Causality and observational studies

Seung-Ho An, University of Arizona

Agenda

- What is causality
- Randomized experiments
- Calculating effects
- Observational studies

Causal: relationship between things where one causes the other

Inference: to derive as a conclusion from facts or premises

Causal inference is the attempt to derive causal connection based on the conditions of the occurrence of an effect

Think of the causal effect as the difference between what happened and what could have happened with/without a treatment (or change in X)

Are the variables causally related?

X -> Y

T -> Y

How do we know whether those variables are causally related?

Do they occur together?

If X goes up, Y goes up

If X happens, Y happens

If T, then change in Y

- If two things happen together a lot, we say they are correlated

Does correlation imply causation?

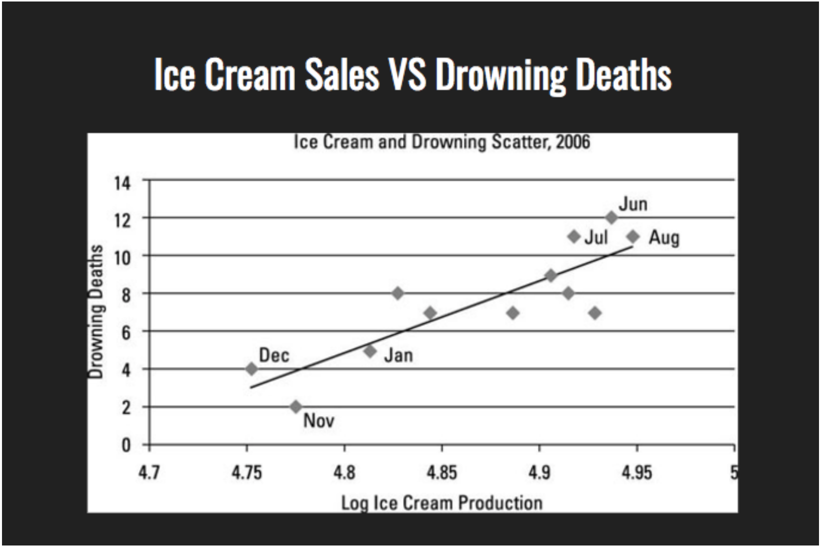

Figure 1: Ice cream sales and drowning incidents

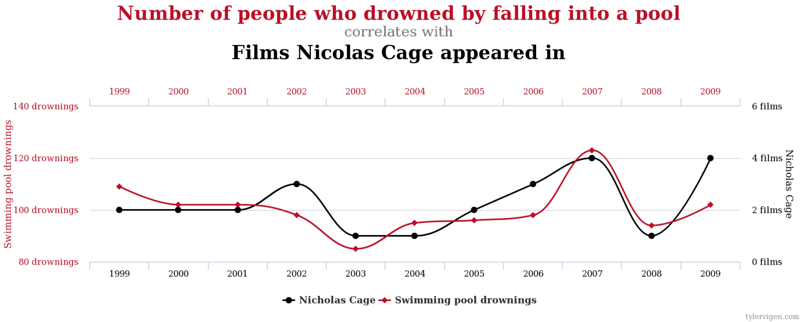

Figure 2: Cage films and pool incidents

Factual and counterfactual?

Factual: What happened

Counterfactual: What would have happened

Fundamental problem of causal inference: We can never observe counterfactuals, must be inferred

Fundamental problem of causal inference in a movie: Sliding Doors (1998) https://www.youtube.com/watch?v=BvUbv4iwbDs

Causal questions in social science research

Does the minimum wage increase the unemployment rate?

- Unemployment rate went up after the minimum wage increased
- Would it have gone up if the minimum wage increase not occurred?

Does race affect one's job prospect?

- Jamal applied for a job but did not get it
- Would Jamal have gotten a job if he were White?

Does having girls affect a judge's ruling in court?

- A judge with a daughter gave a pro-choice ruling
- Would she/he have done that if she/he had a son instead?

Application: Political canvassing study

- Research Q: Can canvassers change minds about optics like transgender rights?
- Experimental setting:
  - Randomly assign canvassers to have a conversation about transgender right or a conversation about recycling

  -Trans rights conversations focused on “perspective taking”
- **Outcome of interest**: support for trans rights policies

	Conversation script (T)	Support for Nondiscrimination law (Y)
Respondent 1	Recycling	No
Respondent 2	Trans rights	Yes

- Key interest: Did the second respondent support the law **because** of the perspective-taking conversation?

Useful to have **compact** notation for referring to **treatment variable**

$$T_i = \begin{cases} 1, & \text{if respondent } i \text{ had trans rights conversation} \\ 0, & \text{if respondent } i \text{ had recycling conversation} \end{cases}$$

Similar notation for the **outcome variable**:

$$Y_i = \begin{cases} 1, & \text{if respondent } i \text{ supports trans nondiscrimination laws} \\ 0, & \text{if respondent } i \text{ doesn't support nondiscrimination laws} \end{cases}$$

i is a placeholder to refer to a generic unit/respondent: Y_{12} is the outcome for the 12th unit

A tale of two respondents (redux)

	Conversation script (T)	Support for Nondiscrimination law (Y)
Respondent 1	Recycling	No
Respondent 2	Trans rights	Yes

becomes. . .

i	T_i	Y_i
Respondent 1	0	0
Respondent 2	1	1

What does "T_i causes Y_i" mean? $\rightsquigarrow$ **counterfactuals**, "what if"

Would respondent i change their support based on the conversation?

Two **potential outcomes**:

- $Y_i(1)$: would respondent i support ND laws if they had trans rights script?
- $Y_i(0)$: would respondent i support ND laws if they had recycling script??

Causal effect: $Y_i(1) - Y_i(0)$

- $Y_i(1) - Y_i(0) = 0 \rightsquigarrow$ script has no effect on policy views
- $Y_i(1) - Y_i(0) =$ -1 $\rightsquigarrow$ trans rights script lower support for laws
- $Y_i(1) - Y_i(0) = +1 \rightsquigarrow$ trans rights script increases support for laws

i	T_i	Y_i	$Y_i(1)$	$Y_i(0)$
Respondent 1	0	0	???	0
Respondent 2	1	1	1	???

- **Fundamental problem of causal inference:**
- We only observe one of the two potential outcomes
- Observe $Y_i(1) - Y_i(0) = 1$ or $Y_i(1) - Y_i(0) = 0$
- To infer causal effect, we need to infer the missing counterfactuals!

How can we figure out counterfactuals?

Find a similar unit! -> **matching** (Mill's method of difference)

Does respondent support law because of the trans rights script?

- $\rightsquigarrow$ find an identical respondent who got the recycling script

NJ increased the minimum wage. Causal effect on unemployment?

- $\rightsquigarrow$ find a state similar to NJ that didn't increase minimum wage

The problem: we cannot match on everything (imperfect matches)

- Say we match i (treated) and j (control)
- $Y_i(1) \neq Y_j(1)$ due to **selection bias**
- Those who take treatment may be different that those who take control
- How can we correct for that? **Randomization!**

2. Randomized experiments

Randomized control trial: each unit's treatment assignment is determined by chance

- Flip a coin; draw red and blue chips from a hat; etc

Randomization ensures **balance** between treatment and control groups

- Treatment and control group are identical **on average**
- Similar on both observable and unobservable characteristics

We will often refer to the **sample size** (number of unites) as `n`.

We often have `n` measurements of some variable, $(Y_1, Y_2, ..., Y_n)$

How many in our sample support nondiscrimination laws?

$$Y_1 + Y_2 + Y_3 + ... + Y_n$$

Notation is a bit chunky so we often use the **Sigma notation**:

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + Y_3 + ... + Y_n$$

$\sum_{i=1}^{n} Y_i$ simply means sum each value from Y_1 to Y_n

The **sample average** or sample mean is simply the sum of all values divided by the number of values

Sigma notation allows us to write this in a compact way:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Suppose we surveyed 6 people and 3 supported gay marriage:

$$\overline{Y} = \frac{(1 + 1 + 1 + 0 + 0 + 0)}{6} = 0.5$$

Quantity of interest

We want to estimate the average causal effects over all units:

$$\text{Sample Average Treatment Effect (SATE)} = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$$
$$= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$$

Why can't we just calculate this quantity directly?

What we can estimate instead:

Difference in means$=\overline{Y}_{treated} - \overline{Y}_{control}$

- $\overline{Y}_{treated}$: observed average outcome for treated group
- $\overline{Y}_{control}$: observed average outcome for control group

When will the difference-in-means is a good estimate of the SATE?

- Under an RCT, treatment and control groups are random samples
- Average in the treatment group will be similar to average if all **treated**:

$$\overline{Y}_{treated} \approx \frac{1}{n} \sum_{i=1}^{n} Y_i(1)$$

- Average in the control group will be similar to average if all **untreated**:

$$\overline{Y}_{control} \approx \frac{1}{n} \sum_{i=1}^{n} Y_i(0)$$

- Implies difference-in-means should be close to SATE:

$$\overline{Y}_{treated} - \overline{Y}_{control} \approx \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - \frac{1}{n} \sum_{i=1}^{n} Y_i(0) = \frac{1}{n} \sum_{i=1}^{n} [Y_i(1) - Y_i(0)] = \text{SATE}$$

Some potential problems with RCTs

Placebo effects:

- Respondents will be affected by any intervention, even if they shouldn't have any effectively

Hawthorne effects:

- Respondents act differently just knowing that they are under study

Can we determine if randomization "worked"?

If it did, we shouldn't see large differences between treatment and control group on **pretreatment variables**

- Pretreatment variables are those that are unaffected by treatment

We can check in the actual data for some pretreatment variable `X`

- $\overline{X}_{treated}$: average value of variable for treated group
- $\overline{X}_{control}$: average value of variable for control group
- Under randomization, $\overline{X}_{treated} - \overline{X}_{control} \approx 0$

An example of balance checking

Q. Would taking a year long stat class increase the likelihood of being promoted?

Before discussing the balance test, what would be the outcome and treatment in the question?

Outcome: individual officer's likelihood of being promoted

Imagining that we are designing a randomized field experiment to answer this question, we can randomly assign TPD officers into treatment and control groups

Treatment: officers who are randomly assigned to take a year long statistic course

Control group: officers who are randomly assigned and not taking a year-long stat course

An example of balance checking (cont)

Balance checking means the differences in pretretment variables should not exist among officers between the two groups on average

What would be examples of pretreatment variables?

These could be gender, age, race, tenure, income, marital status, education, or any other characteristics

- Instead of 1 treatment, we might have multiple **treatment arms**:
  - Control condition
  - Treatment A
  - Treatment B
  - Treatment C, etc
- In this case, we will look at multiple comparisons:
  - $\overline{Y}_{treated,A} - \overline{Y}_{control}$
  - $\overline{Y}_{treated,B} - \overline{Y}_{control}$
  - $\overline{Y}_{treated,A} - \overline{Y}_{treated,B}$
  - $\overline{Y}_{treated,A} - \overline{Y}_{treated,C}$
  - $\overline{Y}_{treated,B} - \overline{Y}_{treated,C}$, etc
- If treatment arms are randomly assigned, these differences will be good estimators for each causal contrast

- Blacklist Season 6 Episode 20 (19:39)

3. Calculating effects

Now we `RetuRn` to `R` to illustrate how to calculate the effects!

```
## reinstall TPDdata if necessary
library(TPDdata)
```

Variable	Description
age	Age of the Respondent (R) in years
female	1=R marked "Female" on voter reg., 0 otherwise
voted_gen_14	1 if R voted in the 2014 general election
voted_gen_12	1 if R voted in the 2012 general election
treat_ind	1 if R assigned to trans rights script, 0 for recycling
racename	name of racial identity indicated on voter file
democrat	1 if R is a registered Democrat
nondiscrim_pre	1 if R supports nondiscrim. law at baseline
nondiscrim_post	1 if R supports nondiscrim. law after 3 months

```
trans
```

```
## # A tibble: 565 x 9
##      age female voted_gen_14 voted_gen~1 treat~2 racen~3 democ~4 nondi~5 nondi~6
##    <dbl>  <dbl>        <dbl>       <dbl>   <dbl> <chr>     <dbl>   <dbl>   <dbl>
##  1    29      0            0           1       0 Africa~       1       1       1
##  2    59      1            1           0       1 Africa~       1       1       1
##  3    35      1            1           1       1 Africa~       1       0       1
##  4    63      1            1           1       1 Africa~       1       0       0
##  5    65      0            1           1       1 Africa~       0       1       0
##  6    51      1            1           1       0 Caucas~       0       1       1
##  7    26      1            1           1       0 Africa~       1       1       0
##  8    62      1            1           1       1 Africa~       1       1       1
##  9    37      0            1           1       0 Caucas~       0       1       0
## 10    51      1            1           1       0 Caucas~       0       0       0
## # ... with 555 more rows, and abbreviated variable names 1: voted_gen_12,
## #   2: treat_ind, 3: racename, 4: democrat, 5: nondiscrim_pre,
## #   6: nondiscrim_post
```

Calculate the average outcomes in each group

```r
treat_mean <- trans |>
  filter(treat_ind == 1) |>
  summarize(nondiscrim_mean = mean(nondiscrim_post))
treat_mean
```

```
## # A tibble: 1 x 1
##   nondiscrim_mean
##             <dbl>
## 1           0.687
```

```r
control_mean <- trans |>
  filter(treat_ind == 0) |>
  summarize(nondiscrim_mean = mean(nondiscrim_post))
control_mean
```

```
## # A tibble: 1 x 1
##   nondiscrim_mean
##             <dbl>
## 1           0.648
```

```r
treat_mean - control_mean
```

```
##   nondiscrim_mean
## 1      0.03896674
```

We will see more ways to do this throughout the semester

Checking balance on numeric covariates

We can use group_by to see how the mean of covariates varies by group:

```
trans |>
  group_by(treat_ind) |>
  summarize(age_mean = mean(age))
```

```
## # A tibble: 2 x 2
##   treat_ind age_mean
##       <dbl>    <dbl>
## 1         0     48.2
## 2         1     48.3
```

Checking balance on categorical covariates

Or we can group by treatment and a categorical control:

```
trans |>
  group_by(treat_ind, racename) |>
  summarize(n = n())
```

```
## `summarise()` has grouped output by 'treat_ind'. You can override using the
## `.groups` argument.

## # A tibble: 9 x 3
## # Groups:   treat_ind [2]
##   treat_ind racename             n
##       <dbl> <chr>            <int>
## 1         0 African American    58
## 2         0 Asian                2
## 3         0 Caucasian           77
## 4         0 Hispanic           150
## 5         1 African American    68
## 6         1 Asian                4
## 7         1 Caucasian           75
## 8         1 Hispanic           130
## 9         1 Native American      1
```

Hard to read!

pivot_wider() takes data from a single column and moves it into multiple columns based on a grouping variable:

```
trans |>
  group_by(treat_ind, racename) |>
  summarize(n = n()) |>
  pivot_wider(
    names_from = treat_ind,
    values_from = n
  )
```

names_from tells us what variable will map onto the columns

values_from tells us what values should be mapped into those columns

```r
trans |>
  group_by(treat_ind, racename) |>
  summarize(n = n()) |>
  pivot_wider(
    names_from = treat_ind,
    values_from = n
  )
```

```
## `summarise()` has grouped output by 'treat_ind'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 5 x 3
##   racename           `0`   `1`
##   <chr>            <int> <int>
## 1 African American    58    68
## 2 Asian                2     4
## 3 Caucasian           77    75
## 4 Hispanic           150   130
## 5 Native American     NA     1
```

Calculating diff-in-means by group

```
trans |>
  mutate(
    treat_ind = if_else(treat_ind == 1, "Treated", "Control"),
    party = if_else(democrat == 1, "Democrat", "Non-Democrat")
  ) |>
  group_by(treat_ind, party) |>
  summarize(nondiscrim_mean = mean(nondiscrim_post)) |>
  pivot_wider(
    names_from = treat_ind,
    values_from = nondiscrim_mean
  ) |>
  mutate(
    diff_in_means = Treated - Control
  )
```

```
## `summarise()` has grouped output by 'treat_ind'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 2 x 4
##   party        Control Treated diff_in_means
##   <chr>          <dbl>   <dbl>         <dbl>
## 1 Democrat       0.704   0.754        0.0498
## 2 Non-Democrat   0.605   0.628        0.0234
```

4. Observational Studies

- Can newspaper endorsements change voters' minds?

Why not compare vote choice of readers of different papers?

- Problem: readers choose papers based on their previous beliefs
- Liberals $\approx$ New York Times, conservatives $\approx$ Wall Street Journal

Our case: British newspapers switching their endorsements

- Some news papers endorsing Tories in 1992 switched to Labour in 1997
- **Treated group:**: readers of Tory -> Labour papers
- **Control group:**: readers of papers who didn't switch

Variable	Description
to_labour	Read a newspaper that switched endorsement to Labour between 1992 and 1997 (1=yes, 0=no)
vote_lab_92	Did respondent vote for Labour in 1992 election (1=yes, 0=no)?
vote_lab_97	Did respondent vote for Labour in 1997 election (1=yes, 0=no)?
age	Age of respondent
male	Does the respondent identify as Male (1=yes, 0=no)
parent_labour	Does the respondent' identify as 's parents vote for Labour (1=yes, 0=no)
work_class	Does the responedent identify as working class (1=yes, 0=no)?

```
library(tidyverse)
library(TPDdata)
newspapers
```

```
## # A tibble: 1,593 x 7
##    to_labour vote_lab_92 vote_lab_97        age  male parent_labour work_class
##        <dbl>       <dbl>       <dbl> <hvn_lbll> <dbl>         <dbl>      <dbl>
##  1         0           1           1         33     0             1          1
##  2         0           1           0         51     0             1          0
##  3         0           0           0         46     0             1          1
##  4         0           1           1         45     1             1          1
##  5         0           1           1         29     0             1          1
##  6         0           1           1         47     1             1          1
##  7         0           1           1         34     1             0          1
##  8         0           1           1         31     0             1          1
##  9         0           1           1         24     1             1          1
## 10         1           1           1         48     0             1          1
## # ... with 1,583 more rows
```

Example of an **observational study**:

- We as researchers observe a naturally assigned treatment
- Very common: often can't randomize for ethical/logistical reasons

Internal validity: are the causal assumption satisfied? Can we interpret this as a causal effect?

- RCTs usually have higher internal validity
- Observational studies less so, because treatment and control groups may differ in ways that are hard to measure

External validity: can the conclusions/estimated effects be generalized beyond this study?

- RCTs weaker here because often very expensive to conduct on representative
- Observational studies often have larger/more representative samples that improve external validity

Confounder: pre-treatment variable affecting both treatment & the outcome

- Leftists (X) more likely to read newspapers switching to Labour (T)
- Leftists (X) also more likely to vote for Labour (Y)

Counfounding bias in the estimated SATE due to these differences

- $\overline{Y}_{control}$ not a good proxy for $Y_i(0)$ in treated group
- one type: **selection bias** from self-selection into treatment

How can we find a good comparison group?

Depends on the data we have available

Three general types of observational study **research designs**:

1. **Cross-sectional design**: compare outcomes treated and control unites at one point in time
2. **Before-and-after design**: compare outcomes before and after a unit has been treated, but need over-time data on treated group
3. **Difference-in-differences design**: use before/after information for the treated and control group; need over-time on treated & control group

Compare treatment and control groups after treatment happens

- Readers of switching papers vs. readers of non-switching papers in 1997

Treatment & control groups assumed identical on average as in RCT

- Sometimes called **unconfoundedness** or **as-if randomized**

Cross-section comparison estimate:

$$\overline{Y}^{after}_{treated} - \overline{Y}^{after}_{control}$$

Could there be counfounders?

```
switched <- newspapers |>
  filter(to_labour ==1) |>
  summarize(mean(vote_lab_97))

no_change <- newspapers |>
  filter(to_labour == 0) |>
  summarize(mean(vote_lab_97))

switched - no_change
```

```
##   mean(vote_lab_97)
## 1         0.1404826
```

Statistical control: adjust for confounders using statistical procedures

- Can help to reduce confounding bias

One type of statistical control: **subclassfication**

- Compare treated and control groups within levels of a confounder
- Remaining effect can't be due to the confounder

Treat to inference: we can only control for observed variables $\approx$ threat of **unmeasured confounding**

```
newspapers |>
  group_by(parent_labour, to_labour) |>
  summarize(avg_vote = mean(vote_lab_97)) |>
  pivot_wider(
    names_from = to_labour,
    values_from = avg_vote
  ) |>
  mutate(diff_by_parent = `1` - `0`)
```

```
## `summarise()` has grouped output by 'parent_labour'. You
## `.groups` argument.
```

```
## # A tibble: 2 x 4
## # Groups:   parent_labour [2]
##   parent_labour   `0`   `1` diff_by_parent
##           <dbl> <dbl> <dbl>          <dbl>
## 1             0 0.279 0.434          0.155
## 2             1 0.597 0.698          0.101
```

Compare readers of party-switching newspapers before & after switch

Advantage: all person-specific features held fixed

- comparing within a person over time

Before-and-after estimate:

$$\overline{Y}^{after}_{treated} - \overline{Y}^{before}_{treated}$$

Threat to inference: **time-varying counfounders**

- time trend: Labour just did better overall in 1997 compared to 1992

```
newspapers |>
  mutate(
    vote_change = vote_lab_97 - vote_lab_92
  ) |>
  summarize(avg_change = mean(vote_change))
```

```
## # A tibble: 1 x 1
##   avg_change
##        <dbl>
## 1      0.119
```

Key idea: use the before-and-after difference of **control group** to infer what would have happened to **treatment group** without treatment

DiD estimate:

$$\underbrace{(\overline{Y}^{after}_{treated} - \overline{Y}^{before}_{treated})}_{\text{trend in treated group}} - \underbrace{(\overline{Y}^{after}_{control} - \overline{Y}^{before}_{control})}_{\text{trend in control group}}$$

Change in treated group above and beyond the change in control group

Parallel time trend assumption

- Changes in vote of readers of non-switching papers roughly the same as changes that readers of switching papers would have been if they read non-switching papers
- Threat to inference: non-parallel trends

```r
newspapers |>
  mutate(
    vote_change = vote_lab_97 - vote_lab_92,
    to_labour = if_else(to_labour ==1, "switched", "unswitched")
  ) |>
  group_by(to_labour) |>
  summarize(avg_change = mean(vote_change)) |>
  pivot_wider(
    names_from = to_labour,
    values_from = avg_change
  ) |>
  mutate(DID = switched - unswitched)
```

```
## # A tibble: 1 x 3
##   switched unswitched    DID
##      <dbl>      <dbl>  <dbl>
## 1    0.190      0.110 0.0796
```

Summarizing approaches

1. **Cross-sectional comparison**
    - Compare treated units with control unites after treatment
    - Assumption: treated and control units are comparable
    - Possible confounding
2. **Before-and-after comparison**
    - Compare the same units before and after treatment
    - Assumption: no time-varying confounding
3. **Differences-in-differences**
    - Assumption: parallel trends assumptions
    - Under this assumption, it accounts for unit-specific and time-varying confounding

All rely on assumptions that can't be verified to handle confounding

RCTs handle confounding by design (gold standard)