Causality, 2

Frank Edwards

9/28/2021

Returning to Pager’s experiment

The causal effect

For observation *i* is equal to callback_crimTRUE_i - callback_crimFALSE_i

The fundamental problem of causal inference is that we only observe one of these outcomes

The counterfactual and potential outcomes

```
##   callback crimrec callback_crimT callback_crimF
## 1        1       1              1             NA
## 2        0       0             NA              0
## 3        1       0             NA              1
## 4        1       0             NA              1
## 5        0       1              0             NA
## 6        0       1              0             NA
```

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent
- If units are equivalent, we can estimate the average treatment effect as a difference in means on the outcome between the treatment and control group

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent
- If units are equivalent, we can estimate the average treatment effect as a difference in means on the outcome between the treatment and control group
- If we don't randomize, we have no assurance that the treated and control groups are equivalent, meaning we don't have a strong case that we've observed the counterfactual

Obtaining a sample average treatment effect

The sample average treatment effect is defined as:

$$\mathrm{SATE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

Obtaining a sample average treatment effect

The sample average treatment effect is defined as:

$$\mathrm{SATE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

In practice, since we only observe $Y_i(1)$ OR $Y_i(0)$, we instead estimate a *difference-in-means* of the outcome between the treatment and control: $mean(Y(1)) - mean(Y(0))$. If assignment has been randomized, these values are identical.

Why we randomize

An experiment on voting and a social pressure

Civic duty: The whole point of democracy is that citizens are active participants in government; that we have a voice in government. Your voice starts with your vote. On August 8, remember your rights and responsibilities as a citizen. Remember to vote. DO YOUR CIVIC DUTY – VOTE

An experiment on voting and a social pressure

Civic duty: The whole point of democracy is that citizens are active participants in government; that we have a voice in government. Your voice starts with your vote. On August 8, remember your rights and responsibilities as a citizen. Remember to vote. DO YOUR CIVIC DUTY – VOTE

Hawthorne effect (surveillance): This year, we're trying to figure out why people do or do not vote. We'll be studying voter turnout in the August 8 primary election. Our analysis will be based on public records, so you will not be contacted again or disturbed in any way. Anything we learn about your voting or not voting will remain confidential and will not be disclosed to anyone else. DO YOUR CIVIC DUTY – VOTE

An experiment on voting and a social pressure

```r
social <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/CAUSALITY/social.csv")
head(social)
```

```
## # A tibble: 6 x 6
##   sex    yearofbirth primary2004 messages   primary2006 hhsize
##   <chr>        <dbl>       <dbl> <chr>            <dbl>  <dbl>
## 1 male          1941           0 Civic Duty           0      2
## 2 female        1947           0 Civic Duty           0      2
## 3 male          1951           0 Hawthorne            1      3
## 4 female        1950           0 Hawthorne            1      3
## 5 female        1982           0 Hawthorne            1      3
## 6 male          1981           0 Control              0      3
```

Obtaining mean voting by treatment/control

```r
control <- social %>%
  filter(messages == "Control") %>%
  summarise(primary2006 = mean(primary2006))

treatment <- social %>%
  filter(messages != "Control") %>%
  summarise(primary2006 = mean(primary2006))

control
```

```
## # A tibble: 1 x 1
##   primary2006
##         <dbl>
## 1       0.297
```

```r
treatment
```

```
## # A tibble: 1 x 1
##   primary2006
##         <dbl>
## 1       0.338
```

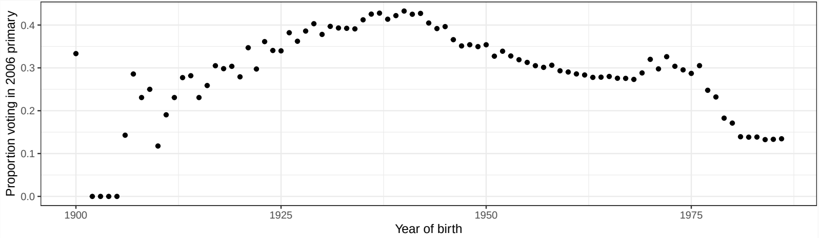
The difference in means (causal effect)

```r
effect <- treatment - control

effect
```

```
##   primary2006
## 1  0.04164458
```

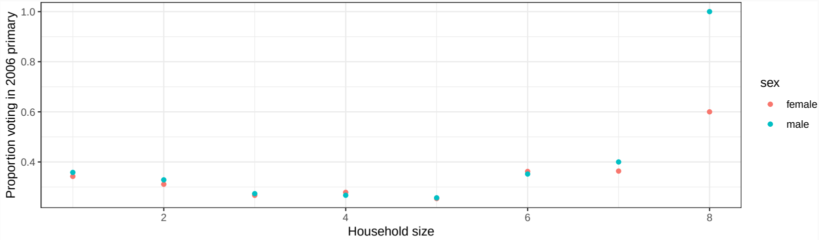
Why randomization matters

```
prop_vot<-social %>%
  group_by(yearofbirth) %>%
  summarise(voting_prop = sum(primary2006)/n())

ggplot(prop_vot,
       aes(x = yearofbirth, y = voting_prop)) +
  geom_point() +
  ylab("Proportion voting in 2006 primary") +
  xlab("Year of birth")
```



Why randomization matters (continued)

```
sex_hh<-social %>%
  group_by(sex, hhsize) %>%
  summarise(voting_prop = sum(primary2006)/n())

ggplot(sex_hh,
       aes(x = hhsize, y = voting_prop, color = sex)) +
  geom_point() +
  xlab("Household size") +
  ylab("Proportion voting in 2006 primary")
```



Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others

Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others
- We note these differences between observed variables and our outcome: primary2006

Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others
- We note these differences between observed variables and our outcome: primary2006
- We didn't measure very much here. They could also differ across unobserved or unobservable variables!

Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others
- We note these differences between observed variables and our outcome: primary2006
- We didn't measure very much here. They could also differ across unobserved or unobservable variables!
- Randomization (given a large enough n) ensures that treatment and control groups are *identical* across all observed and unobserved/unobservable differences prior to treatment

Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others
- We note these differences between observed variables and our outcome: primary2006
- We didn't measure very much here. They could also differ across unobserved or unobservable variables!
- Randomization (given a large enough n) ensures that treatment and control groups are *identical* across all observed and unobserved/unobservable differences prior to treatment
- This condition – statistically identical treatment and control groups – is a necessary condition for causal inference. Randomization is the most straightforward way to achieve this condition.

Causal inference in observational data

Estimating the impact of a minimum wage increase

In 1992, New Jersey raised it's minimum wage from \$4.25 to \$5.05.

Pennsylvania did not.

```
minwage <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/CAUSALITY/minwage.csv")
head(minwage)
```

```
## # A tibble: 6 x 8
##   chain      location wageBefore wageAfter fullBefore fullAfter partBefore partAfter
##   <chr>      <chr>         <dbl>     <dbl>      <dbl>     <dbl>      <dbl>     <dbl>
## 1 wendys     PA             5         5.25         20         0         20        36
## 2 wendys     PA             5.5       4.75          6        28         26         3
## 3 burgerking PA             5         4.75         50        15         35        18
## 4 burgerking PA             5         5            10        26         17         9
## 5 kfc        PA             5.25      5             2         3          8        12
## 6 kfc        PA             5         5             2         2         10         9
```

Describing the data, categoricals

```
table(minwage$chain)
```

```
##
## burgerking        kfc       roys     wendys
##        149         75         88         46
```

```
table(minwage$location)
```

```
##
## centralNJ   northNJ        PA   shoreNJ   southNJ
##        45       146        67        33        67
```

Did NJ minimum wage increase the wages paid to employees?

```r
minwage %>%
  group_by(location) %>%
  summarise(wageBefore_mn = mean(wageBefore),
            wageAfter_mn = mean(wageAfter))
```

```
## # A tibble: 5 x 3
##   location  wageBefore_mn wageAfter_mn
##   <chr>             <dbl>        <dbl>
## 1 centralNJ          4.63         5.09
## 2 northNJ            4.63         5.09
## 3 PA                 4.65         4.61
## 4 shoreNJ            4.64         5.07
## 5 southNJ            4.54         5.06
```

Another way to look at change in wages

```r
minwage %>%
  group_by(location) %>%
  summarise(prop_below_before = mean(wageBefore>=5.05),
            prop_below_after = mean(wageAfter>=5.05))
```

```
## # A tibble: 5 x 3
##   location  prop_below_before prop_below_after
##   <chr>                 <dbl>            <dbl>
## 1 centralNJ            0.133            0.978
## 2 northNJ              0.0753           1
## 3 PA                   0.0597           0.0448
## 4 shoreNJ              0.121            1
## 5 southNJ              0.0746           1
```

Look at our outcome variable

```r
###Compute proportion full time before
###And after

minwage<- minwage %>%
  mutate(prop_ft_pre =
           fullBefore /
           (fullBefore + partBefore))

minwage <- minwage %>%
  mutate(prop_ft_post =
           fullAfter /
           (fullAfter + partAfter))
```

Look at our outcome variable

```r
minwage %>%
  group_by(location) %>%
  summarise(prop_ft_pre = mean(prop_ft_pre),
            prop_ft_post = mean(prop_ft_post))
```

```
## # A tibble: 5 x 3
##   location  prop_ft_pre prop_ft_post
##   <chr>           <dbl>        <dbl>
## 1 centralNJ       0.311        0.251
## 2 northNJ         0.321        0.375
## 3 PA              0.310        0.272
## 4 shoreNJ         0.286        0.345
## 5 southNJ         0.239        0.236
```

Assumption: PA is a no-treatment counterfactual

Estimate the causal effect

```
control <- minwage %>%
  filter(location == "PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post))

treatment <- minwage %>%
  filter(location != "PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post))

treatment - control
```

```
##   prop_ft_post
## 1   0.04811886
```

Is this a valid estimate of the causal effect?

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome
- Selection bias: a unit may choose to participate in a treatment for reasons that are correlated with the outcome

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome
- Selection bias: a unit may choose to participate in a treatment for reasons that are correlated with the outcome

Correlation != Causation

Dealing with confounding

- Randomize treatment!

Dealing with confounding

- Randomize treatment!
- When we can't...

Dealing with confounding

- Randomize treatment!
- When we can’t...
- Statistical control: within-subgroup analysis based on confounder values

Are NJ and PA the same (at least when it comes to fast food jobs?)?

```
minwage %>%
  group_by(location) %>%
  summarise(prop_wendys = mean(chain=="wendys"),
            prop_bk = mean(chain=="burgerking"),
            prop_kfc = mean(chain=="kfc"),
            prop_roys = mean(chain=="roys"))
```

```
## # A tibble: 5 x 5
##   location  prop_wendys prop_bk prop_kfc prop_roys
##   <chr>           <dbl>   <dbl>    <dbl>     <dbl>
## 1 centralNJ      0.0889   0.378    0.244     0.289
## 2 northNJ        0.130    0.459    0.158     0.253
## 3 PA             0.164    0.463    0.149     0.224
## 4 shoreNJ        0.152    0.364    0.303     0.182
## 5 southNJ        0.104    0.328    0.313     0.254
```

Maybe restaurant chain matters? Let’s control for it!

```
control<-minwage %>%
  filter(location==”PA”) %>%
  group_by(chain) %>%
  summarise(prop_ft_post = mean(prop_ft_post))
```

Maybe restaurant chain matters? Let's control for it!

```
treatment<-minwage %>%
  filter(location!="PA") %>%
  group_by(chain) %>%
  summarise(prop_ft_post = mean(prop_ft_post))
```

Maybe restaurant chain matters? Let's control for it!

```
treatment$effect<-treatment$prop_ft_post -
  control$prop_ft_post

treatment
```

```
## # A tibble: 4 x 3
##   chain      prop_ft_post effect
##   <chr>             <dbl>  <dbl>
## 1 burgerking        0.358 0.0364
## 2 kfc               0.328 0.0918
## 3 roys              0.283 0.0697
## 4 wendys            0.260 0.0117
```

Maybe region matters: central and south vs north and shore

```
control<-minwage %>%
  filter(location=="PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post))

treatment<-minwage %>%
  filter(location!="PA") %>%
  group_by(location) %>%
  summarise(prop_ft_post = mean(prop_ft_post))

control
```

```
## # A tibble: 1 x 1
##   prop_ft_post
##          <dbl>
## 1        0.272
```

```
treatment
```

```
## # A tibble: 4 x 2
##   location  prop_ft_post
##   <chr>            <dbl>
## 1 centralNJ        0.251
## 2 northNJ          0.375
## 3 shoreNJ          0.345
## 4 southNJ          0.236
```

Maybe region matters?

```
treatment$effect<-treatment$prop_ft_post -
  control$prop_ft_post

treatment
```

```
## # A tibble: 4 x 3
##   location  prop_ft_post  effect
##   <chr>            <dbl>   <dbl>
## 1 centralNJ        0.251 -0.0210
## 2 northNJ          0.375  0.103
## 3 shoreNJ          0.345  0.0728
## 4 southNJ          0.236 -0.0366
```

Cross-sections and time series

- Longitudinal data: repeated measurements of the same unit on the same variables over time

Cross-sections and time series

- Longitudinal data: repeated measurements of the same unit on the same variables over time
- Cross-sectional data: one measurement of many units

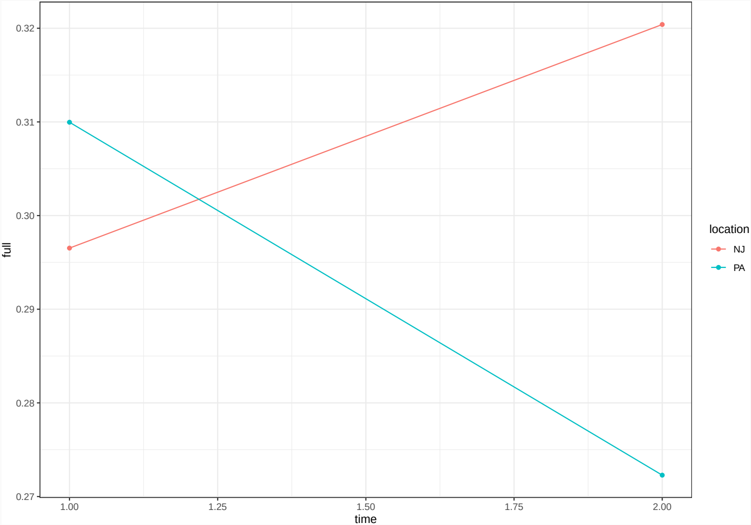
Cross-sections and time series

- Longitudinal data: repeated measurements of the same unit on the same variables over time
- Cross-sectional data: one measurement of many units
- Panel data (or time series cross-sectional data): repeated measurements of many units on the same variables over time

Cross-sections and time series

- Longitudinal data: repeated measurements of the same unit on the same variables over time
- Cross-sectional data: one measurement of many units
- Panel data (or time series cross-sectional data): repeated measurements of many units on the same variables over time
- Key advantages to panel data: variables may differ across units and within-units over time (trends).

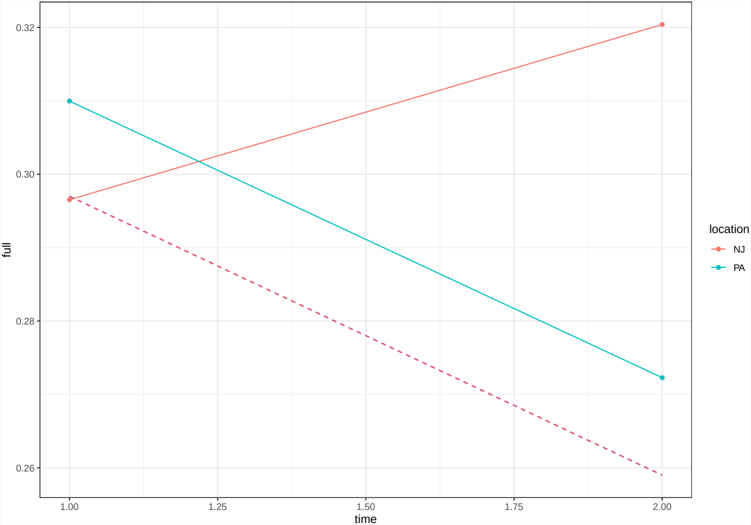
Before and after design (longitudinal)



- What if we treated PA as the counterfactual, and used information about it’s trend in employment to estimate the effect of NJ’s minimum wage increase?

- What if we treated PA as the counterfactual, and used information about it's trend in employment to estimate the effect of NJ's minimum wage increase?
- Assumption: The trend in the outcome over time would have been identical across all units if the treatment had never been imposed (parallel trends)

Difference in Differences (visual)



Estimating the causal effect: Differenc in Differences

Where y_{ij} is the outcome for treatment group $i = 1$ and post-treatment time $j = 1$

$$\mathrm{DiD} = (\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{2,1} - \bar{y}_{2,0})$$

Assuming that the counterfactual outcome for the treatment group has a parallel time trend to that observed for the control group.

Compute the DiD estimator

```
## # A tibble: 2 x 3
##   location prop_ft_pre prop_ft_post
##   <chr>          <dbl>        <dbl>
## 1 NJ             0.297        0.320
## 2 PA             0.310        0.272
```

Compute the DiD estimator

```
## # A tibble: 2 x 3
##   location prop_ft_pre prop_ft_post
##   <chr>          <dbl>        <dbl>
## 1 NJ             0.297        0.320
## 2 PA             0.310        0.272
```

$$\mathrm{DiD} = (\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{2,1} - \bar{y}_{2,0})$$

Compute the DiD estimator

```
## # A tibble: 2 x 3
##   location prop_ft_pre prop_ft_post
##   <chr>          <dbl>        <dbl>
## 1 NJ             0.297        0.320
## 2 PA             0.310        0.272
```

$$\text{DiD} = (\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{2,1} - \bar{y}_{2,0})$$

```
### the DiD Estimator
(0.320 - 0.297) - (0.272 - 0.310)
```

```
## [1] 0.061
```

Descriptive Statistics

Summarizing a variable

Reduce a vector to a single or smaller set of values that tell us something useful

Examples we've already used: - minimum: min() - maximum: max() - median: median() - mean: mean()

Quantiles

- The median is the 0.5 quantile (50th percentile)
- Quantiles are less sensitive to outliers than are other measures (like the mean)
- Quantiles tell you the proportion of a data that falls below some cutpoint

Quantiles: example

```r
quantile(minwage$wageBefore, 0.25)
```

```
##  25%
## 4.25
```

Quantiles: example

```
quantile(minwage$wageBefore, 0.75)
```

```
##    75%
## 4.9875
```

Quantiles: example

```r
quantile(minwage$wageBefore, c(0.05, 0.25, 0.5, .75, 0.95))
```

```
##      5%    25%    50%    75%    95%
## 4.2500 4.2500 4.5000 4.9875 5.2500
```

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable
- It provides a measure of how much each observation of a variable differs from the mean of the variable

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable
- It provides a measure of how much each observation of a variable differs from the mean of the variable
- You can use the sd() function in R

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable
- It provides a measure of how much each observation of a variable differs from the mean of the variable
- You can use the sd() function in R
- The variance (var() function) is the square of the standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable
- It provides a measure of how much each observation of a variable differs from the mean of the variable
- You can use the sd() function in R
- The variance (var() function) is the square of the standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$variance = \sigma^2$$

Compute an SD for these variables

```r
minwage$wageBefore[1:10]
```

```
##  [1] 5.00 5.50 5.00 5.00 5.25 5.00 5.00 5.00 5.00 5.50
```

```r
minwage$fullBefore[1:10]
```

```
##  [1] 20.0  6.0 50.0 10.0  2.0  2.0  2.5 40.0  8.0 10.5
```

Standard deviations and meaningful differences

How meaningful is a ten point difference on a test?

```
### draw 50 random scores from a test with a minimum of zero and maximum of 100
testA<-runif(50, 0, 100)
### draw 50 random scores from a test with a minimum of zero and maximum of 1000
testB<-runif(50, 0, 1000)
```

Visualizing the distributions

The mean

A 10 point jump from the mean

A 1 SD jump from the mean

Homework

- HW4 posted to Slack
- make sure to use `na.rm = TRUE` for mean(), quantile() and other functions
- `group_by()` and `summarize()` are very helpful on this one