The Theorem Every Data Scientist Should Know

04 Jul 2016

Yesterday, I was reading a thread on Quora. The people in this thread where answering the following question: `What are 20 questions to detect fake data scientists?` . The most upvoted answer contained a list of questions that could leave a good number of data scientists off guard.

In that thread, my attention was drawn to one particular question. Not because it was specifically hard but because I doubt many data scientists can answer that question. Yet, most of them, whether they know it or not, are using this concept on a daily basis.

The question was: ***What is the Central Limit Theorem? Why is it important?***

Explain the Theorem Like I'm Five

Let's say you are studying the population of beer drinkers in the US. You'd like to understand the mean age of those people but you don't have time to survey the entire US population.

Instead of surveying the whole population, you collect one sample of 100 beer drinkers in the US. With this data, you are able to calculate an arithmetic mean. Maybe for this sample, the mean age is 35 years old. Say you collect another sample of 100 beer drinkers. For that new sample, the mean age is 39 years old. As you collect more and more means of those samples of 100 beer drinkers, you get what is called a sampling distribution. The sampling distribution is the distribution of the samples mean. In this example, 35 and 39 would be two observations in that sampling distribution.

The statement of the theorem says that the sampling distribution, the distribution of the samples mean you

collected, will approximately take the shape of a bell curve around the population mean. This shape is also known as a normal distribution. Don't get the statement wrong. The CLT is not saying that any population will have a normal distribution. It says the *sampling distribution* will.

As your samples get bigger, the sampling distribution will tend to look more and more like a normal distribution. The Theorem holds true for any populations, regardless of their distribution*. There are some important conditions for the Theorem to hold true but I won't cover them in this post.

Why is it important?

The Central Limit Theorem is at the core of what every data scientist does daily: make statistical inferences about data.

The theorem gives us the ability to quantify the likelihood that our sample will deviate from the population without having to take any new sample to compare it with. We don't need the characteristics about the whole population to understand the likelihood of our sample being representative of it.

The concepts of confidence interval and hypothesis testing are based on the CLT. By knowing that our sample mean will fit somewhere in a normal distribution, we know that 68 percent of the observations lie within one standard deviation from the population mean, 95 percent will lie within two standard deviations and so on.

The CLT is not limited to making inferences from a sample about a population. There are four kinds of inferences we can make based on the CLT

1. We have the information of a **valid sample**. We can make accurate assumptions about it's population.
2. We have the information of the **population**. We can make accurate assumptions about a valid sample from that population.
3. We have the information of a **population and a valid sample**. We can accurately infer if the sample was drawn from that population.
4. We have the information about **two different valid samples**. We can accurately infer if the two samples

where drawn from the same population.

As a data scientist, you should be able to deeply understand this theorem. You should be able to explain it and understand why it's so important. This post skips many important aspects of the theorems such as it's mathematical demonstration, the criteria for it to be valid and the details about the statistical inferences that can be made from it. These elements are material for another post.

Related Posts

Getting your first job in data science 12 Aug 2016

What I Wish I Knew About Data For Startups 04 Aug 2016

The Theorem Every Data Scientist Should Know (Part 2) 13 Jul 2016