CSC380: Principles of Data Science

Data Analysis, Collection, and Visualization 3

Xinchen Yu

Outline

- Data Collection & Basics of Causal Inference
- Data Collection Strategies

Much of the content in this section from Scribbr.com and Shona McCombes

Statistical Analysis

1. Plan research design
2. Collect data (essentially, sampling)
3. Visualize and summarize the data (plots and summary stats)
4. Make inferences from data (i.e., estimate stuff, test hypotheses, …)
5. Interpret results

Have touched on these already…

Statistical Analysis

1. Plan research design
2. Collect data (essentially, sampling)

Will focus on these

3. Visualize and summarize the data (plots and summary stats)
4. Make inferences from data (i.e., estimate stuff, test hypotheses, …)
5. Interpret results

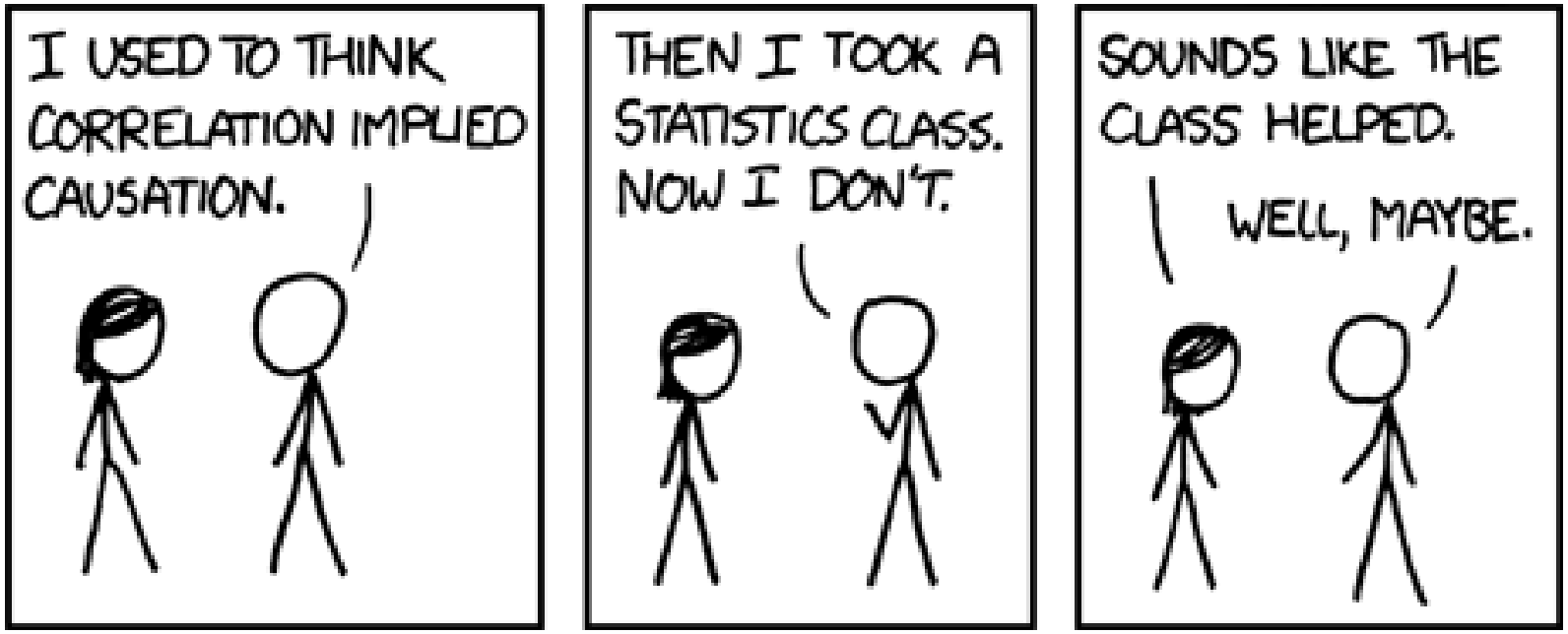
Research Design

Randomized Control. Researcher controls treatment among groups. Used to assess *causal* relationships. Stronger than correlational study but difficult to conduct. (e.g., clinical trials)

Observational. Collect data by "observing" passively. If there are treatments (i.e., vaccines), they are not under control of the researcher.

Natural Experiment. Observe naturally-occurring phenomena. Approximates a controlled study, despite the researcher not having control of any groups. (e.g., different US state policies of COVID protocols and its effect on COVID spread)

Correlation is not Causation



Why does he say “well, maybe” instead of “yes, the class helped”?

https://www.explainxkcd.com/wiki/index.php/552:_Correlation

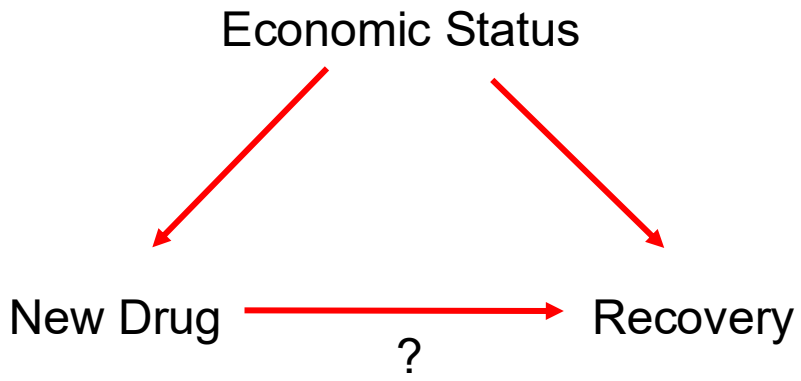
Causal inference: an example

- “Remdisivir becomes the first COVID-19 treatment to receive FDA approval” – CNN, 2020
- FDA’s data shows:

	Mortality rate
Remdisivir	11%
No Remdisivir	20%

Caveat: Remdivisir costs more than $2000 – wealthier patients are more likely to receive it

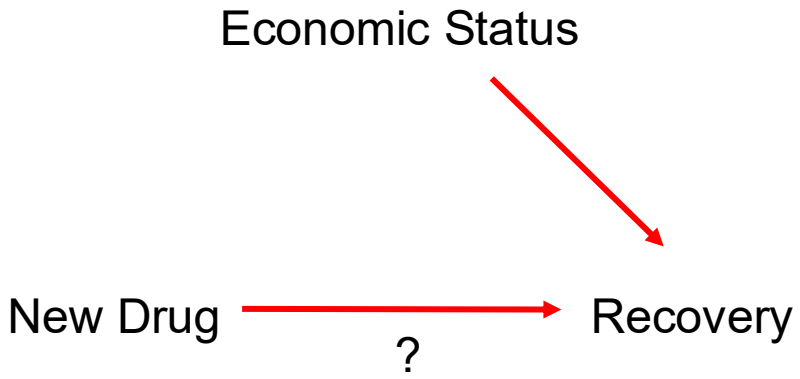
Causal inference: an example



- Weathier people more likely to have better environment for recovery
- Weathier people more likely to take Remdivisir

- Thus, positive correlation between Remdivisir and recovery does not imply Remdivisir's effectiveness
- 'Economic status' is called a *confounder variable*

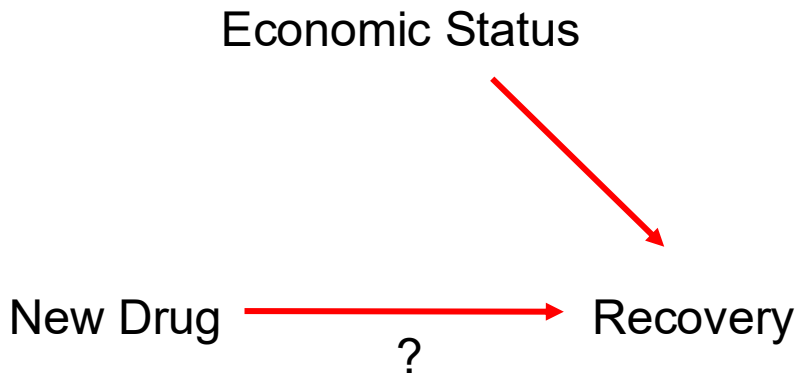
Confounding: an example



- More popular in practice: randomized controlled trials (RCT)
  - For each individual, flip a fair coin
  - If heads, give them treatment
  - If tails, give them placebo

- Estimate the recovery rates in treatment group vs. control group

Confounding: an example



Study of WHO using RCT:

	Mortality rate
Remdisivir	15%
No Remdisivir	15%

"WHO recommends against use of Remdisivir for COVID patients" – CNN, 2020

Randomized Controlled Trials

- Pros:
  - Gold standard in estimating causal effects
  - Eliminate selection bias and confounding in data, making results trustworthy
- Cons:
  - Ethical considerations (who goes to the treatment / control group?)
  - In fast-moving public health emergencies (like AIDS or COVID-19), waiting for definitive RCT results may cause lives

Population vs. Sample

Generally infeasible to collect data from entire *population*

Population Entire group that we want to draw conclusions about.

Can be defined in terms of location, age, income, etc.

Sample Specific group that we collect data from.

Population Parameter vs. Sample Statistic

Population parameter A measure that describes *the whole population.*

Sample statistic A measure that describes the sample and reflects the population parameter.

> ***Example*** *We are studying student* ***political attitudes*** *and ask students to rate themselves on a scale: 1, very liberal, to 7, very conservative. The* ***population parameter*** *of interest is the average political leaning. The sample mean, say 3.2, is our* ***statistic.***

Reasons for Sampling

Necessity It is usually impractical or impossible to collect data from an entire population due to size or inaccessibility.

Sufficiency The number of samples needed to draw reliable conclusion (e.g. estimate avg political leaning up to *0.1*) only depends on *precision*

Cost-effectiveness There are fewer participant, laboratory, equipment, and researcher costs involved.

Manageability Storing data and running statistical analyses is easier on smaller datasets.

Sampling Error

The *sampling error* is the difference between the population parameter and the sample statistic.

- Sampling errors are **normal**, but we want them to be low
- Samples are **random**, so sample statistics are estimates and thus subject to random noise
- **Sample bias** occurs when the sample is not representative of the population (for various reasons)

Sampling Bias

Occurs if data are collected in a way that some members of the population have lower/higher probability of being sampled than others

Sometimes is unavoidable (e.g., not all members are equally accessible) but
(1) we should be aware of it

(2) must be corrected if possible at all

Example We conduct a poll by randomly calling numbers in a phone book. People that have less time are less likely to response. Called **non-response bias**.

Examples of Sample biases

Population	Sample
Advertisements for IT jobs in the Netherlands	The top 50 search results for advertisements for IT jobs in the Netherlands on May 1, 2020
Songs from the Eurovision Song Contest	Winning songs from the Eurovision Song Contest that were performed in English
Undergraduate students in the Netherlands	300 undergraduate students from three Dutch universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Keep in mind: You could easily collect biased data

Sampling Methods

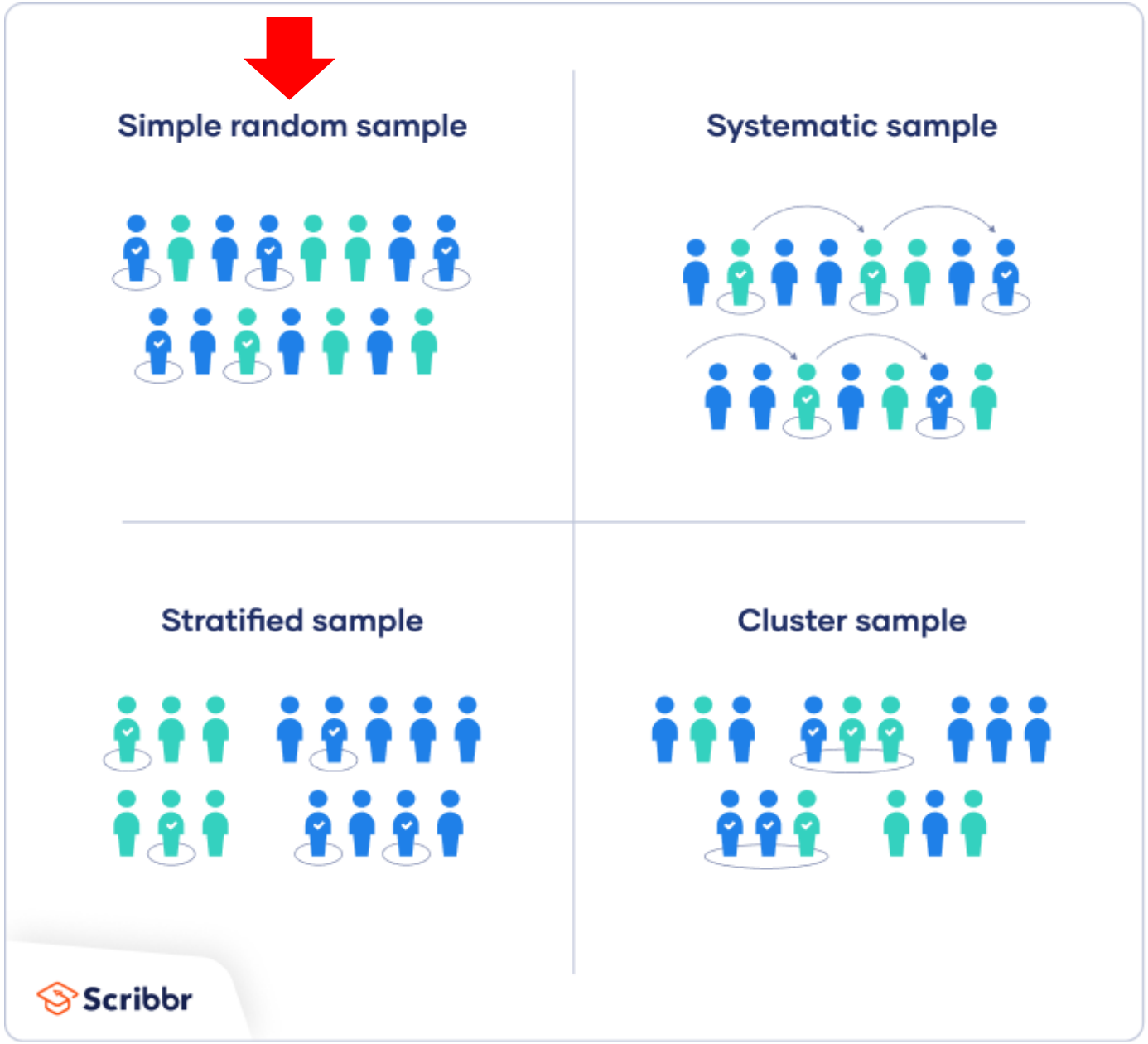
Sampling must be conducted properly, to avoid sample bias

Two primary types of sampling…

Probability Sampling Random selection; allowing strong statistical inferences about the population

Non-Probability Sampling Based on convenience or other criteria to easily collect data (but no random sampling)

Probability Sampling



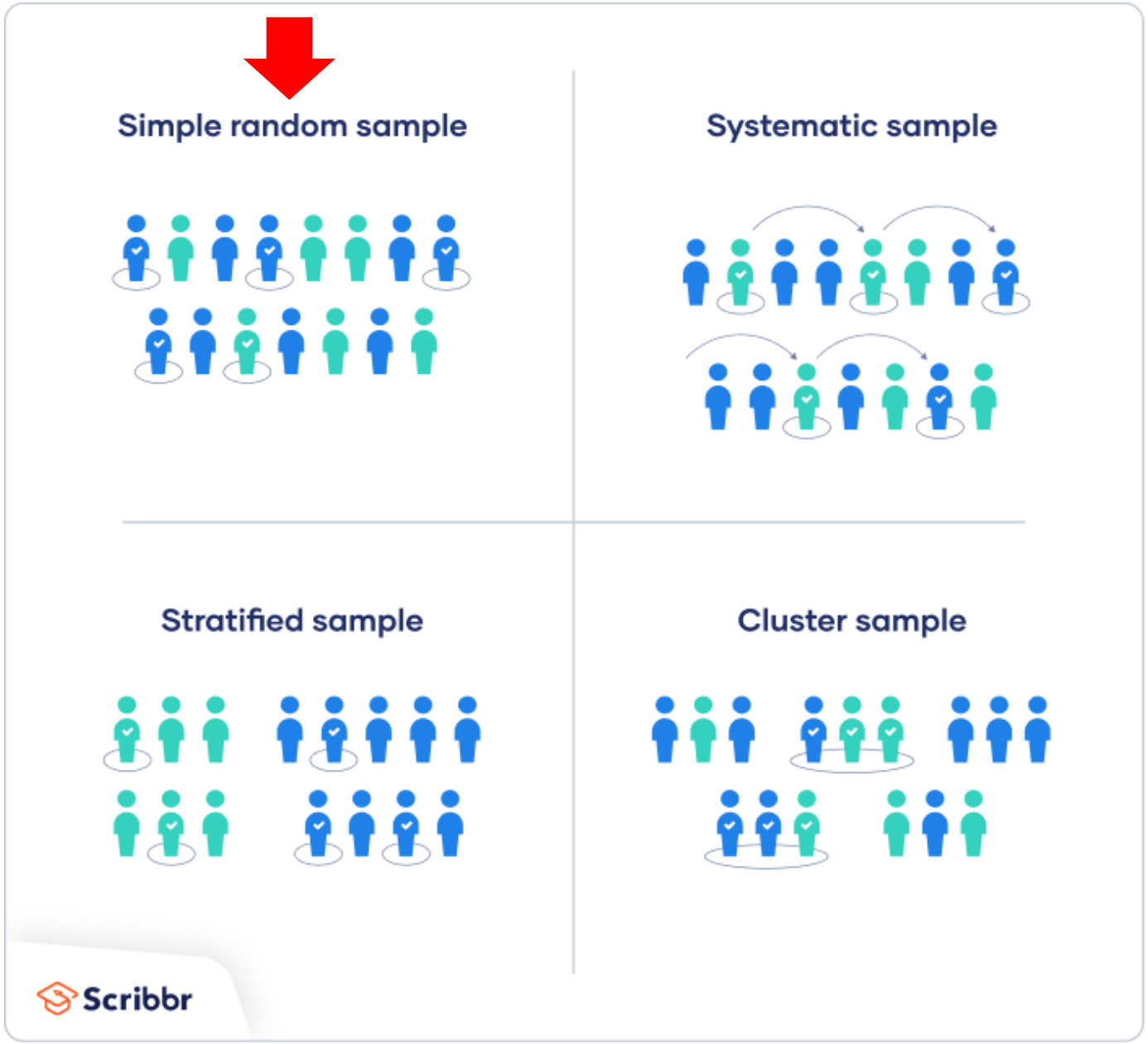
Simple Random Sample (SRS)

Each member of the population has the *same chance* of being selected (i.e., uniform over the population)

Example : American Community Survey (ACS)

Each year the US Census Bureau use *simple random sampling* to select individuals in the US. They follow those individuals for 1 year to draw conclusions about the US population as a whole.

Probability Sampling

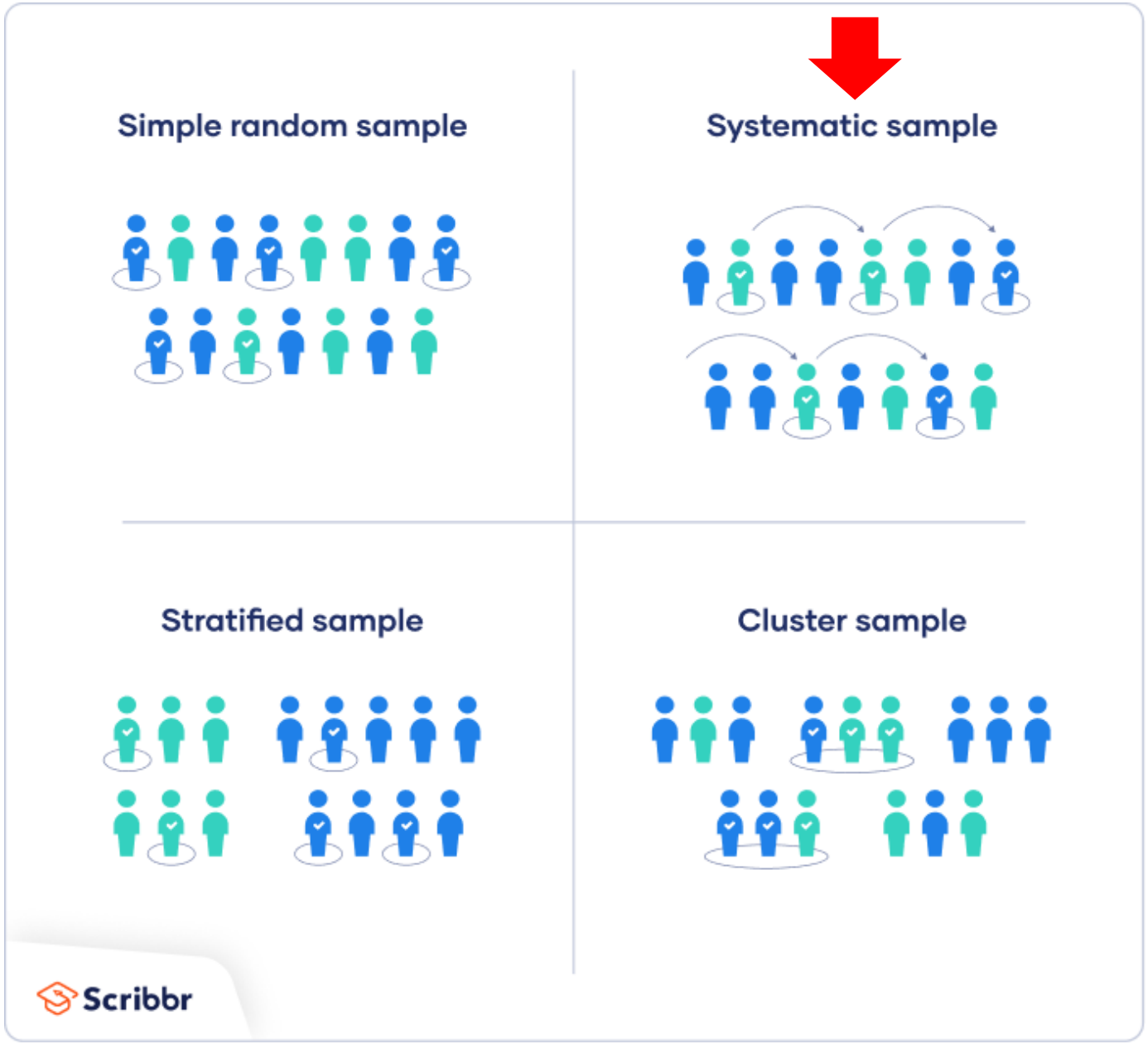


Simple Random Sample (SRS)

Each member of the population has the *same chance* of being selected (i.e., uniform over the population)

- Most straightforward probability sampling method
- Impractical unless you have a complete list of every member of population

Probability Sampling



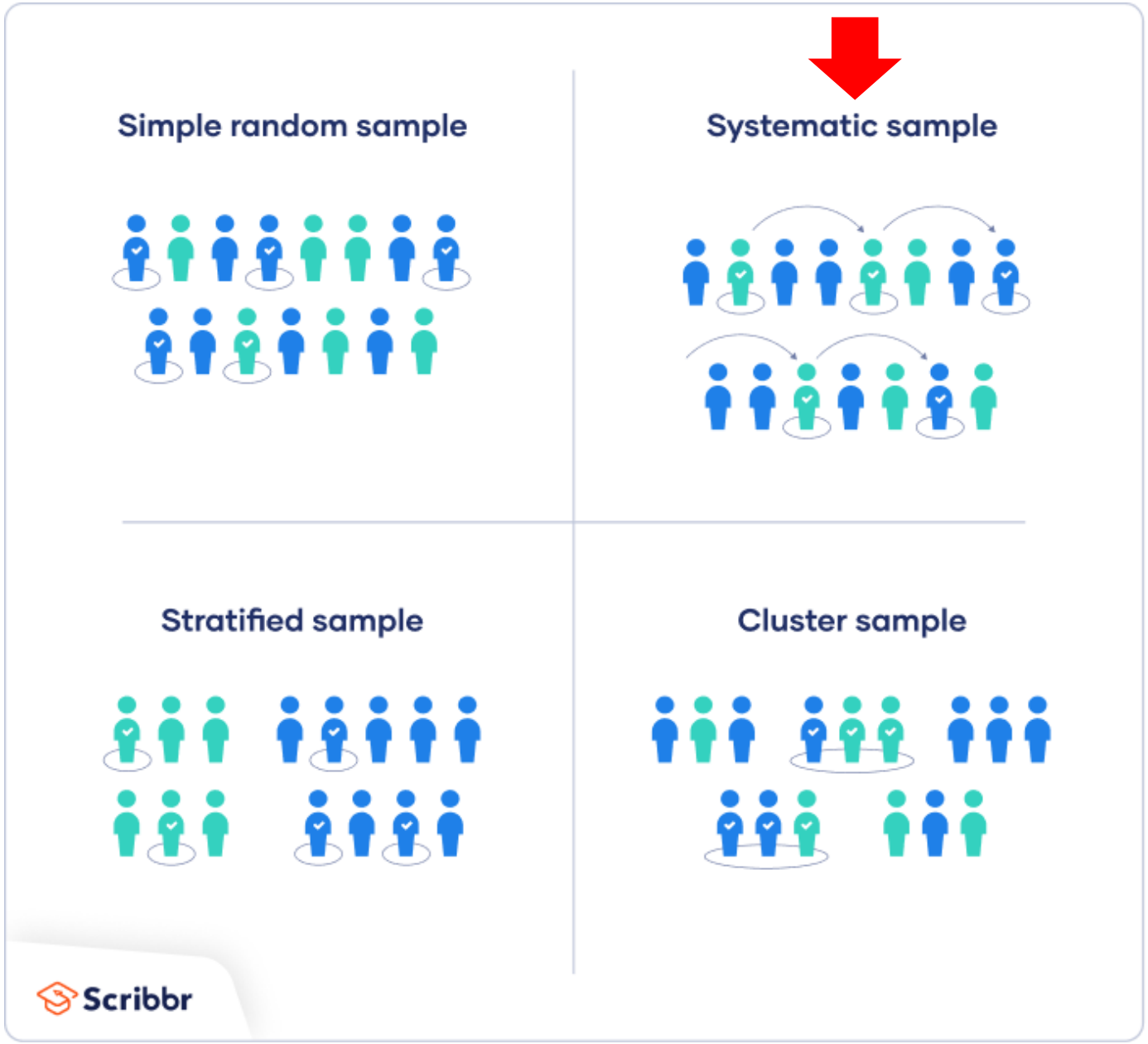
Systematic Sample

Select members of population at a regular interval, determined in advance

Example You own a grocery store and want to study customer satisfaction. You ask *every 20th customer* at checkout about their level of satisfaction.

Note We cannot itemize the whole population in this example, so SRS is not possible.

Probability Sampling

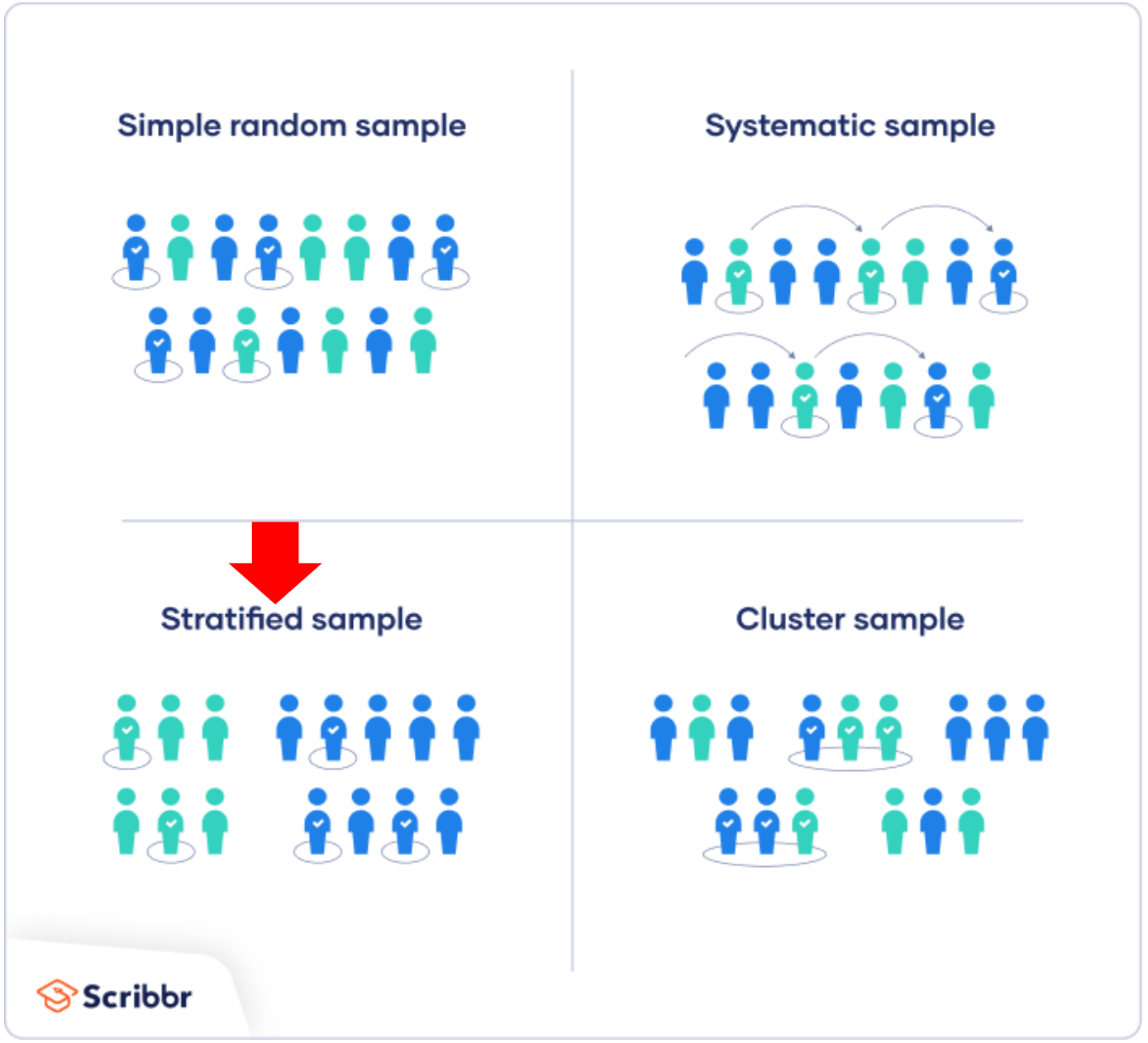


Systematic Sample

Select members of population at a regular interval, determined in advance

- Imitates SRS but is easier in practice
- Can even do systematic sampling when you can't access the entire population in advance
- **Do not** use when there can be a pattern. E.g., survey at the exit of a rollercoaster with N seats but with every N-th customer.

Probability Sampling

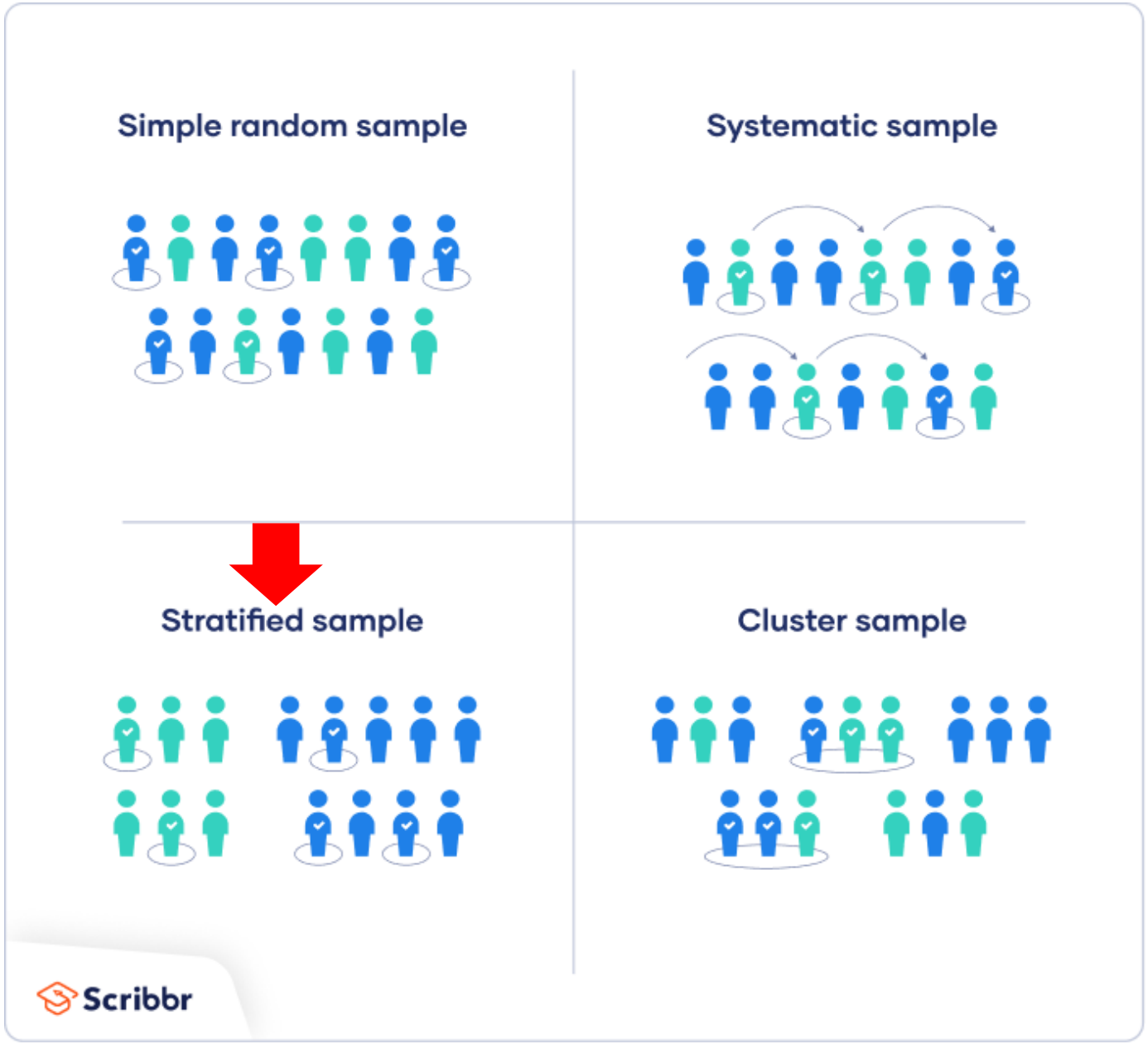


Stratified Sample

Divide population into *homogeneous* subpopulations (strata). Probability sample the strata.

Example We wish to solicit opinions of UA CS freshman by asking 100 of them, but they are about 14% women. SRS could easily fail to capture adequate number of women. We divide into men / women and perform SRS within each group.

Probability Sampling

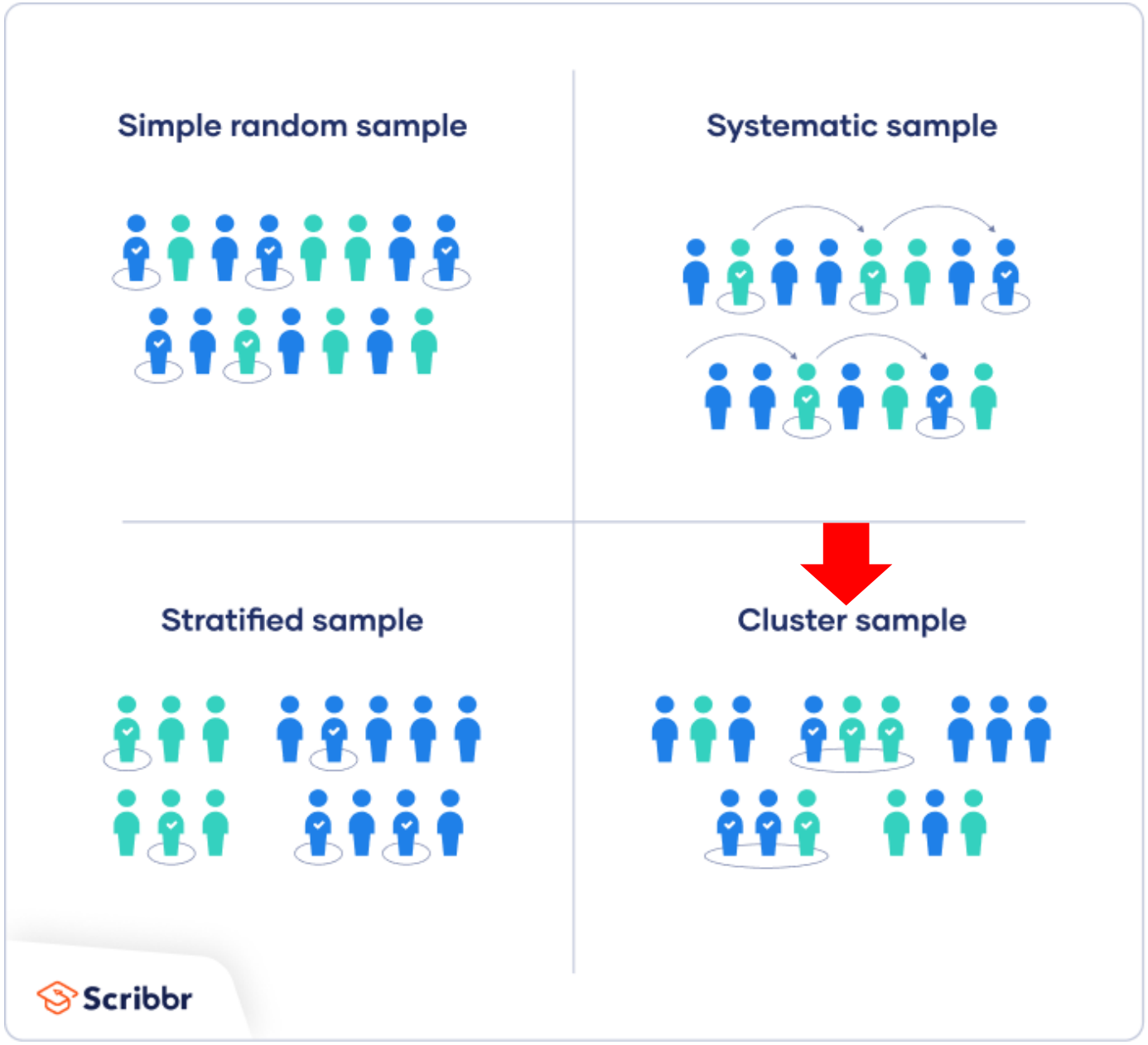


Stratified Sample

Divide population into *homogeneous* subpopulations (strata). Probability sample the strata.

- Use when population is diverse and want to accurately capture characteristic of each group
- Ensures similar variance across subgroups
- Lowers overall variance in the population

Probability Sampling

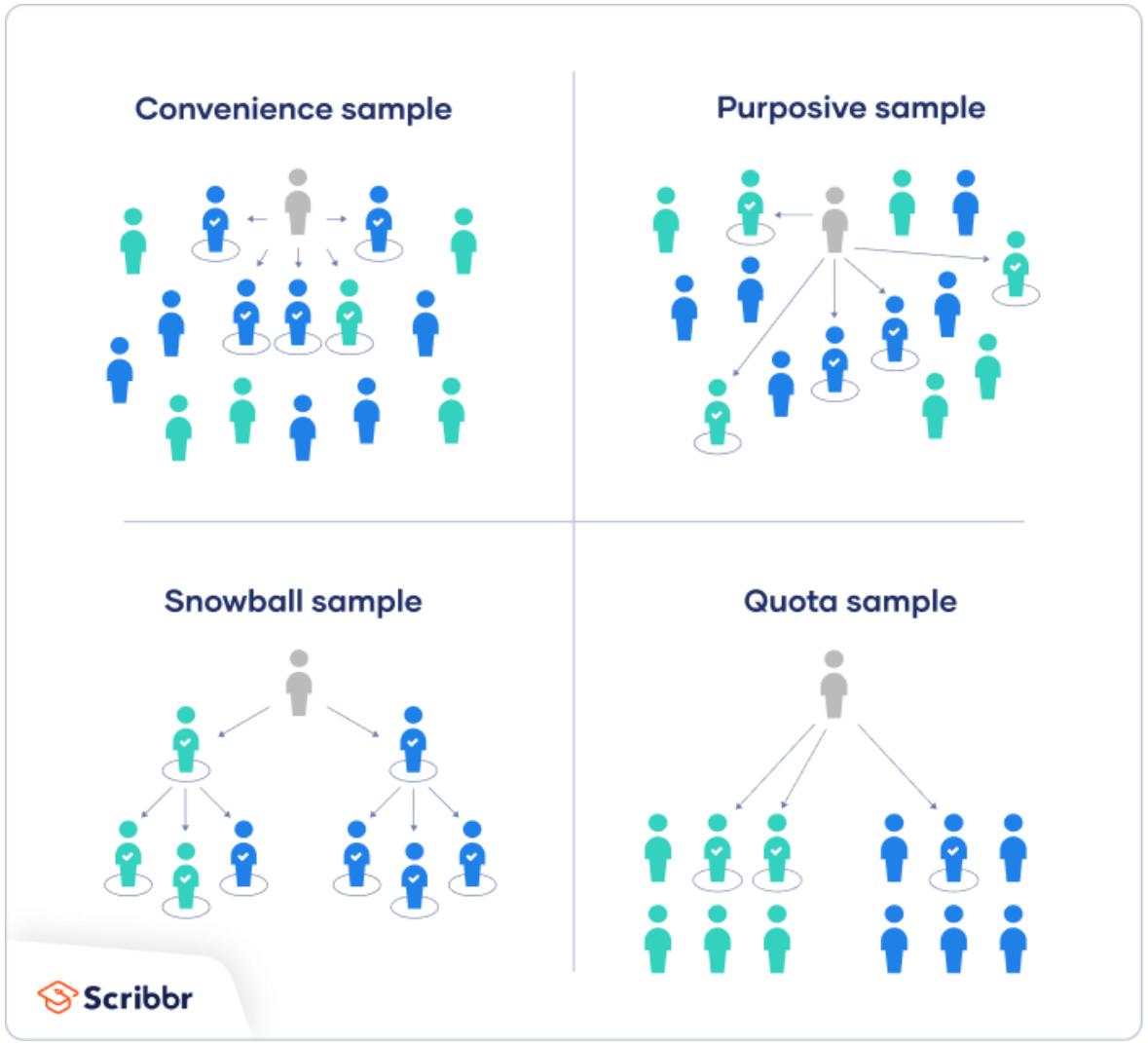


Cluster Sample

Divide population into subgroups (clusters). Randomly select entire clusters.

Example We wish to study the average reading level of *all 7th graders in the city* (population). Create a list of all schools (clusters) then randomly select a subset of schools and test every student.

Non-Probability Sampling



Easier to access data, but higher risk of *sample bias* compared to probability sampling

Usually used to perform *qualitative research* (e.g., gathering student opinions, experiences, etc.)

We will not focus on these, but you should be aware if your data are from non-probability methods