STATISTICS

- collection of methodsfor collecting, displaying, analyzing and drawing conclusion from data.

Descriptive statistics

Descriptive statistics is the branch of statistics that involves organizing, displaying, and describing data.

Inferential statistics

Inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a sample taken from that population.

- average
- maximum
- minimum
- percentage
- liklihood
- variance
- t test
- anova

data types-1

```
- Cross sectional
- Time series
```

data types-2

```
- univariate
- multivariate
```

Variable Types-1

```
- binomial
- multinomial
```

Variable Types-2

```
- ordinal variable
```

Variable Types-3

```
- ratio data
```

Variable Types-4

```
- interval data
```

Measure of central tendancy

- mode (repeated value)
- median (middle Value,outliers don't affect)
- mean (Average value, outliers affects)

parameters

> A parameter is a number that summarizes some aspect of the population as a whole.

population vs samples

> A population is any specific collection of objects of interest. A sample is any subset or subcollection of the population, including the case that the sample consists of the whole population, in which case it is termed a census. **Sample mean** = x = (Σ xi) / n **Population mean** = μ = (Σ Xi) / N

notions and terms

find basic stat (summary) by

```
import pandas as pd
df=pd.read_csv("iris.csv)
print(df.describe())
```

Measurement od dispersion

** variability ,scatter or disperse**

- is about how much value disperse around the mean
- also called
    - standard deviation (std)
    - standard error (se)
    - variance
    - bell curve
- range= minimum_to_maximum

Example – Calculation of variance and standard deviation

Let's calculate the variance of the follow data set: 2, 7, 3, 12, 9.

The first step is to calculate the **mean**. The sum is 33 and there are 5 data points. Therefore, the mean is 33 ÷ 5 = 6.6. Then you take **each** value in data set, **subtract the mean and square the difference**. For instance, for the first value:

(2 - 6.6)2 = 21.16

The **squared differences for all values are added**:

21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.20

The **sum is then divided by the number of data points**:

69.20 ÷5 = 13.84

The **variance** is 13.84. To get the **standard deviation, you calculate the square root of the variance**, which is 3.72.

mean with std is more usefull than only mean by itself

- mean is incomplete without std
- mean only gives small picture

Variable type matters while plotting graphs

- for catagorical type variables (qualitative) use count plot etc.
- for continous variables (quantitative) use scatter plot etc.
