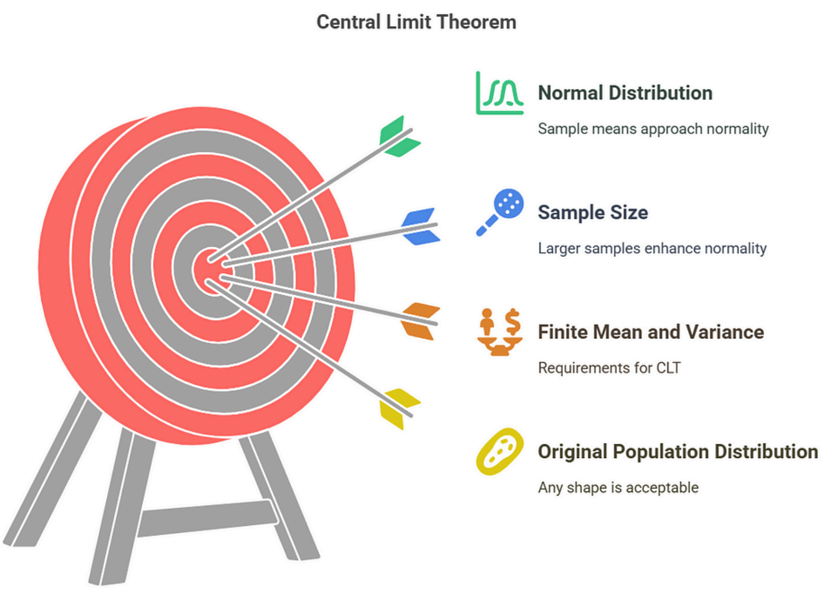
📌 Central Limit Theorem (CLT)

The **Central Limit Theorem (CLT)** states that when we take multiple random samples from any population, the average of those samples will form a normal (bell-shaped) distribution as the sample size increases, no matter the shape of the original population.



🔷 What does CLT say?

1️⃣ The **distribution of sample means** always follows a **normal distribution** 📈, no matter how the original data is distributed!

$$[\bar{X}_1, \bar{X}_2, \bar{X}_3, ...\bar{X}_n] \approx N(\mu, \sigma)$$

2️⃣ The **mean of the sample means** is approximately equal to the **population mean** 🏠:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \approx \mu$$

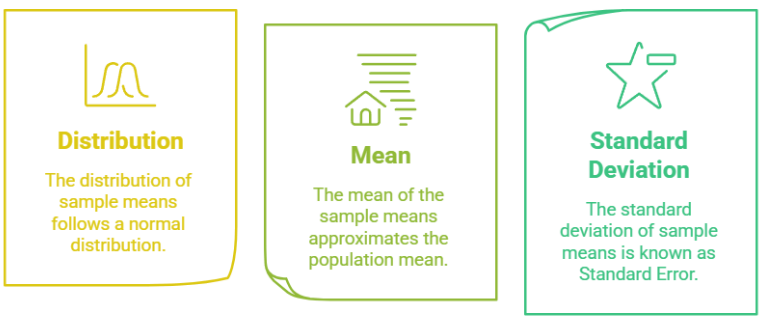
3️⃣ The **standard deviation of sample means** (also called **Standard Error**) is approximately:

$$s \approx \frac{\sigma}{\sqrt{n}}$$

where:

- **s** = Sample standard deviation
- **σ** = Population standard deviation
- **n** = Number of samples

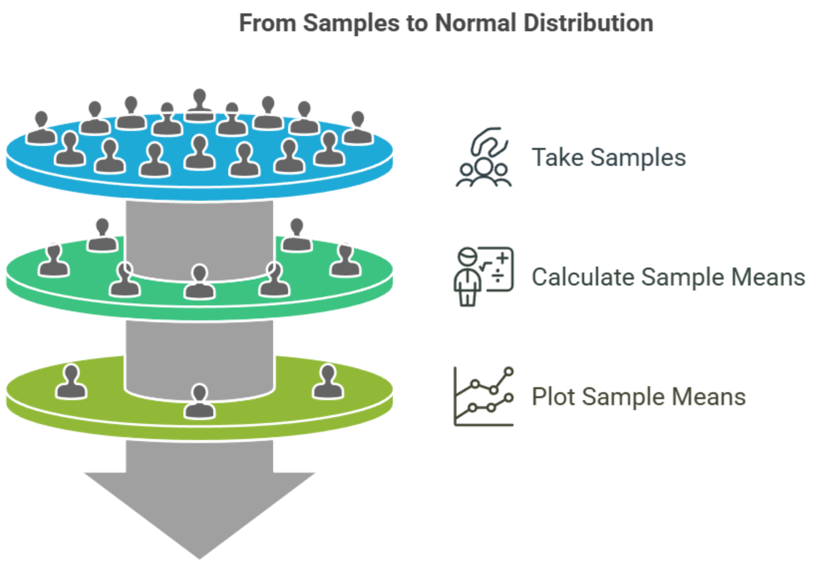
Sample Means Properties



🎯 Simple Example:

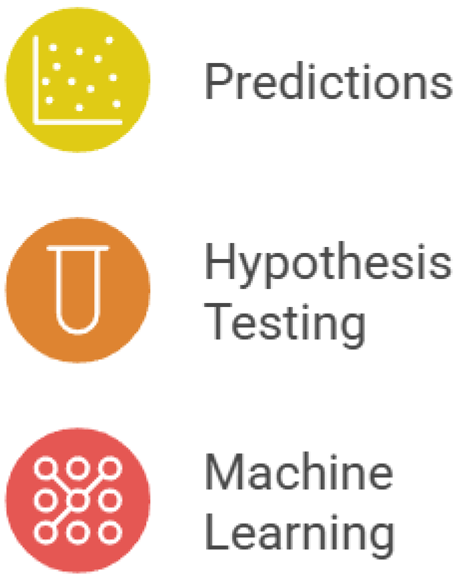
Imagine you want to estimate the **average height** of people in your city 🌆.

- Instead of measuring **everyone**, you take **multiple small samples** (e.g., groups of 1️⃣ 0️⃣ 0️⃣ people each).
- Each group will have an **average height (sample mean)**.
- If you plot these sample means, you'll get a **normal distribution** 🎯, even if the original population was **not normally distributed**!



🔥 Why is CLT Important?

✅ Helps in making **predictions** when data is unknown.
✅ Forms the foundation for **hypothesis testing & confidence intervals.**
✅ Used widely in **Machine Learning & Data Science.**



📢 ESTIMATION IN STATISTICS

📌 1. Point Estimation

👉 Estimating population parameters using a **single value.**
📌 **Formulas:**

- **Mean (μ) $\approx$ Sample Mean ($\bar{X}$)**
- **Population Std (σ) $\approx$ Sample Std (S) $\times \sqrt{n}$**

📌 2. Confidence Interval (CI)

📊 Estimating population parameters **with a range of values.**

Types of CI:

✅ **Z-Statistic (Z):** Used when **σ (population std) is known** & **$n \geq 30$**
✅ **T-Statistic (T):** Used when **σ (population std) is unknown** & **$n < 30$**

📌 **Formulas:**
🔹 **Z-Statistic**

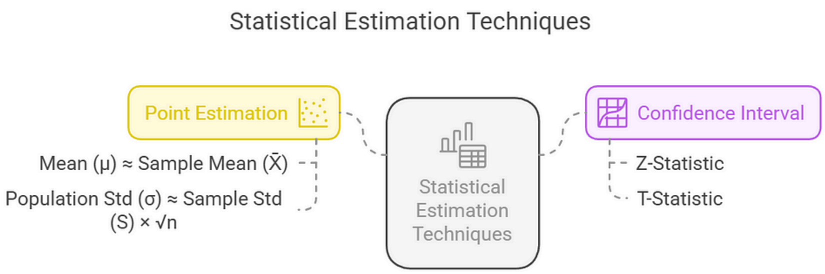
$$\mu = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

🔹 **T-Statistic**

$$\mu = \bar{X} \pm T_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}$$

📌 **Key Terms:**

✔️ **$\bar{X}$ (Sample Mean)** → Average from sample data
✔️ **σ (Population Std Dev)** → Spread of population data
✔️ **n (Sample Size)** → Number of observations
✔️ **α (Significance Level)** → α = **1 - Confidence Level (CL)**
✔️ **S (Sample Std Dev)** → Spread of sample data
✔️ **Z (Z-Value)** → From Z-table
✔️ **T (T-Value)** → From T-table
✔️ **n-1 (Degrees of Freedom)** → Adjusts for small sample sizes



📢 How to perform Estimations:

🔹 1. Collect Samples 🎯

✅ Use **Simple Random Sampling (SRS)** to collect unbiased samples.
✅ Ensure data is **uniformly selected** from the population.

🔹 2. Calculate Sample Statistics 📊

Once you have the sample, compute key statistics:

📌 **Sample Mean ($\bar{X}$)** → Average of sample values

$$\bar{X} = \frac{\sum X_i}{n}$$

📌 **Sample Variance (S^2)** → Spread of sample values

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

📌 **Sample Standard Deviation (S)** → Square root of variance

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

Where:

✔ $\bar{X}$ = Sample Mean

✔ S^2 = Sample Variance

✔ S = Sample Std Deviation

✔ n = Sample Size

✔ X_i = Individual Sample Values

🔹 3. Estimate Population Parameters 📈

Now, use the sample data to estimate the **population parameters** using:

✅ **Point Estimation** → A single best estimate of population parameters.

- **$\mu \approx \bar{X}$**
- **$\sigma \approx S \times \sqrt{n}$**

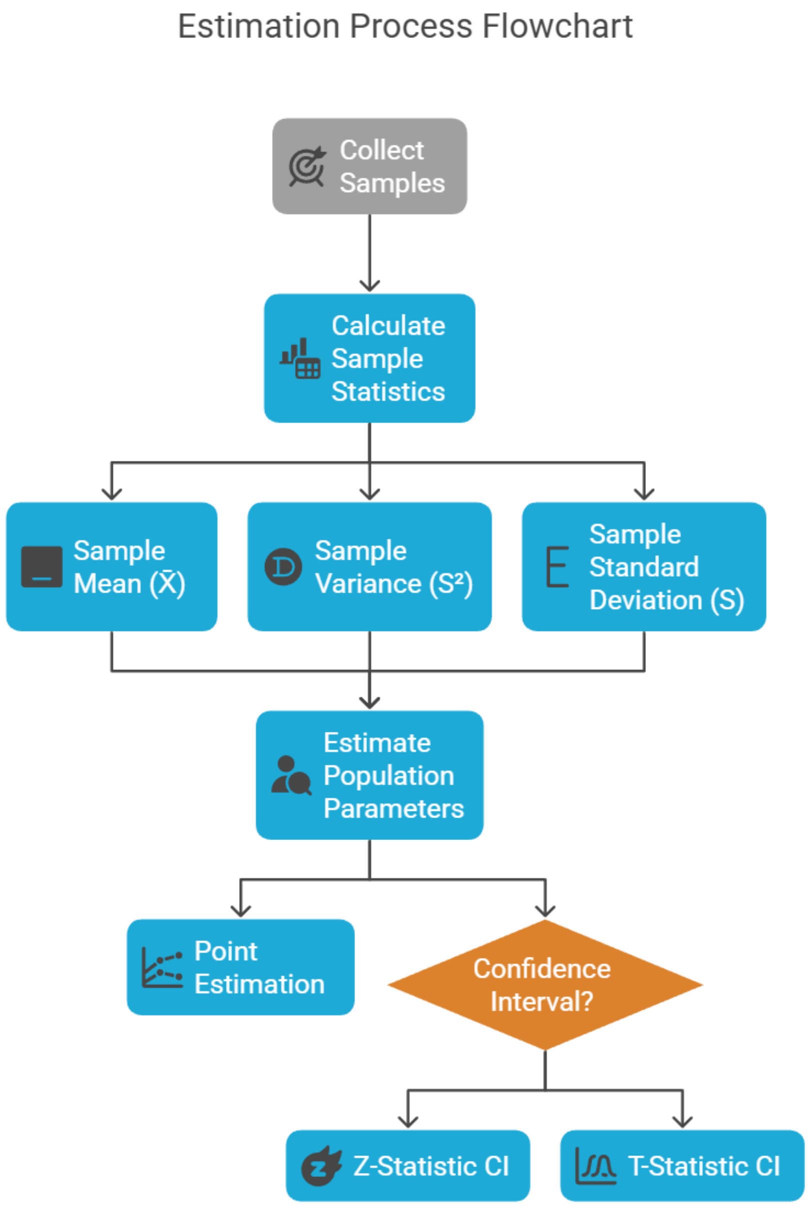
✅ **Confidence Interval (CI)** → A range that likely contains the population mean.

- **Z-Statistic CI** (when **σ** is known & **n ≥ 30**)

$$\mu = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- **T-Statistic CI** (when **σ** is unknown & **n < 30**)

$$\mu = \bar{X} \pm T_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}$$

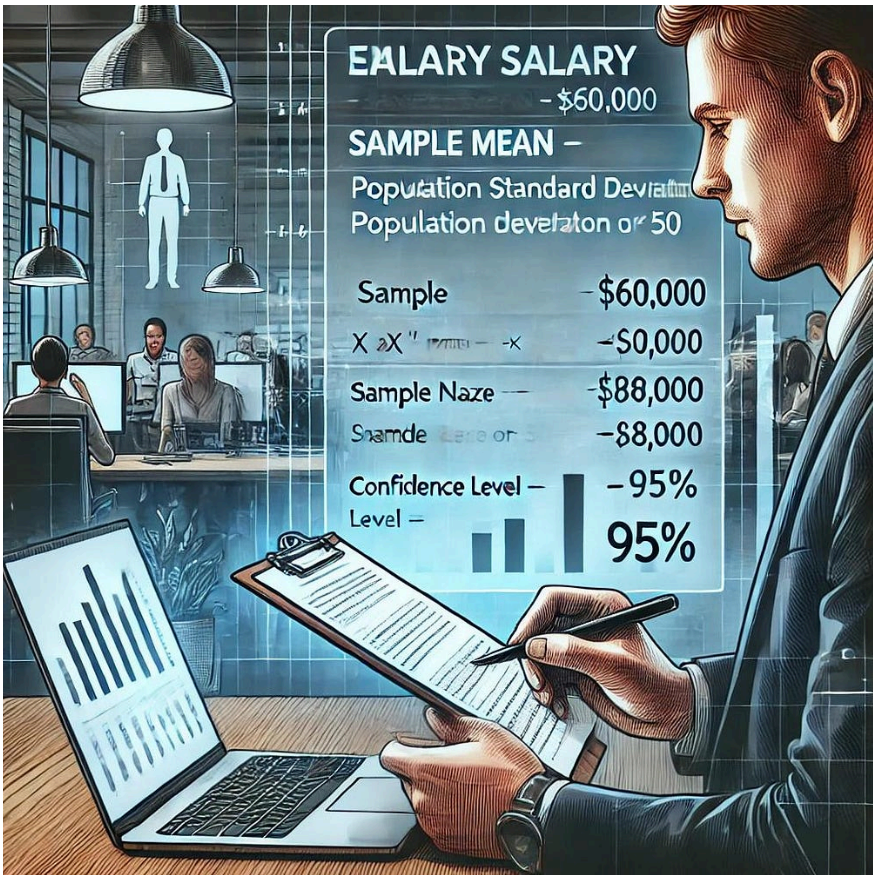


🔹 Example 1: Z-Statistic (Large Sample, σ Known) 🎯

📌 **Question:**

A company wants to estimate the **average salary** of employees. A **random sample of 50 employees** was selected, with:

✔ **Sample Mean ($\bar{X}$) = $60,000**
✔ **Population Standard Deviation (σ) = $8,000**
✔ **Sample Size (n) = 50**
✔ **Confidence Level (CL) = 95%**



🔹 Step 1: Collect the Sample

✅ **Simple Random Sampling (SRS)** is used to select 50 employees.

🔹 Step 2: Compute Sample Statistics

✓ $\bar{X}$ = 60,000
✓ σ = 8,000
✓ n = 50

🔹 Step 3: Estimate Population Parameters

We calculate the **95% Confidence Interval** using the **Z-Statistic formula:**

$$\mu = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

From the Z-table, **Z_(0.05/2) = 1.96** for 95% CI.

$$\mu = 60,000 \pm 1.96 \times \frac{8,000}{\sqrt{50}}$$

$$\mu = 60,000 \pm 1.96 \times \frac{8,000}{7.07}$$

$$\mu = 60,000 \pm 1.96 \times 1,131.4$$

$$\mu = 60,000 \pm 2,218.6$$

✔ **Final 95% CI = [$57,781.4, $62,218.6]**

🔹 **Interpretation:**
We are **95% confident** that the true average salary of employees lies **between $57,781.4 and $62,218.6.**

🔹 Example 2: T-Statistic (Small Sample, σ Unknown) 🎯

📌 **Question:**
A researcher wants to estimate the **average height** of students in a university. A **random sample of 15 students** was taken, with:
✔ **Sample Mean (X̄) = 170 cm**
✔ **Sample Standard Deviation (S) = 6 cm**
✔ **Sample Size (n) = 15**
✔ **Confidence Level (CL) = 95%**

🔹 Step 1: Collect the Sample

✅ **Simple Random Sampling (SRS)** is used to select 15 students.

🔹 Step 2: Compute Sample Statistics

✔ **X̄ = 170 cm**
✔ **S = 6 cm**
✔ **n = 15**
✔ **Degrees of Freedom (df) = n—1 = 14**

🔹 Step 3: Estimate Population Parameters

We calculate the **95% Confidence Interval** using the **T-Statistic formula:**

$$\mu = \bar{X} \pm T_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}$$

From the **T-table**, for **df = 14, T_(0.05/2) ≈ 2.145.**

$$\mu = 170 \pm 2.145 \times \frac{6}{\sqrt{15}}$$

$$\mu = 170 \pm 2.145 \times \frac{6}{3.87}$$

$$\mu = 170 \pm 2.145 \times 1.55$$

$$\mu = 170 \pm 3.33$$

✔ **Final 95% CI = [166.67 cm, 173.33 cm]**

🔹 **Interpretation:**
We are **95% confident** that the true average height of students lies **between 166.67 cm and 173.33 cm.**

📢 Applications of Central Limit Theorem (CLT)

1️⃣ **Confidence Interval Estimation**—Estimates population parameters using sample data when the standard deviation is unknown.

2️⃣ **Hypothesis Testing**—Compares sample means to a hypothesized mean using z-tests/t-tests.

3️⃣ **Regression Analysis**—Ensures regression coefficients follow a normal distribution for valid inference.

4️⃣ **ANOVA (Analysis of Variance)**—Compares multiple group means using the F-distribution.

5️⃣ **Machine Learning Model Evaluation**—Justifies MSE and cross-validation by ensuring error distributions are normal.