Central Limit Theorem: Data Scientist's Secret Weapon (Demo Included)

Arun Prakash Asokan · Follow
5 min read · Feb 11

Hey there! I welcome you to my next article in the Statistical Symphony Series. Do check out my previous articles in the series. It's an attempt at **simplifying Statistics** for everyone ! Let's deep dive into Central Limit Theorem (CLT), one of the most fundamental and important theorems in all of statistics.

Have you ever wondered how scientists can draw conclusions about a big group by only looking at a small portion of it? The secret is in the Central Limit Theorem (CLT).

> "the average of the averages is the average"?

That's the central idea of the Central specify Theorem (CLT).

Let's simplify the central limit theorem. Imagine you have a large bag of marbles and you want to determine the average size of all the marbles.

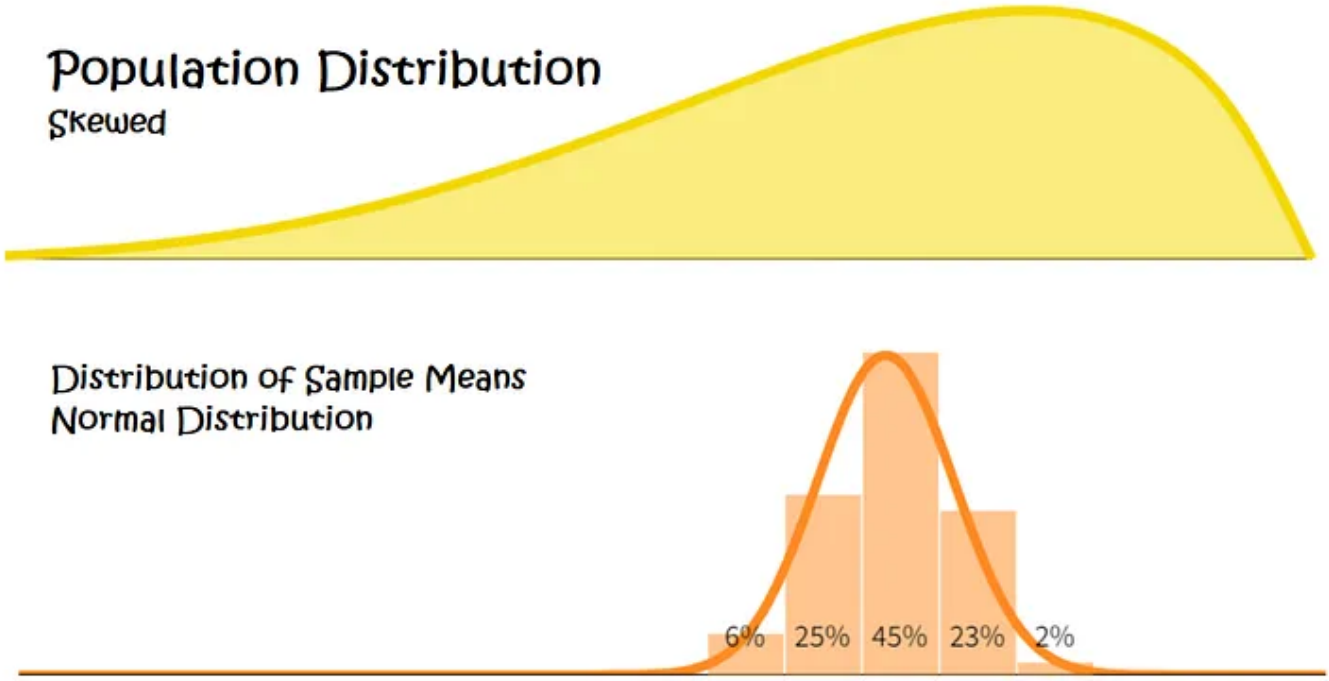
Bag of Marbles in different sizes and colors

Rather than measuring every single marble (which would take a long time), you randomly select a handful of marbles, measure their sizes, and calculate the average. You repeat this process a few times to get multiple averages.

Here's where the magic of the Central Limit Theorem comes in.

> The central limit theorem states that the sampling distribution of a sample mean is approximately normal if the sample size is large enough, even if the population distribution is not normal.

Even if the marbles in each handful aren't representative of the entire bag, the average of the handfuls will still provide a very good estimate of the average size of all the marbles. And the more handfuls you measure, the better your estimate will become.



The reason this works is because the CLT states that as the number of samples (i.e., handful of marbles) increases, the distribution of the sample means (i.e., the average size of each handful) approaches a normal distribution, even if the original distribution of marble sizes is not normal. This is why the average of the averages provides a good estimate of the average size of all the marbles.

A Demo of CLT

The video shows a simulation to showcase the power of Central Limit Theorem. The idea is to generate 20 random numbers from 0 to 9. Find the average. Repeat it a 1000 times and plot the distribution of sample averages.

Watch how the sampling distribution (distribution of averages) forms a normal distribution (bell curve) centered at 4.5.

Demo of CLT.

So what does this mean for data scientists?

Well, the CLT is super useful for making inferences about populations based on samples. In real-world applications, we often can't measure the entire population, so we have to use a sample to make inferences about the population. Thanks to the CLT, we know that if we have a large sufficiency sample, we can be sure that our sample mean will be a good estimate of the population mean. Moreover, using the standard deviation of the sample means & the sample size (number of samples drawn in a sample), we can estimate the standard deviation of the population.

$$\text{Sample mean} = \text{Population mean} = \mu$$

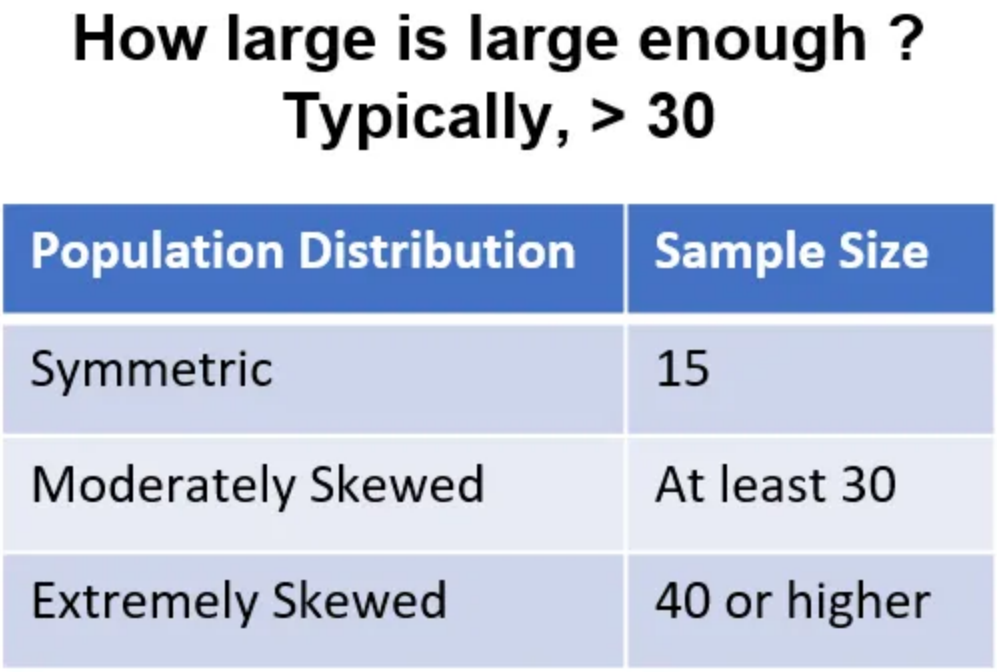
$$\text{Sample standard deviation} = \frac{(\text{Standard deviation})}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

> *One can infer the population parameters — population mean and population standard deviation just with the help of mean of sample means and the standard deviation of the sample means if the sample is large enough.*

How large is large enough ?

If we have a fair idea about the distribution of the population, even a smaller sample size will help. But in most of the instances we know less to nothing about our population. In such instances the thumb rule is to stick to a **minimum of 30 samples**.

> *A sufficiently large sample size can predict the characteristics of a population more accurately.*



Population Distribution	Sample Size
Symmetric	15
Moderately Skewed	At least 30
Extremely Skewed	40 or higher

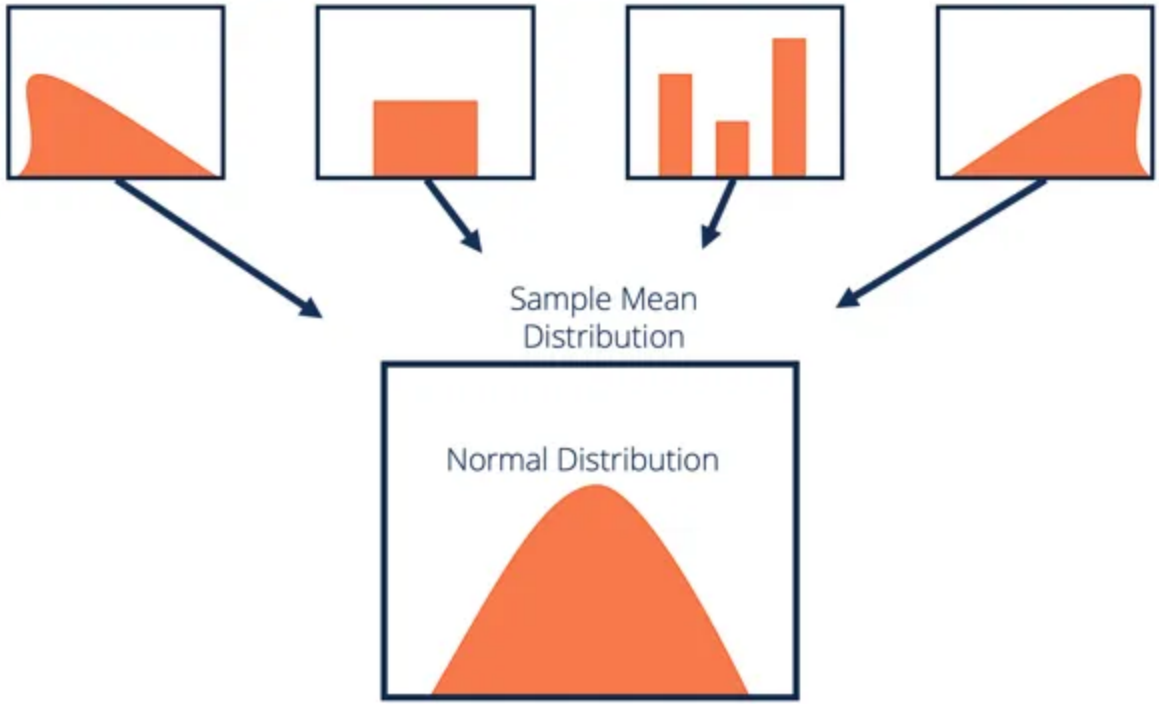
Things to Remember

In order to leverage the power of CLT one should follow the below guidelines.

1. Samples should be selected randomly
2. Sample draws should be independent of each other
3. Sampling with replacement meaning, some samples potentially being common to samples selected on previous occasions from a population
4. Sample size should be no bigger than 10% of the population.
5. Sample size should be at least 30 if one has no clue about the population distribution.

In conclusion, the Central Limit Theorem is a powerful tool in the data scientist's arsenal. It allows us to make inferences about populations based on samples. Hope you're able to appreciate why the average of the averages is the average.

> *Reiterating the key idea — CLT states that, regardless of how the data is distributed in the population, the mean of the sample will be close to the true population mean if the sample size is large enough.*



So, even if the population behaves wildly, it's not the end of the world — with the Central Limit Theorem, you can find some order among the chaos!" :)

Check out the part 2 of CLT where I discuss real world examples with additional concepts like margin of error, confidence interval and detailed breakdown of applying CLT in any real world problem.

Central Limit Theorem in Real Life

Learn how to apply CLT (Margin of Error, Confidence Interval) in any business problem by getting exposed to several...

medium.com

That’s a wrap! Clap if you like :) I’m Arun Prakash Asokan. Do check out other articles on Statistical Symphony Series and stay tuned for the next one! See you soon !

Statistics | Inferential Statistics | Data Analysis | Central Limit Theorem | Population

More from the list: "NLP"

Curated by Himanshu Birla

Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

· 11 min read · Sep 4, 2021

Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

· 6 min read · Sep 3, 2021

Jon Gi... in

The Word2ve

· 15 min rea

View list

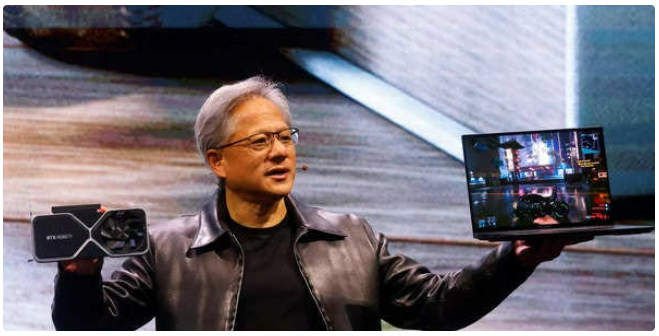
Written by Arun Prakash Asokan

Follow

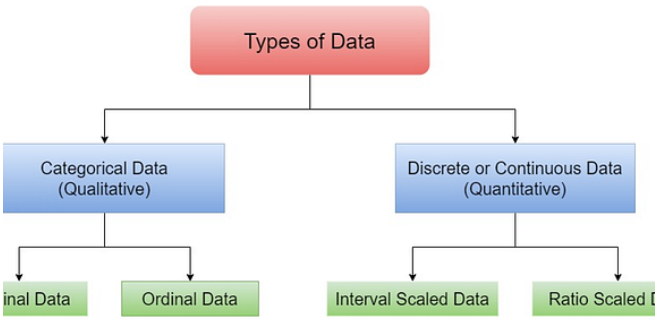
275 Followers

Passionate Data Scientist | AI Intrapreneur | Ardent Teacher | Personal Finance Enthusiast.
Follow me for rich content on AI, Statistics, Tech, Personal Finance

More from Arun Prakash Asokan



Arun Prakash Asokan



Arun Prakash Asokan

Nvidia's AI Revolution. Story behind How it Became a Trillion-Dollar...

$200B in a Day ! Nvidia dominates the GPU market, capturing 88% of all GPUs worldwide.

4 min read · May 30

339 3

Statistical Symphony Series : Types of Data in Statistics !

Learn in & out of types of data & 4 data scales: Nominal, Ordinal, Interval, & Ratio....

7 min read · Jan 26

15

Arun Prakash Asokan

The Law of Large Numbers—Magic Behind Statistics

Have you ever flipped a coin and noticed that you didn't seem to get a even split between...

6 min read · Jun 20

59

Arun Prakash Asokan

Statistical Symphony Series : What is Statistics!

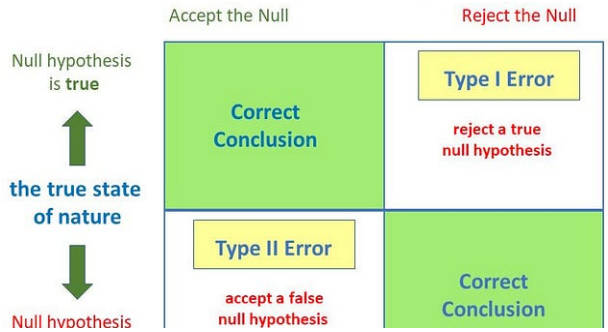
Start the fun ride by understanding what is statistics and its branches, learn about...

7 min read · Jan 26

102 2

See all from Arun Prakash Asokan

Recommended from Medium



Data Analysis

How I Chose Between Factor Analysis and Principal Componen...

Key Takeaways

7 min read · Sep 26

23

misun_song

Hypothesis Test and P-Value

Discerning Real Differences with P-values

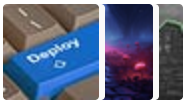
7 min read · Sep 24

73

Lists

Practical Guides to Machine Learning

10 stories · 519 saves



Predictive Modeling w/ Python

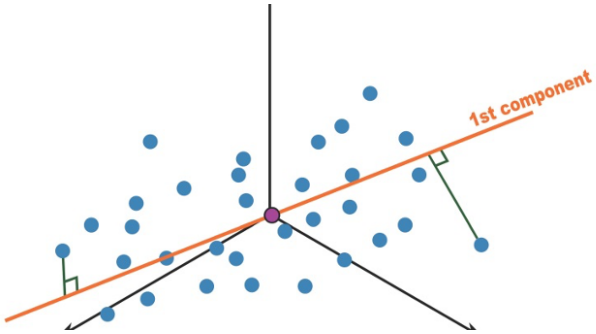
20 stories · 452 saves

New_Reading_List

174 stories · 133 saves

ChatGPT prompts

24 stories · 459 saves



Huda Swati

Understanding Principal Component Analysis (PCA)

What is PCA?

8 min read · Sep 25

8

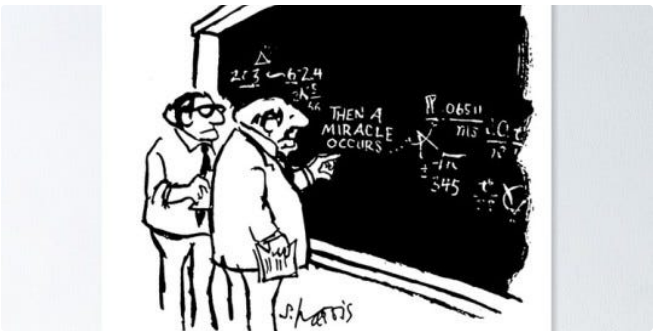
Sruthy Nath

Hypothesis Testing: The Backbone of Statistical Inference

In the world of data analysis and decision-making, statistics plays a pivotal role in...

3 min read · Aug 4

7



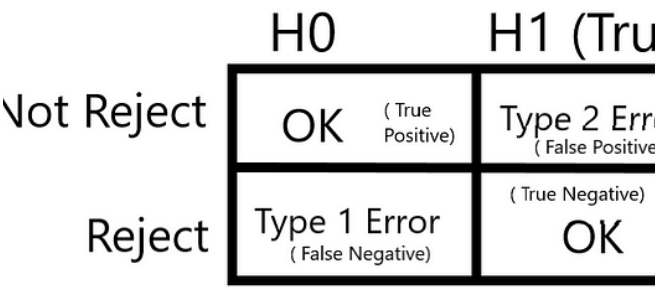
Dr. Marc Jacobs

The importance of applying mathematical models in decision...

Introduction

9 min read · Sep 27

84 4



Kamna Sinha in Data At The Core !

Hypothesis Test And All About P Value,T test,Chi Square Test, Ano...

Hypothesis Test

4 min read · 6 days ago

22

See more recommendations