* mkdir folder1 folder2 folder3
↳ This will creates three folder at once.

Pg: 01

Statistics for Machine Learning

Types of statistics

1) Descriptive statistics ⇒ helps us describe or summarize data in a meaningful way. It includes
   * Measure of central tendency (mean, Median, Mode)
   * Measures of dispersion (variance, standard deviation, Range, IQR)
   * Data visualizations (Histogram, Boxplots, scatterplots)

   (Summarizing past data)

2) Inferential statistics: It allows us to make prediction of generalization about a larger population based on sample data. It includes
   * Hypothesis testing
   * Confidence Intervals
   * Regression Analysis

   (Making prediction from sample data)

Sampling techniques

process of selecting a subset (sample) ~~Simple random sampling~~ from a larger group (population) to make conclusion about the whole ~~world~~.

Probability sampling (Random and unbiased)

1) simple random sampling: Each sample is chosen randomly from a university with 100 students
   eg: picking 100 students randomly from a university with 1000 students.

2) Stratified sampling: The population is divided into sub groups based on characteristics like age, gender, ~~ca~~ income etc and then randomly samples are taken from each group.

Eg:- If a company has 70% full time employees and 30% part time employes, you take accordingly to maintain the same ratio.

10/2

3) Systematic Sampling : Pick every k^{th} - individuals from a list after a random starting point.
Eg: Selecting every 10^{th} customer from a stores visitor-li

4) Cluster Sampling
The population is divided into groups (clusters) and Entire clusters are randomly selected.
Eg:- choosing 5 random schools in a city and surveying all students in those schools.

Non-Probalistic Sampling (Biased and non-Random)

5) Convenience sampling: Sampling people who are easily accessible.
Eg:- interviewing only people at a near by coffee shop instead of random people from the city.

6) Quota Sampling: similar to stratified sampling, but selection is not random - you choose based on convenience.
Eg:- Selecting 50 men and 50 women for a survey, but picking them without randomization.

7) Purporive (Judgemental sampling)
selecting individuals based on researchers judgement.

8) Snowball sampling
used when the population is hard to find.

Population VS Sample

Population: A population refers to the entire set of individuals. eg:- All students in india.

Sample: A sample is a smaller group selected from the population to represent it.

Variable: A variable is any characteristics, number, or quantity that can change or take different values in a dataset. It represent data that we measure or observe.

Example:-

student	Age	height	Gender	Marks (%)
A				
B				
C				

here the colums age, height, gender, marks all are variables.

Types of variables

a) Based on data type

1) Quantitative variables: Variables that represent measurable numerical values. (Numerical) student in a class.
   - ↳ Discrete Variables (countable), eg:- marks, no. of
   - ↳ continuous variables (infinite values within range) eg:- height, weight, temprature.

2) Qualitative values (Categorical): Variables that represent categories or group. (not members)
   - ↳ Nominal variables (no natural order) eg:- M/F, blood grp
   - ↳ Ordinal variables (have a meaningful order) eg:- Highschool < Bachelor < Masters

b) Based on Role analysis

1) Independent Variable (Predictor): A variable that causes of influence changes in another variable.

eg:- In a study on weight loss, diet and exercise are independent Variable

Dependent Variable: A variable that depends on a weight loss, diet and exercise influence change in another variable depends on or is affected by the independent variables. $y = mx + c \Rightarrow y$ = dependent variable (04)

Variable measurement scale

A variable measurement scale defines how data is categorized, order and mathematically processed. It determines the type of statistical analysis we can perform.

Types

1) Nominal Scale :- (Categorical, no order)

→ Represent categories without a natural order.

→ No mathematical operation (like addition and subtraction) can be performed.

→ only counting and grouping are meaningful.

Eg:- Gender (Male, Female), Blood Type (A, B, AB, O), Country (India, USA, Germany)

Allowed operation :- Counting, Mode, Frequency.

Eg:- Number of students in India.

2) Ordinal Scale (Categorical, ordered)

→ Represents categories with a meaningful order, but the difference between them are unknown.

Eg:- Education level (high school < Bachelors, < Masters)
customer satisfaction (Poor < average < good < excellent)
Military Ranks (Private < Sergant < Captain)

→ we rank the values but can't perform precise arithmatic.

Allowed operation : Counting, Mode, Median, Ranking.

Example Question : what is the most common customer satisfaction level.

3) Interval scale :

→ Represents numerical values with equal interval b/w them.

→ Addition and subtraction are meaningful, but ratios are not.
→ No True Zero (Zero doesn't mean nothing)

05

Examples

Temprature (0°C or °F) → (0°C ≠ no temprature, it's just a refernce point)
IQ Score → An IQ of 100 is not "twice as intelligent as" 50)
Year (2020, 2021, 2022) → There's no "0 year zero".

Allowed operations: Addition, subtraction, mean, median, mode.
Example questions: "what is the average temprature in january?

4) Ratio Scale (Numerical, True zero exists)
→ Like the interval scale, but with a true zero (zero means the absence of quantity)
→ All mathematical operation (addition, subtraction, multiplication, division are possible.

Example:
1) hight (0 cm means to height) and same with others like weight, salary, distance etc.

Allowed operation: All mathematical operation (Mean, median, mode, Ratios, Percentages
Example questions: A person earning 50,000 earns twice as much as someone earning $5,000

Measure of Central Tendency

It is used to describe the center of data set.

1) Mean (Average)
The mean is the sum of all values divided by the total number of values.

$$\text{Mean} = \frac{\Sigma x}{N}$$

where x = individual values
N = total number of values.

eg:- age of of Student = 18, 20, 22, 24, 26

$$\text{Mean} = \frac{18+20+22+24+26}{5} = 22$$

2) Median (Middle value)

The median is the middle value when the data is arranged in arranding order.

if N is odd → median is the middle value.

if N is even → the median is the average of the two middle values.

eg:- 18, 20, 22, 24, 26 ⇒ median is 22

eg: 18, 20, 22, 24 ⇒ median is $(20+22)/2 = 21$

3) Mode (Most Frequent element)

The mode is the most frequently occuring value in a data set.

eg:- Test scores 85, 90, 92, 85, 88, 85, 91, 92, 92

→ 85 appears 3 times

→ 92 appears 3 times

→ other number appear only once.

Mode = 85 and 92 (Bimodal dataset)

Measure of Dispersion in statistics

Measures of dispersion describe how spread out or scattered data values are in a dataset. They help us understand variability in the data.

1) Range:- It is the difference between maximum and minimum values in the dataset.

$$\text{Range} = \text{Max}(x) - \text{Min}(x)$$

#2 Variance : Variance tells how far each value is from the mean, on average.

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

x_i = individual data point

μ = mean

N = number of values.

$$\because \mu = \frac{\Sigma x_i}{N}$$

1): deviation from the mean is $x_i - \mu$

2) squaring the deviation $(x_i - \mu)^2$ ⇒ this will ensure all deviations are positive.

$$\therefore \text{Mean} = \frac{\Sigma (x_i - \mu)^2}{N}$$

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

#3 Standard deviation: It is a key concept in statistics that measure how much data points deviate from the mean.

→ It tells us how spread out the data is from the average.

→ A small standard deviation means data is closer to the mean (less spread)

→ A large standard deviation means data is widely spread.

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

Intuition behind Standard deviation

Imagine two classes taking a math test

Student	class A score	class B scores
1	85	60
2	87	90
3	88	55
4	86	95
5	84	90

Class A mean = $\mu_A = 86$
" B " = $\mu_B = 70$

Class A Variance = $\sigma_A^2 = 2$
" B " = $\sigma_B^2 = 350$

S.D of class A = $\sigma_A \approx 1.41$
" " " B = $\sigma_B \approx 18.71$

Interpretation

For class A

→ Scores are closely packed around 86
→ students performed similarly with less variation in marks.

For class B

→ Scores are widely spread from the mean (70)
→ Some students scored very high, while other scored very low.

Larger S.D ⇒ more spread out data.
Smaller S.D ⇒ more consistent data.

#4) Interquartile Range (IQR)

Interquartile range is the measure of how spread out the middle 50% of dataset is. It helps to find the range of values that are not affected by extreme outliers.

Example

Score of 10 student ⇒ 10, 20, 25, 30, 35, 40, 45, 50, 55, 90

Step 1 ⇒ In ascending order

Step 2 ⇒ Find Q1 (1st quartile) → This is the 25th percentile (middle of the lower half)

→ Lower half : [10, 20, 25, 30, 35]
→ Q1 = 25 (median of lower half)

Step 3: Find Q3 (3rd quartile) → This is the 75th percentile (middle of the upper half):

→ upper half [40, 45, 50, 59, 90]

→ Q3 = 50 (median of the upper half)

Step 4: Compute IQR

$$IQR = Q3 - Q1 = 50 - 25 = 25$$

This means the middle 50% of scores lie between 25 and 50, ignoring the extreme values like 10 and 90

Probability distributions

1) Normal or (Gaussian distribution)

It is a continuous prob. distribution that describes data that clusters around a mean (average) value. The normal distribution is defined by its probability density function (PDF), which gives probability density at a particular value x.

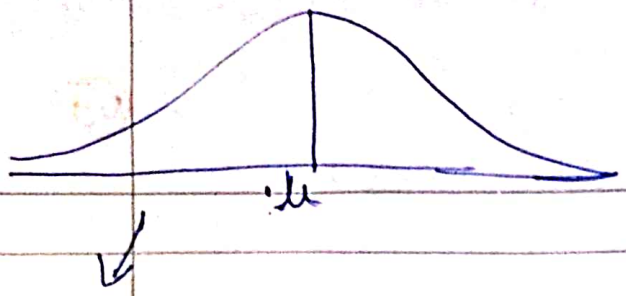
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

∵ the normal distribution is continuous, the probability of any single exact value is zero, Instead we look at the probability of x being within a small range dx, which is given by

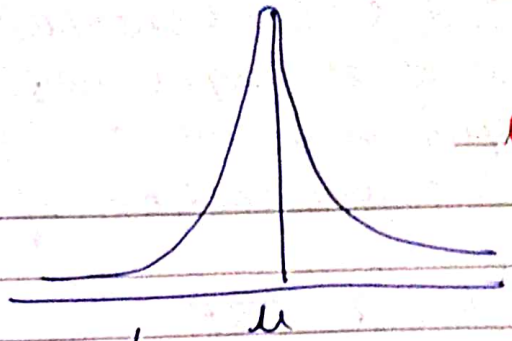
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Higher f(x) value → x is more likely to occur
Lower f(x) value → x is less likely to occur.
Peak at x = μ ⇒ The most frequent or most likely value in the mean.

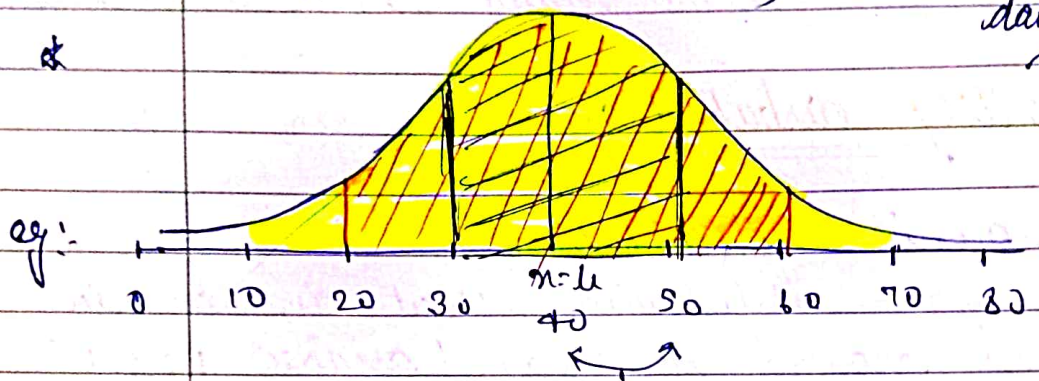
This function integration at any x=d will give an approximate probability that x=d falls within the range a, b.

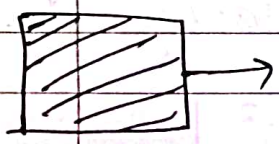


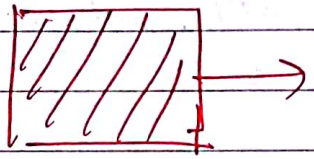
high standard deviation (means maximum data are more scattered)

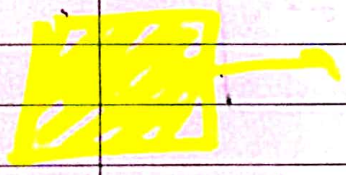


low standard deviation.
↓
means maximum data lies near mean.




→ 65 %.


→ 95 %.


→ 99 %.

→ Bell shaped curve
→ Mean = Median = Mode
→ 1 S.D covers → 65 %, 2 S.D covers → 95 %, 3 S.D → 99.7 %.
→ Total probability = 1
→ Asymptotic behaviour

Proof →

1) $f(x) \geq 0$ (No negative condition) $\forall x \in (-\infty, +\infty)$

2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Step 1 General form of the gaussian function

$$f(x) = Ae^{-B(x-\mu)^2}$$

here A and B are constants

$$\therefore \int_{-\infty}^{\infty} A e^{-B(x-\mu)^2} dx = 1$$

on comparing we get

As we know, $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}, \quad a > 0$

here we set $a = \frac{1}{2\sigma^2}$

Standard gaussian integral formula.

here we set $a = \frac{1}{2\sigma^2}$ so that normal distribution inherently follows a bell curve and to normalize it (make prob. = 1), we need to solve the integral $\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$, by doing so we transform the equation into a standard gaussian integral which has known solution.

A, B

$A = \frac{1}{\sigma\sqrt{2\pi}}, \quad B = \frac{1}{2\sigma^2}$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

2) Uniform distribution

Type of prob distribution where all outcomes in a given range are equally likely. This means that every value within the specified range has the same probability of occurring

The PDF for a continuous uniform distribution in the range $[a, b]$ is given by:

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

a = lower bound
b = upper bound.

$f(x)$ = constant probability, the PMF

$$P(X = x) = \frac{1}{n}, \quad x \in \{x_1, x_2, \ldots x_n\}$$

3) ~~Poisson~~ Binomial Distribution

Binomial distribution is perform when
i) All the trails are independent
ii) Number n of trails are finite.
iii) The prob. p of success is same of each trail.

The prob. of x success in n trails taken in any order is given by addition theorem of prob.

as $P(x) = {}^nC_x p^x q^{n-x}$ (PMF)

where p = probability of success

q = probability of failure.

① $P(x) \geq 0$ } condition

② $\Sigma P(x) = 1$

* mean = $E(x) = np$

* Variance = $Var(X) = np(1-p)$

Proof

$$E(x) = \Sigma x P(x)$$

$$= \Sigma x \, {}^nC_x p^x q^{n-x}$$

$$= \Sigma x \times \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \Sigma \frac{p\, n(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{(x-1)} q^{((n-1)-(x-1))}$$

$$= np \left[\Sigma \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{(x-1)} q^{((n-1)-(x-1))} \right]$$

This equals to 1

$\therefore$ $E(x) = np$

hence $Var(X) = E(x^2) - (E(x))^2 = np(1-p)$

Descrete random variable

A variable which takes finite or at most countable number of values called discrete random variable

Eg:- no of head obtain when two coins are tossed.

continuous random variable

A random variable which can take infinite number of values in an interval is known as C.R.V.

Probability density fn

A fn $f(x)$ is called P.D.F if

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \quad \text{and} \quad f(x) \geq 0$$

Mathematical expectation (Also called as mean)

also called the value $(E(x))$ is the average value you would expect from a random variable over a large number of trials. It represents the long term average of an experiment.

Definition

For a descrik random variable x with perible values $x_1, x_2, \ldots x_n$ and their corrurponding probabilities $P(X = x_1), P(X = x_2) \ldots P(X = x_n)$, the expected value is given by

$$\mu = \boxed{E(X) = \sum x_i P(X = x_i)} \quad \text{or mean}$$

For continuous random, the expekd value is given by

$$\mu = \boxed{E(X) = \int_{-\infty}^{+\infty} x f(x)\,dx} \quad \text{or mean.}$$

eg:- Think of the expekd value as the weighed average of all possible values, where each value is weighed by its probability.

$$E(X) = 1 \cdot \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} \times 6 \times \frac{1}{6}$$

$$= 3.5$$

This means that if we roll the dice many times, the average result will approach 3.5

$$\boxed{\underbrace{E(X)}_{\text{used in theoritical term}} = \underbrace{\mu}_{\text{used in statistical term}}}$$

Now we know what is mean (μ) i.e,

$$\mu = E(X)$$

So, we will now find formula for Variance.

VARIANCE

The formula is:

how?

$$Var(X) = E(X^2) - (E(X))^2$$

As we know, variance measure how spread out the values of a random variable X are around the mean $\mu = E(X)$

∴ From upper definition we have

$$Var(X) = E[(X-\mu)^2]$$

$$\Rightarrow Var(X) = E[X^2 + \mu^2 - 2X\mu]$$

using linear of expectation, we can distribute E over addition/subtraction, we have,

$$E[(X-\mu)^2] = E[X^2] + E[\mu^2] - E[2\mu X]$$

Now, As we know $E[C] = C$ (where C is constant)

$$\therefore E[(X-\mu)^2] = E[X^2] + \mu^2 - 2\mu E[X]$$ ($E[X] \to \mu$)

$$= E[X^2] + \mu^2 - 2\mu^2$$
$$= E[X^2] - \mu^2$$
$$= E[X^2] - (E[X])^2$$

$$\therefore \boxed{Var(X) = E[X^2] - (E[X])^2}$$

Proof of probability density function of Normal distribution

1) As we know, the most important result in prob. and statistics is the Gaussian integral

Very imp.

$$\int_{-\infty}^{+\infty} e^{-ax^2}\, dx = \sqrt{\frac{\pi}{a}}, \quad \text{for } a > 0 \qquad \text{—(i)}$$

2) Now the probability density fn of a normal distribution is

$$f(x) = A e^{-B(x-\mu)^2} \qquad \text{—(ii)}$$

A ⇒ normalization constant (to make sure that the total prob. sum to 1)

B ⇒ is the coffiecient inside the exponent, controlling the spread of distribution.

Now, we know that $E[x] = \mu$, so we will find $E[x^2]$ so ~~compute~~ compute Variance.

$$Var(x) = E[x^2] - (E(x))^2$$

$$E(x) = \int_{-\infty}^{+\infty} x f(x)\, dx$$

$$\therefore E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx$$

Before solving this we first understand the normalization of the fn.

from eq (i) we know any equation of the form $\int_{-\infty}^{+\infty} e^{-ax^2} dx$ the value is $\sqrt{\frac{\pi}{a}}$

$$\therefore \int_{-\infty}^{+\infty} A e^{-B(x-\mu)^2} dx = 1 \quad \text{---(iii)}$$

/ (1b)

of $f(x) = 1$ (for probability distribution function)

on comparing eq.(i) and (iii) we get

$$A \cdot \sqrt{\frac{\pi}{B}} = 1 \quad \text{---(iv)}$$

Now,

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx$$

$$\Rightarrow E(X^2) = \int_{-\infty}^{+\infty} x^2 A e^{-B(x-\mu)^2} dx$$

Let $x - \mu = y$

$\therefore x = y + \mu \quad \therefore dx = dy$

$$\Rightarrow E(X^2) = \int_{-\infty}^{+\infty} (y+\mu)^2 \cdot A \cdot e^{-By^2} dy$$

$$\Rightarrow E(X^2) = \int_{-\infty}^{+\infty} (y^2 + \mu^2 + 2y\mu) \cdot A \cdot e^{-By^2} dy$$

$$\Rightarrow E(X^2) = A\int_{-\infty}^{+\infty} y^2 e^{-By^2} dy + \underbrace{2\mu A \int_{-\infty}^{+\infty} y e^{-By^2} dy}_{\text{becomes 0 since it is a odd fn.}} + A\mu^2 \underbrace{\int e^{-By^2} dy}_{\sqrt{\frac{\pi}{B}} \text{ from eq(i)}}$$

$$\Rightarrow E(X^2) = A\int_{-\infty}^{\infty} y^2 e^{-By^2} dy + \underbrace{A\mu^2\sqrt{\frac{\pi}{B}}}_{\mu^2 \text{ (from eq iv)}}$$

$$\therefore \rightarrow E(x^2) = \underbrace{A \int_{-\infty}^{+\infty} y^2 e^{-By^2} dy}_{\text{use Integral by part and use ILATE i.e, 1st fn is } y^2 \text{ and second fn is } e^{-By^2}} + \mu^2$$

$$\therefore E(x^2) = A \times \frac{1}{2B}\sqrt{\frac{\pi}{B}} + \mu^2$$

$$\boxed{E(x^2) = \frac{1}{2B} + \mu^2} \quad \text{---- (v)}$$

Now, Variance $= \sigma^2 = E(x^2) - (E(x))^2$

$$\sigma^2 = \frac{1}{2B} + \mu^2 - \mu^2$$

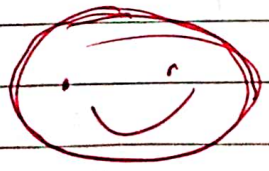
$$\boxed{B = \frac{1}{2\sigma^2}} \rightarrow \text{(vi)}$$

$$A = \sqrt{\frac{B}{\pi}} = \sqrt{\frac{1}{2\sigma^2\pi}}$$

$$\boxed{A = \frac{1}{\sigma\sqrt{2\pi}}}$$

Hence $f(x) = A e^{-B(x-\mu)^2}$

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \Rightarrow \text{PDF}$$



and $\int_{-\infty}^{+\infty} f(x) = 1$

Poisson distribution : is a discrete prob. distribution that expresses the prob. of a given number of events occurring in a fixed interval of time or space, / given that these events occur with a known constant mean rate and are independent of the time since the last event. eg:- number of calls received at a call center per hour.

PMF ⇒ $$P(X=K) = \frac{e^{-\lambda}\lambda^K}{K!}, \quad K = 0, 1, 2, 3, \ldots$$

K ⇒ number of occurance
λ ⇒ (mean rate) expected number of occurance in the given interval.
e ⇒ eulers number.

It is the limiting case of binomial distribution as n is very large and probability p is very small

$$P(X=K) = {}^{n}C_K \, p^K (1-p)^{n-K}$$

$$\lim_{n\to\infty} (1-p)^n = e^{-np}$$

see in limits theory

$$= \frac{n!}{K!\,(n-K)!} p^K e^{-np}$$

$$= \frac{n(n-1)(n-2)\ldots(n-(K-1))\,(n-K)!}{K!\,(n-K)!} p^K e^{-np}$$

$$= \frac{n^K}{K!} p^K e^{-np}$$

Let $np = \lambda$

$$p = \frac{\lambda}{n}$$

$$\therefore \frac{n^K}{K!} \frac{\lambda^n}{n^K} e^{-n\lambda/n}$$

$$\therefore \quad P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Exponential distribution is a continuous prob. distribution used to model the time between events in a poisson process. It is process in which events happens continuously and independently at a constant average rate.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

λ is called the distribution rate.

$$\text{Mean} = E(x) = \int_{-\infty}^{\infty} x f(x)$$

$$= \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{\infty} x f(x)\,dx$$

$$= \lambda\left[\left|-\frac{x e^{-\lambda x}}{\lambda}\right|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\,dx\right]$$

$$= \lambda\left[0 + \frac{1}{\lambda}\cdot\frac{-e^{-\lambda x}}{\lambda}\right]_0^\infty = \boxed{\frac{1}{\lambda}}$$

$$\text{Var} = E(x^2) - (E(x))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \boxed{\frac{1}{\lambda^2}}$$

Normalization and standardization

FEATURE SCALING

Standardization :- Suppose we have two features in an ML model : age (20-80 years) and income ($10,000 - $ 500,

→ here income is way larger than age. If we don't standardize the model focus too much on income.

So, we use $Z = \frac{X - \mu}{\sigma}$ → Original value, → S.D.

↓

It is the standardize value

eg :-

20	10000
40	50000
60	200000
80	500000

After →

-1.34	-0.93
-0.44	-0.72
0.44	0.05
0.34	1.61

centered around 0

This is called standardize normal distribution

Now, the question is why we use $Z = \frac{X - \mu}{\sigma}$

↓ also called z score

there is two reason behind this

(1) Standardization shifts the mean to 0

∴ New Mean $= \frac{1}{n} \sum Z_i$

$= \frac{1}{n} \sum \frac{X_i - \mu}{\sigma}$

$= \frac{1}{n\sigma} \sum X_i - \mu \rightarrow 0$

$= 0$

hence New Mean = 0.

Now, there can be a question that new mean can be zero without the use of σ. This will be explained in the 2nd point

(2) Standarization scales the standard deviation to 1.

New standard deviation $= \sqrt{\frac{1}{n}\Sigma(z_i - 0)^2}$. 21

as. S.D: $\sqrt{\frac{1}{n}\Sigma(x_i - u)^2}$

New S.D: $\sqrt{\frac{1}{n}\Sigma\left(\frac{x_i - u}{\sigma}\right)^2}$

$= \sqrt{\frac{1}{n\sigma^2}\Sigma(x_i - u)^2}$

Now As we know $\sigma^2 = \sqrt{\frac{\Sigma(x_i - u)^2}{n}}$

$\therefore \sigma^2 = \frac{\Sigma(x_i - u)^2}{n}$

$\therefore$ New S.D $= \sqrt{\frac{1}{\sigma^2} \times \sigma^2} = 1$

$\therefore$ New S.D = 1

This ensures all values have a uniform spred, making it easier in ML models to process data efficiently.

Normalization

Normalization scales data between a fixed range typically [0, 1] or [-1, 1]. It don't change the shape of the data like standarization but rather compresses it into a fixed range.

→ Some ML models (like NN, KNN) work better when all input features are within similar range.

→ Normalization ensures no features dominates due to differences in scale.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$ → Min max scaler

X = original data point
X_{min} = min value of data set
X_{max} = max " " " "
X' = Normalized value (b/w 0 and 1)

eg:

Person	Income (₹)	Age
A	10,000	20
B	50,000	40
C	200,000	60
D	500,000	80

Income range between 0 and 1 and comes

Normalized Income
0.00
0.21
0.63
1.00

Normalized Income
0.00
0.33
0.67
1.00

~~Standard normal distribution~~ normalized.

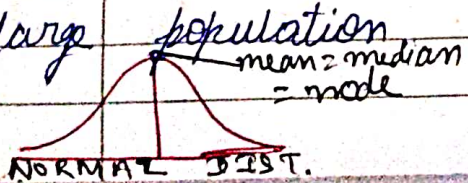
Skewness and Kurtosis

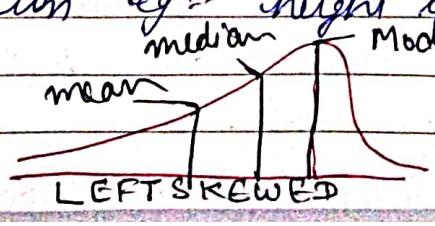
Skewness tells us whether a distribution is symmetrical or asymmetrical around the mean.

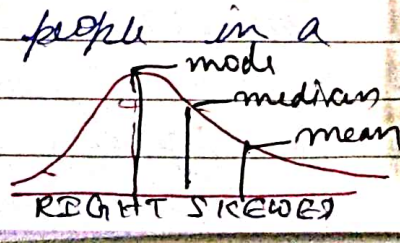
→ Positive Skew :- The right tail is longer, meaning most data points are concentrated on the left. eg :- income distribution, where a few people are significantly more than the majority.

→ Negative Skew :- The left tail is longer, meaning most data points are concentrated on the right. Eg :- Exam scores where most students score high, but a few score very low.

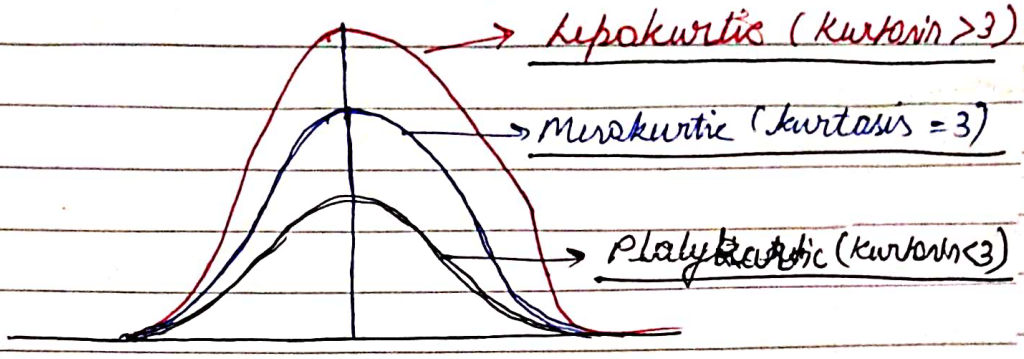
→ Zero Skewness (symmetrical distribution) → The distribution is perfectly balanced on both side of the mean, like normal distribution eg :- heights of people in a large population







Kurtosis: tells us how much of the data is in the tails of the distribution compared to a normal distribution.



Law of large number.

It states that if we repeat an experiment many times, the average of the results will get closer and closer to the true expected value.

Eg:-

Imagine if we flip a coin. The probability of getting head is 50% (0.5). But if we flip the coin just few times (let's say 10 times), you might get heads 7 times or only 3 times. This doesn't match the true prob. Now, imagine flipping the coin 1000 times, the percentage of head will likely be closure to 50%. If you increase the flips to 10,000 times, the percentage will even closure to 50%. This happens because as the number of trails increases, the observed average gets closure to the true probability.

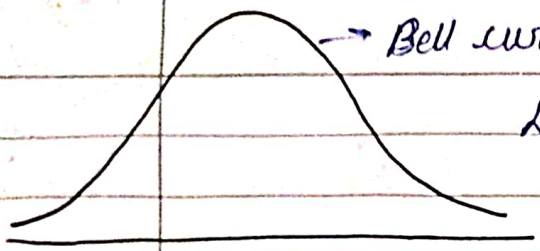
Central limit theorem

That distribution can be any distribution

It states that the distribution of a sample mean will approximately a normal distribution as the sample size becomes larger, regardless of the population actual distribution shape.

→ There must be large number of sample data (>30)
→ average of the sample means and S.D will equal the population mean and S.D
→ used in finance, investings etc.

* practical explanation of central limit theorem



$x = \{65, 72, 83, 55, 64, 67 \ldots\}$ (24)
n

Let say we take different samples randomly.

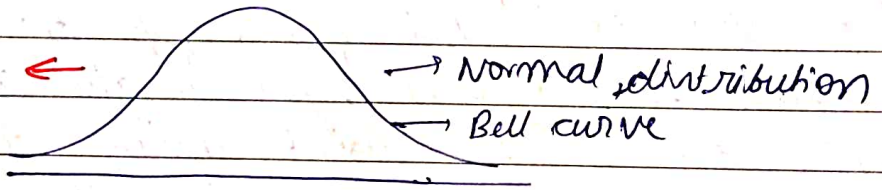
$S_1 = \{x_1, x_2, x_3 \ldots\} \rightarrow \bar{x}_1$,
$S_2 = \{x_1, x_2, x_3 \ldots\} \rightarrow \bar{x}_2$
⋮
$S_n = \{x_1, x_2, x_3, \ldots\} \rightarrow \bar{x}_n$

Now, we have collection of sample means as

$\bar{S} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_n\}$

Now, if we plot this sample mean in histogram then it will form normal distribution

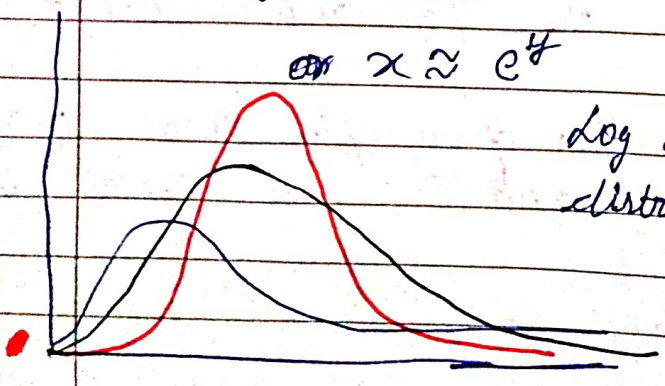
distribution of sample mean ←



Log Normal distribution

A log normal distribution is a probability distribution of a random variable whose logarithm is normally distributed

If $x \approx$ log normal distribution
then
$y \approx \ln(x) \Rightarrow$ Normal distribution,
or $x \approx e^y$



Log normal is a right skewed distribution

Bernoulli Distribution

25

Z - Score

Z-score tells us how far a data is from the mean of the data set, in terms of standard deviation. It helps us to understand whether a value is above or below the mean and by how much.

- Uses

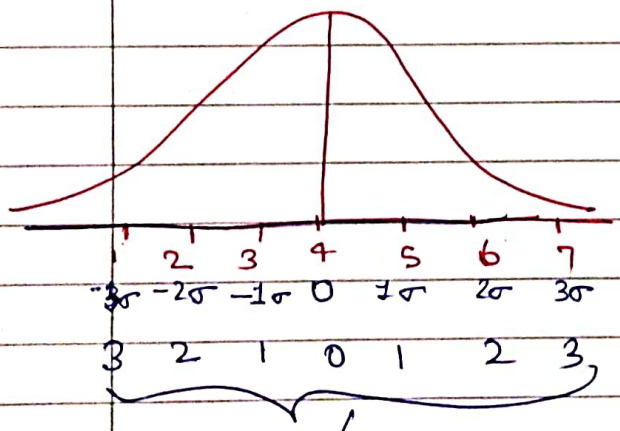
→ It standarizes different datasets, making them comparable.
→ It identifies outliers (extreme value in data)
→ It is used in probability calculations (eg. normal).

$$Z = \frac{X - \mu}{\sigma}$$

μ → The mean
σ → S.D
X → The data point

Intuition: Number of S.D our data is away from the mean.

eg:-



Here $\mu = 4$
$\sigma = 1$

$\therefore Z(1) = \frac{1-4}{1} = -3$

$Z(6) = \frac{6-4}{1} = 2$

$Z(4) = \frac{4-4}{1} = 0$

Standard normal distribution

Percentiles and Quartiles

Percentage : 1, 2, 3, 4, 5

% of the numbers that are odd?

$$\% = \frac{\text{no. of numbers that are odd}}{\text{Total numbers}}$$

$$= \frac{3}{5} = 0.6 = 60\%$$

Percentiles : A percentile is a value below which a certain percentage of observation lie.

Dataset: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9 / 27

① What is the percentile ranking of 10

∴ Percentile ranking of 10 is = $\frac{\text{no. of values below 10}}{n} \times 100$

n ↓ total number of values.

$= \frac{16}{20} \times 100 = 80$ percentile.

Eg: Imagine 100 people took a test, if you are in the
→ 90th percentile → You did better than 90% people.
→ 50 " " → " " " " " 50% "

② What value exists at percentile ranking of 25%?

$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times 21 = 5.25$ → so take average of value at index 5 and 6. which will comes out to be 5.

Five number Summary

1) Minimum (5) Maximum.
2) First Quartile (Q_1)
3) Median
4) Third Quartile

Removing the outlier

outlier

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

Lowerfence ⟷ higher fence

all numbers before it are outlier

all numbers after it are outlier.

Lowerfence $= Q_1 - 1.5(IQR)$
upperfence $= Q_3 + 1.5(IQR)$
and $IQR = Q_3 - Q_1$

Q_3 = third quartile or 75 percentile

$\therefore$ Value $= \frac{75}{100} \times 20 = 15 \rightarrow$ index i.e, 7

Q_1 = first quartile or 25 percentile

$\therefore$ Value $= \frac{25}{100} \times 20 = 5 \rightarrow$ index 3

$\therefore IQR = 7 - 3 = 4.$

$\therefore$ Lowerfence $= 3 - 1.5 \times 4 = 3 - 6 = -3$
upperfence $= 7 + 1.5 \times 4 = 13$

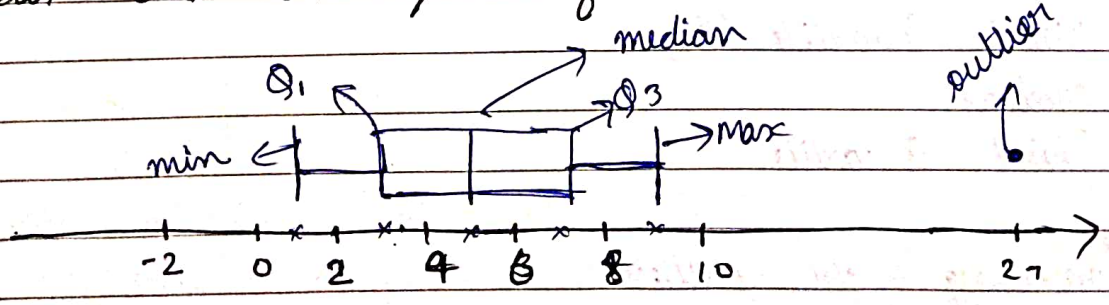
$\therefore$ value above 13 are outlier and value less than -3 are outlier.

$\therefore$ Remaining data is 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9.

$\therefore$ 1) minimum = 1
2) First quartile is $= Q_1 = 3$
3) Median is $\frac{5+5}{2} = 5$

4) third quartile is $Q_3 = 7$
5) maximum is 9

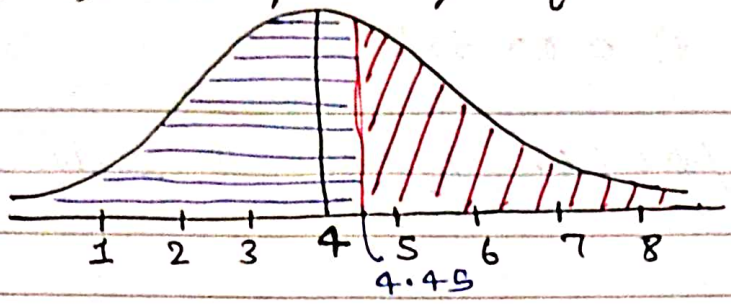
Now we plot the box plot for this.



Box plot

Use of Z-table

Eg) What is the percentage of score falls above 4.45?



⇒ So basically we need to find the area of the red side.

Step 1: Calculate the Z score:

The formula for z score is

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

So, $z = \frac{4.45 - 4}{1} = 0.45$

Step 2: A z table gives the area (probability) to the left of the z-score in a standard normal distribution. Here for $z = 0.45$ it gives the area of the blue part.

According to the ~~left~~ left z table

$$P(Z < 0.45) \approx 0.6736$$

Step 3: Find the percentage above the z-score

Since we want the percentage of score above 4.45, we need:

$$P(Z > 0.45) = 1 - P(Z < 0.45) = 1 - 0.6736 = 0.3264$$

Hence 32.64% of score falls above 4.5 in this normal distribution

Q) In india the avg. IQ is 100, with a standard deviation of 15. What percentage of population would you expect to have an IQ lower than 85?

$\Rightarrow$ Z score = $\frac{85-100}{15} = -1$

for $P(Z = -1) \approx 0.1587$ /30

$\therefore$ percentage of population having IQ lower than 85 is 15.87 %.

~~## Z score~~

Inferential Statistics

Hypothesis testing

It is the statistical method used to make decision or inferences about a population based on sample data. It is commonly used to test claims or ideas (hypothesis) about a population parameter.

In brief

hypothesis is like checking a claims using data.

Imagine, you are a detective and someone says "The coin is not fair! It gives more head than tails.

Now you want to test the claim. You flip the the coin 100 times and record the results. That's hypothesis Testing using data to check if a claim is likely true or false.

* The two hypothesis

In every hypothesis test, we start with two opp. ideas:

i) Null hypothesis (H_0): The default belief. "The coin is fair",

ii) Alternate hypothesis (H_1): The claim we want to check. "The coin is biased (not fair)"

our job is to use our data to decide:
→ Do I have enough evidence to reject the null hypothesis.
→ or should I stick with the null hypothesis.

Simple example:
Let's say we flip a coin 100 times and get 60 heads.
→ If the coin is really fair, you would expect about 50 heads, right?
→ But if we got 60 heads, so is that just luck or is the coin actually biased?

Now comes the P-values to help us decide

What is P-value
The P-Value is a key concept in hypothesis testing it's like a measure of how surprising your data is if the null hypothesis is true.

The P-value is the answer to the question:
"If the coin were truly fair, what is the chance I would see something as extreme as 60 heads (or more)?".

In simple terms:-
→ A small P-value (like 0.001 or 0.03) means: wow! getting 60 heads is pretty rare if the coin is fair,(";
so may be the coin isn't fair.
This means our observed data is quite unlikely if the null hypothesis were true.

→ A big P-value (like 0.4 or 0.6) means: "getting 60 heads is not that weird", so we can't say the coin is biased.
This means our observed data is reasonably likely to occur even if the null hypothesis were true.

We usually compare the P-value to a threshold called alpha (α), often 0.05 (5%)

→ If P-value < 0.05 → Reject the null hypothesis
→ If P-value > 0.05 → Do not reject the null hypothesis.

↳ alpha (α) or significance value.

Coin example

Let say our P value comes out to be 0.03 (3%)
This means:
"There's only a 3% chance I would get 60 or more heads just by luck if the coin were fair."
Since 0.03 < 0.05, we reject the null hypothesis and say: "I have enough evidence to believe the coin might be biased".

Calculation of P-value

"To calculate the P-value, I need a z-score, and to get that, I need the correct mean and standard deviation. But how do I know which distribution to use, and which formulas go with it?"

Answer:- Identify the type of data or situation. The type of problem will tell which probability we should use.

1) Problem type 1: Flipping coins / success-failure (fixed trials).
→ Distribution used ⇒ Binomial
→ Mean formula ⇒ $\mu = np$
→ S.D formula ⇒ $\sigma = \sqrt{n \times p \times (1-p)}$

2) Problem type 2 : Rare events per time or space (eg. calls/min, bacteria/cm²)
→ Distribution used ⇒ Poisson
→ Mean formula ⇒ $\mu = \lambda$
→ S.D formula ⇒ $\sigma = \sqrt{\lambda}$

~~3) Measurements like height~~

3) Problem type 3 : Measurements like height, weight, test scores (continuous data)
→ Distribution used ⇒ Normal
→ Mean formula ⇒ Based on data
→ S.D formula ⇒ Based on data.

4) Problem type 4 :
~~4)~~ Averages from samples (sample mean)
→ Distribution used ⇒ Normal (sampling distribution)
→ Mean formula ⇒ μ = population mean.
→ S.D formula ⇒ $\sigma = \sigma/\sqrt{n}$

5) Problem type 5 :
~~5)~~ Comparing two means or small samples (with unknown σ)
→ Distribution used ⇒ t - distribution.
→ Mean formula ⇒ depends on data.
→ S.D formula ⇒ depends on data.

6) Problem type 6 : comparing categories (like observed vs expected).
→ distribution used ⇒ chi - square
→ Mean formula ⇒ NA
→ S.D formula ⇒ NA.

Since here we are counting how many times we get head out of 100 flips, and each flips has two outcomes (head or tail), we use binomial distribution.

But when the flip is large (like 100), we often approximate it using a normal distribution. (because it is easier to calculate).

Now, we will calculate the mean and S.D.

mean = expected no. of heads = $n \times p$

$= 100 \times 0.5 = 50$

$S.D = \sqrt{n p(1-p)} = 5$

Calculate the Z-score so as to know how far is our data is from the mean of the data set in terms of S.D.

$$\therefore Z = \frac{\text{My value} - \text{Mean}}{S.D}$$

$$= \frac{60 - 50}{5} = 2$$

Now, we check: what is the prob. of getting a Z score this extreme or more?

From Z-table (standard normal distribution) the probability of getting a $Z \geq 2$ is about 0.0228.

Since we are doing a two-tailed test (because we are checking for bias in either direction, more heads or more tails), we double this

$$P\text{-Value} = 2 \times 0.0228 = 0.0456$$

Final decision

→ P-value = 0.0456

→ alpha = 0.05 (alpha)

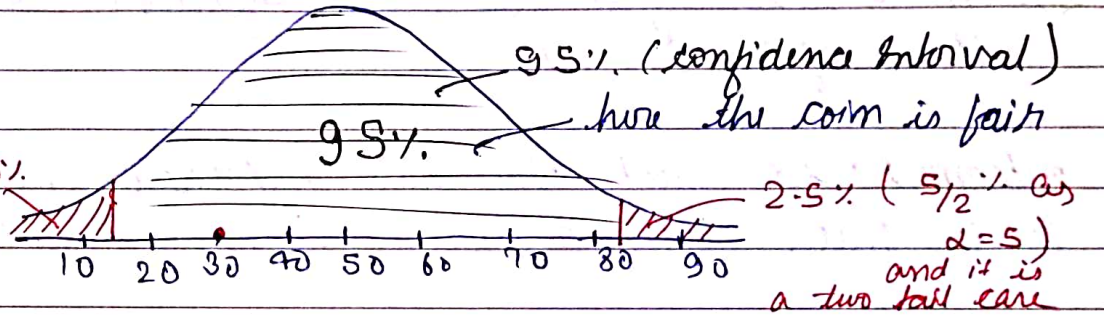
Since $0.0456 < 0.05$, we reject the null hypothesis hence, "there's enough evidence to say the coin might be biased."

Now the question is what is alpha(α) or significance value.

Significance value:- It is the threshold set before a hypothesis test that defines the maximum probability we are willing to accept for making a wrong decision by rejecting a true null hypotheses.

In simpler word,

it is the cutoff point used to decide whether a result is statistically significant or just due to random chance.



Suppose we get 30 head so how do we know that the coin is biased or not.

So for this to define it is always said that our experiment should be nearer to the mean

Now how do we define that how far it can be away from the mean. So for that we use the property of significance value (α).

suppose $\alpha = 0.05$ or 5% (defined by domain expert)

this means that $100 - 5 = 95\%$ is my confidence Interval

Range of values that we believe contains the true value we are trying to estimate.

means the person who is doing the experiment or (someone who understand the field very well) decides how much risk of being wrong they are willing to except before they do the test.

Type 1 and Type 2 Error

Null hypothesis (H_0) = coin is fair

Alternate hypothesis (H_1) = coin is not fair.

Reality check

Null hypothesis is true or Null hypothesis is false.

Decision

Null hypothesis is true or null hypothesis is false

outcome 1 : we reject the null hypothesis when in reality it is false ⇒ True

outcome 2 : we reject the null hypothesis when in reality it is true. ⇒ Type 1 error (False positive)

outcome 3 : we accept the null hypothesis when in reality it is false. ⇒ Type 2 error (False negative)

outcome 4 : we accept the null hypothesis when in reality it is true. True

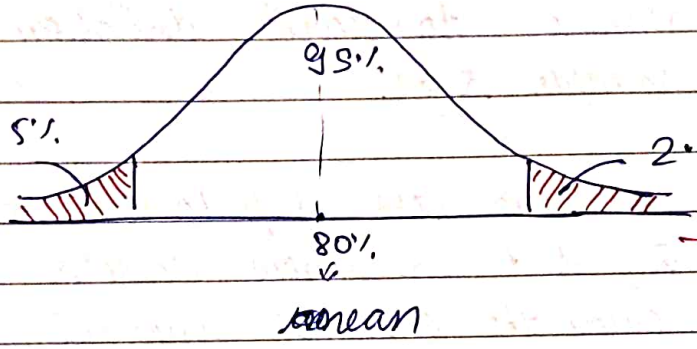
Confusion matrix

	Reality : H_0 is true	Reality : H_0 is false
we reject H_0	Type 1 error	correct (Power of test)
we fail to reject H_0	correct decision	Type 2 error

One tailed and two tailed test

Eg:- Colleges in a state have an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88% with a standard deviation 4%. Does this college has a different placement rate?

⇒ Let say $\alpha = 0.05$



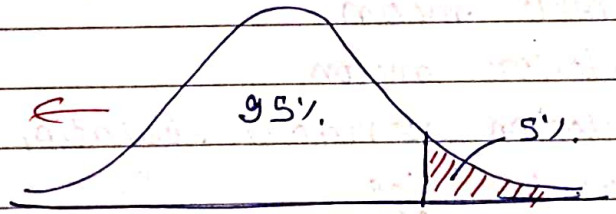
→ two tailed test

Since here we are asked whether the placement rate is greater or smaller than 85%

But if we are asked ~~who~~ does the college have a placement rate greater than 85%? ⇒ one tailed test

$\alpha = 0.05$

one tailed test



⇒ focus on finding only greater.

~~##~~ Confidence Intervals

A confidence interval is a range of values that we believe the true population value (like mean, average etc) falls into, based on sample data.

eg:- We ~~might~~ choose a confidence level, like 95%, which means:

I am 95% confident the true values lies inside the range.

Formula for C.I

$$C.I = \bar{x} \pm Z \times \left(\frac{S}{\sqrt{n}}\right)$$

→ assuming 95% C.I

$\bar{x}$ = sample mean

Z = Z-score based on confidence level (for 95% it is 1.96)

S = sample standard deviation

n = sample size.

Proof

We want to estimate the population mean μ using a sample mean $\bar{x}$, but we know there's some uncertainty. So, we calculate a range (interval) that is likely to contain the true population mean - that's the confidence interval.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

→ This is called Z-test.

$\bar{x}$ ⇒ sample mean

μ ⇒ population mean

σ ⇒ population standard deviation.

n ⇒ sample size

$\sigma/\sqrt{n}$ ⇒ standard error

Z ⇒ how many S.E away our sample mean is from the ⇒ from the definition of z score

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\Rightarrow Z \cdot \frac{\sigma}{\sqrt{n}} = \bar{x} - \mu$$

$$\Rightarrow \mu = \bar{x} - Z \cdot \frac{\sigma}{\sqrt{n}}$$

But that give only one bound. Since we want a range, we use ±z (positive and negative side):

$$\mu \in \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

$$C.I = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

* Now the question is how standard error = $\frac{\sigma}{\sqrt{n}}$ → why $\sqrt{n}$ why not n

To get there I start from basic statistics and properties of variance.

① Let's say we have a population with:
⇒ mean = μ
⇒ standard deviation = σ

② We take random samples of n independent value:
$x_1, x_2, \ldots, x_n$ → consider it as a samples of marks of students in the exam from the population of students.

* Each value has variance
$$Var(x_i) = \sigma^2 \quad (\because \text{it is same for all the sample})$$
HINGLISH
* Saare sample ka apna apna variance

* Variance of the sum of observations:
Let say we add all n observations: HINGLISH
$$S = x_1 + x_2 + x_3 + \ldots + x_n$$ → saare sample ka apna apna sum

Since all x_i are independent, the variance add.

This will be used later when we are calculating it.

$$Var(S) = Var(x_1 + x_2 + x_3 + \ldots + x_n) = n\sigma^2 \quad \text{(i)}$$

Saare sample ka alag alag sum ka variance HINGLISH

* Now we define the sample mean:

$$\bar{x} = \frac{S}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

∴ ~~Sample Variance~~

Variance of sample mean = $Var(\bar{x})$

HINGLISH

Saare alag alag sample ke mean ka variance

$\therefore \bar{x} = \frac{S}{n}$

$$Var(\bar{x}) = Var\left(\frac{S}{n}\right) \quad \text{(i)}$$

Now, As we know,

$Var(X) = E[X^2] - (E(X))^2$ (Page : 14)

$\therefore Var(aX) = E[a^2x^2] - (E(ax))^2$

$= a^2 E[x^2] - a^2 (E(x))^2$

$= a^2 E(x^2) - a^2 (E(x))^2$

$Var(aX) = a^2 [Var(X)]$

similarly:

$$Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X)$$

NOTE

See the central limit theorem, then you will understand how Rahisheran derived this formula by performing the experiment. Pg - 23, 24

$\therefore$ eq (ii) becomes

$$Var(\bar{x}) = \frac{1}{n^2} Var(S)$$

$$Var(\bar{x}) = \frac{1}{n^2} \times \sigma^2 n \quad \text{(from eq i)}$$

$$Var(\bar{x}) = \frac{\sigma^2}{n}$$ ⇒ variance of sample mean

$S.D = \sqrt{Var(\bar{x})}$

$$S.D = \frac{\sigma}{\sqrt{n}}$$ ⇒ standard deviation of sample mean or standard error.

Summary

Let say I have a population of students in a collages. I take one sample of more than 30 students (say, 36). (41)
Calculate the sample mean.
Since the sample mean can vary from sample to sample, we calculate its standard error, which is

$$SE = \frac{\sigma}{\sqrt{n}}$$

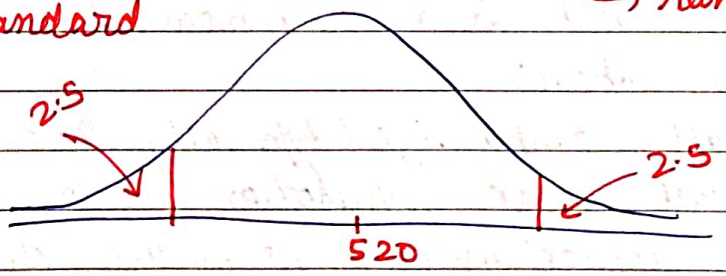
The S.E tells us how much our sample mean might vary from the true population mean. Then we use this S.E to construct a confidence interval around our sample mean - which gives us a range where we are reasonably confident (eg 95%) the true population mean lies.

Q) On the quant test of CAT exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. construct a 95% CI about the mean?

⇒ $\sigma = 100$, $n = 25$, $\alpha = 0.05$, $\bar{x} = 520$

σ = Population standard deviation

$\bar{x}$ → sample mean



① If population std. is given, then we will apply z test.

∴ Point Estimate ± margin of error

② $n \geq 30$ (here we take $n = 25$ for simple calculation)

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

upper bound of C.I $= \bar{x} + Z_{0.05/2} \frac{100}{\sqrt{25}}$

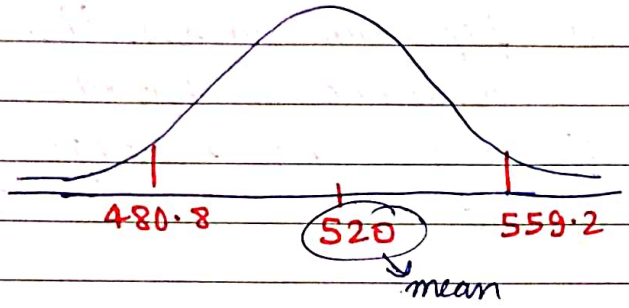
4/2

$= 520 + (1.96)(20)$

↳ from z-table

$= 559.2$

lower bound of C.I $= \bar{x} - Z_{0.05/2} \frac{100}{\sqrt{25}} = 480.8$



* But what if population standard deviation is not given. ⇒ then we use t-test

T-test : A t-test is a statistical test used to compare means when:

→ we have small sample size (typically $n < 30$)
→ when we dont have population S.D,
→ The data is approximately normally distributed.

In simple terms:
"I want to check if the average of my sample is really different from a known value or another sample."

eg:- Let say:
→ I believe student sleep 7 hours on average
→ we take sample of 15 students, and their average sleep is 6.5 hour with sample standard deviation of 0.8 hour

we want to check:
"is this difference (6.5 instead of 7) statistically significant or just random?"

Sample mean ($\bar{x}$) = 6.5
hypothesized population mean (μ) = 7
Sample std. dev (s) = 0.8
sample size (n) = 15
degree of freedom (df) = $n-1 = 14$

Why $n-1$?

Suppose we have 3 numbers and their average is 10

$$\therefore \quad \frac{x_1+x_2+x_3}{3} = 10$$

$$\Rightarrow x_1+x_2+x_3 = 30$$

Now, we can choose any two numbers, like
$x_1 = 8$, $x_2 = 12$
But then the third $x_3 = 10$ must be fixed
hence eventhough we have 3 numbers, only 2 are free to vary.

Now, comes to question.
t-test formula

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{6.5-7}{0.8/\sqrt{15}} \approx -2.43$$

Now, acc. to t-value for dof = 14 at 95% C.I $\rightarrow \sim 2.145$
our $|t| = 2.43$ is greater than 2.145
so, we reject the assumption that average is 7 hour

Q) on the quant test of CAT exam, a sample of 25 test takers has a mean of 520 within a standard deviation of 80. Construct 95% of confidence interval about the mean.

Ans:- $n = 25,\ \bar{x} = 520,\ (S) = 80$

$\sqrt{}$ → Sample S.D

Population S.D not given hence t-test.

Point estimation ± margin of error

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \rightarrow \text{standard error}$$

For t we use D.O.F which is $25-1 = 24$

∴ for $t_{0.025}$ (as $\alpha = 0.05$) and D.O.F = 24

we have $t = 2.064$

$$\text{upper bound} = \bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 520 + 2.064\left(\frac{80}{5}\right)$$

$$= 553.024$$

$$\text{lower bound} = \bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 520 - 2.064\left(\frac{80}{5}\right)$$

$$= 486.97$$

<u>One Sample Z-Test</u>

Q) In the population, the average IQ is 100 within a S.D of 15. Researcher wants to test a new medication to see if there is a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication, has a mean IQ of 140. Did the medication affect the intelligence? $\alpha = 0.05$ (C.I = 95%)

⇒ Population mean (μ) = 100

Population S.D (σ) = 15

Sample size (n) = 30
sample mean ($\bar{x}$) = 140
significance level (α) = 0.05 (for 95% CI, by default)

Test type = Two-tailed (we want to check if IQ is increased or decreased)

Step 1: Set hypothesis
- Null hypothesis (H_0) $\Rightarrow \mu = 100$ (no effect)
- Alternate hypothesis (H_1) $\Rightarrow \mu_1 \neq 100$ (effect present either increase or decrease)

Step 2: Calculate standard Error (SE)

$$S.E = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{30}} \approx 2.738$$

we use this formula for SE when S.D is known.

Step 3: Calculate z score
z-score tells us how many SEs the sample mean from the population mean:

$$Z = \frac{\bar{x} - \mu}{SE} = \frac{140 - 100}{2.738} \approx \frac{40}{2.738} \approx 14.61$$

Step 4: Find critical z-value
since it's a two tailed test and we are using 95% confidence
- critical z-values = ± 1.96

Step 5: compare and conclude
we calculated $z = 14.61$, which is far outside the range of -1.96 to $+1.96$.

Final conclusion
There is strong statistical evidence that the medication affects intelligence, because the sample mean is significantly different from the population mean.

One sample T-test

Z test ⇒ population S.D given

t-test ⇒ population S.D not given.

We will use the same problem as we have taken before in z test, the only change is that this time the population S.D is not given.

Q> Population average IQ = 100, n = 30, $\bar{x}$ = 240, s = 20. did the medication affect intelligence? $\alpha = 0.05$

⇒ Given:- population mean (μ) = 100

sample mean ($\bar{x}$) = 140

sample size (n) = 30

sample standard deviation (s) = 20

degree of freedom (df) = n-1 = 29

$\alpha = 0.05$

Step 1: set up hypothesis (Two tailed)

H_0 : $\mu = 100$ (medication has no effect)

H_1 : $\mu \neq 100$ (medication affect IQ)

Step 2: calculate the test statistic (t-score)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{110 - 100}{20/\sqrt{30}} = \frac{10}{3.651} \approx 2.739$$

Step 3: Find the critical t-value

From the t-distribution table with:

→ df = 29

→ $\alpha = 0.05$ (two tailed)

critical $t \approx \pm 2.045$

Step 4: compare

- Calculated $t = 2.739$
- Critical $t = \pm 2.045$

since, $2.739 > 2.045$, we reject the null hypothesis.

Final conclusion

1) The medication has a statistically significant effect on IQ.

2) since the sample mean (110) > population mean (100), the medication likely improve intelligence.

CHI SQUARE TEST

The chi-square test is a statistical test used to compare observed data with expected data.

Formula

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

where:

O = Observed frequency
E = Expected frequency.

Properties.

1) It is a non-Parametric test.

Non-parametric means

- It doesn't assume anything about the distribution of the data (like normal distribution).
- It is based on frequencies, not means or standard deviation.

So, it's perfect when you data is catagorical - not continuous.

2) chi-square test is performed on categorical variables, which can be:

Type	Example
Nominal	Gender (Male / Female), Yes / NO, color
Ordinal	Education level (High school < Graduate < Postgrad)

- Eventhough ordinal data has some order, in chi-square we just count categories, not treat them as numeric values.

Summary

Feature	chi-Square Test
Type	Non-parametric
Works on	categorical data (nominal, ordinal)
Uses	Frequencies (not means / variance)
Assumptions	No assumption of normality
Typical use case	Test independence or goodness-of-fit.

(96)

Q) In the 2000 indian census, the age of individual in small town were found to be the following.

< 18	18 - 35	> 35
20%	30%	50%

In 2010, age of $n = 500$ individuals were sampled, below are the result

< 18	18 - 35	> 35
121	288	91

using $\alpha = 0.05$, would you conclude the population has changed in the last 10 years.

Ans :⇒ Step 1: Define the hypothesis.

- Null hypothesis (H_0): The age distribution in 2010 is the same as in 2000
- Alternative hypothesis (H_1): The age distribution in 2010 is different from that in 2000.

Step 2: Calculate expected counts.

Multiply the 2000 proportions with the 2010 sample size ($n = 500$)

< 18	18 - 35	> 35
$0.20 \times 500 = 100$	$0.30 \times 500 = 150$	$0.5 \times 500 = 250$

Step 3: Apply chi-square test formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed freq. E_i = Expected freq.

$$\therefore \chi^2 = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{25}$$

$$\chi^2 = 232.49$$

Step 4: Degree of freedom.

df = Number of categories $- 1 = 3 - 1 = 2$

Step-5: critical value at $\alpha = 0.05$ and $df = 2$ from chi-square distribution table

$$\chi^2_{critical} = 5.991$$

Step 6: Conclusion

since,

$$\chi^2_{calculated} = 232.49 \gg \chi^2_{critical} = 5.991$$

we reject the null hypothesis

hence, there is a strong evidence to suggest that the population age distribution has changed between 2000 and 2010.

Relation between confidence interval, significance value and P value

1) confidence Interval = $1 - \alpha$ (significance value)

2) If P- value < significance level (α)
→ reject the null hypothesis (H_0)

If P-value > α
→ accept the null hypothesis (H_0)

Covariance

Covariance is a measure of how two variables change together.

Intuition

→ If both variable increases together, the covariance is +ve.
→ If one increases while the other decreases, the covariance is negative.
→ If they don't seem to change together, the covariance is close to 0.

eg:-

study hour	Exam marks
2	50
4	60
6	70
8	80

+ve covariance

Formula

~~Cov(X,Y) = $\frac{1}{n}$~~

1) Population covariance formula

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

↳ use this when we have the entire population.

2) Sample Co-variance formula

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

↳ use this when we have a sample from the population.

The question is here why divide by $(n-1)$. in the formula of sample covariance.

Ans:- This is due to the Bessel's correction.
But what it is and how it is divided by $(n-1)$.

Population (N)	sample (n)
For mean	
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
For Variance	
$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu_x)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

suppose we have a data of people's ages in a population.

When I randomly take a sample from this population and calculate the sample mean, it often turns out to be close to the population mean. The same happens with sample variance - it becomes closure to the true variance of the population, but only if we correct it properly.
However, if I just take the first few values (like the first 4), or a random slice, the sample mean and variance might be very different from the actual population.

To fix this issue, statistician ran experiments:
They took many samples from the population and tried calculating variance by dividing by

- n
- $n-1$
- $n-2$

and so on

what they found was this!
when they divided the sum of squared differences by n-1, the average of the sample variances across all their samples came closest to the true population variance.

This is why we divide by n-1 to make the sample variance an unbiased estimator of the population variance.

This technique is called bessel's correction.

Since covariance doesn't have fixed limit - its value depends on the scale of the variables.
For example, if our variable in kilograms and rupees, or meters and dollar, the covariance can become very large or very small, making it hard to interpret. So we use the pearson correlation coefficient.

Pearson Correlation coefficient

The pearson correlation coefficient (r) is a statistical measure that tells us how two variables are linearly related. i.e, whether they move together and how closely.

Formula:

$$r = \frac{Cov(X, Y)}{\sigma_x * \sigma_y}$$

Where:
→ Cov(X, Y) = covariance of variables X and Y.
→ σ_x = standard deviation of X.
→ σ_y = standard deviation of Y.

Values of r lies between -1 and 1.

r value	meaning
$+1$	Perfect positive correlation
0	NO linear correlation.
-1	Perfect negative correlation.
closer to 1 and -1	strong relationship
closer to 0	weak relationship.

Eg:-

Student	X (hours studied)	Y (Marks Scored)
A	1	2
B	2	3
C	3	6
D	4	8
E	5	7

Step 1: Calculate means:

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{Y} = \frac{2+3+6+8+7}{5} = \frac{26}{5} = 5.2$$

Step 2: Build the table

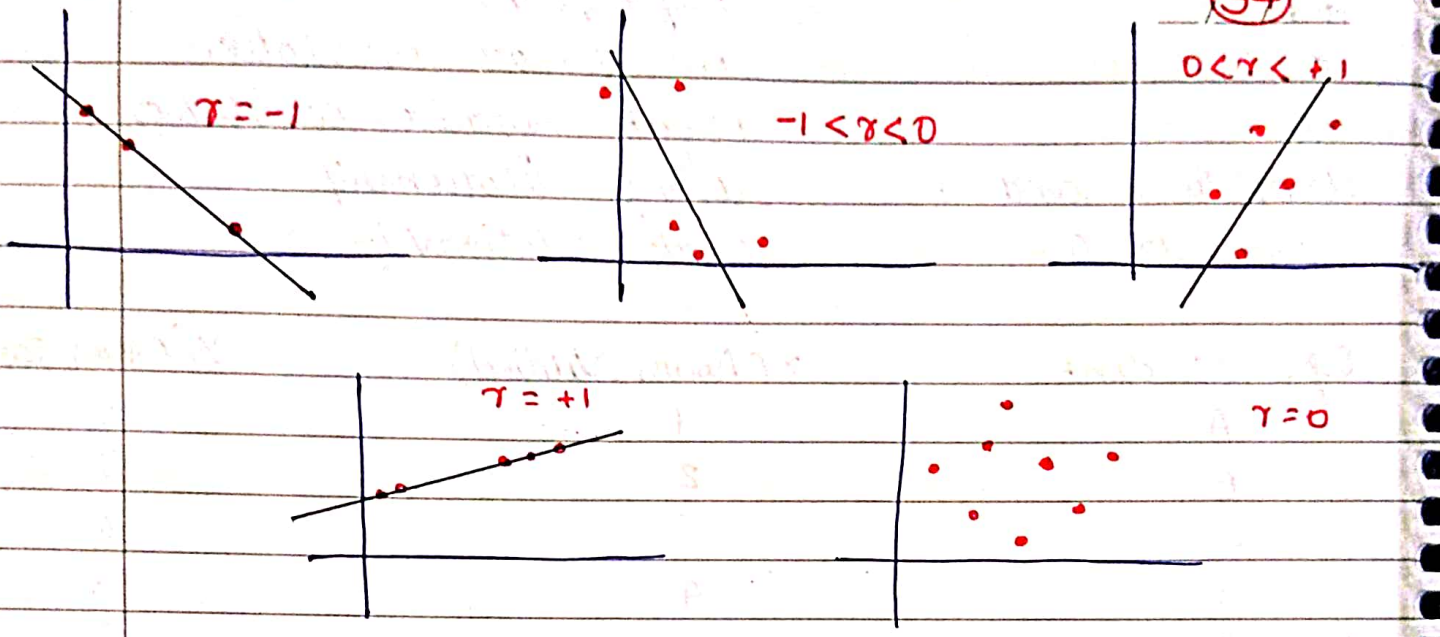
Student	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	1	2	-2	-3.2	6.4	4	10.24
B	2	3	-1	-2.2	2.2	1	4.84
C	3	6	0	0.8	0	0	0.64
D	4	8	1	2.8	2.8	1	7.84
E	5	7	2	1.8	3.6	4	3.24

Step 3: Apply the pearson correlation Formula

$$r = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \cdot \Sigma (y-\bar{y})^2}}$$

$$r \approx 0.916$$

That's a very strong +ve correlation - almost perfect, but with small variations in data.



Spearman coefficient correlation

The spearman's Rank correlation coefficient (denoted by ρ or r_s) is a non-parametric measure of monotonic association between two variables - based on the rank values other than the raw data.

Key Points

- It doesn't assume a linear relationship like Pearson's correlation.
- It is used when data is ordinal, not normally distributed, or when you want to measure monotonic relationships.
- Works well even with outliers or non-linear relationships.

Formula

If there is no similar ranks

$$r_s = \frac{1 - 6\sum d_i^2}{n(n^2-1)}$$

where: d_i = diff. b/w the ranks of the two variable for each observation.

n = number of observations.

Example

Let say we have data for 5 students

Student	Math Rank (X)	English Rank (Y)
A	1	2
B	2	1
C	3	4
D	4	3
E	5	5

Now compute the difference in ranks:

Student	X	Y	$d = X - Y$	d^2
A	1	2	-1	1
B	2	1	1	1
C	3	4	-1	1
D	4	3	1	1
E	5	5	0	0

$$\Sigma d^2 = 1 + 1 + 1 + 1 + 0 = 4$$

Now apply the formula:

$$r_s = 1 - \frac{6 \cdot 4}{5(5^2 - 1)} = 0.8$$

$\therefore r_s = 0.8 \rightarrow$ strong +ve monotonic relationship b/w Math and English ranks.