CSE 291/DSC 291
Information Manipulation: Trustworthiness of Information in Cyberspace

Molly Roberts & Stefan Savage, Winter 2023

The Effects of Misinformation
How do we define an “effect”? 

- For individual i:
  - Treatment: $T_i$
    - Some intervention we are interested in
    - Taking aspirin vs. taking a placebo
    - Being exposed to misinformation vs. being exposed to news
    - Receiving a content warning vs. not receiving it
    - Usually with think of $T$ is binary to make things less complicated
  - Outcome: $Y_i$
    - What we want to assess the impact of treatment on, e.g.
    - Headaches
    - Retweeting
    - Voting

- What is the effect of $T$ on $Y$?
Potential outcomes framework for causal inference

- $Y_i(T_i=1)$: The outcome, $Y_i$, when $T_i=1$.
- $Y_i(T_i=0)$: The outcome, $Y_i$, when $T_i=0$.

The effect of $T_i$ on $Y_i$ is:
- $Y_i(T_i=1) - Y_i(T_i=0)$
- $Y_i(T_i=1)$: potential outcome under treatment
- $Y_i(T_i=0)$: potential outcome under control

The fundamental problem of causal inference: we can never observe both potential outcomes because we cannot rewind time.

The counterfactual: whatever potential outcome we do not observe
Fundamental problem of causal inference

- In an ideal world we would see this in our data:

<table>
<thead>
<tr>
<th>Unit_i</th>
<th>X_i^1</th>
<th>X_i^2</th>
<th>X_i^3</th>
<th>T_i</th>
<th>Y_i(0)</th>
<th>Y_i(1)</th>
<th>Y_i(1) − Y_i(0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>69</td>
<td>75</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>98</td>
<td>0</td>
<td>111</td>
<td>108</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>80</td>
<td>1</td>
<td>92</td>
<td>102</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>98</td>
<td>1</td>
<td>112</td>
<td>111</td>
<td>-1</td>
</tr>
</tbody>
</table>
Fundamental problem of causal inference

- But in the real world we see this:

<table>
<thead>
<tr>
<th>Unit_i</th>
<th>$X_i^1$</th>
<th>$X_i^2$</th>
<th>$X_i^3$</th>
<th>$T_i$</th>
<th>$Y_i(0)$</th>
<th>$Y_i(1)$</th>
<th>$Y_i(1) - Y_i(0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>69</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>98</td>
<td>0</td>
<td>111</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>80</td>
<td>1</td>
<td>?</td>
<td>102</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>98</td>
<td>1</td>
<td>?</td>
<td>111</td>
<td>?</td>
</tr>
</tbody>
</table>
Average Treatment Effect (ATE)

- We can’t get individual treatment effects, so we can settle for the average treatment effect

- \( ATE = E[Y_i(T_i=1)] - E[Y_i(T_i=0)] \)

- For treated units, \( = E[Y_i(T_i=1) | T_i=1] - E[Y_i(T_i=0) | T_i=1] \)

- For control units, \( = E[Y_i(T_i=1) | T_i=0] - E[Y_i(T_i=0) | T_i=0] \)

- But we don’t observe \( E[Y_i(T_i=0) | T_i=1] \) or \( E[Y_i(T_i=1) | T_i=0] \)!

- When will
  - \( E[Y_i(T_i=0) | T_i=1] = E[Y_i(T_i=0) | T_i=0] \)
  - \( E[Y_i(T_i=1) | T_i=0] = E[Y_i(T_i=1) | T_i=1] \)?
Assumptions needed

- **Unconfoundedness (Ignorability)**
  - \( Y(1), Y(0) \) are independent of \( T \)
  - Treatment status is independent of the potential outcomes
  - Violation: omitted variable bias

- **Stable Treatment Value Assumption (SUTVA)**
  - The treatment status of any unit does not affect the potential outcomes of other units (non-interference)
  - The treatments for all units are comparable (no variations in the treatment)

- If these two conditions are met, we can use:
  - \( E[Y_i(T_i=1)|T_i=1] - E[Y_i(T_i=0)|T_i=0] \)
  - to estimate the ATE.
Gold Standard: Randomized Experiment

- Gold standard for achieving unconfoundedness is a randomized experiment
  - Randomization breaks the dependence of $T$ on other variables
  - By design, $Y(1)$ and $Y(0)$ are independent of treatment status.

- Even in an experimental setting, SUTVA may not hold.
Often in the social sciences: “As if random”

- Experiments are hard in social settings
- We often have to rely on “as-if” random treatment assignment, where we can make an argument that we have unconfoundedness and that SUTVA holds.
- Examples: Interrupted time series, differences-in-differences, regression discontinuity, instrumental variables designs
- Or, we can use statistical models to adjust for known, observed confounders. However, these rely on a lot of untestable assumptions.
Some examples of experiments in social science

- Large “Get out the vote” (GOTV) experiments conducted by Alan Gerber and Don Green
- Send mailers to thousands of people with different messages. Estimate effects on voter turnout
- Randomize: In person, mass email, personalized phone call
- For some treatments find (small) effects
- Social treatments particularly effective
Some examples of experiments in social science

- 61 million person experiment
- One treatment group – info about voting and could click “I voted” which also showed them how many of their friends had
- Control group – info about voting, but no friend treatment
Some examples of experiments in social science

**Figure 1** The experiment and direct effects. **a, b,** Examples of the informational message and social message Facebook treatments (a) and their direct effect on voting behaviour (b). Vertical lines indicate s.e.m. (they are too small to be seen for the first two bars).
Aral and Eckles (2019)

- Four-step research agenda for estimating the effect of social media manipulation (treatment) on vote choice and turnout

- 1. Measure exposure to manipulation (measure the treatment)
  - What are the difficulties here?

- 2. Combine exposure data with data on voting behavior
  - What are the difficulties here?

- 3. Assess the effects of exposure on opinions and behavior
  - What do they suggest?

- 4. Compute consequences in aggregate for elections.
  - What does this mean?
Aral and Eckles (2019)

- What legal and ethical problems do they identify that would impede these types of analyses?
- How might these be resolved?
Enders et al (2022)

- Strong correlation between conspiracy theories and misinformation (CTM) and vaccine hesitancy
  - Assumption: exposure to CTM $\rightarrow$ vaccine hesitancy
  - But what if: vaccine hesitancy $\rightarrow$ exposure to CTM?
  - Or what if: political orientation/personality $\rightarrow$ vaccine hesitancy and political orientation/personality $\rightarrow$ exposure to CTM?
  - “Exogenous” vs. “endogenous”
  - Motivated reasoning: seek out information that is consistent with your own views
Enders et al (2022)

- ~2,000 people interviewed on Qualtrics in the U.S.
- Demonstrate the *plausibility* of model B
- Factor analysis (kind of like a principal components analysis) on the matrix of questions about vaccine hesitancy and belief in CTM.
  - Belief in CTM and vaccine hesitancy extremely highly correlated
  - Most exposure to CTM happening on social media (?)
  - But correlation between social media use and vaccine hesitancy not correlated
- Background traits predict both belief in CTM and vaccine hesitancy
- Open ended responses suggest vaccine hesitancy predates COVID CTM
Discussion

▪ What kind of research design would tell us whether exposure to CTM causes vaccine hesitancy?

▪ Where do deep-seated beliefs come from?

▪ Vaccine hesitancy has changed over time, how do we explain that without admitting that beliefs can change too?