Statistics

→ Statistics is the science of collecting, Organizing and Analyzing the data.

data = facts or piece of information.
Ex:- Height, IQ, Daily activities, Weight etc.

Types

Descriptive stats

→ It consists of organizing and summarizing the data.

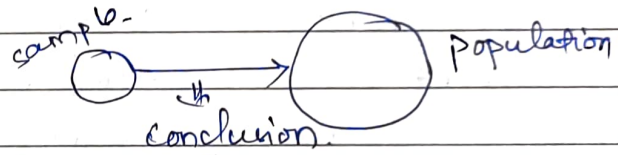
① Measure of central tendancy.
(Mean, Median, Mode.)

② Measure of disportion.
(Variance & standard devi.)

③ Diff. types of distribution of data.
→ Histogram, pdf & pmf, cdf function.

inferential stats

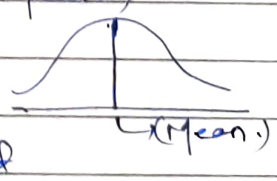
→ It consist of data which you measure to form Conclusion.



Made using Hypothesis testing

Ex:- Z test.
T-test.
Chi-squre test.
Anova test.
f. test.

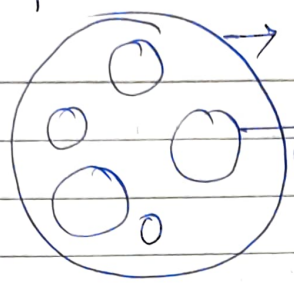
Ex:- Heights collected from the class = [175 cm, 180 cm, 140 cm, 150 cm]
Some descriptive Qns can be formed like.
→ what is the average height. (Mean.)



inferential Qns
→ Are the height in the class is what you expect from the college. Here. class = SAMPLE.
College = Population.

population & sample data

Ex:- Exit pole (Electection.)



→ Population. (N)

→ Sample of 500 for example. (n.)

⇒ Based on the result of the sample. We came to an conclusion about the population.

Types of data

Data

- Quantitative. (numerical value.)
  - Discrete.
    - → whole no. within some range.
    - eg:- No. of bank account.
    - → No. of children.
  - Continous.
    - Any value.
    - eg:- Weight. Height., age. temp., speed. salary.
- Qualitative. (catagorical data.)
  - Ordinal.
    - Ranking poss. like Good Better & best.
  - Nominal.
    - eg:- Gender Blood groups Hair clr pincode.
    - → No ranking is possible.
    - As the values are fixed.

→ We can make nominal data to a ~~ord~~ ordinal data.

Scales of Measurement

① Nominal scale data. :- Qualitative & catagorical data.

→ eg:- Gender, colour, labels etc.

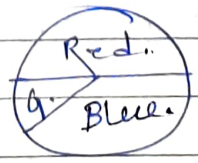
→ Order does not matter.

→ Ranking does not matter.

eg:- Ask 10 people about fav. colour.

5 = Red. (50%)
3 = Blue. (30%)
2 = Green. (20%)

} we can then visualize in Pie chart.



② - Ordinal scale data.

→ categorical data.
→ Ranking & order matters.
→ difference cannot be measured.

eg:- 1st Rank
2nd Rank
3rd Rank.

in this term diff. does not matter

1st Rank = 70 marks. } 30
2nd Rank = 40 marks } 20
3rd Rank = 20 marks

← → But when we convert in terms of values we can measure the dif.

eg:- in a race
1st = 3 min. } -2
2nd = 5 min. } -3
3rd = 8 min.

diff. can't be measured

→ difference cant measured.

⇒ Like this we can convert Ordinal to nominal data.

③ Interval scale data

→ Order matters.
→ Difference can be measured.
→ Ratio can not be measured.
→ No '0' starting Point.

eg:- Temperature.
30 f. → order
60 f. → diff. can be measured
90 f.
120 f.

when temp. is 30 f & then 60 f. then hottness of coldness is not exactly doubled. So ratio can't be measured.

As external factor is dep.

④ Ratio scale data

→ Order matters. (sorting.)
→ Difference can be measured.
→ Ratio can be measured.
→ Can have a '0' starting point.

eg:- Marks in exam.

Qn. → Marital - Status :- Nominal scale data.
→ fav. food based on gender :- "
→ IQ = Ratio scale data.

→ Measure of central tendancy

(1) Mean
(2) Median.
(3) Mode.

Mean

population	Sample
population Mean $(\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$ → population size.	Sample mean $(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$ → sample size

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

⇒ for this set mean = 3.2

Median

$(\mathcal{A} = \{4, 5, 2, 3, 2, 1\}$

⇒ first sort the random variable.

⇒ count the no. of elements.

if count = even, the add the centre 2 element & devide it by 2

if count = odd, then median = center element.

After sorting = $\{1, 2, \boxed{2, 3}, 4, 5\}$

$= 5/2 = 2.5$.

WHY MEDIAN?

→ Mean usually get affected by outliers, which median doesn't.

Ex:- $\{1, 2, 3, 4, 50\}$ → Mean $= \frac{1+2+3+4+50}{5} = \frac{60}{5} = 12$

Median = 3

→ without the outlier so mean would have been 2.

Conclusion:- Median is used to find the central tendancy when outlier is present.

Mode:- Maximum frequency occuring element.

$\{2, 1, 1, 1, 4, 5, 6, 7\}$

Mode = (1).

Mode is use mostly in EDA & feature engineering.

→ One of the feature of F.e is handling the missing value.

Eg:-

Age	Weight	Salary	Gender	Degree
24	70	40K	M	B.E
25	80	70K	F	—
27	95	45K	F	—
24	—	50K	M	PHD
32	—	60K	—	B.E
—	60	—	—	Master
—	65	55K	—	BE
40	72	—	M	BSc.

→ In the above table to fill the missing value we can use Mean, median & Mode.

→ for numerical value use Mean (if no. outlier.) & Median (if outlier.).

→ for categorial value use Mode.

Measure of Dispertion.

(1) Variance (σ^2)

(2) Standard deviation (σ)

Variance

i) Population variance

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

Sample variance

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

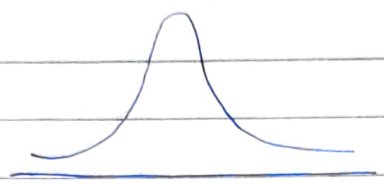
imp. Q:- why (n-1) in sample variance.

Ans:- It is for unbiased estimation of Population variance.

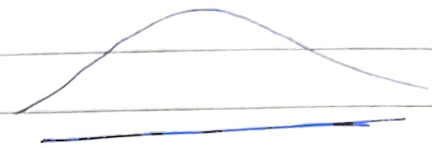
→ Sample variance is calculated using the data derived from a population data. So when we calculate sample variance value falls closer to the sample data which will be a biased one so to get an unbiased answer we use (n-1) instead of (n.)

Ex:-

One variance = 2.5 ⇓



second variance = 7.5 ⇓



→ When variance is less spread is less. & when variance is more spread is more.

Standard deviation population (σ) ⇒ $\sqrt{\text{variance}}$.

Sample stand. dev. (s) = $\sqrt{s^2}$

→ stand deviation used for getting the next & previous values.

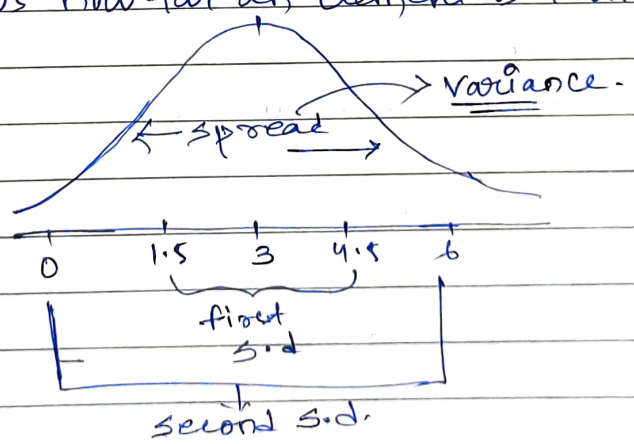
for ex.

if mean = 3 & s.d = 1.5

then ⟹ { 0 1.5 3 4.5 6 }

→ Here 6 is 2 standard deviation away from the mean.

So basically s.d shows how far an element is from mean.



⇒ Random variable :- It is a process of mapping the output of a random process or experiment to a number.

Eg:- Tossing a coin.

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

Histogram & skewness

Best visualisation diagram in case of frequency.

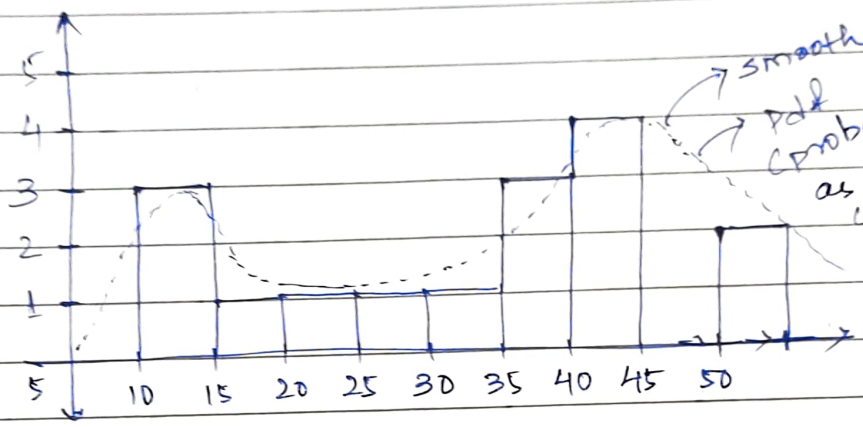
Ages:- {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Now we have to visualize is using the histogram.

Bin size = (50/10) = (5)

No. of Bin = 10 (which i need)

{where 50 = max range
10 = min range}

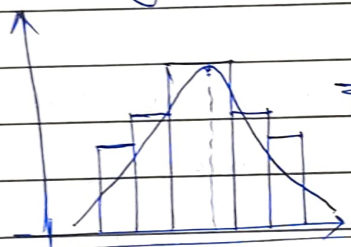


→ smoother curve

→ pdf (probability function) as it is a continuous value.

* Kernel density estimators makes the curve smooth.

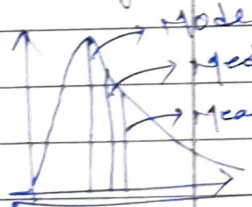
Skewness ⟹

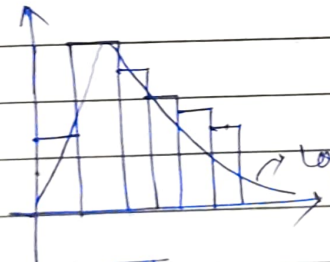


⇒ Normal distribution.
symmentrical distribution.
no skewness.

Here [(Median = mean = Mode.)]

Right skewed



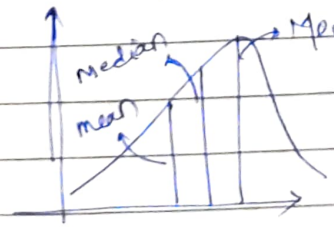


→ Positive skewed.

→ Log normal distribution.

[(Mean ≥ Median ≥ Mode.)]

left skewed
(negative skewed)



[(Mean ≤ median ≤ Mode)]

Percentile and Quartiles

→ Percentile is a value below which a certain percentage of data points lie.

→ $X = \{2, 3, 3, 4, 6, 6, 6, 7, 8, 8, 9, 9, 10, 11, 12\}$

to get the percentile of 10.

$$\text{percentile} = \boxed{\frac{\text{No. of values below 10}}{n} \times 100}$$

$$= \frac{12}{15} \times 100 = 80 \text{ percentile}$$

means 80% of the distribution falls below the value 10

If we go for the opposite.

then value $= \boxed{\frac{\text{percentile}}{100} \times (n+1)}$ → This will give the index of the value.

for example

25 percentile $= \frac{25}{100} \times 16 = 4^{th}$ element $= 4$

** If the output will be a decimal value like 4.5 then add the 4th and 5th value and devide it by 2.

Quartiles

Q_1 = 25 percentile.

Q_2 = Median (50 percentile.)

Q_3 = 75 percentile.

All this concepts used for removing the Outliers.

5 number summary

(1) Minimum value (Q_0)

(2) first Quartile (25 percen.) Q_1

(3) Median (Q_2)

(4) Third Quartile (75 percent.) Q_3

(5) Maximum.

⇒ Remove the Outliers

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$

We can do it by using 5 point summary.
We need to calculate (Lower fence & Higher fence).
Below the lower fence & above the higher fence all the values are outliers.

⇒ Lower fence $= Q_1 - 1.5(IQR)$
Higher fence $= Q_3 + 1.5(IQR)$

IQR = Inter Quartile Range $(Q_3 - Q_1)$

⇒ $Q_1 = (25/100) \times 20 = 5^{th}$ value $= \underline{3}$
$Q_3 = (75/100) \times 20 = 15^{th}$ value $= \underline{7}$

Lower fence $= 3 - 1.5(7-3)$
$= 3 - 1.5(4)$
$= \underline{-3}$

Higher fence $= 7 + 6 = \underline{13}$

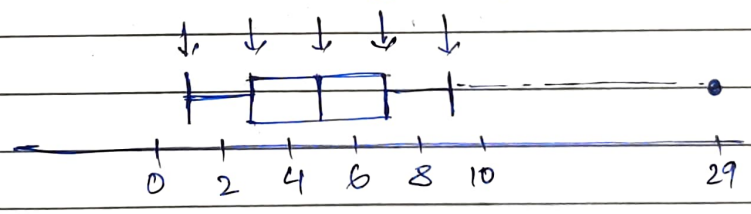
So 29 is an outlier.

Box plot (To visualize the Outlier.)

Requirments:-

(1) Minimum value = 1
(2) $Q_1 = 3$
(3) Median $= Q_2 = 5$
(4) $Q_3 = 7$
(5) Maximum = 9
(after excluding 29.)

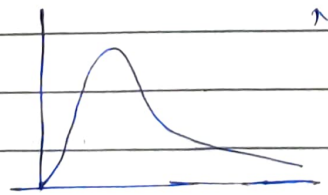
Box plot

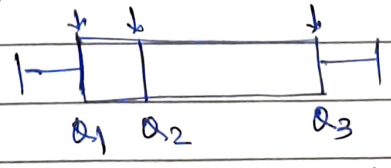


Interview Qn

what will be the boxplot for ~~left~~ Right skewed distribution.

Mean > Median > Mode.





$(Q_3 - Q_2) > (Q_2 - Q_1)$

Qn ⇒ $y = \{-13, -12, -6, -5, 3, 4, 5, 6, 7, 7, 8, 10, 10, 11, 24, 55\}$.

Q_1 = 25 percentile = $(25/100) \times 17 = 4.25$

so value is = $\frac{(-5+3)}{2} = -1$

Q_3 = 75 percentile = $(75/100) \times 17 = 12.75$.

value is = $(10+10)/2 = \underline{10}$ $\qquad IQR = (10+1) = \underline{11}$

Lower fence $= Q_1 - 1.5(IQR)$
$= -1 - 1.5(11)$
$= -17.5$.

Higher fence $= Q_3 + 1.5(IQR)$
$= 10 + 1.5(11)$
$= 26.5$.

So Outlier is $\underline{55}$.

Covariance & correlation.

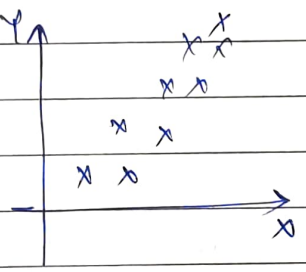
X	Y
2	3
4	5
6	7
8	9

Relation between X & Y

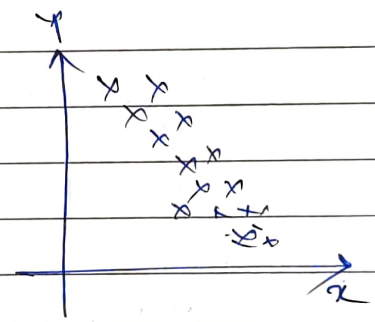
$X \uparrow \quad Y \uparrow$

$X \downarrow \quad Y \downarrow$

for $x \propto Y$



for $(x \propto 1/y)$



Covariance :-

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

Interview Qn Relation between Covariance & variance.

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

$$\Rightarrow \boxed{var(x) = cov(x, \bar{x})}$$

variance means the spread of the data around the mean. while covariance means the dir. relation between two random variable.

when x & y are directly proportional the covariance will be positive.
and when indirectly proportional it will be negative.

eg:-

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

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

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

$$= \frac{4+0+4}{2} = 4 \ (+ve)$$

Advantage

(1) Relation between (x & y)

disadvantage.

→ cov. does not have a specific limit value.

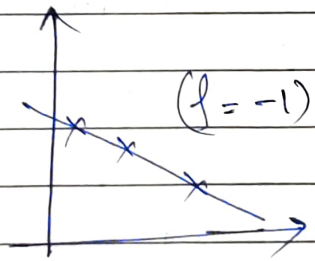
means it will only say the direction like posite or negative,
It can be $+\infty$ or $-\infty$.

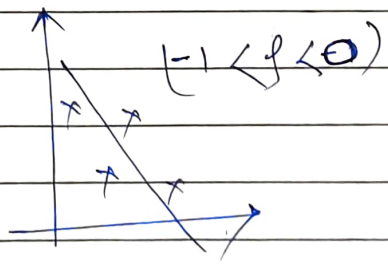
imp. what we need to know is how much two variables are correlated.
So for that we have to use **Pearson Correlation coeff.**

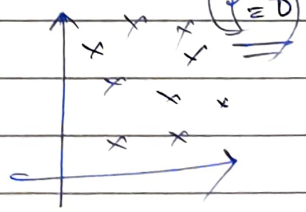
⇒ Its value lies between $(-1, 1)$. The more the value is towards $+1$ the more they are correlated. and vice versa.

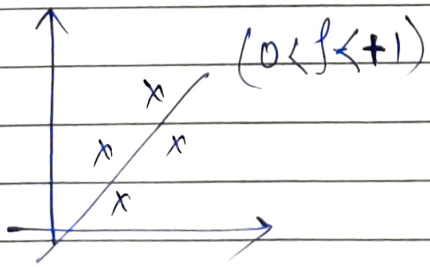
$$\boxed{\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}}$$

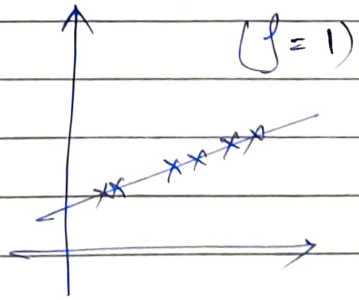
eg:- for x & y it is 0.5
x & z it is 0.8
the x & z are highly related.



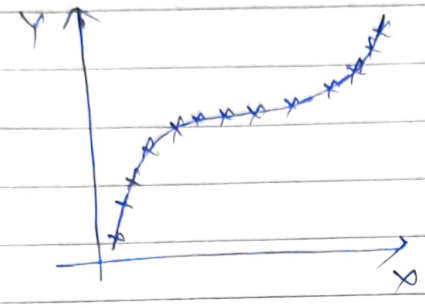








in case pearson correlation it can capture only the linear property. So to capture non linear we need to use Spearman's Rank correlation coefficient.



Here in case of this both x & Y are increing but pearson corr will be around (0.8) & spearman's " " " 1.

bcoz. here x and Y are not linearly increing so pearson corr. will not be 1.

$$\boxed{r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \times \sigma_{R(y)}}}$$

R(x) means Rank of (x)
R(y) " " " (y)

Eg:-

X	Y	R(x)	R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4

→ use this value in the formulae.

All this concepts will be used in feature selection.

eg:- Compare the below factors and find the relation between that and price.

size of house	No. of Room	Location	No. of people.	Hunted
(+ve)	(+ve)	(+ve)	(≈ 0)	(-1)

Relation with price

→ So as we can see No. of people is not related so we can drop it.

* Probability Distribution function.

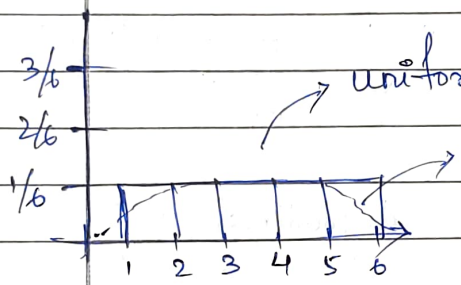
Three parts of this function is
(1) Probability density function.
(2) probability mass function.
(3) Cummulative distribution function.

PMF

→ Discrete Random variable

Ex:- Rolling dice, toss coin etc.

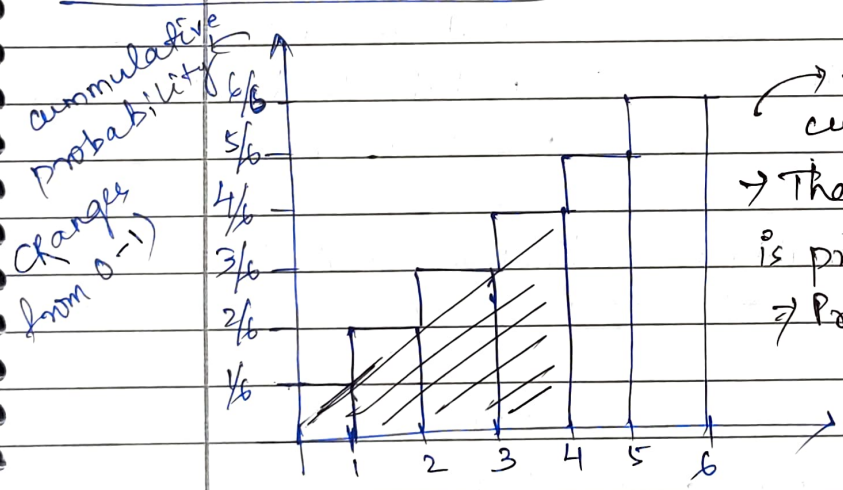
output = {1, 2, 3, 4, 5, 6} probability of getting each of these no. = (1/6)



→ uniform distribution.

→ curve can be smoothen by KDC.

How this is related to CDF (cummulative distrib. function.)



→ from this we can calculate cummulative probability.

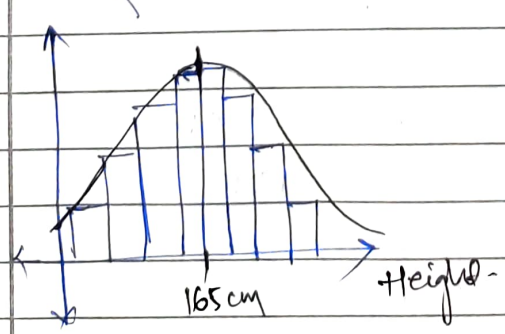
→ The area which is highlighted is probability of 4.

$\Rightarrow P_r(4) = P_r(1) + P_r(2) + P_r(3) + P_r(4)$

$= 4/6$

Probability density function.

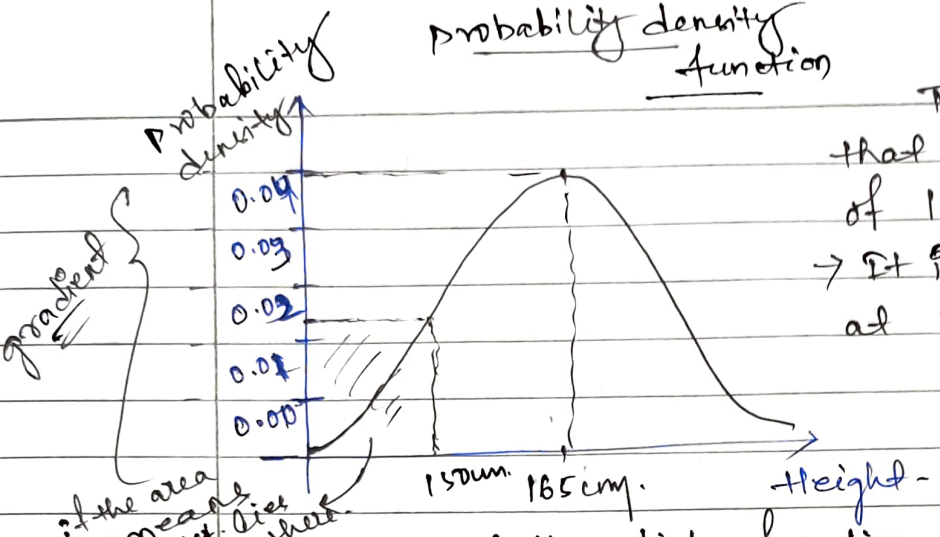
→ Continous Random Variable. (ex:- age, height, IQ, etc.).



Before it was histogram.
But when we converted it to a smoother curve we call it a probability density function.

In this case in X axis it will be probability density.

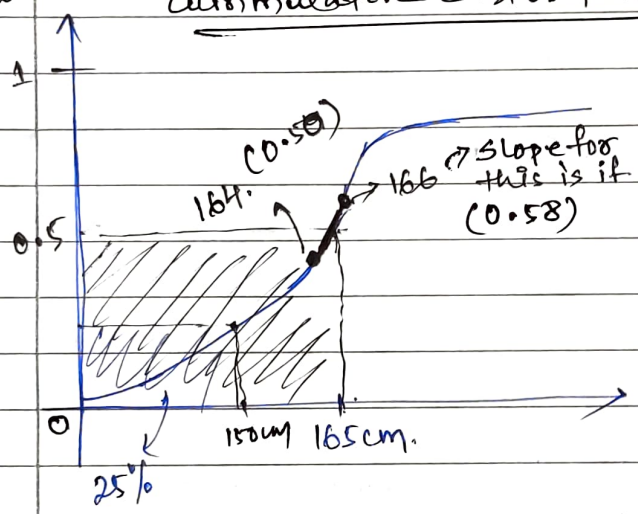
Probability density function



This graph doesnot mean that the probability density of 165 cm is 0.04

→ It is basically the gradient at 165 cm from CDF function.

Cummulative distr. function.



** for continous Random variable. the CDF will be a S curve

$$\text{gradient} = \left(\frac{0.58 - 0.50}{166 - 164}\right)$$

$$= 0.04$$

Thats why it is 0.04 as probability density in the PDF function.

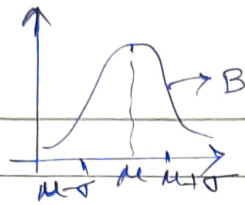
$$\boxed{\text{Probability density} = \text{Gradient of Cumulative Curve}}$$

Types of Distribution.

(1) Normal / Gaussian distri. → PDF
(2) Standard Normal distri → PDF
(3) Log normal distr. → PDF
(4) Power law distr. → PDF
(5) Bernouli distr. → PMF
(6) Binomial distr. → PMF
(7) Poission distr. → PMF.
(8) Uniform distr. → Discreat = PMF
→ conti. = PDF
(9) Exponential dist. → PDF
(10) CHI-square dist → PDF.
(11) f-distri. → PDF.

Normal distribution / Gaussian dist.

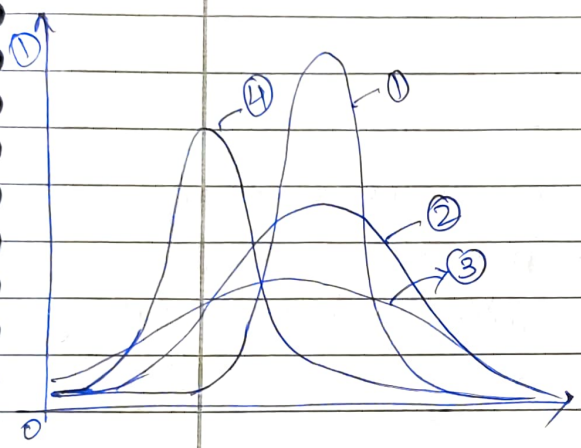
→ continous Random Variable.



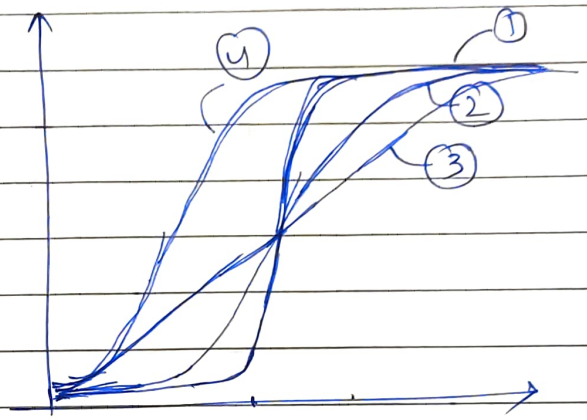
→ Bell curve. (Mean = Mode = Median).

eg:- Height, Weight, age, IRIS dataset.

PDF



CDF



$$X \approx N(\mu, \sigma^2)$$

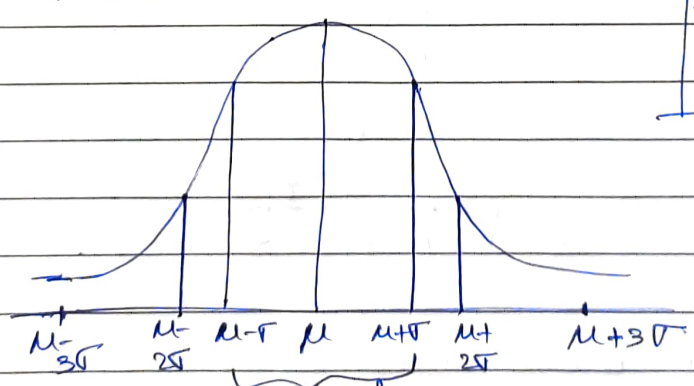
Mean, variance.

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

No need to byheart the formulaes, Just Google it and see in the wikipedia.

$$\boxed{\text{Emporical Rule} = 68-95-99.7\%}$$

Maximum data follows Normal distribution.



68% of data

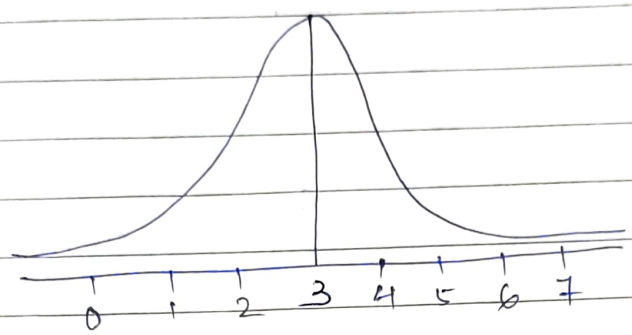
95% of data.

99.7% of data.

Standard Normal distribution

$X = \{1, 2, 3, 4, 5\}$ Here $\mu = 3$ (mean).
$\sigma = 1$

so the data will look like.



$\Downarrow$ Apply some transformation.
by making $\mu = 0$ & $\sigma = 1$
the new distn. which we will get
is standard normal distn.

So we can convert μ & σ we use $\boxed{\text{Z-score}}$

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

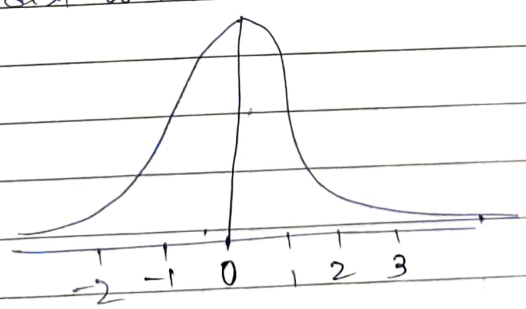
So for every element we need to calculate the Z-score.

$\Rightarrow$ for $1 = \frac{1-3}{1} = -2$ for $4 = \frac{4-3}{1} = 1$

for $2 = \frac{2-3}{1} = -1$ for $5 = \frac{5-3}{1} = 2$

for $3 = 0$ for $6 = \frac{6-3}{1} = 3$

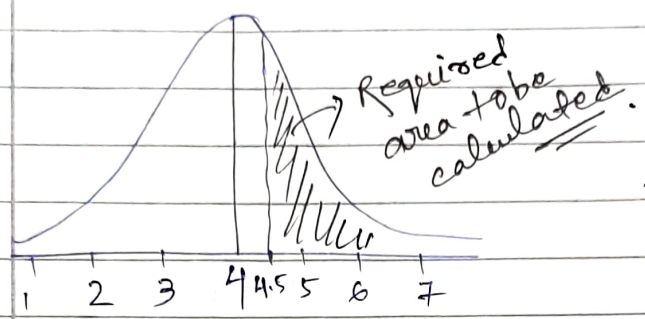
The new dist. will be



Z-score tells you about a value how many standard deviations away from the mean.

Ex:- consider a set where $\mu = 4$ & $\sigma = 1$.
So what percentage of scores lies above 4.5?

Ans:- So from μ & σ we can draw the distribution.

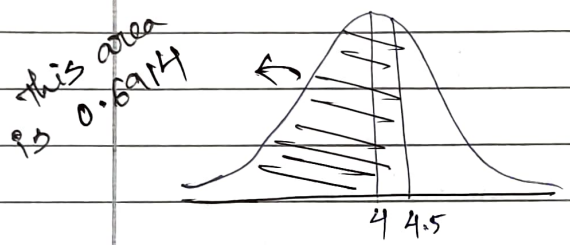


$$Z\text{-score} = \frac{4.5-4}{1} = 0.5$$

$\Rightarrow$ 0.5 standard deviation away from the mean.

Now go and search for Z table in Google to get the area left to the value.

& the value is 0.6914.



So the area left to 4.5 is 0.6914

" " " right to 4.5 "

$= 1 - 0.6914$

$= 0.3086$

$= 30.86\%$

Log-normal distribution

→ continous Random variable.

→ This distribution is always Right skewed.

So basically in machine learning our model should be trained with the data which has a normal distribution for better efficiency.

So to convert this Right skewed distribution to normal one then the new random variable (Y)

$$\boxed{(Y) = \ln(X)}$$

Notation = Lognormal (μ, σ^2)

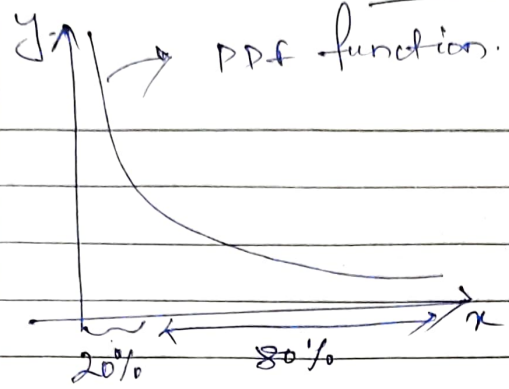
parameters = $\mu \in (-\infty, +\infty)$

$\sigma > 0$

PDF & CDF function (go to Google)

Check from wikipedia.

Power Law Distribution (80-20% Rule)



⇒ This is a functional relationship between two Quantities, where a relative change in one quantity results in a proportional relative change in other quantity.

The quantity varies as a power of another. considering the area of a square in terms of the length of its side, if the length is doubled then area is multiplied by a factor of 4.

Ex:- ① In RCB 20% of the team is responsible for winning 80% of the matches.

② 80% of the sales in amazon is derived from 20% of the product.

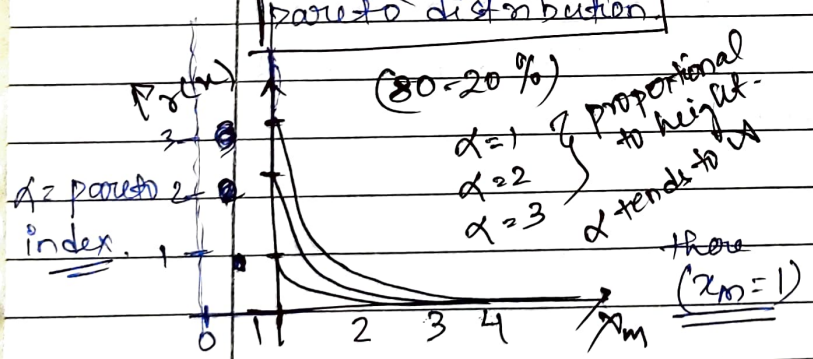
③ 80% of the wealth is distributed among 20% of the people.

④ 80% of project is done by 20% of the people.

⇒ Power law distribution (continous random var.).

Pareto distribution

Exponential distribution



α = where the line touches the y-axis.

parameter: $x_m > 0$, $\gamma > 0$ (real value)

$$\boxed{PDF = \frac{\alpha \cdot x_m^{\alpha}}{x^{\alpha+1}}}$$

$$\boxed{CDF = 1 - \left(\frac{x_m}{x}\right)^{\alpha}}$$

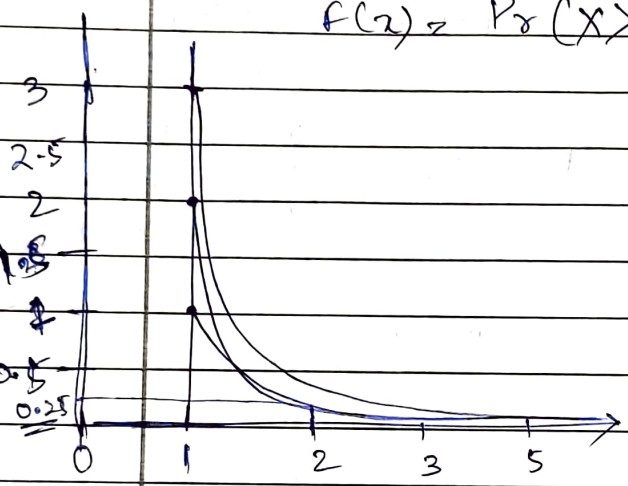
Q. How to convert a pareto distribution to a normal dist.

→ By using $\boxed{\text{Box-cox}}$ Transformation.

$$\boxed{\begin{aligned} \text{formulae :-} \quad Y &= \frac{x^{\alpha}-1}{\alpha} \quad \text{where } \alpha \neq 0 \\ Y &= \ln(x) \quad \text{where } \alpha = 0 \end{aligned}}$$

If X is a random variable with a Pareto (Type-1) distribution then the probability that X is greater than some number x i.e the survival function (Tail function) is given by.

$$\bar{F}(x) = Pr(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 1 & x < x_m. \end{cases}$$



⇒ if $x_m = 2$

then $= \left(\frac{1}{2}\right)^2 = \frac{1}{4} = 0.25$

Pareto distribution is often used to describe the distribution of wealth, income.

Bernouli's distribution (PMF.)

→ discrete Random Variable. (Tossing of coin etc.)

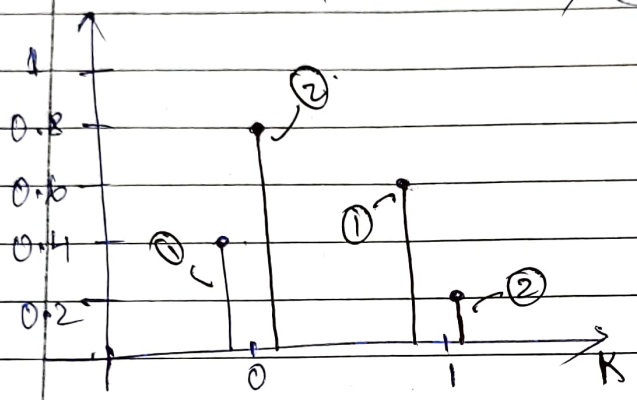
→ Outcome of process is binary = {1, 0}
{success / failure}

Tossing of coin

H = 0.5 ⇒ p

T = 0.5 ⇒ (1−p) = q.

$\boxed{p+q=1}$



parameter = $0 \leq p \leq 1$

$q = 1-p$.

Support = $K \in \{0, 1\}$

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

CDF = $\begin{cases} 0 & \text{if } K<0 \\ 1-p & \text{if } 0 \leq K < 1 \\ 1 & \text{if } K \geq 1 \end{cases}$

function is

$$\boxed{p^K (1-p)^{1-K}} \Rightarrow K\{1, 0\}$$

Binomial distribution.

Binomial distribution with parameters 'n' & 'p' is the discrete probability distribution of the number of successes in a sequence of 'n' independent experiments each asking a (Yes – No) qn. and each with its own Boolean-valued outcome.

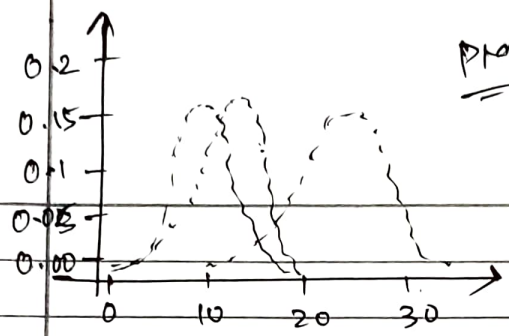
$$\boxed{\begin{array}{l} \text{Success} = p \\ \text{failure} = q = (1-p) \end{array}}$$

✓ for a single trial where n = 1 it is Bernouli's dist.

So basically Binomial distribution is n times of a Bernouli's distr.

$$\boxed{\text{PMF} = {}^{n}C_{K}\, p^K (1-p)^{n-K}}$$

factors : n, p, K



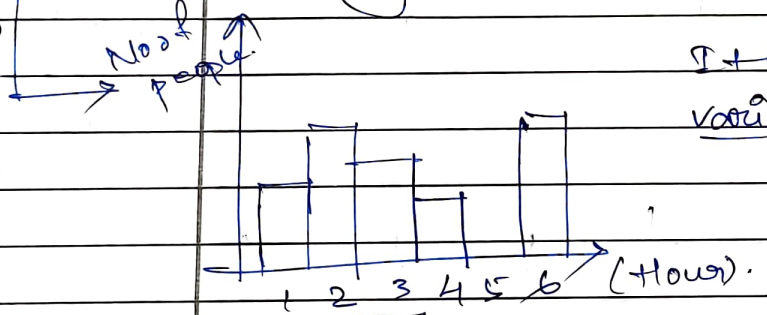
PMF Graph will be like this.

Q: Convert log normal distribution to pareto distribution.

Poission Distribution.

Ex:- No. of people visiting to a bank at every hour.

Basically this expresses the probability of a given number of events occuring in a fixed interval of time



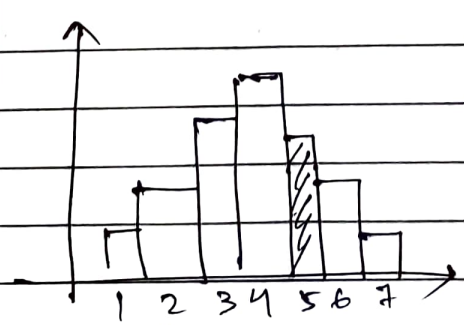
It is a discrete Random variable.

One component $\boxed{\lambda}$ is present in poision distribution.
Suppose $\lambda = 3$ in the above example.
This means The expected no. of people to come at the specific time interval is 3 in this case.

$$\text{PMF} = \underbrace{Pr(x=5)}_{\text{probability of a person visiting bank at } 5^{th} \text{ hour.}} = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$x!$ → x factorial.

Ex:-



consider $\lambda = 3$

$$Pr(x=5) = \frac{e^{-3} \cdot 3^{5}}{5!} = 0.101 = 10.1\%$$

what is the prob. of a person to come at 5th or 6th hour

$$= Pr(x=5) + Pr(x=6) = \frac{e^{-3} \cdot 3^{5}}{5!} + \frac{e^{-3} \cdot 3^{6}}{6!}$$

$= (\quad)$ Ans.

Uniform distribution

(1) continous uni. distribution.

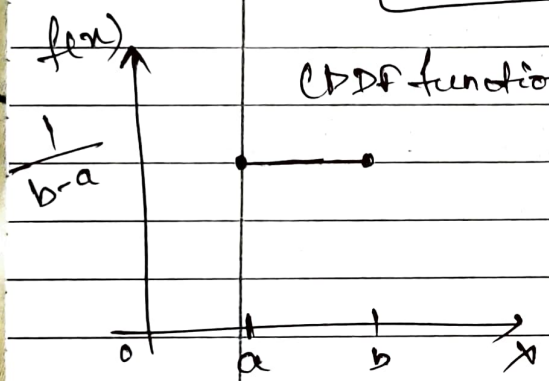
(2) Discrete uni. distribution.

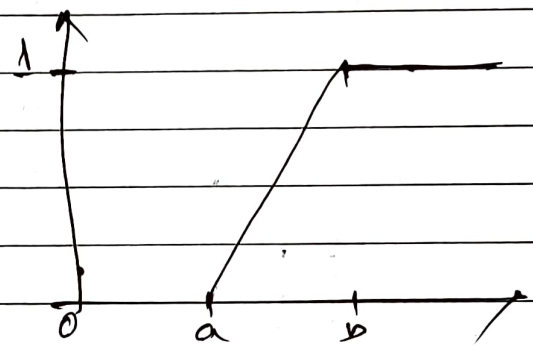
Cont. Uni. distribution.

Ex:- The number of candies sold daily at a shop is uniformly distributed.

→ In this case there is a min. & max. value.

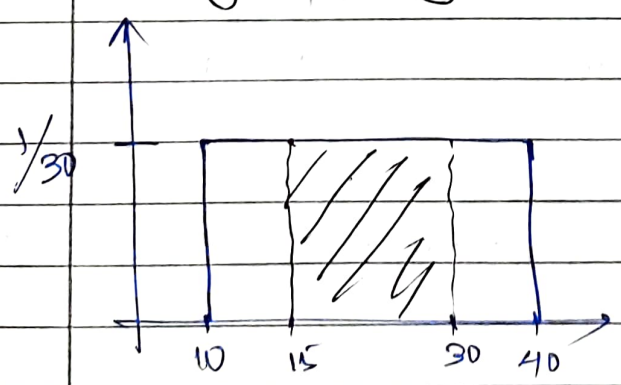
Notation = $U(a,b)$. $(b \gg a)$ (interval.)	= b = max.
parameter = $-\infty < a < b < \infty$	a = min.





$$PDF = \begin{cases} \frac{1}{b-a} & \text{for } x \in (a,b) \\ 0 & \text{otherwise.} \end{cases}$$

Q) The number of candies sold daily at a shop is uniformly distributed with a max. of 40 & min. of 10. what is the probability of daily sales to fall between 15 & 30.



$Pr(15 \leq x \leq 30)$

$= (30-15) \times \frac{1}{b-a}$

$= 15 \times \frac{1}{30} = 0.5$

$= 50\%$.