Statistics :-

Sampling & Population

→ Probability Sampling

→ Non " "

Simple Random Sampling ①

*) Easy to implement.

Systematic Sampling ②

*) Hard to find hidden pattern with this.

Stratified Sampling ③

*) Hard to apply on large population.

*) Considers all groups.

Clustered Sampling ④

Non Probability Sampling

Convenience Sampling

Quota Sampling

Scanned with OKEN Scanner

Measures of Central tendancy :-

Mean Sensitive to outlayers

Median

Mode.

Mostly uses to fill the categorical values

Scanned with OKEN Scanner

Measures of Dispersion :-

Mean absolute Deviation :-

Name	Score		Name	Score	
Mohan	75	5	Mohan	93	23
Andsa	72	2	Andsa	96	26
Sofia	68	2	Sofia	43	27
Joe	65	5	Joe	47	23
Virat	67	3	Virat	51	19
Abdul	73	3	Abdul	90	20
Average	70	3.33	Average	70	23

"average of distance between the actual spreaded points and actual average of overall spreaded points."

Scanned with OKEN Scanner

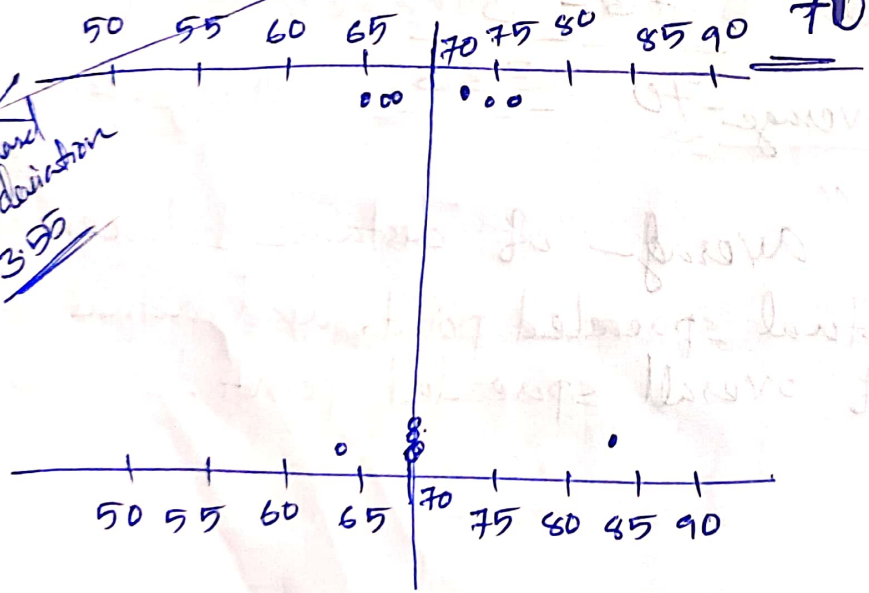
Name	Score	Square
Mohan	75	5 - 25
Andrea	72	2 - 4
Sofia	68	2 - 4
Joe	65	5 - 25
Vivat	67	3 - 9
Abdul	73	3 - 9
	70	3.33

12.66 Variants

Name	Score	Abs Score - avg
Mohan	83	13
Andrea	70	0
Sofia	70	0
Joe	63	7
Vivat	70	0
Abdul	70	0
	70	3.33

Standard deviation

$\sqrt{12.66}$ 3.55



Square
169
0
0
49
0
0
36.33

Variants

$\sqrt{36.33}$ - standard deviation = 6.02

Range.

Maximum value - Minimum value

75 - 65 = 10

83 - 63 = 20

Research on this using medium blogs.

EDA

Data cleaning & visualisation on titanic dataset - 15 graphs.

- Research topic - "Probability"

Scanned with OKEN Scanner

Mean -

SD -

Mean/SD ×100 — CV.

Quartiles.
Percentiles.
Deciles.
~~Quintiles~~
Quartiles

} Quantiles.

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Equations:-

Percentile location =

Percentile/100 × (n+1)

75/100 × (10+1)

3/4 × 11 = 8.25 — Value 96.5

75th percentile

to get the value: lowes point + (difference of data points) × .25

to get the value

lowes point + (difference of data points) × .25

to get the Value.

lowes point + (difference of data points) × .25

25th percentile.

25/100 × 10+1 = 25/100 × 11

= 1/4 × 11

location = = 2.75

Value.

82 + 2 × .75 = 83.5

50th percentile

50/100 × 11

1/2 × 11 = 5.5 - location

91 + 2 × .5 = 92 - Value

Quartiles.

Divide the data points into 4 equals parts.

Percentile

Divide the data points into 100 equal parts.

5 Numbers Summary

- ~~starting~~ Point/Minimum Value.
- Q_1/25th percentile/1st quartile. the value that seperates the lower 25% of the data from the rest of data set
- Q_2/Median/50th percentile/2nd Quartile
- Q_3/75th percentile/3rd quartile. the value that seperates the lowest 75% of the data from the rest of data
- 100th percentile/Maximum Value.

Inter Quartile Range.

This is where 50% of the data exists. $Q_3 - Q_1 = IQR$, Mean 75th Percentile '−' 25th percentile.

Finding the location or Percentile with the Selected point :-

Q - Point = 91.

equation :-

$$\frac{x + .5 \times y}{n} \times 100$$

x - no. of values under the point.

y - No. of same points. respectation

$$= \frac{4 + .5 \times 1}{10} \qquad \frac{4.5}{10} = .45 \times 100$$

$= 45^{th}$ percentile

Q - 96.

$$\frac{7 + .5 \times 1}{10} = \frac{7.5}{10} = .75 \times 100$$

$= 75^{th}$ percentile

Q-99

$- \frac{94.5 \times 1}{10} = .95 \times 10 = 95^{th}$ percentile

$\boxed{\text{This equation excludes maximum Value}}$

Question :-

6, 213, 241, 260, 281, 290, 314, 321, 350, 500 (Find the five numbers summary)

(Min) 1 - 6

(Q_1) 2 - 234

(Q_2) 3 - 285.5

(Q_3) 4 - 328.25

(Max) 5 - 500

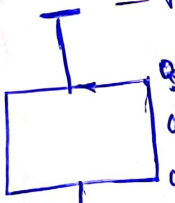
Inter Quartile Range :-

$IQR = Q_3 - Q_1$

$= 94.25$

(Upper limit)

- Max - $Q_3 + 1.5 \times IQR$



Min - $Q_1 - 1.5 \times IQR$
(lower limit)

Max - 469.62 - upper limit

Min - 92.62 - lower limit

35, 40, 78, 85, 88, 94, 98, 102, 150, 160

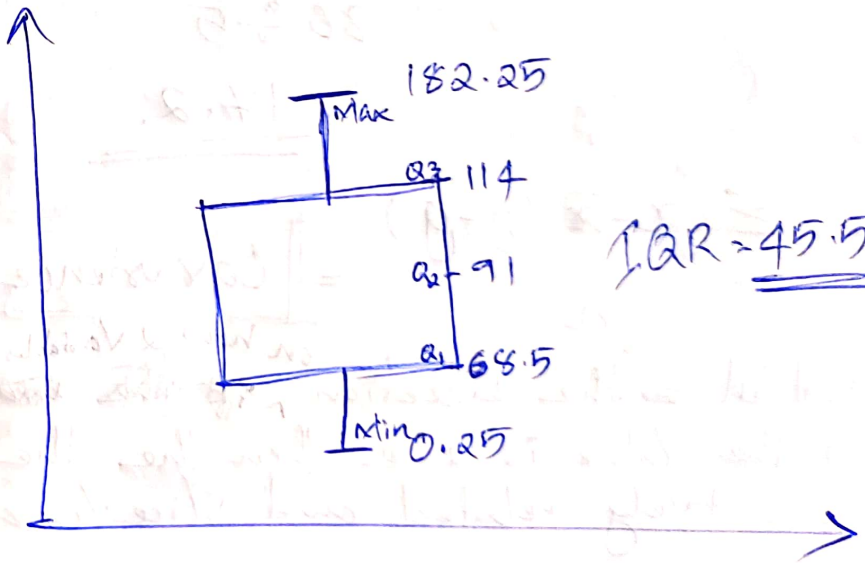
275

8.25

Q_1 - 68.5

Q_2 - 91

Q_3 - 114



IQR = 45.5

(x) ~~Scatter Plot~~ :-

Experience	Salary (y)	$x-\bar{x}$	$y-\bar{y}$	$\bar{x} \times \bar{y}$
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	0
12	12	4	6	24
13	10	5	4	20
$\bar{x} = 8$	$\bar{y} = 6$			86

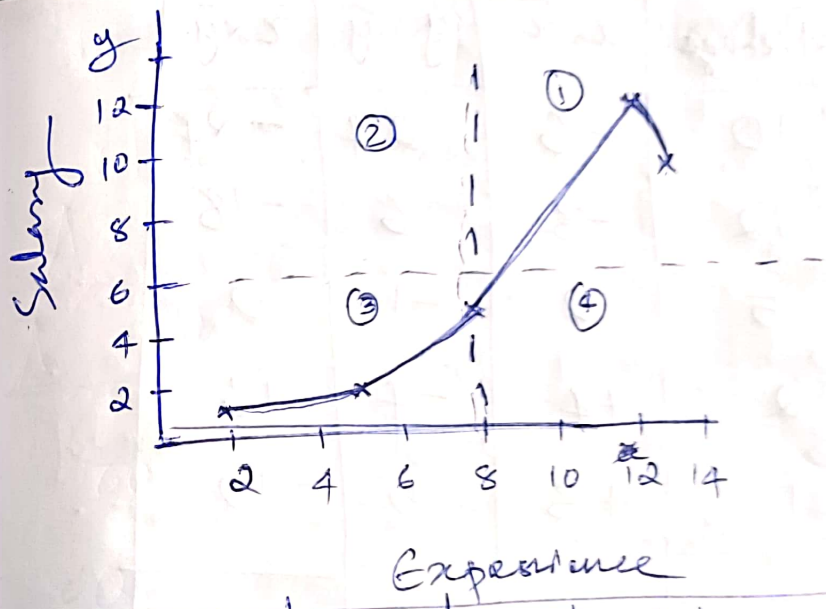
$86 \div 5$

$= 17.2.$

$$\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \boxed{\text{Covarience}}$$

on how 2 Variables are related.

Just tell the direction, if the Value is +ve then the the Values are +vely related and Vice Versa.

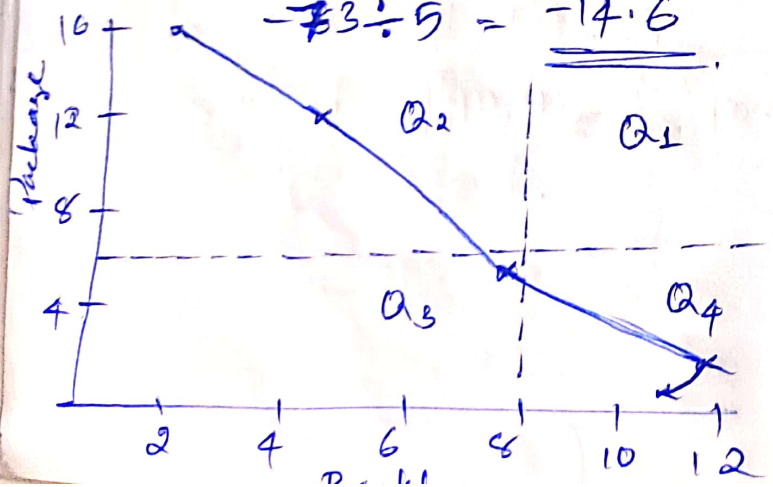


If the 1st & 3rd quadrant has more Points - it shows the graph is positive. and Vice Versa.

Backlog	Package	$x-\bar{x}$	$y-\bar{y}$	$\bar{x}\times\bar{y}$
2	10	-6	24	-24
5	12	-3	6	-18
8	5	0	-1	0
12	2	+4	-4	-16
11	1	+3	-5	-15
$\bar{x}$ 8	$\bar{y}$ 6			-73

$-11 \div 5 = -2.2$ = Covariance

$-73 \div 5 = -14.6$



Range of Covariance is -infinity to +infinity

Correlation co-efficient = $\frac{\text{Covariance}}{\text{std}(x) \times \text{std}(y)}$

(-1 to +1)

Value will be b/w

"Correlation"

Causation

Increment of a value is not the reason of increment of another value. and vice-versa

Probability

2 conditions
① should have multiple outcomes
② Outcome should be uncertain

→ Random Experiment

→ Trial - Single execution of a random experiment

→ Outcome - Single possible ~~value~~ result of a trial

→ Sample space - All possible outcomes.

→ Event - (a subset of sample space). a set that contains atleast 1 particular outcome.

Probability range = 0 — 1

① Rolling a dice.

→ It's a random experiment

→ Rolling a dice once

→ 1 is a single possible outcome

→ 1, 2, 3, 4, 5, 6.

→ Event = [1, 3, 5] - possibility of getting odd numbers.

② Tossing a coin twice.

→ It's a random experiment.

→ Tossing a coin twice

→ HH is a single possible outcome

→ HH, HT, TT, TH.

→ Event = [HH, HT, TH] possibility of getting a Head.

③ Rolling dice twice.

→ Rolling dice twice

→ " "

→ 1, 1

→ Event = [11, 12, 13, 14, 15, 16] ... possibility of getting atleast 1 in each outcome.

→ Sample space :- 11 12 ...

21 ....

31 ...

41 ...

51 ...

61 ....

Scanned with OKEN Scanner

Types of Events :-

1. Simple event – has only 1 outcome
2. Compound event – has multiple outcomes
3. Independent event – occursing of 1 outcome wont affect the other outcome.
4. Dependent events – Occursing of 1 outcome will affect the other outcome.
5. Mutually exclusive event – If they cant occur both simultaneously.

Emperical Probability

Theoritical Probability.

Algebraic Variable :-

~~Set of posib~~

Random Variable :-

Set of all possible values from a Random experiment.

Types of Random Variable :-

→ **Continuous**

~~Values~~ outcomes/values will be Continuous. Set of all values within a range.

eg:- Mark of CGPA

→ **Discreate**

Values will be discreate or Separable. Countable finite values

eg:- Rolling a dice

Probability distribution ~~function~~ :- PDF

All possible values with its probability

← of a Dice :-

All values :-	1	2	3	4	5	6
All probability :-	1/6	1/6	1/6	1/6	1/6	1/6

← of a Coin.

x	H	T
P(x)	1/2	1/2.

Scanned with OKEN Scanner

Value

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Probability

(x)	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Probability Distribution function :- (PDF)

$y = \frac{1}{2}$ if x belongs to 1, 2, 3, 4, 5, 6.
otherwise 0.

→ ~~a~~ ~~types~~ PDF for :- PDF

(*) Continuous - Probability density function

(*) Discreate - Probability mass function PMF

Refer :- Types Probability distribution.

why probability distribution :-
- gets the shape of data.
- gets if its following any famous distribution, ~~as~~ ~~if~~ all its equations are ready

PMF :- A mathematical function that describes the probability distribution of a discreate random variable

PMF of Rolling a dice :-

$$\left.\begin{array}{ll} \frac{1}{6} \text{ if } x = 1 \\ \frac{1}{6} \text{ if } x = 2. \\ \frac{1}{6} \text{ if } x = 3 \\ \frac{1}{6} \text{ if } x = 4 \\ \frac{1}{6} \text{ if } x = 5 \\ \frac{1}{6} \text{ if } x = 6. \\ 0 \text{ otherwise} \end{array}\right\} \text{PMF of rolling a dice.}$$

Scanned with OKEN Scanner

PMF of Sum of rolling a dice:

$\frac{1}{36}$ if $x = (2, 12)$

$\frac{2}{36}$ if $x = (3, 11)$

$\frac{3}{36}$ if $x = (4, 10)$

$\frac{4}{36}$ if $x = (5, 9)$

$\frac{5}{36}$ if $x = (6, 8)$

$\frac{6}{36}$ if $x = (7, 7)$

CDF of PMF (Cumilative distribution function)

This describes the probability that a random variable (X) with a given probability distribution will be found at a value less than or equal to 'x'

$$\boxed{F(x) = P(X <= x)}$$

The PMF of Discrete Variable assigns a Probability to each possible Value of the random Variable. It must satisfy 2 conditions:-

- each Value should by non-negative
- ~~Such~~ Sum of the probability assigned to each Value should be 1.

CDF of rolling a dice:-

$$P(x \leq 4) = P(x=1) + P(x=2) + P(x=3) + P(x=4)$$

	PMF	CDF
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$
6	$\frac{1}{6}$	$\frac{6}{6}$

Scanned with OKEN Scanner

Probability density function :-

Q- Tossing a coin twice - getting a head

Sample space : [HH, HT, TH, TT]

Random Variable - X = (0, 1, 2)

PMF = $\boxed{\frac{1}{4}, \frac{2}{2}, \frac{1}{4}}$

HH	TH	HT	TT
2	1	1	0

that is PMF =

$\frac{1}{4}$ if $x = 0$

$\frac{1}{2}$ if $x = 1$

$\frac{1}{4}$ if $x = 2$

0 otherwise

CDF at $P(X = 2)$

$F(x) = P(X \leq x)$

$= \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = 1$

Probability density function :-

its a mathematical function that describe probability distribution of a continuous ~~var~~ random variable.

Q- Why probability density instead of probability

Ans- It is hard to find a point as it is continous random variable

(*) Y axis of graph represents probability density which is not the probability bcoz the values of x axis are continous and Probability of a particular value of x like 7.91234 is close to '0'.

(*) Probability density tells the probability of a specific number that lies between a number (a, b).

(*) The area b/w 'a' and 'b' gives the Probability of a particular number which lies b/w 'a' and 'b'.

Scanned with OKEN Scanner

How to make graph for Probability density function :-
(PDF)

Major Problem here is to calculate the probability density. So, here we use the ~~pro~~ density ~~[illegible]~~ estimation technique.

It is a statistical technique used to estimate the PDF of a random Variable based on a set of observation or data.

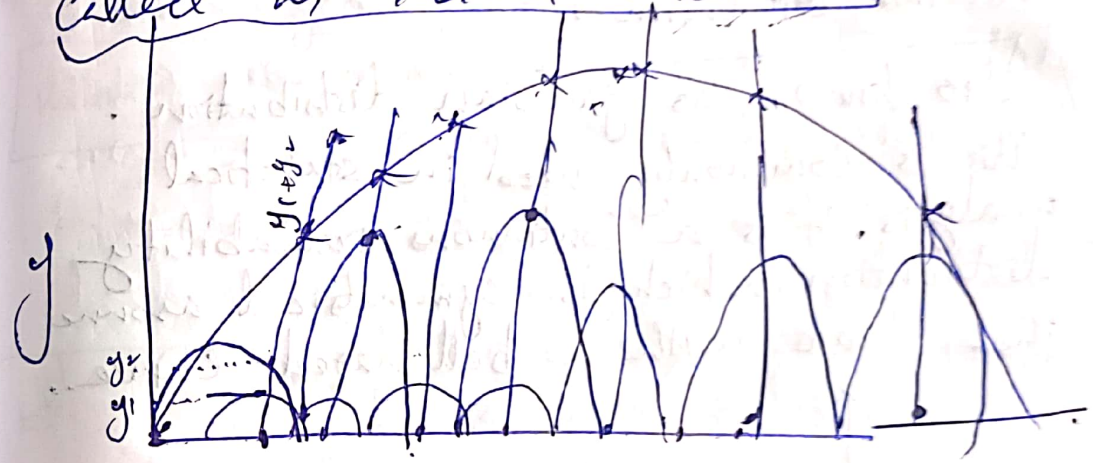
The 2 methods of density estimation are :-

1. Parametric density estimation.
2. Non parametric " "

Assume the data follows a specific Probability distribution such as normal distribution pattern.

This donot make any assumption about the distribution and instead estimate it directly from the data.

Called as Kernel Methond :-



adding the intersection (y_i).

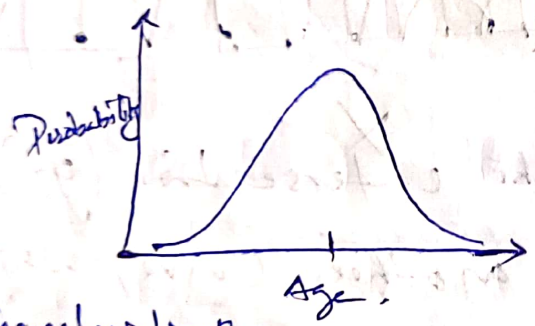
we plot another curve, which is the p.d.f generated using Non-Parametric estimation

Else Case of PDF

Helps in machine learning for feature Selection. eg :- In the dataset of Iris, By doing KDE plot, we Selected column "petal width", "petal length" for feature selection for Seggregation of species.

Normal distribution :-

Also known as gaussian distribution. this is commonly used in statistical analysis. It is a continuous probability distribution which is symmetrical around the mean with a bellshaped curve.



characteristics :-

→ Asymptotic in nature (never touches x axis → touches infinity)

→ lot of points near mean and very few far away

This is characterised by 2 parameters :-

① Mean (μ) - Represents the Centre →

② standard deviation (σ) - Represents the spread of the distribution.

PDF of normal distribution - equation :-

notation of normal distribution $= N(\mu, \sigma)$

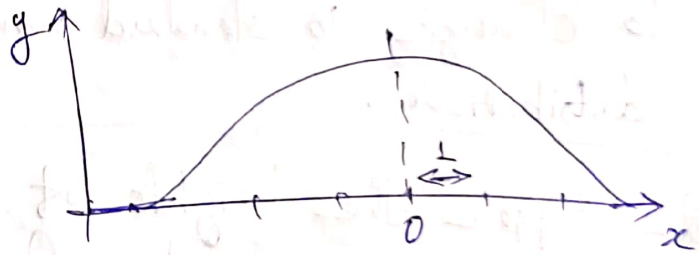
$$y = f(x) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \times \left(\frac{x-\mu}{\sigma}\right)^2}$$

Why is it important :-

(*) Commonality in nature (height, weight).

Standard normal distribution :-

It's a special or standard case of a normal distribution with mean = 0 and standard deviation = 1.



standardising a normal distribution allows us to compare different distribution with each other and to calculate probabilities using standardised tables or software.

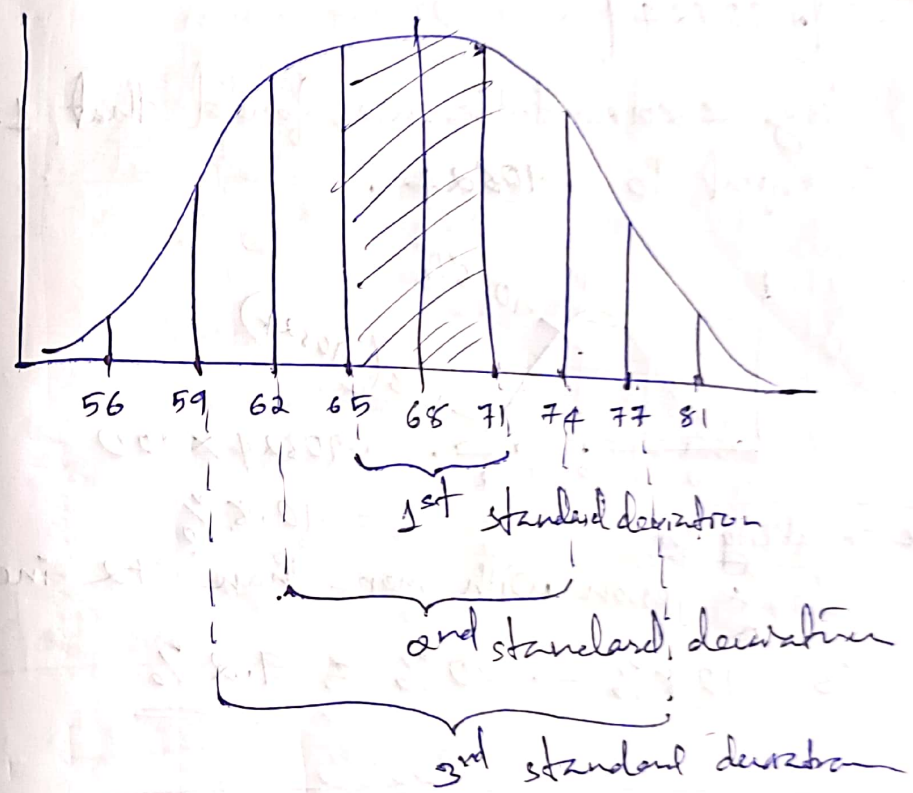
$$y = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2)}$$

Scanned with OKEN Scanner

Converting data to a standard normal distribution allows for comparison across different data sets. By standardizing, we can compare scores from different distributions by seeing how many standard deviations away from the mean a particular point is.

$$\boxed{\left(\frac{x-\mu}{\sigma}\right) \text{ is the equation to change to standard normal distribution}}$$

Q- Suppose the heights of adult males in a certain population follow a normal distribution with a mean of 68 inches and standard deviation of 3 inches. What is the probability that a randomly selected adult male from this population is taller than 72 inches.



$$Z \text{ score} = \frac{x-\mu}{\sigma}$$

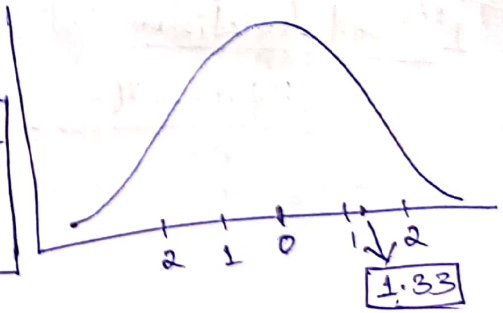
$x = 72$.
$\mu = 68$
$\sigma = 3$

$$= \frac{72-68}{3}$$

$$= 1.33$$

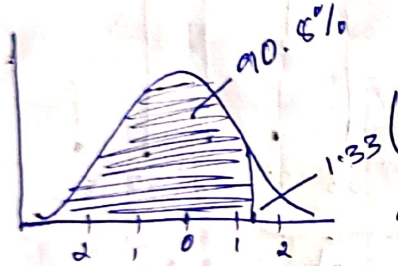
(72)

This point is away from the mean by 1.33 standard deviation.



$\boxed{.90824}$

closing z score table, we found that 1.33 is equal to .90824.



$.90824 \times 100$

$= 90.8\%$

The probability for

So, a person with more than 72 inches is $90.8\% - 100\% = \underline{\underline{9.2\%}}$

Q- For normal distribution, what percentage of Population lie between 1st, 2nd & 3rd standard deviation

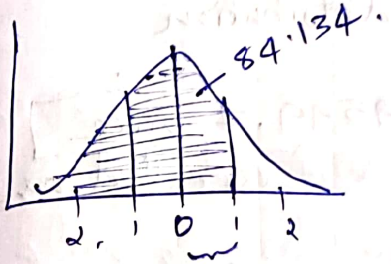
Ans

1st std deviation

$$\frac{\mu + \sigma - \mu}{\sigma} = 1. = z \text{ score}$$

$= .84134$

$= 84.134\%$.



$84.134\% - 50\%$

$= 34.134$.

So, 1st std deviation $= 34.134 + 34.134$

$= \underline{\underline{68.268\%}}$.

2nd std deviation

$$\frac{\mu + 2\sigma - \mu}{2\sigma} = 2 - z \text{ score}$$

$= 97.725\%$

$97.725 - 50\%$

$= 47.725$

So, 2nd std deviation $= 47.725 + 47.725$

$= \underline{\underline{95.45\%}}$

3rd std deviation

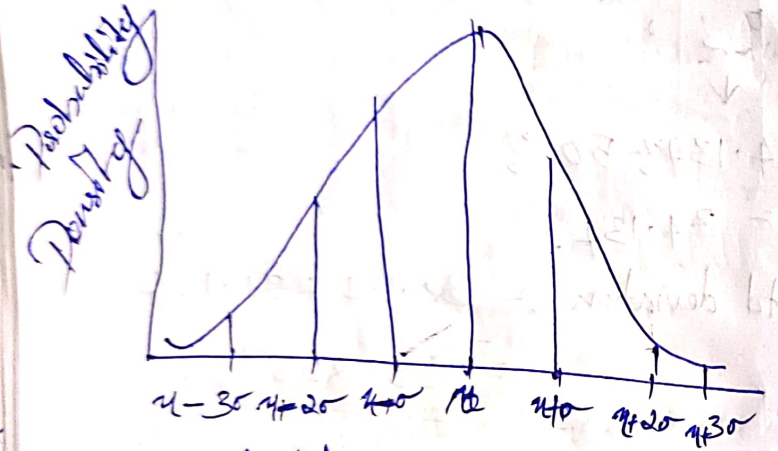
z-score 3 $= 99.865\%$, $99.865\% - 50\% = 49.865\%$

3rd std deviation $= 49.865 + 49.865 = \underline{\underline{99.73\%}}$

Properties of Normal distribution

1- Empirical rule :- $\boxed{68\,95\,99.7 \text{ rule}}$

also called as



1st std deviation — 68.268%

2nd std deviation — 95.45%

3rd std deviation — 99.73%

Usecase of empirical rule :-

(*) Outlayes detection other than boxplot.

$\boxed{\text{Standardisation also called "Zscore normalisation"}}$

We can use standardisation :-

→ when data follows a normal distribution

→ ML algorithms assumes data is normal distributed. eg:- linear regression

→ features have different units or scales and you want to compare them on the same scale.

→ The model we are using is sensitive to outlayers since standardisation doesn't limit the range of data.

2- Symmetricity

It is simmetric about its mean which means that the probability of observing a value above the mean is same as the probability of observing a value below the mean. The bell shaped curve reflects the symmetry

3- Measures of central tendancy are equal

Normalisation :

also known as min-max scaler. Rescales the data to a fixed range typically b/w 0 – 1. The formula used is

$$x' = \frac{x - x_{min}}{X_{max} - X_{min}}$$

Student	Math score	x'	English Score	x'
A	70	→ ·33	85	→ ·25
B	80	→ ·66	95	→ 1
C	90	→ 1	80	→ ·5
D	60	→ 0	70	→ 0

Q – Identify min & max value of each features.

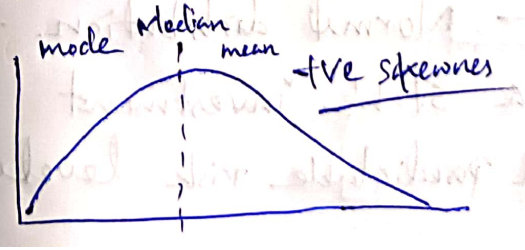
Maths :– $\frac{70-60}{30} = \frac{10}{30} = ·33$.

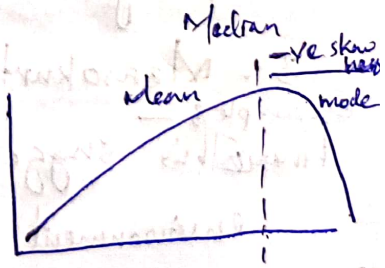
why it is used :–

Normalization ensures that no feature dominates the distance metric used in algorithms like KNN or clustering (k means) which rely heavily on distance means.

Skewness :–

A normal distribution is a bellshaped symmetric distribution that describes how the data is spread out. Skewness indicates the data is not symmetrical. which means it is not normally distributed. [A measure of assymmetry of a probability distribution.] It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.





Scanned with OKEN Scanner

Types of skewness :-

① positive skewness - More data ~~right~~ left side

② negative skewness - More data on ~~left~~ right side

If the skewness lies b/w -0.5 & 0.5 the distribution is moderately skewed. If it lies b/w -1 & +1, the distribution is highly skewed.

Kurtosis

This is the 4th statistical moment. In Probability theory & statistics, It is a measure of tailedness.

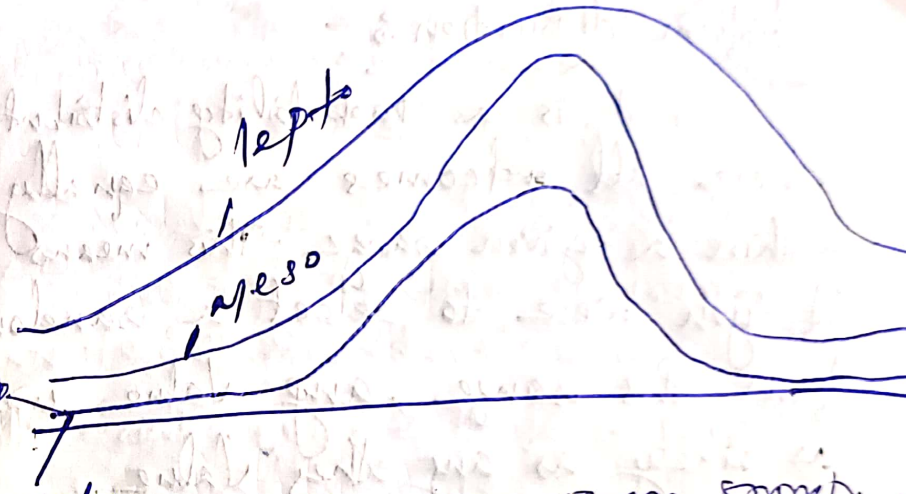
3 types :-

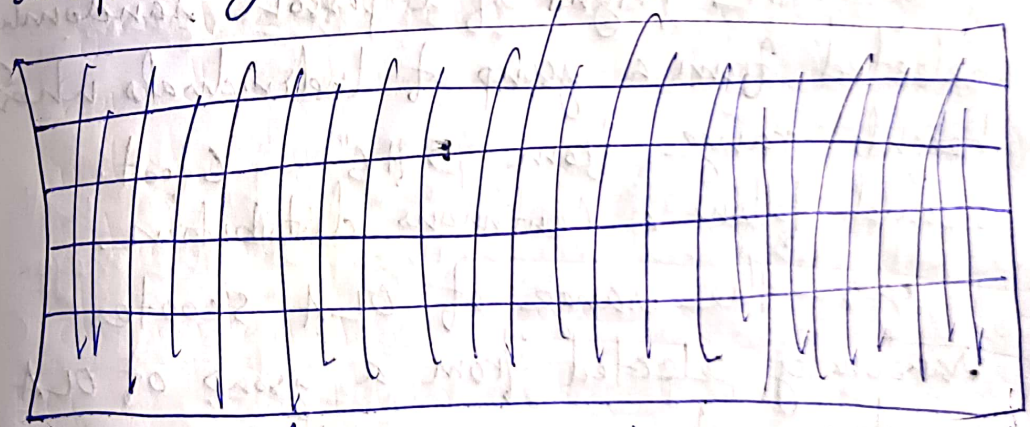
1- leptokurtic - flattened tail. >0 (fatter tails) eg:- Risk Management - Credit score - highlight the

2- platykurtic - Narrowed tail. <0 eg:- Exam score - This suggest a uniform performance among students with fewer extreme

3 - Mesokurtic - Normal distribution = 0

examples :-
Finance This suggest a stable investment environment with predictable risk levels.





Values could indicate extreme risk.

Uniform distribution :-

It is a probability distribution where all outcomes are equally likely within a given range. This means that if you were to select a random value from this range, any value would be as likely as any other value.

eg:- The height of a person randomly selected from a group of individuals whose height range from 5 ft 6" — 6 feet would follow a continuous distribution

eg:- The range of OLA scooters randomly selected from a group of OLA scooters whose ~~range~~ mileage ranges from 150 – 170, would follow a continuous distribution

USE CASE
- Random initialization
- Sampling
- Hyper parameter tuning

Equation :-

$$f(x) = \frac{1}{b-a}, \quad a = \text{Min}, \; b = \text{Max}.$$

Log normal distribution :-

It is a heavy tailed continuous Probability distribution of a random variable whose logarithm is normally distributed.

eg:- When few percentage of users of Instagram's screen time is used for ~~reading~~ reading comments other than just watching the posts. The distribution here is log normal distribution among both scrolless and comment readers.

When your data shows a right skew and most observations are small with a few large outliers.

The log normal distribution has a long right tail. Most of the values are small, but a few large values exist.

Scanned with OKEN Scanner

Pareto distribution :- $y = k \times x^{\alpha}$

80/20 - Powerrule.

In mathematics, power law is a functional relationship between 2 variables where one variable is proportional to the power of the other. This follows the 80/20 rule, where 80% of the effect comes from the 20% of the causes.

Elsecase :-

when your data is extremely skewed with very few items accounting for the majority of distribution.

eg:- wealth, web traffic etc.

Transformation :-

1) logarithmic transformation
2) Square root
3) Reciprocal.

④ Box cox ⑤ Yeo Jhonson

Transformation in machine learning :-

In machine learning, transformations are applied to the data to make it more suitable for modelling, especially for algorithms that assume normality or linearity.

1) Logarithmic transformation :-

It compresses large values and reduces right skew distributions & make it more symmetric.

when you have features with exponential growth & large outliers - $\boxed{\text{else case}}$

2) Square root transformation :-

Reduces the impact of large values, but less agressively than the logarithmic transformation.

$\boxed{np.sqrt(\ )}$

Scanned with OKEN Scanner

3) Reciprocal transformation :-

Inverse the data making large values small. Useful when data need to be compressed. For extremely skewed data, often in situations where outliers dominate.

4) Box-cox transformation :-

Use case :- when you have positively skewed data, when all values are positive.

It finds the optimal power parameter lambda for the transformation which can vary between different datasets. More flexible and can perform transformation similar to square root / log / no transformation depending on lambda.

lambda = 0 (log)
lambda = 1 (No transformation)
lambda = .5 (Square root)

5) Yeo-Jhonson transformation :-

This is a generalisation of the box-cox transformation. The major difference is that Yeo-Jhonson works with both +ve & -ve data. While box-cox works with positive values only. The make Yeo-Jhonson more flexible.

USE CASE - when your data set contains both +ve & -ve values, when you need to transform mixed sign dataset.

eg :- Profit & loss scenario.

Bernoulli Distribution :-

It is a probability distribution that models a Binary outcome, where the outcome can be either success (Represented by value 1) or failure (Represented by value 0)

$$P(x=x) = P^{x}(1-P)^{1-x}$$

eg :- Tossing a coin once.

Scanned with OKEN Scanner

eg:- Rolling a dice - getting '5' is my experiment.

Success = '5' = 1.

failure = others = 0.

$\frac{1}{6}$ = Success-probability

$\frac{5}{6}$ = failure-probability

Binomial Distribution

It is a probability distribution that describes the number of success in a fixed no: of Independent bernoulli trials with 2 Possible outcomes where the probability of success is constant for each trial. It is characterised by 2 parameters.

(*) no: of trials = N.

(*) Probability of Success = P

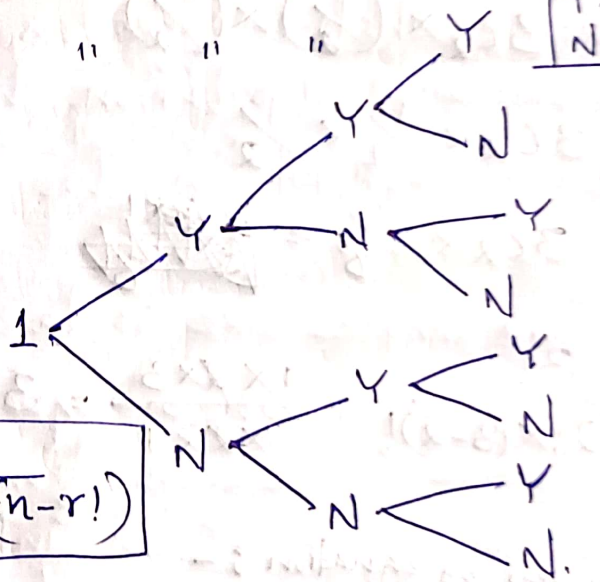
equation =

Q. The probability of anyone watching a Video and then liking it is 0.5. What is the Probability that -

1) No one outof 3 people like it.
2) 1 out of 3 people " "
3) 2 " " " "
4) 3 " " " "

Y = Success
N = failure

Ans -



8 outcome

YYY
YYN
YNY
YNN
NYY
NYN
NNY
NNN.

combination equation

$$\boxed{ncr = \frac{n!}{r! \times (n-r!)}}$$

1) $\frac{1}{8}$
2) $\frac{3}{8}$
3) $\frac{3}{8}$
4) $\frac{1}{8}$

$$\boxed{P(x=x) = {}^{n}c_{x}\, p^{x} (1-p)^{n-x}}$$

Scanned with OKEN Scanner

3rd answer by equation :-

$$nc_x \times p^x (1-p)^{n-x}$$

$$3c_2 \times \left(\tfrac{1}{2}\right)^2 \times \left(\tfrac{1}{2}\right)^{3-2}$$

$$= 3c_2 \times \tfrac{1}{4} \times \tfrac{1}{2}$$

$$3c_2 \times \tfrac{1}{8} =$$

$$\frac{3!}{2! \times (3-2)!} = \frac{1 \times 2 \times 3}{1 \times 2 \times 1} = 3 = \tfrac{3}{8}.$$

4th answer by equation :-

$$3c_3 \times \left(\tfrac{1}{2}\right)^3 \times \left(1-\tfrac{1}{2}\right)^{3-3}$$

$$3c_3 \times \left(\tfrac{1}{8}\right) \times \left(\tfrac{1}{2}\right)^0$$

$$3c_3 \times \tfrac{1}{8} \times 1$$

$$3c_3 \times \tfrac{1}{8} = \frac{3!}{3! \times (3-3)!} = 1$$

$$1 \times \tfrac{1}{8} = \tfrac{1}{8}$$

Criteria for Binomial distribution :-

1. The process consist of n trials.
2. Only 2 exclusive outcomes are possible
   A Success & A failure
3. P(success) = P & P(failure) = 1 − P and its is fixed from trial to trial.
4. The trials are Independent.

applications in Data science.

1. Binary classification problems. (Spam/Not spam)
2. Hypothesis testing.

Sampling distributions :-

The distribution of mean of samples.

Sample 1 = x^1
Sample 2 = x^2
⋮
Sample 100 = x^{100}.

} equal sized Samples.

$(\bar{x}^1, \bar{x}^2 \ldots \bar{x}^{100})$ → Sample distribution.

This distribution can be any statistical variable (mean, variance, std etc.)

Inferential statistics :-

Central limit Theorem :-

It states that, for a sufficiently large sample size, the distribution of the sample mean will tend to follow a normal distribution regardless of the original population distribution shape.

The condition required for the CLT to hold true are :-

1) The ~~sma~~ Sample size is large enough typically greater than or equal to 30 :->
2) The Sample is drawn from a finite population or an infinite population with a finite variance.

~~3) The random variable~~

Mathematical Expression :-

If $x_1, x_2 \ldots x_n$ are independent and identically distributed random variables with mean, ~~mu~~ and variance sigma square. Then the sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution as the sample size 'n' increases.

$$\bar{x} \text{ approximately } N\left(\mu, \frac{\sigma^2}{n}\right)$$

Where $\bar{x}$ - Sample mean.

μ = population mean

σ^2 = population variance

n = sample size.

~~Sample size must be sufficiently large for the CLT to apply. Sample size depends on the population.~~

Characteristics of a population - Parameter
Characteristics of a sample - statistic

A population is the entire group or ~~set~~ set of individuals, objects or events that a researcher wants to study or draw conclusions about.

A sample is a subset of population that is selected for study. It is a smaller group that is intended to be representative

of the larger population.

A parameter is a numerical value that describes a characteristics of a population, denoted ~~such as~~ using greek letters like μ, σ.

A statistic is a numerical value that describes the characteristics of a sample which is a subset of population, denoted as $\bar{x}$.

~~Confidence int~~

Point estimate :-

It is a single value calculated from a sample that serves as the best guess or approximation for an unknown population parameter such as mean or standard deviation. These are often used in statistics when we want to make inferences about a population based on a sample.

Confidence Interval :-

It is a range of values within which we expect a particular population parameter like a mean to fall. It is a way to express the uncertainity around an estimate obtained from a sample of data.

2 ways to calculate confidence interval :-

(1) z procedure :-

We use this when we know the population's standard deviation. (σ known)

Assumption :-

(*) Random sampling :- The data must be collected ~~as a whole~~ using a random sampling method to ensure that the sample is representative of the population.

(*) ~~Known~~ Known population standard deviation :- The population's standard deviation must be known or accurately estimated. (often unknown in most real cases)

(*) The 'z' procedure assumes that, underline population is normal distribution. However, if the Population distribution is not normal, the CLT can be applied if the sample size large enough.

Interpreting Confidence Interval :-

Confidence Interval is a range of values within which a population parameter is estimated to lie with a certain level of confidence.

Confidence level :-

It commonly set at (90, 95, 99) represent the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times.

for eg :- At 95% Confidence Interval means, that if you were to draw 100 different samples from the population and calculate confidence interval for each, approximately 95% of those intervals would contain the true population parameter.

Confidence Interval using 'z' test :-

$$CI = \text{Point Estimate} \pm \text{margin of error}$$

$$= \underset{\text{Point estimate}}{\bar{x}} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

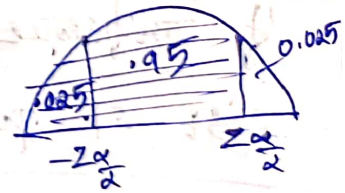
σ - Population standard deviation.

n - no. of samples.

For eg :- 95% Confidence level means $1-\alpha = 95\%$ or $.95$

$$Z_{\frac{\alpha}{2}} = 1 - .95 = .05$$

$$= \frac{.05}{2} = 0.025$$



$$1 - 0.025 = \underline{0.975}$$ (have to look this value in the z table)

So, z score = $\underline{1.96}$

So, the equation will be $= \bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$.

Another eg :-

75% Confidence level

Mean $1-\alpha = .75$

$Z_{\frac{\alpha}{2}} = 1 - .75 = .25$

$\frac{.25}{2} = 0.125$

$1 - 0.125$

$= 0.875$

Z score = $\underline{\underline{1.16}}$

So, the equation =

$$\bar{x} \pm 1.16 \times \frac{\sigma}{\sqrt{n}}$$

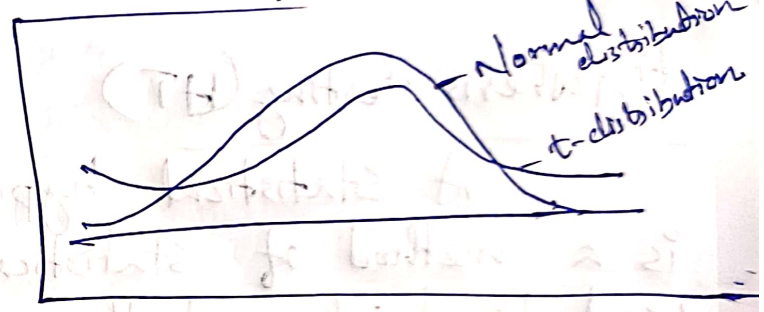
If ~~confidence~~ population std not known :-

CISE T-procedure :-

$$CI = \bar{x} \pm t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

α - std of sample

the t-distribution is specifically designed to account for the additional uncertainity introduced by using the sample standard deviation instead of population standard deviation.



t-distribution is a theoritical distribution.

T-distribution is ~~families~~ similar to normal distribution, but has heavier tails. The shape of the T-distribution is determined by the degrees of freedom, which is closely related to the sample size.

$\boxed{\text{Sample size} - 1 = \text{degree of freedom}}$

As the degree of the freedom increases, the t-distribution approaches the normal distribution. The T-distribution

accounts for the additional uncertainity that arises from estimating the population standard deviation using sample standard deviation.

Hypothesis testing (HT)

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. This allows us to make probabilistic statements about population parameters.

HT consists of 2 contradictory statements :-

(1) Null hypothesis :- H_0

In a null hypothesis, we claim that there is no relationship or difference b/w groups with respect to the value of population parameter.

(2) Alternative Hypothesis (H_1 / H_a) :-

In this we claim that there is some change or relationship between the groups with respect to the value of population parameter.

we begin the hypothesis test by assuming that the null hypothesis to be true and later we retain or reject null hypothesis based on the evidence found in sample data. An alternative hypothesis always contradict null hypothesis. Only anyone of the above hypothesis could be true.

The purpose of hypothesis testing is to gather evidence to either reject or fail to reject the null hypothesis in favour of ~~the~~ alternative hypothesis

which claims there is a significant effect or relationship.

Important Points :-

→ How to decide, what will be Null hypothesis & what will be alternative hypothesis.

→ We try to gather evidence to reject the null hypothesis.

→ It is imp to note that failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true. It just means that, there isn't enough evidence to support the alternative hypothesis.

eg:- Hypothesis tests are similar to jury trials, In a sense. In a jury trial, H_0 is similar to the not guilty verdict and H_a is the guilty verdict. You assume in a jury trial that the defendant isn't guilty unless the prosecution can show beyond a reasonable doubt that he or she is guilty. If the jury says the evidence is beyond a reasonable doubt, they reject H_0, not guilty, in favour of H_a, guilty.

Steps involved in Hypothesis testing :-

(#) Rejection Region approach :

→ formulate a null and alternative hypothesis.

→ ~~Second~~ Select a significance level (α) (This is the probability of rejecting the null hypothesis when it is actually true usually set at .05 or .01.

→ check assumptions

→ Decide which test is appropriate.

→ State the relevant test statistic.

→ Conduct the test
→ Reject or not reject the null hypothesis.
→ Interpret the result.

Q. Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The Company has data on the average Productivity of its employees before implementing the training program. The average productivity was 50 units per day with a known population standard deviation of 5 units. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The Sample has an average Productivity of 53 units per day.

The company wants to know if the new training program has significantly increased productivity?

Answer

Null :- H_0

The training program has not made any impact on Productivity $\boxed{\mu = 50}$

Alternative :- H_a

The training program has made a impact on productivity $\boxed{\mu = >50}$

Significance level $\alpha = 0.05$ $\boxed{5\%}$

standard deviation of Population $= 5$

$n = 30$

mean of Sample $= \bar{x} = 53$

Population mean $= \mu = 50$

So, we Select z test, coz σ is known.

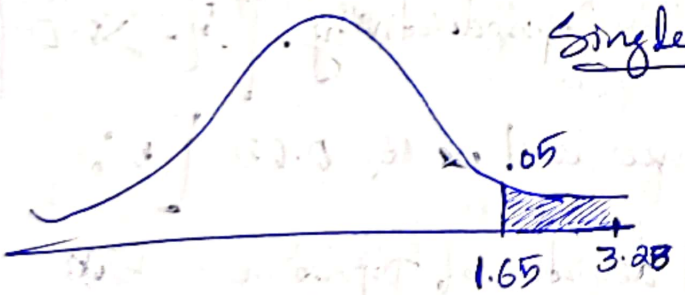
$$z \text{ score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{53 - 50}{5/\sqrt{30}}$$

$$= \frac{3}{0.912}$$

$$= 3.28$$

Single tailed:



(.95 in z table)

As the value falled in the rejection area, we reject the null hypothesis, which means the new training Program is effective.

Q- Suppose a snack food company claims that their Lays Wafes packets contain an average weight of 50 grams per Packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays Wafes packets. The organisation wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays Wafes packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.

Scanned with OKEN Scanner

Answer.

H_0 – Lays wafers packets contains 50 gsms.
$\mu = 50$

H_a – Lays wafers packets doesnt contain exactly 50 grams. $\mu \neq 50$

Two tail test

$\mu = 50$

Significance level $\alpha = .05$ (5%)

Population std deviation – $\sigma = 4$.

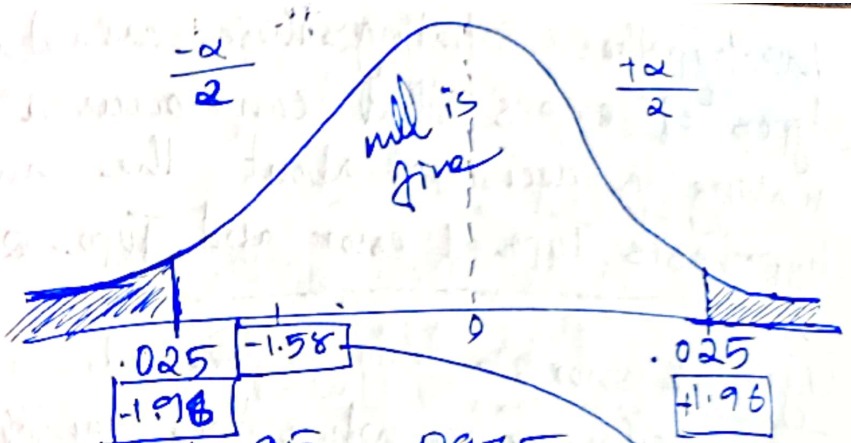
$n = 40$.

$\bar{x} = 49$

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

$$= \frac{49 - 50}{4/\sqrt{40}} \quad \frac{-1}{.6324} = -1.5812$$

Z value



So, $0.25 + .95 = 0.975$

Z score = 1.96

As it occured in the area of not critical – lets conclude that the null is fine.

Significance level :-

Denoted as 'α' is a predetermined threshold used in hypothesis testing to determine weather the null hypothesis should be rejected or not. It is represent the probability of rejecting the null hypothesis when it is actually true.

Scanned with OKEN Scanner

In hypothesis testing, there are two types of errors that can occur when making a decision about the null hypothesis. Type-1 error and Type-2 error.

Type-1 error :-

This occurs when the sample results, lead to the rejection of the null hypothesis when it is in fact true. In other words, It's a mistake of finding a significant effect or relationship when there is none. The probability of committing a Type-1 error is denoted by α (alpha), which is also known as the significance level. By choosing a significance level, researchers can control the risk of making a type-1 error.

Type-2 error :-

This error occurs when based on the sample results. The null hypothesis is not rejected when it is in fact false.

	H_0 True	H_0 False
Reject H_0	Type 1 error	Correct decision
Accept H_0 or Fail to reject	Correct decision	Type-2 error

One sided v/s two sided test :-

A one sided test is used when the researches is interested in testing the effect in a specific direction (either greater than or less than the value specified in the null hypothesis). The alternative hypothesis in a one sided test contains an inequality (either '>' or '<'). eg :- A researches wants to test whether a new medication increases

the average recovery rate compared to the existing medication.

2-sided Test :-

It is used when the researcher is interested in testing the effect in both directions. (i.e, whether the value specified in the null hypothesis is different, either greater or lesser). The alternative hypothesis in a 2 sided test contains a "not equal to" sign (≠).

Example :- A researcher wants to test whether a new medication has a different average recovery rate compared to the existing medication.

P-Value :-

It is the probability of getting a sample as or more extreme (having more evidence against H_0) than our own sample given the null hypothesis (H_0) is true.

eg :- Imagine, you are flipping a coin and you believe it is a fair coin. (This is our null hypothesis). You flip it 10 times and get 8 heads. "you start wondering, Is this coin actually biased or could this just be by chance or luck ?!"

The P-Value tells you how likely it is to get 8 heads (or more) out of 10 flips if the coin were truly fair.

Think of 'a P-Value as a measurement of surprise :-

- If the P-Value is small, It means that the ~~rem~~ result is surprising if the null hypothesis is true. It suggests that what we observe is unlikely to happen just by random chance. So, may be the null hypothesis

is not true.

- If the P-Value is large, it means that the result is not surprising. It suggests that the observed data could easily happen just by random chance. So, we dont have enough reason to doubt the null hypothesis.

eg:- Lets say we have a study that claims a new drug doesn't work (null hypothesis). You conduct an experiment and find that the patients who took the drug did better than those who didn't. You calculate the P-Value to see how likely this difference is if the drug really doesn't work.

If the P-Value is 0.01 (1%), it means that there is only a 1% chance that the observed difference happened by random chance. This suggests that the drug probably does work.

If the P-Value is 0.4 (40%), it means there is a 40% chance the difference is just due to random chance. This suggests that the drug probably doesn't work.

A smaller P-Value less than 'α' [significance level]

$$\boxed{\text{P-Value} <= \alpha}$$

indicates the strong evidence against null hypothesis. So, you reject null hypothesis.

So, a large P-Value > α will be failing to reject the null hypothesis.

Q- $\mu = 50$

$\alpha = 0.05$ (5%)

$n = 30$

$\bar{x} = 53$

$\sigma = 5$

Q1 before 5 pages

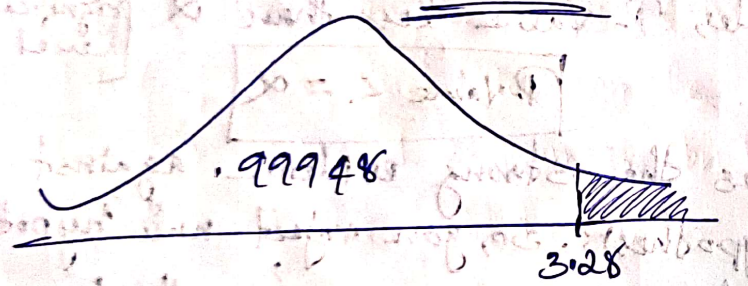
H_0 = Training has no impact on productivity $\mu = 50$

H_a = Training has impact on productivity $\mu = >50$.

$$z \text{ score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{53-50}{5/\sqrt{30}}$$

$$= 3.28$$



look the area of 3.28 in z-table.
the area is .99948.

So P-Value = 1 − .99948

P-Value = 0.00052

$\alpha = .05$

As - P-Value is less than "α", lets reject the null hypothesis.

Q. $\boxed{\text{\#2 before 5 Pages}}$

H_0 - lays packets contains 50 grams.

$\mu = 50$

H_a - lays packets doesn't contain exactly 50 grams.

$\mu \neq 50$

$\mu = 50$

$\alpha = .05$

$\sigma = 4$

$n = 40$

$\bar{x} = 49$

$$z \text{ score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{49-50}{4/\sqrt{40}}$$

$$= -1.5812$$

area = .05705

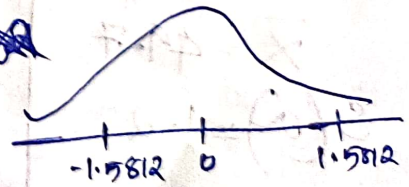
P-Value = 1 − [illegible]

= [illegible]

P-Value = .05705 × 2

= [illegible]

= 0.1141 > α = So, we fail to reject null hypothesis



Single Sample t-test

Q. Suppose a manufactures claims that the average weight of their new chocolate bars is 50 grams, we highly doubt that and want to check this so we drew out a sample of 25 chocolate bars and measured their weight, the sample mean came out to be 49.7 grams and the sample standard deviation was 1.2 grams. Consider the significance level to be 0.05.

$\mu = 50$

$\alpha = .05$

$n = 25$

$\bar{x} = 49.7$

$S(\sigma) = 1.2$

degree of freedom $= n-1 = 24$

$$t\text{-value} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{49.7 - 50}{\frac{1.2}{\sqrt{25}}}$$

$$= -1.25$$

P-Value = .2233

So, P-Value > α, so we fail to reject null hypothesis.

From the python code we got to know that the area to the left of -1.25 is .11, since it is a 2 tail test, P-Value will be .11 × 2 = .22

T-Test

A T-test is a statistical test used in hypothesis testing to compare the means of two samples or to compare a sample mean to a known population mean. The t-test is based on the t-distribution, which is used when the Population standard deviation is unknown and the sample size is small.

There are 3 main types of t-tests :-

One Sample t-test :

The one-sample t-test is used to compare the mean of a single sample to a known population mean. The null hypothesis states that there is no significant difference

between the sample mean and the population mean, while the alternative hypothesis states that there is a significant difference.

(*) Independent two-sample test :-

The Independent two-sample test is used to compare the means of two independent samples. The null hypothesis states that there is no significant difference between the means of the two samples, while the alternative hypothesis states that there is a significant difference.

(*) Paired T-test (dependent two sample t-test) :-

It is used to compare the means of two samples that are dependent or paired, such as Pre-test and post-test scores for the same group of subjects or measurements taken on the same subjects under two different conditions. The null hypothesis states that there is no significant difference between the means of the paired differences, while the alternative hypothesis states that there is a significant difference.

A one sample t-test checks whether a sample mean differs from the population mean.

Assumptions for a single sample t-test :-

1- Normality :-

Population from which the sample is drawn is normally distributed.

2- Independence :-

The observations in the sample must be independent, which means that the value of one observation should not influence the value of another

observation

3.- Random Sampling :-

The sample must be a random and representative subset of the population.

4.- Unknown population std :-

The population std is known.

When we confirm the normality assumption of t-test, if the sample is less than 30, we have to use the Shapiro wilk test. If the P-Value is greater than α, then we can conclude that the samples we drawn from the population assumes normality.

Independent & Sample t-test :-

It is also known as an unpaired t-test. It is a statistical method used to compare the means of two independent groups, to determine if there is a significant difference b/w them.

Assumptions :-

(*) Independence of observation;
(*) Random Sampling
(*) Normality.
(*) Equal Variance. - Varience of two Populations should be approximately equal.

Question :-

Suppose a website owner claims that there is no difference in the average time spent on their website between desktop & mobile users. To test this claim, we collect data from 30 desktop users and 30 mobile users.

Desktop users :-

$n = 30$

$\bar{x}_1 = 18.5$

$\sigma = 3.5$ mins

$\alpha = .05$

Mobile users :-

$n = 30$

$\bar{x}_2 = 14.3$

$\sigma = 2.7$ minutes.

$\alpha = .05$

H_0 = average of desktop users = average of mobile users

$\mu_1 = \mu_2$

$H_a = \mu_1 \neq \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

degree of freedom

$n_1 + n_2 - 2$

$= 30 + 30 - 2$

$= 58$

$$= \frac{18.5 - 14.3}{\sqrt{\frac{3.5^2}{30} + \frac{2.7^2}{30}}}$$

$= 5.21$ - Put it in the python code for finding cdf & P-value

P-value = 1.999

As P-value is > α, we reject null hypothesis.

Paired t-test :-

Q. Let's assume that a fitness centre is evaluating the effectiveness of a new 8-week weight loss program. They enroll 15 participants in the program and measure their weights before and after the program. The goal is to test whether the new weight loss program leads to a significant reduction in the participants' weight.

Before the program :-

[80, 92, 75, 68, 85, 78, 73, 90, 70, 88, 76, 84, 82, 77, 91]

After the program :-

[78, 93, 81, 67, 88, 76, 74, 91, 69, 88, 77, 81, 80, 79, 88]

H_0 = There is no significant difference in the participants, weights before and after the program, ie, the weight loss is not significant.

H_1 = There is a significant reduction in the the participants weights after the program.

Steps :-

1- Calculate the difference b/w 'before' and 'after'.
2- Calculate the mean of the differences
3- Calculate the 'σ' of the differences.
4- Compare the statistics using the formula

$\bar{x}$ diff =	mean of difference
Sd diff / $\sqrt{n}$ =	std of differences / no. of participants.

5- Compare the calculated t-value with the critical t-value from the t-distribution table for n-1 degrees of freedom (14) at a chosen significance level $\alpha = 0.05$.

Difference :- ① Substract each 'after' weight from the corresponding "before" weight.

[80-78, 92-93, 75-81, 91-88] = [2, -1, -6, 1, -3, 2, -1, -1, 1, 0, -1, 3, 2, -2, 3]

② Mean of the difference :- $\bar{d} = \frac{2-1-6+1-3+2-1-1+6-1+3+2-2}{15}$

$= 1/15 \approx 0.067$

③ Standard deviation = 2.463

SE = 0.636

t-statistics = -0.105

P-Value = 0.918.

P-value > α, fail to reject the null hypothesis. No significant reduction in the participants after program.

Q- We want to determine whether there is an association between gender and movie preference (Action, Comedy or Drama). Specifically, we aim to find out if the preferences for these movie genres differ between males and females using chi-square test of independence.

H_0 = There is no association between gender & movie preference.

H_1 = There is an association between gender and movie preference.

Gender	Action	Comedy	Drama	Total
male	20	15	10	45
female	10	25	20	55
Total	30	40	30	100

Scanned with OKEN Scanner

chi-square is a statistical method used to determine if there is a significant association between two categorical variables. It evaluates how the observed frequency distribution of a sample matches the expected distribution based on a hypothesis.

Step-2

Calculate expected frequency

expected frequency for each cell is calculated using the formula

$$E_{ij} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

	Action	comedy	Drama
Male	13.5	18	13.5
female	16.5	22	16.5

$$E_{male\ action} = \frac{45 \times 30}{100} = 13.5$$

$$E_{male\ comedy} = \frac{45 \times 40}{100} = 18$$

$$E_{male\ Drama} = \frac{45 \times 30}{100} = 13.5$$

$$E_{female\ action} = \frac{55 \times 30}{100} = 16.5$$

$$E_{female\ comedy} = \frac{55 \times 40}{100} = 22$$

$$E_{female\ Drama} = \frac{55 \times 30}{100} = 16.5$$

Step-3

compute chi-square statistics.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} — observed frequency

E_{ij} — Expected frequency

$$\text{For male action} = \frac{(20-13.5)^2}{13.5} = 3.12$$

$$\text{For male comedy} = \frac{(15-18)^2}{18} = .5$$

Scanned with OKEN Scanner

For male drama $= \frac{(10-13.5)^2}{13.5} = .907$

for female action $= \frac{(10-16.5)^2}{16.5} = 2.56$

for female Comedy $= \frac{(25-22)^2}{22} = .40$

for female Drama $= \frac{(20-16.5)^2}{16.5} = .74$

$$X^2 = 8.22$$

Step 4

Find the degree of freedom.

$$\underbrace{(R-1)}_{\text{Row}} \times \underbrace{(C-1)}_{\text{Column}}$$

$= 1 \times 2 = 2$.

we found that, with the help of Python code — P Value $= 0.016$

P value $< \alpha$ — So we reject the null hypothesis.

Scanned with OKEN Scanner