

# Kullback-Leibler Divergence

Diptarko Choudhury

# Probability Distribution

A probability distribution is a statistical function that describes all the possible values and probabilities for a random variable within a given range.

The probability distribution is divided into two parts:

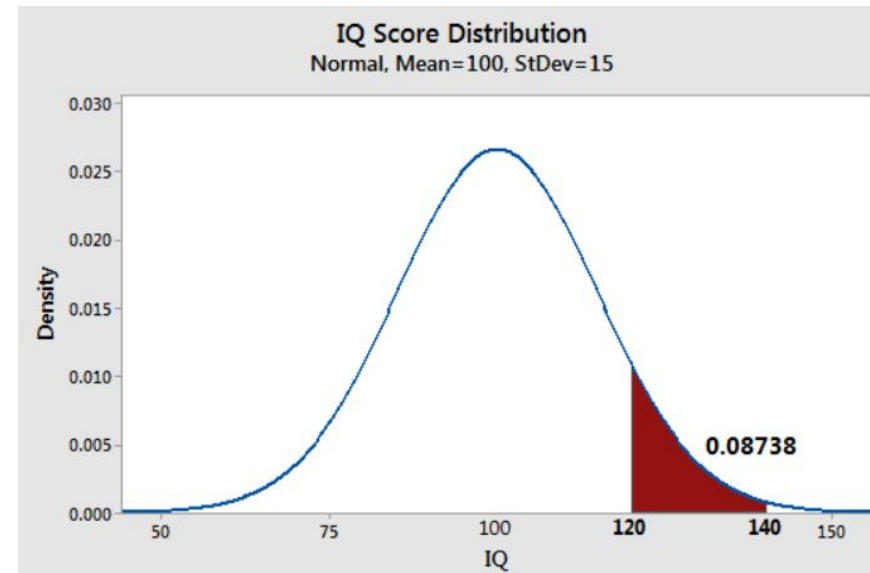
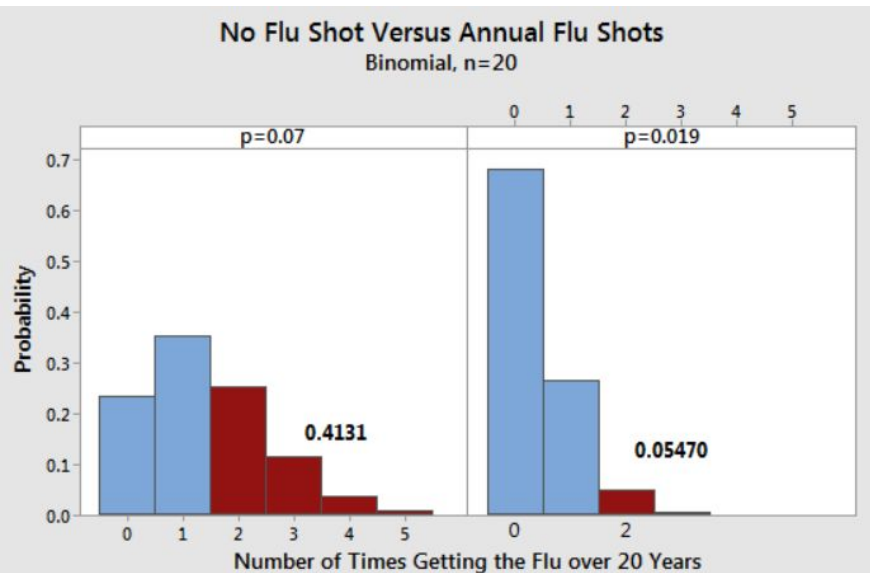
- Discrete Probability Distributions
- Continuous Probability Distributions



# Discrete vs. Continuous Distributions

A discrete probability distribution can assume a discrete number of values.

While continuous distribution assume infinite number of values.



# Continuum Limit and Discretisation

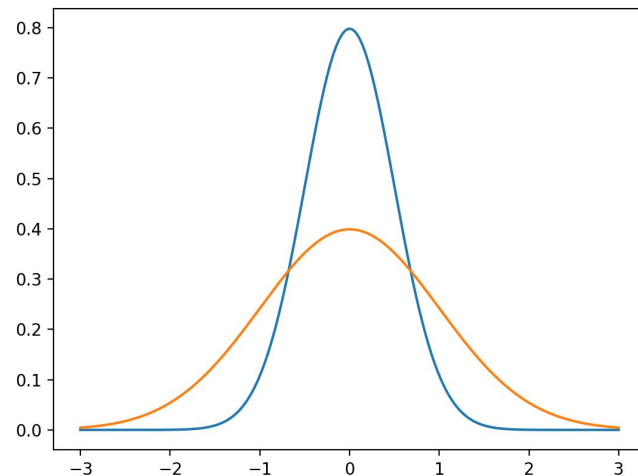
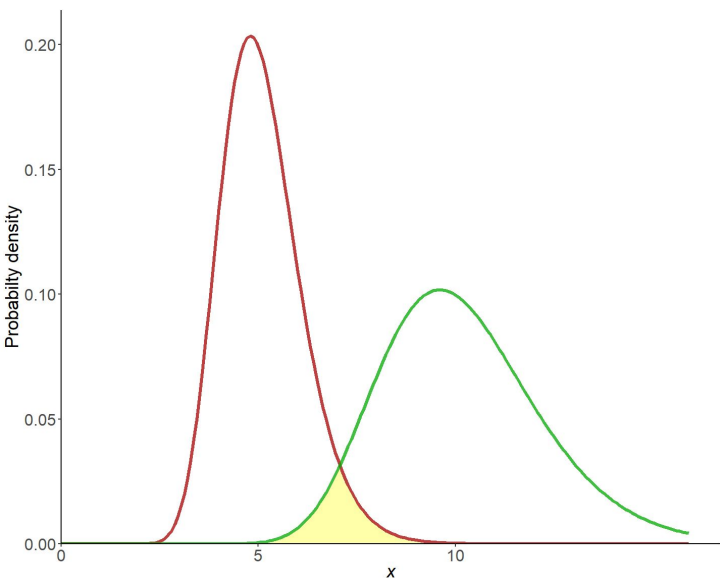
A discrete probability can often be considered a continuous distribution if the number of values the probability function can assume tends to infinity.

A continuous probability distribution can often be converted to a discrete distribution if we discretise or bin the possible outcomes.



# Similarity between two Probability Distributions

Statement: We have two probability distributions  $P(x)$  and  $Q(x)$ . How to quantify the similarity between the two ?



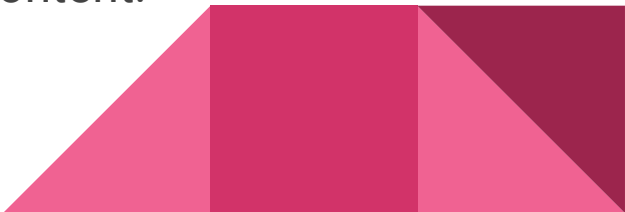
# Entropy

Entropy is a measure of uncertainty or randomness in a set of data.

The information content of an event is inversely proportional to the probability of the event. We can define the information content  $I(x)$  of an event  $x$  with probability  $P(x)$  as follows:

$$I(x) = -\log(P(x))$$

The entropy  $H(X)$  of a discrete random variable  $X$  with probability mass function  $P(x)$  is defined as the expected value of the information content:

$$H(X) = E[I(x)] = \sum [P(x) * I(x)]$$


# Entropy

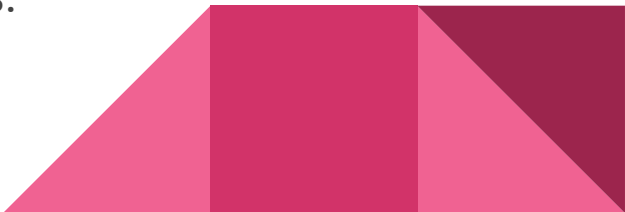
Substituting the formula for information content we get

$$H(X) = \sum [P(x) * I(x)] = \sum [P(x) * -\log(P(x))]$$

This can be rewritten as

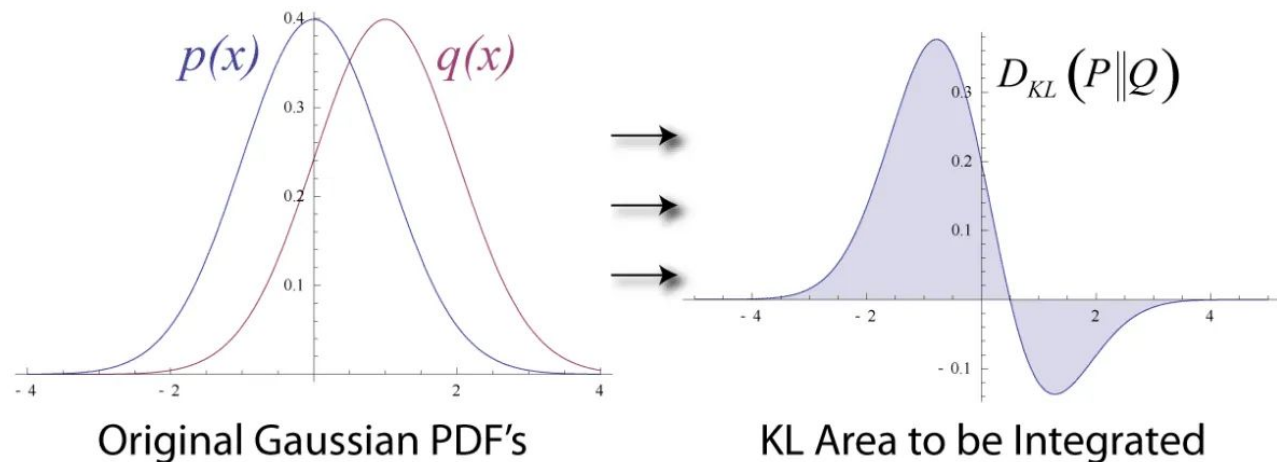
$$H(X) = - \sum [P(x) * \log(P(x))]$$

$H(X)$  is the information content in the probability distribution, now since we have the amount of information contained in a single distribution we can start looking ways to quantify the difference between two distributions.



# KL Divergence

To solve this specific problem we use KL Divergence. Kullback–Leibler divergence or KL Divergence denoted by  $D_{KL}(P||Q)$  is a type of statistical distance: a measure of how one probability distribution  $P$  is different from a second, reference probability distribution  $Q$ .





# KL Divergence

The formula for KL Divergence for discrete probability distributions  $p(x)$  and  $q(x)$  is defined as:

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

The above equation can be further extended to continuous systems by using the continuum limit described before.

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Both  $p(x)$  and  $q(x)$  sum up to 1, and  $p(x) > 0$  and  $q(x) > 0$



# Metrics

A metric on a set  $X$  is a function (called the distance function or simply distance)  $d : X \times X \rightarrow \mathbb{R}^+$  (where  $\mathbb{R}^+$  is the set of non-negative real numbers). For all  $x, y, z$  in  $X$ , this function is required to satisfy the following conditions:

- $d(x, y) \geq 0$
- $d(x, y) = 0$  if and only if  $x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$



# KL Divergence

KL Divergence measures the “distance” between two distributions, it is not a metric due to the following reasons:-

- KL Divergence is not symmetric: KL from  $p(x)$  to  $q(x)$  is generally not the same as the KL from  $q(x)$  to  $p(x)$ .
- KL Divergence need not satisfy the triangular inequality.

Still KL Divergence is a positive quantity and is only zero if both the distributions are absolutely similar.



# Limitations of KL Divergence

One of the biggest drawbacks of KL Divergence is that it is not a metric. Although it is relatively simple to calculate and easy to understand there are multiple occasions where it fails. The following slides describe some of the situations where KL Divergence fails.



# Norm condition

For KL Divergence to work well both the probability distributions  $p(x)$  and  $q(x)$  must be normalised to one. In other words the sum of all the probability distributions must be equal to 1. In many real life distribution this condition is violated the norm is not always work making task of interpreting the KL Divergence difficult.



## Zero values blow up.

We can say that  $\lim_{p \rightarrow 0} p \log p = 0$ . But when  $p \neq 0, q = 0$ ; we define  $D_{\text{KL}}(p||q)$  is defined as  $\infty$ .

This means that if one event  $e$  is possible (i.e.,  $p(e) > 0$ ), and the other predicts it is absolutely impossible (i.e.,  $q(e) = 0$ ), then the two distributions are absolutely different.

In a dataset dominated by large number events whose probability of occurrence is non-zero this statement seems improbable.



# Asymmetry

KL Divergence for the probability distribution  $p(x)$  and  $q(x)$  is defined as

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

The problem with the above definition is that if we switch  $p(x)$  and  $q(x)$  the overall problem statement remains the same but  $D_{KL}(p||q)$  is not equal to  $D_{KL}(q||p)$ .



# Vectorised Implementation in Python

```
def kl_divergence(p, q):  
    return np.sum(np.where(p != 0, p * np.log(p / q), 0))
```





Questions ?

