Course intro & motivation

EDS 222

Tamma Carleton
Fall 2021

Today

Why are we here?

What is statistics? Why do we need it as environmental data scientists?

Details

Into the syllabus weeds 🌱

Getting started: Sample vs. population

- What is a sample? What is a population?
- How are samples constructed?
- Study design: how are samples constructed to fit a question?

Why are we we here?

Statistics:

The science of **collecting**, **manipulating**, and **analyzing** empirical data

Statistics enables us to use environmental data to follow the **scientific method**

The scientific method: Example

Step one: Make an observation (cool, don't need stats)

Ex: When Mt. Pinatubo erupted in 1991, much less sunlight was available for plants...

- 20 Mt of sulfur dioxide $\longrightarrow$ increased stratospheric sulfate aerosols, decreased sunlight reaching Earth's surface
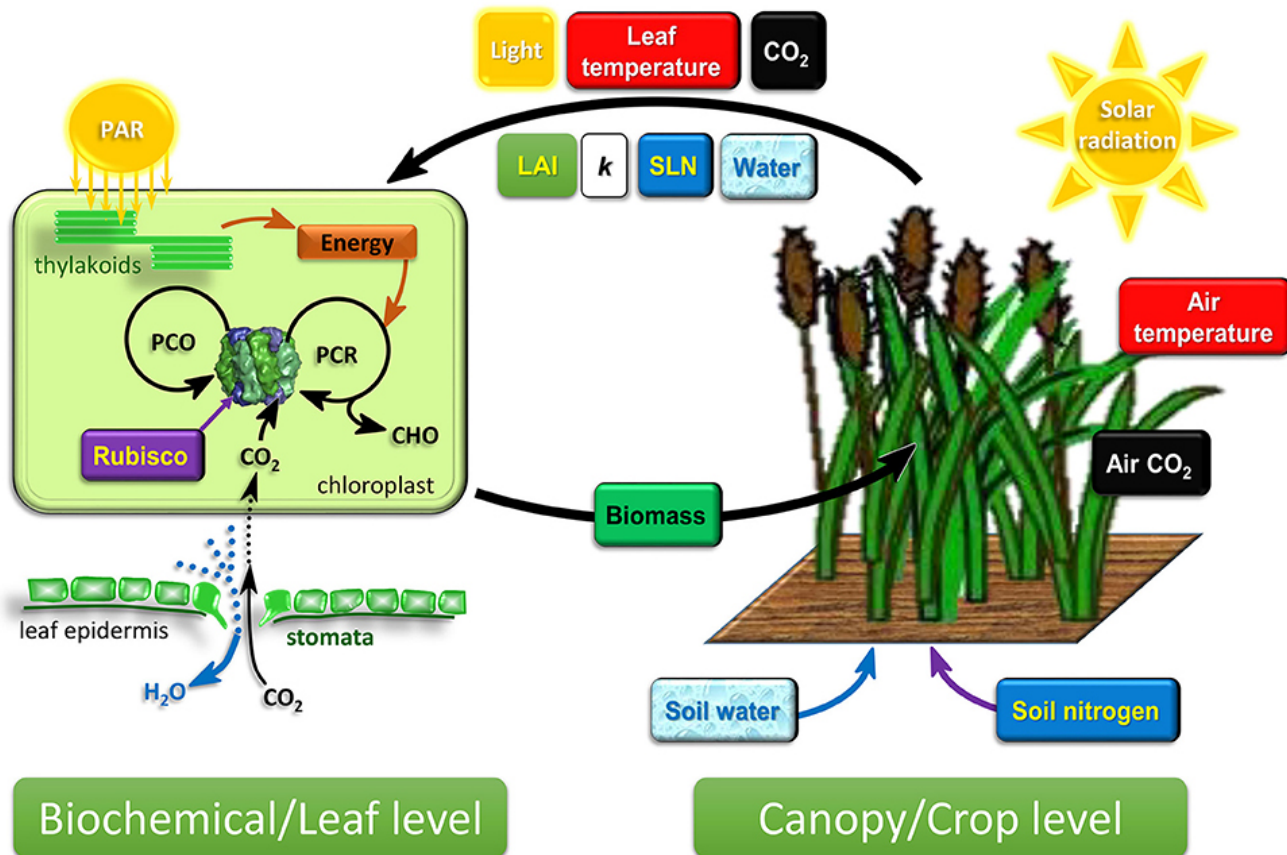

ALBERT GARCIA

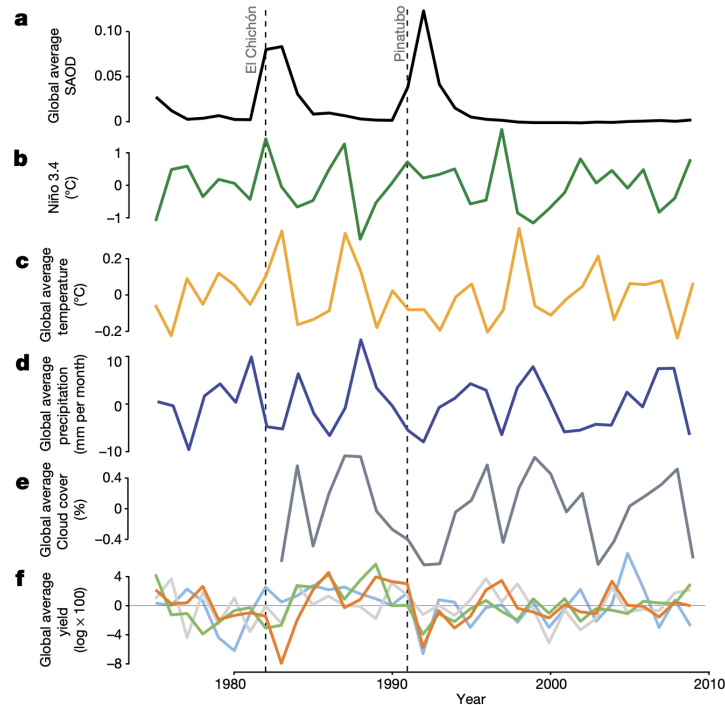**Step two:** Ask a question (easy, also don't need stats)

Ex: I wonder if these aerosols could decrease crop yields?

The scientific method: Example

**Step three:** Form a hypothesis / testable explanation (sounding somewhat stats-like, but still no pre-req's here)

Ex: Did less sunlight from Mt. Pinatubo's aerosols lead to lower crop yields?

The scientific method: Example

**Steps four/five:** Analyze data & test the hypothesis (**‼ NEED STATISTICS ‼**)

Ex: Assemble spatio-temporal dataset, run multivariate linear regression, test statistical significance of regression parameters
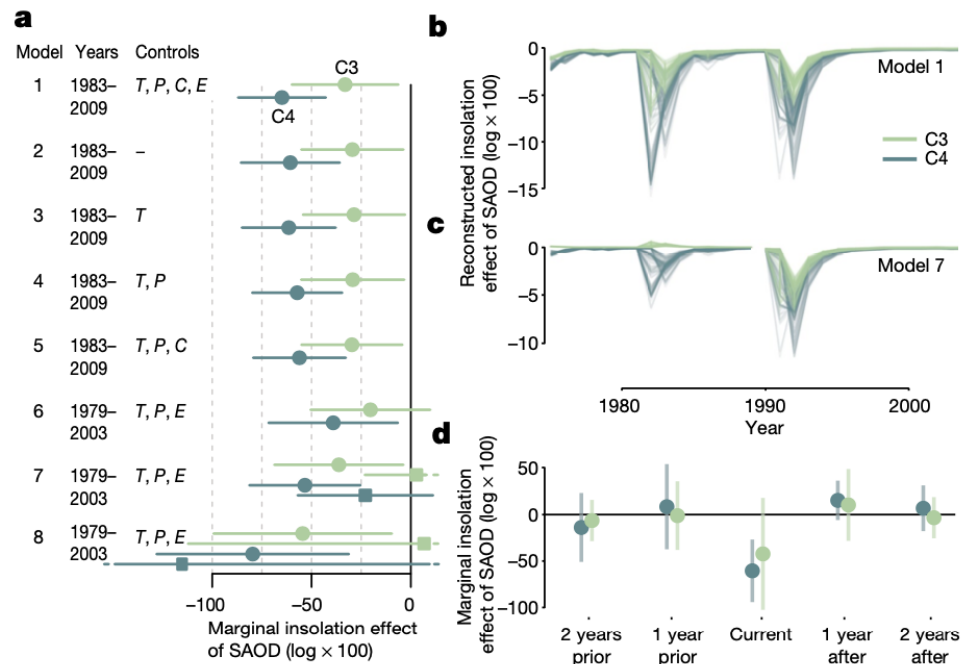
Step six: Report conclusions & what they mean for your question

Ex: Run simulation to consider the implications for solar radiation management (a form of geo-engineering)
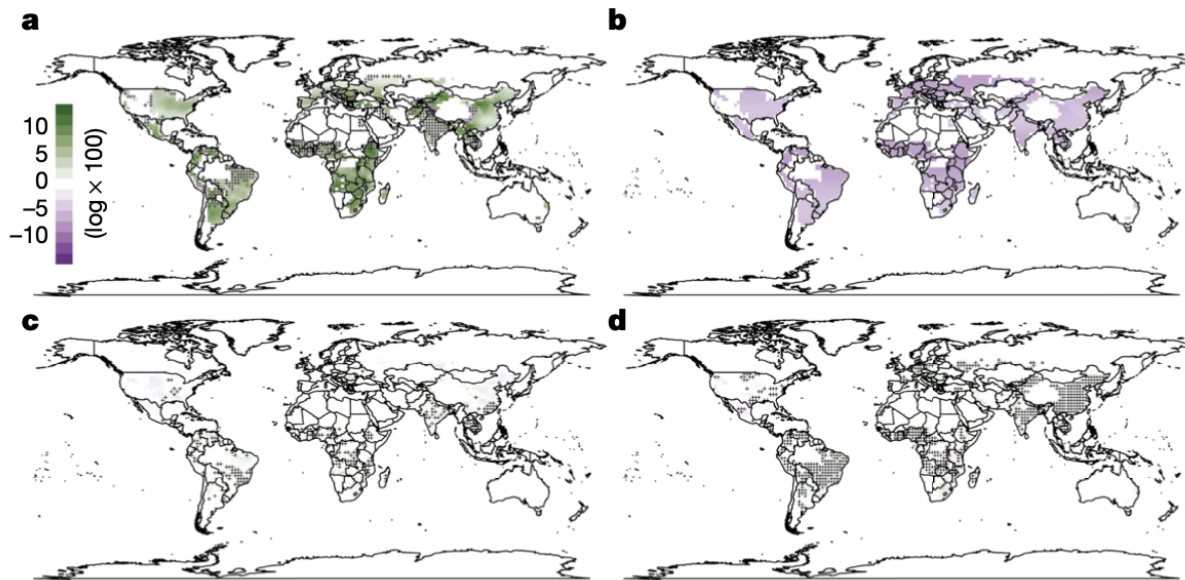
Statistics and the scientific method

- Statistics will enable us to **test hypotheses**, **analyze data**, and **draw conclusions** about the world from the process

- Otherwise, we'd be stuck at **observing** and **forming hypotheses**

- ... and we'd have a lot of unanswered empirical questions!
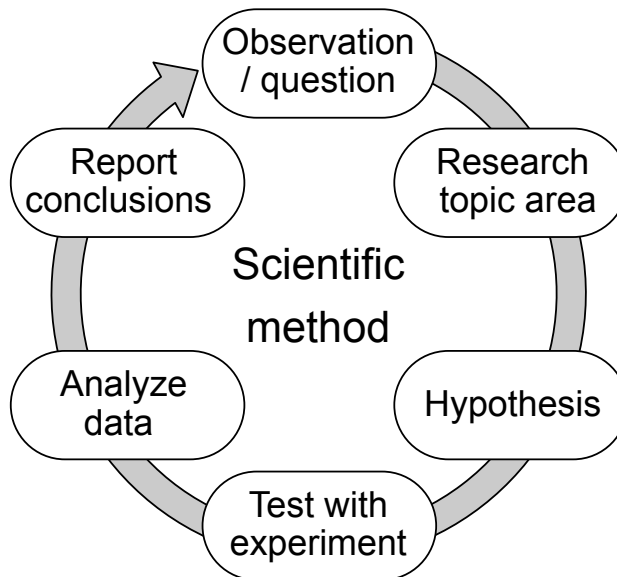
This course within MEDS

This course **is**:

- Designed to build your fundamental statistical toolkit
- Designed to teach you to *apply* statistics in `R`
- Designed to show you key spatio-temporal methods that come up frequently in environmental data science
- **Brand new!** ⟶ Things will not go according to plan. But you can help shape the curriculum!

This course is **not**:

- A programming class ⟶ I will make mistakes. I don't know all the new exciting things. I will learn from you!
- A class in program evaluation/causal inference (👀 EDS 241)
- A class in spatial data analysis (👀 EDS 223)

Syllabus

Syllabus can always be found on our course website.

Sample versus population

Consider a potential research question:

> What is the average mercury content in swordfish in the Atlantic
> Ocean?

Some definitions

- **Population**: The entire target population of interest.
    - Ex: All swordfish in the Atlantic Ocean

- **Census**: A data collection including *all* individuals in the population
    - Ex: Count every single swordfish in the Atlantic Ocean (hard and 💰)

Sample versus population

Consider a potential research question:

> What is the average mercury content in swordfish in the Atlantic Ocean?

Some definitions

- **Sample**: A subset of the target population for which we actually have data
  - Ex: 60 tagged swordfish from a government survey in the Atlantic Ocean

Parameters and statistics

Usually we are interested in a numerical summary of the *population* (e.g., mean, slope, intercept, variance)

- **Parameter**: A numerical summary of the **population**

    - Ex: *average* mercury content in swordfish in the Atlantic Ocean

- **Statistic**: A numerical summary of the **sample**

    - Ex: *average* mercury content of the 60 swordfish collected in a government survey in the Atlantic Ocean

Parameters and statistics

We use *statistics* (from a sample) in hopes of learning about *parameters* (from the population)

- This means that every time you do statistics'', you should be thinking...
  - What is the population of interest?
  - What is my sample?
  - How are they different?

All samples are not created equal

From *IMS*: Suppose we want to estimate time to graduation for Duke undergraduates in the last five years using a sample of recent students.
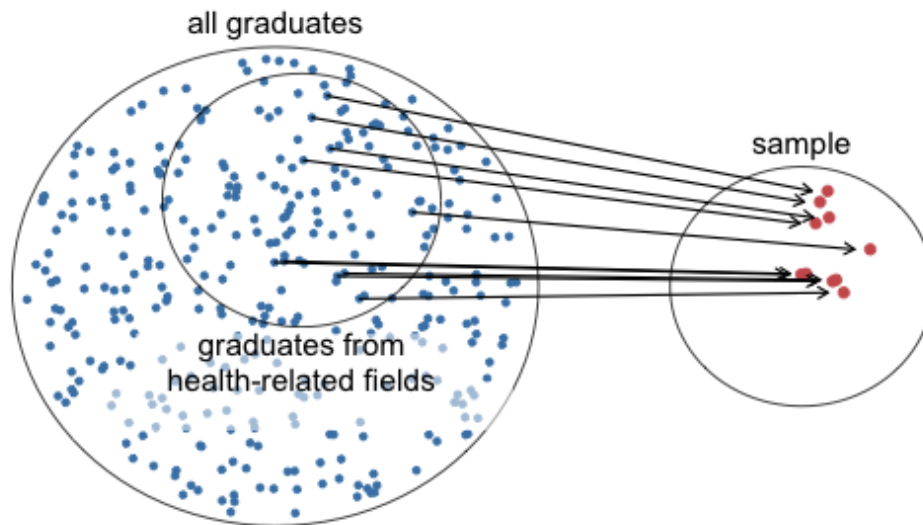
- Q: Who is the population?

Suppose we take a **random** sample (i.e., every individual in the population has the same probability of being selected)
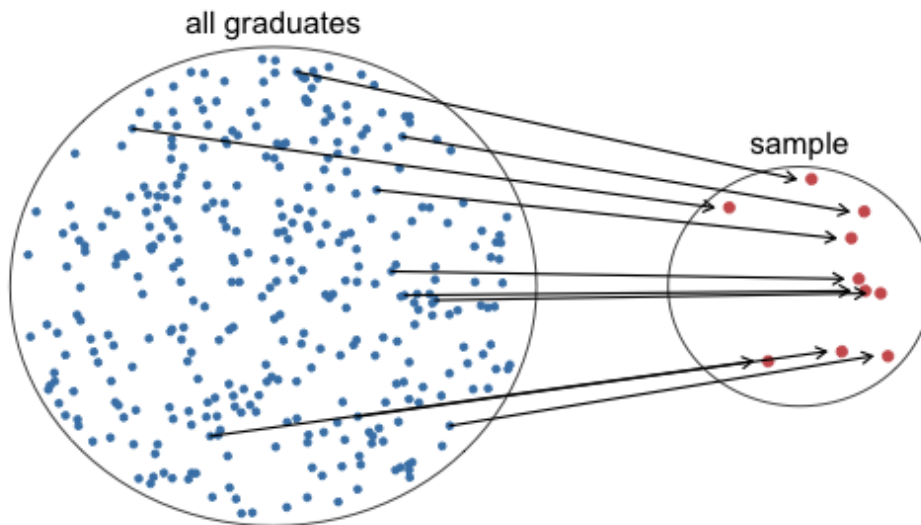
All samples are not created equal

Suppose we ask a nutrition major to pick a few of her friends for the sample.

- What might go wrong here?

Asked to pick a sample of graduates, a nutrition major might inadvertently pick a disproportionate number of graduates from health-related majors.

All samples are not created equal

When a sample is *not* drawn randomly, it is likely your statistic will be a biased estimate of the population parameter

Some other examples of biased sampling:

- **Systematic non-response** (e.g., only people from a certain group respond to the phone survey)

- **Convenience sampling** (e.g., biologists only take forest transects near the edge of a large forested area)

All samples are not created equal

Under-represented groups may be particularly misrepresented due to improper sampling

- Rolf et al., "Representation Matters"

    - Show the value of diverse samples for training machine learning algorithms

- Buolamwini et al., "Gender shades"

    - Facial analysis benchmark datasets (i.e., samples) are overwhelmingly white, leading to misclassification for darker-skinned subjects

Four (random) sampling strategies

Nearly all statistical methods are based on assumptions of randomness. If data are not collected randomly from the population, estimates are likely to be biased.

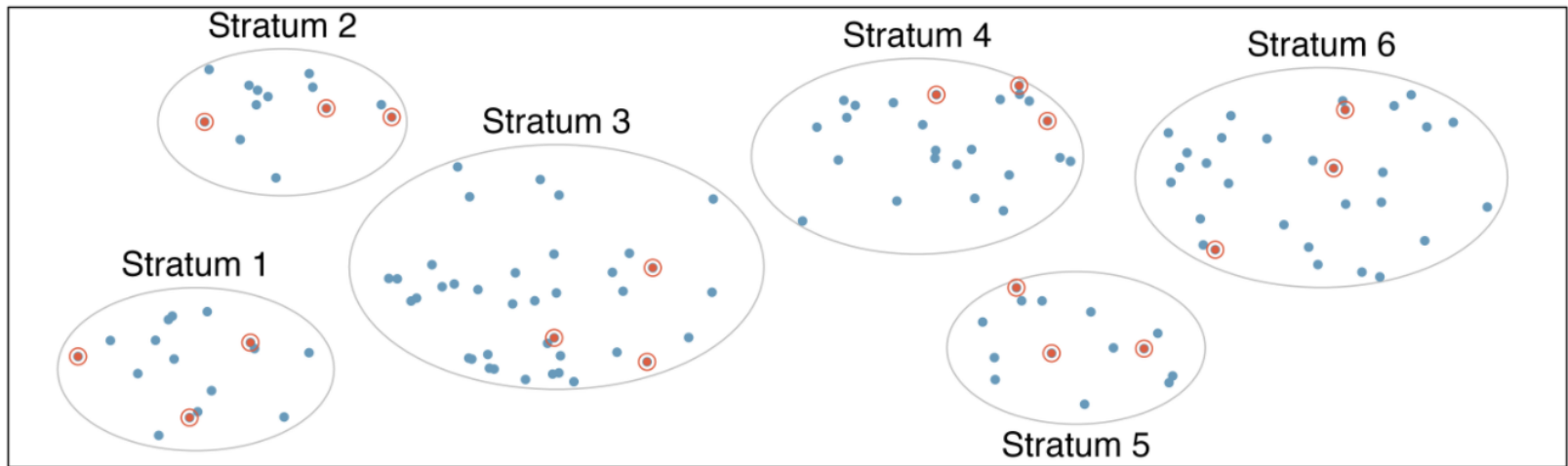## Strategy 1: Simple random sampling

- As simple as it sounds!
- To consider for later classes: What problems might arise if your (simple random) sample is small?
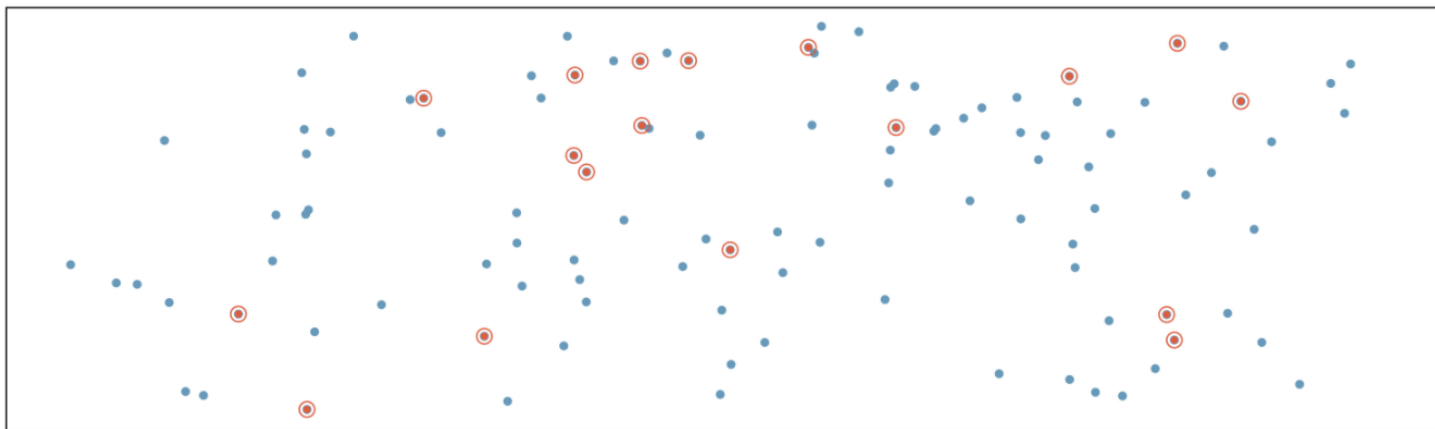
Four (random) sampling strategies

Strategy 2: Stratified sampling

- More complex to analyze sample to construct estimates of population parameters (but still possible)
- Helpful when individuals within a strata are quite similar
- Used often as a method of variance reduction (we'll discuss this later)

Source: IMS

Four (random) sampling strategies

Strategy 3: Cluster random sampling

- Helpful when individuals within a cluster are quite *different* from one another
- Used often when costs of data collection are high per cluster (e.g., Demographic and Health Surveys)
- Also more complex to estimate population parameters
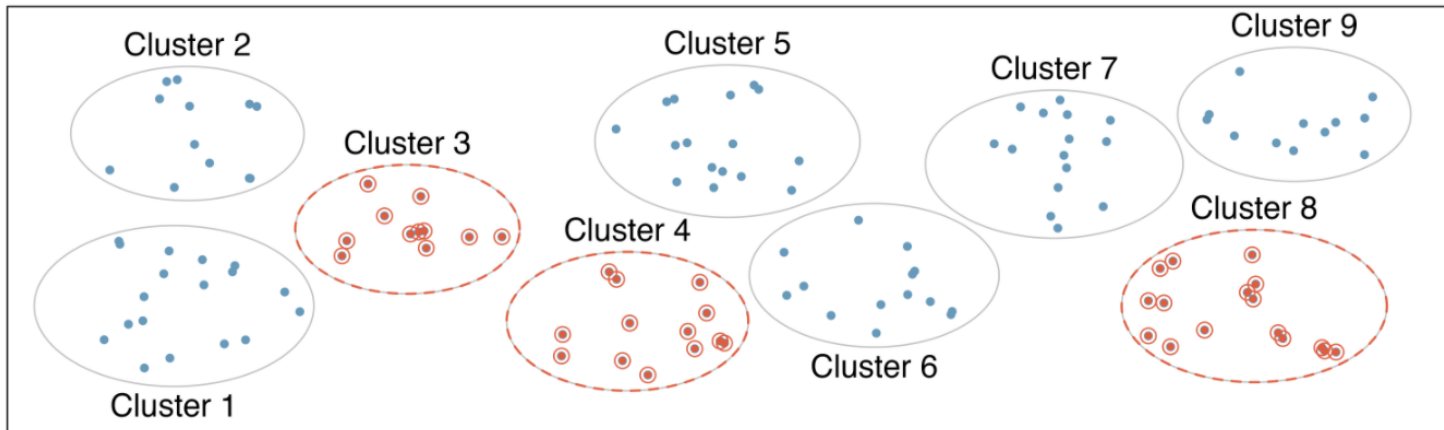
Source: IMS

Four (random) sampling strategies

Strategy 4: Multi-stage sampling

- Very similar to cluster sampling, just take fewer samples (randomly)
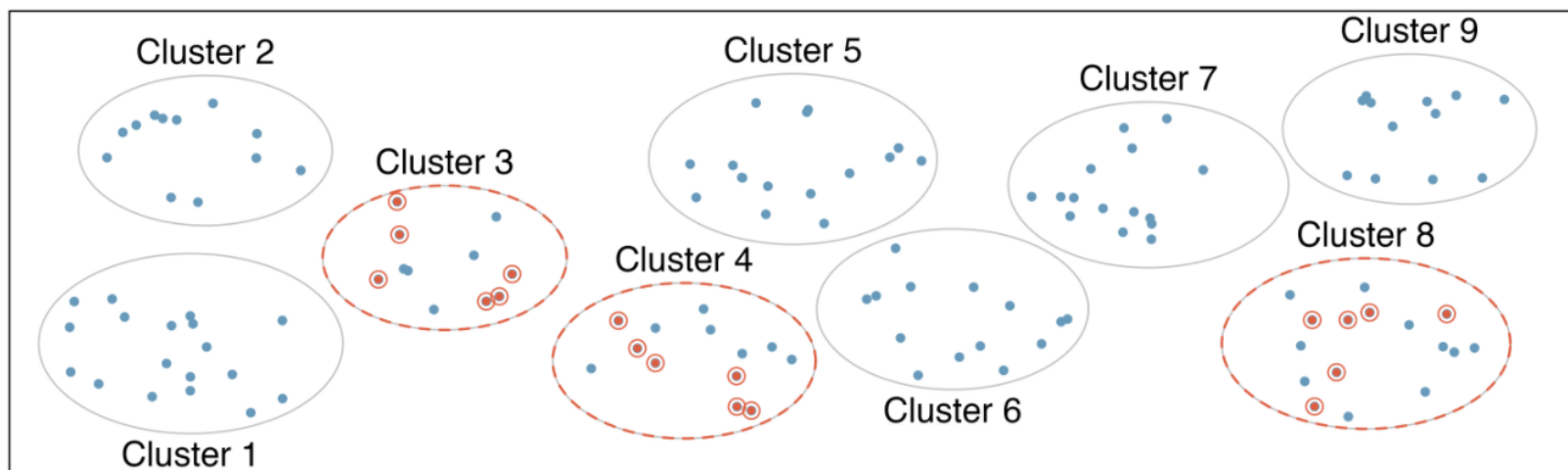
Source: IMS

Study design

When we conduct statistical analyses, where do our samples come from?

- **Experimental** studies

  - Sample is collected to fit the study's needs

- **Observational** studies

  - Sample exists, design your study to make best use of available data

In both cases, the researcher generally aims to *causally* identify a population parameter

- Ex: What is the effect of increased atmospheric aerosols on crop yields? (Proctor et al., 2018)
- Ex: What is the effect of a vaccine booster shot on the risk of COVID-19 hospitalization?
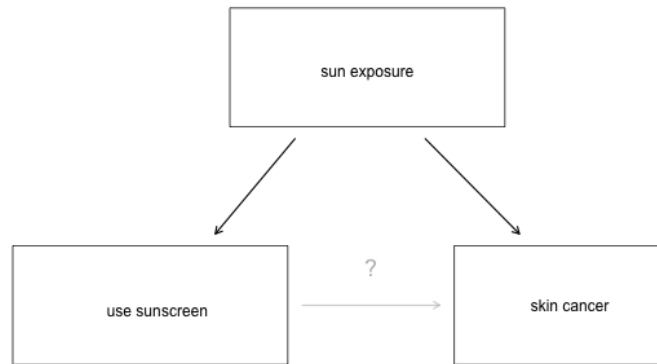- **You will talk a lot more about this in EDS 241!**

[Very] short foray into study design
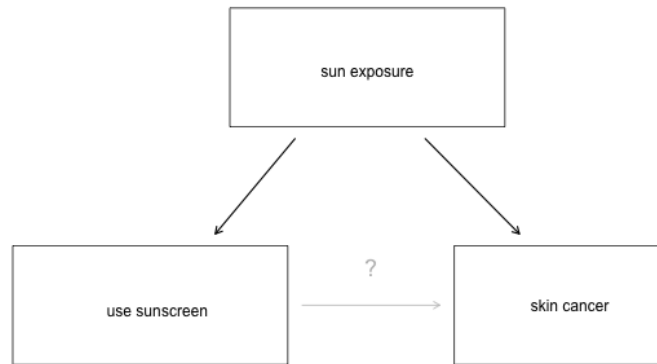
> Q: Does sunscreen lower risk of skin cancer?

- Experiment
  - Randomly sample 50% of individuals and assign them the sunscreen "treatment", require the other 50% to wear no sunscreen
  - Follow individuals for 20 years, compare cancer outcomes

[Very] short foray into study design

> Q: Does sunscreen lower risk of skin cancer?

- Observational study
  - Collect data on sunscreen use, skin cancer, and sun exposure
  - Compare cancer rates for individuals with different sunscreen use habits

Homework 1 (warmup): Logistics

Where are the data?

- All the data for Thursday labs and homeworks will be on Taylor
- See EDS 214 for a refresher on how to access Taylor, compute on Taylor, and pull data to/from Taylor

Homework 1 (warmup): Logistics

Where should I do computation?

- All the data we use will be small enough to load and work with *locally* (e.g., use Cyberduck to pull data down)
- But you're welcome to work on Taylor with an RStudio GUI or Workbench, etc.
- You won't have write access to our class directory, but you do have your own directories on Taylor

Homework 1 (warmup): Logistics

Github classrooms

- I'm learning, you're learning, might be a little bumpy in the beginning
- Please communicate and give me feedback!

Slides created via the R package **xaringan**.