Variance

Variance measures the dispersion of a set of data points around their mean.

$$\text{Population Variance} = \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - M)^2}{N}$$

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

~~Sample Variance~~ ~~Population Variance~~.

* The reason behind n-1 in Sample Variance is in population Variance we will use total N which is 100% of data.

* In Sample Variance we are taking n-1, because, we dont have 100% of Values

* In a population of 1000, if u take Sample as 100, then to calculate ~~Std~~ ~~Std~~ Sample Variance the denominator is 100-1 that is 99.

* we have more uncertainity due to incomplete data.

Reason for squaring the values:-

* the reason for squaring the values because, the main aim of variance is to find the dispersion of data, which is distance, so the distance cannot be negative.

* It amplifies the effect of large differences

* Non-negative values dont cancel out.

Standard deviation (Preferred more than variance)

* It means how much or how far the data has been spreaded with respect to the mean.

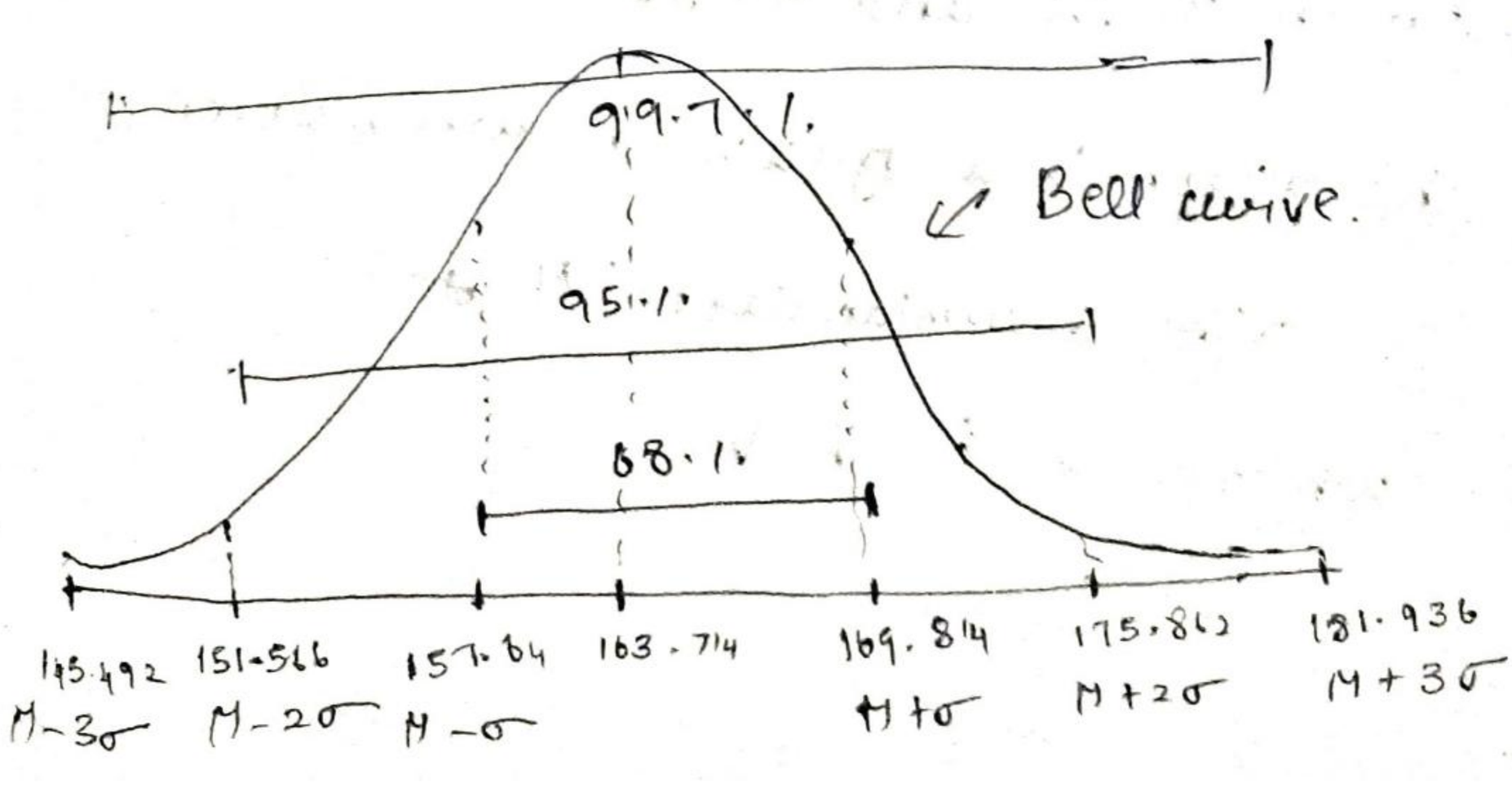
Population
* Stand.dev = $\sqrt{\sigma^2}$

Sample
Std. deviation = $\sqrt{s^2}$

$X = \{160, 165, 168, 170, 172, 175, 178\}$

mean = 169.714

$\sigma = 6.074$



* In that graph the data is normally distributed means it forms a bell curve

* The feature which is normally distributed gives better accuracy.

Emperical formulas:-

$P(M - \alpha \leq x \leq M + \alpha) = 68\%$

means 68% of data distributed between $M - \alpha$ and $M + \alpha$

$P(M - 2\alpha \leq x \leq M + 2\alpha) = 95\%$

means 95% of data distributed between $M - 2\alpha$ and $M + 2\alpha$

$P(M - 3\alpha \leq x \leq M - 3\alpha) = 99.7\%$

means 99.7% of data distributed between $M - 3\alpha$ and $M + 3\alpha$

* The graph is a gaussian distribution

$X \approx$ Gaussian dist (M, σ)

where x is R.V

Coefficient of Variation :-

- It is also called as relative standard deviation
- CV = Standard deviation / mean.
- coefficient of Variation (CV) plays a important role in comparing 2 datasets
- comparing standard deviation is meaningless but the comparison of coefficient of variation is meaning ful.
  - Population formula $C_v = \sigma / \mu$
  - Sample formula $\hat{C_v} = s / \bar{x}$

Example :-

Lets take a example of ~~a~~ pizza price in 2 different as 2 datasets.

Dataset 1

America

currency	Price
$	1.00
$	2.00
$	3.00
$	3.00
$	5.00
$	6.00
$	7.00
$	8.00
	9.00, 11.00

Dataset 2

Mexico

currency	Price
MXN	18.81
"	37.62
"	56.43
"	56.43
"	94.05
"	112.86
"	131.67
"	150.48
	169.29
	206.91

To calculate

1. Find whether it is sample or population.
2. Find mean
3. Find Sample Variance
4. Find Sample Std. dev

This dataset contains just price info of 10 hotels in america and mexico, so it is a sample population.

	Dataset 1	Dataset 2
Mean	5.50	103.46
Sample variance	10.72	3793.69
Sample std deviation	3.27	61.59

∴ We cannot compare 3.27 and 61.59 but we can compare its coefficients.

Sample Coefficient of variation	0.60	0.60

- So we obtained same results
- So both the datasets have same Variability.
- Coefficient of Variation does not have a unit of measurement
- It is universal across datasets.

Percentiles and quantiles:-

heights = {168, 170, 150, 160, 182, 140, 175, 180, 170, 190}

* First we need to sort

Sort = {140, 150, 160, 168, 170, 170, 175, 180, 182, 190}
1 2 3 4 5 6 7 8 9 10

* 5th percentile ~~Value~~ = basically means 5th Value.

★ 5th percentile = 170

— 5th percentile means 50% of data points are less than 170 and 50% of data are more than 170

★ 8th percentile = 180

— It means 80% of data are less than 180 and 20% of data are more than 180.

~~quantiles means 0th 25th 50th 100th percentiles~~

↓

~~4 divisions~~

of

quantile means 25.1.th 50th 75th 100th
1st 2nd 3rd 4th quantiles

Percentile.

Lets consider Amazon delivery time report of the products

$X = \{4d, 5d, 4d, 4.5d, 5d, 5.2d, 5.3d$ etc.....$\}$

Here in this type of cases 95th percentile and 99th percentile plays major role.

* If 95th percentile is 5.6 days, it means 95.1. of products are delivered within 5.6 days from the time of placing order.

* If 99th percentile is 7 days, it means 99.1. of products are delivered within 7 days from time of placing order.

95th .1. = 5.6 days

99th .1. = 7 days

~~So the supply unit should work on to reduce the time btw 95th and 99th .1.~~

~~In 95~~

Considering 95th .1., 5.1. product are not delivered in 5.6 days, it taking extra time, So supply unit of amazon will work on this to reduce time.

Interquartile range :-

- IQR also called as midspread or middle 50%, technically H-spread is the difference between the third quartile and first quartile

$$IQR = Q_3 - Q_1$$
$$75th - 25th$$

- IQR has breakdown point of 25% due to which it often preffered over total range.
- IQR is used to build box plots
- IQR identifies outliers
- IQR gives central tendency of data.

∴ dataset of high IQR has more Variability

∴ dataset of low IQR is preferable.

ex:- $x = \{5, 2, 1, 7, 3, 4\}$

1 Step:- Sort it

$x = \{1, 2, 3, 4, 5, 7\}$ $N = 6$

median = $3+4/2 = 3.5 = Q_2$

Q_1 is median of n smallest values

that is Q_1 is median of ~~first~~ first 3 values

$$Q_1 = 2$$

Q_3 is median of n largest values

$$Q_3 = 5$$

$$IQR = Q_3 - Q_1 = 5 - 2 = 3$$

Median Absolute deviation

~~It calculates the~~

It calculates the spread of the data with respect to median.

$$\text{deviation} = |x_i - \text{median}|$$

Since the modulus is used the value is absolute.