Statistics

1. Data types in statistics
2. Categorical and Numerical
3. Levels of data

* Nominal level
* Ordinal level
* Interval level
* Ratio level

1. Population
2. Sample
3. Population vs Sample
4. Inferential statistics
5. Frequency table: class vs frequency
6. Relative frequency table
7. Bar chart
8. Pie chart
9. Frequency distributed table: CI vs Interval freq
10. Histogram
11. Distribution graph
12. Central tendency
13. Mean- Median-mode
14. Mean vs Median
15. Outliers
16. Skewed data
17. Pos – Neg – normal
18. Bi mode – Uni mode – Multi mode
19. Data dispersion
20. Range : It will not consider middle value
21. Mean deviation: ; Total deviation =0
22. AMD: : graph is dis cont at point zero
23. Variance : value will be raise as well as units also
24. Standard deviation
25. Variance : how a data is varying it self
26. Covariance: how a data is varying w.r.t other variable
27. Pos or neg or neutral
28. Covariance Matrix
29. Pearson Correlation ( r ) -1 to 1 : cov(x,y)/std(x) \* std(y)
30. Percentile
31. Quartile
32. Outliers analysis
33. Box plot
34. How to Deal the outliers

* Drop if outliers <2%
* Fill with Median value
* Cap the values with Q3 and Q1

1. Normal distribution points
2. Emperical rule 68-95-99.7
3. Chebyshev inequality ( if data does not follows normal distribution ) (1-1/k^2)

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Income | Spending percentage | Credit usage points |
| 30 | 100000 | 30 | 40 |
| 31 | 500000 | 55 | 140 |
| 32 | 1000000 | 75 | 250 |

The number of columns : 4 it is a 4 dimensions of data

First observation: (30,100000,30,40)

Every population of data has different range of values and different units

It is very very important to **make all the data points under one scale**

1. Z-standardization ( Z-score)
2. Normalization

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Income | Spending percentage | Credit usage points |
| 30 | 100000 | 30 | 40 |
| 31 | 500000 | 55 | 140 |
| 32 | 1000000 | 75 | 250 |

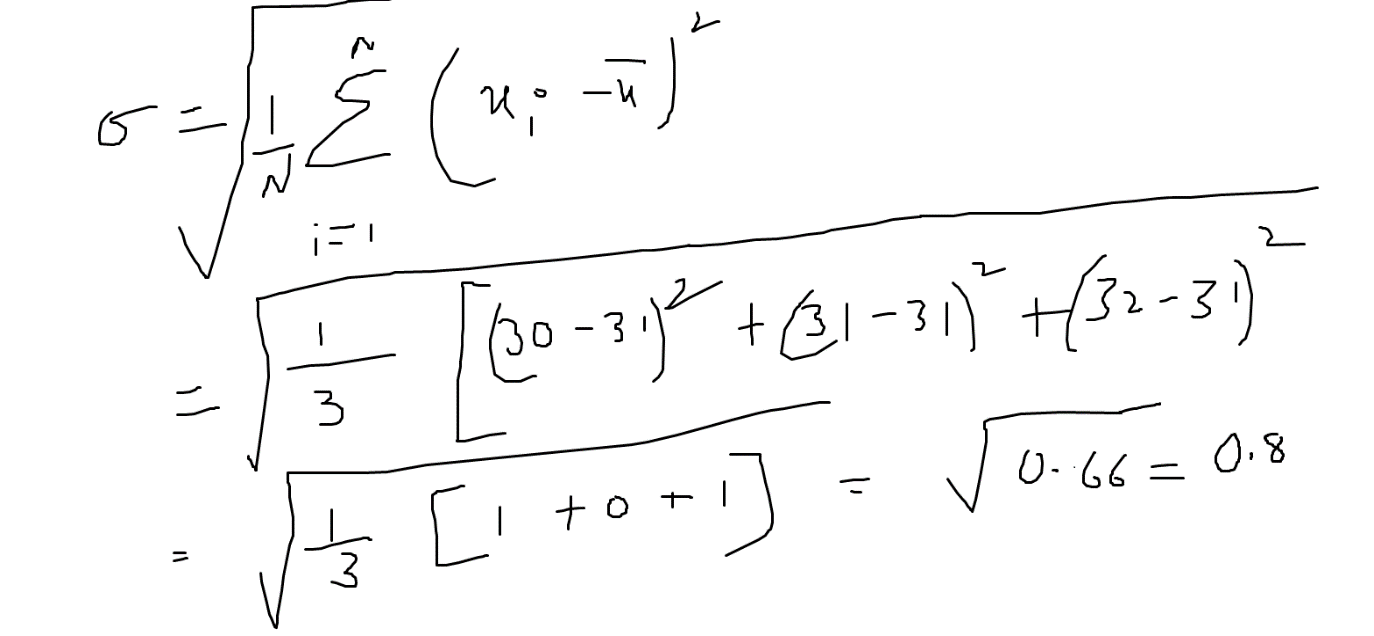
30 31 32 what is the mean = 31

Std = 30-31= -1 ==== > 1

31-31 =0 ==== > 0

32-31=1 ====== > 1

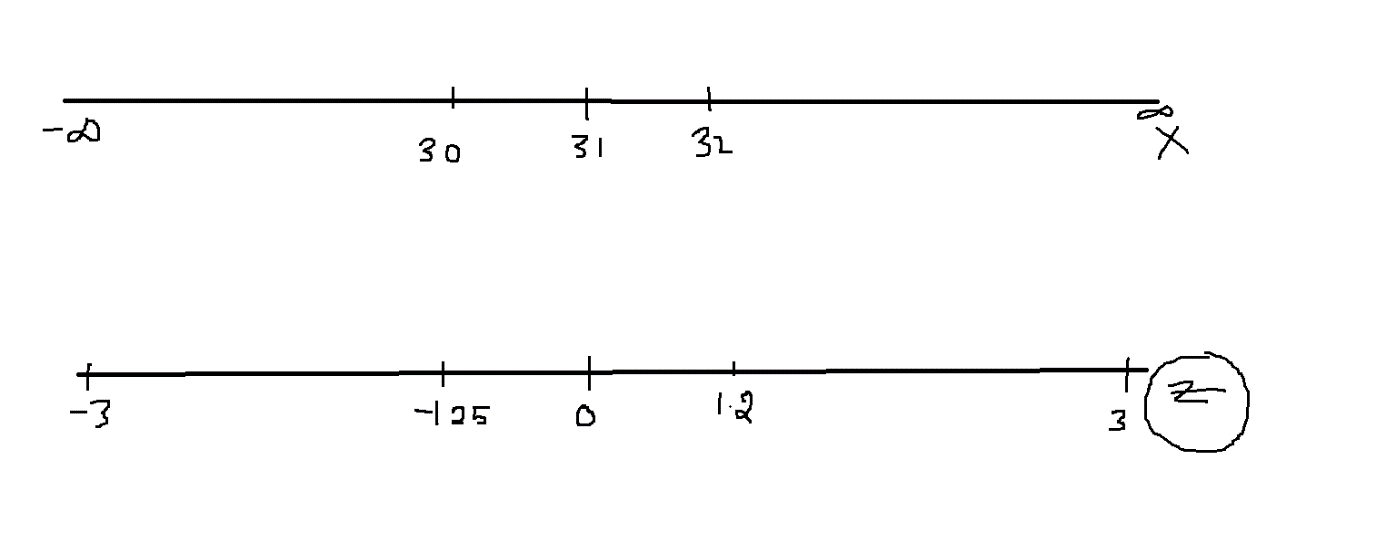
1+0+1 = 2/3= 0.66 = root (0.66)



|  |  |
| --- | --- |
| Age (x)a | Age (Z) = (x-u)/sigma |
| 30 | (30-31)/0.8= -1/0.8= -10/8= -1.25 |
| 31 | 31-31/0.8=0 |
| 32 | 32-31/0.8=1.25 |

Distance metrics 30,100000 31,500000

-1.25 to 2



Now thinking how python will use this formulae’s : Packages

LR ML : 1 week 1min

standardization and z score both are same

sir how to decide which one we will use , normalization or z score

don't we need to verify whether the data follows normal distribution

Read about Normal scale and Log scale

Normal scale and Z-scale

Machine Learning algorithms find the relation between then

Cov(age,Income) = age 30 31 32 income 1lak 5 lak 10lak (100000-30)^2

|  |  |
| --- | --- |
| Age X | Age normalization |
| 30 | 30-30 ====0 |
| 31 |  |
| 32 | 32-30/32-30=1 |