A. Population and Sample

1. Definitions and Differences

- **Population:** The entire group of individuals, objects, or measurements that we are interested in studying. It is denoted by the size N. For example, all 12th-grade students appearing for CUET in India constitute a population.
- **Sample:** A subset or a smaller collection of items selected from the population. It is denoted by the size n. For example, 1,000 CUET aspirants selected from different states.

Key Difference: The population represents the "complete truth" (which is often impossible to measure entirely), while the sample represents the "available evidence."

2. Representative vs. Non-Representative Samples

- **Representative Sample:** a sample that accurately reflects the characteristics of the entire population. To be representative, the sample must be free from **bias**.
- **Non-Representative (Biased) Sample:** A sample that does not reflect the population. For instance, if you only sample students from elite private schools to estimate the average CUET score of all Indian students, your sample is non-representative.

3. Sampling Techniques

To ensure representativeness, we use random sampling methods:

- **Simple Random Sampling (SRS):** Every member of the population has an equal and independent chance of being selected. This is the gold standard of sampling, rooted in the concept of **equally likely outcomes** discussed in probability theory.
- **Systematic Random Sampling:** This involves selecting every k^{th} individual from a list of the population.
  - **Formula for Interval (k):** $k = \frac{N}{n}$
  - **Procedure:** A random starting point between 1 and k is chosen, and then every k^{th} member is picked thereafter.

B. Parameter, Statistics, and Statistical Inferences

1. Parameter vs. Statistic

- **Parameter:** A numerical value that describes a characteristic of the **population**. Parameters are usually denoted by Greek letters.
  - Population Mean: μ
  - Population Standard Deviation: σ
- **Statistic:** A numerical value that describes a characteristic of a **sample**. Statistics are denoted by Roman letters.
  - Sample Mean: $\bar{x}$
  - Sample Standard Deviation: s

2. The Relation and Limitations

The primary goal of inferential statistics is to use the **Sample Statistic** to estimate the **Population Parameter**.

The Limitation: A statistic is only an **estimate**. Because a sample does not include every member of a population, there is always a **Sampling Error**. Even if the sampling is perfectly random, different samples will produce slightly different statistics.

3. Central Limit Theorem (CLT)

(Note: This concept is foundational but not detailed in the sources. Please verify independently.) The CLT is the most important theorem in inferential statistics. It states that for a large enough sample size (usually $n \geq 30$), the **sampling distribution of the sample mean** will be approximately **normal**, regardless of the shape of the population distribution.

- **Mean of Sampling Distribution:** $\mu_{\bar{x}} = \mu$

- **Standard Deviation of Sampling Distribution (Standard Error):** $SE = \frac{\sigma}{\sqrt{n}}$

4. Statistical Significance and Inference

Statistical Inference is the process of using data analysis to deduce properties of an underlying probability distribution.

- **Statistical Significance:** This indicates that the result of a study is unlikely to have occurred by chance. It is usually measured by a **p-value**. If the p-value is less than a predetermined level (usually 0.05), we say the result is "statistically significant."

C. t-Test (One Sample t-test for Small Groups)

The t-test is used when the sample size is small ($n < 30$) and the population standard deviation (σ) is unknown. It was developed by William Sealy Gosset (under the pseudonym "Student").

1. Defining a Hypothesis

A **hypothesis** is a claim or statement about a population parameter.

- **Null Hypothesis (H_0):** The hypothesis of "no effect" or "no difference." We assume this is true until we have enough evidence to reject it. (e.g., $H_0 : \mu = 50$).
- **Alternate Hypothesis (H_a or H_1):** The statement we want to prove. It contradicts the null hypothesis. (e.g., $H_a : \mu \neq 50$ or $H_a : \mu > 50$).

2. Degree of Freedom (df)

The **degree of freedom** represents the number of values in the final calculation of a statistic that are free to vary. For a one-sample t-test, it is calculated as: $df = n - 1$

3. The t-Test Statistic

To test the null hypothesis, we calculate the t-score using the sample data. **Formula for One-Sample t-test:** $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ Where:

- $\bar{x}$ = Sample Mean
- μ = Population Mean (under the null hypothesis)
- s = Sample Standard Deviation
- n = Sample Size

4. Making Inferences

Once the t-value is calculated, it is compared to a **Critical Value** from a t-distribution table using the calculated df and a chosen level of significance (α).

- **Decision Rule:**
  - If $|t_{calculated}| > t_{critical}$, we **reject** the Null Hypothesis.
  - If $|t_{calculated}| \leq t_{critical}$, we **fail to reject** the Null Hypothesis.

Summary Table for Unit VI

Concept	Definition / Formula

Population (N)	Complete set of data points.
Sample (n)	Subset used for observation.
Parameter	Population measure (μ, σ).
Statistic	Sample measure ($\bar{x}, s$).
Standard Error (SE)	$\frac{\sigma}{\sqrt{n}}$ (Measures sample mean variation).
Degrees of Freedom (df)	$n-1$
Null Hypothesis (H_0)	The "Status Quo" or "No Difference" claim.
t-Statistic	$\frac{\bar{x}-\mu}{s/\sqrt{n}}$