Session With Sumit

DATA SCIENCE INTERVIEW PREPARATION SERIES

SKEWNESS AND KURTOSIS

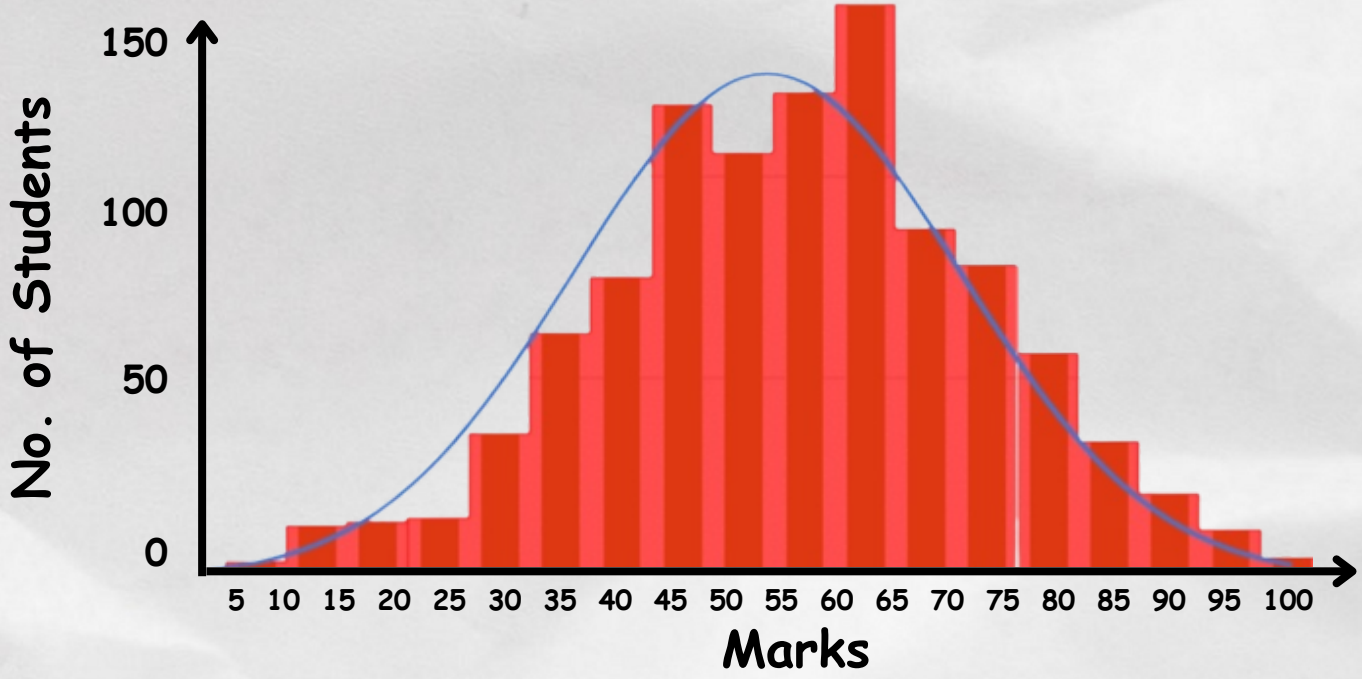
Dholakpur Public School (DPS) plans to introduce special classes for one or two subjects, and up to 50 students need to be selected for each subject. MasterJi is uncertain about which subjects and students to choose. Therefore, Bheem suggests conducting a surprise test in English, Maths, and Science to identify the students and subjects that require special classes. MasterJi agrees with Bheem's idea.

After the examination results were announced, MasterJi was given multiple sheets of results. He was confused and didn't know how to analyze them. He called Bheem and requested him to analyze the sheets by the end of the day. Bheem quickly created a plan and completed the analysis within an hour. MasterJi was amazed and asked how he was able to finish it so quickly. Bheem smiled and said,

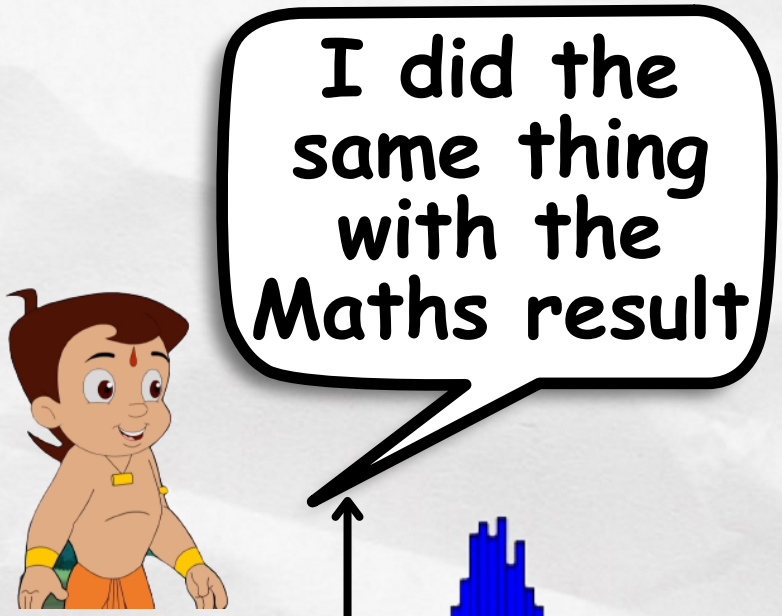
"I used Skewness and Kurtosis."

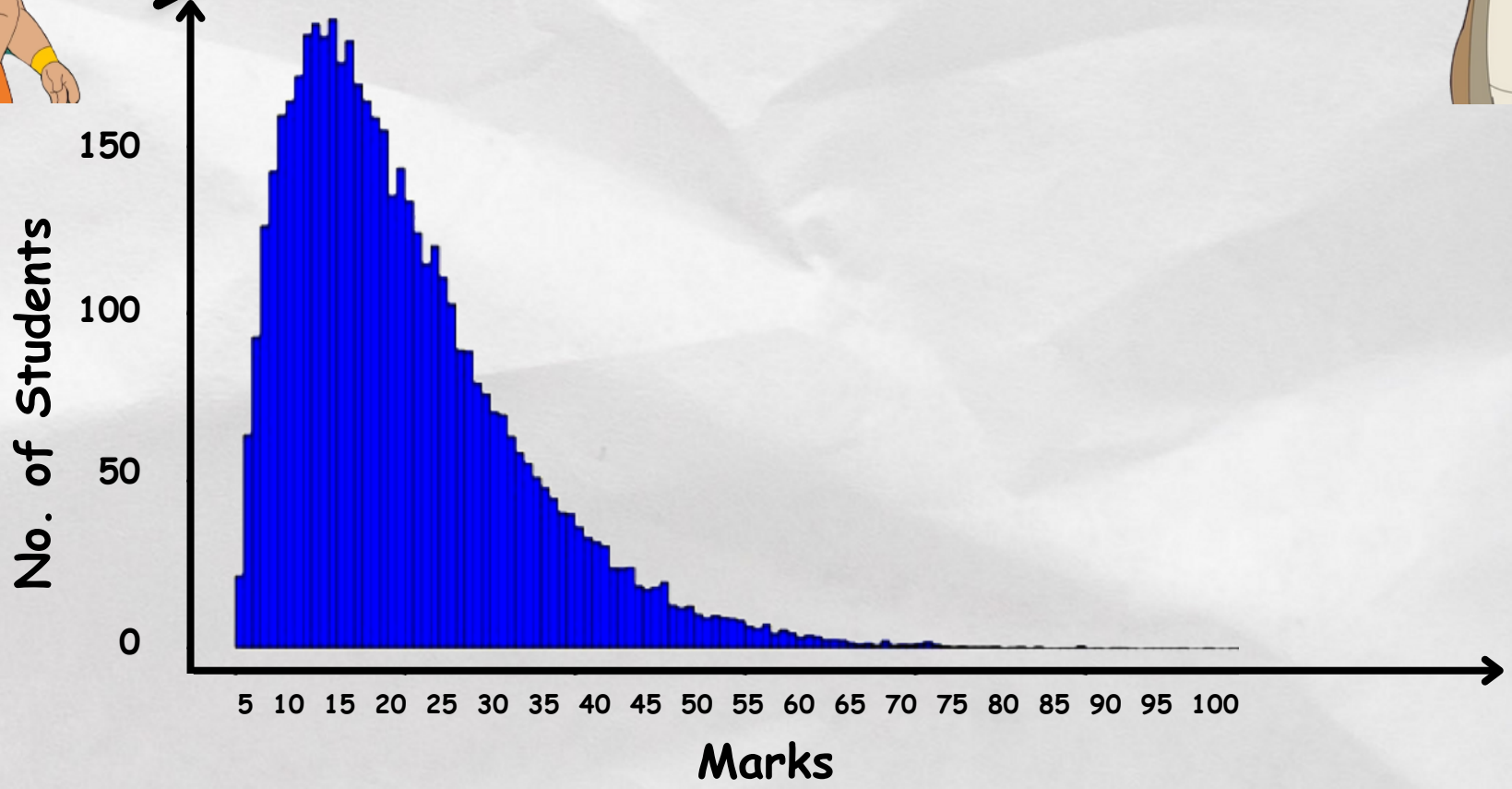
Using the English scores, I have created a histogram in which X-Axis represents the students' marks and Y-Axis represents the frequency.

So, we got the following histogram:-



This graph has Mean, Median and Mode all lying at the center, which means that maximum students have scored marks between 50-60 and this is the average result.

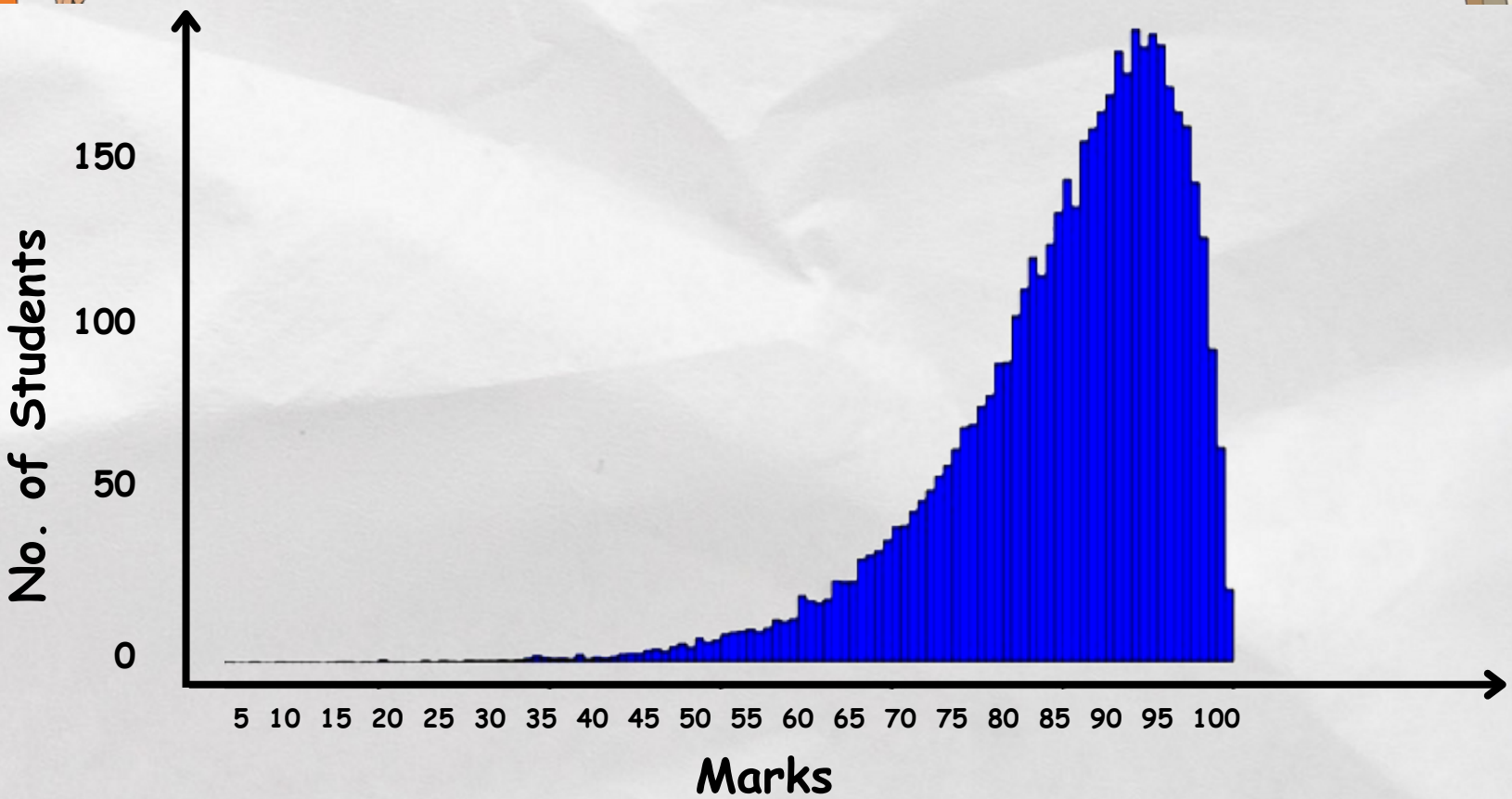




I noticed that most students scored less, causing the graph to shift to the left with few points on the right.

After calculating the Mean, Median, and Mode, I discovered that Mean>Median>Mode

Lets look at science result:-



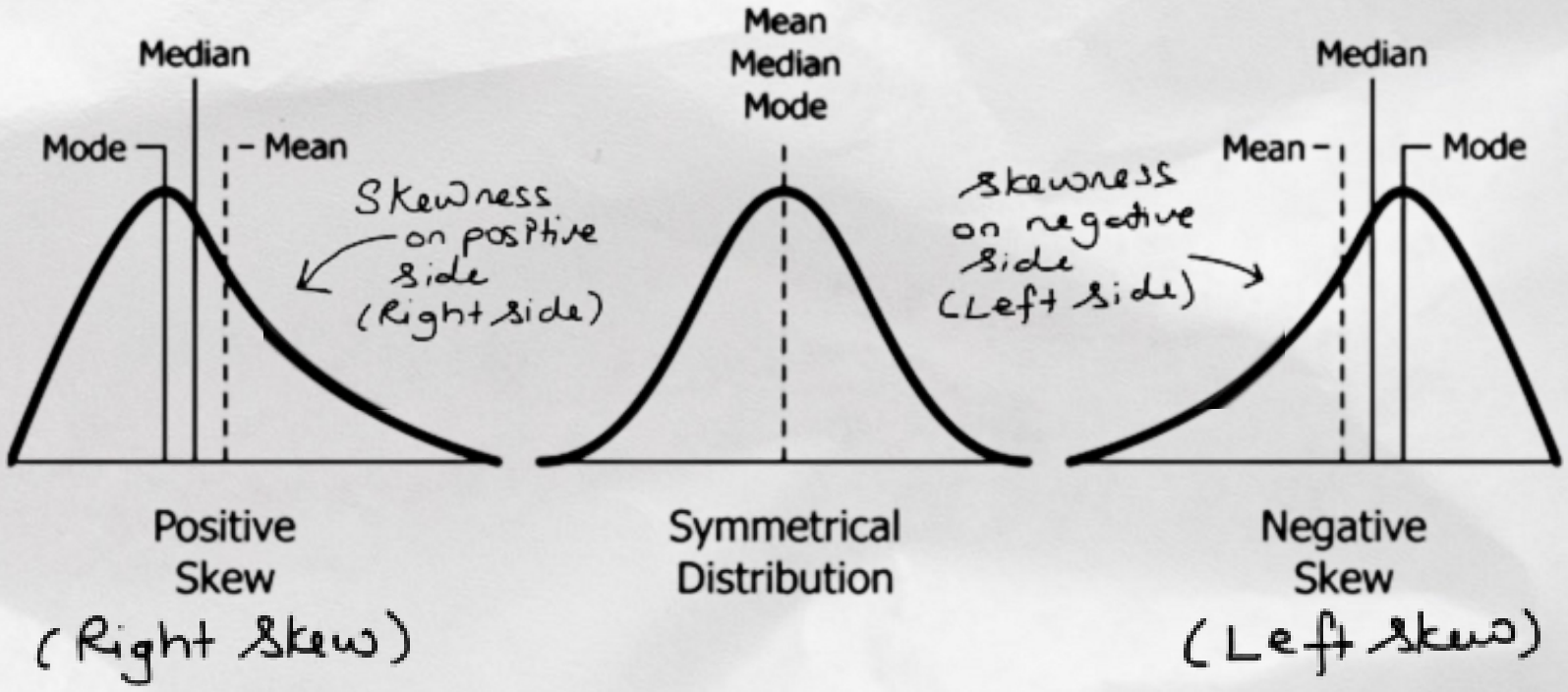
I noticed that there are more data points towards the right and very few towards the left. This is because the majority of the students scored higher marks while only a few could not perform well.

Additionally, here Mean>Median>Mode

But where is the skewness coming in?

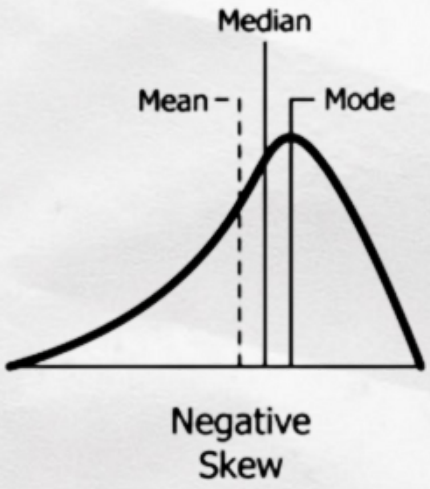
Let me explain to you further with the help of definition now

- Skewness is a measure of the asymmetry of a distribution. Here, asymmetry means that the graph is not equal on both parts.
- We have three types of skewness- Positive, Negative or Symmetry.
- Let's see all the graphs but this time without histograms:-



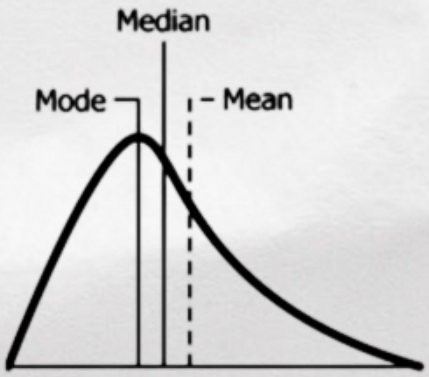
- In the Positive Skew graph, we will see data points are more shifted towards the left and it's not forming a perfect bell shape curve.
- In negative skew, we will see data points are more shifted towards the right.
- But the symmetry curve shows a perfect graph, where maximum points lie in the centre, forming a perfect curve.

Let's revisit the graph:-



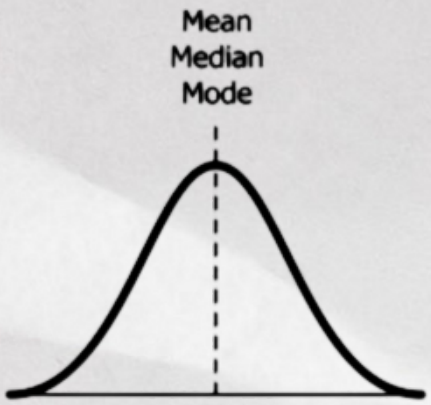
Negative Skew

It is negative skew or left skewed.
Now here our mean is less than the median and mode, which means that the maximum number of students from this group have scored the highest marks and very few scored less marks.
That is why the graph is shifted towards the right. Meaning students understand Science well.



Positive Skew

It is called positive skew or right skewed. In this graph, we can see that mean is more than the median and mode.
Now this graph interprets that the maximum number of students have scored lesser marks and very few students have scored higher marks. So that's why the mode is on the lower side
So, Maths subject needs more attention.



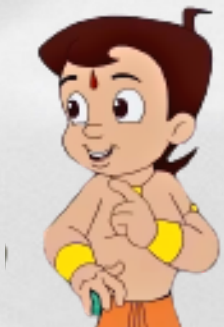
Symmetrical Distribution

The mean, median, and mode are all the same, indicating a perfectly symmetric graph with no skewness. Based on our data, it can be interpreted that students have an average level of proficiency in English. Therefore, English should receive more attention.

Using skewness, I concluded that the Maths and English subjects require special classes.

Let's look at it in more detail.

- Skewness is a measure of the asymmetry of a distribution.
- When a distribution is perfectly symmetrical, it has zero skewness.
- However, if the distribution is asymmetrical, meaning it's "lopsided" in some way, it will have a non-zero skewness value.
- There are two types of skewness: positive skewness and negative skewness.
- Positive skewness occurs when the tail of the distribution extends further to the right than to the left, while negative skewness occurs when the tail extends further to the left than to the right.



This clears how to do interpretation but how do we calculate skewness?

There is a formula to calculate the skewness if the data is given. Here's the formula

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{(x-\bar{x})}{s}\right)^3$$

where n is the number of data points and x is the mean and x is the data and s is standard deviation.

Let's see this with the help of an example:-
Let's say we have a dataset of exam scores for a group of students, and we want to know if the distribution of scores is symmetrical or skewed.

Following are the exam scores of 20 students:
75, 78, 80, 82, 84, 85, 86, 87, 88, 89, 90, 90, 91, 92, 94, 95, 96, 98, 99, 100

To calculate skewness for this dataset, we need to first calculate mean and standard deviation. Here's how we can do that:

1. Calculate the mean of the dataset:
Mean=
(75+78+80+82+84+85+86+87+88+89+90+90+91+92+94+95+96+98+99+100) / 20
= 89.2
2. Calculate the standard deviation of the dataset:
Standard deviation = 8.069

Skewness=(1/20) * [(75-89.2)/8.069]³ + (1/20) * [(78-89.2)/8.069]³ + ... + (1/20) * [(100-89.2)/8.069]^3
= 0.412

Now the calculation is clear but what about it's interpretation on number line

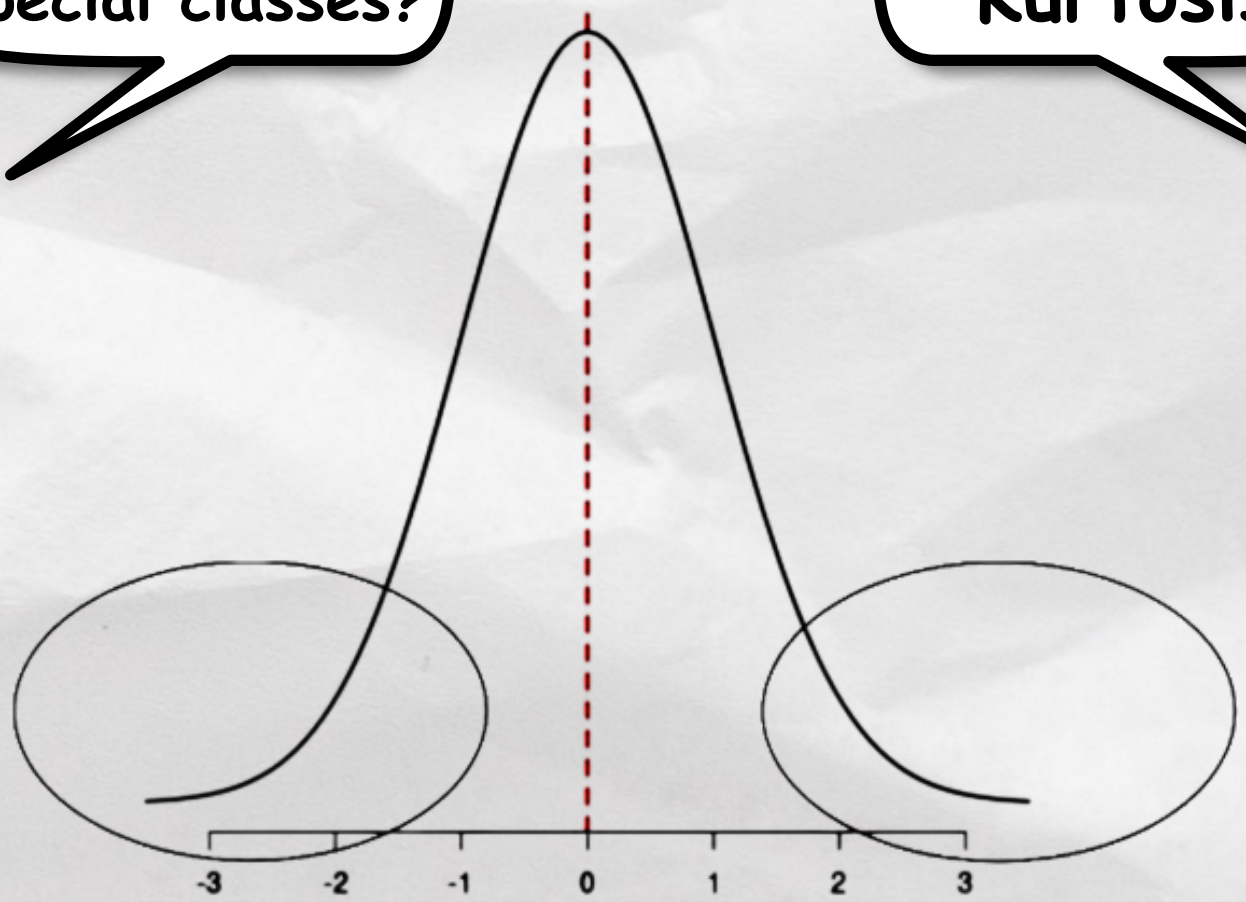
Sure,

Look over the number line

If the value is 0	it is a perfect symmetrical curve with no skewness
If the value is between 0 to 0.5	moderately skewed towards positive
If the value is between -0.5 to 0	moderately skewed towards negative
If the value is between 0.5 to 1	it is positive skewed
If the value is 1+	highly skewed towards positive
If the value is between -1 to -0.5	it is negative skewed
If the value is less than -1	highly negative skewed

But, how do we evaluate which students should enroll in the special classes?

To do that, I used the concept of Kurtosis.

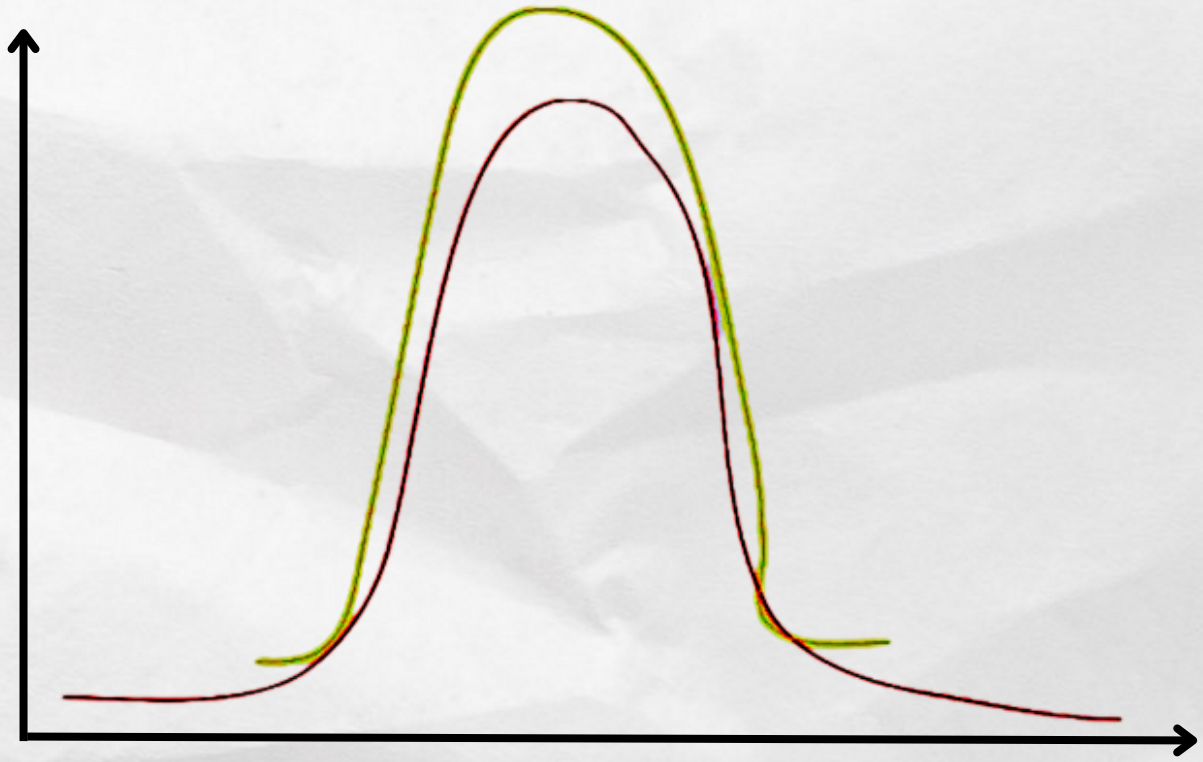


Let's revisit the definition first:-
Kurtosis measures the "peakedness of a tail" in a distribution:

- How much of the data is concentrated in the centre of the distribution versus in the tails. Where tail refers to the points lying in the circle.
- In every data set, you get some points that are extremely low or extremely high and are not in the centre and lying away from the center.

Kurtosis is the measure of these points.

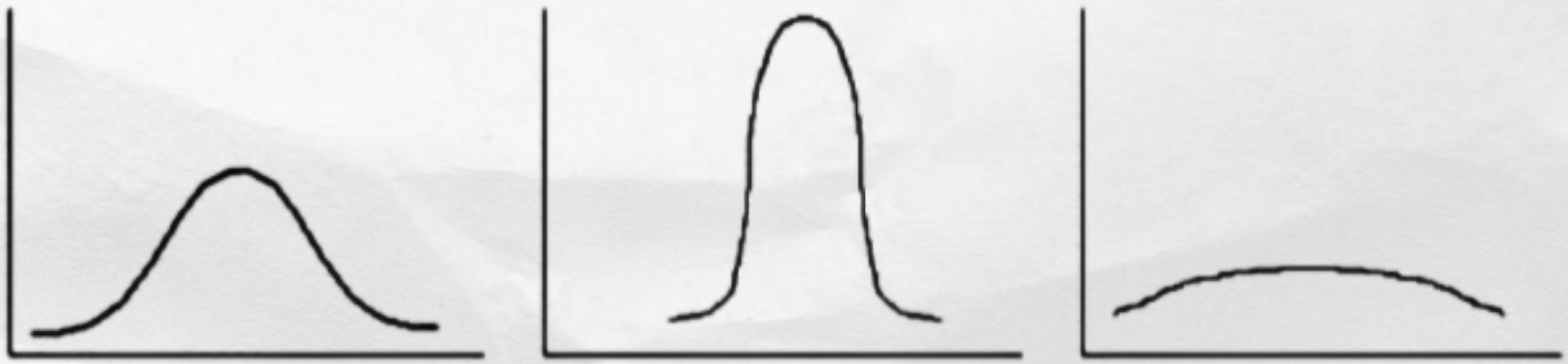
Let's take an example:



In the above example both the curves have the same mean, same standard deviation and same skewness but the difference lies in their tails.

- The red graph has a more flat tail than the yellow graph, which means there are more potential outliers or extreme values at the tail.
- So, kurtosis is the measure of the "tailedness",i.e., the extremity of the tail.
- It measures the heaviness of the tail, i.e., how many data points lie in the tail
- It also tells how flat the tail is for the distribution - a more flat tail signifies a larger presence of outliers.
- In simple terms, kurtosis tells us the amount of variance in a dataset due to the extreme values (i.e., values that are far from the mean) versus the amount of variance due to values that are closer to the mean.

There are three types of kurtosis:



Mesokurtic Curve | Leptokurtic Curve | Platykurtic Curve

These are standard graphs where mean, median and mode are assumed to lie in the center.
These graphs may vary depending on the dataset you are working at.

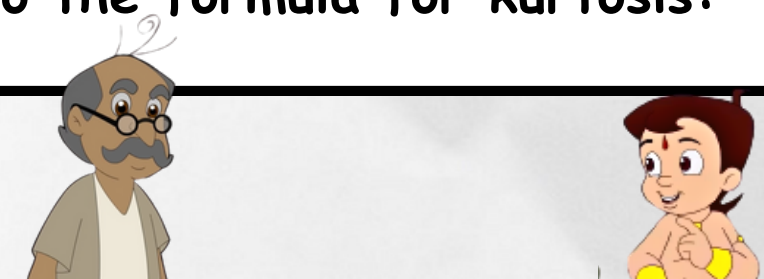
From these graphs, you only need to focus on the flatness of tails and the differences between them.

Before understanding this, let us look at the formula first:-

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{SD(x)}\right)^4$$

Here's an example dataset with 10 values:

5, 10, 15, 20, 25, 30, 35, 40, 45, 50
First, we need Mean = (5+10+15+20+25+30+35+40+45+50) / 10 = 27.5
Standard deviation = sqrt([(5-27.5)^2 + (10-27.5)^2 +...
+(50-27.5)^2] / 9) = 15.8
Now we can plug these values into the formula for kurtosis:
We got Kurtosis=1.63



Now, let's see the interpretation of the values:-

The interpretation of kurtosis values depends on the context of the data being analyzed.

In general, a kurtosis value of 3 is considered to be "normal" or "mesokurtic," meaning that the data follows a normal distribution with moderate peakedness and moderate tails.

Values greater than 3 indicate "leptokurtic" distributions with heavier tails and a sharper peak,

while values less than 3 indicate "platykurtic" distributions with flatter peaks and thinner tails.

Here are some guidelines for interpreting kurtosis values:

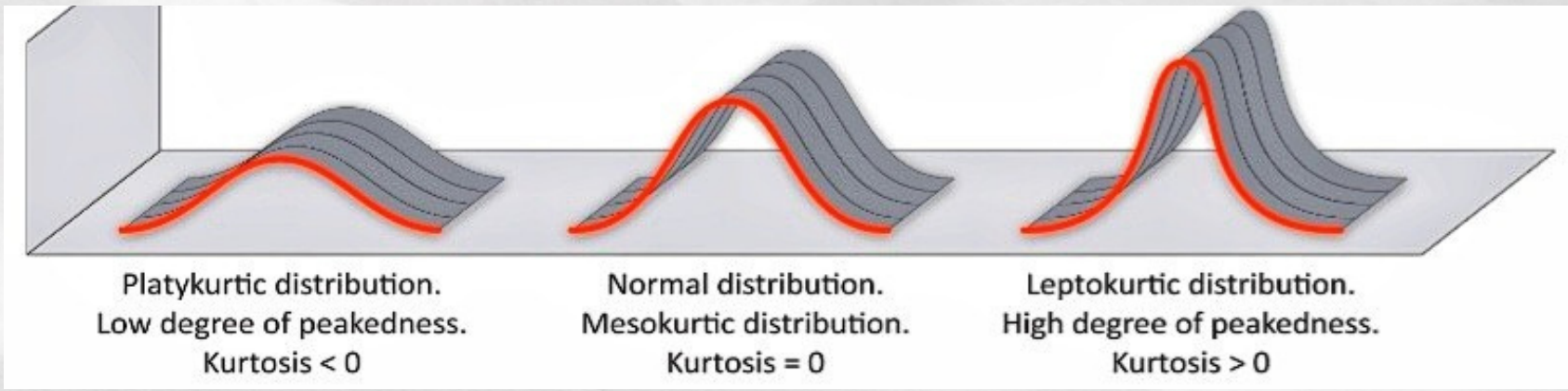
- Kurtosis values between 2 and 3 are considered moderately peaked, and values between 3 and 4 are considered moderately flat.

- Kurtosis values greater than 4 indicate a very peaked or "heavy-tailed" distribution, where extreme values are more common than in a normal distribution. This type of distribution is also called a "leptokurtic" distribution.

- Kurtosis values less than 2 indicate a very flat or "thin-tailed" distribution, where extreme values are less common than in a normal distribution. This type of distribution is also called a "platykurtic" distribution.

For the above dataset, we can interpret that:-

The kurtosis value of this dataset is 1.63, indicating that the dataset is platykurtic (i.e., has low kurtosis) and is flatter than a normal distribution. This means that there are fewer extreme values in the dataset compared to a normal distribution. Visually, the distribution looks flatter and more spread out than a normal distribution.



In some literature, the value of the Mesokurtic distribution, which is an ideal distribution, is stated as 3, while in others, it is stated as 0. This is due to the concept of Excess Kurtosis. To simplify the range, we typically subtract 3 from the final result. Therefore, if the value is 0, it is Mesokurtic; if it is less than 0, it is Platykurtic; and if it is greater than 0, it is Leptokurtic.

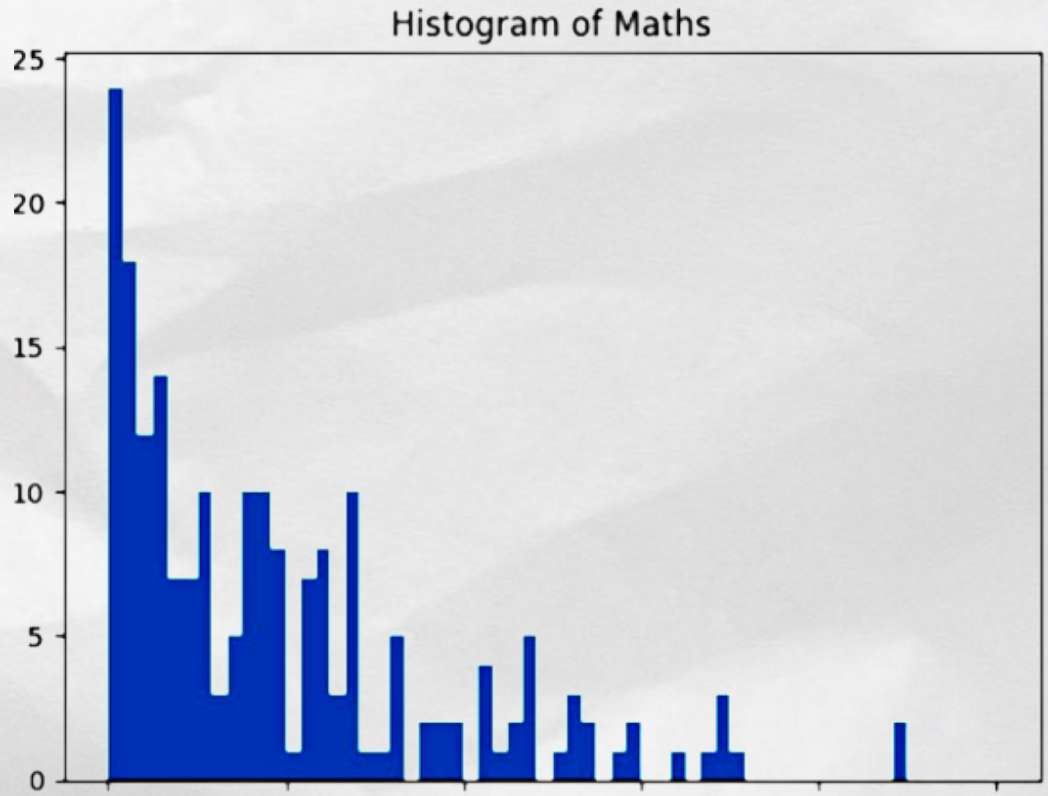
So, the above graph gives more clearer concept.

- In Platykurtic distribution, tail is not heavy or flat. It is even less than Normal Distribution, which implies lesser presence of extreme values or outliers ,
- But in Leptokurtic, it is more flat tail or more heavy tail, which means there's more presence of extreme values or outliers.

We have a data of 200 students with their scores in Maths

Roll no.	Marks
1	11
2	25
3	0
4	7
5	3
...	...
196	54
197	0
198	5
199	19
200	60

Following is the histogram for the data:-

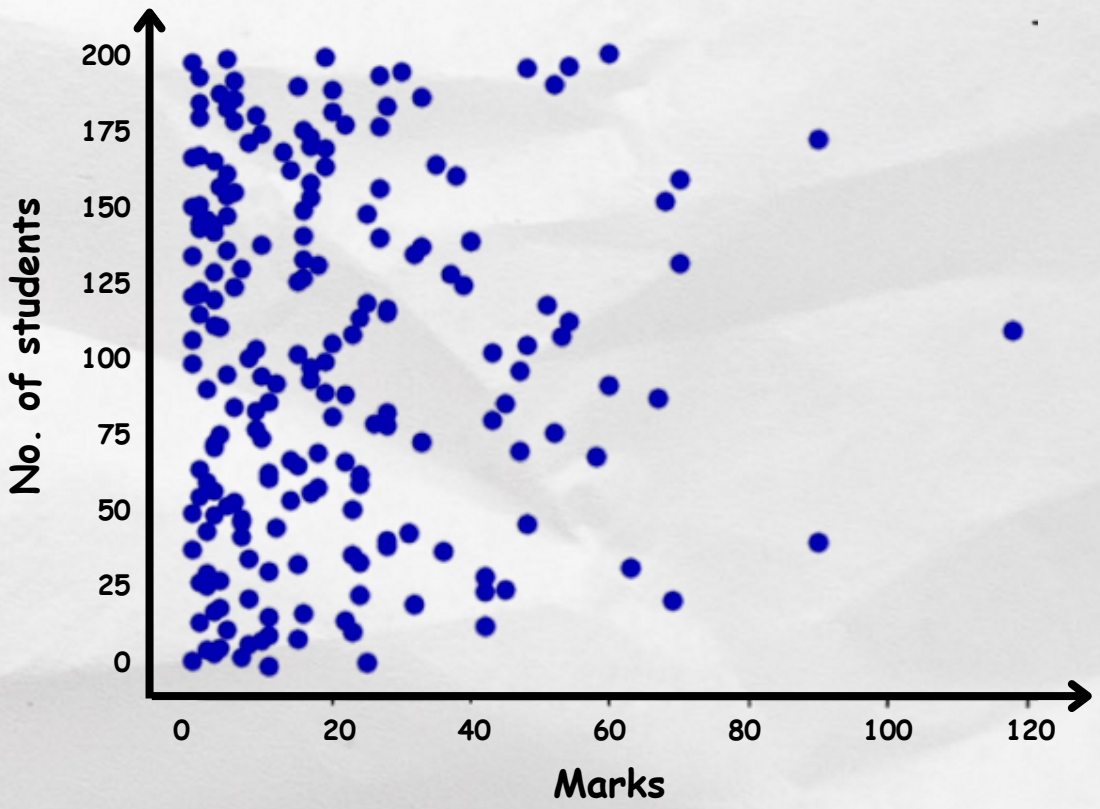


On this data, we calculated the skewness and kurtosis:-

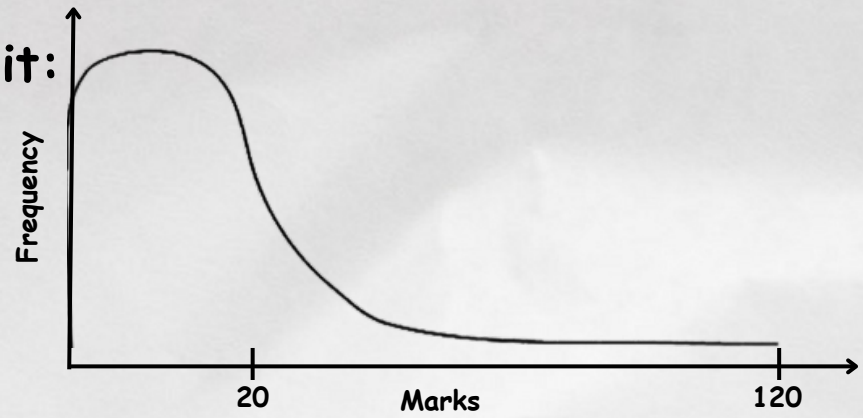
```
skewness = skew(dataset)
kurt = kurtosis(dataset)


print("Skewness:", skewness)
print("Kurtosis:", kurt)


Skewness: 1.672610910605827
Kurtosis: 3.508433814883932
```

Since the value of skewness is more than 1, we can say it is positively skewed. Also, the value of Kurtosis is more than 3, thus we can say it is a leptokurtic distribution.

Now, if we find a high kurtosis value, it means that the distribution of math scores has more extreme values. Having discovered the presence of extreme values, I created a scatter plot to examine how the data is dispersed.



On analyzing the plot, it was clear that the mean of the dataset is 19, hence some students have scored more than 50 and some students have performed extremely well, which explains the heaviness of the tail that we saw in kurtosis value.
So here's the graph for it:



Clearly, you can see the tail is heavy and flat, which means some students have done excellent in exams and have scored extremely higher than 19 (which is my mean)

It became clear that all students scoring less than the mean value, i.e. 19, should enroll in the special classes.

If you enjoyed the content, please support us by sharing it in your network. It takes a lot of effort to create such original content. Tag people who may find this useful and leave your questions and feedback in the comments.

Content Strategist : Hanit Kaur
Graphic Designer : Adithya Prasad
Content Lead : Sumit Shukla

DON'T FORGET TO

SAVE

Follow for more

Session
With
Sumit