INTRO TO COUNTERFACTUAL ANALYSIS

Jesse Lecy

Core Concepts

Definition of a counterfactual

Philosophical Definition: NOT P then NOT Q

Statistical Definition: Y(t)-Y(c) = effect

Meaningful null hypotheses

Construction of the counterfactual

True experiments

Quasi-experiments

PHILOSOPHICAL FOUNDATIONS

A counterfactual assertion is a conditional whose antecedent is false and whose consequent describes **how the world would have been if the antecedent had obtained.**

The counterfactual takes the form of a subjunctive conditional: “If P had obtained, then Q would have obtained”. In understanding and assessing such a statement we are asked to consider how the world would have been if the antecedent condition had obtained. For example, “If the wind had not reached 50 miles per hour, the bridge would not have collapsed” or “If the Security Council had acted, the war would have been averted.”

There is a close relationship between counterfactual reasoning and causal reasoning. If we assert that “P caused Q (in the circumstances Ci)”, it is implied that we would assert: **“If P had not occurred (in circumstances Ci) then Q would not have occurred.”** So a causal judgment implies a set of counterfactual judgments.

Lewis, David K. 1973. Counterfactuals. Cambridge: Harvard University Press.

Translated to Statistics

Pr(A | B) means the probability that A occurs given that B has occurred.

We can augment this notation by incorporating the notion of “how the world would have been if the antecedent had obtained” using an intervention or a “treatment”:

```
Pr( Y = TRUE | Treatment = TRUE ) - Pr( Y = TRUE | Treatment = FALSE )
```

In cases where the outcome is continuous, such as income levels or wheat yield per acre, the notation would only be slightly different:

```
[ mean(Y) | Treatment = TRUE ] - [ mean(Y) | Treatment = FALSE ]
```

Or more succinctly:

```
Treatment Effect = Y(t) - Y(c)
```

world with treatment

world without

Translated to Statistics

The outcome is measured now as a difference of means instead of a change in probabilities of observing success.

Thus, we typically care about the **Average Treatment Effects**

because it is the easiest thing to measure (the average outcome for the treatment and control groups) and most succinct way to communicate program effectiveness in evaluation studies.

Translated to Statistics

The outcome is measured now as a difference of means instead of a change in probabilities of observing success.

Thus, we typically care about the **Average Treatment Effects**

because it is the easiest thing to measure (the average outcome for the treatment and control groups) and most succinct way to communicate program effectiveness in evaluation studies.

Note! The part of this statement that is rarely explicit but really important, **the dosage that represents the typical treatment.**

For example, going to the gym results in more muscle mass. What does “going to the gym” mean? How many visits per week, and time spent each visit? Not to mention activities during the visit.

Temperature of world
WITH
Paris Climate Accord
in year = 2050

Temperature of world
WITHOUT
Paris Climate Accord
in year = 2050

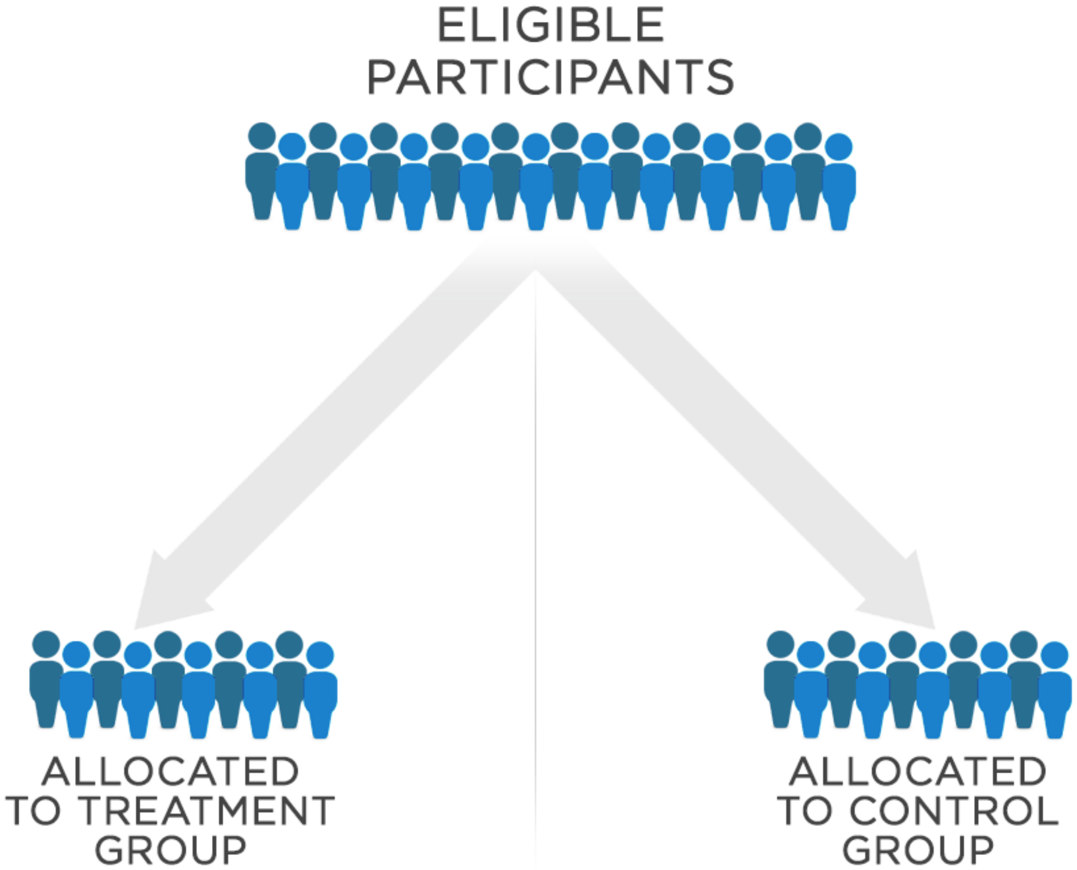
Effect of climate accord = Y(t) – Y(c)

Unfortunately we don't have two worlds for this experiment!

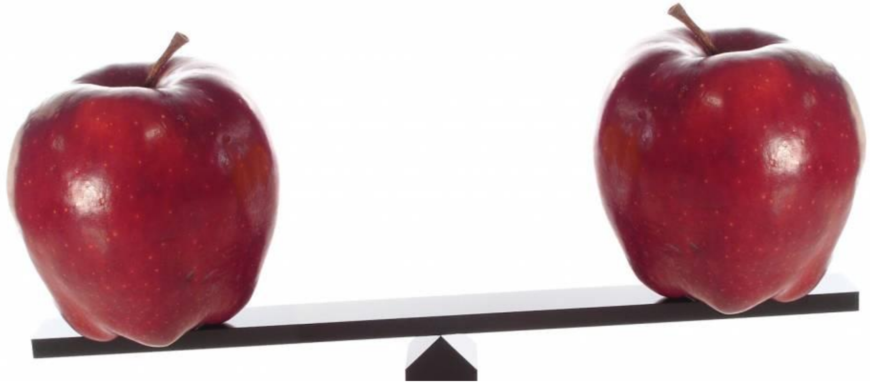
Effect of climate accord = Y(t) – Y(c)

Experimental Design

While we can't turn back time and do it all over with the exact same participants and conditions, we can **create groups that represent states of the world**.



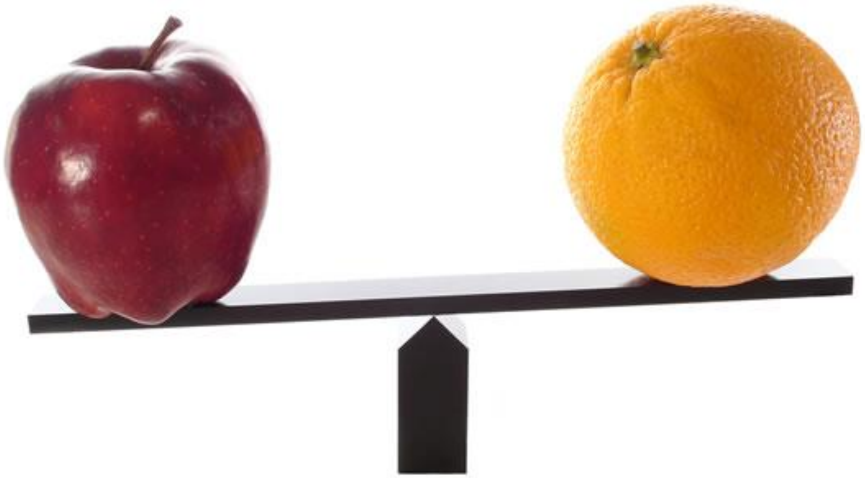
$$\text{Gains from fertilizer} = \text{weight}_{\text{treatment}} - \text{weight}_{\text{control}}$$



apple tree fertilized

apple tree NOT fertilized

Non-experimental studies



tree
fertilized

tree
NOT fertilized

Is the difference due to the treatment (fertilizer), or to differences in the groups?

(confounding factors that disallow causal claims)

SELECTING A MEANINGFUL NULL

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

There have been multiple student suicides over the past year in a specific school district.

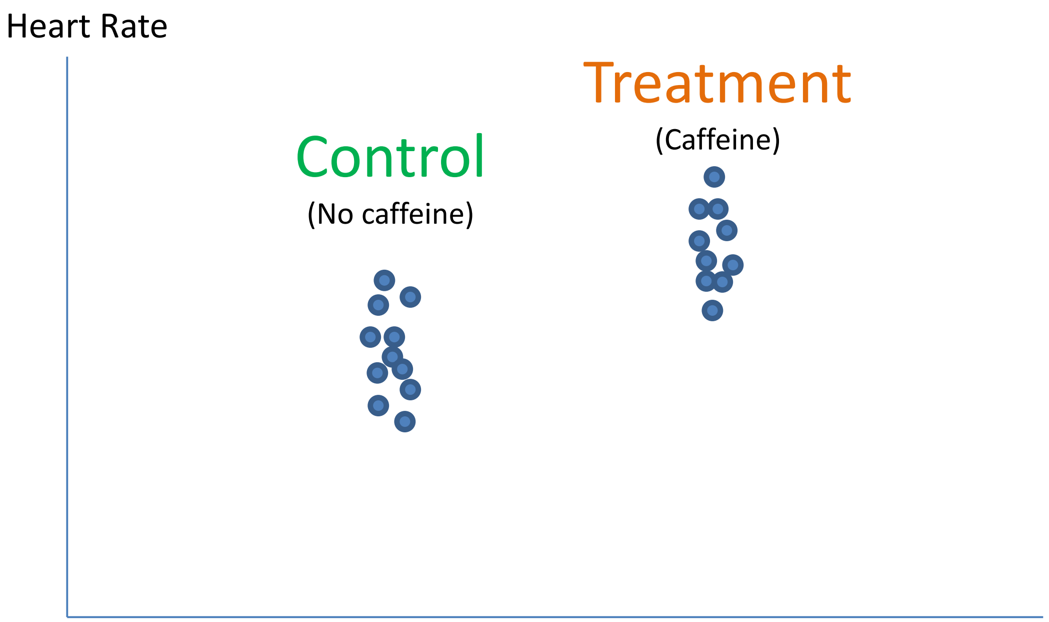
The current district superintendent cut support for school counseling services. Parents are considering filing a lawsuit against the school district because they feel the cuts in spending resulted in loss of support for mental health services, thus leading to increased rates in suicide.

You have been hired as an expert evaluator to build evidence for the case. They would like you to determine whether increases in rates at the school district are notable, and thus potentially linked to the recent cuts in counseling services.

How do you operationalize this research question? Any statistical test requires a **null hypothesis**.

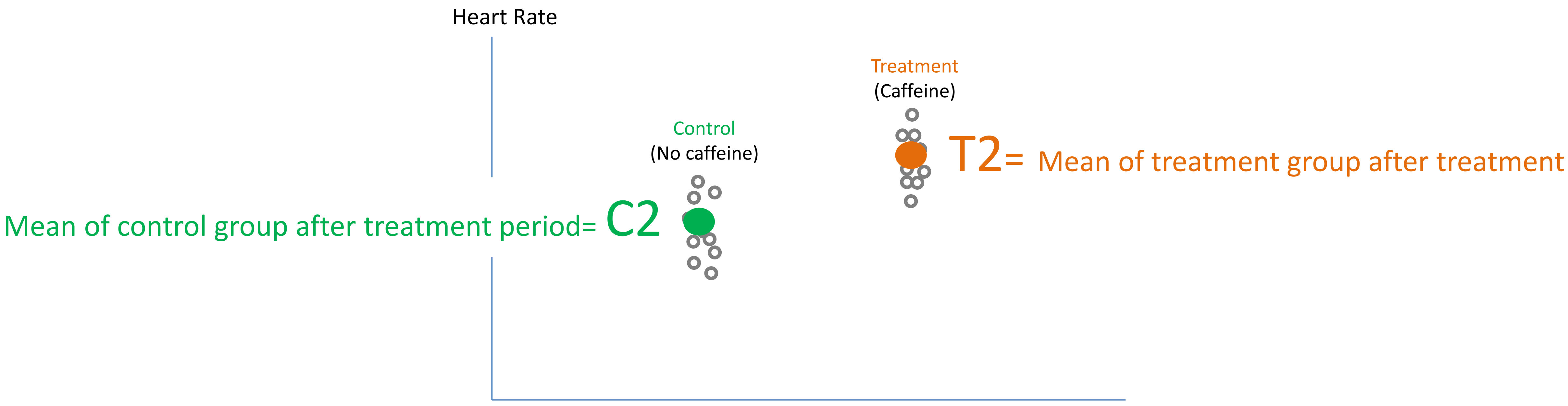
REVIEW: HYPOTHESIS-TESTING

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

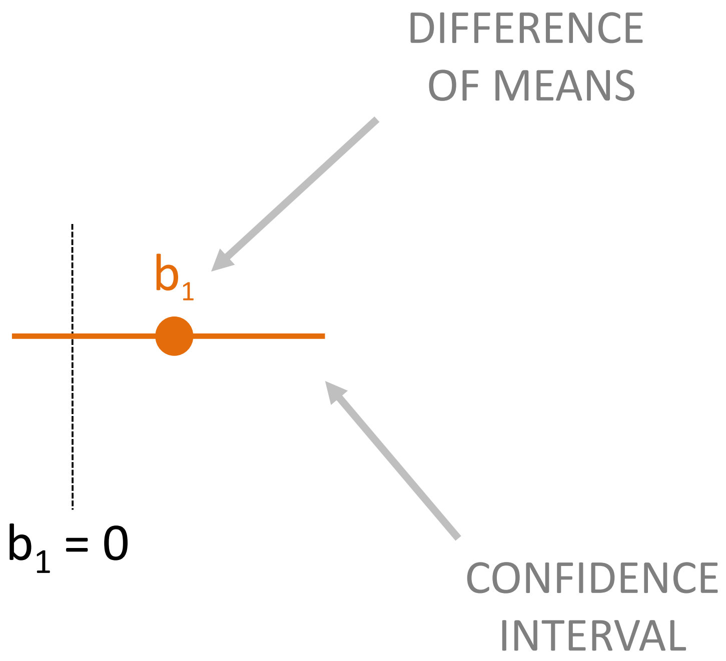


THE PROGRAM EVALUATION FRAMEWORK

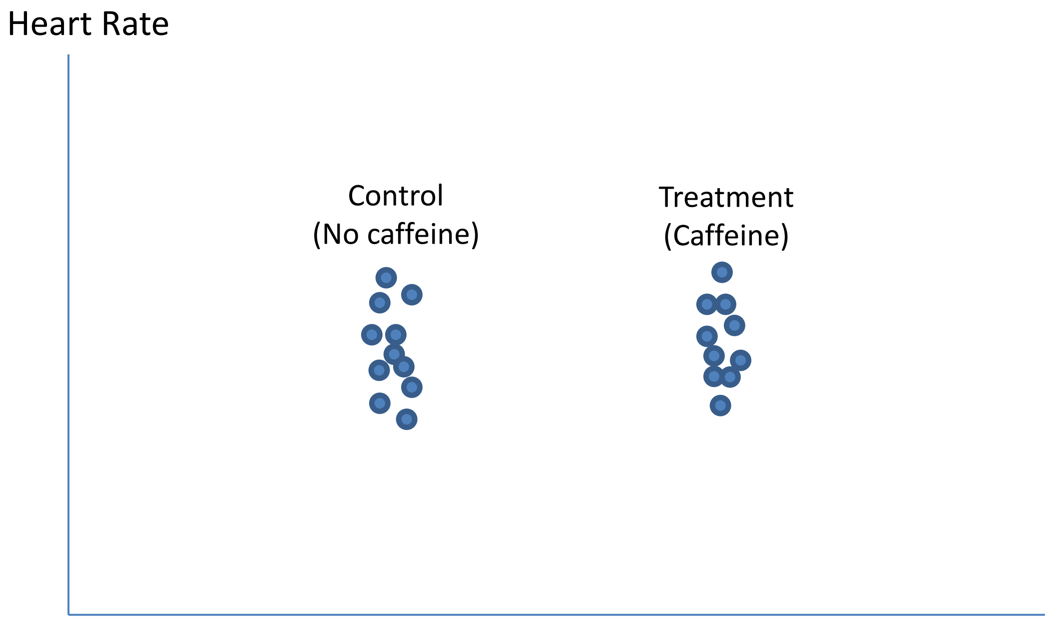
$$Program\ Effect\ =\ T2 - C2$$

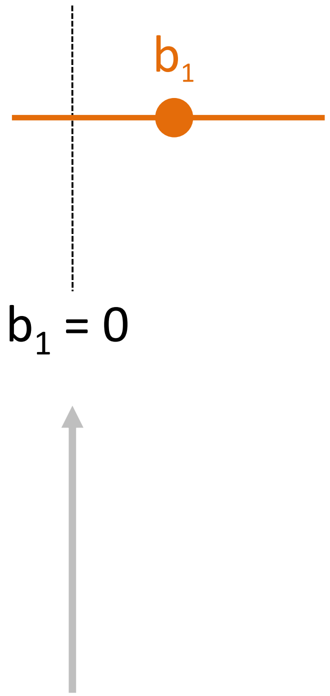


b_1 = MEAN_{treat} - $\text{MEAN}_{control}$



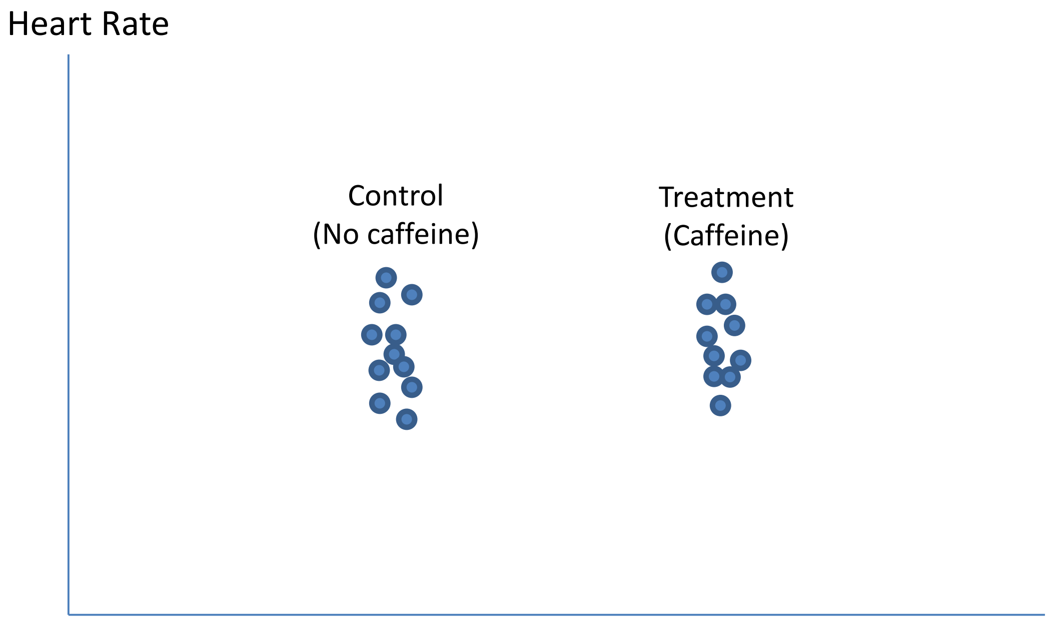
b_1 = MEAN_{treat} - $\text{MEAN}_{control}$



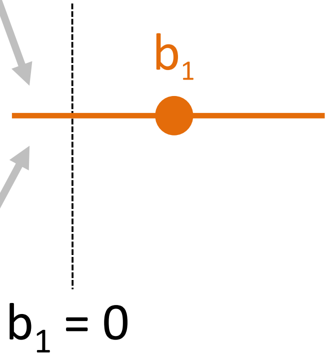


NULL HYPOTHESIS
(b_1=0 means NO IMPACT)

$$b_1 = \text{MEAN}_{treat} - \text{MEAN}_{control}$$



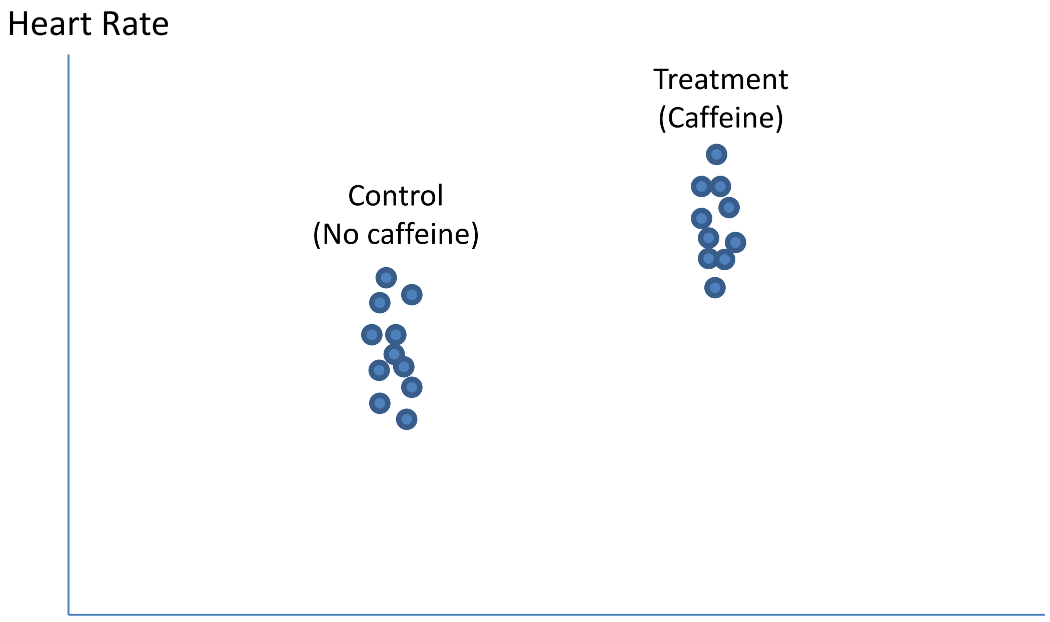
STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

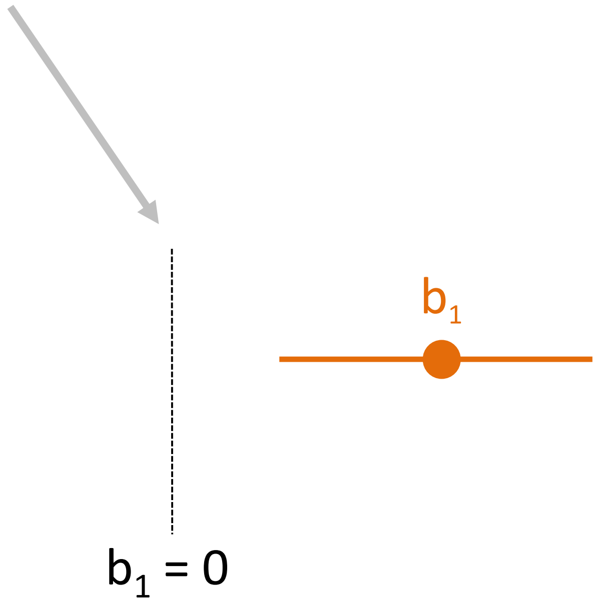


NOT SIGNIFICANT
(NO PROGRAM IMPACT)

STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

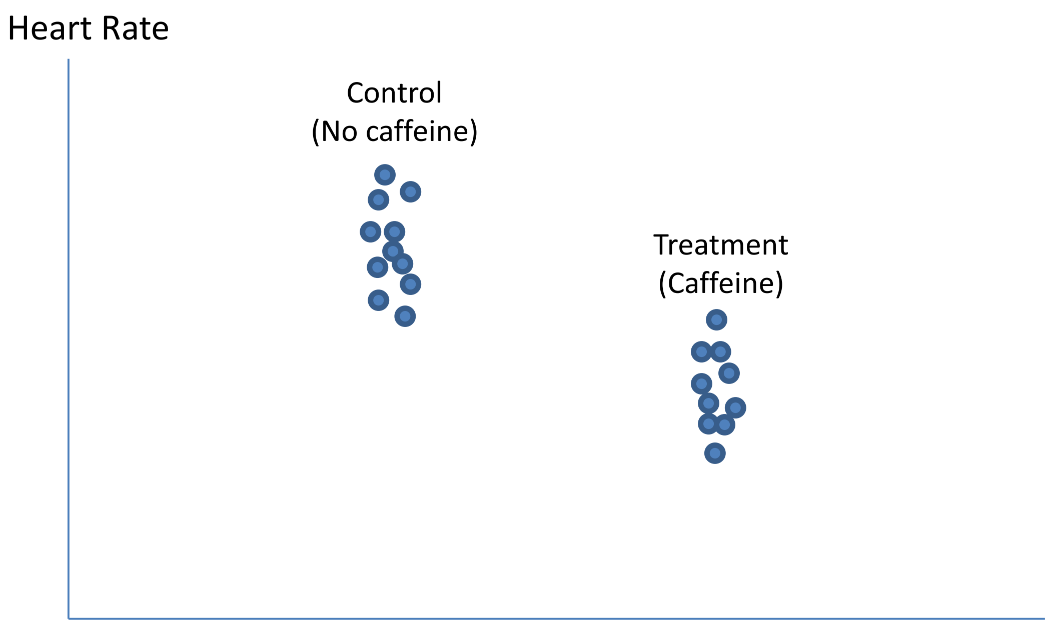
b_1 = MEAN_{treat} - $\text{MEAN}_{control}$



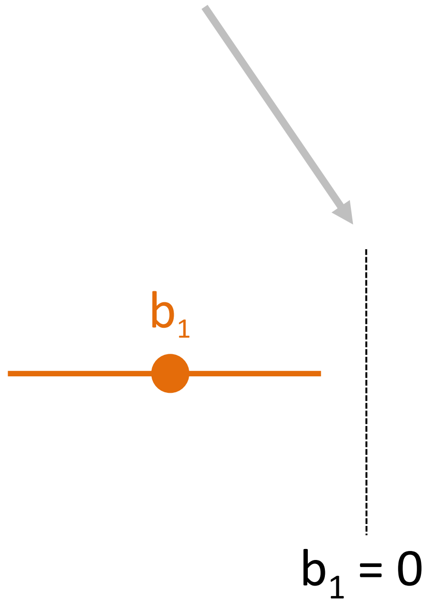


SIGNIFICANT
(POSITIVE PROGRAM IMPACT)

b_1 = MEAN_{treat} - $\text{MEAN}_{control}$



STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

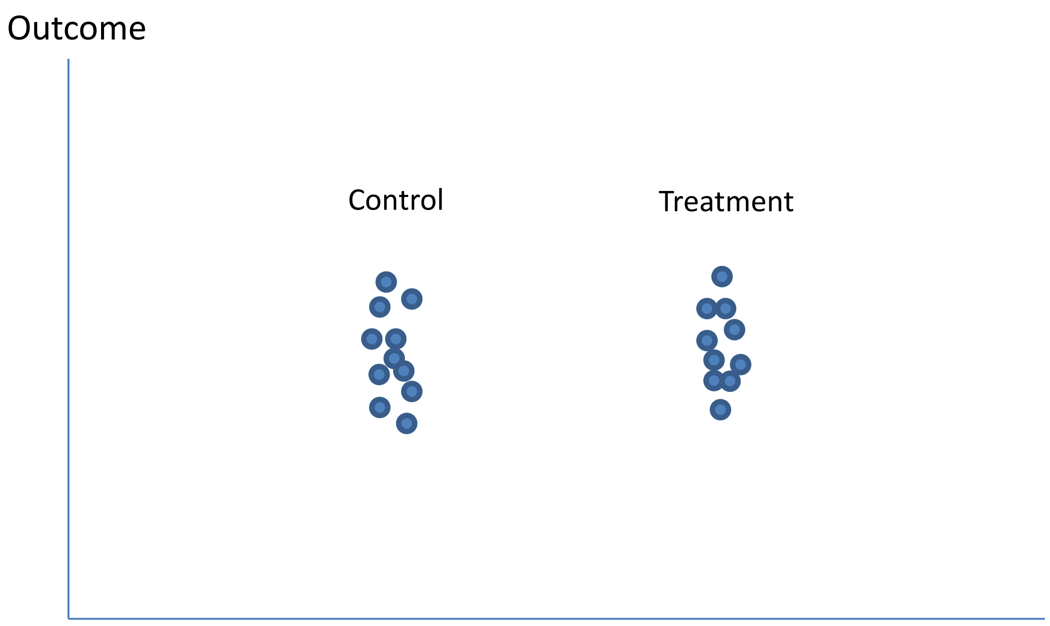


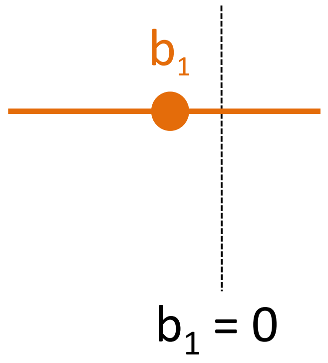
SIGNIFICANT
(NEGATIVE PROGRAM IMPACT)

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

$b_1 = T2 - C2$

No Program Impact

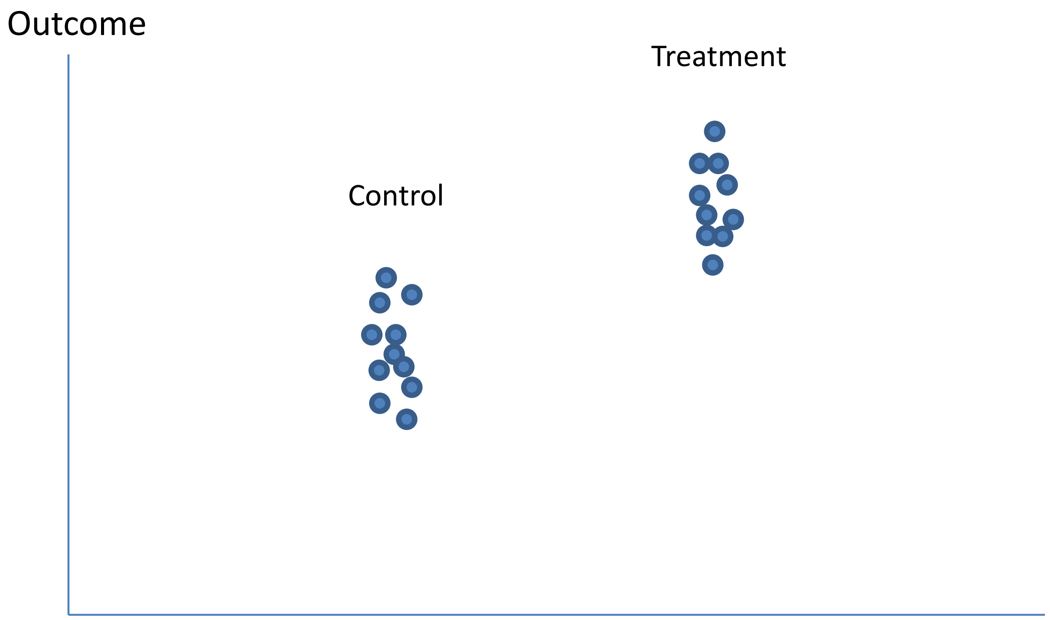


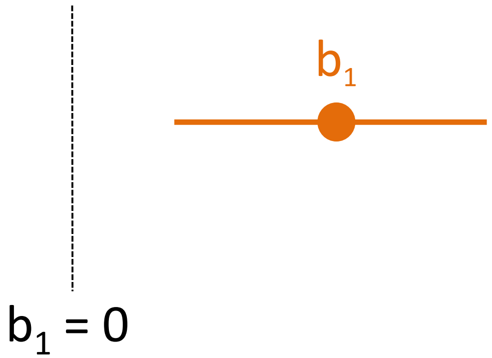


$$b_1 = MEAN_{treat} - MEAN_{control}$$

$$b_1 = T2 - C2$$

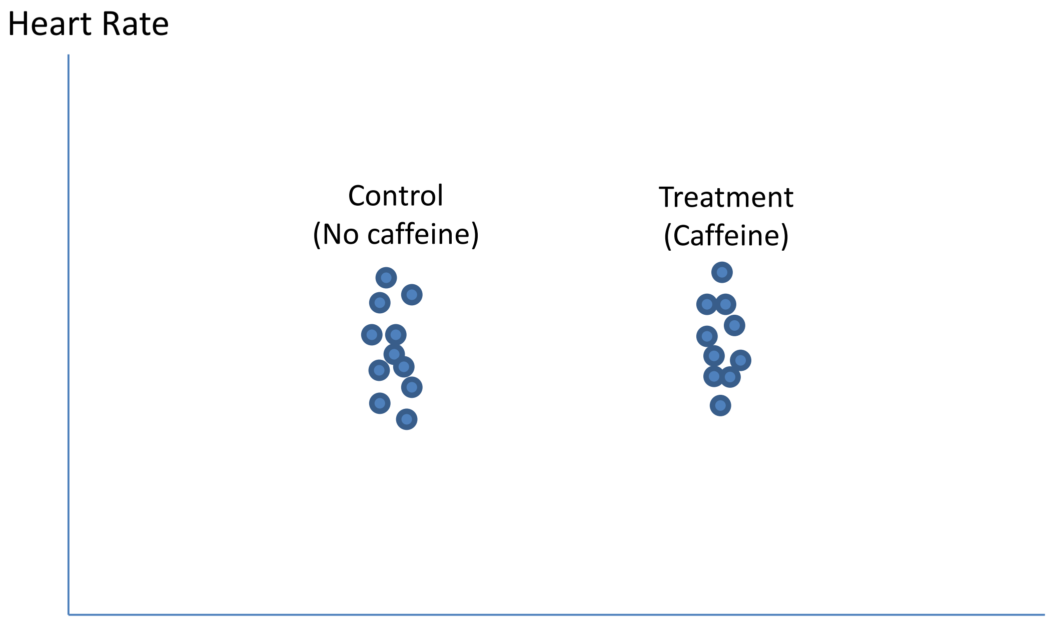
Positive Program Impact

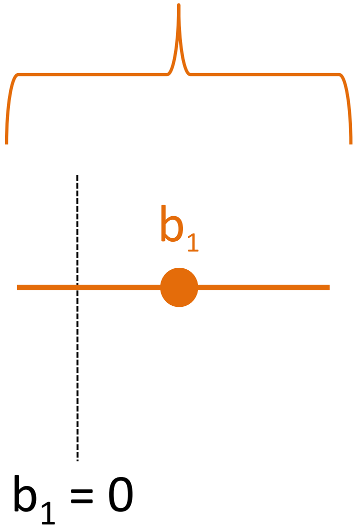




Definition of EFFECT in program eval:
Observed change + confidence interval
(size of observed impact plus accuracy,
can we say with confidence it’s positive)

b_1 = MEAN_{treat} - $\text{MEAN}_{control}$

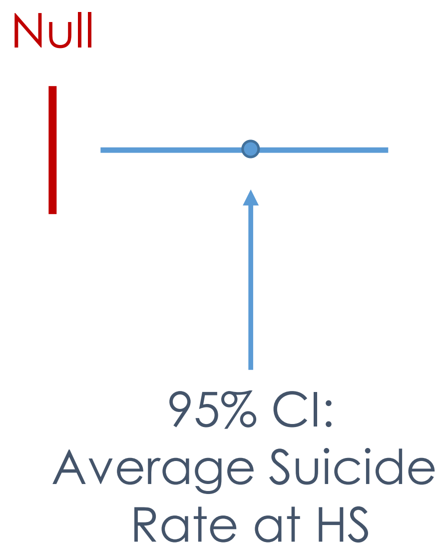




BACK TO THE EXAMPLE

Were suicide rates <u>*HIGH*</u> for a specific high school district in suburban California?

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

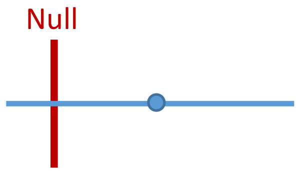


Null Hypothesis:
Population Average

Suicide rates are **HIGHER** than the population average (significant at a 0.05 level)

Were suicide rates *<u>HIGH</u>* for a specific high school district in suburban California?

Null Hypothesis:
Population Average

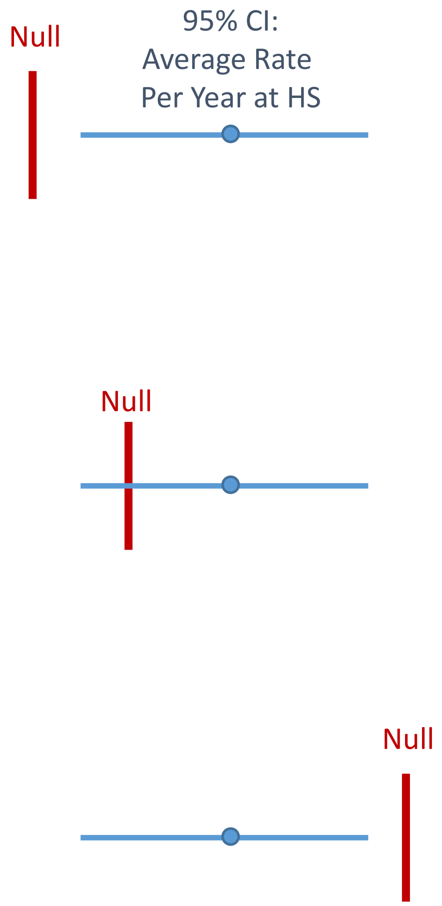


Null Hypothesis:
All HS Students
OR All Californians

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

These are all valid counterfactuals.

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 <u>different than outcomes in the comparison group</u>. The comparison group is defined by the researcher.

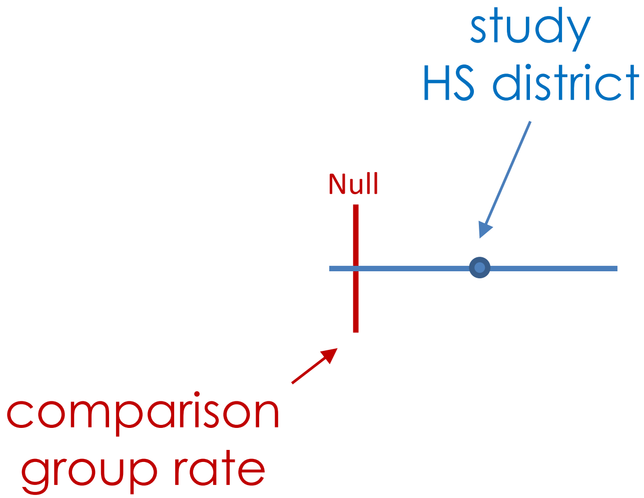
   In some special cases the comparison group is identical (statistically speaking) to the treatment group. In this case we call it a “control” group.

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.

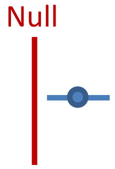
Which is a more meaningful finding?

CASE A:



Suicide rates at the high school are **MUCH LARGER** than expected (triple the comparison group rate)
But **NOT** statistically significant at the alpha=0.05 level

CASE B:



Suicide rates are **SLIGHTLY LARGER** than the comparison group (0.25 cases per year in the district)
And statistically significant

TRUE EXPERIMENTS

Figure 3.1 The Perfect Clone

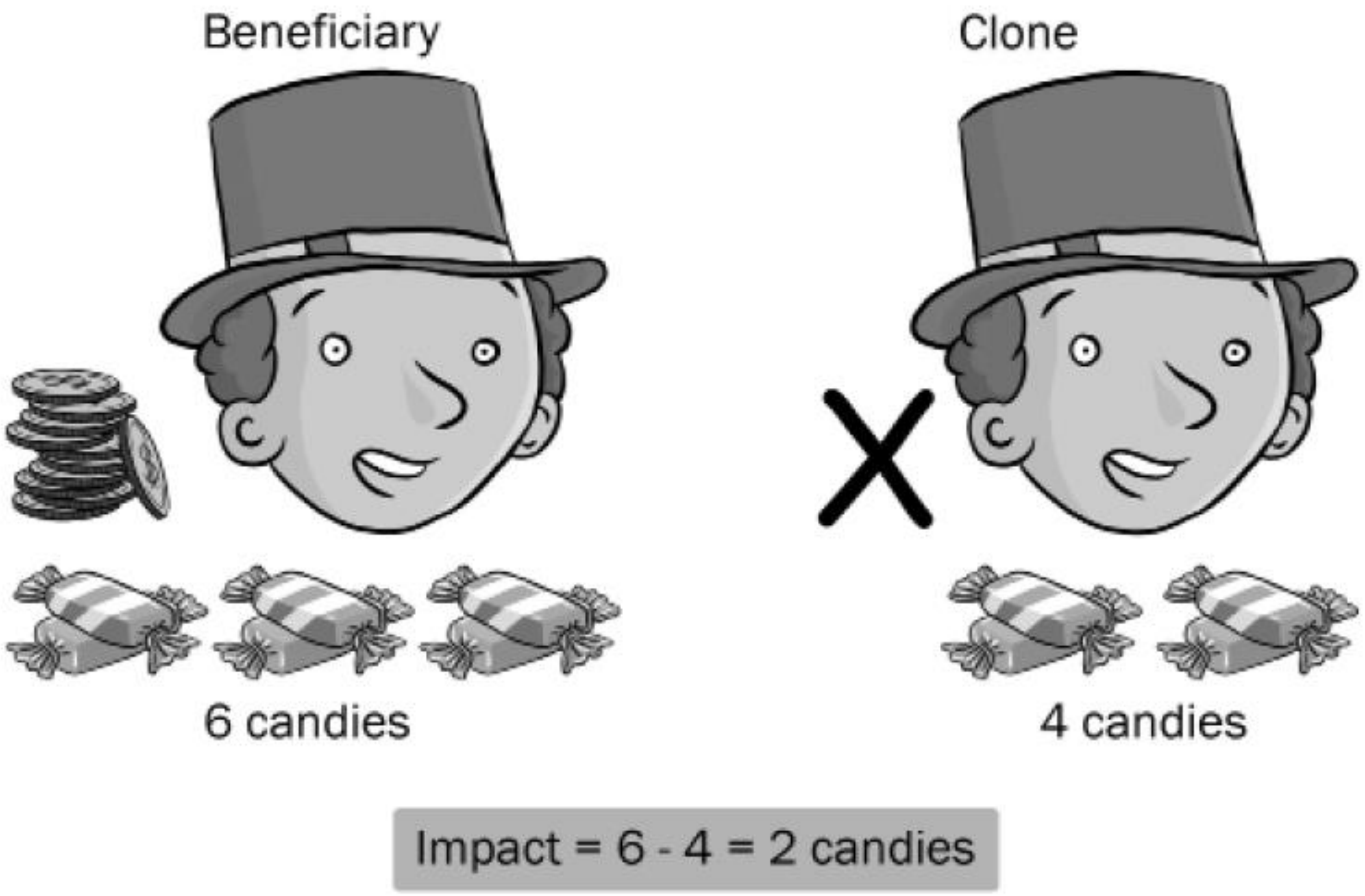


Figure 4.3 Steps in Randomized Assignment to Treatment

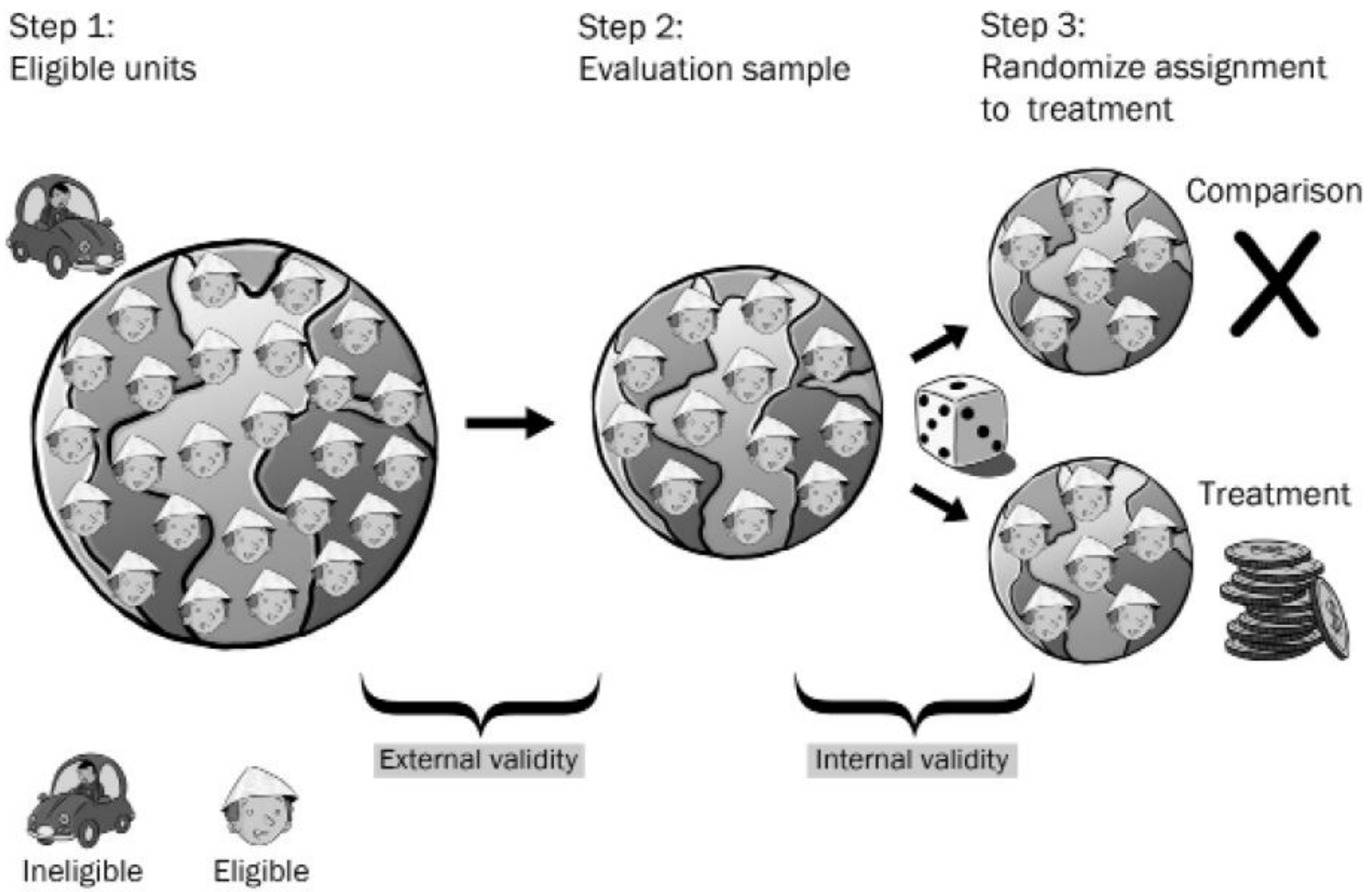
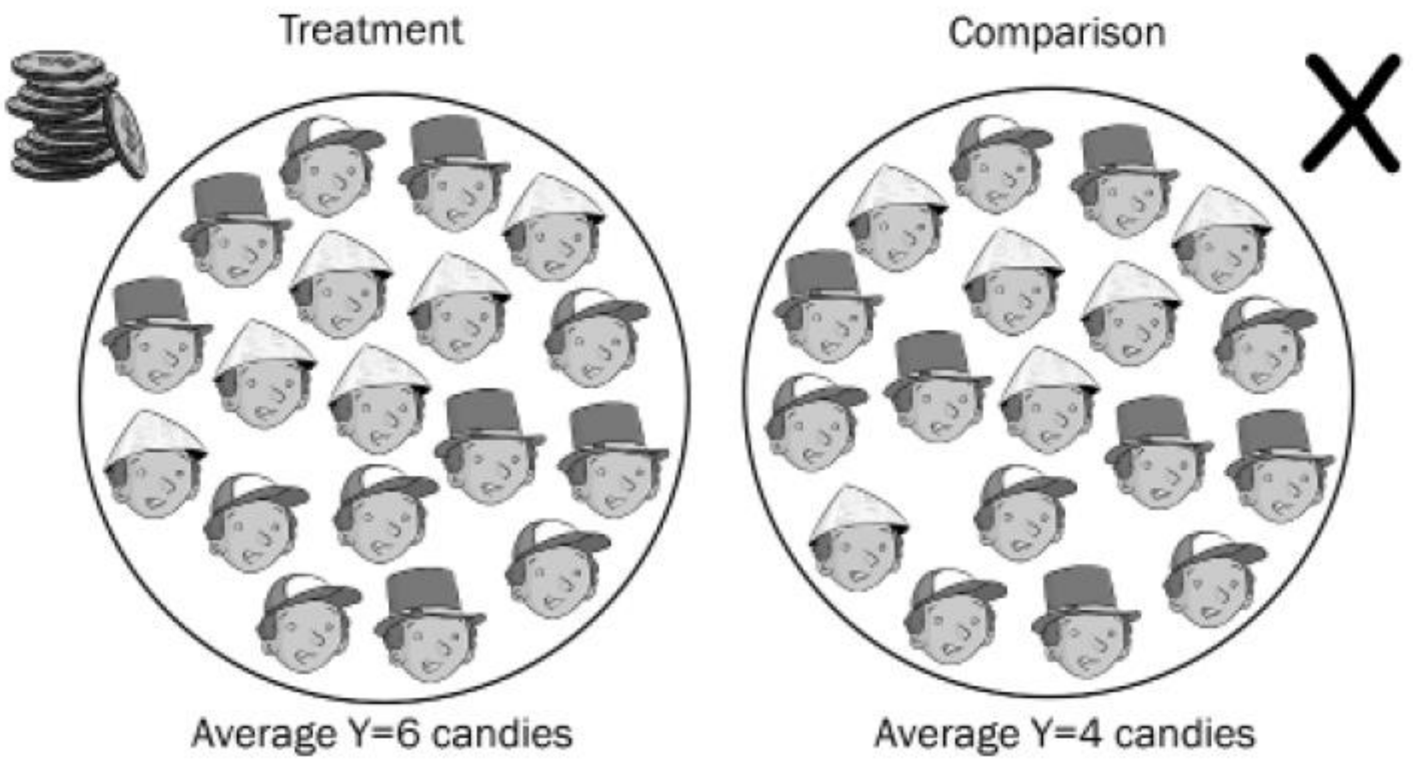


Figure 3.2 A Valid Comparison Group



Impact = 6 - 4 = 2 candies

b_1 = T2 – C2

NON-EXPERIMENTS

Experiments

Quasi-Experiments

Observational Studies

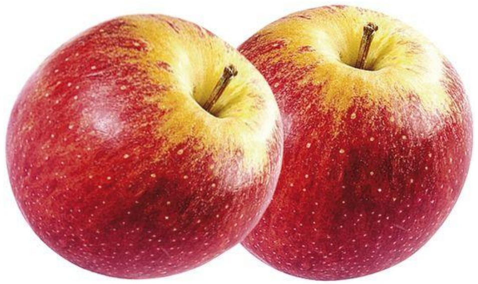
When careful, quasi-experimental methods can produce the same results as experimental methods

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *27*(4), 724-750.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*(2), 207-244.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs.

Y(treatment) – Y(CONTROL) = Y(treatment) – Y(COMPARISON)



This course covers the conditions under which these cases should be equivalent