RESEARCH DESIGN FOR COUNTER-FACTUAL ANALYSIS

*Lecy * NLM 530*

Core Concepts

1. Counter-Factual Analysis for Causal Reasoning

2. Randomization Process

3. Control versus Comparison Group

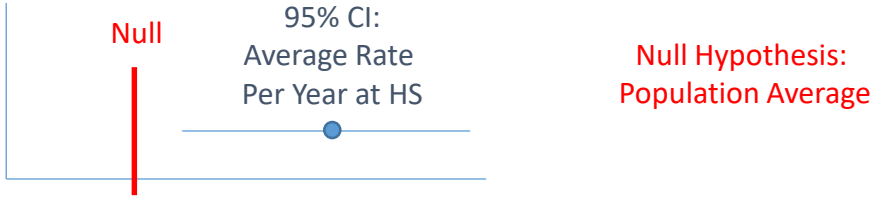
4. Treatment Effects
    1. Average Treatment Effect (ATE)
    2. Intention to Treat (ITT)
    3. Treatment on the Treated (TOT)

5. Calculation of Program Impact (effect size)

COUNTER-FACTUAL ANALYSIS

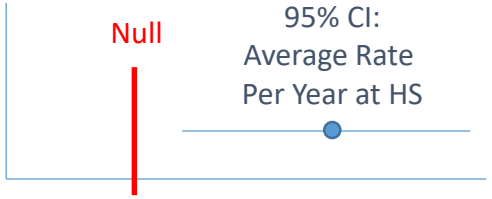
Were suicide rates *HIGH* for a specific high school in suburban California?

Were suicide rates *HIGH* for a specific high school in suburban California?

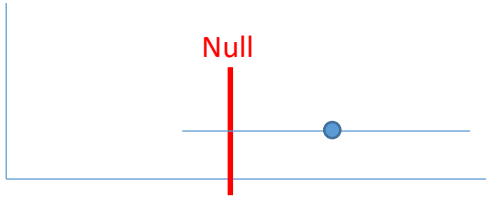


Null Hypothesis:
Population Average

Were suicide rates HIGH for a specific high school in suburban California?



Null Hypothesis:
Population Average

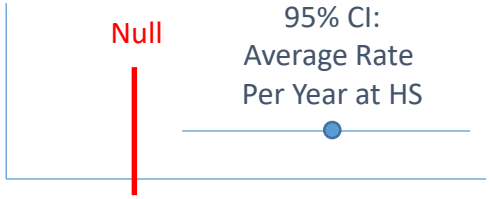


Null Hypothesis:
All HS Students
OR All Californians

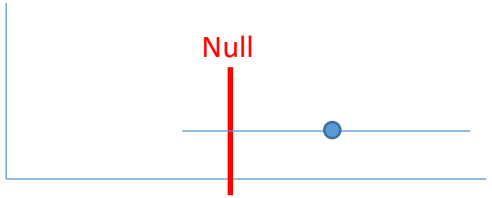
Were suicide rates *HIGH* for a specific high school in suburban California?

These are all valid counter-factuals.

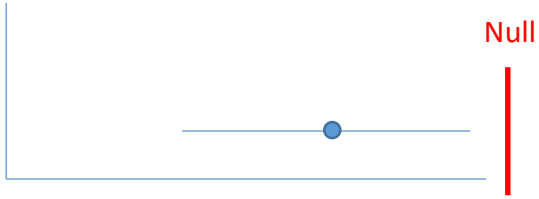
How we define our comparison drives the conclusions.



Null Hypothesis:
Population Average



Null Hypothesis:
All HS Students
OR All Californians



Null Hypothesis:
All Suburban HS
Students

A VALID COUNTER-FACTUAL ALLOWS US TO ANSWER THE FOLLOWING TWO QUESTIONS:

1) **Compared to what?** The program outcomes are different than outcomes in the comparison group. The comparison group is defined by the researcher.

   In some special cases the comparison group is **identical** (statistically speaking) to the treatment group. In this case we call it a “control” group.

A VALID COUNTER-FACTUAL ALLOWS US TO ANSWER THE FOLLOWING TWO QUESTIONS:

1) **Compared to what?** The program outcomes are different than outcomes in the comparison group. The comparison group is defined by the researcher.

   In some special cases the comparison group is identical (statistically speaking) to the treatment group. In this case we call it a “control” group.

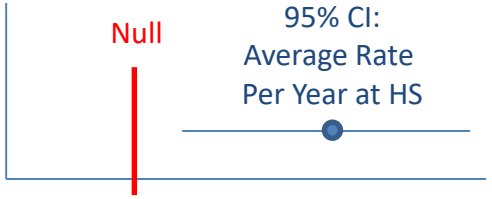
2) **How big is the program effect?** Is the difference meaningful (statistically significant and socially salient)?

   In the simple case the program effects is just the difference of the average outcome of the treatment and control group, but in practice there are many ways we calculate an effect.

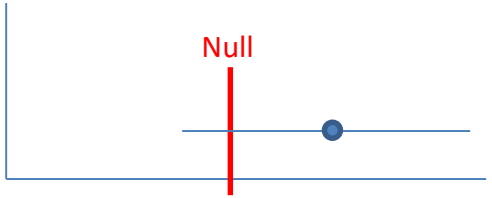
Were suicide rates *HIGH* for a specific high school in suburban California?

These are all valid conter-factuals.

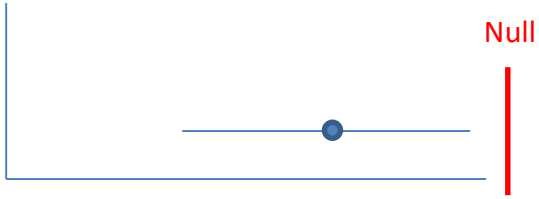
How we define our comparison drives the conclusions.



Null Hypothesis:
Population Average



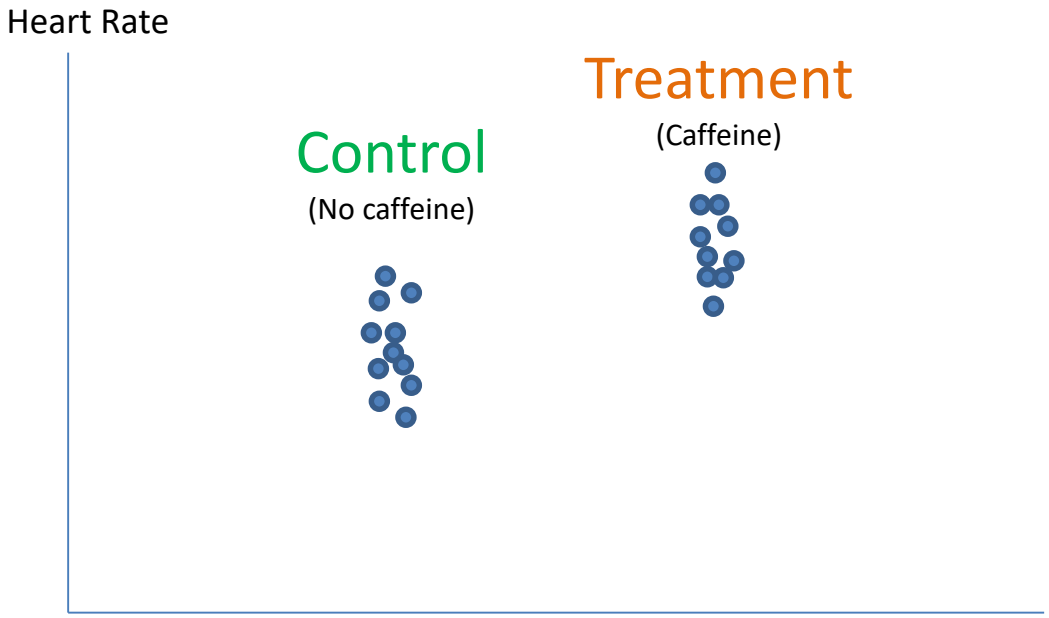
Null Hypothesis:
All HS Students
OR All Californians



Null Hypothesis:
All Suburban HS
Students

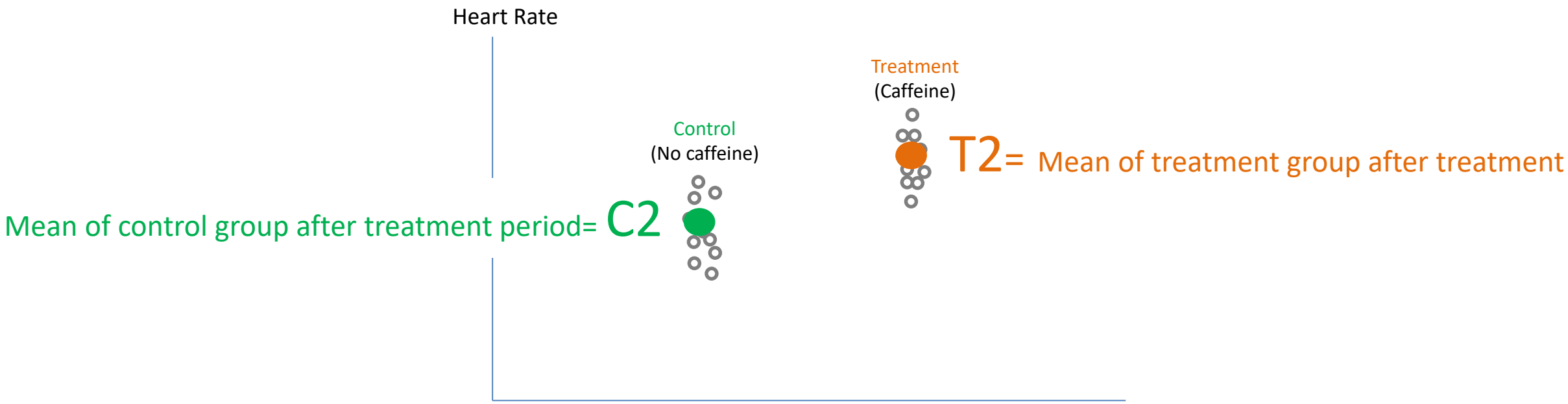
PROGRAM IMPACT: THE “EFFECT”

THE PROGRAM EVALUATION FRAMEWORK: "DISCRETE" TREATMENT GROUPS (YES/NO)

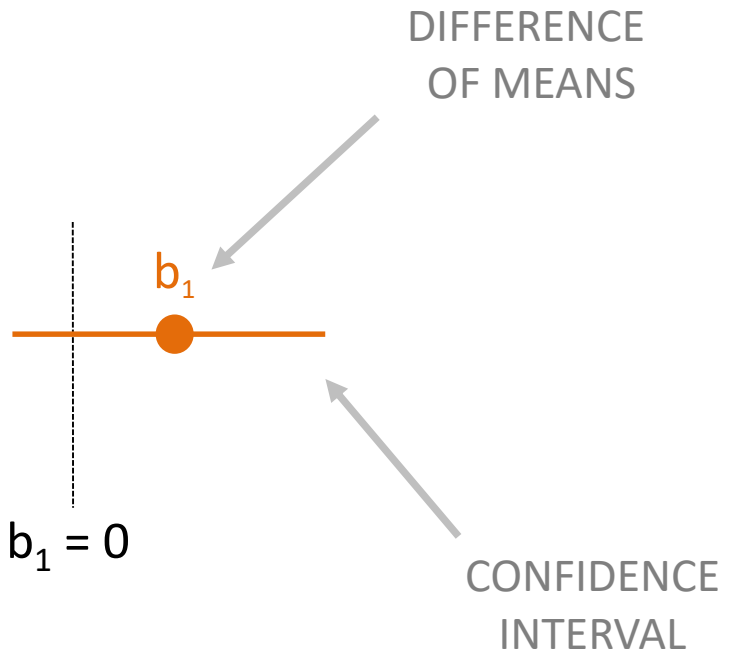


THE PROGRAM EVALUATION FRAMEWORK

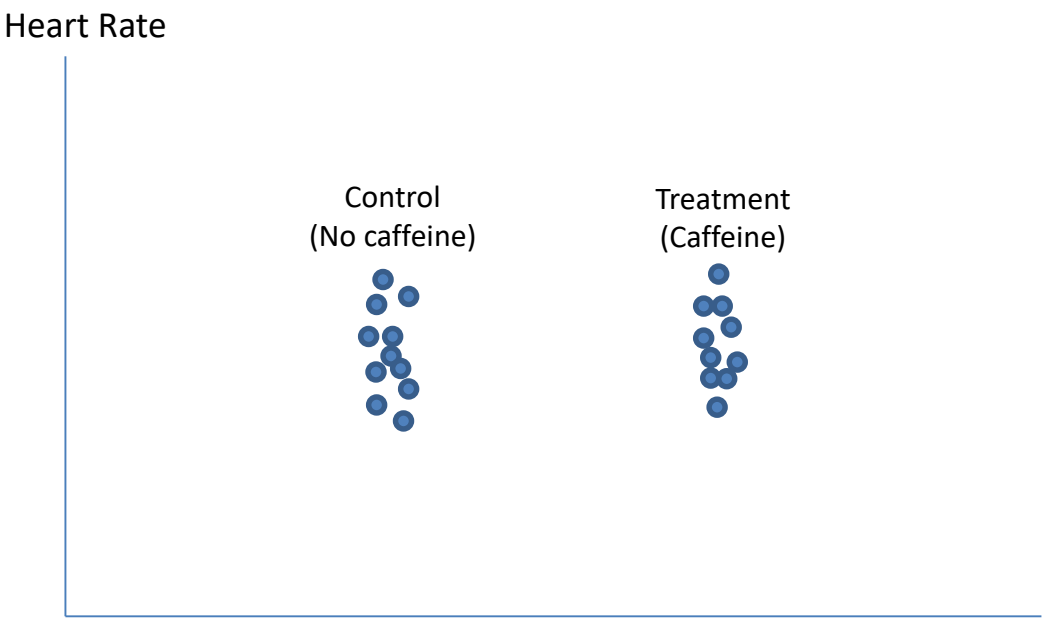
$$Effect = T2 - C2$$

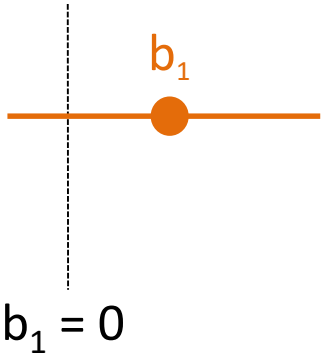


$b_1 = MEAN_{treat} - MEAN_{control}$



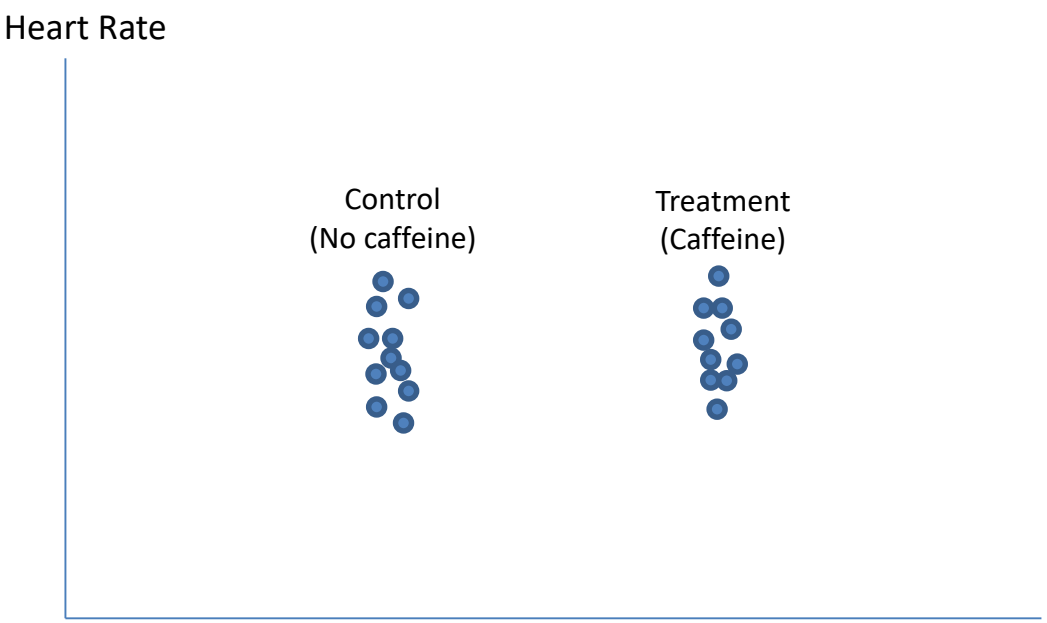
$b_1 = MEAN_{treat} - MEAN_{control}$



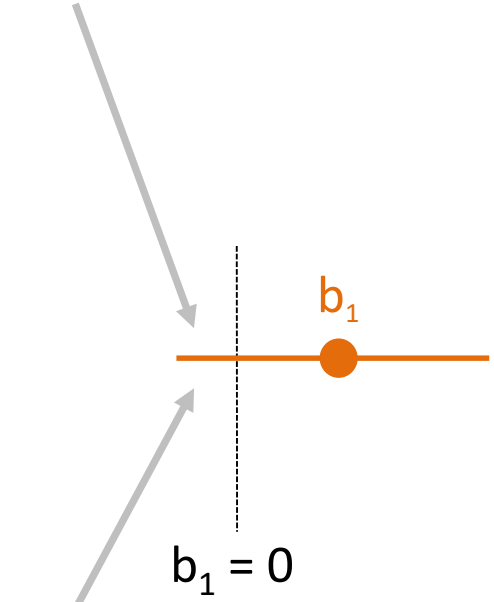


NULL HYPOTHESIS
(b_1=0 means NO IMPACT)

b_1 = $MEAN_{treat}$ - $MEAN_{control}$



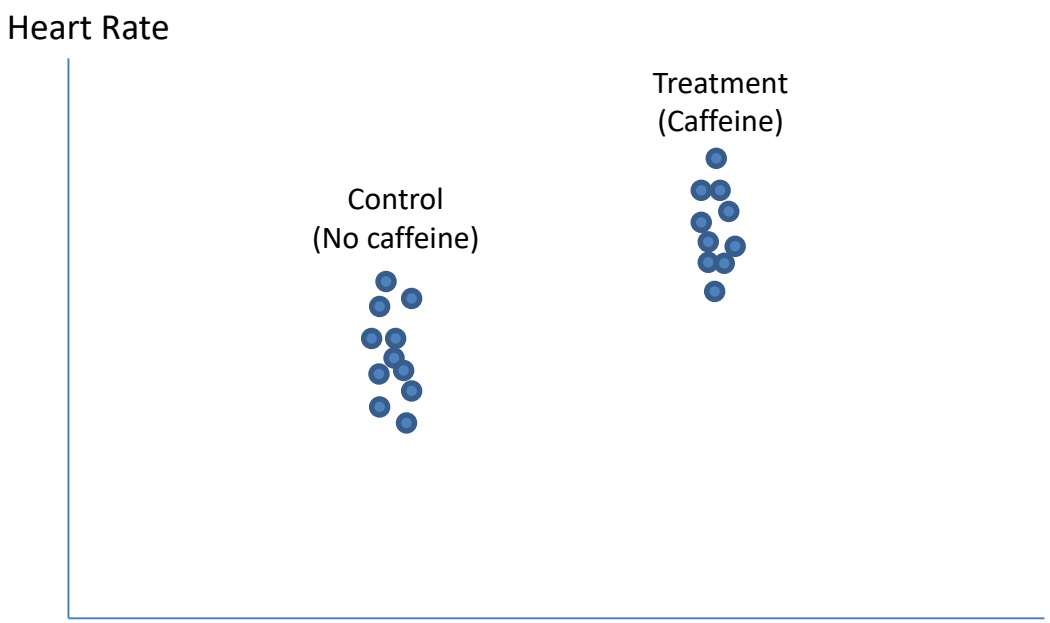
STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

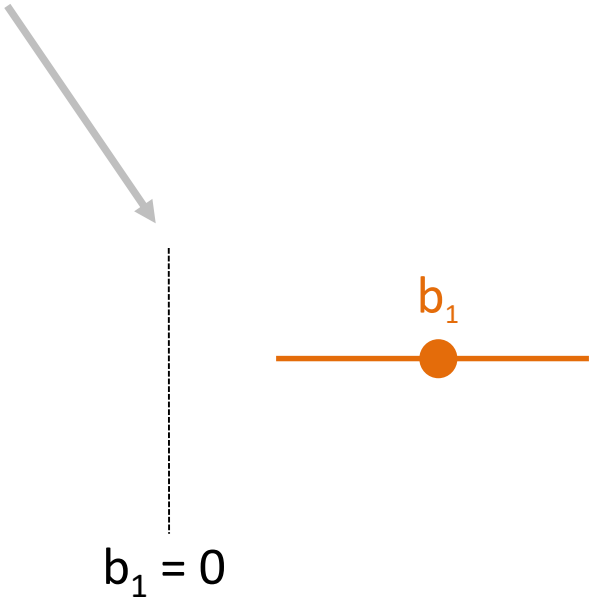


NOT SIGNIFICANT
(NO PROGRAM IMPACT)

STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

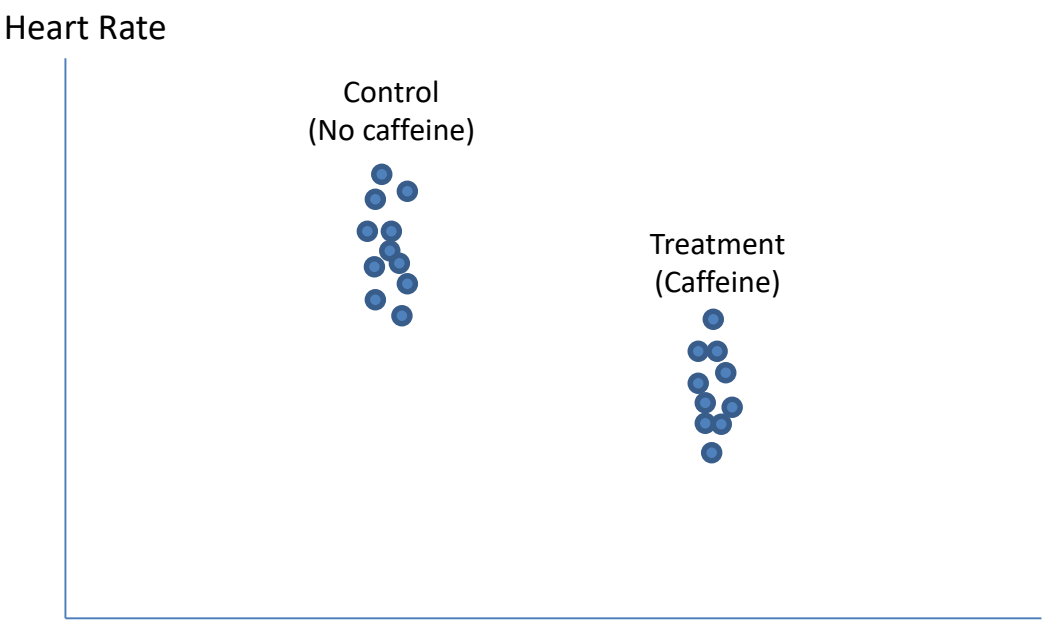
$b_1 = MEAN_{treat} - MEAN_{control}$



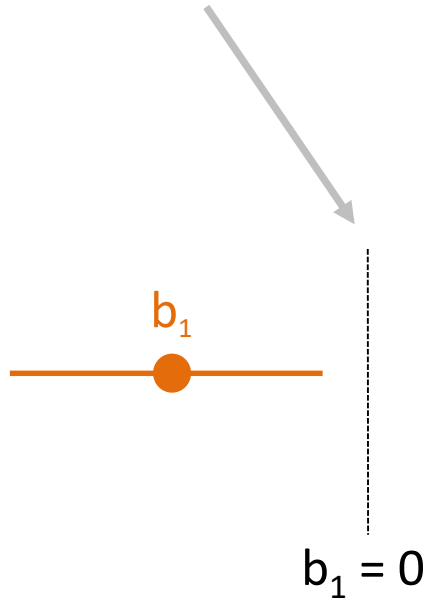


SIGNIFICANT
(POSITIVE PROGRAM IMPACT)

$b_1 = MEAN_{treat} - MEAN_{control}$

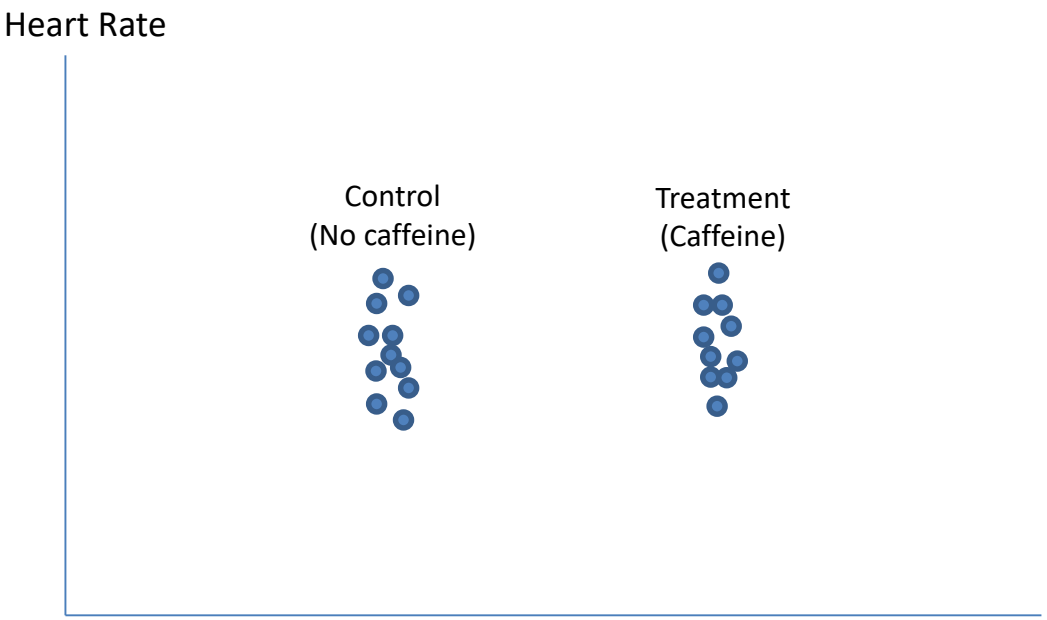


STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

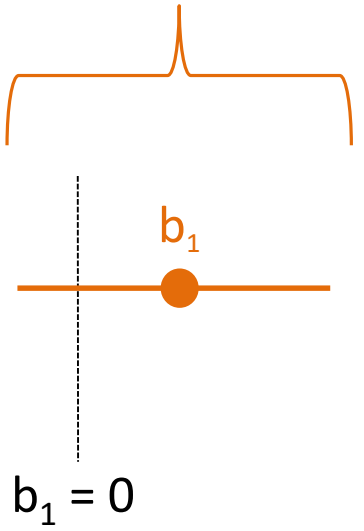


SIGNIFICANT
(NEGATIVE PROGRAM IMPACT)

$b_1 = MEAN_{treat} - MEAN_{control}$



Definition of EFFECT in program eval:
Observed change + confidence interval
(size of observed impact plus accuracy,
can we say with confidence it's positive)



TRUE EXPERIMENTS

Figure 3.1 The Perfect Clone

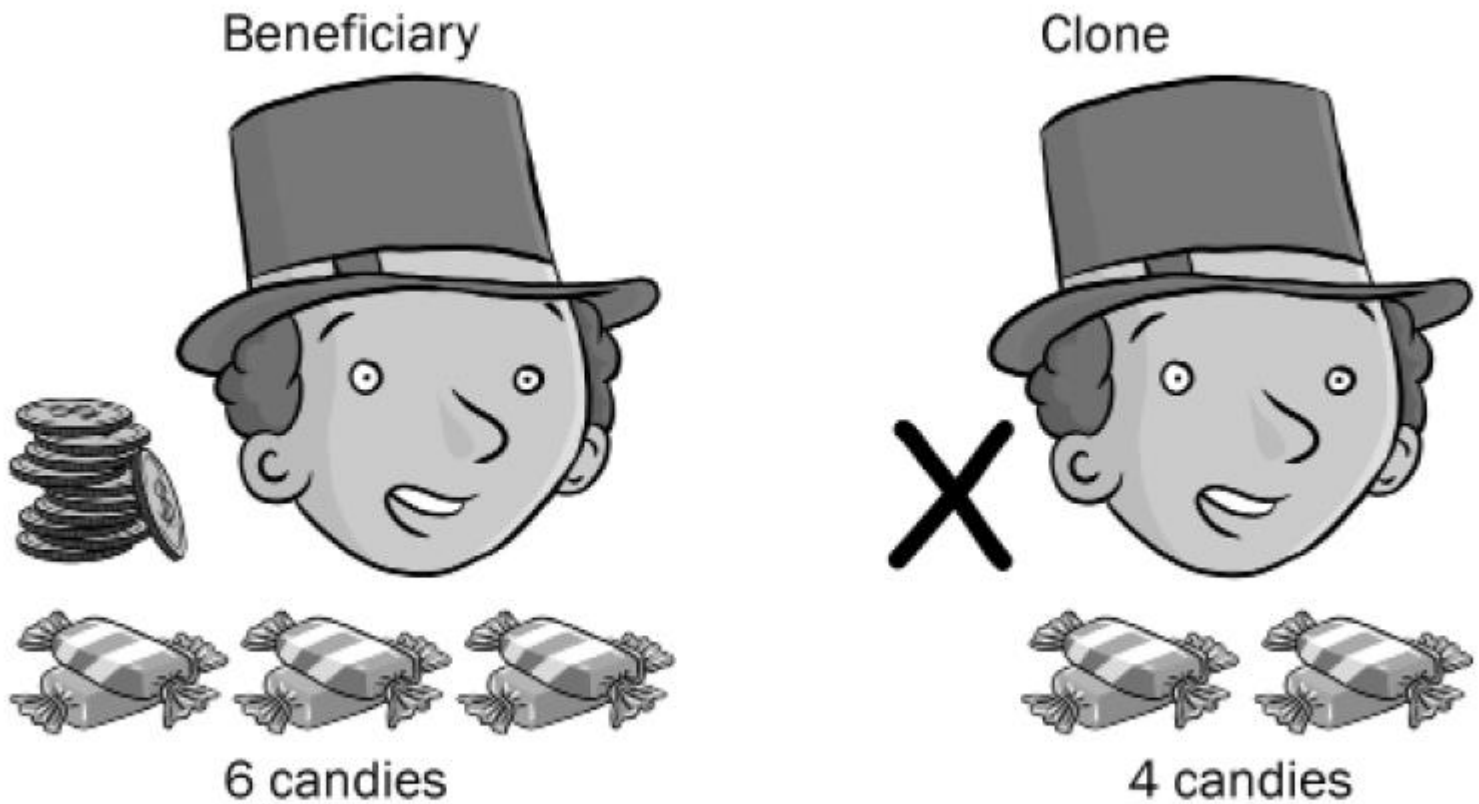


Figure 4.3 Steps in Randomized Assignment to Treatment

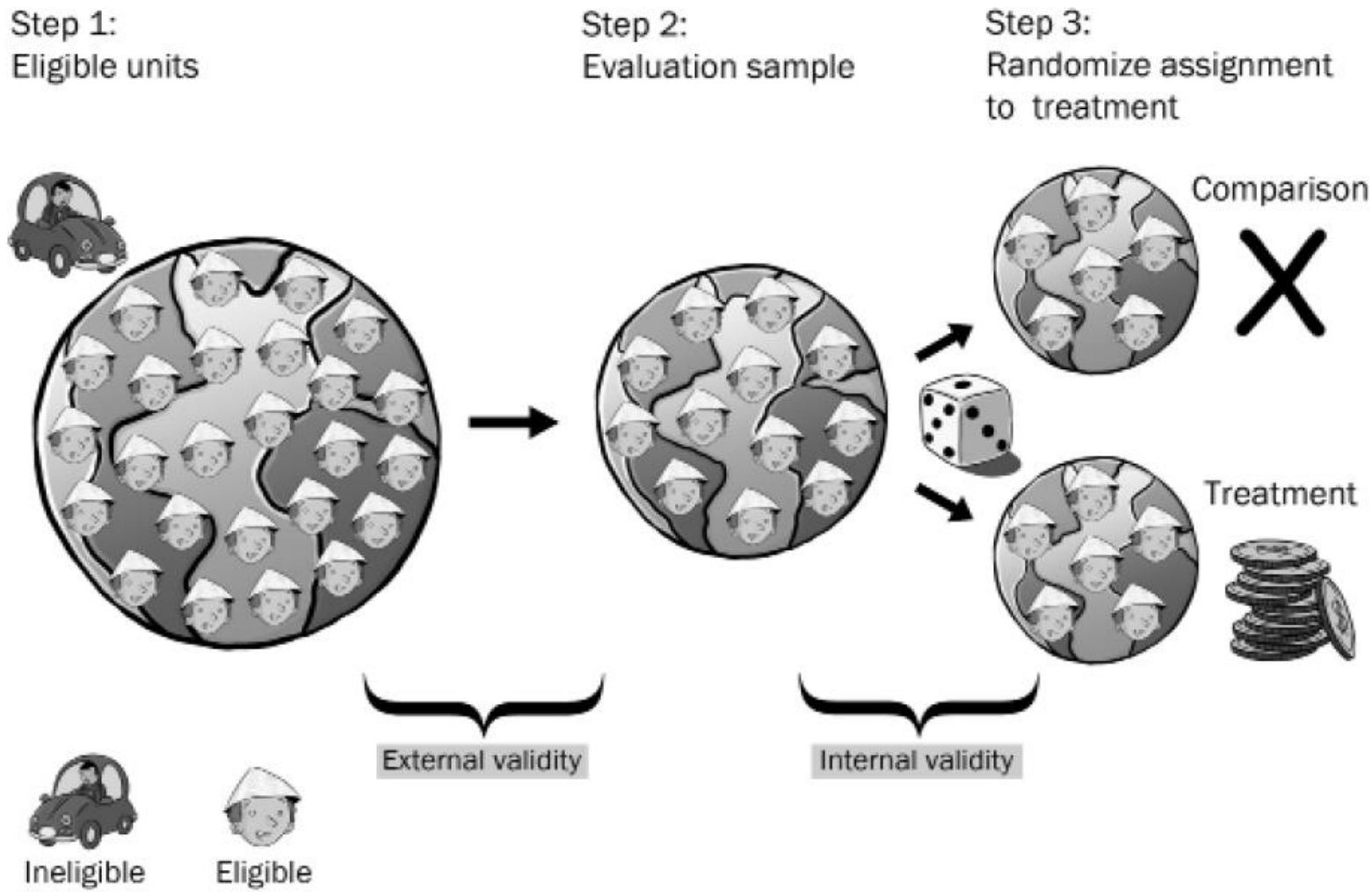
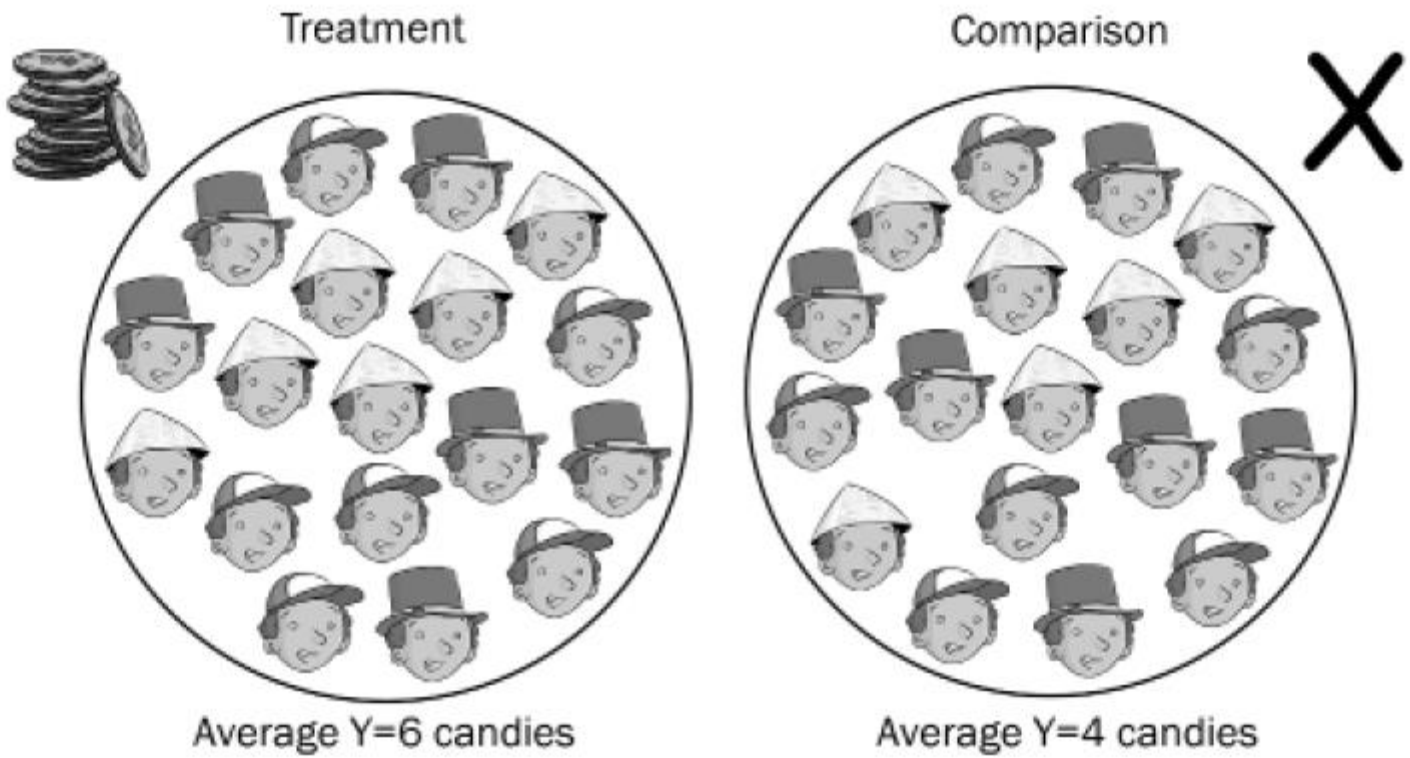


Figure 3.2 A Valid Comparison Group



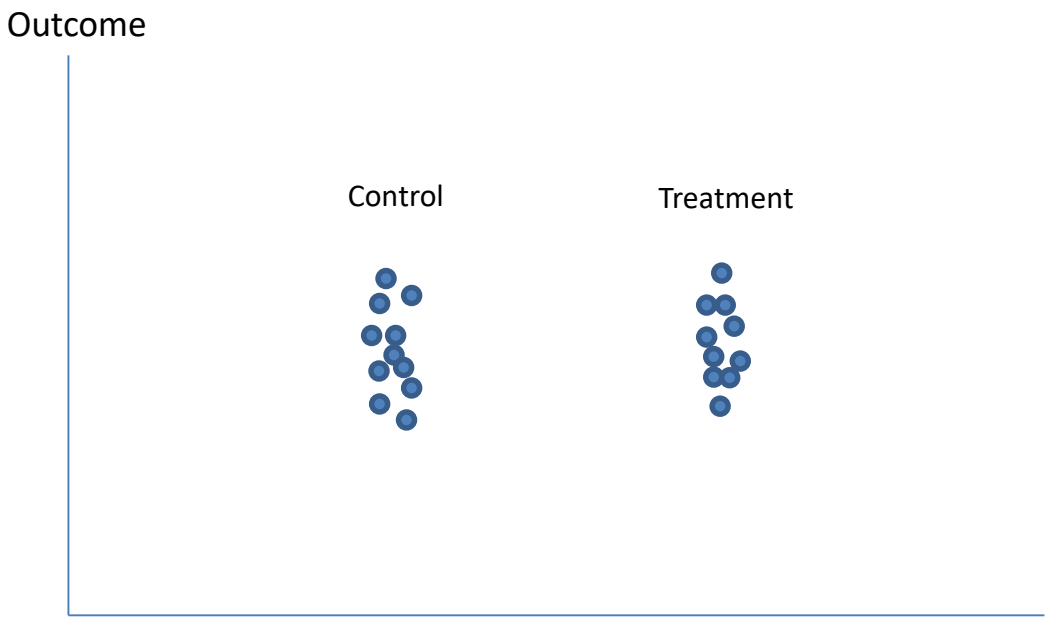
Impact = 6 - 4 = 2 candies

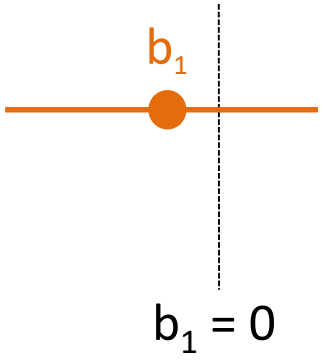
b_1 = T2 – C2

$b_1 = MEAN_{treat} - MEAN_{control}$

$b_1 = T2 - C2$

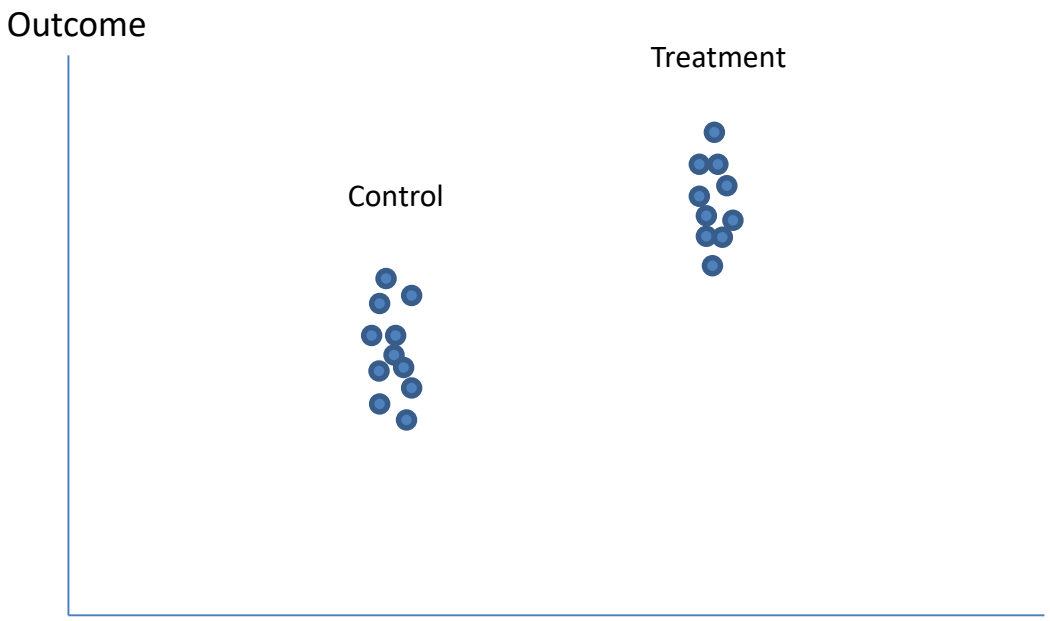
No Program Impact



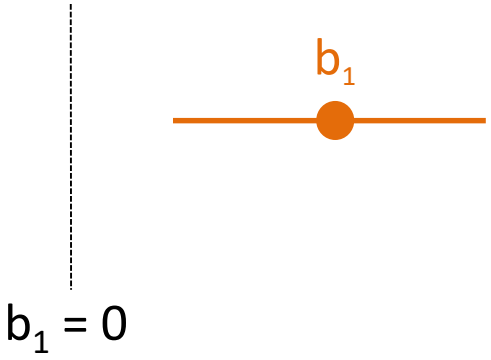


$b_1 = MEAN_{treat} - MEAN_{control}$

$b_1 = T2 - C2$



Positive Program Impact



HAPPY RANDOMIZATION

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

Source: Authors' calculation.

* Significant at the 5 percent level.

Is this problematic?

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 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*.05, 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 "Happy" Randomization

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.

0.05 / 6 = 0.0083

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

RCT versus Natural Experiments:

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

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

   → **Charter Schools Example**

   → **Vietnam Veterans Example**

SELECTION PROBLEM

Microfinance example of bias from selection INTO a study group

Number of people per category

	Not an Entrepreneur	Entrepreneur
No Loan	30	15
Takes a Loan	20	35

Average weekly income per category

	Not an Entrepreneur	Entrepreneur
No Loan	$10	$20
Takes a Loan	$10	$20

$$No\ Microfinance\ Loan: \frac{30 \cdot \$10 + 15 \cdot \$20}{45} = \$13.33$$

$$Yes\ Microfinance\ Loan: \frac{20 \cdot \$10 + 35 \cdot \$20}{55} = \$16.37$$

The loan appears to have an impact even when we know it didn't!

“Selection” Problem

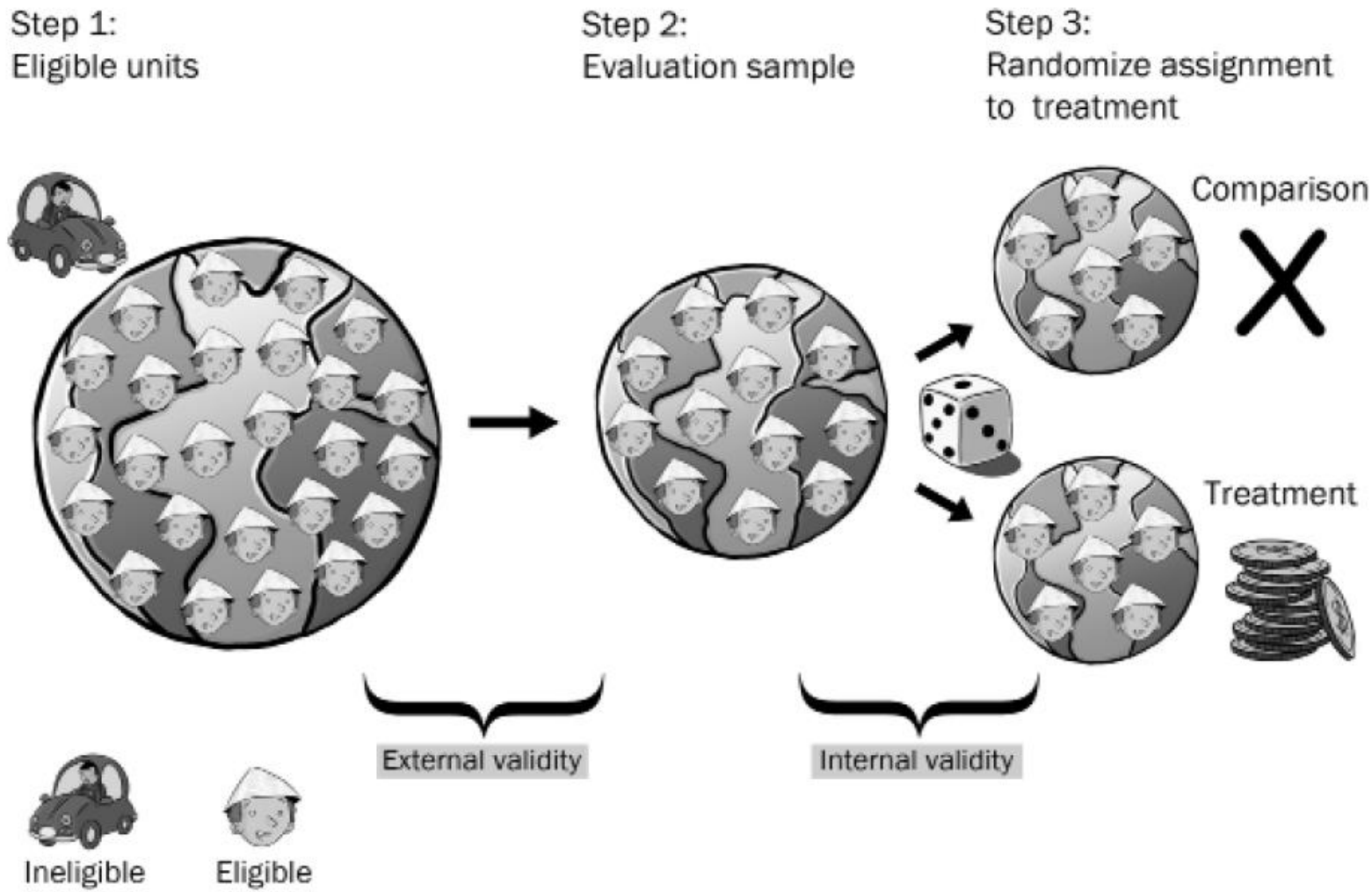
Those that participate in the program (treatment group) are different from those that do not (control group).

This is the biggest problem in impact evaluation!

The Fix:

Use randomization or matching to ensure an apples to apples comparison.

Figure 4.3 Steps in Randomized Assignment to Treatment



Test for "Happy" Randomization

TABLE 2

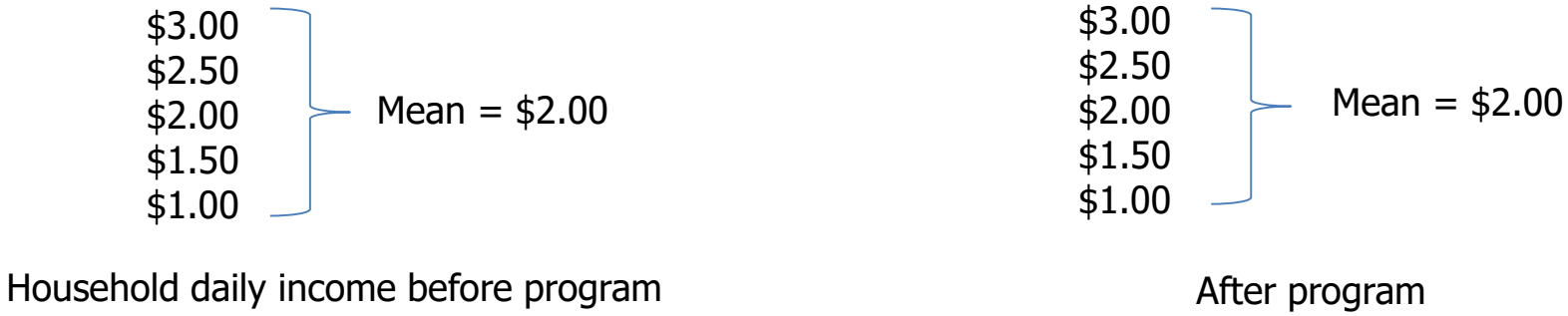
Background Characteristics of Students in Treatment and Control Groups (Total numbers of cases in parentheses)

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.

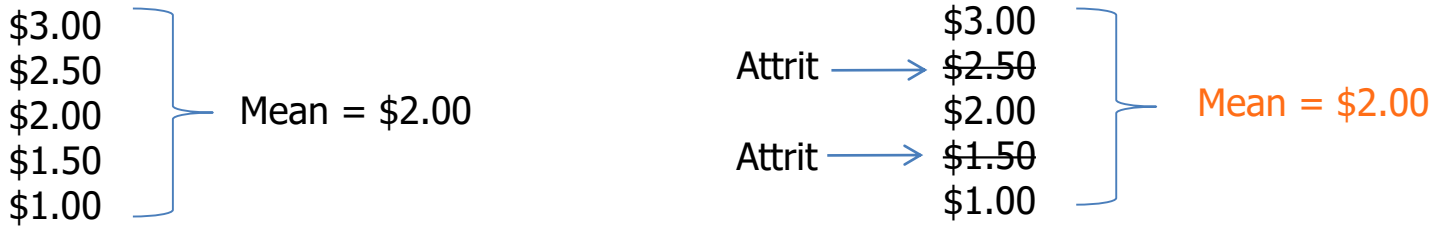
Microfinance example of bias from selection OUT OF a study group

Reflexive design



T2 - T1 = 0, estimate of effect is accurate ("unbiased")

Random Attrition Example



T2 - T1 = 0, Impact study accurately represents program effects

Non-Random Attrition Example

T2 – T1 ≠ 0, We over-estimate program effects

Non-Random Attrition Example

T2 – T1 ≠ 0, We under-estimate program effects

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.

Test for Attrition

TABLE 2

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

	All students in the study			*All students with scores three or four years after application*		
Characteristic	*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.

Can also be tested in another way: do background characteristics of T1 = T2

DIFFERENT INTERPRETATIONS OF PROGRAM 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.

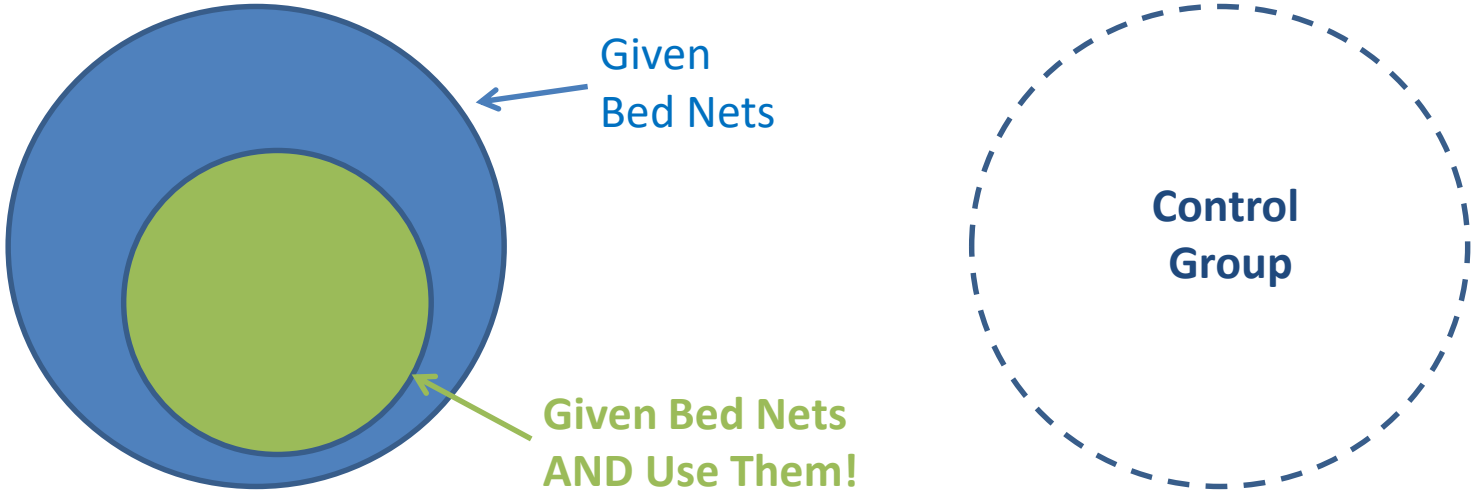
That's because human error can mess with things quite a lot. In fact, the UK's National Health Service (NHS) explained that **when the app was used perfectly all the time**, only five out of every 1,000 women would fall pregnant every year - a rate slightly better than the pill (**99.5 percent**).

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.

Is REAL effectiveness 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:

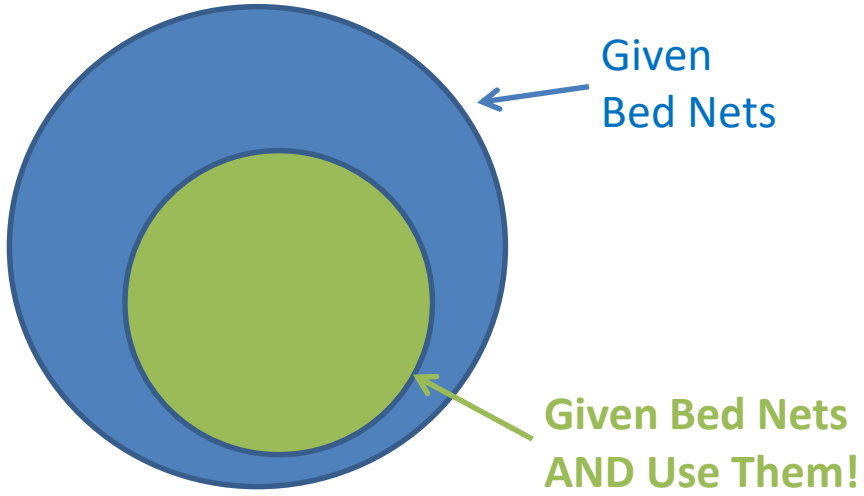


Program Effect = (T2 – T1) – (C2 – C1)

Is Group T 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**



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.

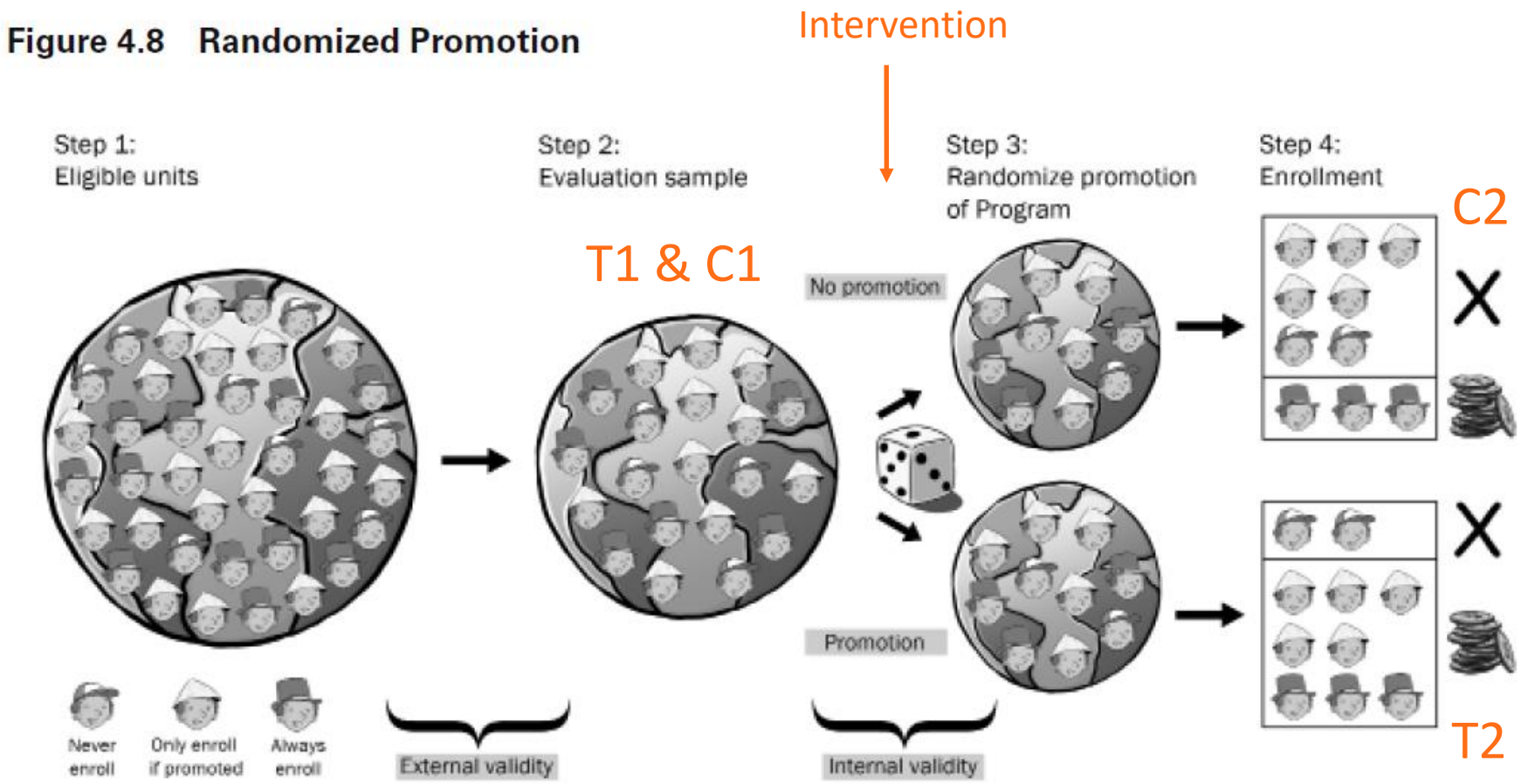
Which is treatment on the treated effect?

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

THREE MAIN TYPES OF COUNTERFACTUAL ANALYSIS

CALCULATION OF TREATMENT EFFECTS:



We compare the group averages for the treatment and control groups after the study. The "effect" represents program impact:

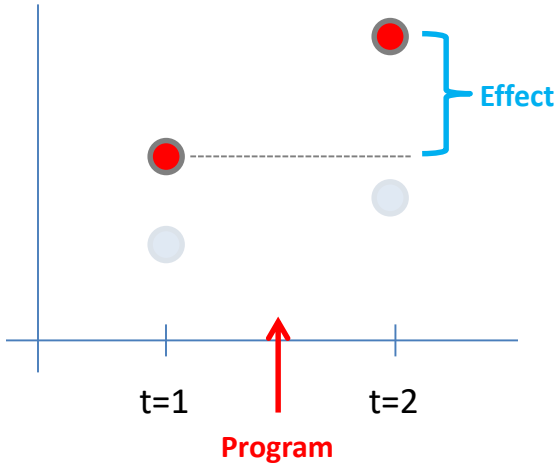
Effect = T2 - C2

VARIETIES OF THE COUNTER-FACTUAL:

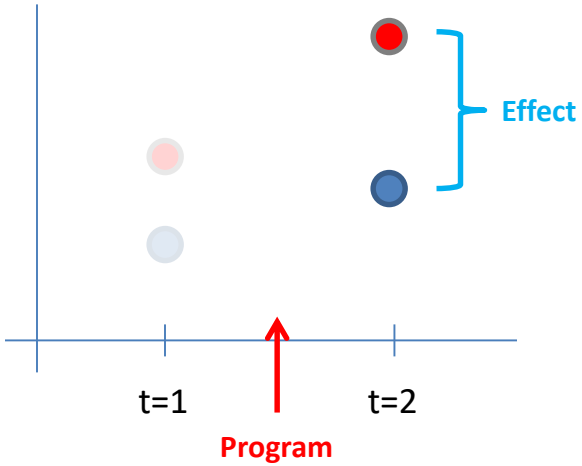
- Treatment Groups, T1=before, T2=after
- Control Groups, C1=before, C2=after

Each counterfactual has a different formula for program effects.

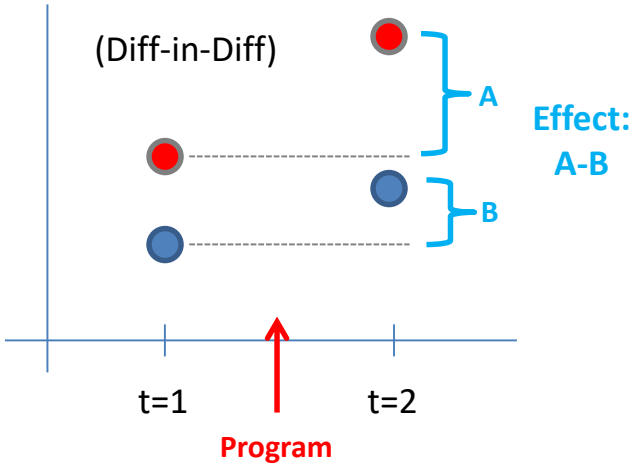
Pre-Post Reflexive (T2-T1)



Post-Only (T2-C2)



Pre-Post
With Control
(T2-T1) – (C2-C1)



THE THREE COUNTERFACTUALS:

1) **Difference-in-Difference:** Gains in the treatment group compared to gains in the comparison group.

   Effect calculation: (T2-T1) – (C2-C1)

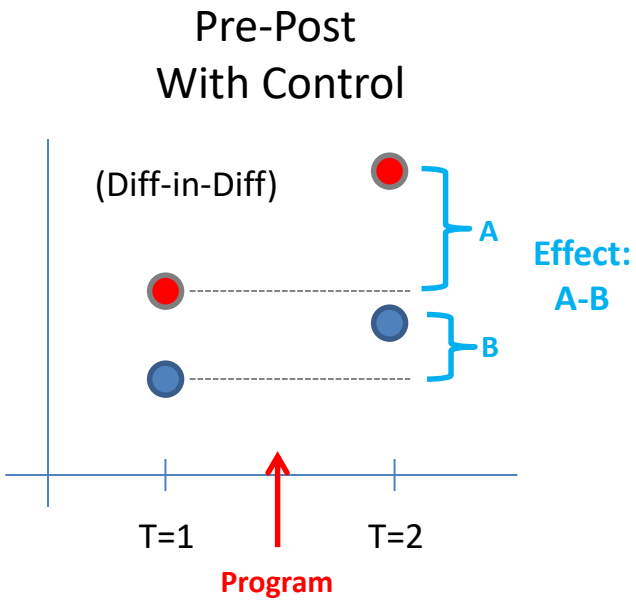
2) **Reflexive Design:** Gains in the treatment group only.

   Effect Calculation: (T2-T1)

3) **Post-Test Only Design:** Differences between the treatment and control groups after the program is delivered.

   Effect Calculation: (T2-C2)

One "strong" counterfactual: Diff-in-Diff



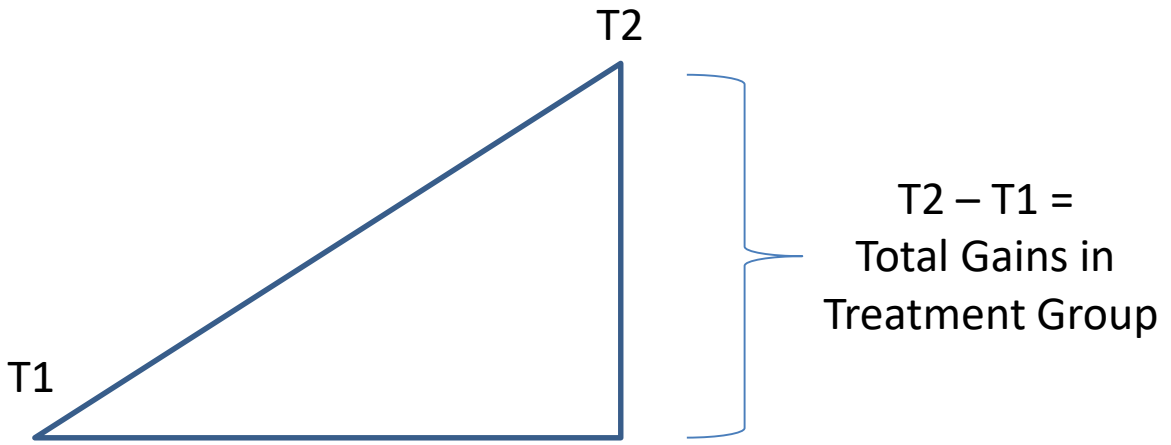
Accounts for initial difference in groups and secular trends

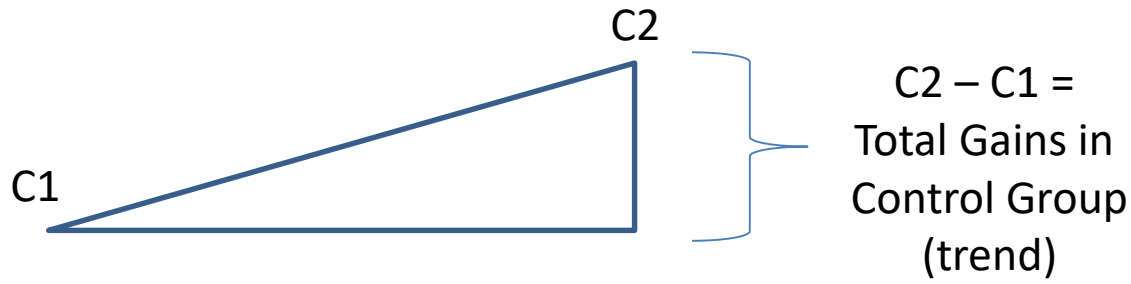
THE DIFFERENCE IN DIFFERENCE ESTIMATOR:

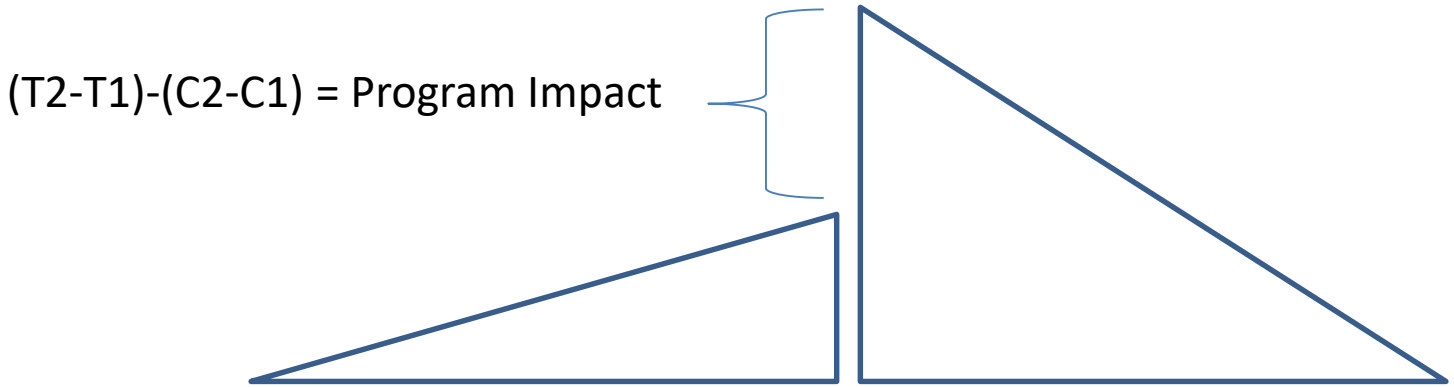
$$\text{Program Impact} = (T2 - T1) - (C2 - C1)$$

Gains during the program period by treatment group

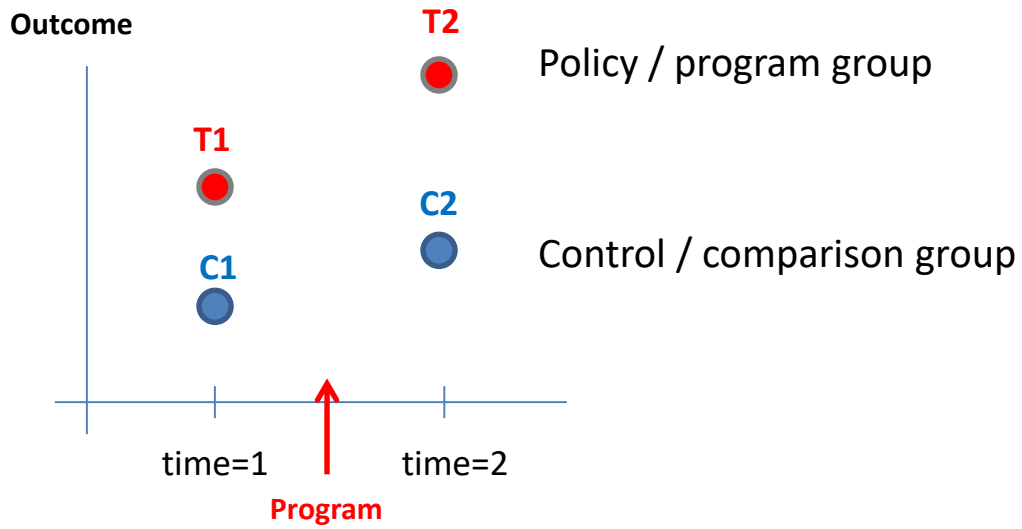
Gains as a result of trend


T2
T1
T2 – T1 =
Total Gains in
Treatment Group


C2
C1
C2 – C1 =
Total Gains in
Control Group
(trend)


(T2-T1)-(C2-C1) = Program Impact

"Control" Versus "Comparison" Groups



Two Considerations:

Does C1 = T1 ?

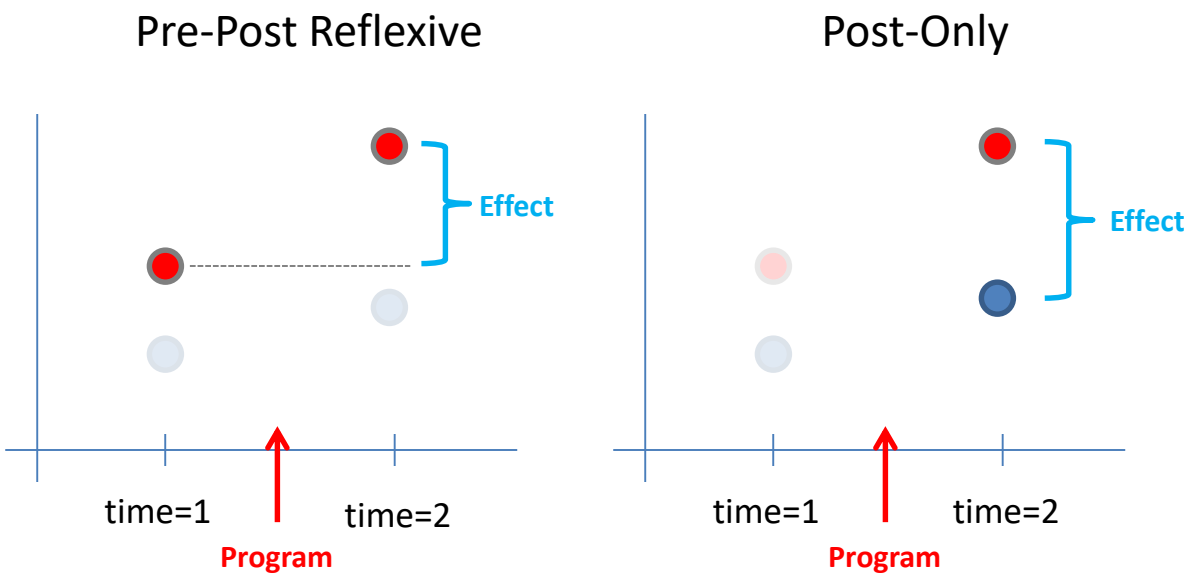
Not usually for comparison groups.

Is C2 – C1 accurate reflection of trend?

In some cases, the comparison group can adequately capture trend.

Single Difference:	Difference in Difference:
Effect = T2 – T1	Effect = (T2 – T1) – (C2 – C1)
	trend
Effect = T2 – C2	

Two “weak” counterfactuals



These are **only** valid (high internal validity) when certain conditions are met.

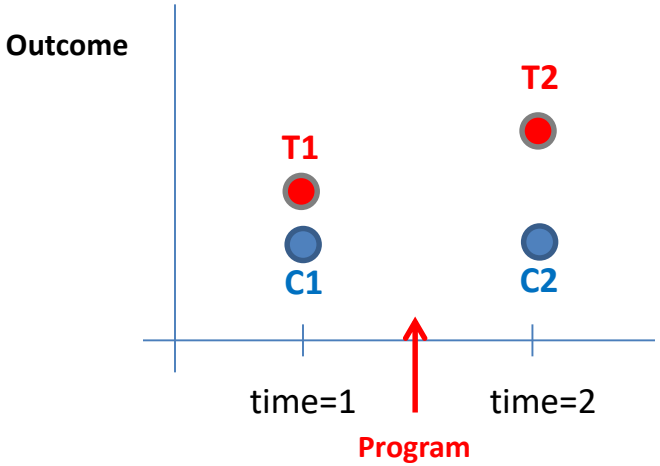
Validity of two "weak" counterfactuals:

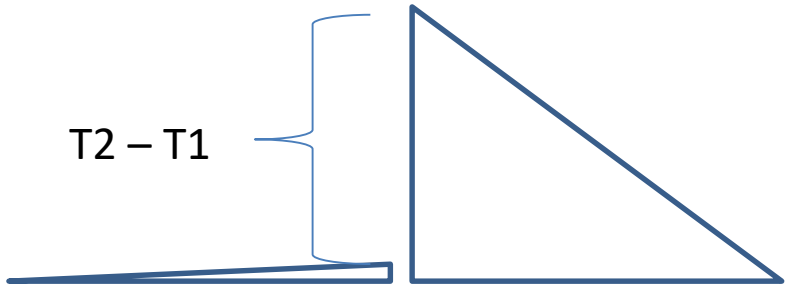
Reflexive (Pre-Post)

$(T2 - T1) - (C2 - C1) = T2 - T1$

IFF

$C2 - C1 = 0$
(no trend)





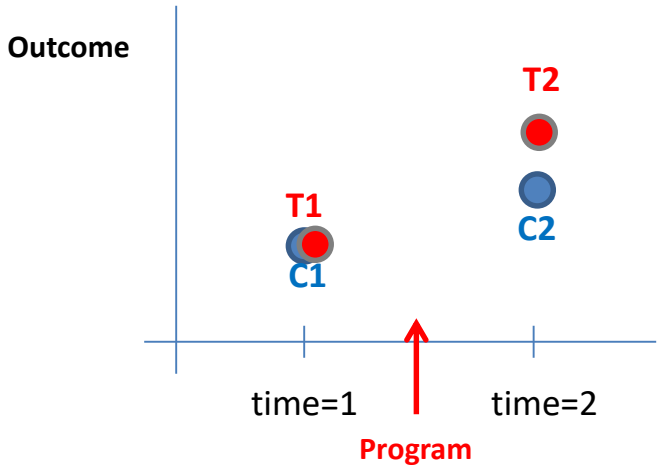
Validity of two "weak" counterfactuals:

Post-Only

$(T2 - T1) - (C2 - C1) = T2 - C2$

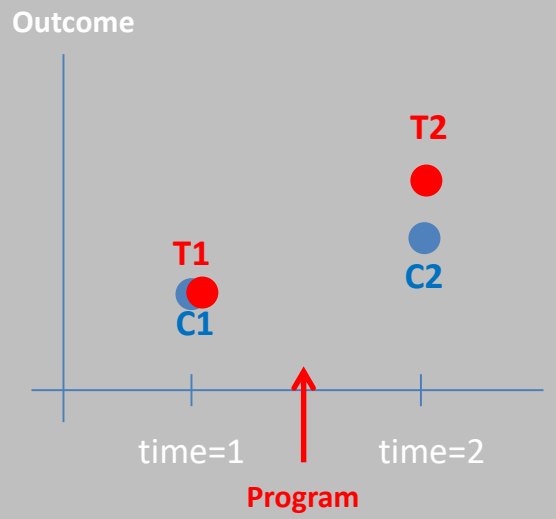
IFF

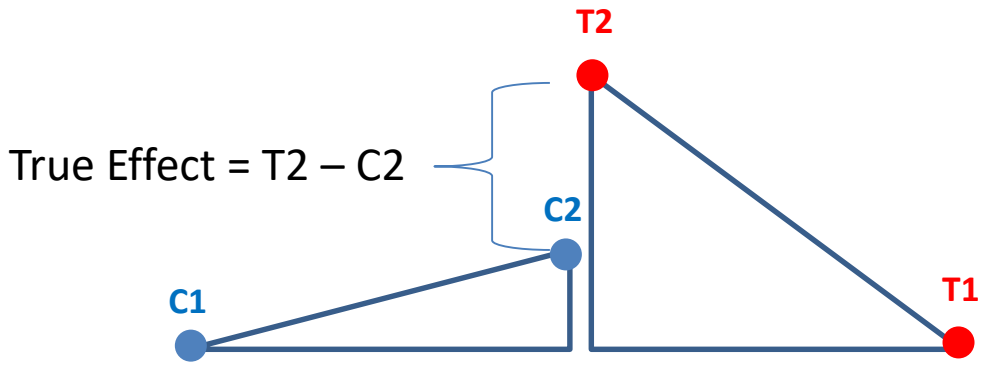
$C1 - T1 = 0$
(equivalent at time 1)



Randomization gives us this. Therefore we can use post-test only estimators in experiments.

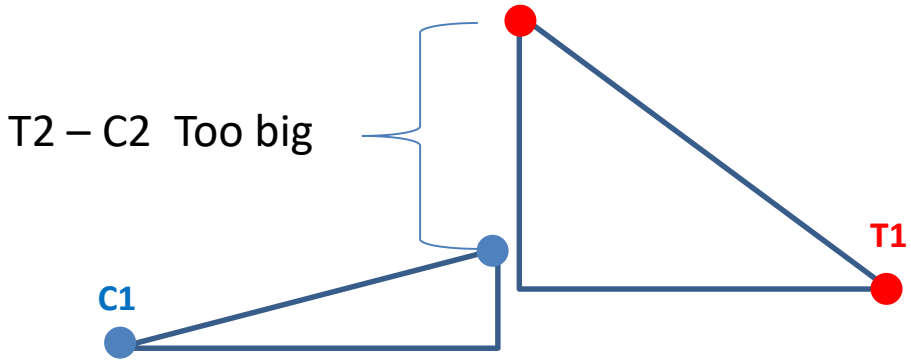
Post-Test Only Measure





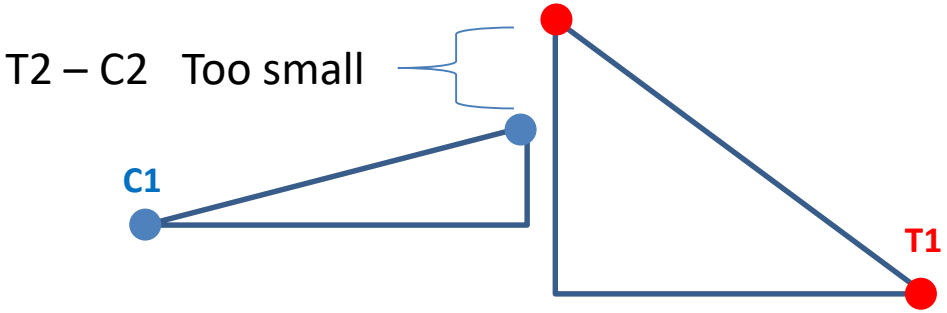
If C1 = T1

Measured effect accurately represents program impact



C1 < T1

Measured effect overstates program impact



C1>T1

Measured effect understates program impact

COUNTERFACTUALS IN PRACTICE: QUASI-EXPERIMENT METHODS

1) **Difference-in-Difference: Effect = (T2-T1) – (C2-C1)**

    1) Difference-in-Difference Analysis
    2) Fixed Effects Models

2) **Reflexive Design: Effect = T2 – T1**

    1) Pre-Post Design with only a Treatment Group
    2) Time Series Regression

3) **Post-Test Only Design: Effect = T2 – C2**

    1) True Experiments (T2-C2)
    2) Post-Test Only Comparison, No Baseline (T2-C2)
    3) Propensity Score Matching
    4) Regression Discontinuity Design

CASE STUDY FROM READINGS

Another Example:

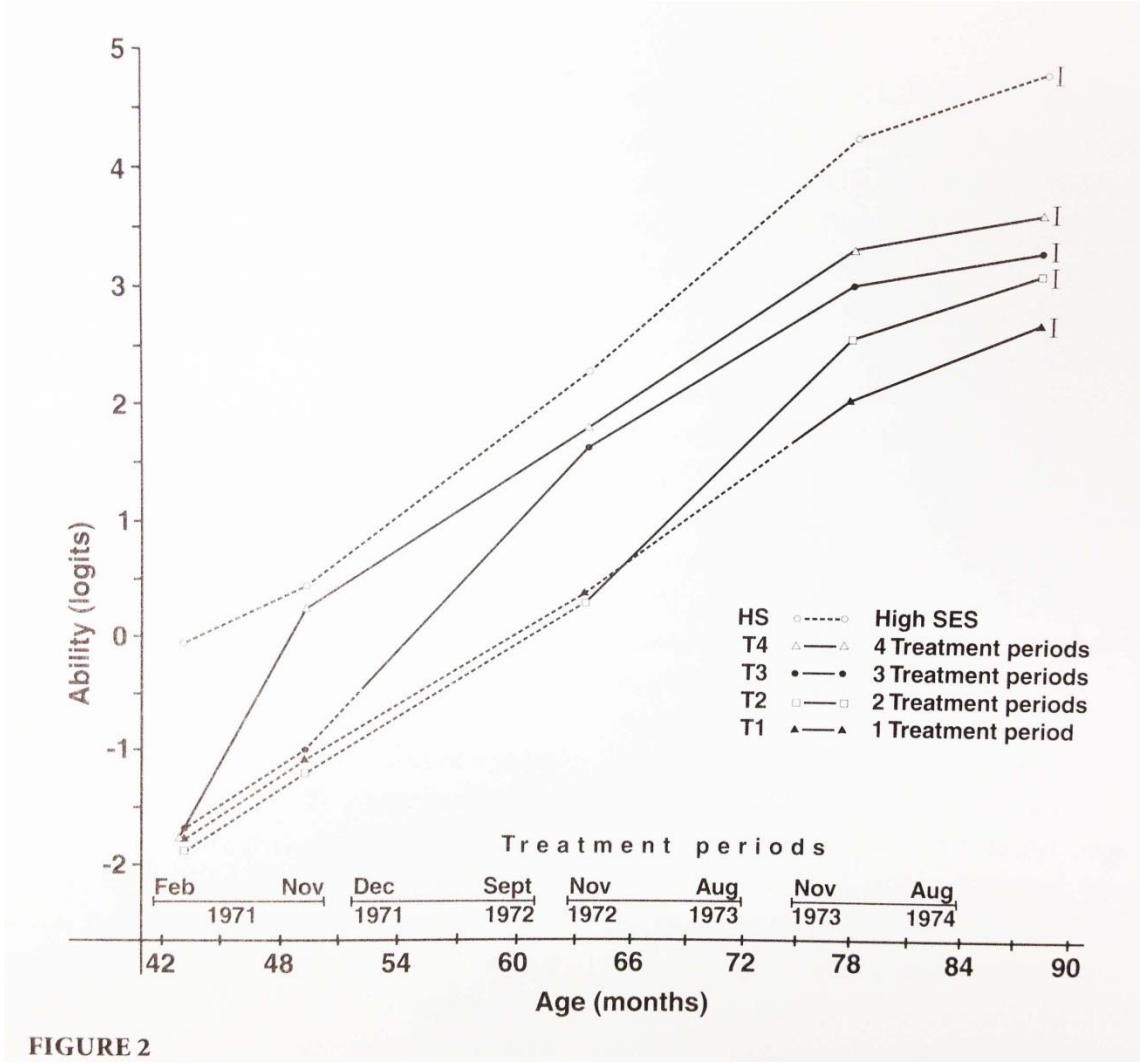
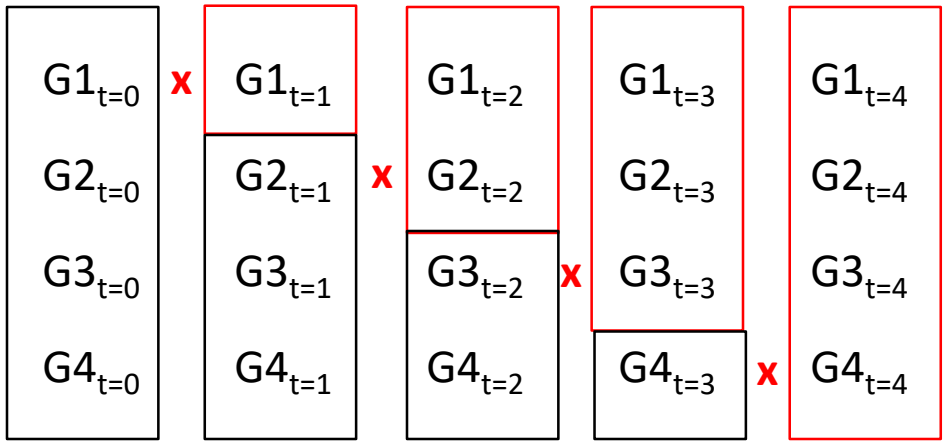


FIGURE 2

Treatment Groups



Control Groups

Treatment Group

Control Group

$G1_{t=0}$ x	$G1_{t=1}$	$G1_{t=2}$	$G1_{t=3}$	$G1_{t=4}$
$G2_{t=0}$	$G2_{t=1}$ x	$G2_{t=2}$	$G2_{t=3}$	$G2_{t=4}$
$G3_{t=0}$	$G3_{t=1}$	$G3_{t=2}$ x	$G3_{t=3}$	$G3_{t=4}$
$G4_{t=0}$	$G4_{t=1}$	$G4_{t=2}$	$G4_{t=3}$ x	$G4_{t=4}$

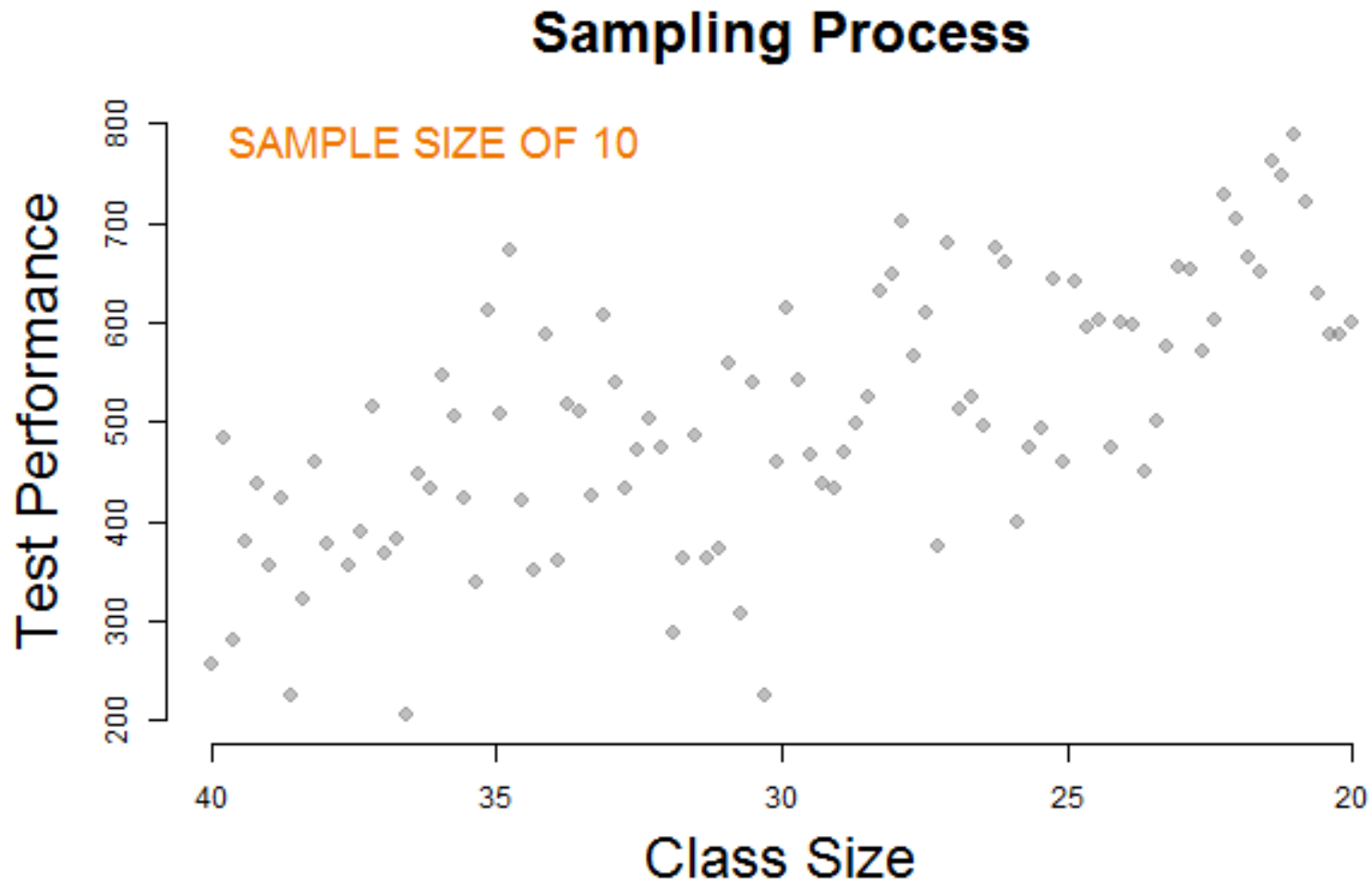
Specific tests: treatment gains for late treatment?

Discussion Questions:

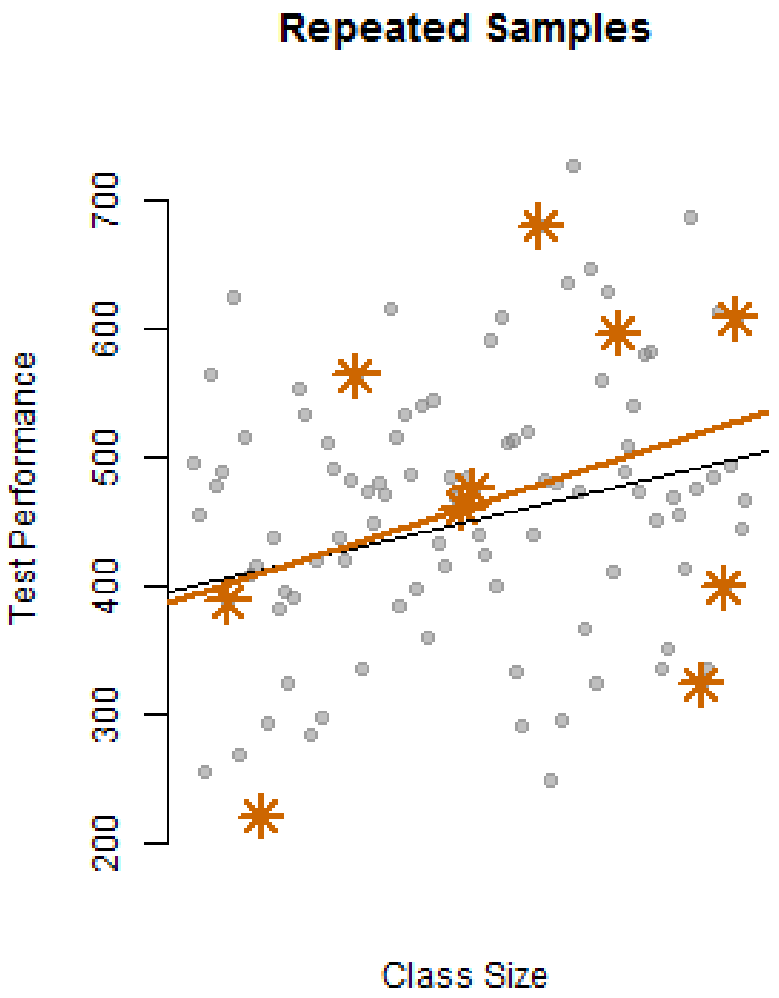
1. Is this an RCT? Do we have an identical “control group”?
2. What role does the high SES group perform?
3. Why do we have four treatment groups?
4. What outcome is measured here? Is it valid and reliable?
5. How would I test whether two treatment periods has the same impact as three periods, but is more cost-effective?
6. Can you identify a weakness in the design or a threat to validity?

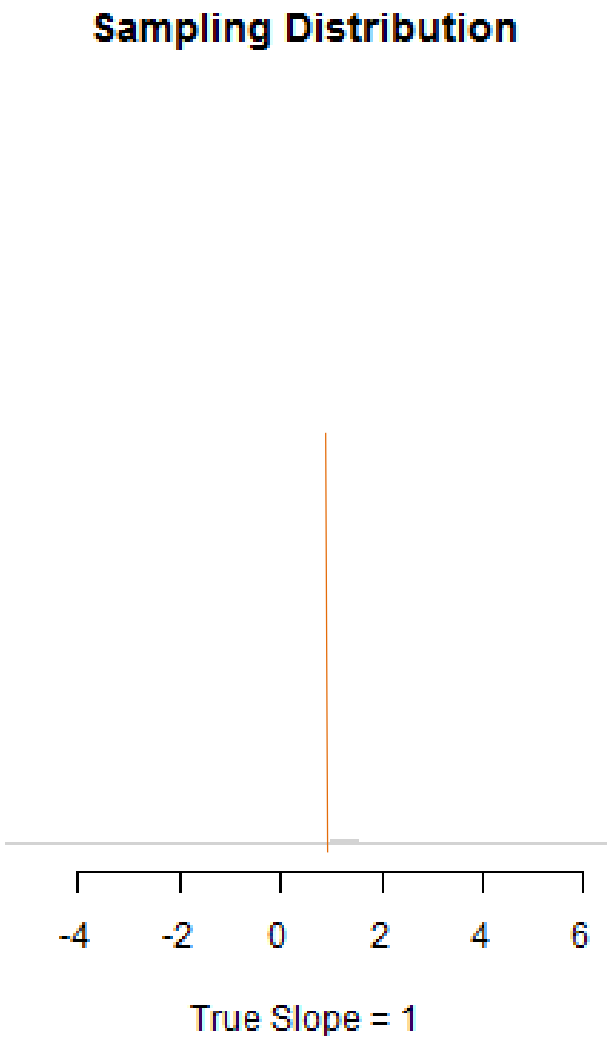
STATS

SAMPLING DISTRIBUTION


Sampling Process
SAMPLE SIZE OF 10
Test Performance
200
300
400
500
600
700
800
40
35
30
25
20
Class Size

SAMPLING DISTRIBUTION

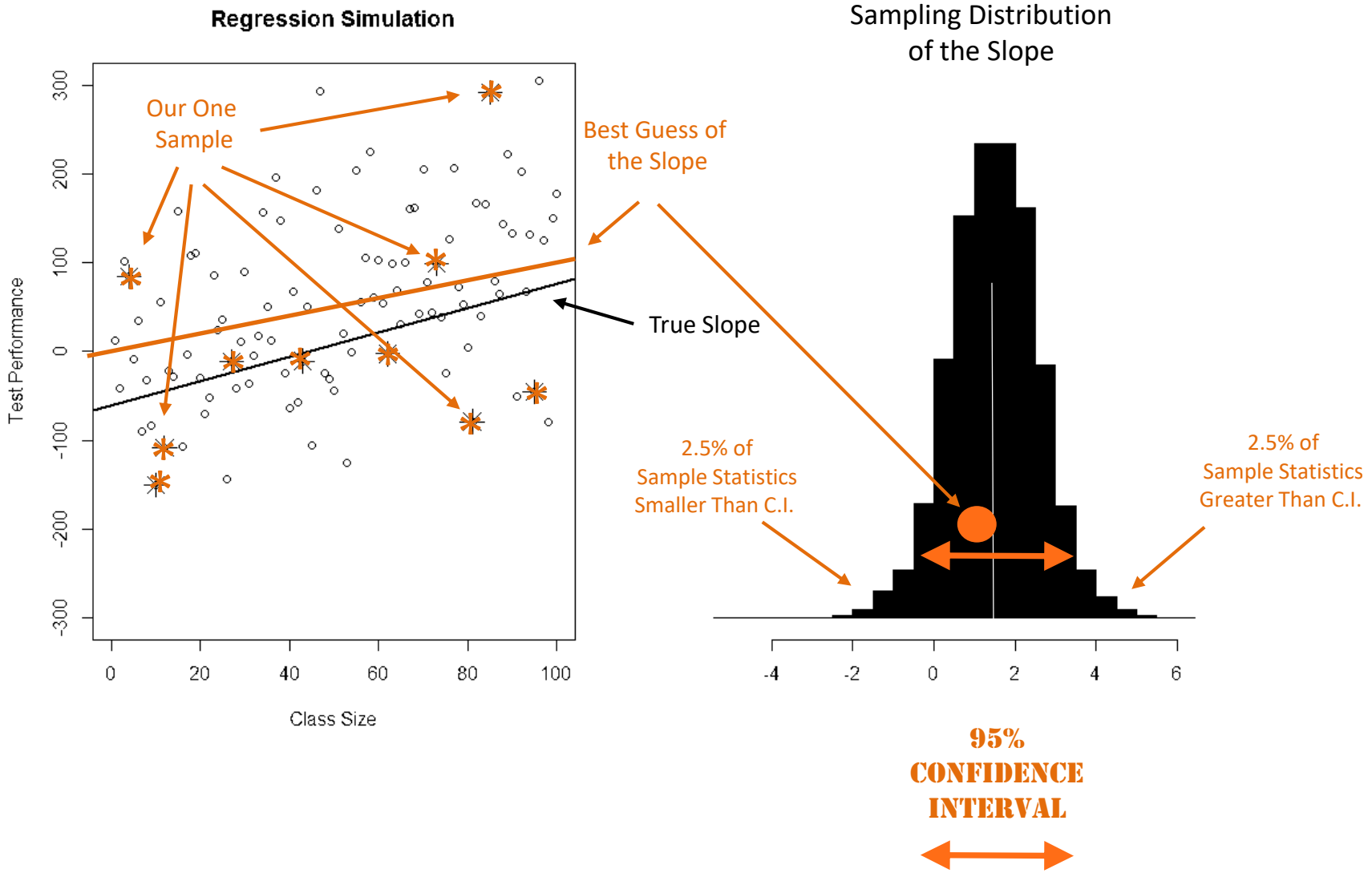


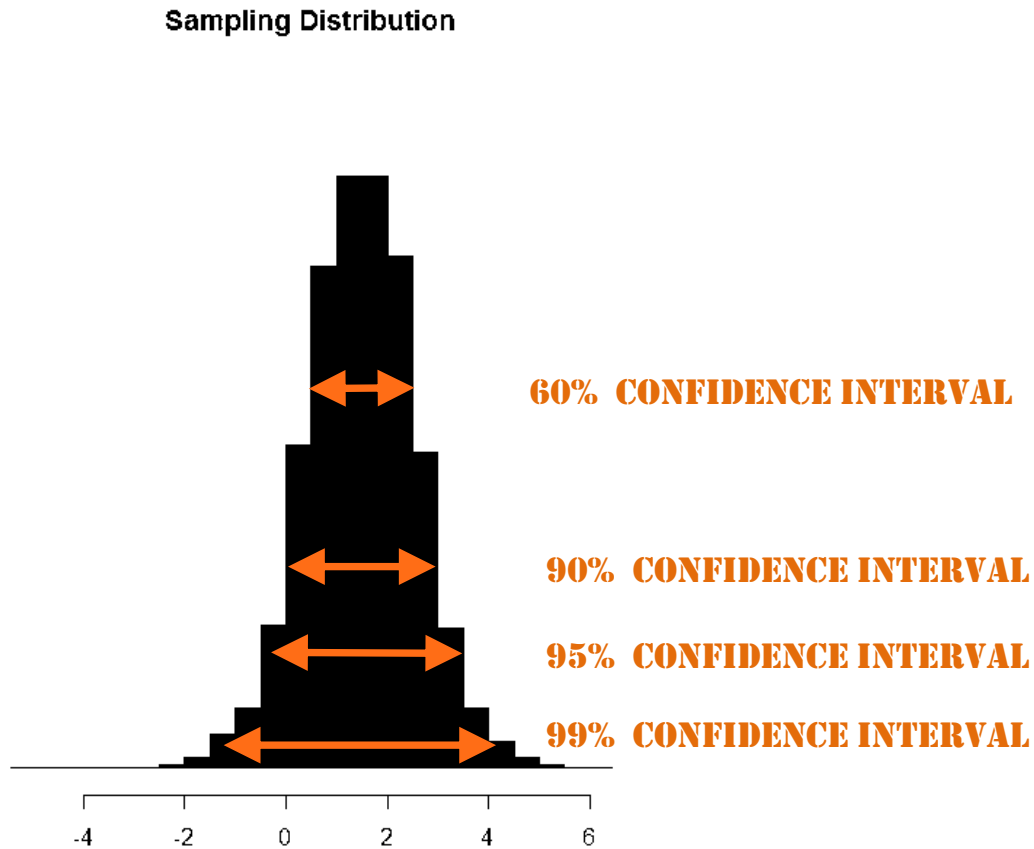


SAMPLE SIZE = 10

CONFIDENCE INTERVALS

95% CONFIDENCE INTERVAL





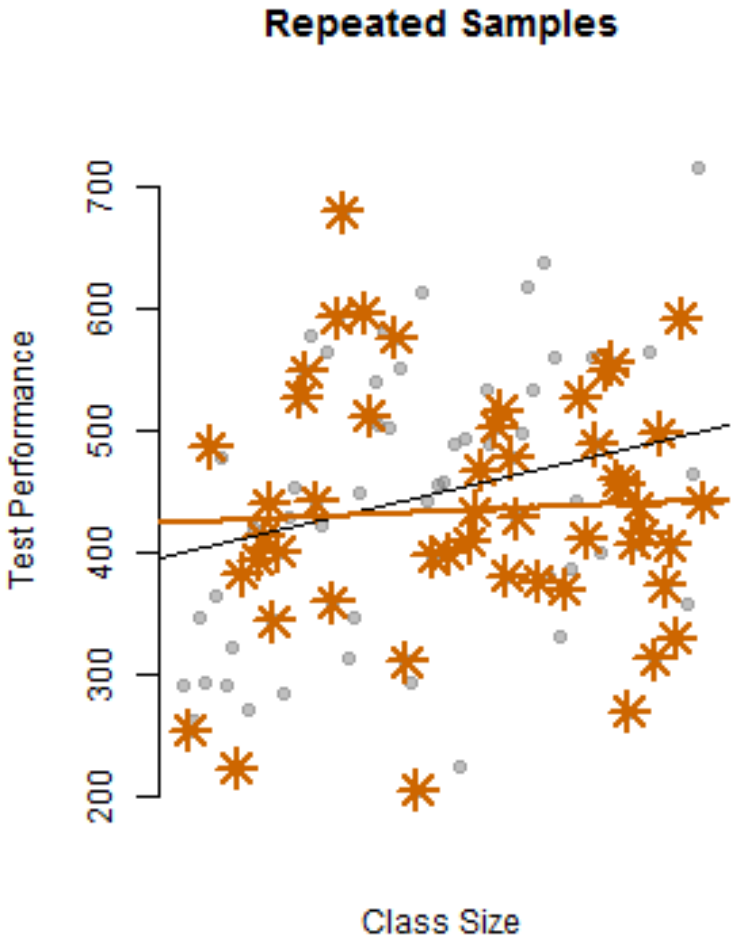
60% CONFIDENCE INTERVAL

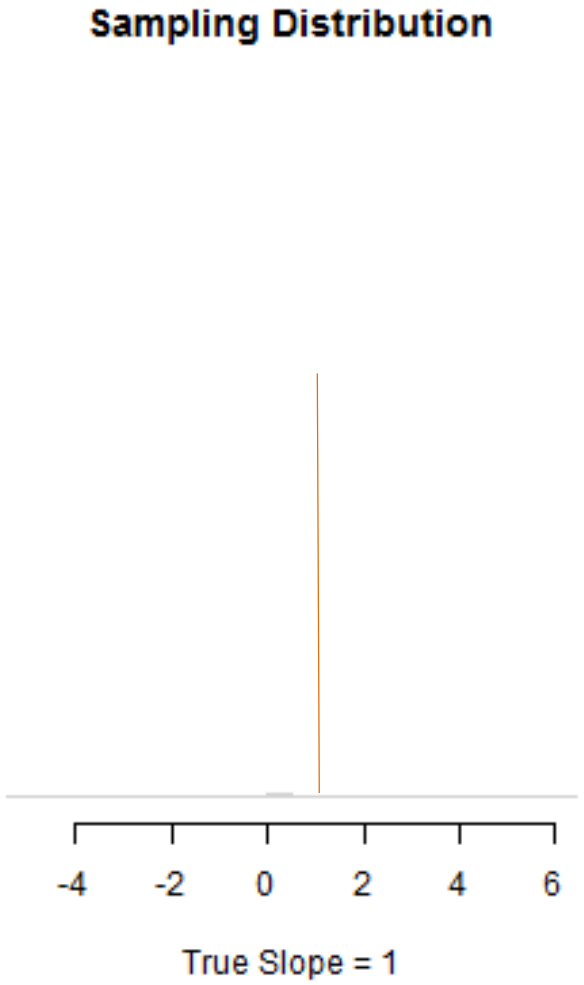
90% CONFIDENCE INTERVAL

95% CONFIDENCE INTERVAL

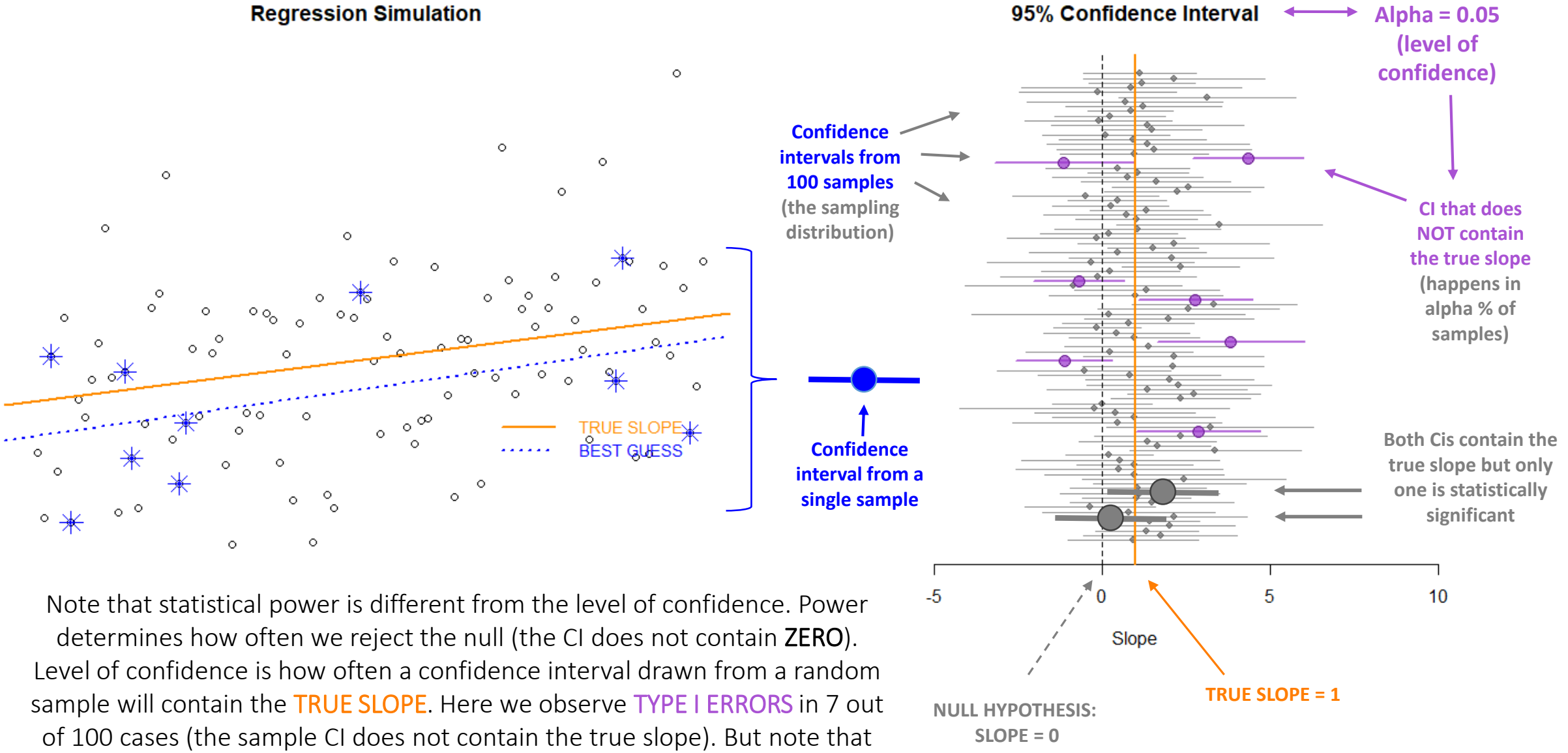
99% CONFIDENCE INTERVAL

SAMPLING DISTRIBUTION





SAMPLE SIZE = 50



Note that statistical power is different from the level of confidence. Power determines how often we reject the null (the CI does not contain ZERO). Level of confidence is how often a confidence interval drawn from a random sample will contain the TRUE SLOPE. Here we observe TYPE I ERRORS in 7 out of 100 cases (the sample CI does not contain the true slope). But note that we would fail to reject the null in more than half of the samples (p-values > alpha), even though the true slope is not zero. This means the sample framework has low power (ability to detect actual effects).

Point estimate plus the standard error (size of the CI) determines the p-value. The p-value is roughly the probability that the slope = 0

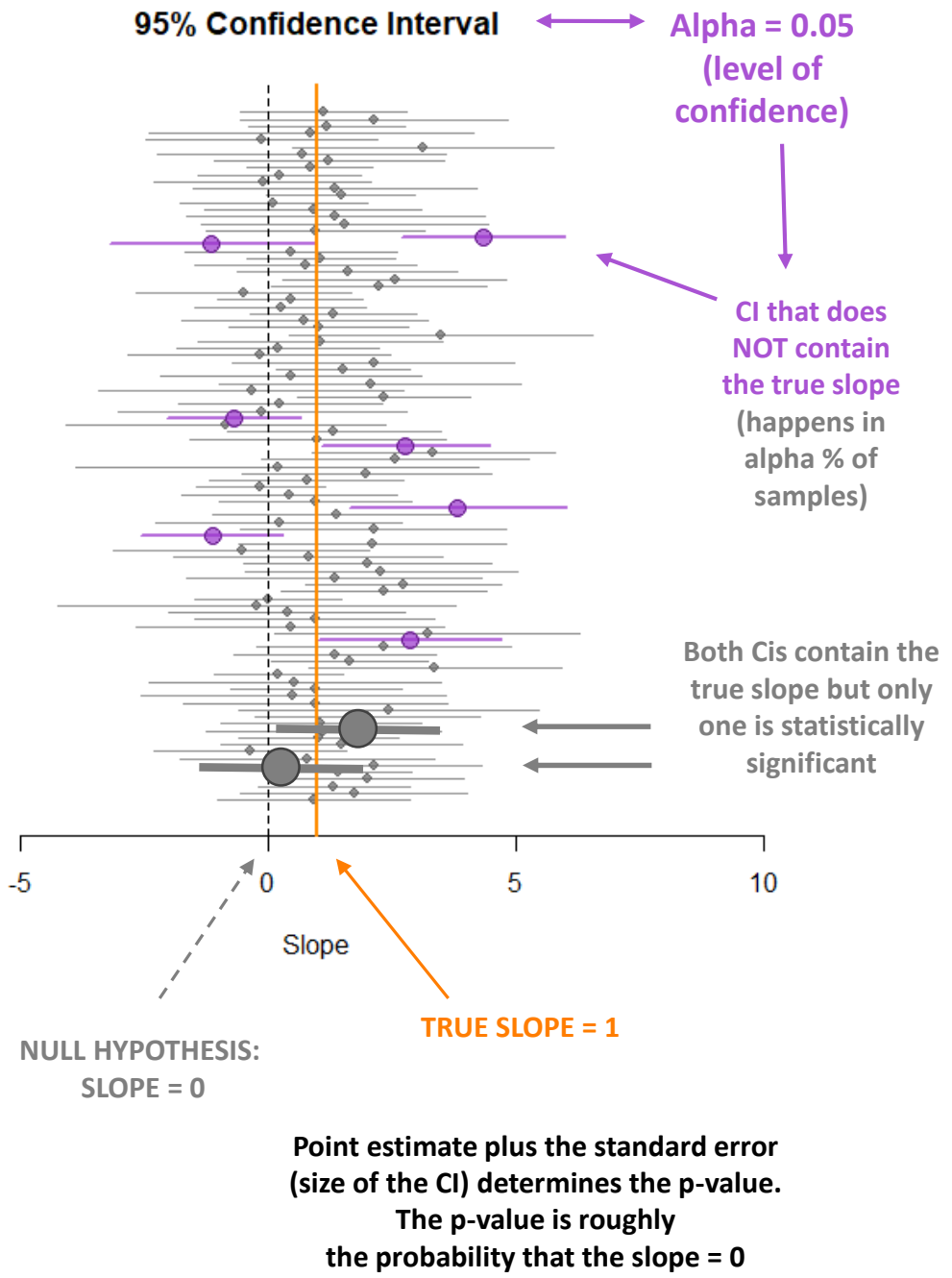
Note that we are using TYPE I ERROR, or a FALSE POSITIVE, in two distinct ways.

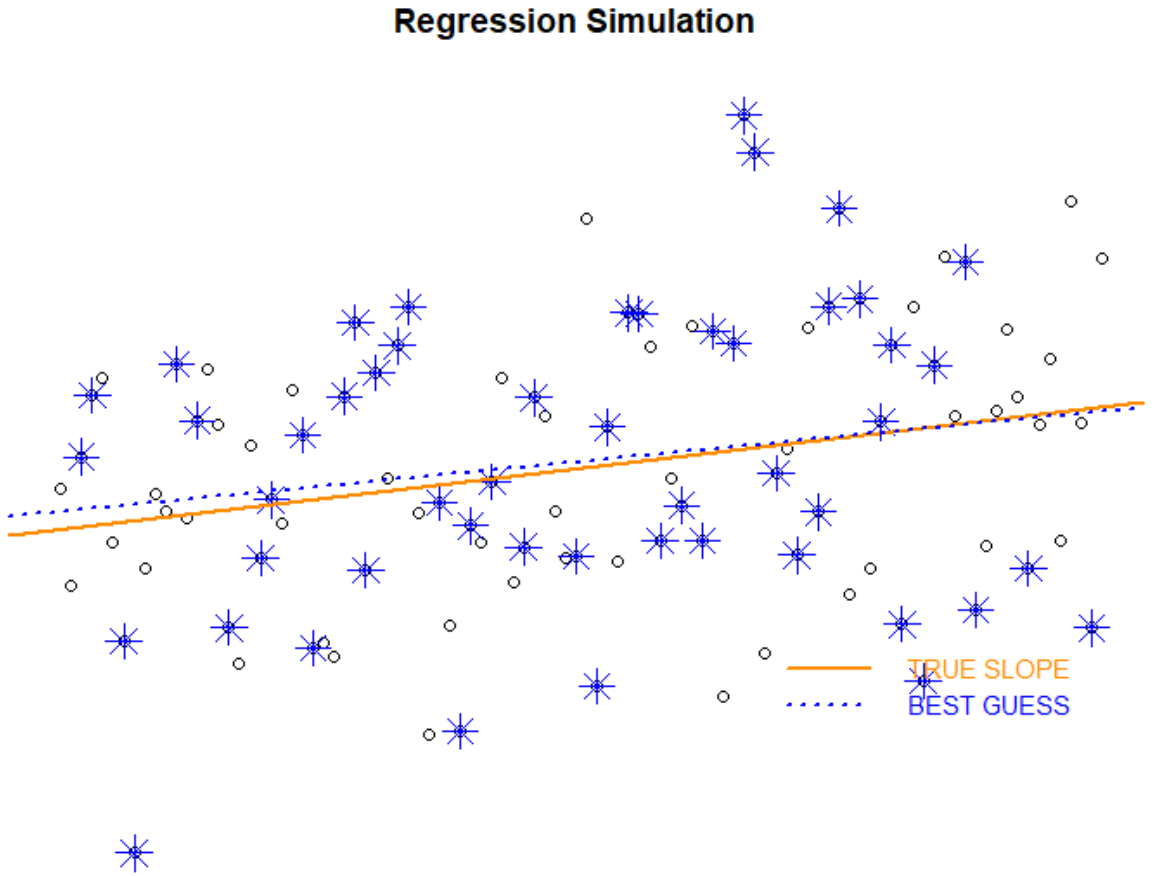
The first type of FALSE POSITIVE is linked to the level of confidence (alpha), which determines how often we will tolerate confidence intervals that fail to contain the true slope. This is the statistical sampling version that uses the **true slope as the NULL**. The purple CIs are all samples that would lead us to conclude that the slope is NOT equal to one (even though it is).

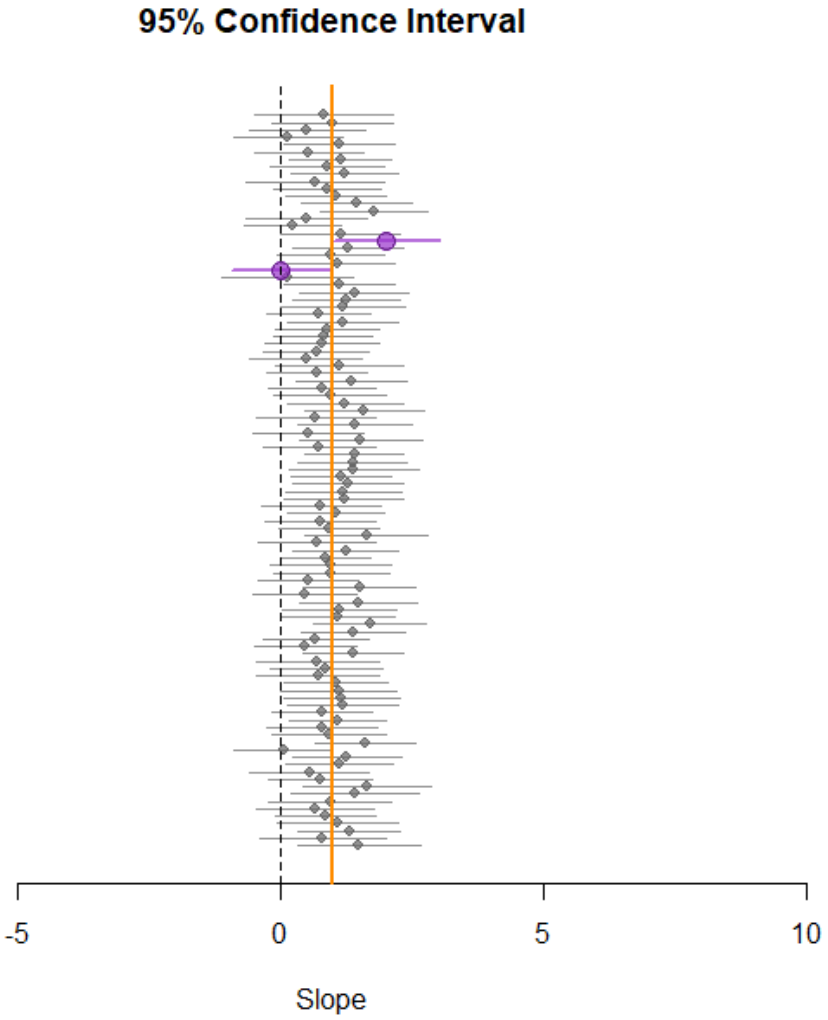
The more important version is the program evaluation context, in which case a TYPE I ERROR or a FALSE POSITIVE means we conclude that a program produced a meaningful change when it did not. This version assumes the **NULL is zero** (no program effect). The p-value produced by a regression model corresponds to this version. Here, since about half of the Cis contain zero, the p-value would be around 0.50.

Similarly, a **TYPE II ERROR** (a **FALSE NEGATIVE**) is when we can't differentiate the true slope from the NULL, even when they differ.

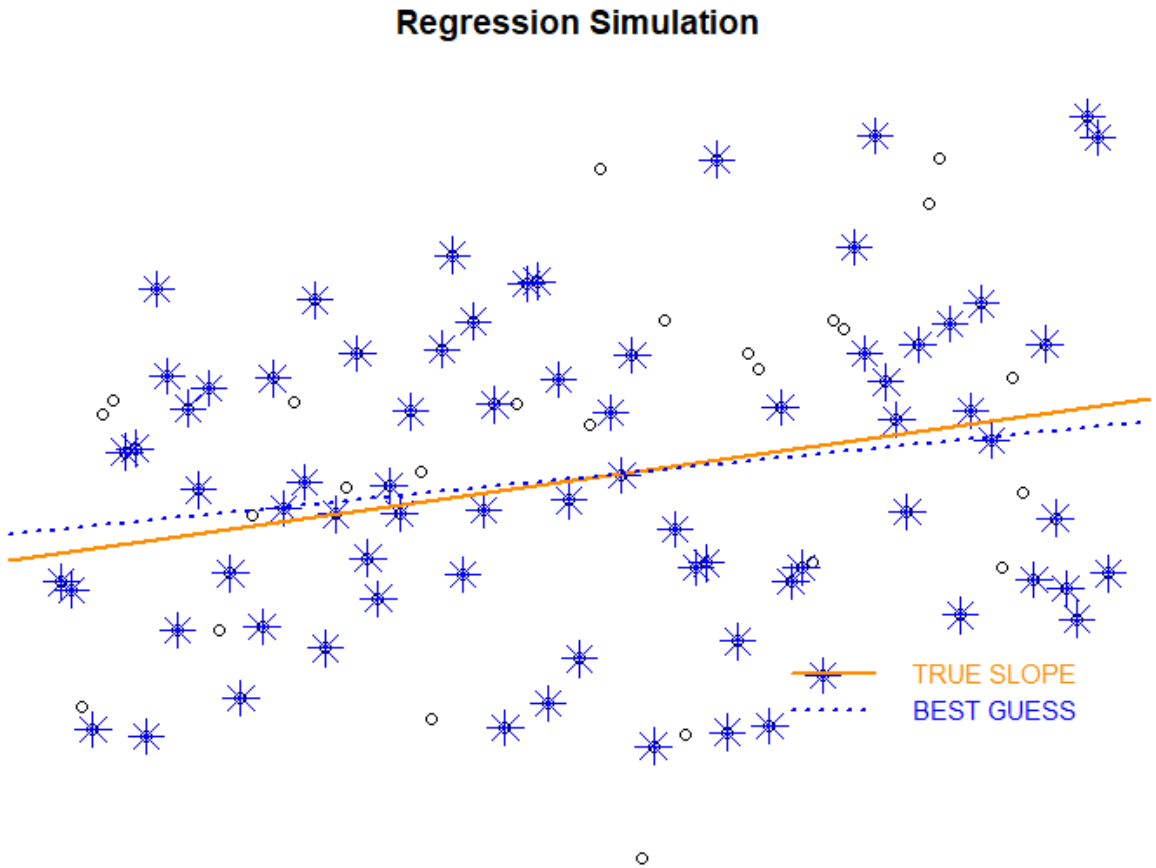
STATISTICAL POWER is the probability of identifying a specific program effect (slope or effect size) using a specific sampling framework. In this case power is very low.

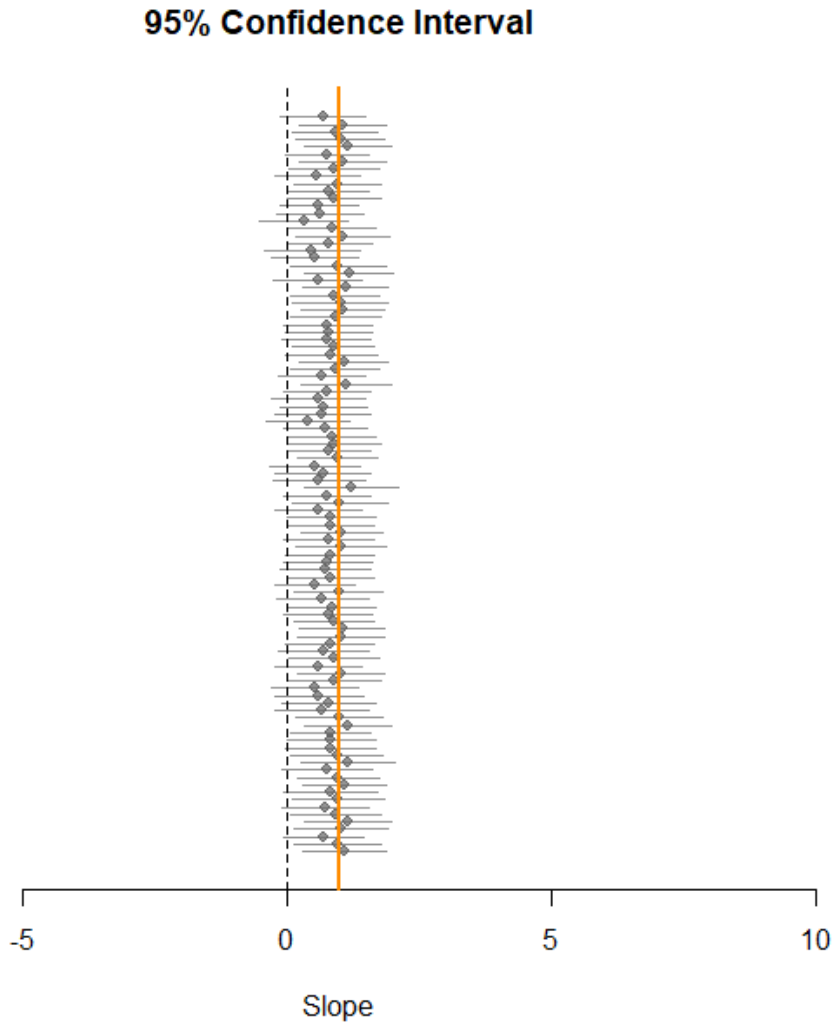


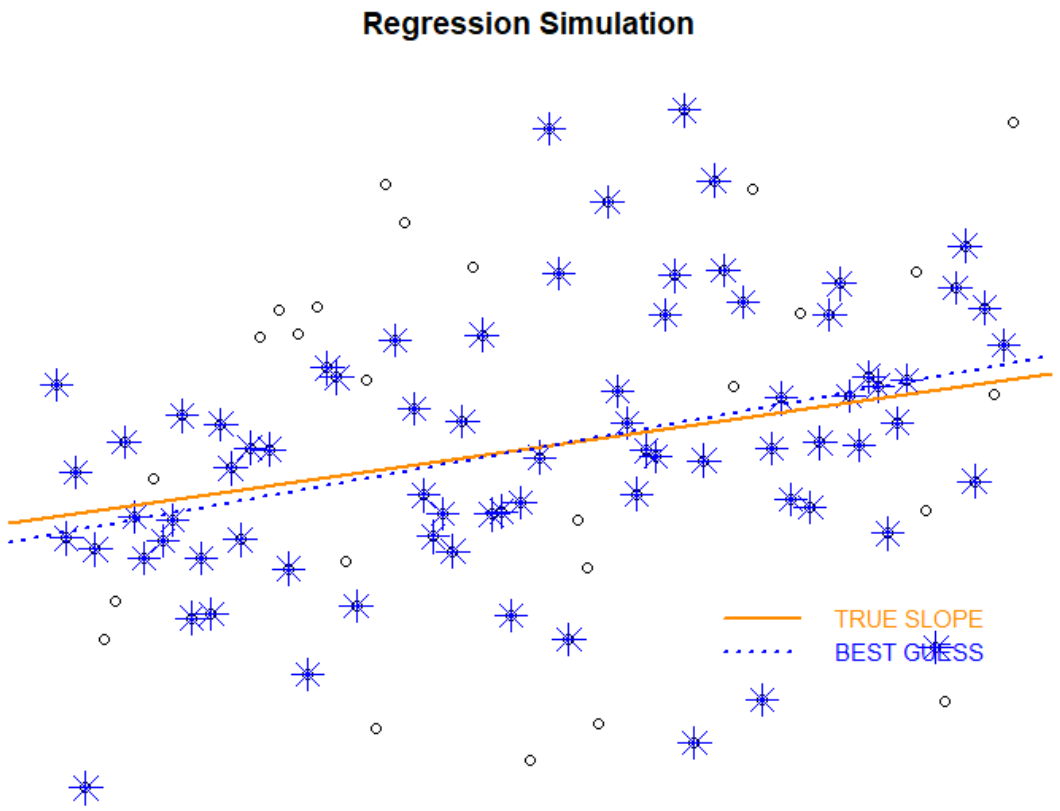


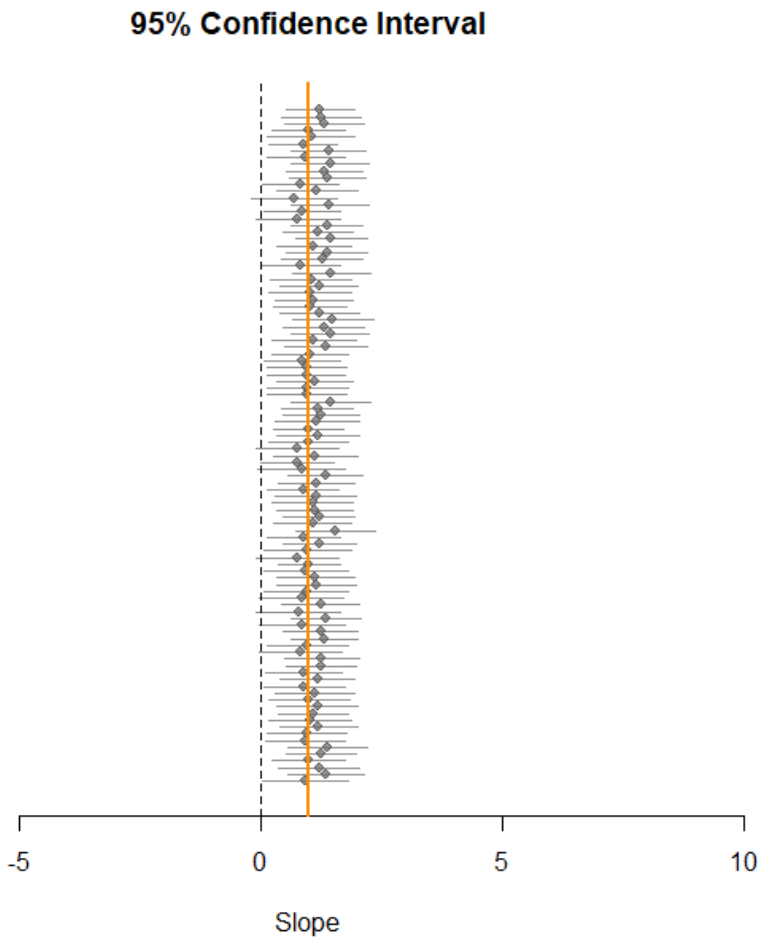


We can increase STATISTICAL POWER by increasing the sample size or adding control variables, thus reducing the standard error and shrinking confidence intervals. Increasing the sample size from 10 to 50 has improved our ability to detect program effects. We are only failing to reject the NULL in 1/3 of cases instead of ½ of cases.

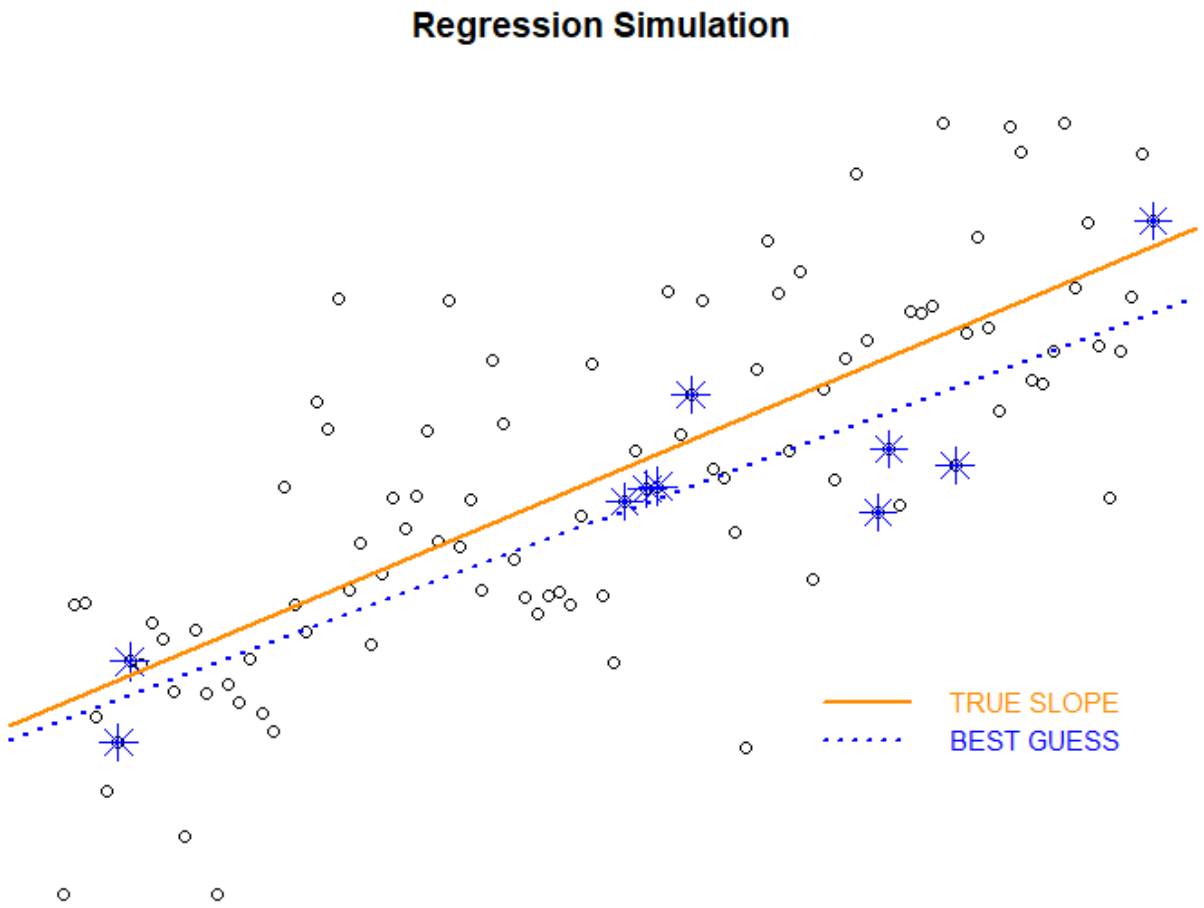

Regression Simulation
TRUE SLOPE
BEST GUESS

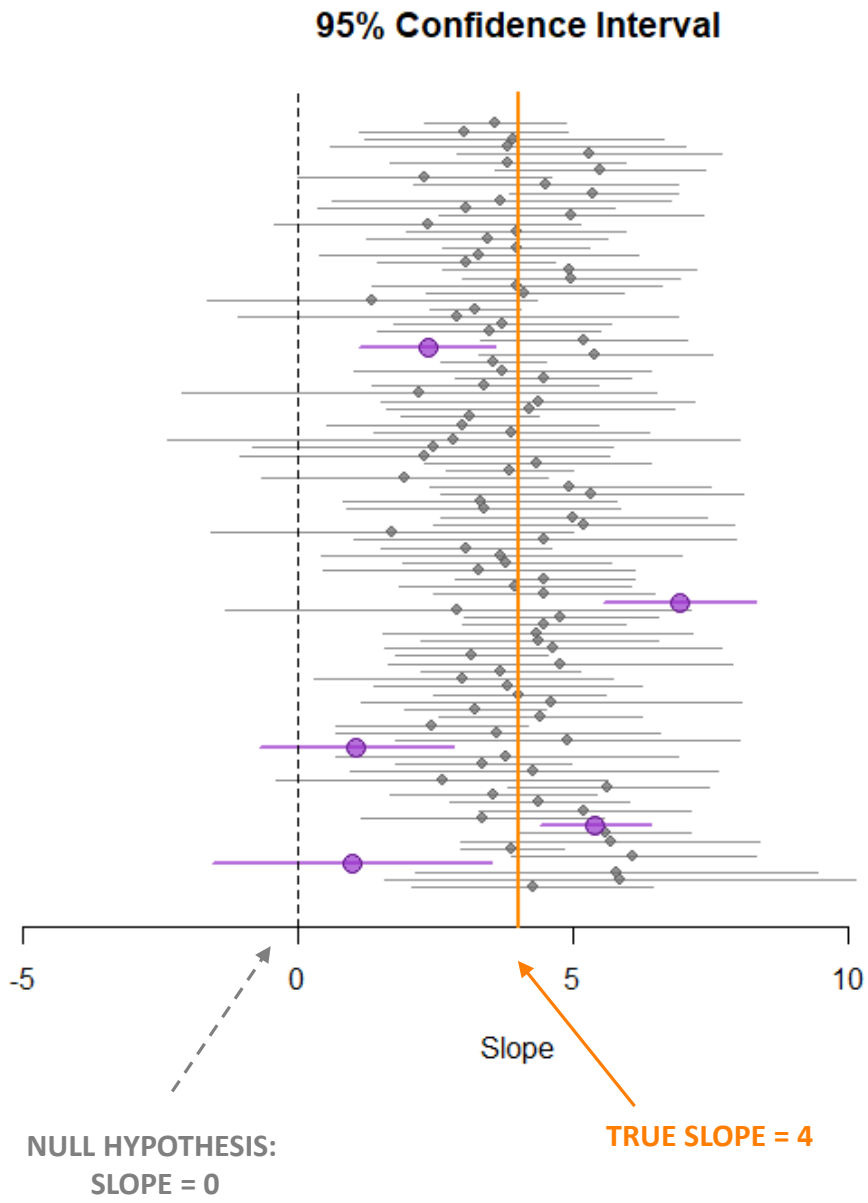

95% Confidence Interval
-5
0
5
10
Slope





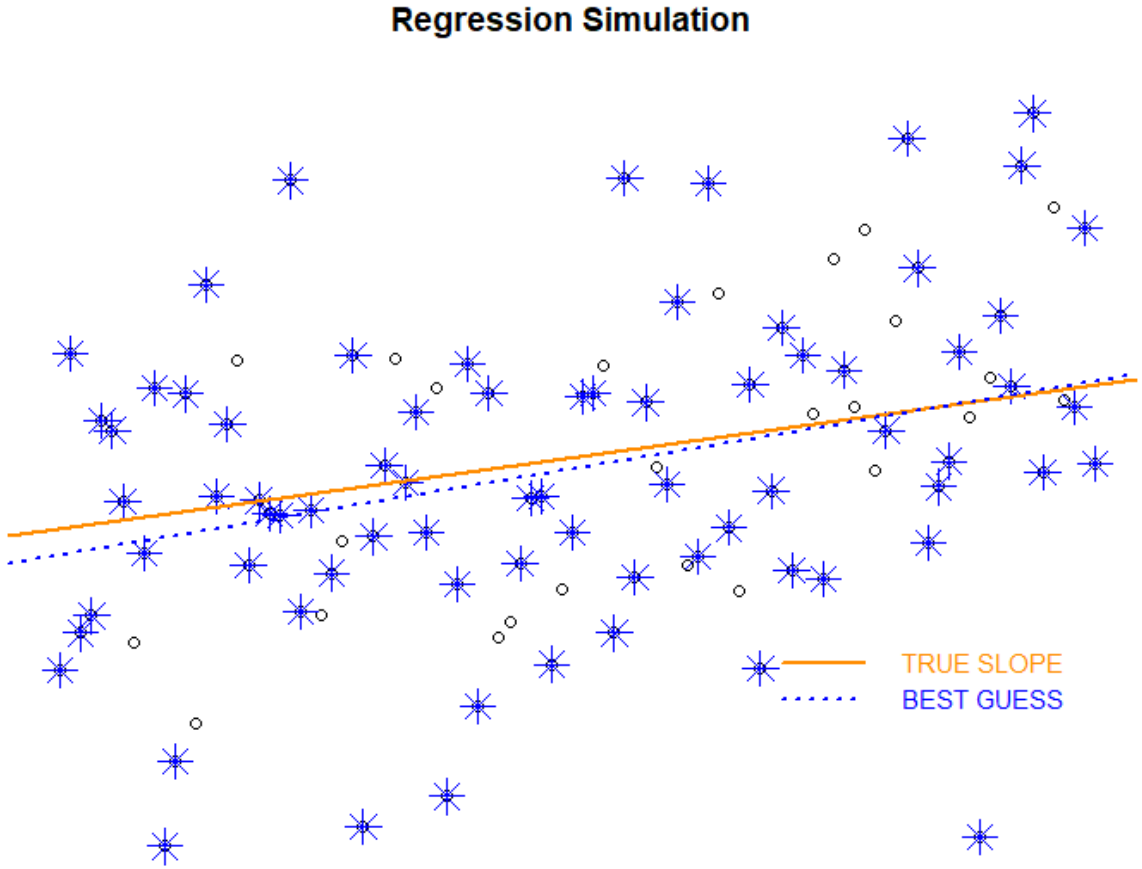
We do even better when the sample size is increased to 75. Now we only fail to identify program effects around 5% to 10% of the time in these case (about 5 or 10 CIs still contain zero).

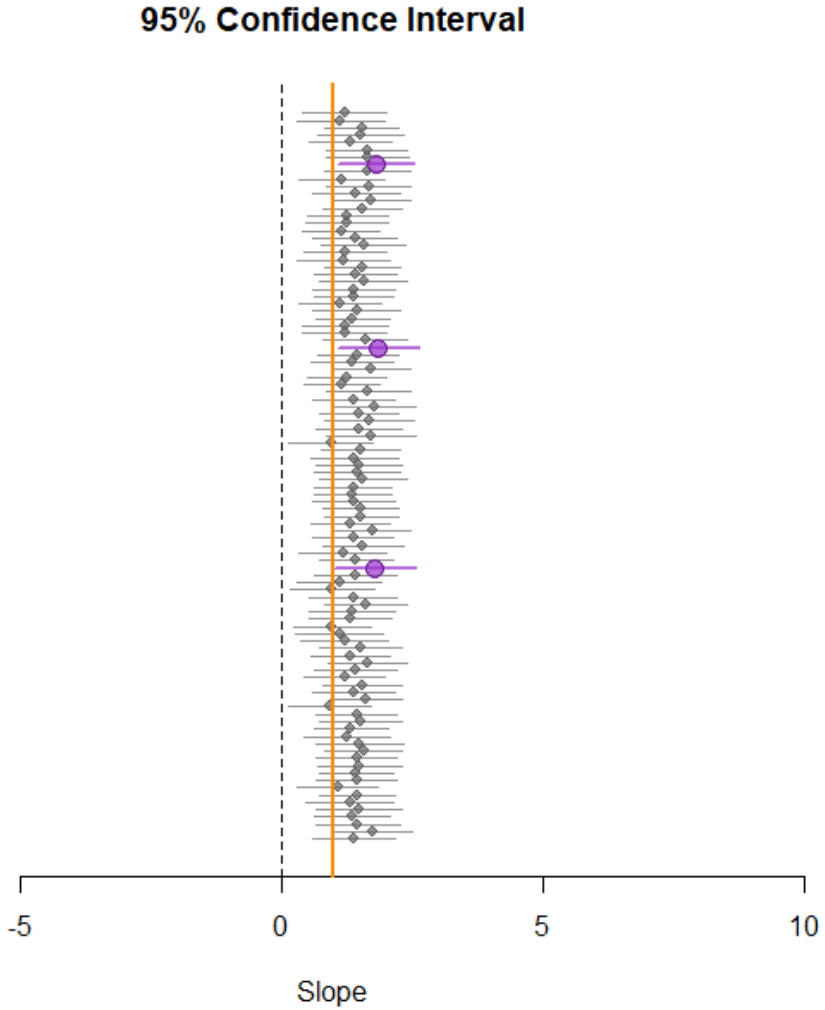




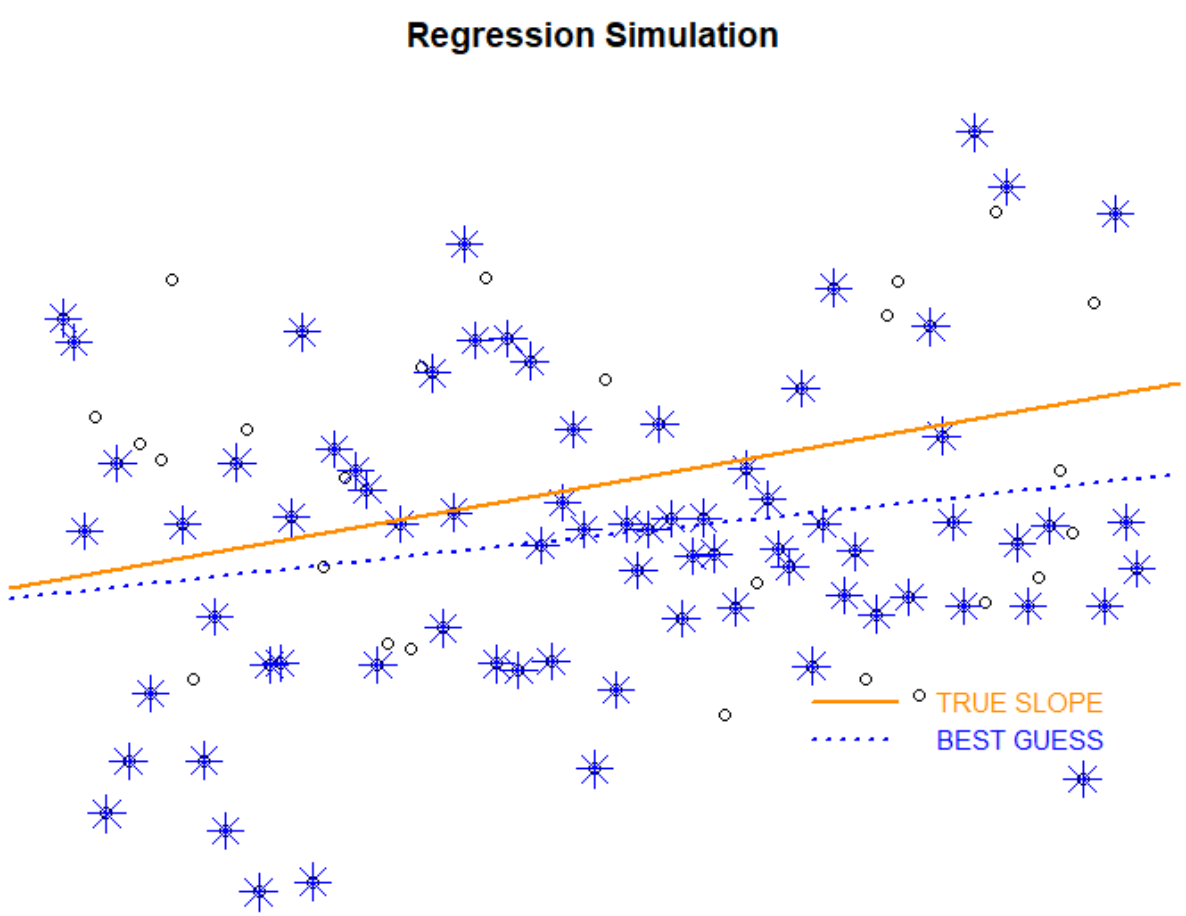
If the sample size remains low (n=10) but the program impact is large (slope of 4 instead of slope of 1) we achieve a lot more statistical power.

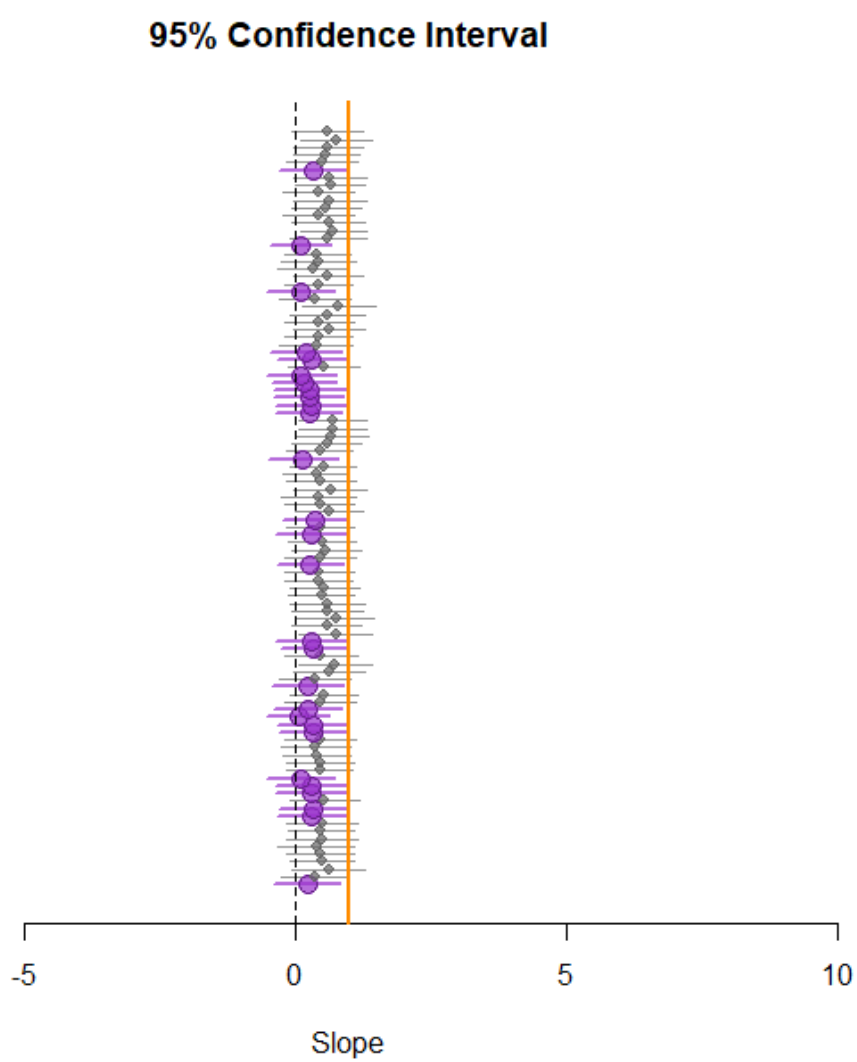
This should be somewhat intuitive - it is easier to detect large changes than small changes.



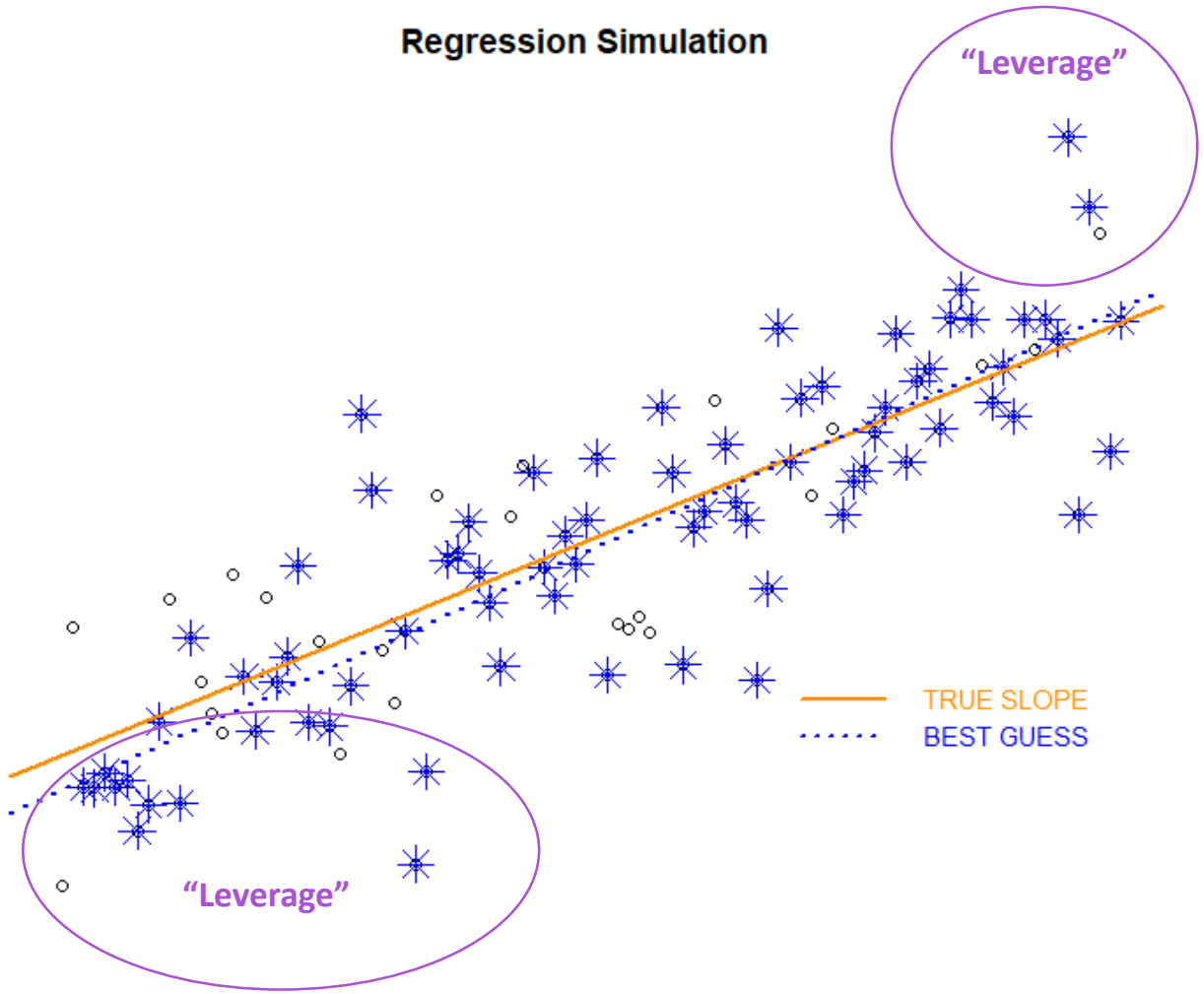


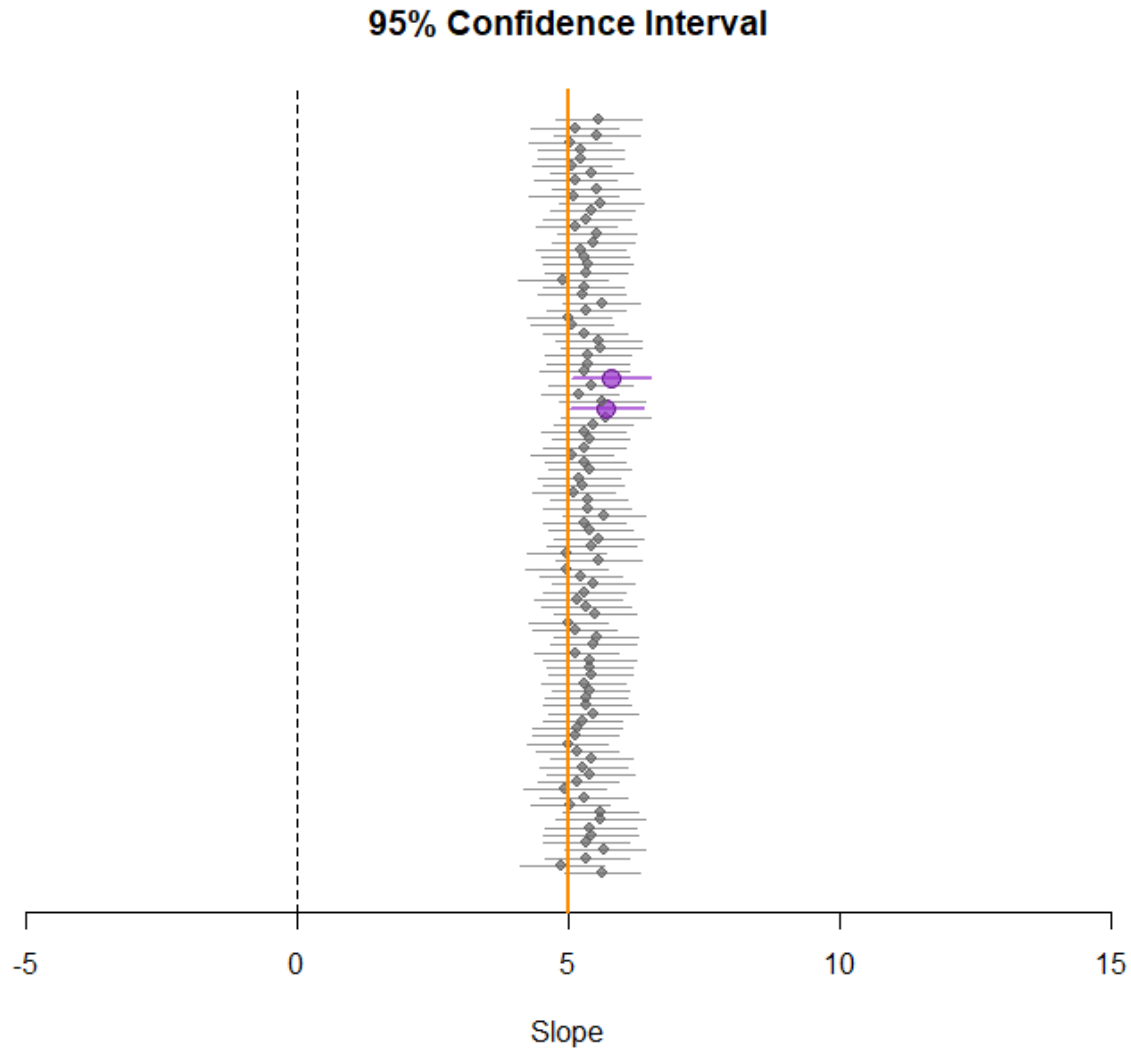
Any idea what is happening here? The sample slope estimates are systematically too large, even though they represent random samples? Hint, it is a type of specification bias.

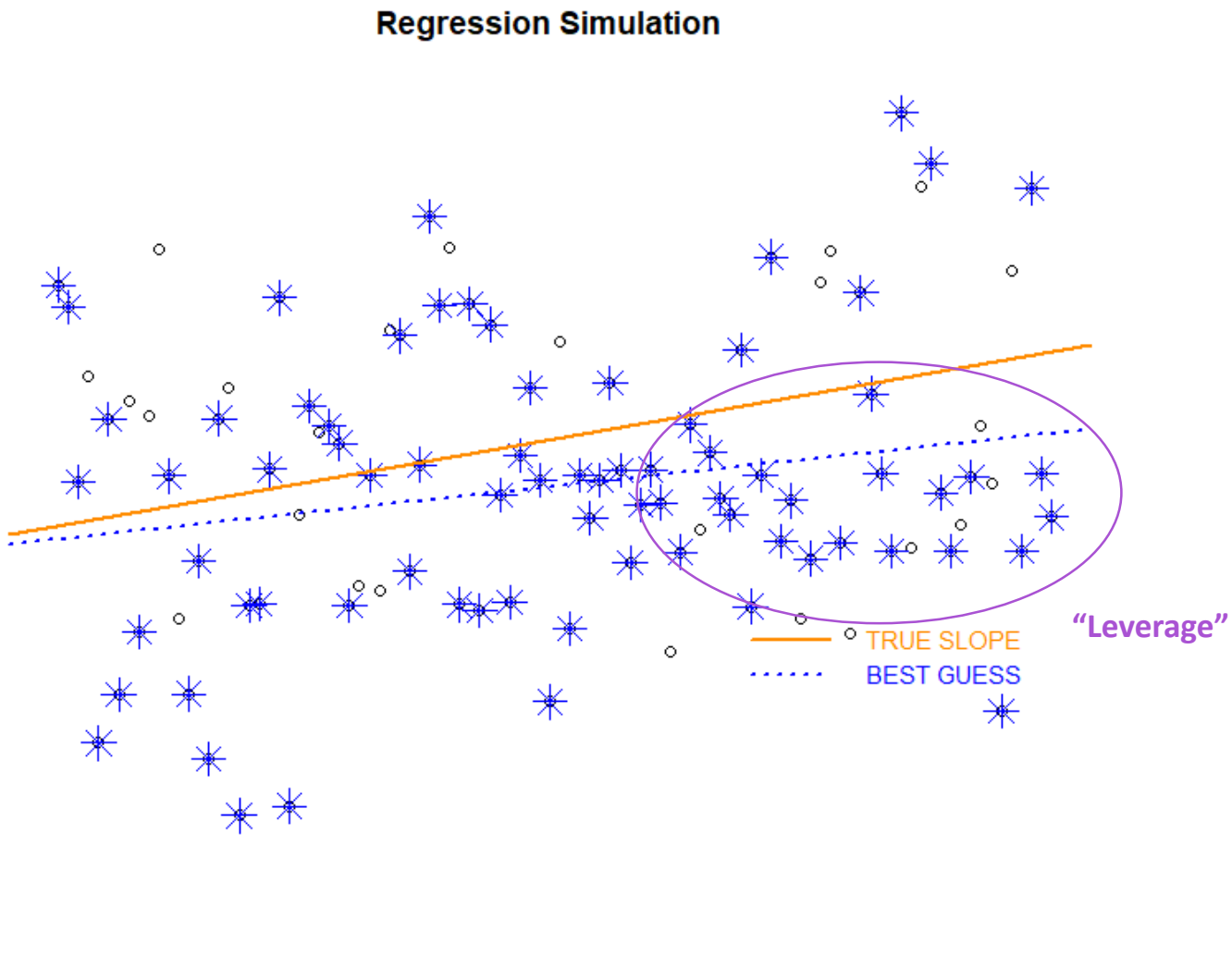


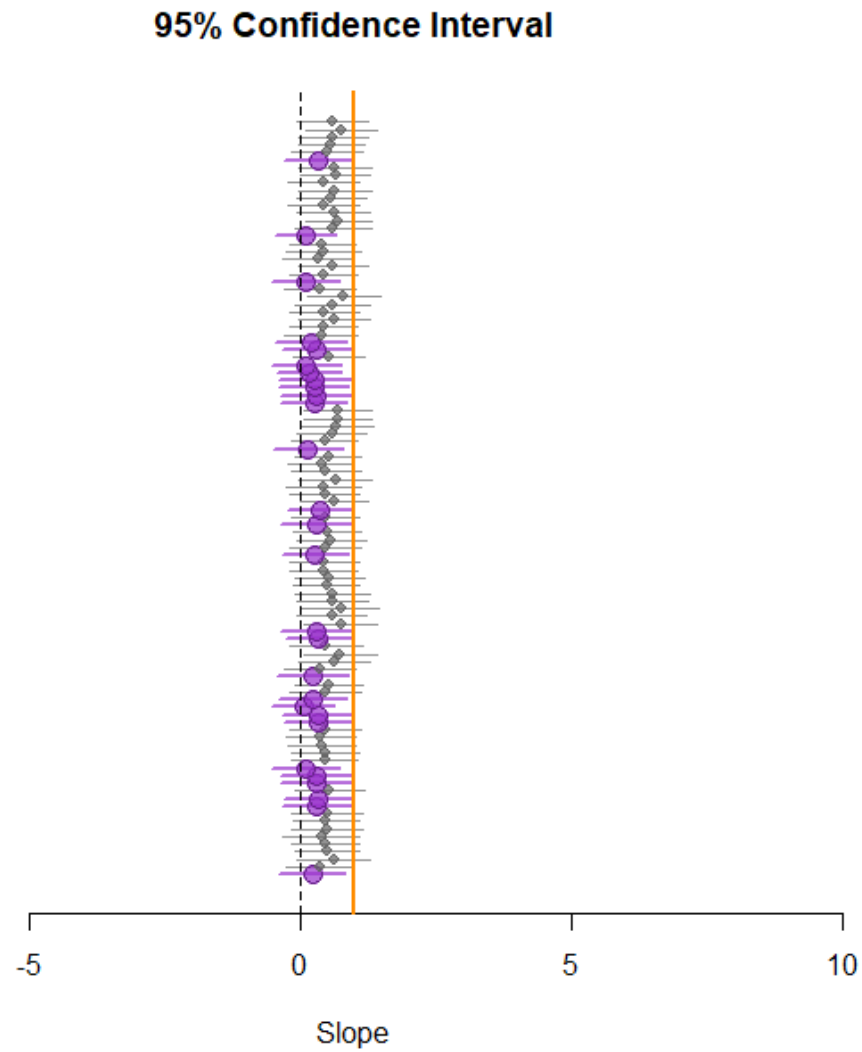


And similarly, here?


Regression Simulation
“Leverage”
“Leverage”
TRUE SLOPE
BEST GUESS


95% Confidence Interval
-5
0
5
10
15
Slope


Regression Simulation
TRUE SLOPE
BEST GUESS
“Leverage”


95% Confidence Interval
-5
0
5
10
Slope

THREATS TO VALIDITY