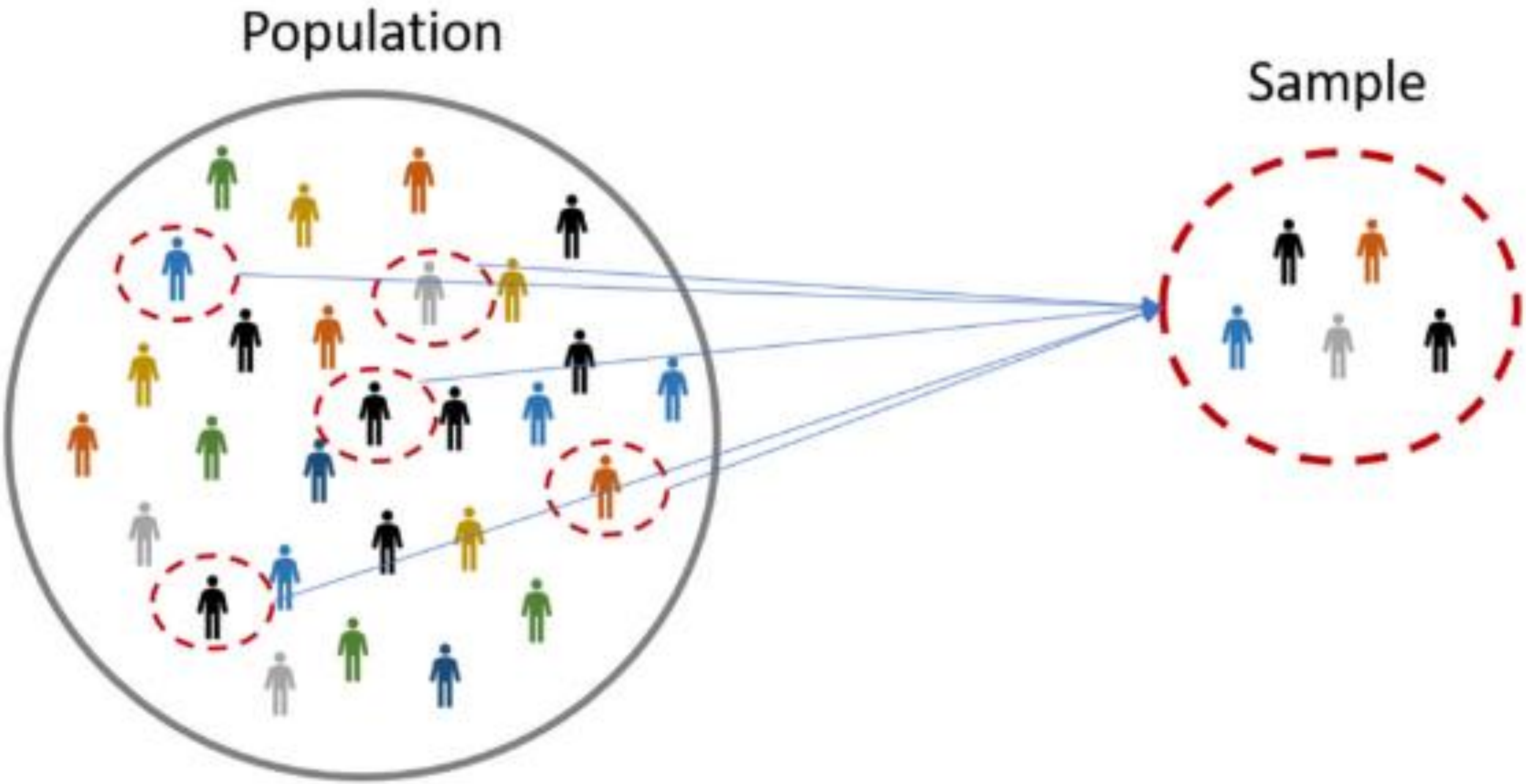
DICE
ANALYTICS

Data Science and Machine Learning

👍 https://www.facebook.com/diceanalytics/

in https://pk.linkedin.com/company/diceanalytics

DATA SAMPLING TECHNIQUES



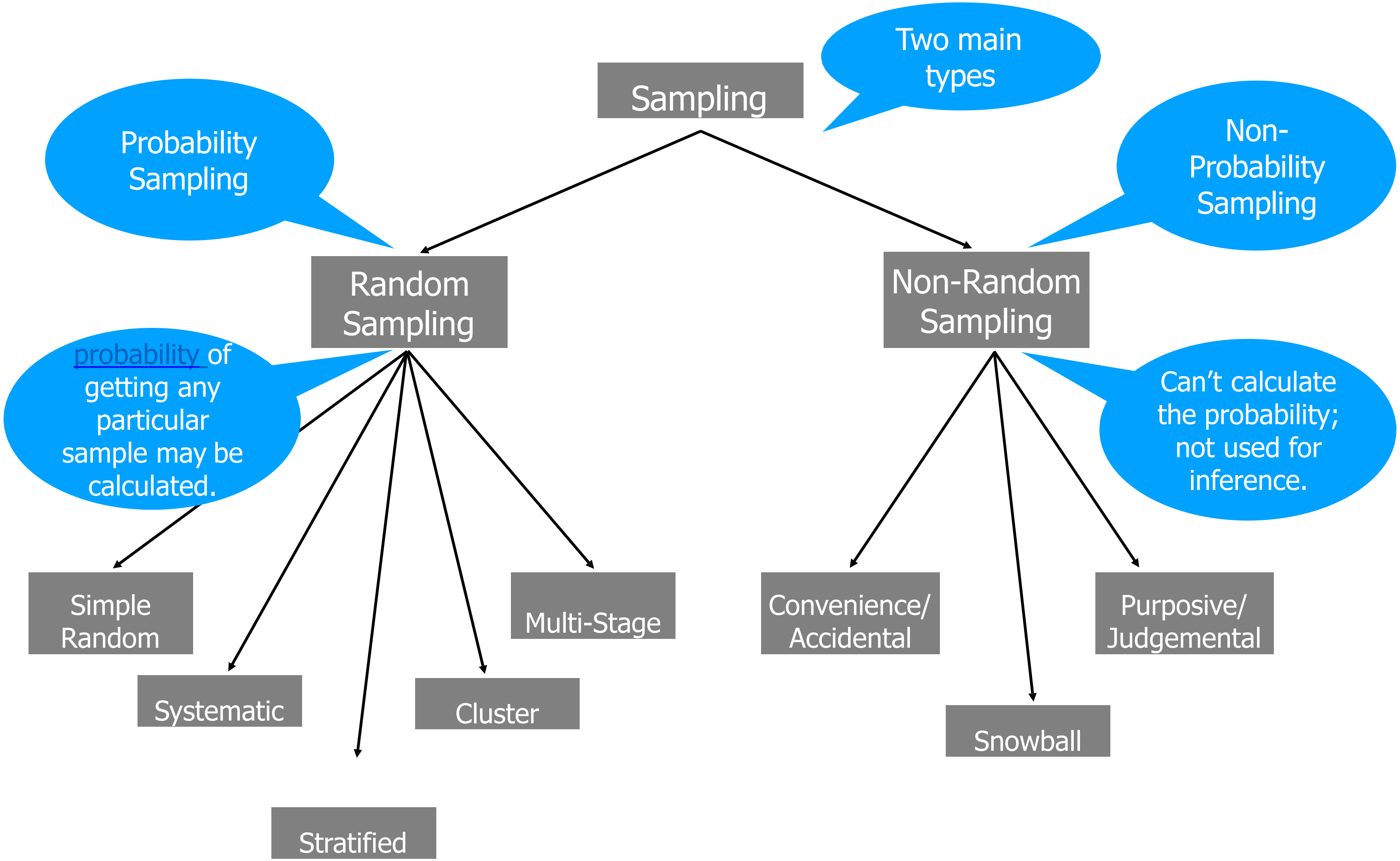
Sampling

CENSUS VS SAMPLE

- *Census*: A **census** is a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.
- Census not mostly possible: time-consuming, expensive, population hardly still, etc.

- *Sample*: A **sample** is a subset of units in a population, selected to represent all units in a population of interest.

TYPES OF SAMPLING



Sampling

Two main types

Random Sampling

Probability Sampling

probability of getting any particular sample may be calculated.

Simple Random

Systematic

Stratified

Cluster

Multi-Stage

Non-Random Sampling

Non-Probability Sampling

Can't calculate the probability; not used for inference.

Convenience/ Accidental

Snowball

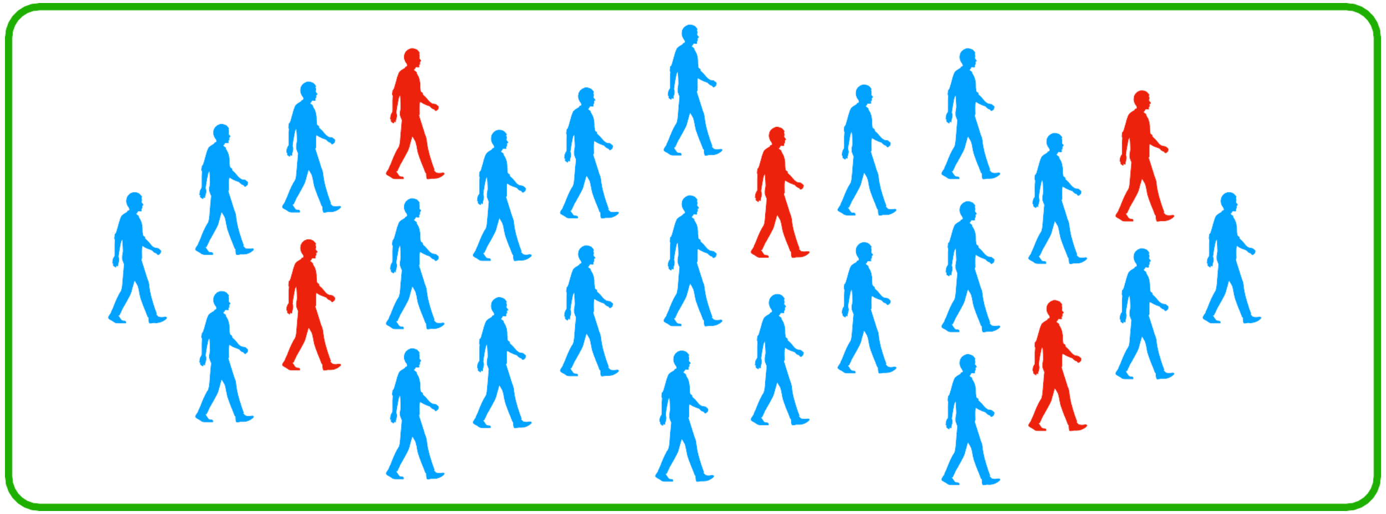
Purposive/ Judgemental

RANDOM SAMPLING

SIMPLE RANDOM SAMPLING (SRS)

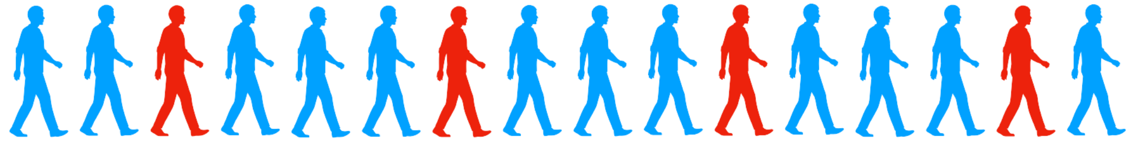
- Select ***n*** observations randomly from entire population

- Each observation is likely to be selected



SYSTEMATIC SAMPLING

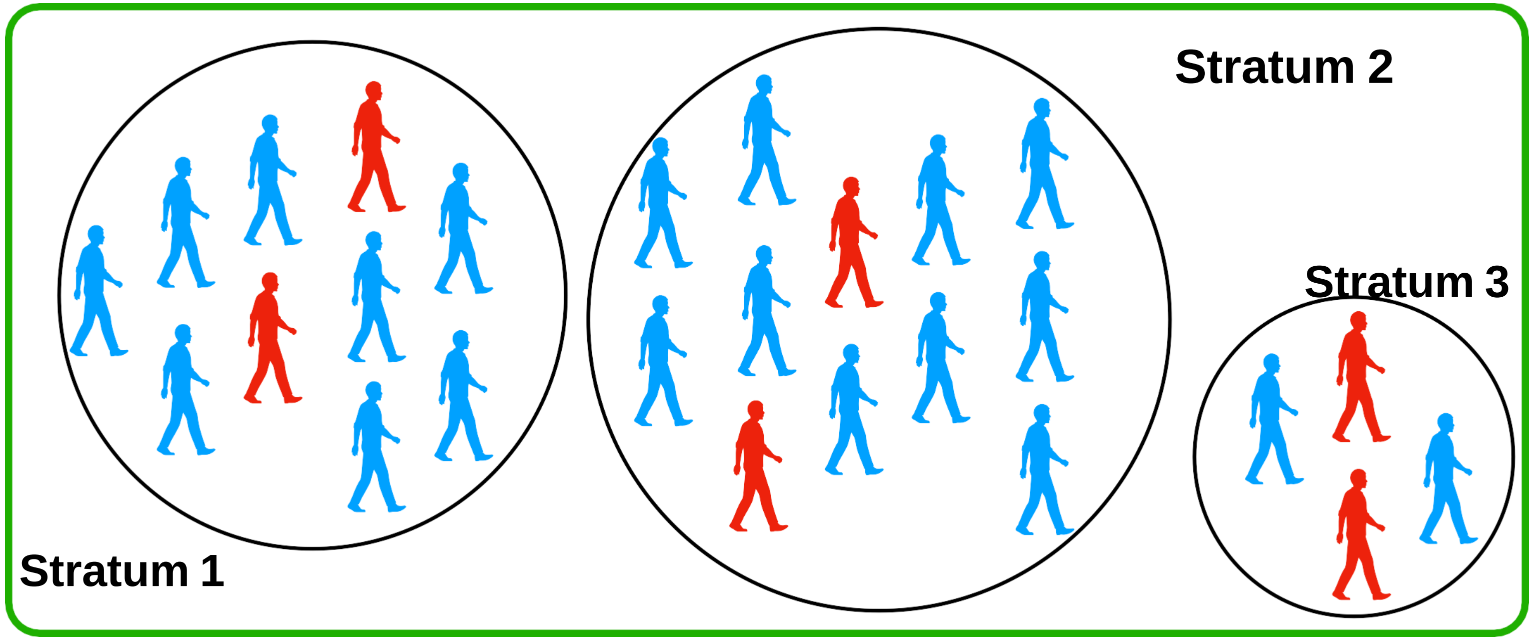
- Arrange the population according to some ordering
- Start randomly and select every **k**th observation



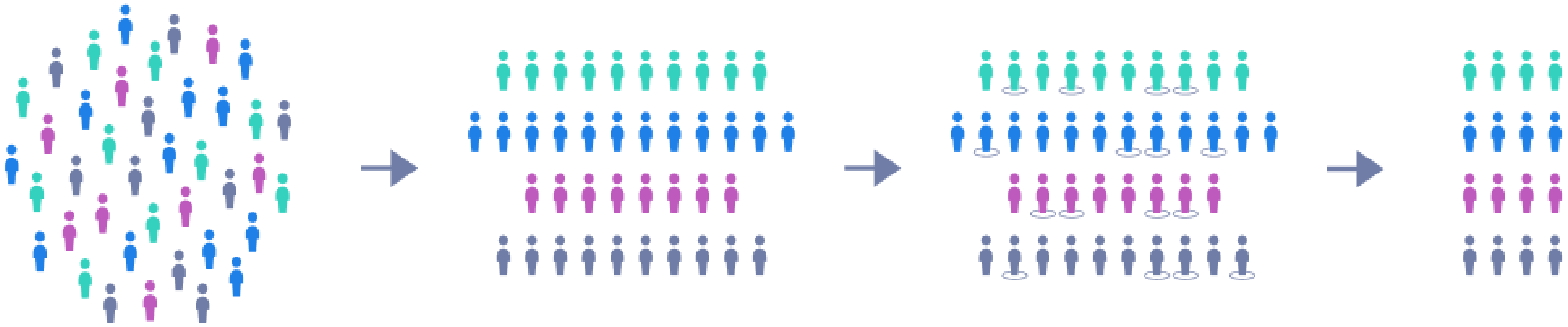
K = 4

STRATIFIED SAMPLING

- Divide population in homogenous groups called ***<u>strata</u>***
- Do Simple Random Sampling (SRS) from each stratum



Stratified sampling



Population

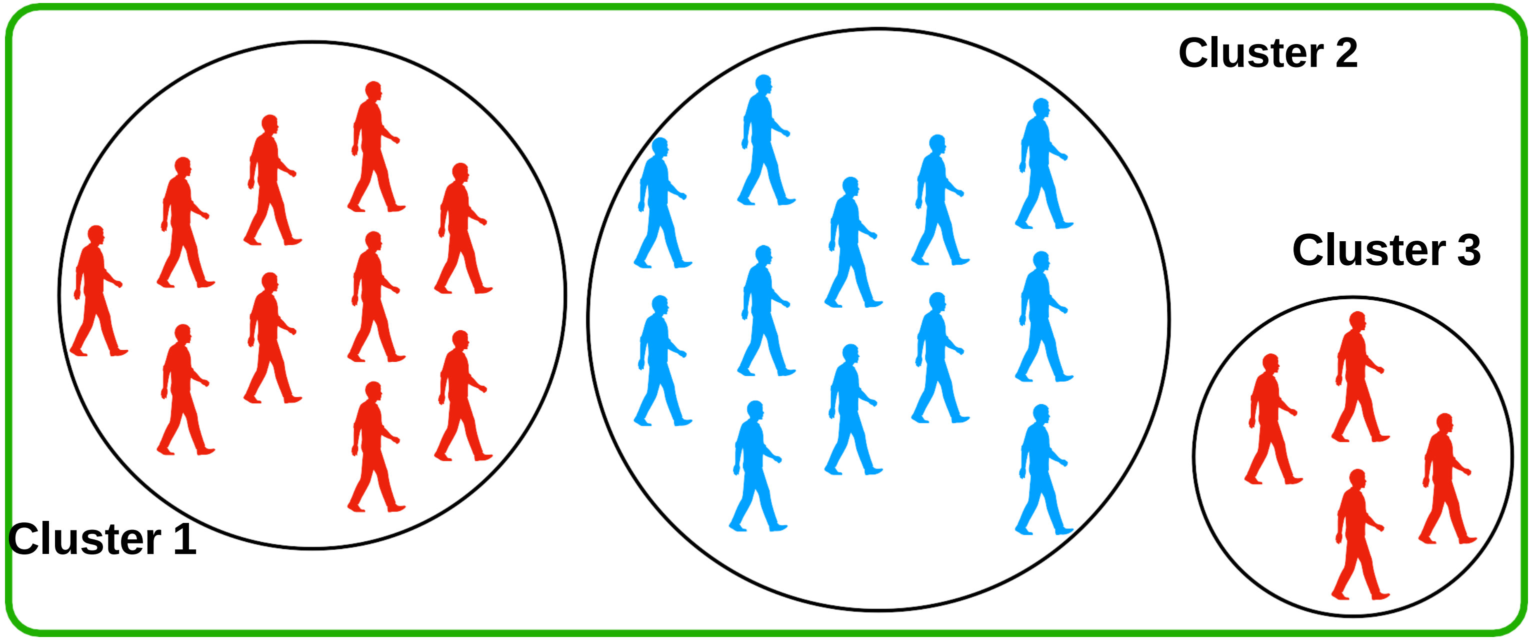
Strata

Random selection

Sample

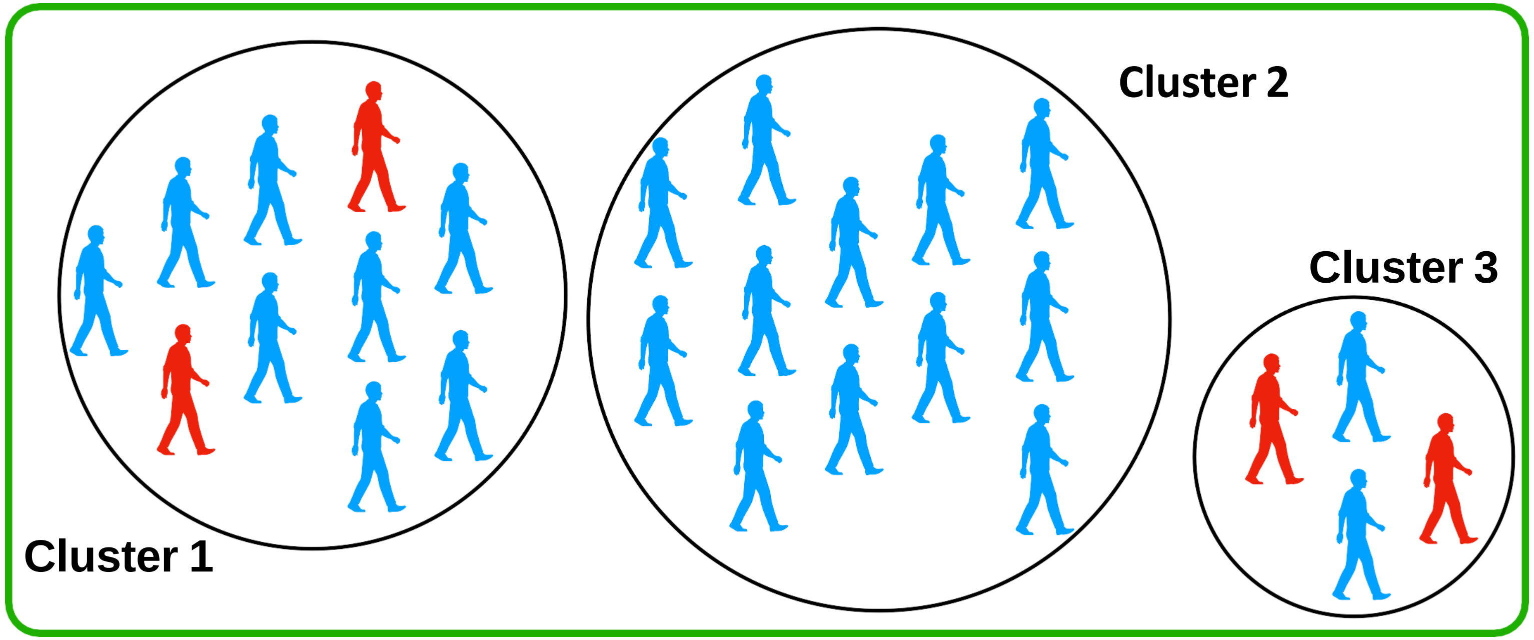
CLUSTER SAMPLING

- Divide population in heterogenous groups called ***clusters***
- Randomly Sample ***k*** clusters; and sample all observations within those clusters



MULTI-STAGE SAMPLING

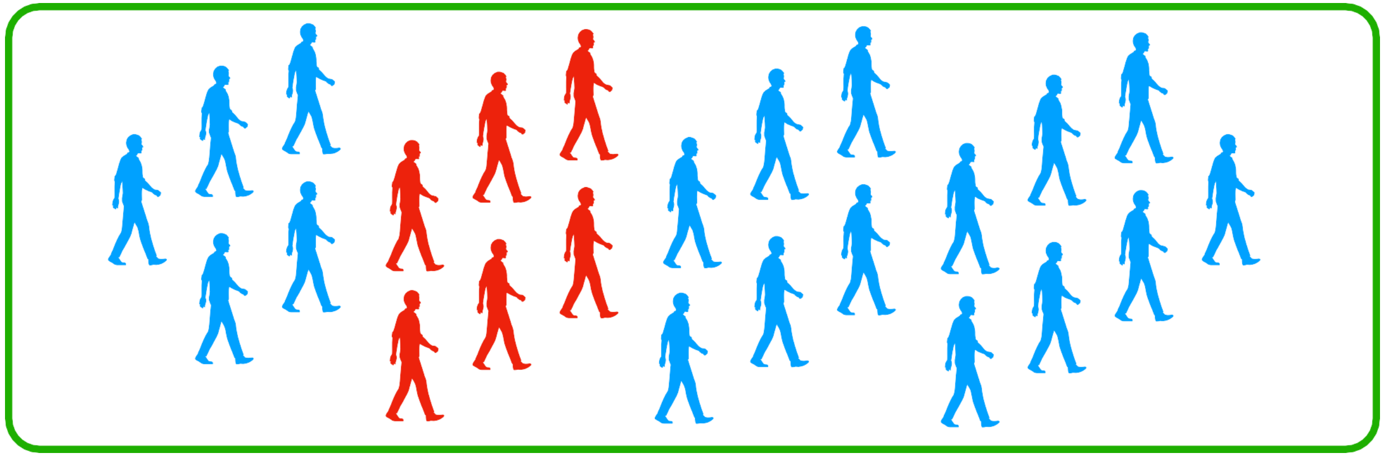
- Divide population in heterogenous groups called ***clusters***
- Randomly Sample ***k*** clusters; and do SRS within those clusters



NON-RANDOM SAMPLING

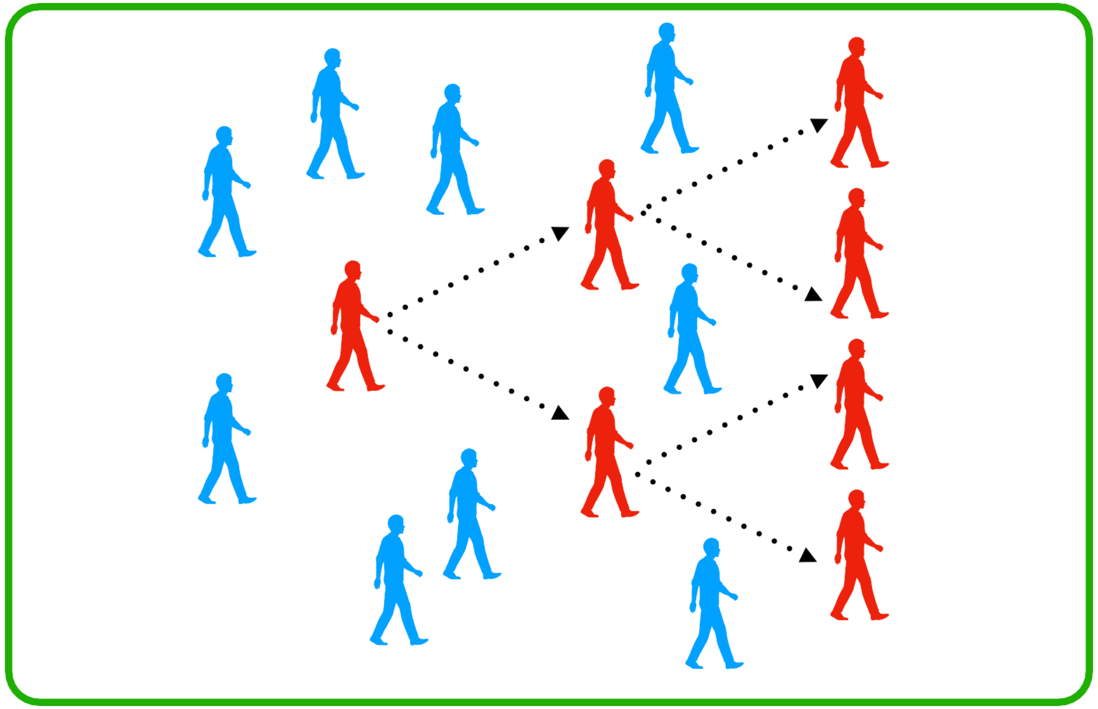
CONVENIENCE/ACCIDENTAL SAMPLING

- Members of the population are chosen based on their relative ease of access.
- To sample friends, co-workers, or shoppers at a single mall, are all examples of convenience sampling.
- Such samples are biased because researchers may unconsciously approach some kinds of respondents and avoid others (Lucas 2014a), and respondents who volunteer for a study may differ in unknown but important ways from others (Wiederman 1999).



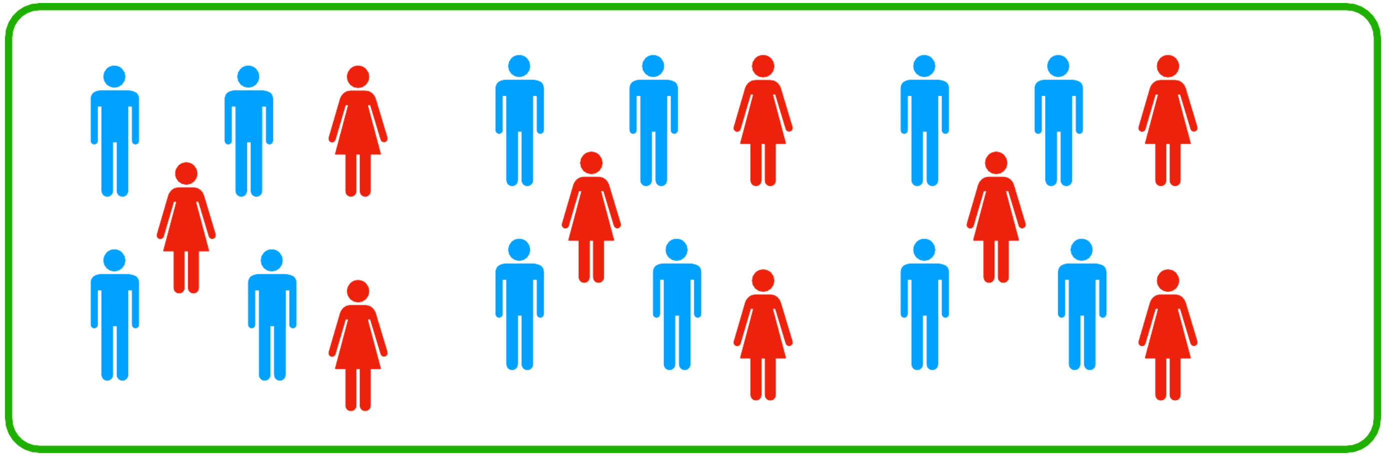
SNOWBALL SAMPLING

- The first respondent refers an acquaintance. The friend also refers a friend, and so on.
- Such samples are biased because they give people with more social connections an unknown but higher chance of selection (Berg 2006), but lead to higher response rates.



PURPOSIVE/JUDGMENTAL SAMPLING

- The researcher chooses the sample based on who they think would be appropriate for the study.
- This is used primarily when there is a limited number of people that have expertise in the area being researched, or when the interest of the research is on a specific field or a small group.



SAMPLING BIAS VS SELECTION BIAS

- *Sampling Bias*: A **bias** in which a **sample** is collected in such a way that some members of the intended population are less likely to be included than others; occurs when you choose your sample which is the 1st step of a research.

- *Selection Bias*: A **bias** introduced by the **selection** of individuals, groups or data for analysis in such a way that proper randomisation is not achieved; occurs when you select which subject goes to the control group and which to the treatment group.

SOURCES OF SAMPLING BIAS

- *Convenience Sample*: Easily accessible people more likely to be included in the sample.
- *Non-Response*: If only particular type(s) of randomly sampled people respond to survey.
- *Voluntary Response*: Happens when sample consists of people who volunteered to respond because they are opinionated.

CORRELATION VS CAUSATION

- *Correlation*: It describes the mutual relationship or connection between an independent and dependent variable.

- *Causation*: Causation, also known as cause and effect, is when an observed event or action (independent variable) appears to have caused a second event or action (dependent variable).

Correlation does not imply Causation!