Descriptive statistics

What is DS ?

As its name suggests, DS uses data to represent population. Like to represent the heights of student in college, we can use mean or some other parameter. By looking at one value we get the insight of whole data, we don't need to see whole data. There are two categories of DS.

a. Central Tendency : It is used to represent centre point of the data. There are three parameters in CT - mean, median and mod.
b. Measure of Variability : It is used to measure the variability in dataset, like how far datapoints are from each other. The parameters used are - variance, standard deviation, range.

Mean :

Advantages :
1. Easy to calculate
2. Many applications in practical life
3. Useful when data is equally distributed

Disadvantages :
1. It is very sensitive to outliers. Even a single outlier can affect the mean, making it unrepresentative of most of the data.
2. Not very good when distribution is skewed.
3. Can not be applied to nominal or ordinal data.
4. When population size is low, on discrete data it is not very useful.
5. Since outlier and skewed distribution and affect it, it is not robust.

Median :

Advantages :
1. Easy to calculate
2. Robust to outliers
3. Robust to skewness
4. Useful in ordinal data

Disadvantages :
1. Sensitive to population size
2. When data is not skewed and distributed well, on small range mean is better
3. On continuous data it is better on high population only

Mod :

Advantages :
1. Easy to calculate
2. Applicable to nominal data
3. Not sensitive to outliers

Disadvantages :
1. When many datapoints have same frequency, it is difficult to find meaningful mode
2. Not good for small sample size
3. Doesn't utilise all data
4. Applicable for only nominal data

5. Not affected by full range of data

Range :

Range is the max value - min value. It defines the range.

Variance :

It defines how values are far from mean. Ultimately, how far values are from each other.

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

Where,
σ^2 is variance, μ is mean, x_i is datapoint and N is number of datapoints.

Standard Deviation :

It is square root of variance. Std is preferred over variance.