$$Statistics - 5$$

1) $Data\ types$

2) $Categorical\ and\ Numerical$

3) $Qualitative\ and\ Quantitative$

4) $Continuoues\ and\ discrete\ type$

5) $Levels\ of\ data$

6) $Nominal\ level$

7) $Ordinal\ level$

8) $intervel\ level$: $It\ does\ not\ have\ true\ zero\ point$

9) $Ratio\ level$: $It\ has\ zero\ point$

10) $Population\ \ and\ sample$

11) $Inferential\ statistics$: $Will\ work\ on\ sample\ and\ estimate\ on\ population$

12) $Descriptive\ statiscs$: $Analyse\ the\ data$ ($analyse\ the\ population$)

13) $Frequency\ table$

14) $Bar\ chart$

15) $Relative\ frequency\ table$

16) $pie\ chart$

17) $frequency\ distribution\ table$

18) $histogram$

19) $Distribution\ plot$

20) $How\ to\ find\ the\ interval$

21) $How\ to\ choose\ class\ width$

22) $Central\ tendency$

22) $Mean - mode - Median$

23) $Median\ vs\ Mean$

24) $Postive\ skew$: $Right\ side\ skew$: $Mean > Median > Mode$

25) $Negative\ skew$: $Left\ side\ skew$: $Mean < Median < Mode$

26) $No\ skew$: $Normal\ distribution$: $Mean = Median = Mode$

$In\ order\ to\ understand\ about\ data\ dispersions$

$\boldsymbol{we\ can\ approach\ in\ two\ ways}$

$\mathbf{1)}\ \boldsymbol{Center\ point\ analysis}$

$\mathbf{2)}\ \boldsymbol{Data\ flow\ analysis}$

$\mathbf{1)}\ \boldsymbol{Range}:$

- $in\ a\ data\ the\ marks\ values\ conisder\ as\ raw\ data$
- $in\ that\ raw\ data\ we\ have\ a\ lowest\ value\ and\ Heighest\ value$
- $Range = H - L$
- $The\ data\ lowest\ value\ has\ 1mark\ ,heighest\ value = 100\ marks$
- $Range = 100 - 1 = 99$

$\boldsymbol{draw\ back}:$

$\boldsymbol{Range\ will\ not\ tell\ about\ the\ middle\ points\ of\ the\ data}$

$\mathbf{2)}\ \boldsymbol{Mean\ deviation}:$

$\boldsymbol{mean\ deviation\ tells\ about\ how\ a\ data\ point\ is\ deviated\ from\ mean}$

$\boldsymbol{for\ example\ consider\ 5\ data\ points}: \mathbf{1, 2, 3, 4, 5}$

$\boldsymbol{The\ mean\ is} = \mathbf{3}$

$\boldsymbol{How\ the\ data\ point\ (observation)\ 1\ is\ deviated\ from\ mean\ value\ 3}: \mathbf{1 - 3 = -2}$

$\boldsymbol{x_1: 1, x_2: 2, x_3: 3, x_4: 4, x_5: 5}$

$\boldsymbol{Mean\ is\ denoted\ with}: \ \boldsymbol{\bar{x}}\ \ \boldsymbol{(sample\ mean)}$

$\boldsymbol{Mean\ is\ denoted\ with}: \ \boldsymbol{\bar{\mu}}\ \ \boldsymbol{(population\ mean)}$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{x_1 - \bar{x} = 1 - 3 = -2\ (-indecates\ back\ side\ of\ mean)}$$

$How\ the\ data\ point\ (observation) x_2: 2\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{x_2 - \bar{x} = 2 - 3 = -1}$$

$How\ the\ data\ point\ (observation) x_3: 3\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{x_3 - \overline{x} = 3 - 3 = 0}$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{x_4 - \overline{x} = 4 - 3 = 1}\ (+\boldsymbol{indicates\ ahead\ of\ mean})$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{x_5 - \overline{x} = 5 - 3 = 2}$$

$\boldsymbol{Total\ deviation}:$

$$= \boldsymbol{-2 - 1 + 0 + 1 + 2}$$

$$= \boldsymbol{(x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) + (x_4 - \overline{x}) + (x_5 - \overline{x})}$$

$$= \boldsymbol{\sum_{i=1}^{5} (x_i - \overline{x})}$$

$$\boldsymbol{Mean\ deviation}: \boldsymbol{\frac{1}{5} * \sum_{i=1}^{5} (x_i - \overline{x})}$$

$$= \boldsymbol{(x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) + \cdots + (x_n - \overline{x})}$$

$$= \boldsymbol{\sum_{i=1}^{n} (x_i - \overline{x})}$$

$$\boldsymbol{Mean\ deviation}: \boldsymbol{\frac{1}{N} * \sum_{i=1}^{N} (x_i - \overline{x})}$$

$\boldsymbol{Draw\ back}:$

$\boldsymbol{What\ is\ the\ Mean\ deviation\ from\ above\ example}:$

$$= \boldsymbol{-2 - 1 + 0 + 1 + 2 = 0}$$

$\boldsymbol{But\ actually\ we\ are\ seeing\ the\ deviation, but\ the\ maths\ says\ No\ deviation}$

$Chiru\ and\ vignesh\ are\ discussing\ why\ this\ is\ happend$

$they\ identified\ because\ of\ the\ neagtive\ values$

$we\ will\ convert\ negative\ to\ postive$

$3)$ ***Absolute Mean deviation :***

- $\boldsymbol{mean\ deviation\ has\ drawback\ of\ total\ devaition\ becomes\ zero\ due\ to\ postive\ and\ negative\ values}$
- $\boldsymbol{will\ convert\ negative\ values\ to\ postive}$
- $\boldsymbol{we\ choose\ a\ method\ mod = |-5| = 5 \quad also\ \ |5 = 5}$

$\boldsymbol{for\ example\ consider\ 5\ data\ points: 1, 2, 3, 4, 5}$

$\boldsymbol{The\ mean\ is = 3}$

$\boldsymbol{How\ the\ data\ point\ (observation)\ 1\ is\ deviated\ from\ mean\ value\ 3: 1 - 3 = -2}$

$\boldsymbol{x_1: 1, x_2: 2, x_3: 3, x_4: 4, x_5: 5}$

$\boldsymbol{Mean\ is\ denoted\ with: \ \bar{x} \ \ (sample\ mean)}$

$\boldsymbol{Mean\ is\ denoted\ with: \ \bar{\mu} \ \ (population\ mean)}$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{|x_1 - \bar{x}| = |1 - 3| = |-2| = 2}$$

$How\ the\ data\ point\ (observation) x_2: 2\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{|x_2 - \bar{x}| = |2 - 3| = |-1| = 1}$$

$How\ the\ data\ point\ (observation) x_3: 3\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{|x_3 - \bar{x}| = |3 - 3| = 0}$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{|x_4 - \bar{x}| = |4 - 3| = 1\ (+indicates\ ahead\ of\ mean)}$$

$How\ the\ data\ point\ (observation) x_1: 1\ is\ deviated\ from\ mean\ value\ 3:$

$$\boldsymbol{|x_5 - \bar{x}| = |5 - 3| = 2}$$

$\boldsymbol{Total\ deviation}:$

$$= \boldsymbol{|-2| + |-1| + |0| + |1| + |2|}$$

$$= \boldsymbol{|(x_1 - \bar{x})| + |(x_2 - \bar{x})| + |(x_3 - \bar{x})| + |(x_4 - \bar{x})| + |(x_5 - \bar{x})|}$$

$$= \boldsymbol{\sum_{i=1}^{5} |(x_i - \bar{x})|}$$

$$\textbf{Absolute Mean deviation}: \frac{1}{5} * \sum_{i=1}^{5} |(x_i - \overline{x})|$$

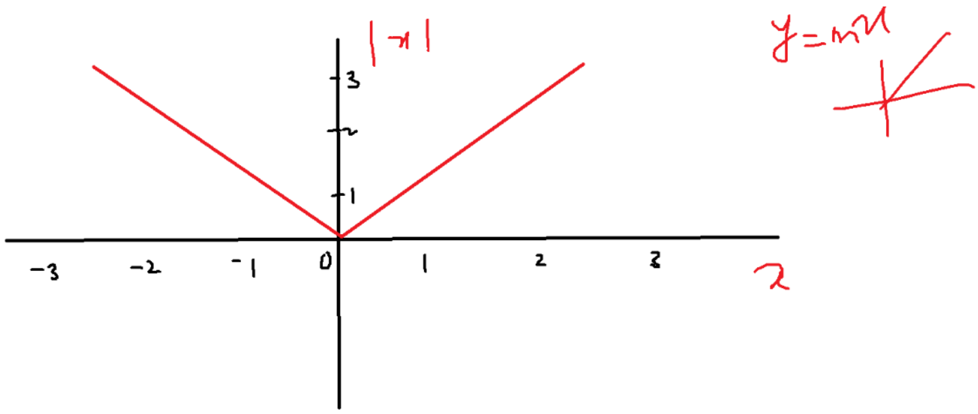
$$\textbf{Absolute Mean deviation}: \frac{1}{N} * \sum_{i=1}^{N} |(x_i - \overline{x})|$$

$\textbf{draw back}$:

$|x|$ $\textbf{graph}$

X	|x|
-1	1
-2	2
-3	3
0	0
1	1
2	2
3	3

$(-1, 1)$, $(-2, 2)$, $(-3, 3)$, $(0, 0)$, $(1, 1)$, $(2, 2)$, $(3, 3)$



$|x|$ $\textbf{graph is not continoues , it does not have sharp point at 0}$

$\textbf{Maths says differentatition fails for not continoues equations}$

$$|x| \ \textbf{differentation} = \frac{1}{|x|} \quad \textbf{or} \quad \frac{x}{|x|}$$

$$\frac{1}{0} \quad \textbf{or} \quad \frac{0}{0} = \textbf{zero division error or undefined}$$

4) ***Variance*** :

- *mean deviation has drawback of total devaition becomes zero due to postive and negative values*
- *Absolute mean deviation has drawback of no smooth curve at point '0'*
  so that the differentation fails
  ALL ML algorithms developed by Maths only
  If any maths equation not holds the properties, we will not consider
- *Our main goal is : To convert negative values to postive*
- *we choose square method* $= (-5)^2 = 25$ *also* $(5)^2 = 25$

for example consider 5 *data points* : 1,2,3,4,5

The mean is $= 3$

$x_1: 1, x_2: 2, x_3: 3, x_4: 4, x_5: 5$

Mean is denoted with : $\bar{x}$ *(sample mean)*

Mean is denoted with : $\bar{\mu}$ *(population mean)*

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

How the data point (observation)x_1: 1 *is deviated from mean value* 3:

$$(\boldsymbol{x_1} - \overline{\boldsymbol{x}})^2 = (\mathbf{1} - \mathbf{3})^2 = (-\mathbf{2})^2 = \mathbf{4}$$

How the data point (observation)x_2: 2 *is deviated from mean value* 3:

$$(\boldsymbol{x_2} - \overline{\boldsymbol{x}})^2 = (\mathbf{2} - \mathbf{3})^2 = (-\mathbf{1})^2 = \mathbf{1}$$

How the data point (observation)x_3: 3 *is deviated from mean value* 3:

$$(\boldsymbol{x_3} - \overline{\boldsymbol{x}})^2 = (\mathbf{3} - \mathbf{3})^2 = (\mathbf{0})^2 = \mathbf{0}$$

How the data point (observation)x_1: 1 *is deviated from mean value* 3:

$$(\boldsymbol{x_4} - \overline{\boldsymbol{x}})^2 = (\mathbf{4} - \mathbf{3})^2 = (-\mathbf{1})^2 = \mathbf{1}$$

How the data point (observation)x_1: 1 *is deviated from mean value* 3:

$$(\boldsymbol{x_5} - \overline{\boldsymbol{x}})^2 = (\mathbf{5} - \mathbf{3})^2 = (-\mathbf{2})^2 = \mathbf{4}$$

Total deviation:

$$= \mathbf{4} + \mathbf{1} + \mathbf{0} + \mathbf{1} + \mathbf{2}$$

$$= (\boldsymbol{x_1} - \overline{\boldsymbol{x}})^2 + (\boldsymbol{x_2} - \overline{\boldsymbol{x}})^2 + (\boldsymbol{x_3} - \overline{\boldsymbol{x}})^2 + (\boldsymbol{x_4} - \overline{\boldsymbol{x}})^2 + (\boldsymbol{x_5} - \overline{\boldsymbol{x}})^2$$

$$= \sum_{i=1}^{5} (x_1 - \overline{x})^2$$

$$\boldsymbol{Variance}: \frac{1}{5} * \sum_{i=1}^{5} (x_1 - \overline{x})^2$$

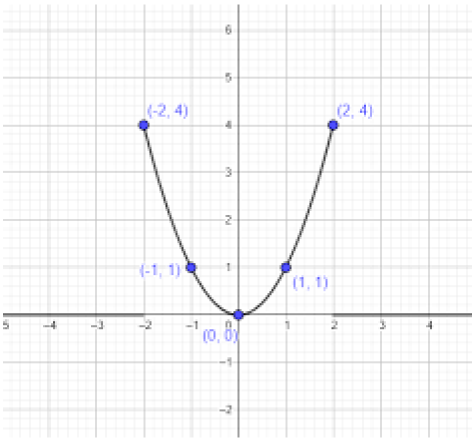
$$\boldsymbol{Variance}: \frac{1}{N} * \sum_{i=1}^{N} (x_1 - \overline{x})^2$$

draw back:

x^2 ***graph***

X	X^2
-3	9
-2	4
-1	1
0	0
1	1
2	4
3	9

$(-3, 9), (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4), (3, 9)$



$$y = x^2 \quad power = 2 \quad Non\ linear$$

$$y = x \ \ power = 1 \ \ Linear$$

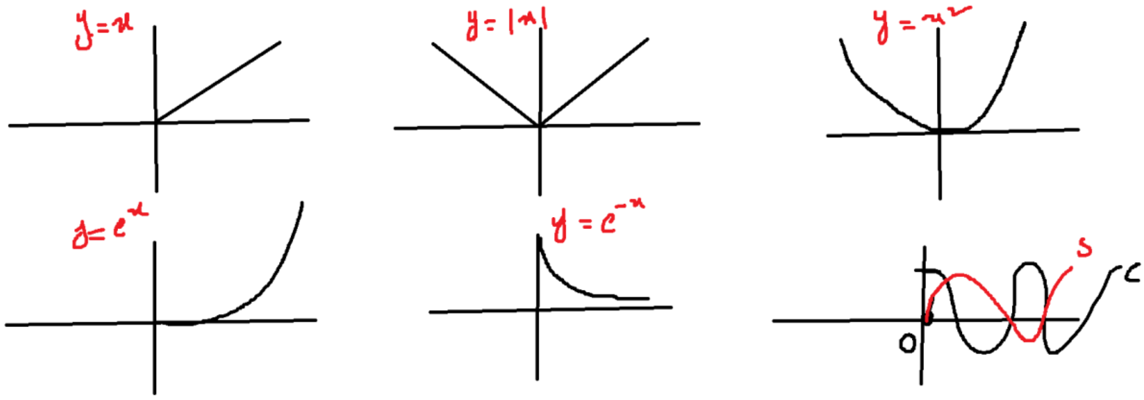
$$y = x^2$$

$$\frac{dy}{dx} = 2x$$

$$The\ curve\ has\ smooth\ every\ where\ x = 0\ , \frac{dy}{dx} = 2 * 0 = 0$$

$$So\ variance\ is\ a\ valid\ metric$$

Type equation here.



$\boldsymbol{Drawback}$:

$\boldsymbol{Variance}$: $\boldsymbol{\frac{1}{N}} * \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2$

$$Step - 1: Calculate\ mean\ \overline{\boldsymbol{x}}$$

$$\boldsymbol{step - 2: calculate\ individual\ deviation: x_i - \overline{x}}$$

$$\boldsymbol{Step - 3: Calculate\ square\ of\ the\ deviation = (x_i - \overline{x})^2}$$

$$\boldsymbol{Step - 4: Calcuate\ total\ deviation}$$

$$\boldsymbol{Step - 5: Calculate\ the\ variance\ by\ dividing\ N}$$

$\boldsymbol{x = km}$	$\boldsymbol{x_i - \overline{x}}$	$(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2$
$\boldsymbol{1km}$	-2	4km^2 : (1km-3km)^2= (-2km)^2
$2km$	-1	1km2
3	0	0km2
4	1	1km2
5	2	4km2
$\overline{\boldsymbol{x}} = \boldsymbol{3}$		

1) $When\ we\ do\ the\ variance\ the\ values\ are\ increasing$
2) $when\ we\ do\ the\ variance\ the\ units\ also\ mention\ as\ square\ units$
3) $This\ leads\ interpreation\ problem$
4) $In\ the\ above\ example\ the\ variance = \frac{10}{5}km^2 = 2km^2$
5) $In\ the\ realtime\ we\ does\ not\ have\ km^2\ units\ ,we\ can\ not\ explain\ in\ the\ proper\ way$
6) $Inorder\ to\ avoid\ the\ draw\ back\ we\ need\ to\ apply\ root\ on\ variance$

Standard Deviation

It is denoted with $\boldsymbol{\sigma}$

$$\boldsymbol{\sigma = \sqrt{variance}}$$

$$\boldsymbol{\sigma = sd = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (x_i - \overline{x})^2}}$$

$Interpreation$:

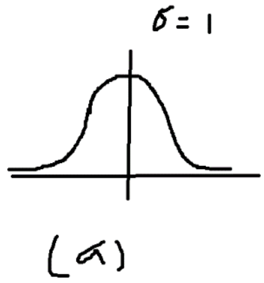
$standard\ deviation\ will\ explain\ on\ of\ average\ a\ data\ point\ how\ much\ deviated\ from\ mean$

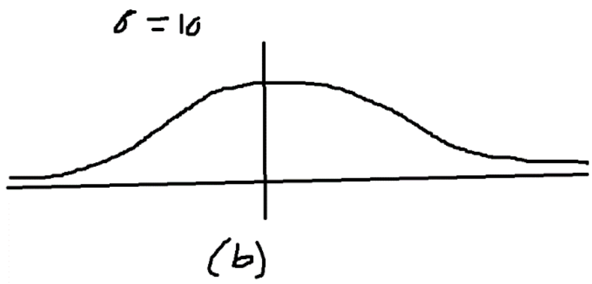
$$for\ example:\ ,4,5 \quad var = 2km^2$$

$$std = \sqrt{\boldsymbol{2km^2}} = 1km$$

$Each\ and\ every\ data\ point\ on\ of\ average\ 1km\ deviated\ from\ mean$

on of average How much a data point is deviated from mean





$Standard\ deviation\ is\ low: The\ data\ points\ are\ close\ to\ mean$
$Standard\ deviation\ is\ hight: The\ data\ points\ are\ far\ from\ mean$

$How\ much\ hyd\ is\ far\ from\ Nagpur:\ 400km$

$How\ much\ a\ data\ point\ is\ far\ from\ mean = standard\ deviation$

- $Range = H - L$
- $\boldsymbol{Mean\ deviation}$: $\frac{\mathbf{1}}{\boldsymbol{N}} * \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})$
- $\boldsymbol{Absolute\ Mean\ deviation}$: $\frac{\mathbf{1}}{\boldsymbol{N}} * \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\boldsymbol{x_i} - \overline{\boldsymbol{x}}|$
- $\boldsymbol{Variance}$: $\frac{\mathbf{1}}{\boldsymbol{N}} * \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2$
- $\boldsymbol{\sigma = sd =} \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}} * \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}$