Lecture - 02

→ Measure of Central tendency.

→ Measure of dispersion.

→ Gaussian (or) Normal distribution.

→ Z Score

→ Standard Normal distribution.

Arithmetic Mean for population and Sample.

Mean (Average)

Population (N)

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

Sample (n)

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

Random Variable (x) = {1, 1, 2, 2, 3, 3, 4, 5, 5, 6}

$$\mu = \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

Central tendency:

It refers to the Measure Used to determine the Centre of distribution of data.

Mean: [Average]

$X = \{10, 20, 30, 40, 50\}$ [without Outlier]

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30$$

Suppose,

$X = \{10, 20, 30, 40, 50, 1000\}$ [with outlier 1000]

$$\mu = \sum_{x_i=1}^{N} \frac{x_i}{N} = \frac{10+20+30+40+50+1000}{6} = \frac{1450}{6}$$

$$= 191.11$$

Huge difference
So use median

Median: [middle]

Suppose, $X = \{10, 20, \underline{30}, 40, 50\}$ [odd elements]

Median = 30

$X = \{10, 20, \underline{30, 40}, 50, 60\}$ [even elements]

$$= \frac{30+40}{2} = \frac{70}{2} = 35$$

Median = 35

Random Variable with some Outliers,

Suppose,

$X = \{10, 20, 30, 40, 50, 60, 500, 600\}$

$$\text{with outliers} = \frac{40+50}{2} = \frac{90}{2} = 45$$

↓

There is no huge change.

[median is acceptable]

Mode: [Reappearing Element]

Let Say,

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

Mode = 3 . [It is well suited for Categorical Value]

Examples ----------→

Ages	Name
15	Najim
16	basith
18	najim
17	najim
20	najim
⋮ → Missing values	———
30	——— → missing values
50	

definitely Use Mean

definitely Use Mode

Measure of dispersion:

Variance

Standard deviation

Measure of dispersion refers to the how the data points are different from each other data point (or) Mean.

data = 2, 2, 2, 2, 2	data = 1, 2, 3, 2, 2
$\mu = \frac{10}{5} = 2$	$\mu = \frac{10}{5} = 2$

Both mean are 2. How to measure the different between the data distribution ? Is achieved by Measure of dispersion. [How the data points are far]

Variance:

→ It is the statistical measures that Quantifies the spread or dispersion of set of data points.

avg of
→ It provides a measure of how much the individual datapoints in the dataset deviate from the mean.

Variance are non-negative one.

Population Variance

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

where,

μ - Mean

x_i - datapoint

N - No of population.

Sample Variance

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

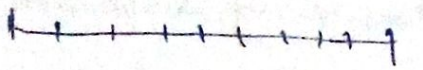
where,

$\bar{x}$ - Mean

x_i - datapoint

n - No of Samples.

Why (n-1) is divide in Sample Variance?



=> Population

=> Sample

Calculate Population Variance, Something comes 10
with same formula
But Sample Variance must comes very smaller than Population Variance like 1. It more affect and huge difference one.

Using this formula for Sample,

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$ [definitely Small Value of variance like 1]

So, Researcher research about this sample Variance, Obiviously $(n-1)$ is concluded.

Before that, Researcher research $(n-1)$, $(n-2)$, $(n-3)$, $(n-4)$ with different experiment. Finally, $(n-1)$ is suited for this Sample Variance.

$\frac{10}{2} \Leftrightarrow \frac{10}{2-1}$ $(n-1)$ is Unbiased estimation (or) degree of freedom (or) Bessels Correction.

$5 \Leftrightarrow (10) \rightarrow$ definetely these one is high.

Similarly,

$$\frac{(x_i - \bar{x})^2}{n} \Leftrightarrow \frac{(x_i - \bar{x})^2}{n-1}$$ [these one is high]

So, definetly the lower variance of sample definetly approximately equal to population variance.

$\sigma^2 \approx s^2$ [Thats why $(n-1)$ is divide]

Standard deviation:

→ It is the measure of how much datapoints are far away from the mean.

→ Variance and SD both are Moreover Similar, The difference is SD is the root of variance.

Why root and SD are more preferable than Variance?

Suppose,

Calculate the variance for every data points with -Out the Square of the Mean and datapoints

i.e: $(x_i - \mu)$.

Some time negative value will be come,

like -3, -2, -1, 0, 1, 2, 3

⇓ In order to avoid, we need to Square.

After that, The variance is some big value because of Squaring. The original value is Changed to Squared Value.

In Squared Value is also preferable, but it takes Some time to solve.

⇓ In order to eliminate we need to get back original value.

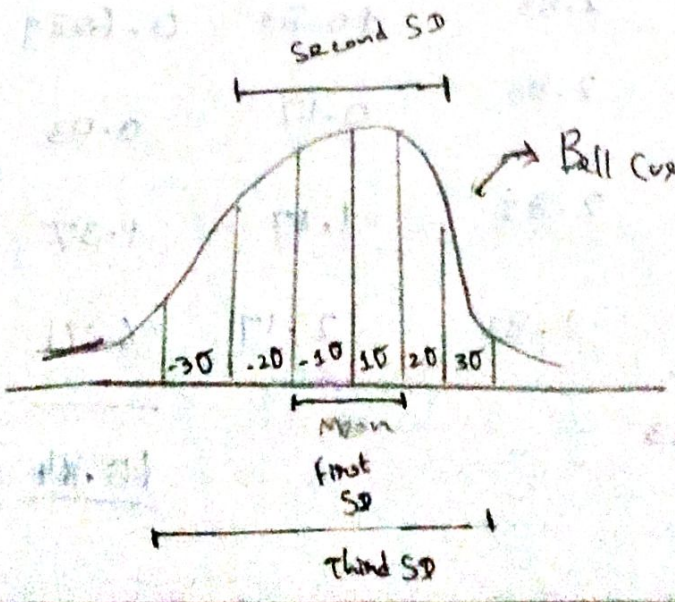
$\sqrt{(\ )^2}$ → first convert to negative to positive then Sqroot applied.

In Simple, SD is the measure of average of how ~~much~~ datapoints are far away from the mean as unit. It is the root of Variance and returns the Original Value. [SD also says 1SD, 2SD, 3SD ...]

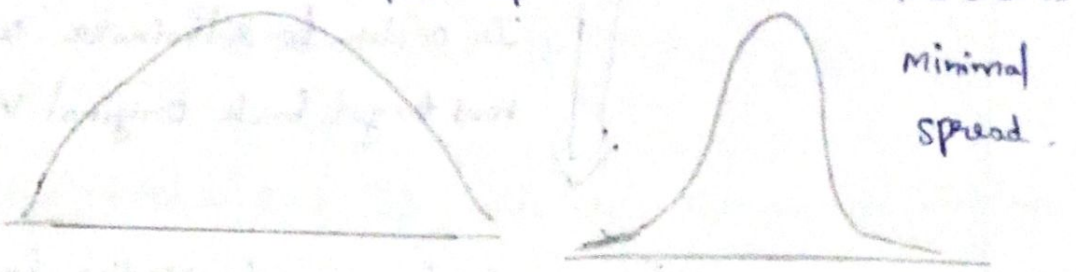
[Unit means like 1SD, 2SD, 3SD]

$\sigma = \sqrt{V.of.pop}$ SD FOR POPULATION

$S = \sqrt{V.of.sample}$ SD FOR ~~VARIENCE~~ SAMPLE



Which one is more Variance ?



Example :

Random Variable,

$$X = \{1, 2, 2, 3, 4, 5\} \quad \text{[Population]}$$

Calculate Variance,

x	μ	$x - \mu$	$(x-\mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	+0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$

$= \frac{10.84}{6}$

Var = 1.81

SD = $\sqrt{1.81}$

SD = 1.345

Percentiles and Quartiles: [FOR OBSERVING OUTLIER]

Percentage:

data = [1, 2, 3, 4, 5]

What is the percentage of numbers that are odd?

$$\text{Percentage} = \frac{\text{No of number that odd}}{\text{Total numbers.}}$$

$$= 3/4 = 0.6 \quad [60\ \%]$$

Percentile:

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

Example,

data = [2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12]

① What is the percentile of ranking of 10?

$x = 10$

$$\text{Percentile Rank of } x = \frac{\text{Number of value below}(x)}{n} \times 100$$

$x = 10$ $\quad$ $n(\text{sample}) = 20$

$$\text{Percentile Rank of } x = \frac{16}{20} \times 100 = 80\%$$

The 80% of data distribution is below 10.

Same way for 11,

$$\text{Percentile Rank of } x = \frac{17}{20} \times 100 = 85\%$$

② What value exists at percentile ranking of 25%?

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

$$= \frac{25}{100} \times (21) = 5.25 \quad \text{[It specifies the index]}$$

$5.25 \Rightarrow (5) \Rightarrow$ 5th position element is 5.

5 exists at percentile ranking of 25%.

Quartile:

→ Quartiles divide a dataset into four equal parts.

→ Each Containing an equal number of the observation.

→ It is used to understand the distribution of data.

→ It denoted, Q_1, Q_2 and Q_3.

Q_1 - 25 %

Q_2 - 50 % [Mean or median]

Q_3 - 75 %.

Five Number Summary:

1. Minimum
2. First Quartile (Q_1)
3. ~~Mean (or)~~ Median (Q_2)
4. Third Quartile (Q_3)
5. Maximum

Using

↓

To remove the Outliers.

Removing the Outlier:

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

↳ It may be outlier.

Lower fence ⟷ Higher fence

The lower and higher fence are actually says that the value should present in the between the fence. Otherwise, It Consider as Outlier.

formula,

$$\text{Lower fence} = Q_1 - 1.5(IQR)$$

$$\text{Higher fence} = Q_3 + 1.5(IQR)$$

InterQuartile Range = (IQR)

In the data distribution, The ranges between the Q_3 and Q_1 is called as Interquartile range. [It means ranges between 75% and 25%].

Formula: $IQR = Q_3 - Q_1$

The above data,

$$25\% = \frac{25}{100} \times (19+1) = \frac{\overset{5}{25}}{\underset{8_1}{100}} \times \overset{1}{20} = 5 \text{ [index]}$$

$$\boxed{25\% = 3}$$

$$75\% = \frac{75}{100} \times (19+1) = \frac{\overset{15}{75}}{\underset{8_1}{100}} \times \overset{1}{20} = 15 \text{ [index]}$$

$$\boxed{75\% = 7}$$

$Q_1 = 3$, $Q_3 = 7$

To find IQR,

$$IQR = 7 - 3 = 4$$

$$\boxed{IQR = 4}$$

lowerfence = $3 - 1.5(4) = -3$

Higherfence = $7 - 1.5(4) = 13$

[lowerfence $\leftrightarrow$ Higherfence]

[$-3 \leftrightarrow 13$] (27 not present in this range)

So, the remaining data are,

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9

So, five summary,

Minimum = 1

$Q_1 = 3$

Median = 5

$Q_3 = 7$

Max = 9

} To construct → Box plot.

Box plot: [Application for Data Visualization]