STATS

Lecture : 1
8. Sep.

Variable:

- Any characteristic whose values can change from one element to another element
- Any characteristic that can assume different value

Data:

Set of information about variables.

Variable

- **Qualitative** → Those variables which can't be numerically measured
- **Quantitative** → Those variables which can be numerically measured

Data

- **Qualitative** → Data about qualitative variables. e.g Blood Group
- **Quantitative** → Data about quantitative variables e.g CGPA, Weight

Quantitative Variables

- **Discrete** → Countable variable (How many?)
- **Continuous** → Measureable Variable (How much?)

(N) Population: → (Statistical Population)
Set of objects under study.

(n) Sample:
Subset of population is called sample.

Sampling:
Process of sample selection

Sampling

Random / Probability	Non-Random / Non-Probability
Each sample has chance to be selected.	Sample is selected on personal judgement
Can be generalization for the population	Each sample does not have the chance to be selected.
A sample selected in random sampling.	Result of non-random samples is only valid for themselves.

Population

Finite → Known number of elements

Infinite → Unknown number of elements

Finite → Real / Hypothetical
Infinite → Real / Hypothetical.

Real
Elements exist physically.

Hypothetical
Elements does not exist physically.

Target

STATS

Lecture: 2
13 Sep

Target Population:

→ Under the study / whole

Sampled Population:

→ Seleded sample / access
→ from which we can draw the conclusion.

Types of Random Sampling:

→ Simple random sampling (Homogenous).
→ Systematic sampling (Sequence, Otem)
→ Stratified sampling (Hetrogenous).
→ Cluster sampling.

Simple Random Sampling:

→ Seleded sample from the population using random method.
→ Best for homogenous / same population.

Stratified Sampling:

→ Breaking hetrogenous population into homogenous population

Systematic Sampling:

→ Select interval

$N = nk$

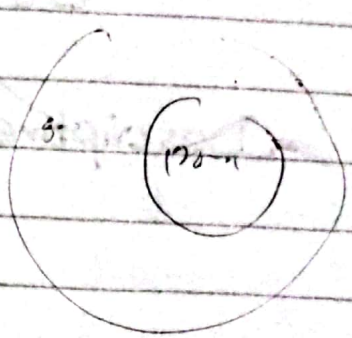
N → Population, n → Sample, k → interval

$$k = \frac{N}{n}$$

$$k = \frac{25}{3}$$

$$k = 8.33$$

$$\boxed{k = 8}$$



→ List of unit is known as sampling frame

Cluster Sampling:

→ When we do not have the data / info about the samples, we seled the samples as clusters.

→ Natural selection — Random selection

Statistic:

→ Any numerical value that is calculated from sample

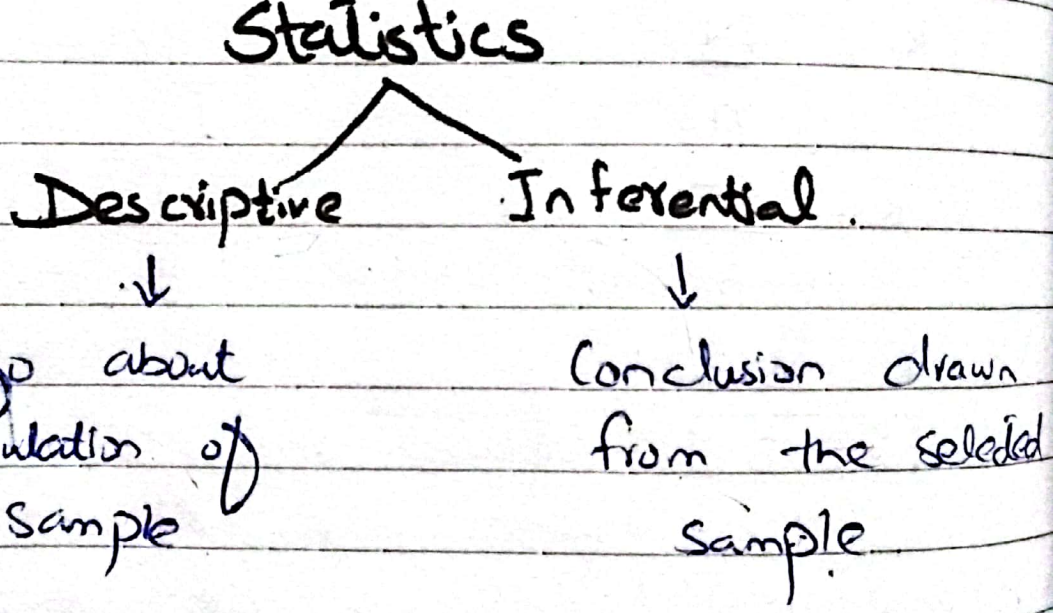
Parameter:

→ Any numerical value that is

calculated from population.

Statistics:

→ Statistics is a science of conducting studies to colled, organize, summarize, analyze and draw conclusion from data



STATS

ASSIGNMENT 1

Lecture 3
15. Sep

Data visualization

What → Introduction
Why → Objects
How → Methods
Which → Best
Application? Dataset.
Report → Summary

Presentation of Data:

Data presentation is the method by which people summarize, organize and communicate information using a variety of tool, such as diagrams, distribution charts, histogram and graphs.

Objective:

→ It condense the raw data into a form suitable for statistical analysis.
→ It remove the complexities and highlights the feature of data.
→ It facilities comparison and drawing conclusion from data.
→ It helps in statistical analysis by separating elements of data set into homogeneous groups.

Method:

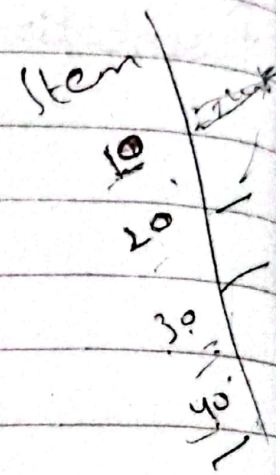
Textual Method:

→ Rearrangement from lowest to highest

→ Stem and leaf plot.

11.3, 9.6, 10.4, 7.5, 8.3, 10.5, 10.0, 9.3, 8.1, 7.7, 7.5, 8.4, 6.3, 8.8.

Stem	Leaf
0	
1	
2	
3	
4	
5	
6	.3
7	.5 .7 .5
8	.3 .1 .4 .8
9	.6 .3
10	.4 .5 .0
11	.3



Tabular Method:

Presenting data in the form of table.

Types of Tabular Method:

Frequency Distribution:

A frequency distribution is a table that distributes a data set into a suitable number of categories (classes). Data presentation in the form of frequency table called grouped data.

$$\begin{cases} K = 1 + 3.22 \log(N) \\ W = \dfrac{\max - \min}{K} \end{cases}$$

2, 4, 8, 4, 8, 1, 2, 32, 12, 6, 5, 7, 5, 5, 3, 4, 24, 19, 4, 14

$$K = 1 + 3.22 \log(20)$$

$$K = 5.18$$

$$\boxed{K = 5}$$

$$W = \frac{\max - \min}{K}$$

$$W = \frac{32 - 1}{5}$$

$$W = \frac{31}{5} \Rightarrow W = 6.2 \Rightarrow \boxed{W = 6}$$

Classes	Tallies	frequency.
1 – 6	||||| ||||| ||	12
7 – 12	||||	4
13 – 18			1
19 – 24	||	2
25 – 30	-	0
31 – 36			1

||||| ||||| ||

||||| ||||| ||

STATS

Lecture 4
20 Sep.

Presentation of Data:

→ Stem and leaf
→ Frequency Distribution (F.D)] ← Tabular Method
  - Commulative F.D
  - Percentage F.D

→ Graphs
  - Histogram
  - Frequency Polygon
  - Frequency Curve
  - Time series plot

Frequency Distribution:

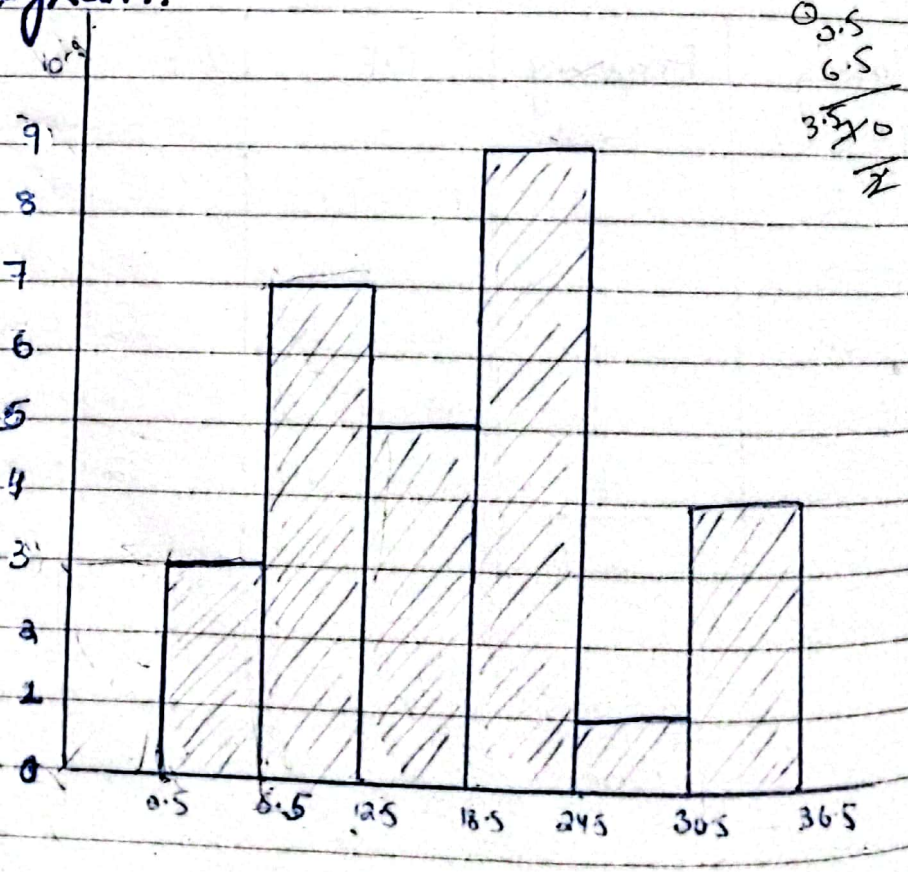
Classes	Frequency	C.F	%.F
1-6	3	3	3/29 x100
7-12	7	3+7=10	7/29 x100
13-18	5	10+5=15	5/29 x100
19-24	9	15+9=24	9/29 x100
25-30	1	24+1=25	1/29 x100
31-36	4	25+4=29	4/29 x100
	29		

Graphs:

Histogram → { F.D with class boundaries
Plot frequency against C.B

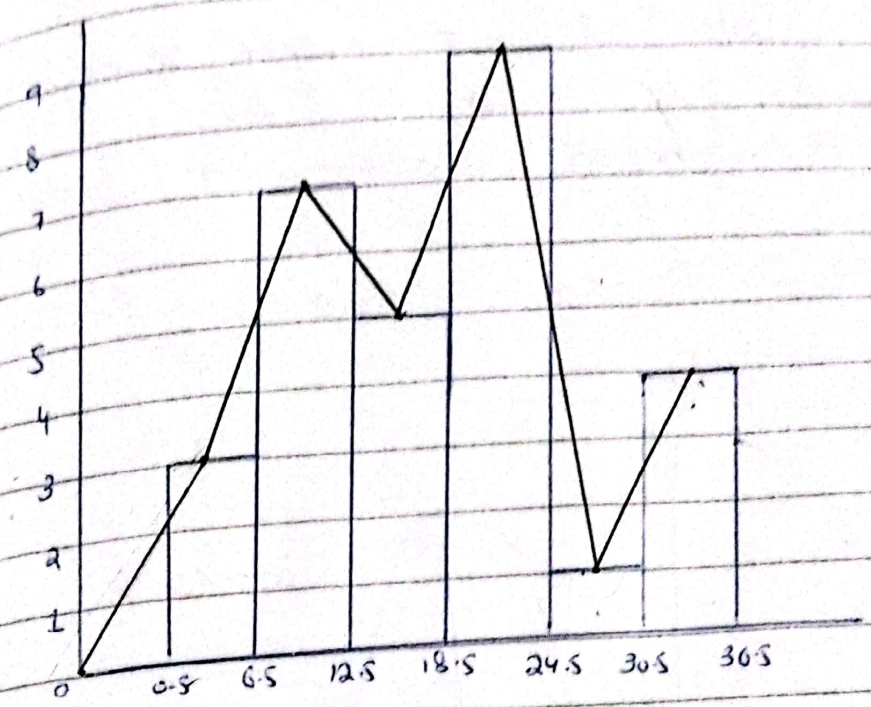
Classes	Frequency	C.B	X (Mid Point)
1 – 6	3	0.5 – 6.5	3.5
7 - 12	7	6.5 - 12.5	9.5
13 - 18	5	12.5 - 18.5	15.5
19 - 24	9	18.5 – 24.5	21.5
25 - 30	1	24.5 – 30.5	27.5
31 - 36	4	30.5 – 36.5	33.5

Histogram:



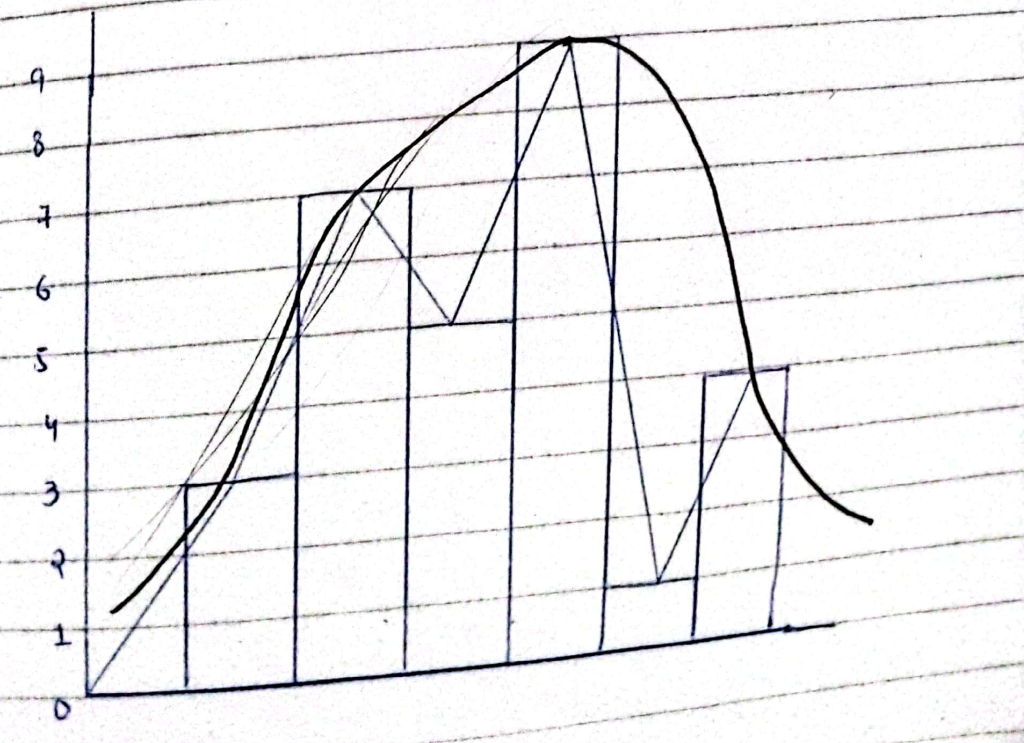
Frequency Polygon:

↓ Histogram or midpoint (x)



Frequency Curve:

→ frequency polygon or histogram



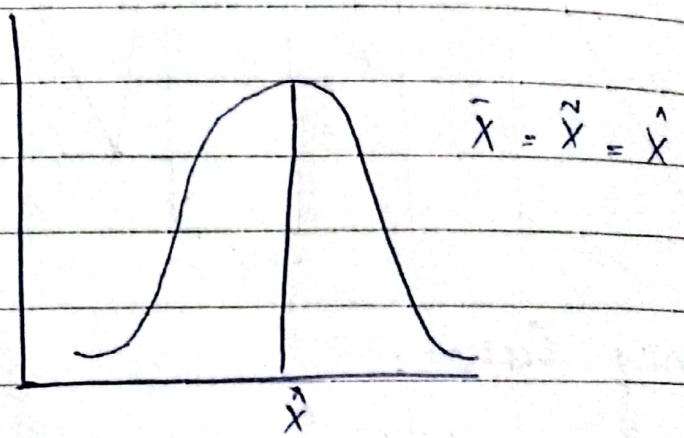
Skewed And Symmetric Distribution

Symmetric Data:

Data set are evenly spread around the center.

Bell shaped curve

Mean = Media = Modo



Skewed Data:

Measure of Central Tendency

Lecture 6
4. Oct

Arithmetic Mean:-

Arithmetic mean is a value which is obtained by dividing the sum of all values by their number.

[Population mean

$\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$ → parameter

Greek Latin

Population size = N

Sample size = n

[Sample mean

$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n}$ → statistic/Variable

$\bar{x}, \bar{y}, s^2, p$

Geometric Mean:-

Geometric mean is a value which is obtained by N^{th} under-root of product of all N values.

→ $G.M = (x_1 \cdot x_2 \cdot x_3 \cdots x_N)^{1/N}$

OR

→ $G.M = \left[\prod_{i=1}^{N} x_i\right]^{1/N}$

OR

→ $G.M = \text{Antilog}\left[\frac{\sum_{i=1}^{N} \log x_i}{N}\right]$

OR

→ $G.M = \text{Antilog}\left[\frac{\log x_1 + \log x_2 + \dots \log x_N}{N}\right]$

Harmonic Mean:-

Harmonic mean is the reciprocal of arithmetic mean of reciprocal of values

→ $$H.M = \frac{N}{\frac{1}{x1} + \frac{1}{x2} + \dots + \frac{1}{xN}}$$

OR

→ $$H.M = \frac{N}{\sum_{i=1}^{N} \left(\frac{1}{xi}\right)}$$

→ Gives more importance to smaller value and less importance to greater value

Example: 70, 37, 55, 63, 59, 68, 56, 54

$$\text{Mean} = \frac{70+37+55+63+59+68+56+54}{8}$$

$$= \frac{462}{8}$$

$$\text{Mean} = 57.75$$

$$G.M = (70 \times 37 \times 55 \times 63 \times 59 \times 68 \times 56 \times 54)^{1/8}$$

$$G.M = (1088793988128 00)^{1/8}$$

$$G.M =$$

Mode:- Most repeated value

Median:- Mid value sorting the dataset into ascending order

Measure of Dispersion :-

→ Absolute Measure

→ Range = $X_{max} - X_{min}$

→ Mean Deviation (M.D)

↓

Population :-

$$M.D = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

∴ μ = Mean

Sample :-

$$M.D = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

∴ $|x_i - \bar{x}|$ = [-ive to +ive]

→ Variance

Population :-

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

Sample :-

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

→ Standard Deviation

$$S.D = \sqrt{\text{Variance}}$$

Population :-

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

Sample :-

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

Relative Measure:-

Coefficient of Range = $\frac{X_{max} - X_{min}}{X_{max} + X_{min}}$

Coefficient of Mean Deviation = $\frac{\text{Mean Deviation}}{\text{Mean}} \times 100$

Coefficient of Variance = $\frac{\text{Standard Deviation}}{\text{Mean}} \times 100$

Example:-

Mean data shows the marks of MTH605. Find all measures of dispersion.

X	$X - \mu$	$\lvert X - \mu \rvert$	$X^2 = \lvert X - \mu \rvert^2$
70	-3.83	3.83	14.66
82	8.17	8.17	66.74
70	-3.83	3.83	14.66
71	-2.83	2.83	8.008
75	1.17	1.17	1.36
75	1.17	1.17	1.36
		21	106.78

$\mu = \frac{\Sigma X}{N} = \frac{70+82+70+71+75+75}{6} \Rightarrow \frac{443}{6} \Rightarrow 73.83$

Range = $X_{max} - X_{min}$

$= 82 - 70 \Rightarrow 12$

→ Absolute Measure:

Mean Deviation (Population) :-

$$M.D = \frac{|x_i - \mu|}{N}$$

$$M.D = \frac{21}{6}$$

$$\boxed{M.D = 3.5}$$

Variance (Population) :-

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

$$\sigma^2 = \frac{106.78}{6}$$

$$\boxed{\sigma^2 = 17.79}$$

Standard Deviation

$$S.D = \sqrt{\text{Variance}}$$

$$S.D = \sqrt{17.79}$$

$$\boxed{S.D = 4.217}$$

→ Relative Measure:

$$\text{Coeff Range} = \frac{X_{max} - X_{min}}{X_{max} + X_{min}}$$

$$\text{Coeff Range} = \frac{82 - 70}{82 + 70}$$

$$\text{Coeff-Range} = \frac{12}{152}$$

$$\boxed{\text{Coeff-Range} = 0.078}$$

Coeff M.D = $\frac{\text{M.D}}{\text{Mean} \rightarrow (\mu)} \times 100$

$= \frac{3.5}{73.83} \times 100$

$= 0.047 \times 100$

$\therefore$ Coeff M.D = 4.74

Coeff Variance = $\frac{\text{S.D}}{\text{Mean} \rightarrow (\mu)} \times 100$

$= \frac{4.217}{73.83} \times 100$

$= 0.057 \times 100$

$\therefore$ Coeff Variance = 5.71

Absolute Measure of Skewness:

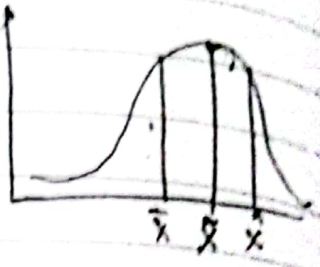
Lecture 20.02

Symmetric:-

Mean = Median = Mode

Mean - Mode = 0

Mean - Median = 0

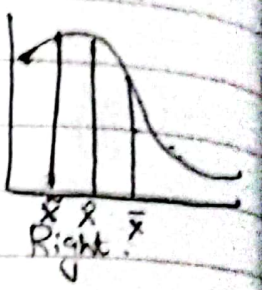


Right Skewed:- +ively skewed

Mean > Median > Mode

Mean - Mode > 0

Mean - Median > 0

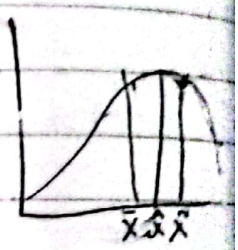


Left Skewed:- -ively skewed

Mean < Median < Mode

Mean - Mode < 0

Mean - Median < 0



Pearson Coefficient of Skewness: (Relative)

$$Skp = \frac{\text{Mean} - \text{Mode}}{S.D} \quad OR \quad Skp = \frac{3(\text{Mean} - \text{Mode})}{S.D}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{or} \quad \frac{\text{Mean} - \text{Median}}{S.D}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\text{Coeff of Skewness} = \gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

Moments:

rth moment about mean :-

$$\mu_r = \frac{\sum_{i=1}^{N} (X - \mu)^r}{N}$$

$$\mu_1 = \frac{\sum_{i=1}^{N} (X_i - \mu)^1}{N} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} = \delta^2$$

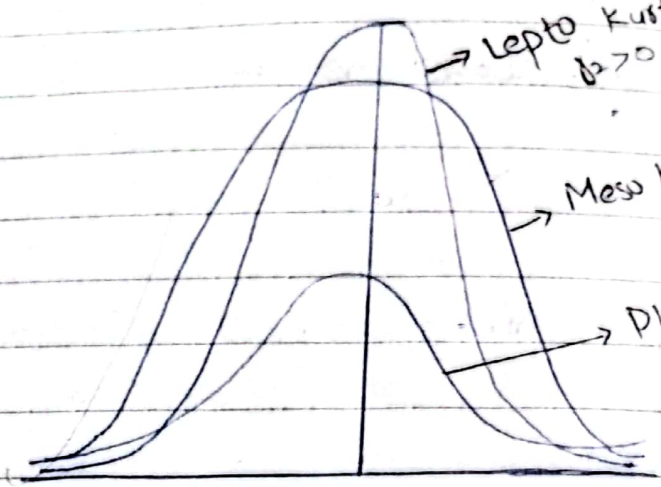
$$\mu_3 = \frac{\sum_{i=1}^{N} (X_i - \mu)^3}{N} = \text{Not specified} = \text{Skewness}$$

$$\mu_4 = \frac{\sum_{i=1}^{N} (X_i - \mu)^4}{N} = \text{Not specified} = \text{Kurtosis}$$

∴ Parameter is a constant value.

Mean is a

Measure of Kurtosis :-



Coefficient of Kurtosis:

$$\gamma_2 = \beta_2 - 3$$

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

$\gamma_2 = 2 - 3 = -1$

γ_2 $\because \mu_2 = \sigma^2$

$$\gamma_2 = \frac{\mu_4}{(\sigma^2)^2} - 3$$

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

→ What will be the degree of skewness in case of Kurtosis

Degree of γ_1 will be zero

γ_1 will be symmetic

Lecture: 9
27 Oct

Probability:

→ Uncertainty & incomplete information
OR [Caused by variation
→ Probability theory is mathematical tool for uncertainty and randomness.
→ Numerical measure of uncertainty.

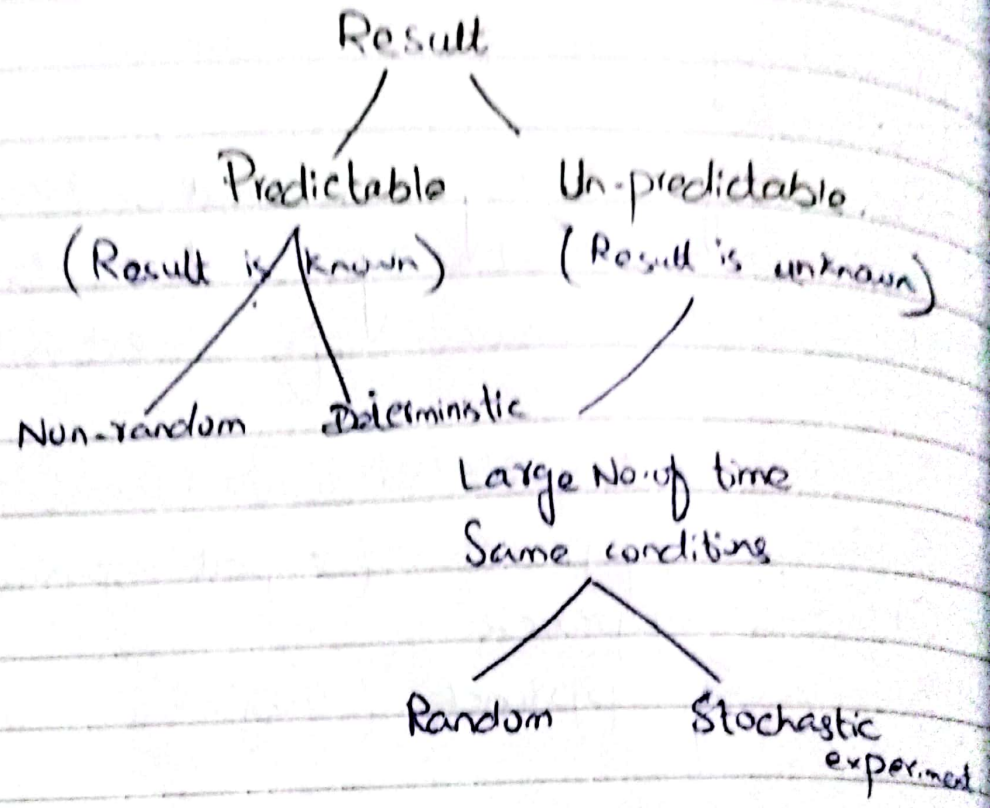
→ Two Approaches to Measure Uncertainty
→ Subjective approach (expert opinion)
→ Objective approach
→ Bayesian approach (Subjective + Objective).

→ Subjective Approaches-
Probability of occurance of any event on the basis of personal judgement.

→ Objective Approaches- (Mathematical Model).
On the basis of previous data, we find the probability for any event to occur.

Experiment:

→ Activity that produces some result
→ It has two outcomes



→ **Sample Space:**
Set of all possible outcomes.

→ **Event:**
Subset of sample space.

$S = \{ win, loss, Draw \}$

$2^n = 2^3 = 2 \times 2 \times 2 = 8$.

$A_1 = \{\ \}$	$A_5 = \{win, loss\}$
$A_2 = \{win\}$	$A_6 = \{win, draw\}$
$A_3 = \{loss\}$	$A_7 = \{loss, draw\}$
$A_4 = \{Draw\}$	$A_8 = \{win, loss, draw\}$

$B = \{ A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8 \}$ → (event space)

→ B is a power set or even space

→ Set of all possible subsets is called event space.