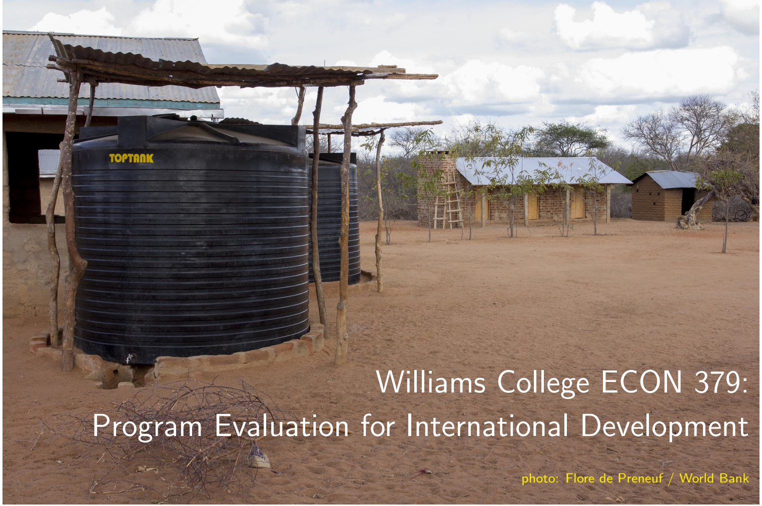
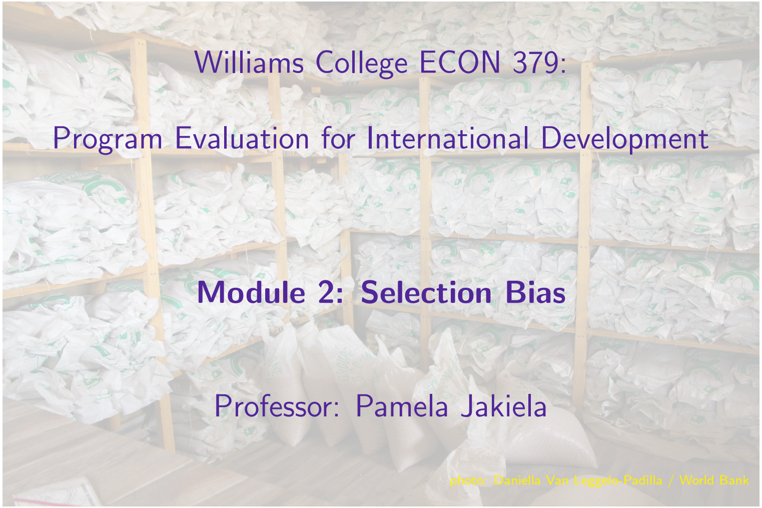


Williams College ECON 379:
Program Evaluation for International Development

photo: Flore de Preneuf / World Bank



Williams College ECON 379:

Program Evaluation for International Development

Module 2: Selection Bias

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

Potential Outcomes

Do Hospitals Make People Healthier?

Your health status is: excellent, very good, good, fair, or poor?

	Hospital	**No Hospital**	**Difference**
Health status	3.21	3.93	-0.72^{***}
	(0.014)	(0.003)	
Observations	7,774	90,049	

A comparison of means suggests hospitals make people worse off: those who had a hospital stay in the last 6 months are, on average, less healthy than those that were not admitted to the hospital

- What's wrong with this picture?

Potential Outcomes

We are interested in the relationship between "**treatment**" and some outcome that may be impacted by the treatment (eg. health)

Outcome of interest:

- Y = outcome we are interested in studying (e.g. health)
- Y_i = value of outcome of interest *for individual i*

For each individual, there are two **potential outcomes**:

- $Y_{0,i}$ = i's outcome if she **doesn't** receive treatment
- $Y_{1,i}$ = i's outcome if she **does** receive treatment

Potential Outcomes: Example

Alejandro has a broken leg.

- $Y_{0,a}$ = If he doesn't go to the hospital, his leg won't heal
- $Y_{1,a}$ = If he goes to the hospital, his leg heals completely

Benicio doesn't have any broken bones. His health is fine.

- $Y_{0,b}$ = If he doesn't go to the hospital, his health is still fine
- $Y_{1,b}$ = If he goes to the hospital, his health is still fine

Potential Outcomes: Example

	Yes Hospital	**No Hospital**
Alejandro	$Y_{1,a}$	$Y_{0,a}$
Benicio	$Y_{1,b}$	$Y_{0,b}$

The Fundamental Problem of Causal Inference

The fundamental problem of causal inference:

We never observe both potential outcomes for the same individual

$\Rightarrow$ Creates a missing data problem whenever we try to compare treated (i.e. those who receive treatment) to untreated

For any individual, we can only observe one potential outcome:

$$Y_i = \begin{cases} Y_{0,i} & \text{if } D_i = 0 \\ Y_{1,i} & \text{if } D_i = 1 \end{cases}$$

where D_i is a treatment indicator (equal to 1 if i was treated)

Defining the Causal Effect of Treatment

$$\boxed{\text{The causal impact of the program on } i \text{ is: } Y_{1,i} - Y_{0,i}}$$

- Each individual either participates in the program or not

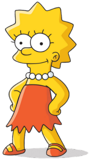
We only observe i's actual outcome:

$$Y_i = Y_{0,i} + \underbrace{(Y_{1,i} - Y_{0,i})}_{impact} D_i$$

Example: Alejandro goes to the hospital, Benicio does not

Establishing Causality

In an ideal world (research-wise), we could clone each treated individual and observe the impacts of treatment on outcomes



vs.

What is the impact of giving Lisa a textbook on her test score?

- Impact = Lisa's score with a book - Lisa's score without a book

In the real world, we either observe Lisa with a textbook or without

- Can't observe the **counterfactual** (other potential outcome)

Establishing Causality

To measure the causal impact of giving Lisa a book on her test score, we need to find a similar child that did not receive a book

vs.

Our estimate of the impact of the book is then the difference in test scores between the **treatment group** and the **comparison group**

- Impact = Lisa's score w/ book - Bart's score w/o book

As this example illustrates, finding a good comparison group is hard

- In applied micro, your research design **is** your counterfactual

Average Causal Effects

What we actually want to know is the **average causal effect**, which is not what we get from a difference in means comparison

Difference in group means

$$= \text{average causal effect of program on participants} + \text{selection bias}$$

Even in a large sample:

- People will choose to participate in a program when they expect it to make them better off (i.e. when $Y_{1,i} - Y_{0,i} > 0$)
- People who choose (or get) to participate maybe different than those who choose not to... even in the absence of the program

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	4
Betty	7	5
Carol	3	1
Diana	4	2

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	4
Betty	7	5
Carol	3	1
Diana	4	2

Alice and Betty take up treatment

Selection Bias: Example

	$Y_{1,i}$	$Y_{0,i}$
Alice	6	
Betty	7	
Carol		1
Diana		2

Alice and Betty take up treatment

$$\Rightarrow \bar{Y}_{treatment} = 6.5$$

Carol and Diana do not participate

$$\Rightarrow \bar{Y}_{comparison} = 1.5$$

$$\boxed{\bar{Y}_{treatment} - \bar{Y}_{comparison} = 6.5 - 1.5 = 5}$$

Selection Bias

Comparing treated/untreated (or participants/non-participants) doesn't normally provide an unbiased estimate of causal impact

- Treated, untreated likely different in absence of program
- Difference in means ($\bar{Y}_T - \bar{Y}_C$) could be biased up or down (relative to true average causal effect of treatment)
- Doesn't matter if treatment effect is constant
- Bias does not disappear in large samples

Mathematical Expectations

The **mathematical expectation** of Y_i, $E[Y_i]$:

- Equivalent to sample average in an infinite population
  - Example: probability coin flipped lands heads
  - Equivalent to fraction heads after a (very) large number of flips

Law of Large Numbers:

- In small samples, average of Y_i might be anything
- Average of Y_i gets very close to $E[Y_i]$ as number of observations (that we are averaging over) gets large

Conditional Expectations

Conditional expectation:

$E[Y_i|X_i = x]$

The conditional expectation of Y_i given a variable $X_i = x$,
is the average of Y_i in an infinite population that has $X_i = x$.

Example:
Let Y_i be height, and let $X_i \in \{0, 1\}$ be an "econ prof dummy"

- $E[Y_i|X_i = 1]$ is the average height among economics professors
- $E[Y_i|X_i = 0]$ is the average height among everybody else

Selection Bias (This Time with Math)

When we compare (many) participants to (many) non-participants:

$$E[\textbf{difference in group means}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

Adding in $\underbrace{-E[Y_{0,i}|D_i = 1] + E[Y_{0,i}|D_i = 1]}_{=0}$, we get:

Difference in group means

$$= \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{average causal effect on participants}} + \underbrace{E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]}_{\text{selection bias}}$$

The Experimental Ideal

How Can We Estimate Causal Impacts?

Experimental approach:

- **Random assignment to treatment:** eligibility for program is determined at random, e.g. via pulling names out of hat

The **Law of Large Numbers**: a sample average can be brought as close as we like to population mean just by enlarging the sample

When treatment status is randomly assigned,
treatment, control groups are random samples of a single population (e.g. the population of all eligible applicants for the program)

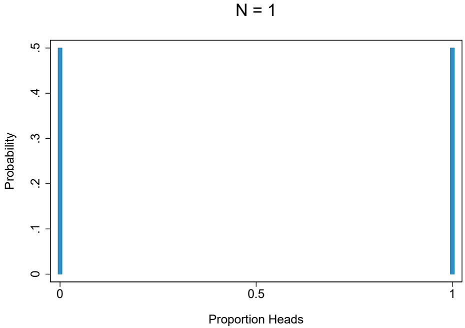
$$\Rightarrow E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected outcomes are equal in the absence of the program

Random Assignment & the Law of Large Numbers

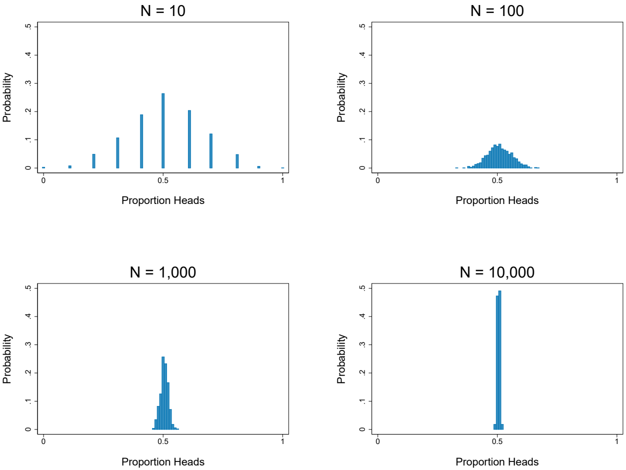
The probability a fair coin lands "heads" is 0.5, but the observed average proportion heads after a single coin flip is either 0 or 1

Random Assignment & the Law of Large Numbers

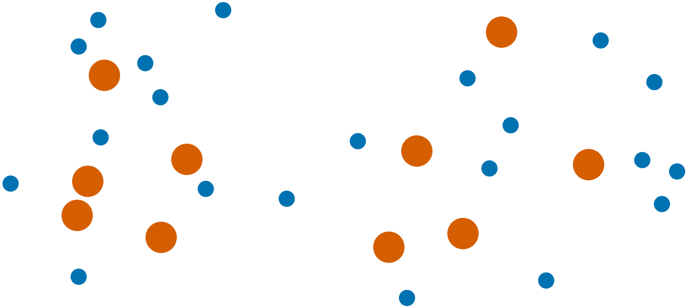


The **Law of Large Numbers**: a sample average can be brought as close as we like to population mean just by increasing sample size

Random Assignment & the Law of Large Numbers



Random Assignment Eliminates Selection Bias



When treatment is randomly assigned:

$$E[Y_{0,i}|D_i = 1] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Random Assignment Eliminates Selection Bias

Difference in means (treatment - control) provides an unbiased estimate of the (casual) **average treatment effect** (or ATE):

$$= E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1] + E[Y_{0,i}|D_i = 1] - E[Y_{0,i}|D_i = 0]$$

$$= \underbrace{E[Y_{1,i}|D_i = 1] - E[Y_{0,i}|D_i = 1]}_{\text{average treatment effect on participants}} + \underbrace{E[Y_{0,i}] - E[Y_{0,i}]}_{=0}$$

$$= \underbrace{E[Y_{1,i}] - E[Y_{0,i}]}_{\text{ATE}}$$

Internal Validity

Excellent news: random assignment eliminates selection bias*

*Some restrictions apply

The **Stable Unit Treatment Value Assumption (SUTVA)**:

- "The potential outcomes for any unit do not vary with the treatments assigned to other units" (Imbens and Rubin 2015)

When is SUTVA likely to be violated?

- When there are spillovers (so i's treatment impacts j)
- Example: vaccination(!) and other health treatments
  - This is why we have "cluster-randomized" trials

The Punchline

When treatment is randomly assigned (at an appropriate level), difference in outcomes between treatment and control groups provides an unbiased estimate of the causal impact of treatment

Randomly assigning treatment status eliminates selection bias (at least in expectation) because treatment, control groups are random samples of same underlying population of eligible units

Randomization: A Short History

Randomized Experiments in Theory

Petrarch (1364):

"If a hundred thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on nature's instincts, I have no doubt as to which half would escape."

van Helmont (who died in 1644):

"Let us take out of the Hospitals, pit of the Camps, or from elsewhere, 200 or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting... we shall see how many Funerals both of us shall have."

Source: Jamison (2019)

Randomization: A Timeline (Part I)

1885 Pierce and Jastrow use randomization in a psychology experiment (varying order in which different stimuli are presented to subjects)

1898 Johannes Fibiger conducts a trial of an anti-diphtheria serum in which every other subject is assigned to treatment (or control)

1923 Neyman suggests the idea of potential outcomes

1925 **Fisher suggests the explicit randomization of treatments (in the context of agriculture experiments)**

1926 Amberson et al. study of sanocrysin treatments for TB: coin flipped to determine which group received treatment, which were controls

1948 Randomized trial of streptomycin treatment for TB conducted by the Medical Research Council of Great Britain

⇒ Randomized evaluations become the norm in medicine

The Lady Tasting Tea

Chapter II of Fisher's *The Design of Experiments* begins:

> *"A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup."*

The Lady Tasting Tea

Null hypothesis (aka H_0)

- Fisher believes that Dr. Bristol cannot taste the difference

A test of the hypothesis:

- *"Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in random order."*

The Lady Tasting Tea: Experimental Design

Rule #1: do not confound your own treatment

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, she will choose 4 cups at random
- Fisher points out that the experimenter could screw this up:

> *"If all those cups made with the milk first had sugar added,*
> *while those made with the tea first had none,*
> *a very obvious difference in flavour would have been introduced*
> *which might well ensure that all those made with sugar*
> *should be classed alike."*

- Gerber and Green refer to this as **excludability**

The Lady Tasting Tea: Experimental Design

Rule #1B: do not accidentally confound your own treatment

- Fisher, in (perhaps) the earliest known scientific subtweet:

  > *"It is not sufficient remedy to insist that 'all the cups must be exactly alike' in every respect except that to be tested. For this is a totally impossible requirement."*

- To minimize likelihood of accidentally confounding your treatment, it's best is to constrain yourself by randomizing
  - Minimizes the likelihood of unfortunate coincidences
  - Highly controversial position at the time, and is still debated in some circles; alternative is to force balance on observables (and then just hope that unobservables don't matter too much)

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

Suppose Dr. Bristol correctly identified all 4 “treated” cups

- How likely is it that this could have occurred by chance?
  - There are $\binom{8}{4} = 70$ possible ways to choose 4 of 8 cups
  - Only one is correct; a subject with no ability to discriminate between treated, untreated cups has a $1/70$ chance of success
  - The p-value associated with this outcome is $1/70 \approx 0.014$, less than the cutoff for the “standard level of significance” of 0.05

The Lady Tasting Tea: a Hypothesis Test

Suppose Dr. Bristol correctly identified 3 “treated” cups

- How likely is it that this could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 of 8 cups
    - There are 17 ways to choose **at least** 3 correct cups
  - The p-value associated with this outcome is $17/70 \approx 0.243$
  - We should not reject the null hypothesis

Only reject H_0 if Dr. Bristol identifies all 4 treated cups

- In the actual experiment, the null hypothesis was rejected

The Lady Tasting Tea: Size and Power

The size of a test is the likelihood of rejecting a true null

- Fisher asserts that tests of size 0.05 are typical

Alternative experiment: what if we had treated 3 out of 6 cups?

- There are $\binom{6}{3} = 20$ possible ways to choose 3 of 6 cups
- Best possible p-value is therefore 0.05

Alternative experiment: what if we had treated 3 out of 8 cups?

- There are $\binom{8}{3} = 56$ possible ways to choose 3 of 8 cups
- Best possible p-value is therefore 0.017

$\Rightarrow$ **Optimal to have equal numbers of treated, untreated cups**

The Lady Tasting Tea: Size and Power

An alternate experiment: an unknown number of treated cups

- Under the null, the probability of getting 8 right is 1 in 2^8
- Probability of getting 7 right is $8/256 = 0.03125$

Design would achieve higher power with the same number of trials

- Possible to reject the hypothesis that Dr. Bristol cannot tell the difference even when her ability to discriminate is imperfect

Ronald Fisher's Contributions to Statistics

Key lesson to take away from "lady tasting tea" anecdote: caffeine breaks with colleagues critical to advancement of science

Other contributions:

1. Introduced the modern randomized trial
2. Introduced the idea of permutation tests

Fisher's permutation-based approach to inference is not the norm in economics; our default is regression analysis and classical statistics

Randomization: A Timeline (Part II)

1942 Launch of Cambridge-Somerville Youth Study of at-risk boys

1962 Perry Preschool (MI) and Early Training Project (TN) experiments randomize assignment of at-risk children to high-quality preschools

1967 NJ Income Maintenance Experiment (proposed by Heather Ross), 4 other negative income tax experiments between 1971 and 1982

1972 Abecedarian Project (NC) randomized intervention for at-risk infants

1974 Rubin introduces the concept of potential outcomes (as we know it)

1994 National Job Corps Study (by Mathematica/US Dept. of Labor)

1995 PROGRESA evaluation launched by Mexican government

1998 Dutch NGO ICS Africa begins randomized trial of “deworming” in Kenyan primary schools... in partnership with Michael Kremer

RCTs in Development Economics: Mexico's Progresa

photo: Curt Carnemark / World Bank

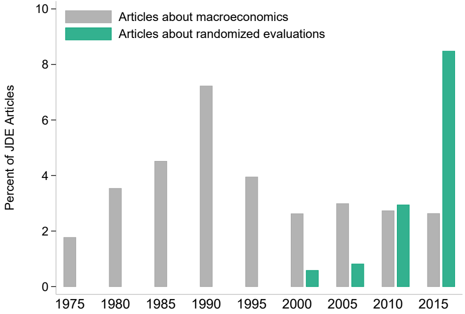
- Mexico piloted conditional cash transfers (CCTs) in mid-1990s
- Economists in gov't pushed for randomized roll out of pilot
- IFPRI reserachers published initial findings in late 1990s

RCTs in Development Economics: Busia, Kenya

photo: Stephanie Skinner / Deworm the World

- Michael Kremer convinces NGO to randomize interventions
- Study of deworming (w/ Miguel) launches RCT movement

RCTs in Development Economics: Trends



Abstracts of 2,695 *Journal of Development Economics* articles
(all articles published prior to 2019, starting form Volume 1 in 1974)

RCTs in Development Economics

In 2019, Michael Kremer, Esther Duflo, and Abhijit Banerjee won the Nobel Prize in economics for their promotion of RCTs and their "experimental approach to alleviating global poverty"

Regression Analysis of RCTs

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

$\uparrow$

equal to population mean because
control group is a random sample

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i|D_i = 0] = E[Y_{0,i}|D_i = 0] = E[Y_{0,i}]$$

Expected value of treatment group mean:

$$\begin{aligned} E[\bar{Y}_T] = E[Y_i|D_i = 1] &= E[Y_{1,i}|D_i = 1] \\ &= E[\delta_i + Y_{0,i}|D_i = 1] \\ &= E[\delta_i|D_i = 1] + E[Y_{0,i}|D_i = 1] \\ &= E[\delta_i] + E[Y_{0,i}] \end{aligned}$$

Treatment Effects Under Random Assignment

Expected value of control group mean:

$$E[\bar{Y}_C] = E[Y_i | D_i = 0] = E[Y_{0,i} | D_i = 0] = E[Y_{0,i}]$$

Expected value of treatment group mean:

expected outcome
in absence of treatment

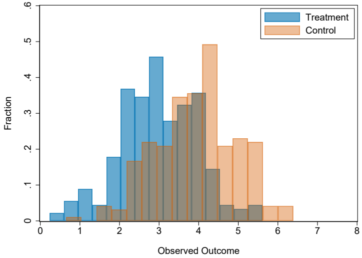
$$\begin{aligned} E[\bar{Y}_T] = E[Y_i | D_i = 1] &= E[Y_{1,i} | D_i = 1] \\ &= E[\delta_i + Y_{0,i} | D_i = 1] \\ &= E[\delta_i | D_i = 1] + E[Y_{0,i} | D_i = 1] \\ &= \underset{\substack{\uparrow \\ \text{average treatment effect}}}{E[\delta_i]} + E[Y_{0,i}] \end{aligned}$$

H_0: $ATE = 0$

Null hypothesis (H_0):

The average treatment effect is zero: $ATE = 0$

Or, equivalently: $\bar{Y}_T = \bar{Y}_C$

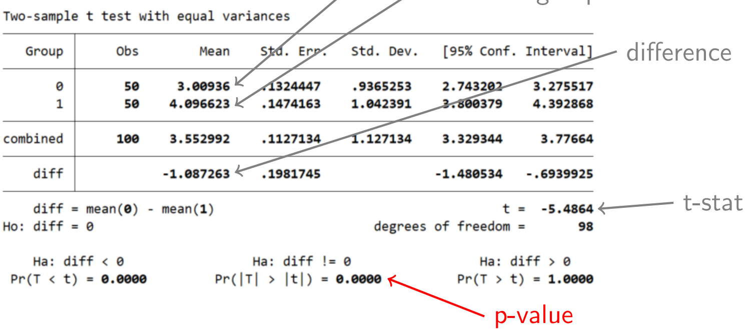


In Stata:

```
ttest y, by(t)
```

Testing the Equality of Means

Stata: `ttest y, by(t)`



```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      50     3.00936    .1324447    .9365253    2.743202    3.275517
       1 |      50    4.096623    .1474163    1.042391    3.800379    4.392868
---------+--------------------------------------------------------------------
combined |     100    3.552992    .1127134    1.127134    3.329344     3.77664
---------+--------------------------------------------------------------------
    diff |           -1.087263    .1981745               -1.480534   -.6939925
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.4864
Ho: diff = 0                                     degrees of freedom =       98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

control group mean

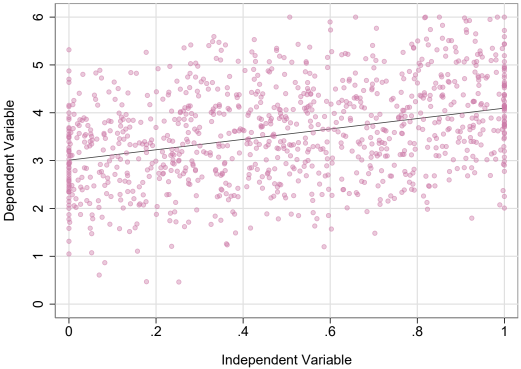
treatment group mean

difference

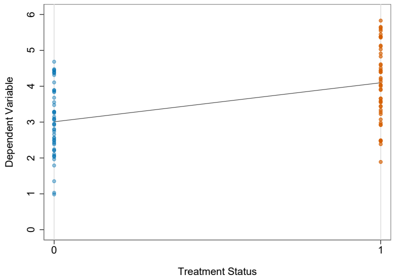
t-stat

p-value

Regression Analysis of RCTs



Regression Analysis of RCTs



Simple regression framework for analyzing RCTs: $Y_i = \alpha + \beta D_i + \varepsilon_i$

- Treatment indicator $D_i = 0, 1 \Rightarrow$ only two possible values of $\hat{Y}_i$

The End!