Variables → like Ages = [22, 25, 27, 30]

Variable → like single Age = 22

→ Variables

A variable is a property that can take on any value.
Eg: Height, Weight

→ Types of variables :-
/Numerical.

1) Quantitative :- measured numerically & perform Airthmatic operations (Add, sub, multi)
Eg :- Weight, height, Age.

2) Qualitative/ categorical var. :- Based on some characteristics we can derive categorical variable.
Eg Gender [M, F], IQ, Blood group, Tshirt (S, L, XL, XXL)

Quantitative

→ Discrete Var. → (whole number)
1) no. of children
2) no. of Bank Acc.
3) no. of vechiles.
of countable units

→ Continuous Var. (any value it can have)
Eg: Height, 172.5,
weight, 55 kg
rainfall: 1.35 inch

⇒ Central Tendency / measure of central tendency

Refers to the measure, used to determine the center of the distribution of data.

Eg: mean, median, mode, Positional values (divides into equal parts like median, Quartile, decile, percentile)

mean ⇒ often called mathematical average.

median ⇒ positional avg.

bcoz median is the positional no. in the sorted data. positional value e.g. data is dividing into 2 parts by median.

with mean in feature engineering.

■ JUNE ■ 2018

15 Friday

⇒ Arithmetic Mean for Population (N) & Sample (n)

① Mean (Average) :- ∑ (Σ)

a) $M = \frac{\sum_{i=1}^{N} xi}{N}$ → in case of Population (N)

b) $\bar{X} = \frac{\sum_{i=1}^{n} xi}{n}$ → in case of sample (n)

a) {1,1,2,2,3,3,4,5,5,6}

$M = \frac{\sum_{i=1}^{n} xi}{N} \Rightarrow \frac{1+1+2+2+3+3+4+5+5+6}{10} \Rightarrow \frac{32}{10} = 3.2$

- M can be $\geq \bar{X}$
- M can never $< \bar{X}$

b) $\bar{X} = \frac{32}{10} = 3.2$

c) with outliers (outliers really have adverse impact on entire distribution)

"Outlier is the no. which is completely different from entire distribution"

{1,1,2,2,3,3,4,5,5,6, 100} → outlier

$M = \frac{32+100}{11} \Rightarrow 12$

$M_0 = 3.2$] huge diff with resp. to mean due to outliers.
$M = 12$]

(2) Outliers in case of Median :-

a) {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100} with 1 outlier

first sort the no.'s.

Median = 3

b) {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112} 2 outliers

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

- In case of mean, while adding outlier
  $\mu = 3.2$
  $\mu_1 = 12$ (with outlier)
- In case of median, while adding outlier
  $Med = 3$
  $Med_1 = 3.5$ (with outlier)

* {It implies, Median works well with outlier}

17 Sunday

(3) Mode with Outliers.

(We can use mode both in integer or categorical way but it works well with categorical variables.

{1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100}

In case of
{1, 2, 2, ... 7, 8, 100, 100, 100, 100}
we take 100 outlier as mode.

Tuesday

→ Measure of Dispersion (Dispersion) ⇓ (Spread)

1. Variance
2. Standard Deviation.
3. Relative dispersion/coefficient of variation.
4. IQR

(1) Variance :- concept of measure of dispersion.

a) Population variance (N)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \Rightarrow \frac{10.84}{6} = 1.81$$

Eg

x	x − μ	(x − μ)²
1	−1.83	3.34
2	− 0.83	0.688
2	0.83	0.688
3	0.17	0.03
4	1.17	1.37
5	2.17	4.71
μ = 2.83		Σ = 10.84

for grouped data (C-I)

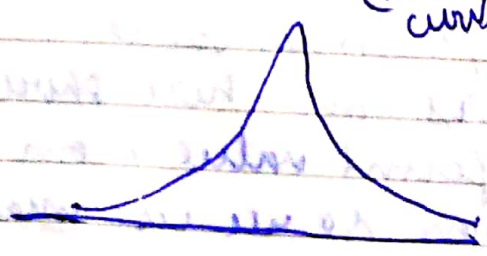
$$\sigma^2 = \sum_{i=1}^{N} f\frac{(x_i - \mu)}{N}$$

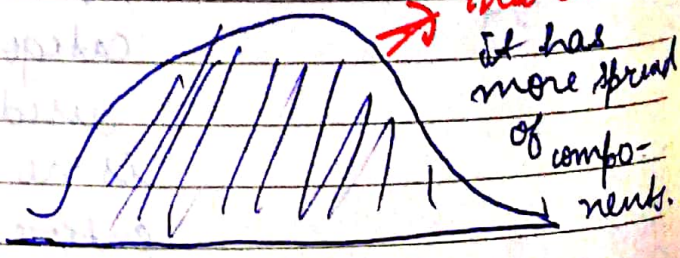
f = frequency.

* "The more variance means the data is more dispersed."

Ex which graph has more variance?

(Bell curve)





Wednesday

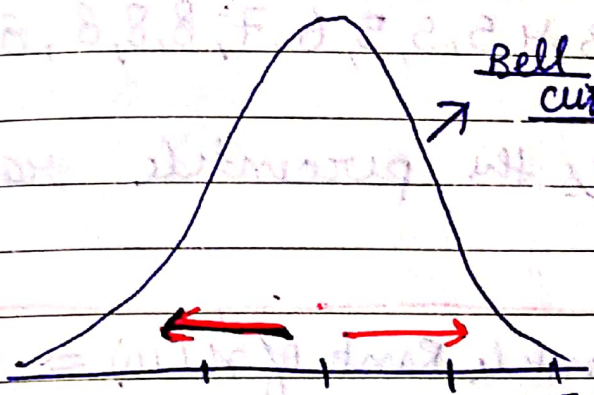
② Standard deviation (σ)

$$\sigma = \sqrt{\text{Variance}} \quad \text{OR} \quad \sigma = \sqrt{\sigma^2}$$

$\sigma = \sqrt{1.81} = 1.345$ (s.d)

2.87
+ 1.34
4.17

2.83
− 1.34
1.49



specifying measure of cent. tendency, center is here for specific distribution.

* Basically variance is rectifying how much spread prsent & S.D tells the particular element is how far from the mean.

⑤ Sample Variance (n)

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

for grouped data:-

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

f = frequency

③

$$C.V = \frac{\sigma}{\mu} \times (100)$$

for Population

for Sample ⇒

Percentile & Quantile (How to find Outliers)

(helps to remove outliers)

A Percentile is a value below which certain percentage of observation lie.

- Ex- 99 percentile :- A student got better marks than 99 % of entire class

Reg

Data: 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

5th

Q1 What is the percentile ranking of 10?

Sol

$n = 20$

Percentile Rank of x (10) = $\frac{\text{no. of values below } x}{n}$

$= \frac{16}{20} \times 100$

$\Rightarrow$ 80 percentile

80%

* means 80 percentage of entire distribution is less than 10

Q2 What value exist at percentile ranking of 25%? (same data: $n = 20$)

Sol

$$\text{Value} = \frac{\text{Percentile} \times (n+1)}{100}$$

$$\text{Value} = \frac{25}{100} \times 21 \Rightarrow 5.24$$

5.24 (index)

value at 5th index

Friday

Five Number Summary is used to determine the Outlier

1. Minimum
2. First Quartile (Q_1)
3. Median
4. Third Quartile (Q_3)
5. Maximum

* Removing the Outlier (i.e.) using Five no. Summary

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

→ When we need to remove outlier, we need to define a lower fence & higher fence.
Value present should b/w lowerfence & higherfence.

- **Lower fence** - After a smaller no. all the no. below the smaller no. will be lower fence
- **Higher fence** - After a higher no. all the no. above the higher no. will be higher fence.

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

IQR = Inter Quartile Range.

$Q_1 = 25\%$.
$Q_3 = 75\%$.

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

What is 25% percentile (Q_1) & 75% (Q_3)

Q_1

- Value (Q_1) $= \frac{25}{100} \times (19+1) \Rightarrow \frac{20}{4}$ th

$\Rightarrow$ **3 = (Q_1)**

- Value (Q_3) $= \frac{3.75}{100} \times 20$ th

$\Rightarrow 15^{th}$

7 = (Q_3)

- IQR $\Rightarrow Q_3 - Q_1$
  $\Rightarrow 7 - 3$
  $\Rightarrow 4$

- Lower fence $= Q_1 - 1.5(IQR)$
  $\Rightarrow 3 - 1.5(4)$
  $\Rightarrow 3 - 6 \Rightarrow$ **-3 = lower fence**

- higher fence $\Rightarrow Q_3 + 1.5(IQR)$
  $\Rightarrow 7 + 1.5(4)$
  $\Rightarrow$ **13 = high fence**

24 Sunday

So, our lower to higher fence range is:-

[lower $\longleftrightarrow$ higher fence] range

$[-3 \longleftrightarrow 13]$

* here, every no. > 13 is outlier & no < -3 is also consider as outliers.

SCHEDULE

Here 27 should be removed as it is greater than 13, belongs to higher fence.

remaining data

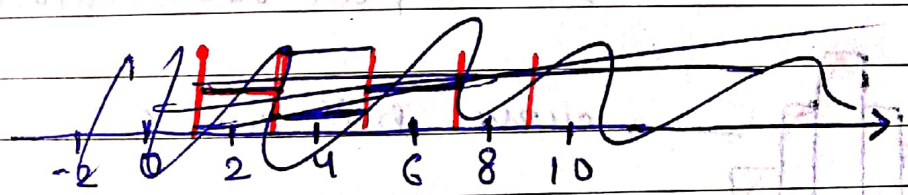
{ 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~ }

minimum value = 1
Q1 = 3
Q3 = 7
maximum = 9
median = 5 $\left(\frac{5+5}{2}\right)$

→ 5 no summary)

→ making box plot from the resultant of 5 no. sum

* "removing an outlier is done by lower fence & higher fence and IQR"



Boxplot basically use to see 'outlier'