350B Political Methodology II, Winter 2016

Randomized Experiments

Jonathan Mummolo

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
    - Average Treatment Effect
    - Standard Error for ATE
    - Hypothesis Testing
    - Covariates
    - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
    - Blocking
    - Power Calculations
    - Threats to Validity and Ethics

Selection Bias

Recall the selection problem when comparing the mean outcomes for the treated and the untreated:

Problem

$$\underbrace{\mathbf{E}[Y|D=1]-\mathbf{E}[Y|D=0]}_{\textit{Difference in Means}} = \mathbf{E}[Y_1|D=1]-\mathbf{E}[Y_0|D=0]$$

$$= \underbrace{\mathbf{E}[Y_1 - Y_0|D=1]}_{\textit{ATT}} + \underbrace{\{\mathbf{E}[Y_0|D=1]-\mathbf{E}[Y_0|D=0]\}}_{\textit{BIAS}}$$

How can we eliminate the bias term?

Selection Bias

Recall the selection problem when comparing the mean outcomes for the treated and the untreated:

Problem

$$\underbrace{\mathbf{E}[Y|D=1]-\mathbf{E}[Y|D=0]}_{\textit{Difference in Means}}=\mathbf{E}[Y_1|D=1]-\mathbf{E}[Y_0|D=0]$$

$$=\underbrace{\mathbf{E}[Y_1-Y_0|D=1]}_{\textit{ATT}}+\underbrace{\{\mathbf{E}[Y_0|D=1]-\mathbf{E}[Y_0|D=0]\}}_{\textit{BIAS}}$$

How can we eliminate the bias term?

- As a result of randomization, the selection bias term will be zero
- The treatment and control group will tend to be similar along all characteristics (identical in expectation), including the potential outcomes under the control condition

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbf{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned} \mathbf{E}[Y|D=1] &= \mathbf{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\ &= \mathbf{E}[Y_1 | D=1] \end{aligned}$$

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbf{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned}
\mathbf{E}[Y|D=1] &= \mathbf{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\
&= \mathbf{E}[Y_1|D=1] \\
&= \mathbf{E}[Y_1]
\end{aligned}$$

$$\mathbf{E}[Y|D=0] =$$

Identification Under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

Problem: $\tau_{ATE} = \mathbf{E}[Y_1 - Y_0]$ *is unobserved. But given random assignment*

$$\begin{aligned}
\mathbf{E}[Y|D=1] &= \mathbf{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=1] \\
&= \mathbf{E}[Y_1|D=1] \\
&= \mathbf{E}[Y_1]
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}[Y|D=0] &= \mathbf{E}[D \cdot Y_1 + (1-D) \cdot Y_0 | D=0] \\
&= \mathbf{E}[Y_0|D=0] \\
&= \mathbf{E}[Y_0]
\end{aligned}$$

$$\tau_{ATE} = \mathbf{E}[Y_1 - Y_0] = \mathbf{E}[Y_1] - \mathbf{E}[Y_0] = \underbrace{\mathbf{E}[Y|D=1] - \mathbf{E}[Y|D=0]}_{\textit{Difference in Means}}$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	0	3	1
2	1	1	1	1
3	2	0	0	0
4	2	1	1	0

What is $\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0]$?

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	0	3	1
2	1	1	1	1
3	2	0	0	0
4	2	1	1	0
$\mathbf{E}[Y_1]$	2			
$\mathbf{E}[Y_0]$		.5		

$$\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0] = 2 - .5 = 1.5$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i
1	3	?	3	1
2	1	?	1	1
3	?	0	0	0
4	?	1	1	0
$\mathbf{E}[Y_1]$	?			
$\mathbf{E}[Y_0]$		?		

What is $\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0]$?

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	?	3	1	?
2	1	?	1	1	?
3	?	0	0	0	?
4	?	1	1	0	?
$\mathbf{E}[Y_1]$	?				
$\mathbf{E}[Y_0]$		?			

What is $\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0]$? In an experiment, the researcher controls the probability of assignment to treatment for all units $P(D_i = 1)$ and by imposing equal probabilities we ensure that treatment assignment is independent of the potential outcomes, i.e. $(Y_1, Y_0) \perp\!\!\!\perp D$.

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4
$\mathbf{E}[Y_1]$	2				
$\mathbf{E}[Y_0]$		.5			

What is $\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0]$? Given that D_i is randomly assigned with probability 1/2, we have $\mathbf{E}[Y|D=1] = \mathbf{E}[Y_1|D=1] = \mathbf{E}[Y_1]$.

All possible randomizations with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
Average $Y|D=1$:	2	2.5	2.5	1.5	1.5	2

So $\mathbf{E}[Y|D=1] = \mathbf{E}[Y_1] = 2$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4
$\mathbf{E}[Y_1]$	2				
$\mathbf{E}[Y_0]$		.5			

By the same logic, we have: $\mathbf{E}[Y|D=0] = \mathbf{E}[Y_0|D=0] = \mathbf{E}[Y_0] = .5$.

Therefore the average treatment effect is identified:

$$\tau_{ATE} = \mathbf{E}[Y_1] - \mathbf{E}[Y_0] = \underbrace{\mathbf{E}[Y|D=1] - \mathbf{E}[Y|D=0]}_{\text{Difference in Means}}$$

Average Treatment Effect (ATE)

Imagine a population with 4 units:

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4
$\mathbf{E}[Y_1]$	2				
$\mathbf{E}[Y_0]$		.5			

Also since $\mathbf{E}[Y|D=0] = \mathbf{E}[Y_0|D=0] = \mathbf{E}[Y_0|D=1] = \mathbf{E}[Y_0]$ we have that

$$\begin{aligned} \tau_{ATT} &= \mathbf{E}[Y_1 - Y_0|D=1] = \mathbf{E}[Y_1|D=1] - \mathbf{E}[Y_0|D=0] \\ &= \mathbf{E}[Y_1] - \mathbf{E}[Y_0] = \mathbf{E}[Y_1 - Y_0] \\ &= \tau_{ATE} \end{aligned}$$

Identification under Random Assignment

Identification Assumption

$(Y_1, Y_0) \perp\!\!\!\perp D$ *(random assignment)*

Identification Result

We have that

$$\mathbf{E}[Y_0|D=0] = \mathbf{E}[Y_0] = \mathbf{E}[Y_0|D=1]$$

and therefore

$$\begin{aligned} \underbrace{E[Y|D=1] - \mathbf{E}[Y|D=0]}_{\textit{Difference in Means}} &= \underbrace{E[Y_1 - Y_0|D=1]}_{ATT} + \underbrace{\{\mathbf{E}[Y_0|D=1] - \mathbf{E}[Y_0|D=0]\}}_{BIAS} \\ &= \underbrace{E[Y_1 - Y_0|D=1]}_{ATT} \end{aligned}$$

As a result,

$$\underbrace{E[Y|D=1] - \mathbf{E}[Y|D=0]}_{\textit{Difference in Means}} = \tau_{ATE} = \tau_{ATT}$$

Identification in Randomized Experiments

Identification Assumption

Given random assignment $(Y_1, Y_0) \perp\!\!\!\perp D$

Identification Result

Let $F_{Y_d}(y)$ *be the cumulative distribution function (CDF) of* Y_d*, then*

$$\begin{aligned} F_{Y_0}(y) &= \Pr(Y_0 \le y) = \Pr(Y_0 \le y | D = 0) \\ &= \Pr(Y \le y | D = 0). \end{aligned}$$

Similarly,

$$F_{Y_1}(y) = \Pr(Y \le y | D = 1).$$

So the effect of the treatment at any quantile $\theta \in [0, 1]$ *is identified:*

$$\alpha_\theta = Q_\theta(Y_1) - Q_\theta(Y_0) = Q_\theta(Y | D = 1) - Q_\theta(Y | D = 0)$$

where $F_{Y_d}(Q_\theta(Y_d)) = \theta$*.*

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Estimation Under Random Assignment

Consider a randomized trial with N individuals.

Estimand

$\tau_{ATE} = \mathbf{E}[Y_1 - Y_0] = \mathbf{E}[Y|D=1] - \mathbf{E}[Y|D=0]$

Estimator

Estimation Under Random Assignment

Consider a randomized trial with N individuals.

Estimand

$\tau_{ATE} = \mathbf{E}[Y_1 - Y_0] = \mathbf{E}[Y|D=1] - \mathbf{E}[Y|D=0]$

Estimator

By the analogy principle we use

$$\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$$

$$\bar{Y}_1 = \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i\,;$$

$$\bar{Y}_0 = \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i$$

with $N_1 = \sum_i D_i$ *and* $N_0 = N - N_1$.

Under random assignment, $\widehat{\tau}$ *is an unbiased and consistent estimator of* τ_{ATE} *(*$\mathbf{E}[\widehat{\tau}] = \tau_{ATE}$ *and* $\widehat{\tau}_N \xrightarrow{p} \tau_{ATE}$*.)*

Unbiasedness Under Random Assignment

One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbf{E}[D] = \frac{N_1}{N}$ and $\mathbf{E}[1 - D] = \frac{N_0}{N}$

Unbiasedness Under Random Assignment

One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbf{E}[D] = \frac{N_1}{N}$ and $\mathbf{E}[1 - D] = \frac{N_0}{N}$

Rewrite the estimators as follows:

$$\widehat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D \cdot Y_1}{N_1/N} - \frac{(1 - D) \cdot Y_0}{N_0/N} \right)$$

Take expectations with respect to the sampling distribution given by the design. Under the Neyman model, Y_1 and Y_0 are fixed and only D_i is random.

Unbiasedness Under Random Assignment

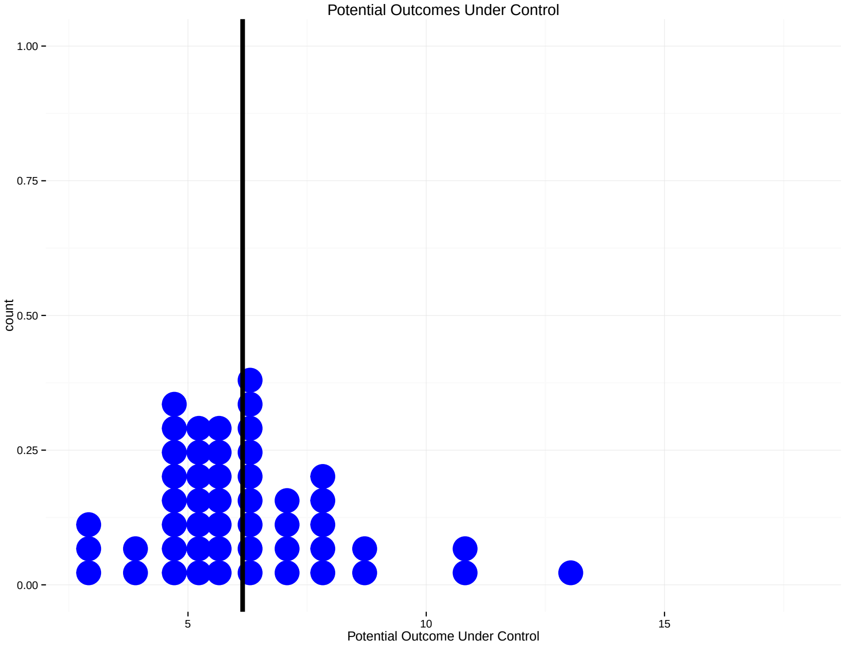
One way of showing that $\widehat{\tau}$ is unbiased is to exploit the fact that under independence of potential outcomes and treatment status, $\mathbf{E}[D]=\frac{N_1}{N}$ and $\mathbf{E}[1-D]=\frac{N_0}{N}$

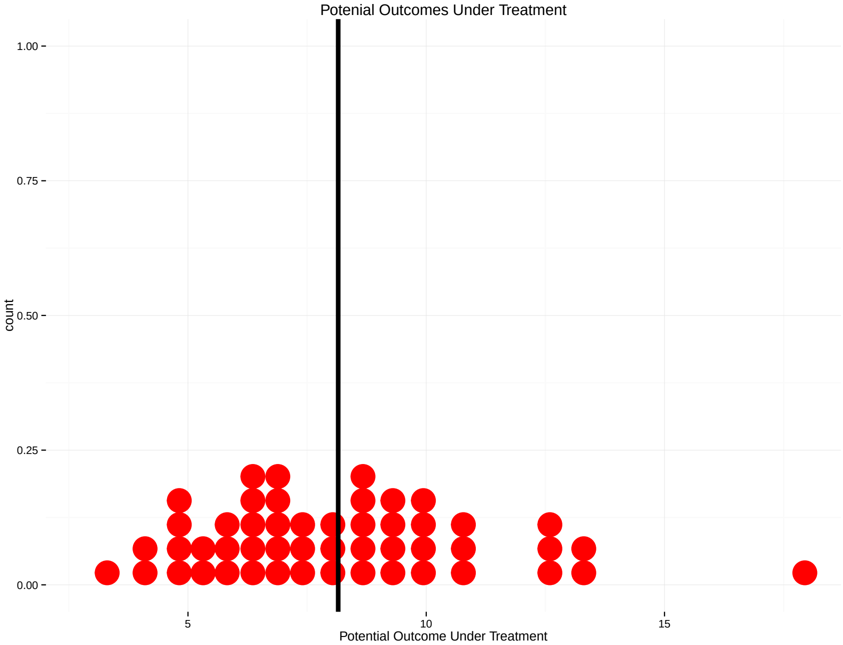
Rewrite the estimators as follows:

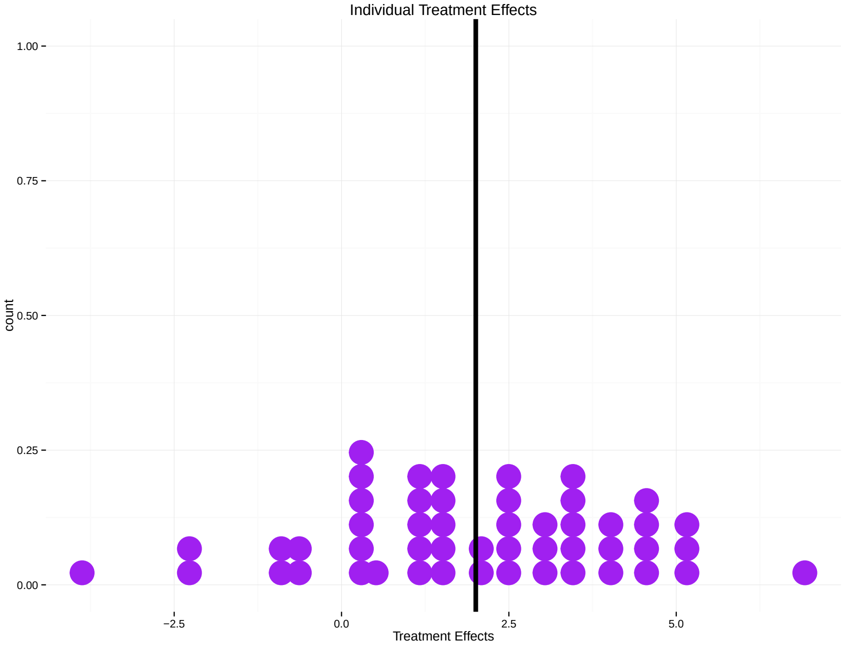
$$\widehat{\tau}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D\cdot Y_1}{N_1/N}-\frac{(1-D)\cdot Y_0}{N_0/N}\right)$$

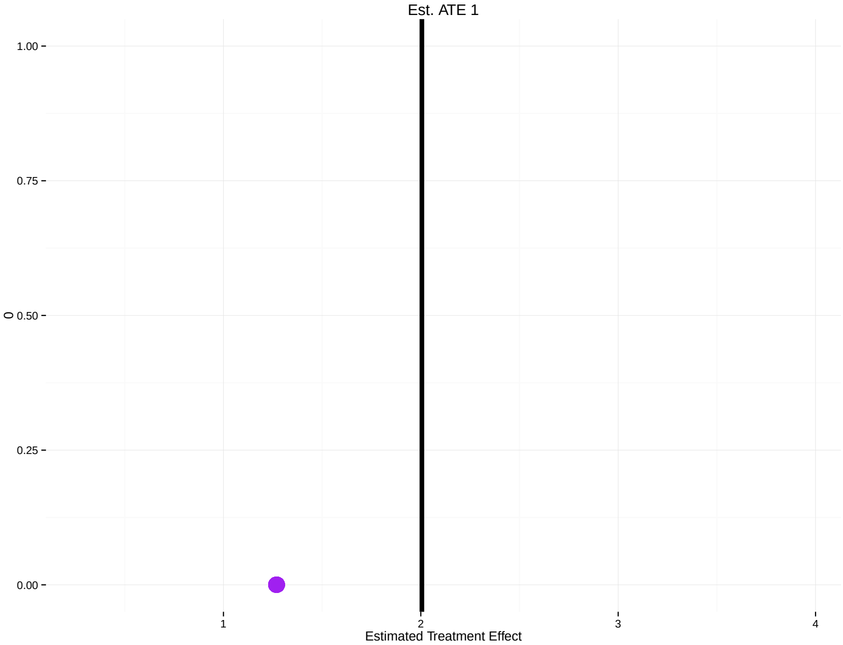
Take expectations with respect to the sampling distribution given by the design. Under the Neyman model, Y_1 and Y_0 are fixed and only D_i is random.

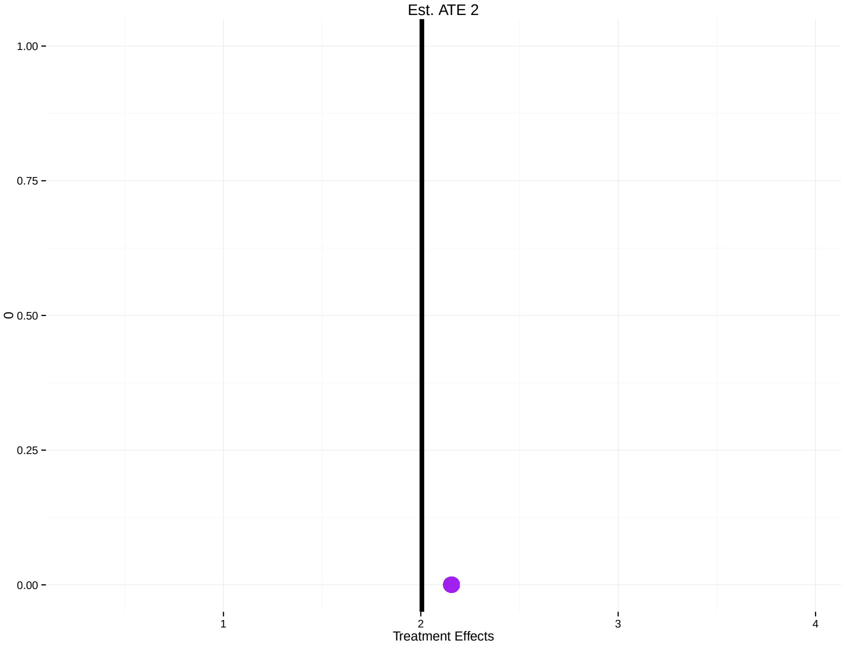
$$\mathbf{E}[\widehat{\tau}]=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbf{E}[D]\cdot Y_1}{N_1/N}-\frac{\mathbf{E}[(1-D)]\cdot Y_0}{N_0/N}\right)=\frac{1}{N}\sum_{i=1}^{N}(Y_1-Y_0)=\tau$$

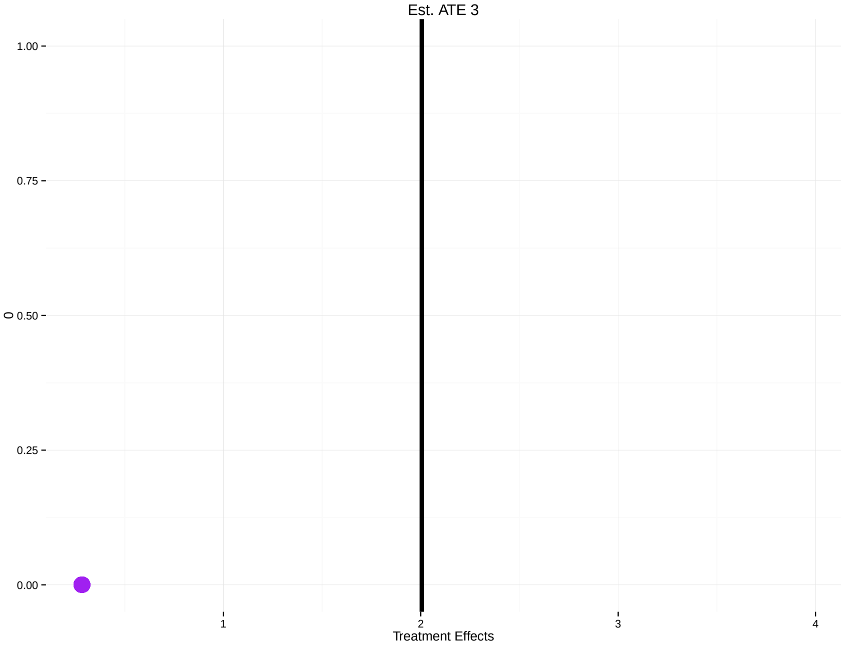

Potential Outcomes Under Control
count
1.00
0.75
0.50
0.25
0.00
5
10
15
Potential Outcome Under Control

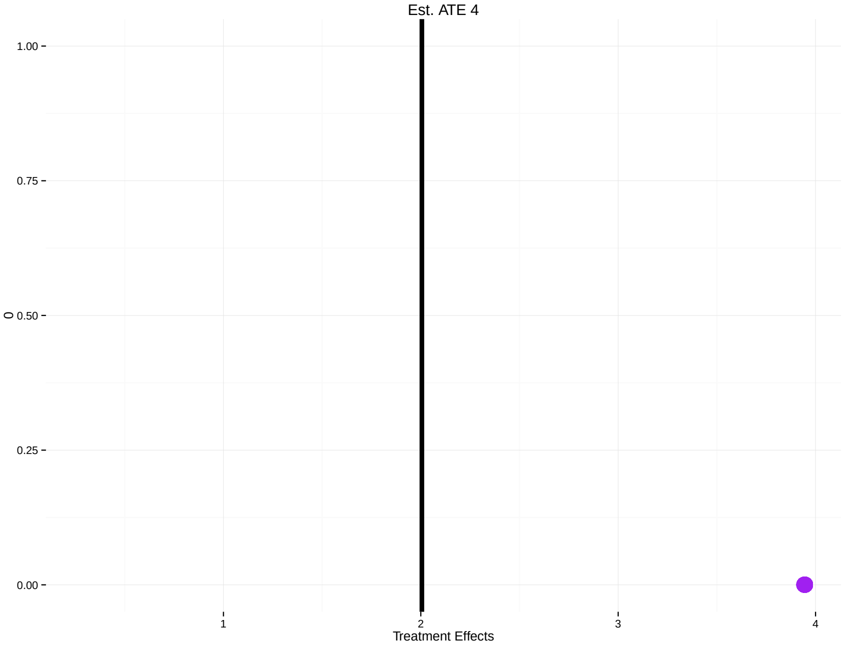

Potenial Outcomes Under Treatment
count
1.00
0.75
0.50
0.25
0.00
5
10
15
Potential Outcome Under Treatment

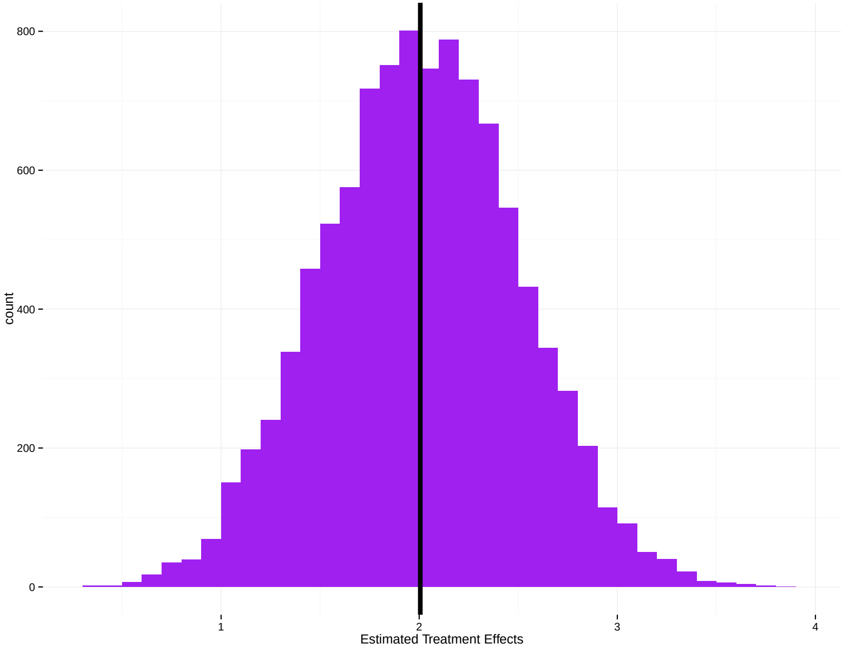

Individual Treatment Effects
count
1.00
0.75
0.50
0.25
0.00
−2.5
0.0
2.5
5.0
Treatment Effects


Est. ATE 1
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
Estimated Treatment Effect


Est. ATE 2
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
Treatment Effects


Est. ATE 3
1.00
0.75
0
0.50
0.25
0.00
1
2
3
4
Treatment Effects


Est. ATE 4
0
1.00
0.75
0.50
0.25
0.00
1
2
3
4
Treatment Effects


800
600
400
200
0
count
1
2
3
4
Estimated Treatment Effects

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

What is the Estimand?

- So far we have emphasized effect estimation, but what about *uncertainty*?

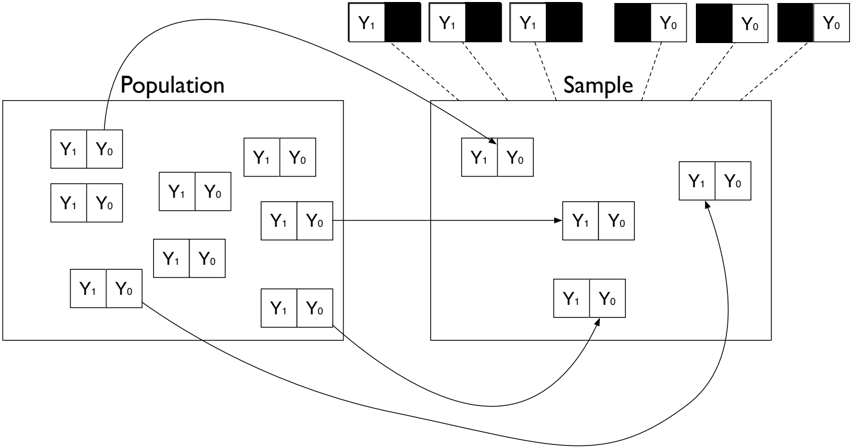
- In the design based literature, variability in our estimates can arise from two sources:

  1. Sampling variation induced by the procedure that selected the units into our sample.

  2. Variation induced by the particular realization of the treatment variable.

- This distinction is important, but often ignored

What is the Estimand?



SATE and PATE

- Typically we focus on estimating the average causal effect in a particular sample: **S**ample **A**verage **T**reatment **E**ffect (SATE)
  - Uncertainty arises only from hypothetical randomizations.
  - Inferences are limited to the sample in our study.
- Might care about the **P**opulation **A**verage **T**reatment **E**ffect (PATE)
  - Requires precise knowledge about the sampling process that selected units from the population into the sample.
  - Need to account for two sources of variation:
    - Variation from the sampling process
    - Variation from treatment assignment.
- Thus, in general, $\mathrm{Var}(\widehat{\mathrm{PATE}}) > \mathrm{Var}(\widehat{\mathrm{SATE}})$.

Standard Error for Sample ATE

The standard error is the standard deviation of a sampling distribution: $SE_{\widehat{\theta}} \equiv \sqrt{\frac{1}{J}\sum_1^J(\widehat{\theta}_j - \overline{\widehat{\theta}})^2}$ (with J possible random assignments).

i	Y_{1i}	Y_{0i}	Y_i	D_i	$P(D_i = 1)$
1	3	0	3	1	2/4
2	1	1	1	1	2/4
3	2	0	0	0	2/4
4	2	1	1	0	2/4

ATE estimates given all possible random assignments with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
$\widehat{ATE}$:	1.5	1.5	2	1	1.5	1.5

The average $\widehat{ATE}$ is 1.5 and therefore the true standard error is

$$SE_{\widehat{ATE}} = \sqrt{\tfrac{1}{6}[(1.5-1.5)^2 + (1.5-1.5)^2 + (2-1.5)^2 + (1-1.5)^2 + (1.5-1.5)^2 + (1.5-1.5)^2]} \approx .28$$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the *estimated* sample ATE is given by

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the *estimated* sample ATE is given by

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Plugging in, we obtain the true standard error of the estimated sample ATE

$$SE_{\widehat{ATE}} = \sqrt{\left(\frac{4-2}{4-1}\right)\frac{.25}{2} + \left(\frac{4-2}{4-1}\right)\frac{.5}{2} + \left(\frac{1}{4-1}\right)2(-.25)} \approx .28$$

Standard Error for Sample ATE

Standard Error for Sample ATE

Given complete randomization of N units with N_1 assigned to treatment and $N_0 = N - N_1$ to control, the true standard error of the *estimated* sample ATE is given by

$$SE_{\widehat{ATE}} \quad = \quad \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_{1i}]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_{0i}]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_{1i}, Y_{0i}]}$$

with population variances and covariance

$$Var[Y_{di}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{di} - \frac{\sum_1^N Y_{di}}{N}\right)^2 = \sigma^2_{Y_d|D_i=d}$$

$$Cov[Y_{1i}, Y_{0i}] \equiv \frac{1}{N}\sum_{1}^{N}\left(Y_{1i} - \frac{\sum_1^N Y_{1i}}{N}\right)\left(Y_{0i} - \frac{\sum_1^N Y_{0i}}{N}\right) = \sigma^2_{Y_1,Y_0}$$

Standard error decreases if:

- N grows
- $Var[Y_1]$, $Var[Y_0]$ decrease
- $Cov[Y_1, Y_0]$ decreases

Conservative Estimator $\widehat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

Conservative Estimator $\widehat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

What about the covariance?

Conservative Estimator $\widehat{SE}_{\widehat{ATE}}$

Conservative Estimator for Standard Error for Sample ATE

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{Var[Y_{1i}]}}{N_1} + \frac{\widehat{Var[Y_{0i}]}}{N_0}}$$

with estimators of the sample variances given by

$$\widehat{Var[Y_{1i}]} \equiv \frac{1}{N_1 - 1} \sum_{i|D_i=1}^{N} \left(Y_{1i} - \frac{\sum_{i|D_i=1}^{N} Y_{1i}}{N_1} \right)^2 = \widehat{\sigma}^2_{Y|D_i=1}$$

$$\widehat{Var[Y_{0i}]} \equiv \frac{1}{N_0 - 1} \sum_{i|D_i=0}^{N} \left(Y_{0i} - \frac{\sum_{i|D_i=0}^{N} Y_{0i}}{N_0} \right)^2 = \widehat{\sigma}^2_{Y|D_i=0}$$

- Conservative compared to the true standard error, i.e. $SE_{\widehat{ATE}} < \widehat{SE}_{\widehat{ATE}}$
- Asymptotically unbiased in two special cases:
  - if τ_i is constant (i.e. $Cor[Y_1, Y_0] = 1$)
  - if we estimate standard error of population average treatment effect ($Cov[Y_1, Y_0]$ is negligible when we sample from a large population)
- Equivalent to standard error for two sample t-test with unequal variances or "robust" standard error in regression of Y on D

Proof: $SE_{\widehat{ATE}} \leq \widehat{SE}_{\widehat{ATE}}$

Upper bound for standard error is when $Cor[Y_1, Y_0] = 1$:

$$Cor[Y_1, Y_0] = \frac{Cov[Y_1, Y_0]}{\sqrt{Var[Y_1]Var[Y_0]}} \leq 1 \Longleftrightarrow Cov[Y_1, Y_0] \leq \sqrt{Var[Y_1]Var[Y_0]}$$

$$\begin{aligned}
SE_{\widehat{ATE}} &= \sqrt{\left(\frac{N-N_1}{N-1}\right)\frac{Var[Y_1]}{N_1} + \left(\frac{N-N_0}{N-1}\right)\frac{Var[Y_0]}{N_0} + \left(\frac{1}{N-1}\right)2Cov[Y_1, Y_0]} \\
&= \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2Cov[Y_1, Y_0]\right)} \\
&\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + 2\sqrt{Var[Y_1]Var[Y_0]}\right)} \\
&\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)}
\end{aligned}$$

Last step follows from the following inequality

$$\begin{aligned}
(\sqrt{Var[Y_1]} - \sqrt{Var[Y_0]})^2 &\geq 0 \\
Var[Y_1] - 2\sqrt{Var[Y_1]Var[Y_0]} + Var[Y_0] &\geq 0 \Longleftrightarrow Var[Y_1] + Var[Y_0] \geq 2\sqrt{Var[Y_1]Var[Y_0]}
\end{aligned}$$

Proof: $SE_{\widehat{ATE}} \leq \widehat{SE}_{\widehat{ATE}}$

$$
\begin{aligned}
SE_{\widehat{ATE}} &\leq \sqrt{\frac{1}{N-1}\left(\frac{N_0}{N_1}Var[Y_1] + \frac{N_1}{N_0}Var[Y_0] + Var[Y_1] + Var[Y_0]\right)} \\
&\leq \sqrt{\frac{N_0^2 Var[Y_1] + N_1^2 Var[Y_0] + N_1 N_0 (Var[Y_1] + Var[Y_0])}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0^2 + N_1 N_0) Var[Y_1] + (N_1^2 + N_1 N_0) Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{(N_0 + N_1) N_0 Var[Y_1]}{(N-1)N_1 N_0} + \frac{(N_1 + N_0) N_1 Var[Y_0]}{(N-1)N_1 N_0}} \\
&\leq \sqrt{\frac{N\, Var[Y_1]}{(N-1)N_1} + \frac{N\, Var[Y_0]}{(N-1)N_0}} \\
&\leq \sqrt{\frac{N}{N-1}\left(\frac{Var[Y_1]}{N_1} + \frac{Var[Y_0]}{N_0}\right)}
\end{aligned}
$$

Proof: $SE_{\widehat{ATE}} \leq \widehat{SE}_{\widehat{ATE}}$

$$SE_{\widehat{ATE}} \quad \leq \quad \sqrt{\frac{N}{N-1}\left(\frac{1}{N_1}Var[Y_1] + \frac{1}{N_0}Var[Y_0]\right)}$$

Now, we need to estimate $Var[Y_1]$ and $Var[Y_0]$. Recall that for simple random sampling without replacement, the unbiased estimator of a population variance (σ^2) is $\hat{\sigma}^2_n(\frac{n}{n-1})(\frac{N-1}{N})$, which can be rewritten as $\hat{\sigma}^2_{n-1}(\frac{N-1}{N})$. In the set-up presented here, we have defined $\widehat{Var[Y_d]}$ to correspond to $\hat{\sigma}^2_{n-1}$ (separately for $d = 1, 0$). Thus, inserting the unbiased estimators in for $Var[Y_1]$ and $Var[Y_0]$, we get:

$$\sqrt{\frac{N}{N-1}\left(\frac{1}{N_1}\widehat{Var[Y_1]}\left(\frac{N-1}{N}\right) + \frac{1}{N_0}\widehat{Var[Y_0]}\left(\frac{N-1}{N}\right)\right)}$$

$$= \sqrt{\left(\frac{\widehat{Var[Y_1]}}{N_1} + \frac{\widehat{Var[Y_0]}}{N_0}\right)}$$

Thus:

$$SE_{\widehat{ATE}} \quad \leq \quad \sqrt{\frac{\widehat{Var[Y_1]}}{N_1} + \frac{\widehat{Var[Y_0]}}{N_0}} = \widehat{SE}_{\widehat{ATE}}$$

So the estimator for the standard error is conservative.

Standard Error for Sample ATE

i	Y_{1i}	Y_{0i}	Y_i
1	3	0	3
2	1	1	1
3	2	0	0
4	2	1	1

$\widehat{SE}_{\widehat{ATE}}$ estimates given all possible assignments with two treated units:

Treated Units:	1 & 2	1 & 3	1 & 4	2 & 3	2 & 4	3 & 4
$\widehat{ATE}$:	1.5	1.5	2	1	1.5	1.5
$\widehat{SE}_{\widehat{ATE}}$:	1.11	.5	.71	.71	.5	.5

The average $\widehat{SE}_{\widehat{ATE}}$ is $\approx .67$ compared to the true standard error of $SE_{\widehat{ATE}} \approx .28$

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16, 199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17, 038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15, 040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16, 180
- Estimated average effect of training:

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:
  - $\widehat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0 = 16,199 - 15,040 = \$1,159$
- Estimated standard error for effect of training:

Example: Effect of Training on Earnings

- Treatment Group:
  - $N_1 = 7,487$
  - Estimated Average Earnings $\bar{Y}_1$: \$16,199
  - Estimated Sample Standard deviation $\widehat{\sigma}_{Y|D_i=1}$: \$17,038
- Control Group :
  - $N_0 = 3,717$
  - Estimated Average Earnings $\bar{Y}_0$: \$15,040
  - Estimated Sample deviation $\widehat{\sigma}_{Y|D_i=0}$: \$16,180
- Estimated average effect of training:
  - $\widehat{\tau}_{ATE} = \bar{Y}_1 - \bar{Y}_0 = 16,199 - 15,040 = \$1,159$
- Estimated standard error for effect of training:
  - $\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\widehat{\sigma}^2_{Y|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y|D_i=0}}{N_0}} = \sqrt{\frac{17{,}038^2}{7{,}487} + \frac{16{,}180^2}{3{,}717}} \approx \$330$
- Is this consistent with a zero average treatment effect $\alpha_{ATE} = 0$?

Testing the Null Hypothesis of Zero Average Effect

- Under the null hypothesis H_0: $\tau_{ATE} = 0$, the average potential outcomes in the population are the same for treatment and control: $\mathbf{E}[Y_1] = \mathbf{E}[Y_0]$.
- Since units are randomly assigned, both the treatment and control groups should therefore have the same sample average earnings
- However, we in fact observe a difference in mean earnings of $\$1,159$
- What is the probability of observing a difference this large if the true average effect of the training were zero (i.e. the null hypothesis were true)?

Testing the Null Hypothesis of Zero Average Effect

- Use a two-sample t-test with unequal variances:

$$t = \frac{\widehat{\tau}}{\sqrt{\frac{\widehat{\sigma}^2_{Y_i|D_i=1}}{N_1} + \frac{\widehat{\sigma}^2_{Y_i|D_i=0}}{N_0}}} = \frac{\$1,159}{\sqrt{\frac{\$17,038^2}{7,487} + \frac{\$16,180^2}{3,717}}} \approx 3.5$$

  - From basic statistical theory, we know that $t_N \xrightarrow{d} \mathcal{N}(0,1)$
  - And for a standard normal distribution, the probability of observing a value of t that is larger than $|t| > 1.96$ is $< .05$
  - So obtaining a value as high as $t = 3.5$ is very unlikely under the null hypothesis of a zero average effect
  - We reject the null hypothesis H$_0$: $\tau_0 = 0$ against the alternative H$_1$: $\tau_0 \neq 0$ at asymptotic 5% significance level whenever $|t| > 1.96$.
  - Inverting the test statistic we can construct a 95% confidence interval

$$\widehat{\tau}_{ATE} \pm 1.96 \cdot \widehat{SE}_{\widehat{ATE}}$$

Testing the Null Hypothesis of Zero Average Effect

R Code

```
> d <- read.dta("jtpa.dta")
> head(d[,c("earnings","assignmt")])
  earnings assignmt
1     1353        1
2     4984        1
3    27707        1
4    31860        1
5    26615        0
>
> meanAsd <- function(x){
+   out <- c(mean(x),sd(x))
+   names(out) <- c("mean","sd")
+   return(out)
+ }
>
> aggregate(earnings˜assignmt,data=d,meanAsd)
  assignmt earnings.mean earnings.sd
1        0      15040.50    16180.25
2        1      16199.94    17038.85
```

Testing the Null Hypothesis of Zero Average Effect

R Code

```
> t.test(earnings˜assignmt,data=d,var.equal=FALSE)

Welch Two Sample t-test

data:  earnings by assignmt
t = -3.5084, df = 7765.599, p-value = 0.0004533
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1807.2427  -511.6239
sample estimates:
mean in group 0 mean in group 1
       15040.50        16199.94
```

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$\begin{aligned} Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\ &= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\ &= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\ &= \alpha + \tau_{Reg} D_i + \epsilon_i \end{aligned}$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$\begin{aligned}
Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\
&= \alpha + \tau_{Reg} D_i + \epsilon_i
\end{aligned}$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:
    - Baseline difference in potential outcomes under control that is correlated with D_i.
    - Individual treatment effects τ_i are correlated with D_i
    - Under random assignment, both correlations are zero in expectation

Regression to Estimate the Average Treatment Effect

Estimator (Regression)

The ATE can be expressed as a regression equation:

$$
\begin{aligned}
Y_i &= D_i\, Y_{1i} + (1 - D_i)\, Y_{0i} \\
&= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= \underbrace{\bar{Y}_0}_{\alpha} + \underbrace{(\bar{Y}_1 - \bar{Y}_0)}_{\tau_{Reg}} D_i + \underbrace{\{(Y_{i0} - \bar{Y}_0) + D_i \cdot [(Y_{i1} - \bar{Y}_1) - (Y_{i0} - \bar{Y}_0)]\}}_{\epsilon} \\
&= \alpha + \tau_{Reg} D_i + \epsilon_i
\end{aligned}
$$

- τ_{Reg} could be biased for τ_{ATE} in two ways:
  - Baseline difference in potential outcomes under control that is correlated with D_i.
  - Individual treatment effects τ_i are correlated with D_i
  - Under random assignment, both correlations are zero in expectation
- Effect heterogeneity implies "heteroskedasticity", i.e. error variance differs by values of D_i.
  - Neyman model implies "robust" standard errors.
- Can use regression in experiments without assuming constant effects.

Regression to Estimate the Average Treatment Effect

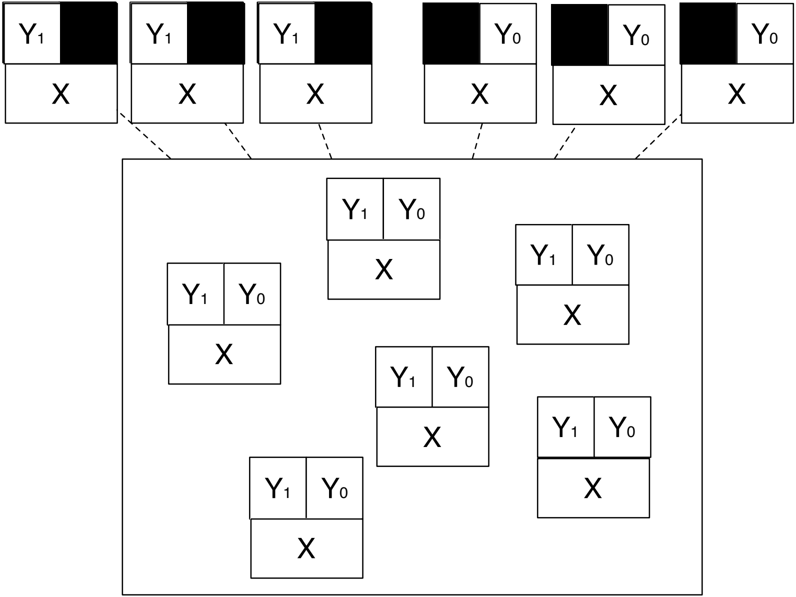
R Code

```
> library(sandwich)
> library(lmtest)
>
> lout <- lm(earnings~assignmt,data=d)
> coeftest(lout,vcov = vcovHC(lout, type = "HC1")) # matches Stata

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 15040.50     265.38 56.6752 < 2.2e-16 ***
assignmt     1159.43     330.46  3.5085 0.0004524 ***
---
```

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Covariates and Experiments



Covariates

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.
- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.
- Under randomization, $f_{X|D}(X|D=1) \overset{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).
- Similarity in distributions of covariates is known as covariate balance.
- If this is not the case, then one of two possibilities:

Covariates

- Randomization is gold standard for causal inference because in expectation it balances observed but also unobserved characteristics between treatment and control group.

- Unlike potential outcomes, you observe baseline covariates for all units. Covariate values are predetermined with respect to the treatment and do not depend on D_i.

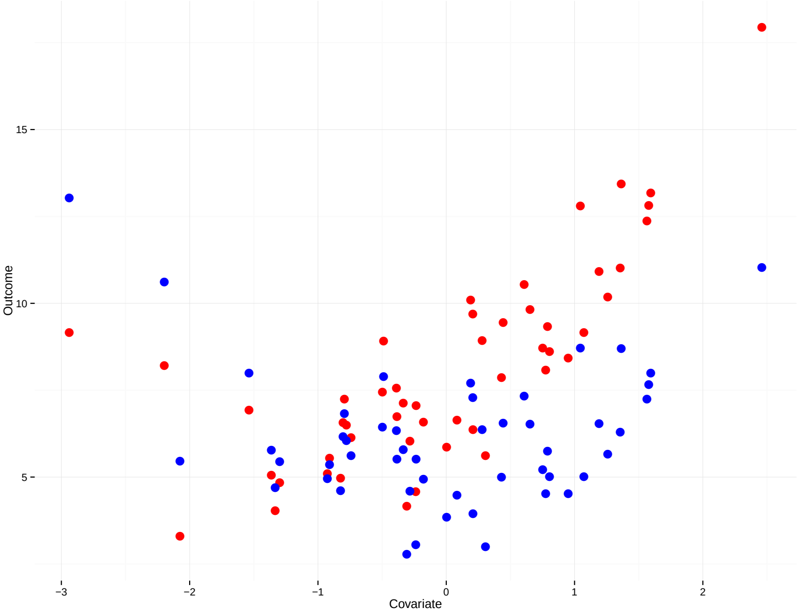
- Under randomization, $f_{X|D}(X|D=1) \overset{d}{=} f_{X|D}(X|D=0)$ (equality in distribution).

- Similarity in distributions of covariates is known as covariate balance.

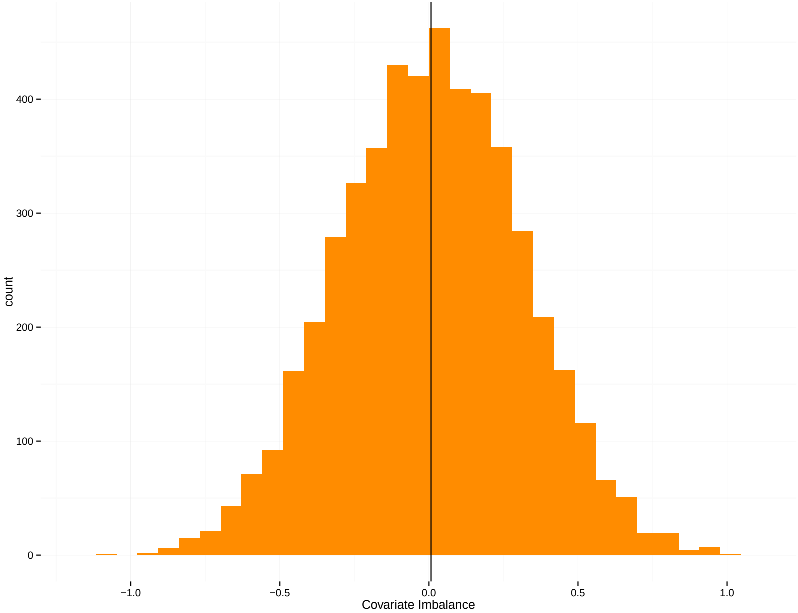
- If this is not the case, then one of two possibilities:
  - Randomization was compromised.
  - Sampling error (bad luck)

- One should always test for covariate balance on important covariates, using so called "balance checks" (eg. t-tests, F-tests, etc.)

Covariates and Experiments



Covariates and Experiments



Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?

Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Why include X_i when experiments “control” for covariates by design?
    - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.

Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?
    - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.
    - Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics, thus making it easier to attribute remaining differences to the treatment.

Regression with Covariates

- Practioners often run some variant of the following model with experimental data:

$$Y_i = \alpha + \tau D_i + X_i \beta + \epsilon_i$$

- Why include X_i when experiments "control" for covariates by design?
  - Correct for chance covariate imbalances that indicate that $\hat{\tau}$ may be far from τ_{ATE}.
  - Increase precision: remove variation in the outcome accounted for by pre-treatment characteristics, thus making it easier to attribute remaining differences to the treatment.
- ATE estimates are robust to model specification (with sufficient N).
  - Never control for **post-treatment** covariates!

Covariate Adjustment with Regression

Freedman (2008) shows that regression of the form:

$$Y_i = \alpha + \tau_{reg} D_i + \beta_1 X_i + \epsilon_i$$

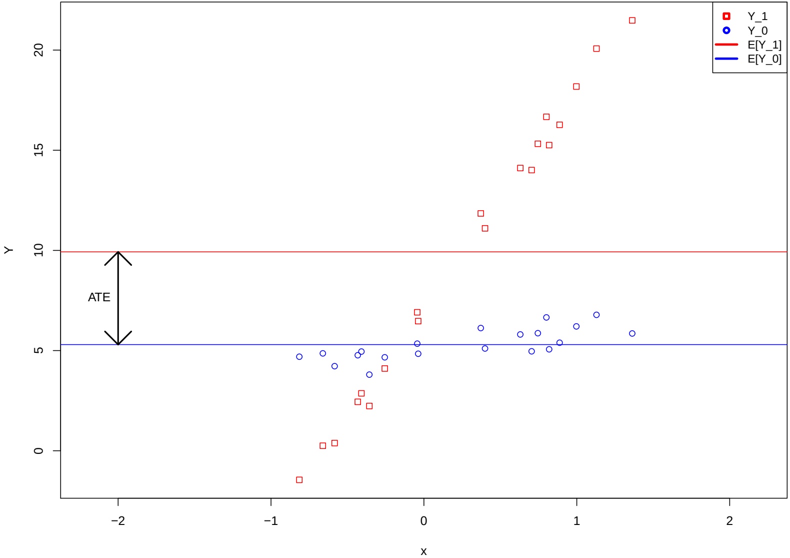
- $\hat{\tau}_{reg}$ is consistent for ATE and has small sample bias (unless model is true)
  - bias is on the order of $1/n$ and diminishes rapidly as N increases
- $\hat{\tau}_{reg}$ will not necessarily improve precision if model is incorrect
  - But harmful to precision only if more than 3/4 of units are assigned to one treatment condition or $\mathrm{Cov}(D_i, Y_1 - Y_0)$ larger than $\mathrm{Cov}(D_i, Y)$.

Lin (2013) shows that regression of the form:

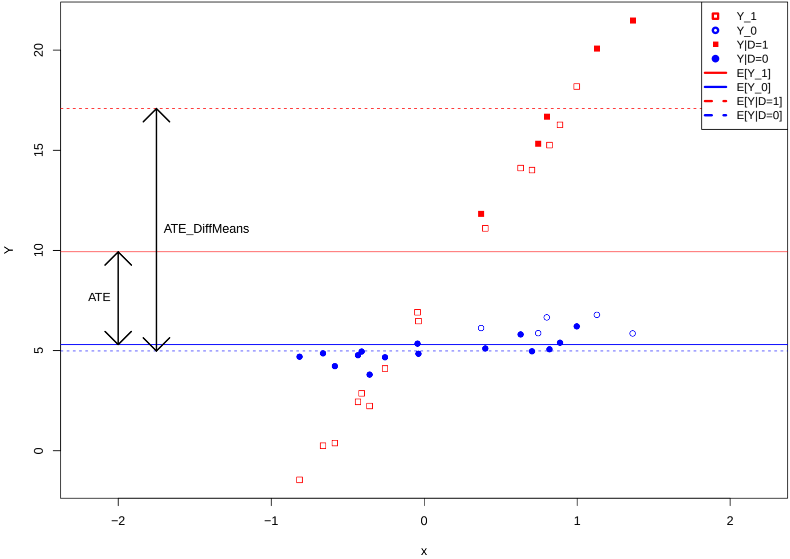
$$Y_i = \alpha + \tau_{interact} D_i + \beta_1 \cdot (X_i - \bar{X}) + \beta_2 \cdot D_i \ \ \cdot (X_i - \bar{X}) + \epsilon_i$$

- $\hat{\tau}_{interact}$ is consistent for ATE and has the same small sample bias
- Cannot hurt asymptotic precision even if model is incorrect and will likely increase precision if covariates are predictive of the outcomes.
- Results hold for multiple covariates

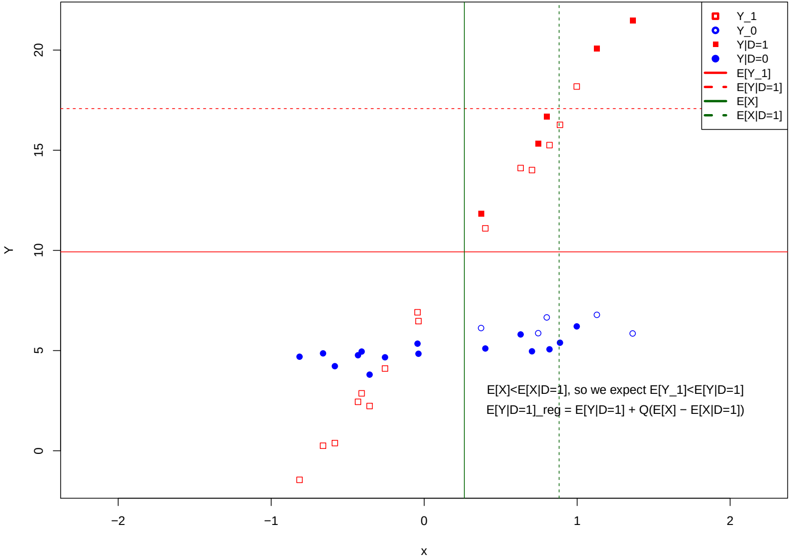
True ATE



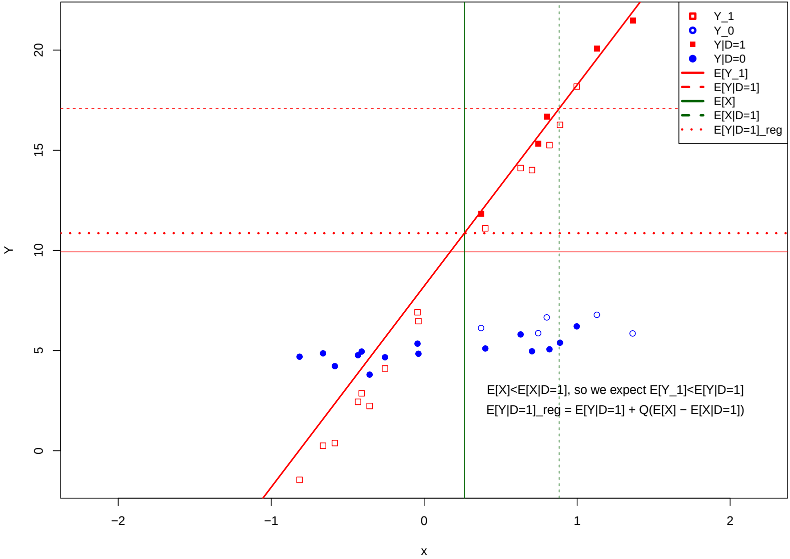
True ATE and Unadjusted Regression Estimator



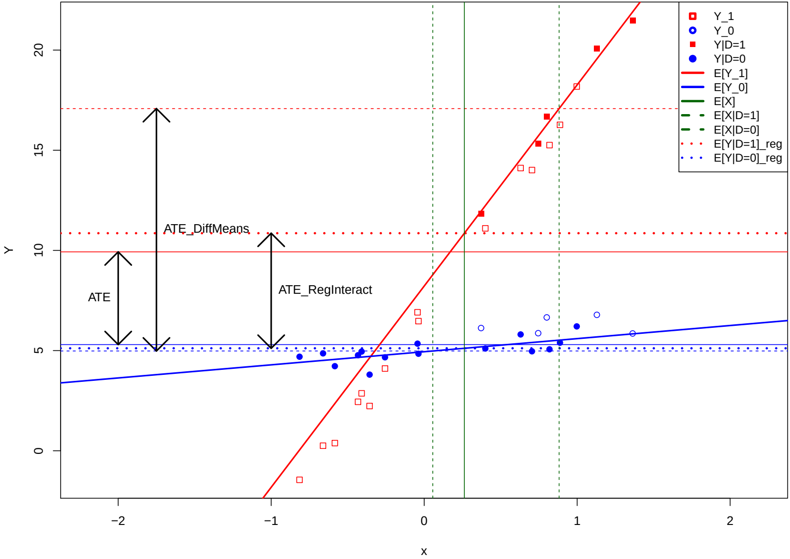
Adjusted Regression Estimator



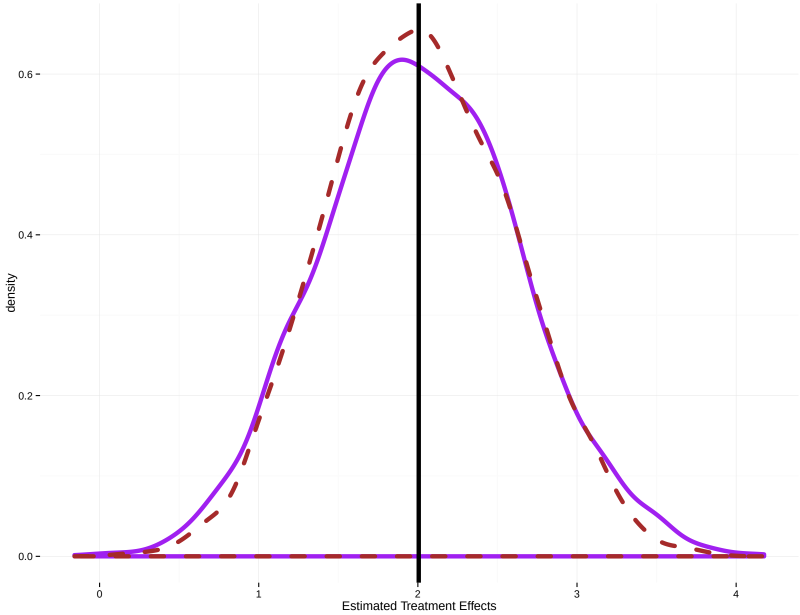
Adjusted Regression Estimator



Adjusted Regression Estimator



Covariate Adjustment with Regression



Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or *Anatomy of Regression*:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or *Anatomy of Regression*:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or *Anatomy of Regression*:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.

Let $\tilde{D}_i$ be the residuals after regressing D_i on X_i. For experimental data, on average, what will $\tilde{D}_i$ be equal to?

Why are Experimental Findings Robust to Alternative Specifications?

Note the following important property of OLS known as the Frisch-Waugh-Lovell (FWL) theorem or *Anatomy of Regression*:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{Var(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates.

Any multivariate regression coefficient can be expressed as the coefficient on a bivariate regression between the outcome and the regressor, after "partialling out" other variables in the model.

Let $\tilde{D}_i$ be the residuals after regressing D_i on X_i. For experimental data, on average, what will $\tilde{D}_i$ be equal to?

Since $\tilde{D}_i \approx D_i$, multivariate regressions will yield similar results to bivariate regressions.

Summary: Covariate Adjustment with Regression

- One does not need to believe in the classical linear model (linearity and constant treatment effects) to tolerate or even advocate OLS covariate adjustment in randomized experiments (agnostic view of regression).
- Covariate adjustment can buy you power (and thus allows for a smaller sample).
- Small sample bias might be a concern in small samples, but usually swamped by efficiency gains.
- Since covariates are controlled for by design, results are typically not model dependent.
- Best if covariate adjustment strategy is *pre-specified* as this rules out fishing.
- Always show the unadjusted estimate for transparency.

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
    - Average Treatment Effect
    - Standard Error for ATE
    - Hypothesis Testing
    - Covariates
    - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
    - Blocking
    - Power Calculations
    - Threats to Validity and Ethics

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0,$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.

Testing in Small Samples: Fisher's Exact Test

- Test of differences in means with large N:

$$H_0 : \mathbf{E}[Y_1] = \mathbf{E}[Y_0], \quad H_1 : \mathbf{E}[Y_1] \neq \mathbf{E}[Y_0] \text{ (weak null)}$$

- Fisher's Exact Test with small N:

$$H_0 : Y_1 = Y_0, \quad H_1 : Y_1 \neq Y_0 \qquad \text{(sharp null of no effect)}$$

- Let Ω be the set of all possible randomization realizations.
- We only observe the outcomes, Y_i, for one realization of the experiment. We calculate $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$.
- Under the sharp null hypothesis, we can compute the value that the difference in means estimator would have taken under any other realization, $\hat{\tau}(\omega)$, for $\omega \in \Omega$.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	?	1
2	1	?	1
3	?	0	0
4	?	1	0
$\widehat{\tau}_{ATE}$			1.5

What do we know given the sharp null $H_0 : Y_1 = Y_0$?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i
1	3	3	1
2	1	1	1
3	0	0	0
4	1	1	0
$\widehat{\tau}_{ATE}$			1.5
$\hat{\tau}(\omega)$			1.5

Given the full schedule of potential outcomes under the sharp null, we can compute the null distribution of ATE_{H_0} across all possible randomization.

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i
1	3	3	1	1
2	1	1	1	0
3	0	0	0	1
4	1	1	0	0
$\widehat{\tau}_{ATE}$			1.5	
$\hat{\tau}(\omega)$			1.5	0.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i
1	3	3	1	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	0	0	1
$\widehat{\tau}_{ATE}$			1.5		
$\hat{\tau}(\omega)$			1.5	0.5	1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0
2	1	1	1	0	0	1
3	0	0	0	1	0	1
4	1	1	0	0	1	0
$\widehat{\tau}_{ATE}$			1.5			
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0
2	1	1	1	0	0	1	1
3	0	0	0	1	0	1	0
4	1	1	0	0	1	0	1
$\widehat{\tau}_{ATE}$			1.5				
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

So $\Pr(\hat{\tau}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed?

Testing in Small Samples: Fisher's Exact Test

i	Y_{1i}	Y_{0i}	D_i	D_i	D_i	D_i	D_i	D_i
1	3	3	1	1	1	0	0	0
2	1	1	1	0	0	1	1	0
3	0	0	0	1	0	1	0	1
4	1	1	0	0	1	0	1	1
$\widehat{\tau}_{ATE}$			1.5					
$\hat{\tau}(\omega)$			1.5	0.5	1.5	-1.5	-.5	-1.5

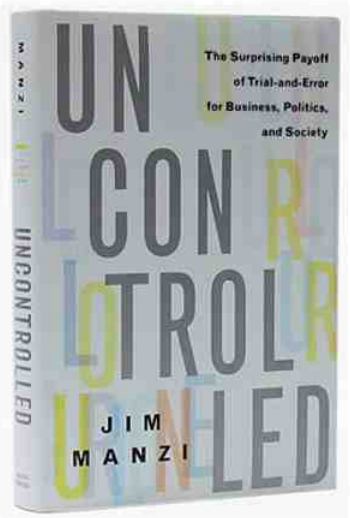
So $\Pr(\hat{\alpha}(\omega) \geq \widehat{\tau}_{ATE}) = 2/6 \approx .33$.

Which assumptions are needed? None! Randomization as "reasoned basis for causal inference" (Fisher 1935)

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
    - Average Treatment Effect
    - Standard Error for ATE
    - Hypothesis Testing
    - Covariates
    - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
    - Blocking
    - Power Calculations
    - Threats to Validity and Ethics

Experiments in Popular Culture


UNCONTROLLED
The Surprising Payoff of Trial-and-Error for Business, Politics, and Society
JIM MANZI
MANZI
UNCONTROLLED

The Rise of Experiments

Large increase in the use of experiments in the social sciences: laboratory, survey, and field experiments (see syllabus)

Abbreviated list of examples:

- *Program Evaluation*: development programs, education programs, weight loss programs, fundraising, deliberative polls, virginity pledging, advertising campaigns, mental exercise for elderly
- *Public policy evaluations*: teacher pay, class size, speed traps, vouchers, alternative sentencing, job training, health insurance subsidies, tax compliance, public housing, jury selection, police interventions
- *Behavioral Research*: persuasion, mobilization, education, income, interpersonal influence, conscientious health behaviors, media exposure, deliberation, discrimination
- *Research on Institutions*: rules for authorizing decisions, rules of succession, monitoring performance, transparency, corruption auditing, electoral systems

Experiments from Political Science and Economics

- Voter mobilization (Nickerson, Gerber and Green)
- Voting mechanisms (Olken)
- Health insurance reform (Finkelstein et al.)
- Race-based discrimination in labor markets (Bertrand and Mullainathan)
- Clientelistic vs programmatic presidential campaigns (Wantchekon)
- Female incumbents (Duflo)
- Information interventions for Elites (Butler)
- Monitoring interventions (Ichino)
- Audience costs (Tomz)
- Many more . . .

Social Pressure Experiment

- Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty.
- Two aspects of this type of utility: intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply.
- Gerber, Green, and Larimer (2008) test these motives in a large scale field experiment by applying varying degrees of intrinsic and extrinsic pressure on voters using a series of mailings to 180,002 households before the August 2006 primary election in Michigan.

Social Pressure Experiment

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: “YOU ARE BEING STUDIED!”

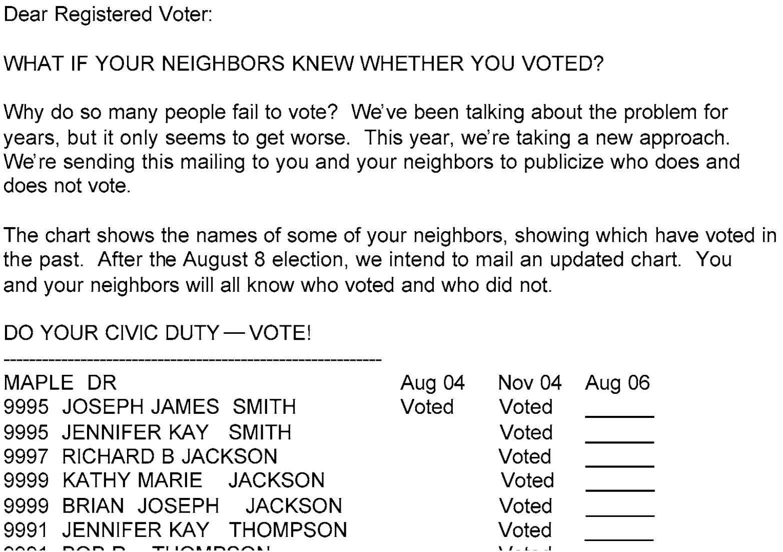
Social Pressure Experiment

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: “YOU ARE BEING STUDIED!”
- **Self**
  - Encouraged to vote.
  - Told that whether one votes is a matter of public record.
  - Shown whether members of their own household voted in the last two elections and promised to send post-card after election indicating whether or not they voted.

Social Pressure Experiment

- **Civic Duty**
  - Encouraged to vote.
- **Hawthorne**
  - Encouraged to vote.
  - Told that researchers would be checking on whether they voted: "YOU ARE BEING STUDIED!"
- **Self**
  - Encouraged to vote.
  - Told that whether one votes is a matter of public record.
  - Shown whether members of their own household voted in the last two elections and promised to send post-card after election indicating whether or not they voted.
- **Neighbors**
  - Like **Self** treatment but in addition recipients are shown whether the neighbors on the block voted in the last two elections.
  - Promised to inform neighbors whether or not subject voted after election.

Example: Social Pressure Experiment



Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______
9999 BRIAN JOSEPH JACKSON		Voted	______
9991 JENNIFER KAY THOMPSON		Voted	______
9991 BOB R THOMPSON		Voted	

Example: Social Pressure Experiment

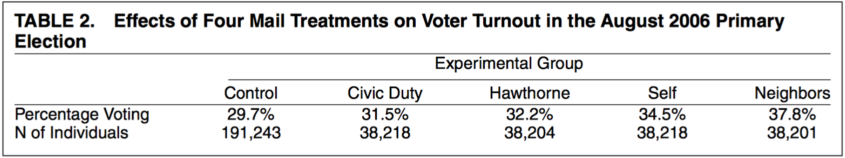


TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Example: Social Pressure Experiment

```
d <- read.dta("gerber.dta")
covars <- c("hh_size","g2002","g2000","p2004","p2002","p2000","sex","yob")
print(aggregate(d[,covars],by=list(d$treatment),mean),digits=3)
```

Example: Social Pressure Experiment

```
d <- read.dta("gerber.dta")
covars <- c("hh_size","g2002","g2000","p2004","p2002","p2000","sex","yob")
print(aggregate(d[,covars],by=list(d$treatment),mean),digits=3)
```

```
     Group.1 hh_size g2002 g2000 p2004 p2002 p2000   sex  yob
1    Control    1.91 0.834 0.866 0.417 0.409 0.265 0.502 1955
2  Hawthorne    1.91 0.836 0.867 0.419 0.412 0.263 0.503 1955
3 Civic Duty    1.91 0.836 0.865 0.416 0.410 0.266 0.503 1955
4  Neighbors    1.91 0.835 0.865 0.423 0.406 0.263 0.505 1955
5       Self    1.91 0.835 0.863 0.421 0.410 0.263 0.501 1955
```

Example: Social Pressure Experiment

```
print(aggregate(d[,covars],by=list(d$treatment),sd),digits=3)
```

Example: Social Pressure Experiment

```
print(aggregate(d[,covars],by=list(d$treatment),sd),digits=3)
```

```
     Group.1 hh_size g2002 g2000 p2004 p2002 p2000   sex  yob
1    Control   0.720 0.294 0.271 0.444 0.435 0.395 0.273 12.9
2  Hawthorne   0.718 0.295 0.270 0.444 0.435 0.393 0.272 12.9
3 Civic Duty   0.729 0.293 0.270 0.444 0.435 0.396 0.275 12.9
4  Neighbors   0.728 0.295 0.273 0.445 0.434 0.393 0.274 13.0
5       Self   0.718 0.294 0.274 0.444 0.434 0.392 0.274 12.8
```

Example: Social Pressure Experiment

```
print(aggregate(d[,c("yob")],by=list(d$treatment),quantile),digits=3)
```

Example: Social Pressure Experiment

```
print(aggregate(d[,c("yob")],by=list(d$treatment),quantile),digits=3)
```

```
    Group.1 x.0% x.25% x.50% x.75% x.100%
1    Control 1900  1946  1957  1964   1986
2  Hawthorne 1908  1946  1957  1964   1984
3 Civic Duty 1906  1947  1957  1964   1986
4  Neighbors 1905  1946  1957  1964   1986
5       Self 1908  1946  1957  1964   1986
```

Example: Social Pressure Experiment

```
form <- as.formula(paste("treatment","~",paste(covars,collapse="+")))
form
treatment ~ hh_size + g2002 + g2000 + p2004 + p2002 + p2000 +
    sex + yob
summary(lm(form,data=d))
```

Example: Social Pressure Experiment

```
form <- as.formula(paste("treatment","~",paste(covars,collapse="+")))
form
treatment ~ hh_size + g2002 + g2000 + p2004 + p2002 + p2000 +
    sex + yob
summary(lm(form,data=d))


              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7944614  0.5496699   3.265   0.0011 **
hh_size     -0.0032727  0.0051836  -0.631   0.5278
g2002        0.0121818  0.0123389   0.987   0.3235
g2000       -0.0233410  0.0133489  -1.749   0.0804 .
p2004        0.0118147  0.0079130   1.493   0.1354
p2002        0.0018055  0.0081488   0.222   0.8247
p2000       -0.0031604  0.0087721  -0.360   0.7186
sex          0.0031331  0.0125052   0.251   0.8022
yob          0.0001671  0.0002815   0.594   0.5528
Residual standard error: 1.449 on 179993 degrees of freedom
Multiple R-squared: 4.004e-05,  Adjusted R-squared: -4.406e-06
F-statistic: 0.9009 on 8 and 179993 DF,  p-value: 0.5145
```

Example: Social Pressure Experiment

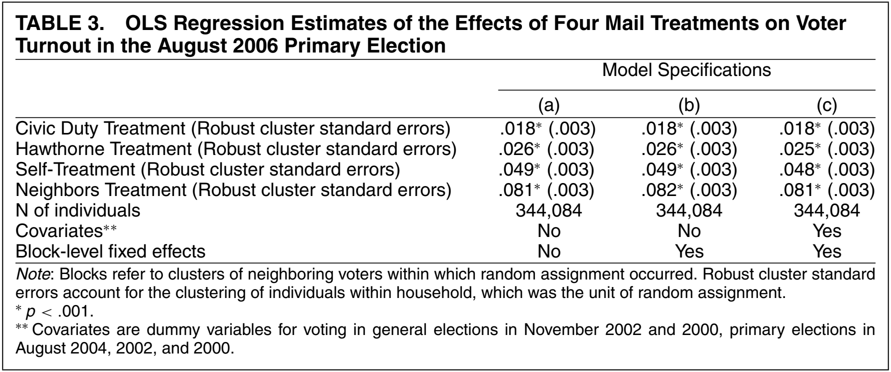


TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Model Specifications		
	(a)	(b)	(c)
Civic Duty Treatment (Robust cluster standard errors)	.018* (.003)	.018* (.003)	.018* (.003)
Hawthorne Treatment (Robust cluster standard errors)	.026* (.003)	.026* (.003)	.025* (.003)
Self-Treatment (Robust cluster standard errors)	.049* (.003)	.049* (.003)	.048* (.003)
Neighbors Treatment (Robust cluster standard errors)	.081* (.003)	.082* (.003)	.081* (.003)
N of individuals	344,084	344,084	344,084
Covariates**	No	No	Yes
Block-level fixed effects	No	Yes	Yes

Note: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.

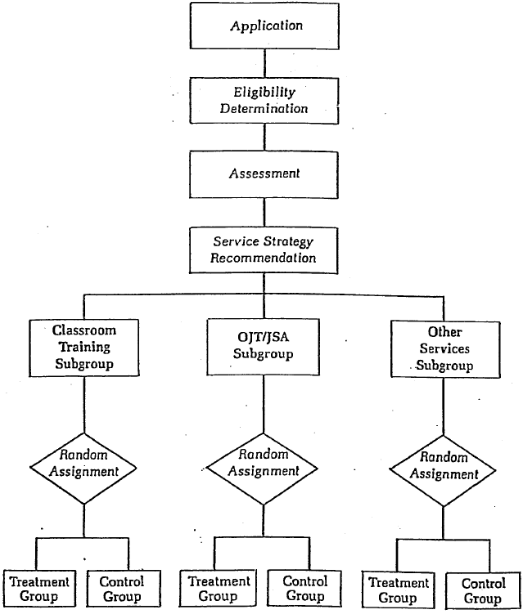
* $p < .001$.

** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

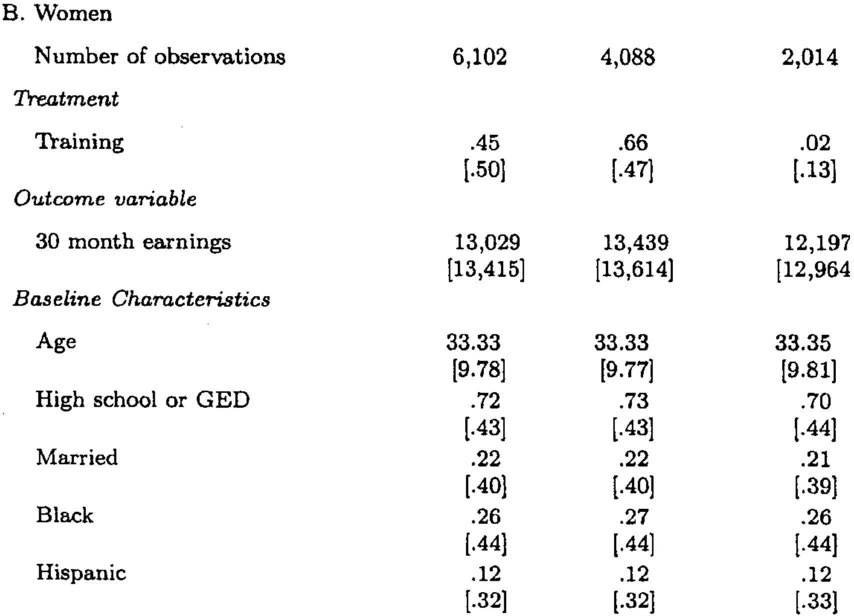
Example: Job Training Partnership Act (JTPA)

- Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country
- Sample: Disadvantaged persons in the labor market (previously unemployed or low earnings)
- D: Assignment to one of three general service strategies
  - classroom training in occupational skills
  - on-the-job training and/or job search assistance
  - other services (eg. probationary employment)
- Y: Earnings 30 months following assignment
- X: Characteristics measured before assignment (age, gender, previous earnings, race, etc.)

Random Assignment Model for JTPA Experiment



Means and Standard Deviations for JTPA Experiment



B. Women			
Number of observations	6,102	4,088	2,014
Treatment			
Training	.45 [.50]	.66 [.47]	.02 [.13]
Outcome variable			
30 month earnings	13,029 [13,415]	13,439 [13,614]	12,197 [12,964
Baseline Characteristics			
Age	33.33 [9.78]	33.33 [9.77]	33.35 [9.81]
High school or GED	.72 [.43]	.73 [.43]	.70 [.44]
Married	.22 [.40]	.22 [.40]	.21 [.39]
Black	.26 [.44]	.27 [.44]	.26 [.44]
Hispanic	.12 [.32]	.12 [.32]	.12 [.33]

Subgroup Effects for JTPA Experiment

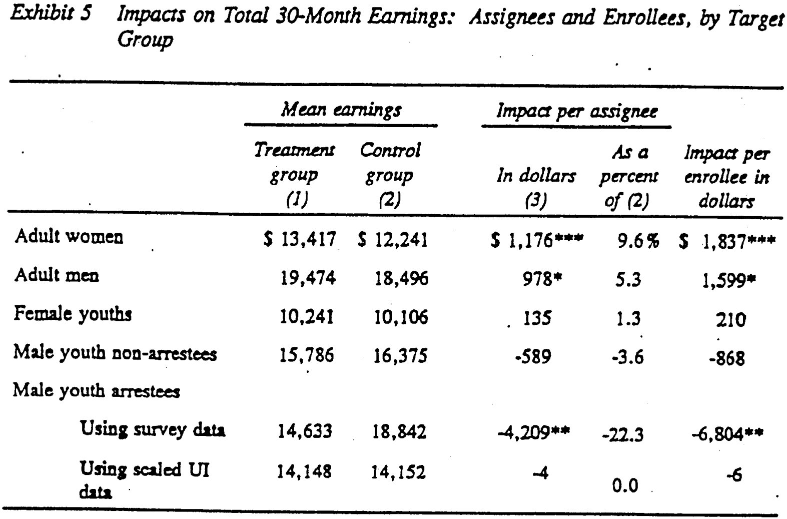


Exhibit 5 *Impacts on Total 30-Month Earnings: Assignees and Enrollees, by Target Group*

	Mean earnings		*Impact per assignee*		
	Treatment group (1)	*Control group (2)*	*In dollars (3)*	*As a percent of (2)*	*Impact per enrollee in dollars*
Adult women	$ 13,417	$ 12,241	$ 1,176***	9.6%	$ 1,837***
Adult men	19,474	18,496	978*	5.3	1,599*
Female youths	10,241	10,106	135	1.3	210
Male youth non-arrestees	15,786	16,375	-589	-3.6	-868
Male youth arrestees					
Using survey data	14,633	18,842	-4,209**	-22.3	-6,804**
Using scaled UI data	14,148	14,152	-4	0.0	-6

A Word about Policy Implications

After the results of the National JTPA study were released, in 1994, funding for JTPA training for the youth were drastically cut:

Spending on JTPA Programs

Year	Youth Training Grants	Adult Training Grants
1993	677	1015
1994	609	988
1995	127	996
1996	127	850
1997	127	895

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Considerations for Experimental Designs

- Unit of analysis and unit of randomization (individuals, groups, institutions, etc)?
  - Choice of analytic level determines what the study has the capacity to demonstrate.
  - Example: randomize school vouchers at the level of the individual or at the level of the community? Do we want to know how students respond to new environment or or how schools respond to competition?
  - Can also help with SUTVA (e.g. interactions within and between schools)

Considerations for Experimental Designs

- Unit of analysis and unit of randomization (individuals, groups, institutions, etc)?
  - Choice of analytic level determines what the study has the capacity to demonstrate.
  - Example: randomize school vouchers at the level of the individual or at the level of the community? Do we want to know how students respond to new environment or or how schools respond to competition?
  - Can also help with SUTVA (e.g. interactions within and between schools)
- How many treatments?
- How many units?
- How many treated and how many controls?
- Is background information available? If so, how can it be used?

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Blocking

- Imagine you have data on the units that you are about to randomly assign. Why leave it to “pure” chance to balance the observed characteristics?
- Idea in blocking is to pre-stratify the sample and then to randomize separately within each stratum to ensure that the groups start out with identical observable characteristics on the blocked factors.
- You effectively run a separate experiment within each stratum, randomization will balance the unobserved attributes
- Why is this helpful?
  - Four subjects with pre-treatment outcomes of $\{2,2,8,8\}$
  - Divided evenly into treatment and control groups and treatment effect is zero
  - Simple random assignment will place $\{2,2\}$ and $\{8,8\}$ together in the same treatment or control group $1/3$ of the time

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

An unbiased estimator for this quantity will be

$$\hat{\tau}_B = \frac{N_f}{N_f + N_m} \cdot \hat{\tau}_f + \frac{N_m}{N_f + N_m} \cdot \hat{\tau}_m$$

Blocking

Imagine you run an experiment where you block on gender. It's possible to think about an ATE composed of two seperate block-specific ATEs:

$$\tau = \frac{N_f}{N_f + N_m} \cdot \tau_f + \frac{N_m}{N_f + N_m} \cdot \tau_m$$

An unbiased estimator for this quantity will be

$$\hat{\tau}_B = \frac{N_f}{N_f + N_m} \cdot \hat{\tau}_f + \frac{N_m}{N_f + N_m} \cdot \hat{\tau}_m$$

or more generally, if there are J strata or blocks, then

$$\hat{\tau}_B = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\tau}_j$$

Blocking

Because the randomizations in each block are independent, the variance of the blocking estimator is simply $(\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y))$:

$$\mathrm{Var}(\hat{\tau}_B) = \left(\frac{N_f}{N_f + N_m}\right)^2 \mathrm{Var}(\hat{\tau}_f) + \left(\frac{N_m}{N_f + N_m}\right)^2 \mathrm{Var}(\hat{\tau}_m)$$

or more generally

$$Var(\hat{\tau}_B) = \sum_{j=1}^{J} \left(\frac{N_j}{N}\right)^2 \mathrm{Var}(\hat{\tau}_j)$$

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block “fixed effects” will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block "fixed effects" will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

If probabilites of treatment, $p_{ij} = P(D_{ij} = 1)$, vary by block, then weight each observation:

$$w_{ij} = \left(\frac{1}{p_{ij}}\right) D_i + \left(\frac{1}{1 - p_{ij}}\right)(1 - D_i)$$

Blocking with Regression

When analyzing a blocked randomized experiment with OLS and the probability of receiving treatment is equal across blocks, then OLS with block "fixed effects" will result in a valid estimator of the ATE:

$$y_i = \tau D_i + \sum_{j=2}^{J} \beta_j \cdot B_{ij} + \epsilon_i$$

where B_j is a dummy for the j-th block (one omitted as reference category).

If probabilites of treatment, $p_{ij} = P(D_{ij} = 1)$, vary by block, then weight each observation:

$$w_{ij} = \left(\frac{1}{p_{ij}}\right) D_i + \left(\frac{1}{1 - p_{ij}}\right)(1 - D_i)$$

Why do this? When treatment probabilities vary by block, then OLS will weight blocks by the variance of the treatment variable in each block. Without correcting for this, OLS will result in biased estimates of ATE!

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i$$

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$\begin{aligned} Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i \qquad (1) \\ Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \qquad (2) \end{aligned}$$

where B_j is a dummy for the j-th block. Then given iid sampling:

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$\begin{aligned} Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i & (1) \\ Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* & (2) \end{aligned}$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$Var[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} (D_i - \bar{D})^2} \qquad \text{with } \widehat{\sigma}_\varepsilon^2 =$$

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$Y_i = \alpha + \tau_{CR} D_i + \varepsilon_i \tag{1}$$

$$Y_i = \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \tag{2}$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$Var[\widehat{\tau}_{CR}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}$$

$$Var[\widehat{\tau}_{BR}] = \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*\,2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}$$

where R_j^2

When Does Blocking Help?

Imagine a model for a complete and blocked randomized design:

$$
\begin{align}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i \tag{1} \\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* \tag{2}
\end{align}
$$

where B_j is a dummy for the j-th block. Then given iid sampling:

$$
\begin{aligned}
Var[\widehat{\tau}_{CR}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} & \text{with } \widehat{\sigma}_\varepsilon^2 &= \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2} \\
Var[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} & \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 &= \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*\,2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon}^*}}{n-k-1}
\end{aligned}
$$

where R_j^2 is R^2 from regression of D on all B_j variables and a constant.

When Does Blocking Help?

$$\begin{aligned} Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i && (3) \\ Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* && (4) \end{aligned}$$

where B_k is a dummy for the k-th block. Then given iid sampling:

$$\begin{aligned} V[\widehat{\tau}_{CR}] &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (D_i - \bar{D})^2} \quad \text{with } \widehat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2} \\ V[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^n (D_i - \bar{D})^2 (1 - R_j^2)} \quad \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^{*2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1} \end{aligned}$$

where R_j^2 is R^2 from regression of D on the B_k dummies and a constant.

So when is $Var[\widehat{\tau}_{BR}] < Var[\widehat{\tau}_{CR}]$?

When Does Blocking Help?

$$
\begin{aligned}
Y_i &= \alpha + \tau_{CR} D_i + \varepsilon_i & (5)\\
Y_i &= \alpha + \tau_{BR} D_i + \sum_{j=2}^{J} \beta_j B_{ij} + \varepsilon_i^* & (6)
\end{aligned}
$$

where B_k is a dummy for the k-th block. Then given iid sampling:

$$
\begin{aligned}
V[\widehat{\tau}_{CR}] &= \frac{\sigma_{\varepsilon}^2}{\sum_{i=1}^{n}(D_i - \bar{D})^2} & \text{with } \widehat{\sigma}_{\varepsilon}^2 &= \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n-2} = \frac{SSR_{\widehat{\varepsilon}}}{n-2}\\
V[\widehat{\tau}_{BR}] &= \frac{\sigma_{\varepsilon^*}^2}{\sum_{i=1}^{n}(D_i - \bar{D})^2(1 - R_j^2)} & \text{with } \widehat{\sigma}_{\varepsilon^*}{}^2 &= \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^{*2}}{n-k-1} = \frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1}
\end{aligned}
$$

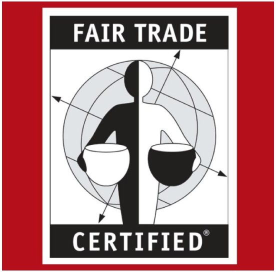
where R_j^2 is R^2 from regression of D on the B_k dummies and a constant.

Since $R_j^2 \approx 0$ $V[\widehat{\tau}_{BR}] < V[\widehat{\tau}_{CR}]$ if $\frac{SSR_{\widehat{\varepsilon^*}}}{n-k-1} < \frac{SSR_{\widehat{\varepsilon}}}{n-2}$

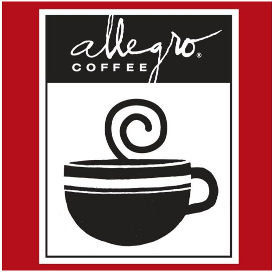
Example: Fair Trade Labeling Experiment

Label Experiment

Treatment



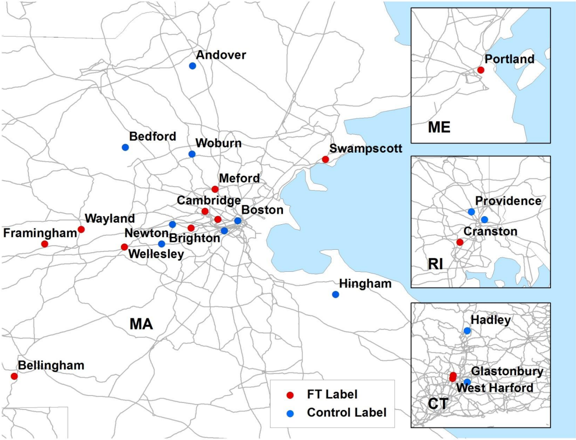
Control



Example: Fair Trade Labeling Experiment

Example: Fair Trade Labeling Experiment

Matched Pairs: Phase 1



Example: Fair Trade Labeling Experiment

R Code

```
> d <- read.dta("FTdata.dta")
> head(d)
  store pair FTweek lnsalesd
1     1    1      1     3.20
2     4    1      0     2.77
3     6    2      1     4.18
4     9    2      0     4.04
5    21    3      1     4.30
6    24    3      0     3.93
```

Example: Fair Trade Labeling Experiment

R Code

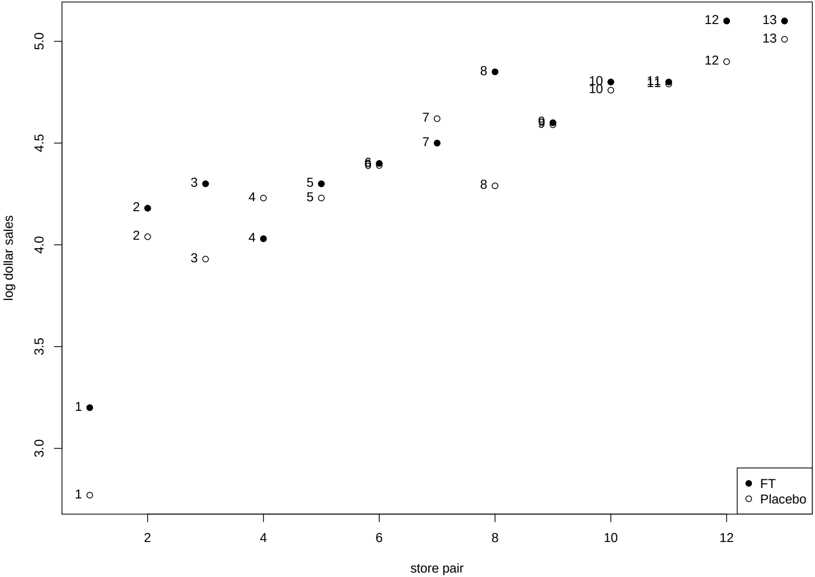
```
> cr.out <- lm(lnsalesd~FTweek,data=d)
> coeftest(cr.out,vcov = vcovHC(cr.out, type = "HC1"))

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.35000    0.16079 27.0537   <2e-16 ***
FTweek       0.12385    0.21424  0.5781   0.5686
---
```

Example: Fair Trade Labeling Experiment

R Code

```
> br.out <- lm(lnsalesd˜FTweek+as.factor(pair),data=d)
> coeftest(br.out,vcov = vcovHC(br.out, type = "HC1"))

t test of coefficients:

                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)       2.923077   0.162144 18.0277 4.671e-10 ***
FTweek            0.123846   0.060176  2.0581 0.0619840 .
as.factor(pair)2  1.125000   0.159549  7.0511 1.335e-05 ***
as.factor(pair)3  1.130000   0.204440  5.5273 0.0001304 ***
as.factor(pair)4  1.145000   0.231925  4.9369 0.0003439 ***
as.factor(pair)5  1.280000   0.161773  7.9123 4.208e-06 ***
as.factor(pair)6  1.410000   0.169987  8.2948 2.591e-06 ***
as.factor(pair)7  1.575000   0.203689  7.7324 5.317e-06 ***
as.factor(pair)8  1.585000   0.277319  5.7154 9.675e-05 ***
as.factor(pair)9  1.610000   0.169987  9.4713 6.420e-07 ***
as.factor(pair)10 1.795000   0.165195 10.8660 1.450e-07 ***
as.factor(pair)11 1.810000   0.169987 10.6479 1.810e-07 ***
as.factor(pair)12 2.015000   0.164183 12.2729 3.763e-08 ***
as.factor(pair)13 2.070000   0.160298 12.9134 2.127e-08 ***
---
```

Example: Fair Trade Labeling Experiment



Example: Fair Trade Labeling Experiment

R Code

```
> summary(lm(lnsalesd~as.factor(pair),data=d))
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.9850     0.1212  24.621 2.72e-12 ***
as.factor(pair)2    1.1250     0.1715   6.562 1.82e-05 ***
as.factor(pair)3    1.1300     0.1715   6.591 1.74e-05 ***
as.factor(pair)4    1.1450     0.1715   6.678 1.52e-05 ***
as.factor(pair)5    1.2800     0.1715   7.466 4.73e-06 ***
as.factor(pair)6    1.4100     0.1715   8.224 1.65e-06 ***
as.factor(pair)7    1.5750     0.1715   9.186 4.77e-07 ***
as.factor(pair)8    1.5850     0.1715   9.245 4.44e-07 ***
as.factor(pair)9    1.6100     0.1715   9.390 3.71e-07 ***
as.factor(pair)10   1.7950     0.1715  10.469 1.05e-07 ***
as.factor(pair)11   1.8100     0.1715  10.557 9.56e-08 ***
as.factor(pair)12   2.0150     0.1715  11.752 2.68e-08 ***
as.factor(pair)13   2.0700     0.1715  12.073 1.94e-08 ***
---
Residual standard error: 0.1715 on 13 degrees of freedom
Multiple R-squared:  0.9474, Adjusted R-squared:  0.8988
F-statistic:  19.5 on 12 and 13 DF,  p-value: 2.356e-06
```

Blocking

- How does blocking help?
  - Increases efficiency if the blocking variables predict outcomes (i.e. they “remove” the variation that is driven by nuisance factors)
  - Blocking on irrelevant predictors can burn up degrees of freedom
  - Can help with small sample bias due to “bad” randomization
  - Is powerful especially in small to medium sized samples

Blocking

- How does blocking help?
    - Increases efficiency if the blocking variables predict outcomes (i.e. they “remove” the variation that is driven by nuisance factors)
    - Blocking on irrelevant predictors can burn up degrees of freedom
    - Can help with small sample bias due to “bad” randomization
    - Is powerful especially in small to medium sized samples
- What to block on?
    - “Block what you can, randomize what you can't”
    - The baseline of the outcome variable and other main predictors
    - Variables desired for subgroup analysis
- How to block?
    - Stratification
    - Pair-matching
    - Check: `blockTools` library

Analysis with Blocking

- **“As ye randomize, so shall ye analyze”** (Senn 2004): Need to account for the method of randomization when performing statistical analysis.
- If using OLS, strata dummies should be included when analyzing results of stratified randomization.
  - If probability of treatment assignment varies across blocks, then weight treated units by probability of being in treatment and controls by the probability of being a control.
- Failure to control for the method of randomization can result in incorrect test size.

Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Relative Sample Sized for Fixed N

If sample sizes are large enough, we can approximate

$$\bar{Y}_1 - \bar{Y}_0 \sim N\left(\mu_1 - \mu_0, \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}\right).$$

Problem

Choose N_1 and N_0, such that $N_1 + N_0 = N$, to minimize the variance of the estimator of the average treatment effect.

Recall that the variance of $\bar{Y}_1 - \bar{Y}_0$ is approximately:

$$var(\bar{Y}_1 - \bar{Y}_0) = \frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}$$

where $p = N_1/N$ is the proportion of treated in the sample.

Relative Sample Sized for Fixed N

Find the value p^* that makes the derivative with respect to p equal to zero:

$$-\frac{\sigma_1^2}{p^{*2}N} + \frac{\sigma_0^2}{(1-p^*)^2N} = 0.$$

Therefore:

$$\frac{1-p^*}{p^*} = \frac{\sigma_0}{\sigma_1},$$

and

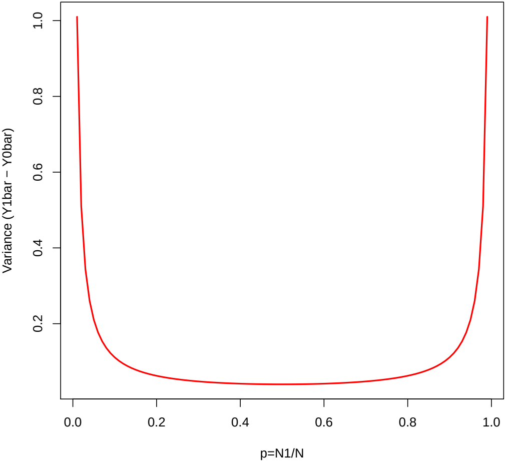
$$p^* = \frac{\sigma_1}{\sigma_1+\sigma_0} = \frac{1}{1+\sigma_0/\sigma_1}$$

A "rule of thumb" for the case $\sigma_1 \approx \sigma_0$ is $p* = 0.5$

For practical reasons it is sometimes better to choose unequal sample sizes (even if $\sigma_1 \approx \sigma_0$). Note: precision erodes slowly until the degree of imbalance becomes extreme ($p < .2$ or $p > .8$), so there is latitude for using an unbalanced allocation.

Variance of ATE as Function of p

Imagine: $\sigma_1^2 = \sigma_0^2 = 1,\ N = 100$



Experimental Design: Power calculations to choose N

- Recall that for a statistical test:
  - Type I error: Rejecting the null if the null is true (α)
  - Type II error: Not rejecting the null if the null is false (Ψ)
- Size of a test is the probability of type I error. Usually 0.05
- Power of a test is one minus the probability of type II error, i.e. the probability of rejecting the null if the null is false
- What does power depend on?

Experimental Design: Power calculations to choose N

- Recall that for a statistical test:
  - Type I error: Rejecting the null if the null is true (α)
  - Type II error: Not rejecting the null if the null is false (Ψ)
- Size of a test is the probability of type I error. Usually 0.05
- Power of a test is one minus the probability of type II error, i.e. the probability of rejecting the null if the null is false
- What does power depend on?
  - True size of the effect (δ)
  - Sample size and proportion of treated (N and p)
  - Variability of outcomes (σ)
  - Desired α level
  - Test statistic
  - Number of treatments

Power calculations with equal and known variances

Suppose that $Y_0 \sim (\mu_0, \sigma_0^2 = \sigma^2)$ and $Y_1 \sim (\mu_1, \sigma_1^2 = \sigma^2)$. Assume also that $p = 0.5$, so $N_0 = N_1 = N/2$. Let $\delta = \mu_1 - \mu_0$. Then, for the t-statistic of equality of means:

$$\frac{\bar{Y}_1 - \bar{Y}_0 - \delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} \sim N(0,1).$$

Therefore:

$$\begin{aligned}
\frac{\bar{Y}_1 - \bar{Y}_0 - \delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} &= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} - \frac{\delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} - \frac{\delta}{\sqrt{\frac{2\sigma^2}{N} + \frac{2\sigma^2}{N}}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} - \frac{\delta}{2\sigma/\sqrt{N}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} - \frac{\delta\sqrt{N}}{2\sigma}
\end{aligned}$$

Therefore:

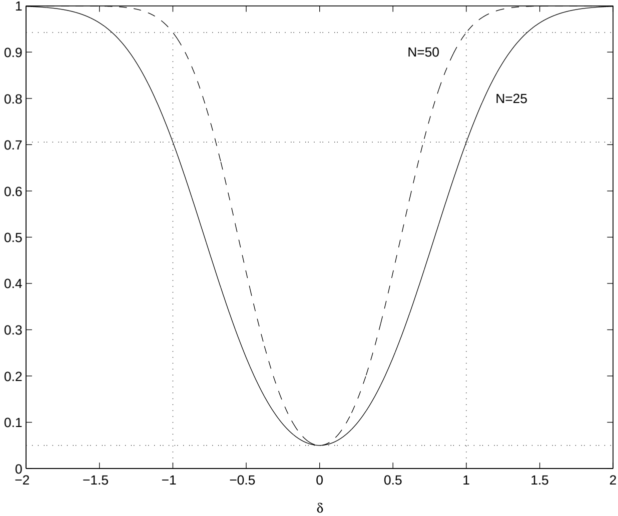
$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}} \sim N\left(\frac{\delta\sqrt{N}}{2\sigma}, 1\right)$$

Power calculations with equal and known variances

The power, i.e. $\Pr(\text{reject } \mu_1 - \mu_0 = 0 | \mu_1 - \mu_0 = \delta)$ is:

$$\begin{aligned}
\Pr(|t| > 1.96) &= \Pr(t < -1.96) + \Pr(t > 1.96) \\
&= \Pr\left(t - \frac{\delta\sqrt{N}}{2\sigma} < -1.96 - \frac{\delta\sqrt{N}}{2\sigma}\right) \\
&+ \Pr\left(t - \frac{\delta\sqrt{N}}{2\sigma} > 1.96 - \frac{\delta\sqrt{N}}{2\sigma}\right) \\
&= \Phi\left(-1.96 - \frac{\delta\sqrt{N}}{2\sigma}\right) + \left(1 - \Phi\left(1.96 - \frac{\delta\sqrt{N}}{2\sigma}\right)\right)
\end{aligned}$$

Power functions for $N = 25$, $N = 50$, and $\sigma^2 = 1$



Note: increasing sample size has a diminishing return for precision.

General formula for the power function ($p \neq 0.5$, $\sigma_0^2 \neq \sigma_1^2$)

$$\begin{aligned}\Pr(\text{reject } \mu_1 - \mu_0 = 0 | \mu_1 - \mu_0 = \delta) \\ &= \Phi\left(-1.96 - \delta \Bigg/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right) \\ &\quad + \left(1 - \Phi\left(1.96 - \delta \Bigg/ \sqrt{\frac{\sigma_1^2}{pN} + \frac{\sigma_0^2}{(1-p)N}}\right)\right).\end{aligned}$$

To choose N we need to specify:

1. δ: minimum detectable effect magnitude
2. Power value (usually 0.80 or higher)
3. σ_1^2 and σ_0^2 (usually $\sigma_1^2 = \sigma_0^2$) (e.g. using previous measures)
4. p: proportion of observations in the treatment group (if $\sigma_1 = \sigma_0$, then the power is maximized by $p = 0.5$)

Formula for Minimum Detectable Effect

Assume $\sigma^2 = \sigma_1^2 = \sigma_0^2$, we can solve for the minimum detectable effect:

$$MDE(\delta) = M_{n-2}\sqrt{\frac{\sigma^2}{Np(1-p)}}$$

where $M_{n-2} = t_{(1-\alpha/2)} + t_{1-\Psi}$ is called the multiplier

- $t_{(1-\alpha/2)}$: critical t-value to reject the null (two-tailed)
- $t_{1-\Psi}$: t-value for t-distribution of the alternative. Depends on desired power $(1-\Psi)$ where Ψ is Pr(type II error)
- E.g. for a two-tailed test with .80 power and $df > 20$ we have approximatly $M_{n-2} = t_{0.975} + t_{.2} = 1.96 + .84 = 2.8$

We can also consider the standardized mean difference effect size ES which is $ES = \frac{\delta}{\sigma}$ and the minimum detectable effect is thus

$$MDES(\delta) = M_{n-2}\sqrt{\frac{1}{Np(1-p)}}$$

Minimum Detectable Effect

Example: Standard deviation is $500 dollars, and average earnings are $2,500 dollars. Here is what we can expect to detect for a given sample size and power.

MDE	MDES	N	SD	Sig	Po	mean Y	MDE/Mean
88.68	0.18	1000	500	0.05	0.8	2500	3.55
125.54	0.25	500	500	0.05	0.8	2500	5.02
282.98	0.57	100	500	0.05	0.8	2500	11.32
404.44	0.81	50	500	0.05	0.8	2500	16.18
585.24	1.17	25	500	0.05	0.8	2500	23.41

- What is the target minimum ES?

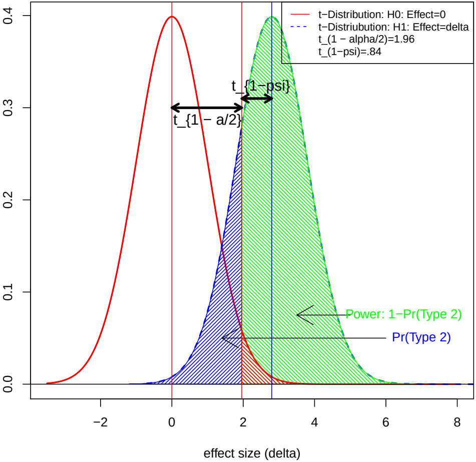
Minimum Detectable Effect

Example: Standard deviation is $500 dollars, and average earnings are $2,500 dollars. Here is what we can expect to detect for a given sample size and power.

MDE	MDES	N	SD	Sig	Po	mean Y	MDE/Mean
88.68	0.18	1000	500	0.05	0.8	2500	3.55
125.54	0.25	500	500	0.05	0.8	2500	5.02
282.98	0.57	100	500	0.05	0.8	2500	11.32
404.44	0.81	50	500	0.05	0.8	2500	16.18
585.24	1.17	25	500	0.05	0.8	2500	23.41

- What is the target minimum ES? Depends on what the benchmark is (theoretical expectations, intervention costs, etc.)
- Popular benchmark for gauging standardized ES is Cohen's (1977) prescription (based on little empirical evidence) that values of 0.20, 0.50, and 0.80 be considered small, moderate, and large.

Multiplier $M_{n-2} = t_{1-\alpha/2} + t_{1-\psi}$



Power Analysis with Blocking

Assuming $\sigma^2 = \sigma_1^2 = \sigma_0^2$, we can solve for the minimum detectable effect:

$$MDES(\delta_{CR}) = M_{n-2}\sqrt{\frac{1}{Np(1-p)}} \tag{7}$$

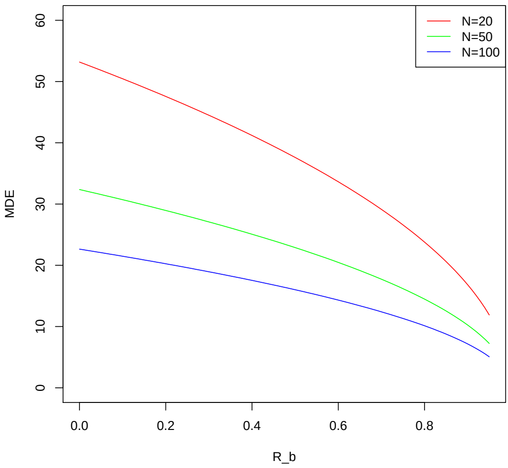
$$MDES(\delta_{BR}) = M_{n-k-1}\sqrt{\frac{1-R_B^2}{Np(1-p)}} \tag{8}$$

- M_{n-2} and M_{n-k-1} are the multipliers
- R_B^2 is the proportion of explained variation in the outcome predicted by the blocks (Regress Y on B_j dummies)
    - The more similar observations are within blocks and the more different blocks are from each other, the higher this predictive power is and the larger the precision gain from blocking.

Power Analysis with Blocking

MDE	MDES	N	SD	Sig	Po	mean Y	MDE/Mean	R
14.52	0.36	26	40	0.05	0.8	90	16.13	0.9
20.53	0.51	26	40	0.05	0.8	90	22.81	0.8
25.15	0.63	26	40	0.05	0.8	90	27.94	0.7
29.04	0.73	26	40	0.05	0.8	90	32.26	0.6
15.93	0.40	22	40	0.05	0.8	90	17.70	0.9
22.53	0.56	22	40	0.05	0.8	90	25.04	0.8
27.60	0.69	22	40	0.05	0.8	90	30.66	0.7
31.87	0.80	22	40	0.05	0.8	90	35.41	0.6

Power Analysis with Blocking $SD = 40$



Outline

1. Identification Under Random Assignment
2. Estimation Under Random Assignment
   - Average Treatment Effect
   - Standard Error for ATE
   - Hypothesis Testing
   - Covariates
   - Hypothesis Testing in Small Samples
3. Examples
4. Further Issues in Experimental Design
   - Blocking
   - Power Calculations
   - Threats to Validity and Ethics

Threats to Internal and External Validity

- Internal validity: can we estimate the treatment effect for our particular sample?
  - Fails when there are differences between treated and controls (other than the treatment itself) that affect the outcome and that we cannot control for
- External validity: can we extrapolate our estimates to other populations?
  - Fails when outside the experimental environment the treatment has a different effect

Most Common Threats to Internal Validity

- Failure of randomization
  - E.g. implementing partners assign their favorites to treatment group, small samples, etc.
    - JTPA: Good balance
- Non-compliance with experimental protocol
  - Failure to treat or "crossover": Some members of the control group receive the treatment and some in the treatment group go untreated
  - Can reduce power significantly
    - JTPA: only about 65% of those assigned to treatment actually enrolled in training (compliance was almost perfect in the control group)
- Attrition
  - Can destroy validity if observed potential outcomes are not representative of all potential outcomes even with randomization
  - E.g. control group subjects are more likely to drop out of a study
    - JTPA: only 3 percent dropped out
- Spillovers
  - Should be dealt with in the design

Most Common Threats to External Validity

- Non-representative sample
  - E.g. laboratory versus field experimentation
  - Subjects are not the same population that will be subject to the policy, known as “randomization bias”
- Non-representative program
  - The treatment differs in actual implementations
  - Scale effects
  - Actual implementations are not randomized (nor full scale)

External Validity? Experimental Sites versus all Sites

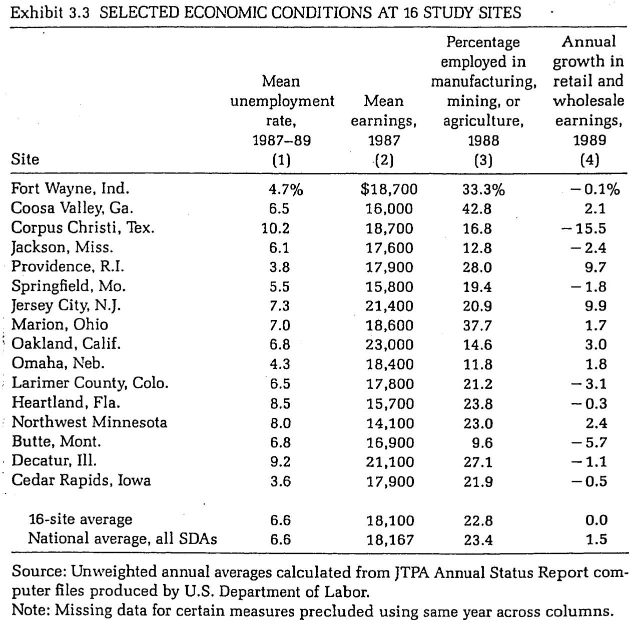


Exhibit 3.3 SELECTED ECONOMIC CONDITIONS AT 16 STUDY SITES

Site	Mean unemployment rate, 1987–89 (1)	Mean earnings, 1987 (2)	Percentage employed in manufacturing, mining, or agriculture, 1988 (3)	Annual growth in retail and wholesale earnings, 1989 (4)
Fort Wayne, Ind.	4.7%	$18,700	33.3%	−0.1%
Coosa Valley, Ga.	6.5	16,000	42.8	2.1
Corpus Christi, Tex.	10.2	18,700	16.8	−15.5
Jackson, Miss.	6.1	17,600	12.8	−2.4
Providence, R.I.	3.8	17,900	28.0	9.7
Springfield, Mo.	5.5	15,800	19.4	−1.8
Jersey City, N.J.	7.3	21,400	20.9	9.9
Marion, Ohio	7.0	18,600	37.7	1.7
Oakland, Calif.	6.8	23,000	14.6	3.0
Omaha, Neb.	4.3	18,400	11.8	1.8
Larimer County, Colo.	6.5	17,800	21.2	−3.1
Heartland, Fla.	8.5	15,700	23.8	−0.3
Northwest Minnesota	8.0	14,100	23.0	2.4
Butte, Mont.	6.8	16,900	9.6	−5.7
Decatur, Ill.	9.2	21,100	27.1	−1.1
Cedar Rapids, Iowa	3.6	17,900	21.9	−0.5
16-site average	6.6	18,100	22.8	0.0
National average, all SDAs	6.6	18,167	23.4	1.5

Source: Unweighted annual averages calculated from JTPA Annual Status Report computer files produced by U.S. Department of Labor.

Note: Missing data for certain measures precluded using same year across columns.

Internal vs. External Validity

Which one is more important?

> One common view is that internal validity comes first. If you do not know the effects of the treatment on the units in your study, you are not well-positioned to infer the effects on units you did not study who live in circumstances you did not study. (Rosenbaum 2010, p. 56)

Randomization addresses internal validity. External validity is often addressed by comparing the results of several internally valid studies conducted in different circumstances and at different times.

The same issues apply in observation studies.

Hardwork is in the Design and Implementation

- Statistics are often easy; the implementation and design are often hard.
- Find partners, manage relationships, identify learning opportunities.
- Designing experiments so that they are incentive-compatible:
    - Free "consulting"
    - Allocating limited resources (e.g. excessively large target groups)
    - Phased randomization as a way to mitigate ethical concerns with denial of treatment
    - Encouragement designs
    - Monitoring
- Potentially high costs.
- Many things can go wrong with complex and large scale experiments.
- Keep it simple in the field!

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

> We received a report that leaders in one community had gathered villagers together after we left and asked people to report how much they had contributed. We moved quickly to prevent any retribution in that village, but also decided to alter the protocol for subsequent games to ensure greater protection for game participants.

Ethics and Experimentation

- Fearon, Humphreys, and Weinstein (2009) used a field experiment to examine if community-driven reconstruction programs foster social reconciliation in post-conflict Liberian villages.
- Outcome: funding raised for collective projects in public goods game played with 24 villagers. Total payout to village is publicly announced.

We received a report that leaders in one community had gathered villagers together after we left and asked people to report how much they had contributed. We moved quickly to prevent any retribution in that village, but also decided to alter the protocol for subsequent games to ensure greater protection for game participants.

These changes included stronger language about the importance of protecting anonymity, random audits of community behavior, facilitation of anonymous reporting of violations of game protocol by participants, and a new opportunity to receive supplemental funds in a postproject lottery if no reports of harassment were received.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
  - Informed consent often done as part of the baseline survey.
  - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
    - Informed consent often done as part of the baseline survey.
    - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.
- **Beneficence**: Avoid knowingly doing harm. Does not mean that all risk can be eliminated, but possible risks must be balanced against overall benefits to society of the research.
    - Note that the existence of a control group might be construed as denying access to some benefit.
    - But without a control group, generating reliable knowledge about the efficacy of the intervention may be impossible.

Ethics and Experimentation

- **Respect for persons**: Participants in most circumstances must give informed consent.
    - Informed consent often done as part of the baseline survey.
    - If risks are minimal and consent will undermine the study, then informed consent rules can be waived.
- **Beneficence**: Avoid knowingly doing harm. Does not mean that all risk can be eliminated, but possible risks must be balanced against overall benefits to society of the research.
    - Note that the existence of a control group might be construed as denying access to some benefit.
    - But without a control group, generating reliable knowledge about the efficacy of the intervention may be impossible.
- **Justice**: Important to avoid situations where one group disproportionately bears the risks and another stands to received all the benefits.
    - Evaluate interventions that are relevant to the subject population

Ethics and Experimentation

- IRB approval is required in almost all circumstances.
- If running an experiment in another country, you need to follow the local regulations on experimental research.
  - Often poorly adapted to social science.
  - Or legally murky whether or not approval is required.
- Still many unanswered questions and lack of consensus on the ethics of field experimentation within Political Science!
  - Be prepared to confront wildly varying opinons on these issues.

Conclusion: Experiments

- Random assignment solves the identification problem for causal inference based on minimal assumptions that we can control as researchers
- Random assignment balances observed and unobserved confounders, which is why it is considered the gold standard for causal inference
- Statistical analysis is simple, transparent, and results are typically not model dependent, since confounders are controlled for "by design"
- Design features can help to improve inferences
- Always important to think about theory and external validity prior to experimentation