

UNIT 5

Sample Survey

(6 Hours) — Probability and Statistics (CACS202)

1. Concept of Population and Sample

1.1 Population and Sample

A **population** (universe) is the complete collection of all units — people, households, firms, items — that are the subject of a study. A **sample** is a part (subset) of the population selected to obtain information about the whole. The number of units in a population is **N** and in a sample **n**.

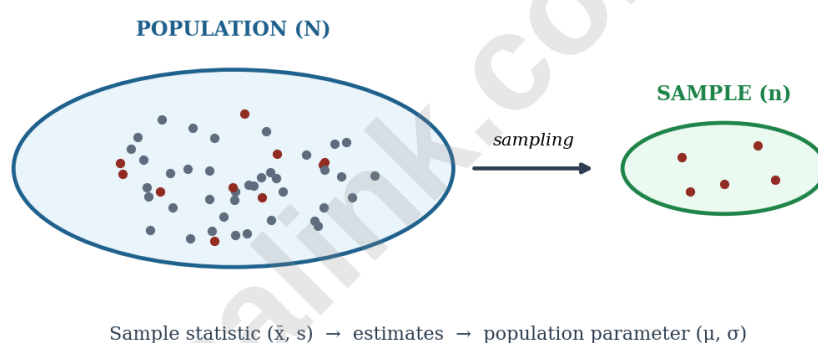


Figure 5.1 — From population to sample and back through inference

Term	Meaning	Example
Finite / infinite population	Countable number of units vs. unlimited units	Students of a college / all possible coin tosses
Parameter	A numerical characteristic of the population (μ , σ , P)	Average income of all households of a ward
Statistic (estimate)	A numerical characteristic computed from a sample ($\bar{x}$, s , p)	Average income of 100 sampled households
Sampling unit	Element or group of elements considered for selection	Household, individual, school
Sampling frame	Complete list of all sampling units from which the sample is drawn	Voter list, student roll
Sampling design	The plan/method of selecting the sample and estimating parameters	Stratified random sampling with $n = 200$

1.2 Need for Sampling

- **Economy** — less cost than a complete enumeration.
- **Speed** — results are available quickly, useful for time-critical decisions.
- **Greater accuracy** — a small, well-supervised sample may give better quality data than a huge census with untrained staff.
- **Destructive testing** — where testing destroys the unit (bulb life, crash tests), a census is impossible.
- **Infinite or very large populations** — a complete count is impractical.
- **Wider scope** — allows detailed investigation of many variables.

1.3 Census and Sample Survey

Census (complete enumeration)	Sample survey
Every unit of the population is studied	Only a selected part is studied
Costly, time-consuming, needs large staff	Cheaper, faster, smaller staff
No sampling error, but non-sampling errors can be large	Both sampling and non-sampling errors exist
Suitable for small populations or when complete information is essential (e.g., national census, every 10 years)	Suitable for large populations and periodic studies (labour force survey, living standards survey)

1.4 Basic Concept of Sampling

Sampling rests on two principles: **statistical regularity** (a reasonably large random sample tends to reflect the characteristics of the population) and **inertia of large numbers** (large groups show more stability). Sampling methods are broadly of two kinds: **probability (random) sampling**, where every unit has a known, non-zero chance of selection and the sampling error can be measured, and **non-probability sampling** (convenience, judgement, quota, snowball), where selection depends on the investigator and the error cannot be measured. Good sampling gives a sample that is **representative, adequate, independent** and unbiased.

2. Organizational Aspects of a Sample Survey

A sample survey is a large organized activity. Its main steps are:

Step	Activity
1. Objectives	State clearly the purpose and the information required
2. Population and frame	Define the target population, sampling unit and prepare or obtain the frame
3. Data to be collected	Decide the variables, definitions, method of collection (interview, mail, online)
4. Questionnaire	Design the questionnaire or schedule and pre-test it (pilot survey)
5. Sampling design	Choose the sampling method and determine the sample size
6. Field work	Recruit and train enumerators/supervisors, assign areas, collect data, control quality
7. Scrutiny and processing	Edit, code, enter, tabulate and check data for errors
8. Analysis and report	Estimate parameters and errors, interpret, write the report

2.1 Questionnaire Design

A **questionnaire** is a set of questions used to collect data from respondents. Its quality strongly affects the quality of data.

- **Structure:** introduction and purpose, respondent identification, main questions arranged logically from general to specific and easy to sensitive, instructions, and thanks.
- **Question types:** **closed-ended** (yes/no, multiple choice, rating scales such as Likert) are easy to analyse; **open-ended** give rich detail but are hard to code; **contingency** questions are asked only if a previous answer applies.
- **Rules:** use simple, short, clear words; ask one thing at a time (avoid *double-barrelled* questions); avoid leading, loaded or ambiguous questions; make response options exhaustive and mutually exclusive; avoid sensitive or embarrassing wording; keep it as short as possible.
- **Pre-testing:** try the questionnaire on a small group and revise.
- **Ethics:** obtain consent and assure confidentiality.

Example 5.1 — Fix the question

Improve: “Don't you agree that the college library is excellent and its staff are helpful?”

Solution

The question is **leading** and **double-barrelled**. Ask two neutral questions: (1) “How would you rate the college library?” (Very poor ... Excellent) and (2) “How would you rate the helpfulness of library staff?” (Very poor ... Excellent).

3. Sample Selection and Determination of Sample Size

3.1 Sample Selection

Sample selection involves defining the population, preparing the frame, choosing a sampling method and drawing the units (using a random-number table, lottery method or computer-generated random numbers). Methods are studied in Unit 6. The choice depends on the objective, population homogeneity, available frame, cost and required precision.

3.2 Determination of Sample Size

Sample size depends on the **desired precision** (margin of error e), **confidence level** (Z value: 1.645 for 90%, 1.96 for 95%, 2.576 for 99%), the **variability** of the population and the **cost/time** available. Larger samples reduce sampling error but increase cost.

Situation	Formula
Estimating a mean (σ known or estimated)	$n = Z^2 \sigma^2 / e^2$
Estimating a proportion	$n = Z^2 p q / e^2$ (use $p = 0.5$ if unknown for maximum n)
Finite population correction	$n' = n / (1 + n/N)$ (or $n / [1 + (n - 1)/N]$)

Example 5.2 — Sample size for a mean

A researcher wishes to estimate the average monthly mobile expense of students within $\pm$ Rs. 3 at 95% confidence. From a pilot study $\sigma =$ Rs. 15. Find n .

Solution

$$n = Z^2 \sigma^2 / e^2 = (1.96 \times 15 / 3)^2 = (9.8)^2 = 96.04 \rightarrow n = 97 \text{ students (rounded up).}$$

Example 5.3 — Sample size for a proportion

Estimate the proportion of students using a certain app within $\pm 5\%$ at 95% confidence, where p is unknown. The population has $N = 2,000$ students.

Solution

$$\text{Take } p = 0.5: n = (1.96)^2 \times 0.5 \times 0.5 / (0.05)^2 = 384.16 \rightarrow 385.$$

$$\text{With finite population correction: } n' = 385 / [1 + (385 - 1)/2000] = 323 \text{ students.}$$

$$(\text{If prior knowledge gives } p = 0.3 \text{ and } e = 4\%: n = 1.96^2 \times 0.3 \times 0.7 / 0.04^2 = 504.)$$

4. Sampling and Non-Sampling Errors

Estimates from a survey differ from true values because of two kinds of errors:

	Sampling error	Non-sampling error
Meaning	Difference between the sample estimate and the population parameter due to observing only a part	Errors arising from causes other than sampling, in any survey including a census
Cause	Chance variation in the selected sample; unrepresentative sample	Faulty design, non-response, measurement, processing mistakes
Behaviour	Decreases as sample size increases; zero in a census	May increase with sample size and is present in a census
Measurement	Can be measured (standard error)	Difficult to measure and control
Examples	Random sampling error; bias from method of selection	Non-response, wrong answers, interviewer bias, coding and tabulation errors, incomplete frame

Controlling errors: sampling error is reduced by increasing n and using an efficient design (e.g., stratification). Non-sampling error is reduced by a good frame and questionnaire, pre-testing, careful training and supervision, **call-backs** and follow-ups for non-response, verification of data and using proper editing and data-entry checks.

*Note: The **standard error** of the sample mean is $\sigma/\sqrt{n}$; quadrupling the sample size only halves it.*

Key Takeaways

- ✓ A population is the entire group under study, a sample is a part of it; parameters describe populations, statistics describe samples.
- ✓ Sampling is preferred to census for economy, speed, destructive testing and very large populations.
- ✓ A survey needs planning: objectives, frame, questionnaire, design, field work, processing and report.
- ✓ Good questions are short, clear, neutral and single-issue; pre-test them.
- ✓ Sample size for mean: $n = Z^2\sigma^2/e^2$; for proportion: $n = Z^2pq/e^2$; sampling error is measurable, non-sampling error is not.

— End of Unit 5 —