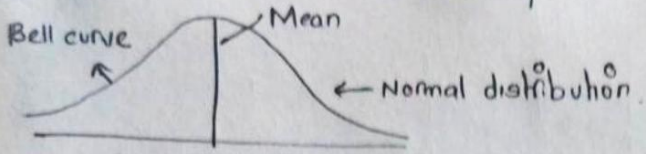
Mean - Measure of Central tendency.



It is disturbed by outliers.

Median - It is also a measure of central tendency. It is not affected by the outliers (not vastly affected).

Mode - Frequency distribution of categorical value. Category which have the highest frequency will be the mode.

EFFECTS OF OUTLIERS ON SPREAD AND CENTRE.

DATASET - 26 15 20.5 31 -350 31 30.5.

Measure.	With Outlier	Without Outlier	Test
Mean	-28	25.667.	Affected.
Median	26	28.25.	Resistant
Mode	31	31	Resistant
Range	381	16	Affected.
Standard deviation			It will also be affected because it contains mean in the formula.

Percentile →

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

So 10 percentile = 1. 20 percentile = 2.

So, if we find out 50^{th} percentile 5, then 50 percentage of numbers are less or equal to 5.

90^{th} percentile 9, 90% (percentage) of data is less or equal to 9. And only 10% of data is greater than 9.

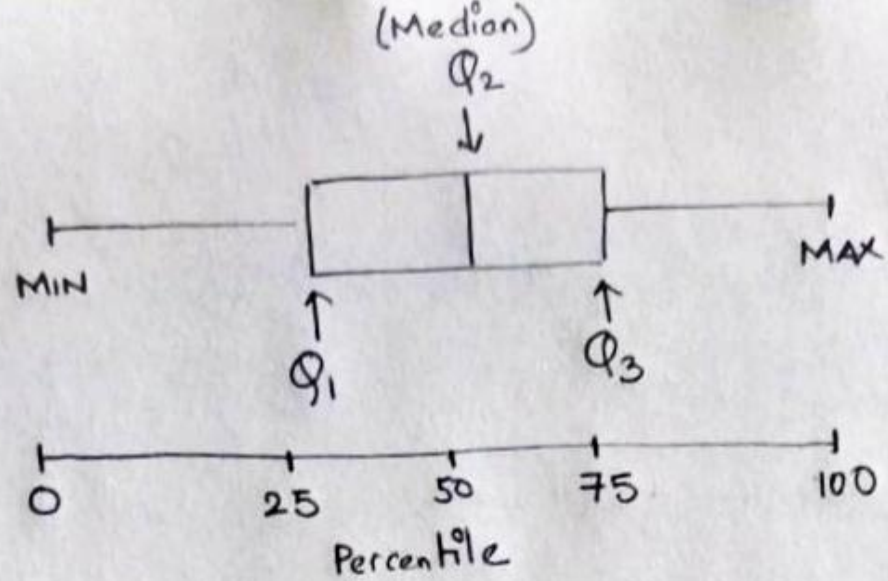
Height = {168, 170, 150, 160, 182, 140, 175, 180, 170, 190}.

Sort in Increasing order = {140, 150, 160, 168, 170, 170, 175, 180, 182, 190}

1 2 3 4 5 6 7 8 9 10

Suppose, 50% percentile, 50% of data is less than or equal to 170.

Box Plot



Eg - 18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22

Sort → 15 18 20 | 25 29 32 | 34 38 41 | 46 54 76

22 (Q_1), 33 (Q_2), 43 (Q_3)

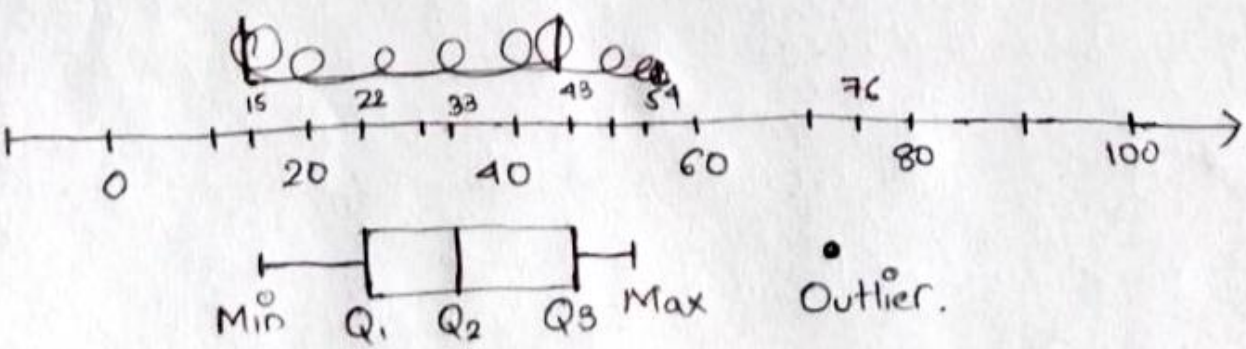
15 → Min, 54 → Max, 76 → Outlier.

Find outlier, $Q_1 - 1.5(IQR)$, $Q_3 + 1.5(IQR)$ = $\cancel{Q}$ $1.5(IQR)$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 43 - 22 \\ &= 21. \end{aligned}$$

$$\text{Outlier range} = [22 - 1.5(21), 43 + 1.5(21)] = [-9.5, 74.5]$$

So, 76 is an outlier.



Standard Deviation (σ) – SD is a measure of how spread our numbers are.
Its symbol is σ.
It is a square root of Variance. $= \sqrt{\text{Variance}}$.

$$SD = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

Variance (σ^2) – The average of the square difference from Mean.

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

3 steps – i) Find out the mean.
ii) Then from each number - subtract the Mean and square the result
iii) Work out on the average of those square.

Eg – Heights – 600mm 470mm, 170mm, 430mm and 300mm.

$$\text{Mean} = \frac{600+470+170+430+300}{5}$$
$$= \frac{1970}{5}$$
$$= 394$$

So mean/average height is 394mm.

$$\text{Variance}, \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$
$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$
$$= \frac{108520}{5} = 21704.$$

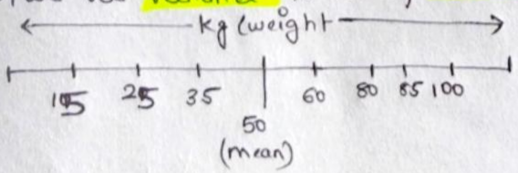
So variance is 21704.

Standard deviation $= \sigma = \sqrt{21704} = 147.3$
So SD is 147.

In case of $n-1$, $\frac{108520}{4} = 27130$ $\sigma = \sqrt{21730}$
$= 147.4 = 147$

Standard deviation tells us how do close the values are in dataset are to mean. Small SD, means small variability in a dataset. In other word, all data point are around mean which is less spread out.
High SD, means high variability in a dataset / high spread out.

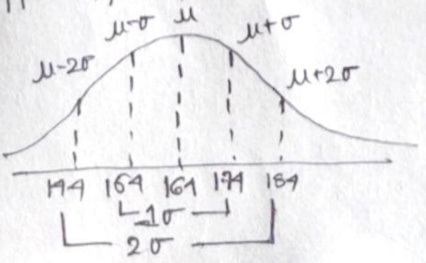
So, we use variance is check how spread the data is.



So, if we need to find variance, mean - data point
$= -35 - 15 - 25 + 10 + 30 + 35$
$= -75 \quad +75 = 0$
So, we squared the deviation from mean.

Why do divide by $n-1$?
Variance is the average square deviation from population mean.

Suppose, you know the mean (μ) = 164 and SD (σ) = 10, then.



$1\sigma \rightarrow$ 1 standard deviation, $2\sigma \rightarrow$ 2nd standard deviation
So, the next value of 164, if we move to right will be 174. And if we move to the left then it will be 164.

Covariance – Quantifying a relationship between variables. Direction of a relationship.

Size	Price.
1200 sq/m	₹1L.
1800 sq/m	₹2L
2500 sq/m	₹3L.

Quantify a relationship between Size and Price.

Size ↑ Price ↑.

Size ↓ Price ↓

$$\text{cov}(\underset{x}{\text{size}}, \underset{y}{\text{price}}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

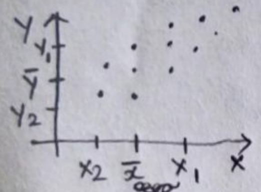
$$\text{Variance}(x),\ \text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})*(x-\bar{x})$$

$$\text{Cov}(x,x) = \text{var}(x)$$

So, through covariance we will get a value, if $x\uparrow, y\uparrow = \square$ +ve. Covariance

$x\uparrow, y\downarrow = \square$ −ve. Covariance.

Suppose $X\uparrow\ Y\uparrow$.

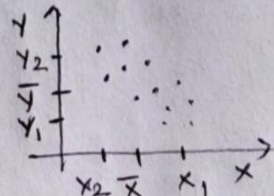


$= (x_1-\bar{x}) * (y_1-\bar{y})$
$= (+ve) * (+ve)$
$= +ve.$

$= (y_2-\bar{y}) * (x_2-\bar{x})$
$= (-ve)*(-ve)$
$\doteq +ve.$

So, always covariance is +ve when $x\uparrow, y\uparrow$.

Suppose $X\uparrow, Y\downarrow$



$= (x_1-\bar{x})*(y_1-\bar{y})$
$= (+ve) * (-ve)$
$= -ve.$

$= (x_2-\bar{x})*(y_2-\bar{y})$
$= (-ve)*(+ve)$
$= -ve.$

Covariance find the direction of relationship. But we don't know exact value of strength.

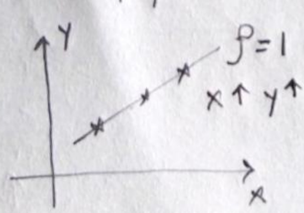
So, always covariance is −ve when $X\uparrow Y\downarrow$.

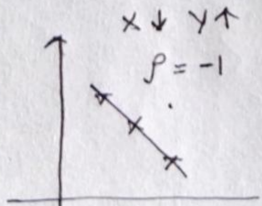
Pearson Correlation Coefficient → Strength of the relationship between variables and direction also of the relationship.

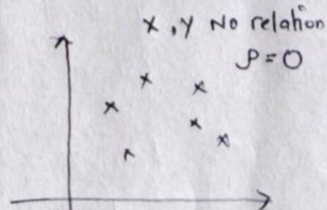
$$P(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x\ \sigma_y}$$

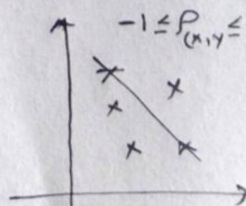
Range is → $-1 \le P(x,y) \le 1$.

Cov(x,y) → covariance (x,y)
σ_x → Standard deviation of x.
σ_y → SD of y.

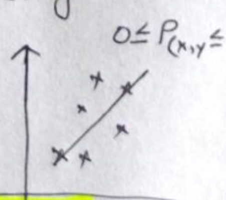








All points are not on straight line.



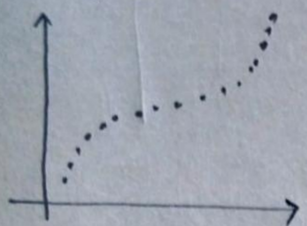
Suppose we have 3 variables a, b, c.
a and b have pearson correlation coefficient of 1. In short they are same, then we will remove one of the variable/feature.

We also use spearman rank Correlation (Refer to Numerical Measure-Mean)

Pearson work only good for linear relationship. But for others we use Spearman rank correlation coefficient. (Refer to page 14)

Spearman's Rank Correlation coefficient

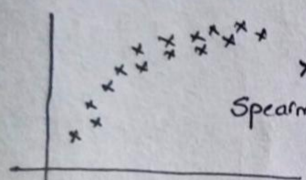
If the plot is non-linear relationship.



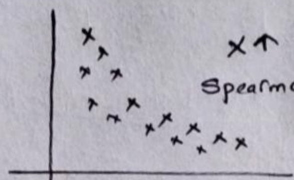
Spearman correlation = 1.
Pearson correlation = 0.88

$$\text{Pearson} = \frac{\text{cov}(x, y)}{\sigma_x \, \sigma_y} \qquad \text{Spearman} = \frac{\text{Cov}(\text{rank}_x, \text{rank}_y)}{\sigma_{\text{rank}_x} \, \sigma_{\text{rank}_y}}$$

- Pearson correlation assesses only linear relationship. whereas Spearman correlation assesses monotonic relationship (whether linear or not).
- Spearman correlation is from range -1 to 1.
- If the two variables are very similar, then spearman correlation will be 1.
- If the two variables are dissimilar / fully opposed, then spearman correlation will be -1



x ↑ y ↑.
Spearman correlation = 0.92



x ↑ y ↓.
Spearman correlation = -0.92

x → Independent Var
y → Dependent Var.

Sign of Spearman correlation Indicates the direction of association between x and y

If Spearman correlation = 0, then there is no tendency of y increase or decrease when x is increased.

Steps in the methods are -
i) Sort the data and assigned the rank.
ii) sort the second column and assign the rank.
iii) create a difference column based on rank.
iv) Create final column based on difference by squaring

IQ	Hours of TV	sort →	IQ (x_i)	Hours of TV (y_i)	rank (x_i)	rank (y_i)	d_i	d_i^2
106	7		86	0	1	1	0	0
100	27		97	20	2	6	-4	16
86	2		99	28	3	8	-5	25
101	50		100	27	4	7	-3	9
99	28		101	50	5	10	-5	25
103	29		103	29	6	9	-3	9
97	20		106	7	7	3	4	16
113	12		110	17	8	5	3	9
112	6		112	6	9	2	7	49
110	17		113	12	10	4	6	36

$\Sigma d_i^2 = 194$.

$n = 10$.

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

Shortcut formula.

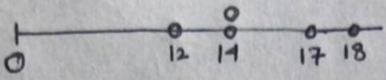
$$= 1 - \frac{6(194)}{10(10^2-1)}$$

$= -0.175$.

So, $\rho = -0.175$ which is very close to 0. So there is no correlation between IQ and watching TV for hours.

MOMENTS

[12 14 14 17 18].

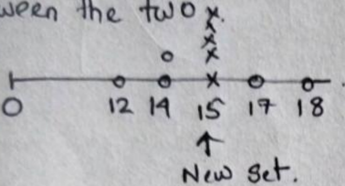


$\frac{\Sigma x_i}{n}$ → Average distance from 0 → **First moment** (linear distance)

$\mu_1 = \frac{\Sigma x_i}{n} = \frac{12+14+14+17+18}{5} = \frac{75}{5} = 15.$

So, this **dataset is averagely distance by 15 from 0**.

But there will be other set also which will have mean 15. So how can we differentiate between the two.



Use **Second moment**,

$\frac{\Sigma x_i^2}{n}$ → Average square distance from 0 → Second moment (crude)

$\mu_2' = \frac{(15)^2+(15)^2+(15)^2+(15)^2+(15)^2}{5}$ $\quad \mu_2' = \frac{(12)^2+(14)^2+(14)^2+(17)^2+18^2}{5} = 229.8$

$= 225.$ Whenever greater spread, then greater second moment!

$\underset{(15)}{\mu_2'} \neq \underset{(12,14,14,17,18)}{\mu_2'}$ Greater value because the number will have high numbers. (sequence)

To remove the effect of first moment, to second moment.

$\frac{\Sigma (x_i - \mu_1')^2}{n}$ (mean) → Average square distance from mean → **Second moment (centered)**

$\mu_2' = \frac{(15-15)^2 + \dots + (15-15)^2}{5} = 0.$

$\mu_2' = \frac{(12-15)^2+(14-15)^2+(14-15)^2+(17-15)^2+(18-15)^2}{5} = 4.8$

So in (15…15) series, $\mu_2' = 0$. It variance is 0.

In (12, 14, …18) series, $\mu_2' = 4.8$. It variance is 4.8.

High Order Moment

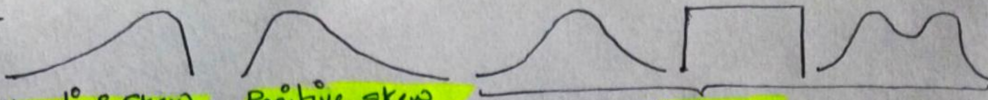
μ → population mean, σ – population SD.

If we use sample, then instead of μ, use $\bar{x}$

	Centered	Standardised	
① $\frac{\Sigma x}{n}$ ← **Mean** population			
② $\frac{\Sigma x^2}{n}$	$\frac{\Sigma (x-\mu)^2}{n}$ ← **Variance** population		
③ $\frac{\Sigma x^3}{n}$ ⟹	$\frac{\Sigma (x-\mu)^3}{n}$ ⟶	$\frac{1}{n} \frac{\Sigma (x-\mu)^3}{\sigma^3}$	← **Skewness** population
④ $\frac{\Sigma x^4}{n}$	$\frac{\Sigma (x-\mu)^4}{n}$	$\frac{1}{n} \frac{\Sigma (x-\mu)^4}{\sigma^4}$	← **Kurtosis** population

Crude (Distance from 0) $\xrightarrow{\text{Remove effect of 1st moment}}$ Centered from mean $\xrightarrow{\text{For 3rd \& 4th moment, remove 1st \& 2nd moment}}$

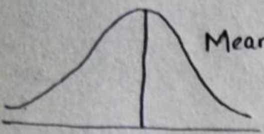
Skewness –



– Which way the tail is pointing.

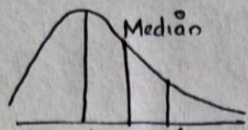
For negative skew:
mean < median < mode

Symmetrical distribution

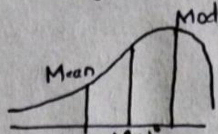


Mean = median
= mode.

Positive skew



Negative skew



For positive skew:
mode < median < mean

The greater the skew, Greater is the distance between mean, mode & median

To calculate value – Use pearson method = ① Mode skewness ⇒ $\text{skew} = \dfrac{\text{mean} - \text{mode}}{\text{std dev}}$

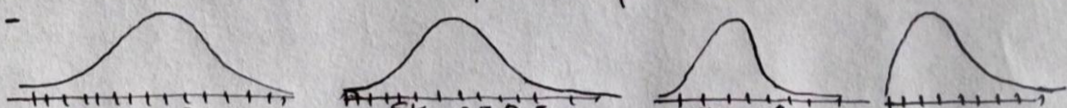
② Median skewness $= \dfrac{3(\text{mean} - \text{median})}{\text{std dev}}$

If the dataset is small, the mode do not work out because many number have equal high frequency.

Another method to find skewness →

Third (centralised) moment $= \dfrac{1}{n} \dfrac{\Sigma(x-\mu)^3}{\sigma^3}$ population skew | $\dfrac{n}{(n-1)(n-2)} \dfrac{\Sigma(x-\bar{x})^3}{s^3}$ sample skew

Some example –



skewness numberline.

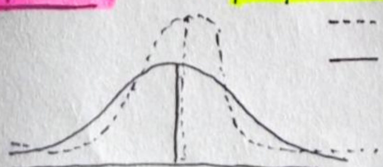
< -1   -1   -0.5   0   0.5   1   > 1.

$-1 > x > 0.5$ ⊢ Approximate symmetric ⊣

⊢ Moderately skewed $0.5 < x < 1$ ⊣

$x < -1$ Highly skewed $x > 1$

Kurtosis – The "peakedness" of a distribution.



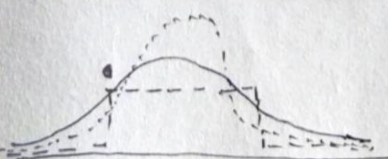
---- $\mu(\text{mean}) = 0$, $\sigma(\text{SD}) = 5$, skew = 0

—— $\mu = 0$, $\sigma = 5$, skew = 0

So, kurtosis will help to differentiate between this two distribution | ---- is more peaked and has fatter tails.

Kurtosis ranges from 1 to infinity (normal distribution = 3)

Fourth (standardised) moment = $\dfrac{1}{n}\left(\dfrac{\Sigma(x-\mu)^4}{\sigma^4}\right)$ population kurtosis



Excess kurtosis = kurtosis − 3.

Excess kurtosis ranges from −2 to infinity (normal distribution = 0)

Note – Outlier contributes greatly to kurtosis.

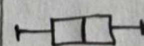
— Normal distribution, kurtosis = 3, mesokurtic

····· kurtosis > 3, lepokurtic

- - - kurtosis < 3, platykurtic

– Thickness of tails, distribution with high kurtosis will likely have a higher peak as well.

NUMERICAL MEASURE OVERVIEW

MEASURE	OVERVIEW	FORMULA
Mean	Measure of central tendency. Sum of the data values divided by data count. Affected by outliers.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	Also a measure of central tendency. It is an observation/value at the middle when the data is sorted in decending order. NOT AFFECTED by outliers.	
Mode	Most frequenty observed variable in a dataset. Might or Might not affected by outliers.	
Quartile	There are several quartile of an observations. Data is sorted first in ascending order. First quartile → Lower quartile → 25% of data. Second quartile → Median → 50% of data. Third quartile → Upper quartile → 75% of data.	
Percentile	N^{th} percentile of an observation variable is the value that cuts off the first n percent of the data. values when it is sorted in ascending order.	
Range	The range of an observation variable is the difference between its largest and smallest value. It measure how far apart the entire data spread in terms of value. Affected by outliers.	Range = Largest value − smallest value
IQR	Interquartile range of an observation variable is the difference of its upper and lower quartile. It measure of how far apart the middle portion of data spread in value.	IQR = Upper quartile − Lower quartile
Box plot	Graphical representation based on quartiles, smallest and larget values.	
Variance	It is a numerical measure of how the data values is dispersed around the mean. Affected by outliers.	$\sigma^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$
Standard deviation	Standard deviation is the square root of variance	$\sigma = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x}_i)^2}$
Covariance	If we have two variables x and y then how two are linearly related. It can be positive or negative.	cov = $\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$
Correlation coefficient	Strength of the relationship between variable. i) Pearson → Works most on linear relationship. ii) Spearman → work good both on linear & non linear relationship	
Central moment	Central moment. 1st central moment → moment about mean. 2nd central moment is Variance.	
Skewness Kurtosis	3rd moment is skewness. How data distribution is skewed. 4th moment is kurtosis. Tail shape of data. Normal distribution have zero Kurtosis.	