What is Data Science? 🡺

Exploring data by doing some analysis on it is called data science

Data🡺

Data is of 3 types

1. Structured
2. Semi-structured
3. Unstructured

Structured Data🡺

1. Nominal data 🡺categorical data/qualitative data/non-parametric data
2. Ordinal data 🡺 categorical data/qualitative data/non-parametric data
3. Interval data🡺numeric data/quantitative data/parametric data
4. Ratio level data 🡺 numeric data/quantitative data/parametric data

Nominal data🡺

Classifying the data based on properties or characteristics is called nominal data. The data is categorized.

Ordinal Data 🡺

Survey data

Interval data🡺

Continues data.Ex: temperature at different time

Ratio level data 🡺

Discrete data/whole number data

Data🡺statistics🡺insight

Statistics🡺

Statistics is classified into 3 categories.

1. Descriptive (structured/semi-structured data)
2. Predictive modelling (structured/semi-structured data)
3. Machine learning (unstructured data)

The role of statistics in data science is classified as below 🡺

1. Classification
2. Collection
3. Organization
4. Summarization🡺 can be done in two ways(Descriptive statistics) or (inferential statistics🡺sample data🡺do analysis🡺apply result to population data)
5. Interpreting the results

Descriptive statistics 🡺

Describe the population.

1. Measure of center tendency

* Mean
* Median
* mode

1. Measure of dispersion or variation

* Range
* Variance
* Standard deviation

Measure of center tendency

Mean: mean is average

1,4,5,6,7

Mean is : 1+4+5+6+7=23/5=4.6

Median :

Data should be sorted in ascending or descending order.

Median of odd number = (n+1)/2 th position

10,5,1,3,28,6,9,13

Sorting : 1,3,5,6,9,10,13,28

Median : ~~1,3,5~~,6,9,~~10,13,28~~ 🡺 if two number then it would be 6+9/2=7.5

Mode: The number which is repeating more number of times.

1,1,3,1,1,5,1,6 🡺 1 I uni mode

1,2,3,4,5 🡺 no mode because none of the numbers are repeating

1,2,2,3,3,4,5🡺 bimode 2 & 3 (if two numbers are repeating equal number of times)

1,1,1,2,3,3,3,4,5,6,6,6🡺 tri mode 1,3,6 (if three numbers are repeating equal number of times)

Disadvantage of mean.median,mode : We cant find min,max value.

Measure of dispersion or variation

Range🡺 max value-min\_value

1,3,5,7,9 🡺 9-1=8

Variance🡺 Each individual element variating from mean value.

1,3,5,7,9

Mean=8

Variance = squre of (1-8)+ squre of (3-8)+ squre of (5-8)+ squre of (7-8)+ squre of (9-8) =49+25+9+1+1=85

Standard deviation: SD = sqrt (variance)

Group and frequency🡺

1,3,5,7,9,11,13,15,20,18,17

|  |  |
| --- | --- |
| Group | frequency |
| 1-5 | 3 |
| 6-10 | 2 |
| 11-15 | 3 |
| 16-20 | 3 |

1,6,11,16 are called lower class limits

5,10,15,20 are called upper class limits

Avarage of lower and upper class limit is called as class mid point

Difference between two consecutive lower limits or two consecutive upper limit is called as class width.

The class width should be equal in all the group.

Lower class limit and upper class can not be same.

Class boundary🡺

Relative frequency🡺

Cumulative frequency🡺

Tabular descriptive status🡺