Machine Learning.

=> what is the Statistics And its application.

Statistics is a field that deals with collection organization, analysis, interpretation and presentation of the data

So it's application is to make the decision based on the given data.

Types of Statistics =>

There are two types of Statistics-

(i) Descriptive
(ii) Inferential

(i) Descriptive =>
  (i) measure of central tendency
    (i) mean.
    (ii) median.
    (iii) mode.
  (ii) measure of Dispersion
    (i) variance.
    (ii) Standard deviation.

classmate
Date
Page

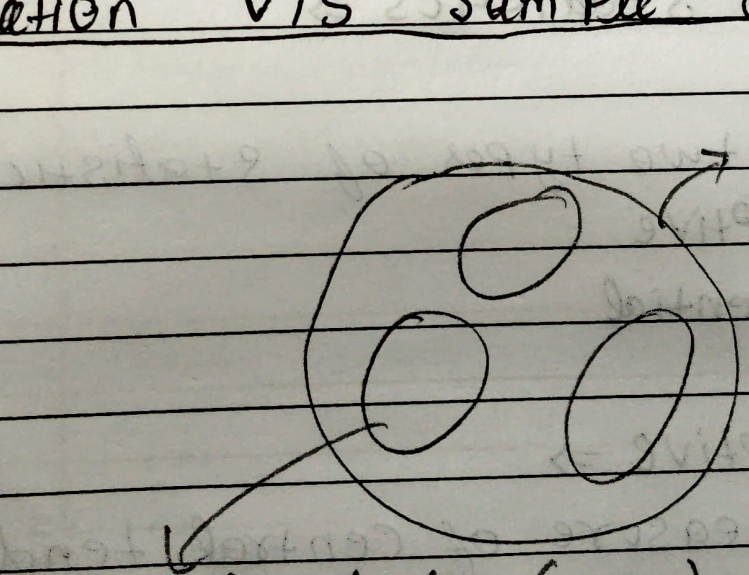
(11) Inferential ⇒

{Z-test, t-test, χ^2 test ...}

Sample data ← Collect Data → {Using some experiments}

Collect Data → Conclusion of Inferences → other data → {Population data}

⇒ Population VIS Sample data ⇒



① ⇒ measure of central tendency

(i) mean
(ii) Median
(iii) Mode

① **Mean⇒**

Population data

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

Sample data.

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

(i) age = {1, 3, 4, 5}.

$$\mu = \frac{1+3+4+5}{4} = \frac{13}{4} = 3.25$$

→ outlier (big number)

(ii) **median** ⇒ age = {4, 3, 1, 5, 100}.

↳ first sort the number.

{1, 3, 4, 5, 100}.

↳ median → 4 to fight with outliers impact

(iii) mode =>

age = {4, 3, 2, 1, 1, 4, 4, 5, 2, 100}.

In mode we select the element which have maximum frequency.

O/P => 4

(2) Measure of Dispersion:

(i) variance
(ii) Standard Deviation.

age 1 = {2, 2, 4, 4} → {Spread is less}

$\mu_1 = \frac{2+2+4+4}{4}$

=> 3

{the data point are near to the mean}

{Spread is more} ↑

age 2 = {1, 1, 5, 5}.

$\mu_2 = \frac{1+1+5+5}{4}$

=> 3

{the data points are far away from the mean}

① · variance

Population data $\{N\}$.

(points) (population mean)

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

(N: Population size)

age 1 = $\{2, 2, 4, 4\}$.

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

$\Rightarrow 3$

x_i	μ	$(x_i - \mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1
		4

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

$$\sigma^2 = \frac{4}{4} = \boxed{1}$$

{↳ less dispersion}

age 2 = $\{1, 1, 5, 5\}$.

$$\mu = \frac{1+1+5+5}{4}$$

$\Rightarrow 3$

x_i	μ	$(x_i - \mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4
		16

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{4}$$

$$\sigma^2 = \frac{16}{4} = \boxed{4}$$

{more dispersion}

(10) Samell data {n}.

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

(x_i → point, $\bar{x}$ → mean (Samell), n → Samell size)

{ why (n-1) ⇒ Imp inwrs. }

So if we use the $\left\{ s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \right\}$.

Imp { it is understimating the true Population variance.

and if we use the $\left\{ s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \right\}$ (→ known as bessel correction)

it is comes with unbaise understimation and it is focousing the different-different data points.

(11) Standard Deviation.

$$S.D = \sqrt{\sigma^2}$$

→ Population / Samrle standard deviation

↓ How far data Points is away from the mean.

=> **Variable:**

variable is a property that can take up any value

eg: age = 25 , Height = 7''.2'

Gender = Mall

Types of variable:

- (i) Quantitative
  - Discrete Quantitative: a whole no not any fraction value {eg:- {3, 5, 6, 18, 80}}
  - Continuous Quantitative: any value can be whole no and fraction as well. {eg: 2, 2.5, 5.5, 6.8}
- (ii) Qualitative / Categorical
  - eg: Gender → mall, female
  - colors: → red, blue, green, ...

Random variable => (X)

it is a function whose values are derived from the "process or experiments"

$$X = \begin{cases} 0 , H \\ 1 , T \end{cases}$$ Tossing a coin.

Random variable

↓ Discrete Random

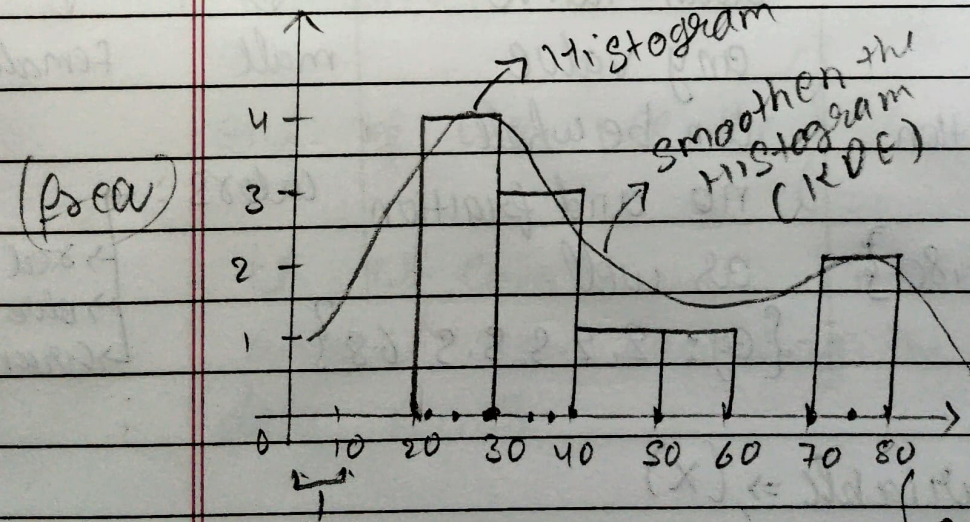
eg: Tossing a coin
rolling a dice

↓ Continuous Random variable.

eg: Tomorrow how many inches it's going to rain.

Histograms → It's count the frequency of element

$X = \{23, 24, 28, 30, 34, 36, 40, 50, 60, 75, 80\}$



20 - 30 = 4

30 - 40 = 3

41 - 50 = 1

51 - 60 = 1

61 - 70 = 0

71 - 80 = 2

KDE => Kernel density Estimation

and it gives you the probability density function

Percentiles and Quartiles.

Percentage:

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

odd number = 3

Percentage of odd number $= \frac{3 \times 100}{6} = 50\%$

Percentiles:

A percentile is a value below which certain percentage of observations lie.

$\{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10\}$. , $x = 9$

Percentaile of x = $\frac{\#\text{ value below } x}{n} \times 100$

below x values are = 11

$= \frac{11}{14} \times 100 = 78.57\%$ of value 9

↙

Percentaile ranking.

So basically it means that the 78.57% of the entair distribution is less than 9

if Percentile is give and we find out the 25 percentile of the given sample value is.

$$\text{value}_{\text{(Position)}} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$\text{value}_{\text{(Position)}} = \frac{25}{100} \times (14+1)$$

$$= \frac{25 \times 15}{100} = 3.75$$

but we don't have 3.75, so in this case take the avg of lower and higher value of the range like in the data points 3 & 4 are present.

$$\text{so } \frac{3+4}{2} = 3.5$$

$$\Rightarrow 3.75 \approx 3.50$$

② Quartiles →

- 25 Percentile = 1st Quartile.
- 50 Percentile = 2nd Quartile.
- 75 Percentile = 3rd Quartile.

5 Number Summary.

outlier.

eg:- {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

(i) minimum
(ii) 1st Quartile (25%)
(iii) median
(iv) 3rd Quartile (75%)
(v) maximum.

IQR => Inter Quartile Range

$Q_3 - Q_1$ → 1st Quartile

3rd Quartile

Removing the outlier. => So in case of removing the outlier we have to calculate the lower fenie and higher fence

(i) {if any value is less than lower fence then it is removed as outliers} and same as

(ii) {if any value is greater that Higher fence then it is removed as outlier.}

#

Lower fence = $Q_1 - 1.5(IQR)$

Higher fence = $Q_3 + 1.5(IQR)$

$\Rightarrow Q_1 \Rightarrow \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (20) = 5^{th}$ position value $= 3$

$\Rightarrow Q_3 \Rightarrow \frac{75}{100} \times 20 = 15^{th}$ position value $= 7$

$IQR = 7 - 3 = 4$

$\Rightarrow$ Lower fence $\Rightarrow Q_1 - 1.5(IQR)$

$\Rightarrow 3 - 1.5 \times 4$

$\Rightarrow 3 - 6 = -3$

$\Rightarrow$ Higher fence $\Rightarrow Q_3 + 1.5(IQR)$

$\Rightarrow 7 + 1.5(4)$

$= 13$

So the lower fence is -3 and Higher fence is 13.

the values which is less than -3 and greater that 13 is considered as outliers.

In our example the 27 is present which is greater that 13 so it is considered as an outlier. and remove all the outliers.

Values = { 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

- minimum = 1
- 1st Quartile = 3
- median = 5
- 3rd Quartile = 7
- maximum value = 9

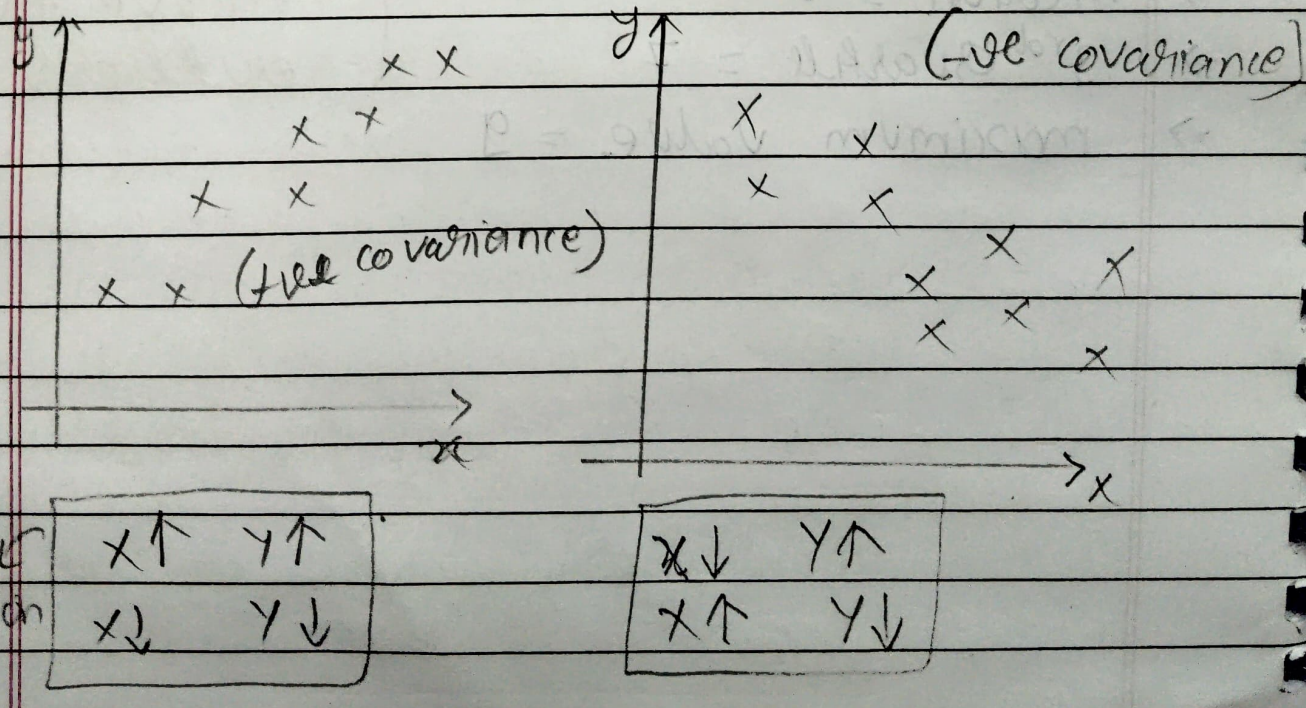
} we use the Box plot to remove the outliers.

Covariance and correlation:

covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable associated with changes in another variable

Covariance ⇒

covariance is a measure of how much two random variables changes together. If the variable tends to increase and decrease together, the covariance is (+ve). If one tends to increase when the other decreases, the covariance is (-ve).



relation

classmate
Date
Page

#

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

#

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

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

$$Cov(x,x) = var(x)$$

$Cov(x,x)$ → Covariance of x with respect to x

$var(x)$ → variance of x

x_i → Data points of random variable x
$\bar{x}$ → sample mean of x
y_i → Data points of random variable y
$\bar{y}$ → sample mean of y

Advantage:

① Quantify the relationship between x and y

Disadvantage:

① covariance does not have a specific limit value

$$Cov(X,Y) \Rightarrow -\infty \text{ to } \infty$$

Correlation

- Pearson correlation coefficient
- Spearman Rank Correlation

① Pearson correlation coefficient:

It's limit the value $[-1, 1]$

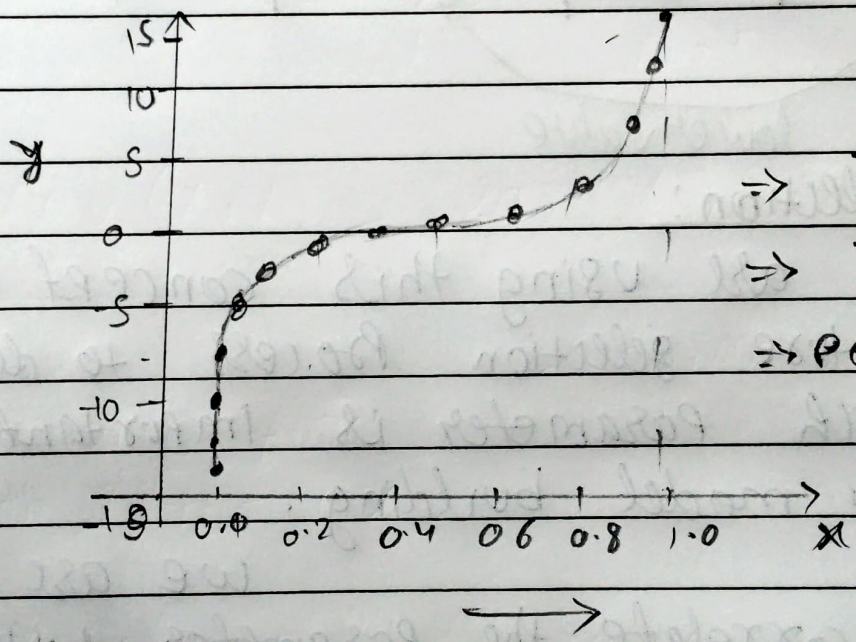
$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

→ Standard deviation.

(i) The more the value towards (+1) the more (+ve) correlated x & y is.

(ii) The more the value towards (-1) the more (-ve) correlated it is (x, y)

(2) Spearman Rank correlation:



⇒ x↑ y↑

⇒ x↓ y↓

⇒ pearson corr...

= 0.88

So the Pearson correlation is not correlating the non-linear data points.

$$r_s = \frac{cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$$

(rank)

Rank ⇒ So basically Rank is considerd the lower value is lower rank.

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

lowervalue (R(x) → R(y))

lowervalue (R(y) → R(x))

Feature Selection:

we are using this concept in our feature selection process to determine which parameter is important for the our model building.

we are trying to correlate the parameter with each other with respect to the output.

eg:

Size of House	No of rooms	location	No. of people living	Haunted	price
↑	↑	↑	(≈0)	↓	↑
(+ve)			remove/drop this feature	-ve	correlation