

★ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



Understanding KL Divergence Intuitively

A constructive approach to measuring distribution differences.



[Mohammed Mohammed](#) · Follow

Published in Towards Data Science

6 min read · Feb 13, 2024

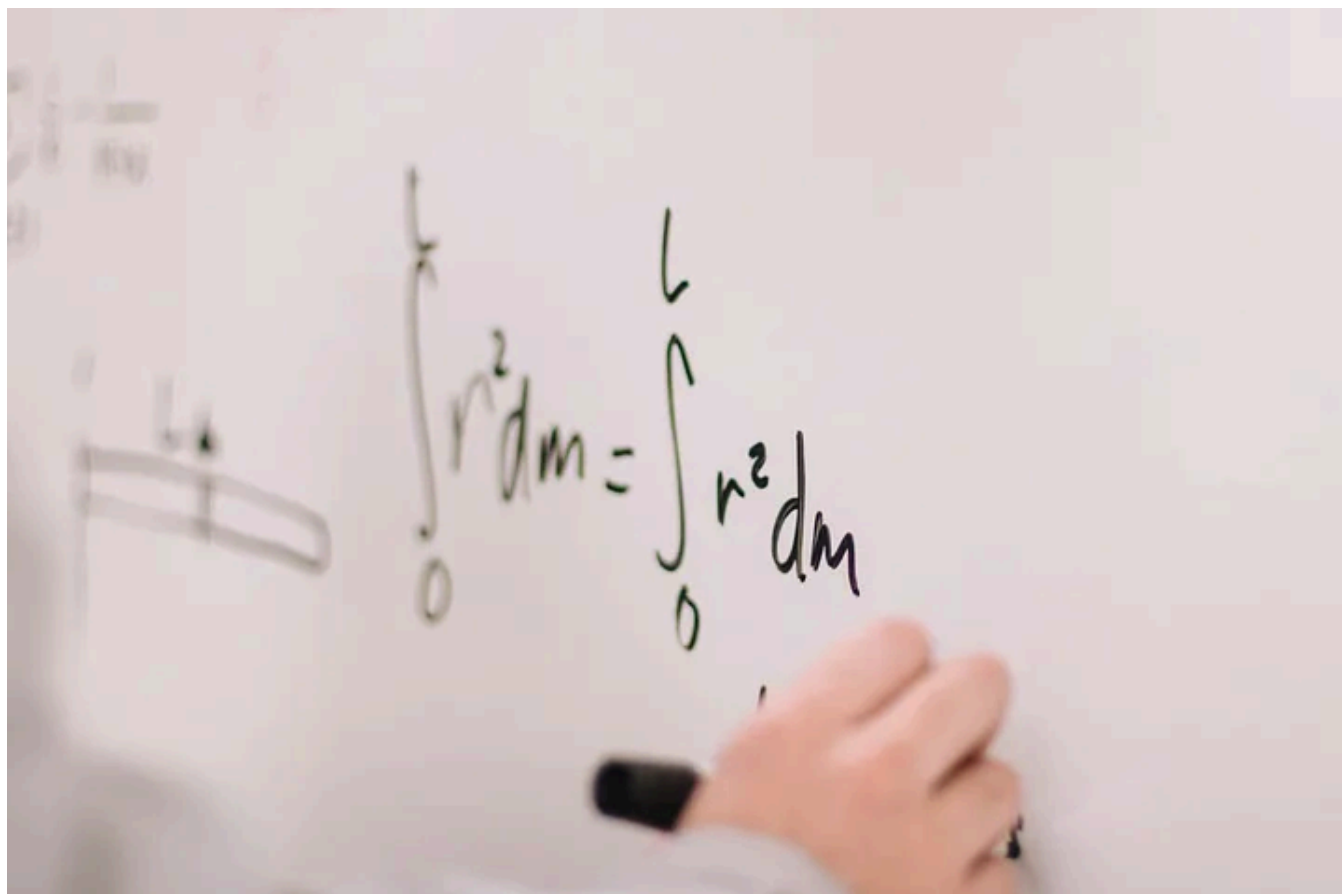


Photo by [Jeswin Thomas](#) on [Unsplash](#)

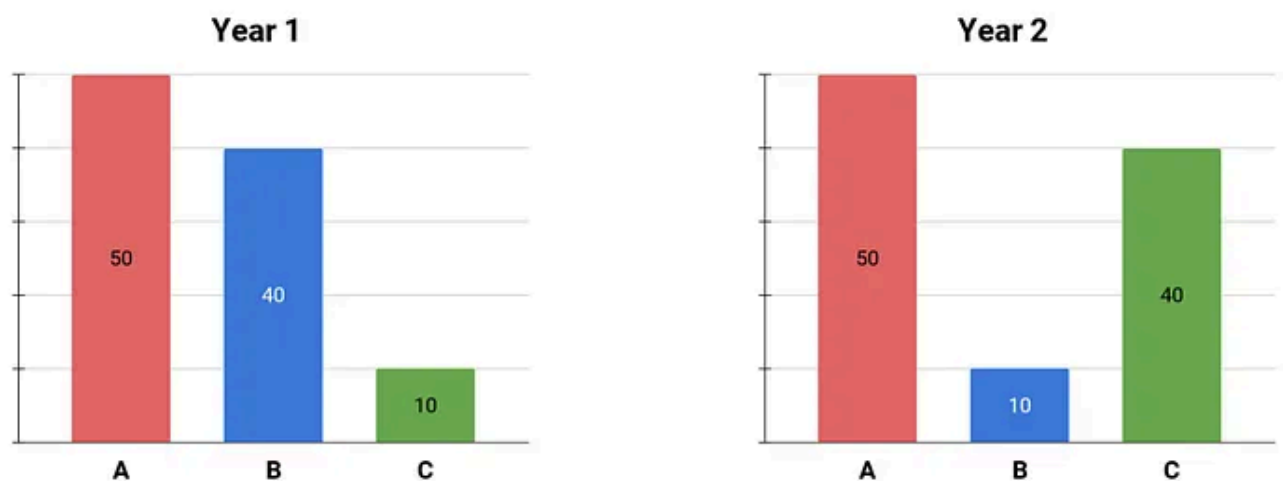
Today, we will be discussing KL divergence, a very popular metric used in data science to measure the difference between two distributions. But before delving into

the technicalities, let's address a common barrier to understanding math and statistics.

Often, the challenge lies in the approach. Many perceive these subjects as a collection of formulas presented as divine truths, leaving learners struggling to interpret their meanings. Take the KL Divergence formula, for instance — it can seem intimidating at first glance, leading to frustration and a sense of defeat. However, this isn't how mathematics evolved in the real world. Every formula we encounter is a product of human ingenuity, crafted to solve specific problems.

In this article, we'll adopt a different perspective, treating math as a creative process. Instead of starting with formulas, we'll begin with problems, asking: *"What problem do we need to solve, and how can we develop a metric to address it?"* This shift in approach can offer a more intuitive understanding of concepts like KL Divergence.

Enough theory — let's tackle KL Divergence head-on. Imagine you're a kindergarten teacher, annually surveying students about their favorite fruit, they can choose either apple, banana, or cantaloupe. You poll all of your students in your class year after year, you get the percentages and you draw them on these plots.



Consider two consecutive years: in year one, 50% preferred apples, 40% favored bananas, and 10% chose cantaloupe. In year two, the apple preference remained at 50%, but the distribution shifted — now, 10% preferred bananas, and 40% favored

cantaloupe. The question we want to answer is: how different is the distribution in year two compared to year one?

Even before diving into math, we recognize a crucial criterion for our metric. Since we seek to measure the disparity between the two distributions, our metric (which we'll later define as KL Divergence) must be **asymmetric**. In other words, swapping the distributions should yield different results, reflecting the distinct reference points in each scenario.

Now let's get into this construction process. If we were tasked with devising this metric, how would we begin? One approach would be to focus on the elements — let's call them A, B, and C — within each distribution and measure the ratio between their probabilities across the two years. In this discussion, we'll denote the distributions as P and Q, with Q representing the reference distribution (year one).

For instance, P(a) represents the proportion of year two students who liked apples (50%), and Q(a) represents the proportion of year one students with the same preference (also 50%). When we divide these values, we obtain 1, indicating no change in the proportion of apple preferences from year to year. Similarly, we calculate $P(b)/Q(b) = 1/4$, signifying a decrease in banana preferences, and $P(c)/Q(c) = 4$, indicating a fourfold increase in cantaloupe preferences from year one to year two.

That's a good first step. In the interest of just keeping things simple in mathematics, what if we averaged these three ratios? Each ratio reflects a change between elements in our distributions. By adding them and dividing by three, we arrive at a preliminary metric:

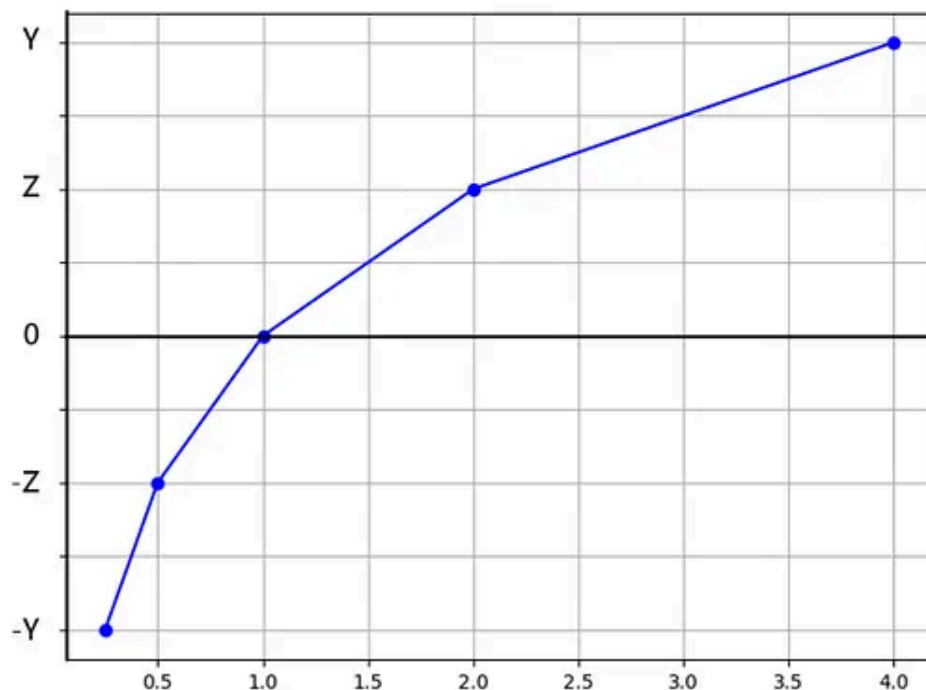
$$\frac{(1 + 1/4 + 4)}{3} = 1.75$$

This metric provides an indication of the difference between the two distributions. However, let's address a flaw introduced by this method. We know that averages can be skewed by large numbers. In our case, the ratios $1/4$ and 4 represent opposing yet equal influences. However, when averaged, the influence of 4 dominates, potentially inflating our metric. Thus, a simple average might not be the ideal solution.

To rectify this, let's explore a transformation. Can we find a function, denoted as F , to apply to these ratios (1, $\frac{1}{4}$, 4) that satisfies the requirement of treating opposing influences equally? We seek a function where, if we input 4, we obtain a certain value (y), and if we input $\frac{1}{4}$, we get ($-y$). To know this function we're simply going to map values of the function and we'll see what kind of function we know about could fit that shape.

Suppose $F(4) = y$ and $F(\frac{1}{4}) = -y$. This property isn't unique to the numbers 4 and $\frac{1}{4}$; it holds for any pair of reciprocal numbers. For instance, if $F(2) = z$, then $F(\frac{1}{2}) = -z$. Adding another point, $F(1) = F(1/1) = x$, we find that x should equal 0.

Plotting these points, we observe a distinctive pattern emerge:



I'm sure many of us would agree that the general shape resembles a logarithmic curve, suggesting that we can use $\log(x)$ as our function F . Instead of simply calculating $P(x)/Q(x)$, we'll apply a log transformation, resulting in $\log(P(x)/Q(x))$. This transformation helps eliminate the issue of large numbers skewing averages. If we sum the log transformations for the three fruits and take the average, it would look like this:

$$\frac{\sum_x \log\left(\frac{P(x)}{Q(x)}\right)}{n}$$

What if this was our metric, is there any issue with that?

One possible concern is that we want our metric to prioritize popular x values in our current distribution. In simpler terms, if in year two, 50 students like apples, 10 like bananas, and 40 like cantaloupe, we should weigh changes in apples and cantaloupe more heavily than changes in bananas because only 10 students care about them, therefore it won't affect the current population anyway.

Currently, the weight we're assigning to each change is $1/n$, where n represents the total number of elements.

$$\sum_x \frac{1}{n} \log\left(\frac{P(x)}{Q(x)}\right)$$

Instead of this equal weighting, let's use a probabilistic weighting based on the proportion of students that like a particular fruit in the current distribution, denoted by $P(x)$.

$$\sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

The only change I have made is replaced the equal weighting on each of these items we care about with a probabilistic weighting where we care about it as much as its frequency in the current distribution, things that are very popular get a lot of priority, things that are not popular right now (even if they were popular in the past distribution) do not contribute as much to this KL Divergence.

This formula represents the accepted definition of the KL Divergence. The notation often appears as $KL(P||Q)$, indicating how much P has changed relative to Q .

$$KL(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Now remember we wanted our metric to be asymmetric. Did we satisfy that? Switching P and Q in the formula yields different results, aligning with our requirement for an asymmetric metric.

Summary

Firstly I do hope you understand the KL Divergence here but more importantly I hope it wasn't as scary as if we started from the formula on the very first and then we tried our best to kind of understand why it looked the way it does.

Other things I would say here is that this is the discrete form of the KL Divergence, suitable for discrete categories like the ones we've discussed. For continuous distributions, the principle remains the same, except we replace the sum with an integral (\int).

NOTE: Unless otherwise noted, all images are by the author.

Probability

Statistics

Mathematics

Math

KL Divergence

