DegreeLive

B.Com Honours

Semester I

Calicut University

Essential Statistics for Business Analytics

Course Code: COM1MN109 • Module 1 Notes

1. Introduction to Sampling and Sampling Theory

In the digital age, businesses collect massive amounts of data daily. However, analyzing entire populations (censuses) is often practically impossible due to cost, time, and logistical constraints. Sampling theory provides the mathematical framework to select a representative subset of a population, analyze it, and make reliable statistical inferences about the entire population. This module covers population vs. sample, parameters vs. statistics, probability and non-probability sampling techniques, sampling and non-sampling errors, confidence intervals, and the Central Limit Theorem (CLT).

Population vs. Sample

- **Population:** The complete collection of all elements (individuals, transactions, items) under study. E.g., all credit card holders of a bank.
- **Sample:** A representative subset of the population selected for analysis. E.g., 1,000 credit card holders selected at random.
- **Parameter vs. Statistic:** A *parameter* is a numerical characteristic of a population (e.g., population mean μ). A *statistic* is a numerical characteristic of a sample (e.g., sample mean $\bar{x}$).

Sampling Errors vs. Non-Sampling Errors

Sampling Error

The difference between a sample statistic and the true population parameter that arises purely due to chance in selecting the sample. It is unavoidable but can be reduced by increasing the sample size.

Non-Sampling Error

Errors that occur at any stage of data collection or processing (e.g., biased questionnaire, respondent lying, data entry errors, non-response). Can occur in both census and sample surveys; cannot be reduced by increasing sample size.

Probability vs. Non-Probability Sampling

Sampling techniques are divided based on whether selection is random and mathematically equal:

Type	Technique Name	Description and Business Use
Probability	Simple Random Sampling (SRS)	Every member of the population has an equal chance of selection. E.g., computer-generated random numbers.
Probability	Stratified Random Sampling	Dividing population into homogeneous groups (strata) based on features (e.g., income), then sampling randomly from each.
Probability	Systematic Sampling	Selecting every k-th item from a ordered list after a random starting point. E.g., checking every 10th product on assembly line.
Probability	Cluster / Multi-stage	Dividing population into heterogeneous clusters (e.g., geographic regions), randomly selecting clusters, and sampling within.
Non-Probability	Convenience & Quota	Convenience: selecting easily accessible items. Quota: selecting items to meet pre-determined demographic targets (non-random).

2. Central Limit Theorem (CLT) and Confidence Intervals

The Central Limit Theorem is the foundation of inferential statistics, justifying the use of normal distribution curves in business forecasting.

Central Limit Theorem (CLT)

Statement: As the sample size n increases (typically $n \geq 30$), the sampling distribution of the sample mean ($\bar{x}$) approaches a normal distribution, regardless of the shape of the parent population distribution. The mean of this sampling distribution equals the population mean (μ), and its standard deviation (known as the **Standard Error**) is given by:

[ext{Standard Error (SE)} = □rac{sigma}{sqrt{n}}]

Where σ is the population standard deviation. CLT allows analysts to compute probabilities and test hypotheses for large samples even when the underlying population distribution is skewed or unknown.

Confidence Intervals

While a point estimate provides a single value (e.g., sample mean), an interval estimate provides a range of values within which the true population parameter is expected to lie with a specified probability (confidence level, e.g., 95% or 99%).

For large samples, the Confidence Interval for the population mean is calculated as:

[ext{Confidence Interval} = □ar{x} pm z left(□rac{sigma}{sqrt{n}} ight)]

Where: **z** is the critical value from the normal distribution table (z = 1.96 for 95% confidence level, z = 2.58 for 99% confidence level).