Lesson 1: Data collection

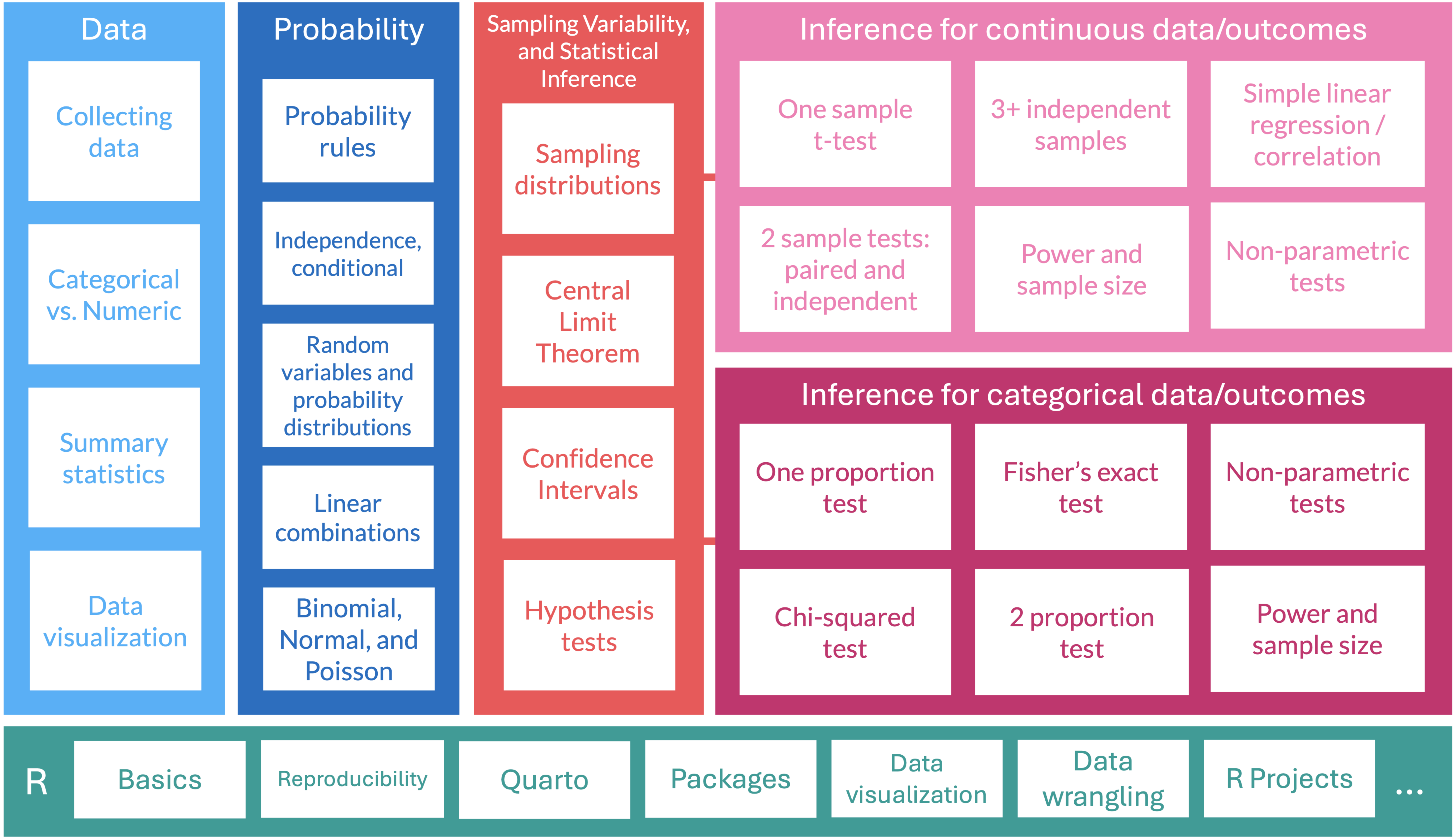
Nicky Wakim, Meike Niederhausen

2024-09-30

Learning Objectives

1. Define and compare a target population and its sample.
2. Explain different sampling methods and understand their advantages.
3. Define and compare experiments and observational studies.

Where are we?



Data

Collecting data

Categorical vs. Numeric

Summary statistics

Data visualization

Probability

Probability rules

Independence, conditional

Random variables and probability distributions

Linear combinations

Binomial, Normal, and Poisson

Sampling Variability, and Statistical Inference

Sampling distributions

Central Limit Theorem

Confidence Intervals

Hypothesis tests

Inference for continuous data/outcomes

One sample t-test

3+ independent samples

Simple linear regression / correlation

2 sample tests: paired and independent

Power and sample size

Non-parametric tests

Inference for categorical data/outcomes

One proportion test

Fisher’s exact test

Non-parametric tests

Chi-squared test

2 proportion test

Power and sample size

R

Basics

Reproducibility

Quarto

Packages

Data visualization

Data wrangling

R Projects

…

Learning Objectives

1. Define and compare a target population and its sample.

2. Explain different sampling methods and understand their advantages.

3. Define and compare experiments and observational studies.

Poll Everywhere Question 1

Asking a question

- Data provide evidence that help us answer questions
- But we need to start with a clearly articulated question

- In this class, we will be working towards **articulating our questions** and **systemically answering them with data**

- How do we formulate a clearly articulated question?

Examples of questions

1. Do bluefin tuna from the Atlantic Ocean have particularly high levels of mercury, such that they are unsafe for human consumption?

2. For infants predisposed to developing a peanut allergy, is there evidence that introducing peanut products early in life is an effective strategy for reducing the risk of developing a peanut allergy?

3. Does a recently developed drug designed to treat glioblastoma, a form of brain cancer, appear more effective at inducing tumor shrinkage than the drug currently on the market?

Population vs. sample

(Target) Population

- Group of interest being studied
- Group from which the sample is selected
  - Studies often have *inclusion* and/or *exclusion* criteria
- Almost always too expensive or logistically impossible to collect data for every case in a population

Sample

- Group on which data are collected
- Often a **small subset** of the population
- Easier to collect data on
- If we do it right, we might be able to answer our question about the target population

Identifying the target population

Let’s focus on the second research question:

2. For infants predisposed to developing a peanut allergy, is there evidence that introducing peanut products early in life is an effective strategy for reducing the risk of developing a peanut allergy?

Poll Everywhere Question 2

Identifying the target population

Let’s focus on the second research question:

2. For infants predisposed to developing a peanut allergy, is there evidence that introducing peanut products early in life is an effective strategy for reducing the risk of developing a peanut allergy?

What is the target population here?

- Infants predisposed to developing a peanut allergy
- We could get more specific with “Infants aged 0 to 5 years old” or “Infants aged 0 to 5 years old who have eczema, egg allergy, or both”
    - In this case we are defining exactly how old “infants” are and what “predisposed” means

From target population to sample

- Once we have a well articulated target populaton, we have inclusion or exclusion criteria for individuals
- Now we can start sampling from our target population…

Learning Objectives

1. Define and compare a target population and its sample.
2. Explain different sampling methods and understand their advantages.
3. Define and compare experiments and observational studies.

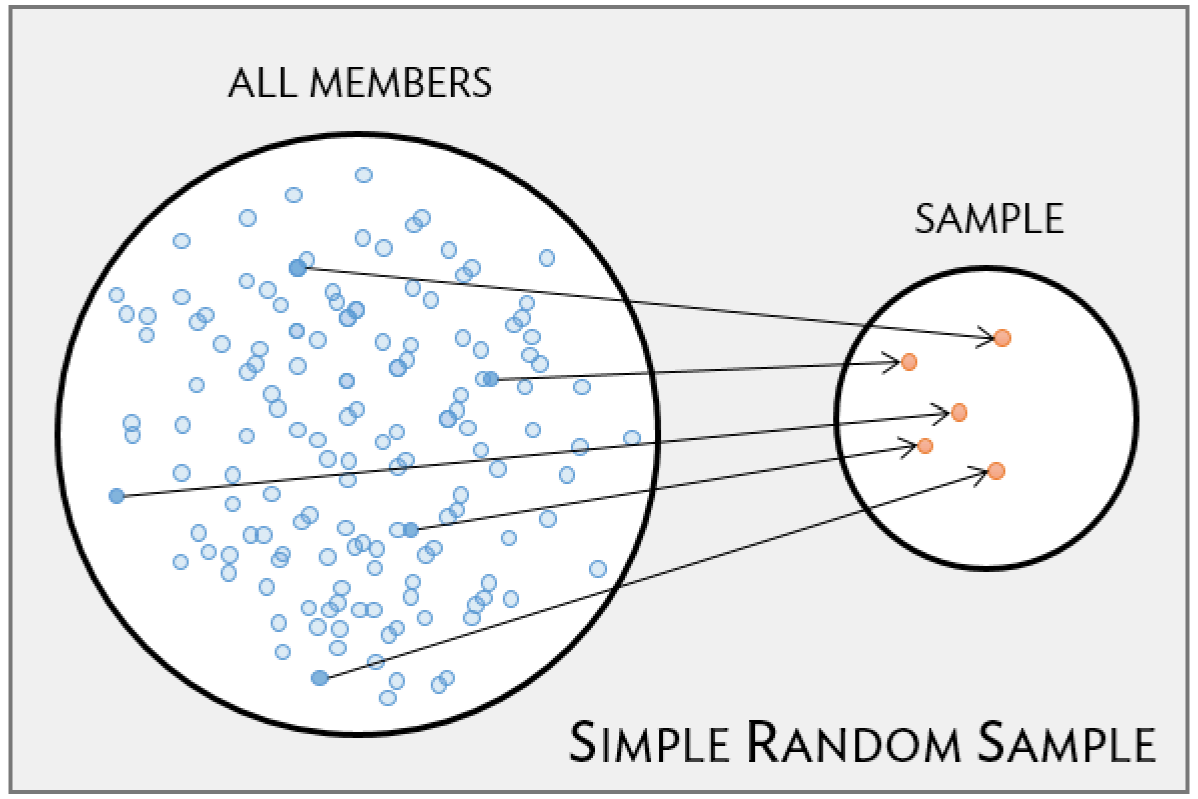
Sampling

- Goal is to get a **representative** sample of the population: the characteristics of the sample are similar to the characteristics of the population
- There are different ways to sample from a target population
- Types of sampling that we discuss
  - Simple random sample (SRS)
  - Convenient sample
  - Stratified sample
  - Cluster sample
  - Multistage sample

Sampling methods: Basic approaches

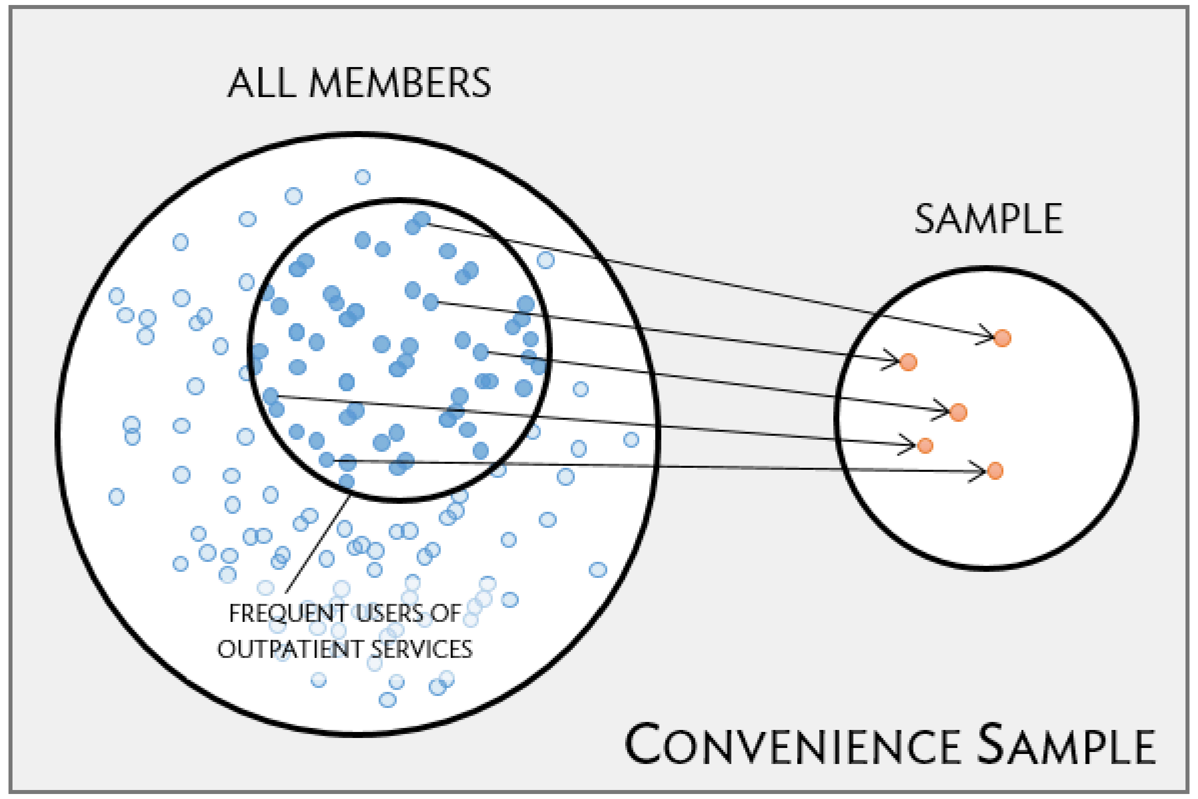
Simple random sample (SRS)

- Each individual of a population has the *same chance* of being sampled
- Randomly sampled
- Considered best way to sample



Convenience sample

- Easily accessible individuals are *more likely* to be included in the sample than other individuals
- A common "pitfall"

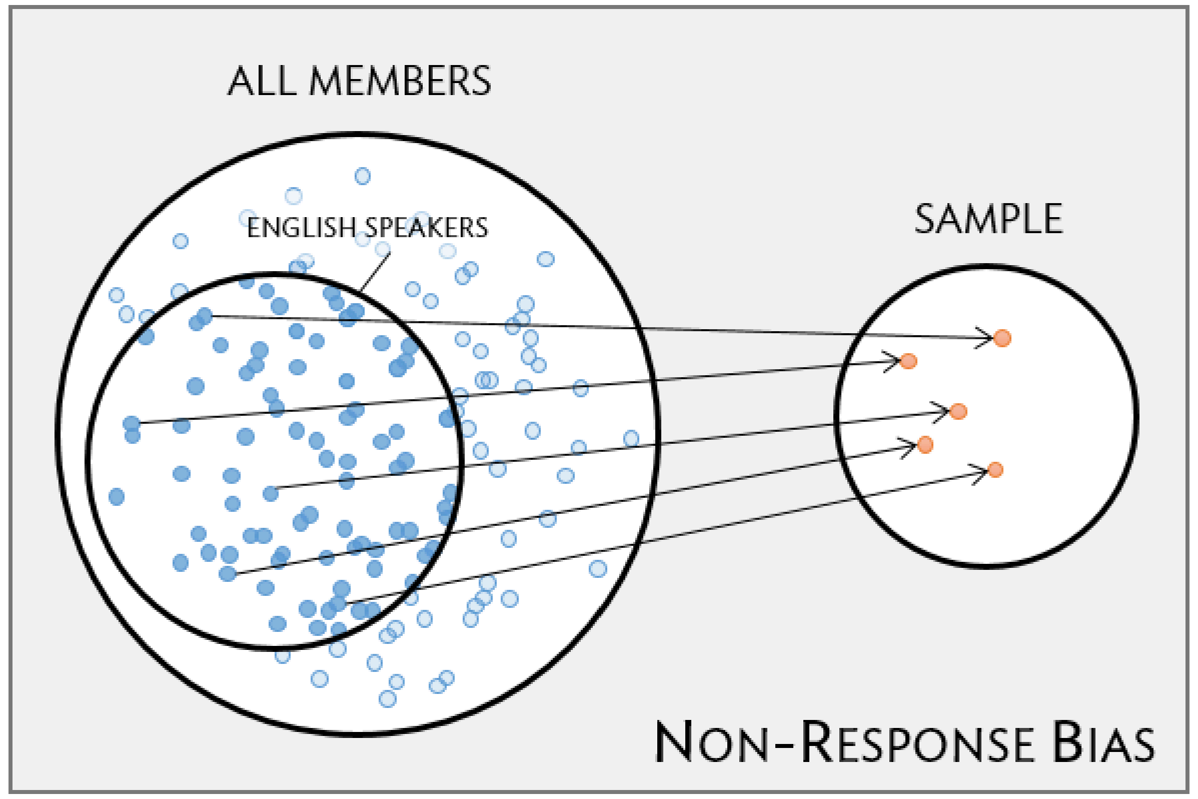


Potential bias with sampling

Good sampling plans don't guarantee samples representative of the population

Non-response bias

- Are all groups within a population being reached?
- Unrepresentative sample can skew results
- Example: survey only administered in English can lead to non-response bias that under-represents individuals who do not fluently speak English
  - Here, bias is stemming from an oversight on the way we are administering our survey (not from the sampling mechanism itself)



Another potential when sampling

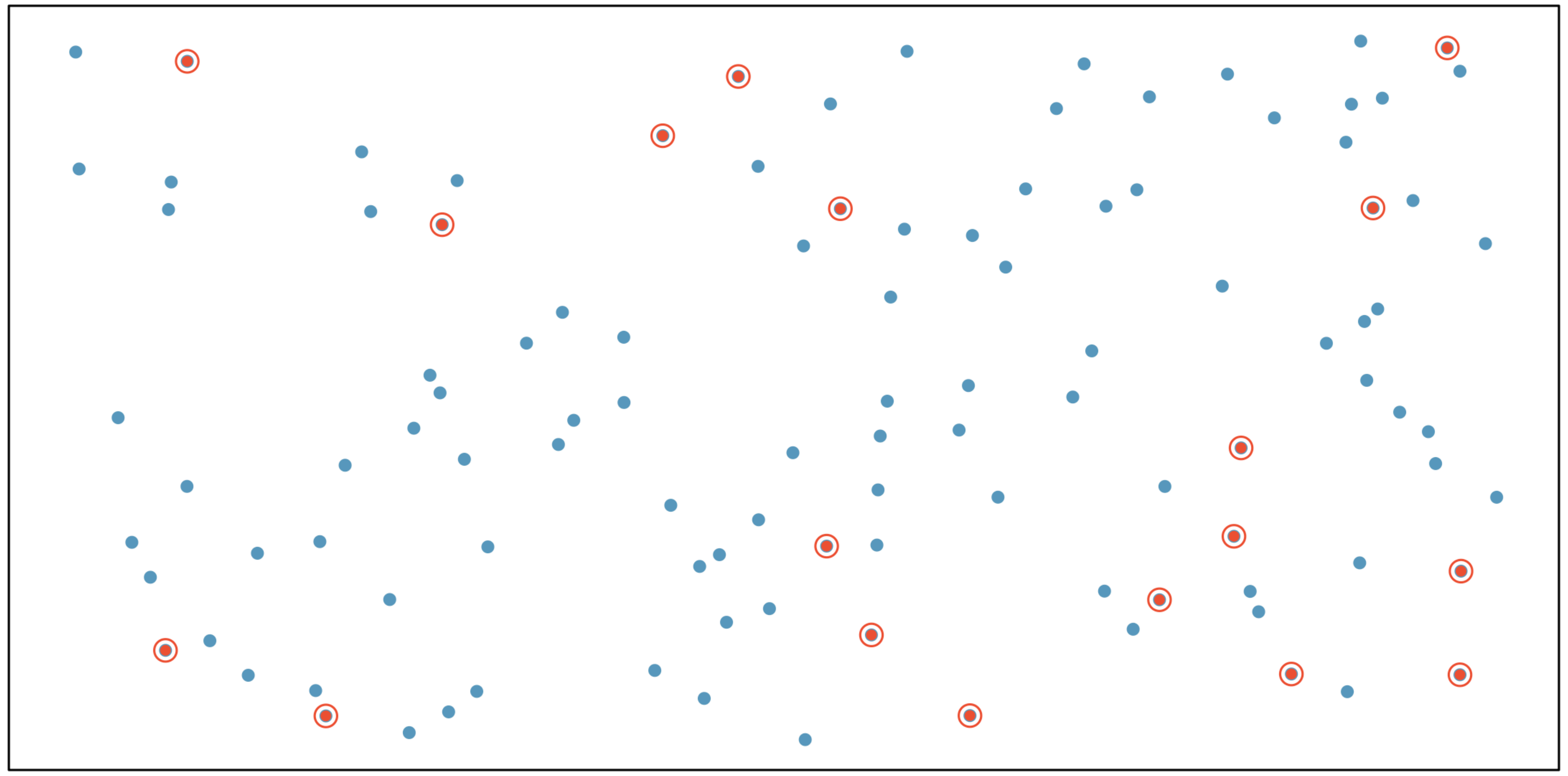
“Random” samples can be unrepresentative by random chance

- In a simple random sample (SRS), each case in the population has an equal chance of being included in the sample
- But by random chance alone a random sample might contain a higher proportion of one group over another
- Example: a SRS might by chance include 70% men (unlikely, but theoretically possible)

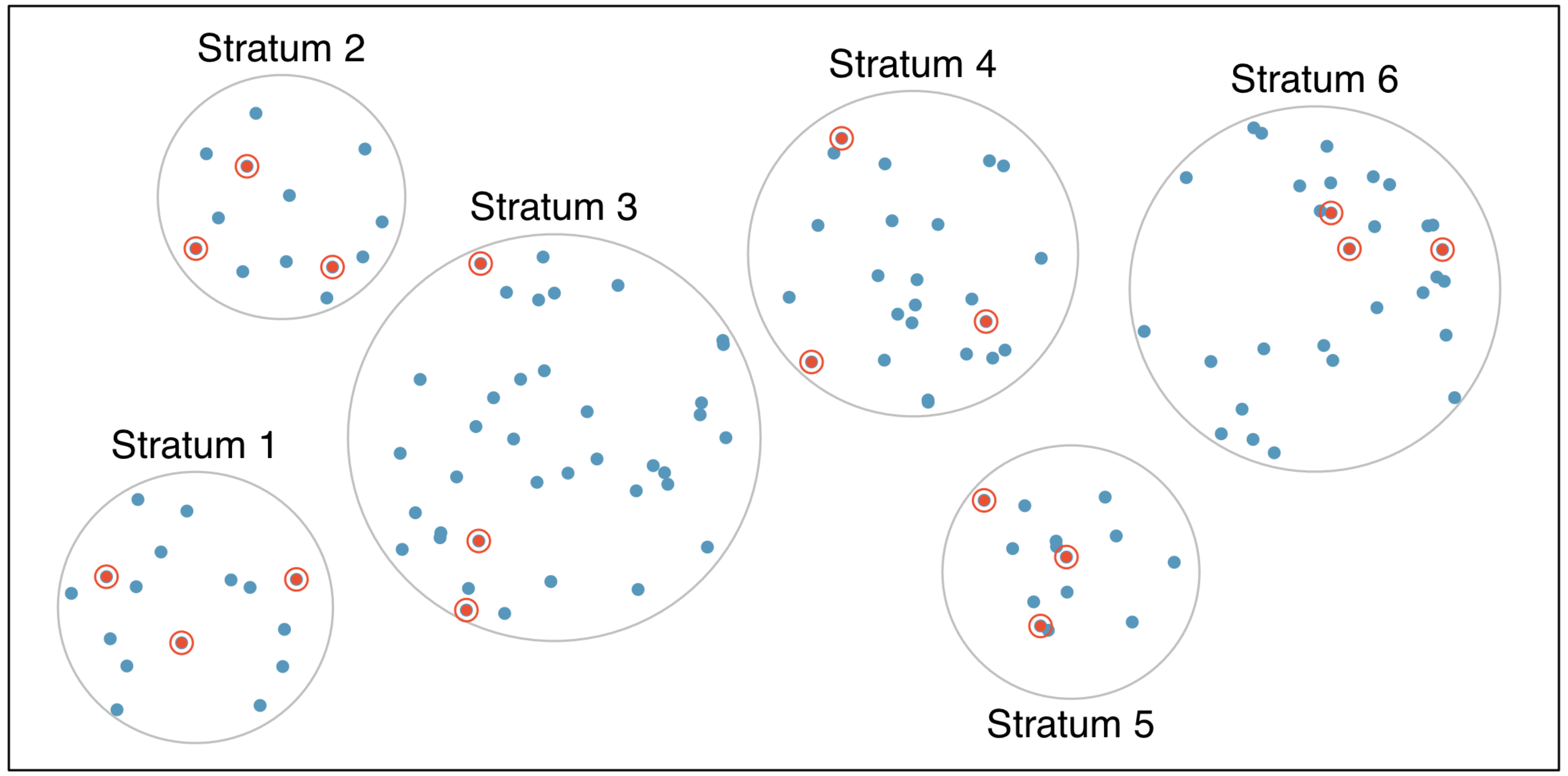
- If it is important for our data to represent specific groups (to answer our research question), then we may consider more complex sampling methods

Sampling methods (3/4)

- **Simple random sample (SRS)**
  - Each individual of a population has the *same chance* of being sampled
  - *Statistical methods taught in this class assume a SRS!*

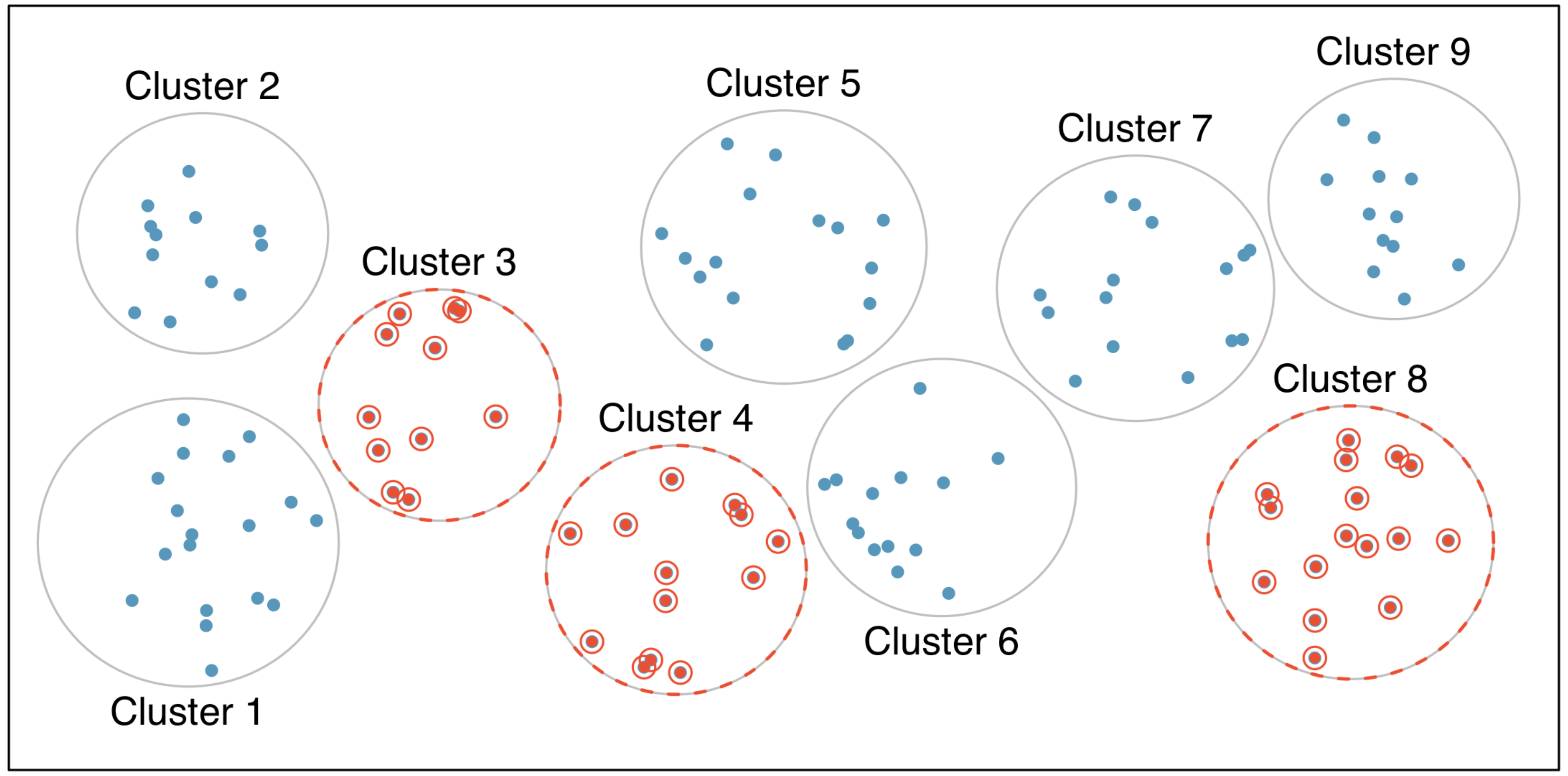


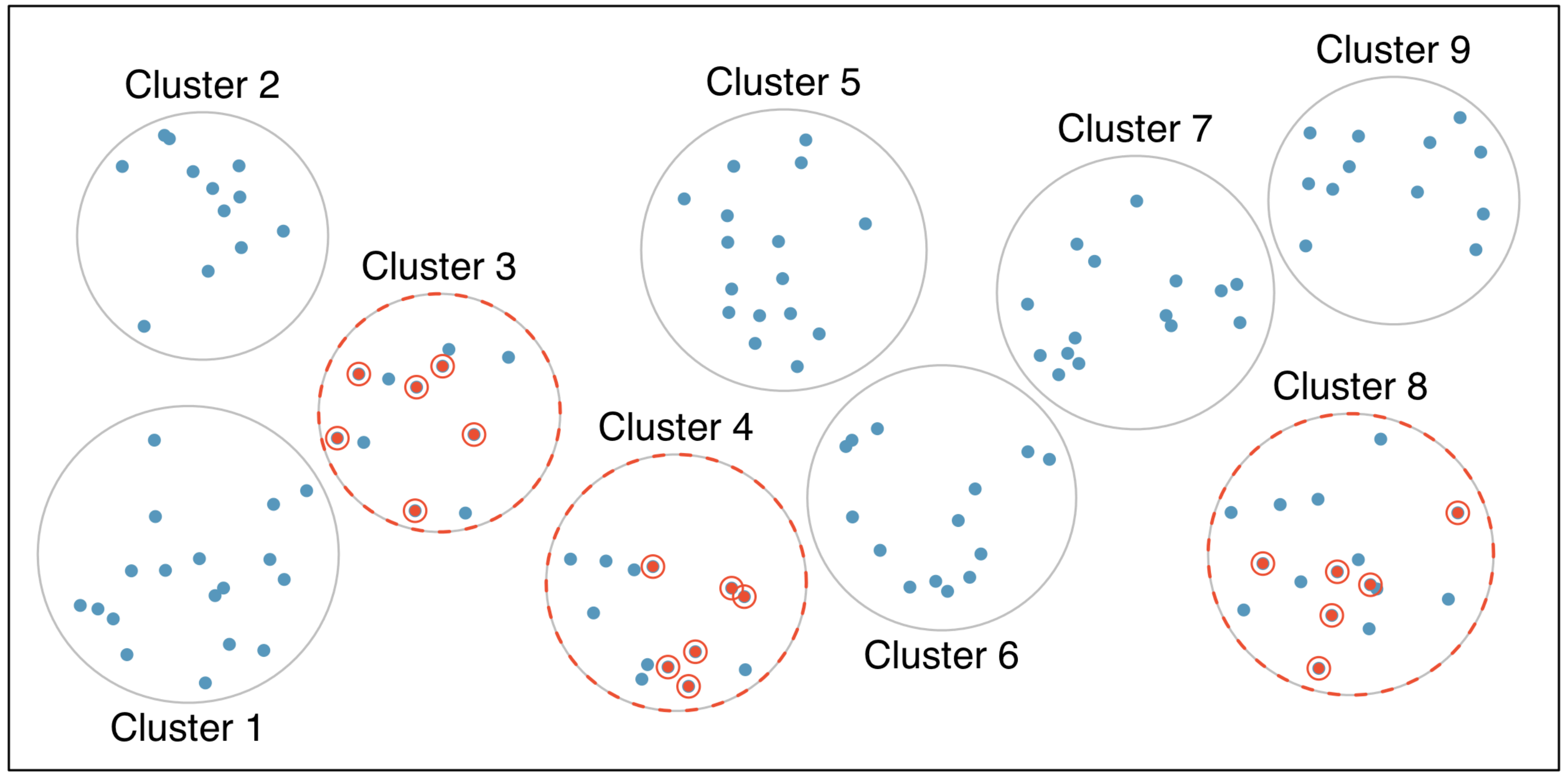
- **Stratified sampling**
  - Divide population into groups (strata) before selecting cases within each stratum (often via SRS)
  - Usually cases within a strata are similar, but are different from other strata with respect to the outcome of interest, such as gender or age groups



Sampling methods (4/4)

- **Cluster sample**
  - First divide population into groups (clusters)
  - Then sample a fixed number of clusters, and include *all* observations from chosen clusters
  - Clusters are often hospitals, clinicians, schools, etc., where each cluster will have similar services/ policies/ etc.
  - Cases within clusters usually very diverse
- **Multistage sample**
  - Similar to a cluster sample, but select a random sample within each selected cluster instead of all individuals





Poll Everywhere Question 3

Learning Objectives

1. Define and compare a target population and its sample.
2. Explain different sampling methods and understand their advantages.
3. Define and compare experiments and observational studies.

Two basic study designs

Experiment

Researchers directly influence how data arise

- Such as: assigning groups of individuals to different treatments and assessing how the outcome varies across treatment groups
- Three major parts to an experiment
  - Control
  - Randomization
  - Replication

Observational study

Researchers merely observe and record data, without interfering with how the data arise

- For example, to investigate why certain diseases develop, researchers might collect data by conducting surveys, reviewing medical records, or following a cohort of many similar individuals.
- Often the only available way to study your research question
  - Due to ethical considerations, funds, or availability of data

Experiments (1/2)

- Researchers assign individuals to different **treatment** or **intervention groups**
  - **Control group**: often receive a **placebo** or usual care
  - Different treatment groups are often called **study arms**
- **Randomization**
  - Group assignment is usually random to ensure similar (balanced) study arms for all variables (observed and unobserved)
  - Randomization allows study arm differences in outcomes to be attributed to treatment rather than variability in patient characteristics
    - Treatment is the only systematic difference between groups
    - Establish causality
  - Different than random sampling! Once we have the sample, then we randomize!!

Experiments (2/2)

- **Replication**
  - Accomplished by collecting a sufficiently large sample
  - Results usually more reliable with a large sample size
    - Often less variability
    - More likely to be representative of population
- Some studies are not ethical to carry out as experiments

Observational studies

- Data are observed and recorded without interference
- Often done via surveys, electronic health records, or medical chart reviews
- Cohorts
- Associations between variables can be established, **but not causality**
  - Individuals with different characteristics may also differ in other ways that influence response
- Confounding variables (lurking variable)
  - Variables associated with both the explanatory and response variables

Observational studies: prospective vs. retrospective studies

Some studies can have prospective and retrospective data!

Prospective

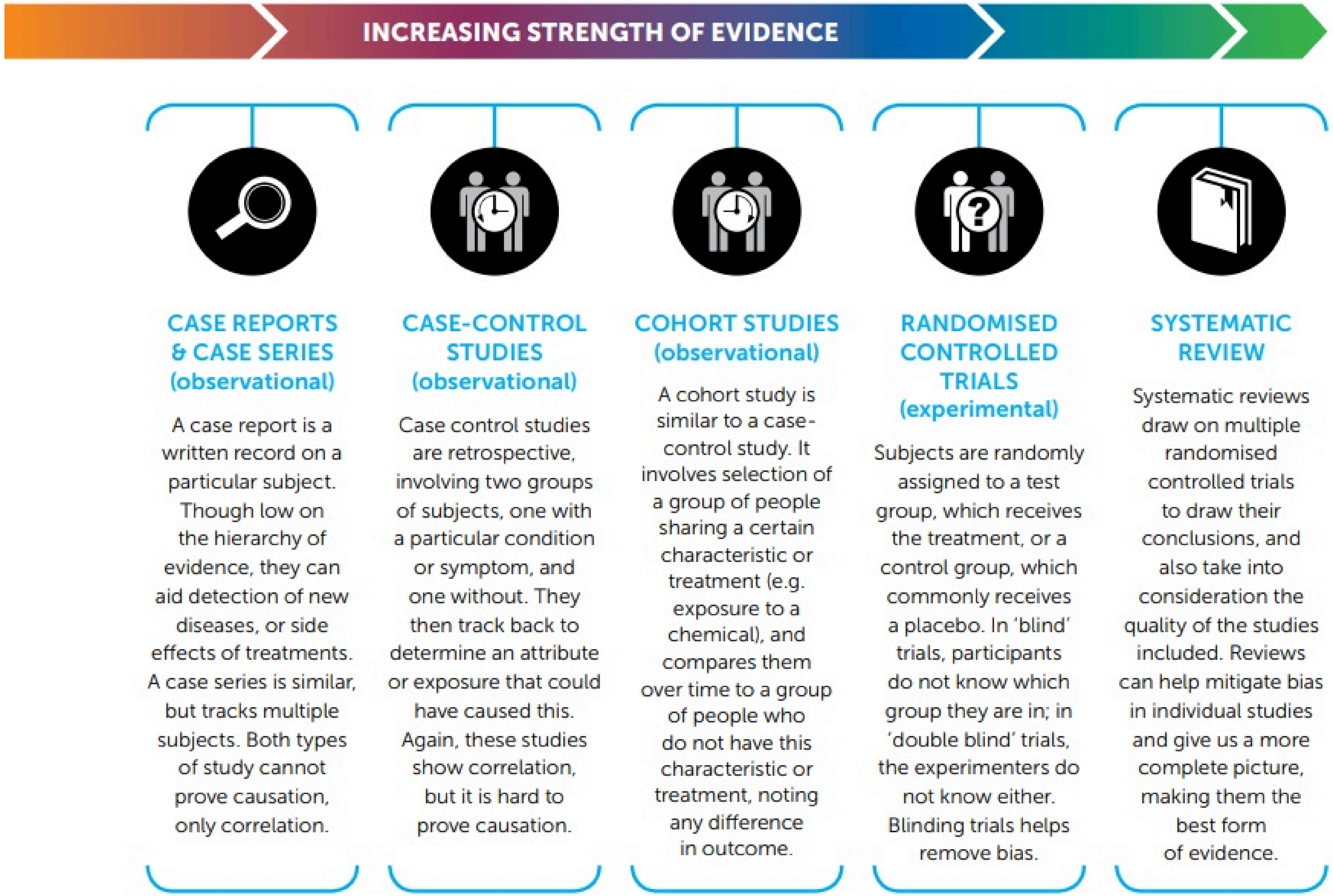
- Identifies participants and collects information **at scheduled times or as events unfold**.

Retrospective

- Collect data **after events have taken place**, such as from medical records

Example: The Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) enrolled participants with lung or colorectal cancer, collected information about diagnosis, treatment, and previous health behavior (retrospective), but also maintained contact with participants to gather data about long-term outcomes (prospective).

Comparing study designs



CASE REPORTS & CASE SERIES (observational)

A case report is a written record on a particular subject. Though low on the hierarchy of evidence, they can aid detection of new diseases, or side effects of treatments. A case series is similar, but tracks multiple subjects. Both types of study cannot prove causation, only correlation.

CASE-CONTROL STUDIES (observational)

Case control studies are retrospective, involving two groups of subjects, one with a particular condition or symptom, and one without. They then track back to determine an attribute or exposure that could have caused this. Again, these studies show correlation, but it is hard to prove causation.

COHORT STUDIES (observational)

A cohort study is similar to a case-control study. It involves selection of a group of people sharing a certain characteristic or treatment (e.g. exposure to a chemical), and compares them over time to a group of people who do not have this characteristic or treatment, noting any difference in outcome.

RANDOMISED CONTROLLED TRIALS (experimental)

Subjects are randomly assigned to a test group, which receives the treatment, or a control group, which commonly receives a placebo. In 'blind' trials, participants do not know which group they are in; in 'double blind' trials, the experimenters do not know either. Blinding trials helps remove bias.

SYSTEMATIC REVIEW

Systematic reviews draw on multiple randomised controlled trials to draw their conclusions, and also take into consideration the quality of the studies included. Reviews can help mitigate bias in individual studies and give us a more complete picture, making them the best form of evidence.

Used with permission © Compound Interest 2015 – www.compoundchem.com

Science Media Centre

Poll Everywhere Question 4