Research Design, Causality, Data

Michael Ash

Lecture 1
Causality

Statistical correlation is not causality

Data

Research Design

Notation
Can we say anything about causality?

- Narrative
Can we say anything about causality?

- Narrative
- Appropriate counterfactual
Can we say anything about causality?

- Narrative
- Appropriate counterfactual
- Statistical correlation (accounting for multiple variables)
Can we say anything about causality?

- Narrative
- Appropriate counterfactual
- Statistical correlation (accounting for multiple variables)
- Reality → Model → Model → Reality
Do we always need causality?

- Confusing a cause with an effect, e.g., welfare causes poverty
Do we always need causality?

- Confusing a cause with an effect, e.g., welfare causes poverty
- Forecasting: why or why not?
Types of Variables

- categorical, discrete
  - nominal (unordered categories, including yes/no)
  - ordinal (ordered, e.g., Likert scale)

- measurement, continuous
  - interval (arbitrary scale without “zero”, e.g., SAT score)
  - ratio (meaningful “zero”, e.g., weight, income)
Organizing data

- Data, dataset, database
- Observation or case, row, record
- Variable, column, field
Organization of data

- Cross-Sectional Data
- Time Series Data
- Panel Data

In each case, what indexes the unit of observation?
Reminders

- Always keep track of the shape of the data
- Always tell your reader the **unit of observation**
Defining and understanding the counterfactual

- "Program Evaluation" Model
  - Treatment of interest is $T$.
    - Easy to think of as discrete, i.e., yes/no, but could be continuous (need more groups).
    - **A source of variation is needed**: some groups are treated, some are not.
    - Counterfactual is what would have happened in the absence of treatment.
  - Outcome $Y$ is agreed on, measurable.
1. Randomized controlled experiment (classic, ideal randomized design)

<table>
<thead>
<tr>
<th>Randomization</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>R</td>
<td>$\bar{Y}_{1,\text{before}}$</td>
<td>$\bar{Y}_{1,\text{after}}$</td>
</tr>
<tr>
<td>Group 2</td>
<td>R</td>
<td>$\bar{Y}_{2,\text{before}}$</td>
<td>$\bar{Y}_{2,\text{after}}$</td>
</tr>
</tbody>
</table>

Treatment has impact if $(\bar{Y}_{1,\text{after}} - \bar{Y}_{1,\text{before}}) > (\bar{Y}_{2,\text{after}} - \bar{Y}_{2,\text{before}})$

Effect is change in Group 1.
Counterfactual is change in Group 2.
2. Quasi-experimental designs

2a. Non-randomized comparison group

<table>
<thead>
<tr>
<th>Randomization</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>$\bar{Y}_{1,\text{before}}$</td>
<td>$T$</td>
<td>$\bar{Y}_{1,\text{after}}$</td>
</tr>
<tr>
<td>Group 2</td>
<td>$\bar{Y}_{2,\text{before}}$</td>
<td></td>
<td>$\bar{Y}_{2,\text{after}}$</td>
</tr>
</tbody>
</table>

Treatment has impact if

\[
(\bar{Y}_{1,\text{after}} - \bar{Y}_{1,\text{before}}) > (\bar{Y}_{2,\text{after}} - \bar{Y}_{2,\text{before}})
\]

Effect is change in Group 1.
Counterfactual is change in Group 2.
2b. No comparison group

<table>
<thead>
<tr>
<th>Randomization</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>$\bar{Y}_{1,\text{before}}$</td>
<td>$T$</td>
<td>$\bar{Y}_{1,\text{after}}$</td>
</tr>
</tbody>
</table>

Treatment has impact if $(\bar{Y}_{1,\text{after}} - \bar{Y}_{1,\text{before}}) > 0$

Effect is change in Group 1.

Counterfactual is level of Group 1 before treatment.
2c. Comparison group but only posttest

<table>
<thead>
<tr>
<th>Randomization</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td>T</td>
<td>$\bar{Y}_{1,\text{after}}$</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td>$\bar{Y}_{2,\text{after}}$</td>
</tr>
</tbody>
</table>

Treatment has impact if $\bar{Y}_{1,\text{after}} > \bar{Y}_{2,\text{after}}$

Effect is level of Group 1.
Counterfactual is level of Group 2.
2d. Only posttest

<table>
<thead>
<tr>
<th>Randomization</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td>$T$</td>
<td>$\bar{Y}_{1,\text{after}}$</td>
</tr>
</tbody>
</table>

Treatment has impact if $\bar{Y}_{1,\text{after}} > \text{something}$

Counterfactual is implicit.
Can you explain the shortcoming(s) of each of the non-ideal research designs?
What assumptions are necessary to make the non-ideal research designs deliver causal results?
Experimental (1) versus Observational (2a-2d)
In every case, there was *potentially* variation in the outcomes.
Organization of data, revisited

The organization of data makes the source of variation explicit. What questions can we answer?

- Cross-Sectional Data
- Time Series Data
- Panel Data
What is the shortcoming of the data for determining...

1. The effect of class size on student learning with detailed data on the socioeconomic background and academic performance of every student in an experimental small classroom?

2. The effect of cigarette taxes on smoking with data on cigarette sales from every county in New Jersey in 2001?

3. The demand for nurses in Massachusetts over the next decade with data on the number of nurses per patient in each State over the previous decade?
Notation: A shared responsibility

The summation sign, $\sum_{i=1}^{n}$, means “add up for each value from 1 to $n$”.

Interpret

\[
\sum_{i=1}^{n} x_i
\]

with the following data:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

In this case $n = 7$ (there are seven numbers), and so $i$ runs from 1 to 7. For example, $x_3 = 5$ (when $i$ is 3, the third value that $x$ takes on is 5. We can think of $x_i$ as some variable, say the number of hours worked on a project by student $i$. 

Michael Ash  
Research Design, Causality, Data
$n$ represents the total number of values to add up.

$i$ is the index that takes on the value 1, then 2, etc., up to $n$.

$$
\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n
$$

In this case, there are seven numbers to add.

$$
\sum_{i=1}^{7} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7
$$

Filling in each of the numbers,

$$
\sum_{i=1}^{7} x_i = 4 + 3 + 5 + 3 + 8 + 3 + 4 = 30
$$

which means that students worked a total of 30 hours on the project.
U.S. population in millions (Source: Census Bureau)

<table>
<thead>
<tr>
<th></th>
<th>1990</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>22.3</td>
<td>35.3</td>
</tr>
<tr>
<td>Total</td>
<td>248.7</td>
<td>281.4</td>
</tr>
</tbody>
</table>

What percent Hispanic was the United States in 1990?
What percent Hispanic was the United States in 2000?
What was the percent increase in the total population between 1990 and 2000?
What was the percent increase in the Hispanic population between 1990 and 2000?
What was the percentage-point increase in the Hispanic share of the population between 1990 and 2000?
What was the percent increase in the Hispanic share of the population between 1990 and 2000?
What percent Hispanic was the United States in 1990?

\[
\frac{22.3}{248.7} = 9.0 \text{ percent}
\]

What percent Hispanic was the United States in 2000?

\[
\frac{35.3}{281.4} = 12.5 \text{ percent}
\]

What was the percent increase in the total population between 1990 and 2000?

\[
\frac{281.4}{248.7} - 1 = 13 \text{ percent}
\]
What was the percent increase in the Hispanic population between 1990 and 2000?

$$\frac{35.3}{22.3} - 1 = 58 \text{ percent}$$

What was the percentage-point increase in the Hispanic share of the population between 1990 and 2000?

$$12.5 \text{ percent} - 9.0 \text{ percent} = 3.5 \text{ percentage points}$$

What was the percent increase in the Hispanic share of the population between 1990 and 2000?

$$\frac{12.5}{9.0} - 1 = 39 \text{ percent}$$
Using natural logarithm

Differences in natural logs (\(\ln(\cdot)\)) are good approximations of percent changes.

\[
\frac{281.4}{248.7} - 1 = 13.1 \text{ percent}
\]

\[
\ln(281.4) - \ln(248.7) = 5.640 - 5.516 = .124
\]
_propose a research question in policy and administration and discuss how you might answer it with data.\"