Statistics or Stats

It is a branch of mathematics that deals with collection of data,Organizing/presentation of data, Analysis of the data to extract some meaningful informations(Insights). ----> To make a proper decision

---

It is divided into two parts:

1. Descripitve Stats (Mathematical operations) It is a branch of stats that involves collection and interpretion of data.

-> Measure of Central Tendency: It is used to indicate where does the center of my data lies

- Mean
- Median
- Mode

-> Measure of dispersion: It is used to indicate how much of my data is spread out/ Dispersed in all direction.

- Range
- Variance
- Standard deviation
- Percentile
- Quartile
- Inter quartile range(IQR)
- Outliers
- Correlation
- Skewness(inclination)
- Kurtosis

Inferential Stats

Descripitve Stats

Measure of Centeral Tendencies

- Mean : Average of data

```
import statistics as stats
import numpy as np


# Mathematical formula
'''
data = [5,2,7,8,5,9,1,3,5]
mean = sum of all values/ Total no of values
mean = 45/9
mean = 5
'''
data = [5,2,7,8,5,9,1,3,5]
result = stats.mean(data)
print(result)


# Gives float value in numpy
sol = np.mean(data)
print(sol)
```

Median : Middle value of a sorted data

```
# Mathematical formula for even number
'''
data = [5,11,12,2,3,6,7,9]
len of data = 8
1) Sort the value in a ascending order
data = [2,3,5,6,7,8,11,12]

2) Median = sum of 2 middle value/2
median = 13/2
median = 6.5
'''


# Python implementation
data = [5,11,12,2,3,6,7,9]
middleValue = stats.median(data)
middleValue


# For odd number of values
data = [5,11,12,2,3,6,9]
middleValue = stats.median(data)
middleValue
```

Mode : The value having highest frequency/most occuring value/ most repeated value/ most common value

```
# Mathematical way
'''
data = [1,4,11,12,13,23,23,23,56,78,1,1,23,4,11,19]
how many times undique value is repeating
1 - 3
4 - 2
11 - 2
12 - 1
13 - 1
23 - 4 # 23 is having max frequency
56 - 1
78 - 1
19 - 1
# Mode of my data is : 23
'''
# python implementation
data = [1,4,11,12,13,23,23,23,56,78,1,1,23,4,11,19]
stats.mode(data)


# import pandas as pd
data = [1,4,11,12,13,23,23,23,56,78,1,1,23,4,11,19]
for i in data:
  print(data.count(i))



import pandas as pd
data = [1,4,11,12,13,23,23,23,56,78,1,1,23,4,11,19]
df = pd.Series(data)
df.value_counts()


# Multi modal case : Having more than one mode
data = [4,23,1,1,11,12,13,23,23,56,78,1,1,23,4,11,19]
stats.mode(data)
```

Note:

- If we have more than 4000 data - We will simply drop the null values for correct analysis
- More data does not mean better analysis, More correct means better analysis

---

- When we have continuous data(temp,weight,lenth) -> We will use (mean/median) to fill null values
- When we have categorical data(Fruits,Ratings) -> We will use mode to fill null values

Measure of dispersion

Start coding or generate with AI.

Start coding or generate with AI.