Class 3
Machine Learning With Python

Measures Of Central Tendency

The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution.

In statistics, a **central tendency** (or **measure of central tendency**) is a central or typical value for a probability distribution.

It may also be called a **center** or **location** of the distribution. Measures of central tendency are often called ***averages***.

Measures Of **Central Tendency**

Methods Of study/ Calculation of the central value in a given mathematical series.

A single Value that represent the value or whole series.

Question:- why should we assume or go with a *Value* when we have whole data.?
Ans:-
1) Easy to understand.
2) Easy to compare.
3) Data series.
4) Universally accepted.
5) Easy to calculate.

- Suppose if we know swift car gives 15 km/l ->*AVERAGE* , then we can get to know how far it can cover in given fuel.
- If we know Swift average is 25km/l and Honda city average is 20km/l, so comparison would be easy

It is further divided into two *categories:-*

*1.**Arithmetic Calculation**.*

- **Arithmetic Mean (AM).**
- **Geometric Mean (GM).**
- **Harmonic Mean (HM).**

*2.**Positional Averages**.*

- **Mode**
- **Median**

GROUPED AND UNGROUPED DATA

Ungrouped data which is also known as raw data is data that has not been placed in any group or category after collection. Data is categorized in numbers or characteristics therefore, the data which has not been put in any of the categories is ungrouped.

For example:- when conducting census and you want to analyze how many women above the age of 45 are in a particular area, you first need to know how many people reside in that area.

The number of individuals residing in that area is ungrouped data or raw information because nothing has been categorized. We can therefore conclude that ungrouped data is data used to show information on an individual member of a sample or population.

Some of the advantages of ungrouped data are as follows;

Most people can easily interpret it.

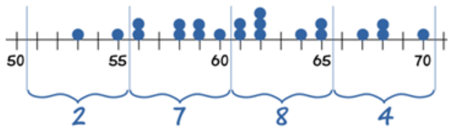
When the sample size is small, it is easy to calculate the mean, mode and median.

It does not require technical expertise to analyze it.

Grouped data is the type of data which is classified into groups after collection. The raw data is categorized into various groups and a table is created. The primary purpose of the table is to show the data points occurring in each group. For instance, when a test is done, the results are the data in this scenario and there are many ways to group this data.

For example:- How many runners take how many seconds to complete the track.

SECONDS	**FREQUENCY**
51-55	2
56-60	7
61-65	8
66-70	4



This type of table called as **FREQUENCY DISTRIBUTION TABLE**

Non-overlapping Classes / Discrete:

If the values of a variable in a collection of data are positive integers less than or equal to 50, we can group the data in five non-overlapping intervals: 1 – 10, 11 – 20, 21 – 30, 31 – 40, 41 – 50. These are the groups of values of the variable. In case of non-overlapping intervals, 1 – 10 is the group containing the values of the variable that are greater than or equal to 1 but less than or equal to10. Similarly, 11 – 20 is the group containing the values of the variable that are greater than or equal to 11 but less than or equal to 20.

Overlapping Classes / Continuous:

If the values of a variable in a collection of data are positive integers less than 50, we can group the data in five overlapping class intervals: 0 – 10, 10 – 20, 20 – 30, 30 – 40, 40 – 50. These are the groups of values of the variable. In case of overlapping intervals, 0 – 10 is the group containing the values of the variable that are greater than or equal to 0 but less than 10. Similarly, 10 – 20 is the group containing the values of the variable that are greater than or equal to 10 but less than 20.

When Intervals overlap then we always put the data in the **lower bound**

Measures Of Central Tendency

ARITHMETIC MEAN (AM)

Represented as :-

Population - μ

Sample - $\bar{x}$ **“x-bar”**

$\bar{x}$ = ∑ (Sum of observations / Number of observation

$$(\sum x_i)/n$$

For Example:-

Ques:- Weights of 6 boys in a group are 63, 57, 39, 41, 45, 45. Find the mean weight.

Solution:-

Number of observations = 6

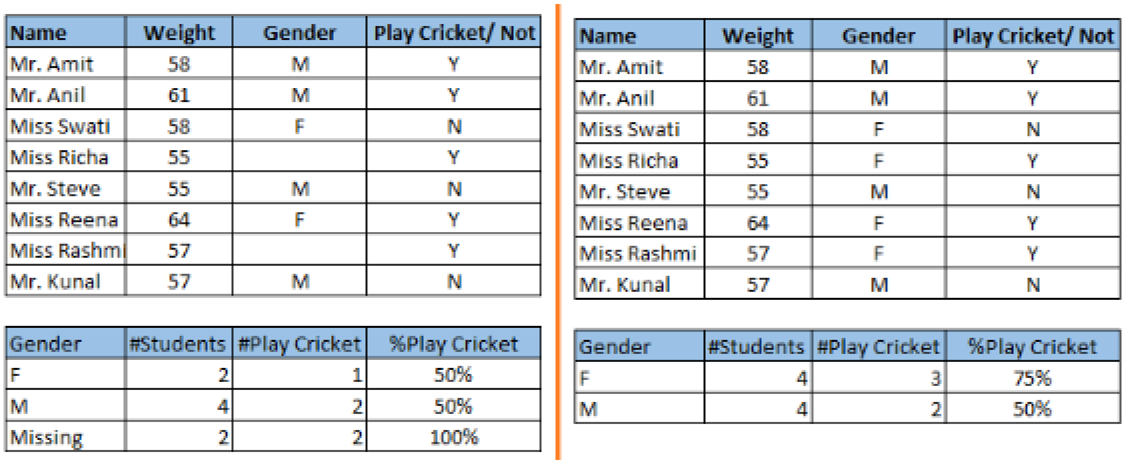
Sum of all the observations = 63 + 57 + 39 + 41 + 45 + 45 = 290

Therefore, arithmetic mean = 290/6 = 48.3

*****IMPORTANT QUESTION*****

Ques:- What is the Difference between *MEAN / AVERAGE .?*
Ans :- All **Averages** are **Mean** , however all **Mean** are not **Averages.**

Ques:- How does Missing values impact mean and why "Missing Value Treatment is required.?
Ans:- Missing data in the data set can reduce the fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.



Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashm	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Notice the missing values in the image shown above: In the left scenario, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

AVERAGE OF AVERAGES

What is Average.?

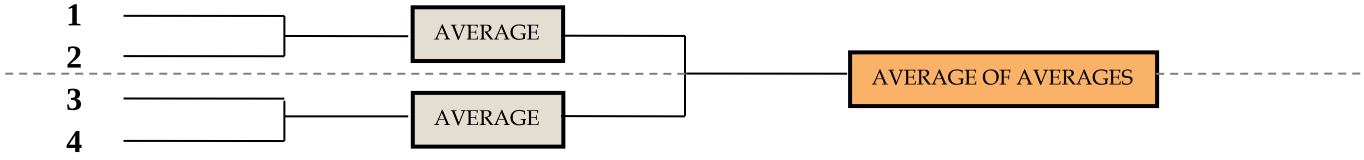
An **Average** is a single number taken as representative of a list of numbers. Different concepts of average are used in different contexts. Often "average" refers to the arithmetic mean, the sum of the numbers divided by how many numbers are being averaged.

What is Average Of Averages .?

Average Of Averages are nothing but calculating the Average of two or more Averages.

Example:-

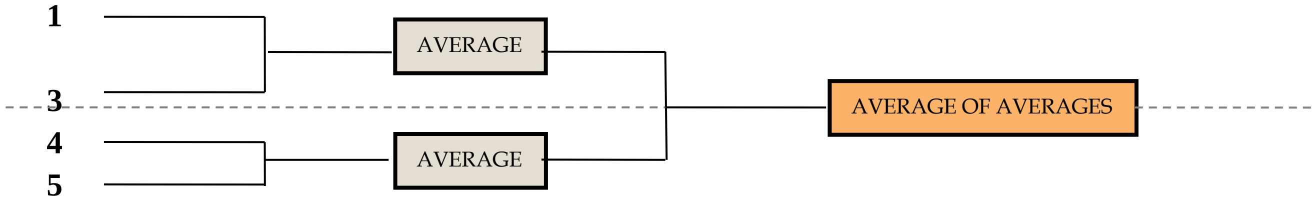
Population



In above image, Average Of Average should lie in the exact middle of sample 2 and 3 because all samples have equal size,

What if samples are not in equal size.?

Population



As we can see in Above image, Average of Averages lie somewhere nearby sample 3, as sample do not have equal size.

- Average of Averages is riskier business when sample are not in equal size and this phenomenon is called **Simpson's Paradox**.

Mean of Grouped Data

An estimate, x̅, of the mean of the population from which the data are drawn can be calculated from the grouped data as:

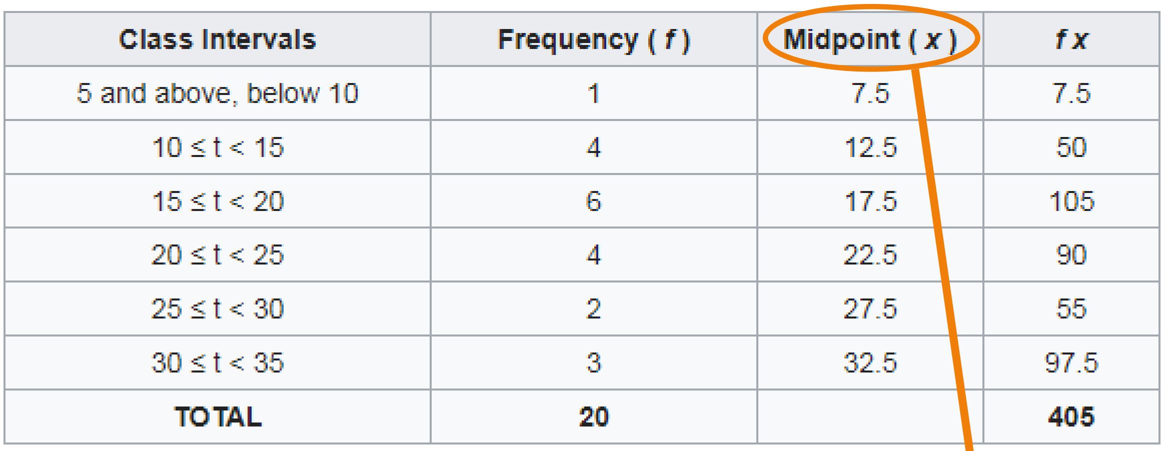
$$\boxed{\bar{x} = \frac{\sum f x}{\sum f}.}$$

Note:-

x refers to the midpoint of the class intervals,

f is the class frequency.

fx = (**f**) x (**x**)



Class Intervals	Frequency (*f*)	Midpoint (*x*)	*f x*
5 and above, below 10	1	7.5	7.5
10 ≤ t < 15	4	12.5	50
15 ≤ t < 20	6	17.5	105
20 ≤ t < 25	4	22.5	90
25 ≤ t < 30	2	27.5	55
30 ≤ t < 35	3	32.5	97.5
TOTAL	**20**		**405**

Thus, the mean of the grouped data is:-

$\overline{x} = \Sigma fx/ \Sigma f$

$= 405 / 20$

$= 20.25$

Mid-point are also called **Class Marks**

This method is called **Grouped Direct Method**

Some Important terms

Closed Intervals**:-**When Lower bound and Higher bound is defined such as:-

Class Intervals
0-10
10-20
20-30

Open intervals**:-***When Lower bound and Higher bound us not-defined such as:-*

Class Intervals
low-10
10-20
20-High

ASSUMED MEAN METHOD

Sometimes when the numerical values of xi and fi are large, finding the product of xi and fi becomes tedious and time consuming. So, for such situations, let us think of a method of reducing these calculations.

Nothing can be done with the fi's, but each xi can be changed to a smaller number to make easier calculations.

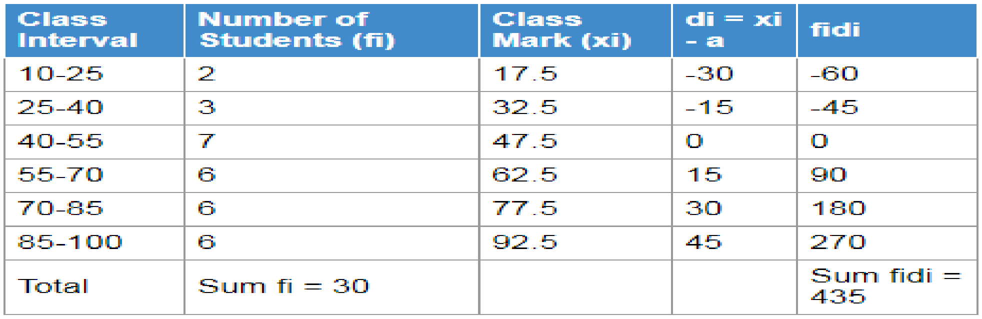
The first step is to choose one among the xi's as the **assumed mean**, and denote it by '**a**'. Also, to further reduce our calculation work.

Let us choose a = 47.5.

The next step is to find the difference di between a and each of the xi's, that is, the deviation of 'a' from each of the xi's.

i.e., di = xi – a = xi – 47.5

The third step is to find the product of di with the corresponding fi, and take the sum of all the fi di's.



Class Interval	Number of Students (fi)	Class Mark (xi)	di = xi - a	fidi
10-25	2	17.5	-30	-60
25-40	3	32.5	-15	-45
40-55	7	47.5	0	0
55-70	6	62.5	15	90
70-85	6	77.5	30	180
85-100	6	92.5	45	270
Total	Sum fi = 30			Sum fidi = 435

So, the Mean of Deviation :-

$$\overline{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{\sum_{i=1}^{n} f_i}$$

x̄ = 47.5 + (435/30) = **62**