$H_1 \rightarrow$ coin is not fair

③ Experiment → 100 times tossed.



④ Significance Value : $\alpha = 0.05$ (given)

$P <$ Significance Value

→ Reject the Null hypothesis

else fail to reject the null hypotheses

Date / /

→ Hypothesis Testing and Statistical Analysis

① z test } ⇒ Average ⇒ z table → z score
② t test } t table & (p value)
③ CHI SQUARE ⇒ Categorical data
④ ANNOVA ⇒ Variance

z test → i) Population St.D. Known. ii) $n \geq 30$

eg 1 The avg heights of all residents in a city is 168cm. with a $\sigma = 3.9$. A doctor believes the mean to be diff. He measured the height of 36 individual and found the avg height to be 169.5cm.

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

Ans $\mu = 168$ cm , $\sigma = 3.9$, $n = 36$

$\bar{X} = 169.5$ cm , CI = 0.95

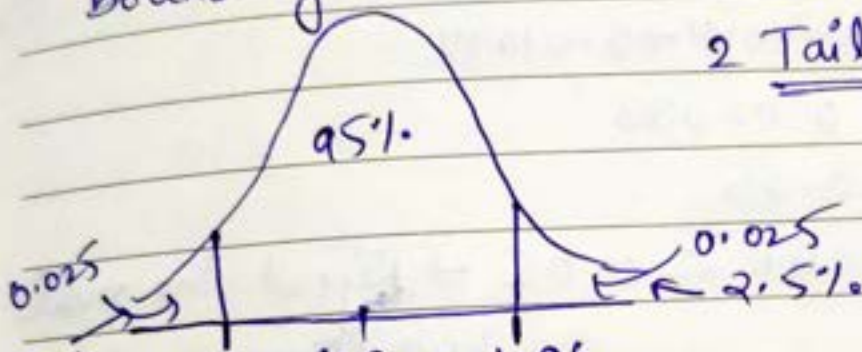
$\alpha = 1 - CI = 1 - 0.95 = 0.05$

(α : significance value)

Date / /

a) Null hypothesis $H_0 = \mu = 168$ cm
Alternate " $H_1 = \mu \neq 168$ cm.

b) Based on CI we will draw Decision boundary.



$1 - 0.025 = 0.9750 \Rightarrow$ z-score (from table)

$\Downarrow$

Area = +1.96.

if z is less than -1.96 or greater than +1.96, Reject the Null hypothesis.

* z-test

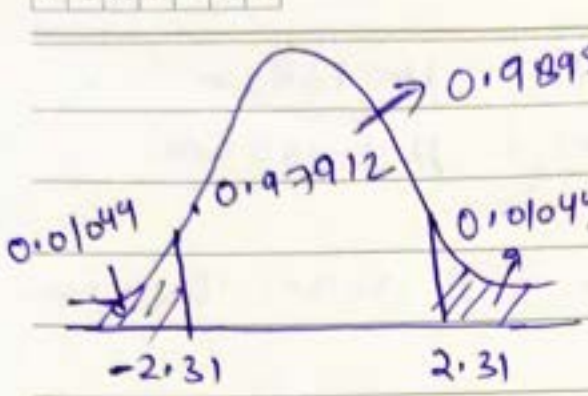
$$Z_d = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{3.9/\sqrt{36}}$$

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

$2.31 > 1.96$, Reject the null hypothesis.

$P < 0.05$

Date / /



P value = 0.01044 + 0.01044
= 0.02088

if $P < 0.05$

$0.02088 < 0.05 \Rightarrow$ Reject the null hypothesis

* Final conclusion is that the Avg. height $\neq 168$ cm.

- Based on doctor sample data the avg height of seems to increase.

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

a) State null and alternate hypothesis

Date / /

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

Sol: $\mu = 5$ yr, $\sigma = 0.50$, $n = 40$

$\bar{X} = 4.8$ yr, S.L $= 0.02$

$\alpha = 1 - CI$

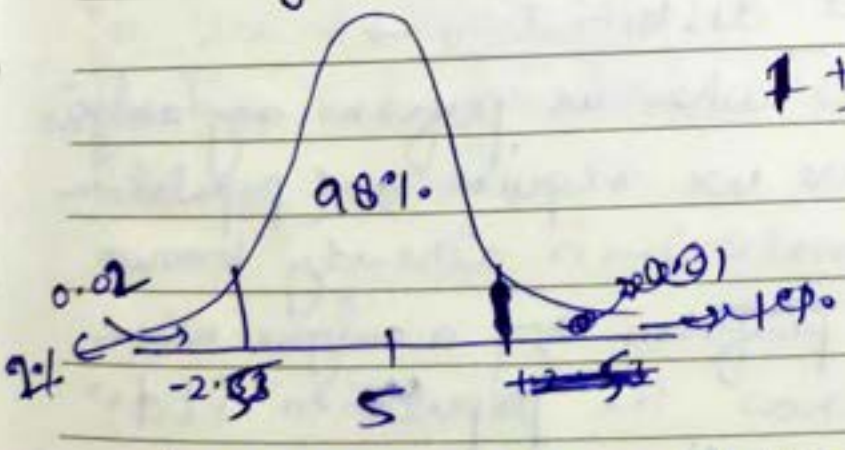
$0.02 = 1 - CI \Rightarrow CI = 1 - 0.02$

$CI = 0.98$

a) Null hypothesis $(H_0) = 5$ yr $= \mu$

Alternate " $(H_1) = \mu \neq 5$ yr (1 tail test)

b) Based on CI we should draw decision boundary.



$1 - 0.01 = 0.99 \Rightarrow$ Z-score (from table)

$\Rightarrow$ Area $= \pm 2.33$

Date / /

if $-2.33 > Z$ ~~$+2.33$~~, Reject the null hypothesis

* z test

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}} = \frac{-0.2}{0.07905}$$

$$= -2.5300$$

Area under the curve with z score

$-2.5300 \Rightarrow 0.0570$

p-value = 0.0570

Compare p value with significance value (α)

$0.0570 < 0.02 \Rightarrow$ false

* Conclusion → we accept the null hypothesis or we failed to reject the " ".

→ Student t distribution →

In z stats when we perform any analysis using z-score we require σ (population standard deviation) → is already known. Now we do perform any analysis when we don't know the population std?

⇓

Student's t distribution

t stats

Date

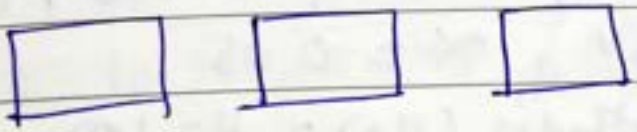
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad , \quad \boxed{t = \frac{\bar{x} - \mu}{s/\sqrt{n}}}$$

$s \rightarrow$ Sample std. deviation

Degree of freedom

$$dof = n - 1$$

eg There are 3 people in a room. and there are 3 chairs in the room to sit.



if first person going to choose a chair he will have 3 choice to him after getting the chair only 2 chairs are empty. than for second one there will be 2 choices for him and for last one he will not have any choice.

Date / /

Q In the population the avg IQ is 100. A team of researchers want to test a new medication to see the if it has either a +ve or -ve effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a Std. deviation of 20. Did the medication affect intelligence? CI = 95%., $\alpha = 0.05$

Ans $\mu = 100$, $n = 30$, $\bar{x} = 140$, $S = 20$

CI = 95%., $\alpha = 0.05$

(1) Null hypothesis $(H_0) = \mu = 100$

Alternate " $(H_1) = \mu \neq 100$ (2 tail test)

(2) $\alpha = 0.05$

(3) Degree of freedom

$dof = n - 1 = 30 - 1$

$= 29$

(4) Decision Rule

Date / /

if t test is less than -2.045 or greater than 2.045, reject the null hypothesis

⑤ Calculate Test Statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{140-100}{20/\sqrt{30}} = \frac{40}{3.65}$$

$$= 10.96$$

$$t = 10.96$$

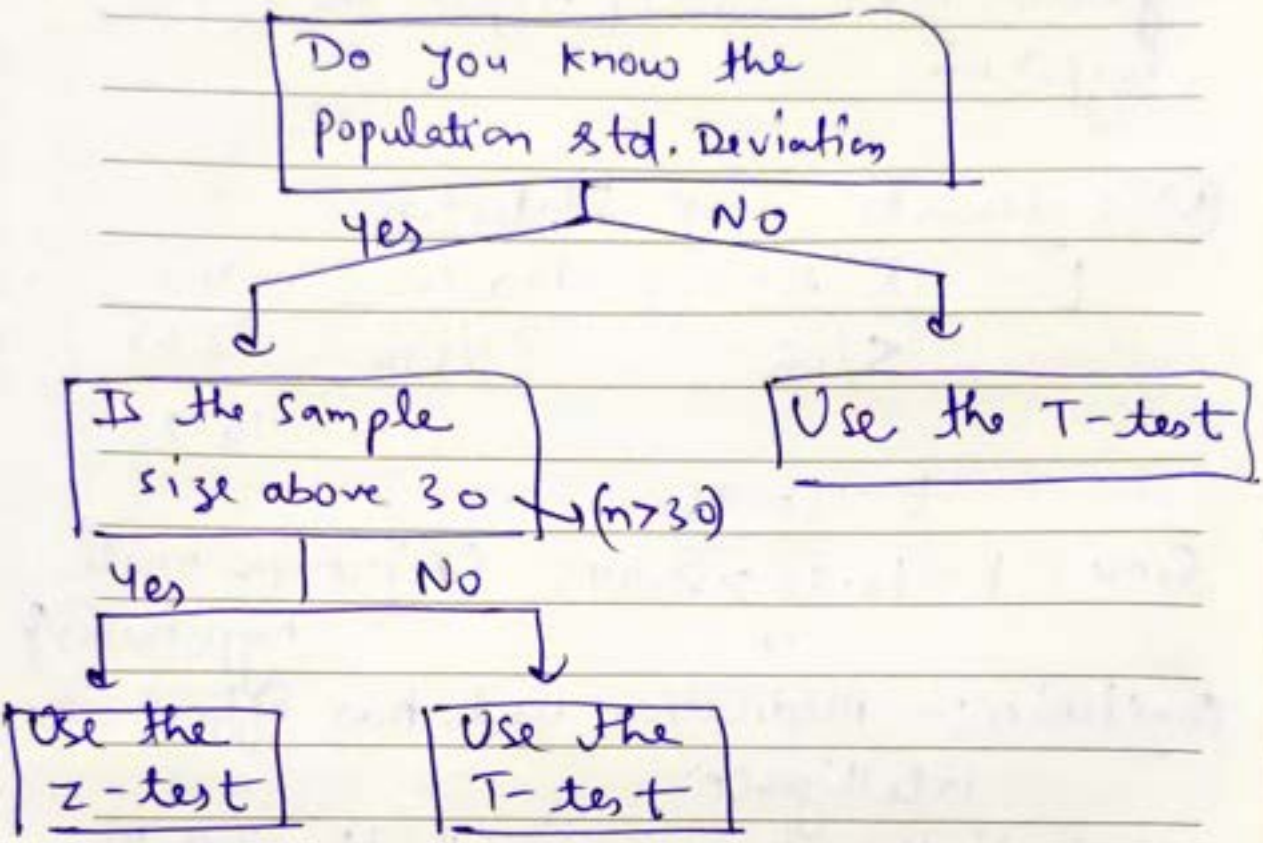
Since, $t = 10.96 > 2.045$ {Reject the null hypothesis}

Conclusion :- Medication used has affect the intelligence

→ Medication has increased the Intelligence

Date / /

→ When to use T-test Vs Z-test



→ Type 1 & Type 2 Error :-

Reality → Null hypothesis is True or Null hypothesis is False

Decision → Null hypothesis is True or Null hypothesis is false

Date

Outcome 1: We reject the Null hypothesis when in reality it is false
→ Good.

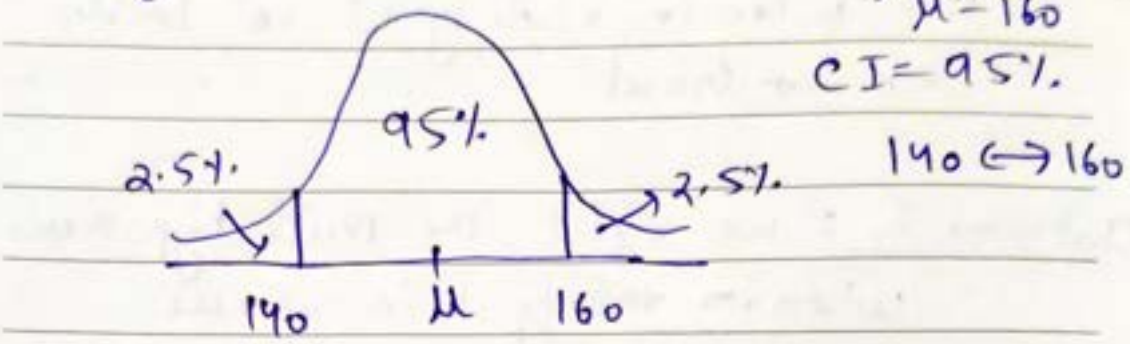
Outcome 2: We reject the Null hypothesis when in reality it is True
→ Type 1 Error.

Outcome 3: We accept the Null hypothesis, when in reality it is false.
→ Type 2 Error

Outcome 4: We accept the Null hypothesis, when in reality it is True
→ Good.

Date / /

→ Confidence Intervals and Margin of Error



Point Estimate → A value of any statistical that estimates the value of an unknown population parameter is called Point Estimate.

$\bar{x} \longrightarrow \mu$

$\bar{x} = 2.95$ $\quad\quad$ $\mu = 3.$

Confidence Interval → We construct a confidence interval to help estimate what the actual value of the unknown population mean is

Point Estimate ± Margin of Error.

z test $\quad \boxed{\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}} \quad \alpha = 0.05$

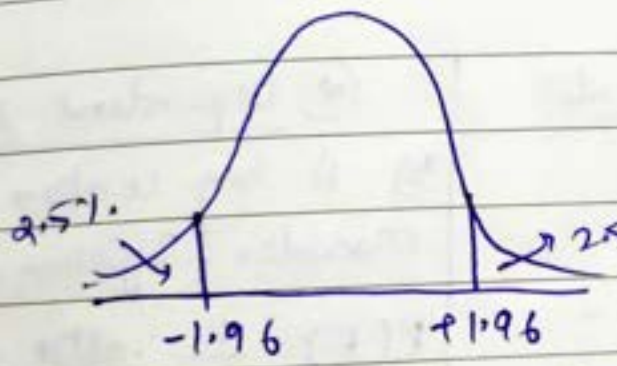
$Z_{0.05/2} \Rightarrow Z_{0.025}.$

Date

Q On the verbal section of CAT exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% CI about the mean?

Ans $\bar{x} = 520$, $\sigma = 100$, $n = 25$, $CI = 0.95$

$\alpha = 0.05$



$1 - 0.25 = 0.9750$

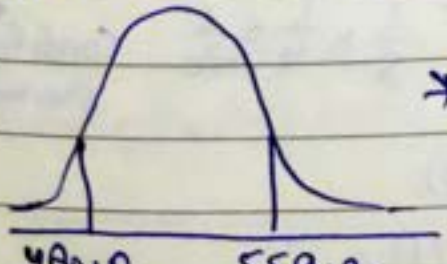
from z table

$z = 1.96$

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

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

Upper CI $= 520 + 1.96 \times \frac{100}{\sqrt{25}} = 559.2$



* Conclusion → I am 95% confident that the mean CAT exam lies b/w 480.8 & 559.2.

Date / /

⇒ Baye's Theorem → Bayesian statistics is an approach to data analysis and parameter estimation based on Baye's theorem.

Baye's theorem

Possibility → Independent Events
→ Dependent Events.

① Independent Events

Eg · Rolling a dice
{1, 2, 3, 4, 5, 6}
P(1) = 1/6, P(2) = 1/6 - -

eg Tossing a coin
P(H) = 0.5, P(T) = 0.5

→ In this all the events are independent of other

② Dependent Event

eg A bag contain 5 marbles (3 yellow 2 Red)

P(R) = 2/5 . after getting out a red marble now P(Y) = 3/4 as a 1 red marble is taken out.

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

= 2/5 × 3/4 = 6/20

Conditional Probability.

P(A and B) = P(B and A)

P(A) × P(B/A) = P(B) × P(A/B)

Date

$$P\left(\frac{B}{A}\right) = \frac{P(B) \times P(A/B)}{P(A)}$$ ⇒ Baye's theorm

$$P(A/B) = \frac{P(A) \times P(B/A)}{P(B)}$$ ⇒ Baye's theorm

A, B → events.

P(A/B) → Probability of A given B is true.

P(B/A) → Probability of B given A is true.

P(A), P(B) → Independent Probabilities of A and B.

eg DATASET

Size of house	No. of Rooms (Independent feature)	Location (Independent feature)	Price (O/P dependent feature)
x_1	x_2	x_3	y

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

⇓

Baye's theorom

Date / /

→ CHI SQUARE TEST

The chi square test for Goodness of fit test claims about population proportions.
- It is a non-parametric test that is performed on categorical data.
[ordinal and nominal data].

① There is a population of male who likes diff. colour of bikes.

	Theory	Sample	* Goodness of fit test
Yellow Bike	1/3	22	
Red Bike	1/3	17	
Orange Bike	1/3	59	
	↓	↳ observed categorical distribution	

Theory categorical distribution

Goodness of fit test

② In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of people are left handed.

Ans

$\frac{12^3}{100} \times \frac{75^3}{1} = 9$

	Obs.	Exp.
Left Hand	11	9
Right Hand	64	66

Date

- CHI SQUARE for Goodness of fit →

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

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

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

< 50	50-75	>75
140	160	200

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

Ans 2020 Expected

< 50	50-75	>75
0.2×500	0.3×500	0.5×500
100	150	250

① Null hypothesis H_0: the data meets the expectation
Alternate " (H_1): the data doesn't meet expectation.

② $\alpha = 0.05$, $CI = 95\%$.

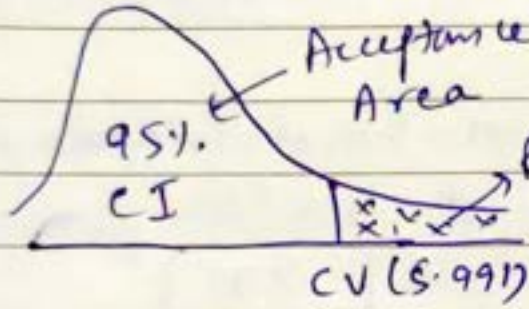
Date / /

(3) Degree of freedom

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

$K \rightarrow$ no of categories (< 50, $50-75$, > 75)

(4) Decision Boundary



using chi square table CV will be 5.991

If X^2 is greater than 5.99, Reject H_0 else we fail to reject the null hypothesis

(5) Calculate Chi square test statistics.

$$X^2 = \frac{\Sigma (O-E)^2}{E}$$

2020 sample

<50	50-75	>75
140	160	200

2020 expected

<50	50-75	>75
100	150	250

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

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

Date / /

$= 16 + 0.66 + 10$

$= 26.66$

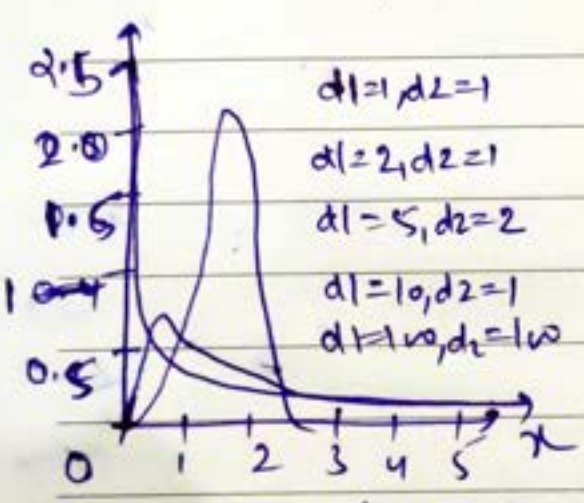
$X^2 = 26.66$

~~If~~ X^2 Hence, $26.66 > 5.99$, Reject H_0.

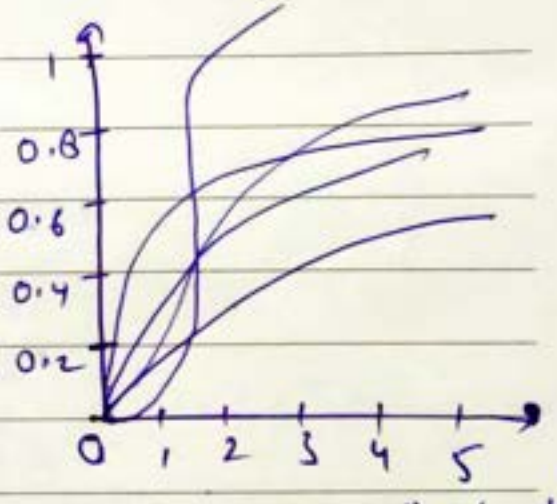
*Conclusion → The weight of 2020 population are diff. than those expected in the 2020 population.

Date / /

→ **F distribution** :- In prob. or statistics, the F-distribution or F-ratio, also k/a Snedecor's F distribution or the Fisher-Snedecor distribution (after Ronald Fisher and George W. Snedecor) is a continuous prob. distribution that arises frequently as the null distribution of a test statistic, most notably in the analysis of variance (ANOVA) and other f-tests.



Probability density fun^n



Cummulative distribution func^n.

→ Parameters → $d_1, d_2 > 0$ → degree of freedom

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

Date / /

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

↓ Beta function

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

→ F distribution with d_1 and d_2 degree of freedom in the distribution of

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

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

F- test [Variance Ratio Test] d_1 & d_2 → Respective degree of freedom.

Date / /

→ f test [Variance Ratio Test] :-

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

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

→ can we consider based on the data worker B is more stable and efficient

$\alpha = 0.05$

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

Alternate $H_1 \quad \sigma_1^2 \neq \sigma_2^2$

② Calculation of Variance

$$S_1^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Date / /

Wooker A			Worker B		
X_1	$\bar{X}_1$	$(X_1-\bar{X}_1)^2$	X_2	$\bar{X}_2$	$(X_2-\bar{X}_2)^2$
40	37	9	39	37	4
30	37	49	38	37	1
38	37	1	41	37	16
41	37	16	33	37	16
38	37	1	32	37	25
35	37	4	39	37	4
		$\Sigma(X_1-\bar{X}_1)^2$	40	37	9
$\bar{X}_1=37$		$=80$	34	37	9
			$\bar{X}_2=37$		$\Sigma(X_2-\bar{X}_2)^2$
					$=84$

$$S_1^2 = \frac{80}{n-1} = \frac{80}{6-1} = \frac{80}{5} = 13$$

$$S_2^2 = \frac{84}{n-1} = \frac{84}{8-1} = \frac{84}{7} = 12$$

→ Calculation of Variance Ratio Test (F test)

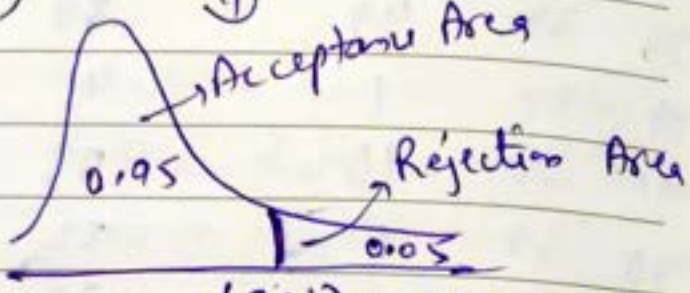
$$F = \frac{S_1^2}{S_2^2} = \frac{13}{12} = 1.33$$

Date / /

* Decision Rule

$df_1 = 6-1 = 5$
$df_2 = 8-1 = 7$
$\alpha = 0.05$

} using F table



Decision :—

F test is greater than 3.9715, Reject the null hypothesis
1.33 < 3.9715, we fail to Reject the null hypothesis.

Conclusion :—

→ Worker B is not more stable or efficient when compared to worker A.

Date / /

⇒ Analysis of Variance (ANOVA)

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

ANOVA

① Factors (Variable)
② Levels

Medicine (factor)

eg [Dosage] 5mg 10mg 15mg → Levels

→ Mode of Payments (factor)

GPay PhonPe IMPS NEFT

Levels.

⇒ Analysis of Variance (ANOVA) →

→ Assumptions of ANOVA :-

① Normality of Sampling Distribution of Mean
- The distribution of sample mean is normally distributed.

② Absence of Outliers
- Outline score need to be removed from

Date / /

the datasets

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 Samples are independent and random

→ Types of ANOVA :-

(3 types)

① One way ANOVA - One factor with atleast 2 levels, these levels are independent

eg Doctors wants to test a new medication to decrease headache. They split the ~~population~~ participants in 3 conditions [10mg, 20mg, 30mg]. Doctors ask the participants to rate the headache [1-10]

Meditation → factor

10 mg	20 mg	30 mg
5	7	2
3	4	6
-	-	-
-	-	-

Date / /

(2) Repeated Measures ANOVA → One factor with atleast 2 levels, levels are dependent.

eg

Running → Factor

Day 1	Day 2	Day 3 → levels
8 km	5 km	4 km
7 km	4 km	9 km

(3) Factorial ANOVA - Two or more factors, each of which with atleast 2 levels, these levels can be independent or dependent.

Eg

Running → Factor

Levels ↓		Day 1	Day 2	Day 3 → levels
Gender ↓ Factor	Male	8	5	4
		9	4	3
	Female	2	4	6
		7	8	3

Date

→ Hypothesis Testing in ANOVA (Partitioning of Variance in the ANOVA)

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

Alternate " H_1 : Atleast one of the sample mean is not equal.

$$\boxed{\mu_1 \neq \mu_2 \neq \mu_3 = \dots \neq \mu_k}$$

not applicable

Test Statistics

$$F \text{ test} = \frac{\text{Variance b/w sample}}{\text{Variance within sample}}$$

σ^2 variance b/w sample

$H_0 \Rightarrow \bar{X}_1 = \bar{X}_2 = \bar{X}_3$

$H_1 \rightarrow$ Atleast one sample mean is not equal

	X_1	X_2	X_3
Variance within sample	1	6	5
	2	7	6
	4	3	3
	5	2	2
	3	1	4
	$\bar{X}_1 = \frac{15}{5} = 3$	$\bar{X}_2 = 19/5$	$\bar{X}_3 = 4$

Date / /

One Way ANOVA

One factor with atleast 2 levels, levels are independent.

Q. Doctor wants to test a new medicine which reduces headache. They split the participants into 3 condⁿ [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache b/w [1 - 10]. Are there any differences b/w the 3 conditions using $\alpha = 0.05$?

Ans

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 and Alternate Hypotheses :-

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

H_1 : not all μ are equal.

Date / /

② Significance $\alpha = 0.05$, $CI = 0.95$

③ Calculate degree of freedom.

$N = 21$, $a = 3$, $n = 7$

$N \rightarrow$ total no. of participants ($9 \times 3 = 27$)

$n \rightarrow$ participants in a single level [eg 15mg]

$a \rightarrow$ factors (30mg, 15mg, 45mg) = 3.

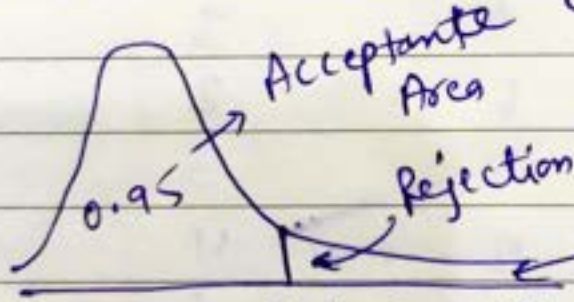
$df_{b/w} = a - 1 = 3 - 1 = 2$

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

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

df_1, df_2
$(2, 18)$
⇓
f-table
$\alpha = 0.05$
⇓
Critical Value

④ Decision Boundary



CV = 3.5546 (using f-table)

Decision Rule

If F is greater than 3.5546, Reject the Null hypotheses

Date / /

⑧ Calculate F test statistic

	SS	df	MS	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.96	20		

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

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

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

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

$SS_{b/w} = \frac{57^2+47^2+21^2}{7} - \left[\frac{57^2+47^2+21^2}{21}\right]$

$= 98.67$

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

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

$= 853$

$SS_{within} = 853 - \left[\frac{57^2+47^2+21^2}{7}\right]$

$= 10.29$

Date

$$f\ test = \frac{MS\ b/w}{MS\ within} = \frac{98.67/2}{10.29/2} = \frac{49.34}{0.54}$$

$$F = 86.56$$

If f is greater than 3.5546, Reject the null hypothesis

$86.56 > 3.5546$, Reject the H_0