One assignment Sunday It is On List

140 Python

9years 22 Cognizant : 32 Lakhs

4 years : 17.5 CTC

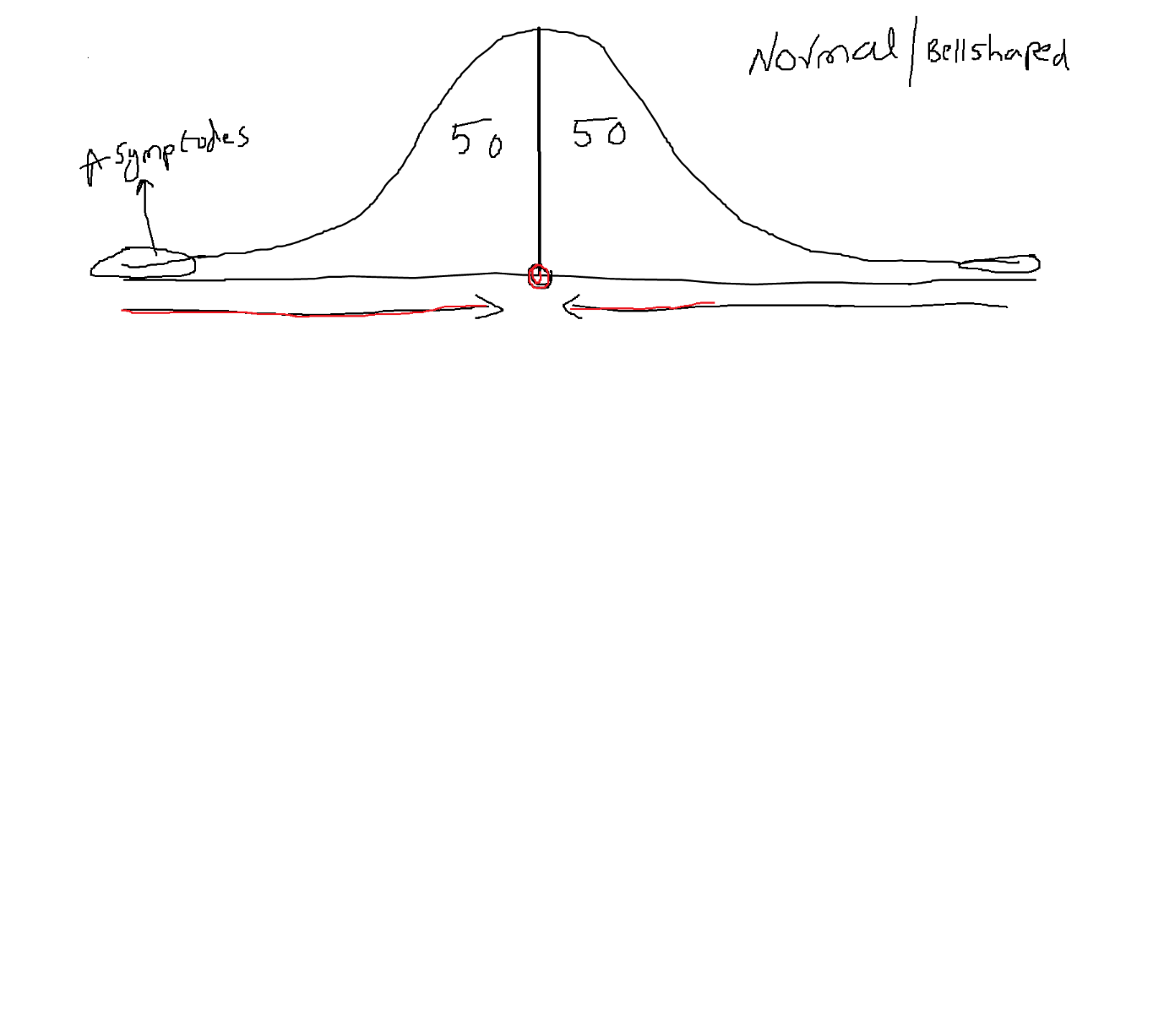
Session-1: Data types

Levels of data

Session-2: Categorical data vs Numerical data

Frequency table Frequency distribution table

Bar chart /Pie chart Histogram / Distribution plot



1. Central tendency :

* Mean
* Median
* Mode

1. Data dispersion / distribution / flow

* Range
* Mean deviation
* Absolute mean deviation
* Variance
* Standard deviation

Mean:

The virat kohli batting average in ODI 50

Tech guy : grand father should

If he goes to the ground to play cricket, he’ll do min 50 runs in every match

10th class marks

Telugu : 91 Hindi : 80 English : 95 Maths : 90 Science = 90 Social : 95

Every subject 100

91+80+95+90+90+95= 541/600

74 marks === Bi.p.c 11k rank B.P

160 147 1400

1st English ==== fail

5 years girls

Median:

* Middle values
* 10,9,15,17,21,8,5,3: even count
* 3 5 8 9 (9.5) 10 15 17 21 = (9+10)/2
* 10,9,15,17,21,8,5,3,21 : odd count
* 3 5 8 9 10 15 17 21 21 : 10 median
* 50 percentile data exactly half
* Percentage vs percentile GATE CAT GMAT

Mean vs Median:

Suppose USA ppl wants to know

4 members income

1lakh 2lakh 50k 75k

Average= (1L+2L+0.5L+0.75L)/4 = 1 lakh 10 thousand

5 members he added Ambani income : 100cr

1lakh 2lakh 50k 75k 100cr

Average= (1L+2L+0.5L+0.75L+100cr)/4= 25crs

One observation is very huge either positive side or negative side will impact entire analysis

This observations are called as outliers

Zim: poor

50k 75k 1lakh 2lakh median=75k

50k 75k 1lak 2lakh 100cr Median= 1lakh

* Median will not affect when outliers are present
* Mean will affect when outliers are present

75l 80l 1cr =====

75l 80l 1cr 10cr

the best thing is that the outliers are always at the first or end in case of decending and ascending order which will definitely be eliminated. Median will not affect.

Mode:

Most occurrence observation in the data

1,1,1,2,2,2,2,3,3,3,3,3,4,4,5,5= 3

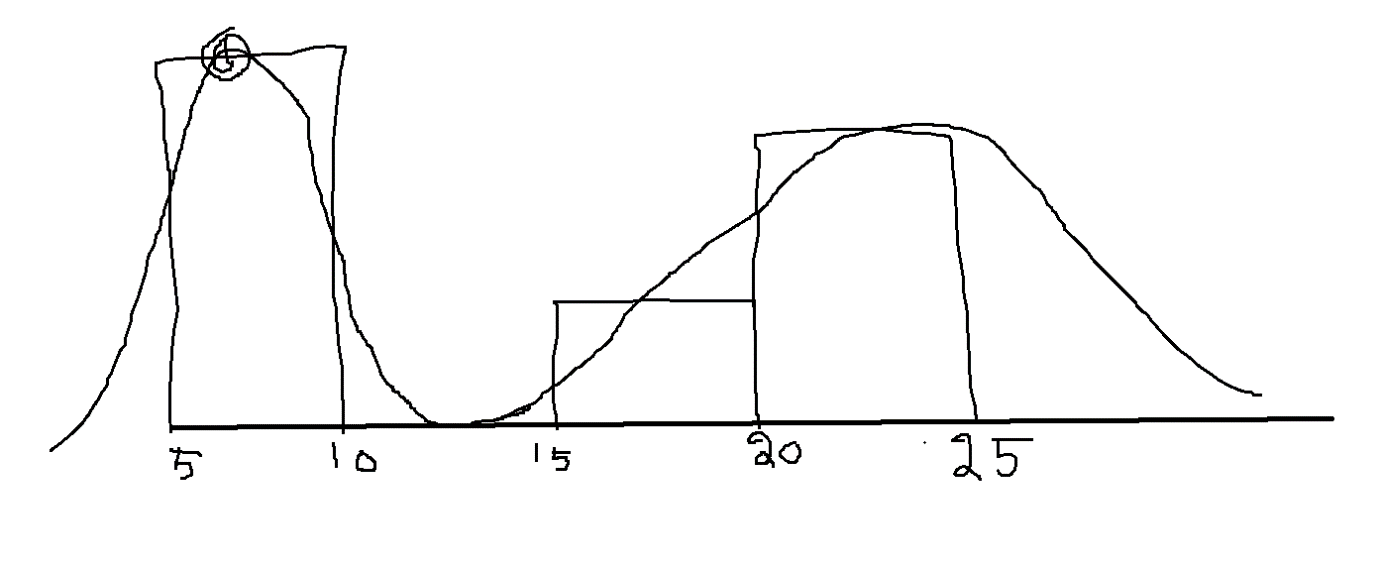
10,10,10,16,16,21,21,21,21,21,5,7,8

5-10 : 6

10-15: 0

15-20: 2

20-25: 5



Raw data ==== > CI ==== > Interval count ==== > Histogram ==== > Distribution plot

Interval count is nothing but repeated values only

The spikes in data distribution will provide mode information

If data has single spike : uni mode

If data has double spike: bi-mode

If data has many spike : Multi mode

In any data distribution

Always see the real line , exactly half : 50 percentile It is a Median point

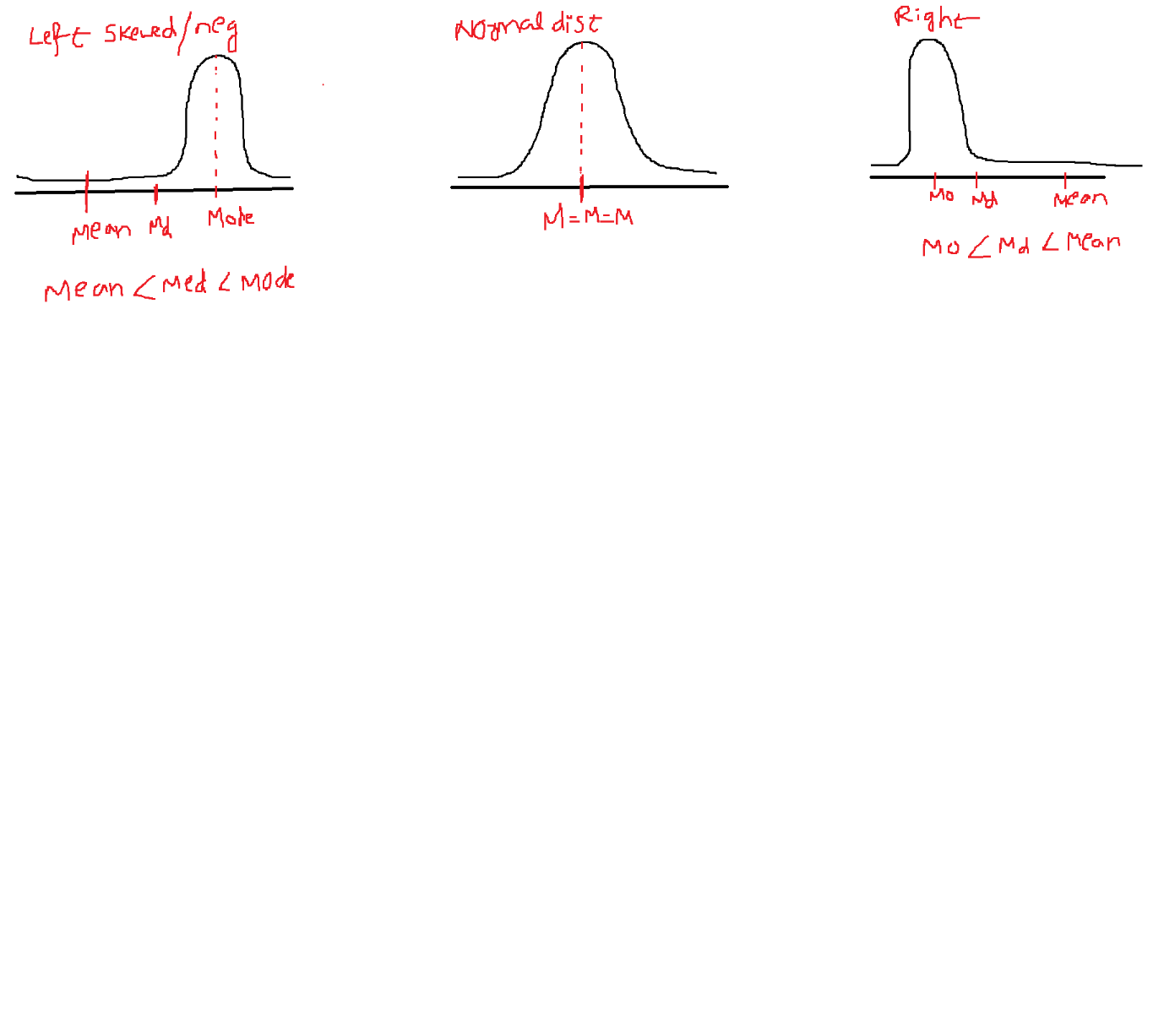
Always see the distribution spike : Mode

Added ambani incr ==== mean in crs

Outliers values will pull the mean to that side

Pulling ========== skewed

* Right skewed or positive skewed
* Left skewed or negative skewed
* No skew or Normal



For Right skewed or Positive skewed : Mean> Median> Mode or Mode<Median<Mean

For Left skewed or Negative Skewed: Mode>Median>Mean or Mean < Median < Mode

For No skew or Normal distribution : Mean=Median=mode

Sir in normal distribution what we are representing on x-axis and y-axis?

Any distribution : CI vs Frequency From Histogram

All maths developed by assumption data follows Normal distribution

We need to convert skewed to no skew : outliers are causing this skew

Outlier analysis will convert skew to no skew

P3