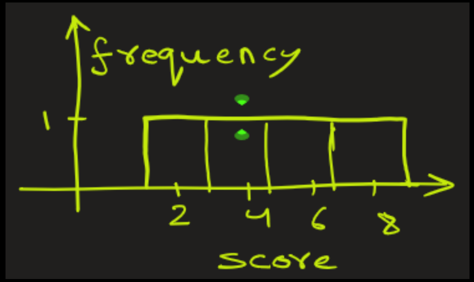
Central Limit Theorem (CLT)

Problem: A professor administered an 8-point quiz to a small class of four students — A, B, C, and D — whose scores are:

Student	Score
A	2
B	6
C	4
D	8

Assume these four students represent the entire **population.**

Here histogram graph of the population is uniform:

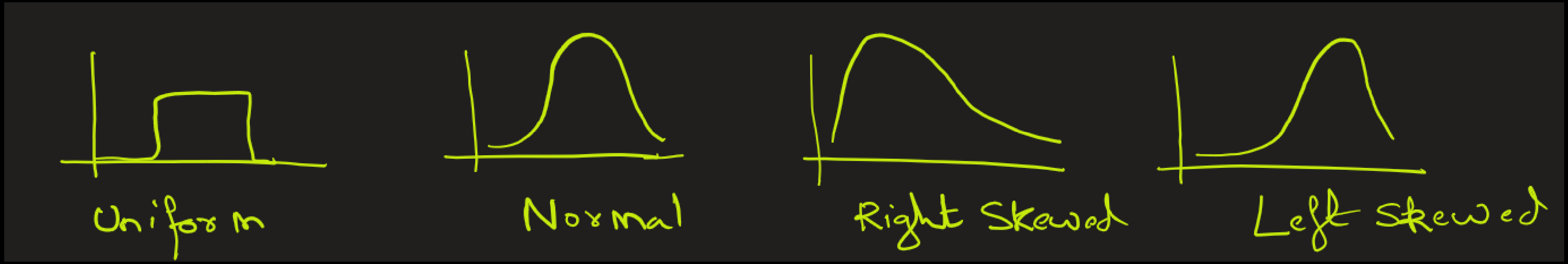


Suppose I take ordered samples of size = 2 with replacement from above population and compute their mean. (See the table.)

Ordered sample means that sample scores of (2,4) and (4,2) are considered different: (2,4) ≠ (4,2)

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

What is the shape of the **sampling distribution of the above sample means ?**



Central Limit Theorem (CLT): Theory

What is the Central Limit Theorem (CLT)?

Let $X_1, X_2, X_3, \ldots, X_n$ be a sequence of **independent and identically distributed (i.i.d.)** random variables with

- finite mean $\mu = E[X_i]$, and
- finite variance $\sigma^2 = Var(X_i)$.

Then, as the sample size $n \to \infty$, the **sampling distribution of the standardized sample mean**

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

approaches a **standard normal distribution**, i.e.

$$Z \xrightarrow{d} N(0,1)$$

In simple terms:

No matter what the shape of the original population distribution is, the **distribution of the sample means** will **approximate a Normal Distribution as the sample size (n) gets larger**

Central Limit Theorem (CLT)

Understanding CLT through an example:
Suppose you have data from a population that has uniform distribution (UD): All values within a range are equal likely

1) You take a sample of 4 points from UD and calculate its mean. Then plot the mean on the histogram

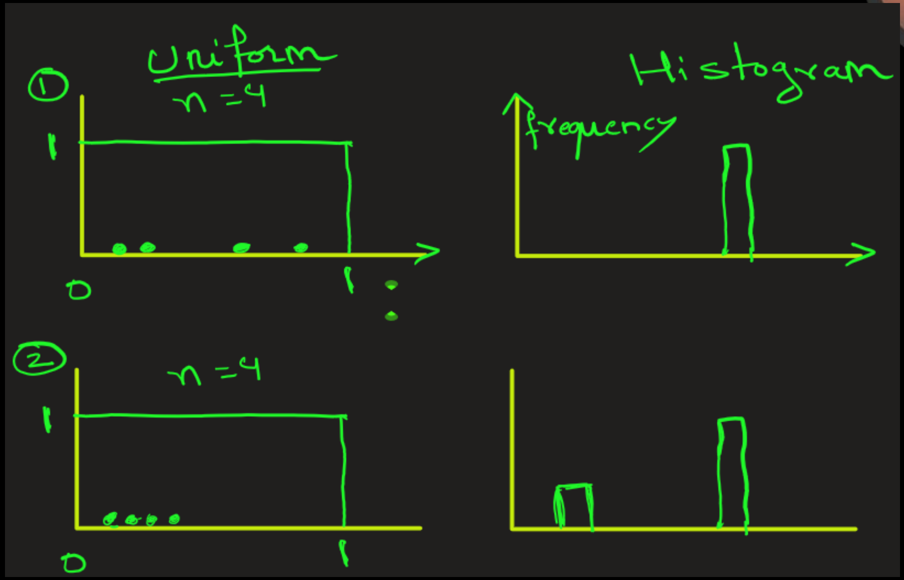
2) You take another sample of 4 points from UD and calculate its mean. Then plot the mean on the histogram

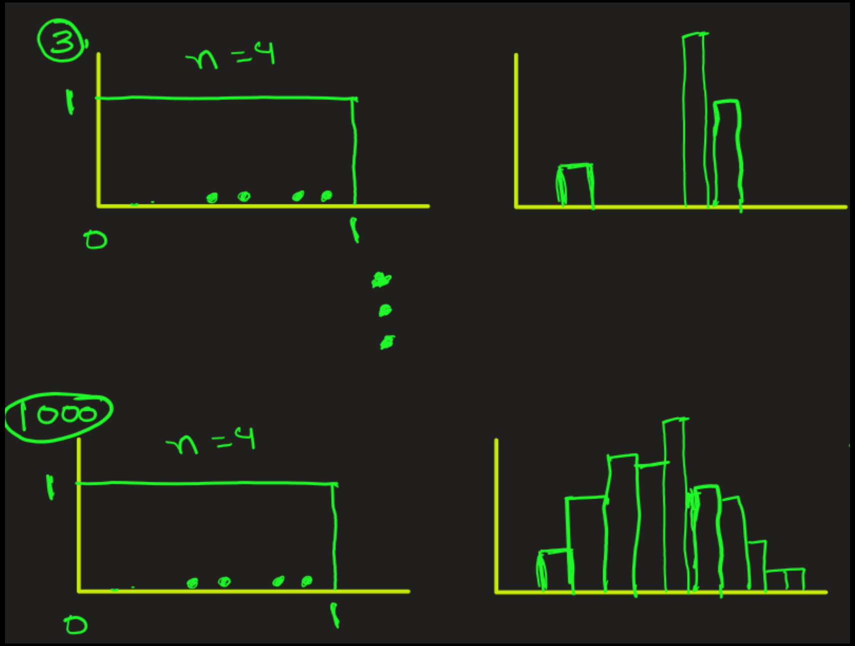
3) You take another sample of 4 points from UD and calculate its mean. Then plot the mean on the histogram

(And you keep doing this for 1000 times)

1000) At the end histogram will take a shape of Normal Distribution.

This is the essential idea: **No matter the distribution of your population, the distribution of the sample means would always come closer to normal distribution as sample size increases.**

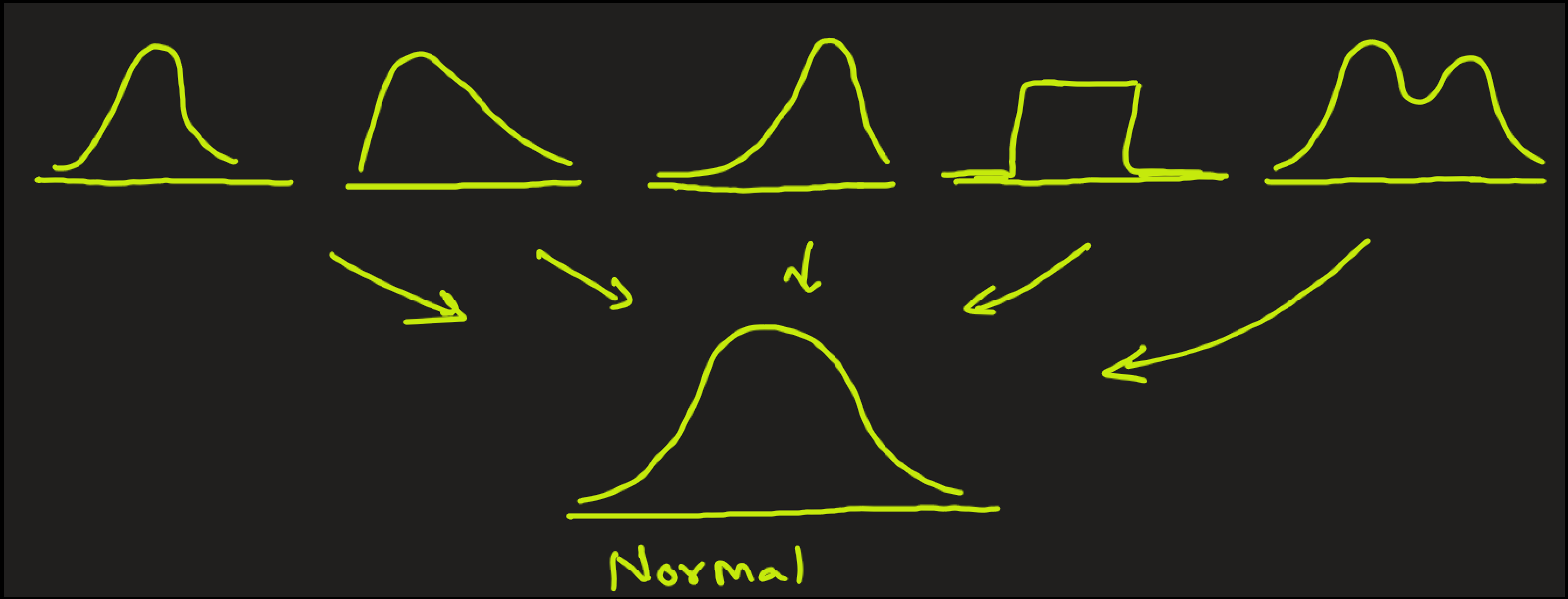




Central Limit Theorem (CLT)

Key idea of CLT:

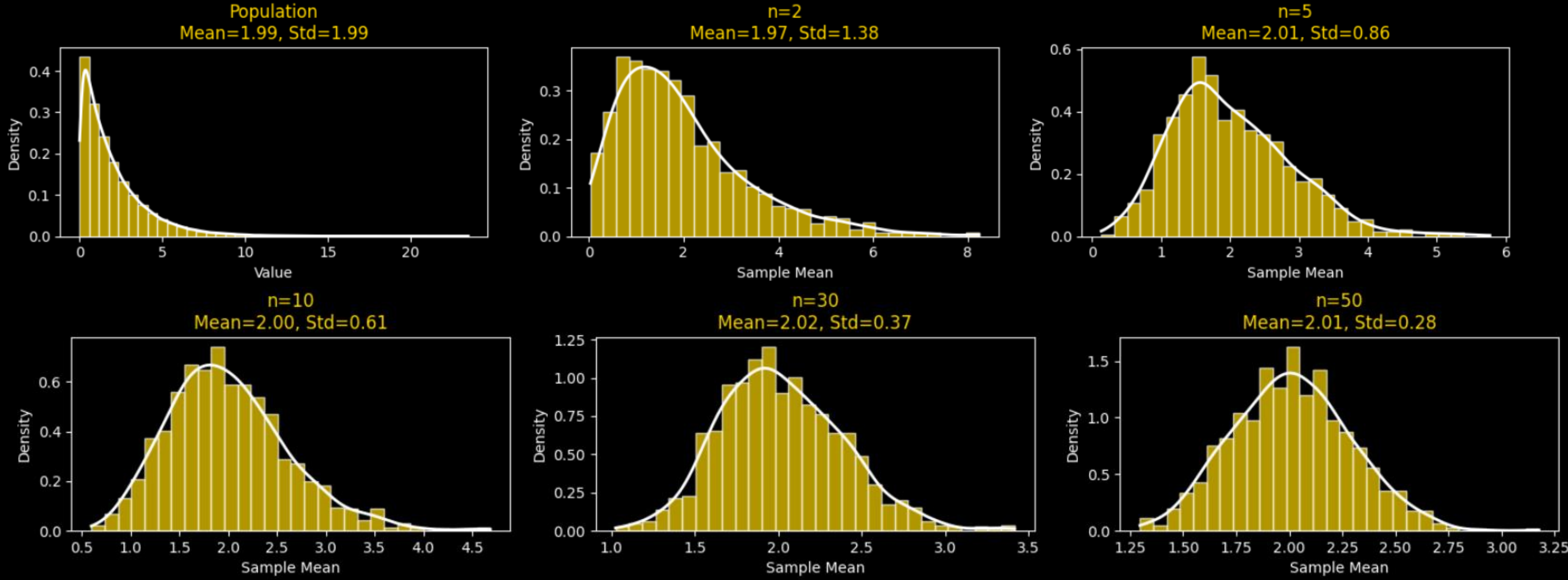
- The **population** can be anything: normal, skewed, uniform, bimodal or any weird shape.
- If you **randomly sample** from it repeatedly and calculate the **mean** each time, then the **distribution of those means** will **always tend toward a normal distribution**.

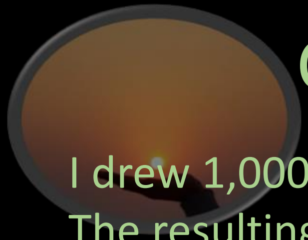


CLT: The Effect Of Sample Size n On The Distribution Of Samples Mean

1) Start with a population that has a mean of $\mu = 2$ and can follow any type of distribution.
2) From this population, randomly draw n=2 data points. Then calculate the mean of those n=2 values.
4) To get reliable results, repeat this process 1,000 times for this n=2. This gives us 1,000 sample means, each representing one possible estimate of the population mean.
5) Finally, plot the distribution of these 1,000 sample means for sample size n = 2.
6) Repeat steps 2 to 5 for n = 5, 10, 30, and 50

Observation: As n becomes large, 50, **the distribution of the sample means is normal with mean = μ = 2 and standard deviation = $\sigma / \sqrt{n}$ = $2 / \sqrt{50}$. The standard deviation decreases with increasing n.**





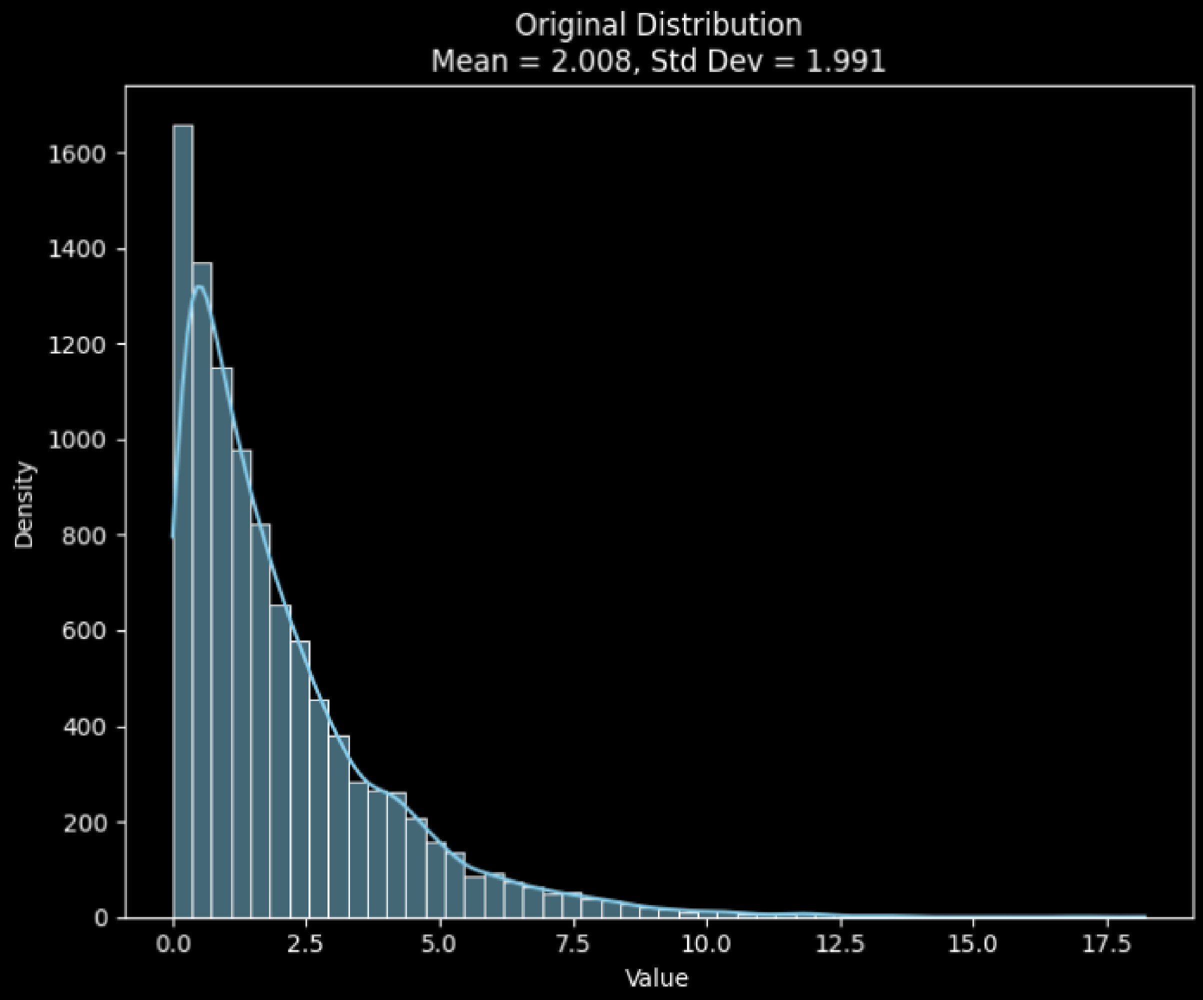
Central Limit Theorem (CLT): Effect of Population Distribution

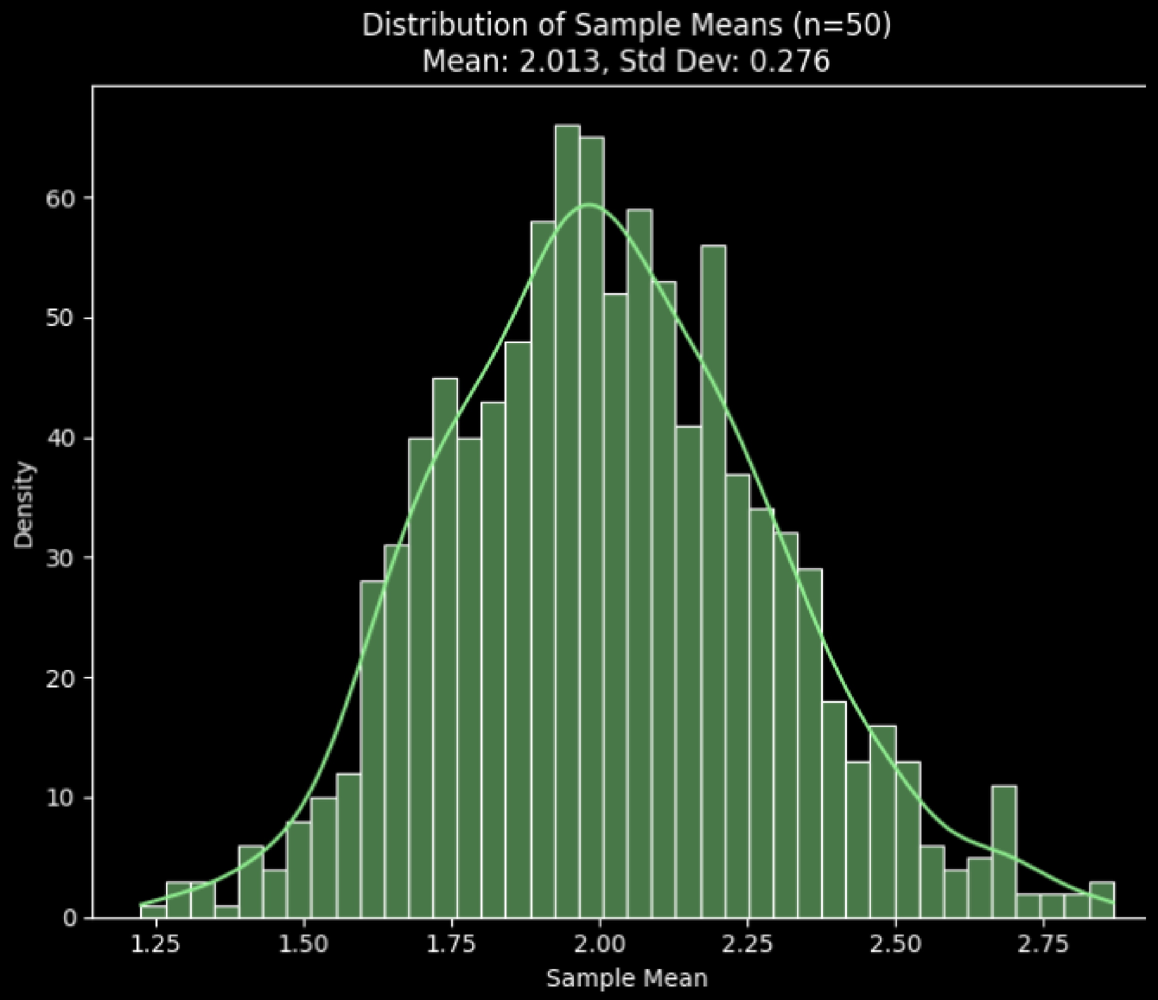
I drew 1,000 random samples from an **exponential distribution** and calculated the mean of each sample. The resulting distribution of these sample means followed an approximately **normal distribution**.

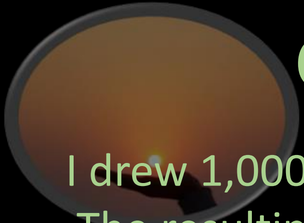
$$\text{Mean of sample means} = \mu$$

$$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$$

Exponential Distribution







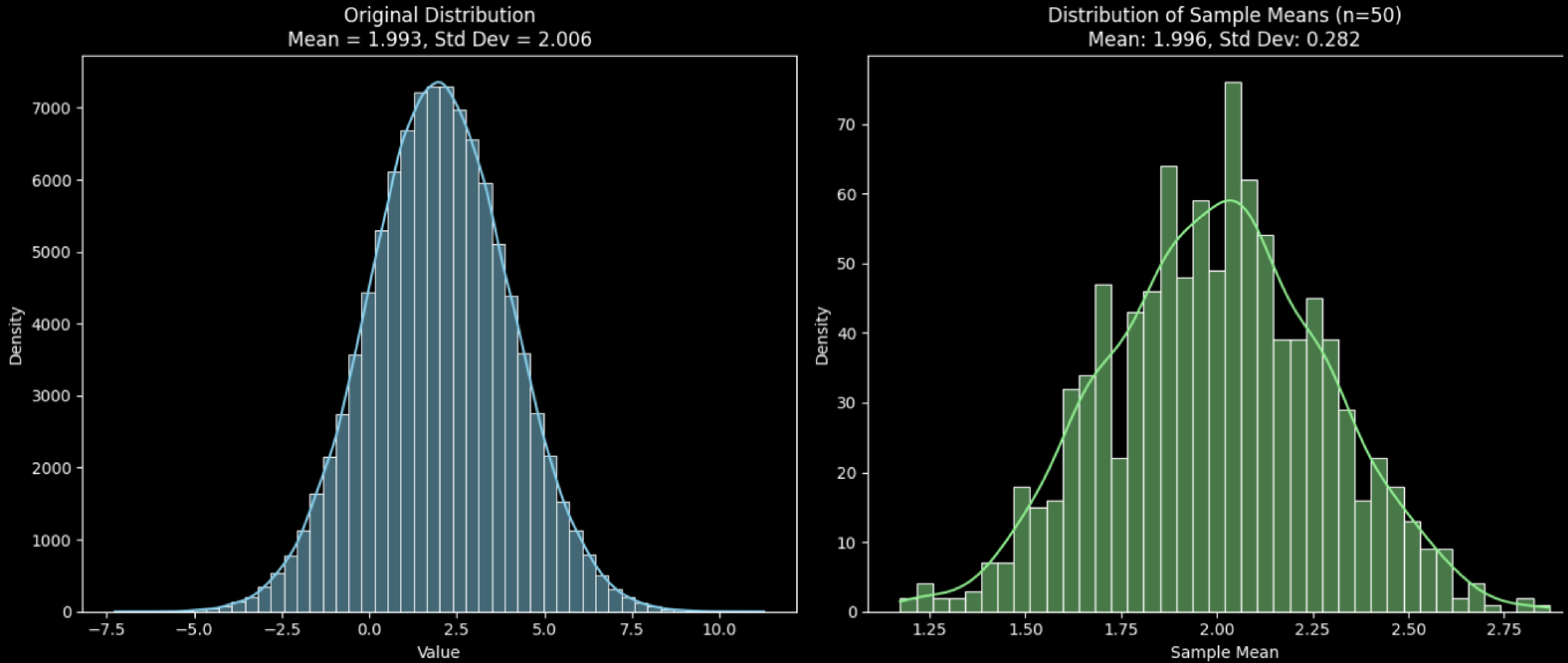
Central Limit Theorem (CLT): Effect of Population Distribution

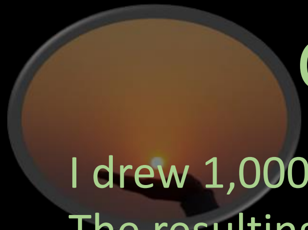
I drew 1,000 random samples from **normal distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

$$\text{Mean of sample means} = \mu$$

$$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$$

Normal Distribution





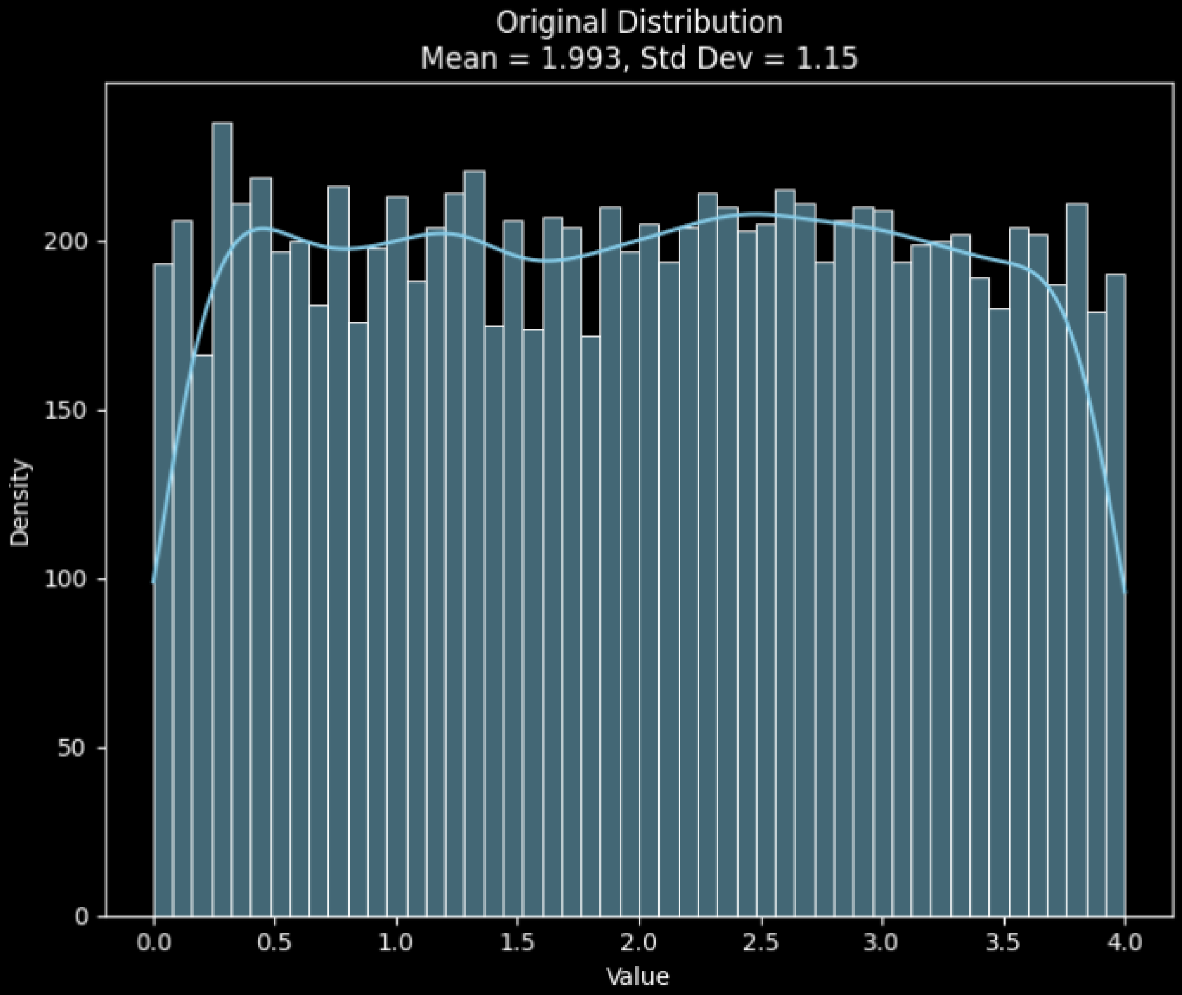
Central Limit Theorem (CLT): Effect of Population Distribution

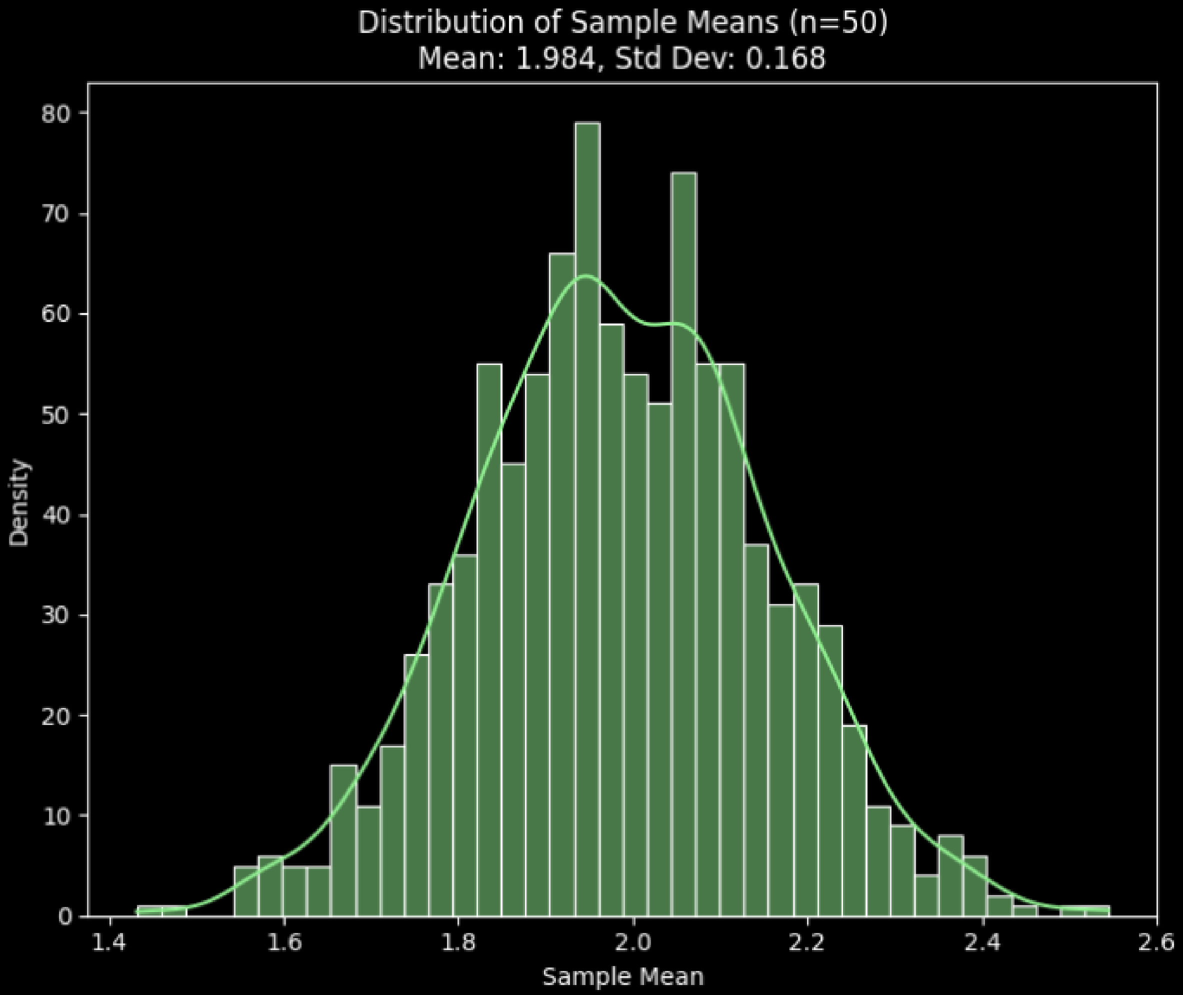
I drew 1,000 random samples from **uniform distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

$\text{Mean of sample means} = \mu$

$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$

Uniform Distribution





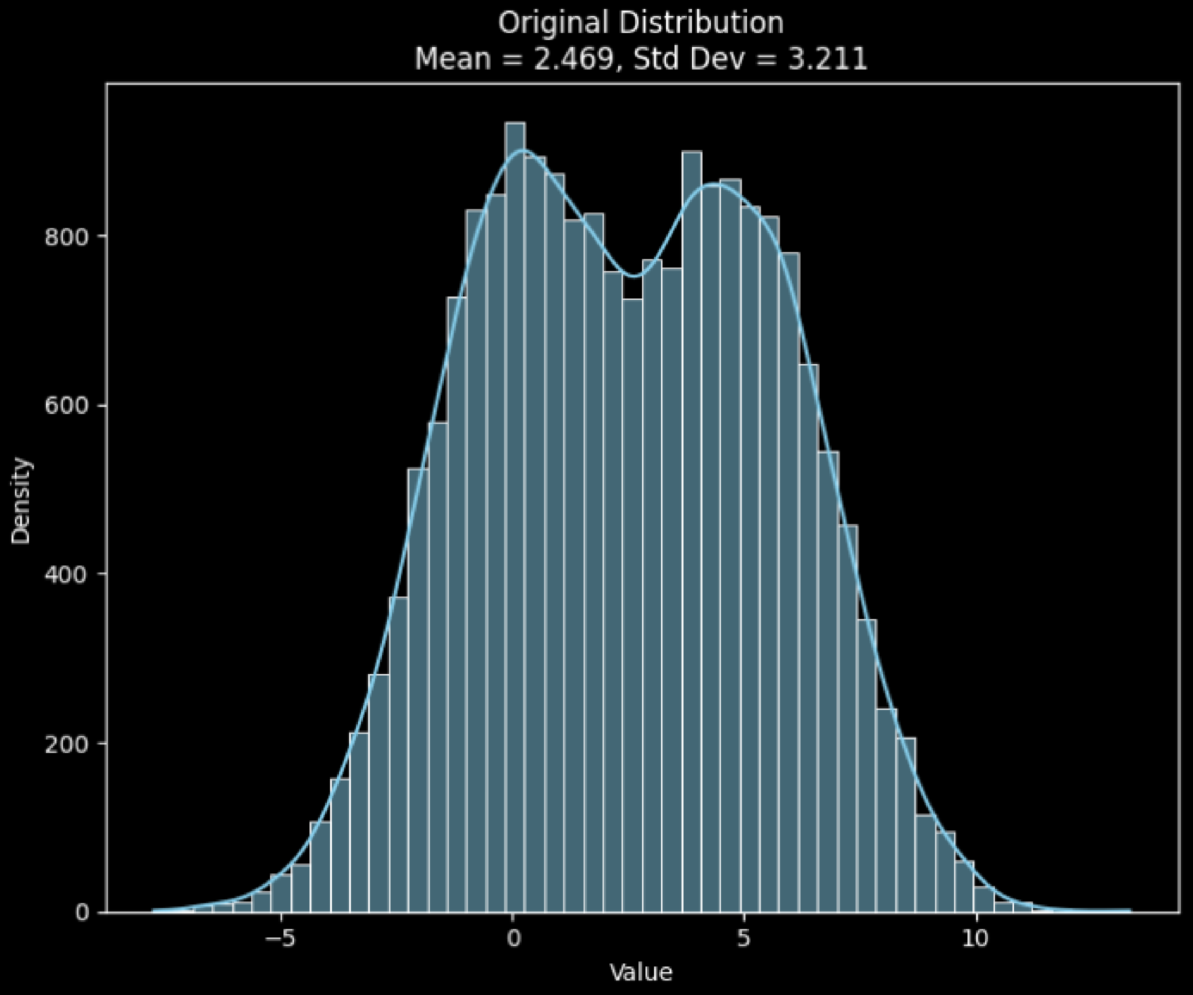
Central Limit Theorem (CLT): Effect of Population Distribution

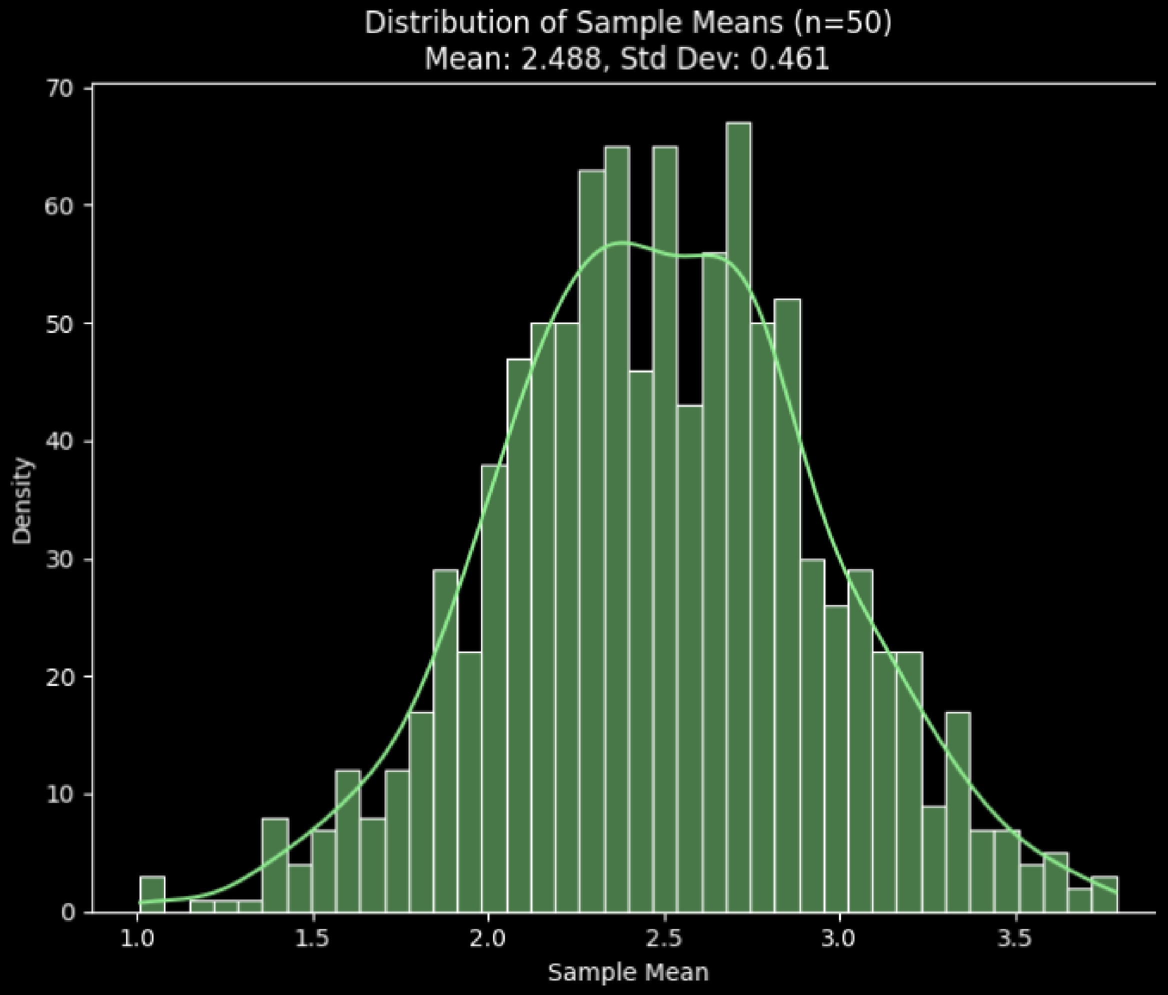
I drew 1,000 random samples from **multi-modal distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

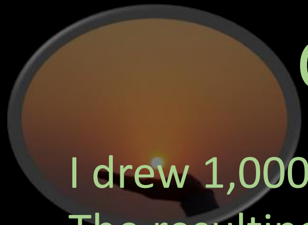
$\text{Mean of sample means} = \mu$

$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$

Multi-modal Distribution







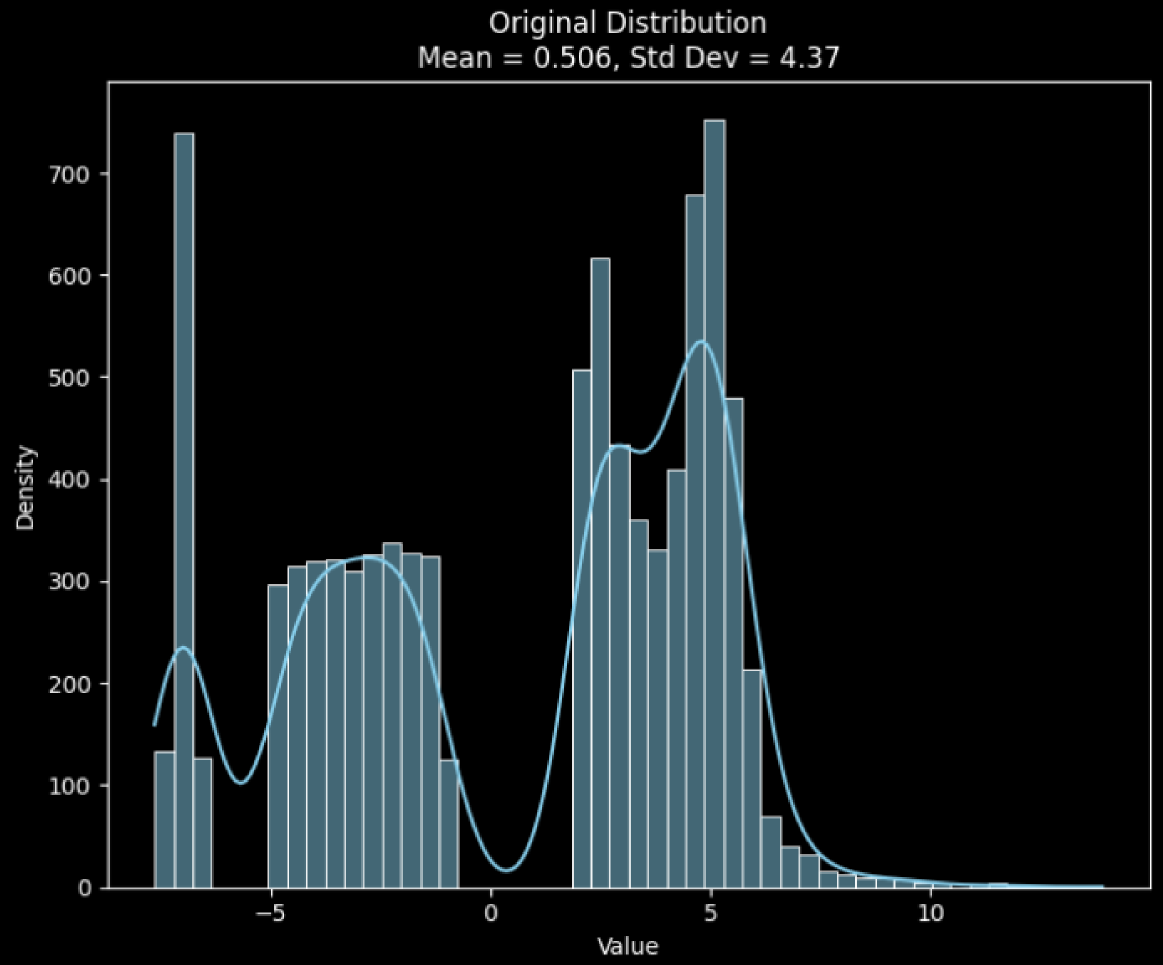
Central Limit Theorem (CLT): Effect of Population Distribution

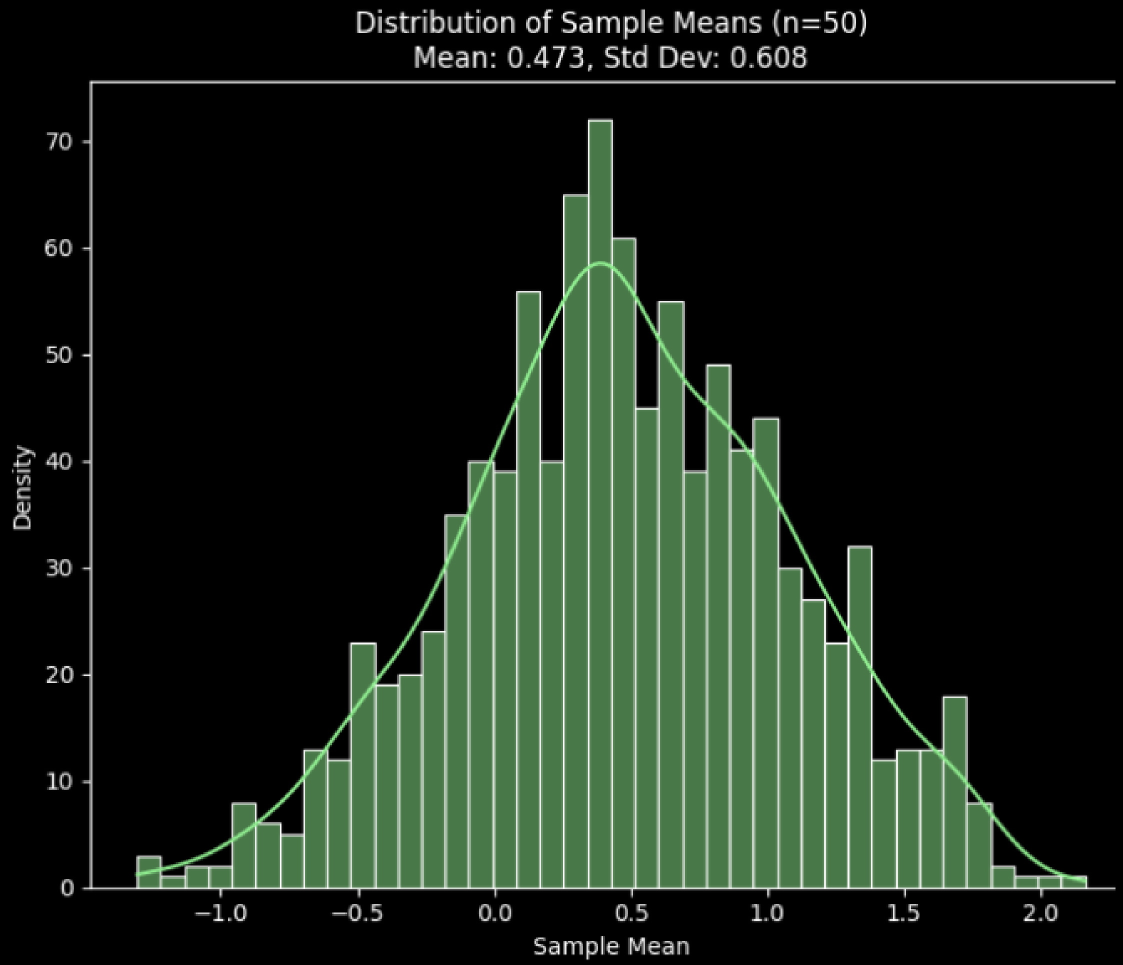
I drew 1,000 random samples from **weird distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

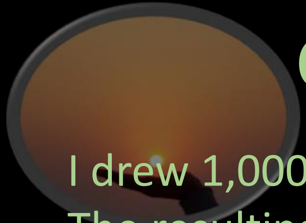
$\text{Mean of sample means} = \mu$

$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$

Weird-shape Distribution







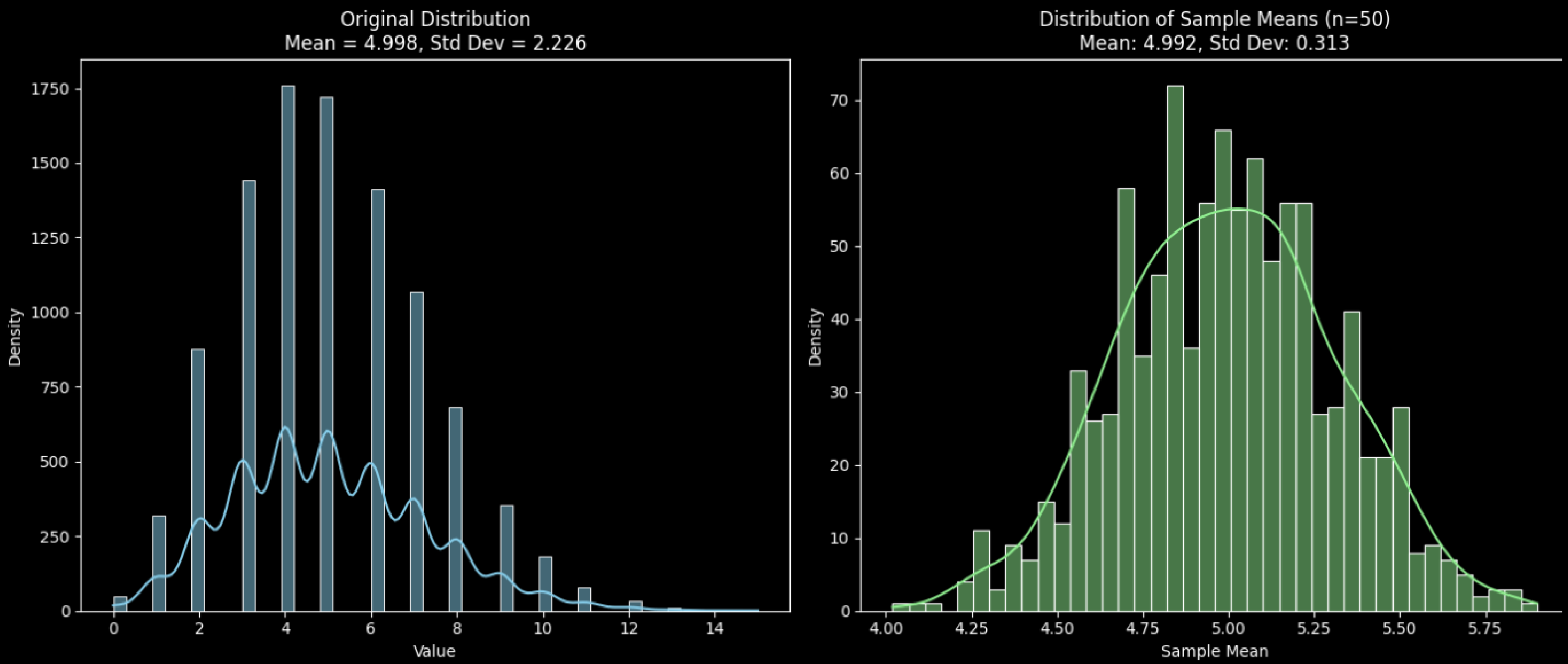
Central Limit Theorem (CLT): Effect of Population Distribution

I drew 1,000 random samples from **Poisson distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

$$\text{Mean of sample means} = \mu$$

$$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$$

Poisson Distribution



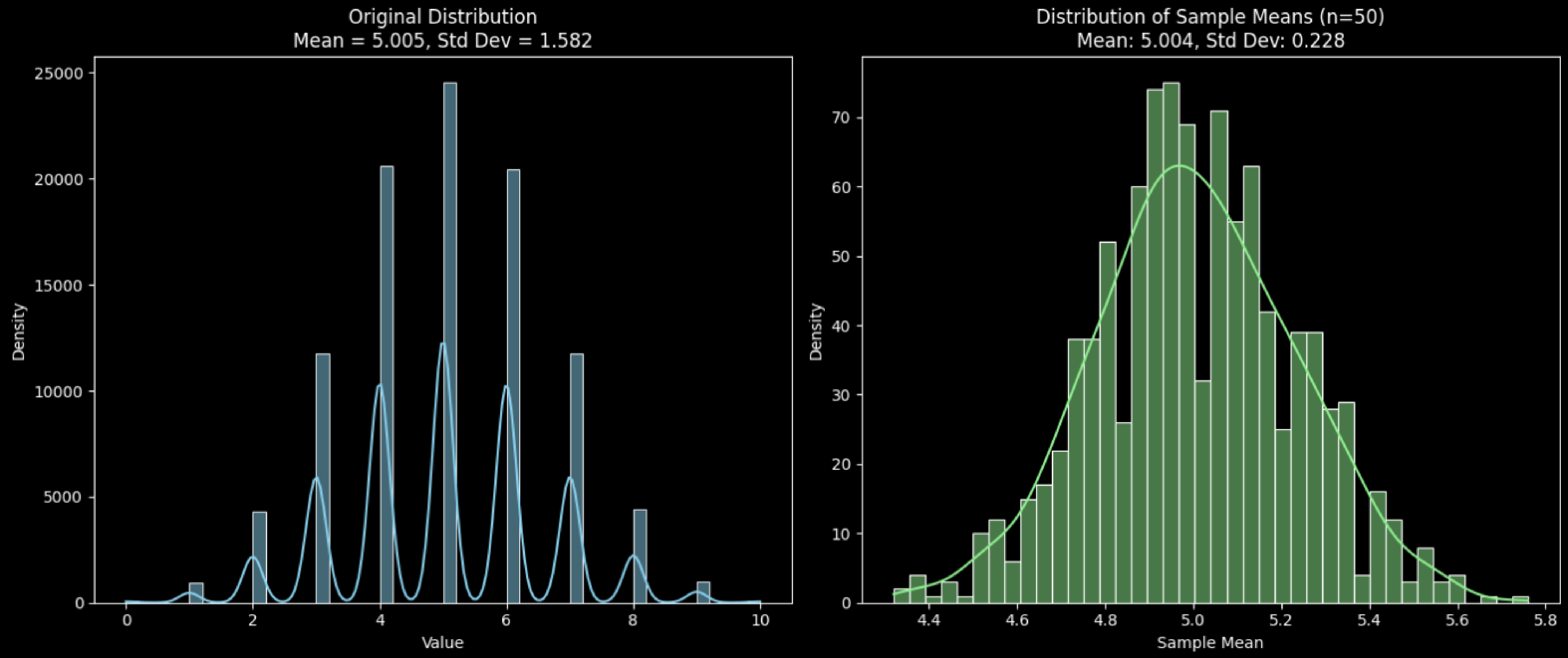
Central Limit Theorem (CLT): Effect of Population Distribution

I drew 1,000 random samples from **Binomial distribution** and calculated the mean of each sample.
The resulting distribution of these sample means followed an approximately **normal distribution**.

$$\text{Mean of sample means} = \mu$$

$$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$$

Binomial Distribution



Central Limit Theorem (CLT): The Essence

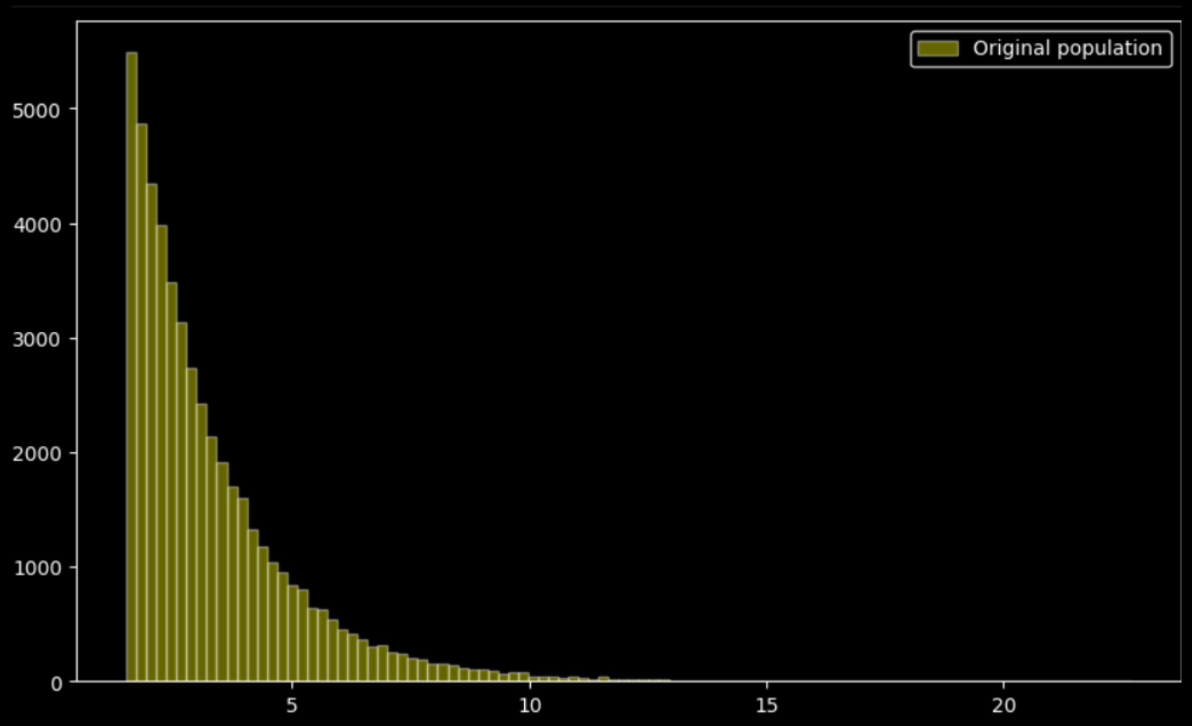
Two important observation about distribution of sample means:

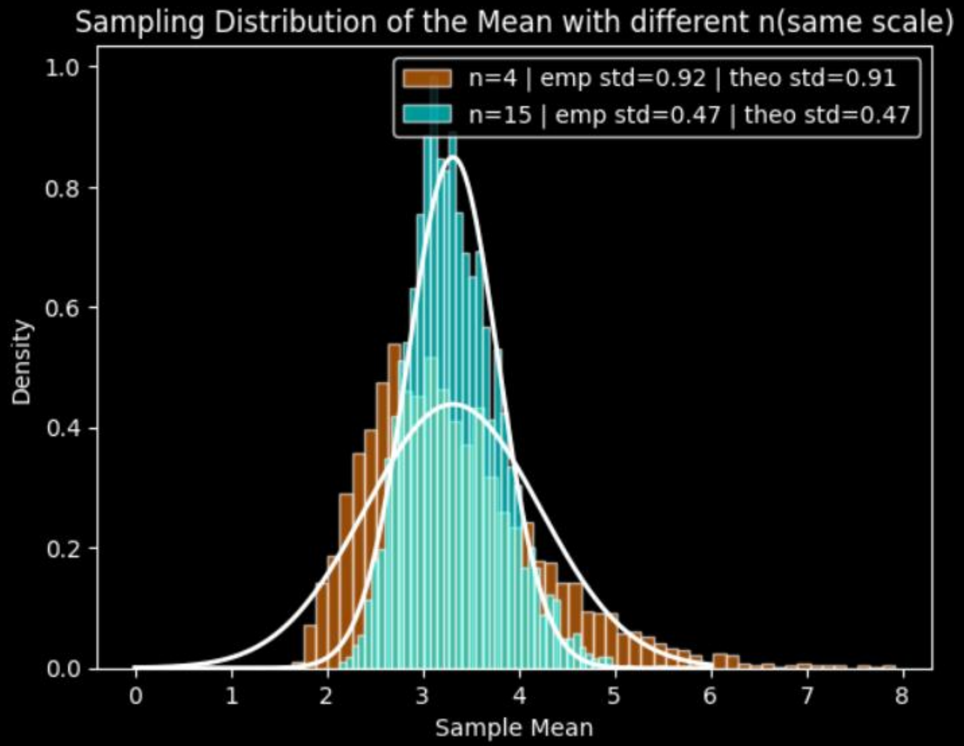
If population mean is μ and population standard deviation is σ, then as you take larger and larger **samples of size n**, the distribution of sample means will be **normal** and have following 2 characteristic:

$$\text{Mean of sample means} = \mu$$

$$\text{Standard deviation of sample means} = \frac{\sigma}{\sqrt{n}}$$

So, the spread (standard deviation) **shrinks** as n grows — that's why the histograms get **narrower and smoother**.





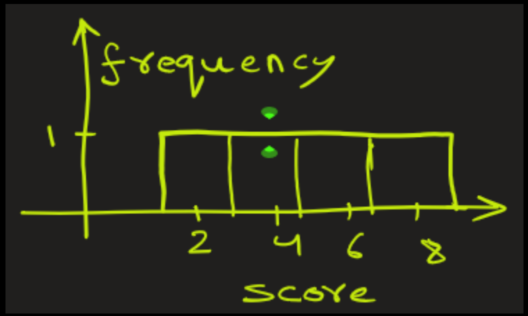
Central Limit Theorem (CLT): Application

Problem: A professor administered an 8-point quiz to a small class of four students — A, B, C, and D — whose scores are:

Student	Score
A	2
B	6
C	4
D	8

Assume these four students represent the entire **population.**

Here histogram graph of the population is uniform:



How many ordered samples of size = 2 with replacement can be drawn from above population ?
Verify the theoretical relationships between the population parameters and the sampling distribution parameters:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where - $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ are the mean and standard deviation of the sampling distribution of sample means, respectively.
- μ and σ are the population mean and standard deviation, respectively,
- n = 2 is the sample size.

Finally, **examine the shape** of the **sampling distribution of the sample means**, and compare it to the shape of the **population distribution**.

Central Limit Theorem (CLT): Application

Ans:

Here the population is [2, 6, 4, 8]. These are scores of students.

The **mean of the population μ** is (2+6+4+8) / 4 = 5.0

The **standard deviation of the population σ** is

$$\sigma = \sqrt{\frac{(2-5)^2 + (6-5)^2 + (4-5)^2 + (8-5)^2}{4}}$$

$$= 2.236$$

Now, if all ordered samples of size 2 are taken with replacement and the mean of each sample is found, the distribution of the samples would be as shown:

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

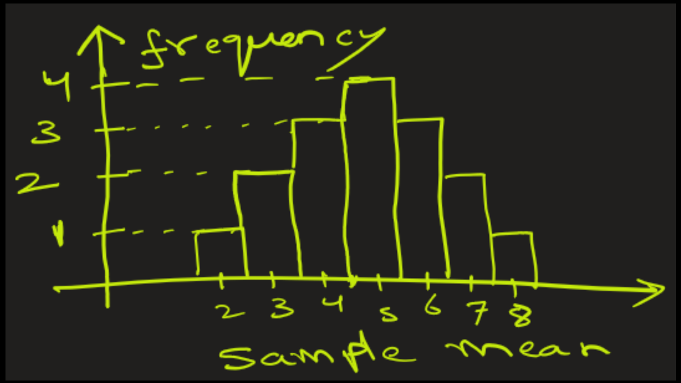
The sample means are [2, 3, 4, 5, 3, 4, 5, 6, 4, 5, 6, 7, 5, 6, 7, 8]

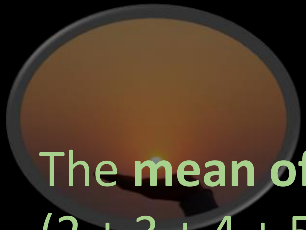
Central Limit Theorem (CLT): Application

A frequency distribution of sample means is as follows

Sample Mean	Frequency
2	1
3	2
4	3
5	4
6	3
7	2
8	1

From this we can draw the graph of the sample means which is a Bell curve(i.e. Normal) as shown:





Central Limit Theorem (CLT): Application

The **mean of the sample means**, is
(2 + 3 + 4 + 5 + 3 + 4 + 5 + 6 + 4 + 5 + 6 + 7 + 5 + 6 + 7 + 8) / 16 = 5
This is same as the population mean.

The **standard deviation of the sample means** is

$$\sigma_{\bar{x}} = \sqrt{\frac{(2-5)^2 + 2(3-5)^2 + 3(4-5)^2 + 4(5-5)^2 + 3(6-5)^2 + 2(7-5)^2 + (8-5)^2}{16}}$$

$$= 1.581$$

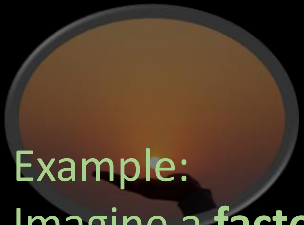
This is same as $\frac{\sigma}{\sqrt{n}}$ = 2.236/$\sqrt{2}$ = 1.581

Conclusion:

- **The shape** of the **sampling distribution of the sample means** is bell curve.
- The **mean and standard deviation of the sampling distribution of sample means** satisfy following relation

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

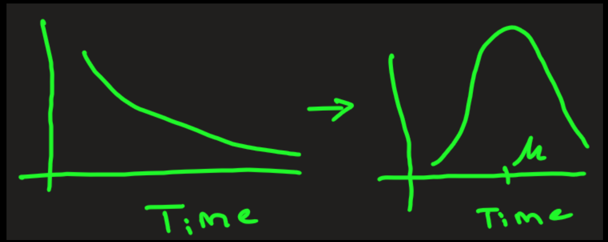


Central Limit Theorem (CLT): Application

Example:
Imagine a **factory** that produces **light bulbs**.
Each light bulb has a lifetime (in hours), and they want to determine average lifetime of bulbs.

But it's impossible to test *every single bulb* — that would destroy them all!
So, instead, the quality control team takes **samples** — say, 30 bulbs each day — and measures their lifetimes.

Even though, the **population distribution** of lifetimes might be **skewed** (some bulbs last much longer than average, some much shorter), the **Central Limit Theorem** says that if enough samples are taken, say maybe for 50 days, the **distribution of sample means** (average lifetimes from those 30 bulbs) will be **approximately normal**.



Because of CLT, the company can **use the normal distribution** to make predictions about average lifetime.
They can **calculate confidence intervals** and **perform hypothesis tests** on the average bulb life.

Without CLT, we'd need to know the true population distribution (which we almost never do).

EXTRA

Example:

Suppose the population is **daily spending by customers** at a store:
It’s highly **skewed** because a few people spend a LOT.
Mean daily spending: $50, variance: known.

Now:
Take 1 random customer → that spending is **not normal**.
But take a **sample of 100 customers**, find the **average spending** → that **average** is nearly **normally distributed**.

If you repeat this process (many samples of 100), the **histogram of those averages** will look **bell-shaped**, centered at the true mean ($50).