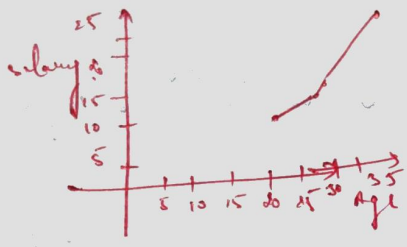
Statistics

★ Statistical analysis is meant to collect & study info available in large quantities.

→ Branch of maths (where computn is done on bulk of data) using charts & graphs.

↓

Age	Salary
25	10000
29	12000
31	16000
35	25000

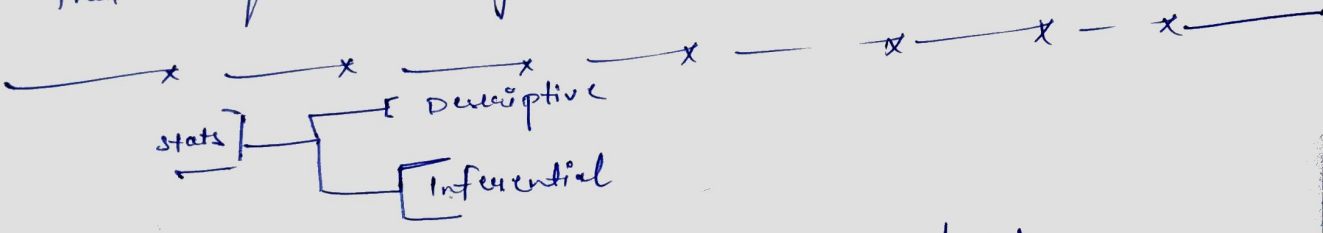


→ Trend (↑ in salary w Time)

★ Measurements: Data collected for analysis. | Sampling -

↓

Then analysis is done for measurement.



DESCRIPTIVE STATS: - • describes characteristics of data.

(What is data telling you?)

consist 3 measurements → Central Tendency (Mean, Median, Mode)

→ Measure of variability (spread) [Std, variance]

→ frequency distribut. [count]

CT → center of data,

MV → describe, dispersion of data set

FD → Describe - occurance of data within dataset.

$$SD(\sigma) = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \quad \{\mu \rightarrow \text{mean}\}$$

INFERENTIAL STATS:- Helps us to make conclusions / predictions about a big group using data from its small sub-group (sample).

2 MAIN AREAS →
- Estimating parameter
- Hypothesis Testing.

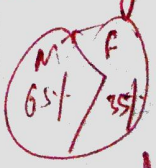
DATA [raw facts need to be analyzed]

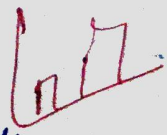
- Qualitative (categorical)
  - Nominal
  - Ordinal
- Quantitative (Numerical)
  - Discrete
  - Continuous

Qualitative / Categorical Data:-
- describes the data that fits into categories.
- Cate-info includes cate-variables which describe features as person's gender, hometown, etc.
- Not numerical
- sometime these have ~~###~~ numeric values but for no sense.
  eg → DOB: 19/7/05
  - Postal code

CD eg → Gender, Color, City

NOMINAL DATA. (Name only)
- info which helps to label the vars w/o numeric values.
- also called nominal scale - can't be measured / ~~scaled~~ ordered.
- Nom data examined using the grouping method.
  Data grouped into categories → frequence / % can be calculated.
- represented by pie charts.





ND → Blood Type, color.

ORDINAL DATA → have order.

eg → economics status, Grading (A, B), cust. ratings, Tshirt size

represented by :- Bar chart.

Quantitative Data →
- numerical data
ie → How much, how many.
→ km, N, W

Discrete [finite num of poss values] can't be subdivided meaningfully. (Whole Nums)
eg → no. of workers.

Continuous - data that can be calculated. that can be selected within a range. Has infinite no. of probable values.
ex → Temperature Range.

SAMPLING

Population
The universe of objects that is required to be analyzed.

SAMPLE
Subset of Populat°.
size of sample < size of Populat°

★ Why Sampling Important :-

> Gathering data from populat° → not possible, Sampling is applicable in such cases.
> sampling makes info faster.
> measuring every d-pt is not cost effective.
> can easily analyze the data using samples.
> smaller data set = smaller data collect° errors.

Sampling
- Probability sampling
- Non-Probability Sampling

Probability Sampling :-

Every member of popul° has an equal chance of being selected.

eg:→ Simple random sampling
- → Stratified Random s.
- → Cluster sampling
- → Systematic Sampling

Pros
- unbiased & representative
- results are statistically valid

Cons
- Time consuming & costly
- Needs a complete popul° list.

We use when:-
- to make statistical inferences
- To avoid biases.
- accuracy & reliability is important.

Non-Probability Sampling

- Not everyone has a chance of being selected.

→ samples are selected on basis of judgement or convience of accessing data.

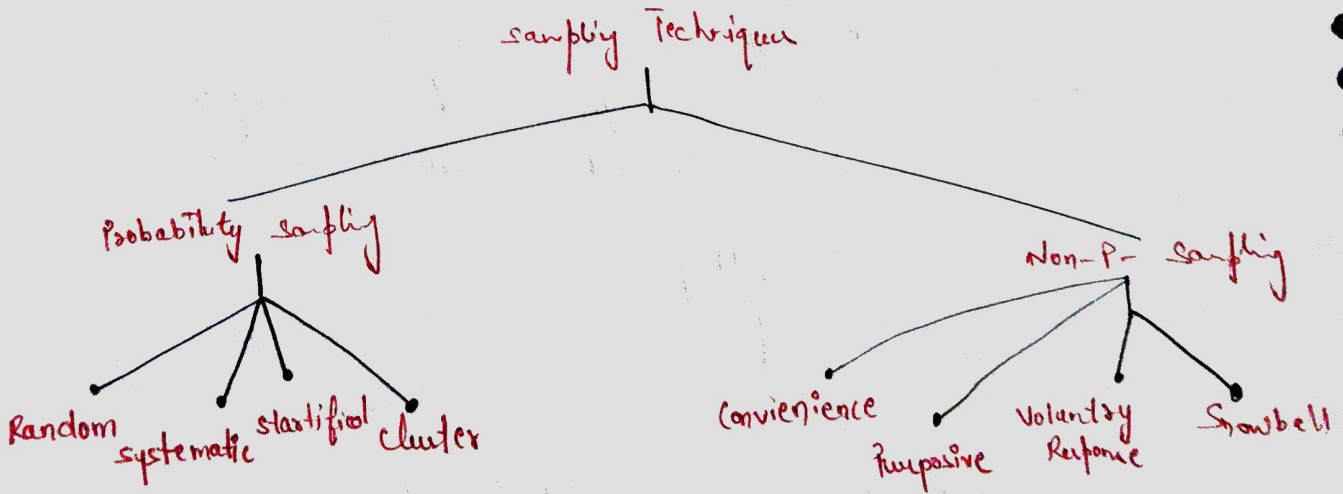
We use when,
- money, time, access → limited
- when we need quick insights

eg:
- → Convienience sampling
- → Purposive Sampling
- → voulantry response sampling
- → snowball sampling.

Pros	Cons
• fast & cheap	← High chance of bias
• Easy to execute	→ results can't be generalized
• useful when less popul° data available	

Cluster Random Sampling :- (PS).

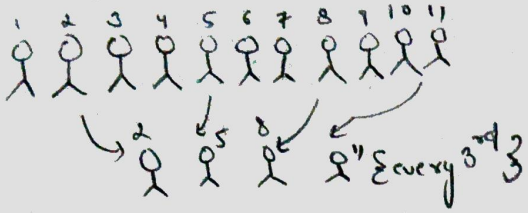
> Divide popul° into groups.
> then randomly select the group from all grps.



PROB. SAMPLING

1.) Random Sampling: randomly choose a member from the population.
> every member or set of number has an equal chance of being selected.

2.) Systematic Sampling:- > similar to random sampling.
. Pick every K^{th} person. (from a ordered set)



3.) Stratified Sampling:- - first divide population into groups
Homogeneous - Then from each group, select numbers randomly

eg: student from each year (1st, 2nd, 3rd & 4th)

stratum → similar groups.

4. Cluster Random Sampling:- Divide population into groups
Hetro - Then randomly select the subgroups from all the groups.

eg - Diff schools.

---

non-p-Sampling

1.) Convinience Sampling: , include member who are easy to reach for researcher.

2.) Purposive Sampling: ~~Simple~~ > select a sample based on purpose of the research.
> researcher select sample using their expertise & knowledge.

3.) Voluntary Response Sampling: > based on ease of access
> member volunteers by themselves instead researchers selecting the participants.

4.) Snowball Sampling: Existing participants refer or recruit new participants, forming a chain.

Population / Sample	Term	Notation	Formula
(P) $x_1, x_2, x_3 \ldots x_n$	Size	N	num of items.
	Mean	μ	$\sum_{i=1}^{i=N} x_i / N$
	Variance	σ^2	$\sum_{i=1}^{i=N} (x_i - \mu)^2 / N$
(S) $x_1, x_2, x_3 \ldots x_n$	S. Size	n	num. of items in sample
	S. Mean	$\bar{x}$	$\sum_{i=1}^{n} x_i / N$
	S. Variance	s^2	$\sum_{1}^{n} (x_i - \bar{x})^2 / (n-1)$

Bessel's Correction ($n-1$ denominator in sample formula)

> This method corrects the bias in the estimation of population variance.

$x_i = 5, 10, 15, 20, 30$

$N = 5$

$\mu = \frac{80}{5}$, $\mu = 16$

~~$\sigma^2 = (-11 - 6 - 1 + 4 + 14)/5$~~

~~$s^2 = (-16 + 16)^2/5$~~

$\sigma^2 = 112 + 36 + 1 + 16 + 196/5$

$\sigma^2 = 370/5$

$\sigma^2 = 74$ (variance)

$\sigma/s.t = \sqrt{74}$

Des. Stats → Measures of Central Tendency ... (Measures the central value of data set.)

> Gives the idea about the concentration of the value in the central part of the distribution.

- Mean
- Median
- Mode

MEAN (sensitive to outliers)

$$\bar{X} = \frac{\Sigma X}{N}$$

MEDIAN → The middle number

→ order all d. pts & pick from middle.

→ if there are 2 dp in middle, take mean of those two.

n is odd

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

n is even

$$\frac{\left(\frac{n}{2}\right)^{th} + \left[\left(\frac{n}{2}\right)+1\right]^{th}}{2} \left(\frac{m_1 + m_2}{2}\right)$$

MODE The value which occurs more frequent.

- can have more than one mode
- used in categorical data set.

Measure of Dispersion ...

→ Tells us where the center of data lies. describes, how much data values vary or how they are spread from the avg.

1. **Range** (max − min) → used to construct control chart in quality assurance / useful when to focus on Xtreme values of data set.
2. **Variance** (V / σ^2) $\frac{\Sigma (x-\mu)^2}{N}$ (μ → Mean) measures the dispersion of data around mean
   if v closer to mean → low variance [slightly dispersed]
   v far to mean → high variance [highly dispersed]
3. **Inter Quartile range**: IQR ($Q_3 - Q_1$) → measures the middle 50% of data. (around mean).
   → Ignores extreme small or large values (outliers).
   → Helpful to detect outliers present in the dataset.
   (Q_3: 75^{th} %ile, Q_1: 25^{th} %ile)
4. **Std** $= \sqrt{V}$

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

Indicates how far the data points is dispersed from mean. (σ)

5. **Mean Deviation** → how data is deviated from mean value

$$MD = \frac{\Sigma |x-\mu|}{N}$$

PROBABILITY always b/w 0/1

Prob. is measure of how likely is something is to happen.

Trial:- Each performance in random trial.

Event:- outcome of trial P(E)

Sample Space:- Set of all possible outcomes in a random experiment P(S).

Random Experiment:- To perform more than once. (outcome can't be predicted)

Addⁿ Rule of Probability:- helps you find the prob that atleast one of two (or more) events happens.

used for 'OR' questions

CASE - A

Events are mutually exclusive

→ means they can not occur at the same time.

$$P(A \cup B) = P(A) + P(B)$$

CASE - B

Events are not mutually exclusive

→ means they can occur at the same time.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independent Events:- Two events are independent when:

→ the occurrence of one does not affect the occurrence of other.

Independence cond's for events A & B.

i) $P\left(\frac{A}{B}\right) = P(A)$

ii) $P\left(\frac{B}{A}\right) = P(B)$

iii) $P(A \& B) = P(A) * P(B)$

Cumulative Probability:- means the total prob. of all outcomes upto a certain point.

> It tells us the chance that a random variable is less than or equal to a certain value.

$P(a \leq X \leq b)$ a range is given

Conditional Probability:- means the chance of something happening, given that something else has already happened.

$P(B|A) = P(A \cap B) / P(A)$

rules of mult → $P(A \cap B) = P(A|B) \times P(B)$ or $P(A \cap B) = P(B|A) \times P(A)$

BAYES THEOREM

extension of conditional probability

cond probality help is : $P(A|B)$

Bayes theorem says, if we know $P(A|B)$, we can determine $P(B|A)$, given that $P(A)$ & $P(B)$ are known to us.

formula: $$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

Intution Behind Conditional Prob. formula

$P(A|B) = \frac{P(A \cap B)}{P(B)}$, is based on the concept of reducing our sample space.

B is occured & B is our new sample space, we don't consider sample where B is not occuring.

$P(A \cap B)$ → a occured & B occured.

BAYES THEOREM

$$P(A|B) = \underset{\text{posterior}}{P(A)} \times \frac{P(B|A)}{P(B)} \quad \begin{matrix} \text{prior} \\ \text{likelihood} \\ \text{marginal} \end{matrix}$$

BT → fundamental concept in field of prob. & statistics. that describes: how to update the probabilities of hypothesis when given evidence.

- Provides a way to revise existing predictions or theories (update prob.) given new evidence.

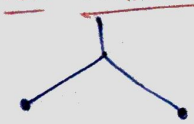
Probability Distribution :- Tell us how the values of data set are spread. and how likely each value is to occur.

> Describes the pattern of data & probabilities of all possible outcomes.

Types:-

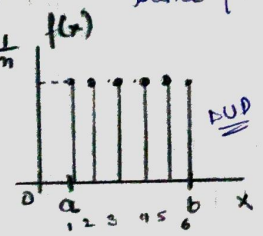
1) uniform Distribution.
2) binomial
3) Poisson
4) Normal / Gaussian
5) calculating prob. with Z-score normal Distribution.

uniform distribution :- all possible outcomes are likely to happen.

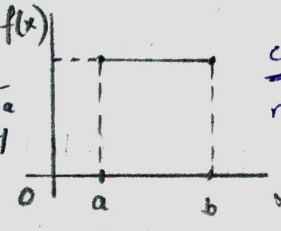


discrete uniform — each int. value b/w a & b has the same prob. to occur.

continous uniform → any values (int or decimal) b/w a & b has the same possibity to happen.



$P(2) = 1/6$
$P(3.4) = 0$
$P(X=x) = \frac{1}{n}$



$P(x) = \frac{1}{(b-a)}$ if $a < x < b$

a → lower limit, b → upper limit

binomial distribution :- tell the prob. of getting a certain num. of success in a fixed number of independent trials.

where, each trial has only two outcomes — success or failure.

short models, how many times success happens out of fixed num of tries.

> N-trials
> trials are independent
- P(success) = p
- P(failure) = 1 − p
- Random var X: total num. of success in 'n' trial.

$$P(x) = {}^{n}C_{x} p^{x} (1-p)^{n-x}$$

$${}^{n}C_{x} \cdot p^{x} (1-p)^{n-x}$$

Poisson Distribution:-

Tells you the prob. of a certain number of events happening in a fixed period of time - space.

In short:- how many times will something happen in a fixed time or area.

def: Dis. prob dist. of num of events occuring in given t-period, given avg no. time the event occurs over t period.

let $x \rightarrow$ random discrete var [num of events over a time]
ex $\rightarrow$ call/hr at call center.
$\lambda \rightarrow$ expected value (avg) of X, eg- avg. num. of call in call center.

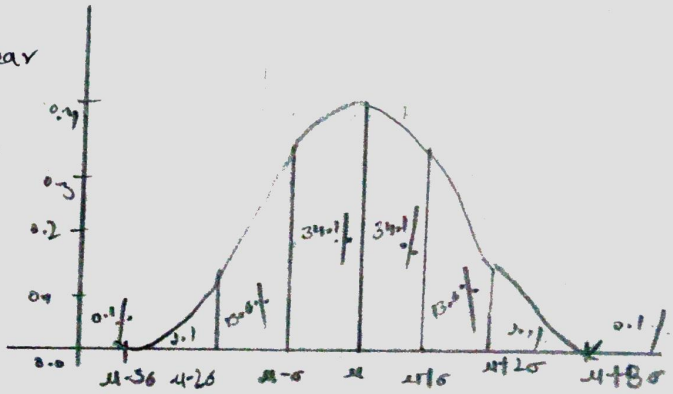
if X follows a possion dist. $P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$ [e = 2.718] maths constant. Euler's Num

normal/ distribution:- Gaussian

a prob. distribution that is symmetric about the mean. showing the data around the mean, are more frequent in occurance than data far from mean.

for all normal distribut's:
- 68.2% of Observat's will appear within +,- 1std of mean
- 95% of observat'n will fall within +- 2 std
- 99.7 with +- 3σ (std).

(0.3%) → outliers

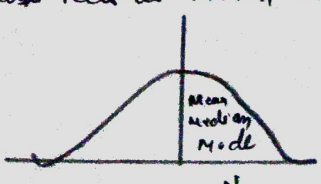


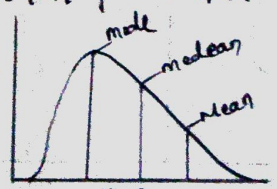
SKEWNESS

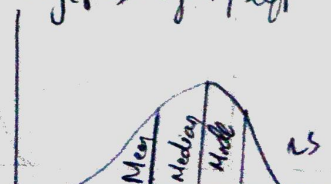
measures the distribut'n of data. Indicates whether data is distributed symmetric or not.
→ if symmetric → data is distributed normally.

→ Types → Symmetric → +ve Skewness → -ve skewness

skewness tells us which side the tail of the data is longer → right / left.







Transformation Techniques (T to Normal Dist)

- log Transfo
- sq root
- cube root
- square
- cube
- Reciprocal

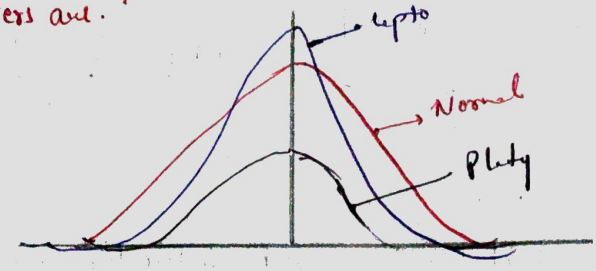
} used pandas / numpy

Kurtosis :- measures the thickness of the distribution of data.
- degree of tailedness of the data measured by Kurtosis.

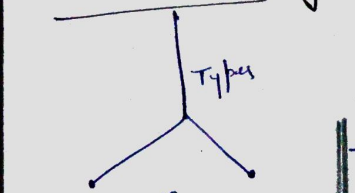
- Tells, extent to which the distribut" is more / less prone than normal dist.

Simply:- It measures how concentrated your data around the mean & how extreme your outliers are.

Types Platykurtic
Meso Kurtic
Lepto Kurtic



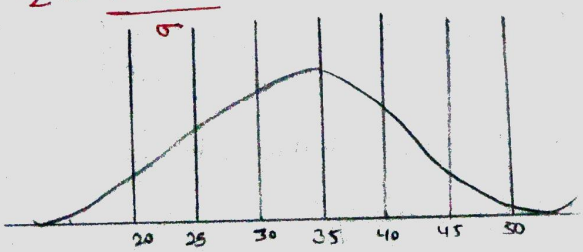
Feature Scaling:- Data preprocessing step where we bring all numerical features to a similar scale (range or distribut"). so that no single feature dominates other due to its magnitude.

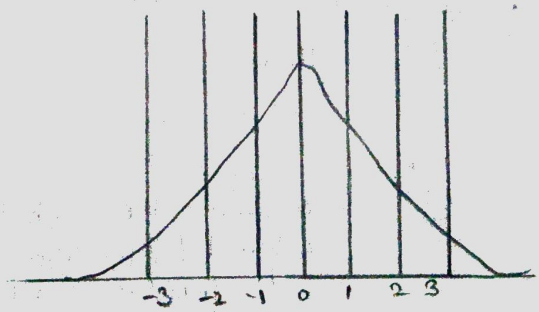


Calculating prob. with z-score for Normal Distribut". (Standardization)

i) find out how far the value of x is from μ. (interms of SD),
ii) The number named as "z-score", standard scores or z-value.

$$z = \frac{x - \mu}{\sigma}$$





Normal Distri (X)

i) $\mu - 3\sigma$: $z = \frac{x - \mu}{\sigma}$, $z = \frac{\mu - 3\sigma - \mu}{\sigma}$, $z = -3$

ii) $\mu - 2\sigma$: $z = \frac{\mu - 2\sigma - \mu}{\sigma}$, $z = \frac{-2\sigma}{\sigma}$, $z = -2$

iii) $\mu - \sigma$: $z = \frac{\mu - \sigma - \mu}{\sigma}$, $z = -1$

iv) μ: $z = \frac{\mu - \mu}{\sigma}$, $z = 0$

Standard Normal Dist (z)

v) $\mu + \sigma$: $z = \frac{\mu + \sigma - \mu}{\sigma}$, $z = 1$

vi) $\mu + 2\sigma$: $z = \frac{\mu + 2\sigma - \mu}{\sigma}$, $z = 2$

Normalization (Min-Max Scaling):-

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

When to use:
- need bounded values eg (neural networks)
- sensitive to outliers.

Covariance: statistical term that refers to a relat"nship b/w two random variables, on how one var. changes would impact the other one.

- covariance value can change from $-\infty$ to ∞.
- +ve co-var shows direct relat"nship & represented by +ve num.
  if one var. is ↑ then also other is ↑.
- -ve co-var shows inverse relat". A↑ → B↓
- value doesn't represent the magnitude of the relat"nship. only direct" matters.

$$\text{covariance}(x,y) = \frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

∴ for population

for sample, ∴ N-1 → n-1

where x_i → data of value x

y_i → ——— y

$\bar{x}$ → mean of x

$\bar{y}$ → — of y

N → number of data values

Correlation: Similar to covariance, also for measuring how to vars are related or moving together.

- ranges from +1 to -1.
- +ve correlat" → A↑ → B↑
- -ve correlat" → A↑ → B↓ (inverse)
- Both direct" & magnitude are useful., bigger num → strong relat".

$$\rho_{xy} = \text{correlat"}(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\,\sqrt{var(y)}}$$

($\sqrt{var(x)}$ = std(x), $\sqrt{var(y)}$ = std(y))

Correlat" value	Meaning
0 - 0.3	Weak linear relat"
0.3 - 0.7	Moderate L. relat"
0.7 - 1	Strong linear relat"

Hypothesis Testing:-

A statistical method that is used in making statistical decisions using experimental data.

- Hypothesis Testing is basically an assumption that we make about the population parameter. i) used to make decisions about populat" using sample Data"
  ii) Hyp-Test helps to decide, based on sample Data, whether claim about popu. is likely T/F?

Why we use it? → an essential procedure in stats.
It evaluates two mutually exclusive statements about a populat" to determine which statement is better supported by sample data.

Hypo. Testing

Null Hypothesis (H_0)

Its a general statement or default position that there is no relat" b/w two measured phenomenon among groups.
In other words its a basic assumption made based on domain or problem knowledge
→ assumes no diff / effect
ex- a company product" is 50 unit/day.

Alternative Hypothesis. (H_1, / H_a)

It the hypothesis used in hypothesis testing that is contrary to null hypo. usually taken to be that observat"s are the result of a real effect (with some chance of variat" supposed)
→ assumes some effect / some diff exists
ex- company's product" != 50 unit/day

Tailed Tests: tail → ends of prop. distribut" curve.
When we perfor Hypothesis testing, we check where the sample statistic falls. - in the left or right tail, or both tails.

Depending on the type of hypothesis, we use
- one tailed Test
- two tailed Test

One tailed Test: where, region of reject" is on only one side of sampling distribut".
eg → marks > 800

Two tailed Test: the critical area of distribut" is two sided. whether a sample is > or <
than a certain range of values.
eg → marks != 800

p-value :- p-value or calculated prob. of finding the observed, or more extreme, results when the null hypothesis (H_0) of study question is true. -

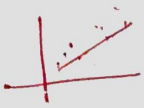
Simply: It measures the prob. of that your data happended by random chances.

Significance value(α) :- A threshold you choose before testing to decide how small the p-value must be to reject null hypoth.

$\alpha = 0.5$ (most common)

$\alpha = 0.01$ (stricter)

$\alpha = 0.10$ (less strict)

Data - Type	What we see	Type of test
1 categorical (c)		1-sample test
1-Numerical (n)		t test
1N + 1C		t test / ANOVA
2N		Correlation test
2C		Chi Square test.

z-test used to compare the mean of the two given sample & infer whether they are from the same distribution or not.

We do implement z test when the sample size less than 30.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$