Statistics in Data Science

Class # 03

Today’s Agenda

1. Measure of Dispersion (Variance and Standard Deviation)
2. Percentiles
3. Quartiles
4. Five Number Summary (Max , Min , First Quartile , Median , Third Quartile)
5. How to detect an outlier
6. Box Plot

Measure of Dispersion

Dispersion means spread , means how well spread your Data

Example: {1,2,3,3,1}

Second Data {2,2,2,2,2}

Variance:

Population Variance Formula

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

Example

Data : {1,2,2,3,4,5}

Mu = 17/6 = 2.83

X-mu = -1.83 , -0.83 , 0.83 , 0.17, 1.17 , 2.17

(X-mu)2 = 3.34 , 0.6889 . 0.6889 , 0.03 , 1.37 , 4.71

Addition of all values of (X-mu)2 = 10.84

Putting in the formula of Variance = 10.84/6 = 1.81

Standard Deviation:

Formula : root of Variance

Root of (1.81) = 1.34

Conclusion

Variance tells us how spread our Data is and std tells us how far the next value from mean

Percentiles and Quartiles

To learn about Percentiles First revise the percentage concept

Example:

1,2,3,4,5

Question: What is the percentage of the odd Number ?

Percentile:

Definition:

A percentile is a value below which a certain percentage lie

Suppose we are saying 25 percentile of a specific value that means 25% of the values are below from the specific value

Example:

Data : 2,2,3,4,5,5,5,5,6,7,8,8,8,8,8,9,9,10,11,11,12

Q) What is the percentile ranking of 10?

Formula: (no. of values below x / no. of values in Data) x 100

Ans :

X is 10 = 16/20 = 0.80 = 80 %

Practice Question:

What is the percentile value of 11??

Removing the Outliers

Five Number Summary

1. Maximum
2. Minimum
3. Q1
4. Median
5. Q3

Example:

Data : 1 , 2 , 2 , 2 , 3 , 3 , 4 , 5 , 5 , 5 , 6 , 6 , 6 , 6 , 7 , 8 , 8 , 9 , 27

Q1: 25% = 3

Q3 : 75% = 7

IQR = 7 - 3 = 4

Continue...

Lower Fence: Q1 -1.5 (IQR)

Putting the values:

3 - 1.5 (4)

3 - 6 = **-3**

Continue...

Higher Fence: Q3 +1.5 (IQR)

Putting the values:

7 + 1.5 (4)

7 + 6 = **13**

Continue :

Higher Fence ⟺ Lower Fence

{ -3 ⟺ 13 }

...

Data : 1 , 2 , 2 , 2 , 3 , 3 , 4 , 5 , 5 , 5 , 6 , 6 , 6 , 6 , 7 , 8 , 8 , 9 ,

27 has been Removed

Final Answer:

Minimum Value = 1

Q1 = 3

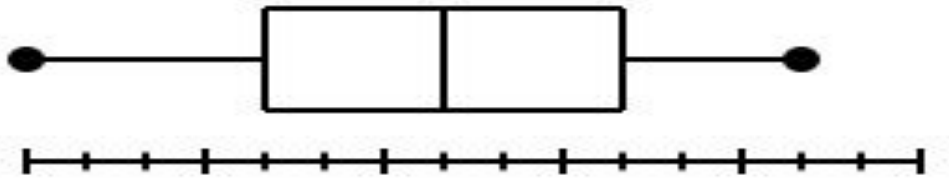
Q3 = 7

Median = 5

Maximum value = 9

Box Plot

...



Box Plot is Basically used to determine the Outlier