

Basic Statistics for Machine Learning

We are going to cover the basics of statistics that are required for machine learning procedures. Why are statistics required for machine learning is primarily as we need, measurable techniques in the planning of training and testing for our machine learning algorithms, meaning that we need to be able to mathematically validate if we are able to get an algorithm up and running and ready for its implementation in the real world.

We are going to cover the following components (Not necessarily in the same order) in our endeavor to understand statistics specifically for their implementation in machine learning.

- Data scaling
- Data sampling
- Missing value imputation
- Outlier detection
- Central Tendency
- Data Distributions (primarily Normal Distributions)

Now, how important are Statistics?

Here are some of the excerpts from a [news article](#) related to the recent Coronavirus pandemic:

On **December 30, 2019**, **BlueDot**, a Toronto-based startup that uses a platform built around artificial intelligence, machine learning, and big data to track and predict the outbreak and spread of infectious diseases, alerted its private sector and government clients about a cluster of “unusual pneumonia” cases happening around a market in Wuhan, China.

That was the first recognition of the novel coronavirus that has come to be known as COVID-19. It would be another nine days before the World Health Organization released its statement alerting people to the emergence of a novel coronavirus.

So how were they able to do this

The simpler answer will be using statistics, and specifically using one of the above-mentioned techniques that we are going to cover in detail, **Outlier detection**.

