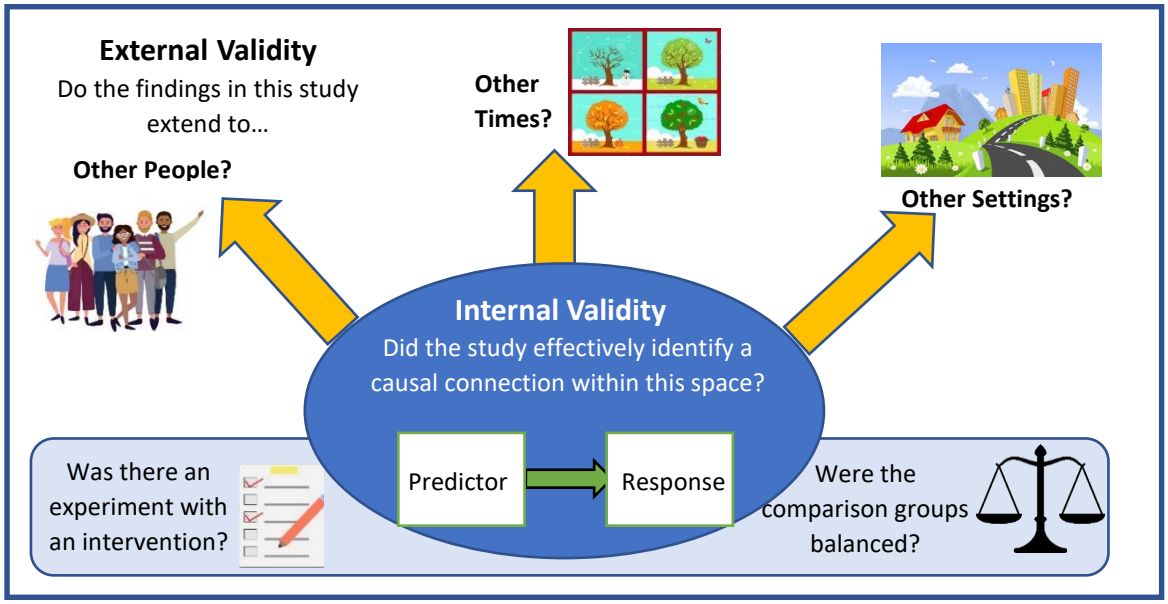
Chapter 9: Evaluating External Validity

Evaluating Validity

- **Validity** refers to the strength of the claims that a statistical study makes.
- Validity can be divided into two categories:
  - **Internal validity**: Is there evidence for a causal link between two variables within this study? Think looking "internally."
    - Does smoking *cause* cancer?
    - Does this medication *directly decrease* LDL cholesterol levels?
    - Internal validity is relevant to assess when exploring *multivariate* questions.
  - **External validity**: Is there evidence that findings in this study generalize to a broader population, setting, and time? Think looking "externally."
    - We surveyed 500 people, and 56% approve of the President's performance. How well do these 500 people *represent* the greater U.S. population?
    - External validity is relevant to assess for both univariate and multivariate questions.
  - The **design** of a statistical study helps us determine a study's *internal validity*.
  - The **sampling procedures**, **setting**, and **timing** help us determine a study's *external validity*.



Practice: One study found that students who took handwritten notes in one particular class performed better on the Final Exam. Decide whether each question below is targeting the study's internal or external validity.

Is taking notes actually leading to more learning? Or is it just that students who take notes happen to be better learners in other ways?	If we repeated this study in a different section of the course with a different instructor, would we still see the same result?
Internal Validity	External Validity

Statistical Power vs. External Validity

- Sampling is a key part of external validity, but sampling is also related to the statistical power of an inference too.

How large is the sample?	How representative of the population is this sample?
Sample size affects the **power** of our inference. The larger the sample size, the easier it will be to detect a departure from the null hypothesis, or estimate the true parameter	*Representation* affects **external validity.** This determines how appropriate it is to generalize to the population.

Power and External Validity are *separate* things to assess in a study!

- Power
  - Larger sample sizes result in smaller standard errors
  - This gives us a better chance of detecting small deviations from the null hypothesis with more confidence, resulting in a lower p-value when the alternative is true.
  - Studies with poor power are more likely to make a Type II error (fail to reject the null when it truly is false).
- External Validity
  - In contrast, external validity is *not* concerned with how large our sample is—only whether the sample we have resembles the population we want to generalize to.

Practice: We're completing a study to estimate the amount of time that University of Illinois students (of all academic levels/programs) spend on school each week. Consider the following sampling plans: What potential issues can you think of for each that may limit the external validity of the claims we make from each?

Take a sample of students taking STAT 100 during the Fall	Conduct a poll on the UIUC reddit page.	Ask students on the Quad one afternoon to fill out a survey

STAT 100: certain majors excluded. Possibly biased toward freshmen/sophomores.

Reddit: Biased toward STEM programs, or possibly more introverted types

Quad: Possibly biased toward more outdoorsy people, undergrads, people living on campus

External Validity – Determining Generalizability

- **Sampling**
  - **A Simple Random Sample (SRS)** (sometimes just called "*random selection"*) means that every member of the population has an *equal* chance to be chosen for the sample at all times in the sampling process.
    - In this scheme, sampling remains *independent*, meaning that possibility of one person being chosen does not affect the chances that someone else is chosen.
    - Unfortunately, simple random sampling is often not possible due to inescapable biases in most contexts:
    - Undercoverage **Bias:** Some in the population don't have an equal chance (or more often, no chance at all) of being selected for the sample. That is because it is often difficult to have contact information for every person or unit in the population. It costs time and money.
    - Volunteer **Bias** *(may also be called Self-select or Non-response Bias)*: The sample is composed of subjects who *chose* to participate and others who chose to ignore or forgot. This is expected with human populations since we can't force participation!

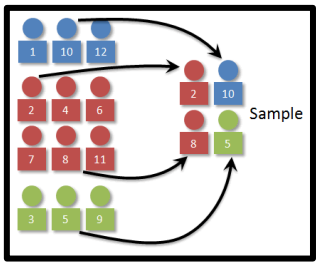
Practice: A pollster is contacting people for a survey on public transportation by collecting responses in a busy downtown square. However, not everyone in the population of interest passes through that square. ***What type of bias would that be?***

Undercoverage (certain people in population have no chance of being asked)

Practice: On the Quad, some students complete the questionnaire, while others decline the questionnaire. ***What type of bias would that be?***

Volunteer (certain people who are invited choose to complete it while others don't)

- **Quota Sampling/Weighted Samples**
  - One way to address undercoverage or volunteer biases is to get target numbers from different key subgroups (e.g., equal male and female, proportional ages, etc.). This is called "Quota" sampling.
  - Examples: Age, Gender, Household Income



  - An alternative to quota sampling is to take a "weighted sample," where we adjust the weight of responses from demographic brackets that are oversampled in order to get what we believe is a more accurate estimate.
  - This type of sampling is often used for public opinion surveys where we know certain subgroups are just harder to contact.

  - Curious to know more? Here is a video on sampling with polls! https://www.youtube.com/watch?v=fzzX9jHDK4k

- **Convenience sampling**
  - This non-random sampling method is quite common in people-centered research. Think online polls or surveys in which there is no attempt to get quotas, or to weight the final responses to represent population demographics.
  - Any time a sampling method has an undercoverage bias or volunteer bias, and that bias is not compensated by some type of weighted or quota sampling scheme, it is technically a convenience sample.
  - People should be cautious of using convenience samples to make statistical generalizations to a larger population.
  - Snowball **Sampling** is a common component of convenience sampling that relies on word of mouth (think shares on social media) to get participants. This would be a *huge* threat to independence in that people participate based on whether they know someone who did. The sample might be an echo chamber.

External Validity Threats

- When evaluating the validity of a study, researchers often frame known or potential problems as "validity threats." The first external validity threat we have defined is known as a "participant selection" threat.
- **PARTICIPANT SELECTION** – Does this group of participants represent the population?
    - We need to make the case that our sample **represents** the population at large and identify what elements of our sampling scheme introduce a threat to representativeness.
    - In particular, we should note in what ways undercoverage bias or a volunteer bias may threaten how well our sample represents the population.
- Besides participant selection, there are also other limitations that can affect our ability to generalize our study's findings to other places and other times.
    - **SETTING LIMITATIONS** – Is the setting representative of all settings we wish to generalize to?
        - **Physical Environment**: What space or what features were involved in this study. A particular doctor's office? An outdoor vs. indoor location? A biased/limited range of external factors? *Do these features generalize?*
        - **Social Environment**: What is the social context for this study, and might that matter with what was studied: Were particular doctors or nurses involved?
        - **Context**: What other contextual factors did this study take place within? A particular weather event or season? The materials used in the study or instrumentation*?*
        - Note that setting threats to external validity are different than that of internal validity.
            - Setting threats for ***internal validity*** have to do with confounding effects between groups. For example, did my treatment and control group complete their participation in different rooms?
            - Setting threats for ***external validity*** are related to setting encapsulating my whole study. Perhaps my participants completed the treatment in a lab, but would these results generalize to household settings?

    - **HISTORICAL SUSTAINABILITY** – Do these results generalize to other times?
        - This threat should be taken into account when dealing with external factors that may change over time—questions linked to culture, lifestyle habits, entertainment, etc.
        - For example: A poll about Americans' views about government surveillance or terrorist prevention before the terrorist attacks on September 11th 2001 may no longer generalize to Americans' views after that event.

Chapter 9 Additional Practice

Practice: What do you think of each sampling plan? What limitations might apply to the participant selection in each situation?

> A poll on msn.com asks American users 18 and over whether they plan to vote in the upcoming Midterm elections. After voting, the website encourages people to share the link of the poll with their friends on social media.

Snowball sampling accelerates bias by getting more people with common views (also undercoverage via the audience of msn.com)

> A clinic is surveying the 874 patients from the past year to assess satisfaction with their recent visits. 209 (24%) of them complete the survey. According to clinic data, 68% of the respondents said they would likely visit again or recommend to family members.

High volunteer bias. Those who responded may be more likely to have positive experiences.

> A university selects 100 graduating seniors by randomly selecting their email addresses from among those who have applied for graduation. These 100 students are asked to complete an exit interview for $20. At the conclusion, a total of 75 of the 100 contacted students completed the exit interview.

Small volunteer bias, but overall likely good representation.

Practice: Researchers in 1985 studied whether the program one watches on TV before bed may assist someone in falling asleep faster. Randomly selected landline phones in America's top 10 most populated cities were called, and callers asked to speak to "heads of household". The researchers concluded that people who watched the news were more likely to fall asleep faster than those watching any other TV platform.

What threats to external validity are present in this study? *Some threats could be related to more than one threat category—that's ok! These categories are simply entry points to finding possible limitations, not mutually exclusive categories.*

Participant Selection Threat? **Setting Threat?** **Historical Sustainability Threat?**

Generalizes only to larger cities. Bias toward men (through "head of household" language). Nature of news programming might have changed considerably since 1985.

Chapter 9 Learning Goals

After this chapter, you should be able to…

- Distinguish questions of external validity from questions of power
- Recognize and distinguish between an undercoverage bias and a volunteer bias in sampling methods
- Identify a simple random sampling method (aka “random sampling” or “random selection”) from a contextual description and its external validity value in avoiding undercoverage and volunteer biases.
- Identify quota and weighted sampling methods from a contextual description and recognize their value in minimizing undercoverage and volunteer biases
- Identify a convenience sampling method from a contextual description and recognize its relatively weak external validity
- Recognize limitations to a statistical claim’s external validity by using the lenses of participant selection, setting limitations, and historical sustainability.