EXAFLUENCE
Data Driven Influence

Central Limit Theorem – The Undervalued Hero of Machine Learning

Kumaravel C

June 2023

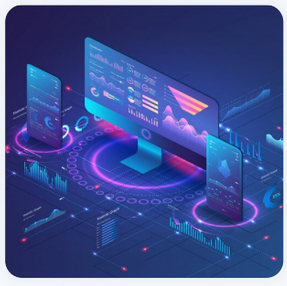
Introduction:

At Exafluence, we not only use ML algorithms to train models but also appreciate the principles of the statistical world that work under the hood. Be it plotting a “multiple boxplot” chart between a categorical and a continuous variable, or training a multi-class classification ML model, we have taken pride in understanding in depth why we do what we do. In this brief text that follows, we would like to felicitate one of the most undervalued heroes of the ML world. The suite of analytics capabilities at built at Exafluence (called ExfAnalytics) - Dashboards, ML/AI models, ExfSpectrum and ExfGenAI are testaments of the Central Limit Theorem at work.

Why is descriptive statistics essential for a data scientist?

Both Descriptive and Inferential statistics form part of applied statistics. In descriptive statistics, we summarize each numeric column of a dataset by calculating the count of entries, average, variance, quartiles, minimum and maximum of the column. This gives us information about the central tendency, spread and variability of the column. Similarly, we summarize a categorical column by listing out the unique values it has along with their frequency. This data can be represented in numeric forms or in the form of a table or a chart (called visualization). In simple terms, we are "describing" the dataset to help us understand what kind of data we are dealing with for our analysis.

Why is inferential statistics necessary to understand for a data scientist?



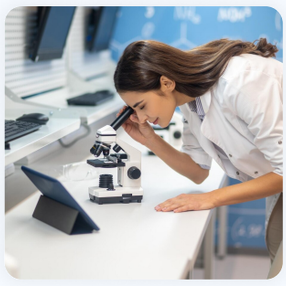
In inferential statistics, we take a sample (random or stratified) and make inferences about the population using our findings from the sample. A population is the universal set of data satisfying the conditions we are interested in. For the obvious reason of size, working on a sample is more convenient than working on the whole population. Moreover, we might not always have the access to a population of data.

How are Central Limit Theorem and ML related?

The use of descriptive statistics in Machine Learning (ML) is not hard to understand. Every data scientist carries out descriptive statistical activities to get a strong hold on the data before proceeding to train ML models. What we usually forget is that, quite often we build our models on a sample of data and interpret the results to arrive at recommendations for business teams. We are able to extend these findings from a "sample of the population" to the "population" because of the validity of inferential statistics. This validity is established through the Central Limit Theorem (CLT) which is the cornerstone of Machine Learning. As per the CLT, the distribution of means of samples (called sampling distribution) is approximately normal. The larger the size of the samples, the more normal and narrower the sampling distribution becomes. This holds true irrespective of the distribution of the same variable in the original population. The sample size should be greater than 30 records.

What happens if the distribution of the variable in the population is not normal?

The real power of the CLT lies in the fact that we don't need to know the distribution of the population in advance. It need not be normal. An ML expert trains a model on a sample dataset and uses it to make predictions for the whole population. The whole population is nothing but any combination of the input variables we provide our model with in the future for carrying out a prediction. As simple as that. Probably as elusive too as that.



How does CLT help in hypothesis tests?

Non parametric hypothesis tests don't need the data to be normally distributed. They often work on medians as the central tendency. On the other hand, parametric hypothesis tests of means require the data to be normally distributed. We are good because we have our friendly Central Limit Theorem. Examples of such tests are t-tests, one way ANOVA, etc. Can you guess which scenarios have a median working better as a central tendency representation than a mean? Uff, the outliers never leave us in peace! However, if the sample size is considerably big, we would rather go for one of the parametric tests.

Conclusion

Next time you build an ML model and use it to do a prediction that improves sales by 10%, you know who to thank. The CLT!