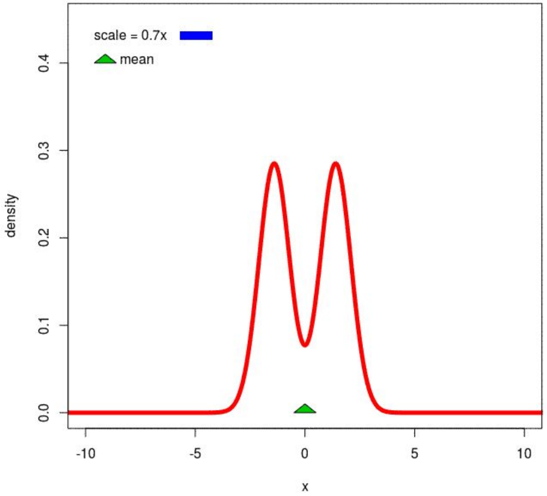
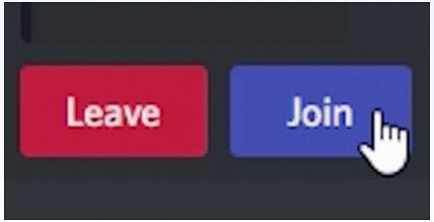
Newton School of Technology



2

Module 4 - Inferential Statistics

Central Limit Theorem, Bootstrapping



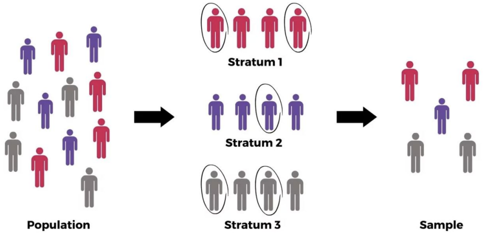
Quick Recap

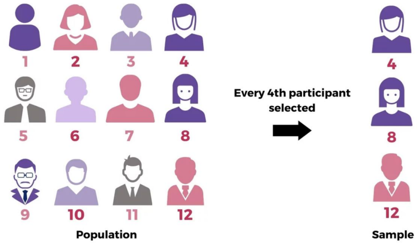
A GOOD SAMPLE IS ⟵ UNBIASED SAMPLE

One that is representative of the entire population gives each thing an equal chance of being chosen.

Simple Random Sampling (SRS): Every item in the population has an equal chance of being selected (Randomly).

Stratified Sampling: The population is divided into subgroups (strata), and random samples are taken from each subgroup.





Systematic Sampling: Every *k*th item is selected from a list after choosing a random starting point.

How trustworthy is Sample?
Are multiple Samples Same?
Maybe one Sample can be more "TRUSTWORTHY"

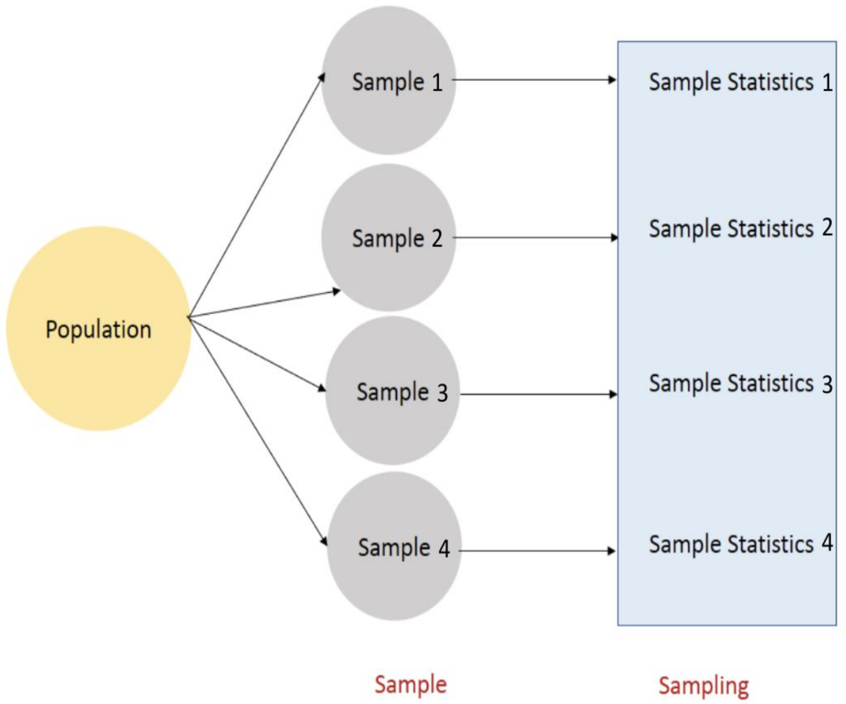
Sampling Distribution

This is the **distribution of a statistic** (like sample mean or proportion) **across many samples** from the same population.

✅ It tells you about the **variability of the sample statistic**.

✅ As the sample size increases, the estimated parameters gets closer to true parameter.

Population - N data points.
Sample - n data points.



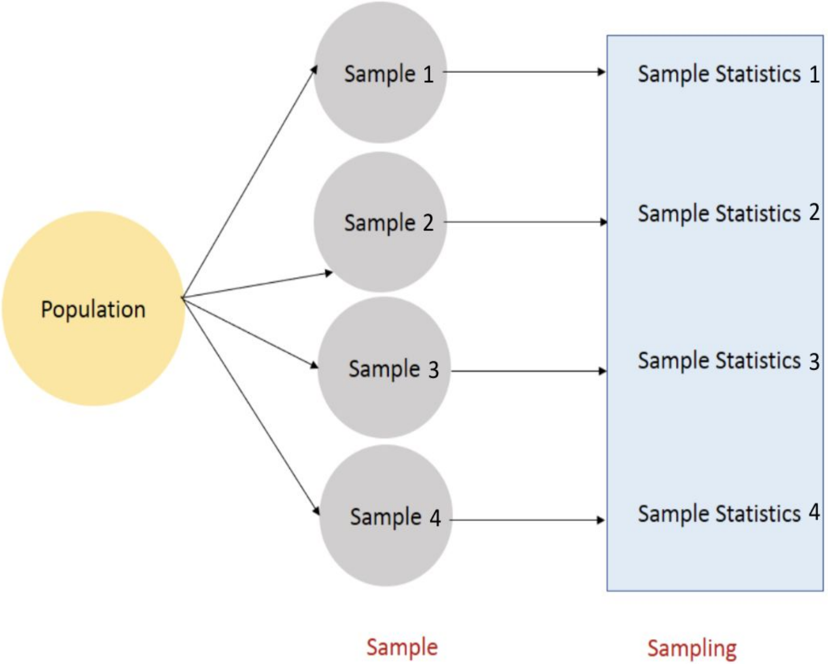
Sampling Distribution

This is the **distribution of a statistic** (like sample mean or proportion) **across many samples** from the same population.

✅ It tells you about the **variability of the sample statistic**.

✅ As the sample size increases, the estimated parameters gets closer to true parameter.

Law of Large Numbers



*We will learn about it later in detail.

Sampling Distribution

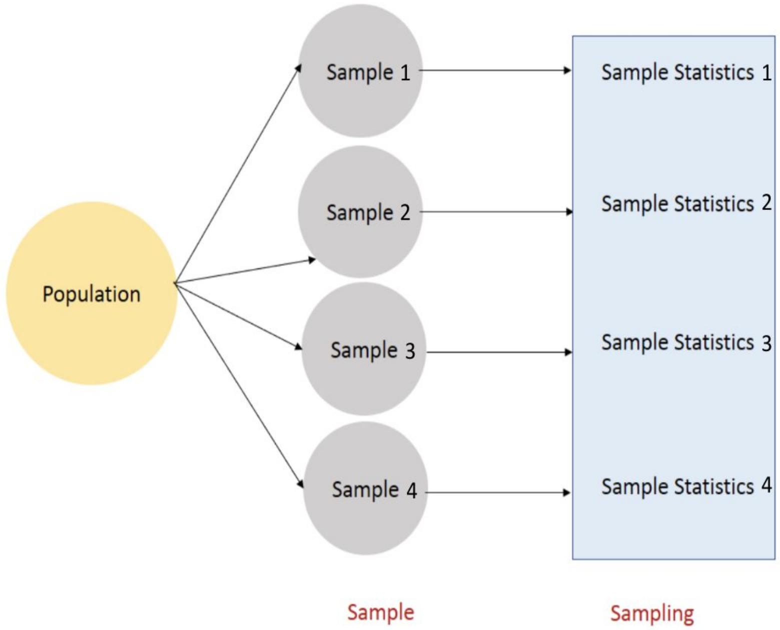
Newton School of Technology

This is the **distribution of a statistic** (like sample mean or proportion) **across many samples** from the same population.

> For example: You take **many samples of 50 people each**, compute the **mean height** in each sample — the distribution of those means is the **sampling distribution of the sample mean**.

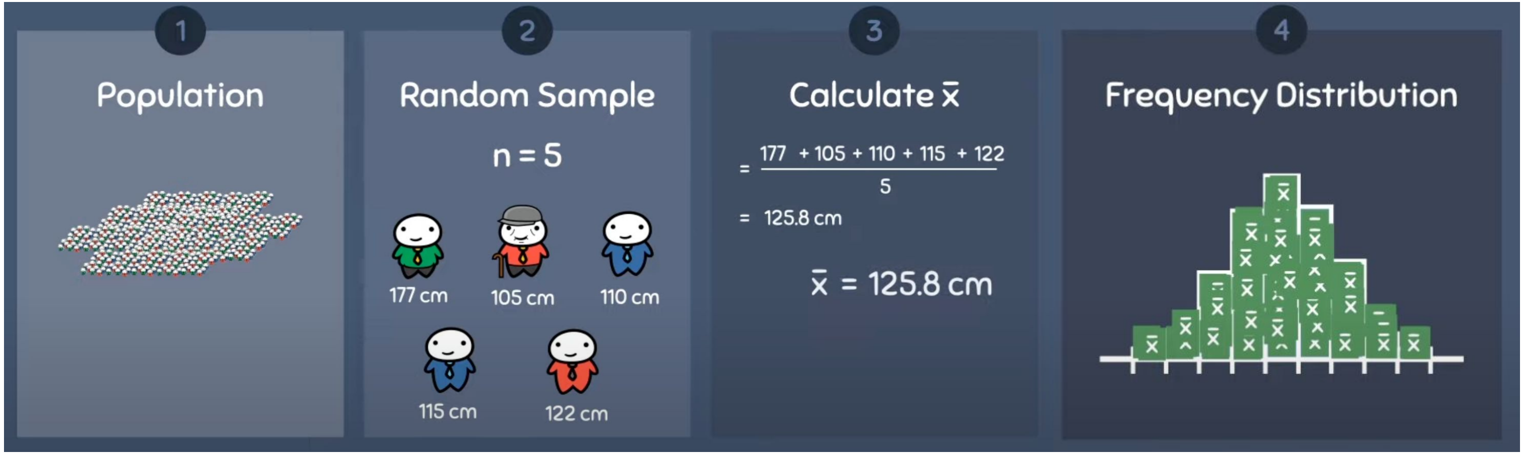
✅ It tells you about the **variability of the sample statistic**.

✅ Foundation of Advance concepts that we will learn later.



Sampling Distribution of Sample Mean

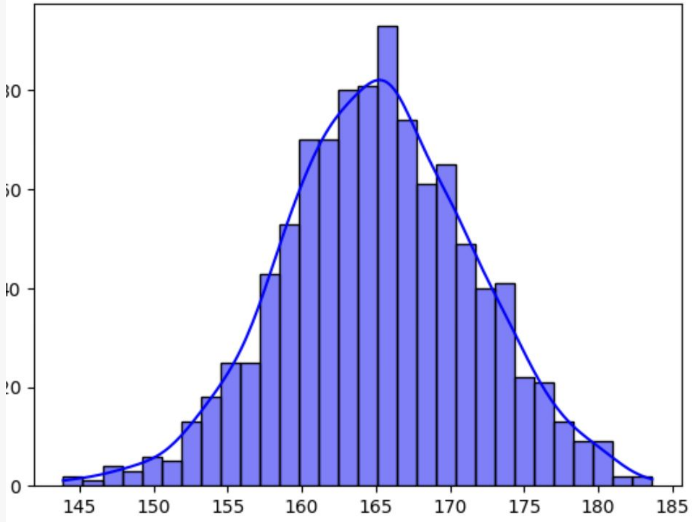
1. **Simulate a population** of 10,000 people where height follows a normal distribution (mean = 165 cm, std = 15 cm).
2. **Draw a random sample of size 5**, and calculate the sample mean height.
3. **Repeat** this process 1000 times and store all the sample means.
4. **Plot the histogram** of these 1000 sample means.
   Repeat for n=20, 30, 50.



Sampling Distribution – Simulation 1

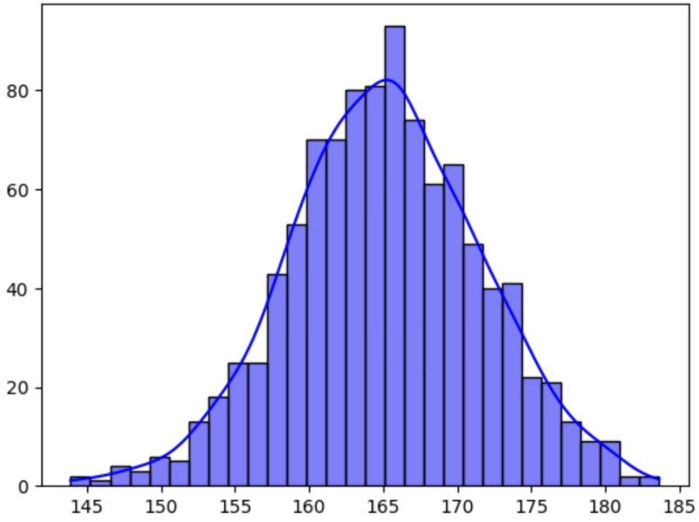
Newton School of Technology

```python
import seaborn as sns
import random
import numpy as np
import matplotlib.pyplot as plt
from statistics import mean

# Step 1: Simulating for 10000
population = np.random.normal(165, 15, 10000)

# Step 2: Function to collect 1000 sample means (sample size = 5)
n = 5
reps = 1000
sample_means = []

for _ in range(reps):
    sample = random.sample(list(population), n)
    sample_means.append(mean(sample))

# Step 3: Plot histogram of the 1000 sample means
sns.histplot(sample_means, bins=30, kde=True, color='skyblue')
plt.show()
```



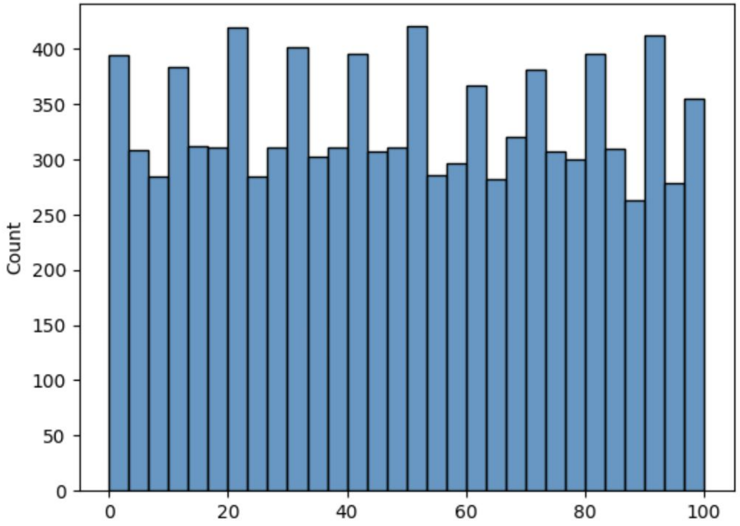
Sampling Distribution – Simulation 1

```
import random
import seaborn as sns
from scipy.stats import norm

# Step 1: Simulate population of 10,000 people (mean=165, std=15)
population = norm.rvs(loc=165, scale=15, size=10000)

# Step 2: Function to collect sample means
def get_sample_means(pop, n, reps=1000):
    return [sum(random.sample(list(pop), n)) / n for _ in range(reps)]

# Step 3: Try different sample sizes n = 5, 20, 50
n= 5
means = get_sample_means(population, n)
sns.histplot(means, bins=30, kde=True, color='skyblue')
```



Sampling Distribution – Not Normal Population

This time Population is not Normal.



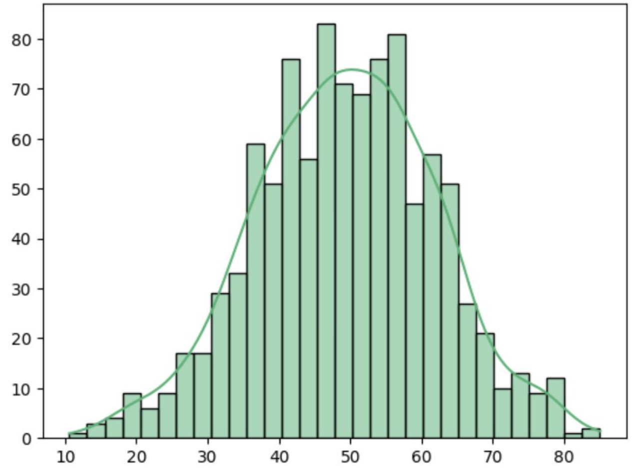
Let say this is how the population is distributed.
How do you think the sampling Distribution will look?

```
population = [random.randint(0, 100) for _ in range(10000)]
sns.histplot(population, bins = 30)
```

Sampling Distribution – Simulation 2

Newton School of Technology

```python
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from statistics import mean

# Step 1: Simulate data with a uniform distribution (range 0 to 100)
population = np.random.uniform(0, 100, 10000)  # Generating 10000 samples

# Step 2: Collect 1000 sample means (each sample has 5 observations)
n = 5  # sample size
reps = 1000  # number of trials
sample_means = []

for _ in range(reps):
    sample = np.random.choice(population, size=n, replace=False)
    sample_means.append(mean(sample))

# Step 3: Plot histogram of the 1000 sample means
sns.histplot(sample_means, bins=30, kde=True, color='skyblue')
plt.show()
```



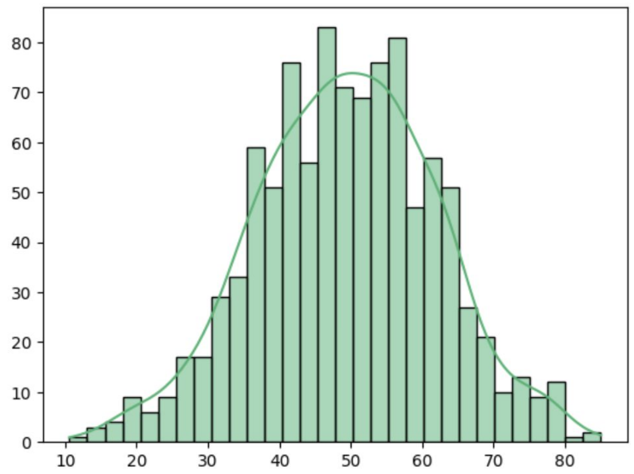
✅ Doesn't matter even if my Population is Normal or not, the sampling distribution of the mean looks approximately Normal!

❓ What happens if we increase the sample size to 20, 30, or 50? Will the distribution become even narrower and more Normal?

Sampling Distribution – Simulation 2

Newton School of Technology

```python
import random
import seaborn as sns
from scipy.stats import uniform

# Step 1: Random Test Scores between 0 to 100
population = [random.randint(0, 100) for _ in range(10000)]

# Step 2: Function to get sampling distribution of means
def get_sample_means(pop, sample_size, repeats=1000):
    return [sum(random.sample(pop, sample_size)) / sample_size for _ in range(repeats)]

# # Step 3: Plot for different sample sizes
n = 5
means = get_sample_means(population, n)
sns.histplot(means, bins=30, kde=True, color='mediumseagreen')
```



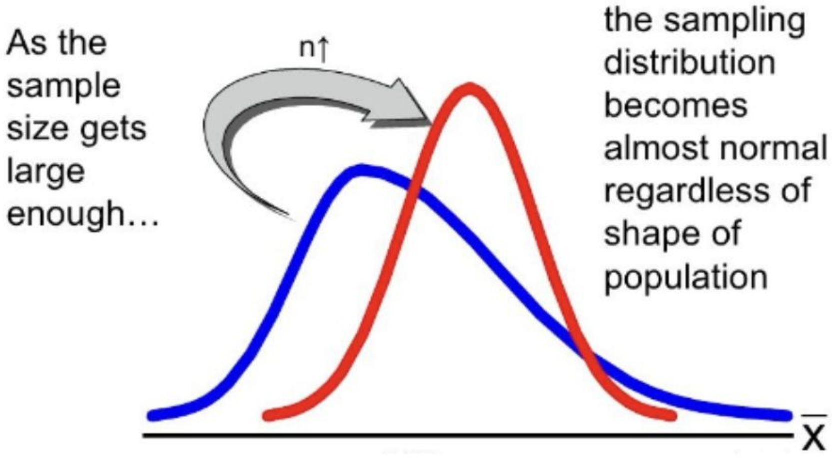
✅ Doesn't matter even if my Population is Normal or not, the sampling distribution of the mean looks approximately Normal!

❓ What happens if we increase the sample size to 20, 30, or 50? Will the distribution become even narrower and more Normal?

Central Limit Theorem

Newton School of Technology

CLT: The distribution of sample means follows a Normal Distribution, even if individual values don't, as long as $n \geq 30$.



Why CLT is important:

- Pfizer cannot test on every human, so they test small groups.
- Sample means help estimate the true average effectiveness.

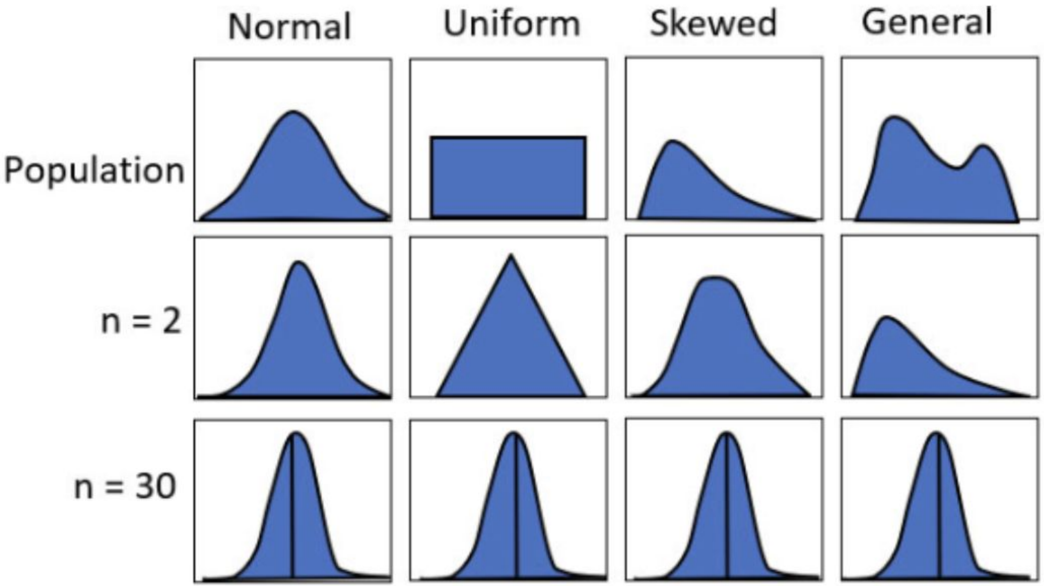
CLT – Continued

Key Properties:

Sample means follow a Normal Distribution (even if population isn't Normal).

Mean of sample means ≈ Population Mean (unbiased estimates).

Larger samples ($n \geq 30$) → More stable & accurate estimates.



CLT – Conditions

Newton School of Technology

To apply the central limit theorem, the following conditions must be met:

1. **Randomization**:
   - Data should be randomly sampled, ensuring every population member has an equal chance of being included.
2. **Independence**:
   - Each sample value should be independent, with one event's occurrence not affecting another.
   - Commonly met in probability sampling methods, which independently select observations.
3. **Large Sample Condition**:
   - A sample size of 30 or more is generally considered "sufficiently large."
   - This threshold can vary slightly based on the population distribution's shape.

Practice – 1

Newton School of Technology

For each population distribution described below, **which of the following would likely produce a sampling distribution that is approximately normal?**

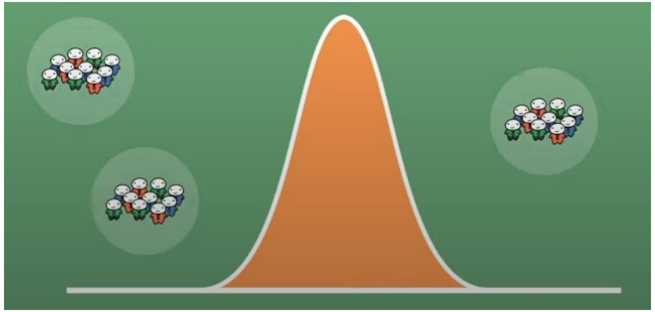
a) Rectangular population distribution, sample size = 15
b) Bimodal population distribution, sample size = 29
c) Skewed population distribution, sample size = 40
d) Triangular population distribution, sample size = 35
e) Normal population distribution, sample size = 20
f) Normal population distribution, sample size = 30

Ans: **c, d, e, f**

Sampling Dist. vs Population Dist.

Newton School of Technology

SAMPLING DISTRIBUTION



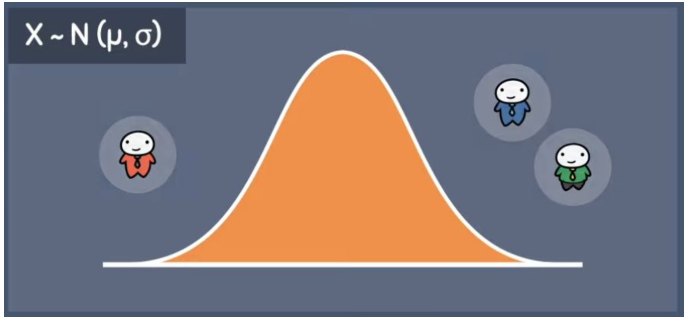
$\mu_{\bar{x}} = \mu$

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

$$z = \frac{x - \mu}{\sigma / \sqrt{n}}$$

$\sigma_{\bar{x}} < \sigma$

POPULATION DISTRIBUTION



μ

σ

$$z = \frac{x - \mu}{\sigma}$$

Standard Error

Standard Error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- Standard Error (SE) measures how much the sample mean fluctuates from the true mean.

Why SE Matters?

- Standard Error tells us **how close** our sample mean is likely to be to the true population mean.

Key Observations:

- Larger samples (higher n) → Lower SE → More accuracy.
- Higher variability (higher σ) → Higher SE → Less accuracy.

Real-life Example:

- Testing 10 vaccines → SE is large, the sample mean is less reliable.
- Testing 1000 vaccines → SE is small, sample mean is very close to 20 hours.

📢 Pfizer can reduce error in estimates by increasing sample size.

Practice – 2

The average time a laptop battery lasts is **6 hours** with a standard deviation of **1.2 hours**. If a sample of **25 laptops** is tested, what is the probability that their **average battery life** is **less than 5.5 hours**?

Solution

Newton School of Technology

Step 1: Compute Standard Error (SE)

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{25}} = \frac{1.2}{5} = 0.24$$

Step 2: Convert to Z-score

$$Z = \frac{\bar{X} - \mu}{SE} = \frac{5.5 - 6}{0.24} = \frac{-0.5}{0.24} \approx -2.08$$

Step 3: Find Probability from Z-table

$$P(Z < -2.08) \approx 0.0188$$

Practice – 3

Suppose the population of student study hours follows a normal distribution with a mean (μ) of 6 hours and standard deviation (σ) of 2 hours.

If we take multiple random samples of size 25 and compute the sample mean for each:

Q1: What will be the mean of the sampling distribution of the sample mean?
Q2: What will be the standard deviation of the sampling distribution (i.e., the standard error)?

Q1:

What is the **mean of the sampling distribution?**

$$\mu_{\bar{x}} = \mu = 6 \text{ hours}$$

Q2:

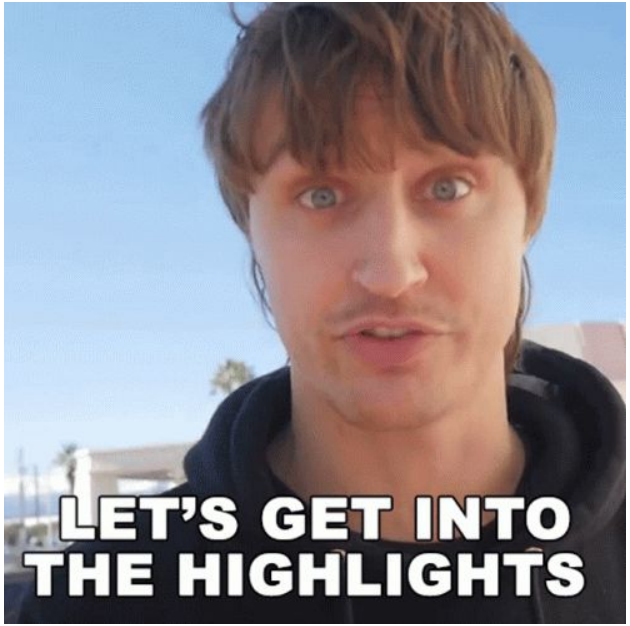
What is the **standard error (SE)**?

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{25}} = \frac{2}{5} = 0.4$$

Key Takeaways

Newton School of Technology

1. **Central Limit Theorem (CLT)**
   - Sample means tend to follow a **Normal Distribution**, even if the original population isn't normal (as long as $n \geq 30$).
   - Mean of sample means ≈ population mean.
   - Larger samples → smaller spread (lower standard error) → more accurate estimates.
2. **Standard Error (SE)**
   - SE measures how much sample means vary from the population mean.
   - Formula: $SE = \sigma / \sqrt{n}$
   - Bigger **n** → smaller SE → more reliable estimates.
3. **Bootstrapping**
   - Resampling **with replacement** from a single sample to simulate the process of repeated sampling.
   - Helps approximate the sampling distribution **without collecting new data**.
   - Boosts confidence in estimates, especially when actual data collection is costly or limited.
4. **Sampling Distribution vs. Population Distribution**
   - Population distribution: actual data distribution
   - Sampling distribution: distribution of **statistic** (e.g., mean) across many samples



References | Homework

Exercises:

Beginner: Try simulations using Python/R for 5–10 bootstrap samples
Intermediate: Plot histograms from 100 bootstrap samples
Advanced: Simulate CLT with different population shapes and increasing n

Additional Resources

1. Khan Academy - CLT Explanation
2. Seeing Theory - CLT Simulation
3. Blog: Bootstrapping in Statistics

Take-Home

- Run your own bootstrap on small dataset (e.g., your test scores)
- Observe how the mean varies across resamples
- **Think:** How does the shape of the distribution change with sample size?