

Research Methodology — Sampling

Study notes with graphical illustrations · sampling techniques, terminology and comparison

1. What is Sampling?

In any research study it is usually impossible — in terms of time, cost, and access — to study every member of the group we are interested in. **Sampling** is the process of selecting a smaller, manageable subset of individuals or items from a larger group so that the findings from the subset can be used to draw conclusions about the whole group.

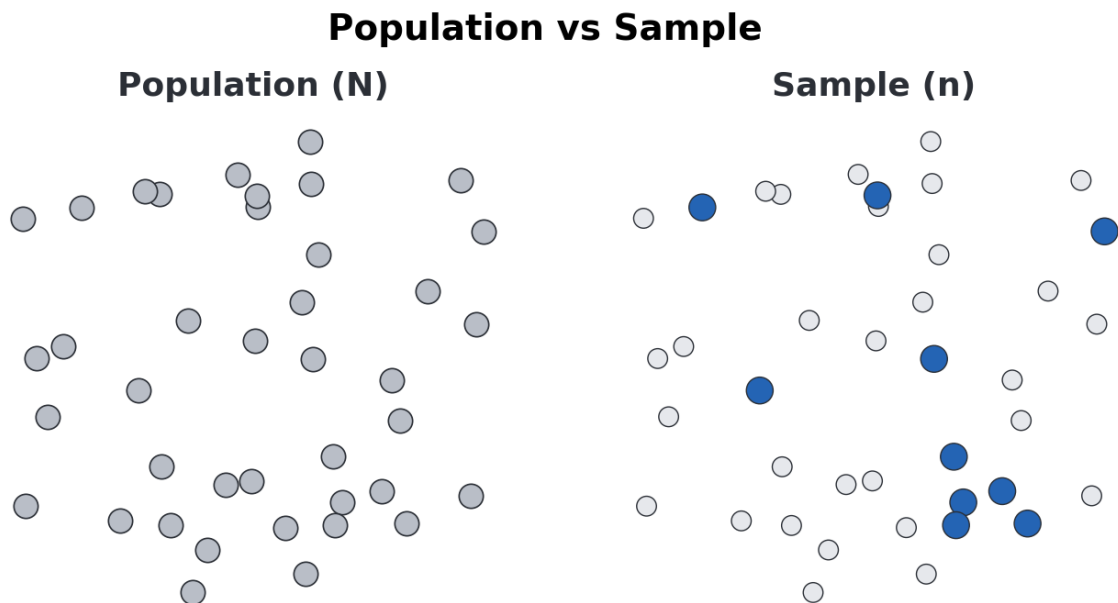


Figure 1: A sample (n) is a subset drawn from the full population (N).

Population (N): the complete set of individuals, items, or events having the characteristic(s) the researcher wants to study.

Sample (n): a subset of the population that is actually observed/measured.

Sampling frame: the actual list or source from which the sample is drawn (e.g. a voter list, hospital records, a list of registered farmers).

Sampling unit: the individual element (person, plant, plot, household) that is selected.

Parameter vs Statistic: a *parameter* describes the population (e.g. population mean μ); a *statistic* describes the sample (e.g. sample mean, denoted “ $\bar{x}$ ”) and is used to estimate the parameter.

Why sample instead of studying the whole population (census)?

- Saves time, cost, and manpower compared to a complete census.
- Often more practical — some testing is destructive (e.g. testing seed germination, crushing concrete cubes).
- Can achieve greater accuracy per unit through more careful, focused data collection.
- Feasible when the population is very large or infinite (e.g. all fish in a river).

2. Classification of Sampling Methods

Sampling techniques fall into two broad families, distinguished by whether the probability of any given unit being selected is known in advance.

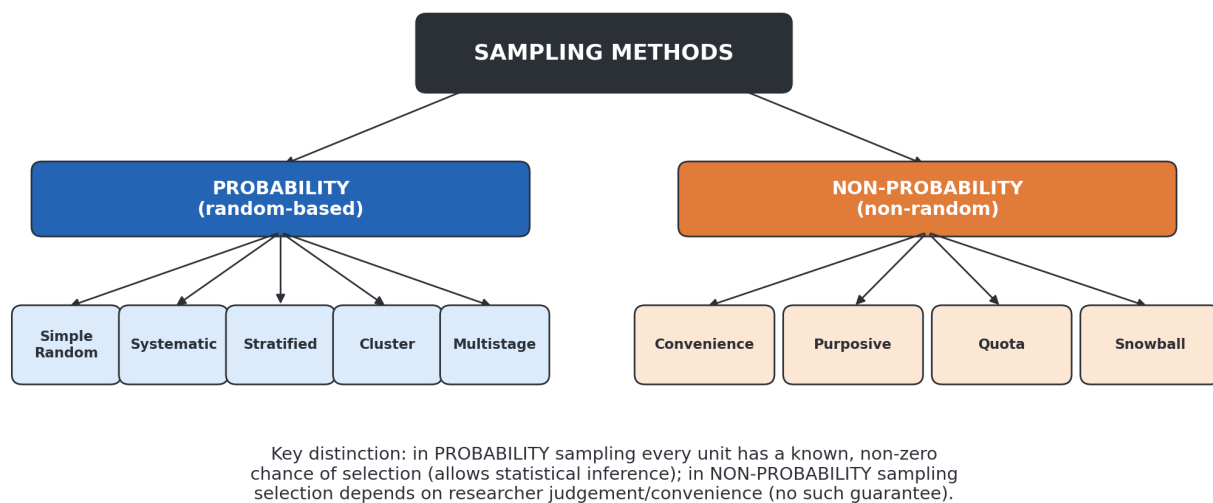


Figure 2: Classification of sampling methods.

3. Probability Sampling

Every unit in the population has a known, non-zero chance of being selected, and selection is done using a random mechanism. This allows the researcher to calculate sampling error and generalize results to the population with a known confidence level.

3.1 Simple Random Sampling (SRS)

Every unit has an equal and independent chance of being chosen. Typically implemented using a random number table, computer-generated random numbers, or the lottery method.

Simple Random Sampling



Figure 3: In SRS, selection is unbiased and every unit is equally likely to be picked.

- **Advantage:** free of researcher bias; simplest to analyse statistically.

- **Limitation:** needs a complete, accurate sampling frame; may not represent small sub-groups well by chance.

3.2 Systematic Sampling

Units are arranged in a list/sequence; a random starting point is chosen, then every k^{th} unit is selected thereafter, where k (the sampling interval) = Population size (N) $\div$ Sample size (n).

Systematic Sampling (every k^{th} unit selected after a random start)

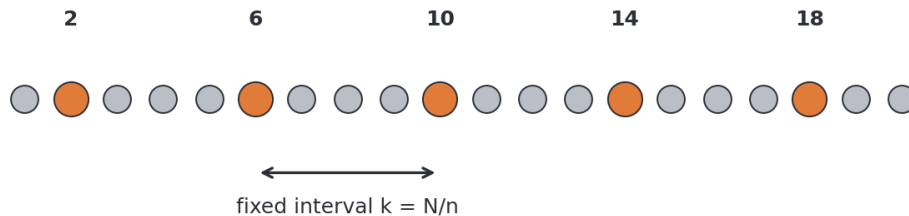


Figure 4: Systematic sampling — a fixed interval k is applied after a random start.

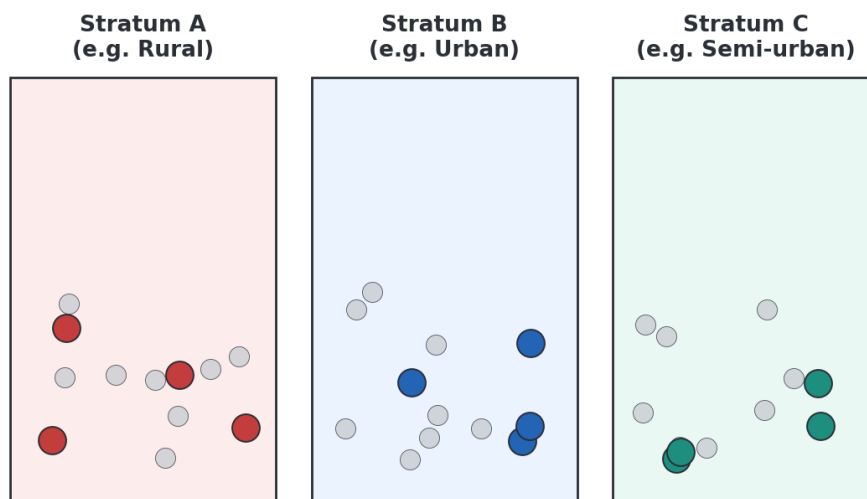
$$k = N / n \quad (\text{sampling interval})$$

- **Advantage:** quick, simple, spreads the sample evenly across the list.
- **Limitation:** risky if the list has a hidden periodic pattern matching interval k .

3.3 Stratified Sampling

The population is first divided into distinct, non-overlapping sub-groups called **strata** that are internally homogeneous (e.g. by region, age group, soil type, industry sector). A random sample is then drawn independently from each stratum — proportionately (matching each stratum's share of the population) or equally.

Stratified Sampling



Population divided into homogeneous strata; random sample drawn independently from EACH stratum

Figure 5: Stratified sampling — independent random draws within each stratum.

- **Advantage:** ensures representation of every important sub-group; increases precision when strata differ from each other.
- **Limitation:** requires prior knowledge to correctly define the strata.

3.4 Cluster Sampling

The population is divided into naturally occurring groups called **clusters** (e.g. villages, schools, city blocks) that are each a small-scale mirror of the whole population. A random sample of whole clusters is selected, and all (or a further random sample of) units within the chosen clusters are studied.

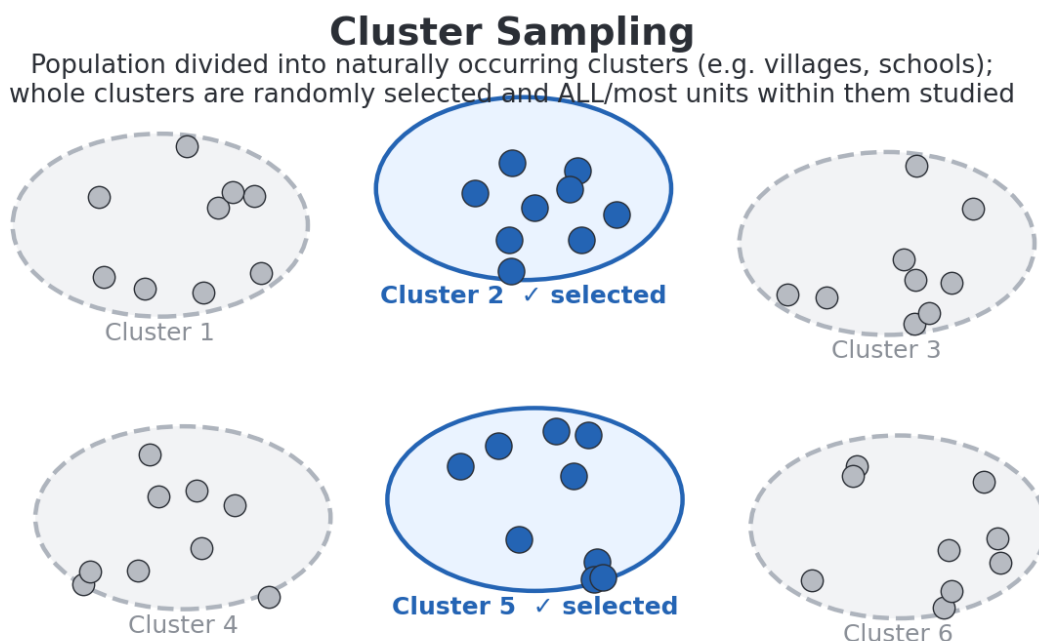


Figure 6: Cluster sampling — entire clusters are randomly chosen rather than individuals.

- **Advantage:** economical and practical when the population is spread over a large geographic area; no need for a full population list, only a list of clusters.
- **Limitation:** higher sampling error if clusters are not internally diverse (i.e. units within a cluster resemble each other).

3.5 Multistage Sampling

An extension of cluster sampling done in successive stages/steps — e.g. randomly select districts, then randomly select villages within chosen districts, then randomly select households within chosen villages. Commonly used in large national surveys (e.g. NSSO, NFHS).

4. Non-Probability Sampling

Selection does not follow randomization; probability of selection cannot be calculated. Faster and cheaper, but carries a higher risk of selection bias and results cannot be statistically generalized with confidence to the full population.

Non-Probability Sampling Techniques

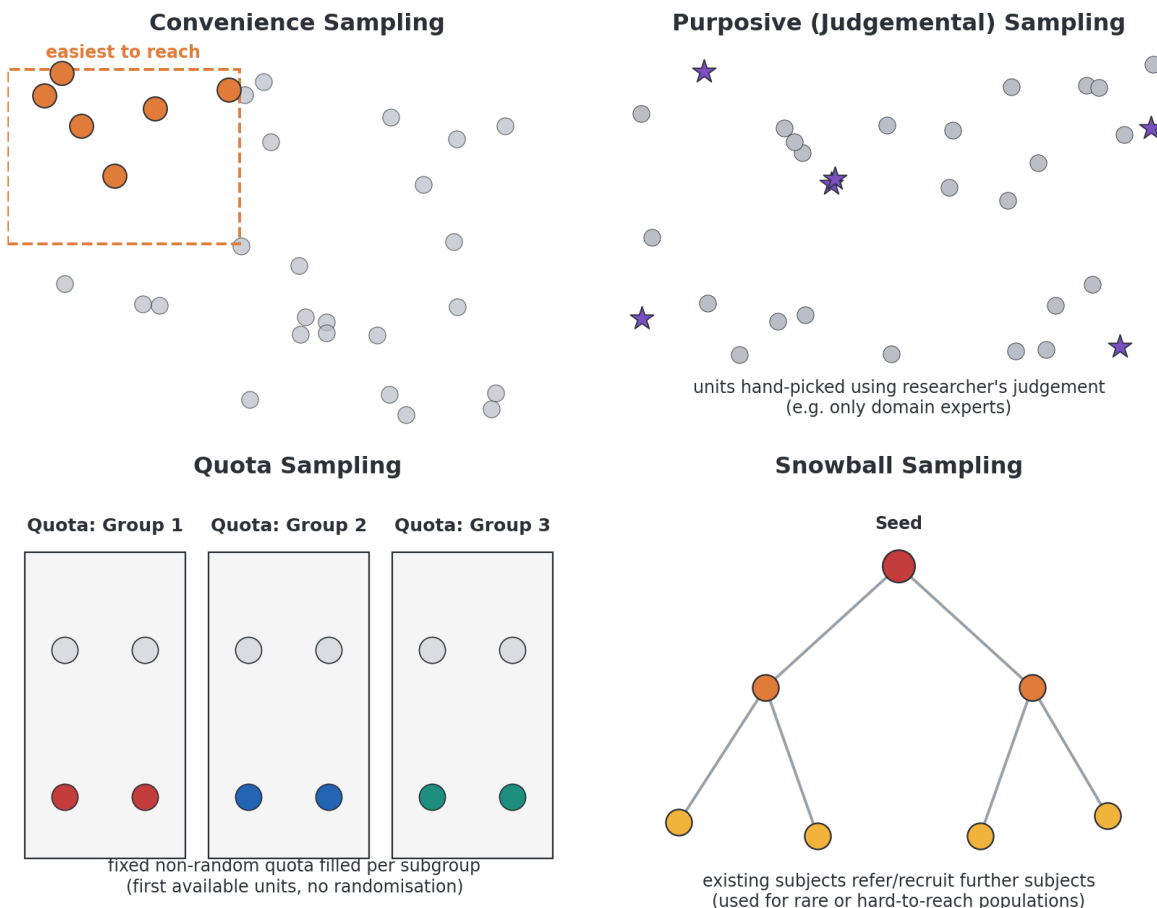


Figure 7: The four common non-probability sampling techniques.

4.1 Convenience Sampling

Units are selected simply because they are easiest to access (e.g. surveying people passing by on a street, the first patients who arrive at a clinic). Quick and inexpensive, but highly prone to bias.

4.2 Purposive (Judgemental) Sampling

The researcher deliberately selects units believed to be most informative or representative based on their own judgement/expertise (e.g. selecting only subject-matter experts for an interview study).

4.3 Quota Sampling

The population is divided into subgroups, and a fixed, non-random **quota** of units is filled for each subgroup by the researcher/enumerator (e.g. "interview 20 men and 20 women in this locality") — resembles stratified sampling but selection within each subgroup is not random.

4.4 Snowball Sampling

Used for rare, hidden, or hard-to-reach populations. A small initial group of subjects (the "seed") is identified, and each existing subject then refers/recruits further subjects from their own network, so the sample grows

like a rolling snowball.

5. Probability vs Non-Probability Sampling

Aspect	Probability Sampling	Non-Probability Sampling
Basis of selection	Random / chance-based	Researcher judgement / convenience
Selection probability	Known and calculable	Unknown
Risk of bias	Low	Comparatively high
Generalization to population	Statistically valid	Limited / not statistically valid
Cost & time	Generally higher	Generally lower
Typical use	Quantitative, confirmatory research	Exploratory, qualitative research
Examples	SRS, systematic, stratified, cluster, multistage	Convenience, purposive, quota, snowball

6. Sample Size and Sampling Error

Sampling error is the difference between a sample statistic and the true population parameter, arising purely because a sample (not the full population) was observed. It decreases as sample size increases, and as population variability decreases.

Non-sampling error arises from sources other than the sampling process itself — e.g. measurement error, non-response, data entry mistakes, poorly worded questions — and can occur even in a full census.

A commonly used formula for sample size (simple random sampling, infinite/large population):

$$n = (Z^2 \times p(1-p)) / e^2$$

where **Z** = Z-value for the desired confidence level (e.g. 1.96 for 95% confidence), **p** = estimated proportion of the attribute in the population (0.5 used if unknown, as it maximizes required sample size), and **e** = acceptable margin of error.

Factors affecting the required sample size:

- Desired level of precision (smaller margin of error → larger sample needed).
- Confidence level required (higher confidence → larger sample needed).
- Variability (heterogeneity) within the population (more variable → larger sample needed).
- Available resources — time, budget, and personnel.

Summary: choose **probability sampling** whenever statistical generalization to the population is the goal; reserve **non-probability sampling** for exploratory work, pilot studies, or when a sampling frame genuinely cannot be constructed.