STATISTICS

BY HIMANSHU SHUKLA (STATISTICAL EXPERT)

Sir R. A. Fisher

"The science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data."

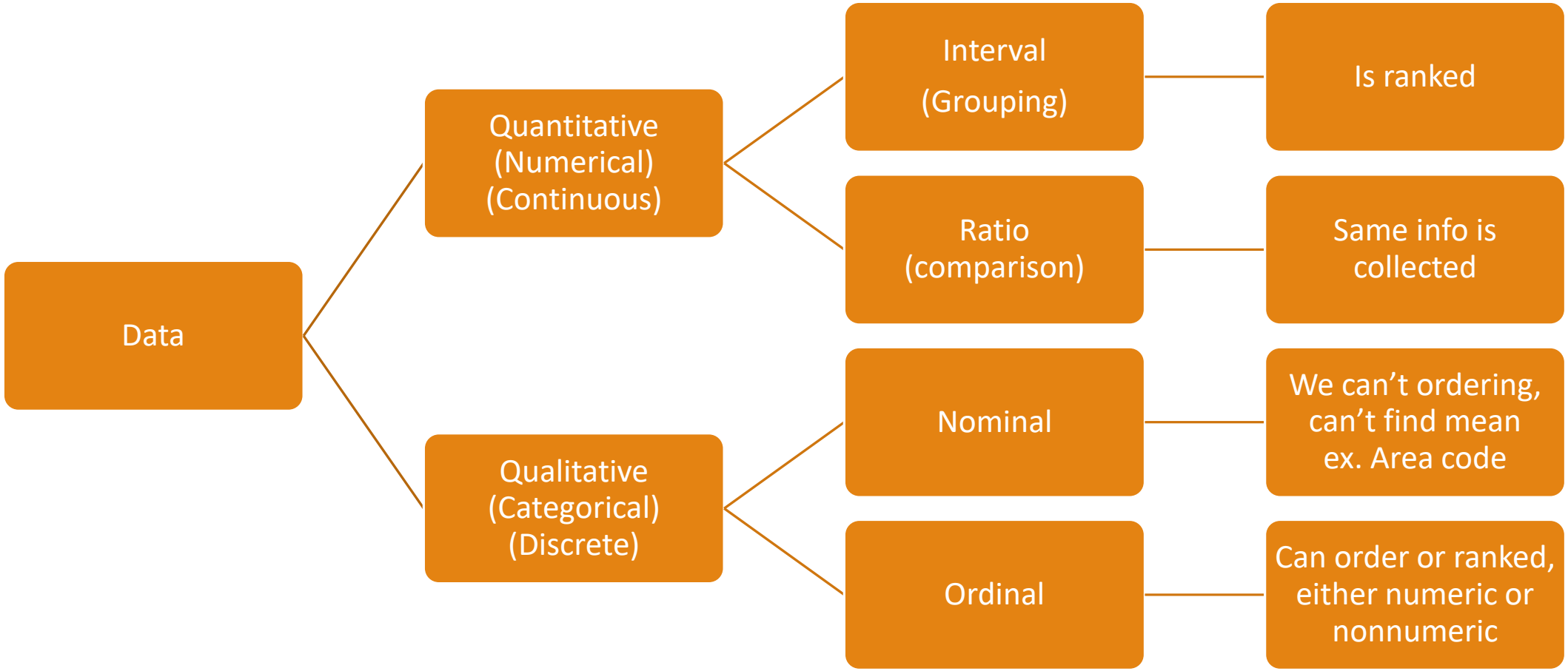
Main Division of Statistics->

Mathematical or Theoretical statistics	• Statistical Distributions, experimental designs, sampling designs, etc.
Statistical methods or functions	• Collection, tabulation, analysis and interpretation of data, etc.
Descriptive statistics	• Classification and diagrammatic representation of data.
Inferential Statistics	• To draw conclusion about population on the basis of sample drawn from it. • In this we check hypothesis.
Applied Statistics	• It mainly covers population, census, national income, production, business statistics, industrial statistics, quality control, biostatistics, etc.

Some times We say statistics “Bekar hai”

- Statistics is true on an average only.
- Statistics deals with the masses, not an individual. No statistics is applicable for a single observation.
- Statistical results are correct in a general sense. They are always subject to certain amount of error.
- Statistics is only a means to draw conclusions about masses or population but nota panacea to all sort of problems.
- Statistics deals with quantitative data only. Even qualitative information is converted into numerical data by the method of ranking, scoring or scaling.

DATA : Primary and Secondary

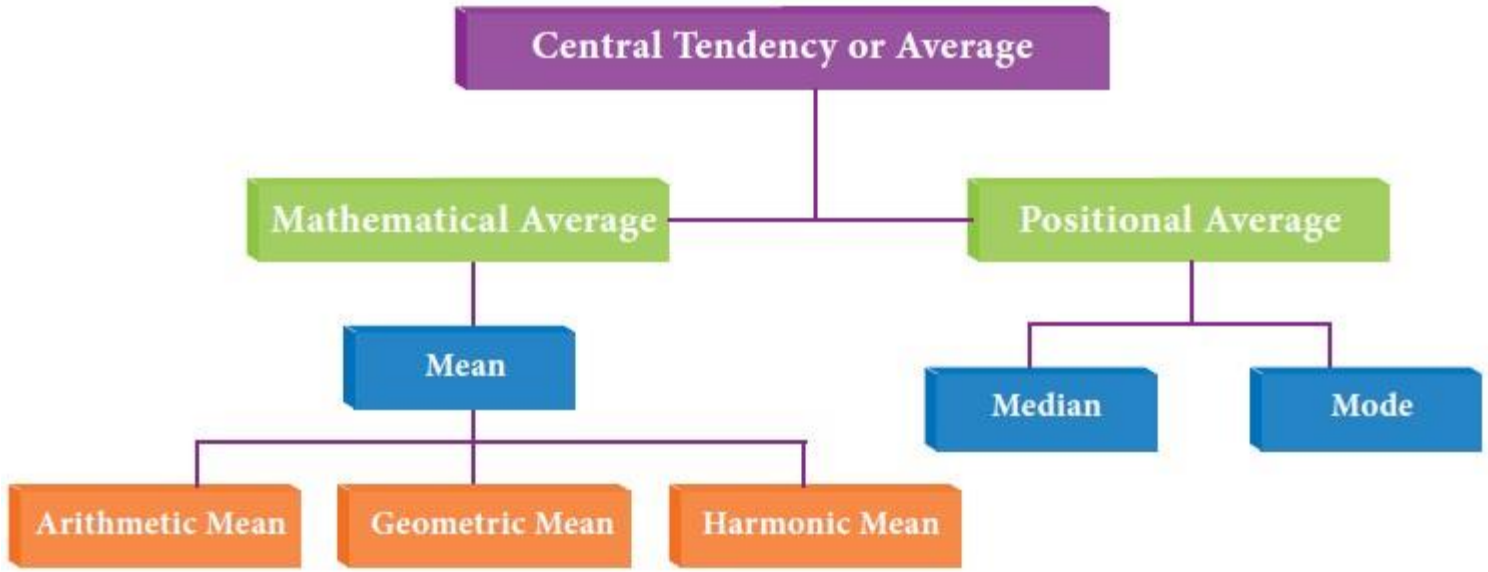


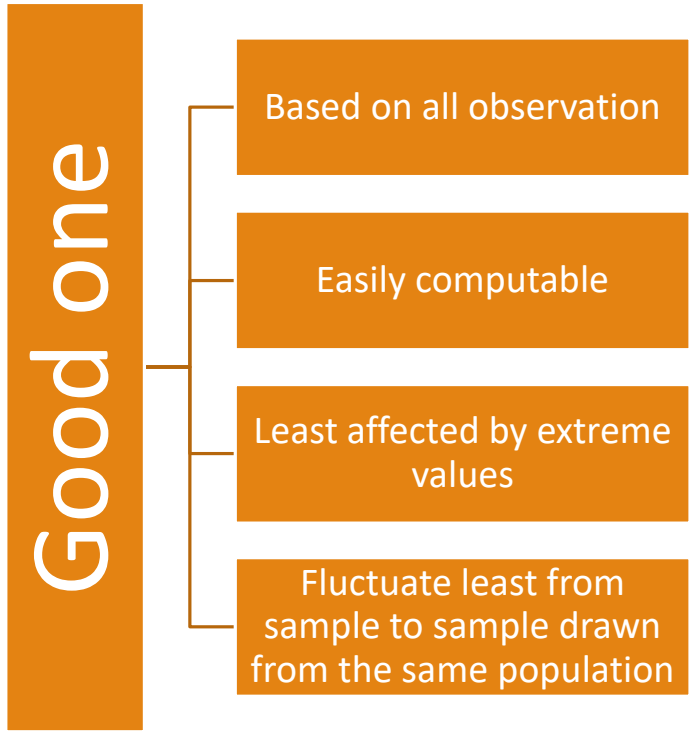
Descriptive Statistics

- Used to describe the basic feature of the data in a study.
- Provide simple summaries about the sample and the measures.
- What the data shows .
  - Central tendency
  - Variance
  - Skewness
  - kurtosis.

Measure of Central Tendency

It is a single value with the range of data which represents a group of individual value in a simple and concise manner. So that the mind can get quick understand. Since the value lies within the range of data, it is known as a measure of central tendency.





Problem in Average

- It represent a group of whole not an individual.
- Often an average is a value which does not exist in the set of data.
- Sometimes an average gives a value which is not feasible eg. Average size of a family is 3.62.

$$Arithmatic\ Mean\ (A.M.) = \frac{a+b}{2}$$

$$Geometric\ Mean\ (G.M.) = \sqrt{ab}$$

$$Harmonic\ Mean\ (A.M.) = \frac{2ab}{a+b}$$

- *Note: we generally refer A.M. for average. Because G.M. have boundation of data in data zero and –ve values should not be percent otherwise we can not calculate G.M. while G.M. is least effected by extreme values. It is basically used in case of Index numbers problems. If 0 is present in data we cant calculate H.M. also H.M. is used when the values are pertaining to the rate of change per unit time such as speed, number of items produced per day etc.*

Relation between

A.M., G.M., H.M.

$$G^2 = A\,H$$

Mean

MEAN($\overline{x}$)

FORMULA:

Ungrouped Data :

$$\overline{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

Group Data:

$$\overline{x} = \frac{\sum fx}{n}$$

where: f = **frequency in each class**

x = **midpoint of each class**

n = **total number of scores**

THE ARITHMETIC MEAN OF GROUPED DATA - EXAMPLE

Selling Price ($ thousands)	Frequency (f)	Midpoint (M)	fM
15 up to 18	8	$16.5	$ 132.0
18 up to 21	23	19.5	448.5
21 up to 24	17	22.5	382.5
24 up to 27	18	25.5	459.0
27 up to 30	8	28.5	228.0
30 up to 33	4	31.5	126.0
33 up to 36	2	34.5	69.0
Total	80		$1,845.0

Solving for the arithmetic mean using formula (3–12), we get:

$$\overline{X} = \frac{\Sigma fM}{n} = \frac{\$1,845}{80} = \$23.1 \text{ (thousands)}$$

Mean in python

```
In [11]: test_scores = [82,93,96,78,64,108,99,54,88,91,89,98]

In [12]: print(len(test_scores))
         print(sum(test_scores))

12
1024

In [14]: count = 12
         sum_scores = 1024
         mean = sum_scores / count
         print(mean)

85.33333333333333
```

Use the NumPy `mean()` method to find the average speed:

```
import numpy

speed =
[99,86,87,88,111,86,103,87,94,78,77,85,86]

x = numpy.mean(speed)

print(x)
```

Demerits of Mean:

- **Too much affected by extreme values.**
- **mostly it does not correspond to any value of the set of observations.**
- **It does not convey any info. About the spread or trend of data.**
- **It is not suitable measure of central values in case of highly skewed distribution.**

Median

Median is the value of the variable that divides the ordered set of values into two equal halves.

Median

First, arrange the observations in an ascending order.

If the number of observations (n) is odd: the median is the value at position

$$\left(\frac{n+1}{2}\right)$$

If the number of observations (n) is even:

1. Find the value at position $\left(\frac{n}{2}\right)$
2. Find the value at position $\left(\frac{n+1}{2}\right)$
3. Find the average of the two values to get the median.

1, 3, 3, **6**, 7, 8, 9

Median = 6

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

Median = (4 + 5) ÷ 2

= 4.5

Weight in grams	Number of apples	Cumulative Frequency
410 – 420	14	14
420 – 430	20	34
430 – 440	42	76
440 – 450	54	130
450 – 460	45	175
460 – 470	18	193
470 – 480	7	200
Total	**N = 200**	

$$\frac{N}{2} = \frac{200}{2} = 100.$$

Median class is 440 – 450

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$l = 440,\ \frac{N}{2} = 100,\ m = 76,\ f = 54,\ c = 10$

$$\text{Median} = 440 + \frac{100 - 76}{54} \times 10$$

$$= 440 + \frac{24}{54} \times 10 = 440 + 4.44 = 444.44$$

The median weight of the apple is 444.44 grams

Median in Python

- import numpy
- x = [99,86,87,88,111,86,103,87,94,78,77,85,86]
- sorted_x=sorted(x)
- sorted_x
- [77, 78, 85, 86, 86, 86, 87, 87, 88, 94, 99, 103, 111]
- m=(len(sorted_x)+1)/2
- median= sorted_x[int(m)]
- median
- 87

```
import numpy
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
x = numpy.median(speed)
print(x)
```

Advantage:

- It is not influenced by extreme values. But it is position average.
- For data from skewed distributions, the median is better than the mean because it isn't influenced by extremely large values.

Disadvantage:

- Data should be arranged.
- A slight change in the series may bring drastic change in median value.
- It is not suitable for further mathematical treatment except its use in mean deviation.

Mode

Drawback: It cant use mathematically in Quantitative data

- Maximum frequency in a distribution.
- Some distributions have equal peaks and hence ode is not necessarily unique. There will be as many modes of distribution as the number of peaks in it.
- The mode is the least used of the measures of central tendency
- The mode will be the best measure of central tendency (as it is the only one appropriate to use) when dealing with nominal data.

The Mode value is the value that appears the most number of times:
99, 86, 87, 88, 111, 86, 103, 87, 94, 78, 77, 85, 86 = 86

The SciPy module has a method for this.

Example
Use the SciPy mode() method to find the number that appears the most:

```
from scipy import stats
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
x = stats.mode(speed)
print(x)
```

The mode() method returns a Mode Result object that contains the mode number (86), and count (how many times the mode number appeared (3)).

With a single-mode sample. In python 3.8 multimode() function is present in Statistics

Advantages of the mode

1. mode is easy to determine
2. It is easy to understand
3. mode is not affected by extremes of values
4. When data are not complete, mode cannot be difficult to estimate
5. mode is very easy to compute

Disadvantages of mode accuracy

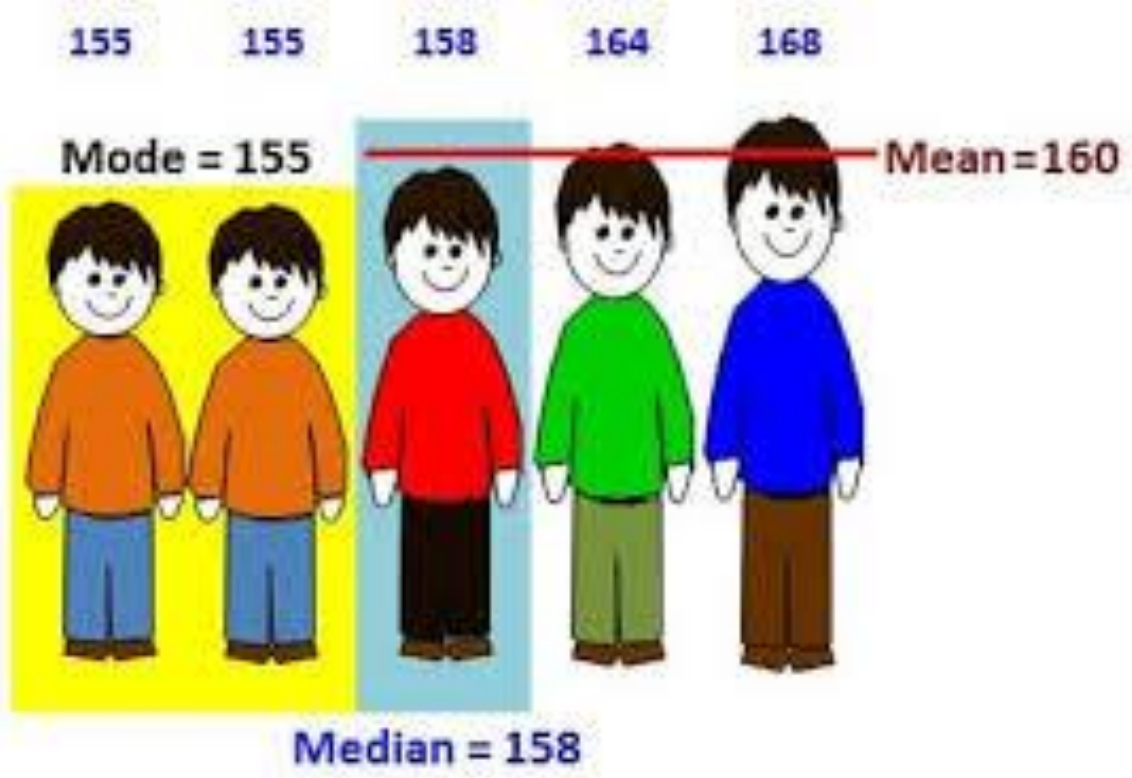
1. It is not a very good measure of accuracy
2. It is relevant in further statistical calculation
3. It represents a very poor average
4. It is difficult to calculate, especially when more than one mode or large numbers are involved.
5. There may be uncertainty in the exact location
6. Arrangement of data is always tedious

In Machine Learning (and in mathematics) there are often three values that interests us:

Mean - The average value

Median - The mid point value

Mode - The most common value

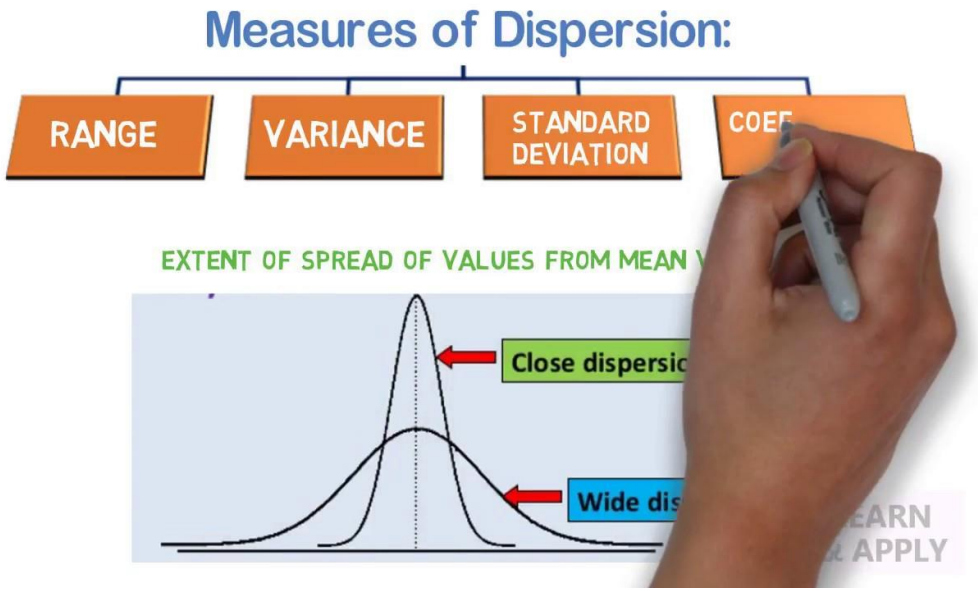


Measures of Dispersion

Measure of dispersion help to gauge the spread of data about a central value. These measures help to determine how stretched or squeezed the given data is.

There are many Measures of Dispersion found that help us to get more insights into the data:

1. **Range**
2. **IQR and Quartile Deviation**
3. **Variance**
4. **Standard Deviation**
5. **Mean deviation**
6. **Skewness**
7. **Kurtosis**
8. **IQR**



1. Range:

Range is the measure of the difference between the largest and smallest value of the data variability. The range is the simplest form of Measures of Dispersion.
Example: 1,2,3,4,5,6,7
•Range = Highest value – Lowest value = (7 – 1) = 6

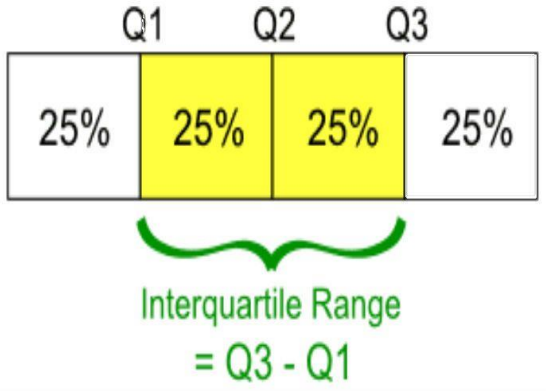
```
import numpy

speed = [2.74, 1.23, 2.63, 2.22, 3, 1.98]

Range= max(speed)- min(speed)
```

2. Quartile Deviation and IQR: Quartile deviation is the half of the difference of third quartile (Q3) and first quartile (Q1) i.e. half of the interquartile range (IQR). **(Q3 – Q1) / 2 = IQR / 2**

Decision making

The data set having higher value of quartile deviation has higher variability.

```
# Import numpy library
import numpy as np
data = [32, 36, 46, 47, 56, 69, 75, 79, 79, 88, 89, 91, 92, 93, 96, 97, 101, 105, 112, 116]
# First quartile (Q1)
Q1 = np.percentile(data, 25, interpolation = 'midpoint')
# Third quartile (Q3)
Q3 = np.percentile(data, 75, interpolation = 'midpoint')
# Interquaritle range (IQR)
IQR = Q3 - Q1
# Quartile Deviation
qd = IQR / 2
print(qd)
```




```
from scipy import stats
data = [32, 36, 46, 47, 56, 69, 75, 79, 79, 88, 89, 91, 92, 93, 96, 97,101, 105, 112, 116]
# Interquartile range (IQR)
IQR = stats.iqr(data, interpolation = 'midpoint')
print(IQR)
```

3. Variance (σ2): The average squared deviation from the mean of the given data set is known as the variance.

- It can be calculated by obtaining the sum of the squared distance of each term in the distribution from the Mean, and then dividing this by the total number of the terms in the distribution.
- This measure of dispersion checks the spread of the data about the mean.
- It basically shows how far a number, for example, a student's mark in an exam, is from the Mean of the entire class.
- **Formula:** $(\sigma 2) = \sum (X - \mu)2 / N$

- A low value for variance indicates that the data are clustered together and are not spread apart widely, whereas a high value would indicate that the data in the given set are much more spread apart from the average value.
- Variance will be always positive value.
- Variance is sensitive to extreme values.

Variance in Python

```
# variance() function of Statistics Module

# Importing Statistics module
import statistics

# Creating a sample of data
sample = [2.74, 1.23, 2.63, 2.22, 3, 1.98]

# Prints variance of the sample set

# Function will automatically calculate
# it's mean and set it as xbar
print("Variance of sample set is % s"%(statistics.variance(sample)))
```

Note: What is %s and %D Python?
%s is used as a placeholder for string values you want to inject into a formatted string. %d is used as a placeholder for numeric or decimal values. For example (for python 3) print ('%s is %d years old' % ('Joe', 42))

4. Standard Deviation: Standard Deviation can be represented as the square root of Variance. To find the standard deviation of any data, you need to find the variance first. Standard Deviation is considered the best measure of dispersion.

Formula:

Standard Deviation = √σ

```
# Python code to demonstrate stdev() function

# importing Statistics module
import statistics

# creating a simple data - set
sample = [1, 2, 3, 4, 5]

# Prints standard deviation
# xbar is set to default value of 1
print("Standard Deviation of sample is % s "
                % (statistics.stdev(sample)))
```

5. Mean Deviation

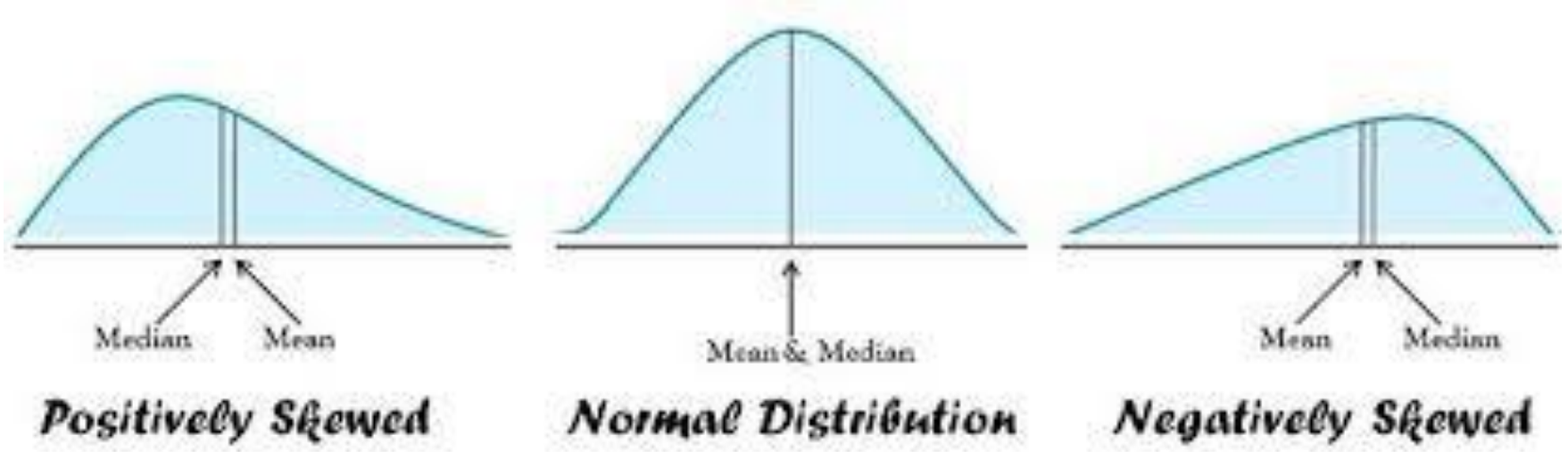
It is defined as the average of absolute deviations taken from an average, usually the median.

$$M.A.D(M) = \frac{\sum_{i=1}^{n} f_i |x_i - M|}{N}$$

6. Skewness

Skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

$$\textbf{Skewness} = \frac{3\,(\textit{Mean} - \textit{Median})}{\textit{Standard Deviation}}$$



Skewness in python

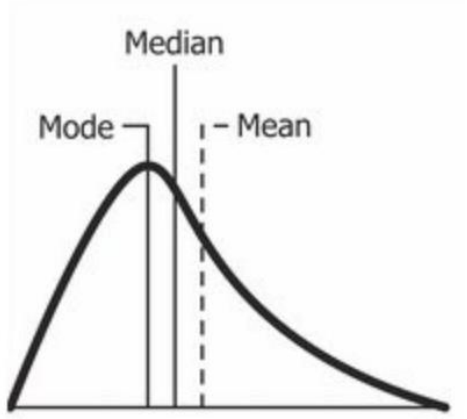
function calculates the skewness of the data set.

scipy.stats.skew(array, axis=0)

skewness along the index axis

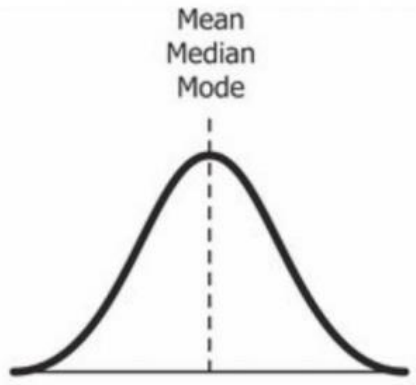
df.skew(axis = 0, skipna = True)

skewness < 0 : more weight in the right tail of the distribution i.e positively skewed



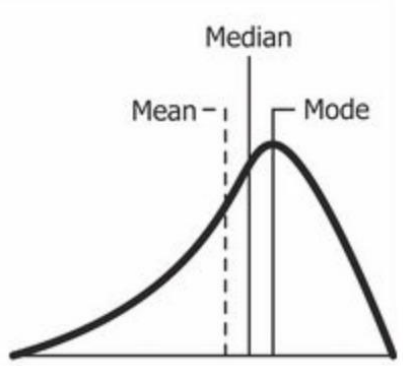
mode<median<mean.

skewness = 0 : normally distributed.



mode=median=mean

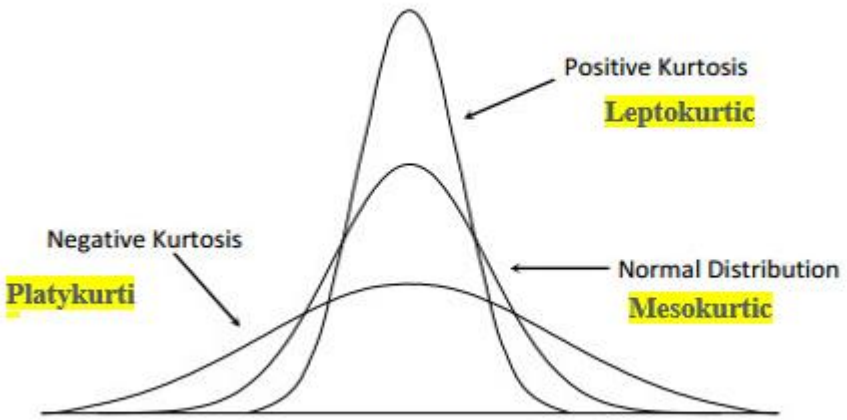
skewness > 0 : more weight in the left tail of the distribution i.e negatively skewed



mode>median>mean

Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.
Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

```
from scipy.stats import kurtosis
sample = [1,2,3,4,5]
kurtosis(sample)
```

Leptokurtic (kurtosis > 3) means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.

Mesokurtic (kurtosis = 3) Distribution is normal distribution.

platykurtic (kurtosis < 3) A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

The end