Week 4, Class 7

Causality

Sean Westwood

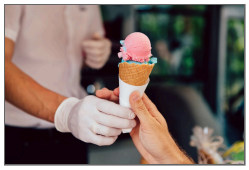
Today's Learning Objectives

By the End of Class You Will:

- Understand the three requirements for establishing causality
- Distinguish between association and causation
- Recognize the importance of temporal ordering in causal claims
- Identify confounding variables and spurious correlations
- Use AI assistance to identify alternative explanations

What Does It Take to Show Causation?

Ice Cream Crime



In 2013, “researchers” noticed that counties with higher ice cream sales had higher rates of violent crime. Should we ban ice cream to reduce crime?

Your intuition says: Probably not…
But why not? What’s wrong with this reasoning?

This is a classic example of a **spurious correlation.**

- Hot weather increases both ice cream sales AND violent crime. Temperature is the **confounding variable** that explains both phenomena.
- Correlation ≠ Causation. Just because two things happen together doesn't mean one causes the other.

The Three Requirements for Causality

To claim that X causes Y, you need:

1. **Association**: X and Y are related (when X changes, Y changes)
2. **Temporal Ordering**: X happens before Y (cause before effect)
3. **No Confounding**: No other variable(s) Z causes both X and Y

All three are required. If any one is missing, you cannot make a causal claim.

Requirement 1: Association

Demonstrating Relationship

Association means: When X changes, Y systematically changes too
Examples of association:

- Higher campaign spending correlates with more votes
- More education correlates with higher income
- Social media use correlates with political polarization

Testing Associations with Data: Example 1 - Poverty and Savings

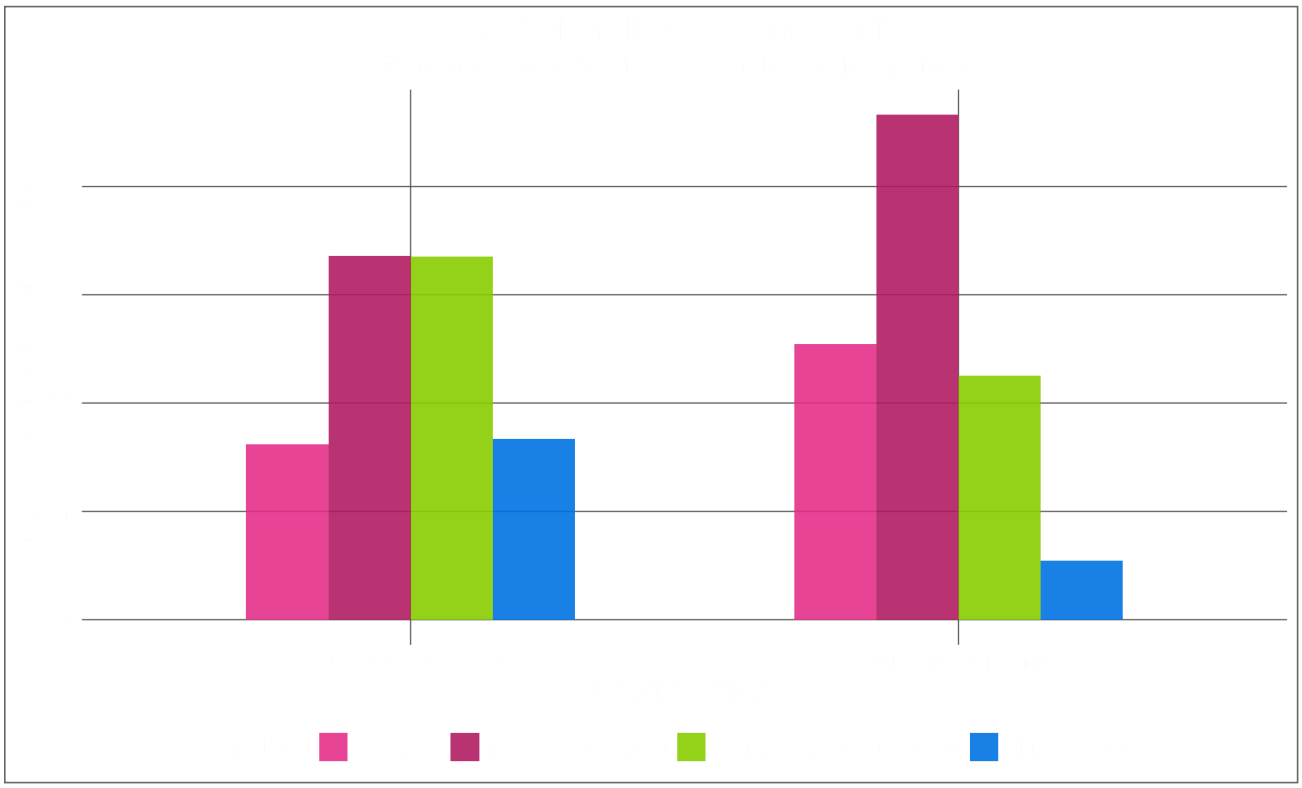
```r
library(tidyverse)

# Load poverty and savings data
poverty <- read_csv("../../data/poverty.csv")

# Test association: Does savings vary by poverty status?
poverty_savings <- poverty %>%
  mutate(
    poverty_status = case_when(
      income_less20k == 1 ~ "Below Poverty Line",
      income_less20k == 0 ~ "Above Poverty Line"
    ),
    # Create savings categories based on account amounts
    savings_level = case_when(
      is.na(accts_amt) | accts_amt == 0 ~ "No Savings",
      accts_amt > 0 & accts_amt <= 1000 ~ "Low Savings ($1-$1,000)",
      accts_amt > 1000 & accts_amt <= 10000 ~ "Medium Savings ($1,001-$10,000)",
      accts_amt > 10000 ~ "High Savings ($10,000+)"
    ),
    # Reorder factor levels to: No, Low, Medium, High
    savings_level = factor(savings_level,
                           levels = c("No Savings",
                                      "Low Savings ($1-$1,000)",
                                      "Medium Savings ($1,001-$10,000)",
                                      "High Savings ($10,000+)"))
  ) %>%
  filter(!is.na(income_less20k)) %>%
```

```
# A tibble: 8 × 4
# Groups:   poverty_status [2]
  poverty_status     savings_level                  count proportion
  <chr>              <fct>                          <int>      <dbl>
1 Above Poverty Line No Savings                      253     0.162
2 Above Poverty Line Low Savings ($1-$1,000)         526     0.336
3 Above Poverty Line Medium Savings ($1,001-$10,000) 525     0.335
4 Above Poverty Line High Savings ($10,000+)         261     0.167
5 Below Poverty Line No Savings                      281     0.254
6 Below Poverty Line Low Savings ($1-$1,000)         515     0.466
7 Below Poverty Line Medium Savings ($1,001-$10,000) 249     0.225
8 Below Poverty Line High Savings ($10,000+)          60     0.0543
```

Visualizing Association: Poverty and Savings



- Different poverty groups show different savings distributions, suggesting poverty and savings are strongly associated

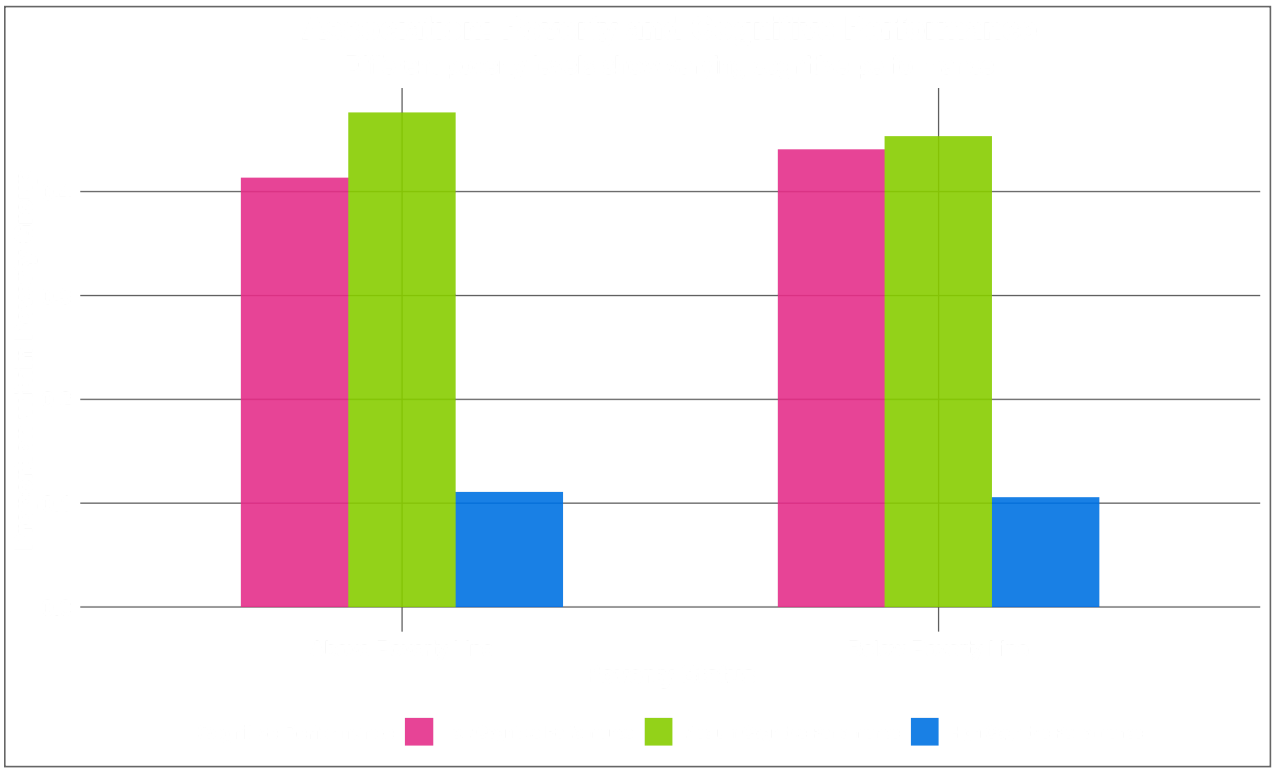
Testing Associations with Data: Example 2 - Poverty and Cognitive Performance

```r
# Test association: Does cognitive performance vary by poverty status?
poverty_cognitive <- poverty %>%
  mutate(
    poverty_status = case_when(
      income_less20k == 1 ~ "Below Poverty Line",
      income_less20k == 0 ~ "Above Poverty Line"
    ),
    # Create education proxy based on cognitive performance (stroop test times)
    cognitive_level = case_when(
      stroop_time <= 7.3 ~ "High Cognitive Performance",
      stroop_time > 7.3 & stroop_time <= 7.6 ~ "Medium Cognitive Performance",
      stroop_time > 7.6 ~ "Low Cognitive Performance"
    ),
    # Reorder factor levels to: Low, Medium, High
    cognitive_level = factor(cognitive_level,
                             levels = c("Low Cognitive Performance",
                                        "Medium Cognitive Performance",
                                        "High Cognitive Performance"))
  ) %>%
  filter(!is.na(stroop_time), !is.na(income_less20k)) %>%
  group_by(poverty_status, cognitive_level) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(poverty_status) %>%
  mutate(proportion = count / sum(count)) %>%
  arrange(poverty_status, cognitive_level)

poverty_cognitive
```

```
# A tibble: 6 × 4
# Groups:   poverty_status [2]
  poverty_status     cognitive_level              count proportion
  <chr>              <fct>                        <int>      <dbl>
1 Above Poverty Line Low Cognitive Performance      647      0.413
2 Above Poverty Line Medium Cognitive Performance   745      0.476
3 Above Poverty Line High Cognitive Performance     173      0.111
4 Below Poverty Line Low Cognitive Performance      487      0.441
5 Below Poverty Line Medium Cognitive Performance   501      0.453
6 Below Poverty Line High Cognitive Performance     117      0.106
```

Visualizing Association: Poverty and Cognitive Performance



Weak (no?) association: Less clear relationship between poverty and cognitive performance compared to poverty and savings

Requirement 2: Temporal Ordering

Cause Must Come Before Effect

Seems obvious, but it's often unclear in real data

Examples where timing matters:

Poverty & Crime

OR

Knowledge & Voting

OR

Spending & Votes

OR

The Challenge of Simultaneous Measurement

```
# Most data is cross-sectional - measured at the same time
poverty %>%
  select(treatment, income_less20k, accts_amt, stroop_time, cash) %>%
  head(5)
```

```
# A tibble: 5 × 5
  treatment     income_less20k accts_amt stroop_time  cash
  <chr>                 <dbl>     <dbl>       <dbl> <dbl>
1 Before Payday             0      3000        7.61    30
2 Before Payday             1       800        7.27    75
3 After Payday              0     15000        7.71   110
4 After Payday              0        NA        7.35   160
5 Before Payday             0     40000        7.37    40
```

Problem: We see poverty, savings, cognitive performance, and cash measured simultaneously. We can't tell which came first:

- Did poverty lead to lower savings, or did lack of savings lead to poverty?
- Did poverty affect cognitive performance, or did poor cognitive performance lead to poverty?
- Did having more cash improve both savings and cognitive performance?

Cross-sectional data cannot establish temporal ordering

Solutions for Temporal Ordering

Longitudinal Data: Follow the same people over time
Natural Timing: Use events with clear before/after structure
Historical Records: Trace sequences of events (often not credible)

Example: To study if negative ads cause lower campaign donations:

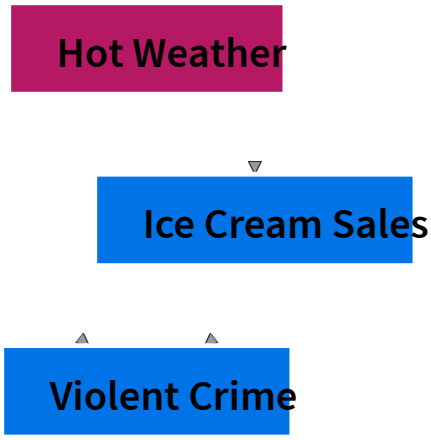
- **Weak**: Survey voters after election about ads and donations
- **Strong**: Measure donations before/after negative ad campaigns begin

Requirement 3: No Confounding

The Biggest Challenge

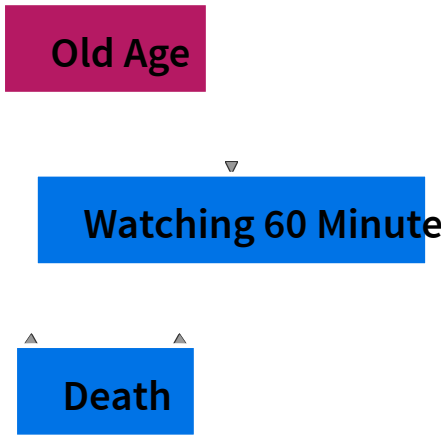
Confounding occurs when: A third variable Z causes both X and Y, creating a spurious association

Death by ice cream



Hot weather causes both → spurious correlation

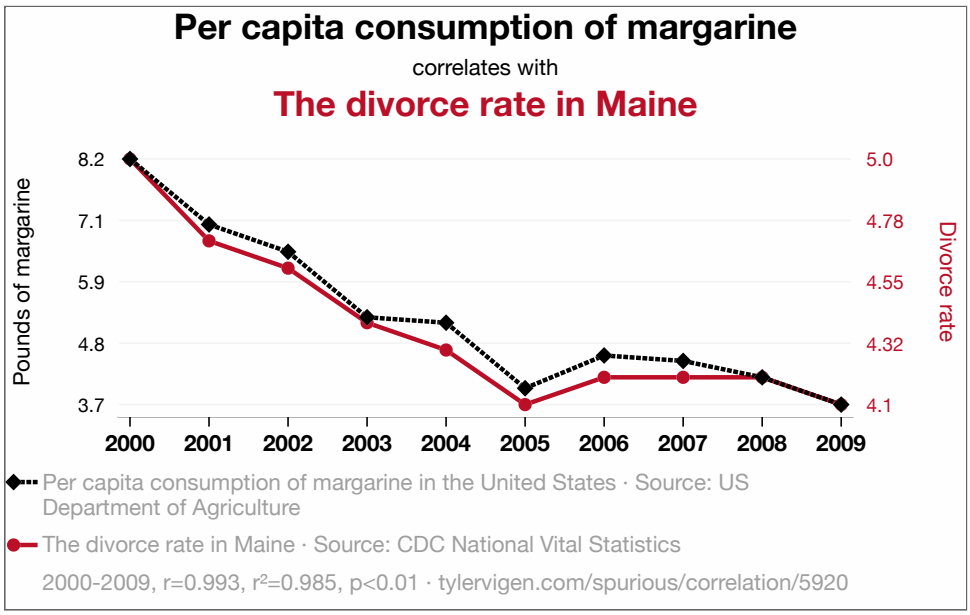
Death by 60 Minutes



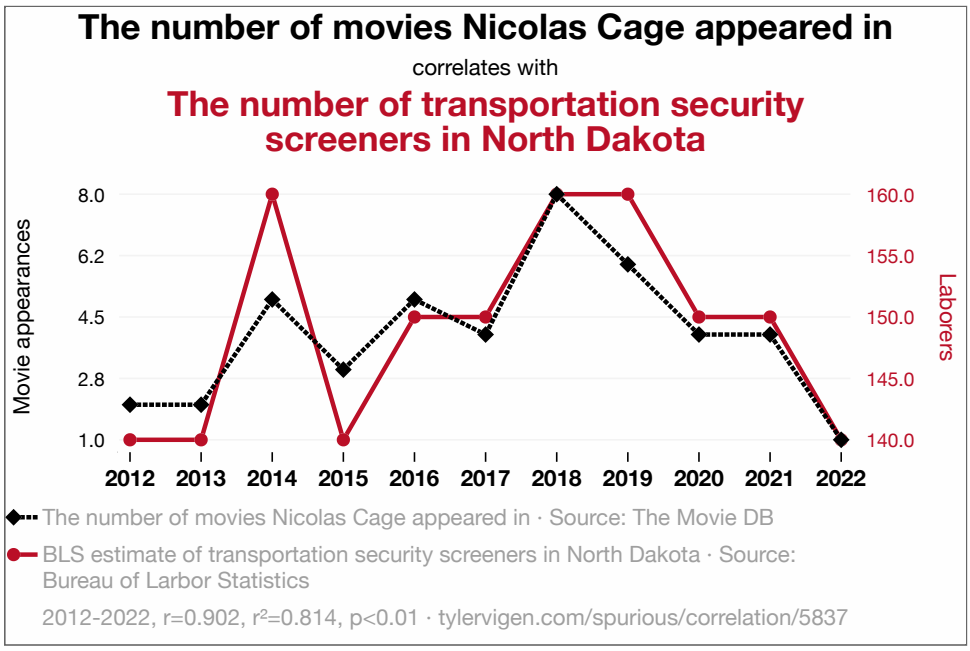
Old age causes both → spurious correlation

Spurious Correlations

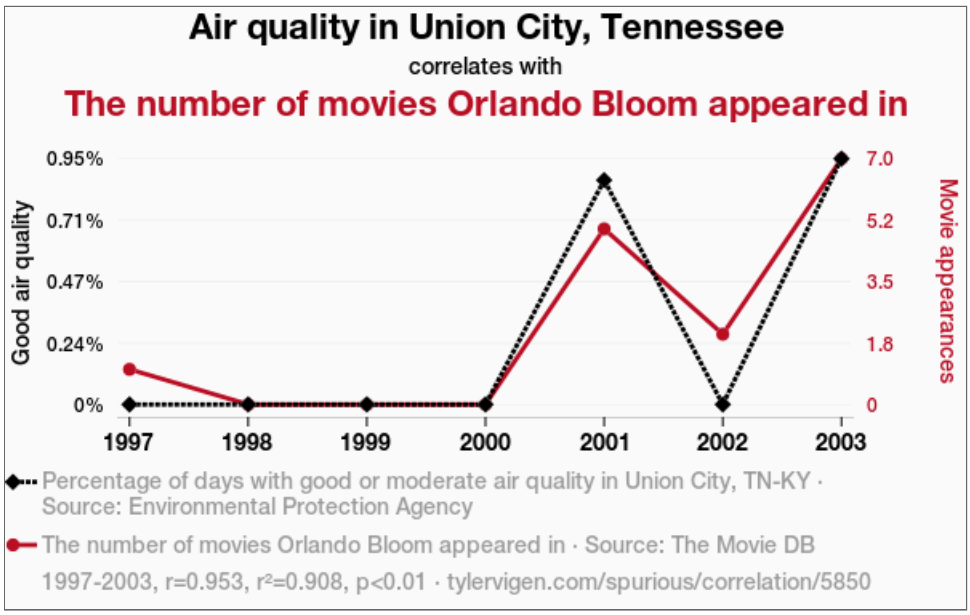
Spurious Correlation: Divorce and Margarine



Spurious Correlation: TSA and Nicolas Cage



Spurious Correlation: Air Quality and Orlando Bloom



Lesson: Always ask “What else could explain this relationship?”

Identifying Confounders

Common confounders in political science:

- **Socioeconomic status**: Affects voting, education, health, political views
- **Geographic region**: Affects culture, economics, political preferences
- **Age/generation**: Affects technology use, political attitudes, life experiences
- **Media environment**: Affects information, political knowledge, opinions

Always ask: “What else could explain both my cause and my effect?”

Modern Applications: Campaign Finance and Vote Share

A Contemporary Causal Question

Question: “Does campaign spending cause candidates to win more votes?”
The Association: Clear relationship between spending and vote share

```r
# Load real campaign finance data
campaign_data <- read_csv("../../data/campaign.csv") %>%
  # Clean and prepare the data
  filter(State=="NH") %>%
  # Convert to more interpretable units
  mutate(
    spending_thousands = total.raised.candidate / 1000,
    vote_share = total.votes ,
    incumbent = ifelse(is.na(prev.elect), 0, 1)  # Create incumbent indicator
  )

# Show the association
round(cor(campaign_data$spending_thousands, campaign_data$vote_share, use = "complete.obs"),2)
```

```
[1] 0.38
```

The correlation coefficient measures the strength and direction of a linear relationship between two variables, ranging from −1 (perfect negative) to +1 (perfect positive).

Is This Confounded

```r
# Investigate confounders: Do incumbents spend more AND get more votes?
campaign_data %>%
  group_by(incumbent) %>%
  summarise(
    avg_spending = mean(spending_thousands, na.rm = TRUE),
    avg_vote_share = mean(vote_share, na.rm = TRUE),
    count = n(),
    .groups = "drop"
  ) %>%
  mutate(
    incumbent = ifelse(incumbent == 1, "Incumbent", "Challenger"),
    avg_spending = round(avg_spending, 1),
    avg_vote_share = round(avg_vote_share, 3)
  )
```

```
# A tibble: 2 × 4
  incumbent  avg_spending avg_vote_share count
  <chr>             <dbl>          <dbl> <int>
1 Challenger         53.6           19.2    36
2 Incumbent          87             21.0     9
```

Incumbents spend more AND get more votes

- **Spurious reasoning**: “Spending causes votes”
- **Reality**: “Being incumbent causes both more spending and more votes”

Using AI to Identify Alternative Explanations

AI as Your Confounding Detective

Effective prompt for identifying confounders:

> “I found that cities with more coffee shops have higher voter turnout. Before concluding that coffee shops increase civic engagement, help me brainstorm what other variables might cause both coffee shop density AND voter turnout. Think about demographics, economics, and culture.”

AI for Causal Critique

Research design critique prompt:

> “Evaluate this causal claim: ‘Social media use reduces political knowledge because people who spend more time on social media score lower on political knowledge tests.’ What are the three requirements for causality, and does this study meet them? What alternative explanations should I consider?”

Common Mistakes in Causal Reasoning

Post Hoc, Ergo Propter Hoc

“After this, therefore because of this”
The Error: Assuming that because B followed A, that A caused B
Examples:

- “I wore my lucky shirt and my team won”
- “The economy improved after the new president took office”
- “Crime dropped after we hired more police”

Remember: Temporal ordering is necessary but not sufficient for causation

Selection Bias Disguised as Causation

The Problem: Comparing groups that chose to be different

Examples:

- “Private school students have higher test scores” (ignoring family background)
- “People who exercise live longer” (ignoring health consciousness)
- “Countries with democracy are more prosperous” (ignoring development level)

Reverse Causation

Getting the direction of causation backwards

Examples:

- Does happiness cause success, or success cause happiness?
- Do good policies cause economic growth, or economic growth enable good policies?

Best Practices for Causal Claims

Red Flags in Causal Arguments

Be skeptical when you see:

- “Studies show X causes Y” (without mentioning study design)
- Causal claims from cross-sectional data
- No discussion of alternative explanations
- Dramatic policy recommendations from single studies

Strengthening Causal Arguments

Use multiple lines of evidence:

- Replicate findings across different populations
- Test with different research designs
- Look for natural experiments
- Check for temporal consistency

Acknowledge limitations:

- Be explicit about what you can and cannot conclude
- Discuss alternative explanations
- Admit when evidence is merely suggestive

Looking Ahead

Next Week Preview

Modern Causal Inference:

- The fundamental problem of causal inference
- Randomized controlled trials as the gold standard
- Natural experiments and quasi-experimental designs
- Average Treatment Effects and their interpretation

Key Concepts to Remember

- **Correlation ≠ Causation** - but it's the first requirement
- **Three requirements**: Association, temporal ordering, no confounding
- **Confounding is everywhere** - always ask "what else could explain this?"
- **John Snow's method**: Map patterns, establish timing, rule out alternatives
- **AI can help** - but only if you ask the right questions

Questions?

Key takeaway: Establishing causation is hard work. It requires careful thinking, good research design, and systematic investigation of alternative explanations. Next week: We'll learn about the most powerful tools for making causal claims in the modern social sciences.

Speaker notes