Sample Statistic
- It describes the characteristic of a sample.
- Sample statistic itself is a random variable.

Sample
A subgroup of population.

Methods of Sampling
- Simple Random Sampling
  - Each item of the population under study has equal probability of being selected.
  - There is no guarantee of selection of items from a particular category.
- Stratified Random Sampling
  - Uses a classification system.
  - Separates the population into strata (small groups) based on one or more distinguishing characteristics.
  - Take random sample from each stratum.
  - It guarantees the selection of items from a particular category.
- Systematic Sampling
  - Select every kth number.
  - Resulting sample should be approximately random.

Sampling error
Sample – Corresponding Statistic Population Parameter.

Sampling Distribution
Probability distribution of all possible sample statistics computed from a set of equal size samples randomly drawn.

Standard Error (SE) of Sample Mean
- Standard deviation of the distribution of sample means.
  \[ \sigma_s = \frac{\sigma}{\sqrt{n}} \]
- If \( \sigma \) is not known then;
  \[ s_s = \frac{s}{\sqrt{n}} \]
- As \( n \uparrow \), \( \bar{X} \) approaches \( \mu \) and S.E \( \downarrow \).

Student’s T-Distribution
- Bell shaped.
- Shape is defined by df
- df is based on ‘sample size’.
- Symmetrical about it’s mean.
- Less peaked than normal distribution.
- Has fatter tails.
- More probability in tails i.e., more observations are away from the centre of the distribution & more outliers.
Central Limit Theorem (CLT)
For a random sample of size 'n' with:
- population mean $\mu$,
- finite variance (population variance divided by sample size) $\sigma^2$, the sampling distribution of sample mean $\bar{X}$ approaches a normal probability distribution with mean $\mu$ & variance as 'n' becomes large.

Properties of CLT
- For $n \geq 30 \Rightarrow$ sampling distribution of mean is approx. normal.
- Mean of distribution of all possible samples = population mean $\mu$.
- Variance of distribution $= \frac{\sigma^2}{n}$

Confidence Interval (CI) Estimates
- Results in a range of values within which actual parameter value will fall.
- $PE \pm (\text{reliability factor} \times SE)$.
- $\alpha =$ level of significance.
- $1-\alpha =$ degree of confidence.

Point Estimate (PE)
- Single (sample) value used to estimate population parameter.
$\bar{X} = \frac{\sum X}{n}$

Estimator: Formula used to compute PE.

Desirable properties of an estimator
Unbiased
Expected value of estimator equals parameter e.g., $E(\bar{X}) = \mu$ i.e., sampling error is zero.

Efficient
If $\text{var}(\bar{X}_1) < \text{var}(\bar{X}_2)$ of the same parameter then $\bar{X}_1$ is efficient than $\bar{X}_2$

Consistent
As $n \rightarrow \infty$, value of estimator approaches parameter & sample error approaches '0' e.g., $\bar{X} \rightarrow \mu$ & $SE \rightarrow 0$

CLT applies only when sample is random.
Biases

- **Data Mining Bias**: Statistical significance of the pattern is overestimated because the results were found through data mining.
- **Sample Selection Bias**: Systematically excluding some data from analysis. • It makes the sample non-random.
- **Look–ahead Bias**: Using sample data that wasn’t available on the test date.
- **Time-period Bias**: Time period over which the data is gathered is either too short or too long.

**Data Mining**
- Using the same data to find patterns until the one that ‘works’ is discovered.

**Survivorship Bias**
- Most common form of sample selection bias.
- Excluding weak performances.
- Surviving sample is not random.

**Warning Signs of Data Mining**
- Evidence of testing many different, mostly unreported variables.
- Lack of economic theory consistent with empirical results.

---

### Issues Regarding Selection of Appropriate Sample Size

- As \( n \) increases, \( s.e. \) decreases & hence C.I is narrow.

### Limitations of Large Sample Size

- Large sample may include observations from more than one population.
- Cost may increase more relative to an increase in precision.

---

### Distribution

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Variance</th>
<th>Sample</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Non normal</td>
<td>Known</td>
<td>Unknown</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>

*The z-statistic is theoretically acceptable here, but use of the t-statistic is more conservative.*