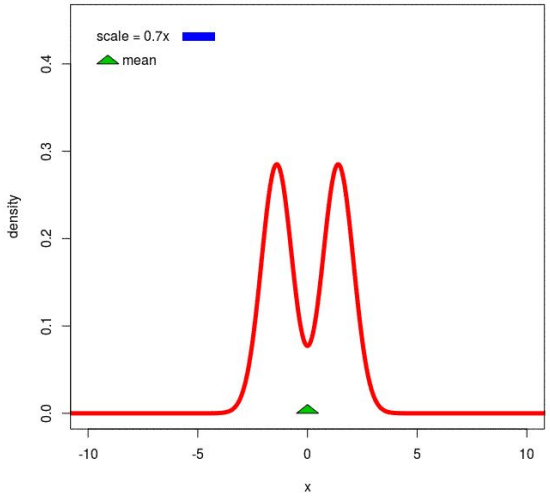
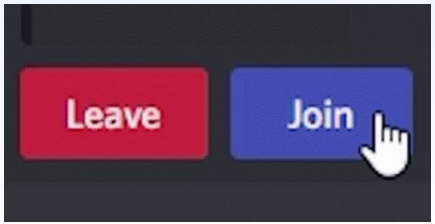


Newton School of Technology

Module *4* - Inferential Statistics

1

Sampling Techniques, Parameter Estimation, Sampling Distribution



Describing Data → Making Decisions

Newton School of Technology

What is Descriptive Statistics? - Summarizes and organizes data we already have

Helps us answer:

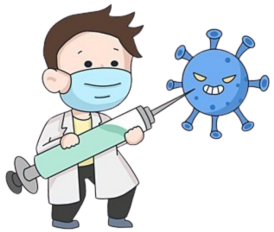
- “What does the data look like?”
- “What is typical or common?”
- “Is the data tightly packed or spread out?”

What if we don’t have full data.
Can we draw any conclusions beyond the data collected.

Suppose we found that 9 out of 10 students in this class love statistics…
Can we say 90% of all students love statistics too? 🤔
That’s where **inferential statistics** enters the picture.

Inferential Statistics – Motivation

Newton School of Technology

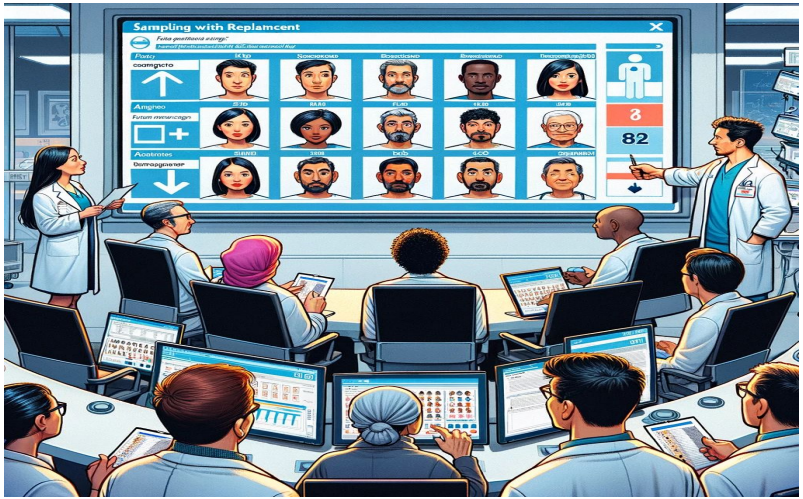


Pfizer developed a new covid vaccine to treat deadly coronavirus

You as Clinical trial specialist, need to devise a plan to release this to market

How can we achieve this?

To access the efficacy of the vaccine, we can conduct clinical trials involving thousands of participants.



Inferential Statistics – Motivation

Newton School of Technology

Is Testing Everyone Practical?

Impractical, extremely expensive, and time-consuming
Ethical issues and potential health risks

How can Pfizer get reliable data cost-effectively and ethically, without testing the entire population?

Use a Sample: Instead of testing everyone, select a smaller group of people that represents the entire population.

Sampling Examples

Newton School of Technology

1. Imagine you want to know the average height of people in India. Would you measure everyone?
2. When you're cooking soup, do you drink the whole pot to check if it needs more salt?
3. How do news channels predict election results before counting is over?
4. A/B Testing in Tech Companies

Population vs Sample

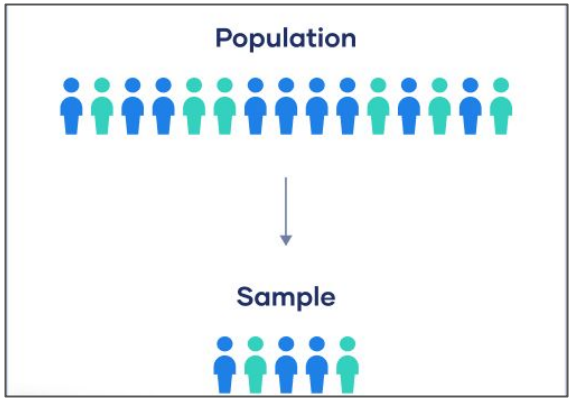
Newton School of Technology

Population: The entire set of items we study.

Example: All people globally who could receive Pfizer's vaccine.

Sample: A subset of the population, used for analysis.

Example: A smaller group chosen to represent the larger population.



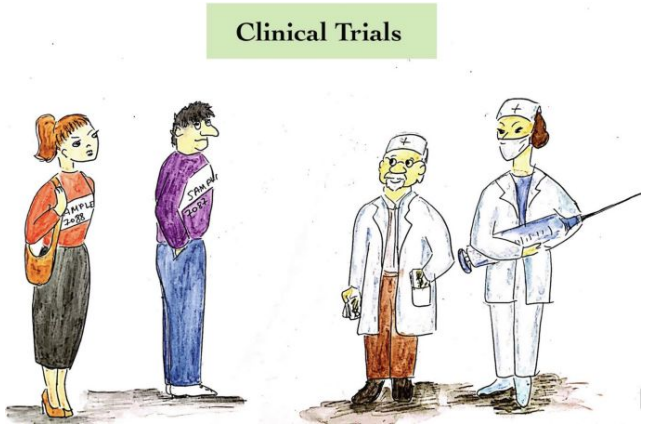
What is a right sample?

Newton School of Technology

Imagine if Pfizer **only tested** its vaccine on **young, healthy volunteers**.
What Happens?

- Results might show very **high efficacy**, but only for that narrow demographic.
- **Misleading Conclusion:**
  - The vaccine appears extremely effective overall, yet **it might fail** to protect older adults or those with health conditions.

Hence, we need **good sample**. A good sample **captures the diversity of the entire population** (age, risk factors, etc.).



Quiz Time

Newton School of Technology

Q1) A company wants to find the most popular ice cream flavor among teenagers. They set up a poll at a luxury ice cream parlor in a high-end mall.

Question:

"What kind of bias could occur here? Do you think this represents all teenagers fairly?"

Q2) A fitness app wants to collect user feedback. They send a survey only to users who completed 30 workouts in the last month.

Question for students:

"Who is being excluded here? Will this give a full picture of what users think?"

Quiz Time Answers

Newton School of Technology

Q1) A company wants to find the most popular ice cream flavor among teenagers. They set up a poll at a luxury ice cream parlor in a high-end mall.

Question:
"What kind of bias could occur here? Do you think this represents all teenagers fairly?"

Answer- The poll is conducted at a luxury ice cream parlor, which likely attracts a certain group-teenagers who will be able to afford it,the results will not represent who may prefer more affordable

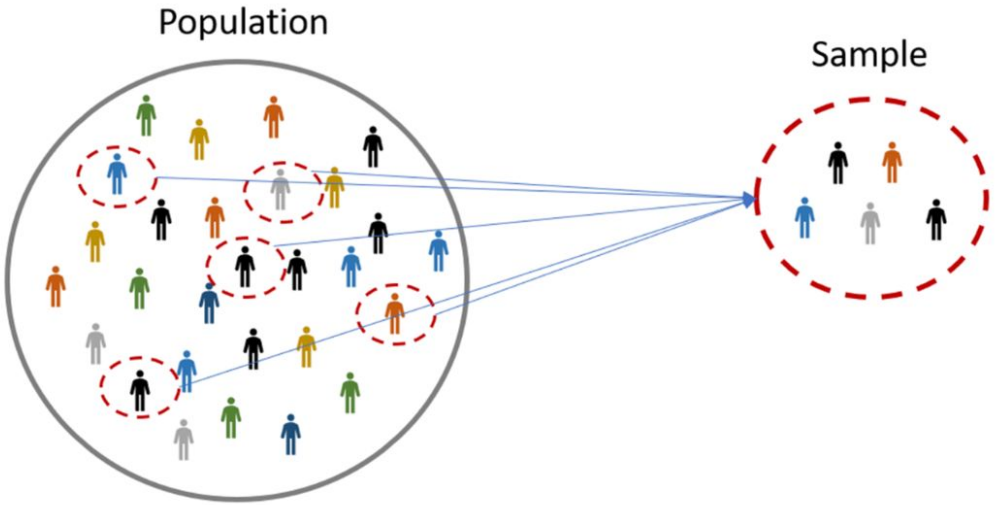
Q2) A fitness app wants to collect user feedback. They send a survey only to users who completed 30 workouts in the last month.

Question for students:
"Who is being excluded here? Will this give a full picture of what users think?"

Answer- Surveying only highly active users excludes those who are inactive or semi-active—the very users whose feedback could highlight app problems or reasons for low engagement.

Simple Random Sampling (SRS)

Newton School of Technology



- Everyone has an **equal chance of selection** (e.g., computer-generated lottery).
- Pros: **Easy, fair, unbiased**
- If we pick people completely at random... is there anything that could still go wrong?

Survey in your campus? Mostly Engineering students!!!

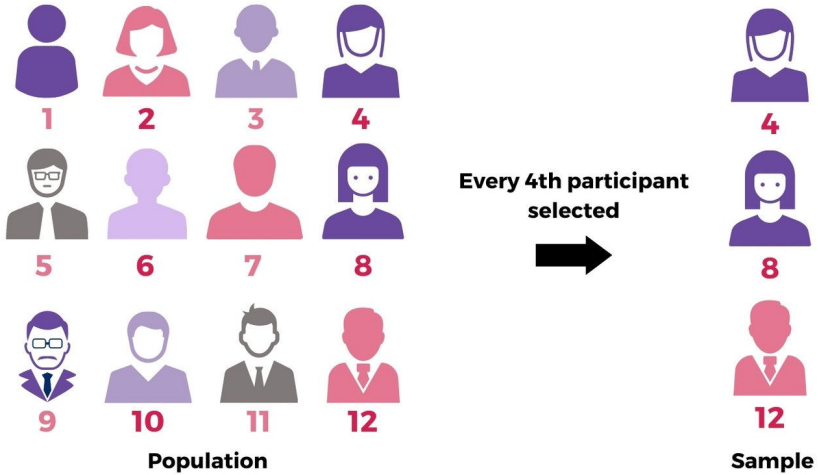
Stratified Sampling

Newton School of Technology

- Divide the population **into groups (strata) based on characteristics** (age, health, etc.).
- Randomly select participants within each stratum.
- Pros: **Highly representative, captures diverse segments**
- What could go wrong in this kind of sampling?
  What if we divide the groups wrongly—or miss an important group altogether?

Systematic Sampling

Newton School of Technology



- **Select every n^{th} individual** from a list or queue after choosing a random starting point.
- Ensure the list is **not ordered** in a way that introduces bias.
- Pros: Simple to implement, time-efficient, **requires less planning than stratified sampling**
- Cons: If the population list isn't randomized, it might not capture certain subgroups accurately.

Common Sampling Techniques

Newton School of Technology

Feature	Simple Random Sampling	Stratified Sampling	Systematic Sampling
Definition	Every individual in the population has an equal chance of selection	Population is divided into subgroups (strata); random samples taken from each	Selection at regular intervals from an ordered list
When to Use	Population is fairly homogeneous and a full list is available	Population has distinct subgroups that need proportional representation	A complete list exists and a fast, evenly spread sample is needed
How It Works	Use random number generator or lottery method	Divide by characteristics (e.g., age, gender), sample each group randomly	Choose a random starting point, then pick every k-th item
Bias Risk	Low, if truly random and list is complete	Very low, especially if strata are well defined	Moderate — may introduce bias if there's a hidden pattern in the list
Pros	Easy to understand, unbiased if truly random	High accuracy, ensures representation from all subgroups	Simple, quick, good for large populations
Example	Selecting 50 students randomly from the entire college	Sampling 10 students from each department in a university	Surveying every 5th customer entering a store

Quiz Time

Newton School of Technology

Q1) What sampling technique should we use if 38% of the sample is selected from college-educated individuals and 62% from non-college-educated individuals, to match the population proportions?

Q2) What sampling technique should we use if a supermarket selects every 10th or 15th customer entering the store for a study?

Q3) What sampling technique should we use if we randomly select 25 employee names out of a hat from a company of 250 employees?

Quiz Time- Answers

Newton School of Technology

Q1) Stratified Sampling–The population is divided into strata (college-educated and non-college-educated), and sampling is done proportionally within each stratum. This ensures that the sample mirrors the population structure.

Q2) Systematic Sampling– In systematic sampling, participants are selected at regular intervals (e.g., every 10th or 15th customer). This method is simple to implement but assumes there's no hidden pattern in the order of customers.

Q3) Simple Random Sampling– Each employee has an equal chance of being selected. Randomly drawing names out of a hat is a classic example of simple random sampling.

Measuring & Interpreting Results

Newton School of Technology

Once we have our sample, we must figure out how to **measure and interpret the outcomes.**

But why do we measure and interpret?

We rarely know these true population values (parameters) because we **can't measure everyone**. Instead, we collect a sample and compute "sample estimates" like mean or variance to **approximate the parameters**.

Concept	Sample	Population
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Proportion	$\hat{p}$	p

Population Mean vs Sample Mean

Newton School of Technology

- **Population Mean (μ):**

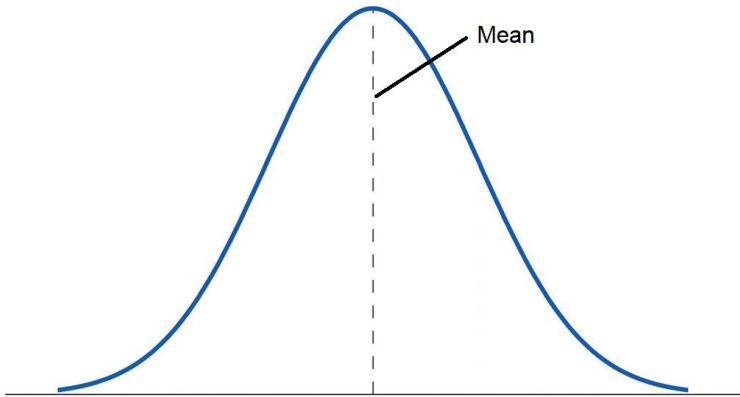
  The **average of all values** in the population.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Sample Mean ($\bar{x}$):**

  The **average of values** in the sample.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



Population Variance vs Sample Variance

Newton School of Technology

Do all people respond to the vaccine the same way?”--**Variance measures spread** in outcomes.

- **Population Variance (σ²):**

  The **average of the squared differences** from the population mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- **Sample Variance (s²):**

  The **average of the squared differences** from the sample mean, adjusted for bias.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Understanding Sample Variance

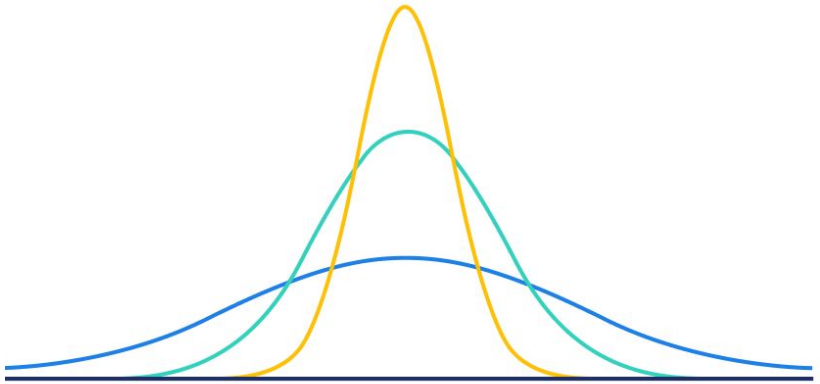
Newton School of Technology

Why do we divide **sample variance by (n−1)?**

Let's consider a Sample Data:
x = [5, 7, 3, 7, 8]
n = 5

What will be the value of mean and variance?

Mean = **6** and Variance = **3.2**



If I tell you that n = 5 and tell you the mean is 6, how many values can you choose freely?

Degrees of Freedom

Newton School of Technology

Sample size: $n = 5$ **Given mean:** $\bar{x} = 6$ **Total sum:** $5 \times 6 = 30$

Suppose you choose: $x_1 = 5,\ x_2 = 7,\ x_3 = 3,\ x_4 = 7$

Sum = $5 + 7 + 3 + 7 = 22$

To keep the total 30, the 5th value must be: $x_5 = 30 - 22 = 8$

Even though there are 5 values, only **4 are free to vary**.

The 5th value is **determined** by the rest to maintain the mean.

Dividing by **n** underestimates the true variance (it's too optimistic).
Dividing by **(n−1) corrects this bias**, giving a more **accurate estimate** of the **population variance**.
If we divide by n-1 it will be an unbiased estimate which is called **Degrees of Freedom**

Limitation of a Single Trial

Newton School of Technology

Is 1 Trial Sufficient to Understand Vaccine effectiveness?

A single sample might not capture the full diversity of the population.

Multiple samples reveal whether our results are consistent or just a fluke.

Different samples will naturally yield **slightly different estimates** (e.g., 88% vs. 90%)

Observing these variations helps us measure **sampling variability**

Interpreting Differences in Sample Outcomes

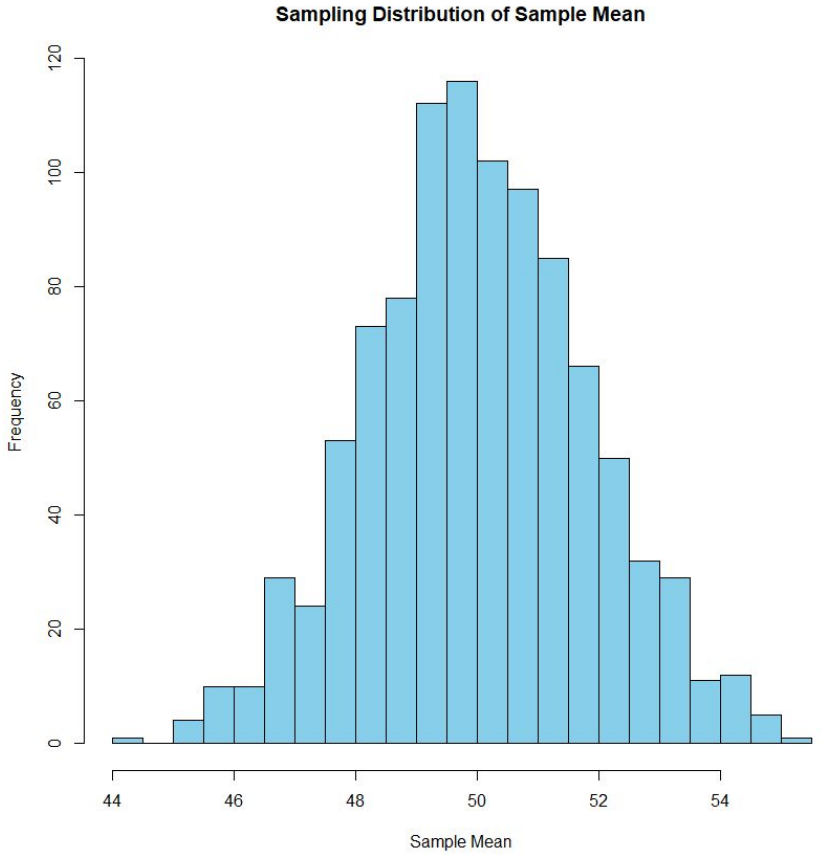
Newton School of Technology

How do we make sense of the fact that different samples give different results?

We can generate **Sampling distribution.**

What is Sampling Distribution?

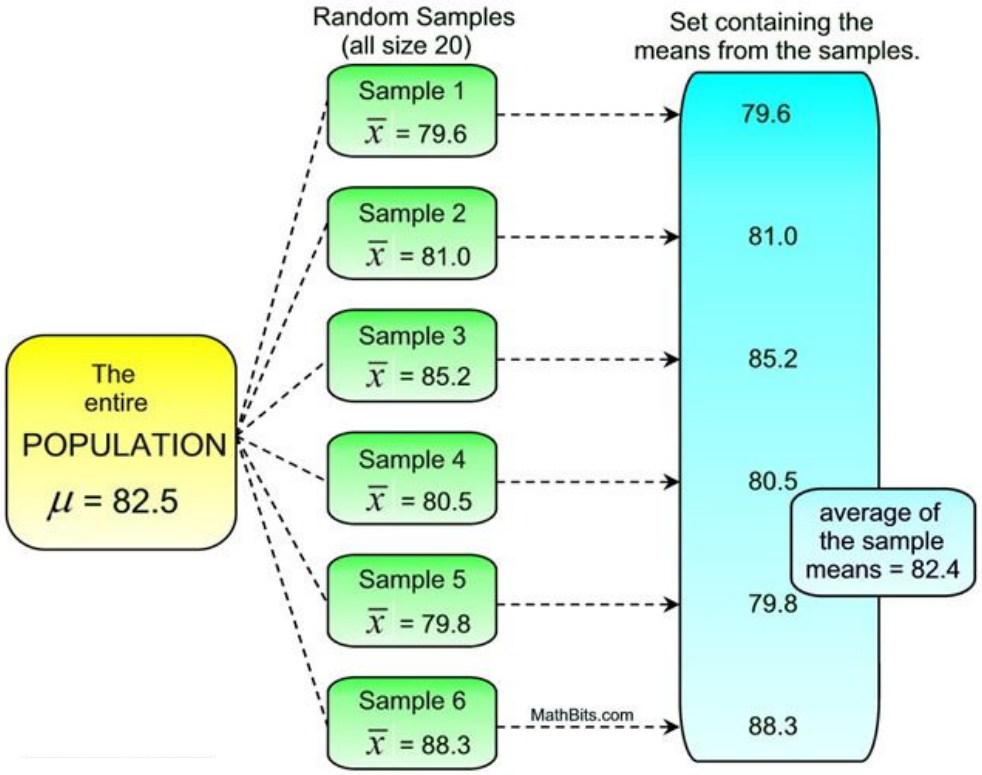
It is the **distribution of a statistic** across many possible samples which tells us how much the sample mean might **vary from sample to sample**.



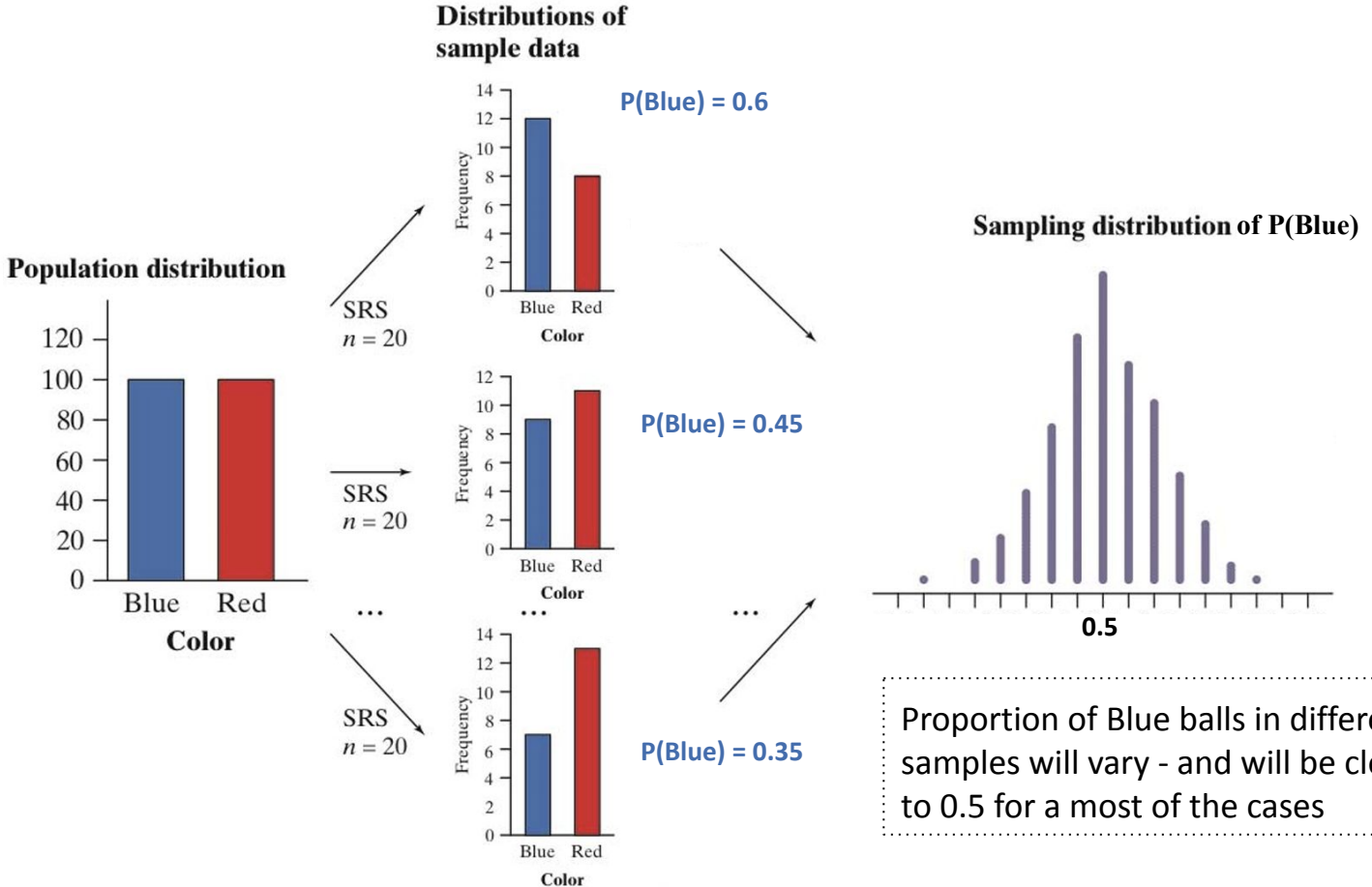
Sampling Distribution

Steps to Build a Sampling Distribution from given samples:

- Step 1: **List the Samples**
- Step 2: **Compute the Statistic for Each Sample**
- Step 3: **Plot & analyze the Sampling Distribution**



Sampling Distribution



Proportion of Blue balls in different samples will vary - and will be close to 0.5 for a most of the cases

Question-1

Newton School of Technology

Q1) You are given the following 6 random samples (each of size 3) drawn from a population:

Sample 1: [4, 6, 8]
Sample 2: [6, 6, 10]
Sample 3: [8, 10, 12]
Sample 4: [4, 10, 12]
Sample 5: [6, 8, 10]
Sample 6: [4, 6, 12]

Use the sample means to construct the sampling distribution.

Question-1 Answer

Newton School of Technology

Q1) You are given the following 6 random samples (each of size 3) drawn from a population:

Sample 1: [4, 6, 8]
Sample 2: [6, 6, 10]
Sample 3: [8, 10, 12]
Sample 4: [4, 10, 12]
Sample 5: [6, 8, 10]
Sample 6: [4, 6, 12]

Use the sample means to construct the sampling distribution.

Colab link
https://colab.research.google.com/drive/1KpBcim3sPIwoJkcM-SNgbs-vZUPNdyIZ#scrollTo=oWcSbn4lhlwq

Question-2

Newton School of Technology

Write Python code to create 30 different random samples using **np.random.randint.**

Each sample should contain 10 integer values.

You can choose any suitable range, for example, between 10 and 50.

Question-2 Answer

Newton School of Technology

Write Python code to create 30 different random samples using np.random.randint.

Each sample should contain 10 integer values.

You can choose any suitable range, for example, between 10 and 50.

Colab link
https://colab.research.google.com/drive/13oMxErq_yrb9rkfdhnuns0LYf2dHCFRB#scrollTo=OiL4SZ1GjpsV

Summary

Newton School of Technology

1. **Inferential statistics** helps us make **broader decisions** from **sample data** about the **entire population**.
2. A **sample** is a **subset** of the **population**, used to draw **conclusions** about the **whole**.
3. Common **types of sampling** (e.g., **Simple Random**, **Stratified**, **Systematic**) ensure a **representative sample**.
4. We measure **sample statistics** (e.g., **mean**, **variance**) to interpret and estimate **population parameters**.
5. A **single trial** may overlook **entire variability**, risking **incomplete** or **biased** outcomes.
6. **Sampling variability** arises because **different samples** yield **different estimates**.
7. A **sampling distribution** illustrates how a **statistic** (mean, variance, etc.) **varies** across **many samples**.

Newton School of Technology

See You Guys in Next Session :)