# Advance Statistics

Distribution

1 Normal Distribution

2 Standard normal distribution

3 Z score

4 Log Normal Distribution

5 Bernoulli Distribution

6 Binomial Distribution

Practical

1 Mean, Median, Mode

2 Variance, Standard Deviation

3 Histogram, pdf, Bar Plot. Violent plot

4 IQR

5 Log Normal Distribution

1 Distributions

Ages = {24,26,27,28,30,32------} Discreate continuous data

Curve

Fig 1

How do we see this data set in visualize way?

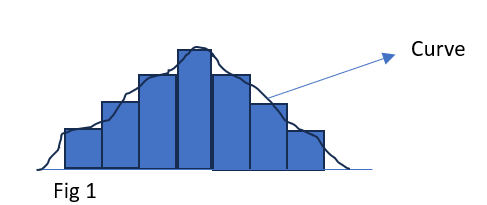
If I have some set of data probably,

I want to plot the data somewhere.

The best and easy way that one can probably think about histogram.

We have already seen how to create histogram you will be able to create diagrams like in Fig 1.

Building bar left to right with group (0-10, 10 – 20, 20 -30). In this way we will be able to get building like this

and finally, we do, smooth the curve in the histogram. This curve looks like bell shape curve, So considering this let’s see first distribution call Gaussian distribution or Normal Distribution.

Why is distribution basically used?

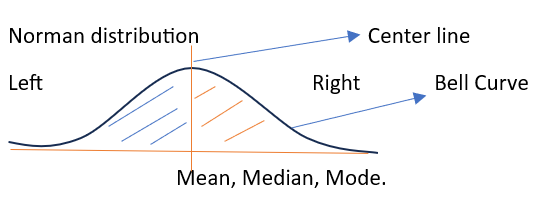
Why is this different kind of distribution used?

Main purpose of distribution is to have basic idea about dataset, when we discuss about gaussian or normal distribution, we can figure out some basic properties.

Norman distribution Center line

Left Right Bell Curve Property of bell curve is that both side

Mean, Median, Mode.

Property of bell curve is that both side bell curve is exactly symmetrical.

Which means area of each side is equal. Which means data represent each side is equal.

This kind of distribution, we can derive lot of conclusions,

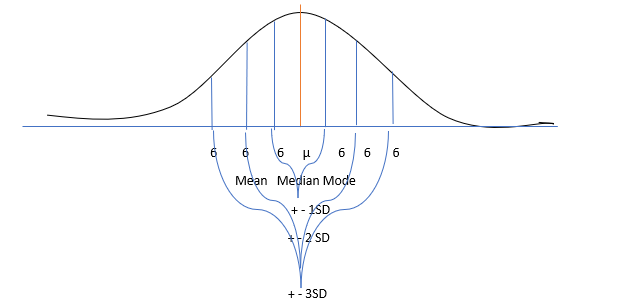
6 6 6 µ 6 6 6

Mean Median Mode

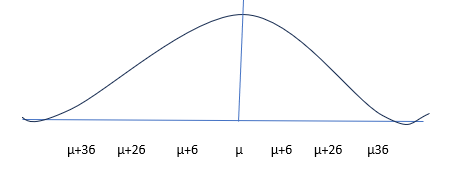
+ - 1SD

+ - 2 SD

+ - 3SD



µ+36 µ+26 µ+6 µ µ+6 µ+26 µ36

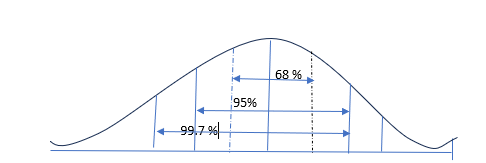


68 %

95%

99.7 %

Empirical Formula,



With this consideration, in the data set {100 data parts}. With this curve normal distribution 68 points lies in between 1 SD right and left. And 95% data points lies in between 2 SD right and left. So, on 99.7 SD data points lies between 99.7 %.

E.g,

Heights – Normal Distribution

Domain Expert -> {Doctors} it distribution shapes like this,

1, Weight also follows in Normal distribution / Gaussian Distribution).

2 IRIS DATSET – Petal length sepal length it follows Gaussian distribution.

Eg,

µ=4 ; SD=1 with information can we create distribution ?

Yes.

In terms of SD, you may think 4.5 lies between 4 and 5 or + 0.5 SD to right .

0 1 2 3 4 5 6 7 8

µ

Similar, where does 4.47 fall

It also falls between 4 and 5 but, it is hard to find .75 where exactly lies between 4 to 5.

For this we can use Z score,

Z score will basically help you find out wherever I talk about a value how much SD away frm the mean

The formula

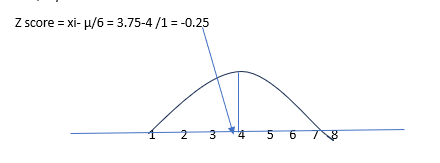
Z score = xi- µ/6 = 4.75 -4/1 = 0.75 SD

-0.75 SD to the right

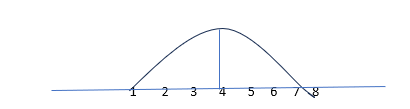
E.G., Try to find out where does 3.75 falls?

Z score = xi- µ/6 = 3.75-4 /1 = -0.25

1 2 3 4 5 6 7 8



If I apply Z score to every value what will happened.



Z score = (xi-4)/1 = (1-4)/1 = -3

Z score = (xi-4)/1 = (2-4)/1 = -2

Z score = (xi -4)/1 = (3-4)/1 = -1

Z score = (xi-4)/1 = (4-4)/1 = 0

Same up to item = 8

Normal Distribution [ 1, 2, 3, 4, 5, 6, 7, 8]

Z score distribution { -3, -2, -1, 0, 1, 2, 3} (After apply Z score). Standard Normal Distribution.

The properties of Standard Normal Distribution are µ = 0, SD (6) = 1

Why do we do this?

What is the use of doing this?

Let’s see by Practical application datasets.

Age in year Salary in Rs. Weight Kg

24 40k 70

25 80k 80

26 60k 55

27 70k 45

Units are completely different our main target should be that we should try to bring up in a form probably in this form where the mean is 0 and SD is equals to 1 at that point of time, I can apply Standard normal distribution.

In the Age data set, applying the Z score convert this into Standard Normal Distribution.

Similarly, Applying Z score in salary data convert to standard normal distribution. And same to Weight data too.

This process basically called standardization. Standardization means z score has been applied.

Standardization is a process where distribution convert into standard normal distribution, where mean is 0 and standard deviation is 1.

What is normalization?

Standardizing property (µ= 0, SD=1) but in normalization entire age value shift {0 to 1}.

How do we normalization?

Formula to normalization is minimize scalar. One must provide (0 to 1).

Age

24

25

26

27

If you want to shift Age distribution – 1 +1

Where do we apply Normalization?

Let’s see with this example.

How classify images by CNN.

Understand every image has a pixel.

Whenever you are doing image training image classification or object detection

In this case ever image has a pixel suppose I have 4 crosses

0-255

/255

Above Min Max scale 0-1 or normalization.

Where the minimum value 0 is assigned to 0 and the maximum value 255 is converted to 1

So, when we do this automatically, we can apply this kind of min max scale called normalization.

# min max scalar has a different power formula but just take the value each pixel divided by 255.

So, when we do this specific division, divided by 255, all your values will be getting changed between 0 to 1 and this is another type of normalization process.

Let’s see one practical example of Z score {India vs SA}

ODI series – 2021

The series Average of 2021 = 250

SD = 10

Team final score = 240

ODI series – 2020

Series average of team scoring 2020 = 260

SD of the score of all the match = 12

Team final score = 245

Compare both the series, in which year final score was better?

Checking this people may confuse.

In this case we will apply Z score

2021 = Z score = (xi-u)/6 = 240-250 /10 = -1

2020 = Z score = 245-260/12 = -15/12= 1.2

In 2021 In 2020

U =250, x =240, SD=10, Z score = -1 u=260, Xi =245, SD=12, Z score = -1.25

220 230 240 250 260 270 280 224 236 248 260 272 284 296

|  |  |
| --- | --- |
| U =250, x =240, SD=10, Z score = -1 | u=260, Xi =245, SD=12, Z score = -1.25 |

In 2021 SD is less; in 2020 SD is more.

Stat interview question

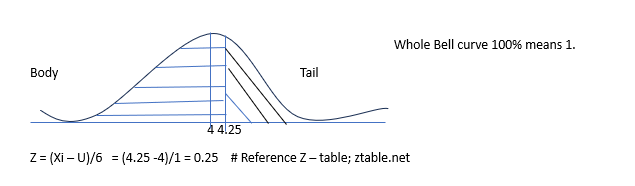
What percentage of score fall above 4.25?

Whole Bell curve 100% means 1.

Body Tail

4 4.25

Z = (Xi – U)/6 = (4.25 -4)/1 = 0.25 # Reference Z – table; ztable.net



1. left Area; 1- 0.5987 (from ztable.net)

0.4013x100 = 40.13 %

2nd Question

In India the average IQ is 100 with a standard deviation of 15. What percentage of the population would you expect to an IQ lower that 85

Z = (85-100)/100 = -15/15 = -1

55 70 85 100 115 130 145

1 – 0.8434 = .15866