HOCHSCHULE
DER MEDIEN

Study design

Prof. Dr. Jan Kirenz
HdM Stuttgart

How was data collected?

Is our data reliable and helps to achieve our goals?

Sampling principles and strategies

The difference between populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last five years, what is the average time to complete a degree for Duke undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Populations and samples

- Each research question refers to a **target population**
- Collecting data for an entire population is called a **census**.
- A **sample** is the data you have. Ideally, a sample is a small fraction of the population.

Parameters and statistics

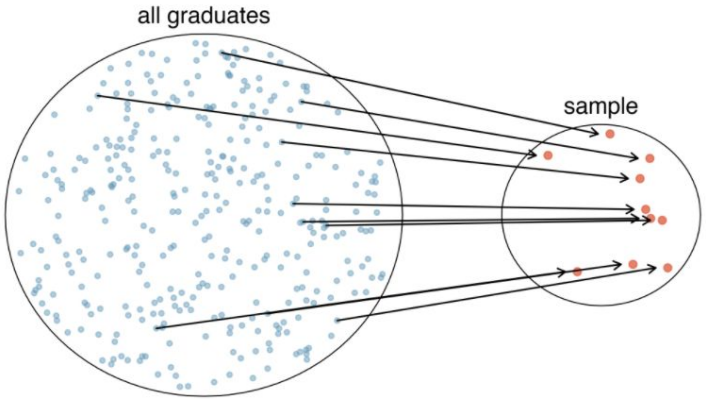
- **Statistic**: when a number is being calculated on a sample of data
- **Parameter**: number is calculated or considered for calculation on the entire population

Anecdotal evidence

Consider the following possible responses to the three research questions:

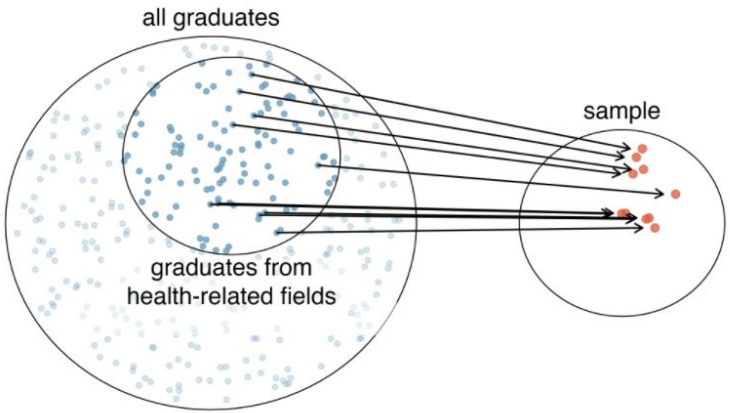
1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from HdM, so it must take longer to graduate at HdM than at many other universities.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Sampling from a population



10 graduates are randomly selected from the population to be included in the sample.

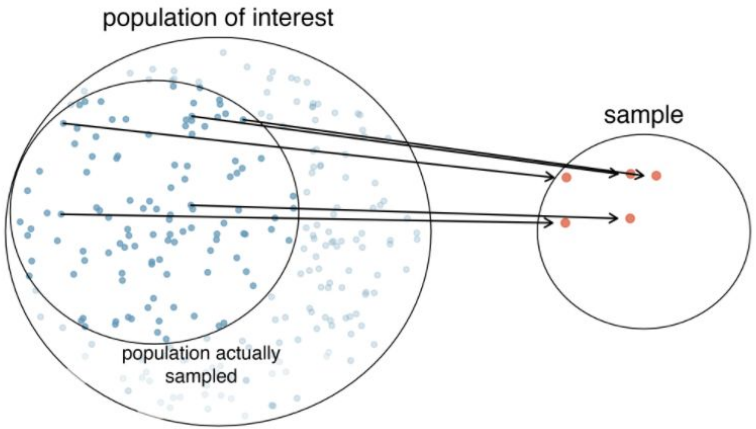
Biased sample



Biased sample:

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

Random sample



Random sample:

Each case in the population has an equal chance of being included and the cases in the sample are not related to each other.

Possible problem: **non-response rate** can skew results (non-response bias) and may lead to a **convenience sample**

Guided practice

- We can easily access ratings for products, sellers, and companies through websites.
- These ratings are based only on those people who go out of their way to provide a rating.
- If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product? Why or why not

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

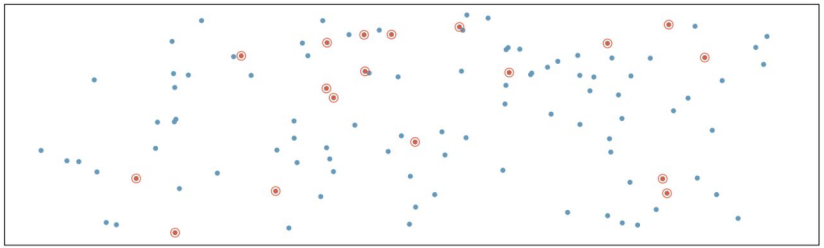
III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

Sampling methods

Simple random and stratified sampling



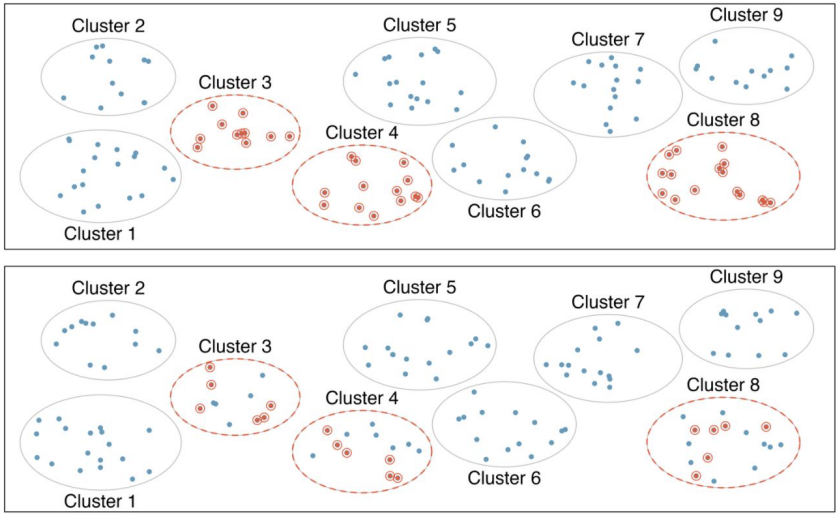
Simple random sampling:

each case in the population has an equal chance of being included in the final sample and knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling

The population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum.

Cluster and multistage sampling



Cluster sampling:

In a cluster sample, we break up the population into many groups, called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample

Multistage sampling

A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we would collect a random sample within each selected cluster.

Example

- Suppose we are interested in estimating the **employee satisfaction rate** in our factories in China.
- We learn that there are **30 production sites** in China, each more or less similar to the next, but the distances between the sites are substantial.
- Our goal is to interview **150 employees**.
- What sampling method should be employed?

Practice

- Our sales department has requested a consumer household survey be conducted in a suburban area of a city.
- The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments.
- Which approach would likely be the *least* effective?

(a) Simple random sampling

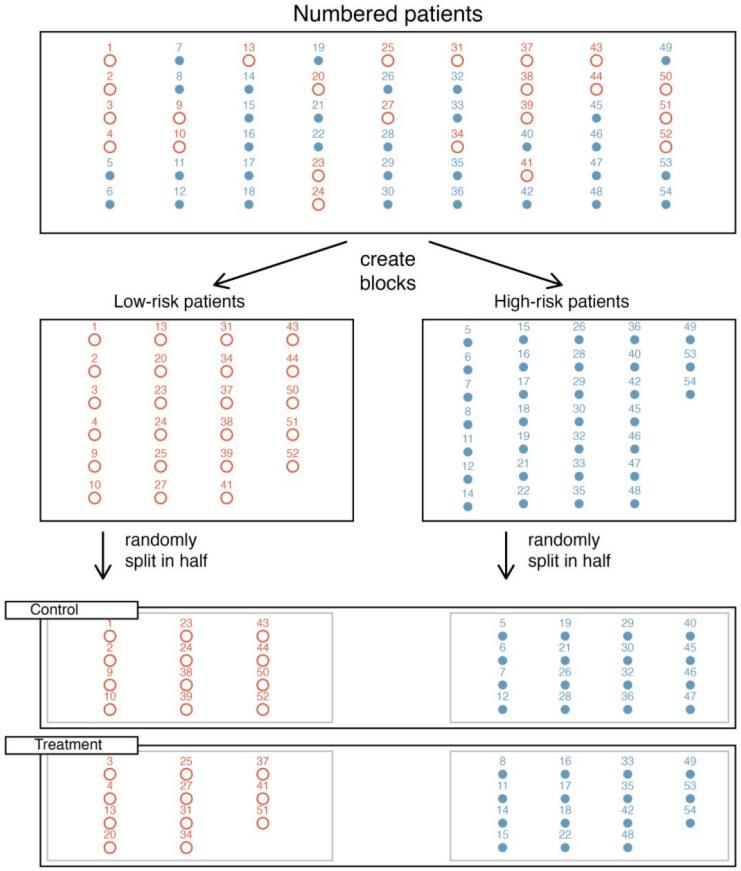
(b) Cluster sampling

(c) Stratified sampling

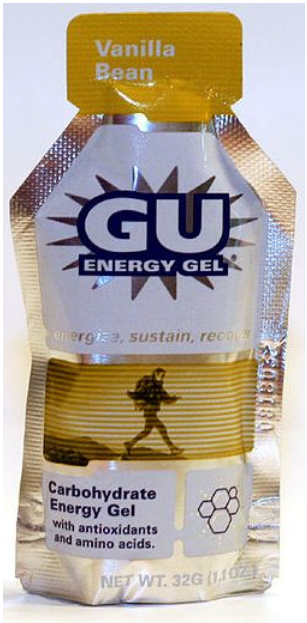
Experiments

Principles of experimental design

1. **Control**: Compare treatment of interest to a control group.
2. **Randomize**: Randomly assign subjects to treatments, and randomly sample from the population whenever possible (helps prevent confounding).
3. **Replicate**: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. **Blocking**: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.


Numbered patients
create blocks
Low-risk patients
High-risk patients
randomly split in half
randomly split in half
Control
Treatment

More on Blocking



- We would like to design an experiment to investigate if energy gels makes you run faster:
  - **Treatment**: energy gel
  - **Control**: no energy gel

- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Why is this important? Can you think of other variables to block for?

Practice

- A study is designed to test the effect of light level and noise level on exam performance of students.
- The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group.
- Which of the below is correct?

A. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

B. There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

C. There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

D. There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Difference Between Blocking and Explanatory Variables

- **Factors** are conditions we can impose on the experimental units.
- **Blocking** variables are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

More Experimental Design Terminology...

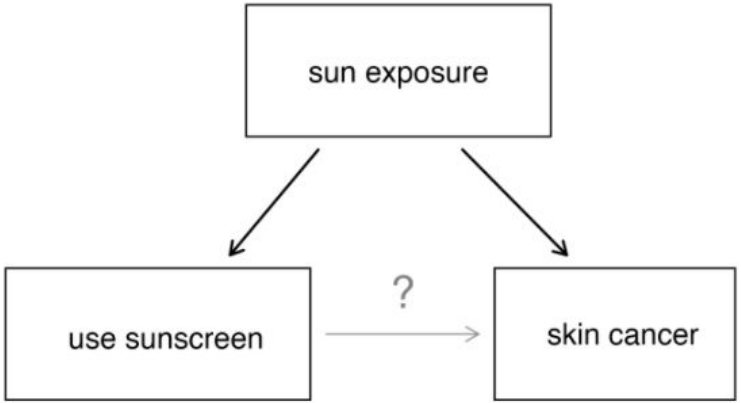
- **Placebo**: fake treatment, often used as the control group for medical studies
- **Placebo effect**: experimental units showing improvement simply because they believe they are receiving a special treatment
- **Blinding**: when experimental units do not know whether they are in the control or treatment group
- **Double-blind:** when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Observational study

Data where no treatment has been explicitly applied (or explicitly withheld) is called observational data.

Guided practice

- Suppose an observational study tracked sunscreen use and skin cancer
- and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer.
- Does this mean sunscreen causes skin cancer?


sun exposure
use sunscreen
?
skin cancer

Observational studies come in two forms

A **prospective** study identifies individuals and collects information as events unfold.

Retrospective studies collect data after events have taken place.

Practice

What is the main difference between observational studies and experiments?

A. Experiments take place in a lab while observational studies do not need to.
B. In an observational study we only look at what happened in the past.
C. Most experiments use random assignment while observational studies do not.
D. Observational studies are completely useless since no causal inference can be made based on their findings.

Summary

		Assignment of Explanatory Variable			
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)		
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	→	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	→	Conclusions might not be generalizable because of volunteer bias.
		↓	↓		
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.		

Terms you need to know

anecdotal evidence
bias
blind
blocking
census
cluster
cluster sampling
confounding variable
control
control group
convenience sample
double-blind
experiment
multistage sample
non-response bias
non-response rate
observational data
parameter
placebo simple
placebo effect
population
prospective study
pseudoreplication
randomized experiment
replication
replication crisis
representative
retrospective study
sample
sample bias
random sample
simple random sampling
statistic
strata
stratified sampling
treatment group

Resources

The slides are based on the excellent book “Introduction to Modern Statistics” by Mine Çetinkaya-Rundel and Johanna Hardin.

The online version of the book can be **accessed for free**:

https://openintro-ims.netlify.app/data-design.html