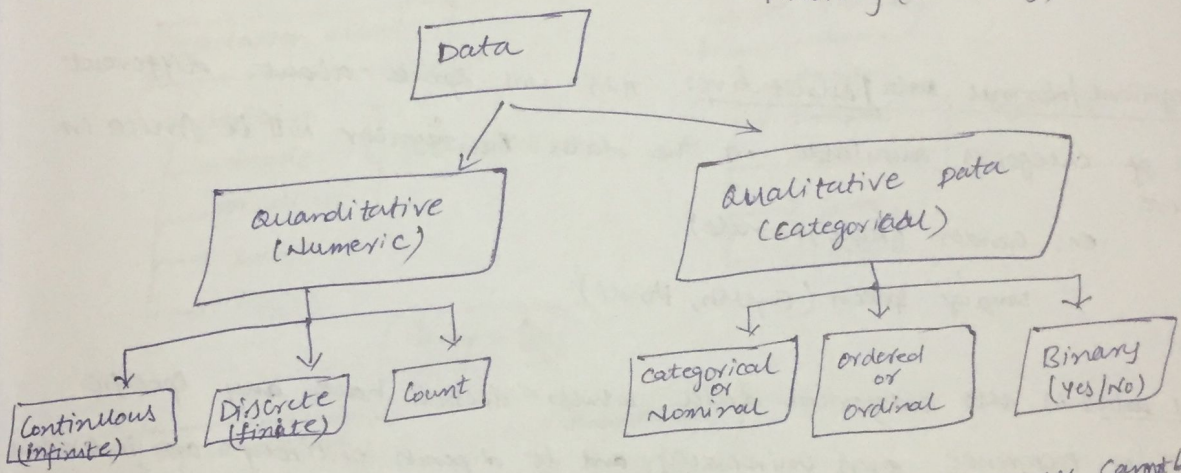
Statistics

Data: Collection of facts which are un organised which is in raw format. This data has to be processed which will be meaningfull.

Information: Collection of un organised facts which is processed and arranged in organised way to give meaningfull information.
(or)
Data that has been processed will have meaningfull insights.
Processing (cleansing, ETL)



Continous: variable will have ~~range of~~ infinite values. and this values cannot be ~~decimal~~. countable & usualy represented in delimals.

Ex: ~~height of~~ Students ~~in~~ a class
~~weight of~~ Students in a class
~~Temperature~~ in ~~this month~~

Ex: → Age which is continious if the measure of granularity is in minutes and seconds.

→ Population which is continious for every sec there is a new born or dead

→ Temperature which is continious every hour it keeps on changing

How ever on circumstance this continous variable can be discret variable.
Ex: Dividing Students on basis of Age (categorical & Discrete)

Discrete: variable will have there values as whole numbers. Such as counts of data.

Ex: No of lang a student can speak (1, 2, 3)
Political parties in a state (2, 3)
Religions in a country (2, 3)
Different hair colors (2, 3)

Count Data: Data that represents counts of data. (i.e Discrete data)

Categorical / Nominal Data / Qualitative: This will speak about different types of categories available in the data. The number will be finite in nature.

Ex: Gender (Male, Female)
Language Spoken (English, Hindi)

Nominal Data: is also categorical data which doesn't have any specific ranking or preference given universally and it depends entirely on individual interest. The represented data will not have any preference.

Ex: Choose a color (Red, Green, Blue) you like
Choose a Lang (Hindi, French, English) you like
Which fruit you like (Mango, Apple, Strawberry)
Choose a car (BMW, Benz, Ferari) which you like.

Ordinal Data: Which will have ordered set of data.

How likely is your math teacher teaches.

Ex: worst, bad, average, good, very good.
→ (It is ordered)

Population & Sample

Population (N): is the collection of items of interest to our expirement / study which is denoted by "N".

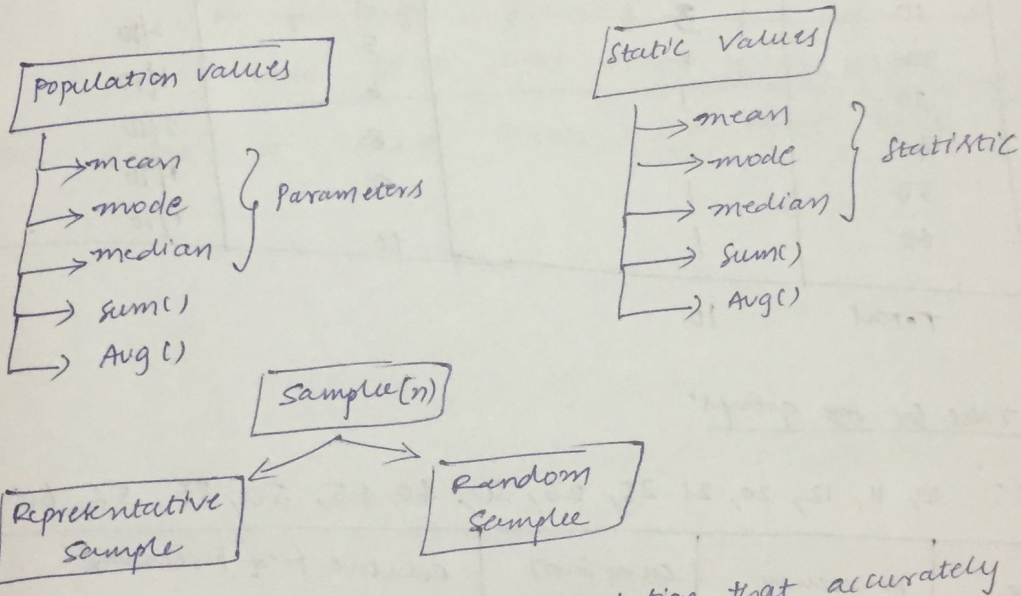
Ex: All the college students (N)

Sample (n): is a sub set of population which is denoted by "n"

Ex: 10 students from each class (n)

n = Sample = Size of the sample. = no of units taken in Sample.

Values refer to population are called "Parameters" and values refer to samples are called "Statistic".



Representative Sample: is a sub set of population that accurately reflects the actual population.

Random Sample: is picking the items in equall proportinate (or) equal chance w.r.t population. This method is less likely used in real world and much complicated to pick the sample in equall proportinate.

Frequency Distribution

~~Discrete~~ ~~Continious~~

Frequency distribution: is a table/graph that displays the frequency/count of various outcomes in a sample.

Ex: 10, 10, 20, 50, 60, 10, 20, 30, 40, 40. (Total 10 data points)

Frequency Table for above dataset

Students	Frequency/count	~~Relative~~ Cummulative Sum	Relative Frequency Percentage
10	~~3~~	3	3/10
20	2	5	2/10
30	1	6	1/10
40	2	8	2/10
50	1	9	1/10
60	1	10	1/10
Total	10		

Frequency Table for Age groups:

Ex: Age: 10, 11, 12, 20, 21 25, 26, 30, 40, 45, 50, 55, 56, 60

Age	Frequency	Cumm Sum	Relative freq Percentage
0 - 10	1	1	1/14
11 - 20	3	4	3/14
21 - 30	4	8	4/14
31 - 40	1	9	1/14
41 - 50	2	11	2/14
51 - 60	3	14	3/14

the frequency (Count)

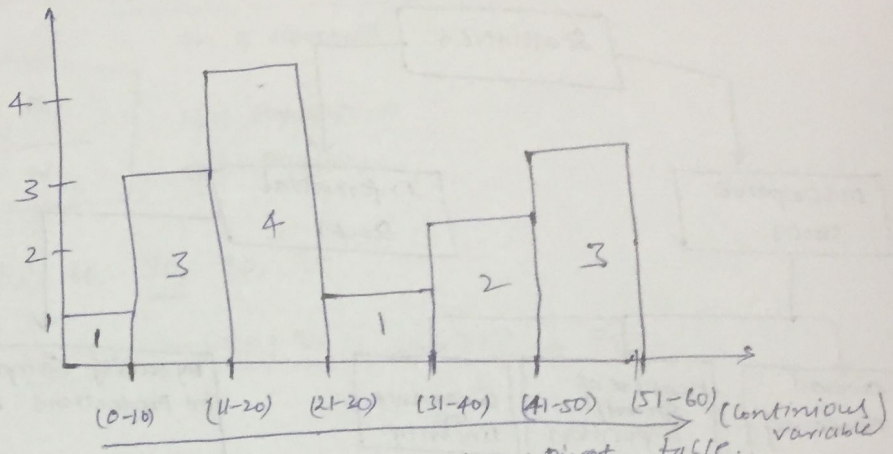
10 data points)

lative requency	Percent
/10	
/10	
10	
10	
0	
0	

56, 60

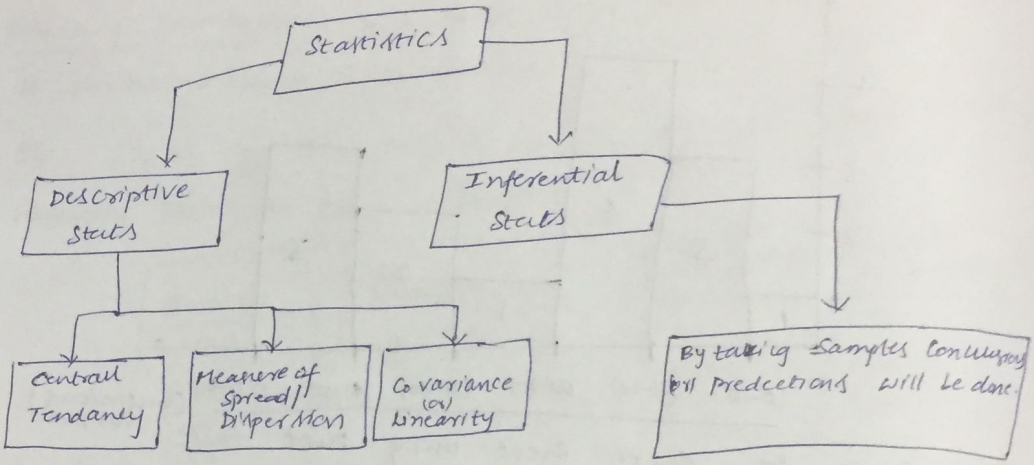
Usually frequency tables are plotted in Histograms.

Histogram: Graphically plotting the distribution of items from the sample/ population.



plot the same in Excel sheet using pivot table.

Statistics: is a collection of data and analyse the data and interpret the data in graphs.



- mean ($\mu, \bar{x}$)
- mode
- median

} Central Tendency

- Variance
- Standard Deviation
- ~~Co variance~~
- Coefficient of Variation
- Skewness & Quartiles

} Measure of Spread

- Co variance
- Co relation coefficient

} Co variance

Descriptive Statistics: Usually will have less population and hence calculation (or) stats will be applied for the entire data set.

Central Tendancy: Measure of central value in a data set can be measured with below formulaes.

mean (μ) average $\frac{\text{Sum of all the numbers}}{\text{no of elements}}$

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

N = Population

n = Sample

Marks = 50, 60, 70, 80, 90

$$\text{Mean} = \frac{50+60+70+80+90}{5} = \frac{350}{5} = 70.$$

Set = 5, 10, 100, 90, 500, 600, 700, 1400, 1500

Set1 = 5, 10, 15, 20, 25, 30, 70, 80

mean = 31.875 (this mean doesn't make any sense out of the data set) this is because of data points 70, 80 which are called outliers.

outliers: will be far away from the central point (mean) hence mean value will be effected.

mean denoted by μ (population) $\bar{x}$ (sample)

* Mean is not a good measure in presence of outliers in such cases the measure of center can be calculated by median.

the mid point element of sorted data set will be median.

Median: Divide the dataset into two equall halfes and the central element will be median.

The data set has to be sorted before making in to equall halfs.

Even data set $S1 = 1, 2, \boxed{3, 4}, 5, 6$

$\frac{3+4}{2} = 3.5$ is the median.

odd data set $S2 = 8, 5, 4, \boxed{6}, 1, 3, 2$

$\text{sort}(S2) = 1, 2, 3, 4, 5, 6, 8$

~~median = 6. X~~

Sort the data set

$S2 = 1, 2, 3, 4, 5, 6, 8$

median = 4 ✓

* Median is the positional measure, it doesn't matter for outliers.
* If mean is not near to median it indicates there are outliers.

Mode: The most frequently occuring dat point will be mode.

$S1 = 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3$

mode = 3 (since '3' is the most repetative element out of other elements)

* If mean, median & mode are near to each other then the data set is good for analysis.

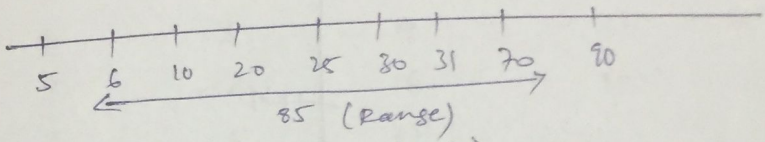
Measure of Dispersion / Spread / variance

Range: calculating the range max(val) − min(val)

S = 40, 10, 20, 25, 70, 5, 6, 30, 90, 22, 31

∴ max(val) − min(val) = 90 − 5 = 85

Plot the data set:



R = 85 (Denoted by R).

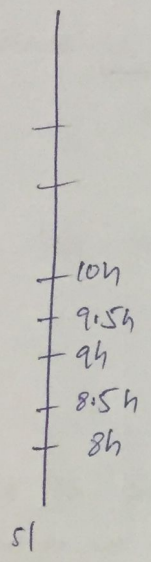
Variance: is used to calculate the distance of each data point from the mean.

Assume we have two working hours data sets.

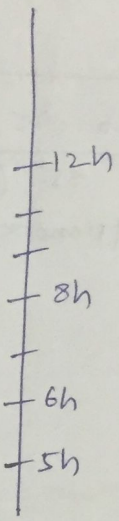
S1 = 8.5h, 10h, 8h, 9h, 9.5h (working hours of weekly)

S2 = 12h, 5h, 14h, 6h, 8h

$\bar{x}(S1) = 9$ $\bar{x}(S2) = 9$



S1



S2

Both the means of above data sets are same but there is a variance/dispersion in the data set 2.

Data set 1 values are near to the mean value

Data set 2 values are dispersed to the mean value

To identify the variance between the data sets we will calculate variance (σ^2) Population s^2 (Sample).

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

$(n-1)$ here ~~n~~ n.

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$	$\frac{\sum x-\bar{x}}{n}$
8.5	9	−0.5	0.25	
10	9	1	1	
8	9	−1	1	0.5
9	9	0	0	
9.5	9	0.5	0.25	

$\sigma^2(S1) = 0.5$

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$	$\Sigma(x-\bar{x})^2$
12	9	3	9	
5	9	−4	16	12
14	9	5	25	
6	9	−3	9	
8	9	−1	1	

$\sigma^2(S2) = 12$

The mean = 9 for both data sets but using variance we were able to identify the variance.

Note: Since variance is calculated by square of numbers the variance value will be dis proportionate with the actual data set points hence standard deviation is used. Interpreting the value of variance with mean is difficult hence Stand Deviation.

Standard Deviation (σ): is the squere root of variance

$$\sigma = \sqrt{\sigma^2}$$

Which is used to calculate how far the data points each are spread out from the mean point.

* Standard deviation close to "Zero" it indicates the data points are tightly coupled/near with to mean.
* If the SD is away from "Zero"/high it indicates the data points are away from mean.

Ex: Profit Percentage of companies of different months.

A = 43, 44, 0, 25, 20, 35, -8, 13, -10, -8, 32, 11, -8, 21, 15 Mean = 15
B = 17, 15, 12, 17, 15, 18, 12, 15, 12, 13, 18, 18, 14, 14, 15 Mean = 15

Data set 'A' data points are dispersed away from mean = 15
Data set 'B' data points are near to it's mean = 15

* In this kind of tricky situation where both means are same and this can be identified by Standard Deviation

SD(A) = 19.47 (Dispersion is away from mean)

SD(B) = 2.2 (Dispersion is near to mean since SD is near zero)

*From this we can identify company 'A' profits are zero & -ve in few months and it is not constantly performing.
Company 'B' is constantly performing profits and this is proved by SD.

Coefficient
to it's mea
→ this notio

CV is use

Mean	
SD	
CV	

In order
it doesn'
So here

Coefficient of Variation: is the percentage of standard deviation to it's mean.

→ This nothing but how much % of SD(s) w.r.t mean ($\bar{x}$)

$$CV = \frac{s}{\bar{x}} \times 100$$

CV is used to compare ~~the~~ two or more SD's in meaningfull way.

	Test1	Test2
Mean	50	100
SD	10	15
CV	20.0	15.0

In order to compare the variability of SD b/w two data sets it doesn't make sense since means are different and far away. So here comes CV to compare the variance of SD's.

Skewness: is the measure of asymmetry which indicates whether the data points are concentrated on which side.

* Skewness will help in identifying the slope of distribution.

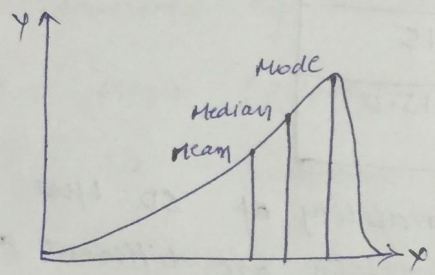
Right Skew / Positive Skew

Left Skew / Negative Skew

No / Zero Skew

Left Skew: In Left Skew the outliers are away from the y-axis.

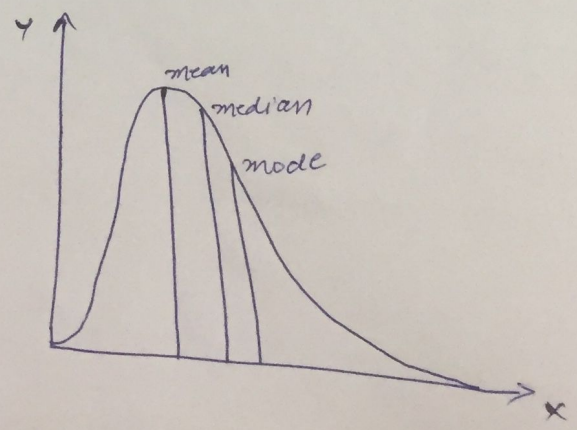
mean < median < mode



Left Skew (Negative Skew).

Right Skew: In Right Skew the outliers / distribution is towards the y-axis.

mean > median > mode



indicates
side.
distribution.

from the y-axis.

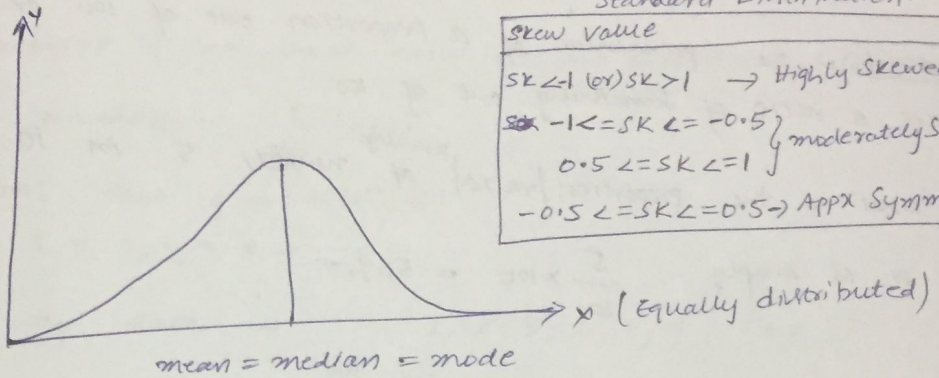
is towards

No-Skew: If there is zero skew then the distribution is a Normal distribution curve. (Bell curve)

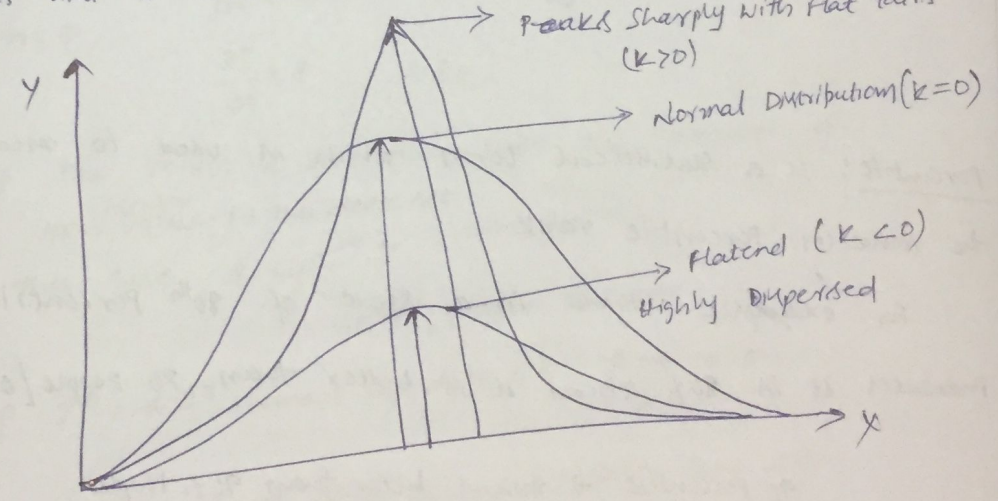
mean = mode = median

Skewness Table w.r.t Standard Distribution.

Skew value	
SK < -1 (or) SK > 1	→ Highly Skewed
-1 <= SK <= -0.5	moderately Skewed
0.5 <= SK <= 1	moderately Skewed
-0.5 <= SK <= 0.5	→ Appx Symmetric



Kurtosis: is used to measure whether the curve is normal, flat or peaked



Kurtosis is measured with 'K'.

→ If the value of 'K' is near to "Zero" it is Normaly Distributed

→ If the value of 'K' is -ve then the curve is Flat.

→ If the value of 'K' is +ve or K>0 then curve is Peak. (Large Number)

Covariance & Correlation Coefficient

Percentage & Percentile:

Percentage (%) is the representation of a proportion out of 100. or it depicts a ratio of something out of 100.

Ex: What is the proportion/ratio/percentage of number 5 in 10.

It is simply $\frac{5}{10} \times 100 = 50\%$.

What is the proportion of 4 in 80

$$\frac{4}{80} \times 100 = 5\%$$

Percentile: is a statistical term which is used to measure the rank (or) Percentile rank.

For example there is a store of 80th Percentile it indicates it is 80% ahead it is better than 80 people/observations.

90 Percentile → means better than 90% people.

By using Percentile rank can be determined.

90 Percentile = 10 rank

95 Percentile = 5 rank

100 Percentile = 1 rank

Percentile is used to calculate/identify at which percentage the data point is standing when compared to other data points.

Definition: A Percentile is defined as a point below which certain percentage of observations lie.

Percentile is calculated based on all observations.

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

Calculation: Score of 8 students.

$n = 5, 4, 2, 6, 9, 8, 1$

① Sort the data → 1, 2, 4, 5, 6, 8, 9 $(n = 7)$

② I want to calculate → 33 percentile

③ $P = 33$, $n = 7$

$$\frac{P}{100}(n+1)$$

$$\frac{33}{100} \times 8 = 2.64$$

Case 1: If the result is decimal take the average of lower and upper integer value in the data set.

In this case 2.64 → 2 (lower Integer)
→ 3 (upper Integer)

Avg of 2nd and 3rd element of data set $\frac{2+4}{2} = 3$

Take the data in excel sheet and calculate percentile.

Quartiles: divide the data in to four segments according into where the number falls.

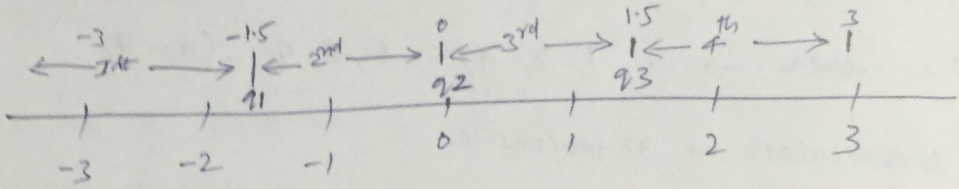
Quartiles are directly based on Percentiles.

Quartiles are basically 3 types

(Lower Quartile) quartile-1 (q1): is the 25th percentile data.

(Second Quartile) quartile-2 (q2): is the 50th percentile data

(Upper Quartile) quartile-3 (q3) is the 75th percentile data.



q1 = least 25% of data.

q2 = least 50% of data

q3 = least 75% of data / Highest 25% of data

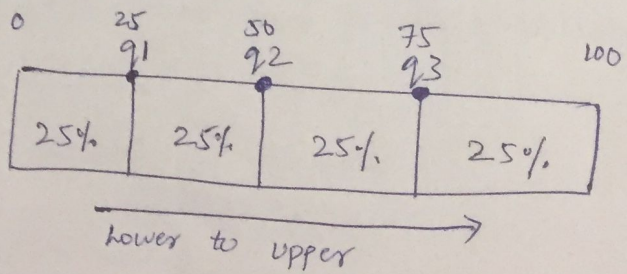
~~q4 = Highest 25% of data~~

q1 is the median of lower half of data set.

q2 is the median of whole data set

q3 is the median of upper half of data set.

Inter quartile: Upper quartile - Lower quartile.



Inter quartile = $q_3 - q_1$

nents according into

Percentiles.

ntile data.

entile data

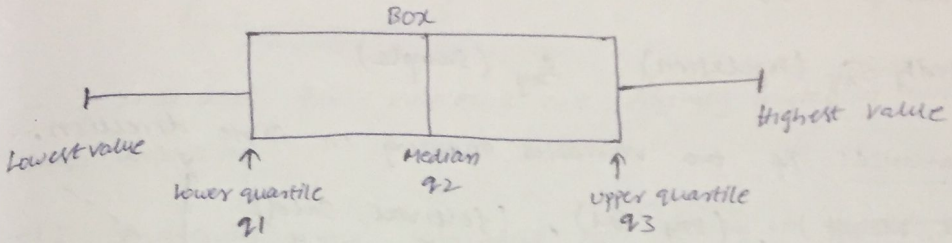
centile data.

→ 3
1

1

3

Box and Whisker Plots: Quartiles are basically represented graphically using box & whisker plots.



Find the quartile in excel sheet using quartile(A1:A10:1)

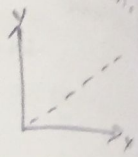
Covariance & Correlation coefficient

Co variance: is the measure of relationship between two random variables. i.e measure of Linearity b/w variables.

Denoted by σ_{xy} (population) S_{xy} (sample)

Positive Covariance: If two variables moving in same direction.

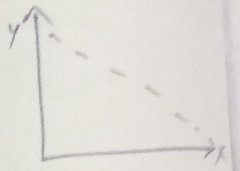
Ex: (height, weight), (Exp, Sal), (festival, Sales)
x y x y x y



Negative Covariance: variables moving in opposite direction.

Ex: (no of claims of Insurance company, profits)

(Age, hair color), (Age, Fitness)



Zero Covariance: If two variables are independent of each other (or) there is no relation b/w variables then zero covariance.

Ex: (Petrol Price, water bottle Price)

- **Dis Advantage of Covariance:** Interpretation of covariance is difficult.

Ex: +ve covariance can be (0.005, 5, 10, 33,44,555)

With the above values we wouldn't be able to interpret.