Statistics

→ What is Statistics:

It is science of collecting, organizing, analysing, data for better decision making.

→ What is data:

facts or Pieces of information that can be measured.

* eg:

marks of 10th class student : {89, 90, 98, 77, ... }

age of students : {24, 24, 21, ... }

→ Types of stats:

1. Descriptive states:

It consist of organizing & summarizing data.

2. Inferential states:

using Data, we can make conclusion using some techniques.

* eg:

1st Sem students marks: [89, 90, 87, 77, 65, 95, ...]

avg. of 1 Sem students = descriptive stats

1st Sem students marks = 7th sem students marks &

= Inferential stats

→ Sample and Population:

Population = N :
Whole dataset is known as population

Sample = n :
Small subsets of data taken from population is known as sample.

→ Sampling Techniques:

(1) Simple Random Sampling:
Every member of population has an equal chance of getting selected in sample (n):

(2) Stratified Sampling:
Spliting data into non-overlapping groups.

example: age group 20-40
40-60 etc.
Gender → Male
Female.

(3) Systematic Sampling:
every 5 or 6 num people taking as a sample.

(4) Convenience Sampling:
particular filed for getting sample

→ Variables
It is a property that can hold/store/take any value.

Age: $ 8, 10, 15, 20, 25, ... y
Marks: $ 76, 80, 95, ... y

→ Types of Variable:

(1) Qualitative:
categorical values (based on some characteristics, we can derive categorical values)

eg: IQ : 0-10 → low
10-50 → Avg.
50-90 → good.
gender, blood group, marital status, etc.

(2) Quantitative:
Numerical Value (measurable numerically)

ex: height → measured in cm or ft.
Weight → measured in kg.

Quantitative

Discrete (int)
(Whole num.)

ex: No. of students
No. of bank Acc.

Continuous (float)
(Decimal num.)

ex: Height
Weight

(1) Blood presure → continuous
(2) Marital status → Qualitative
(3) River length → continuous
(4) Song length → continuous
(5) Gender → Qualitative

→ Variable Measurement Scales:

(1) Ordinal :
Ordered [only order matter]
ex: rank, graduation, ...

(2) Nominal :
categorical values
ex: colour, classes, degrees, ...

(3) Interval :
[NO zero / absolute point] ordered as well as value matter.
ex: zero means not nothing like (0°C)

(4) Ratio:

Zero means nothing.

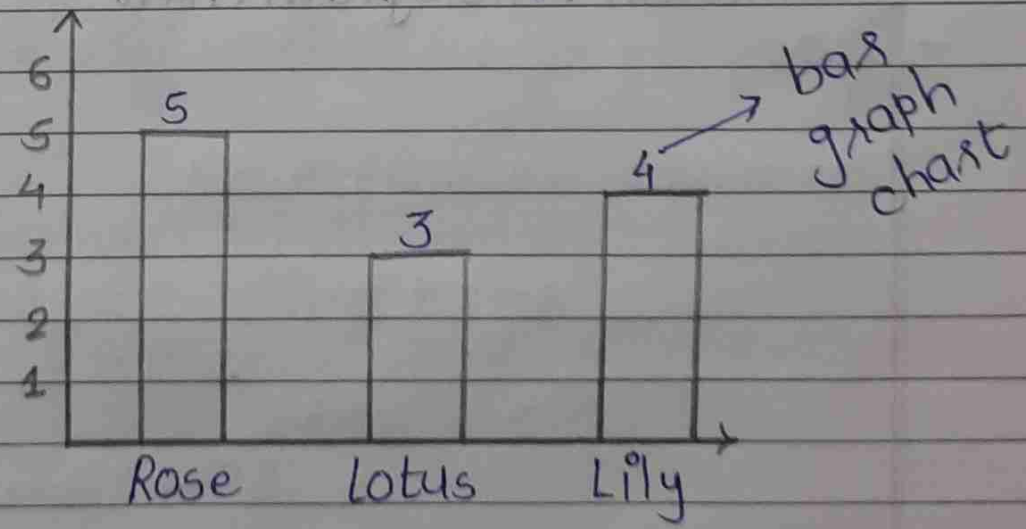
ex: 20 kg : 40 kg = 1:2

→ frequency:

data : flowers

[Rose, lily, lotus, Rose, Rose, Rose, Rose, lotus, lily, lotus, lily, lotus]

flowers	frequency	cum. frequency
Rose	5	5
lily	3	8
lotus	4	12
	12	

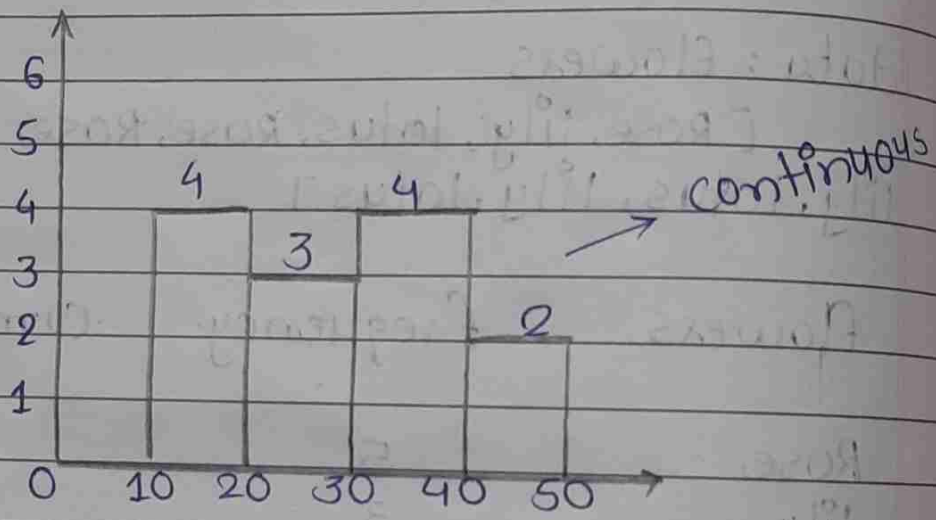


→ Histogram:

marks = [12, 15, 12, 15, 21, 27, 28, 34, 35, 36, 39, 42, 45]

grouping (bines):

10-20 = 4
20-30 = 3
30-40 = 4
40-50 = 2.



→ categorical: bar chart
Numerical: histogram chart

→ Descriptive Statistics:

1. Measure of central tendency:
2. Measure of dispersion:
3. Distribution

1. Measure of central tendency:

(a.) Mean = Avg.

It refers to the measured used to determine the center of the distribution of the data.

→ formula:

Population

$$\mu = \frac{\sum x_i}{N}$$

Sample

$$\bar{x} = \frac{\sum x_i}{n}$$

(b) Median: Middle value

↳ ascending order
↳ data

even odd

→ even:

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

dataset : {11, 12, 13, 14, 15, 16}

$n = 6$

$$\frac{\left(\frac{6}{2}\right)^{th} + \left[\left(\frac{6}{2}\right)^{th}+1\right]}{2}$$

$= \frac{3^{rd} + 4^{th}}{2}$

$= \frac{13+14}{2}$

$= 27/2 = 13.5$ is Median

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

dataset: {11, 12, 13, 14, 15}

$n = 5$

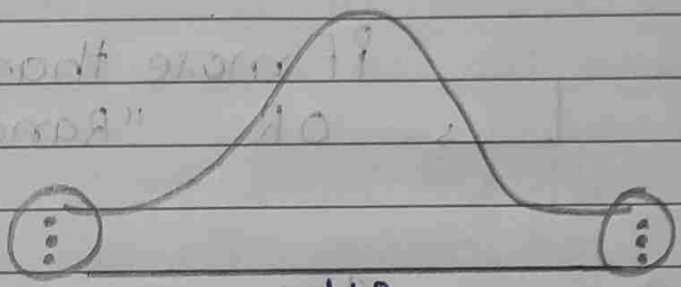
$= \left(\frac{5+1}{2}\right)^{th}$

$= 6/2 = 3^{rd}$

$= 13$ is Median.

M.IMP Median works well with outliers.

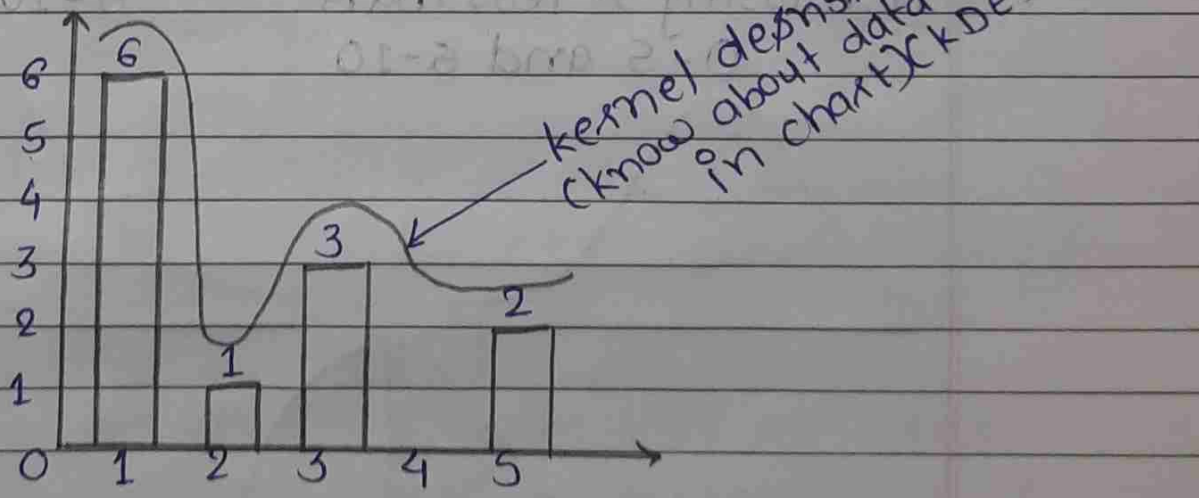
→ outliers:
a datapoint who does not follow pattern or trend of the dataset then it is considered as outliers [are extreme points]



→ Mode:
Most frequent value (repeated)

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

1 = 6 3 = 3
2 = 1 5 = 2



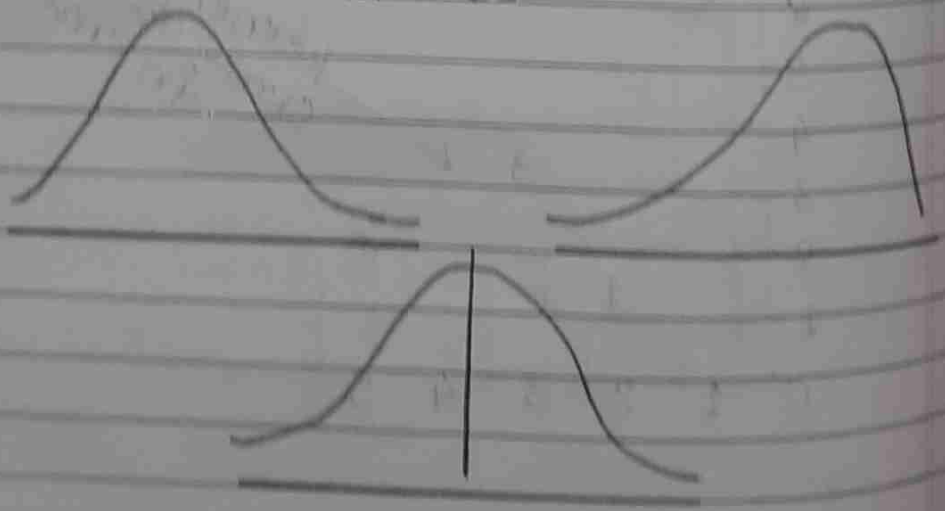
→ for categorical missing data:

0-5% → Mode
  ↳ new category "missing" "unknown"
    if more than 5%
  ↳ OR "Random"

→ for numerical values:

Gaussian / Normal Distribution
↓
mean
if missing % less than 0-5 and 5-10

Skewed
↓
median
if missing 10% less than 0-5% and 5-10%.



mean = median = mode

2. Measure of Dispersion: (Measure of spread

(a) Range:

(b) variance:

It measures how far the numbers in a dataset are from the mean (Avg.)

(how each value differs from a datas in a mean)

High variance → More spread (far from the mea
low variance → closer mean.

Population	Sample
$\sigma^2 = \sum_{i=1}^{N} \frac{(xi - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(xi - \bar{x})^2}{n-1}$
	(Bessel's correction)

example:

xi	$xi - \mu$	$(xi - \mu)^2$
1	-1.83	3.34
2	-0.83	0.69
2	-0.83	0.69
3	0.17	0.02
4	1.17	1.36
5	2.17	4.70

$(xi - \mu)^2 = 10.8$

$\mu = \frac{1+2+2+3+4+5}{6}$

$\therefore \mu = 2.83$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (xi - \mu)^2}{N} = \frac{10.8}{6} \quad \therefore \sigma^2 = 1.8$$

xi → every data point
μ → mean of population
N → Total of population
Σ → Summation

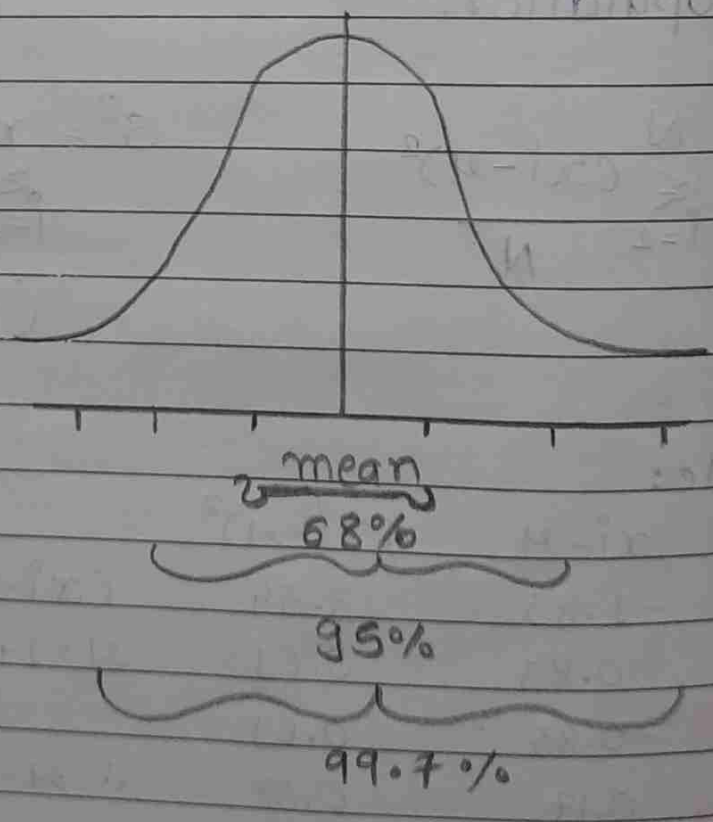
→ Standered Deviation:
(unit same easily comparable)

Population

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

Sample

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



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

Standered Deviation means square root of variance.

It gives measure of spread that is in the same units as the original data, making it easier to interpret

→ Percentage:

5 subject → Maths : 88, S.S : 75, Sci : 90, Eng : 80
G.K : 99 → out of hundred marks

$$= \frac{88+75+90+80+99}{500} \times 100 = 86.4$$

→ percentile:

It is a value below which a certain % of observation lie.

data is always in ascending order.

dataset = {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

$n = 20$

percentile rank of 10 = $\frac{\text{no. of values below } x}{n} \times 100$

$= \frac{16}{20} \times 100 = \frac{4}{5} \times 100 = 80$

80% values are below 10.

eg: 2

percentaile rank of 11 = $\frac{17}{20} \times 100 = 85$

85% values are below 11.

→ What value exists at percentaile rank of 25 ?

$$\text{Value} = \left(\frac{\text{percentile}}{100} \times n\right) + 1$$

$$= \left(\frac{25}{100} \times 20\right) + 1$$

$$= 5+1$$

$$= 6 \text{ index value.}$$

$$\therefore 5$$

percentile rank of 75

$$\text{value} = \left(\frac{75}{100} \times 20\right) + 1 = 15 + 1 = 16^{th} \text{ index}$$

$$\therefore 9$$

→ 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. Maximum [Q_4]

dataset = [1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 15, 27]

lower fence = $Q_1 - 1.5\ (IQR)$

Higher fence = $Q_3 + 1.5\ (IQR)$

$IQR = Q_3 - Q_1$ → Inter Quartile Range

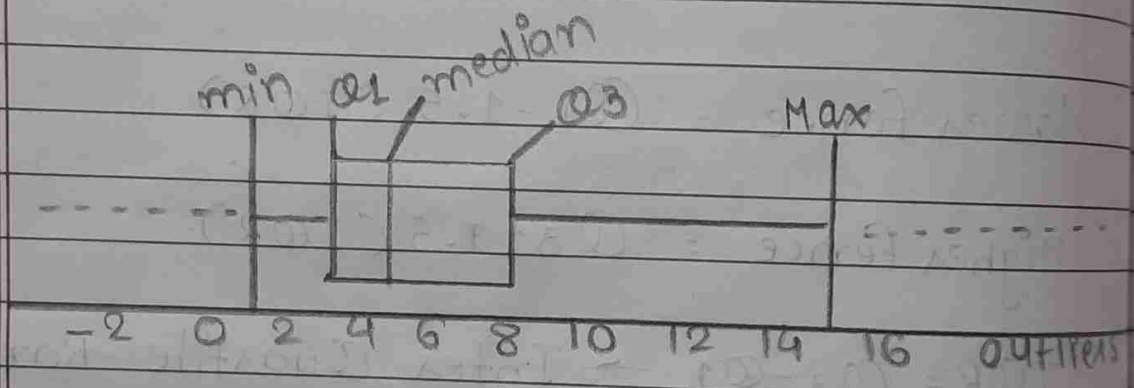
$Q_1 = 3$
$Q_3 = 8$

$IQR = 8 - 3 = 5$
lower fence = $3 - 1.5(5)$
$= -4.5$

Higher fence = $8 + 7.5$
$= 15.5$

(1) min = 1
(2) Q1 = 3
(3) Median = 5
(4) Q3 = 8
(5) Max = 15.

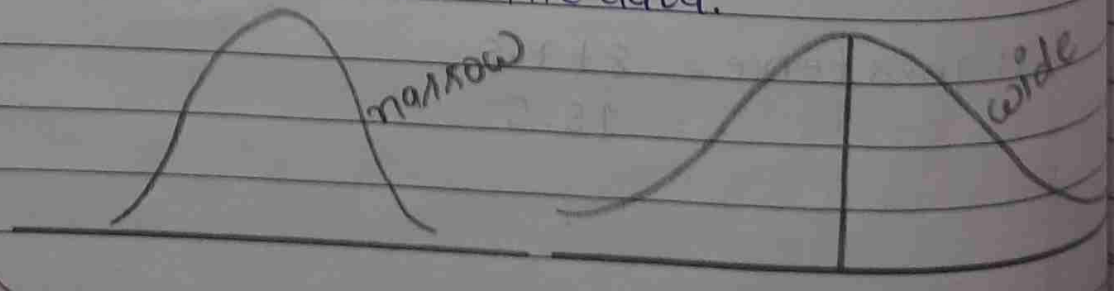
→ Box plot:



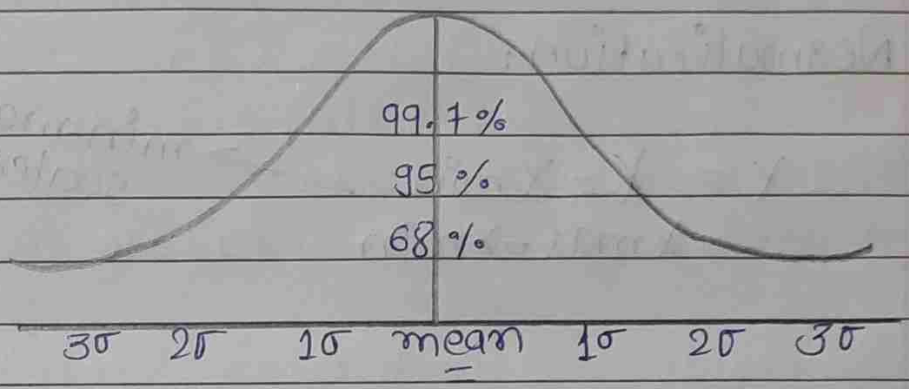
→ Data Distribution:

It refers to a way in which values or data points are spread or arranged.

It shows how often different values occur in data set and describes the overall pattern of the data.



→ Gaussian / Normal Distribution:



mean = median = mode

$\mu \pm 1\sigma = 68\%$
$\mu \pm 2\sigma = 95\%$
$\mu \pm 3\sigma = 99.7\%$
} Empirical Rule.

→ components of data distribution:

(1) center: Where our middle data lie. (mean, midian, mode
(2) Spread: (variance, & standered deviation)
(3) Shape: for overall from of data.
(4) outliers: Case extreme point)

→ Z-score:

$$\text{Z-score} = \frac{x_i - \mu}{\sigma} \quad (\because \mu = 0, \sigma = 1)$$

→ Standard normal Distribution:
When we find z-score of all data point and When we plot it is called standard normal Distribution.

at that time mu=0 and sigma=1.

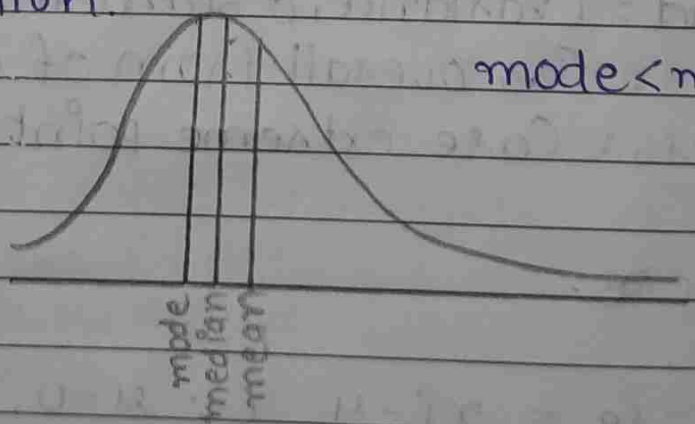
→ Normalization:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$ → min-max scales

→ Positively skewed distribution:

In positively skewed distribution, most values are concentrated on the low end, with a long trial extending to the right. A few high values pulls the average to the right of the median.

→ Skewness: A distribution or argumentay that deviates from the symmetrical bell is also called right skewed or right tailed distribution.

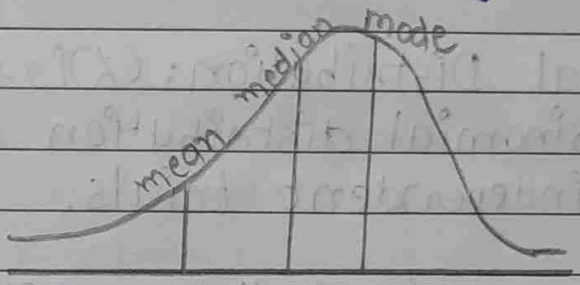


usefull for data with rare but significant high values such as income levels where few individuals earn much more than rest

The mean is greater than the median due to outliers on the higher end.

→ Negatively Skewed distribution:

Most values are clustered at higher end with a few values creating a long tail to the left.



mode > median > mean.

used for datasets where values are typically higher, but a few lower values exists (retirement age).

The mean is generally less than median due to leftward tail.

Also called left skewed or left tailed

It helps detect cases where data points are generally higher but occasionally much lower.

→ Exponential Distribution: (continuous)

λ = constant rate.

It describes the time below events in a process where events occur independently & at a constant rate λ.

$f(x) = \lambda e^{-\lambda x} \quad x \geq 0$

→ Bernoulli Distribution (Des Discrete)
It models a single experiment
two possible outcomes.
Success (x=1) ; failure (x=0)

→ Binomial Distribution: (Discrete)
Binomial distribution extends Be
to n independent trials.

$$P(X=k) \rightarrow \binom{n}{k} p^k (1-p)^{n-k}$$

(50)
X = random variable (no. of success ou
n trial
n = total no. of trial (100).
k = No. of sucess ($0 \leq k \leq n$)
p = P(sucess in 1 trial) 0.5

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$n=5, k=3, P=0.5$

$$P(3) = \frac{5!}{2!3!} \times (0.5)^{(3)} \times (0.5)^{(2)}$$

→ uniform distribution: (continuous)

$[a, b]$

$$f(x) = \frac{1}{b-a} \quad a \le x \le b.$$

The probability of any value within the range $[a, b]$ in same.

→ uniform distribution: (Discrete)

all outcomes are equally likely.

$$P(x) = \frac{1}{n}$$

$n \rightarrow$ Total no.

→ Confidence Interval:

eg.: $\bar{x} = 50 \rightarrow$ Point estimate → 4

$\pm 5 \rightarrow$ Margine of error

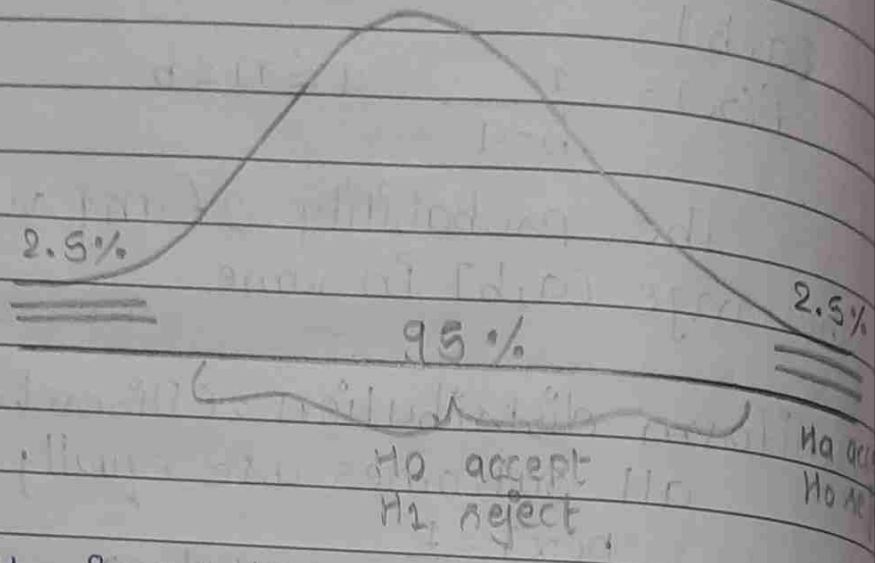
$50 \rightarrow$ point estimate

$45-55 \rightarrow$ It's called confidence Interval

It is a range of values within which we expect a particular population parameter to fall.

confidence Interval = point estimate $\pm$ Margin of error

→ confidence level:



→ Hypothesis Testing:

1 experience → 100 - toss coin.

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

Hypothesis testing allows us to make probabilistic statements about population parameters.

(1) Null Hypothesis: (H_0)

The null hypothesis assumes that there is no significant relationship or effect between two variables. [In simple terms → it says nothing new is happening]

It serves as a starting point for H.T & represents 'static quo' or the assumption of no effect until proven otherwise.

The purpose of H.T. is together evidence to right or fail Null hypothesis in favour of alternate hypothesis, which claims there is significant effect or relationship.

(2) Alternate hypothesis: (Ha or H_1)

It is a statement, that contradicts the NH & claims there is significance effect or ~~relath~~ relationship.

→ Rejection Region Method:

(1) Ho & Ha
(2) α → value → significance level → 95%
↳ loss of significance → 5%.
(3) Assumptions: if Normal Distribution ≥ 30
(4) decide test: z test / t-test
(5) Value:
(6) Test conduct
(7) Reject / Accepts
(8) state Results

→ Two Types of Errors:

	Type-1	Type-2
	Ho True	Ho false
Reject Ho	Type-1	correct
accept Ho	correct	Type-2

Type-1 : False positive
↳ Ho rejected → Ho True (correct)

Rejecting Ho when Ho is actually co

Type-2: False Negative

accept Ho when Ho is actually incor

→ P-value:

It is a measure of the strength the evidence against the null hypothesis

$P > \underset{\alpha}{0.05}$ → Null accept

$P < 0.05$ → Null reject

$P < 0.01$ → strong evidence

$0.01 \leq P < 0.05$ → Moderate evidence

$0.05 \leq P \leq 0.1 \rightarrow$ weak evidence

$P \gtrsim 0.1 \rightarrow$ NO - evidence.

$P < \alpha \rightarrow H_0$ reject

$P > \alpha \rightarrow H_0$ accept.