Statistical Foundations II

Department of Government
London School of Economics and Political Science

Administrative Stuff

Administrative Stuff

1. Summative Essay Deadline
   - Current: Tuesday MT Week 11
   - Option A: Tuesday LT Week 1
   - Option B: Tuesday LT Week 2

Administrative Stuff

1. Summative Essay Deadline
   - Current: Tuesday MT Week 11
   - Option A: Tuesday LT Week 1
   - Option B: Tuesday LT Week 2
2. Topics for Weeks 6–11?

Definitions

1. **Unit**: A physical object at a particular point in time
2. **Treatment**: An intervention, whose effect(s) we wish to assess relative to some other (non-)intervention
3. **Outcome**: The variable we are trying to explain
4. **ATE**: The comparison between average potential outcomes under each intervention

Banerjee et al

What are the following in this experiment:

1. **Unit**: ?
2. **Treatment**: ?
3. **Outcome**: ?
4. **ATE**: ?

What else should we know about this experiment?

Randomization Inference I

- The randomization (or permutation) distribution is an empirical sampling distribution
- It conveys the variation we would observe in $\widehat{ATE}$ if a null hypothesis, $H_0 : ATE = 0$ was true
- If this null hypothesis is true, then treatment had no effect; the variation in permuted ATEs therefore only reflects sampling variance

Randomization Distribution

The randomization distribution is the vector of all possible ATEs that could be observed in the dataset under rerandomization:

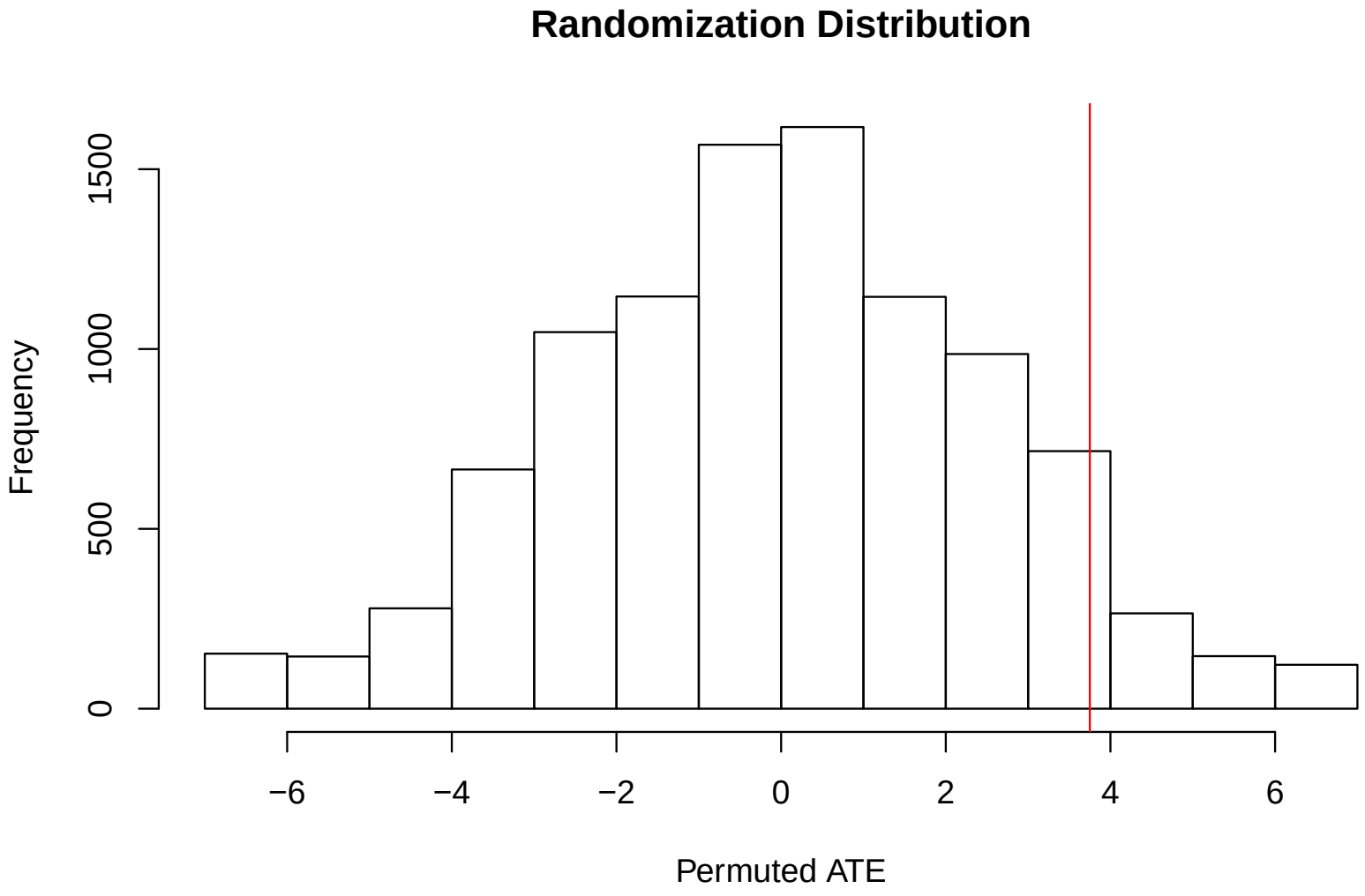
Randomization	ATE
1	3.25
2	-1.50
3	0.75
4	...
...	...

In a two-condition experiment, the number of possible permutations is given by $\binom{n}{n_1}$.

Randomization Inference II

Randomization inference works as follows:

1. Generate every possible randomization scheme
   - Or sample from all possible randomizations
2. Calculate ATE under each randomization
3. The distribution of those estimates is the randomization distribution
4. Its variance is $\widehat{Var}(ATE)$
5. Proportion of values further from 0 than the observed $\widehat{ATE}$ is the p-value for a test of the null hypothesis ($H_0 : ATE = 0$)


Randomization Distribution
Frequency
0
500
1000
1500
−6
−4
−2
0
2
4
6
Permuted ATE

Randomization Inference in R

```
# construct data
d <- data.frame(x = c(0,0,0,0,1,1,1,1),
                y = c(5,7,9,4,11,4,13,12))

# calculate ATE from each randomization
set.seed(1)    # set random number seed
n <- 10000     # number of randomizations
rd <- replicate(n, coef(lm(d$y ~ sample(d$x, 8)))[2L])

# visualize the randomization distribution
hist(rd)
abline(v = coef(lm(y~x, data = d))[2L], col = "red")

# one-tailed significance test
sum(rd >= coef(lm(y ~ x, data = d))[2L])/n
# two-tailed significance test
sum(abs(rd) >= coef(lm(y ~ x, data = d))[2L])/n
```

Parametric Analysis Stata/R

R:

```
t.test(outcome ~ treatment, data = data)
lm(outcome ~ factor(treatment), data = data)
```

Stata:

```
ttest outcome, by(treatment)
reg outcome i.treatment
```

Questions?

Intuition about Variance

- Basic intuition:
  - Bigger sample $\rightarrow$ smaller SEs
  - Smaller variance $\rightarrow$ smaller SEs
- Other design features also matter
- Why do we care?

Statistical Power

- Power analysis is used to determine sample size before conducting an experiment
- Type I and Type II Errors

	H_0 False ($\lvert ATE \rvert > 0$)	H_0 True ($ATE = 0$)
Reject H_0	**True positive**	Type I Error
Accept H_0	Type II Error	True zero

  - True positive rate $(1 - \kappa)$ is power
  - False positive rate is the significance threshold (α)

Doing a Power Analysis

- μ, Treatment group mean outcomes
- n, Sample size
- σ, Outcome variance
- α Statistical significance threshold
- ϕ, a sampling distribution

$$\textit{Power} = 1 - \kappa = \phi\left(\frac{|\mu_1 - \mu_0|\sqrt{n}}{2\sigma} - \phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

(You don't need to know this formula!)

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, "true" ATE, variance of outcome measure, power $(1 - \kappa)$, and α.

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, "true" ATE, variance of outcome measure, power $(1 - \kappa)$, and α.

In essence: some non-zero effect sizes are not detectable by a study of a given sample size.

Intuition about Power

Minimum detectable effect is the smallest effect we could detect given sample size, "true" ATE, variance of outcome measure, power $(1 - \kappa)$, and α.

In essence: some non-zero effect sizes are not detectable by a study of a given sample size.

In underpowered study, we will be unlikely to detect true small effects. And most effects are small! [1]

[1] Gelman, A. and Weakliem, D. 2009. "Of Beauty, Sex and Power." *American Scientist* 97(4): 310–16

Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- Intuition: How large is the effect in standard deviations of the outcome?
  - Know if effects are large or small
  - Compare effects across studies

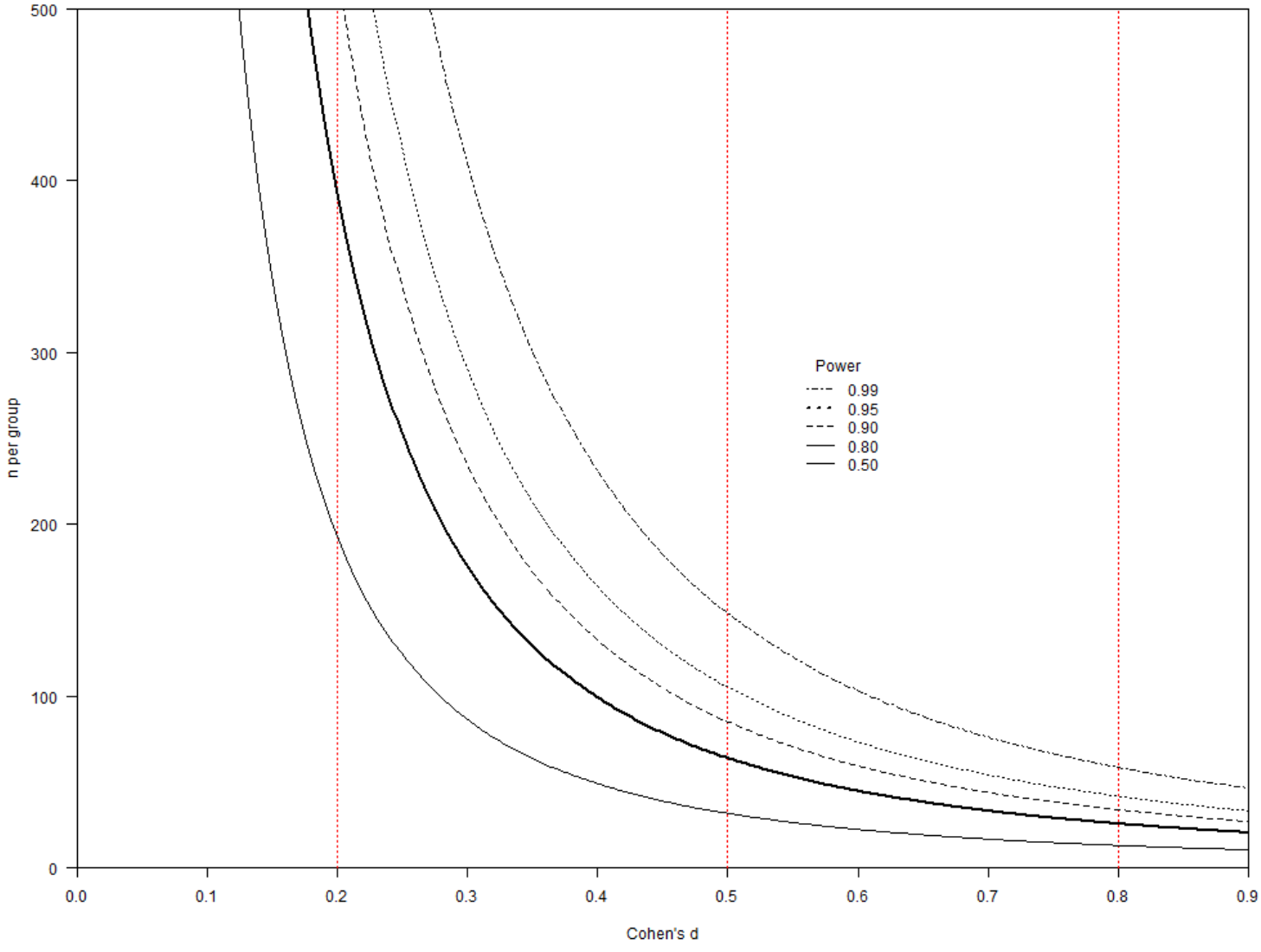
Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- Intuition: How large is the effect in standard deviations of the outcome?
  - Know if effects are large or small
  - Compare effects across studies
- Cohen's d:
  $d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, where $s = \sqrt{\frac{(n_1-1)s_1^2+(n_0-1)s_0^2}{n_1+n_0-2}}$

Intuition about Power

- It can help to think in terms of "standardized effect sizes"
- Intuition: How large is the effect in standard deviations of the outcome?
  - Know if effects are large or small
  - Compare effects across studies
- Cohen's d:
  $d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, where $s = \sqrt{\frac{(n_1-1)s_1^2+(n_0-1)s_0^2}{n_1+n_0-2}}$
- Small: 0.2; Medium: 0.5; Large: 0.8

Intuition about Power



Power analysis in R I

```
power.t.test(
  # sample size (leave blank!)
  n = ,

  # minimum detectable effect size
  delta = 0.4, sd = 1,

  # alpha and power (1-kappa)
  sig.level = 0.05, power = 0.8,

  # two-tailed vs. one-tailed test
  alternative = "two.sided"
)
```

Power analysis in R II

```
# Given a sample size, what is the MDE?
power.t.test(n = 50, power = 0.8)

# Given a sample size and MDE, what is power?
power.t.test(n = 50, delta = 0.2)
```

Increasing/Decreasing Power

Increases Power

- Bigger sample
- Precise measures
- Covariates?

Decreases Power

- Attrition
- Noncompliance
- Clustering

Covariates in Experiments

Covariates in Experiments

- Identification of a causal effect only requires randomization
- We don't need to include covariates in analysis!

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

$$Y = \beta_0 + \beta_1 X + \beta_{2\_J} Z + \epsilon \tag{2}$$

Covariates in Experiments

- Identification of a causal effect only requires randomization
- We don't need to include covariates in analysis!

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

$$Y = \beta_0 + \beta_1 X + \beta_{2\_J} Z + \epsilon \tag{2}$$

Covariates in Experiments

- Identification of a causal effect only requires randomization
- We don't need to include covariates in analysis!

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$
$$Y = \beta_0 + \beta_1 X + \beta_{2\_J} Z + \epsilon \tag{2}$$

- Independence of potential outcomes from treatment assignment is an *asymptotic* property of randomization!

Block Randomization I

Block Randomization I

- Basic idea: randomization occurs within strata defined *before* treatment assignment

Block Randomization I

- Basic idea: randomization occurs within strata defined *before* treatment assignment
- CATE is estimate for each stratum; aggregated to SATE

Block Randomization I

- Basic idea: randomization occurs within strata defined *before* treatment assignment
- CATE is estimate for each stratum; aggregated to SATE
- Why?
  - Eliminate chance imbalances
  - Optimized for estimating CATEs
  - More precise SATE estimate

Block Randomization I

- Basic idea: randomization occurs within strata defined *before* treatment assignment
- CATE is estimate for each stratum; aggregated to SATE
- Why?
  - Eliminate chance imbalances
  - Optimized for estimating CATEs
  - More precise SATE estimate

Stratification:Sampling::Blocking:Experiments

Exp.	Control				Treatment			
1	M	M	M	M	F	F	F	F
2	M	M	M	F	M	F	F	F
3	M	M	F	F	M	M	F	F
4	M	F	F	F	M	M	M	F
5	F	F	F	F	M	M	M	M

Obs.	X_{1i}	X_{2i}	D_i
1	Male	Old	0
2	Male	Old	1
3	Male	Young	1
4	Male	Young	0
5	Female	Old	1
6	Female	Old	0
7	Female	Young	0
8	Female	Young	1

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
  - Most valuable in small samples

Block Randomization II

- Blocking ensures ignorability of all covariates used to construct the blocks
- Incorporates covariates explicitly into the *design*
- When is blocking *statistically* useful?
  - If those covariates affect values of potential outcomes, blocking reduces the variance of the SATE
  - Most valuable in small samples
  - Not valuable if all blocks have similar potential outcomes

Statistical Properties I

Complete randomization:

$$SATE = \frac{1}{n_1}\sum Y_{1i} - \frac{1}{n_0}\sum Y_{0i}$$

Block randomization:

$$SATE_{blocked} = \sum_{1}^{J}\left(\frac{n_j}{n}\right)(\widehat{CATE}_j)$$

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	
8	Female	Young	1	9	

Obs.	X_{1i}	X_{2i}	D_i	Y_i	CATE
1	Male	Old	0	5	5
2	Male	Old	1	10	
3	Male	Young	1	4	3
4	Male	Young	0	1	
5	Female	Old	1	6	4
6	Female	Old	0	2	
7	Female	Young	0	6	3
8	Female	Young	1	9	

SATE Estimation

$$\begin{aligned} SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\ &= 3.75 \end{aligned}$$

SATE Estimation

$$\begin{aligned} SATE &= \left(\frac{2}{8} * 5\right) + \left(\frac{2}{8} * 3\right) + \left(\frac{2}{8} * 4\right) + \left(\frac{2}{8} * 3\right) \\ &= 3.75 \end{aligned}$$

The blocked and unblocked estimates are the same here because $Pr(Treatment)$ is constant across blocks and blocks are all the same size.

SATE Estimation

- We can use weighted regression to estimate this in an OLS framework
- Weights are the inverse prob. of being treated w/in block
  - Pr(Treated) by block: $p_{ij} = Pr(D_i = 1|J = j)$
  - Weight (Treated): $w_{ij} = \frac{1}{p_{ij}}$
  - Weight (Control): $w_{ij} = \frac{1}{1 - p_{ij}}$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(CATE_j)}$$

Statistical Properties II

Complete randomization:

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{n_0} + \frac{\widehat{Var}(Y_1)}{n_1}}$$

Block randomization:

$$\widehat{SE}_{SATE_{blocked}} = \sqrt{\sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2 \widehat{Var}(CATE_j)}$$

When is the blocked design more efficient?

Questions?

Baseline Outcome Measure

- Recall our key definition:

  The observation of units after, and possibly before, a randomly assigned intervention in a controlled setting, which tests one or more precise causal expectations

Baseline Outcome Measure

- Recall our key definition:

  The observation of units after, ***and possibly before****, a randomly assigned intervention in a controlled setting, which tests one or more precise causal expectations*

Baseline Outcome Measure

- Recall our key definition:

  The observation of units after, ***and possibly before****, a randomly assigned intervention in a controlled setting, which tests one or more precise causal expectations*

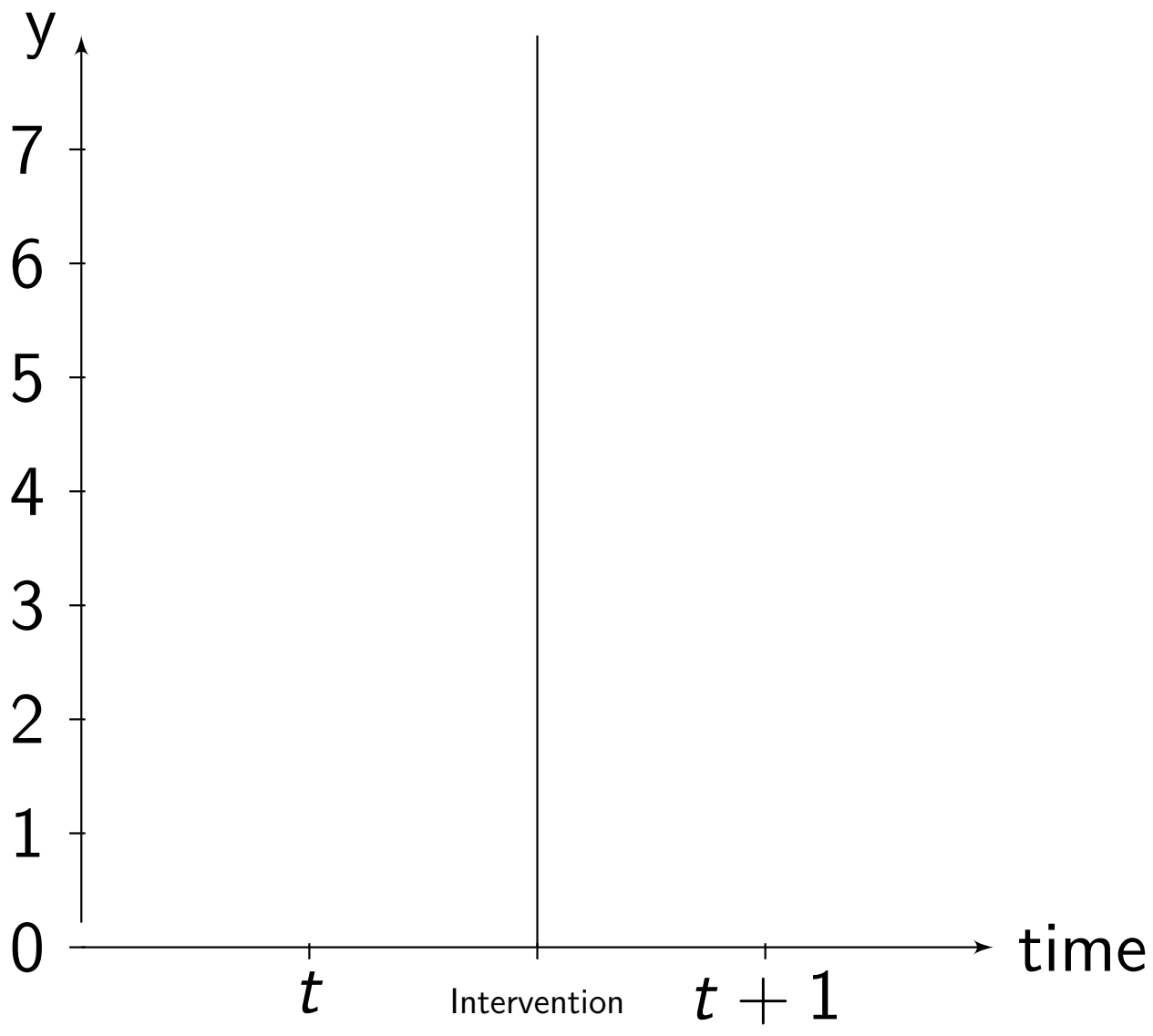
- Pretreatment measures of the outcome can be particularly helpful!

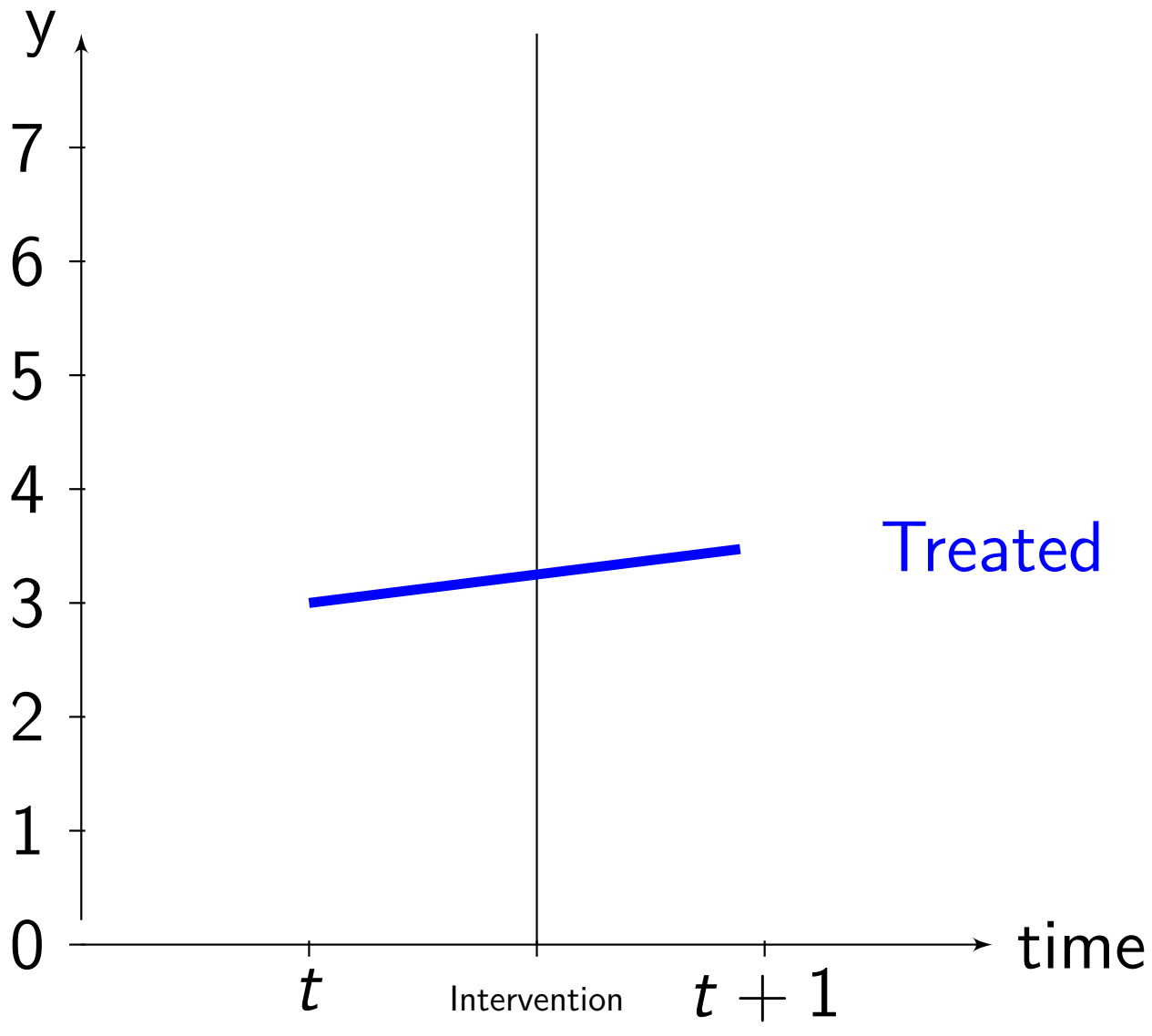
Baseline Outcome Measure

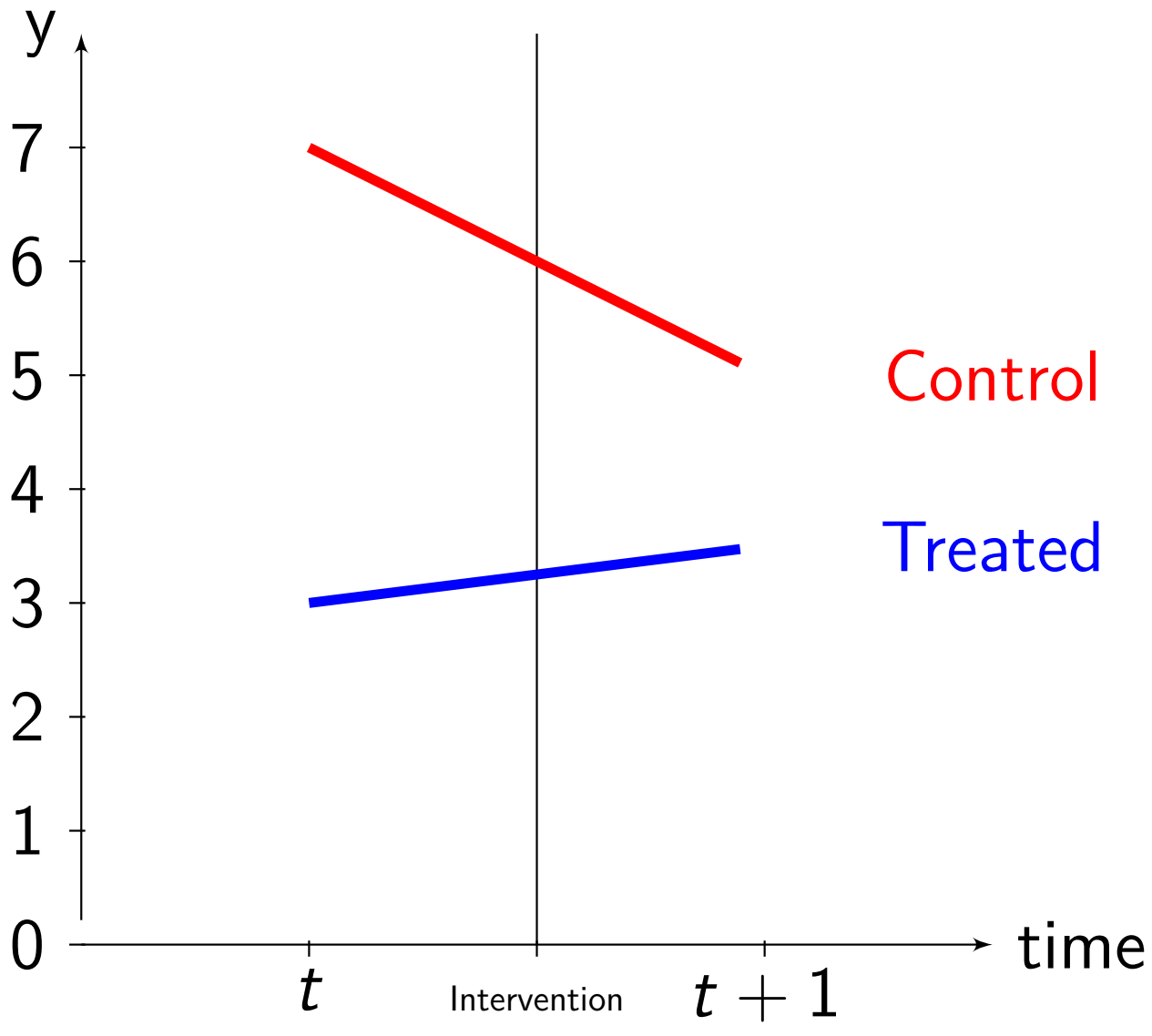
- This changes our estimator of *ATE* from simple *mean-difference* to *difference-in-differences* (DID)

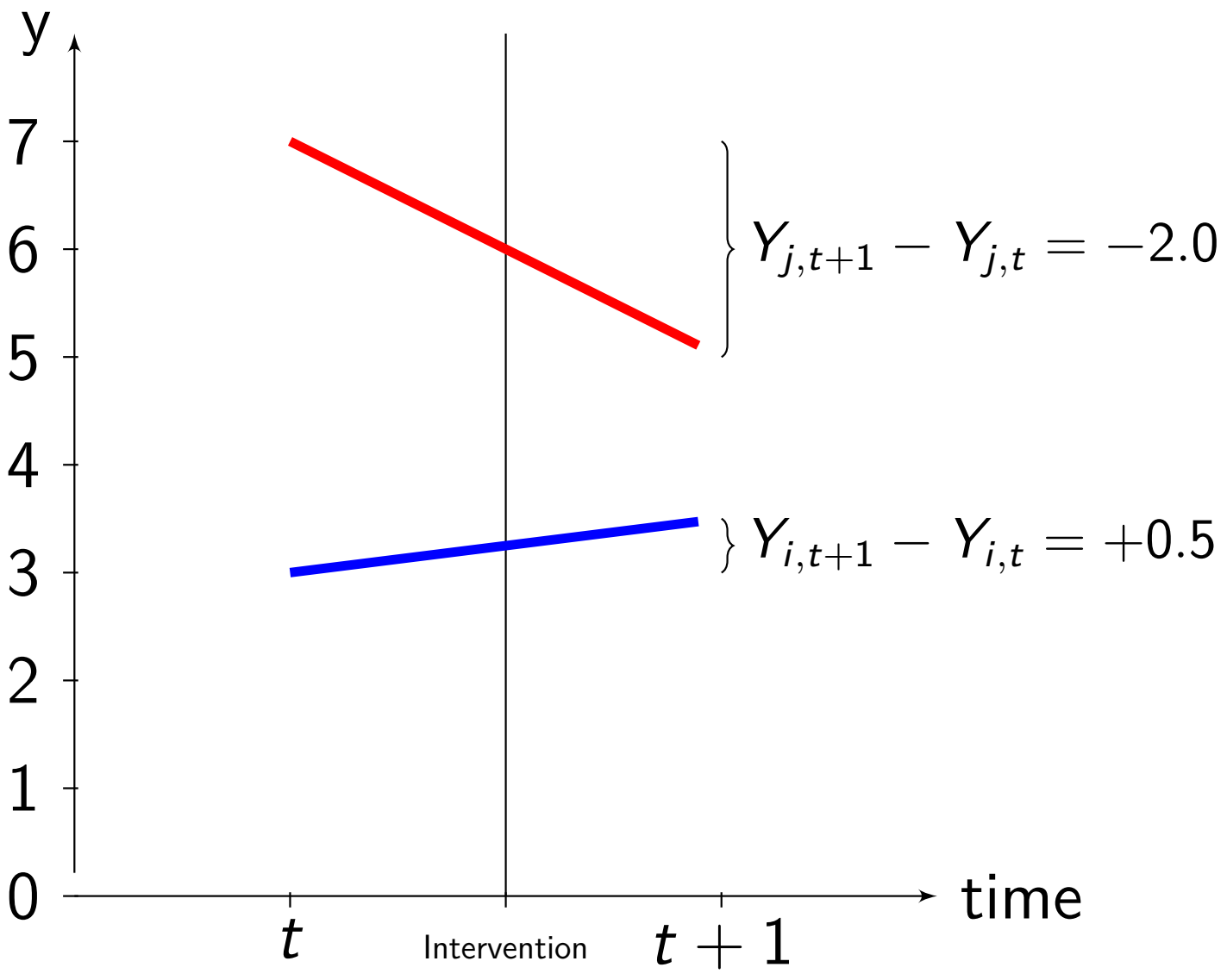
$$(\hat{Y}_{0,t+1} - \hat{Y}_{0,t}) - (\hat{Y}_{j,t+1} - \hat{Y}_{j,t})$$

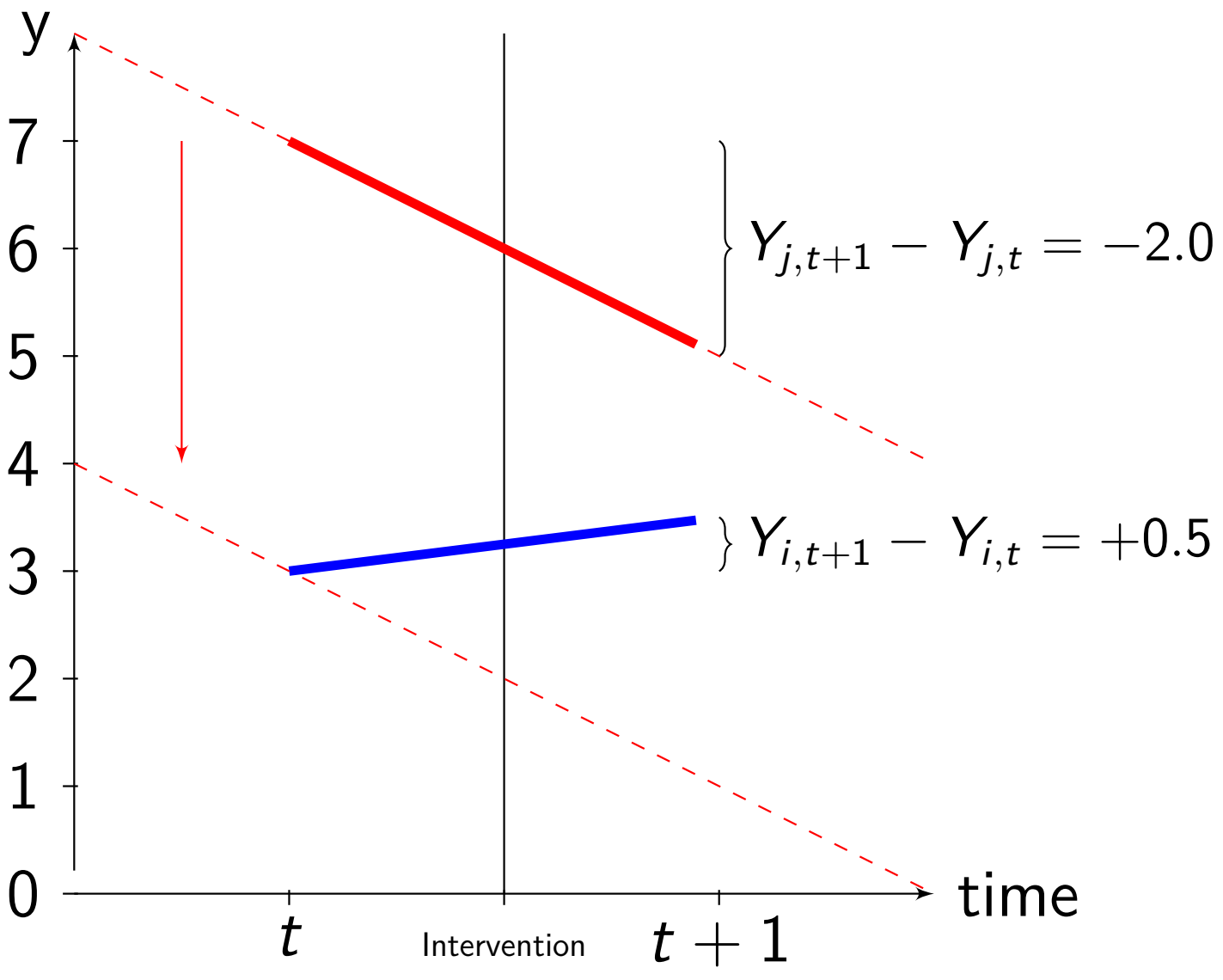
- Advantageous because variance for paired samples decreases as correlation between Y_0 and Y_1 increases

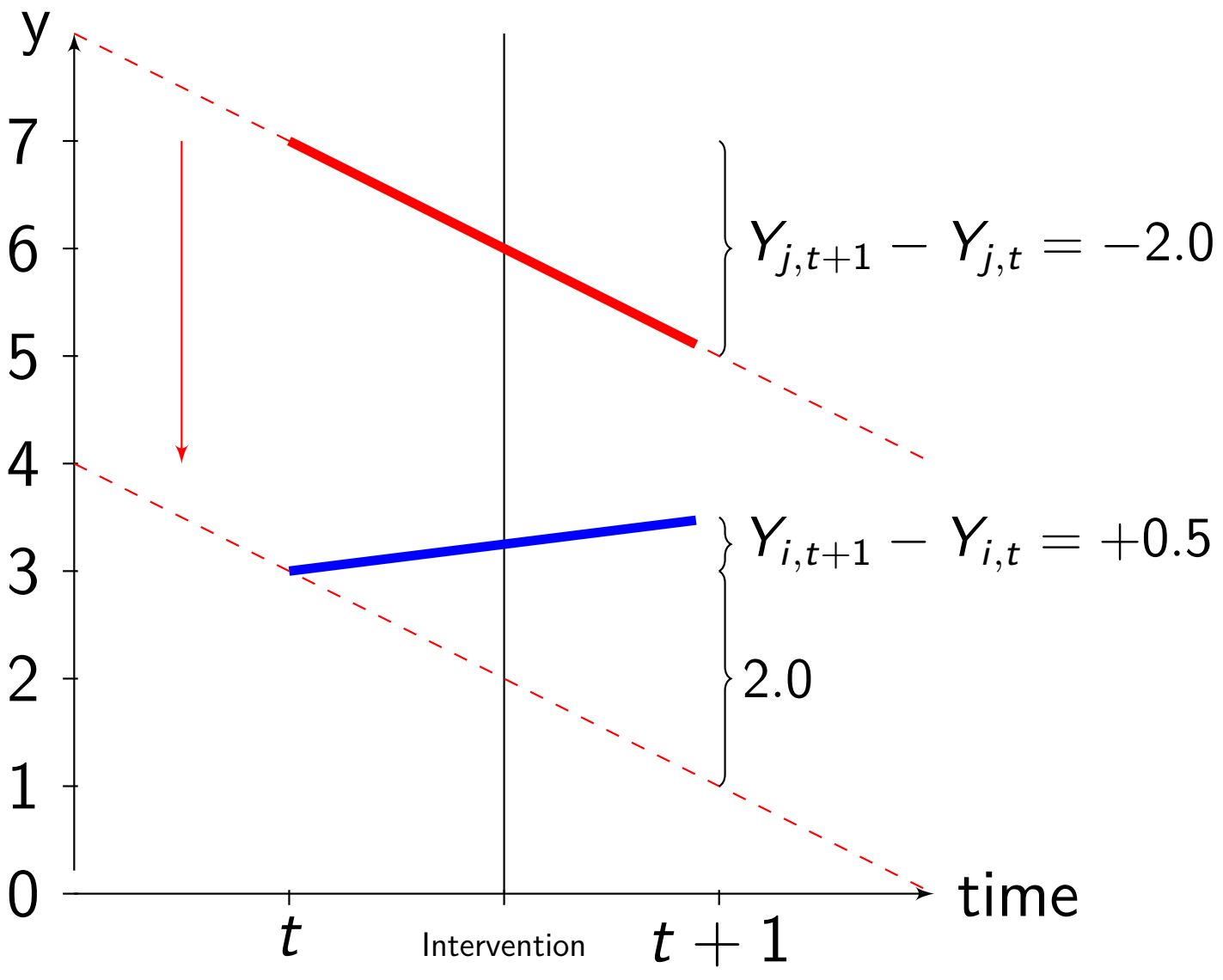

y
7
6
5
4
3
2
1
0
time
t
Intervention
$t+1$

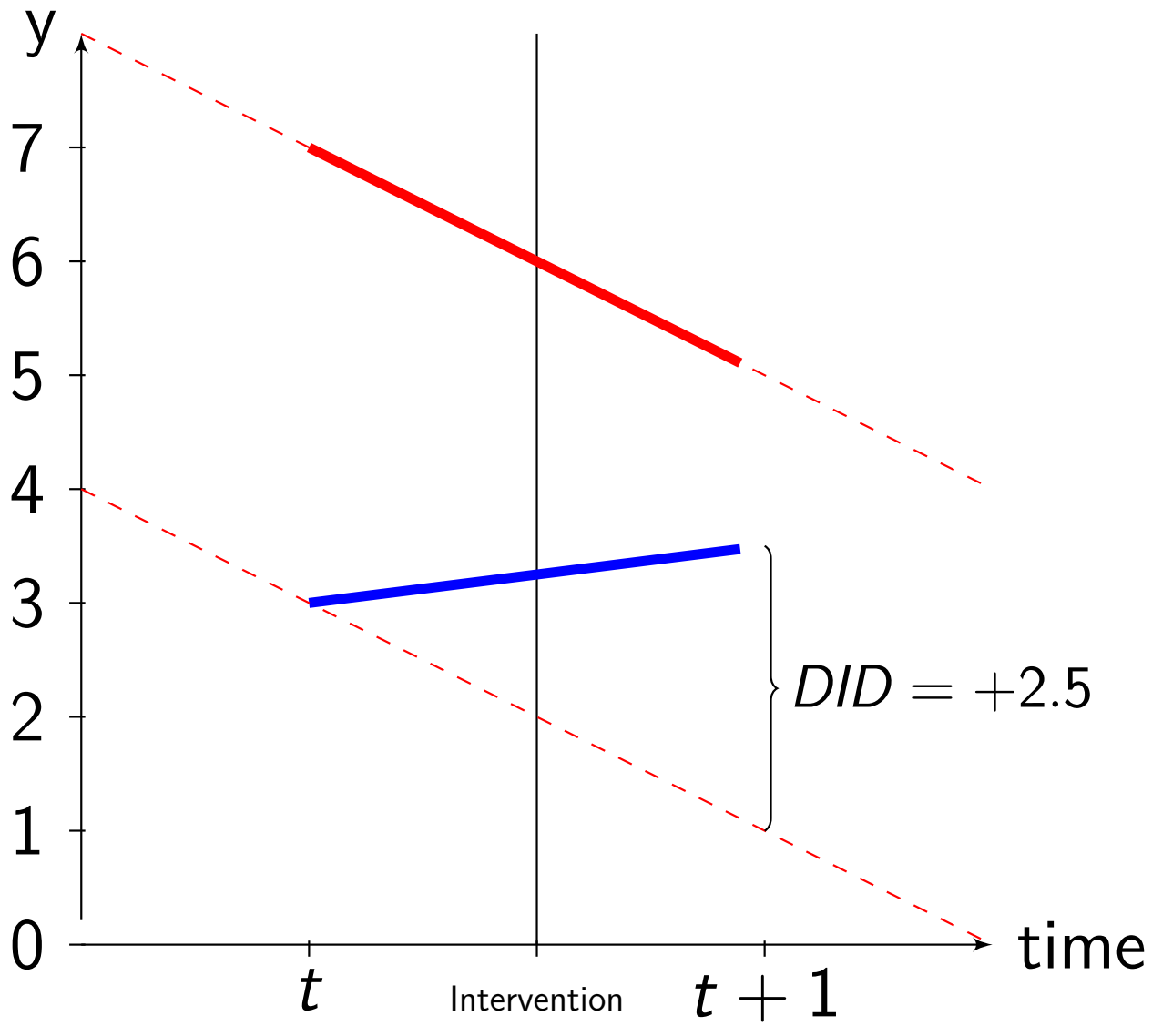

y
7
6
5
4
3
2
1
0
Treated
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
Control
Treated
time
t
Intervention
$t+1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
time
t
Intervention
$t + 1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
time
t
Intervention
$t + 1$


y
7
6
5
4
3
2
1
0
$Y_{j,t+1} - Y_{j,t} = -2.0$
$Y_{i,t+1} - Y_{i,t} = +0.5$
2.0
time
t
Intervention
$t + 1$


y
7
6
5
4
3
2
1
0
time
t
Intervention
$t+1$
$DID = +2.5$

Statistical Advantages I

In post-treatment-only designs:

$$\widehat{ATE}_{Diff} = \frac{\sum_{i=1}^{n_1}(x_{i,1,t+1})}{n_1} - \frac{\sum_{i=1}^{n_0}(x_{i,0,t+1})}{n_0}$$

The variance of this estimate is:

$$Var(\widehat{ATE}_{Diff}) = Var(\bar{Y}_{1,t+1}) + Var(\bar{Y}_{0,t+1})$$

Statistical Advantages II

In pre/post-treatment designs:

$$\widehat{ATE}_{DID} = \frac{\sum_{i=1}^{n_1}(x_{i,1,t+1} - x_{i,1,t})}{n_1} - \frac{\sum_{i=1}^{n_0}(x_{i,0,t+1} - x_{i,0,t})}{n_0}$$

The variance of this estimate is:

$$\begin{aligned} Var(\widehat{ATE}_{DID}) &= Var(\bar{Y}_{1,t+1} - \bar{Y}_{1,t}) + Var(\bar{Y}_{0,t+1} - \bar{Y}_{0,t}) \\ &= \Big(Var(\bar{Y}_{1,t+1}) + Var(\bar{Y}_{1,t}) - Cov(\bar{Y}_{1,t+1}, \bar{Y}_{1,t})\Big) \\ &+ \Big(Var(\bar{Y}_{0,t+1}) + Var(\bar{Y}_{0,t}) - Cov(\bar{Y}_{0,t+1}, \bar{Y}_{0,t})\Big) \end{aligned}$$

```
# create some fake data
set.seed(54321)
n <- 400L
y0 <- rnorm(n)
x <- rbinom(n, 1L, 0.5)

# high Cor(y0, y1)
y1a <- y0 + 0.25*x + rnorm(n, sd = 0.25)
summary(lm(y1a ~ x))
summary(lm(I(y1a-y0) ~ x))

# low Cor(y0, y1)
y1b <- y0 + 0.25*x + rnorm(n, sd = 2)
summary(lm(y1b ~ x))
summary(lm(I(y1b-y0) ~ x))
```

Practicalities

- Blocked randomization and use of pre-treatment measures only works in some circumstances
- Need to observe covariates pre-treatment in order to block on them
  - Challenging in a cross-sectional design
- The cost of gathering pre-treatment data might also outweigh the gain in precision
  - May introduce other biases

Questions?

Clustering

- Everything so far assumes units are *independent*
- Sometimes units are obviously not independent
  - e.g., Students within classrooms
- Non-independence limits our ability to randomize at the unit level and reduces statistical power