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

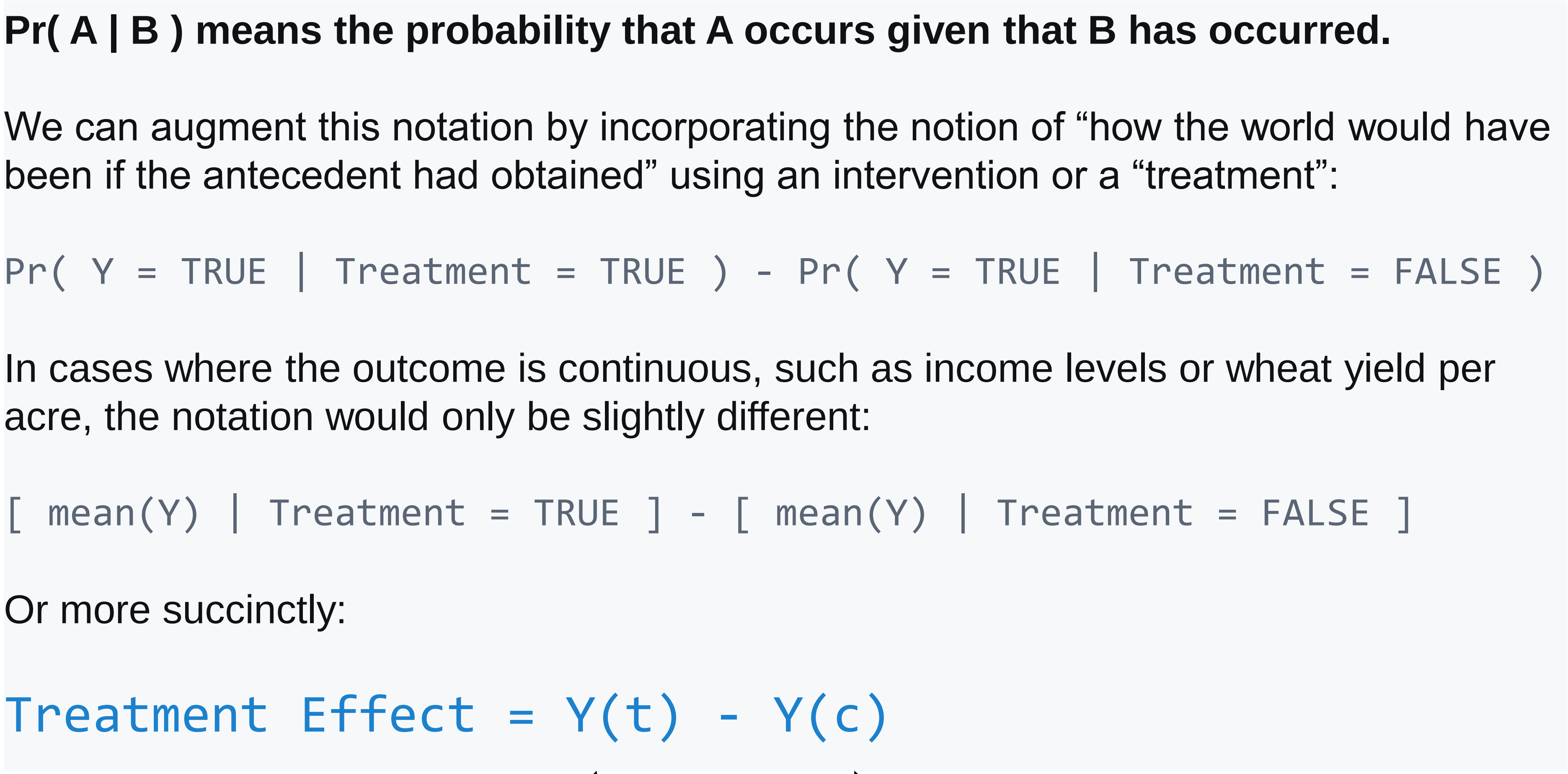
A counterfactual assertion is a conditional whose antecedent is false and whose consequent describes **how the world would have been if the antecedent had obtained.**

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

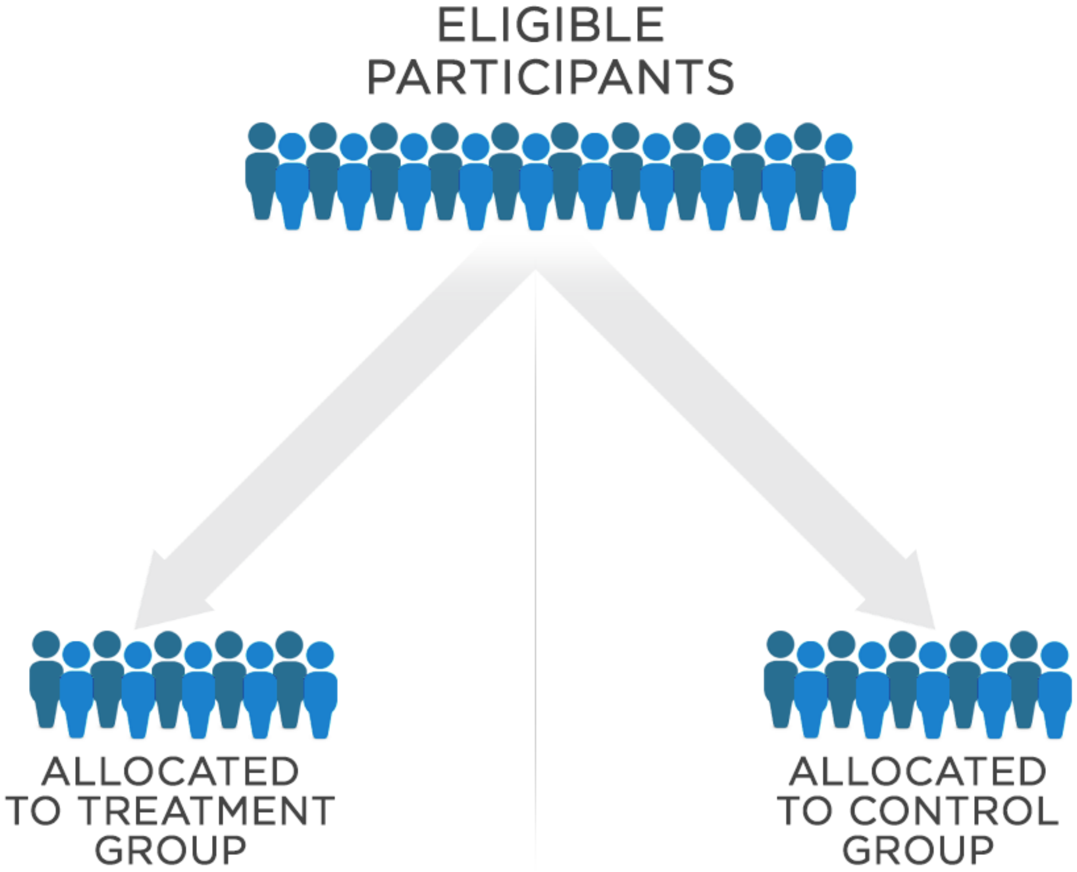
Effect of climate accord = $Y(t) - Y(c)$

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

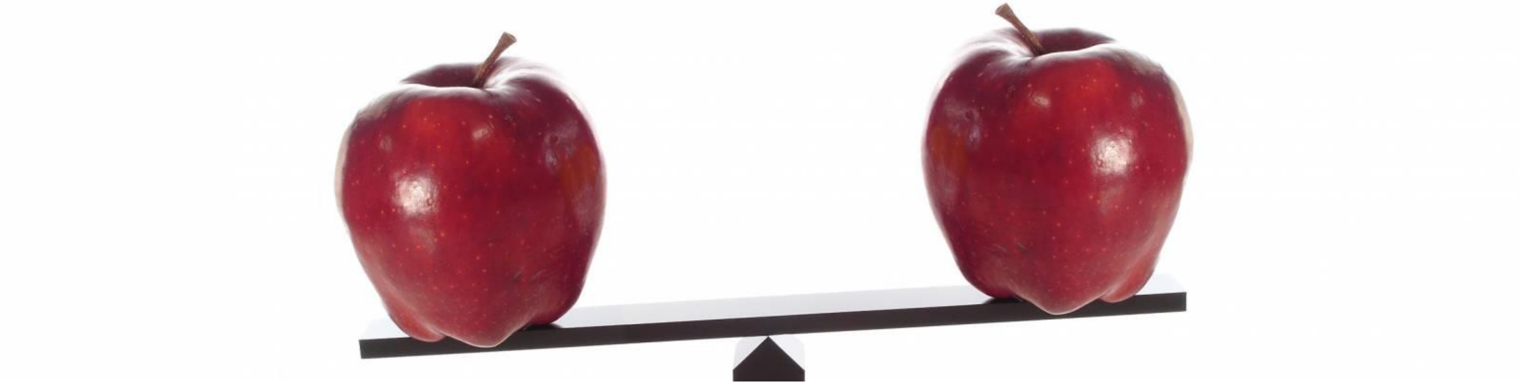
Effect of climate accord = $Y(t) - Y(c)$

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**.



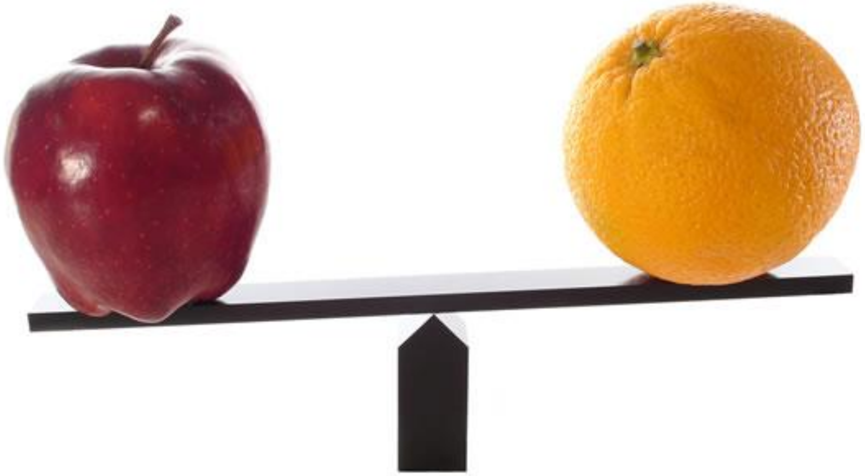
$$\text{Gains from fertilizer} = \text{weight}_{\text{treatment}} - \text{weight}_{\text{control}}$$



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.

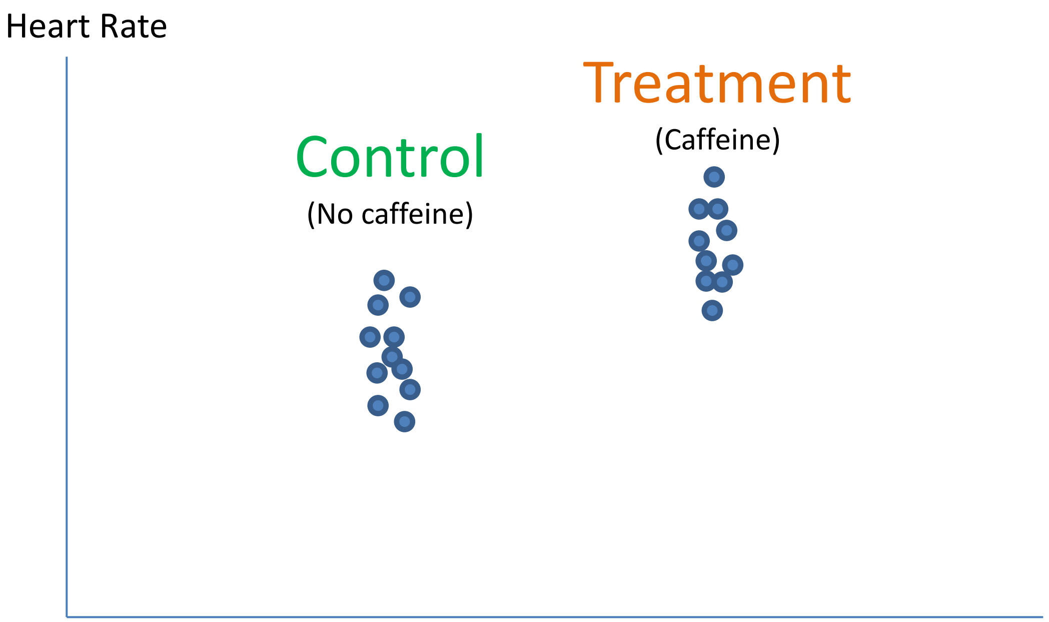
The current district superintendent cut support for school counseling services. Parents are considering filing a lawsuit against the school district because they feel the cuts in spending resulted in loss of support for mental health services, thus leading to increased rates in suicide.

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**.

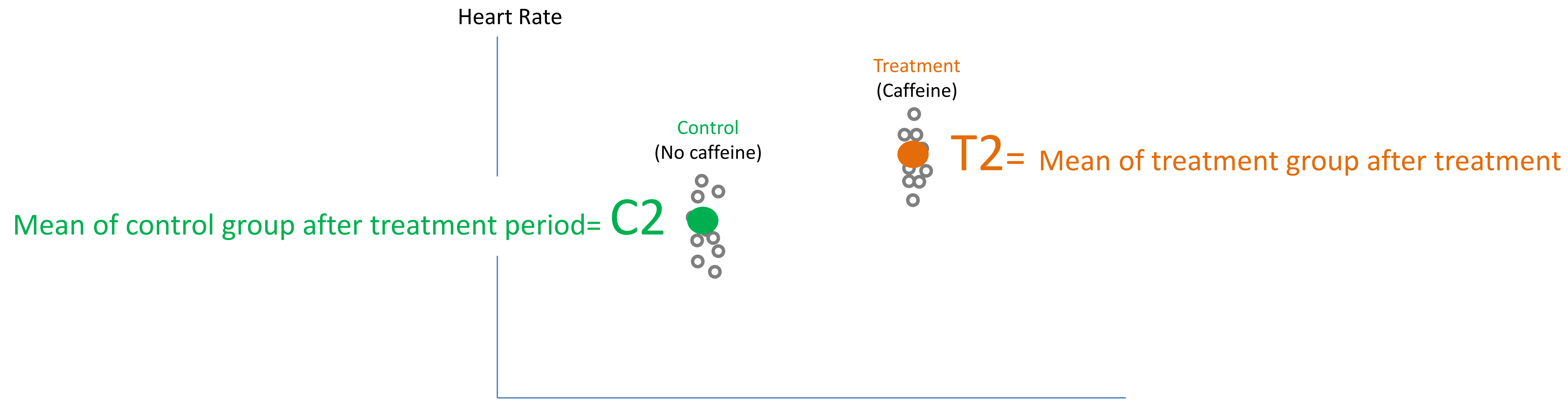
REVIEW:
HYPOTHESIS-TESTING

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

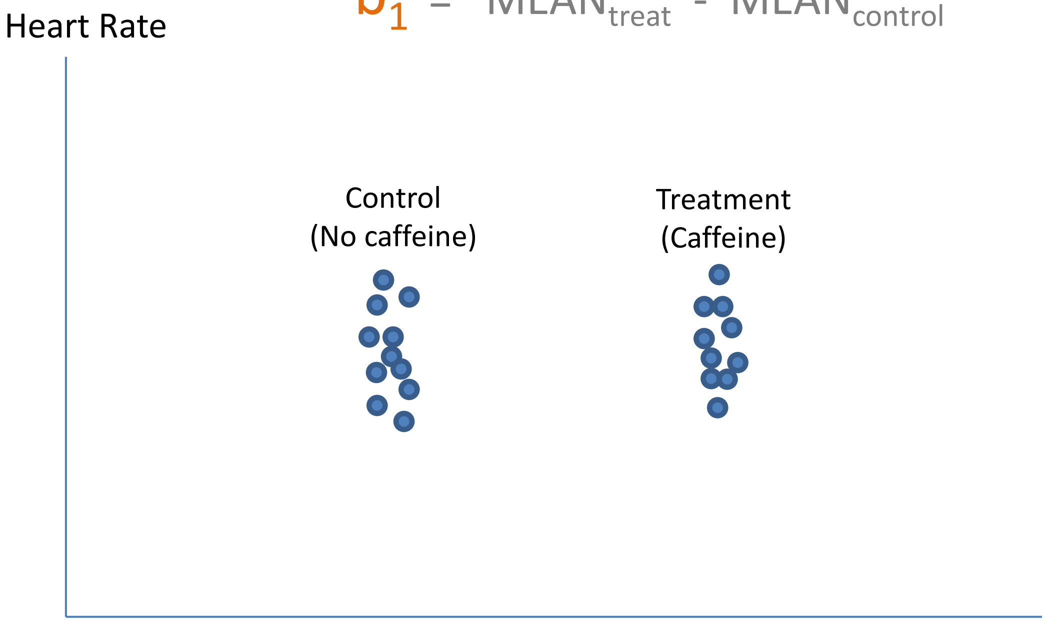


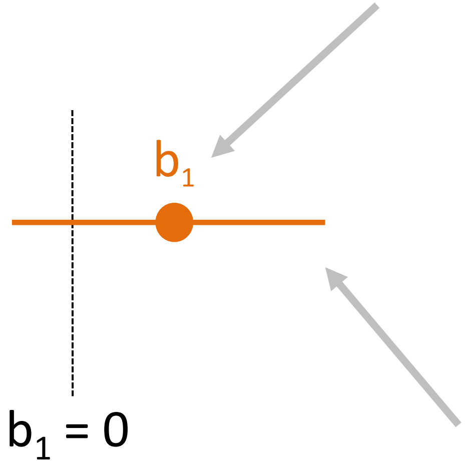
THE PROGRAM EVALUATION FRAMEWORK

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



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

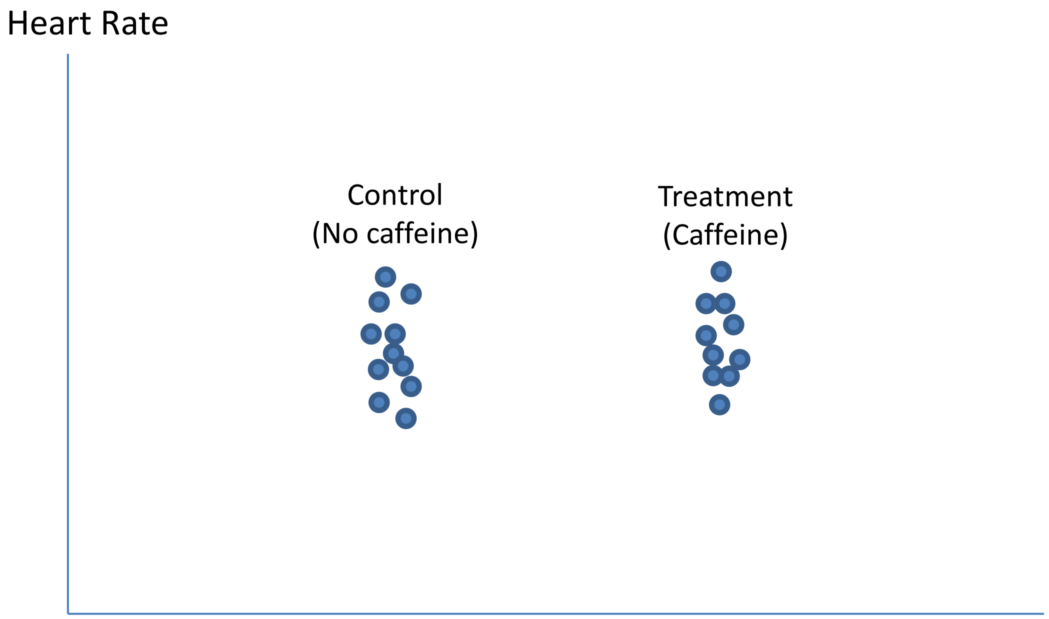


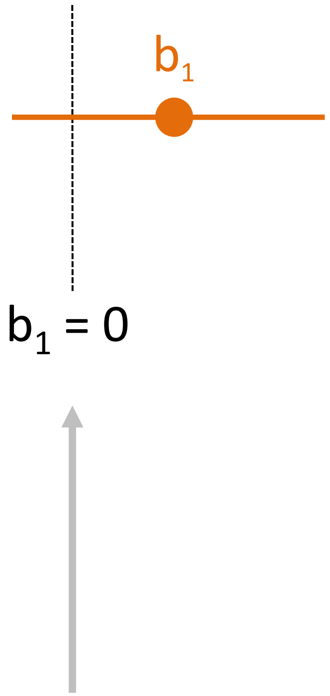


DIFFERENCE OF MEANS

CONFIDENCE INTERVAL

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

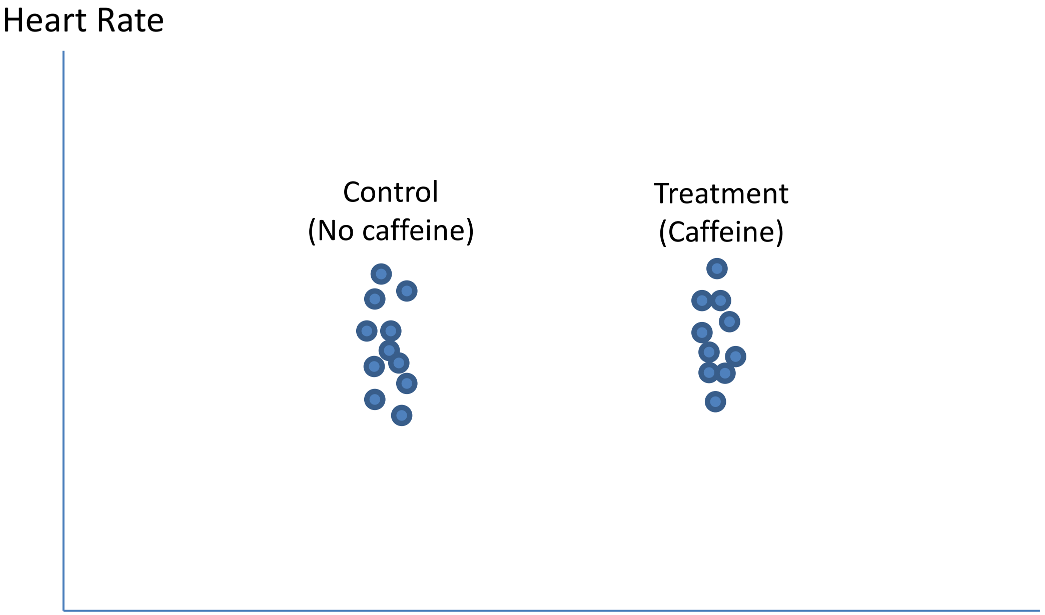


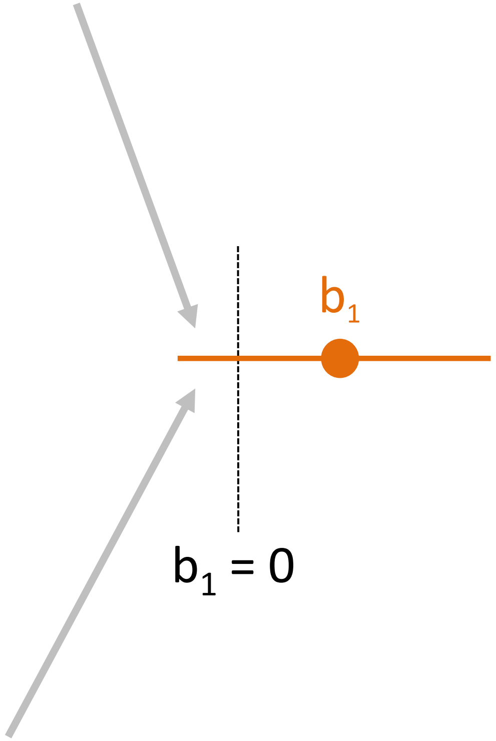


NULL HYPOTHESIS
(b_1=0 means NO IMPACT)

STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

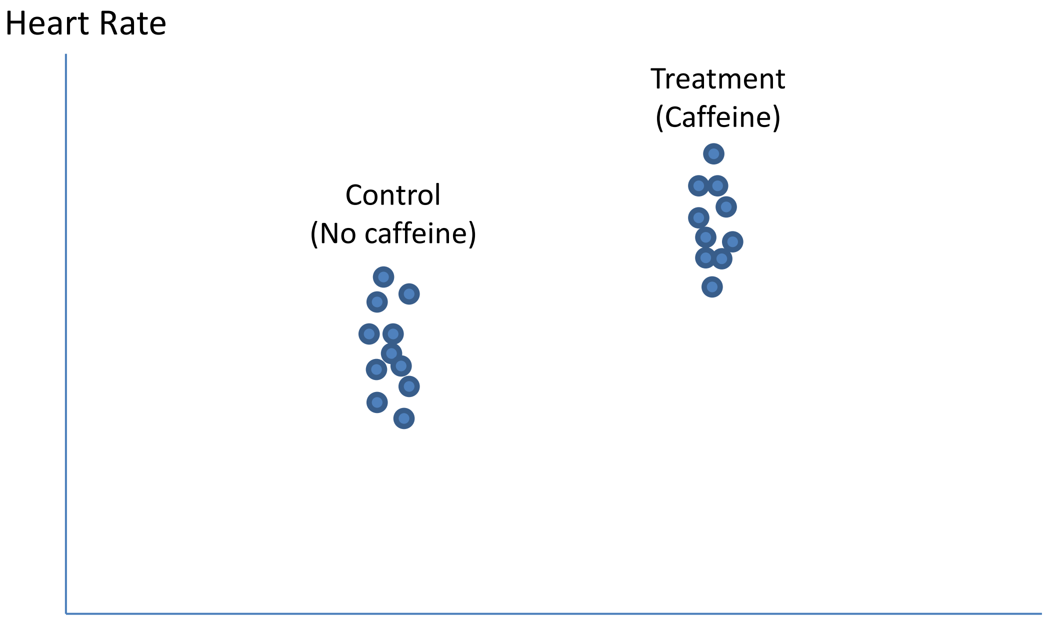
$$b_1 = \text{MEAN}_{treat} - \text{MEAN}_{control}$$



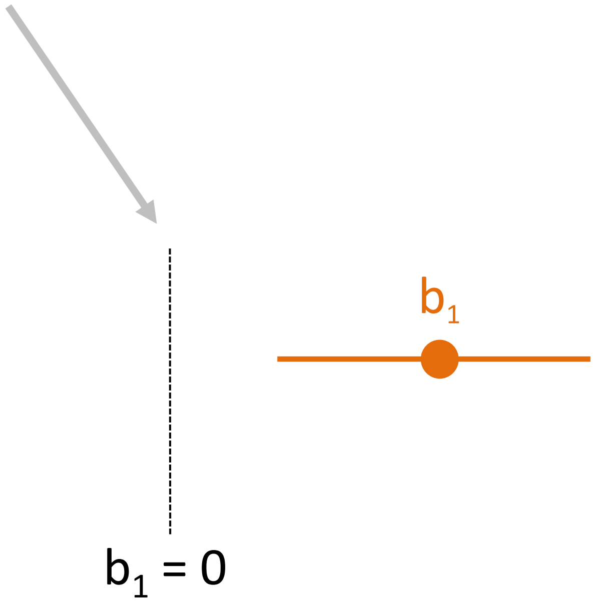


NOT SIGNIFICANT
(NO PROGRAM IMPACT)

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



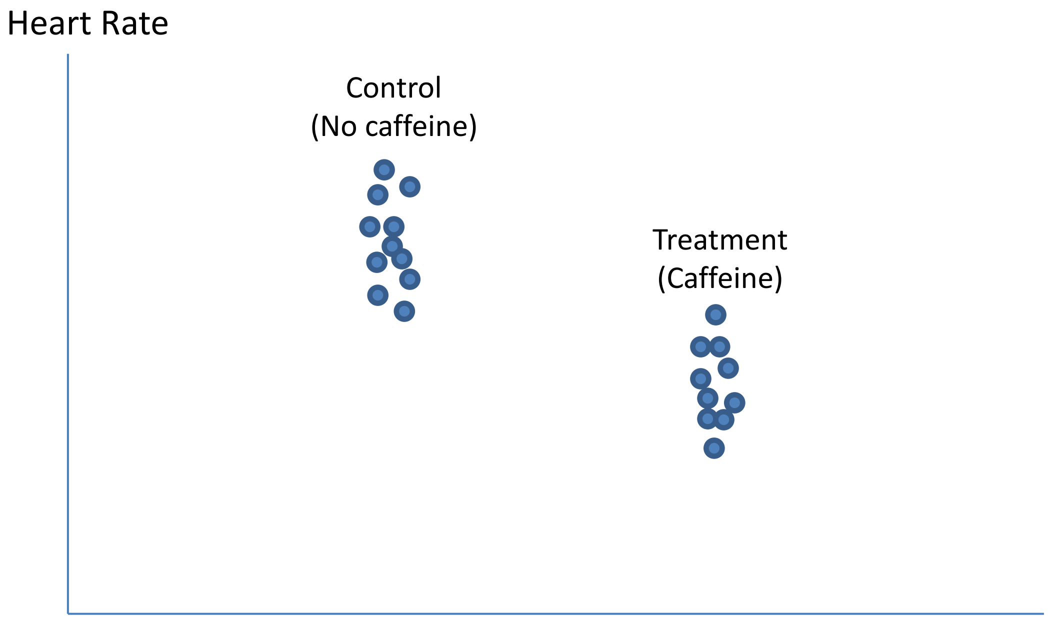
STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

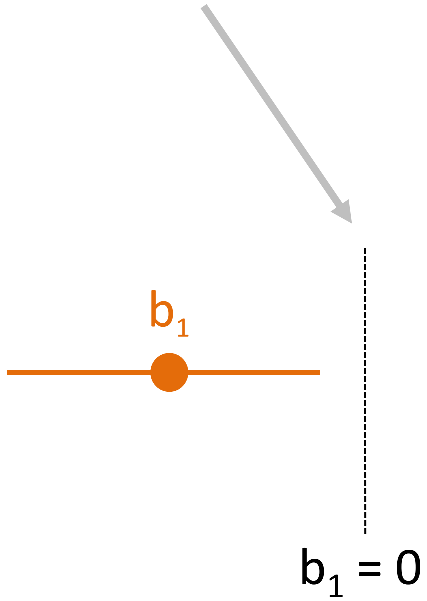


SIGNIFICANT
(POSITIVE PROGRAM IMPACT)

STATISTICAL SIGNIFICANCE
(CONF. INT. CONTAINS ZERO?)

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



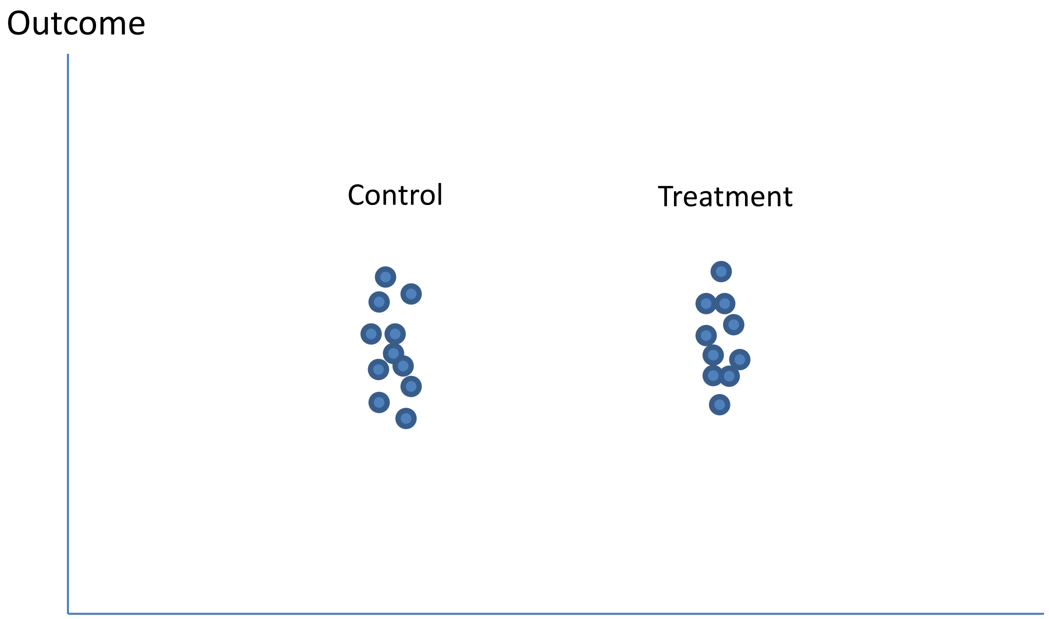


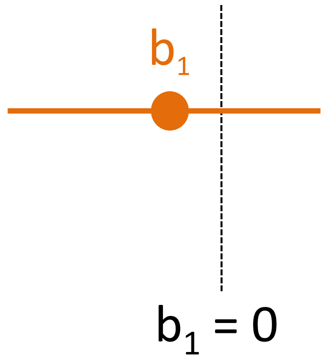
SIGNIFICANT
(NEGATIVE PROGRAM IMPACT)

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

$b_1 = \text{T2} – \text{C2}$

No Program Impact

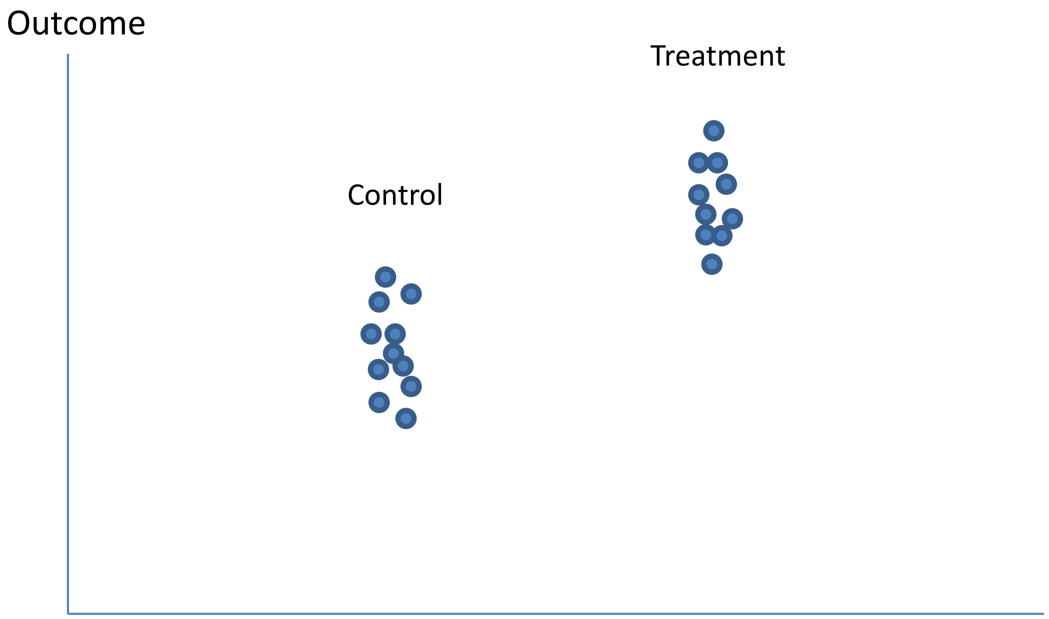


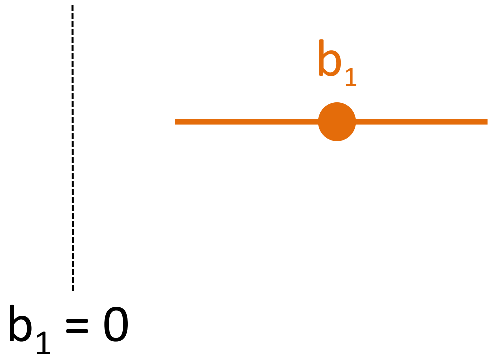


Positive Program Impact

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

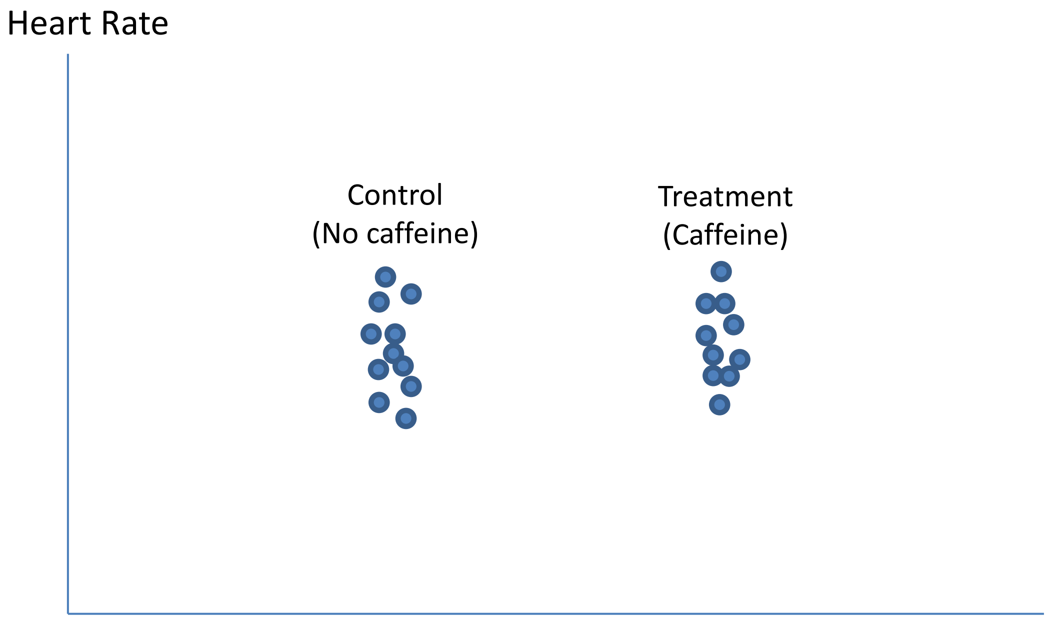
$$b_1 = T2 - C2$$

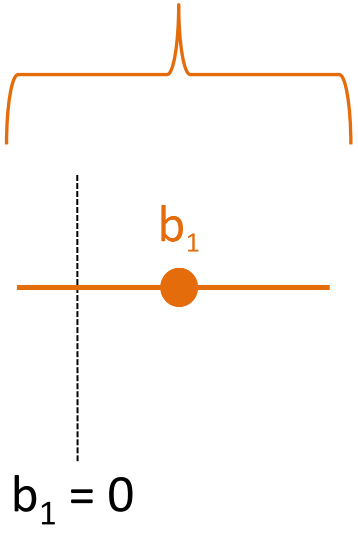




Definition of EFFECT in program eval:
Observed change + confidence interval
(size of observed impact plus accuracy,
can we say with confidence it's positive)

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

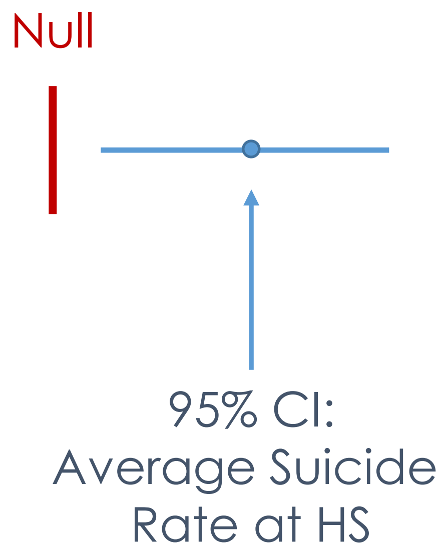




BACK TO THE EXAMPLE

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

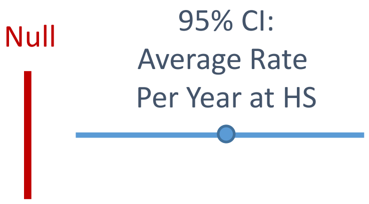
Were suicide rates *HIGH* for a specific high school district in suburban California?



Null Hypothesis:
Population Average

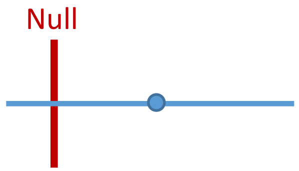
Suicide rates are **HIGHER** than the population average (significant at a 0.05 level)

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



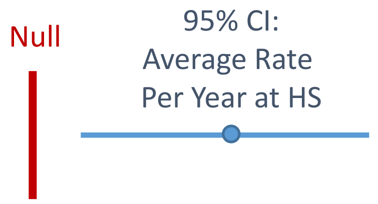
Null Hypothesis:
Population Average

Suicide rates are **NO DIFFERENT** than the population average (NOT significant at a 0.05 level)



Null Hypothesis:
All HS Students
OR All Californians

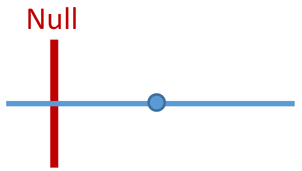
Were suicide rates *HIGH* for a specific high school district in suburban California?



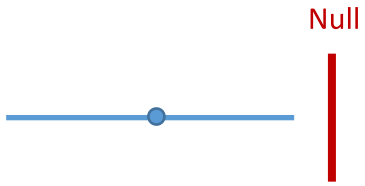
Null Hypothesis:
Population Average

These are all valid counterfactuals.

How we define our comparison drives the conclusions.



Null Hypothesis:
All HS Students
OR All Californians



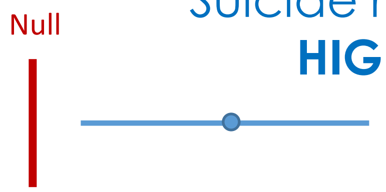
Suicide rates are **LOWER** than the population average (significant at a 0.05 level)

Null Hypothesis:
All Suburban HS Students

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

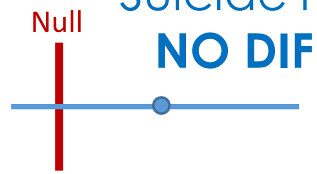
The conclusions we reach are not at all dependent upon the model or lack of significance – they are driven entirely by the selection of the counterfactual.

All of these are reasonable counterfactuals. Which best answers the research question?



Suicide rates are **HIGHER**

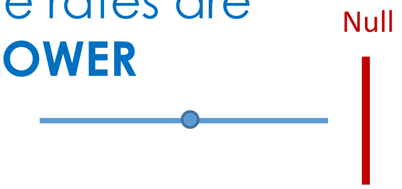
Null Hypothesis:
Population Average



Suicide rates are **NO DIFFERENT**

Null Hypothesis:
All HS Students
OR All Californians

Suicide rates are **LOWER**



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 <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.

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.

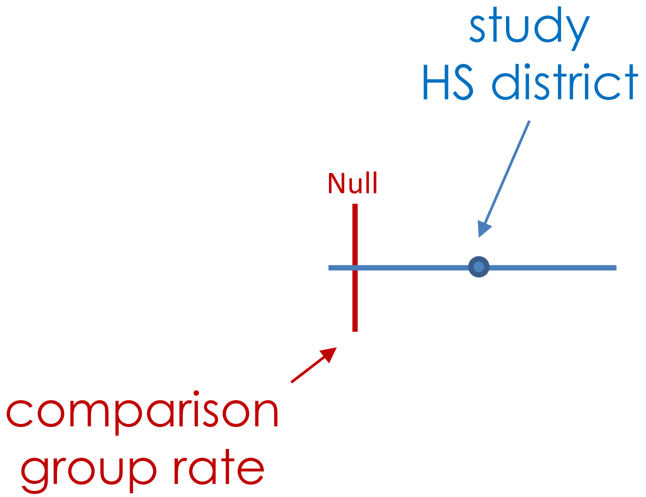
   In some special cases the comparison group is identical (statistically speaking) to the treatment group. In this case we call it a "control" group.

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.

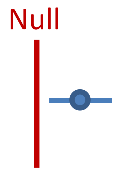
Which is a more meaningful finding?

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

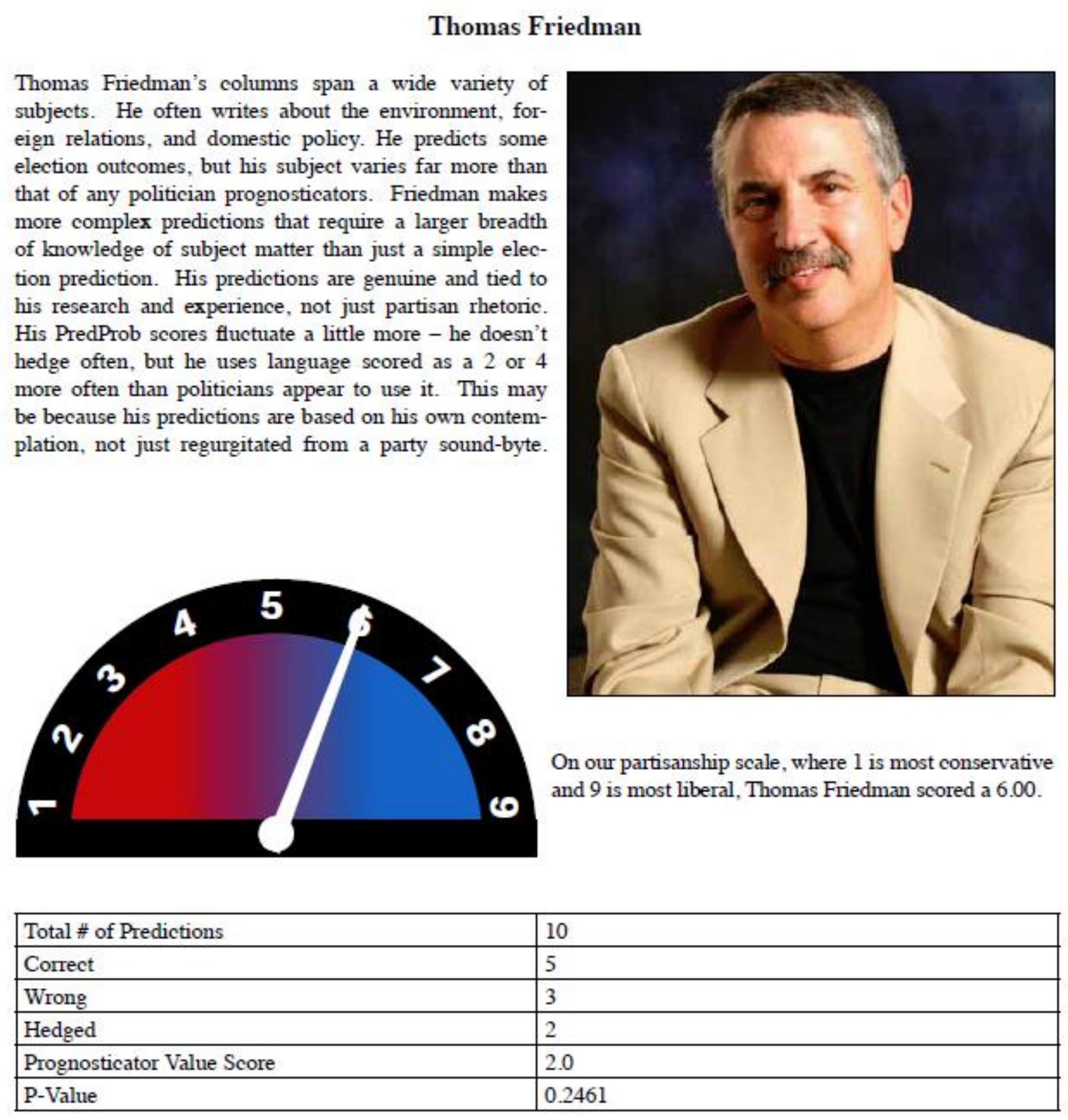
IT CAN BE HARDER THAN YOU THINK TO IDENTIFY A MEANINGFUL COUNTERFACTUAL!

Donaldson, H., Doubleday, R., Hefferman, S., Klondar, E., & Tummarello, K. (2011). Are Talking Heads Blowing Hot Air? An Analysis of the Accuracy of Forecasts in the Political Media. *Hamilton College, Public Policy*, *501*.

A class at Hamilton College led by public policy professor P. Gary Wyckoff has analyzed the predictions of 26 prognosticators between September 2007 and December 2008. The Hamilton students sampled the predictions of 26 individuals who wrote columns in major print media and who appeared on the three major Sunday news shows – Face the Nation, Meet the Press, and This Week.

They evaluated the accuracy of 472 predictions made during the 16-month period. They used a scale of -10 to +10 to rate the accuracy of each.

https://www.poynter.org/reporting-editing/2011/claim-krugman-is-top-prognosticator-cal-thomas-is-the-worst/



Thomas Friedman

Thomas Friedman's columns span a wide variety of subjects. He often writes about the environment, foreign relations, and domestic policy. He predicts some election outcomes, but his subject varies far more than that of any politician prognosticators. Friedman makes more complex predictions that require a larger breadth of knowledge of subject matter than just a simple election prediction. His predictions are genuine and tied to his research and experience, not just partisan rhetoric. His PredProb scores fluctuate a little more – he doesn't hedge often, but he uses language scored as a 2 or 4 more often than politicians appear to use it. This may be because his predictions are based on his own contemplation, not just regurgitated from a party sound-byte.

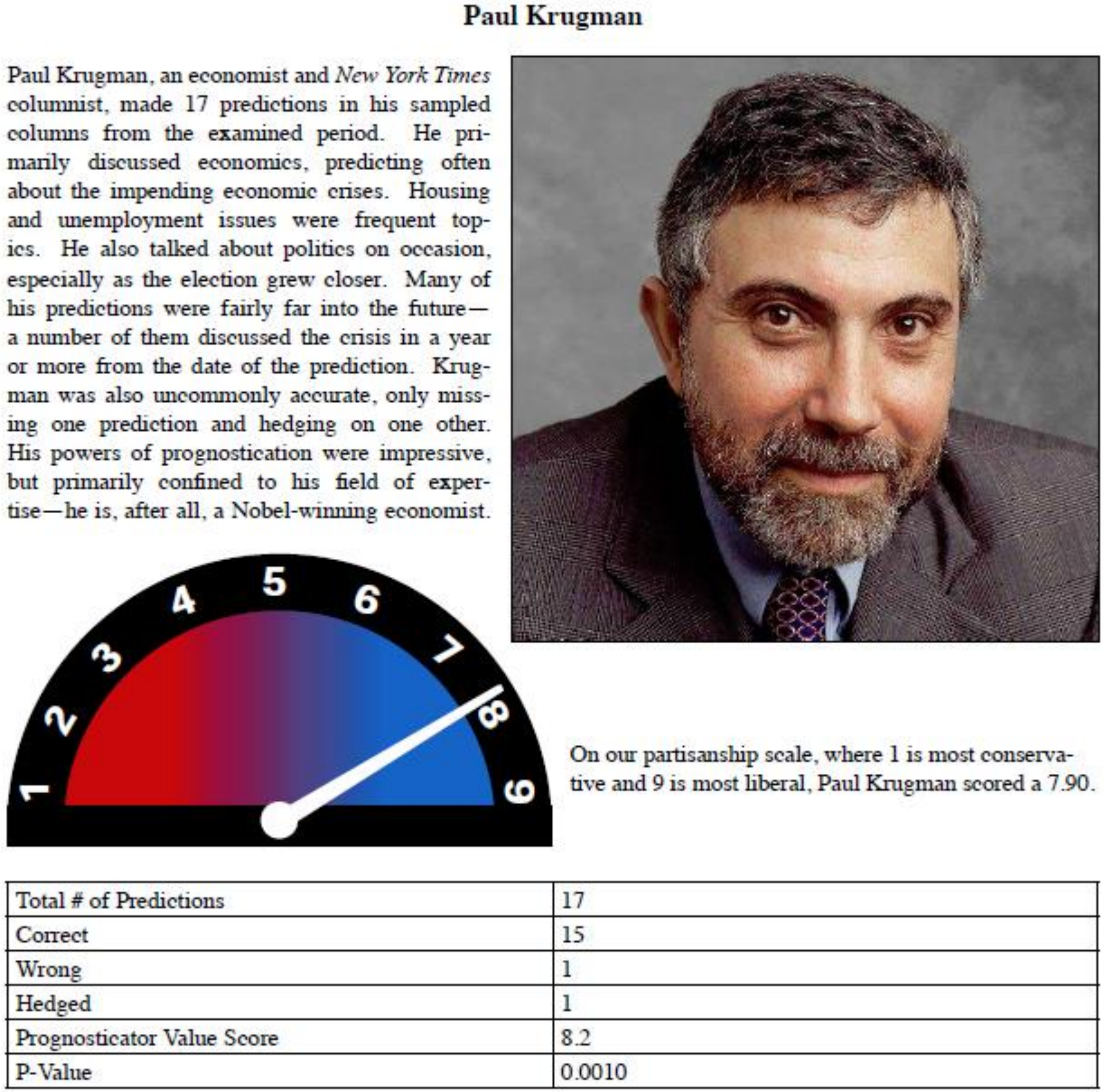
On our partisanship scale, where 1 is most conservative and 9 is most liberal, Thomas Friedman scored a 6.00.

Total # of Predictions	10
Correct	5
Wrong	3
Hedged	2
Prognosticator Value Score	2.0
P-Value	0.2461

Scored 2 out of 10 on the prognostication scale:

p-value of 0.2461

(NOT statistically significant)



Paul Krugman

Paul Krugman, an economist and *New York Times* columnist, made 17 predictions in his sampled columns from the examined period. He primarily discussed economics, predicting often about the impending economic crises. Housing and unemployment issues were frequent topics. He also talked about politics on occasion, especially as the election grew closer. Many of his predictions were fairly far into the future—a number of them discussed the crisis in a year or more from the date of the prediction. Krugman was also uncommonly accurate, only missing one prediction and hedging on one other. His powers of prognostication were impressive, but primarily confined to his field of expertise—he is, after all, a Nobel-winning economist.

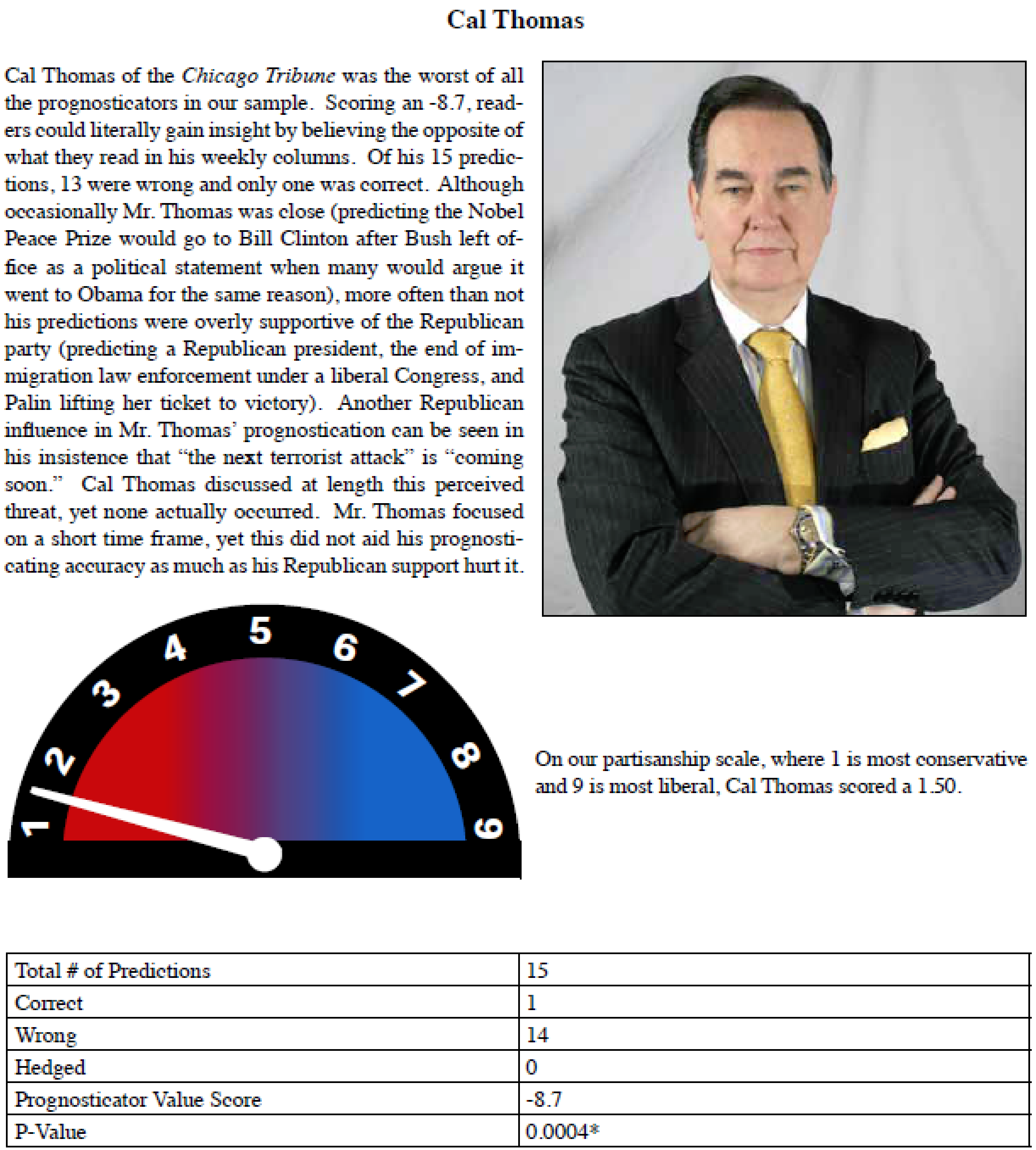
On our partisanship scale, where 1 is most conservative and 9 is most liberal, Paul Krugman scored a 7.90.

Total # of Predictions	17
Correct	15
Wrong	1
Hedged	1
Prognosticator Value Score	8.2
P-Value	0.0010

Scored 8.2 out of 10 on the prognostication scale:

p-value of 0.001

(statistically significant)



Cal Thomas

Cal Thomas of the *Chicago Tribune* was the worst of all the prognosticators in our sample. Scoring an -8.7, readers could literally gain insight by believing the opposite of what they read in his weekly columns. Of his 15 predictions, 13 were wrong and only one was correct. Although occasionally Mr. Thomas was close (predicting the Nobel Peace Prize would go to Bill Clinton after Bush left office as a political statement when many would argue it went to Obama for the same reason), more often than not his predictions were overly supportive of the Republican party (predicting a Republican president, the end of immigration law enforcement under a liberal Congress, and Palin lifting her ticket to victory). Another Republican influence in Mr. Thomas' prognostication can be seen in his insistence that "the next terrorist attack" is "coming soon." Cal Thomas discussed at length this perceived threat, yet none actually occurred. Mr. Thomas focused on a short time frame, yet this did not aid his prognosticating accuracy as much as his Republican support hurt it.

On our partisanship scale, where 1 is most conservative and 9 is most liberal, Cal Thomas scored a 1.50.

Total # of Predictions	15
Correct	1
Wrong	14
Hedged	0
Prognosticator Value Score	-8.7
P-Value	0.0004*

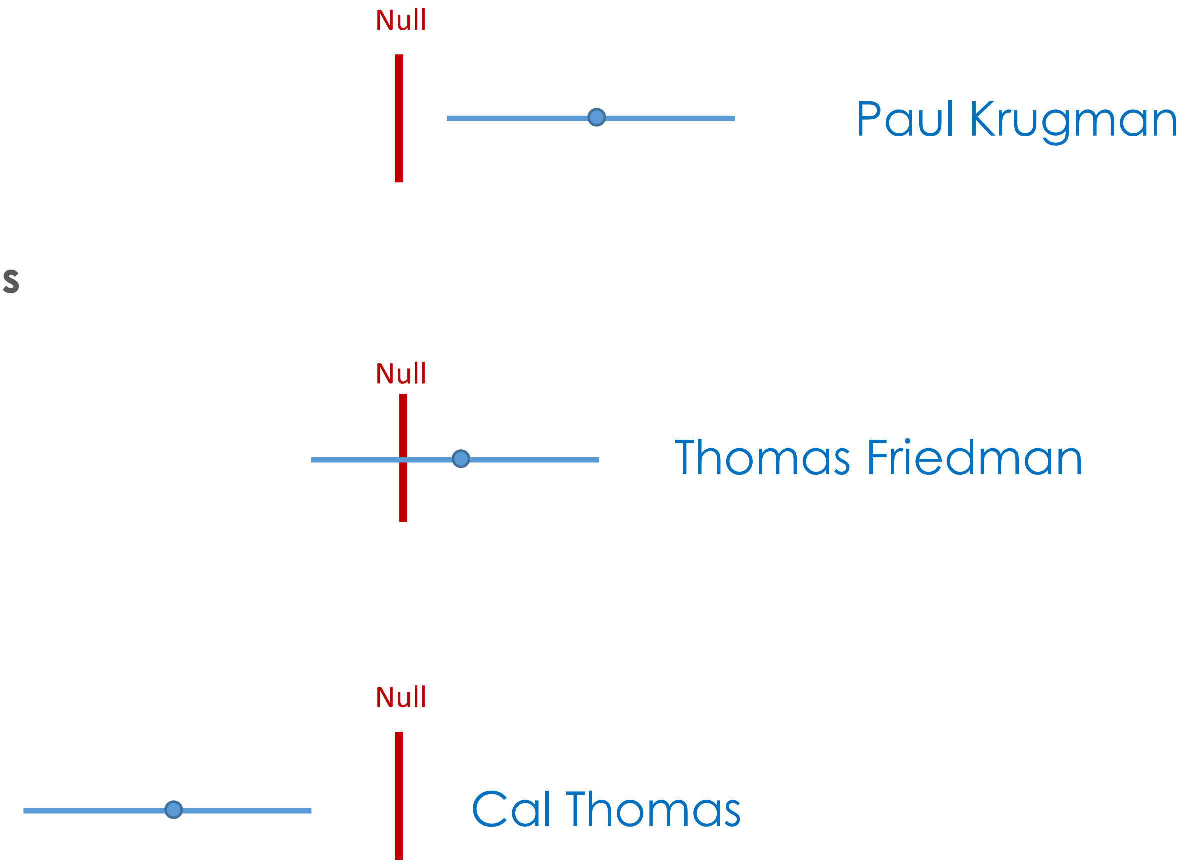
Scored -8.7 out of -10 on the prognostication scale:

p-value of 0.0004

(statistically significant)

What does the null hypothesis represent in this study? How are they calculating statistical significance?

Is this a strong / valid counterfactual?



Null Hypothesis:
Zero on the prognosticator scale

Donaldson, H., Doubleday, R., Hefferman, S., Klondar, E., & Tummarello, K. (2011). Are Talking Heads Blowing Hot Air? An Analysis of the Accuracy of Forecasts in the Political Media. *Hamilton College, Public Policy, 501*.

A class at Hamilton College led by public policy professor P. Gary Wyckoff has analyzed the predictions of 26 prognosticators between September 2007 and December 2008. The Hamilton students sampled the predictions of 26 individuals who wrote columns in major print media and who appeared on the three major Sunday news shows – Face the Nation, Meet the Press, and This Week.

They evaluated the accuracy of 472 predictions made during the 16-month period. They used a scale of -10 to +10 to rate the accuracy of each.

The students found that only nine of the prognosticators they studied could predict more accurately than a coin flip. Two were significantly less accurate, and the remaining 14 were not statistically any better or worse than a coin flip.

Is a coin flip a good null model for a prediction made on the news? What are some other counterfactuals they could have used?

Another Example:

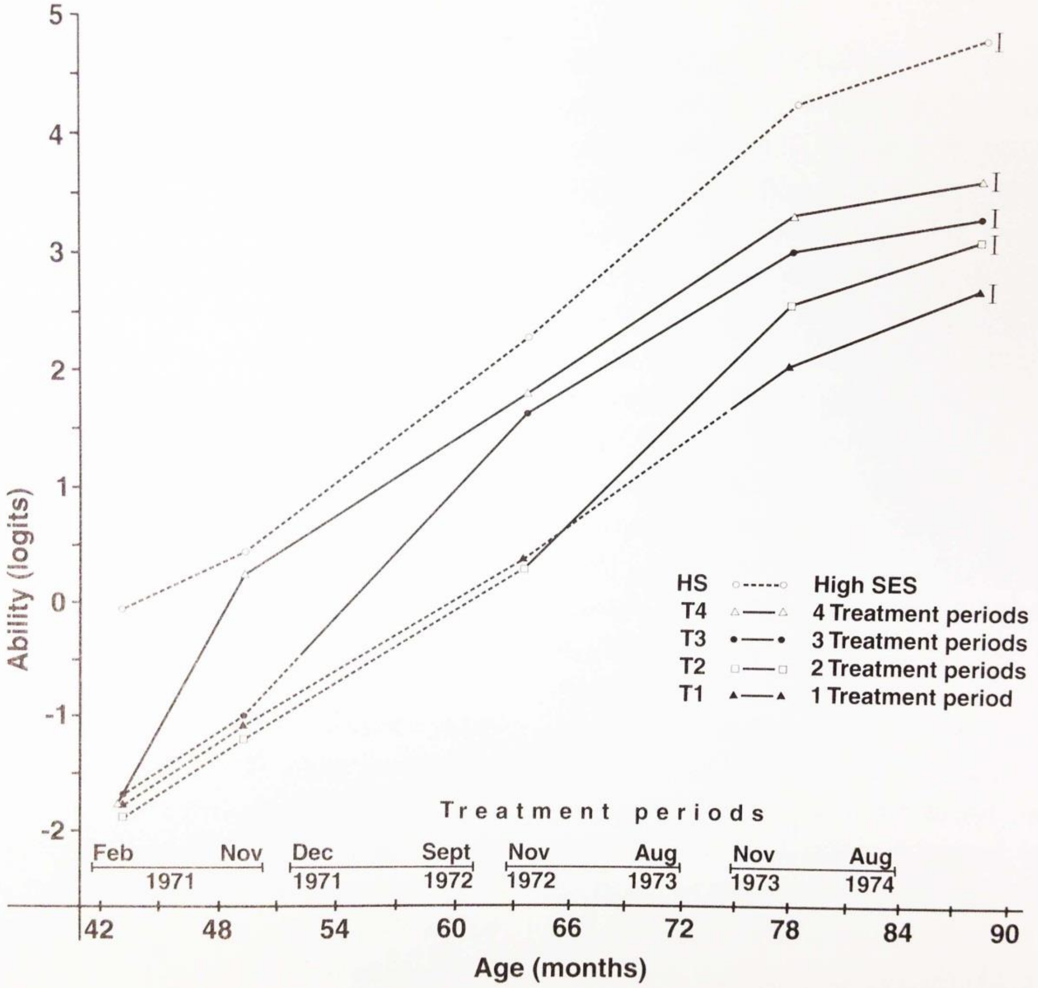


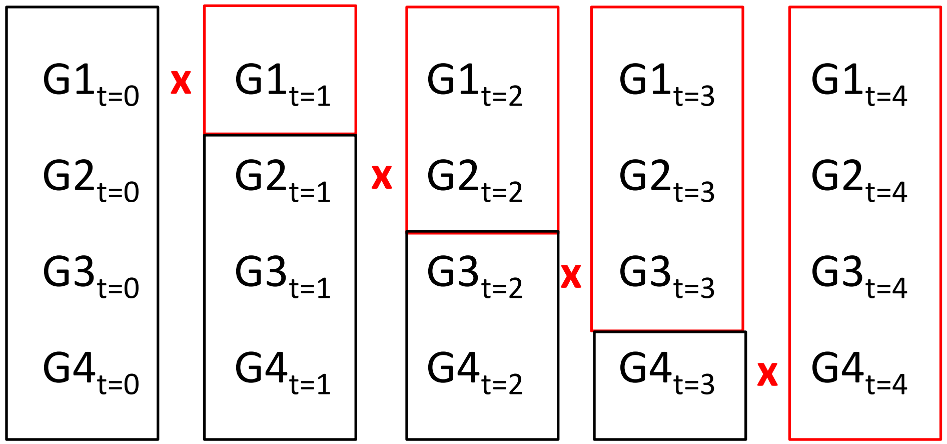
FIGURE 2

What was the counterfactual in this study?

Which group was treated?

Which was the control?

Treatment Groups



$G1_{t=0}$	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	$G4_{t=4}$

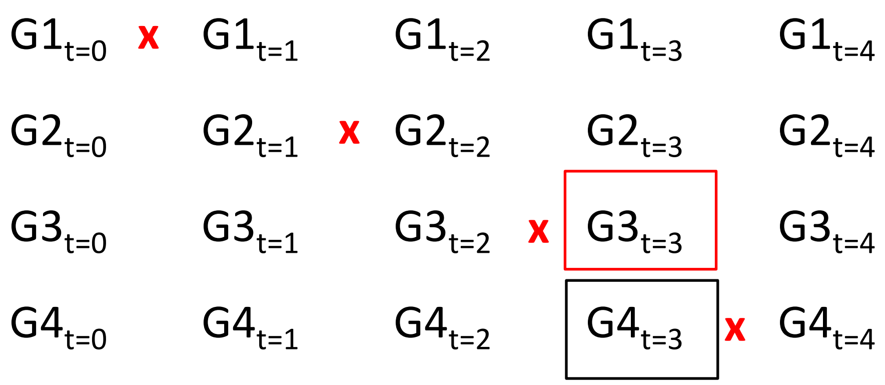
G1 to G4 are group IDs, *t=i* are time periods 1 to 4

Control Groups

Read: *Bingham, R., & Felbinger, C. (2002). Evaluation in practice: A methodological approach. CQ Press.* CH-05: Improving Cognitive Ability in Chronically Deprived Children [pdf]

Treatment Group

Control Group



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

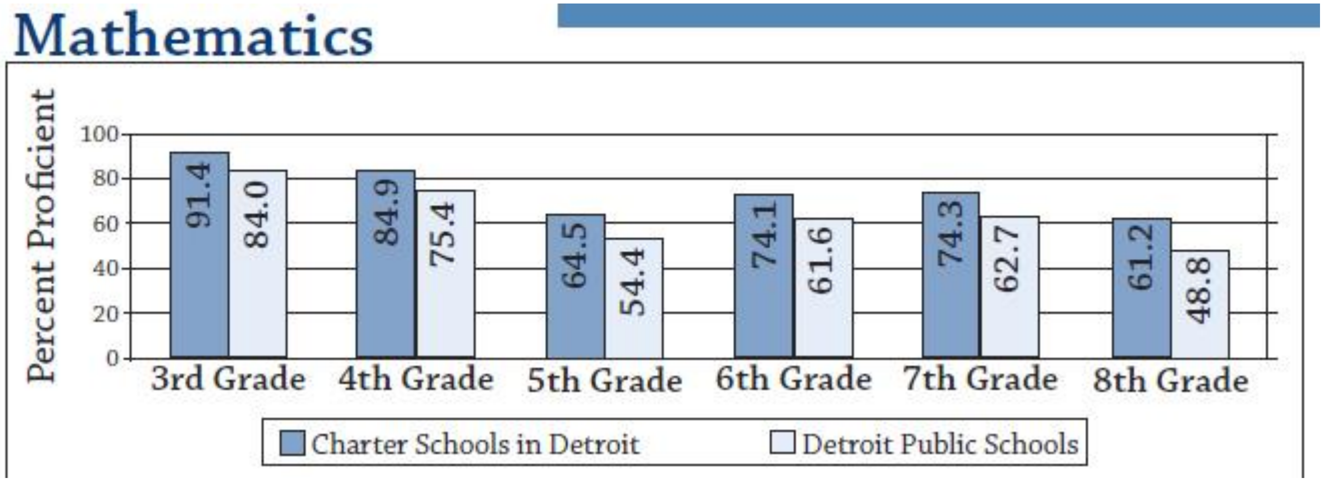
$G2_{t=0}$ $G2_{t=1}$ x $G2_{t=2}$ $G2_{t=3}$ $G2_{t=4}$

$G3_{t=0}$ $G3_{t=1}$ $G3_{t=2}$ x $\boxed{G3_{t=3}}$ $G3_{t=4}$

$G4_{t=0}$ $G4_{t=1}$ $G4_{t=2}$ $\boxed{G4_{t=3}}$ x $G4_{t=4}$

Specific tests: treatment gains for late treatment?

Example: Do Charter Schools Outperform Public Schools?



“Don't get me wrong. I am not opposed to charter schools on principle. My beef with charter schools is that **most skim the most motivated students out of the poorest commun**ities, and many have disproportionately small numbers of children who need special education or who are English-language learners.

The typical charter, operating in this way**, increases the burden on the regular public schools, while privileging the lucky few**. Continuing on this path will further disable public education in the cities and hand over the most successful students to private entrepreneurs.” ~ D.R.

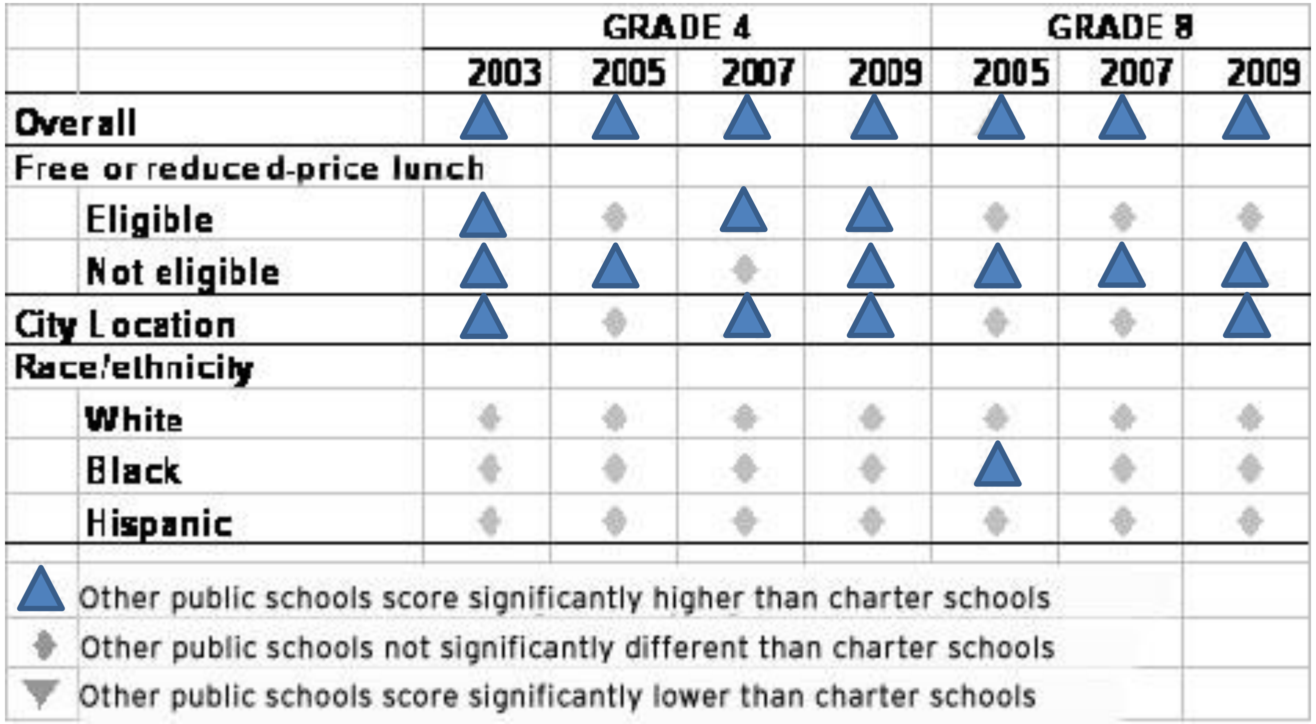
http://www.rightmichigan.com/story/2011/6/21/23927/4600

http://blogs.edweek.org/edweek/Bridging-Differences/2009/11/obama-and-duncan-are-wrong-abo.html

Obama and Duncan Are Wrong About Charters

By Diane Ravitch on November 16, 2009 1:12 PM

Updated analysis by Diane Ravitch

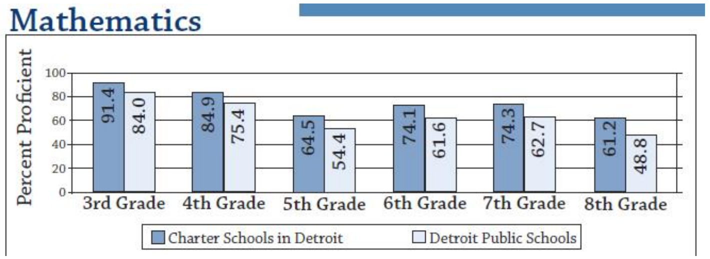


	GRADE 4				GRADE 8		
	2003	2005	2007	2009	2005	2007	2009
Overall	▲	▲	▲	▲	▲	▲	▲
Free or reduced-price lunch							
Eligible	▲	◆	▲	▲	◆	◆	◆
Not eligible	▲	▲	◆	▲	▲	▲	▲
City Location	▲	◆	▲	▲	◆	◆	▲
Race/ethnicity							
White	◆	◆	◆	◆	◆	◆	◆
Black	◆	◆	◆	◆	▲	◆	◆
Hispanic	◆	◆	◆	◆	◆	◆	◆

▲ Other public schools score significantly higher than charter schools

◆ Other public schools not significantly different than charter schools

▼ Other public schools score significantly lower than charter schools

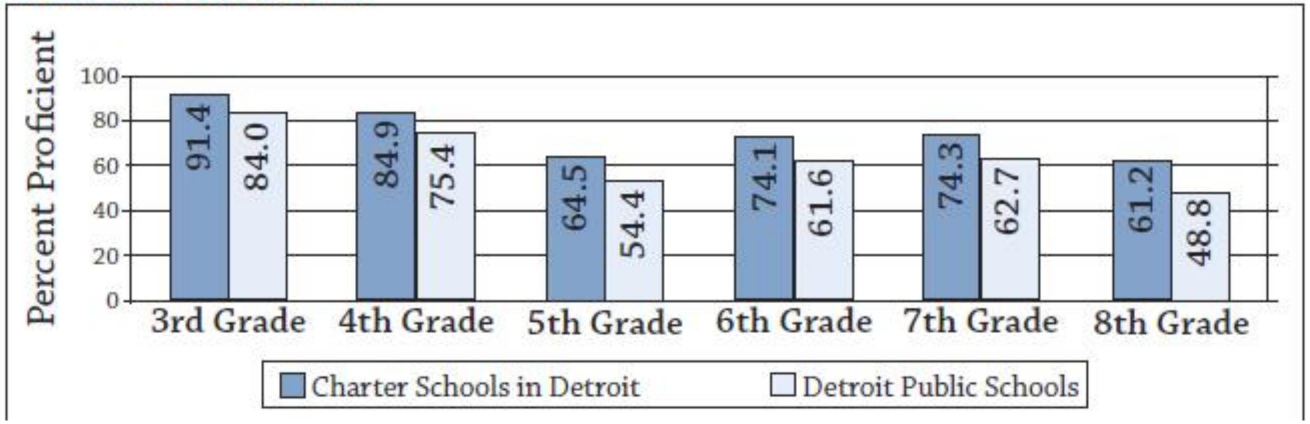


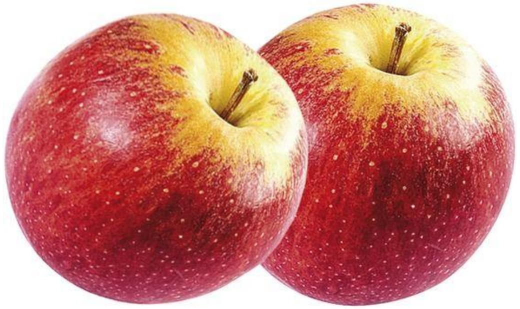
Raw comparison of charters to publics – charters look better (higher test scores on average)

After we control for the type of population the schools serve we see that charters are more likely located in better communities – when they serve the same students as public schools they often perform worse.

http://blogs.edweek.org/edweek/Bridging-Differences/2009/11/obama-and-duncan-are-wrong-abo.html

Mathematics



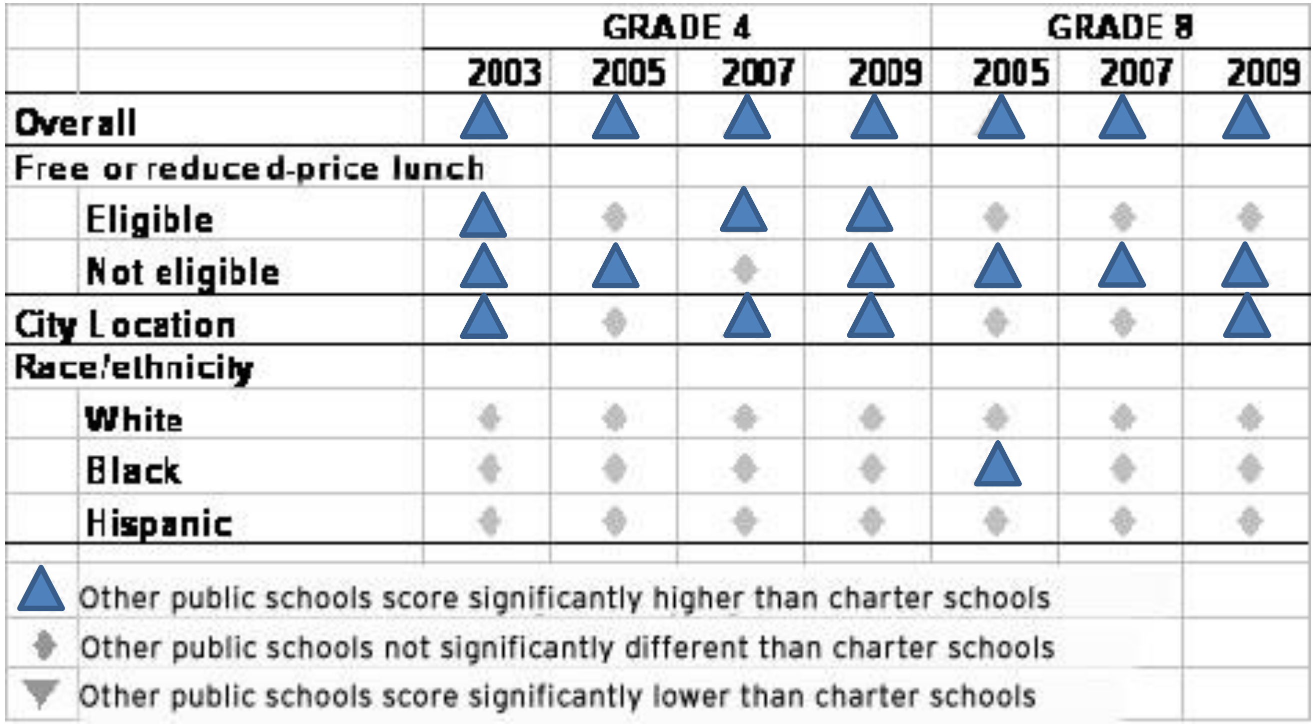


?

Mathematically charter schools outperform pubic on test score.

But do they have a causal effect on student performance?

What's the appropriate counterfactual?



	GRADE 4				GRADE 8		
	2003	2005	2007	2009	2005	2007	2009
Overall	▲	▲	▲	▲	▲	▲	▲
Free or reduced-price lunch							
Eligible	▲	◆	▲	▲	◆	◆	◆
Not eligible	▲	▲	◆	▲	▲	▲	▲
City Location	▲	◆	▲	▲	◆	◆	▲
Race/ethnicity							
White	◆	◆	◆	◆	◆	◆	◆
Black	◆	◆	◆	◆	▲	◆	◆
Hispanic	◆	◆	◆	◆	◆	◆	◆

▲ Other public schools score significantly higher than charter schools

◆ Other public schools not significantly different than charter schools

▼ Other public schools score significantly lower than charter schools

TRUE EXPERIMENTS

Figure 3.1 The Perfect Clone

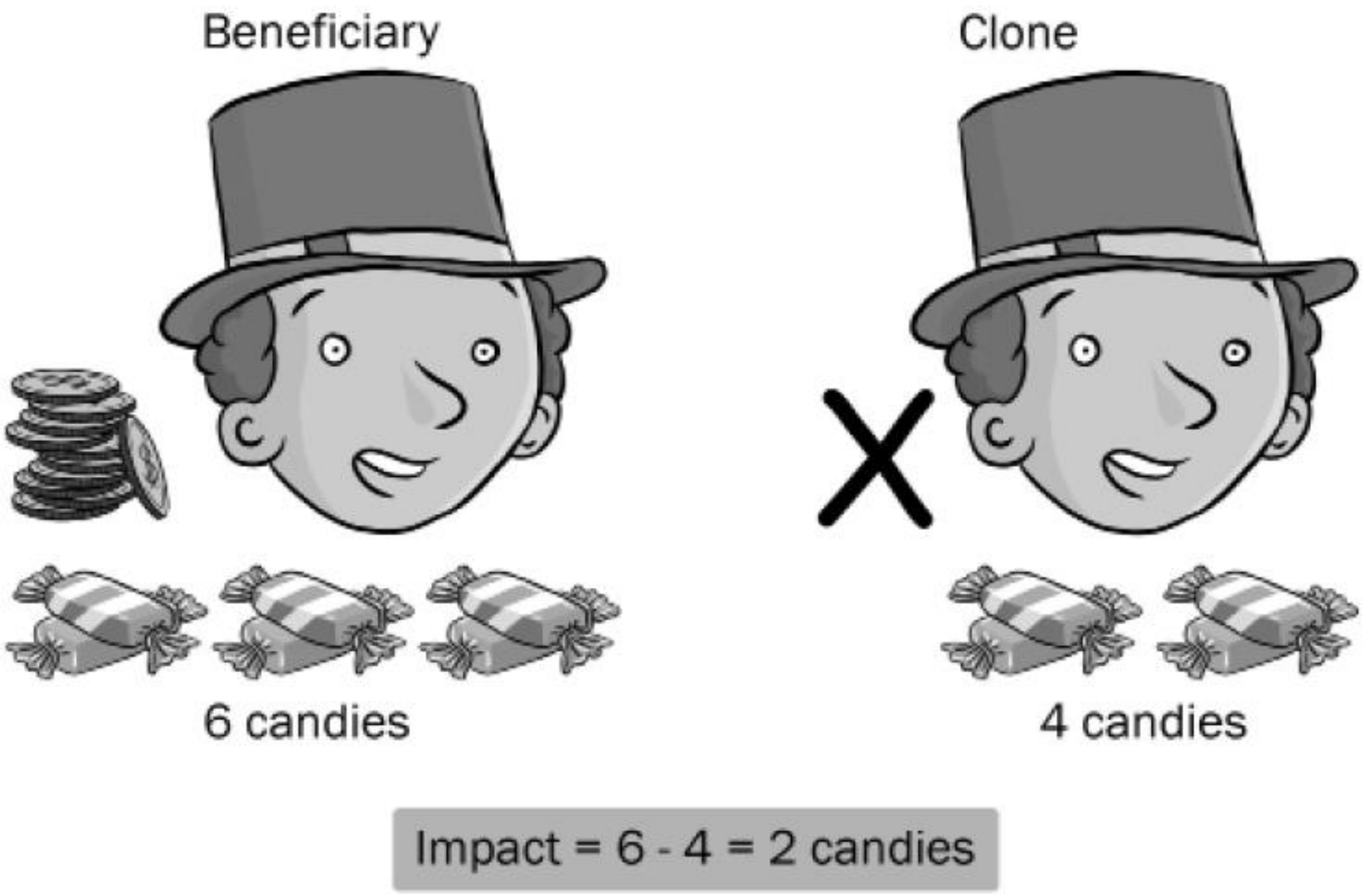


Figure 4.3 Steps in Randomized Assignment to Treatment

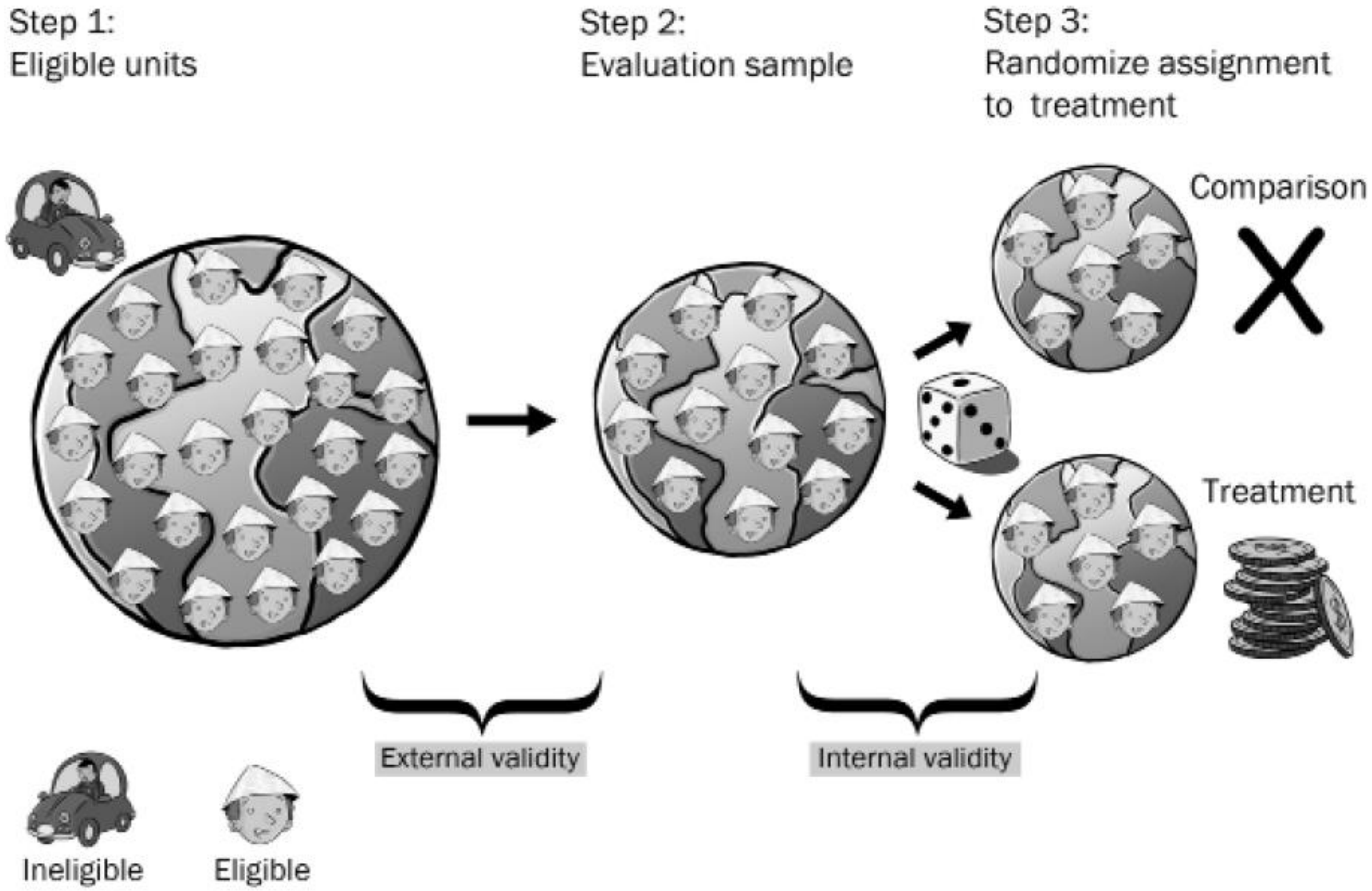
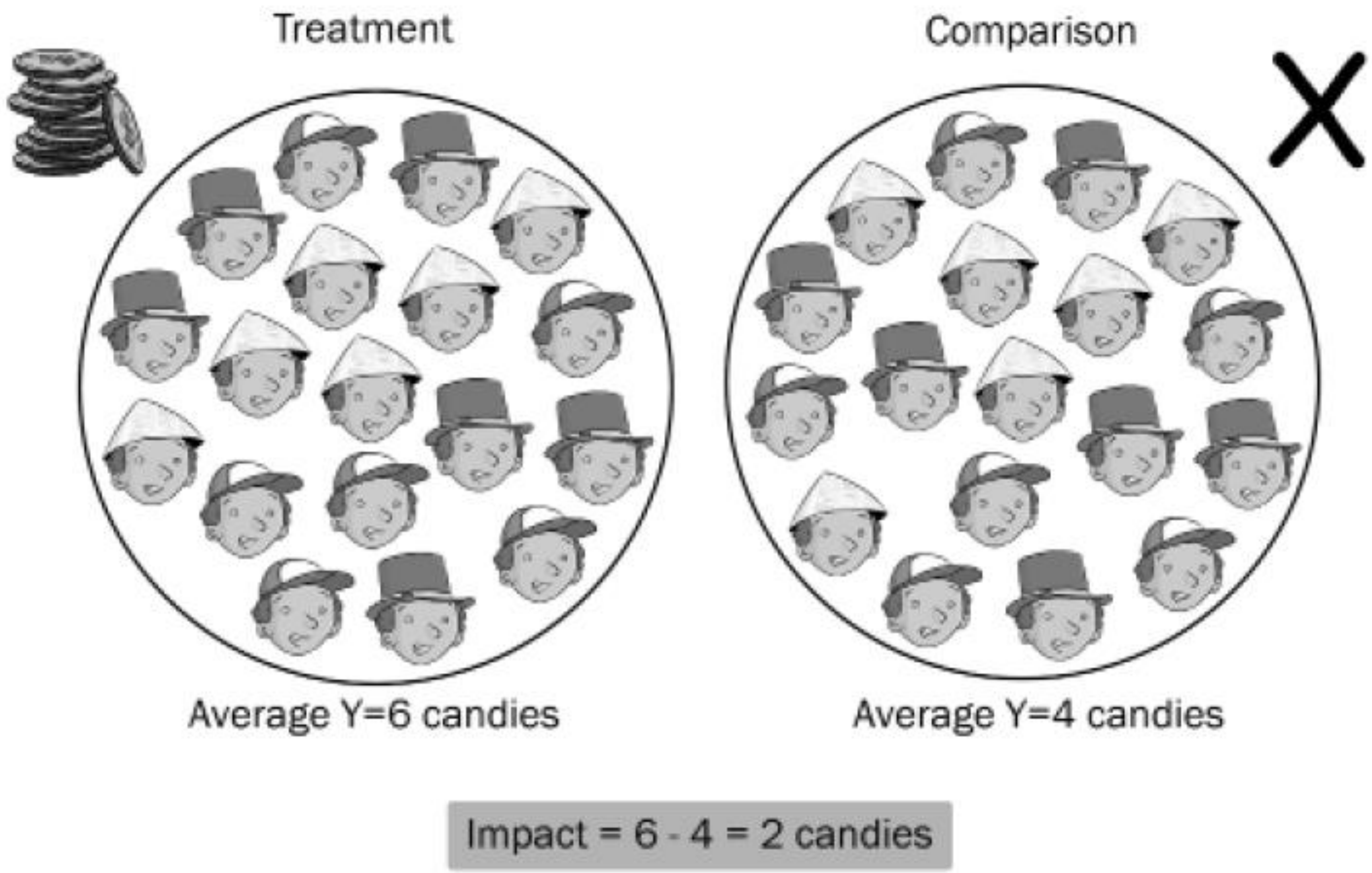


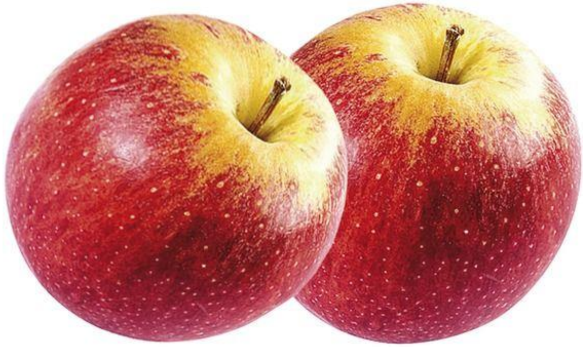
Figure 3.2 A Valid Comparison Group



$$b_1 = T2 - C2$$

NON-EXPERIMENTS

Experiments



☑

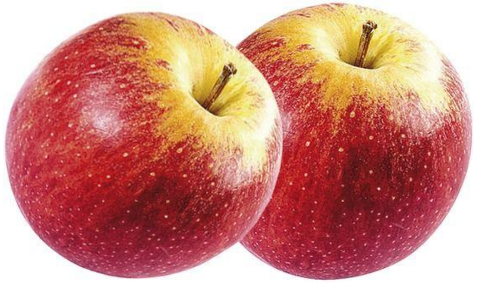
Quasi-Experiments

☑

Observational Studies

☒

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



☑

☑

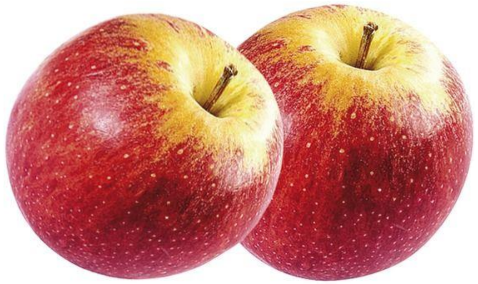
Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis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

Estimated Program Effect

$$Y(\text{treatment}) - Y(\text{CONTROL}) \overset{?}{=} Y(\text{treatment}) - Y(\text{COMPARISON})$$

 ☑

☑

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