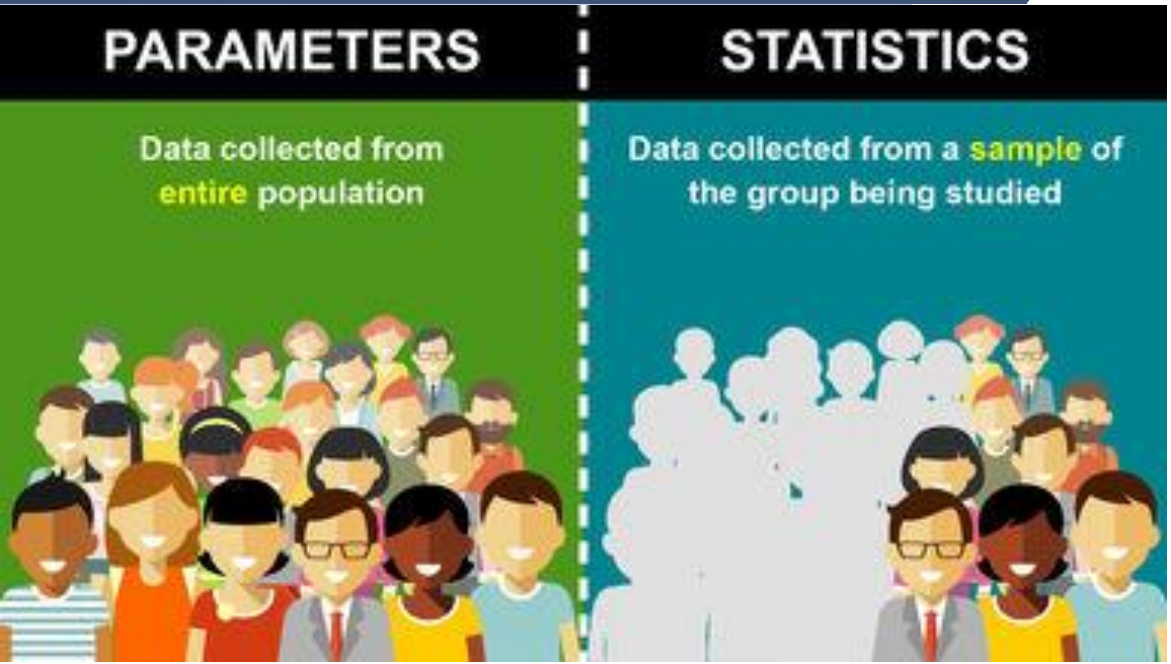
DATA SCIENCE WITH MACHINE LEARNING

A JOURNEY FROM DATA TO DECISIONS

ZAHID HUSSAIN
SR. LECTURER DHA SUFFA UNIVERSITY

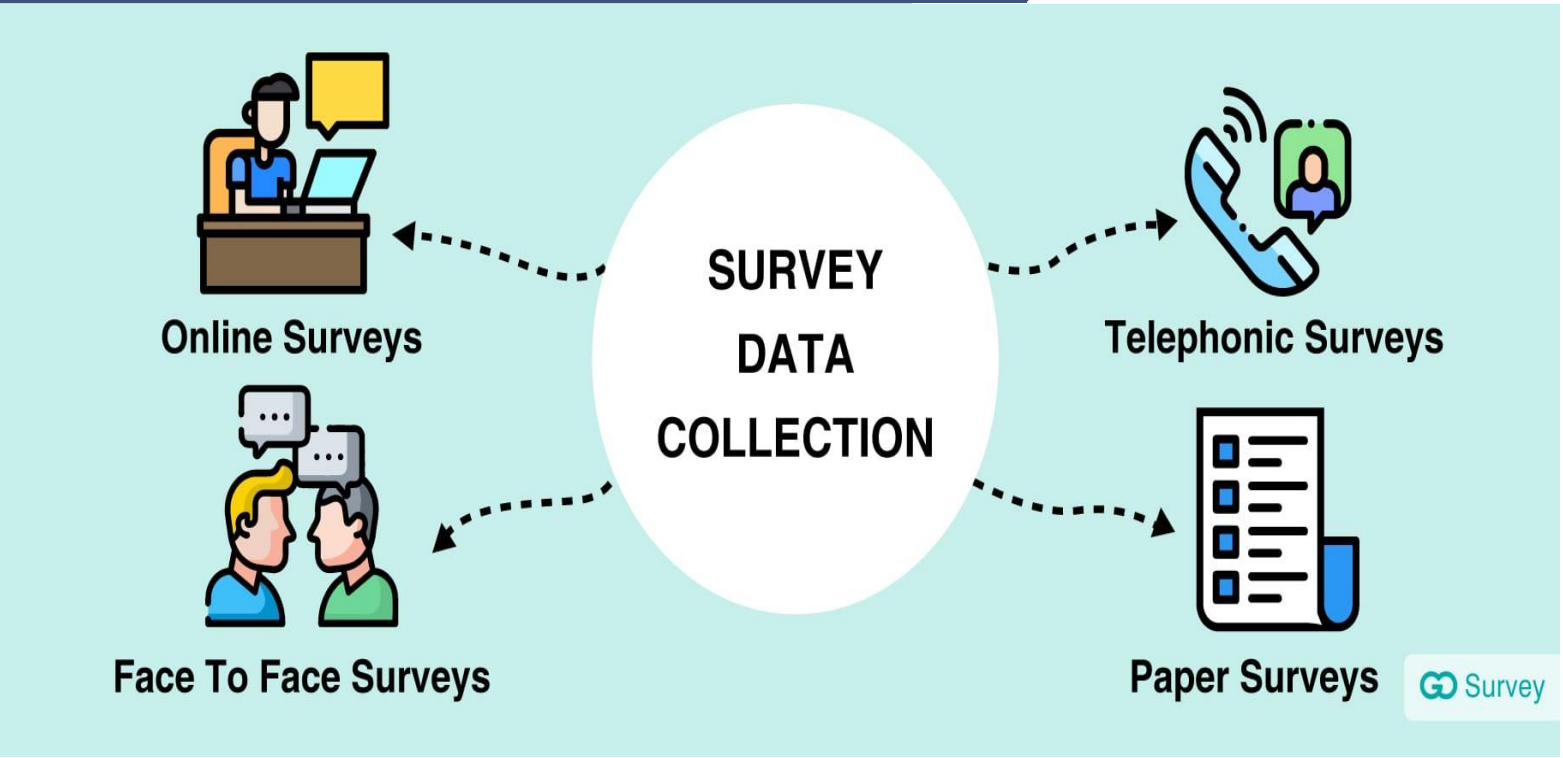
DEPARTMENT OF COMPUTER SCIENCE

Parameters vs. Statistics



- Parameters are often written using Greek letters like μ (pronounced “mew”) or σ, pronounced “sigma.”
- Statistics are written using Roman letters like $\bar{x}$ and s.

Collecting Data by Surveys



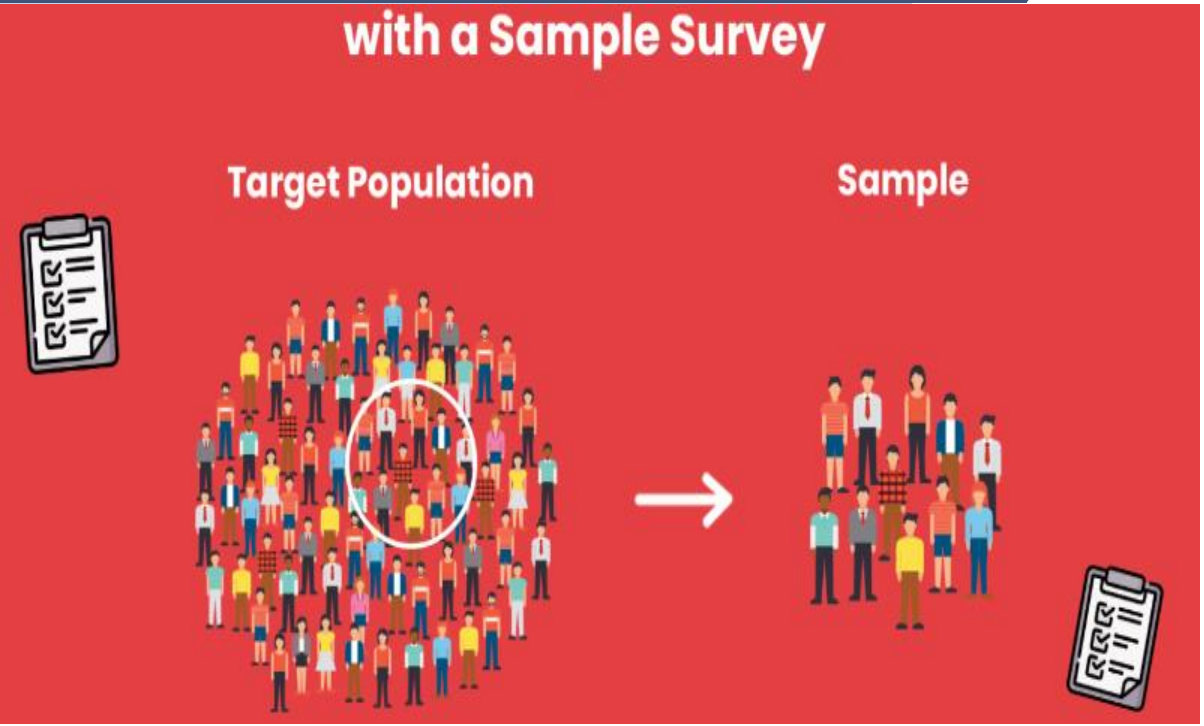
There are four main ways to obtain data – Census

- The main advantage of using a survey to obtain information is that your **conclusions will have 100% certainty.**
- The disadvantages of conducting a census are that it may be **difficult or impossible to obtain all the information, and costs may be prohibitive.**

There are four main ways to obtain data – Existing Data

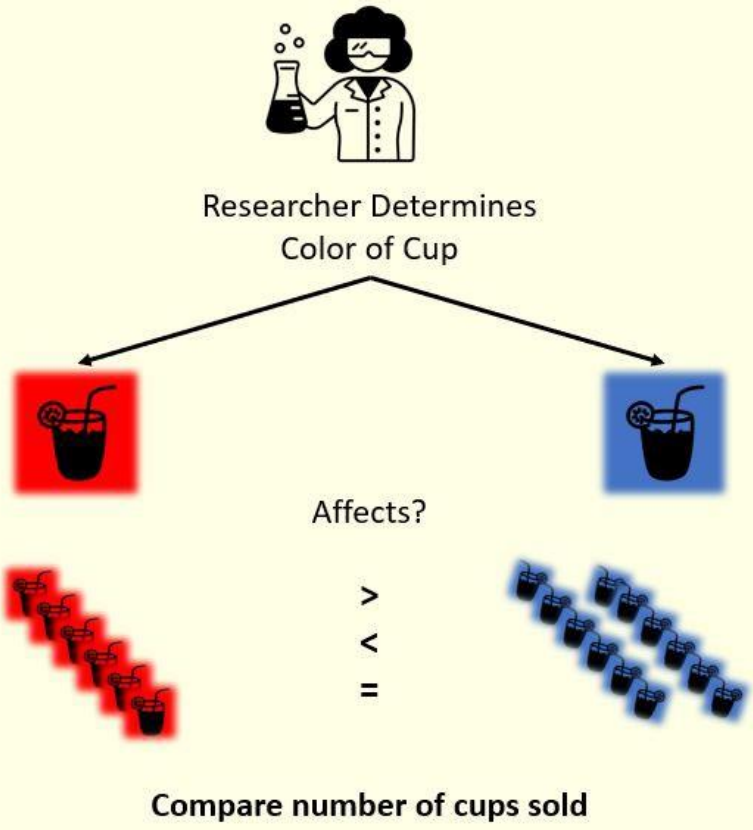
- The advantage of finding an existing source of data is obviously the **savings in both time and money.**
- A disadvantage is that it can often be **difficult to find the exact data you need.**

There are four main ways to obtain data – A survey sample



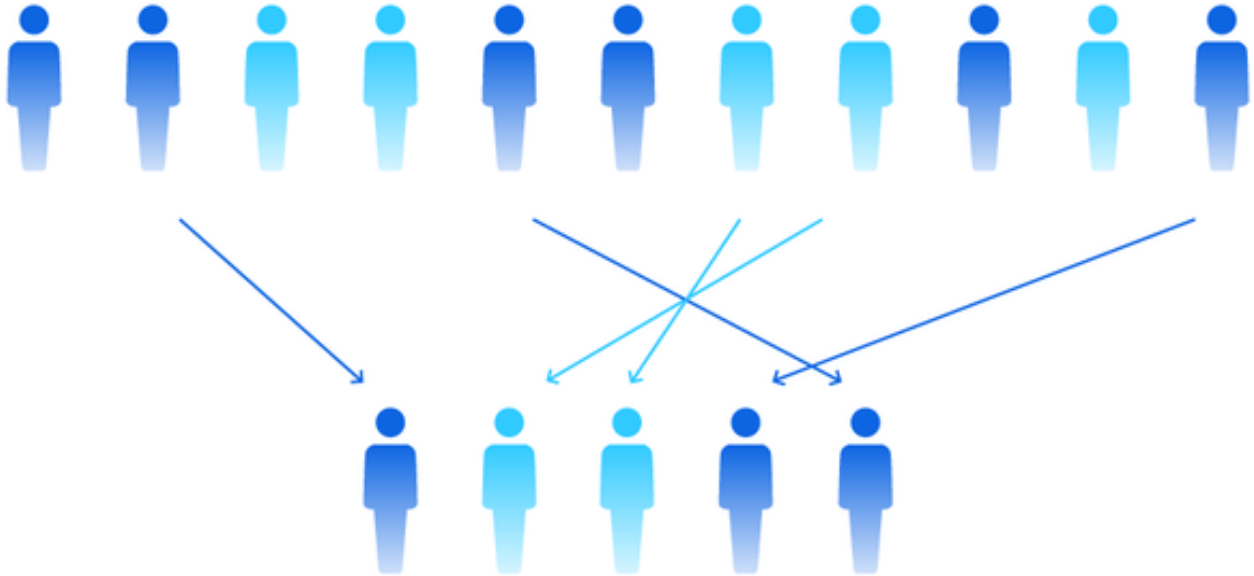
- The advantage of using a survey is **the savings in both time and money** of not having to get information from every individual in the population.
- The disadvantage of a survey sample, and this is extremely important, is that **choosing an appropriate sample could be difficult.**

There are four main ways to obtain data – A designed experiment



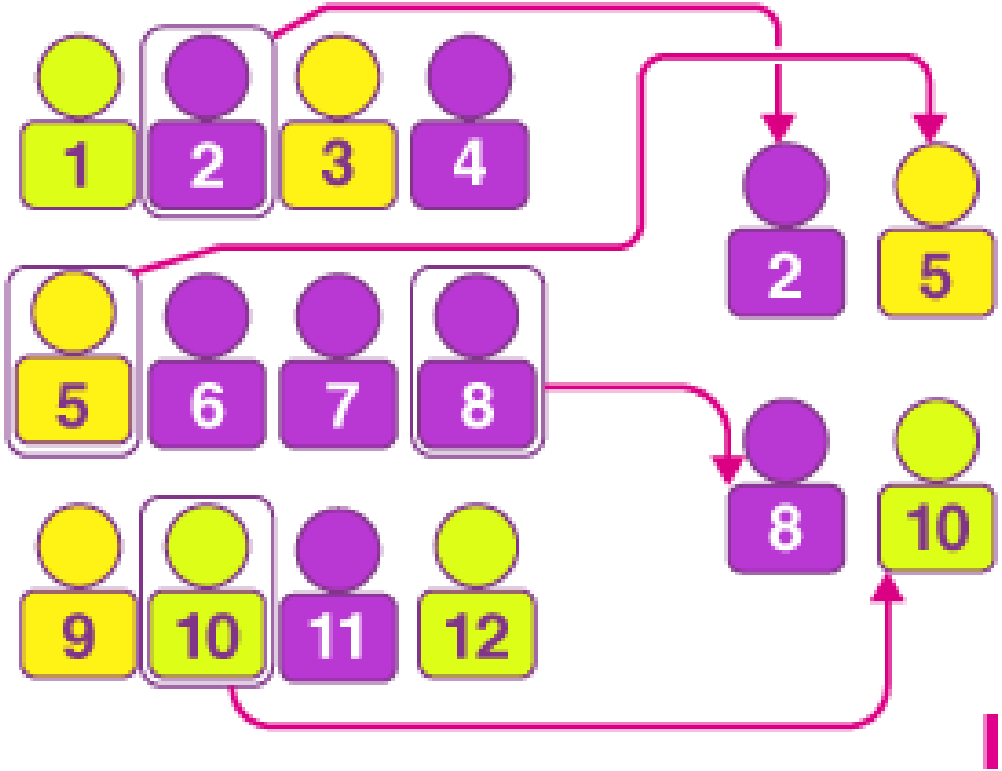
- In an experiment, **information from the treated group is often compared with a control (untreated) group**.
- **Variables from the individuals and the treatments** can easily be controlled in an experiment.
- A major advantage of an experiment is that you can **analyze individual factors**.
- Disadvantages of experiments are that they cannot be conducted when the variables cannot be controlled and in cases for moral/ethical reasons.

Types of Sampling -
A representative sample



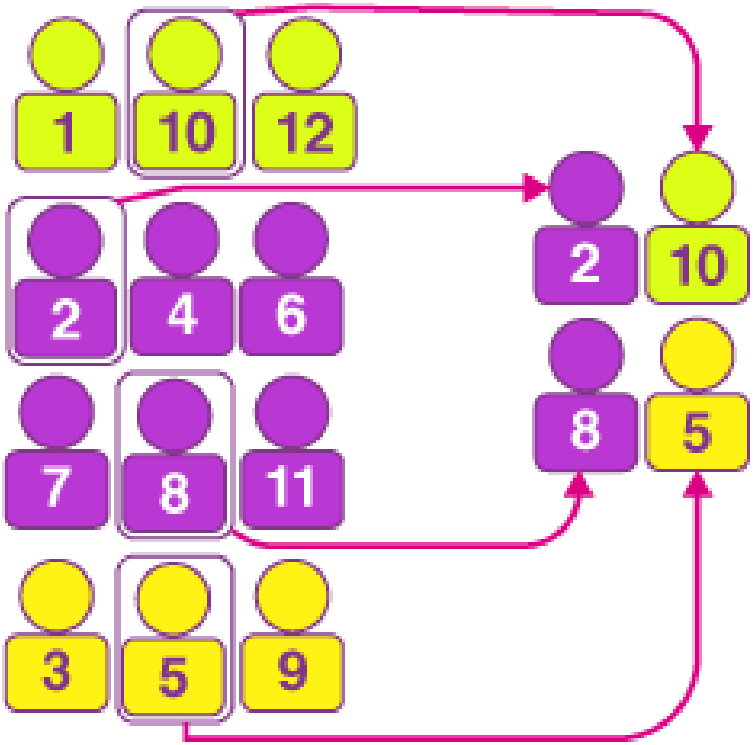
- **A representative sample** is one that has the same relevant characteristics as the population and does not favor one group of the population over another.

Types of Sampling - A random sample



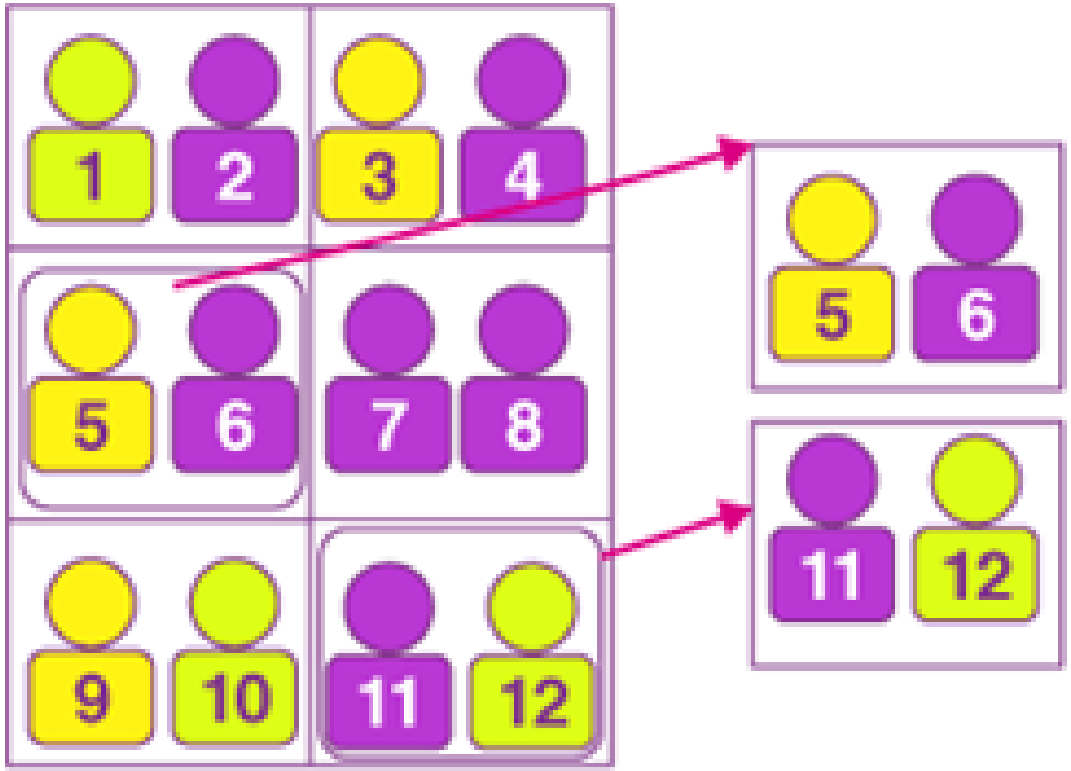
- **A random sample** is one in which every member of the population has an equal chance of being selected.

Types of Sampling - A stratified sample



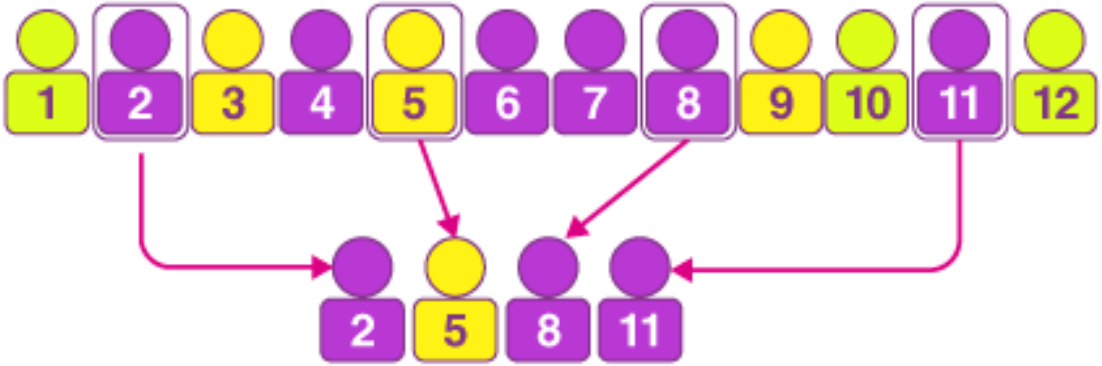
- **A stratified sample** is one in which members of the **population are divided into two or more subgroups, called strata**, that share similar characteristics like age, gender, or ethnicity.
- A random sample from each stratum is then drawn.

Types of Sampling - A cluster sample



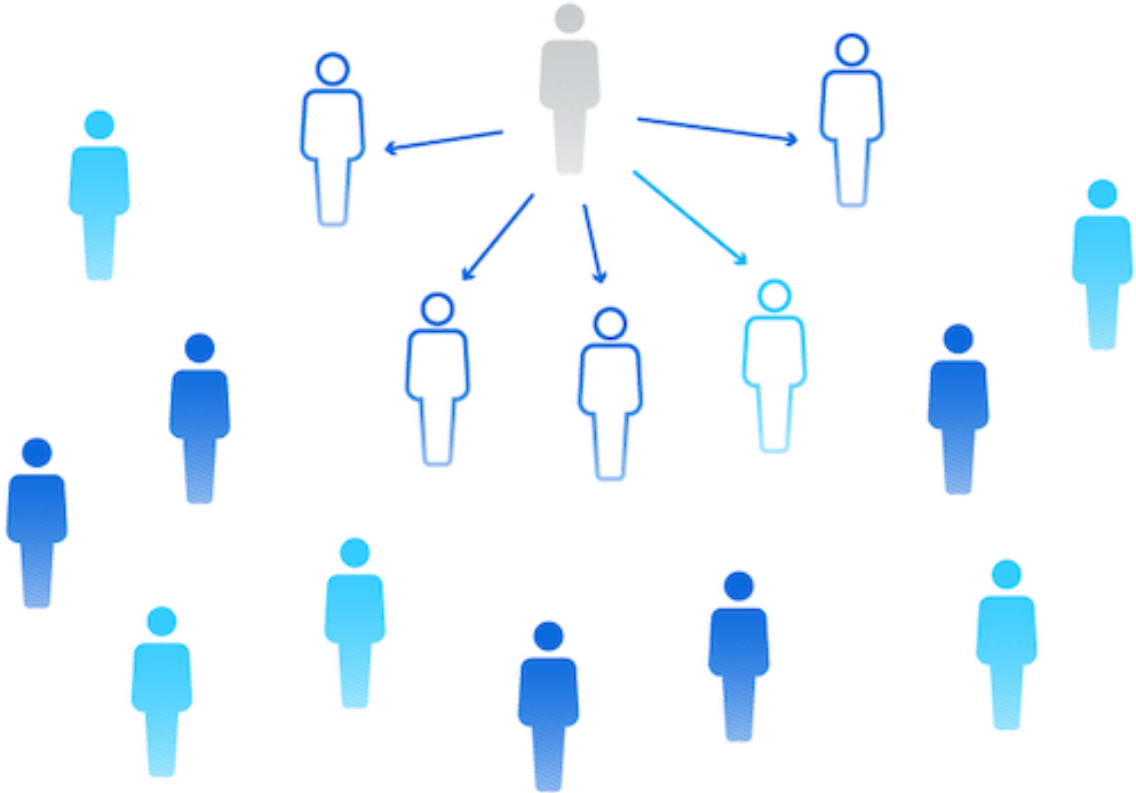
- **A cluster sample** is one chosen by dividing the population into groups, called clusters, that are each similar to the entire population.
- The researcher then randomly selects some of the clusters.
- The sample consists of the data collected from every member of each cluster selected.

Types of Sampling - A systematic sample



- **A systematic sample** is one chosen by selecting every nth member of the population.
- **Systematic sampling** is easy to detect because it always produces the same sample for the same n.
- To get a different sample you will need a different n value.

Types of Sampling - A convenience sample



- A convenience sample is one in which the sample is “convenient” to select.
- It is so named because it is convenient for the researcher.