Asking causal questions

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

January 13, 2025

Recap

Last time: Crash course on R essentials

- Getting data into R: working directories, `read_csv()`
- Essential data manipulation
  - → Subsetting with `filter()` and `select()`
  - → Creating and editing columns with `mutate()` + helpers like `if_else()` and `case_when()`
  - → Groupwise calculations with `group_by()` and `summarize()`
- Data visualization in ggplot2

All this treated as assumed knowledge from here on out

Today’s agenda

This week + next: Conceptual foundations of statistical analysis of causality

What we’ll cover today

1. Causal statements as counterfactual statements
2. Asking causal questions that statistics can (try to) answer

Causality and counterfactuals

Translating causal statements

“Trump won the election because Biden didn’t drop out in 2023.”

“If Biden had dropped out in 2023, Trump would not have won the election.”

“I voted in the 2018 midterm because I saw an Instagram story about the election from a band I follow.”

“If I had not seen an Instagram story about the election from a band I follow, I would not have voted in the 2018 midterm.”

“I’m antsy today because I drank a cold brew and didn’t eat breakfast.”

“If I had not drunk a cold brew, and/or if I had eaten breakfast, I would not be antsy today.”

The trouble(s) with counterfactuals

"If Biden had dropped out in 2023, Trump would not have won the election."

The philosophical problem: How can this statement be "true"?

- David Lewis (not the one we know): it's true if, in alternate world closest to ours in which Biden dropped out in 2023, Trump won the election
  - → This is controversial (see the Stanford Encyclopedia entry)
- Here in PSCI 2301 we'll ignore the deep philosophical issues

The trouble(s) with counterfactuals

“If Biden had dropped out in 2023, Trump would not have won the election.”

The practical problem: How do we gather/evaluate evidence for this?

- We only see what happened, this is a statement about what didn’t
- Many paths to the same outcome
  - → Different Democratic candidate beats Trump
  - → Republicans nominate a different candidate who goes on to win
  - → Aliens destroy Earth to build an interstellar bypass and the election doesn’t happen
- Not only can’t we be certain, we can’t realistically quantify just how uncertain we are

From individual effects to average effects

“If Biden had dropped out in 2023, Trump would not have won the election.”

→ Leaving statements like this for the speculative fiction writers

“If I had not seen an Instagram story about the election from a band I follow, I would not have voted in the 2018 midterm.”

→ Can’t assess this statement for <u>me</u> specifically

…but can think about the <u>average</u> effect of social media influence on voting

Causal statements about average effects

“On average, seeing a get-out-the-vote message from a social media influencer increases a registered voter’s likelihood of voting by 10%.”

“Take a large random sample of registered voters. Imagine we could observe them under two scenarios.

In Scenario A, none of the people in the sample see a get-out-the-vote message from a social media influencer. In Scenario B, all of the people in the sample see a get-out-the-vote message from a social media influencer.

Voter turnout in the sample would be 10% higher in Scenario B than in Scenario A.”

This is an informal statement of the potential outcomes model of causality, which we’ll formalize next week.

Causal statements about average effects

Can’t assess: “If I had not seen an Instagram story about the election from a band I follow, I would not have voted in the 2018 midterm.”

Can assess: “On average, seeing a get-out-the-vote message from a social media influencer increases a registered voter’s likelihood of voting by 10%.”

Why the difference?

We can’t observe voting behavior after influencer message and after no influencer message for the **same person**

But we can observe voting behavior for a **group of people** who got influencer messages and for a — hopefully similar — group who didn’t

Asking causal questions

Data can’t tell you “why”

Why do people vote?

A good question, and a good thing to be curious about!

But not one we can (directly) answer with data/statistics

- Basically every social outcome has multiple overlapping influences
- Impossible to measure all of them, let alone include all in one model
- Have to break down the big questions into smaller, manageable ones

Ingredients of a causal analysis: Outcome variable

Starting point: Identify the **outcome** you want to explain variation in

Sometimes called the **dependent variable** or the **response**

Think about variation: Why do *some people* vote and *other people* don't?

If no variation, statistics won't help us

- "Why do people breathe?" is a question for biology, not stats
- "Why did Trump win in 2024?" at a minimum needs refinement
    - → Why did some people vote for Trump and others didn't?
    - → Why did some communities shift more towards Trump compared to 2020, and others did less so?

Ingredients of a causal analysis: Units

Must define the **unit of analysis** at which you’ll measure the outcome

Imagine what each row of your eventual data frame will constitute

Why did some people vote for Trump and others didn’t?
→ Unit of analysis is the individual

Why did some communities shift more/less towards Trump vs 2020?
→ Unit of analysis might be county, city, metro area, state…

Appropriate unit partly depends on outcome, partly on data availability

Ingredients of a causal analysis: Population

Statistics can only draw causal inferences about groups, not individual units

So you must define the **population** of units to draw inferences about

Statistical ideal is to have a random sample of units from the population

Ingredients of a causal analysis: Population

Why did some people vote for Trump and others didn’t?

Unit is the individual, but what’s the population?

- Americans who voted in the 2024 election?
- Americans who were registered to vote in the 2024 election?
- Americans who were eligible to vote in the 2024 election?
- Americans of voting age as of the 2024 election?

Ideal population definition depends on the broader research question or theory you’re trying to address

Some methods will only allow us to draw inferences about certain subpopulations

Ingredients of a causal analysis: Treatment variable

From "why?" to "what is the effect of...?"

Last critical thing is to specify the **treatment** whose average effect on the outcome you want to estimate (sometimes called **independent variable**)

Must plausibly be manipulable among units in the population

Consider when units are individuals, outcome is voting for Trump in 2024. Which "treatments" are plausibly manipulable?

- Whether they see a GOTV message from a social media influencer?
- Whether they have voted in the past?
- Their annual income?
- Whether they identify as a Republican?
- Whether they are Black?

Ingredients of a causal analysis: Treatment variable

From “why?” to “what is the effect of...?”

Stats deals with variation → need to consider the baseline for comparison against your treatment

Sometimes called the **control** (though don’t confuse w/ “control variables”)

Potential comparisons for “influencer get-out-the-vote message” treatment

- No influencer message at all
- Non-political message from influencer
- Get-out-the-vote message from government bureaucrat

Appropriate comparison group depends on the broader research question or theory you’re trying to address

In-class exercise

First identify a “Why?” question you’re interested in

Then work through a potential causal analysis to study it statistically

Make sure to define:

- Outcome of interest
- Unit of analysis
- Population
- Treatment variable (must be manipulable!)
- Comparison group

Wrapping up

What we did today

1. Defined causal statements in terms of counterfactuals
    - “If [treatment] had been different, then [outcome] would have been different”
    - Statistics can only address average effects
2. Laid out ingredients of a causal analysis
    - Unit of analysis and population of interest
    - Outcome variable
    - Treatment variable and comparison group

~~Why does [outcome] happen?~~

What is the average effect of [treatment], compared to [control/baseline], on [outcome]?

To do for next time

1. Read “Correlation, Causation, and Confusion” and “Introduction to Causality” if you didn’t already
2. Start thinking in terms of *outcomes* and *treatments* for your final project