Statistics

Thursday, November 7, 2024 7:42 PM

Definition :- Statistics is the science of collecting, organizing and analyzing data.

Data : "facts or pieces of information"

Ex- Heights of students, IQ of students

Types of Statistics

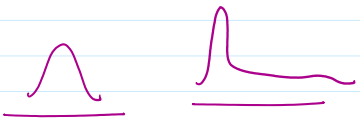
Descriptive Statistics

Defn :- It consists of organizing and summarizing data

1) Measure of central tendency [Mean, Median, Mode]

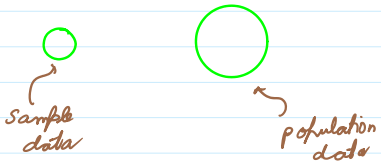
2) Measure of diversion [Variance, standard deviation]

3) Different type of distribution of data



Inferential Statistics

Defn :- It consists of using data you have measured to form conclusion.



1) Z test
2) T test
3) CHI square test

} hypothesis testing

Ex- Let say there are 20 statistics classes at your college. And you have collected the heights of students in the class. Height s are recorded 175cm , 180cm , 140cm, 140cm, 135cm, 160cm, 135cm, 190cm.

Descriptive question : What is the average height of the entire class → which means → mean

$175+180+140+140+135+160+135+190/8$

Inferential question : Are the height of the students in class similar to what you expect in the entire college

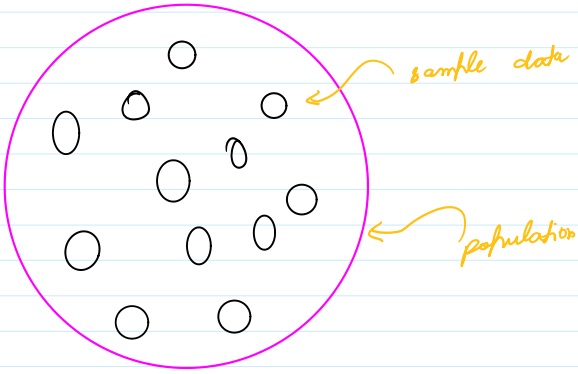
students in class → sample data

entire college → population data

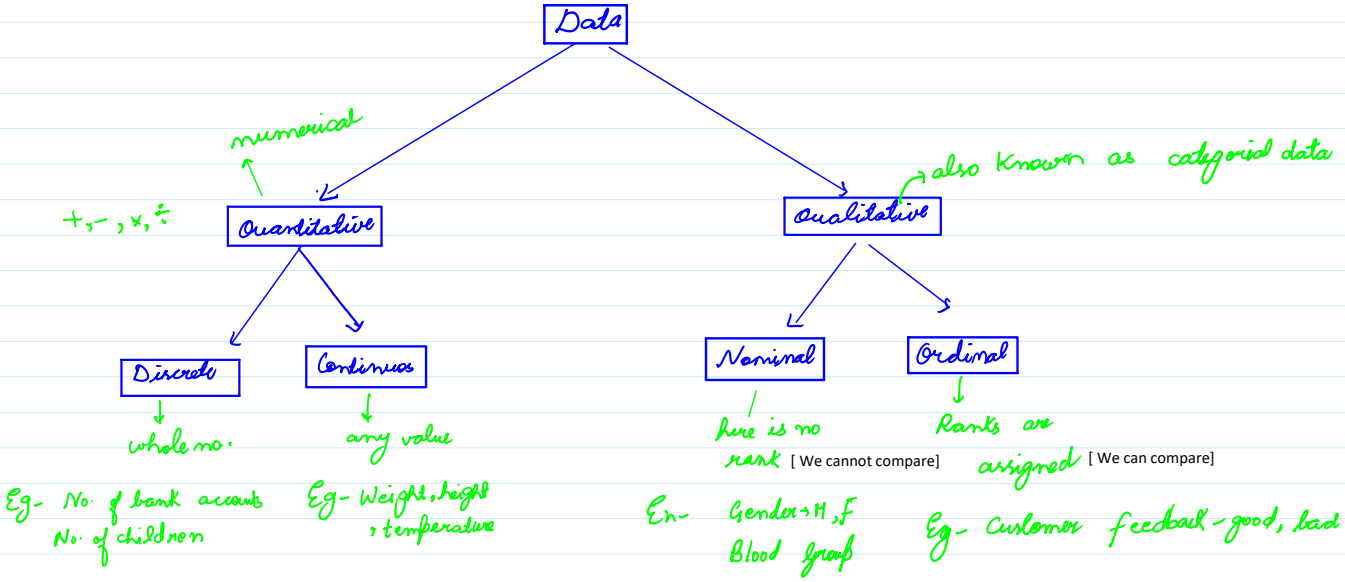
Population Data: The group you are in interest in studying

Sample Data : A subset of population

Ex-



Yha hamm raye lete hai sbki or uske basis pr hamm conclusion nikalte hai



DMC	DC	ISI	BUI	FWI	Classes	Region
2.4	4.6	1.3	3.4	0.5	Fire	0
4.1	3.4	1	3.9	0.4	Not fire	1
2.6	6.9	0.3	2.7	0.4	Fire	0

Scale of Measurement:

1) Nominal Scale data
2) Ordinal Scale data
3) Internal Scale data
4) Ratio Scale data

1. Nominal Scale Data

→ Includes qualitative / categorical data
→ Ex- Gender, Colors
→ Order does not matter

Survey of favourite color

Red → 5 Person

O D V

Red → 5 Person

Blue → 3 Person

orange → 2 Person

Order means there is no point to compare red is greater than blue and orange but we can conclude this 50% red color person opt, 30% blue color person opt and remaining 20% color orange opt

2. Ordinal Scale Data

→ Ranking is important
→ order matters
→ Difference cannot be measured.

Ex- Aftb with review Best, Good, Better

ranking → 1 → Best
2 → Good
3 → Better

Ex- In a Race

difference cannot be measure → 1st → Rahul
2nd → Shivam
3rd → Loveria

3. Interval Scale Data

→ Order matters
→ Difference can be measured
→ Ratio cannot be measured
→ No "0" starting point

Ex- Temperature variable

30 F°
60 F°
90 F°

30 : 90 = 1 : 3

yha hamm yeah nhi keh skte 90°f mai 3 times garmi lag rahi hogi 30°f comparison mai

No 0 starting point → yeah negative se bhi start hob skta hai

4. Ratio Scale Data

→ Order matters
→ Difference are measurable including ratio
→ Contain "0" starting point

Ex- Students marks of a class

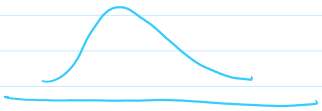
Ex- Students marks of a class

Random Quiz ?

1) Length of different rivers in the world → Ratio Scale Data
2) favourite food based on gender → Nominal scale data

Measure of Central Tendency

1) Mean / Average
2) Median
3) Mode



1. Mean

Population (N) Sample (n)

$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\text{Sample } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\text{Population Mean } (\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$$

$$= \frac{[1+1+2+2+3+3+4+5+5+6]}{16} = 3.2$$

2. Median

$X = \{4, 5, 2, 3, 2, 1\}$

Step 1: sort the Random variable $\{1, 2, 3, 4, 5, 6\}$

Step 2: No. of elements count = 6

Step 3: if count == even

$\{1, \boxed{2, 3}, 4, 5, 6\}$

$\text{Median} = \frac{2+3}{2} = 2.5$

Step 4: if count == odd

$\{1, 2, \textcircled{3}, 4, 5\}$ ← median

Why we use median

$X = \{1, 2, 3, 4, 5\}$

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

$X = \{1, 2, 3, 4, 5, \textcircled{100}\}$ → this large no. is called outlier

$\bar{x} = \frac{1+2+3+4+5+100}{6}$

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

$= 3$

$\bar{x} = \frac{1+2+3+4+5+100}{6}$

~ 19

Important

yha ekk large no. 100 add krne par kitna zyada shift hoh gya mean mai ekdum isliye study median

$X = \{1, 2, 3, 4, 5\}$

Median = 3

$X_2 = \{1, 2, 3, 4, 5, 100\}$

Median = $\frac{3+4}{2} = 3.5$

3 → 3.5

↳ not large shift in median after adding outliers

- Median is used to findt the central tendency when outliers is present

3. Mode → Maximum frequency

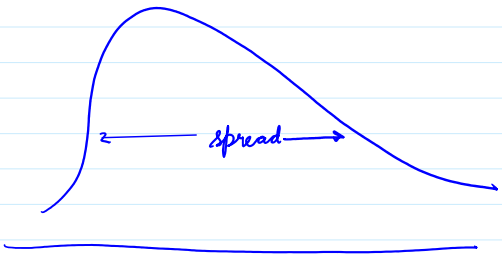
$X = \{1, 1, 1, 7, 2, 3\}$

Mode = 1

- We are learning mean , median and mode toh yeah kha prr use hoga industry main EDA and feature Engineering

Go throught the Measure_of_Central_Tendency_Code In google Colab

Measure of Dispersion [Spread of the data]



1. Variance

Population variance

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

x_i → Data points

Sample Variance

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

Ques : Why we divide Sample variance by n-1 ?
Ans. So that we can create an unbiased estimator of the population variance

this scenerio called

x_i → Data points

$\bar{x}$ → Sample mean

x_i → Data points
μ → Population mean
N → Population size

this scenerio called Bessel correction

x_i → Data points
$\bar{x}$ → Sample mean
n → Sample size

Ex- $\{1, 2, 3, 4, 5\}$

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

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

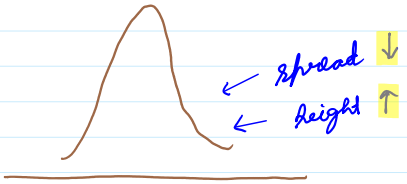
x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
		10

$$\delta^2 = \frac{10}{5-1} = \frac{10}{4} = 2.5$$

what does means let us understand

$x = \{$ $\}$ → $\delta^2 = 2.5$

$y = \{$ $\}$ → $\delta^2 = 7.5$



we can conclude this from δ^2

2. Standard Deviation

Population standard deviation

$\delta = \sqrt{\text{variance}}$

Sample standard deviation

$std = \sqrt{s^2}$

s^2 = sample variance

- Explore python codes based on variance ,standard deviation,

Random Variable

- Random Variable is a process of mapping the output of a random process or experiments to a number

Ex- Tossing a coin

yeah result hai head/tail outcome ka

Y [ ]

Ex- Tossing a coin

yeah result hai head/tail outcome ka

$$X = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tails} \end{cases}$$

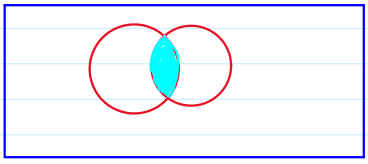
Sets

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7\}$

1. **Intersection**

$A \cap B = \{3, 4, 5, 6, 7\}$

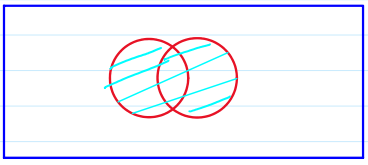


2. **Union**

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7, 9, 10\}$

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 16\}$

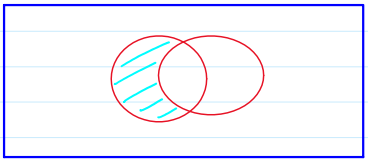


3. **Difference**

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7\}$

$A - B = \{1, 2, 8\}$



4. **Subset**

A C B ×
B C A ✓

5. **Superset**

A ⊃ B
B ⊃ A

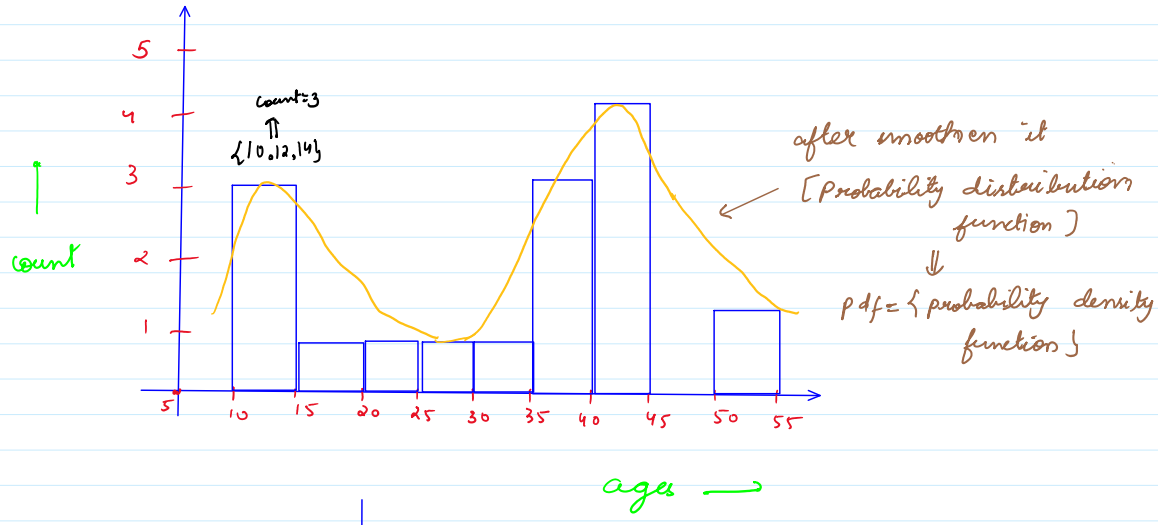
Histogram And Skewnes

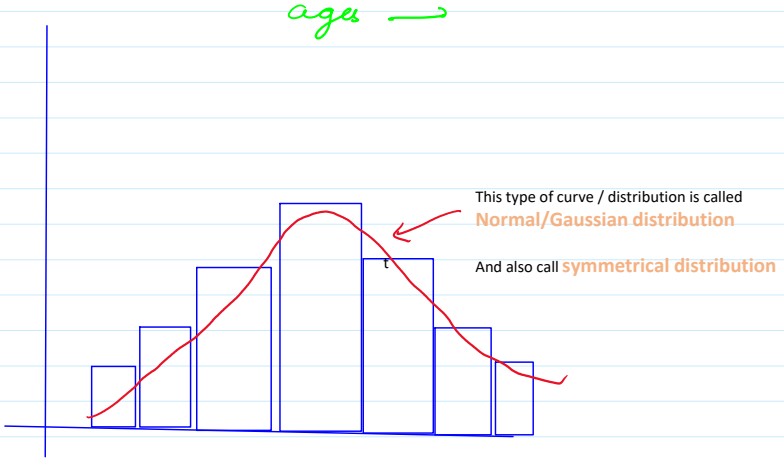
Histogram is related to frequency.

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

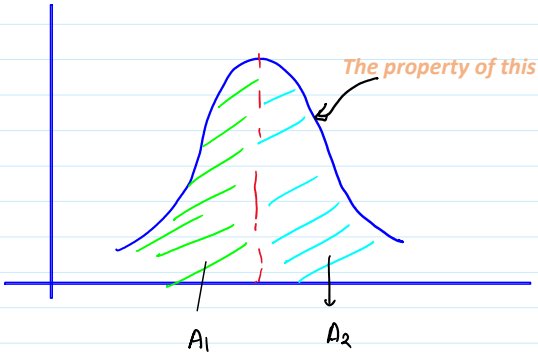
$\frac{50}{10} = 5$ → bin size { No. of bins = 10}

$\frac{50}{20} = 2.5$ bin size { No. of bins = 20}





1. **Superset**

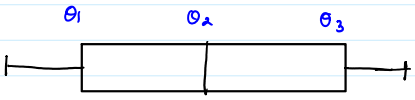


The property of this curve is mean , median and mode is at the center

mean = median = mode

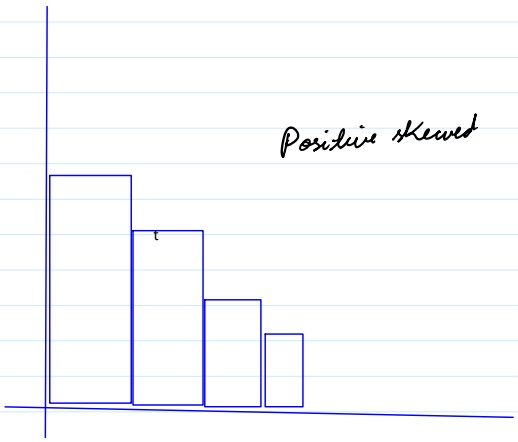
$A_1 = A_2$ area

Box plot:→ used for finding outliers



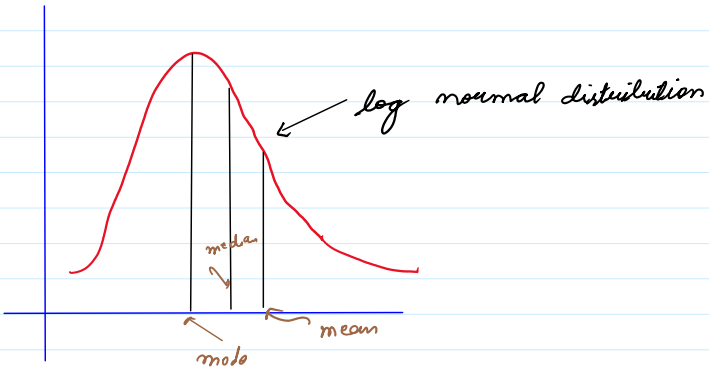
$Q_3 - Q_2 \simeq Q_2 - Q_1$

2. **Right Skewed**



Positive skewed

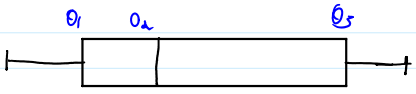
$\Rightarrow$



In this what is the relation b/w mean. median mode

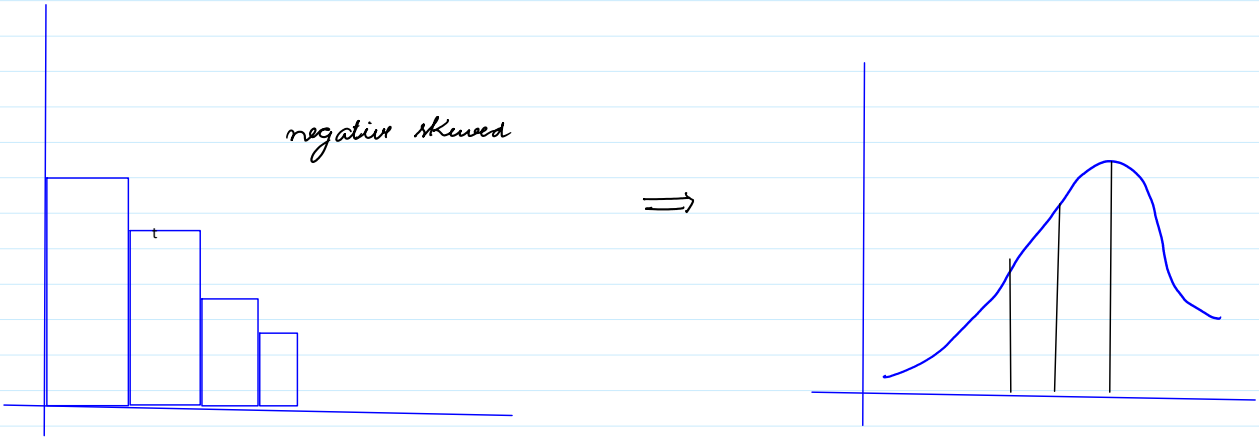
mean > median > mode

Box plot:



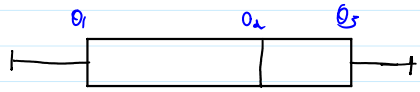
$Q_3 - Q_2 \geq Q_2 - Q_1$

3. **Left Skewed/Negative Skewed**



Mean < median ≤ mode

Box plot:



$Q_2 - Q_1 \geq Q_3 - Q_2$

Covariance And Correlation

What is the Relationship between X & Y

X	Y
2	3
4	5
6	7
8	9

X↑	Y↑
X↑	Y↓
X↓	Y↑
X↓	Y↓

← we can find this relationship / dependency using correlation & covariance

Covariance

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

$x_i \rightarrow$ Datapoints of x

$\bar{x} \rightarrow$ sample mean

$y_i \rightarrow$ Datapoints of y

$\bar{y} \rightarrow$ sample mean of y

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

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

$= cov(x,x) \Rightarrow$ It will tell about Spread

Observe the covariance and try to find the the relationship

$Cov(x,y)$

x↑	y↑	→ +ve covariance
x↑	y↓	→ -ve covariance

$Cov(x, y)$

$x \uparrow$	$y \uparrow$
$x \downarrow$	$y \downarrow$

→ +ve covariance

$x \uparrow$	$y \downarrow$
$x \downarrow$	$y \uparrow$

→ -ve covariance

Ques :

x	y
2	3
4	5
6	7
$\bar{x} = 4$	$\bar{y} = 5$

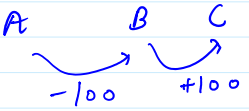
$$cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \frac{[(2-4)(3-4) + (4-4)(5-5) + (6-4)(7-5)]}{3-1}$$

$$= 4 \quad \text{+ve}$$

Conclusion → x & y having a positive covariance

Advantages	Disadvantages
➢ Relationship between X and Y +ve or -ve value	➢ Covariance does not have a specific limit [Interview Question]



To fix the specific limit of covariance we used :-

2. Pearson Correlation Coefficient [-1, 1] ← output

$$\rho_{x,y} = \frac{cov(x, y)}{\delta x \cdot \delta y}$$

- The more the calue towards +1 the more +ve correlation it is (x,y).
- The more the value towards -1 the more -ve correlated it is (x,y)

3. Spearman Rank Correlation [-1, +1]

Rank

$$\rho_s = \frac{cov(R(x), R(y))}{\delta(R(x)) \cdot \delta(R(y))}$$

Rank means more higher the value more higher the rank

x	y	$R(x)$	$R(y)$
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	2
0	7	6	2
8	1	1	6

Where we really used these above in Machine Learning i.e In Feature Selection

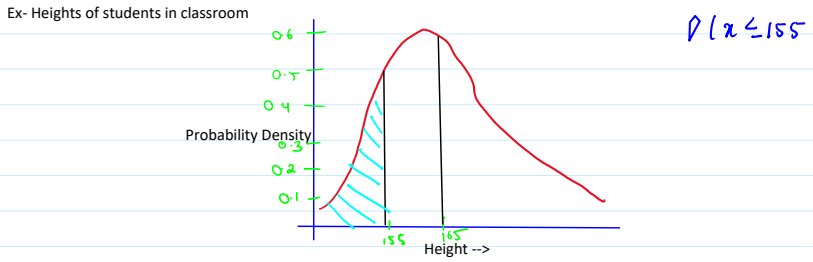
Code

```
import seaborn as sns
df=sns.load_dataset('healthexp') #load_dataset is a method in java and healthexp is a inbuilt datase
df.head()

import numpy as np
np.cov() # cov mean covariance
##Corelations using Pearson correlation coefficient
df.corr(method='pearson')
```

Probability Distribution Function / Density Function

1. Probability Density Function [PDF]

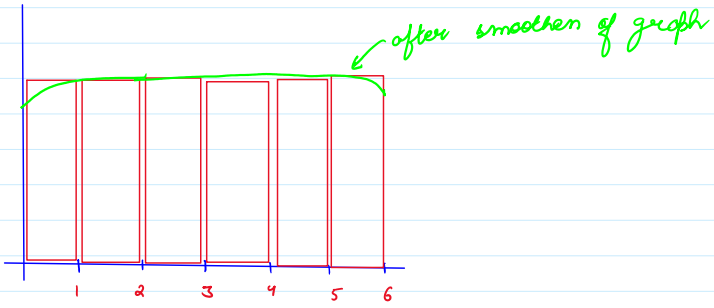


$P[x \leq 155$

2. Continuos Random Variable

Variable → Discrete Random variable

Ex- Rolling a Dice $\{1,2,3,4,5,6\}$



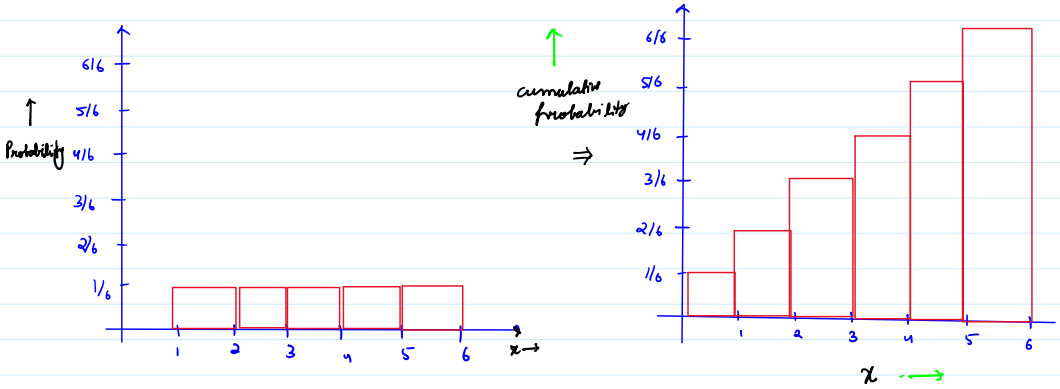
$P_H(1) = \frac{1}{6}$ $P_H(2) = \frac{1}{6}$

$P_H(X \leq 4) = P_H(X=1) + P_H(X=2) + P_H(X=3) + P_H(X=4)$

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

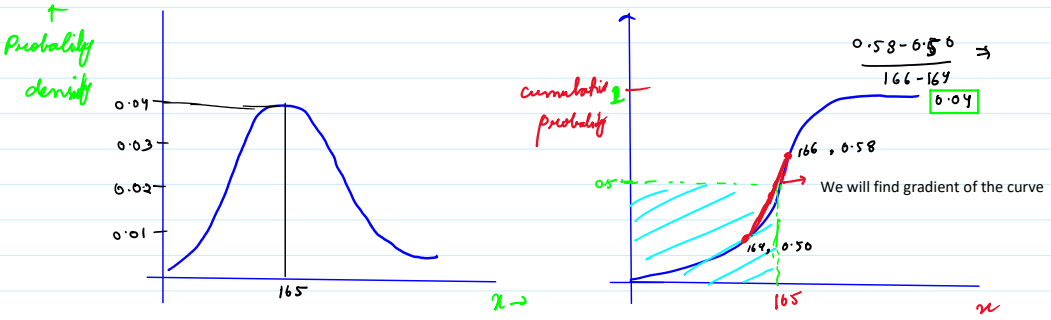
3. Cumulative Distribution Function [CDF]

Rolling a Dice $\Rightarrow \{1, 2, 3, 4, 5, 6\}$



Probability distribution Function

① Distribution of continuous random variable



Probability density => Gradient of Cumulative Curve

Types of Probability Distribution

1. **Normal / Gaussian Distribution** [pdf]
2. **Bernouli Distribution** [Pmf]
3. **Uniform Distribution** [Pmf]
4. **Poisson Distribution** [P mf]
5. **Log Normal Distribution** [PdF]
6. **Binomial Distribution** [pmf]

➢ Bernouli Distribution

In probability theory and statistics, the Bernoulli distribution, named after Swiss mathematics Jacob Bernoulli, is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q=1-p. Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes-no question.

→ Discrete Random variable (pmf)

a yes-no question.

→ Discrete Random variable (pmf)
→ outcomes are Binary

Ex - Tossing a coin {H, T}

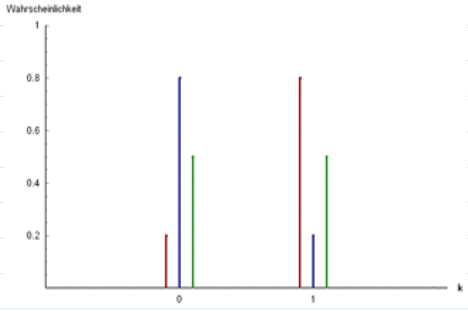
$P_r(head) = 0.5 = p$

$P_r(tail) = 1-p = q$

② whether the person will Pass / fail

$P_r(Pass) = 0.7 = p$

$P_r(fail) = 1-p = 0.3 = q$



$0 \le p \le 1$

$q = 1 - p$

$K = \{0, 1\}$ ← outcome

Probability Mass function

$$PMF = p^K * (1-p)^{1-K} \quad K \in \{0,1\}$$

if (K=1)

$$P_r(K=1) = p^1 (1-p)^{1-0}$$

$$= p$$

if K = 0

$$P_r(K=0) = p^0 * (1-p)^{1-0}$$

$$= (1-p) = q$$

We can write this in simplified Manner

$$\text{Probability} = \begin{cases} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{cases}$$

➢ **Mean of Bernoulli Distribution**

$$E(K) = \sum_{k=0}^{K} k \cdot p(k)$$

➢ **Median of Bernoulli Distribution**

$$\text{Median} = \begin{cases} 0 & \text{if } p < \frac{1}{2} \\ [0,1) & \text{if } p = \frac{1}{2} \\ 1 & \text{if } p > \frac{1}{2} \end{cases}$$

➢ **Variance of Bernoulli Distribution**

$$Var = p*(1-p)$$

$$= pq$$

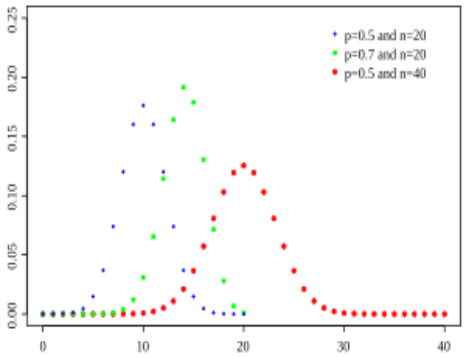
Standard Deviation of Bernoulli Distribution

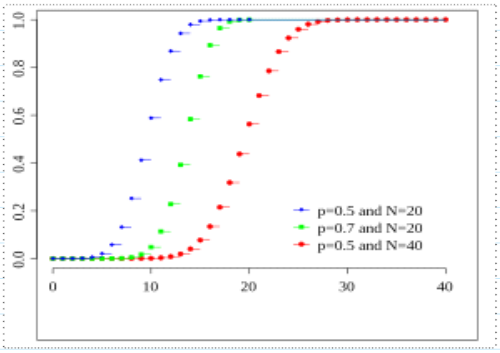
$$std = \sqrt{Var} = \sqrt{pq}$$

➢ **Binomial Distribution**

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence

of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome:Success(with probability p) or failure q=1-p. A single success/failure experiment is also called a Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial , i.e. n=1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.





→ Discrete random variable
→ Every experiment outcomes is binary
→ These experiment is performed for n trials.

Notation : $B(n, p)$

Parameters : $n \in \{0, 1, 2, 3 \dots\}$ → no. of trials

$p \in [0, 1]$ → success probability for each trial

$q = 1 - p$

Suffix : $k \in \{0, 1, 2, \dots n\}$ → No. of success

$$P_r(k, n, p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

Mean of Binomial Distribution

$$\text{Mean} = np$$

Variance of Binomial Distribution

$$\text{Var} = npq$$

Standard deviation of Binomial Distribution

$$sd = \sqrt{pq}$$

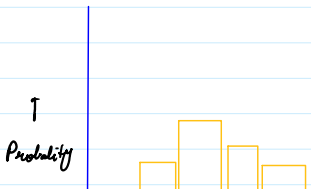
➢ Poisson Distribution

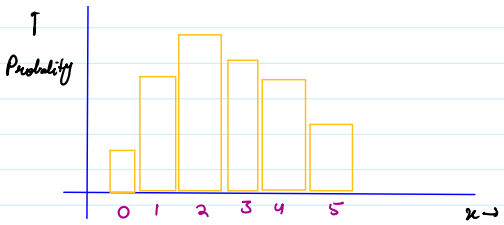
→ Discrete Random variables (pmf)
→ Describes the number of events occuring in a fixed time interval.

Ex

Ex- No. Of People visiting hospital/bank/airport every hour

$\lambda = 3$ **Expected no. of event occur at every time interval**





formula :-

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Ques : Based on the graph what is the probability that the patient will come at 5th hour ?

$[\because \lambda = 3]$ given

$$P(x=5) = \frac{e^{-3} 3^5}{5!}$$

$$= 10.1\ \%$$

Mean of Poisson Distribution

$$\text{Mean} = E(x) = \mu = \lambda t$$

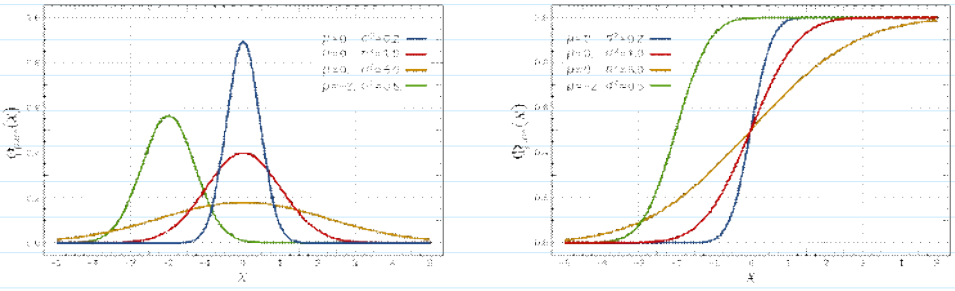
(μ : notation)

λ = Expected no. of envents to occur at every time interval

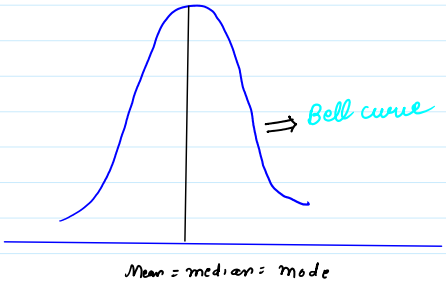
t=Time interval

➢ Normal/Gaussain Distribution [pdf]

In statistics , a normal distribution or Gaussain distribution is a type of continuos probability distribution for a real -valued random variable



In this curve is most symmetrical



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

Mean of Normal Distribution

Mean = μ : Average

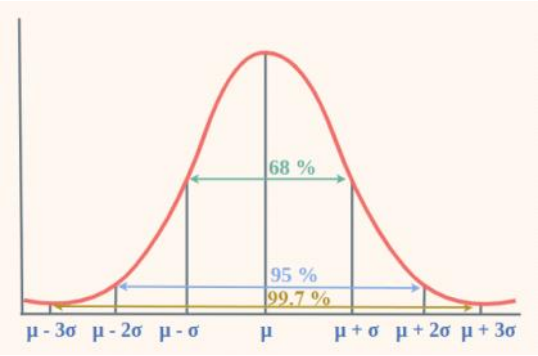
Variance of Normal Distribution

$Var = \sigma^2$

Standard deviation of Normal Distribution

$\sigma = \sqrt{Var}$

Empirical Rule for normal distribution → V. Imp → Interview



The **empirical rule**, also known as the **68-95-99.7 rule**, describes where most values lie in a normal distribution:

- Around **68%** of values are within **1 standard deviation** from the mean.
- Around **95%** of values are within **2 standard deviations** from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.

Probability

$P_R(\mu-\sigma \le x \le \mu+\sigma) \approx 68\%$

$P_R(\mu-2\sigma \le x \le \mu+2\sigma) \approx 95\%$

$P_R(\mu-3\sigma \le x \le \mu+3\sigma) \approx 99.7\%$

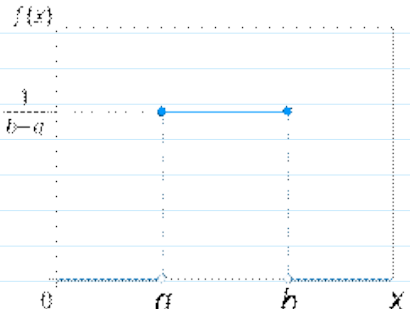
Ex- Weight of students
Heights of students

Uniform distribution

1. Continuos Uniform Distribution
2. Discrete Uniform Distribution

1. Continuos Uniform distribution

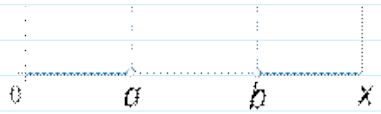
In probability theory and statistics, the continuos uniform distribution or rectangular distribution is a family of symmetric probability distribution. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters, a and b, which are the minimum and maximum values.



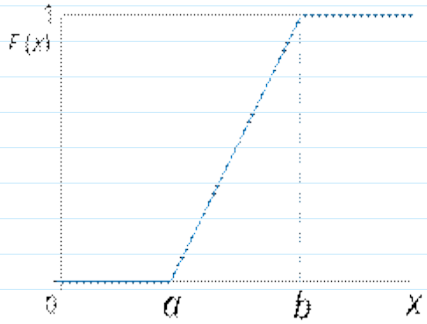
Notation :- $U(a,b)$

Parameters :- $-\infty < a < b < \infty$

$$pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



$$pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



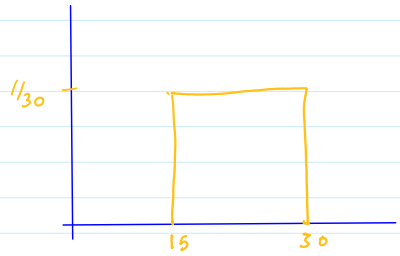
$$cdf = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$

$Mean = \frac{1}{2}(a+b)$ $\quad$ $Median = \frac{1}{2}(a+b)$ $\quad$ $Variance = \frac{1}{12}(b-a)^2$

Ex- The no. of candies sold daily at a shop is uniformly distributed with a maximum of 40 and a minimum of 10.

a) What is the probability of daily sales to fall between 15 and 30 ?

ans :



$$Pr(15 \le n \le 30) = (x_2 - x_1) \times \frac{1}{b-a}$$

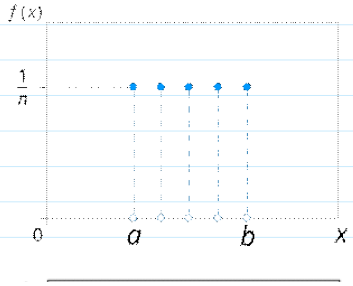
$$= 30 - 15 \times \frac{1}{30}$$

$$= 0.5 \simeq 50\%$$

$$Pr(n \ge 20) = \frac{40-20}{30} \times 1 = \frac{20}{30} = 0.6$$

2. Discrete Uniform distribution (pmf) ✓

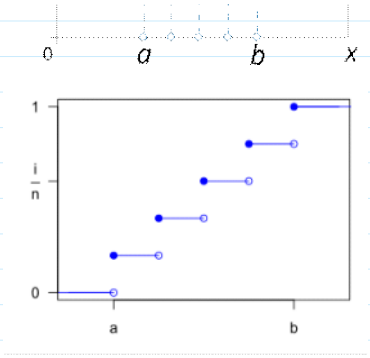
In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution wherein each of some finite whole number *n* of outcome values are equally likely to be observed. Thus every one of the *n* outcome values has equal probability 1/*n*. Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen".



Ex- Rolling a dice $= \{1, 2, 3, 4, 5, 6\}$

$\frac{1}{n} \Rightarrow$ $\boxed{n = b-a+1}$

Notation :-



Notation :-

Parameters a, b with $b \geq a$

PMF $\frac{1}{n}$

Mean $= \frac{a+b}{2}$

Median $= \frac{a+b}{2}$

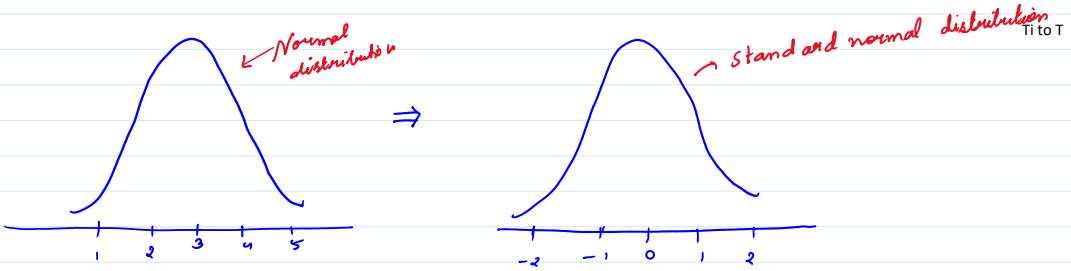
Standard Normal distribution and Z-Score [z- stats]

$X = \{1, 2, 3, 4, 5\}$ — Normally distributed assume

Random variable

μ = Mean = 3

$\sigma = 1.411 \approx 1$ [assume]



Ti to T

How to convert normal distribution to standard normal distribution uske liye use hota hai z -score

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

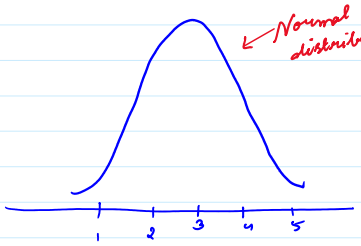
$= \frac{1-3}{1} = -2$

$= \frac{2-3}{1} = -1$

$3 \rightarrow -2$ (graph 1 → graph 2)

$2 \rightarrow -1$

Ques : What percentage of data is falling above 4.5 with $\mu = 4$ and $\sigma = 1$



$\Rightarrow$ Z-score $= \frac{4.5-4}{1} \Rightarrow 0.5$

Area under the curve (≥ 4.5)

$= 1 - 0.69146$ (from table)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702

from table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

Real World Problem : In India the average IQ is 100, with a standard deviation of 15 . What is the percentage of the populati on would you expect to have an IQ lower than 85%

$\mu = 100$ $\sigma = 15$

$x_i =$

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

$$= \frac{85-100}{15}$$

$$= -1$$

Area = 0.15 866

= 15.866

Central Limit Theorem [Interview Ques]

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is larger enough. Regardless of whether the population has a normal, Poisson, binomial , or any other ditribution, the sampling distribution of the mean will be normal .

[Interview Question]

$X \sim N(\mu, \sigma)$

σ = population std

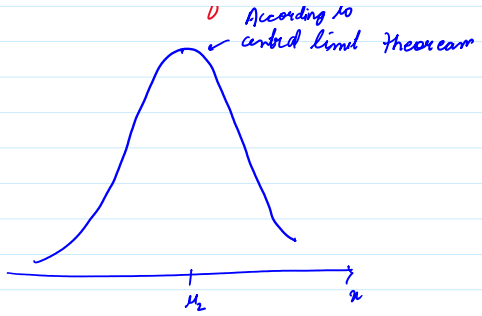
μ = population mean

n = sample size

Sample distribution of mean

According to central limit theorem

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



$$\bar{x} \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

n can be any value

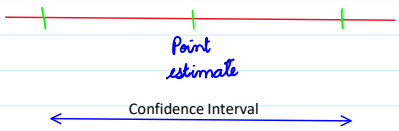
Estimate It is an observed numerical value used to estimate an unknown population parameter

1. Point Estimate Single numerical value used to estimate the unknown population parameter.

Ex- Sample mean is a point estimate of population mean.

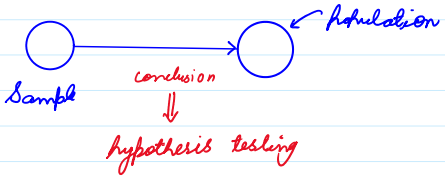
$\mu \leftarrow$ point Estimate $\bar{x}$

2. Interval Estimate Range of values used to estimate the unknown population parameters . Interval estimates of population parameters are called confidence intervals.



A hypothesis and Hypothesis testing Mechanism

Inferential stats :- conclusion or inferences



Hypothesis testing Mechanism

① Null hypothesis (H_0) → The assumption you are beginning with

② Alternate hypothesis :- Opposite of null hypothesis

③ Experiments → proof collect

④ Accept the null hypothesis or reject the null hypothesis

Ex- Colleges at District A states its passed average percentage of students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90% with a standard deviation of 4% . Does this scholl have a different passed percentage

What is null and alternate hypothesis

null hypothesis $\mu = 85$ (By default)

alternate hypothesis $\mu \neq 85$

P value

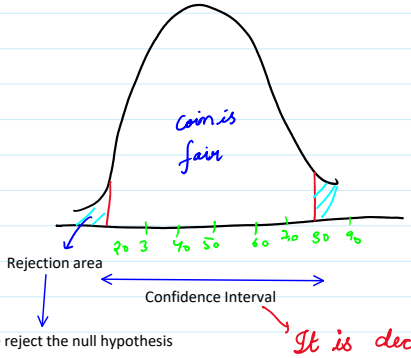
The p value is a number , calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

Ex- Tossing a fair coin

1) Null hypothesis → coin is fair

2) Alternate hypothesis → coin is not fair

Experiment:



It is decided by domain industry expert

If P < significance value → reject the null hypothesis

Hypothesis testing And Statistical Analysis

1) Z test
2) t test
} ⇒ Average
3) chi square ⇒ Categorical data
4) ANOVA ⇒ variance

Z test : Used when → (i) population std } given
(ii) $n \geq 30$

1. The average heights of all residents in a city is 168cm. A doctor believes the mean to be different. He measured the height of 36 individualsand found the average height to be 169.5cm.
   a. State null and alternate hypothesis
   b. At a 95% confidence level, is there enough evidence to reject the null hypothesis.

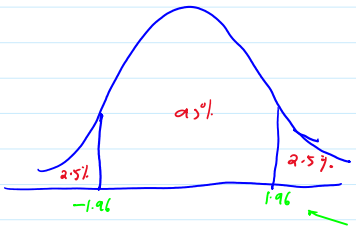
given $\mu = 168$ cm, $\sigma = 3$, $n = 36$ $\bar{x} = 169$ cm

Null hypothesis $(h_0) = \mu = 168$cm

alternate hypothes $(h_1) = \mu \neq 168$cm

$CI = 0.95$ $\alpha = 1 - 0.95 = 0.05$

Decision boundary



using z-table

$1 - 0.025 = 0.9750$ see the z-table and find 1.96

Statistical analysis

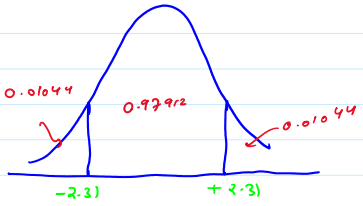
① By z test

$$Z_{test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}} = 2.31$$ it should be in range of 1.96

↓

So we reject null hypothesis

② By p-value



1 - Area under the curve [using table]

$1 - 0.9556$

p value $= 0.01044 + 0.01044 = 0.02088$

if P-value < significance value

$0.02088 < 0.05$ → so reject