Day 4 - Study design

Introduction

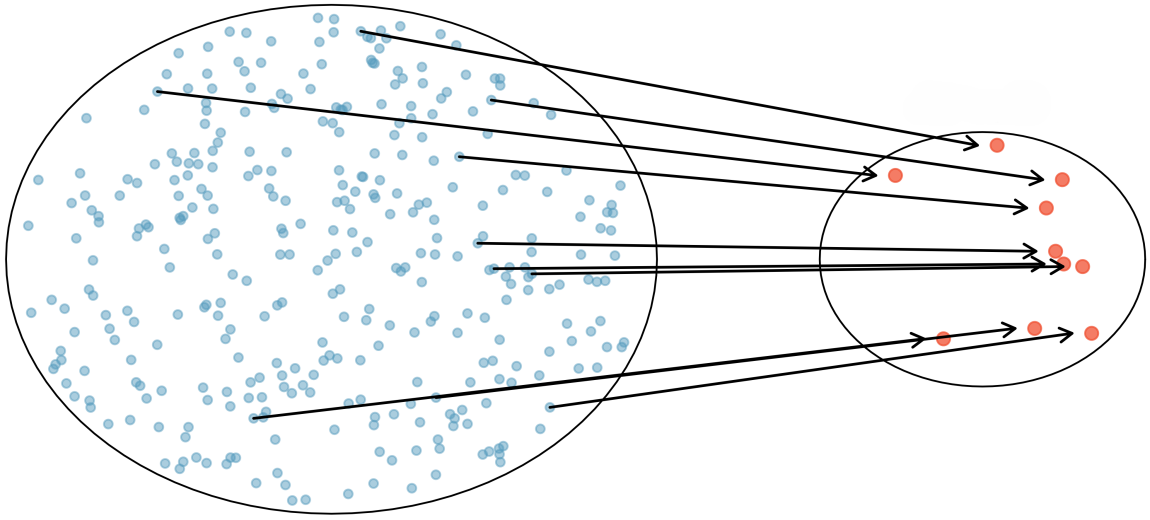
In this section, we consider how data come to be, focusing on how the way data are collected influence the types of inference we can make. At the conclusion of this activity, you should understand the following terms:

- population
- sample
- statistic
- parameter
- representative
- random selection
- sampling bias
- nonresponse bias
- response bias
- convenience sample
- randomized experiment
- confounding variables

Populations and samples

Note

A **population** is the entire collection of individuals about whom we wish to make inferences. Collecting data for an entire population is called a **census**. Often times, we are unable to collect a census. So instead, we select a **sample** from the population and use that sample to draw inferences about the population.



Note

_____ are summaries of characteristics about the population.

_____ are summaries of characteristics about the sample.

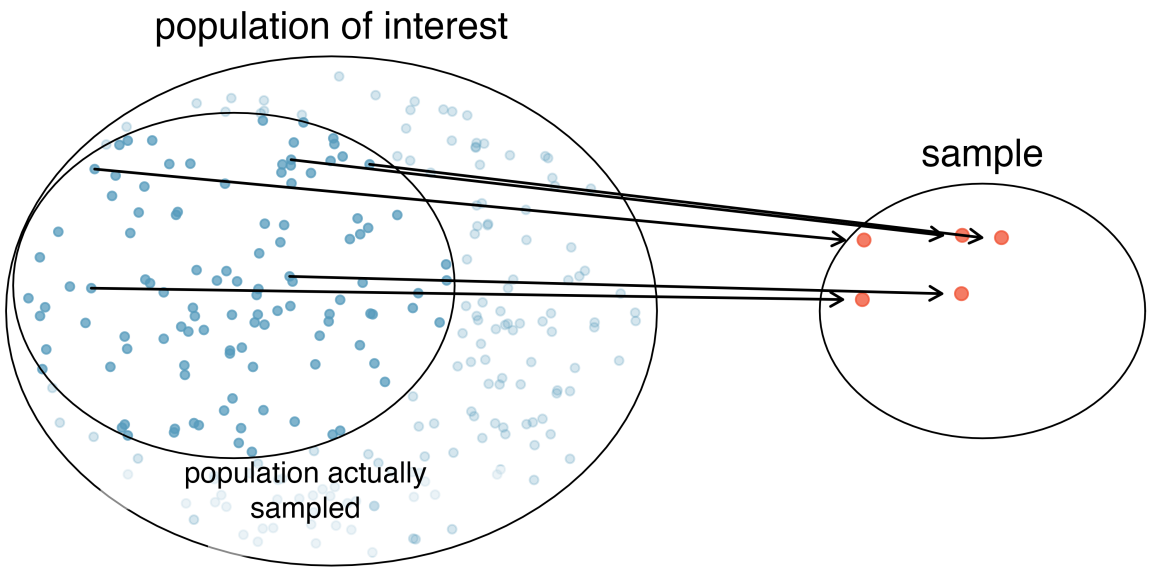
In order to use a sample to make inferences about a population, what must be true?

Important

To obtain representative samples, we often rely on **randomly selecting** individuals from the population of interest.

Sampling bias

> **Note**
>
> **Sampling bias** occurs when the method of collecting a sample results in a sample that is not representative of the population, on average.



_____ occurs when a subset of the individuals randomly selected to be in the sample do not respond.

- Ex. You distribute a poll to Middlebury students inquiring about whether they cheat on exams and only half respond to the survey.

_____ occurs when there is anything in the survey design that influences the responses (leading questions, questions about sensitive issues, etc.)

- Ex. Of the students who responded to the above poll, some may lie about cheating (depending on whether the survey is anonymous, etc.).

A sample in which individuals are selected based on their ease of inclusion is called a

_____.

- Ex. Rather than randomly selecting students from Middlebury, you distribute the poll to your STAT 116 class.

Causation revisited

! Review

Randomized experiments allow researchers to determine causal relationships by distributing potential **confounding variables** between the **treatment** groups equally, on average.

A meta analysis of 21 studies on the use of sunscreen and risk of melanoma reported that eight of the 21 studies found increased risk of melanoma with use of sunscreen.[1] For each of these studies, no individuals were randomly assigned to use sunscreen. Should we conclude that use of sunscreen may cause skin cancer?

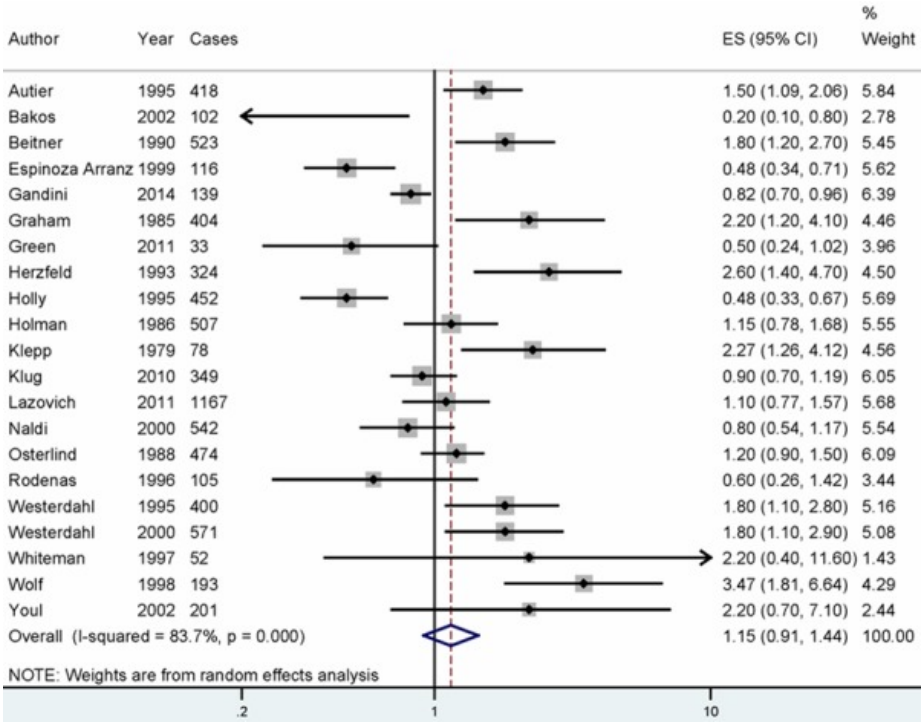


Figure 2 from Xie, F. et al. (2015).

[1] Moderna clinical trial

Scope of inference

! Review

The **scope of inference** describes (1) to whom you can generalize your results (if any) and (2) whether the inference describes a causal relationship.

		Assignment of explanatory variable	
		Random assignment	Non-random assignment
Selection of observational units from the population	Random sampling		
	Non-random sampling		

Practice: variable review and scope of inference

1. The pharmaceutical company Moderna Therapeutics, working in conjunction with the National Institutes of Health, conducted Phase 3 clinical trials of a vaccine for COVID-19 in the fall of 2022[2]. US clinical research sites enrolled 30,000 volunteers without COVID-19 to participate. Participants were randomly assigned to receive either the candidate vaccine or a saline placebo. They were then followed to assess whether or not they developed COVID-19. The trial was double-blind, so neither the investigators nor the participants knew who was assigned to which group.

   a) Identify the explanatory and response variables in this study, and describe whether they are categorical or numerical.

   - **Explanatory:** Whether the participant received the vaccine (categorical)
   - **Response:** Whether the participant developed COVID-19 (categorical)

   b) Are we able to generalize these results beyond the participants in the study? Why or why not?

   Since the participants self-selected, we would not expect this sample to be representative of a larger population. Therefore, we cannot generalize these results beyond this study.

   c) Are we able to determine that the differences observed between patients who received the vaccine and those that received the placebo are due to whether they received the vaccine? Why or why not?

   Since the participants were randomly assigned to whether they received the vaccine, we would expect potential confounding variables to be approximately evenly distributed between the two groups (on average) and can therefore determine a causal relationship between developing COVID-19 and whether the participant received the vaccine.

2. White-nose syndrome (WNS) is a deadly condition affecting bats in North America, caused by a fungus known as *Pseudogymnoascus destructans*. To combat the spread of WNS, researchers at the U.S. Geological Survey (USGS) and U.S. Fish and Wildlife Service (FWS) developed a vaccine to be distributed to large caves in which bats hibernate, called hibernacula. To test the viability of their vaccine, researchers administered the vaccine to a subset of little brown bats that were easy to access in two large hibernacula in Wisconsin. Researchers followed these bats until the following year and found reduced viral loads (measured in copies/mL) compared to unvaccinated bats in the hibernacula.[3]

   a) Identify the explanatory and response variables in this study, and describe whether they are categorical or numerical.

[2] Moderna clinical trial
[3] White-nose syndrome article

- **Explanatory:** whether the bat received the vaccine (categorical)
- **Response:** the viral load (numerical)

b) Are we able to generalize these results beyond the bats in these two hibernacula? Why or why not?

The bats were selected based on convenience (proximity to the mouth of the cave). Therefore, we would not expect these bats to be representative of any larger population of bats and cannot generalize these results.

c) Are we able to determine that the differences observed between the bats that received the vaccine and those that did not are due to whether they received the vaccine? Why or why not?

The bats were not randomly assigned to whether they received the vaccine. Therefore, we cannot conclude that the observed differences in viral load are due to whether the bat received the vaccine.

3. In 2008, the Los Angeles Times published a headline that included "Hospitals … Riskier than a Casino in Event of Cardiac Arrest.[4]" The article, based on a study published in the *New England Journal of Medicine*[5], states that the widespread availability of defibrillators and bystanders in public places like casinos leads to a higher survival rate than hospitals in the case of cardiac arrest.

a) Identify the explanatory and response variables in this study, and describe whether they are categorical or numerical.

- **Explanatory:** Whether participants were in a hospital ("location" is also acceptable)
- **Response:** Whether participants survived the cardiac arrest event (**not** the survival rate, which is a summary statistic)

b) What is a potential confounding variable in this study?

The severity of the cardiac arrest event. We would certainly expect the severity of the cardiac arrest event to be related to whether the subject survives (response), and also related to whether the subject is in the hospital to begin with (explanatory).

c) Are we able to conclude that being in a hospital during cardiac arrest caused the lower survival rate? Why or why not?

[4] LA Times article

[5] Chan, P.S., Krumholz, H.M., Nichol, G., Nallamothu, B.K. 2008. Delayed time to defibrillation after in-hopital cardiac arrest. *New England Journal of Medicine* 358(1) 9-17.

> Certainly not! The location of the cardiac was not randomly assigned, so we cannot determine a causal relationship.

d) If you were having a heart attack and your goal was to survive, would you go to a hospital or casino?

> Hospital :)

Extra practice: Populations, Parameters, and Scope of Inference

For each scenario below, identify:

- the **population** of interest
- the **sample** actually obtained
- the **parameter**
- the **statistic**
- the **scope of inference**

For the scope of inference, be sure to justify your claim.

Flu vaccination attitudes among university students

The following is an excerpt from the abstract of a paper by Dopelt et al. (2025) on the attitudes of college students towards flu vaccination:

> Influenza poses a significant public health threat, causing millions of severe cases and hundreds of thousands of deaths annually. Vaccination remains the most effective measure to reduce transmission, complications, and the strain on healthcare systems. Despite its importance, vaccination uptake remains suboptimal among college students, a key population for infectious disease transmission and a target group for public health interventions. The primary objective of this study was to examine the psychological and contextual determinants of influenza vaccination among college students in Israel, utilizing the Theory of Planned Behavior (TPB) as the guiding framework. A cross-sectional online survey was conducted among 591 students at Ashkelon Academic College between April and May 2023. The questionnaire included validated items assessing attitudes toward vaccination, subjective norms, perceived behavioral control, and vaccination history.

You may suppose that researchers are interested in estimating the proportion of students that have or plan to be vaccinated against the flu; of the 591 students in the survey, 76 were vaccinated, 266 intended to get vaccinated, 47 were undecided, and 202 did not intend to get vaccinated.

- Population of interest: college students in Israel
- Sample: 591 self-selected students at one college
- Parameter: the true proportion of all Ashkelon Academic College students that feel positively about flu vaccination.
- Statistic: $\hat{p} = \frac{76+266}{76+266+47+202} = \frac{342}{591}$
- Scope of inference: Since participants were not randomly selected for this study (they self-selected online), we cannot generalize these results beyond the 591 participants. There really is no notion of causation here, since there is only one variable.