Presentation

On

Descriptive statistics & its application

Name :- Arindam Karmakar

Roll :- 10871022002

Subject:-Basic Data Science

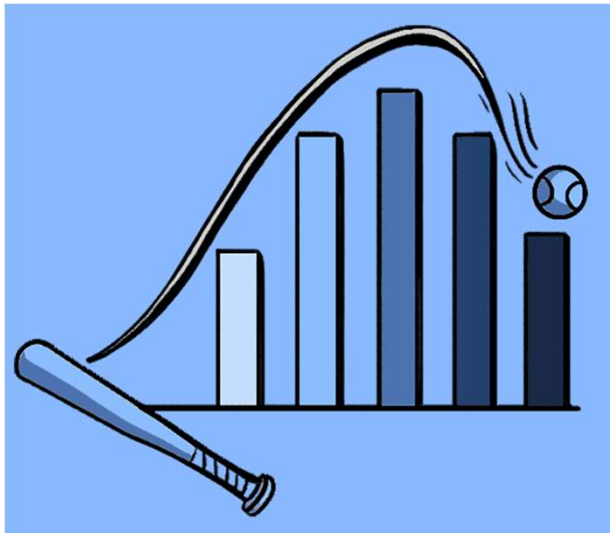
Paper Code :- MCAN-E304F

Dept : MCA

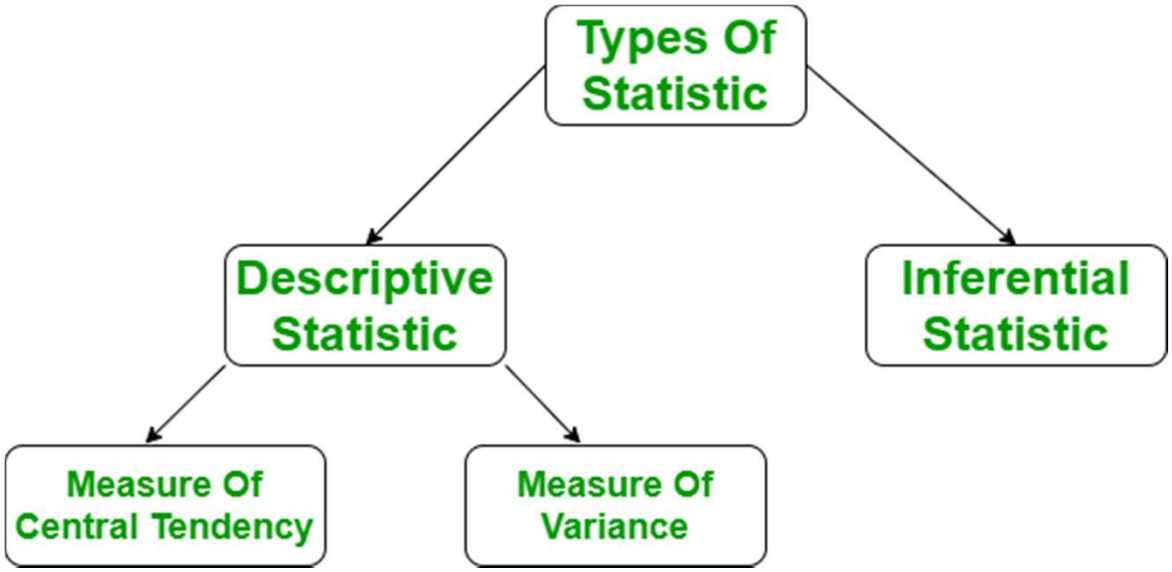
College : Asansol Enginnering College

Descriptive statistics

Descriptive statistics refers to a set of methods used to summarize and describe the main features of a dataset, such as its central tendency, variability, and distribution. These methods provide an overview of the data and help identify patterns and relationships.

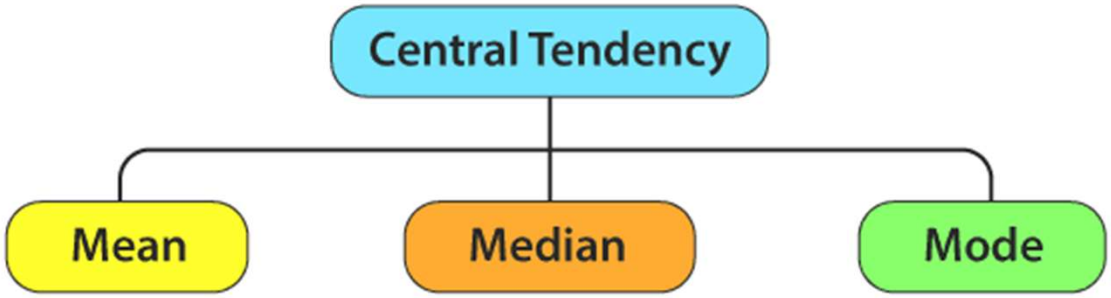


Types of Descriptive Statistics



Measures of Central Tendency

It represents the whole set of data by a single value. It gives us the location of the central points. There are three main measures of central tendency



Here we take a dataset and try to do the operations

	Data Set	Page	Of Cases	Of Variable	Text Format
0	Milk Production	4	199	7	P004.txt
1	Right-To-Work Law	5	38	7	P005.txt
2	Domestic Immigration	7 8	48	6	P007-8.txt
3	Egyptian Skulls	9	150	5	P009.txt
4	New York Rivers	10	20	5	P010.txt
5	Industrial Production	11	59	8	P011.txt
6	Space Shuttle Challenger	12	23	2	P012.txt
7	Cost Of health Care	14	52	8	P014.txt

Mean : It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

$$\text{mean} = \sum xi/n$$

Where,
X= Observations
N= number of terms

If we calculate the mean of the column " Of Cases" from our data set, Then,

X=199+38+48+150+20+59+23+52
N=8

Mean=589/8
=73.625

Median : It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is the median and if it is even then the median would be the average of two central elements.

To find the median:

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

If we calculate the median of the column " Of Cases" from our data set, Then,

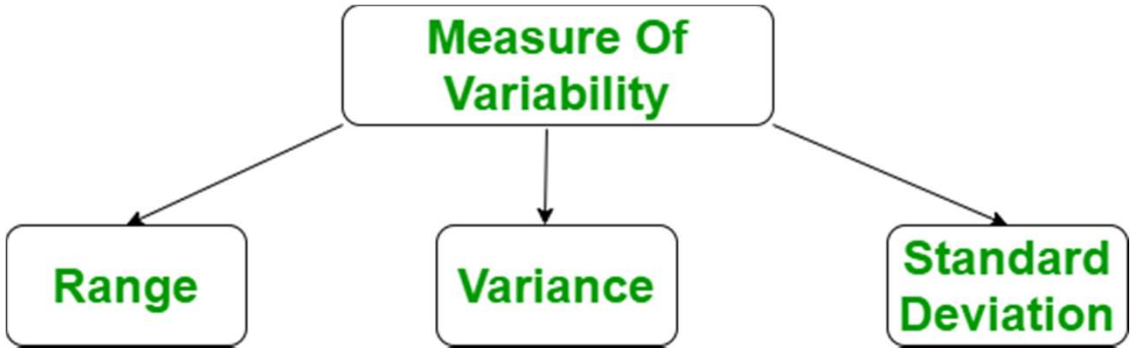
Median = 50

Mode : It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

Mode = ModeResult(mode=20, count=1)

Measure of Dispersion

Measures of variability are also termed measures of dispersion as it helps to gain insights about the dispersion or the spread of the observations at hand. Some of the measures which are used to calculate the measures of dispersion in the observations of the variables are as follows:



Range : The range describes the difference between the largest and smallest data point in our data set. The bigger the range, the more the spread of data and vice versa.

Range = Largest Data Value – Smallest Data Value

If we find the Range of the column "Of cases" from our previous data set,
Then we get :

Range = 199 – 20
=179

Variance : It is defined as an average squared deviation from the mean. It is calculated by finding the difference between every data point and the average which is also known as the mean, squaring them, adding all of them, and then dividing by the number of data points present in our data s

Variance : 4228

Standard Deviation : It is defined as the square root of the variance. It is calculated by finding the Mean, then subtracting each number from the Mean which is also known as the average, and squaring the result. Adding all the values and then dividing by the no of terms followed by the square root.

Standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

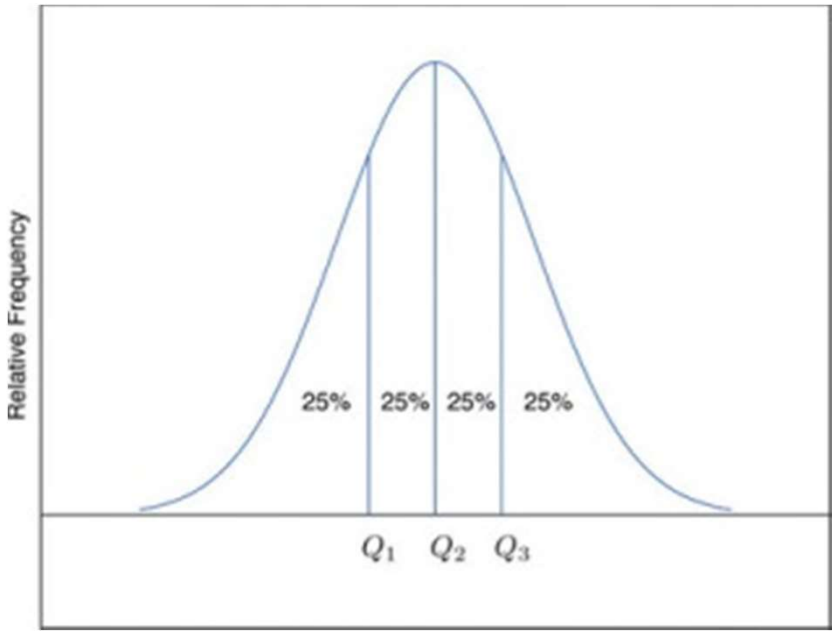
x = Observation under consideration
N = number of terms
mu = Mean

If we find the Standard Deviation of the column “Of cases” from our previous data set,
Then we get :

Standard Deviation = 60.82544183974334

Relative position

The percentile rank and z-score of a measurement indicate its relative position with regard to the other measurements in a data set. The three quartiles divide a data set into fourths. The five-number summary and its associated box plot summarize the location and distribution of the data.



References

- https://www.geeksforgeeks.org/descriptive-statistic/
- https://www1.aucegypt.edu/faculty/hadi/RABE5/
- https://byjus.com/maths/variance-and-standard-deviation/


Thank
You!