CSC380: Principles of Data Science

Data Analysis, Collection, and Visualization 3

Xinchen Yu

Acknowledgement: Built on Jason Pacheco, Kwang-Sung Jun, Chicheng Zhang's slides

Midterm solutions

9. Given a distribution D with unknown mean μ and variance σ^2, and a set of n iid samples $X_1, \ldots, X_n$ drawn from it. Define $\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ as an estimator of μ.
(a) (4 points) Is $\tilde{\mu}_n$ an unbiased estimator of μ? Justify your answer.
(b) (6 points) Let $n = 4$. What is the bias, variance, and Mean Square Error (MSE) of $\tilde{\mu}_4$, respectively? Note: For variance, you can compute $Var[\tilde{\mu}_4]$, in other words, $Var[\frac{X_1+X_2+X_3+X_4}{3}]$. (You can have μ, σ^2 or numbers in the results).

Lecture statistics 3, page 7

$$\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i \qquad E[\tilde{\mu}_n] = E\left[\frac{1}{n-1}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E[X_i]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \mu \quad = \frac{n\mu}{n-1}$$

$\tilde{\mu}_n$ is not an unbiased estimator of μ.

Midterm solutions

9. Given a distribution D with unknown mean μ and variance σ^2, and a set of n iid samples $X_1, \ldots, X_n$ drawn from it. Define $\tilde{\mu}_n = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ as an estimator of μ.
(a) (4 points) Is $\tilde{\mu}_n$ an unbiased estimator of μ? Justify your answer.
(b) (6 points) Let $n = 4$. What is the bias, variance, and Mean Square Error (MSE) of $\tilde{\mu}_4$, respectively? Note: For variance, you can compute $Var[\tilde{\mu}_4]$, in other words, $Var[\frac{X_1+X_2+X_3+X_4}{3}]$. (You can have μ, σ^2 or numbers in the results).

$$\tilde{\mu}_4 = \frac{1}{3}(X_1 + X_2 + X_3 + X_4)$$

$$\mathbf{Bias}(\tilde{\mu}_4) = E[\tilde{\mu}_4] - \mu$$

$$= \frac{4\mu}{3} - \mu$$

$$= \frac{\mu}{3}$$

$$\mathbf{Var}[\tilde{\mu}_4] = \mathbf{Var}\left[\frac{1}{3}(X_1 + X_2 + X_3 + X_4)\right]$$

$$= \frac{1}{9}\mathbf{Var}[X_1 + X_2 + X_3 + X_4]$$

Since the X_i are iid:

$$= \frac{1}{9}(\mathbf{Var}[X_1] + \mathbf{Var}[X_2] + \mathbf{Var}[X_3] + \mathbf{Var}[X_4])$$

$$= \frac{1}{9}(4\sigma^2)$$

$$\mathbf{MSE}(\tilde{\mu}_4) = \mathbf{Var}[\tilde{\mu}_4] + \mathbf{Bias}(\tilde{\mu}_4)^2 = \frac{4\sigma^2 + \mu^2}{9}$$

Lecture statistics 3, page 8 and 18

Participation

- A total of 10 points (10% of final grade).

- Ask or answer questions in-person in / after the class: +1 point
  - Don’t forget to let me know your names after class :)

- Ask OR answer 1 question in the lecture OR on piazza: +1 point
  - Also apply to participations before midterm

- Attend office hour once: +1 point

Outline

- Data Visualization
- Data Summarization
- Data Collection and Sampling

Motivation

Not understanding how data are collected is one of the top reasons behind bad data science…

…we will not do data collection or experimental design, but should be familiar with the basics

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

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. Helena, Montana banned smoking ban in all public spaces for six months –before/after this ban)

- **Case Studies and Surveys**. Analysis based on previously-collected data. (e.g., Analysis of US census data, or US current population survey (CPS))

Recall: Explanatory, response variable

Example: Say we study the relationship between Smoking vs Cancer

Independent variable: variables that are manipulated or are changed by researchers and whose effects are measured and compared.

= explanatory variable

Dependent variable: the variable that depends on independent variable (or speculated to do so).

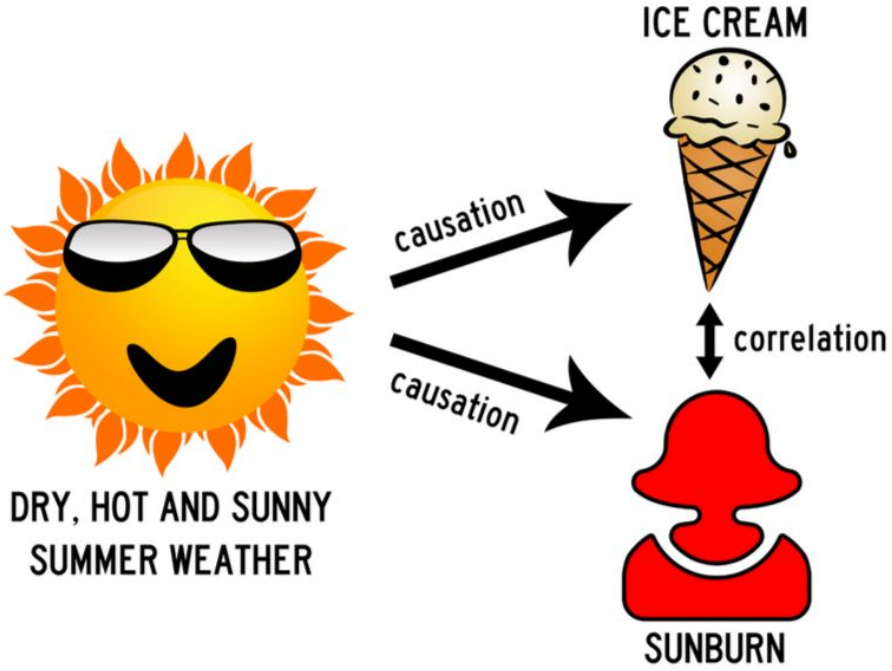
= response variable

Recall: Explanatory, response variable

Research question	Independent variable	Dependent variable(s)
Do tomatoes grow fastest under fluorescent, incandescent, or natural light?	• Type of light the tomato plant is grown under	• The rate of growth of the tomato plant
What is the effect of intermittent fasting on blood sugar levels?	• Presence or absence of intermittent fasting	• Blood sugar levels
Is medical marijuana effective for pain reduction in people with chronic pain?	• Presence or absence of medical marijuana use	• Frequency of pain • Intensity of pain
To what extent does remote working increase job satisfaction?	• Type of work environment (remote or in office)	• Job satisfaction self-reports

Causation vs. Correlation

Studies generally try to show *either* correlation (association) or causation, but they are not the same…



Confounding Variables

A variable that influences the *response* but is unaccounted for in data collection

Example: You are studying whether **birth order** affects **Down's Syndrome** in the child. You collect samples of children, their birth order, and cases of Down's syndrome.

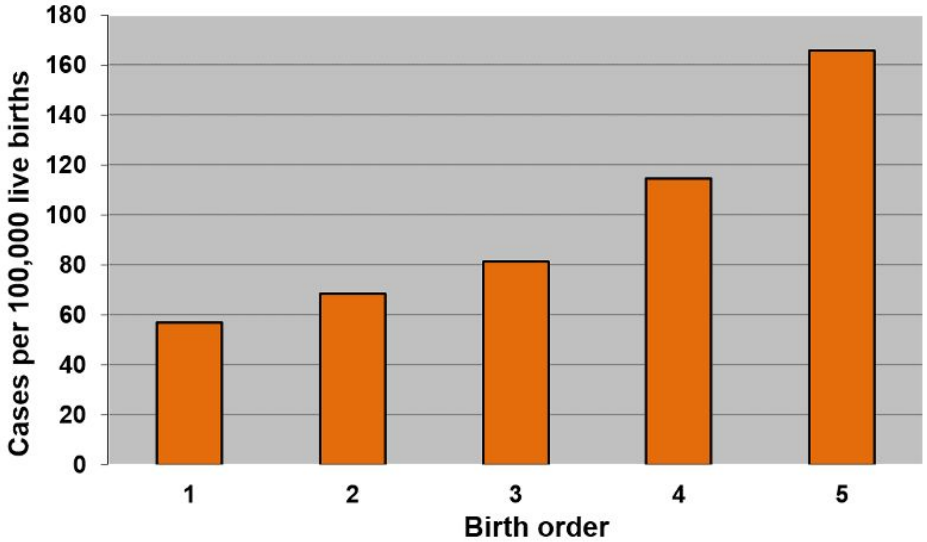
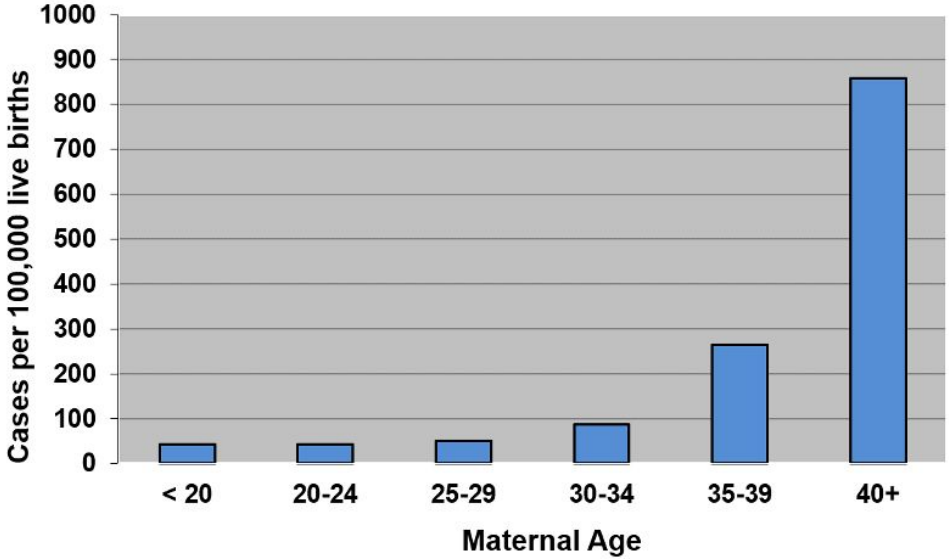


chart from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module11-Confounding-EMM/PH717-Module11-Confounding-EMM3.html

Confounding Variables

- You went on to collecting the maternal age data.



So.. both maternal age and birth order is associated with Down's syndrome?

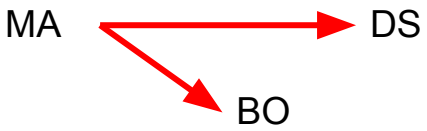
Confounding Variables

A variable that influences the *response* but is unaccounted for in data collection

Example: You are studying whether **birth order** affects **Down's Syndrome** in the child. You collect samples of children, their birth order, and cases of Down's syndrome.

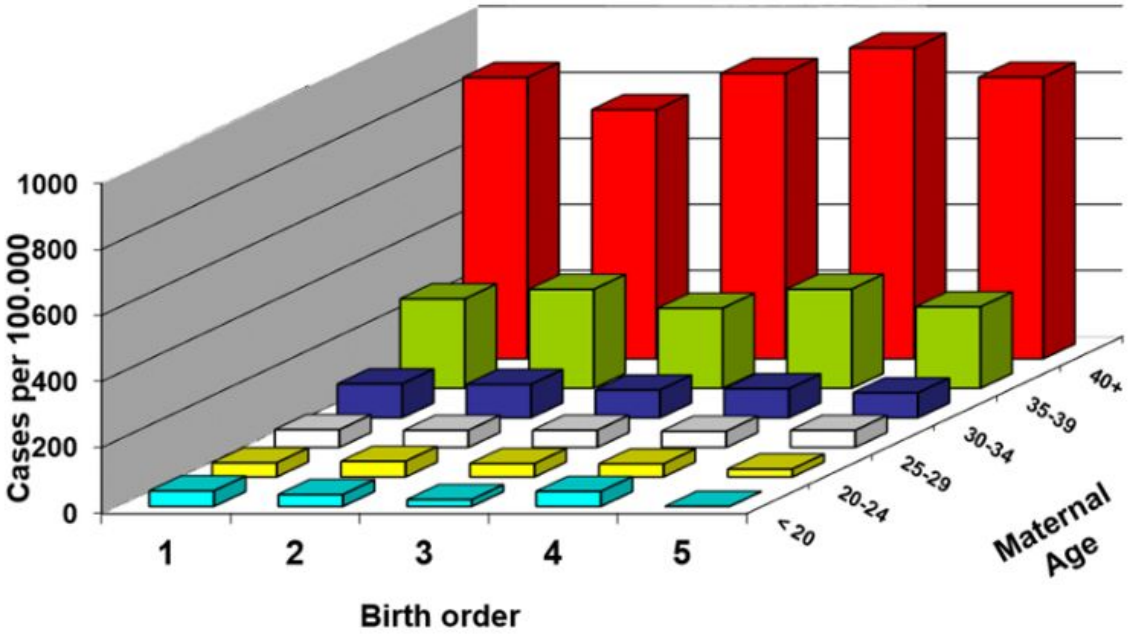
Explanation: Maternal age (confounder) was not recorded. Two scenarios:

1. Higher maternal age is directly associated with Down's syndrome, regardless of birth order.
2. Maternal age directly assoc. with birth order (mother is older with later children), but not directly associated with Down's syndrome.



Controlling for Confounders

Stratified Sampling: Divide population into smaller groups. Previous example can divide population of children by maternal age at birth and collect data from each stratum



Approach

1. **Control** confounders: design treatments
2. **Randomize** the assignment of subjects to treatments (to eliminate bias due to systematic differences in categories)
3. **Replicate** experiment on many subjects, to obtain statistically meaningful results

Example: Pfizer COVID Phase 3 Vaccine Trials

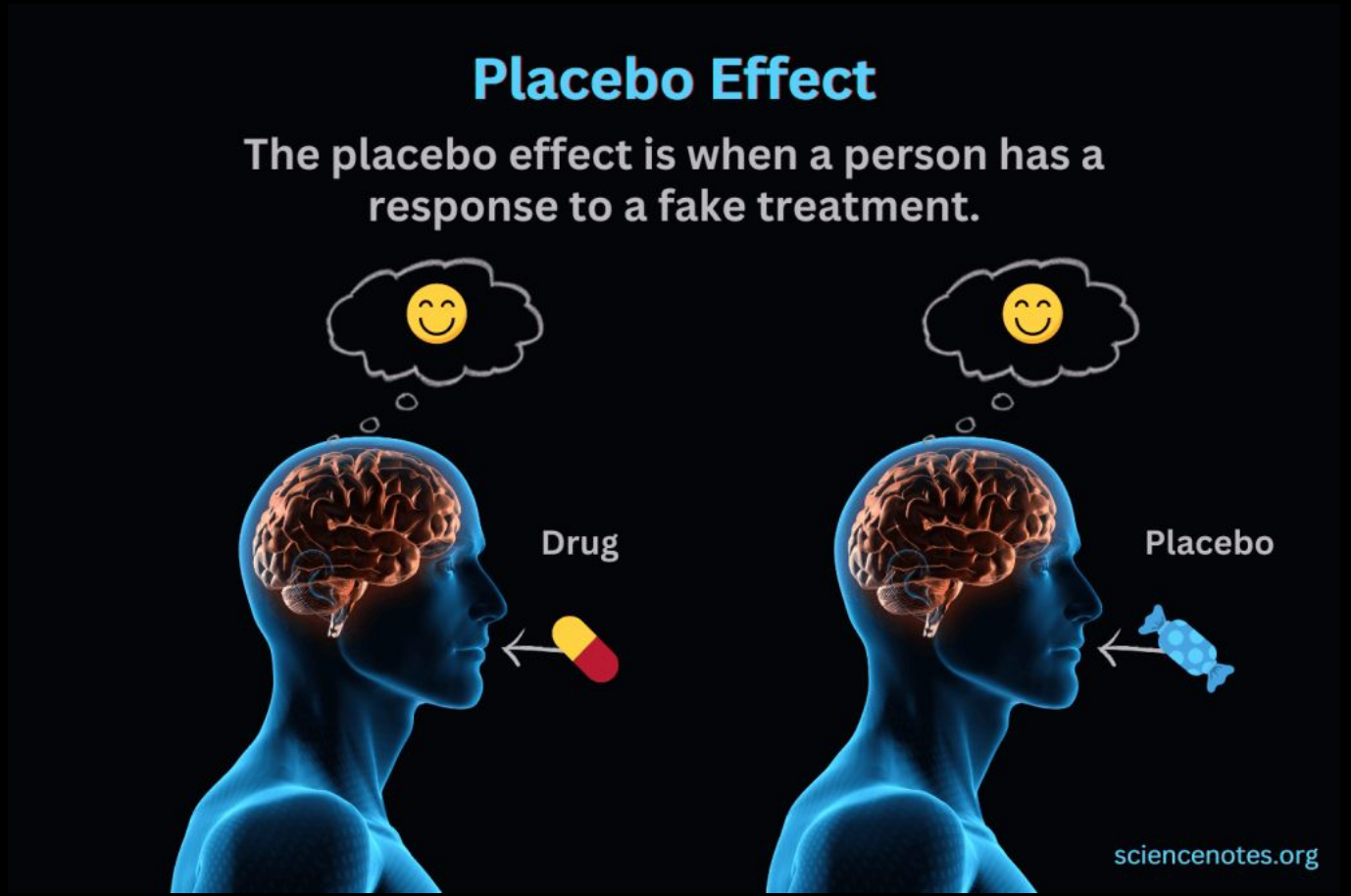
the tendency of any medication or treatment, even an inert or ineffective one, to exhibit results simply because the recipient believes that it will work.

1. **Placebo Control:** Subjects are randomly selected to receive either the vaccine or an injection of saline solution

2. **Randomize:** Stratified sampling with age strata: 12-15yrs, 16-55yrs, 55+yrs

3. **Replicate:** Experiment is repeated at multiple sites in several countries

Placebo Effect



Example: Pfizer COVID Phase 3 Vaccine Trials

The landmark phase 3 clinical trial enrolled **46,331** participants at **153** clinical trial sites around the world.

Trial Geography

Our trial sites are located in **Argentina, Brazil, Germany, Turkey, South Africa** and the **United States.**

Participant Diversity

Approximately **42%** of overall and **30%** of U.S. participants have diverse backgrounds.

Participants	Overall Study	U.S. Only
Asian	5%	6%
Black	10%	10%
Hispanic/Latinx	26%	13%
Native American	1.0%	1.3%

49.1% of participants are male and **50.9%** are female

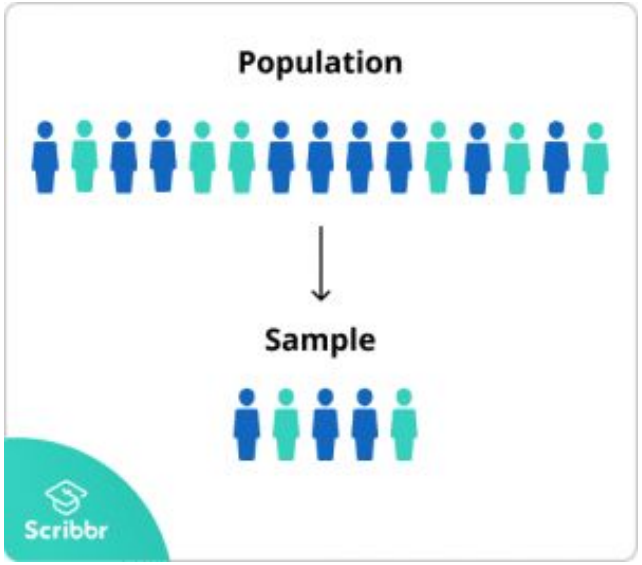
Participant Age

Ages 12-15	2,260
Ages 16-17	754
Ages 18-55	25,427
Ages 56+	17,879

Data Collection

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.

Reasons for Sampling

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

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

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

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.***

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 Methods

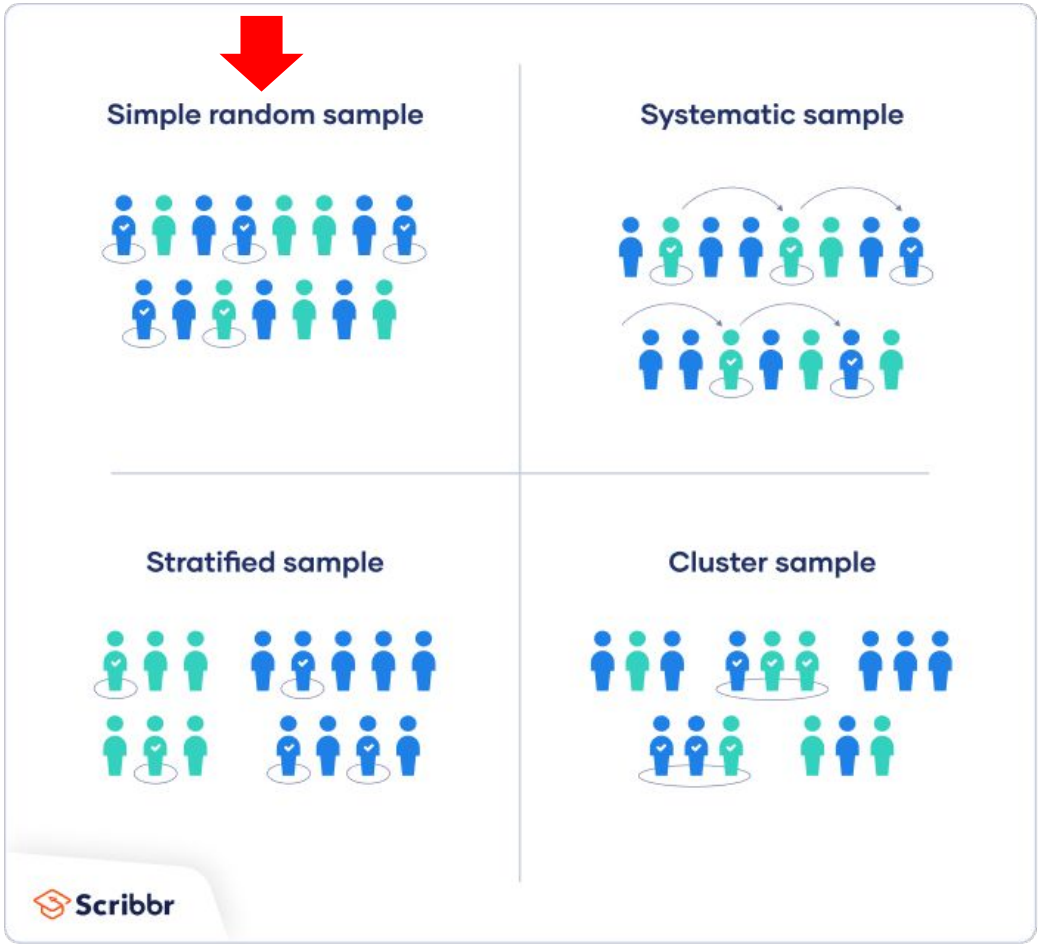
Sampling must be conducted properly, to avoid sample bias

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



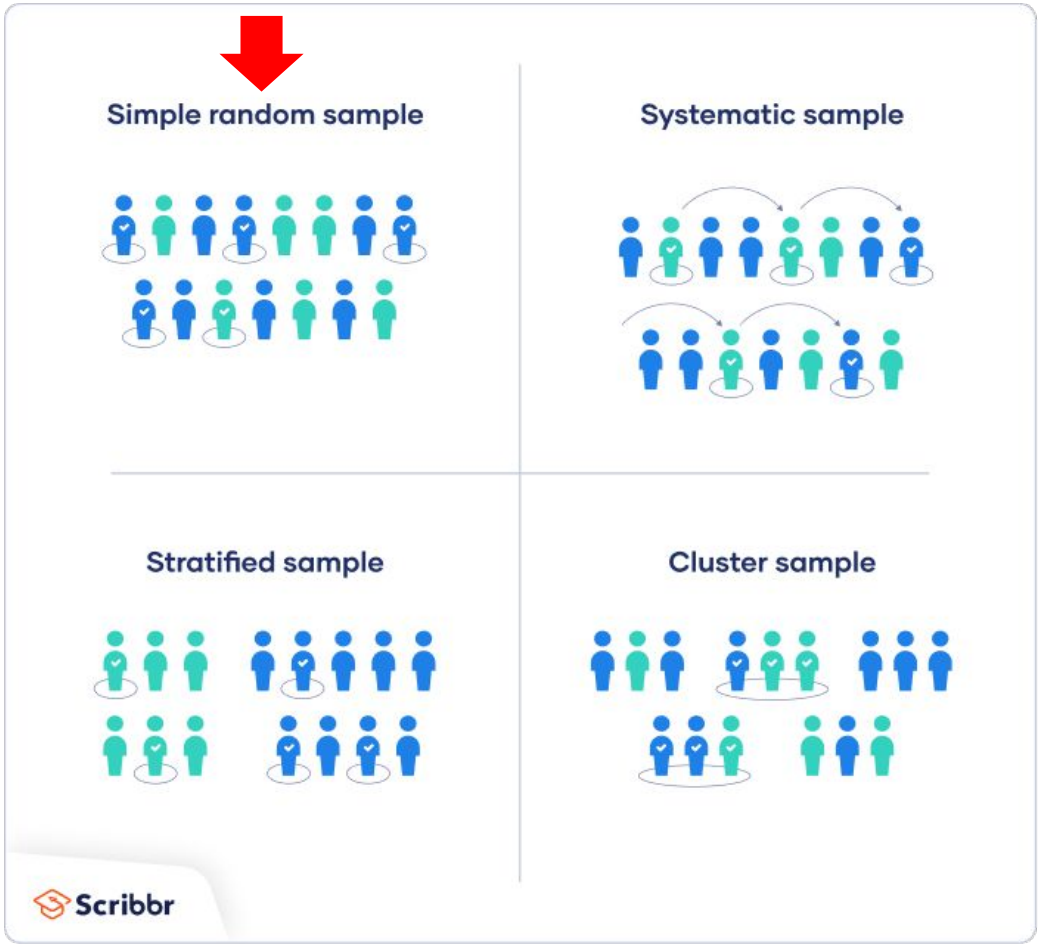
Simple Random Sample (SRS)

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

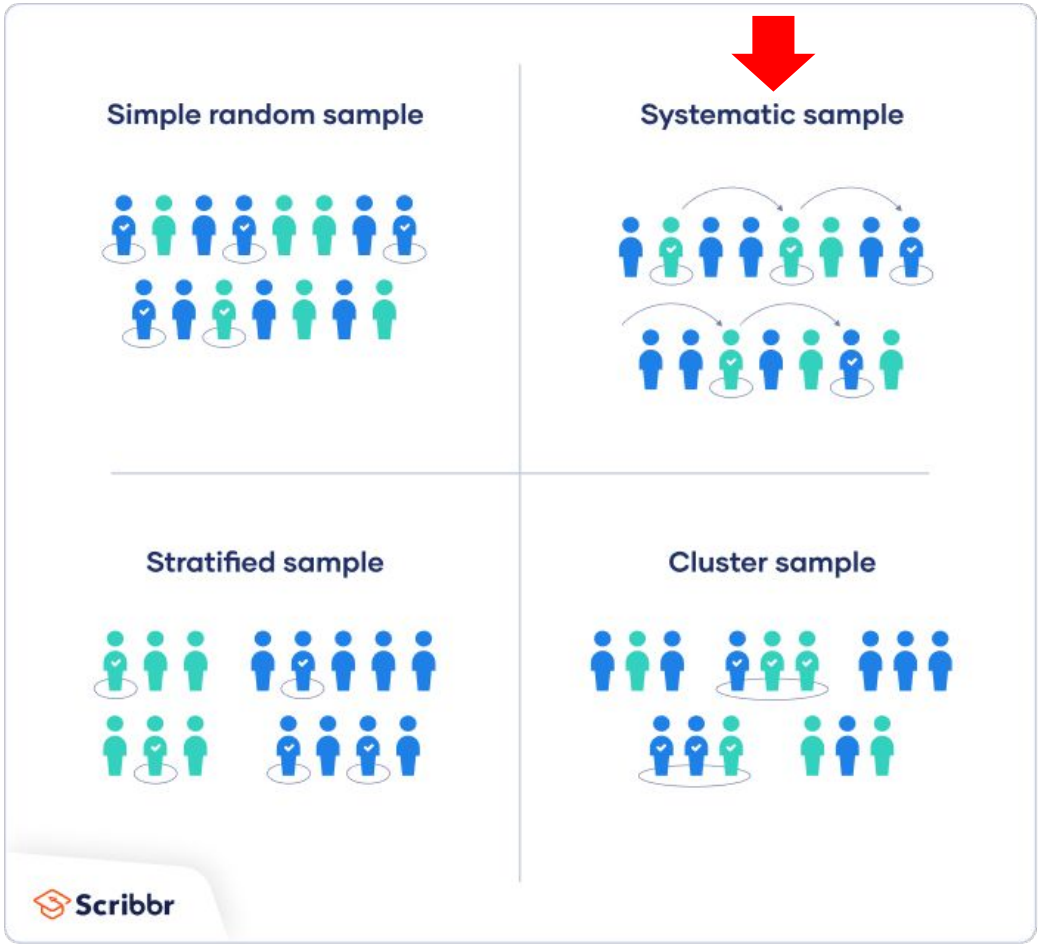


Simple Random Sample (SRS)

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



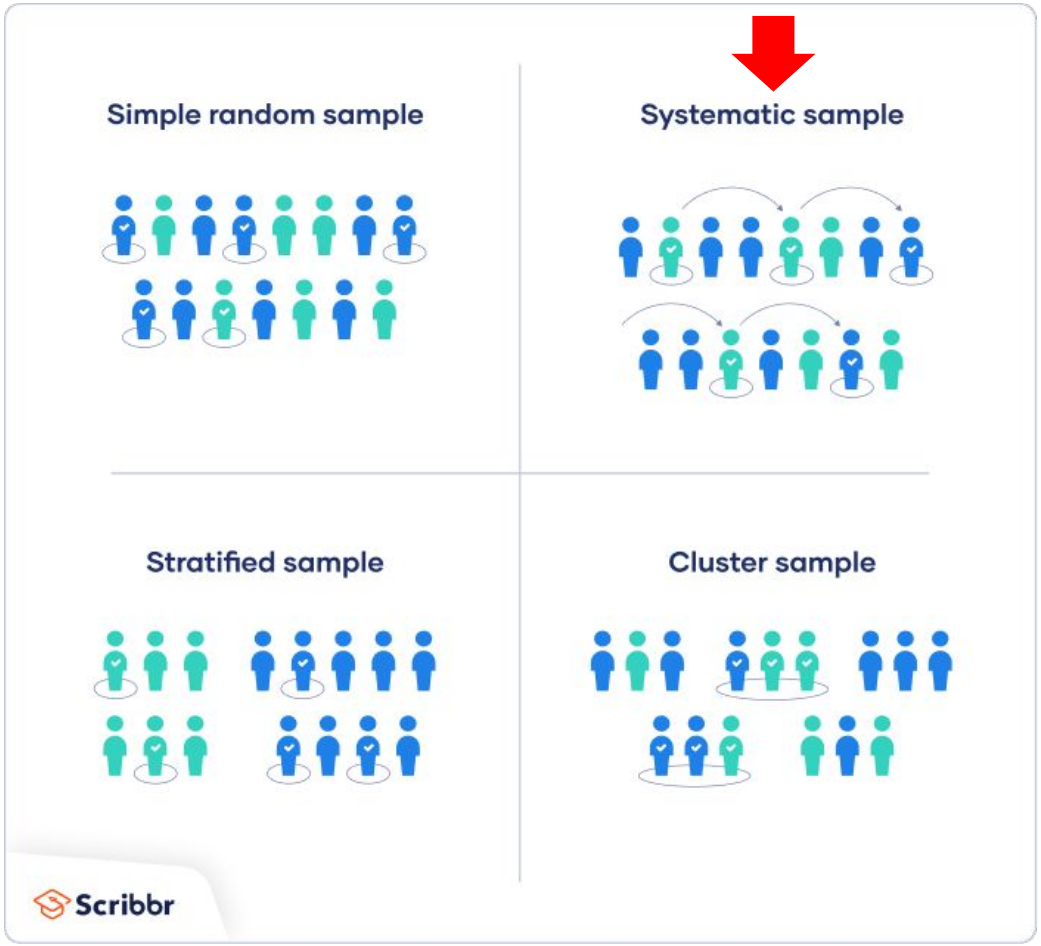
Systematic Sample

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



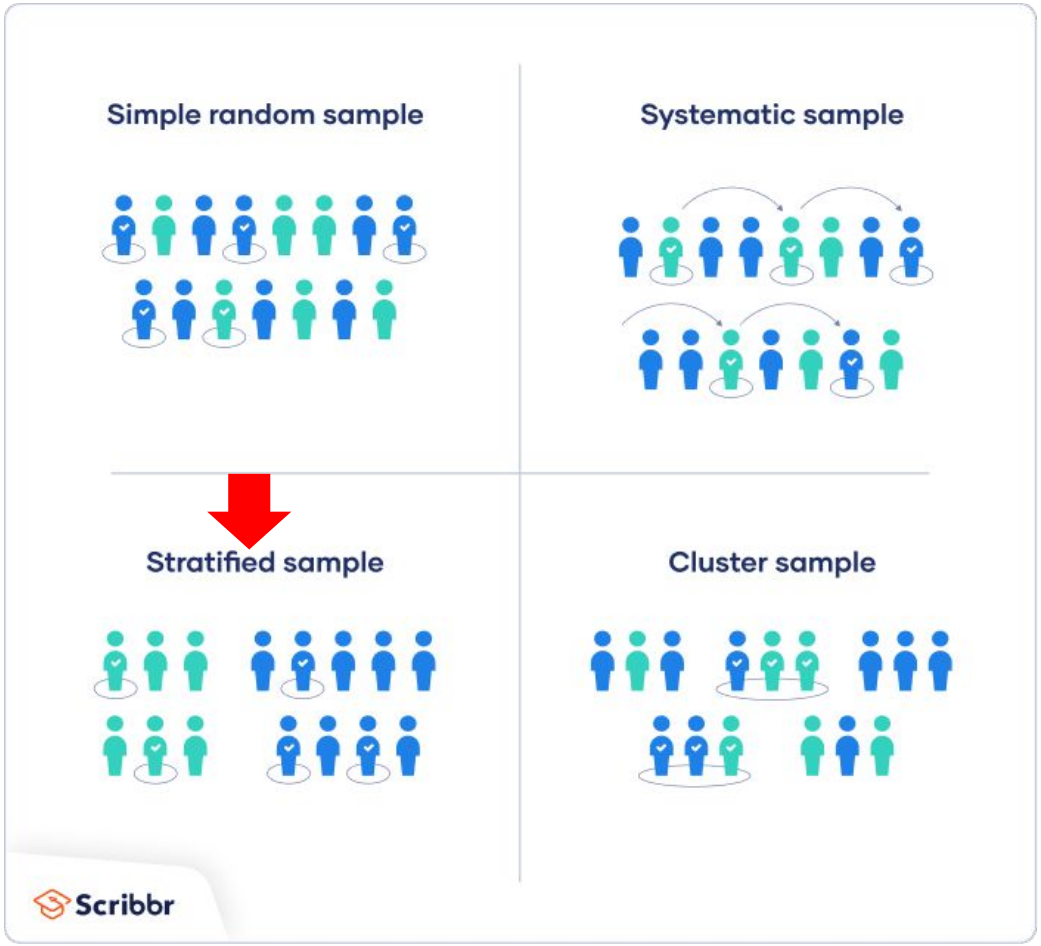
Systematic Sample

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

- Imitates SRS but is easier in practice
- **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.

Alternative: use a Bernoulli(p) (e.g., p=1/20)

Probability Sampling

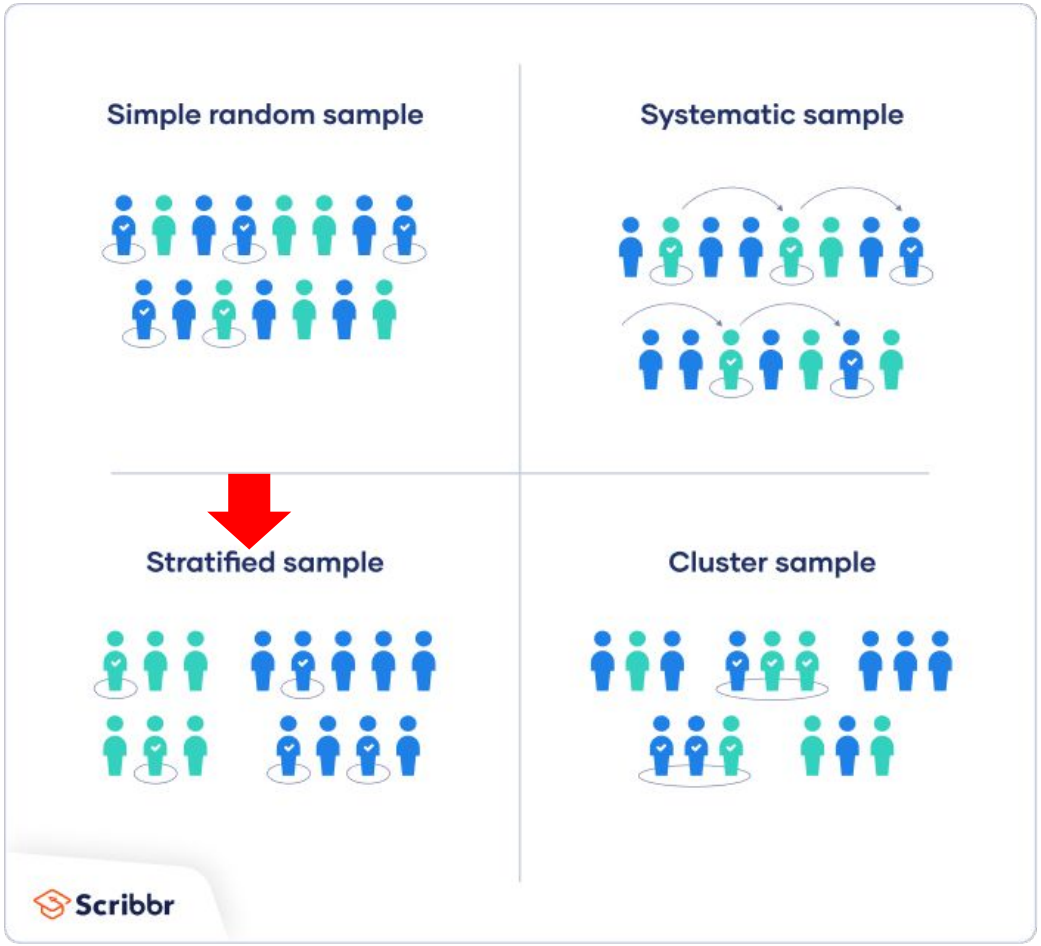


Stratified Sample

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

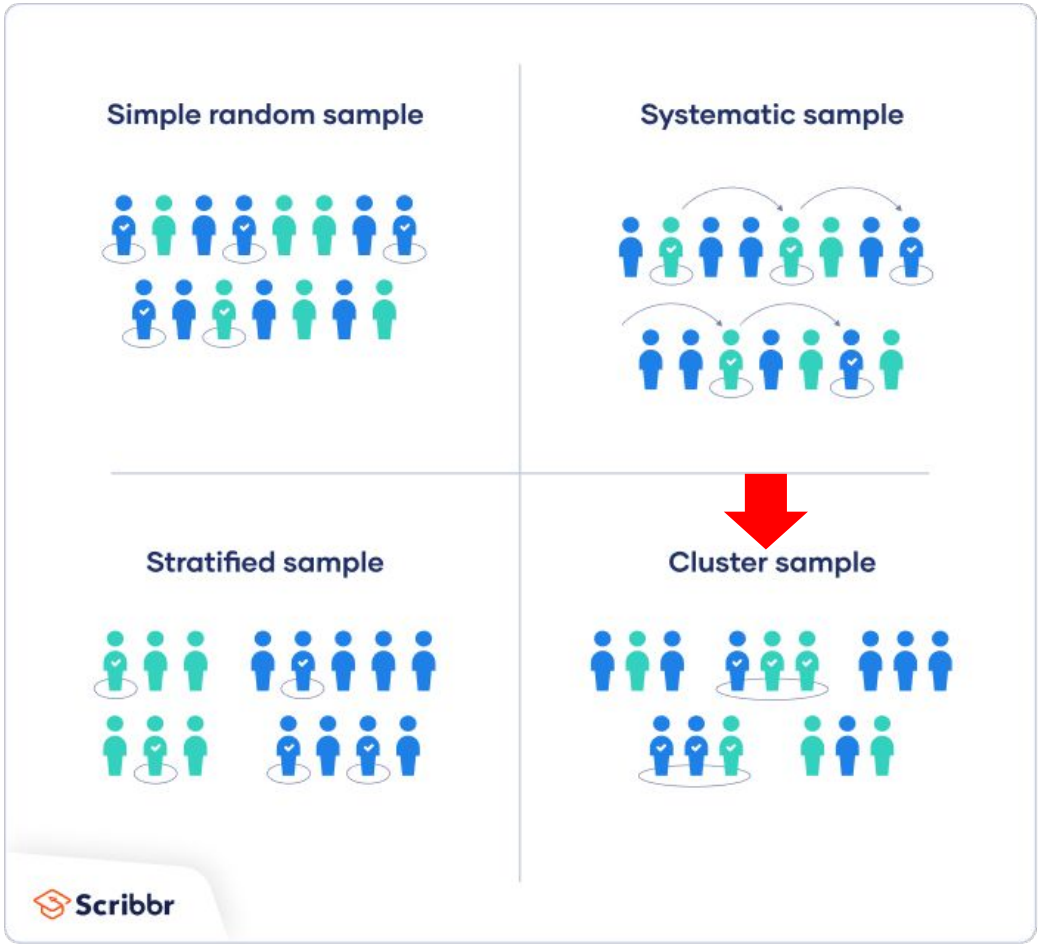


Stratified Sample

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

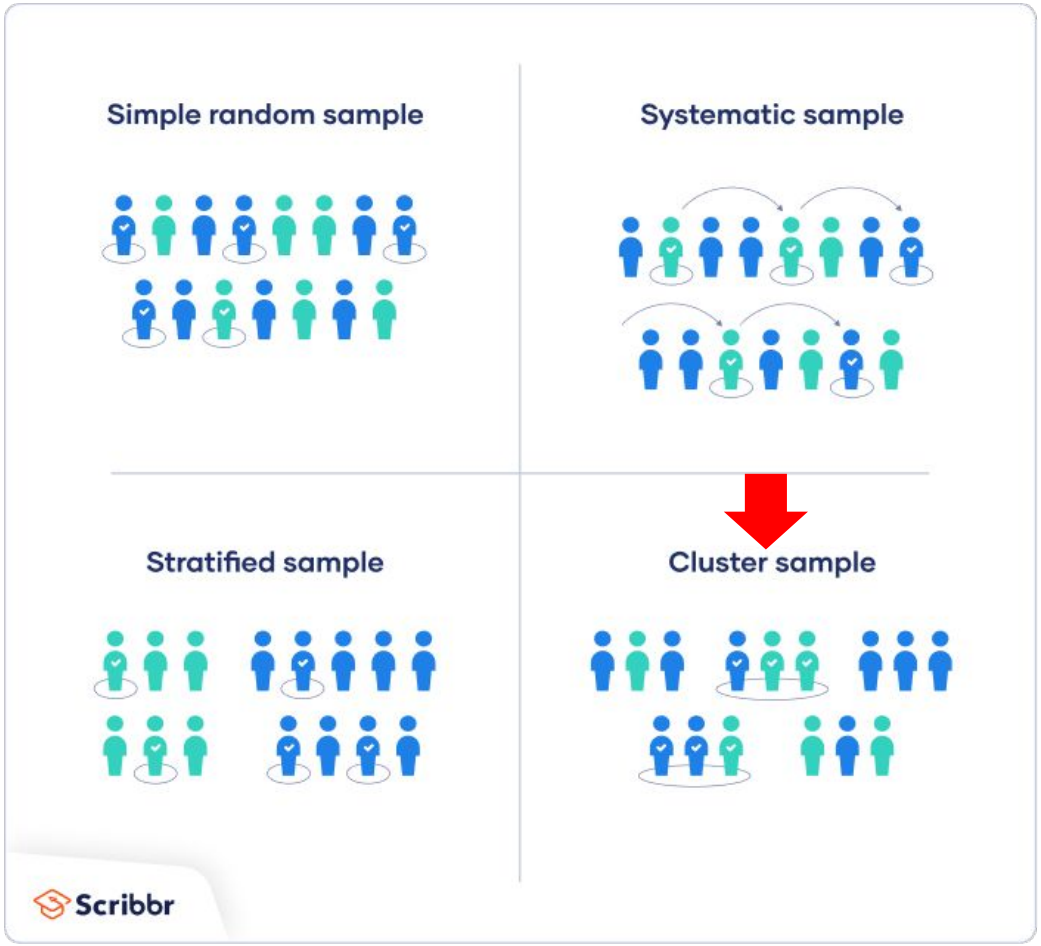


Cluster Sample

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.

Probability Sampling

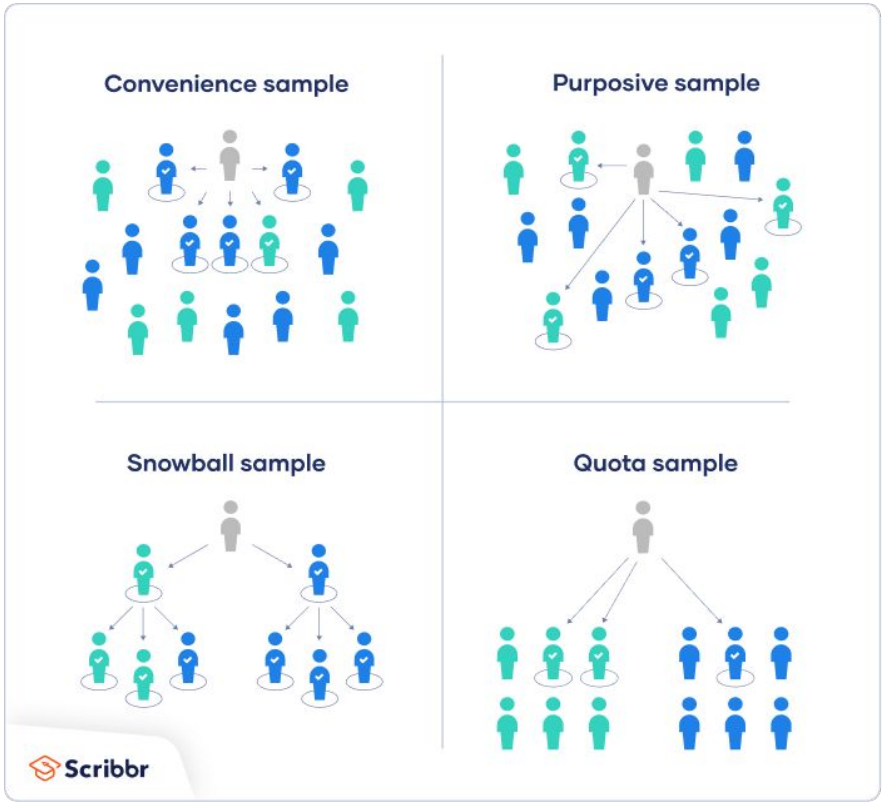


Cluster Sample

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

- This is *single-stage* cluster sampling
- *Multi-stage* avoids sampling every member of a group
- Related to stratified sampling, but groups are not homogeneous

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

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) you 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**.

Common Types of Sampling Bias

Self-selection: Possible whenever members under study have control over whether to participate.

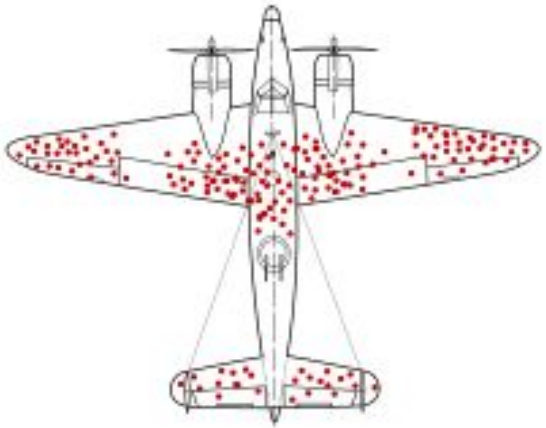
- E.g., online or phone-in poll—user can choose whether to initiate participation.

Exclusion: Excluding certain groups from the sample.

- E.g., longitudinal data collection for 2 years in a town where we exclude groups that move in or out.

Survivorship: Only *surviving* subjects are selected.

- E.g., studying improvements on the retention rate of CS majors by surveying CS graduates.



(Wikipedia)

Review

- Data Visualization
  - matplotlib.pyplot; see the documentation & tutorials

- Data Summarization
  - scipy for more advanced functionality
  - Anscomb's quartet: importance of visualization.

- Research Design
  - Randomized control, observational.
  - Correlation vs causation
  - Confounding variables: could cause correlation that may disappear after observing the confounding variable.
  -
- Data Collection and Sampling
  - Sampling methods: SRS, systematic, stratified, cluster.
  - Look out for the bias: self-selection, exclusion, survivorship.

Example of Bias in a Simple Random Sample

SRS is least prone to bias, but not always…

You want to study procrastination and social anxiety levels in undergraduate students at your university using a simple random sample. You assign a number to every student in the research participant database from 1 to 1500 and use a random number generator to select 120 numbers.

What is the cause of bias in this simple random sample?

https://www.scribbr.com/methodology/sampling-bias/

Example of Bias in a Simple Random Sample

SRS is least prone to bias, but not always…

You want to study procrastination and social anxiety levels in undergraduate students at your university using a simple random sample. You assign a number to every student in the research participant database from 1 to 1500 and use a random number generator to select 120 numbers.

Although you used a random sample, not every member of your target population –undergraduate students at your university – had a chance of being selected. Your sample misses anyone who did not sign up to be contacted about participating in research. This may bias your sample towards people who have less social anxiety and are more willing to participate in research.

https://www.scribbr.com/methodology/sampling-bias/