Statistics

Definition –

A branch of applied maths that involves the collection description, presentation, analysis and interpretation of numerical data. is called statistics.

Type of Data –

① structured data

② Unstructured data

Stages of stats –

① collection of data

② Organizing of data

③ Presentation of data

④ Analysis of data

⑤ Intelepretation of data

Type of stats -

① Descriptive stats

Available data sample or population on it we perform action like anatysny. discribe, summerize. it called descriptive stats.

② Inferential stats -

On describe data we perform interepetation like hypothesis testing on the data for example z-test, t-test, F-test, chi-squeue test. is called inferential statistics

① Descriptive stats.

① Univariate Des. stats.

② Bivariate Des. stats.

③ multi variate Des. stats.

Population & Sample

Entire data available is call population.

Population = N

Sample = n

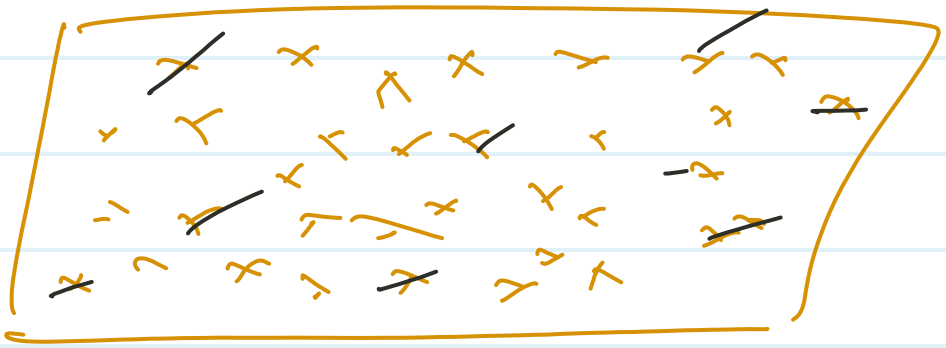
Sampling methods –

Type of Sampling method

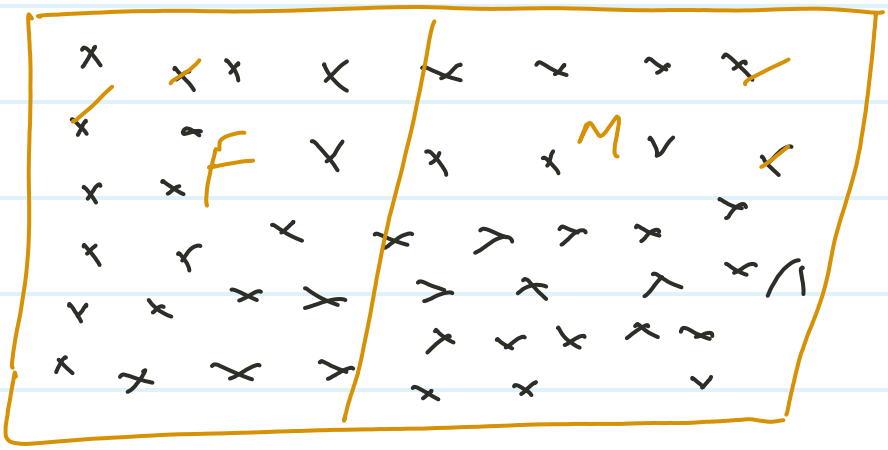
1. Probability Sampling
2. Non-probability sampling.

Probability sampling

① Simple random Sampling. –



② stratified Sampling



> 18

Agenda.

③ Systematic Sampling.-

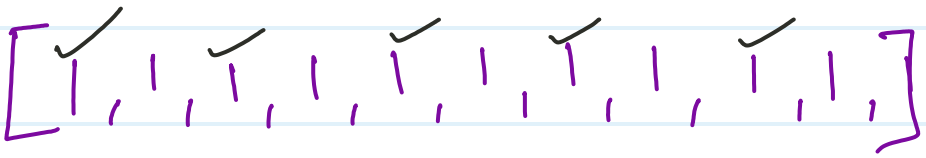
n^{th} select

population - 10000
sample - 50

$$= \frac{10000}{50}$$

$$= 200.$$

1^{st}, 201^{st}, 301^{st},



② Non-probability Sampling -

Convinience Sampling -

Research -

How to calculate sample size?

✱ Cochran formula

$$n_0 = \frac{z^2 pq}{e^2}$$ – infinite populations

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$ – Required population.

e = margin of error 5%

P = Population proportion 50%.

q = $1 - P$

Z = value from Z-table

C.I.		α
90 %	-	0.1
95 %	-	0.05
99 %	-	0.01

Z-table

$$\alpha = 0.05$$

$$1 - 0.05 = 0.95$$

$$Z = 1.6 + 0.05$$

$$Z = 1.65$$

$$= \frac{(1.65)^2 \times 0.5 \times (1 - 0.5)}{(0.05)^2}$$

$$= \frac{2.72 \times 0.5 \times 0.5}{0.025}$$

$$n_0 = \frac{0.68}{0.025} = 27.2$$

Required population $N = 10000$

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

$$= \frac{27.2}{1 + \frac{(27.2 - 1)}{10000}}$$

$$\Rightarrow 27.12$$

$$28 / 27$$

If population = 100000

Sample =

Descriptive stats -

1. measure of Center tendency
2. measure of Dispersion or variation
3. measure of position
4. measure of shape

✱ Measure of Center Tendency

(1) mean -

[2, 3, 4, 5, 6]

$$\text{mean} = \frac{2+3+4+5+6}{5} = \frac{20}{5}$$

$$= 4$$

population mean = μ

Sample mean = $\bar{x}$

② median –

I. case [6, 2, 4, 5, 3]

data sort

[2, 3, 4, 5, 6]

↑

median = 4

II - case

[2, 3, 5, 7, 8, 9, 11, 13]

$$\text{median} = \frac{7+8}{2} \Rightarrow 7.5$$

③ mode –

(i) uni-modal
(ii) Bi-modal
(iii) multi-modal

① [2, 3, 4, 5, 5, 6]

mode = 5

② [2, 2, 4, 5, 6, 6]

mode = 2, 6

③ [2, 2, 4, 4, 5, 5, 6, 7, 8]

mode = 2, 4, 5

→ main application

→ outlier handeling

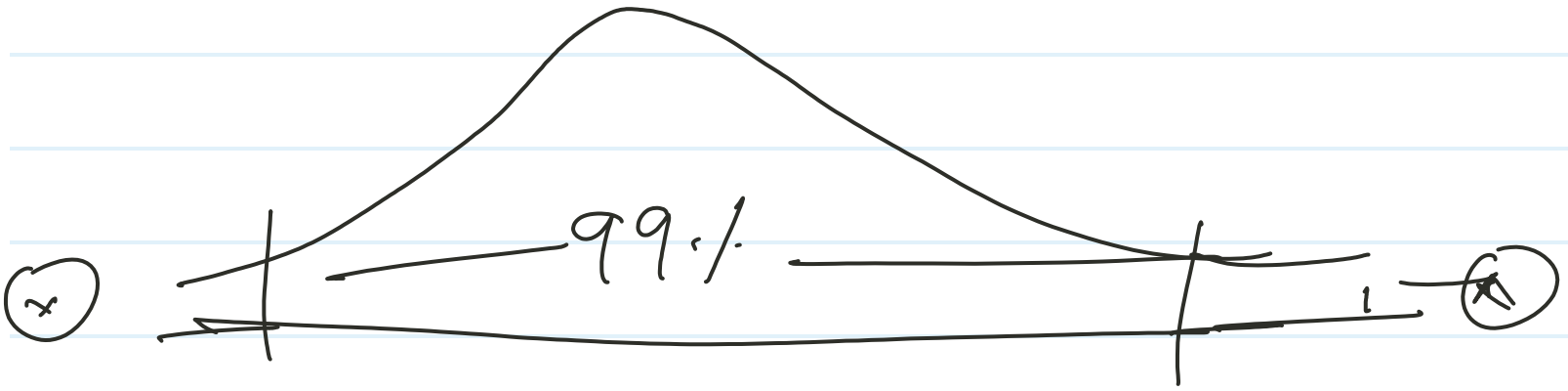
→ missing value handel

X	Y
2	2
3	3
~	4
7	45 ✓
~	7
11	5
13	

2, 3, 4, 5, 7, 45

4.5

2, 3, 4, 5, 7, 4.5



Result

P

F

P

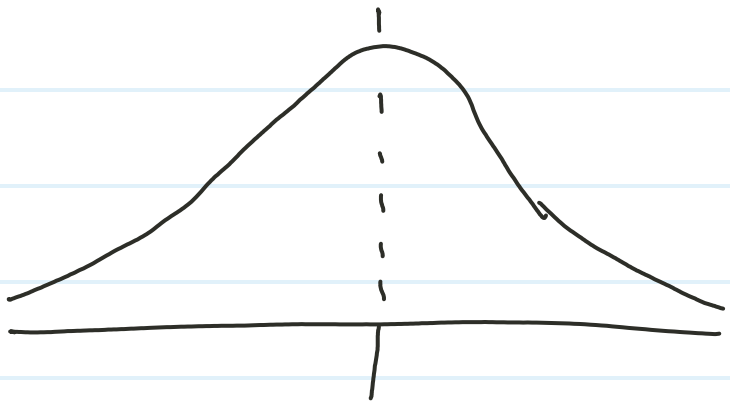
F

P

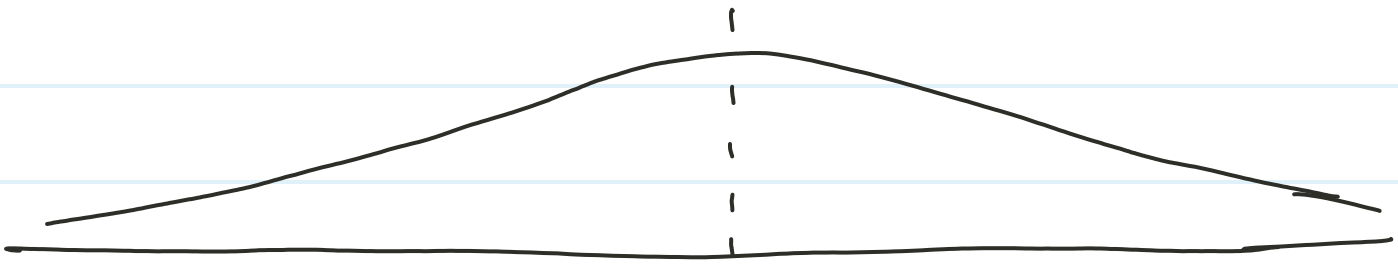
F

P

✱ measure of Dispersion or variance



— low variability



High variability

① mean absolute deviation -

The mean absolute devi. of a dataset is the avg. distance blw each data point and the mean.

$$\Rightarrow \quad \frac{1}{N} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Data - [10, 15, 15, 17, 18, 21]

$$\text{mean} = \frac{96}{6} = 16$$

$|10-16| = |-6| = 6$

$|15-16| = |-1| = 1$

$|15-16| = 1$

$|17-16| = 1$

$|18-16| = 2$

$|21-16| = 5$

$\frac{16}{16}$

$$\Rightarrow \frac{16}{6} \Rightarrow 2.67$$

② Variance -

It tells the degree of spread in dataset.

High variability - datapoint spread widely

low variability - data point close to mean.

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

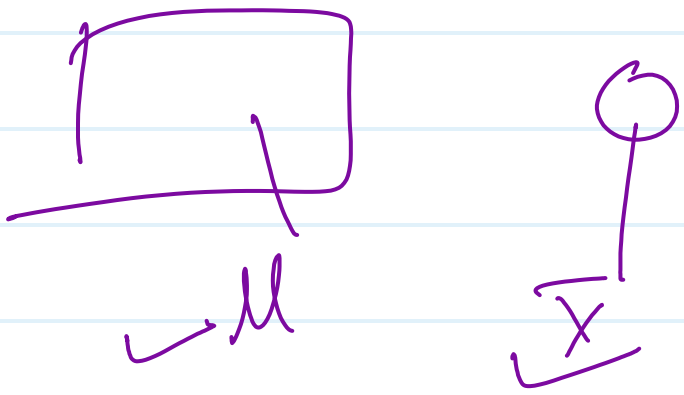
sample $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Note- $n-1$ is a degree of freedom

OR

Besils correction

For keep away result from biased.



③ standard deviation -

The square root of variance is called st. dev..

The farthes the data points from the higher the deviation

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

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

③ Range -

$[1, 2, 5, 6, 11, 15, 19, 25, 30]$

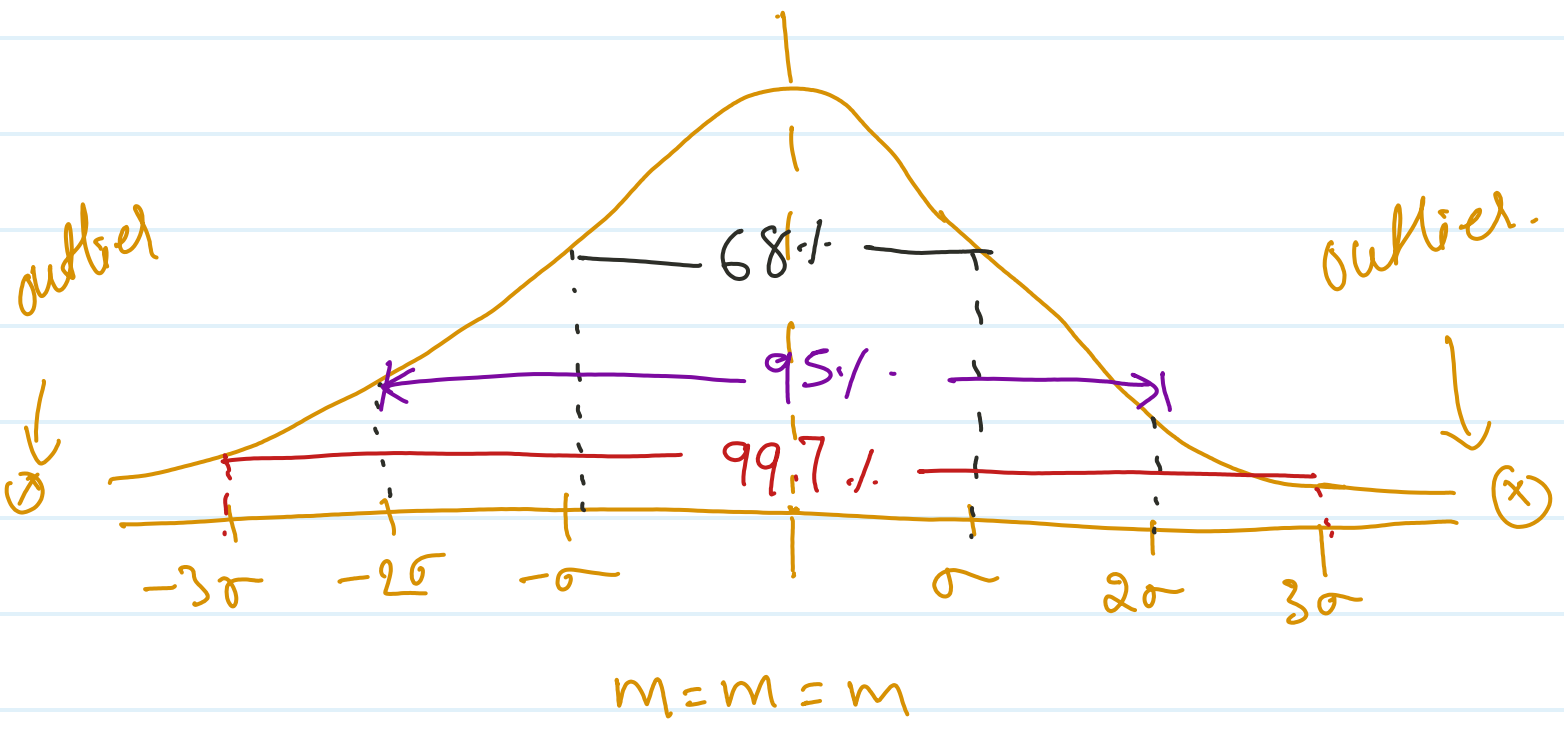
max - 30

mn - 1

range = max - mn

= 30 - 1 = 29

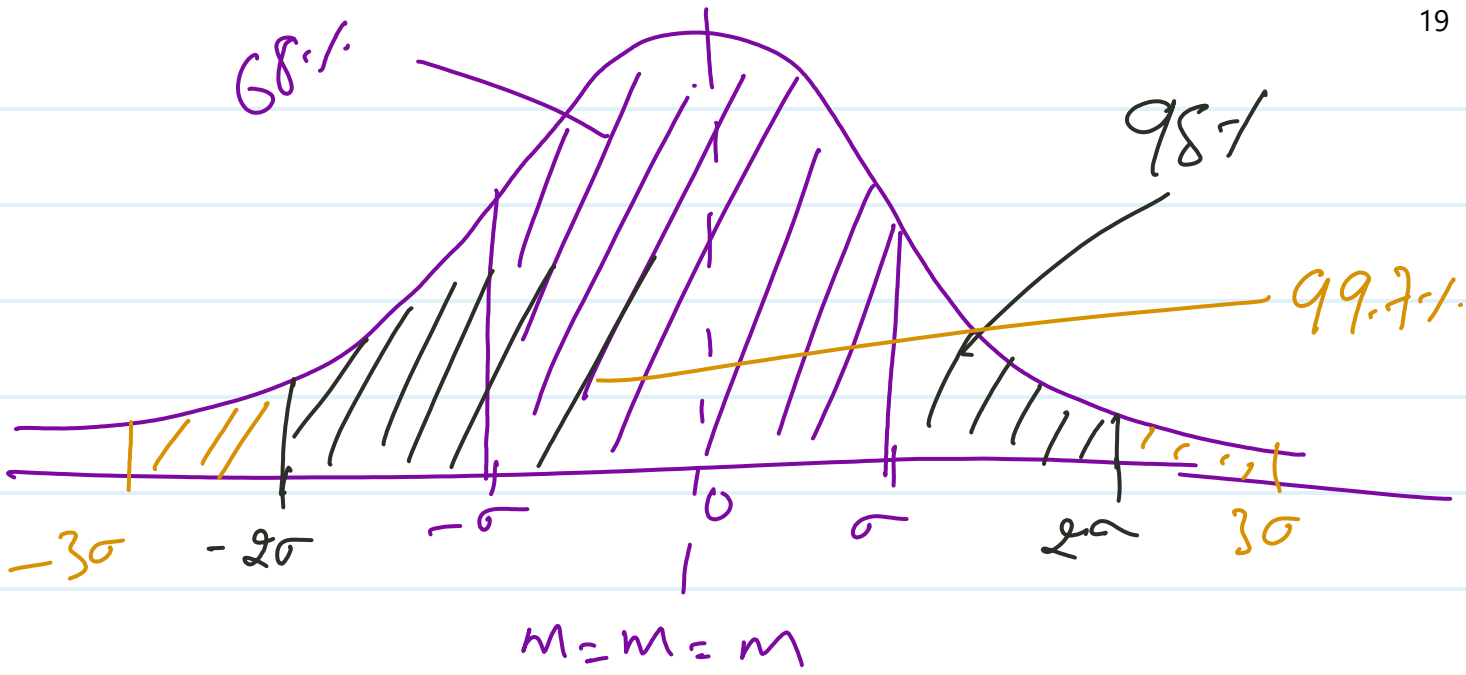
Emperical Rule



Normal Distribution Curve

68 – 95 – 99.7

σ 2σ 3σ


68.1.
95.1
99.7.1.
-3σ
-2σ
-σ
0
σ
2σ
3σ
M=M=M