TESTING THE VALIDITY OF THE COUNTERFACTUAL

Jesse Lecy

Core Concepts

The Selection Problem

Tests for Group Equivalence

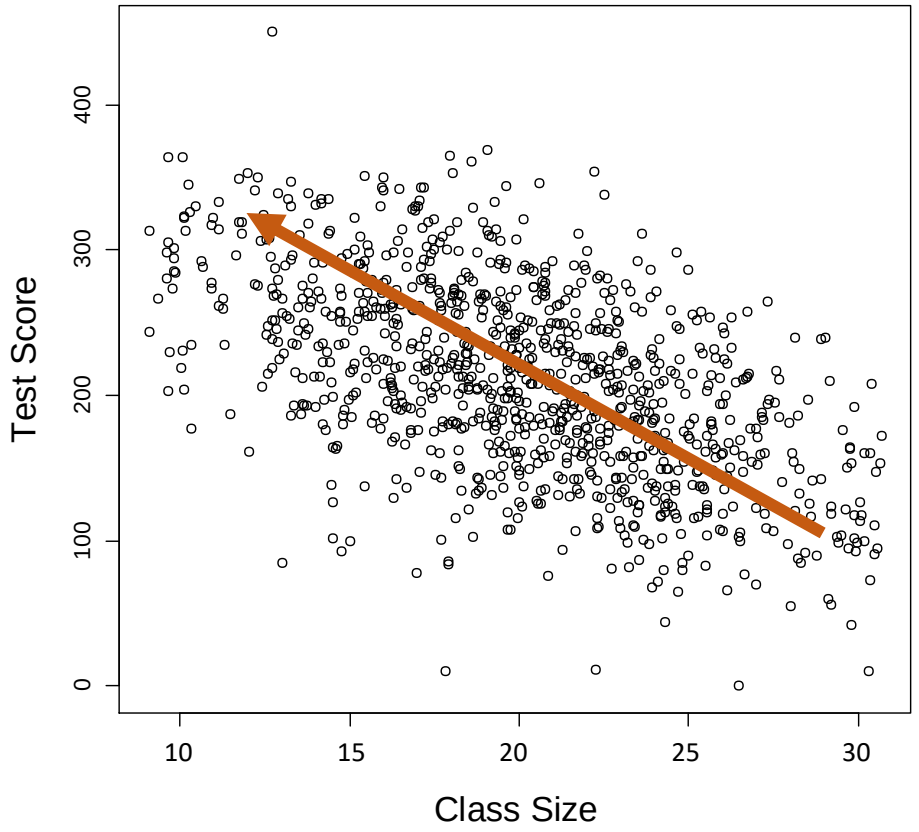
Omnibus Test

Bonferroni Correction

Tests for Nonrandom Attrition

NATURE GIVES US CORRELATIONS: THE SELECTION PROBLEM IN EVALUATION RESEARCH

If we reduce classroom size we should see student performance improve. Right?



THE MATH OF CITIES

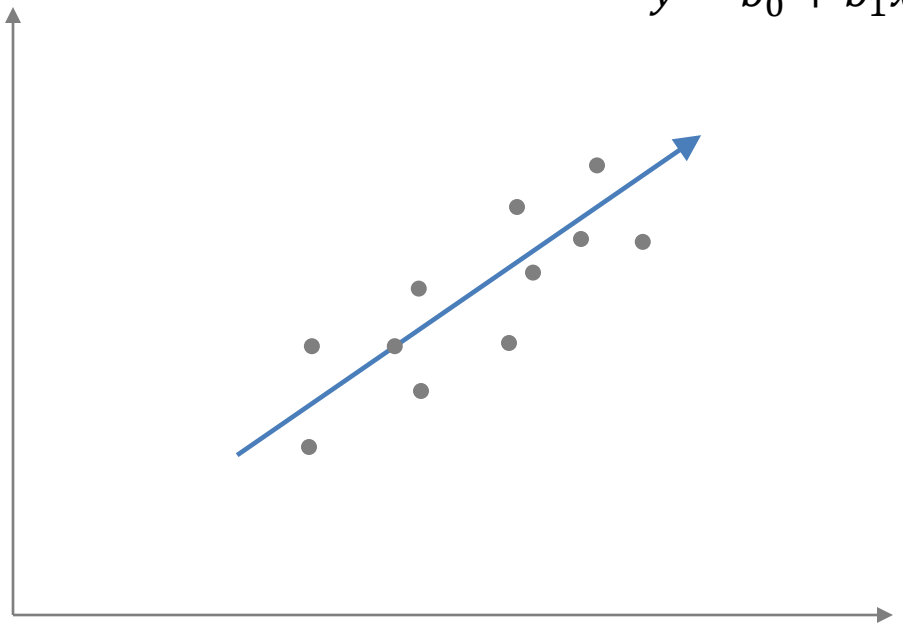
Listen: 4:30 – 15:30

http://www.radiolab.org/2010/oct/08/its-alive/

THE MATH OF CITIES

Patents / Crime / Salary
(outcomes)

$$y = b_0 + b_1 x$$



Walking Speed?

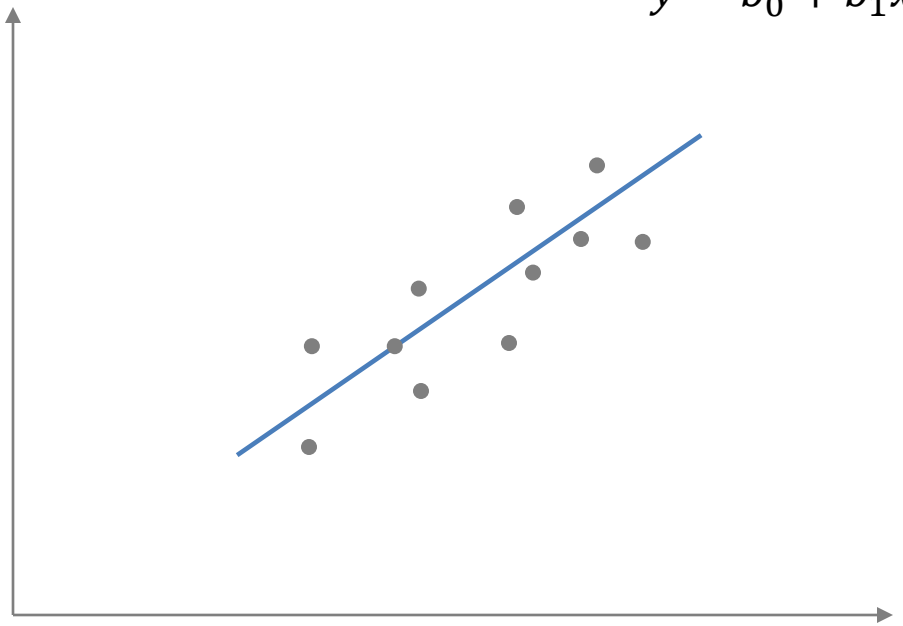
THE MATH OF CITIES

If you increase the
WALKING SPEED of a city,
would you increase the
number of patents
produced?

THE MATH OF CITIES

Patents / Crime / Salary
(outcomes)

$$y = b_0 + b_1 x$$

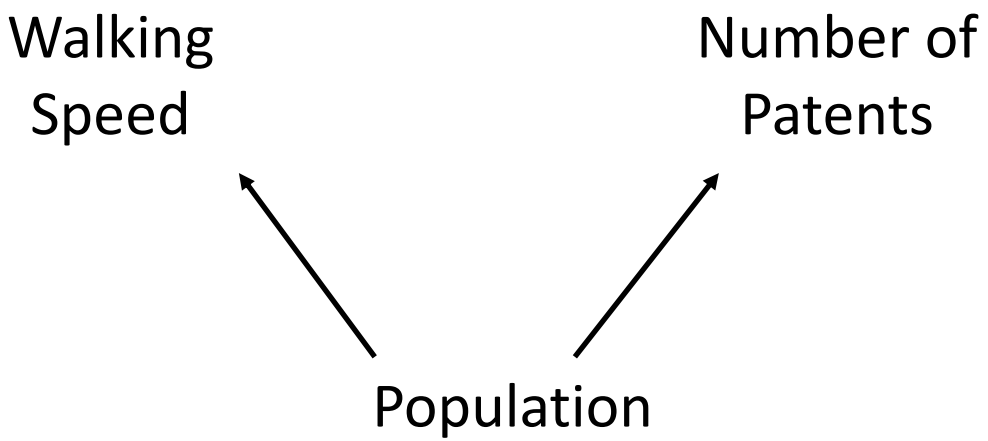


Population

THE MATH OF CITIES

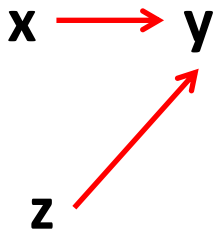
If you increase the POPULATION of a city, would you increase the number of patents produced?

The problem with correlations:

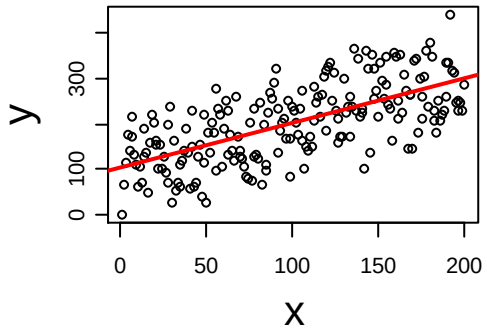


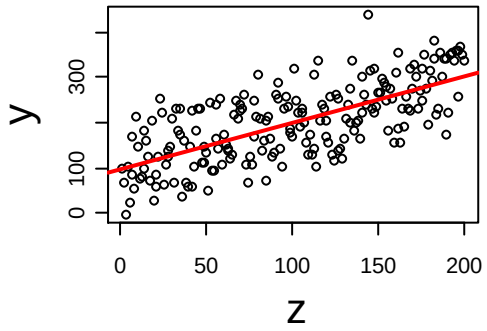
NATURE GIVES US CORRELATIONS

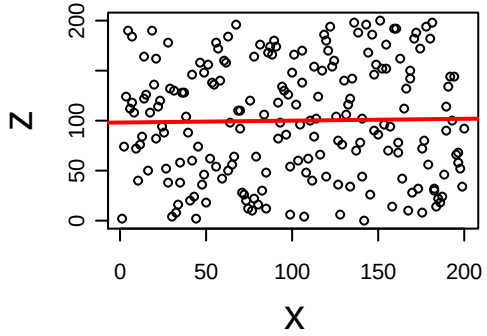
Example #1



X and Z both impact Y

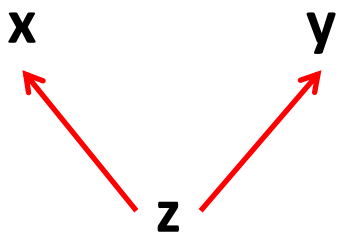




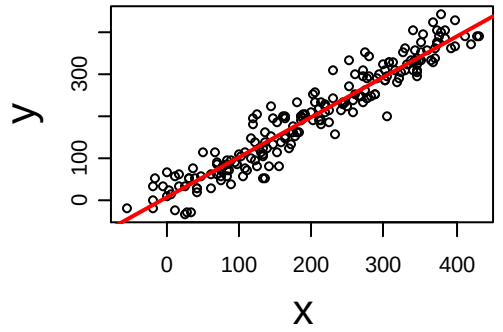


no correlation

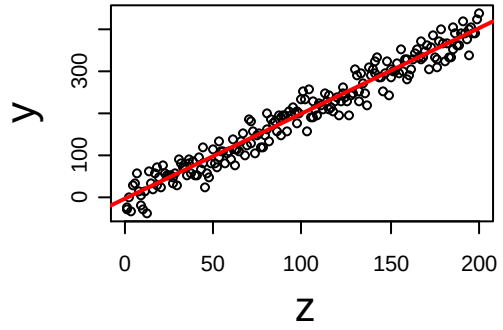
Example #2

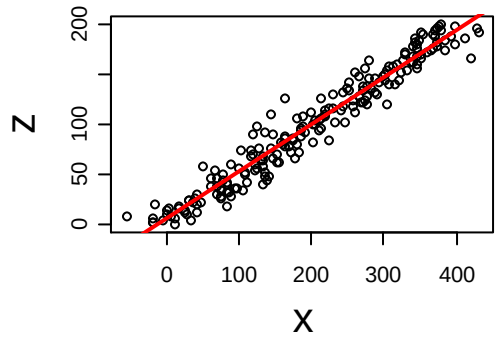


Z impacts both X and Y

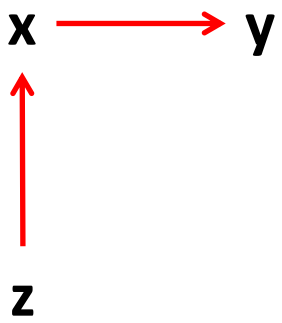


No causal relationship but high correlation

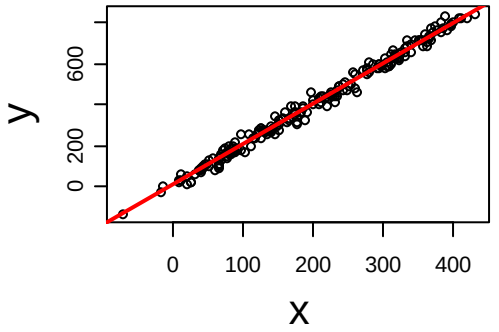


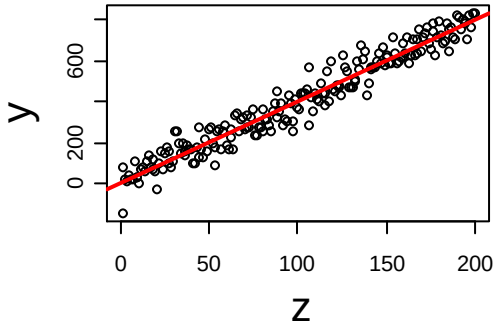


Example #3

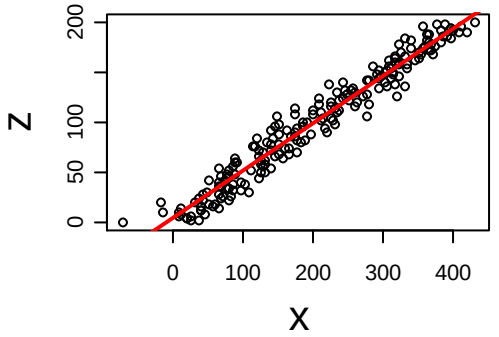


Causal chain



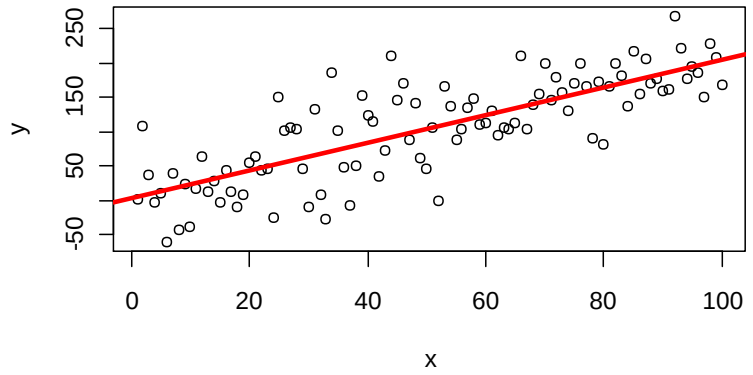


no direct causal relationship but highly correlated



Reverse causality

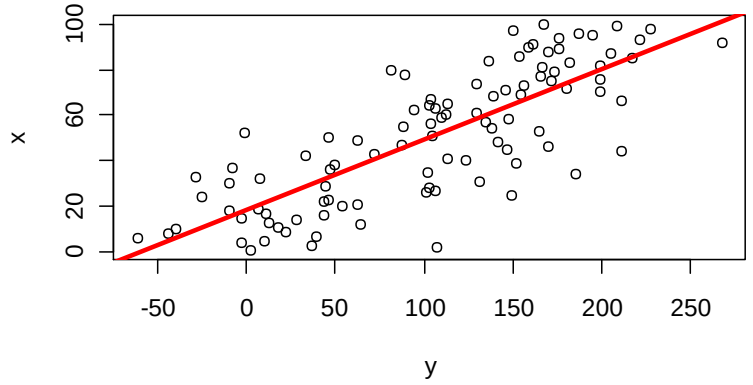
x ⟶ y



$Y = b_0 + b_1X + e$

Both models are highly significant, how do we know which one is the causal relationship?

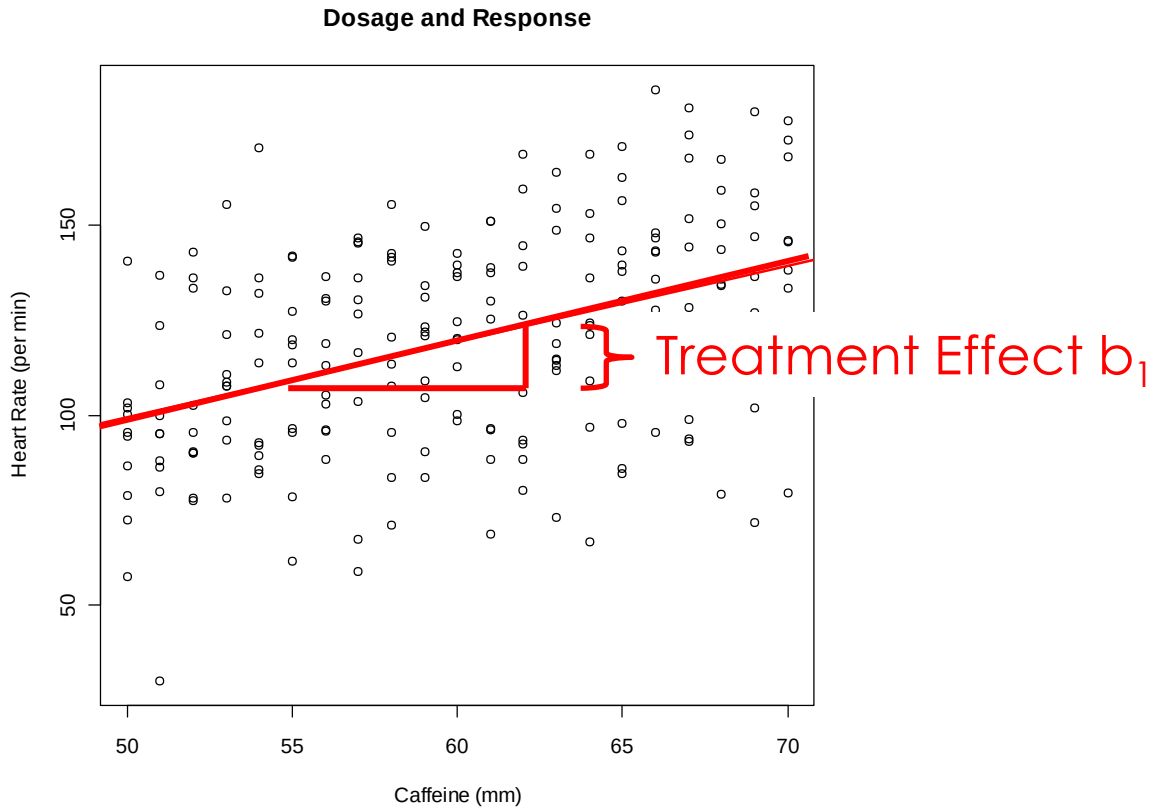
x ⟵ y



$X = b_0 + b_1Y + e$

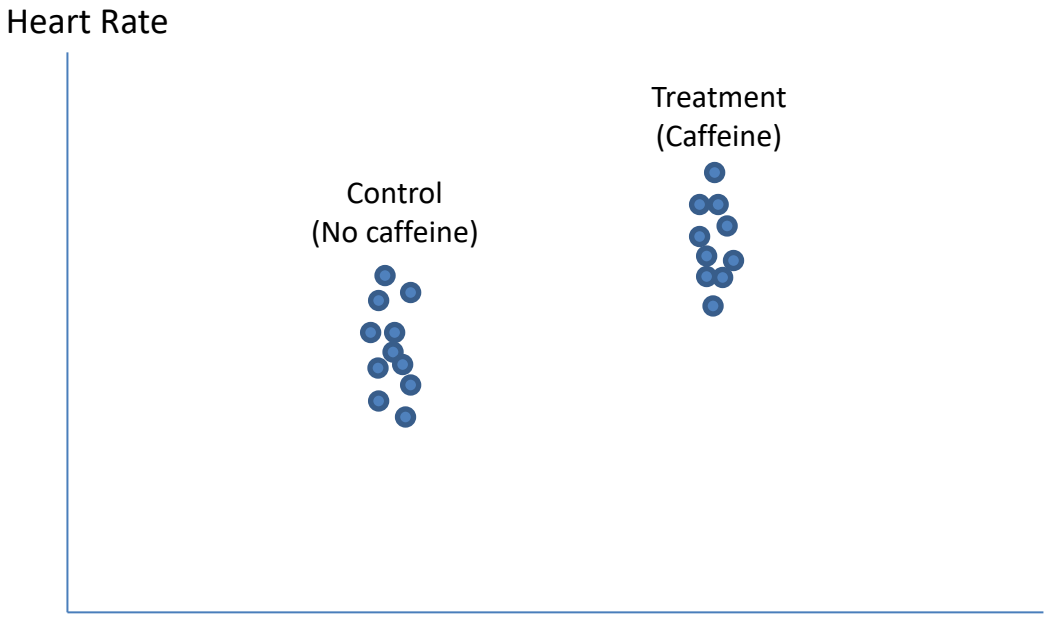
THE PROGRAM EVALUATION FRAMEWORK

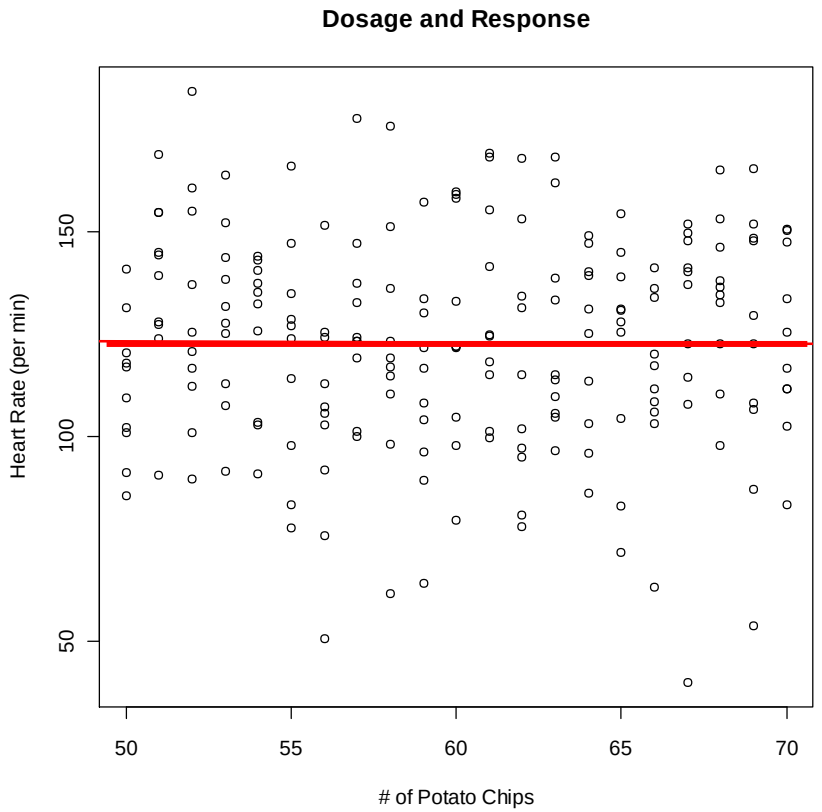
$$Heartrate = b_0 + b_1 \cdot Caffeine + \varepsilon$$



DISCRETE TREATMENT CASE: MODEL IS THE SAME EXCEPT CAFFEINE IS A DUMMY NOT A LEVEL

$$Heartrate = b_0 + b_1 \cdot Caffeine + \varepsilon$$

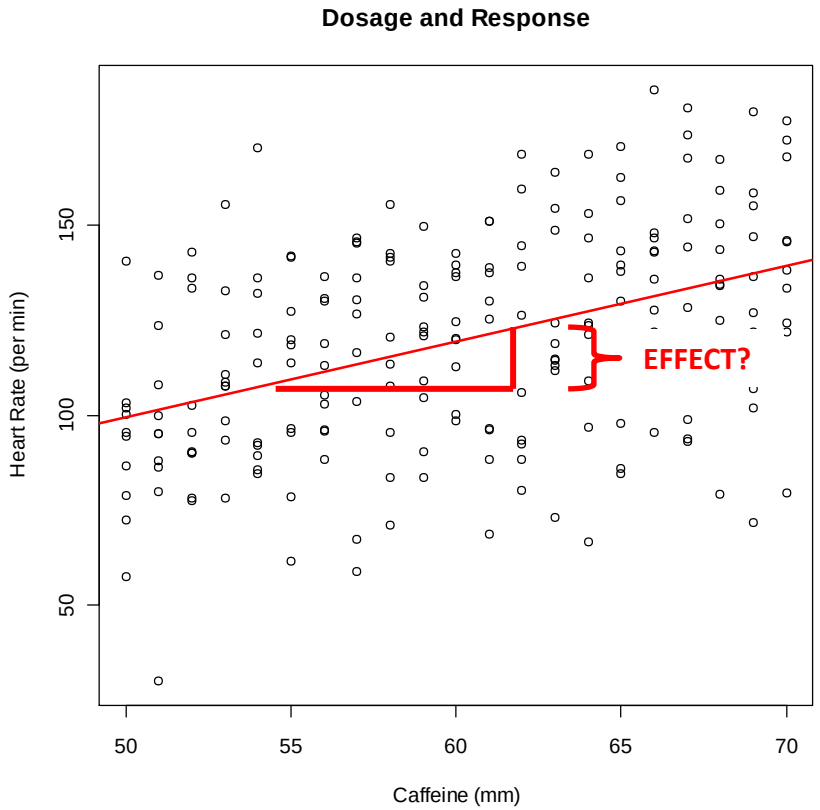


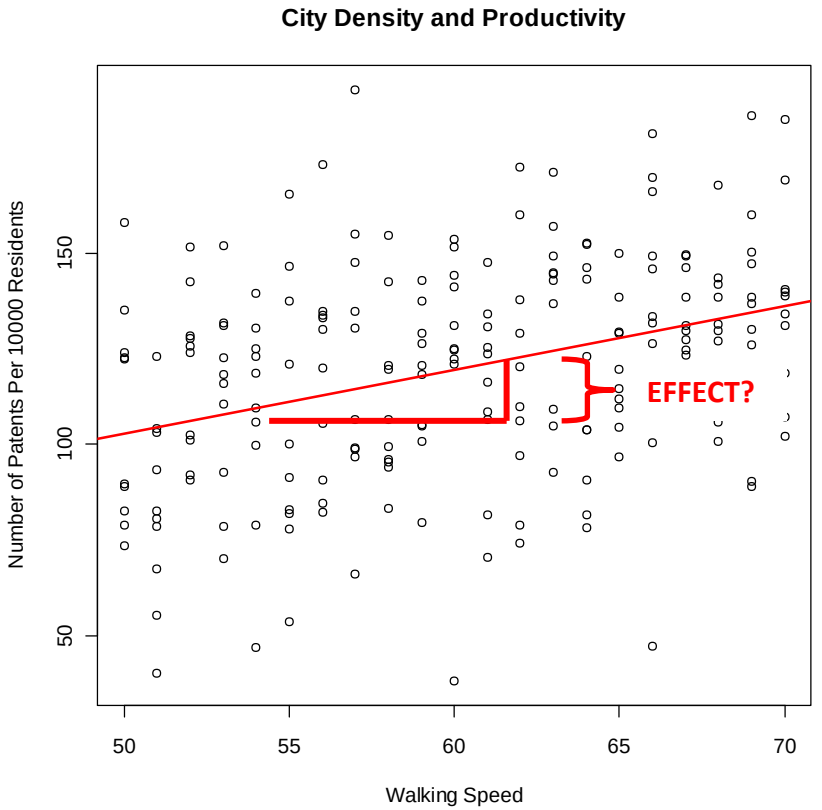


No effect of treatment on outcome

HOW DO WE KNOW WHEN THE INTERPRETATION IS CAUSAL?

When is b_1 an impact, and when is it just a relationship in data?





THE SELECTION PROBLEM IN EVALUATION RESEARCH

Microfinance example of bias from selection INTO a study group

Number of each "type" of person in the study

	NOT Entrepreneurial	Entrepreneurial
No Loan	30	15
Takes a Loan	20	35

You are more likely to take a loan if you know you are good at business

Average weekly income after loan period

	NOT Entrepreneurial	Entrepreneur
No Loan	$10	$20
Takes a Loan	$10	$20

Income not impacted by the loan

Takes Loan?

$$NO: \quad \frac{30 \cdot \$10 + 15 \cdot \$20}{45} = \$13.33$$

$$YES: \quad \frac{20 \cdot \$10 + 35 \cdot \$20}{55} = \$16.37$$

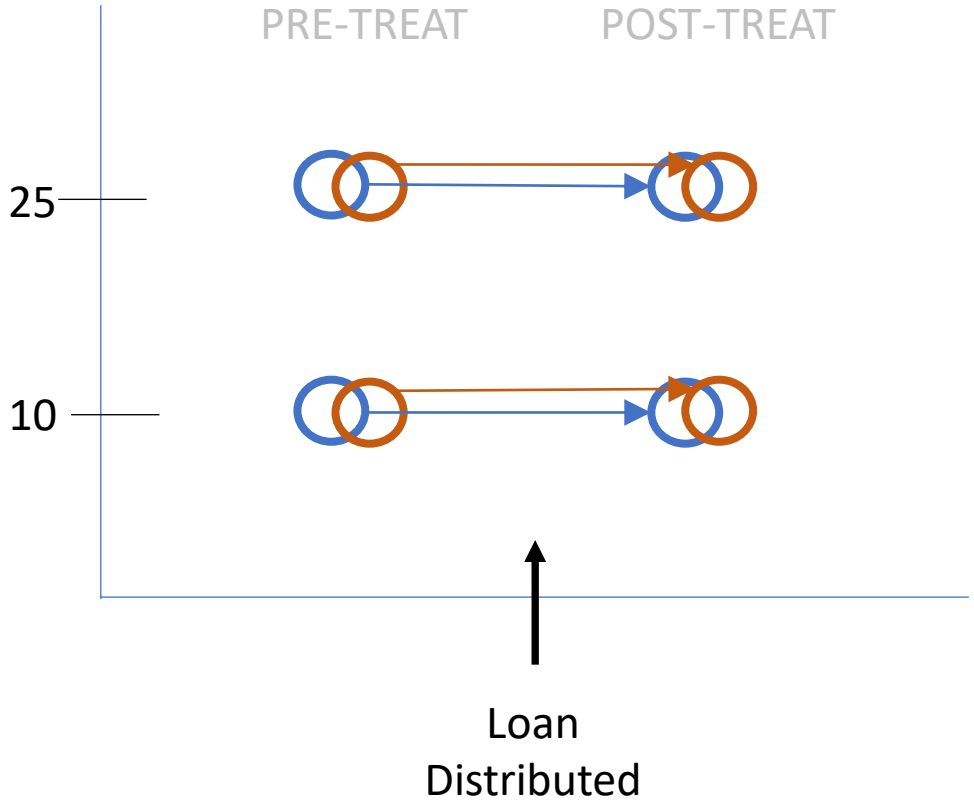
The loan appears to have an impact!

Even though we know it didn't.

NO EVIDENCE OF IMPACT

Takes Loan / Entrepreneurial (35 people)
NO Loan / Entrepreneurial (15 people)

Takes Loan / NOT Entrepreneurial (20 people)
NO Loan / NOT Entrepreneurial (30 people)



The calculations above are an example of the POST-TEST ONLY estimator (T2-C2). This estimator requires that the groups are balanced or statistically equivalent prior to the intervention. Which would not be the case here (the pre-treatment calculations would be the same as post-treatment since they incomes don't change as a result of the loan, so these differences would be present prior to the treatment as well).

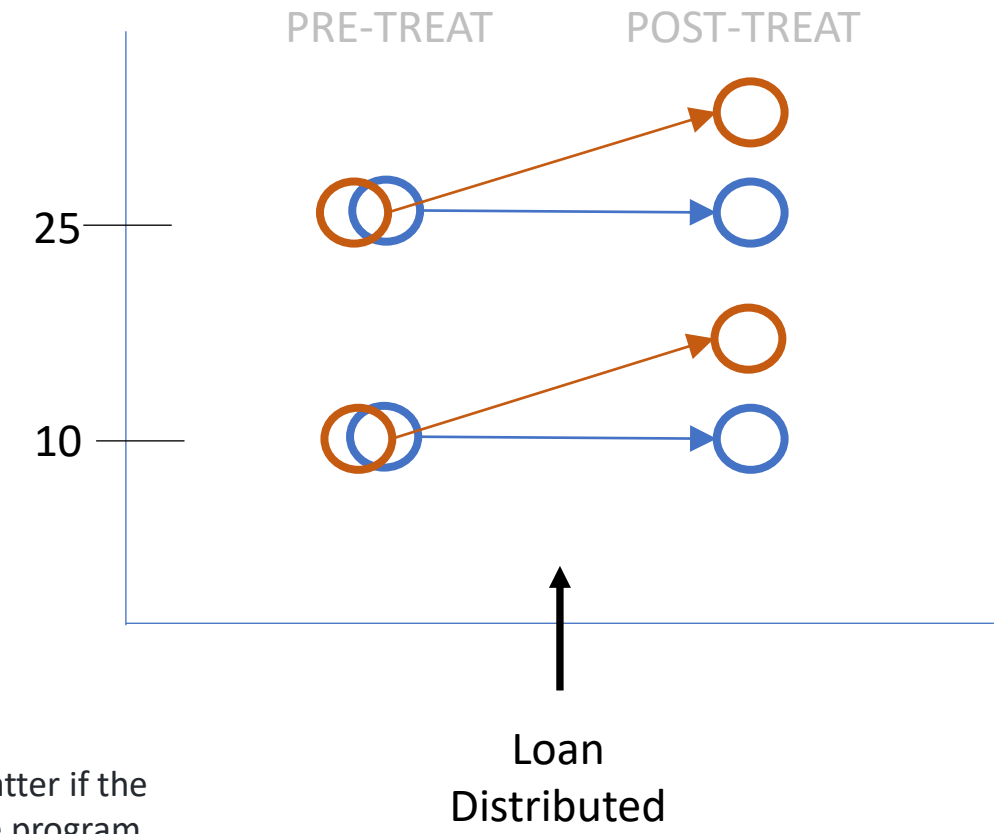
$$NO: \quad \frac{30 \cdot \$10 + 15 \cdot \$20}{45} = \$13.33$$

$$YES: \quad \frac{20 \cdot \$10 + 35 \cdot \$20}{55} = \$16.37$$

EVIDENCE OF IMPACT

Takes Loan / Entrepreneurial (35 people)
NO Loan / Entrepreneurial (15 people)

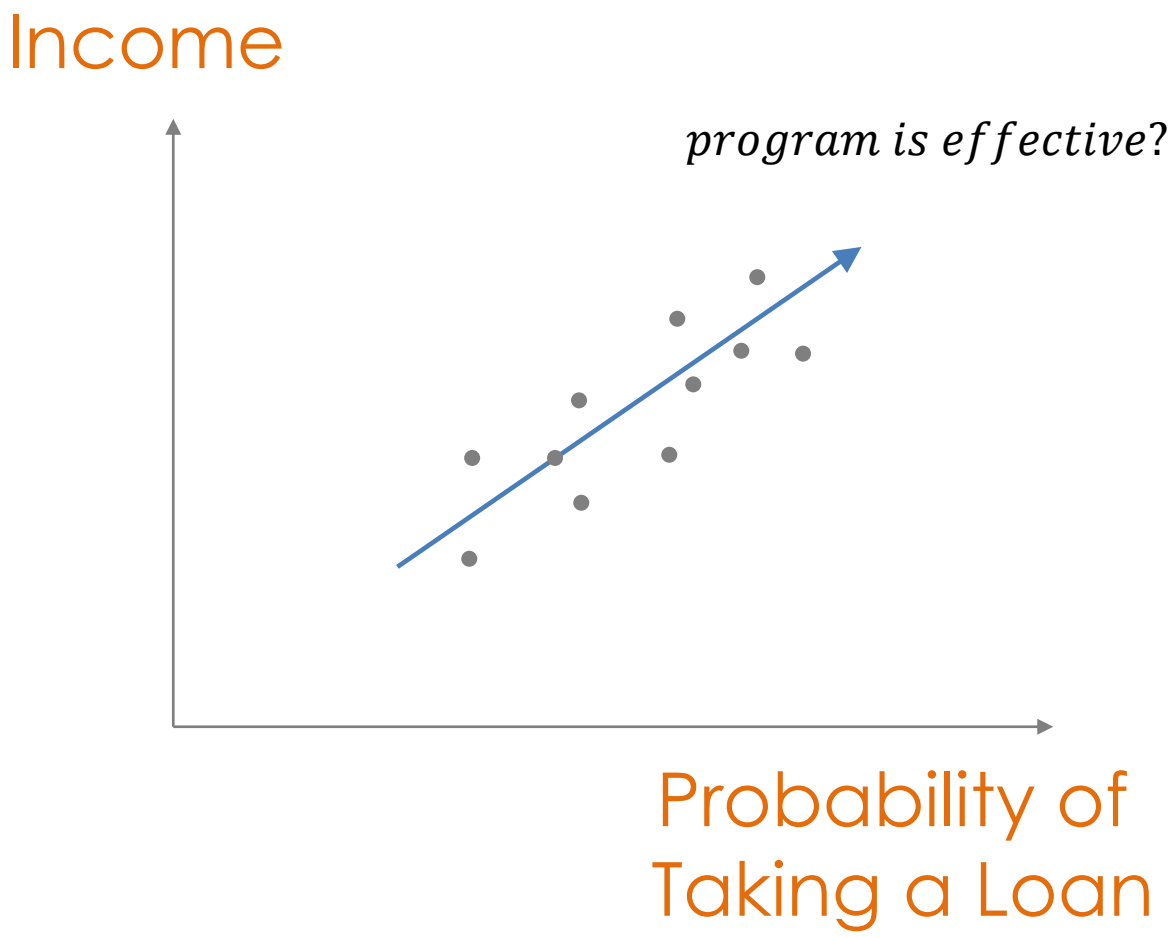
Takes Loan / NOT Entrepreneurial (20 people)
NO Loan / NOT Entrepreneurial (30 people)

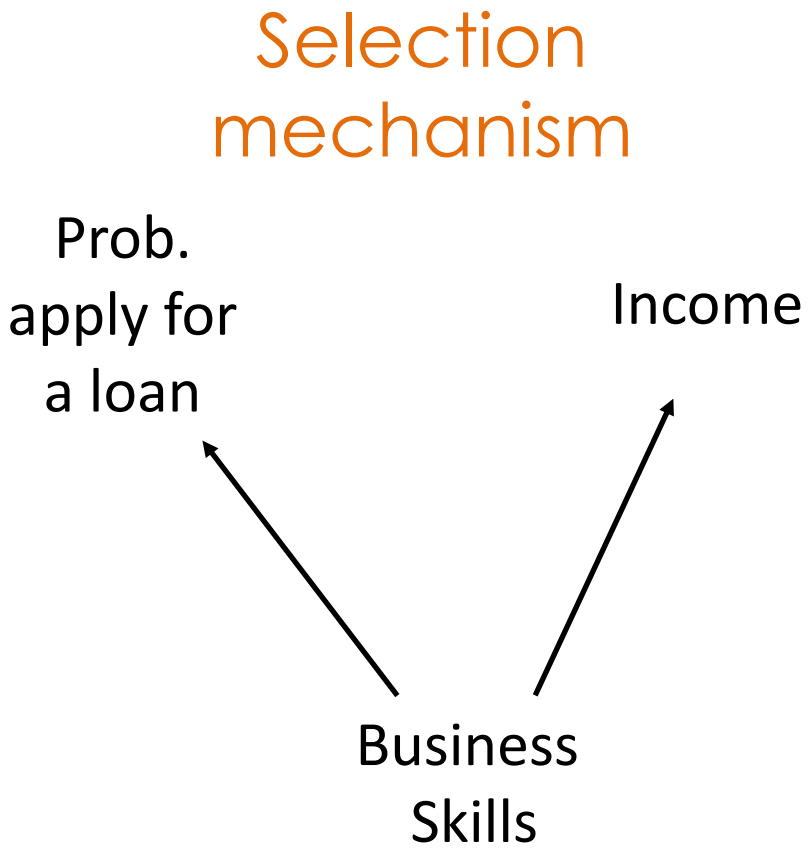


If we use the DIFF-IN-DIFF estimator ([T2-T1] - [C2-C1]) then it wouldn't matter if the groups are not identical in the pre-treatment period. It will still capture the program impact correctly. Note that we need pre-treatment data for this estimator, which is often the challenge. It's more robust but also more data-intensive.

The main point is developing intuition for when you can interpret differences in the study groups as program impact or a program "effect", versus when differences arrive because of lack of study group equivalence and thus they capture selection and not impact.

CORRELATION VERSION OF LAST EXAMPLE:





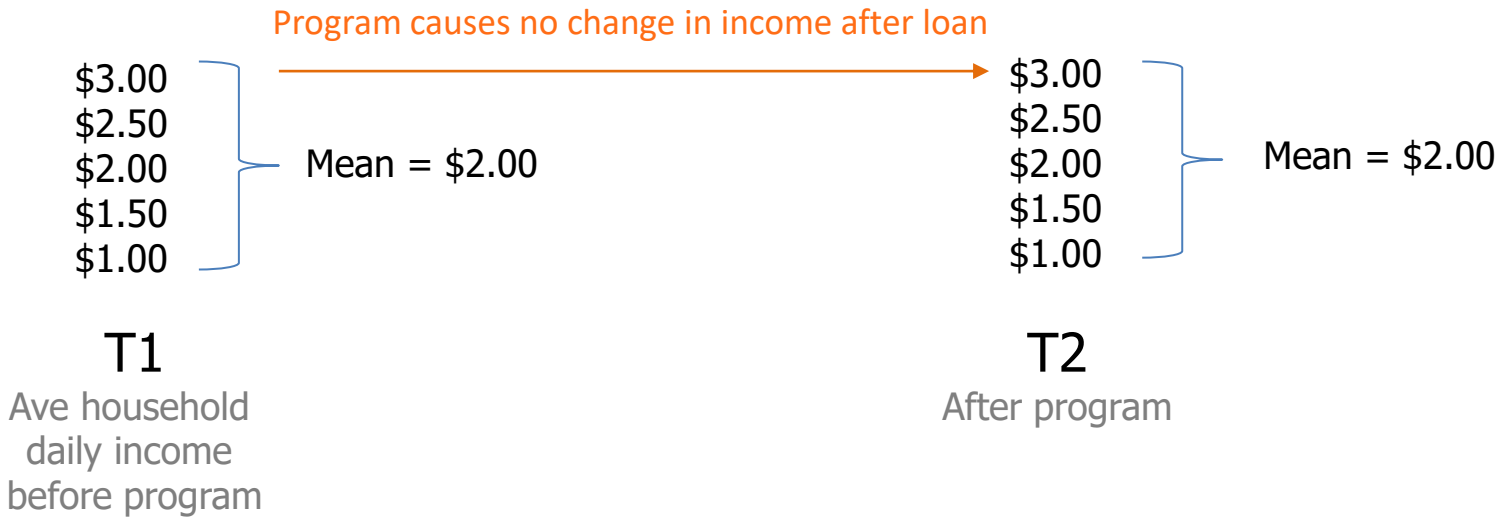
“Selection” Problem

Those that participate in the program
are different from those that do not participate.

This is the biggest problem in impact evaluation!

Microfinance example of bias from selection OUT OF a study group

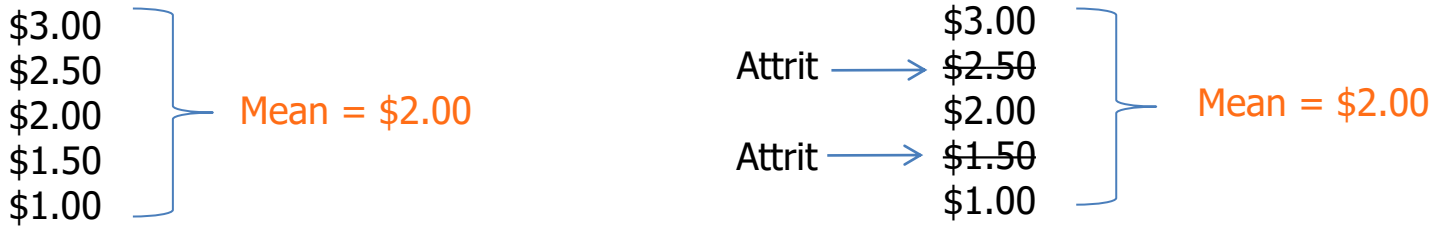
Reflexive design



$$T2 - T1 = 0$$

causal estimate is unbiased

Random Attrition Example

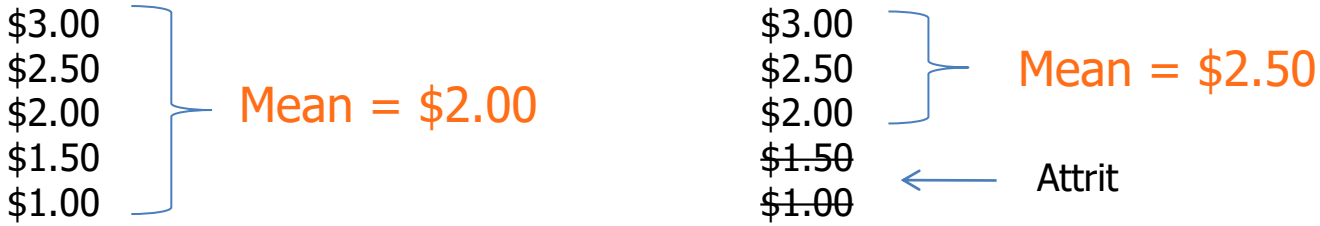


T2 - T1 = 0

Impact study accurately represents program effects

Program is not determined to be effective (no change)

Non-Random Attrition Example

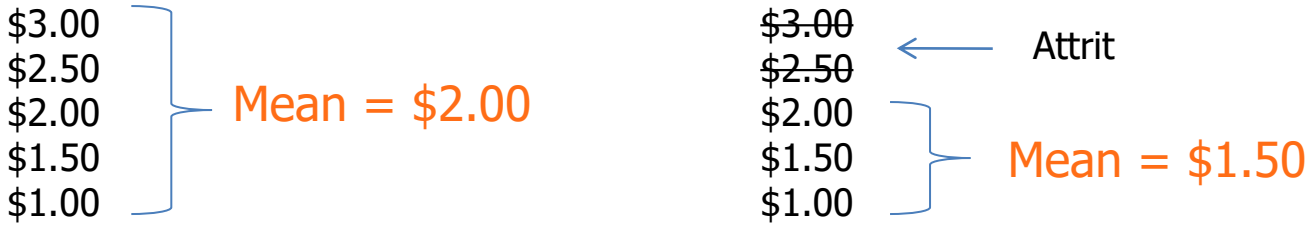


$$T2 - T1 \neq 0$$

We over-estimate program effects

Program appears to be effective

Non-Random Attrition Example



$$T2 - T1 \neq 0$$

We under-estimate program effects

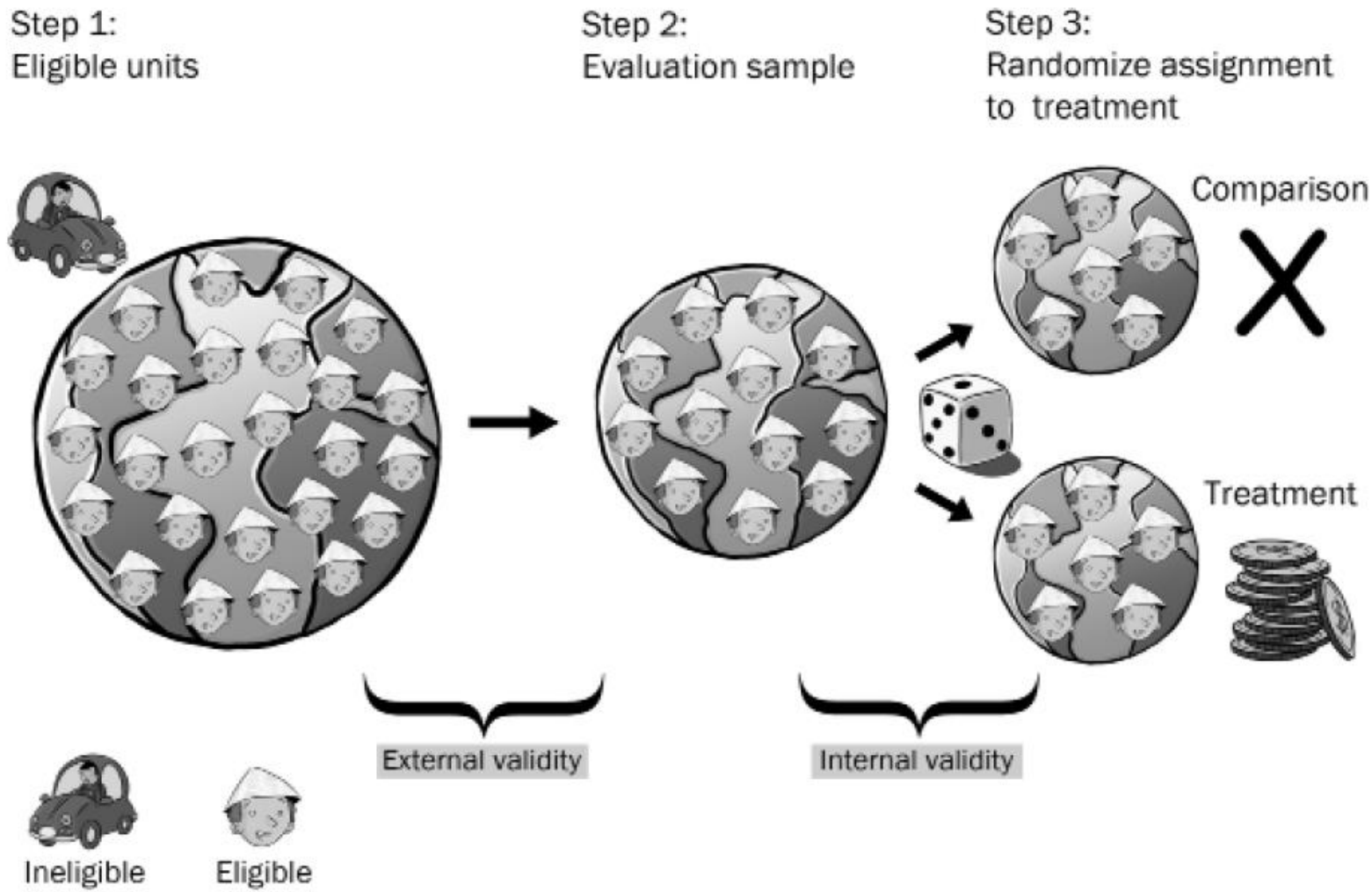
Program appears to harm families

RANDOMIZED CONTROL TRIALS (RCT'S): THE "GOLD STANDARD" FOR INTERNAL VALIDITY

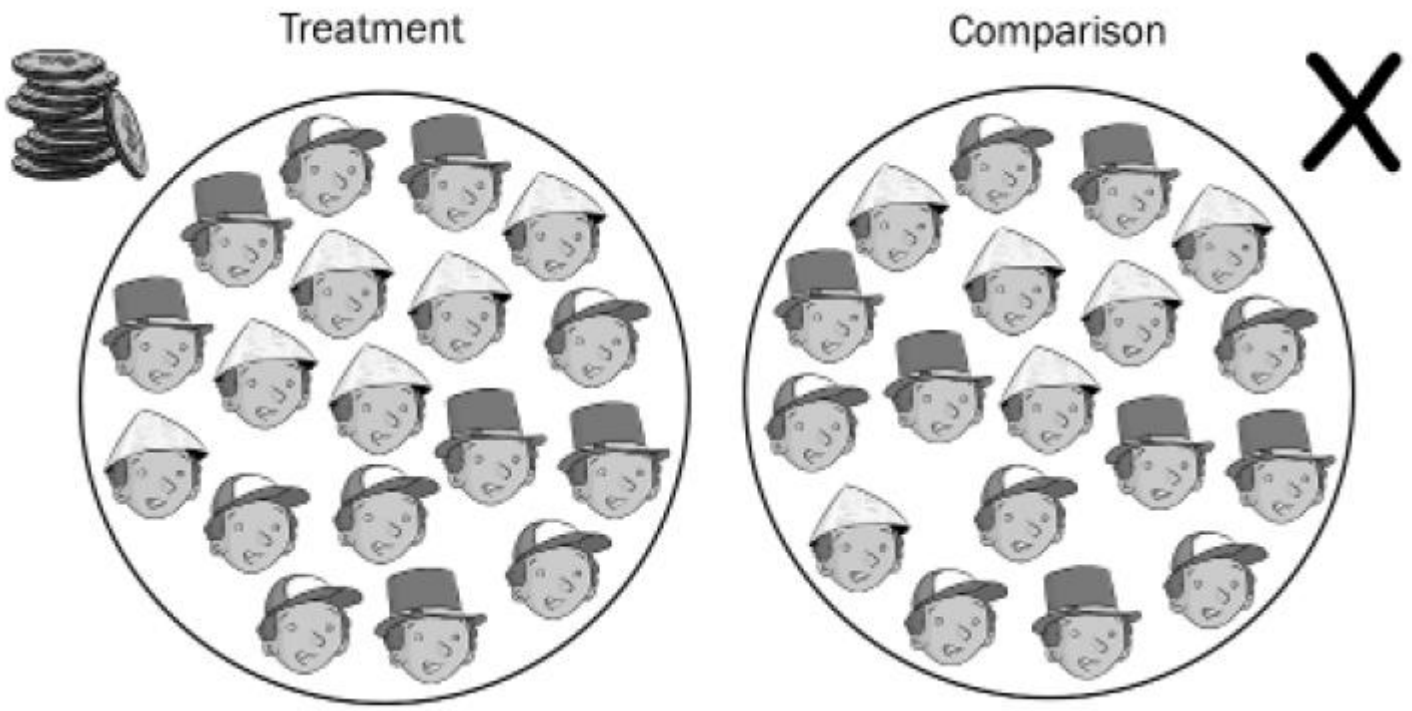
Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice.* The World Bank.

>> Chapter 4. Randomized Selection Methods

Figure 4.3 Steps in Randomized Assignment to Treatment



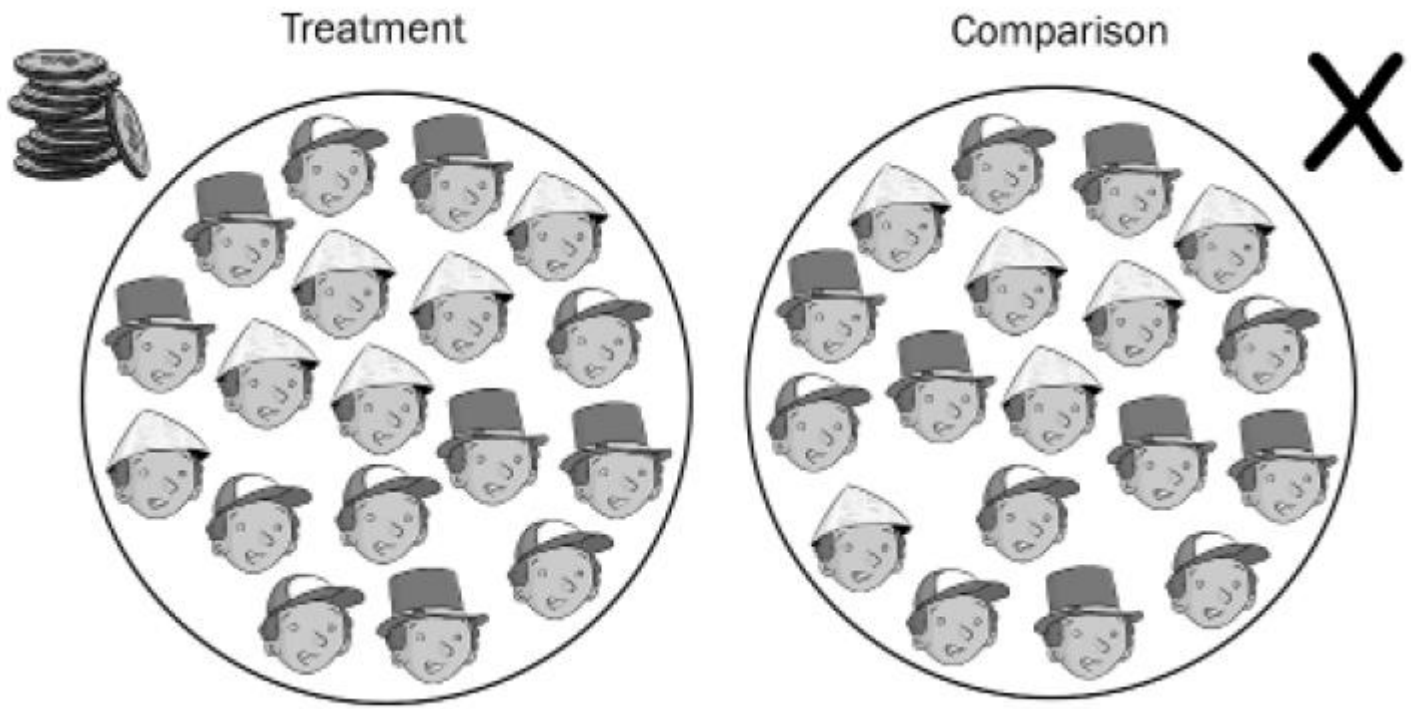
Our counterfactual framework is
valid / robust when the
groups only DIFFER BY THE TREATMENT
but are OTHERWISE "IDENTICAL"

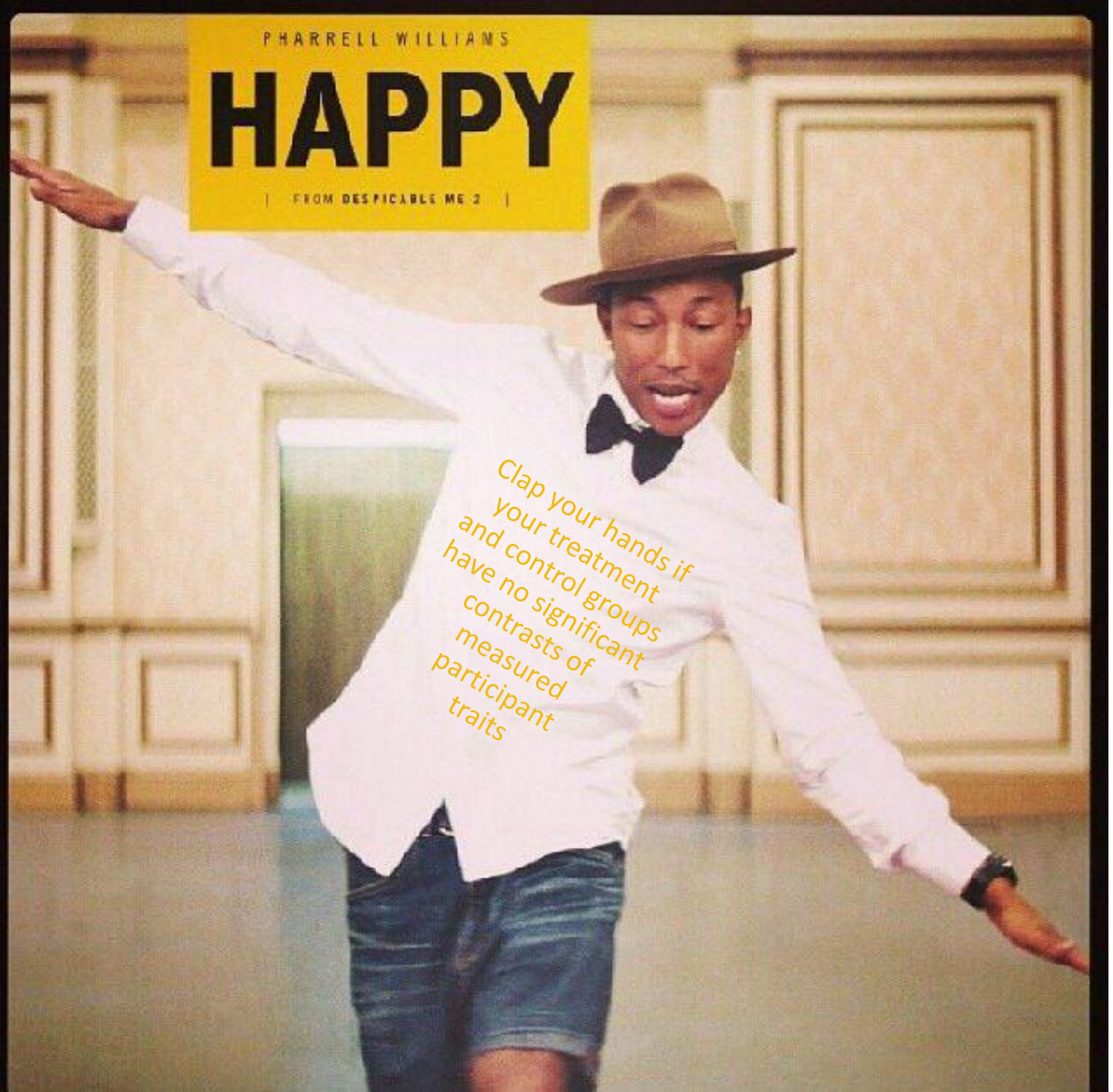


When true, we can interprett the differences in group outcomes
after the treatment period to be caused by the treatment

How do we test the criteria:

groups only OTHERWISE "IDENTICAL" ???





"HAPPY" RANDOMIZATION

If we have a group of 100 people and we randomly assign them to two groups, 50 people each, how often would we expect the average weight of each group to differ?

If we have a group of 100 people and we randomly assign them to two groups, 50 people each, how often would we expect the average weight of each group to differ?

MATHEMATICALLY: ALWAYS !!!

weights will never be exactly identical

so what do we mean by “different”?

If we have a group of 100 people and we randomly assign them to two groups, 50 people each, how often would we expect the average weight of each group to differ?

STATISTICALLY:

TEST OF GROUP MEANS

Use a t-test and select a
level of confidence
that we are comfortable with

If we have a group of 100 people and we randomly assign them to two groups, 50 people each, how often would we expect the average weight of each group to differ?

STATISTICALLY:

TEST OF GROUP MEANS

What does alpha=0.05 mean?

[--- 95% confidence interval ---]

If we have a group of 100 people and we randomly assign them to two groups, 50 people each, how often would we expect the average weight of each group to differ?

STATISTICALLY:

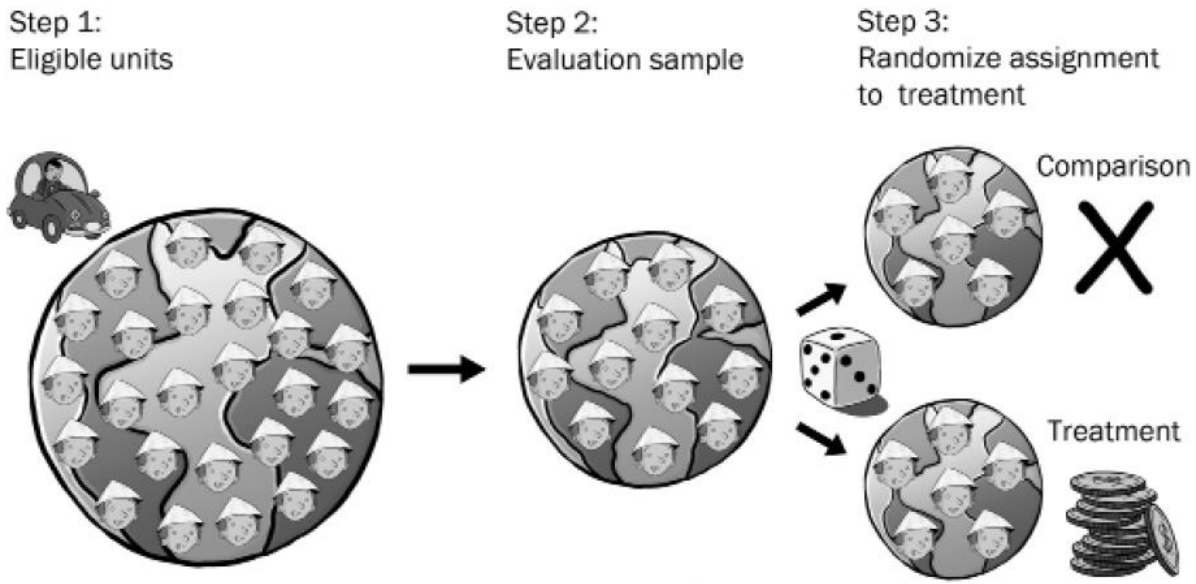
TEST OF GROUP MEANS

5 times out of 100 the two samples are drawn from the same population (or weight distribution), but we will still consider them to be different.

What does alpha=0.05 mean?

[--- 95% confidence interval ---]

How often will randomization "fail"?



By definition of a test at a 95% confidence level, each measured characteristic like weight will differ

5 out of 100 times

"Unhappy Randomization"

Unhappy randomization is not failed randomization (process applied improperly), rather just bad luck of the draw

Table 4.1 Case 3—Balance between Treatment and Comparison Villages at Baseline

Household characteristics	Treatment villages (N = 2964)	Comparison villages (N = 2664)	Difference	*t*-stat
Health expenditures ($ yearly per capita)	14.48	14.57	−0.09	−0.39
Head of household's age (years)	41.6	42.3	−0.7	−1.2
Spouse's age (years)	36.8	36.8	0.0	0.38
Head of household's education (years)	2.9	2.8	0.1	2.16*
Spouse's education (years)	2.7	2.6	0.1	0.006
Head of household is female = 1	0.07	0.07	−0.0	−0.66
Indigenous = 1	0.42	0.42	0.0	0.21
Number of household members	5.7	5.7	0.0	1.21
Has bathroom = 1	0.57	0.56	0.01	1.04
Hectares of land	1.67	1.71	−0.04	−1.35
Distance to hospital (km)	109	106	3	1.02

"contrasts"

Source: Authors' calculation.

* Significant at the 5 percent level.

The most important table in every study: comparisons of treatment and control group characteristics

For the counterfactual to be **valid**, the groups can ONLY differ by the treatment, not by any measured traits.

Is this problematic?

What is the appropriate test for "identical" or equivalent groups?

We should observe no differences in measured traits.
Assume a 95% confidence interval.

Bonferroni Correction:

When we want to be 95% confident that two groups are the same, and we can measure those groups using a set of contrasts, then our decision rule is no longer to reject the null (that the groups are the same) if the p-value < 0.05. A “contrast” is a comparison of means of any measured characteristic between two groups.

If we have a 5% chance of observing a p-value of less than 0.05 for each contrast, then the probability of observing at least one contrast with a p-value that small is greater than 5%! It is actually n*0.05 (minus prob of observing multiple < 0.05 at same time) where n is the number of contrasts.

So if we want to be 95% confident that the groups are different (not just the contrasts), we have to adjust our decision rule to α/n.

For example, if we have 10 contrasts, then our decision rule is now 0.05/10, or 0.005. The p-value of at least one contrast must be below 0.005 for us to conclude that the groups are different.

```
x1 <- rbinom( 10000, 6, 0.05 )
table( x1 ) / 10000
y1 <- rbinom( 10000, 6, 0.05/6 )
table( y1 ) / 10000
```

Test for Group Equivalence

seven contrasts reported

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

Characteristic	All students in the study: Choice students	All students in the study: Control students	All students in the study: p value[a]	All students with scores three or four years after application: Choice students	All students with scores three or four years after application: Control students	All students with scores three or four years after application: p value[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Smallest p-value in table

New alpha = (0.05 / 7) = 0.0071

0.04 > 0.0071

Do not reject :: Groups are equivalent

RCT versus Natural Experiments:

1. **RCT assumes complete control over the assignment process**

2. **Natural Experiments often utilize randomization:**

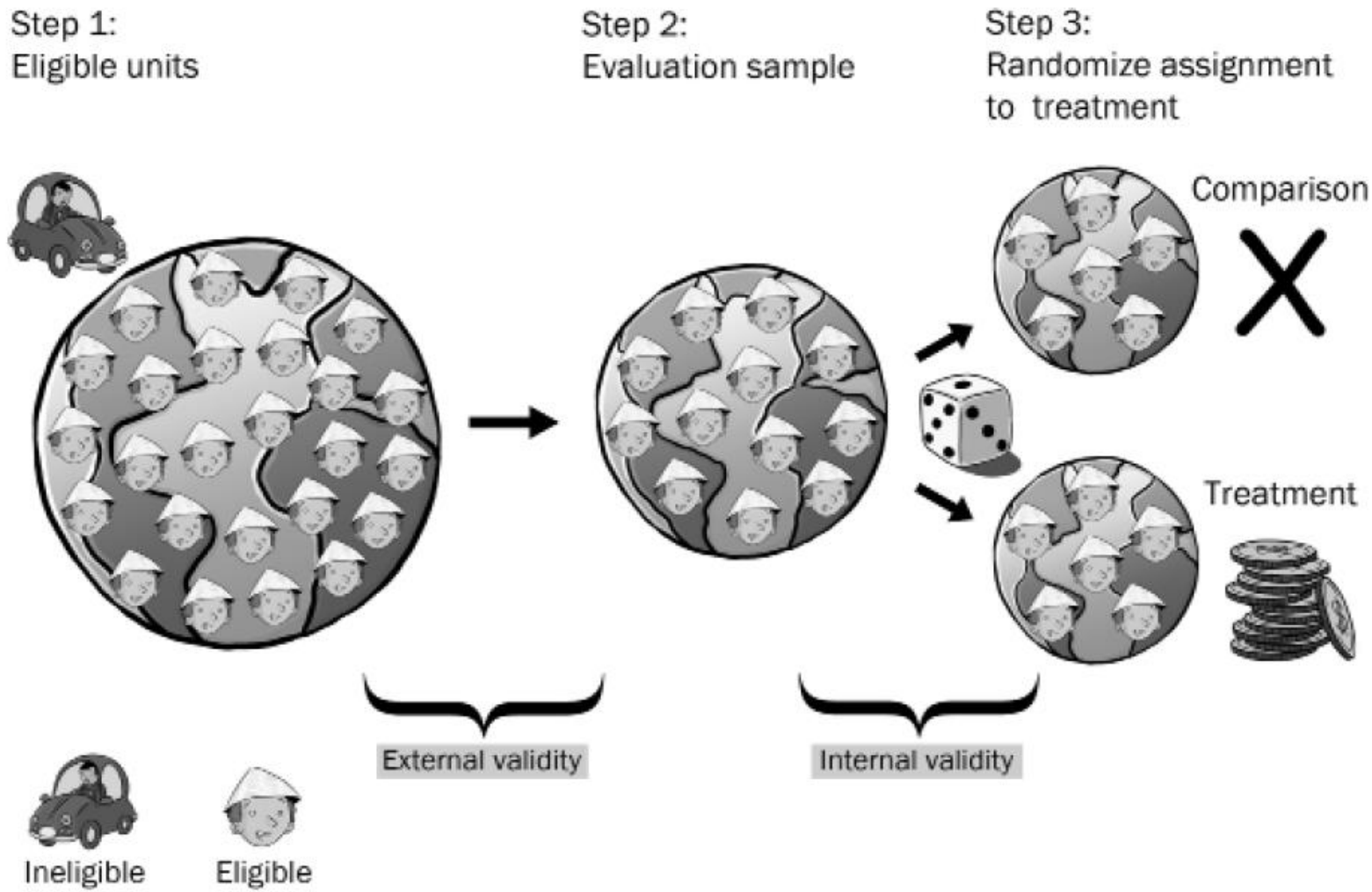
   → **Charter School lotteries**

   → **Vietnam draft**

3. **Quasi-Experimental techniques can use other methods to create group equivalence (for example, matching)**

ATTRITION

Figure 4.3 Steps in Randomized Assignment to Treatment



Tests for Selection-Into Study Group

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

Characteristic	*All students in the study*			*All students with scores three or four years after application*		
	Choice students	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Non-Random Attrition

If the people that leave a program or study are different than those that stay, the calculation of effects will be biased.

The Fix:

Examine characteristics of those that stay versus those that leave.

Non-Random attrition tests for selection OUT of the study group

TABLE 2

Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

Characteristic	*All students in the study*			*All students with scores three or four years after application*		
	Choice students	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

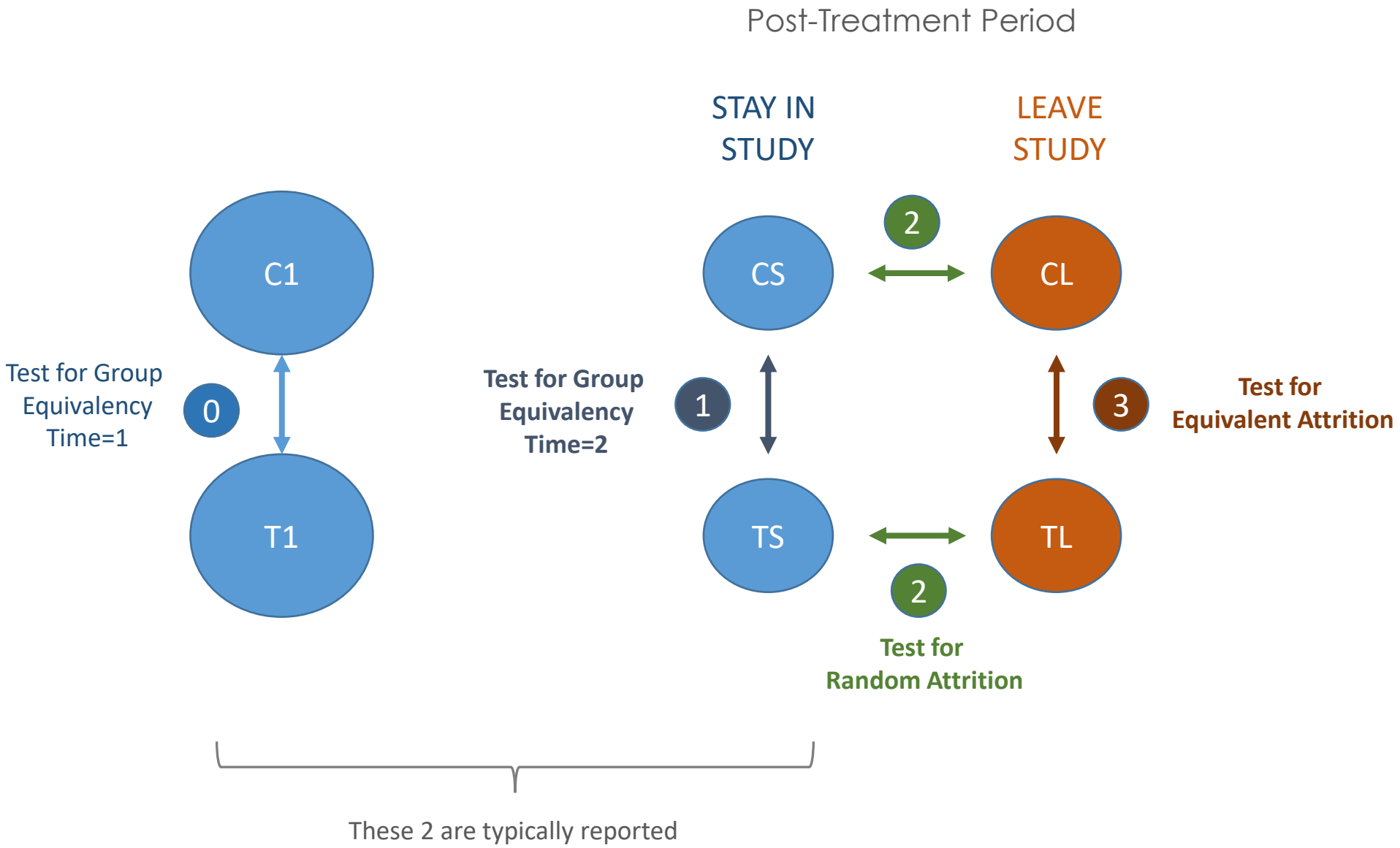
a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Do group traits differ after attrition occurs?

Attrition is natural, the question is whether it is random (will not change the groups) or non-random (will change the groups)

Can also be tested by comparing traits of those that stay to those that leave.

Attrition Tests



Most studies will report group equivalence before and after attrition occurs (tests 0 and 1 in the diagram). But you could test for random attrition in two other ways as well.

I am using the term "random attrition" a little imprecisely. You could have non-random attrition, but as long as it's non-random in the same way in both groups it would still result in balanced groups (all high performers leave from each group, for example).

Test 1 looks at ***group balance*** after attrition. It would be agnostic to the type of attrition, but as long as it did not impact balance it should not matter.

Test 2 is the only real test for ***random attrition***. Are those that stay the same as those that left?

Test 3 would ensure that both groups experience the ***same type*** of attrition, but it does not necessarily have to be random.

For any of these tests you would subset the data to isolate the two groups, create a dummy variable for one group, then run the regression:

$Y = b0 + b1(dummy) + e$

If b1 is significant it means the groups are not balanced (1), attrition is non-random (2), or attrition is non-equivalent (3).

Y can be the pre-treatment outcome, or it can represent a series of study participant traits, like what you did in Lab 2.

In all cases, though, the goal is to ensure that ***attrition does NOT deleteriously impact the integrity of the counterfactual.***

TABLE 2

Background Characteristics of Students in Treatment and Control Groups
(Total numbers of cases in parentheses)

Characteristic	*All students in the study*			*All students with scores three or four years after application*		
	Choice students	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

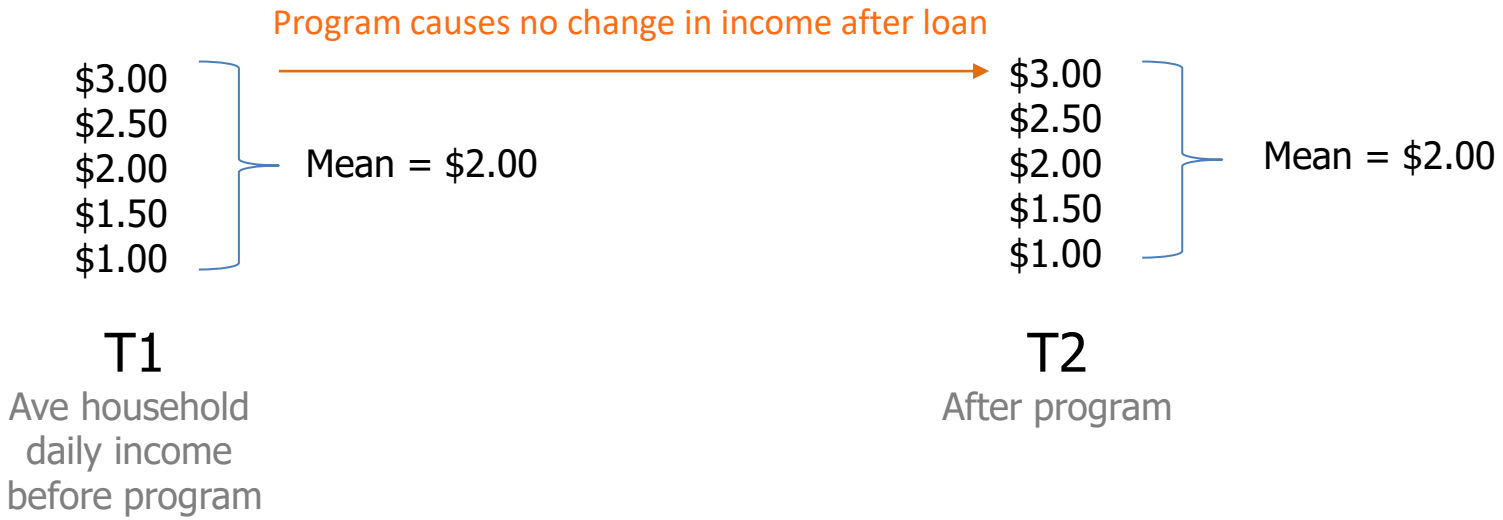
a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

Test for Group Equivalency Time=1 0

Test for Group Equivalency Time=2 1

Microfinance example of bias from selection OUT OF a study group

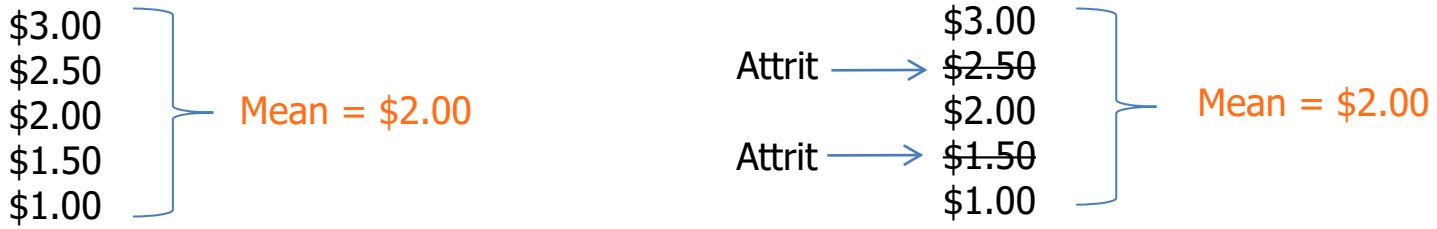
Reflexive design



$$T2 - T1 = 0$$

causal estimate is unbiased

Random Attrition Example

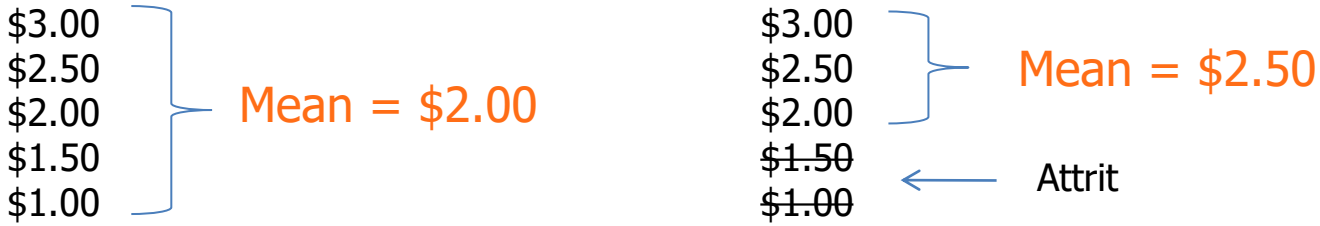


T2 - T1 = 0

Impact study accurately represents program effects

Program is not determined to be effective (no change)

Non-Random Attrition Example

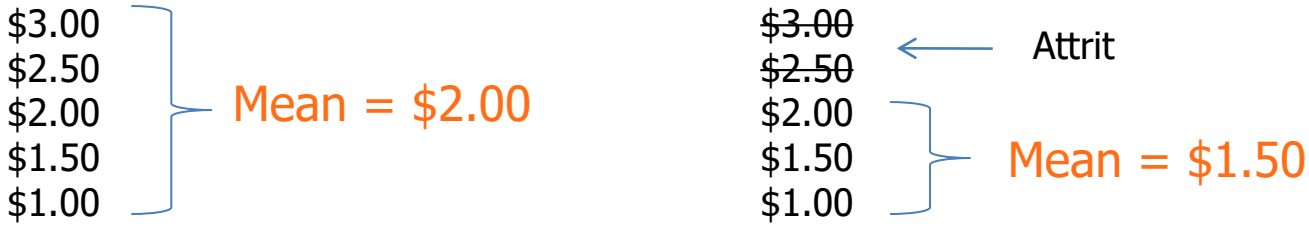


$$T2 - T1 \neq 0$$

We over-estimate program effects

Program appears to be effective

Non-Random Attrition Example



$$T2 - T1 \neq 0$$

We under-estimate program effects

Program appears to harm families

WHAT DO WE MEAN BY “TREATED”?

HOW NON-COMPLIANCE CHANGES OUR MEASURE OF EFFECTS

One of The Physicists Behind The Higgs Boson Has Made an Algorithm to Replace The Pill

It's up to 99.5% effective at stopping pregnancy.

Those more effective methods are ones that don't require people to remember to take a pill, put on a condom, or record their temperature daily, such as intrauterine contraception or implants.

That's because human error can mess with things quite a lot. In fact, the UK's National Health Service (NHS) explained that **when the app was used perfectly all the time**, only five out of every 1,000 women would fall pregnant every year - a rate slightly better than the pill (**99.5 percent**).

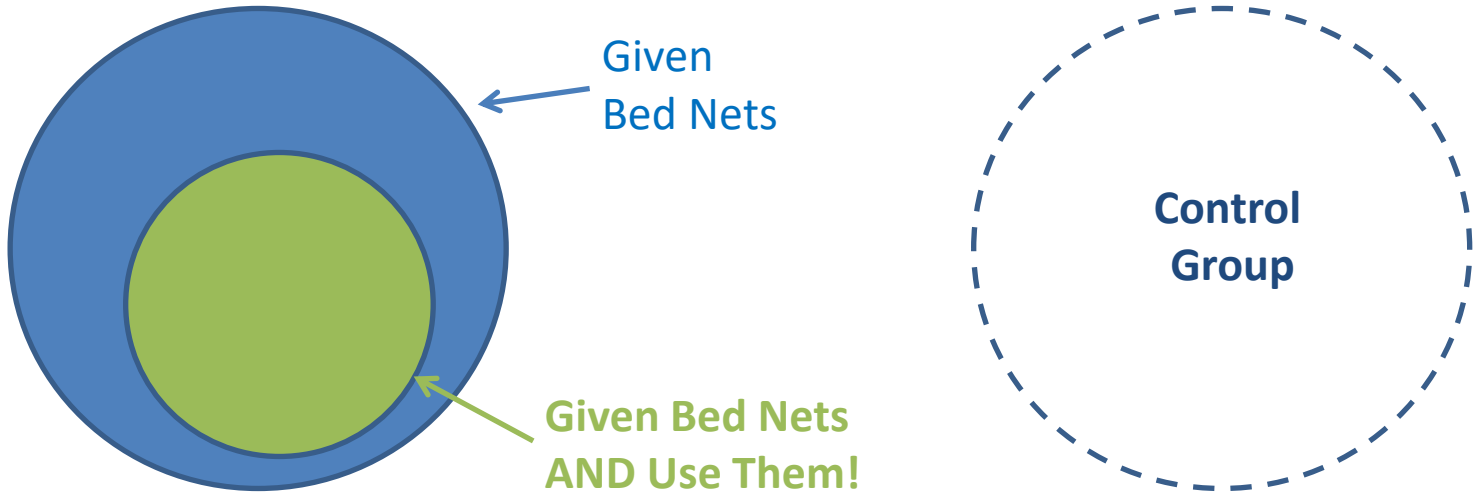
But for **"typical use" - where the app isn't used entirely correctly every day** - it's more likely that seven out of every 100 women would experience accidental pregnancies, which is around **93 percent** efficiency.

What is the TRUE effectiveness of the app?

99.5 percent, or 93 percent?

https://www.sciencealert.com/one-of-the-physicists-behind-the-higgs-boson-has-made-an-algorithm-to-replace-the-pill

Estimation of the counter-factual:

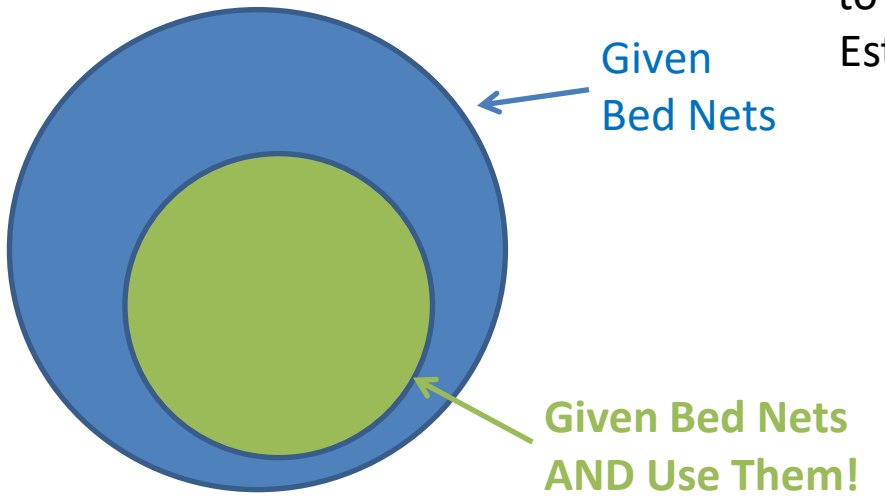


Program Effect = T2 – C2

Is Group T (for treatment) those that are GIVEN bed nets, or those that USE them?

TERMINOLOGY:

- "Average" treatment effects
  - **Treatment on the Treated (TOT) Effects**
  - **Intention to Treat (ITT) Effects**



"Intention to Treat" Estimate

"Treatment on the Treated" Estimate

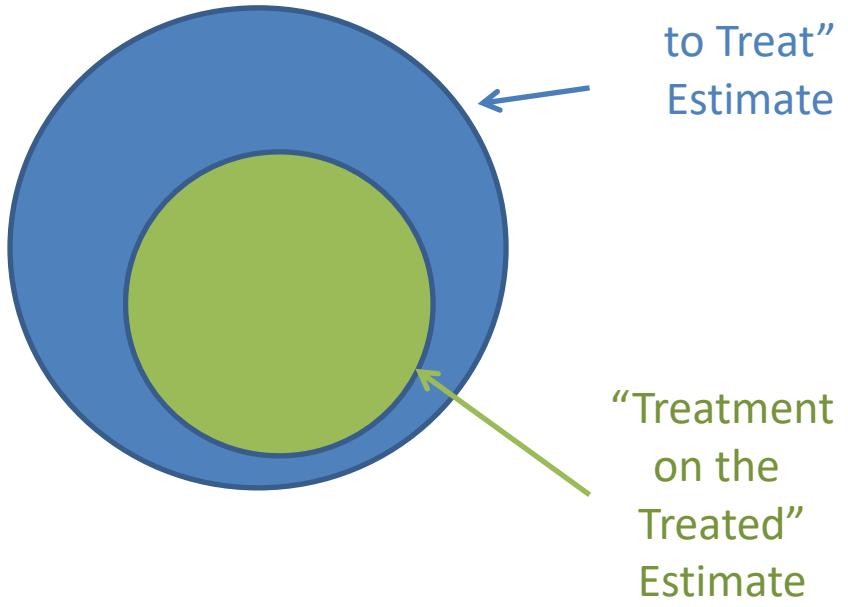
2 Valid Ways to Measure the Treatment

WHY DO WE NEED TWO MEASURES?

THE TOT MEASURE is the optimistic or best-case scenario. It tells us how effective the program or intervention is when followed with HIGH FIDELITY.

THE ITT MEASURE is the more cynical or realistic version. We never expect that programs work exactly as designed. The ITT is closer to a measure of how the laboratory equipment works once it's in the field. It is also a better estimate of how much change we can expect at the population level.

The different between the TOT and the ITT tells us how many gains can be made by improving program implementation! So both are useful and important!



One of The Physicists Behind The Higgs Boson Has Made an Algorithm to Replace The Pill

It's up to 99.5% effective at stopping pregnancy.

Those more effective methods are ones that don't require people to remember to take a pill, put on a condom, or record their temperature daily, such as intrauterine contraception or implants.

That's because human error can mess with things quite a lot. In fact, the UK's National Health Service (NHS) explained that **when the app was used perfectly all the time**, only five out of every 1,000 women would fall pregnant every year - a rate slightly better than the pill (**99.5 percent**).

But for **"typical use" - where the app isn't used entirely correctly every day** - it's more likely that seven out of every 100 women would experience accidental pregnancies, which is around **93 percent** efficiency.

Which is treatment on the treated effect?

Which is the intention-to-treat effect?

https://www.sciencealert.com/one-of-the-physicists-behind-the-higgs-boson-has-made-an-algorithm-to-replace-the-pill

NOTE !!!

These two measures are **NOT** about attrition.

They are about **refusing the treatment**, or failing to follow the properly-prescribed treatment regiment (or one school burning down during the study).

Non-compliance means participants refuse to receive or comply with the treatment. But they do not refuse to participate in the study – **we can still measure their performance in the second time period.**

Attrition means people left the study. You can accept the treatment and still attrit. Or you can refuse treatment and refuse participation. Either way, **attrition means we can't measure the outcome in the post-treatment study period**.