Lecture - 03

Distribution,

→ Normal (or) Gaussian distribution.

→ Standard Normal distribution.

→ Z score.

→ Log Normal distribution

→ Bernaullis distribution.

→ Binomial distribution.

Practical,

→ Mean, Median, Mode.

→ Varience, Standard deviation.

→ Histogram, Pdf, Barplot, Violin plot.

→ IQR

→ Log Normal distribution.

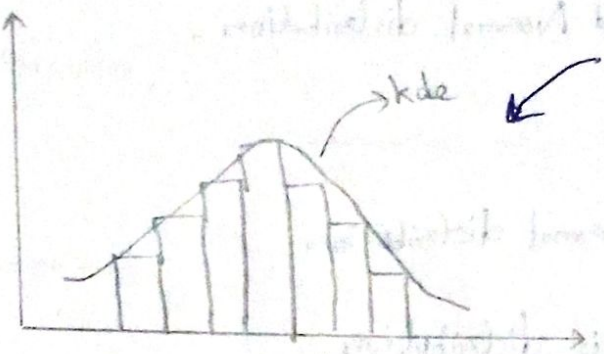
What is distribution?

How the data points are actually distributed in the visualized graph.

Example,

age: {24, 26, 28, 15, 16, 17, 34 . . . . . }

Plot histogram for age Something look like,



age, height, marks, weight alone in these form by domain expert.

These type of distribution is known as Normal (or) Gaussian distribution.

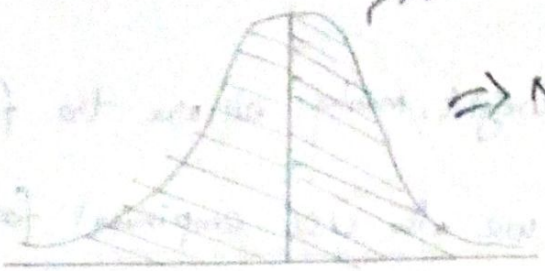
In world, most of the data are follows the Gaussian distribution.

Gaussian / Normal Distribution:

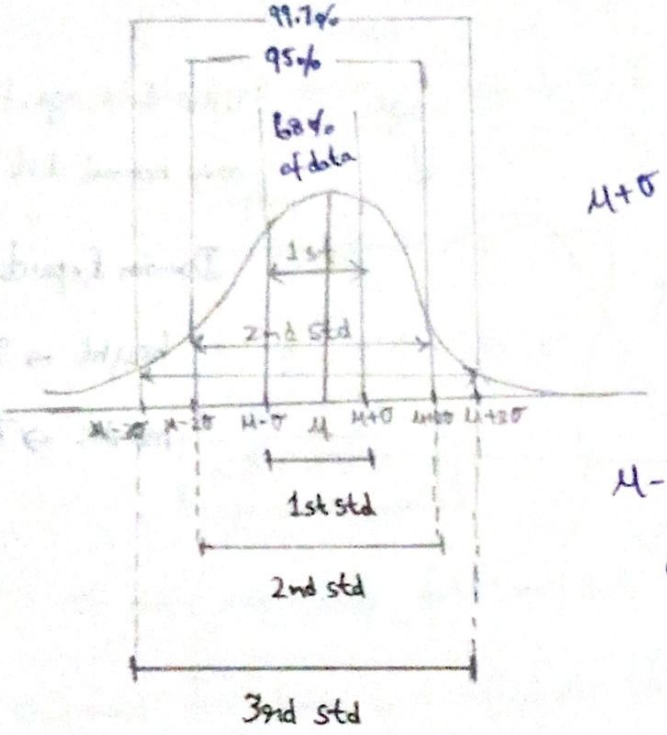
→ It is the distribution it looks like an bell-curve.

→ Ball curve

⇒ Normal distribution



↳ It may be Mean, Median (or) Mode



μ+σ ⇒ one standard deviation right from the mean.

μ-σ ⇒ one standard deviation left from the mean

Empirical formula: [This rule is applicable for normal dist]

This formula says the 68-95-99.7 % Rule

This rule is applied on the normal distribution.

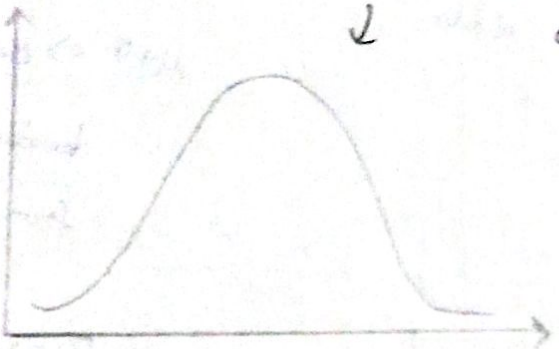
68% → 1st std

95% → 2nd std

99.7% → 3rd Std

Example,

age, height, weight, marks all are the form of normal distribution. we also use empirical formula for feature.



age
↓

who told age, height, are normal dist?

Domain Expert,

height → Doctor.

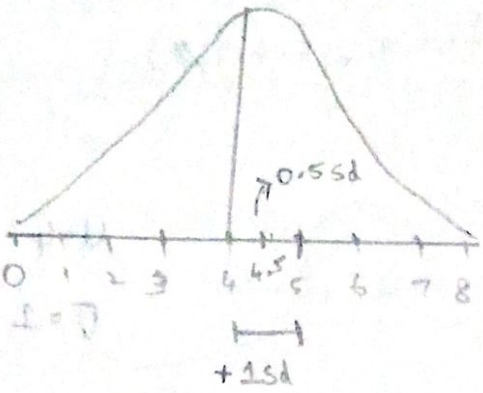
marks → Teacher.

Z-Score:

→ Z-Score refers to find out the how much Standard deviation of the data (or) Value from the mean.

→ Example,

$\mu = 4, \sigma = 1$



How much Standard deviation of 4.5 ?

Simply,

$Sd = \sigma = 1$

$1 - 0.5 = 0.5$

$\boxed{Sd = +0.5}$

This method

It only applicable for Small data points and Small no of standard deviation and difficult to find.

So, Using Z-Score to find out the Sd.

formula $\Rightarrow \boxed{Z_{score} = \frac{x_i - \mu}{\sigma}}$

To find 4.75,

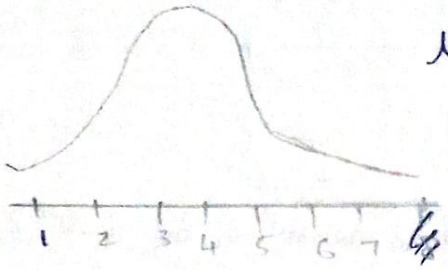
$Z_{score} = \frac{4.75 - 4}{1} = +0.75$ sd.

Suppose (-) to consider the left sd from the Mean.

[obiviously to find left or right Sd use Z-score]

Example,

data: {1, 2, 3, 4, 5, 6, 7, 8}



$\mu = 4$

$\sigma = 1$

To find out Z-score for every element,

$Z(1) = \frac{1-4}{1} = -3$

$Z(2) = \frac{2-4}{1} = -2$

$Z(3) = \frac{3-4}{1} = -1$

$Z(4) = \frac{4-4}{1} = 0$

$Z(5) = \frac{5-4}{1} = 1$

$Z(6) = \frac{6-4}{1} = 2$

$Z(7) = \frac{7-4}{1} = 3$

data = {1, 2, 3, 4, 5, 6, 7}

⇓ After Applying Z-score for each.

data = {−3, −2, −1, 0, 1, 2, 3}

↓ −3 Sd

↓ −2 Sd from mean

↓ +3 Sd from mean

Note that,

data = {-3, -2, -1, 0, 1, 2, 3}

⇓

definitely $\mu = 0$, $\sigma = 1$

If $\mu = 0$, $\sigma = 1$, Then the datapoints are fall in Standard normal distribution.

The given data is the distribution of Normal distribution, the we apply Z-score. The data are Converted to the Standard Normal distribution.

Standard Normal distribution:

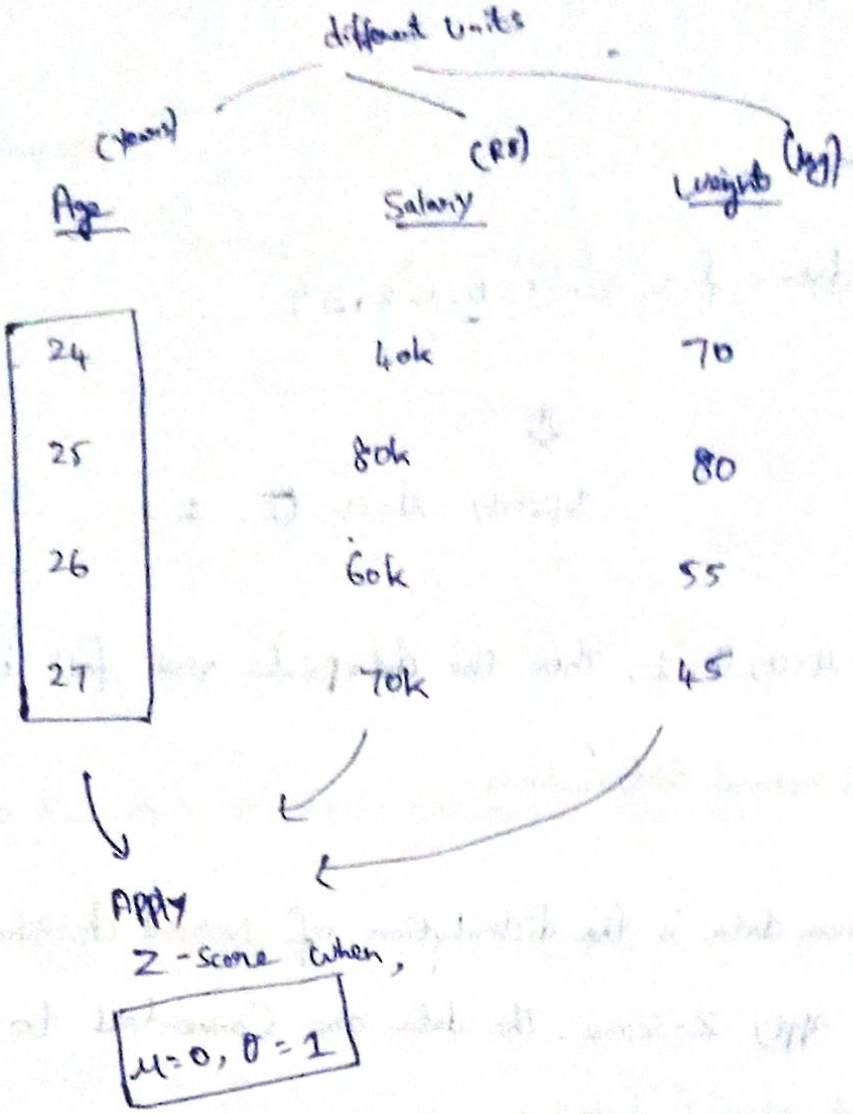
Property:

$\mu = 0$, $\sigma = 1$

Random Variable $(x) = X \sim SND(\mu = 0, \sigma = 1)$

Real time Example:

Consider the dataset with feature age, Salary, Weight.



This applying z-score process is known as Standardization.

Standardization: [Uses z-score]

Standardization is an important technique that is mostly performed pre-processing step. To standardize the range of features of an input data set.

Many of algorithm need Standization because of features have different units to scale.

Example,

Weight

Height

Consider Weight and height in the dataset.

To build machine learning model, Height is dominated with Weight. Maybe height will be broad range of value. To avoid So, we need to standardize the value for height and weight. Range: (-) to (+)

Normalization: [Uses Min max Scaler]

Normalization is also method for used in feature Scaling and pre-processing steps. It Using Min-Max Scaler. [There are so many formula applied in Minmax Scaler based on needs]

Range = 0 to 1

Example, In CNN (Image classification)

Every pixel range from 0 to 255

Using minmax scaler to convert range to 0 to 1.

(00)

$\frac{0}{255} = 0, \frac{1}{255} = \ldots, \frac{255}{255} = 1 \quad [0, \ldots, 1]$

Practical Examination:

ODI Series for Cricket of 2020 and 2021.

2020	2021
Series average score = 250	Series average score = 260
Standard deviation = 10	Standard deviation = 12
Team final score = 240	Team final score = 245

Compare to both the scenes in which year team final score ~~prey~~ was better (or) perform well?

Ans,

In 2020,

$$Z_{score} = \frac{240 - 250}{10} = \frac{-10}{10} = -1$$

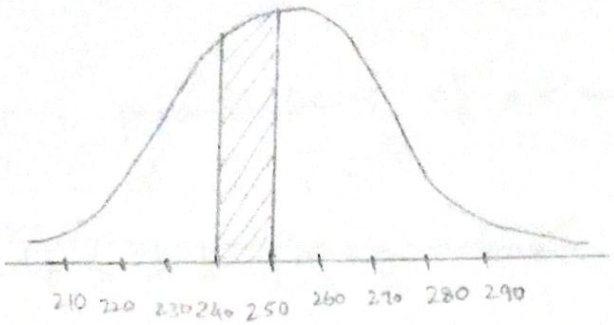
$$Z_{score} \text{ of } 2020 = 1$$

In 2021,

$$Z_{score} = \frac{245 - 260}{12} = \frac{-15}{12_4} = \frac{5}{4} = -1.25$$

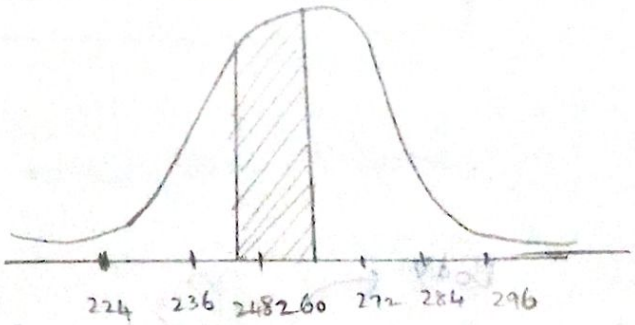
In 2020,

$\mu = 250 \quad x_i = 240 \quad \sigma = 10$



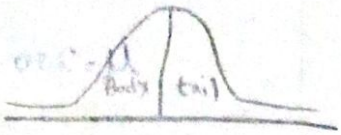
In 2021,

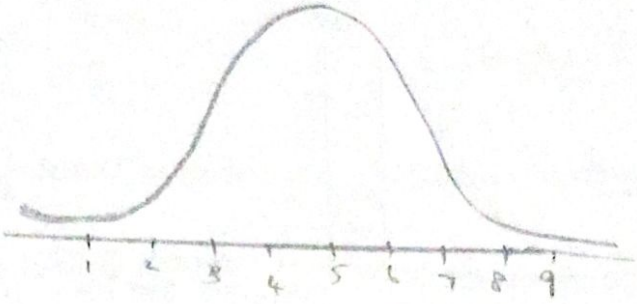
$\mu = 260 \quad , \quad x_i = 245 \quad , \quad \sigma = 12$



Based on the area Spread in 2021 is high Compared to 2020. It means the data ranges may be high between the area. So, definitely 2021 was better score Compare than 2020.

Stats Interview Question:

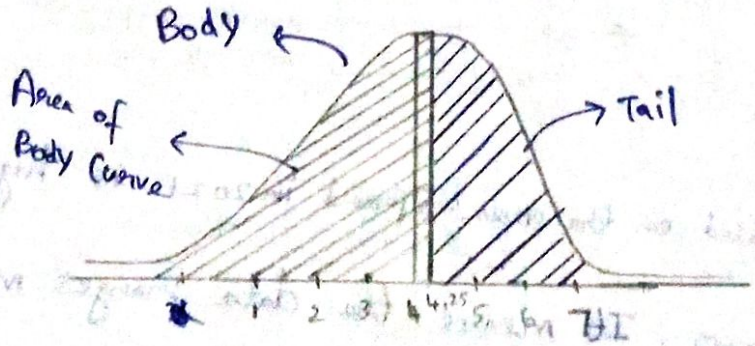




(1) Question:

What Percentage of Scores fall above 4.25?

Consider $\mu = 4$, $\sigma = 1$



To find Scores fall above 4.25?

Two ways,

(1). First using basic Simple Percentage,

~~This is not efficient method, It depends very on incorrect result. Only support for~~ some data.

$$\text{Percentage} = \frac{4.25}{7} \times 100 = 60.71428\ \%$$

$$= 60\ \%$$

4.25 is present 60 % of entire data.

Our Question is to fall above 4.25 ?

So simple,

$$100\ \% - 60\ \% = 40\%$$

Scores above 4.25 = 40 %

② . Second method Using Z-score,

$$Z_{score} = \frac{4.25 - 4}{1} = 0.25$$

~~Right score~~

See Z_{score} table with respect to ~~Z score~~.

0.2 and 0.05 = 0.59871 ~~0.40129~~

(left area region)

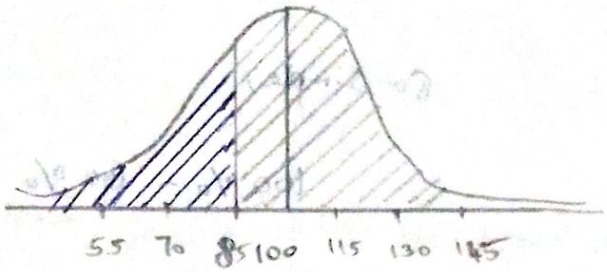
We need above 4.25,

$$1 - \text{leftarea} = 1 - 0.59871$$

$$= 0.40129 \Rightarrow 40\ \% :$$

(2) Question:

In India the average IQ is 100, with a standard deviation of 15. what percentage of the population would you expect to have an IQ lower than 85 ?



$\mu = 100$, $\sigma = 15$

$$Z_{score} = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$

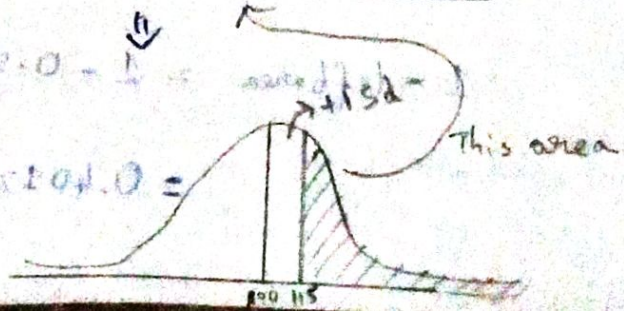
In Z table (-1.0) is 0.15866 [direct look in Ztable]

(00)

See (1.0) in Ztable is 0.84134

So,

$$1 - 0.84134 = 0.15866$$



In Gaussian distribution, Both the sides are equal.

The same way both the Sd size area also same.

So, The area is 0.15866 is nothing but 15.8%.

Real time practical using Python,

```
import numpy as np
import matplotlib.pyplot as plt
% matplotlib inline
import seaborn as sns
import statistics

df = sns.load_dataset('tips')
df.head()

#Mean, mode, median

np.mean(df['total_bill'])
np.median(df['total_bill'])
Statistics.mode(df['total_bill']) # well suited for Categorical
                                    Value
# Visualize boxplot

sns.boxplot(df['total_bill'])
```

Visualize histplot

```
Sns.histplot(df['total_bill'])
```

```
Sns.histplot(df['total_bill'], kde = True)
```
pdf applied.

pdf used to wheather the distribution is gaussian (or) not.

```
Sns.Cookplot(df[parameter])
```

Find Percentile of feature values.

Q_1 Q_3

IQR = $Q_3 - Q_1$

```
np.percentile(df['total_bill'], [25, 75])
```

```
np.percentile(df['total_bill'], [25, 99]
```