Published in Towards Data Science

Rizwan Alam Follow

May 5, 2020 · 5 min read ★ · Listen

Normalization vs Standardization Explained

Stop using them interchangeably!

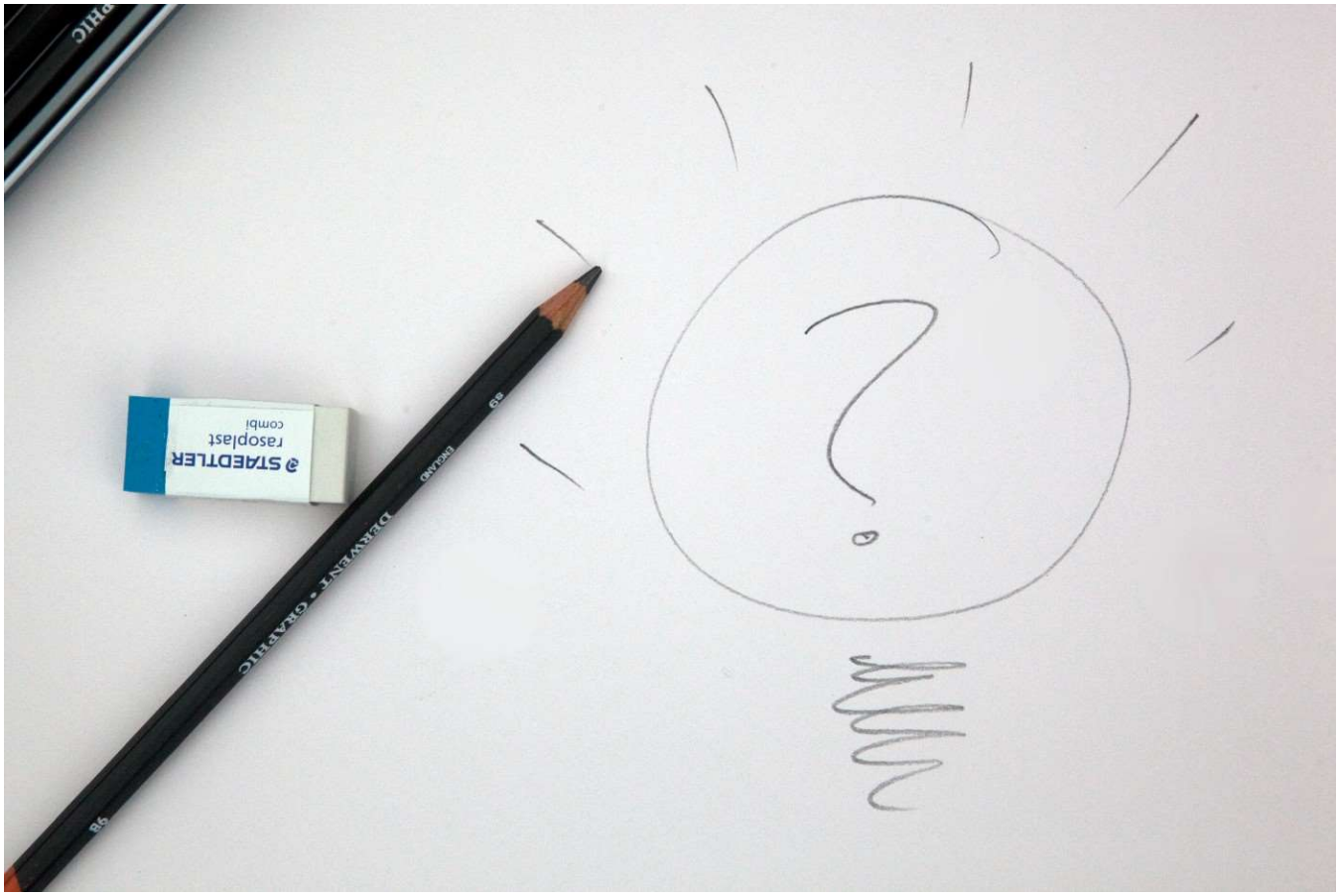


Photo by Mark Fletcher-Brown on Unsplash

The term normalization and standardization are used a lot in statistics and data science. We sometimes use them interchangeably. People usually get confused between these two terms. But there is a subtle difference between these two. And that's where

using much of technical ng examples, and the difference between them. So bear with me for five minutes it will be worth your time.

What is Normalization?

It is a scaling technique method in which data points are shifted and rescaled so that they end up in a range of 0 to 1. It is also known as **min-max scaling**.

The formula for calculating normalized score:

X new = (X — X min)/ (X max — X min)

Here, Xmax and Xmin are the maximum and minimum values of the feature respectively.

· If X=Xmin; then Xnew =0

Since numerator will become Xmin –Xmin, which is nothing but 0.

· If X=Xmax ; then Xnew =1

In this case, both numerator and denominator will be equal and cancel each other to give us the value of Xnew =1.

It's too complicated, isn't it? Let's take an **example** and clear this out.

In CAT (an aptitude test conducted by IIM's for the selection in top B-schools in India)too many applications are received. So they can't examine all candidates at the same time, therefore the exam is conducted in shifts or even on different days. In different shifts, the question paper set is different. Although the questions are set in such a way that the difficulty level of each shift remains the same but still there might be a possibility that difficulty level of shift varies. So it will be unfair to the candidates who got a difficult set of questions. To make it fair for all candidates the score of candidates is normalized.

Let's say the exam is conducted in two shifts **shift A** and **shift B** and questions in shift A were relatively easy as compared to that of shift B. Because questions were relatively



So we cannot compare t in shift A to a candidate who scored the same in shift B.

Hence we normalize the scores

The normalized score of a candidate who scored 150 in shift A will be calculated as follows

For simplicity sake let's name it Xa

Xa =150-Xmin/(Xmin-Xmax)

Xmax = 280

Xmin = 80

Putting these values we get

Xa= 150–80/(280–80)

Xa =0.35

The normalized score of a candidate who scored 150 in shift B will be

For simplicity sake let's name it Xb

Xb=150-Xmin/(Xmin-Xmax)

Xmax = 250

Xmin = 50

Putting these values we get

Xb = 150–50/(250–50)

Xb =0.5

Here we can see Xb>Xa



Xa *300 =105

Xb *300 = 150

I think now it will be clearer to understand what normalization is and why we need it.

All right, let's jump to the second one

What is Standardization?

Standardization is another scaling method where the values are centered around mean with a unit standard deviation. It means if we will calculate mean and standard deviation of standard scores it will be 0 and 1 respectively.

The formula for standardized values:

$$z = \frac{x - \mu}{\sigma}$$

Where,

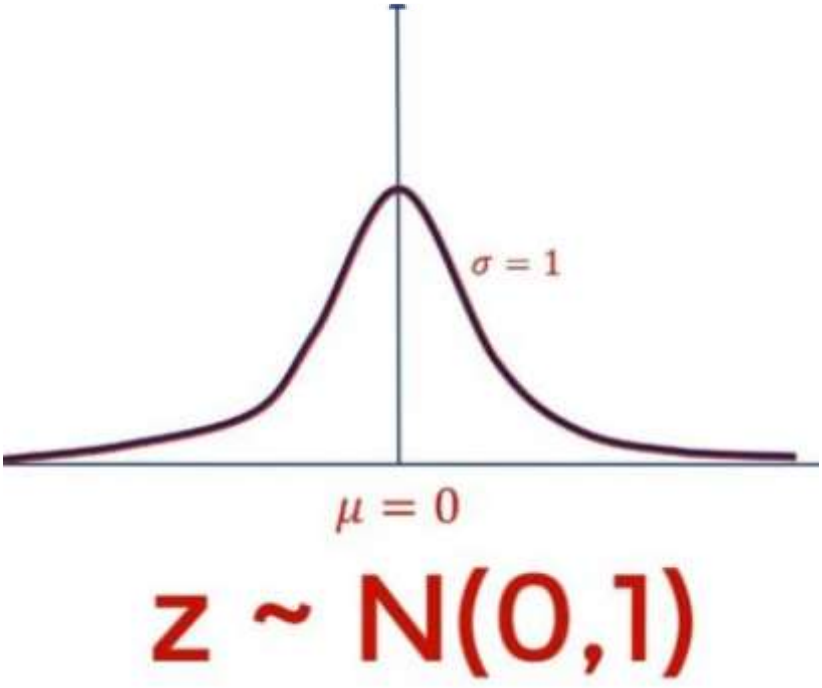
μ= mean of the given distribution

σ = standard deviation of the given distribution

This Z is called standard score and it represents the number of standard deviations above or below the mean that a specific observation falls.

i.e. If Z=2, it means that the observation lies two standard deviations above the mean.

If we plot these standard scores it will be a normal distribution with mean at 0 and the standard deviation equals to 1.



https://365datascience.com/wp-content/uploads/2018/10/image4-9.jpg

The standard deviation with mean = 0 and σ =1 is also known as standard normal distribution and is denoted by N(0,1).

Getting too technical, isn't it?

Let's solve an **example** for better understanding.

Let's assume you and your friend study in different universities where the grading system is different. You get your score of 85 in a test. The mean grade of the class is 75 and a standard deviation of 5. Your friend got a grade of 615 and the mean grade of the class is 600 with a standard deviation of 50. How you are going to say who is performing better? As grade 85 can't be compared to 615.

Here comes the role of standardization as it allows us to compare the scores with different metrics directly and make a statement about them.

Your Z score = 85–75/5 = 2

It means you are 2 standard deviations above the average grade.

By looking at standard s erforming much better that he or she is the class.

Wasn't that easy? I promised you.

Now the big question arises

Which is better normalization or standardization?

Well, that depends on the type of data you are using.

Normalization is preferred over standardization when our data doesn't follow a normal distribution. It can be useful in those machine learning algorithms that do not assume any distribution of data like the k-nearest neighbor and neural networks.

Standardization is good to use when our data follows a normal distribution. It can be used in a machine learning algorithm where we make assumptions about the distribution of data like linear regression etc

Point to be noted that unlike normalization, standardization doesn't have a bounding range i.e. 0 to 1.

It's also not influenced by maximum and minimum values in our data so if our data contains outliers it's good to go.

Final words:

I hope you got a good idea about normalization and standardization. If you like my work then do appreciate me for posting more such good content related to statistics and data science. Please share it on social media platforms.