1. Stats

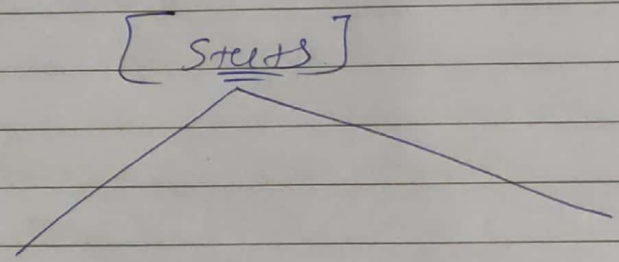
* What is Stats?

→ Stats is a scinie of collecting, analyzing & organizing data.

* Data :-

→ facts or pieces of information that can be measured.

* Types of Stats



Descriptive Stats
↓
[consist of organizing & summarizing data.]

Inferential Stats
↓
[using data, we can make conclusion using techniques.]

2.

*-) ## Sample & Population

Population, N / sample, n

↓

-) whole dataset is known as population.

* sampling techniques

1. simple Random sampling
2. Stratified sampling
3. Systematic sampling
4. Convenience sampling

* Variables

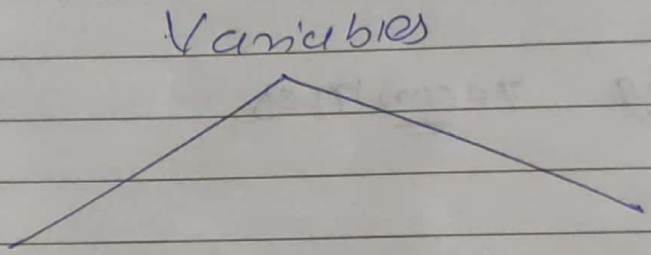
-) It is a property that can hold or take ~~only~~ various value.

Eg:- {8, 20, 5, 7, 600, 4078, ...}

3.

marks : { 76, 80, 95, ... }

* Types of variables



1) Qualitative

→ categorical values

Eg

IQ: 0 - 10 → low

10 - 50 → Avg

50 - 70 → good

2) Quantitative

→ Numerical value

Eg

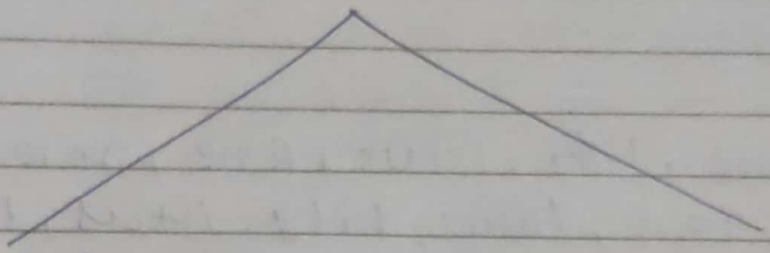
Height : { 162, 159 }

weight : { 59, 65 }

4.

There are two types of Quantitative.

2) Quantitative



1) Discrete (int)

→ whole no.

Eg.

→ no. of students
→ no. of peoples.

continuous (float)

→ Decimal no.

Eg.

→ Height
{165.2, 167.9, ...}

→ weight
{165.5, 60.9, ...}

* Variable measurement scales.

① ordinal :- ordered [rank, graduation]

② Nominal :- categorical values [colors, classes]

③ Interval :- [No zero / abs datepoint]

④ Ratio :- zero means nothing

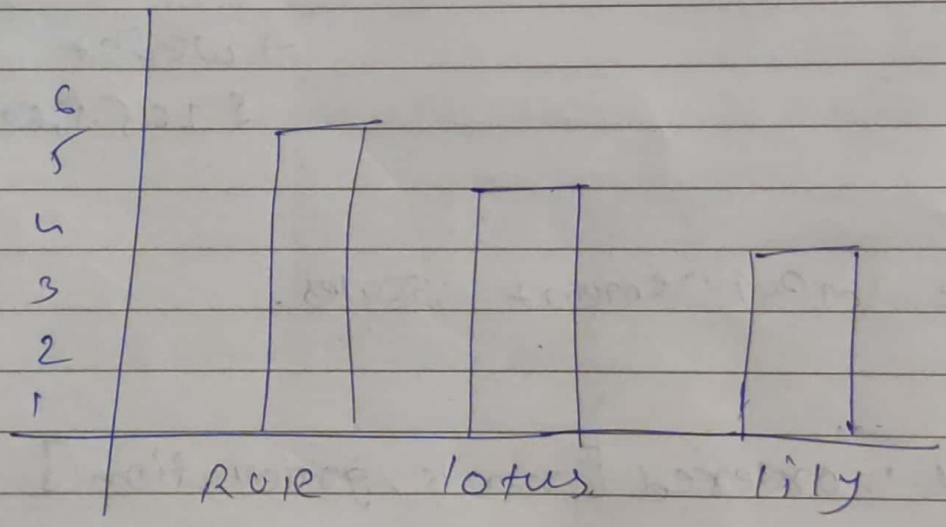
5.

Frequency

Eg.

{Rose, lilly, lotus, Rose, rose, rose, rose, rose, lotus, lily, lotus, lily, lotus}

Flowers	F	cf
Rose	5	5
lily	3	8
lotus	4	12
	12	



bar graph / Chart

6.

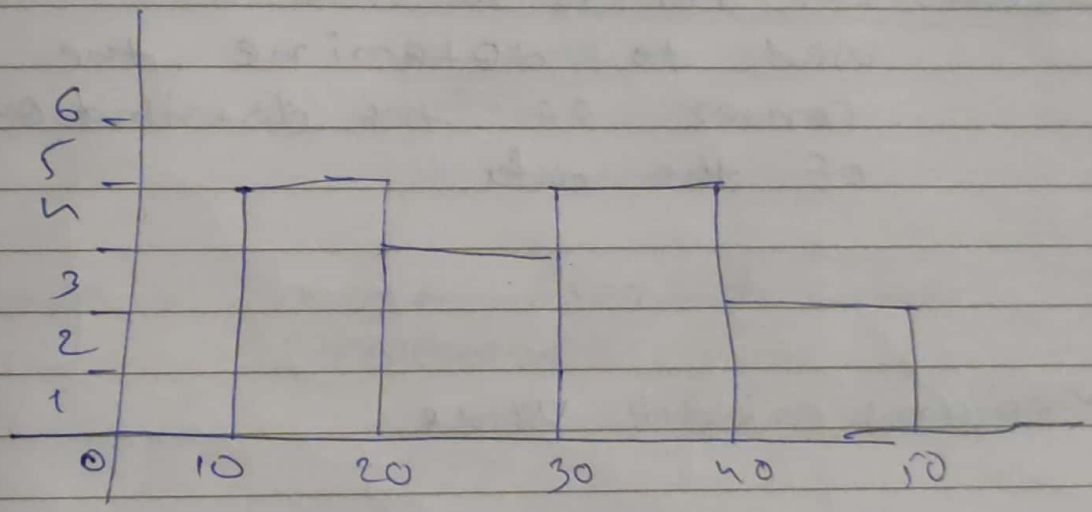
histogram

marks: [12, 15, 17, 15, 21, 27, 35, 34, 36, 39, 42, 45]

bins

(10-20)	4
(20-30)	3
(30-40)	4
(40-50)	2

→ continuous



7.

measure of central Tendency

Avg → mean

PoP

sum

$$\mu = \frac{\Sigma x_i}{N} \qquad \bar{x} = \frac{\Sigma x_i}{n}$$

→ mean :- It refers to the measure used to determine the center of the distribution of the data.

→ median : middle value

→ in ascending order

→ data

even odd

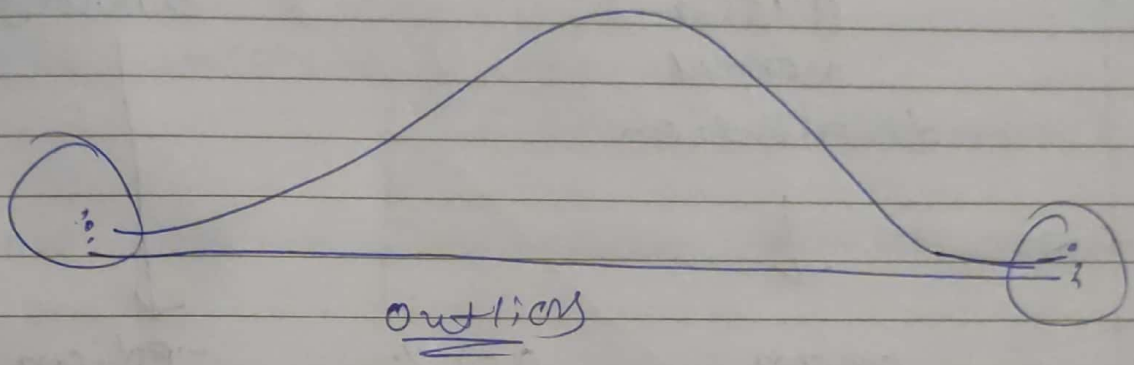
$$\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)^{th}}{2}$$

$$\longrightarrow \frac{(n+1)^{th}}{2}$$

8.

* outliers :- a datapoint who doesn't follow patterns or trend of the dataset than it is considered as outlier.

[are extrem points]



outliers

* mode :- most frequent value (repeated)

* for categorical missing data use (mode)

* ~~for numerical values~~

9.

* for numerical values.

numerical values

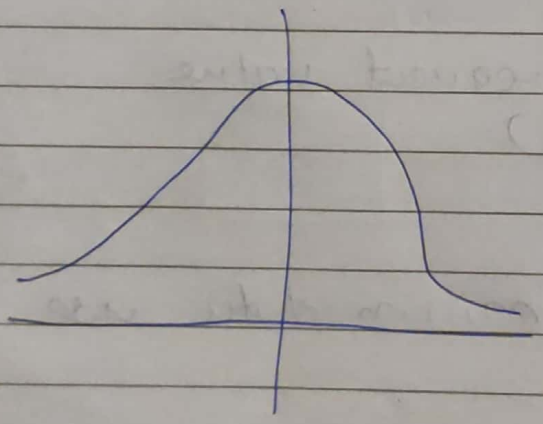
gussian / normal distribution

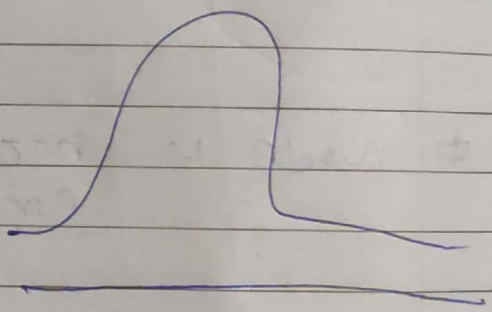
skewed

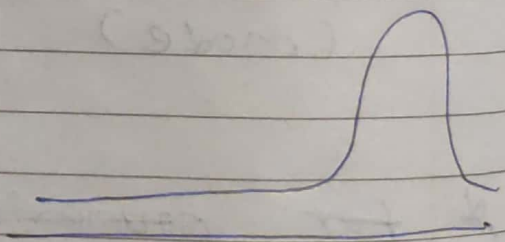
mean

0-5 %.
10 %.

median







men = median = mode

10.

* Variance

→ It measures how far the numbers in a dataset are from the mean (avg)

→ High variance → more spread. (far from the mean)

→ low variance → closer to mean

PoP

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

summation

Sample

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

x_i = every data pt
μ = mean of PoP
N = total PoP
Σ = summation

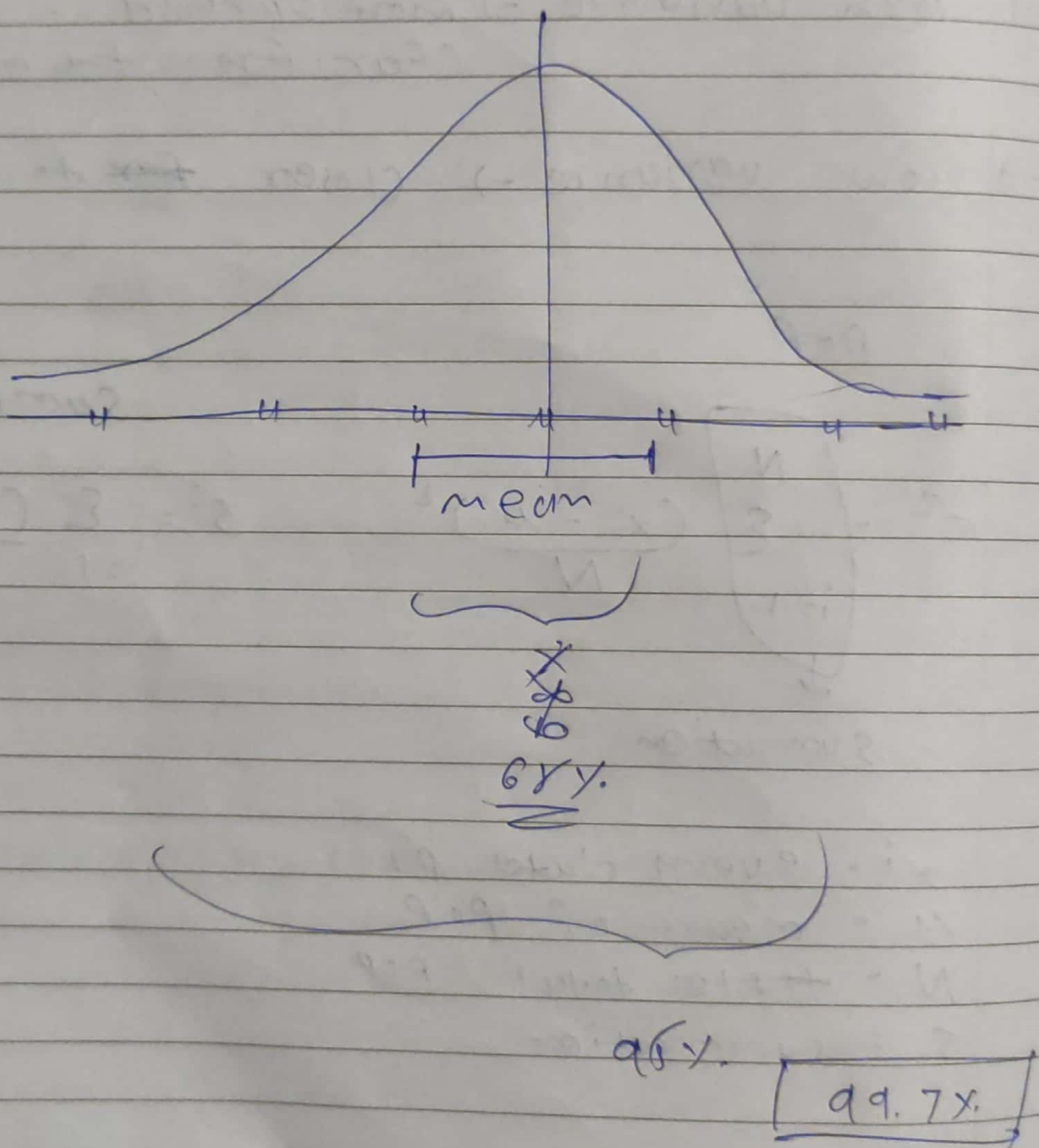
* standard deviation

pop

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

sample

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



12.

$\mu \pm 1\sigma$: 68%.
$\mu \pm 2\sigma$: 95%.
$\mu \pm 3\sigma$: 99.7%.

→ square root of variance

→ it gives measure of spred that is in the same units as the original data, making it easier to interpret.

five number summary

① minimum Q_0

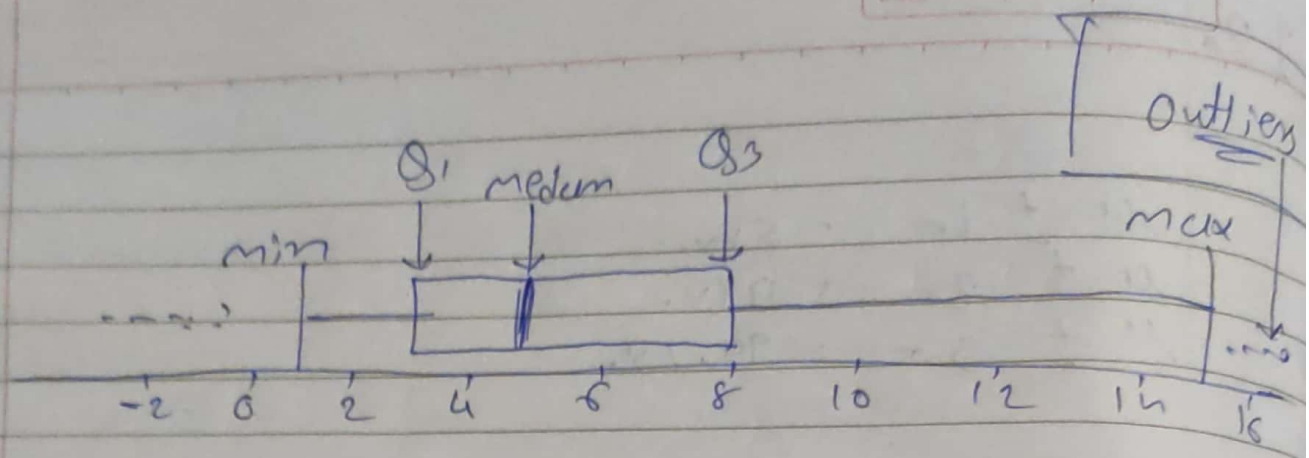
② 25 percentile ⟶ first quartile $[Q_1]$

③ median ⟶ 50 percentile Q_2

④ 75 percentile ⟶ third quartile $[Q_3]$

⑤ Maximam Q_4

13.



Box plot

* Z-score

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

$\mu = 0$

$\sigma = 1$

* normalization

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}$$

14.

* Exponential distribution

→ It describes the time b/w events in a process where events occur independently & at a constant rate λ.

$$f(x) = \lambda e^{-\lambda x}$$

$$x \geq 0$$

* Bernoulli distribution (discrete)

→ It models a single experiment with two possible outcomes.

success ($x=1$) ; failure ($x=0$)

* Uniform distribution (continuous)

$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

→ The probability of any value within the range $[a, b]$ is same

15.

* uniform distribution (discrete)

all outcomes are equally likely

$P(x) = \frac{1}{n}$ ⟶ total no.

Hypothesis testing

→ A statistical hypothesis that is test is a method of statistical inference used to decide whether the data at hand is sufficient to support a particular hypothesis.

* Rejection Region method

① H_0 & H_a
② α → value
③ assuptions
④ decide test → { → Z-test, → T-Test }
⑤ value
⑥ Test conduct
⑦ Reject / Accept
⑧ state results

16.

* 2 types of ~~error~~ errors

	Type 1 ✓ Ho true	Type -2 ✗ Ho false
reject Ho	Type-1	correct
accept Ho	correct	Type-2

=) Type=2 false +ve

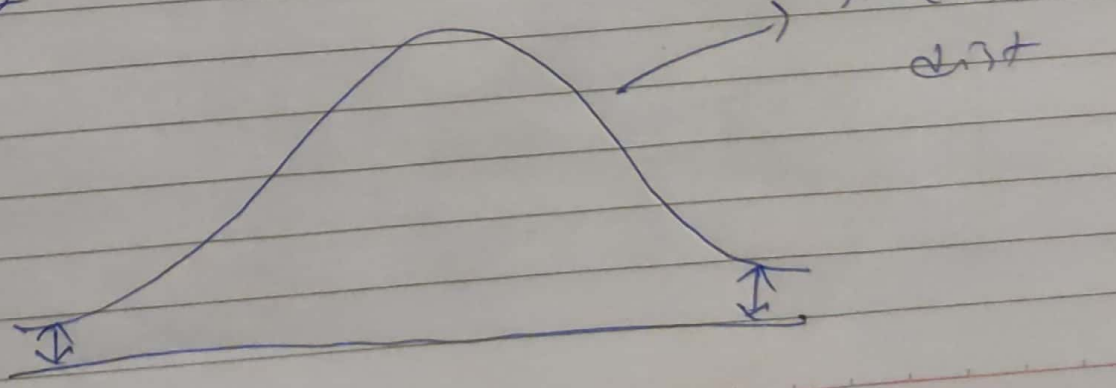
↳) Ho reject ⟶ Ho ture (coorct)

∴ reject Ho when Ho is actualy correct

=> Type=2 false -ve

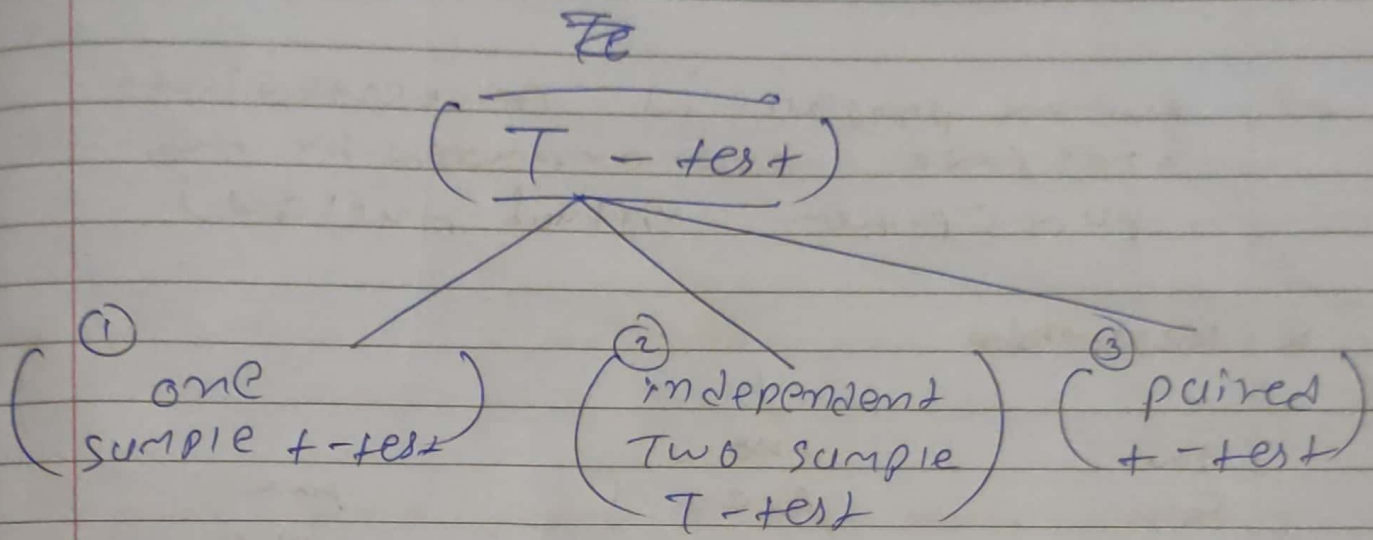
∴ accept Ho when Ho is actually incorrect

T-test



17.

→ 3 types



① one sample t-test

→ compares the mean of a single sample to a known μ

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

② Independent Two sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\left[\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]} \rightarrow \text{Standard Error}$$

③ Paired t-test

$$t = \frac{\bar{d}}{sd/\sqrt{n}}$$

$\bar{d}$ => 2 -> mean diff

sd = std -> diff

~~18~~ 19.

Chi-square test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O => observed frequency
E => Expected frequency

$$E = \frac{\text{Row Total} \times \text{column total}}{\text{Grand Total}}$$

$$df = (r-1) \times (c-1)$$

r = no. of rows
c = no. of cols