**Statistics**

**Mean:** Summation of total number of values by total number of values or Sum of all the values by total number of values

Ex: data values = [1,2,3,4,5,6]

(1 + 2 + 3 + 4 + 5 + 6) / 6

**Median:** Sum of the first value and the last value by total.

1+6/6

**Mode:** most repetitive value or most repetitive multiple values (here we need to mention indentation of which we want for multiple repetitive)

Dataset[‘column’].mode()[0] 🡪 0 represent to choose Highest repetitive value

**Variance:** Difference between the mean of the data set to the average distance of the data points (or) Average distance from each point to the mean of the data set

**Calculate the mean:**

Mean = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 21 / 6 = 3.5

**Find the difference between each value and the mean:**

For the first data point (1): Difference = 3.5 - 1 = 2.5

For the second data point (2): Difference = 3.5 - 2 = 1.5

For the third data point (3): Difference = 3.5 - 3 = 0.5

For the fourth data point (4): Difference = 3.5 - 4 = -0.5

For the fifth data point (5): Difference = 3.5 - 5 = -1.5

For the sixth data point (6): Difference = 3.5 - 6 = -2.5

Square each difference:

(2.5)^2 = 6.25

(1.5)^2 = 2.25

(0.5)^2 = 0.25

(-0.5)^2 = 0.25

(-1.5)^2 = 2.25

(-2.5)^2 = 6.25

**Calculate the sum of the squared differences:**

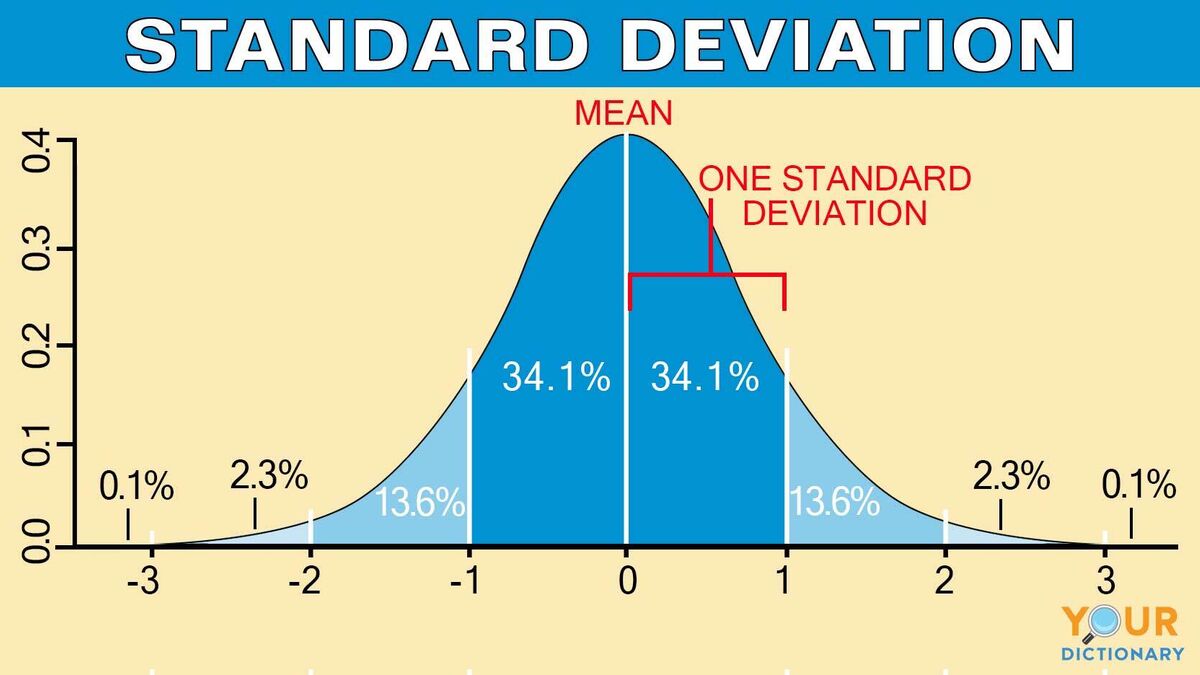
Sum = 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 = 17.5

Finally, calculate the variance:

**Variance** = Sum / Number of data points = 17.5 / 6 ≈ 2.92

**Standard Deviation:**

Spread of the data values form the mean in the bell curve.



If standard deviation = 0 mans no spread

The bigger the standard deviation the distance it has from the mean the smaller the standard deviation, which represents the value is closer to the mean.

**Hypothesis Testing:**

**Null Hypothesis:** Considering variable A and B if there is no relation between variable A And variable B, then it is called as Null Hypothesis. If I change variable A, there is no changes happen in variable B.

Null = Zero Relation

Then I will try to support this statement. And if I get succeed in supporting this statement, then We can accept null hypothesis.

**Alternate Hypothesis:** If we fail to support the Null Hypothesis with proper evidence, or by lack of evidence to support null hypothesis. Then we accept alternative hypothesis.  
Her alternative hypothesis says there is a relation between variable A and variable B If I change anything in variable A There is a chance of changes happen in variable B.

**Errors:**

Our goal is to accept or reject the Null Hypothesis in the concept of hypothesis testing but here there is a chance that we make mistakes. So, making mistakes we can consider in two different types of errors.

**Type 1 Error:** One is null hypothesis true, but we rejected it (estimated as False).

we denote it by **α.**

**Type 2 Error:** Another one is null hypothesis is false, but we Accepted it (estimated as True).

we denote it by **β.**

**T-Test:** It is used for continuous variable to do hypothesis testing.

Ex: Age, Prices

**Chi Square Test:** It is used for categorical variable / Binary classified variable.

Ex: yes/No, Male/Female.

**Anova Test:** It is used for multi classification problems. For multiple categorical variables.

Ex: Good, Average, and Poor.

**Under Fitting and Over Fitting:**



**Under Fitting = Bias**

**Over Fitting = Variance**

**Bias:** In this situation, model unable to find the important patterns in the data, giving error in understanding the Training data.

**Variance:** In this situation model oversimplify the data and also accept noise in the data which might be gives good accuracy and understanding the testing data. But it fails in understanding the training data

So, model has to perform well in training data as well as in testing data so that the model is ready for prediction. Then only we can get the maximum accuracy for prediction. So here model should not be oversimplifying the data by accepting the noise, or it should not miss the important patterns of the data. Model should always balance these two situations.