NAVIGANT

ENERGY

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Issues in Measuring Behavioral Program Energy Savings Using Matching Methods

Bill Provencher
Navigant Consulting and University of Wisconsin

BECC 2014

Outline

- Context: The Gold Standard
- Why matching instead of Good Ol' Regression
- What to match on?
- Matching with bias correction
- Which matching method is best?
- Evaluating the evidence for selection bias
- Rethinking the Gold Standard in a constrained environment

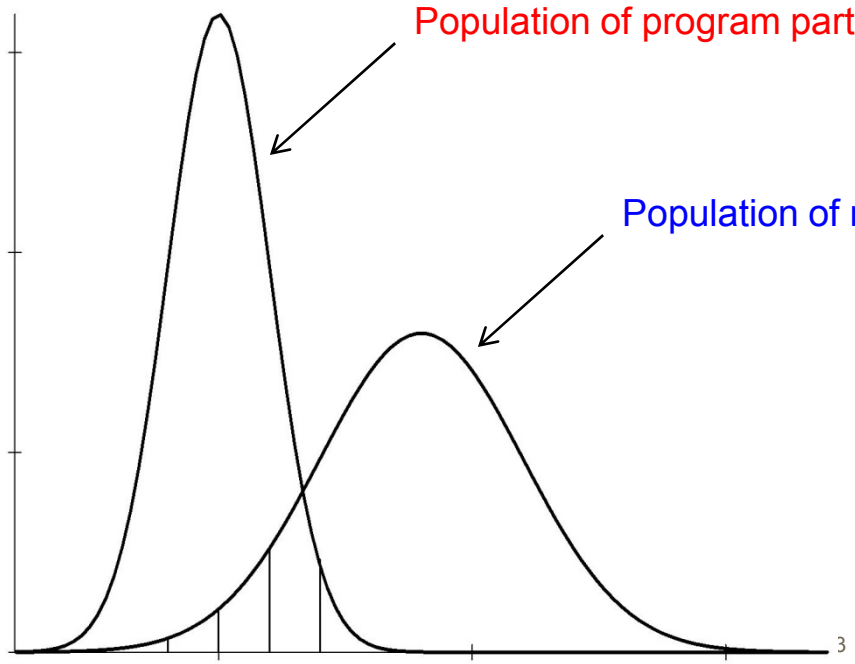
The Gold Standard

Well understood that RCT is the “gold standard” of program evaluation.

- Several substantial advantages to an RCT
- But what if an RCT is not possible?
  - Is it really not possible???
  - We are econometricians…
- What if the program horse is out of the barn?

Matched comparison group "looks like" the group of program participants

- The comparison group provides the counterfactual
- "looks like" applies to the distributions of covariates X that are deemed correlated with the outcome variable Y (usually energy use at time t)

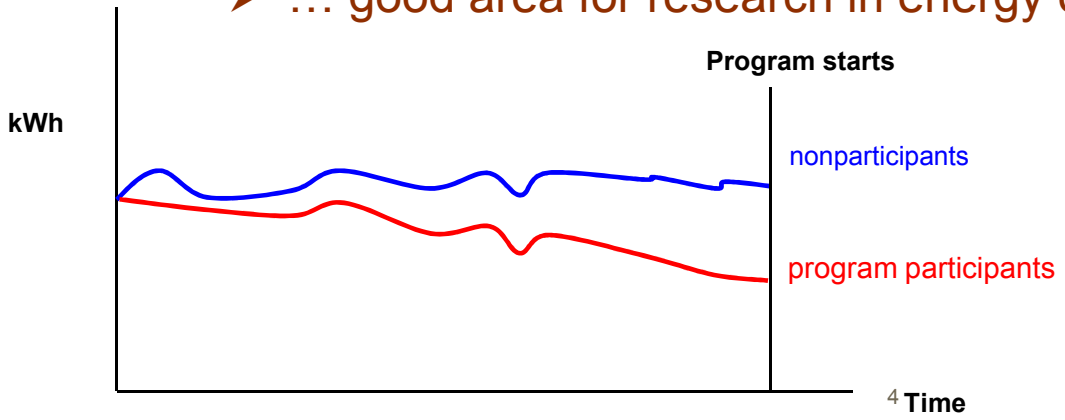


"Overlapnorm". Via Wikipedia - http://en.wikipedia.org/wiki/File:Overlapnorm.jpg#mediaviewer/File:Overlapnorm.jpg

NAVIGANT ENERGY

Why find customers who look like the program participants?

- Mitigates against the potential for model specification bias
- Does NOT have a theoretical claim for addressing self-selection bias
- Angrist and Pischke (2009) argue that standard regression techniques are just as good (e.g. pg. 70: "differences between matching and regression estimators are unlikely to be of major empirical importance").
  - …Other experts argue for matching methods....
  - …and Angrist and Pischke aren't consultants…
  - … good area for research in energy evaluation.



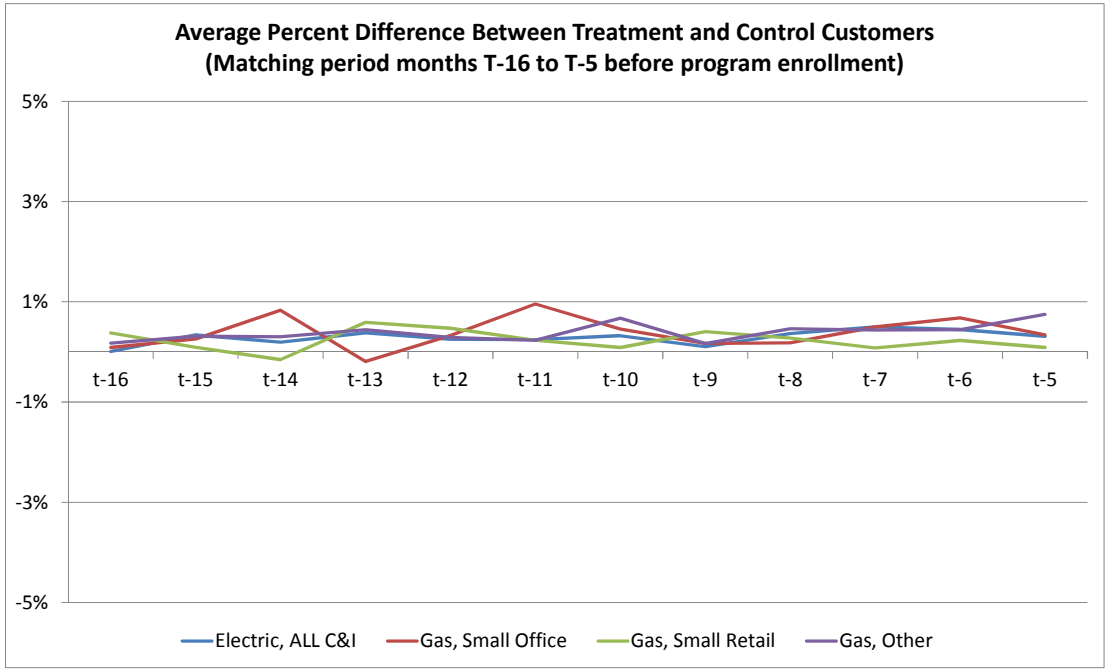
NAVIGANT ENERGY

What variables to match on?

- Proposition: Only one variable that truly matters: past (pre-program) energy use
  - Highly correlated with current energy use
  - At the customer level, a (nearly) sufficient statistic
  - If concerned about other variables, there is a way to account for them…(next slide)

Using Regression as bias correction

- The analyst should NOT use the simple matching estimator
  - Estimates are typically biased and inefficient
- Correct using regression analysis
  - Ho et al (2007)
  - Abadie and Imbens (2011)



NAVIGANT ENERGY

Which matching method to use?

- There are many approaches
  - Propensity score matching with/without calipers
  - Nearest neighbor matching with/without calipers
    - Using various weighting schemes, such as Mahalanobis matching
- No strong theoretical claim for one over another (Imbens and Wooldridge 2008)
- Proposition: with regression bias correction, and given matching is fundamentally based on past energy use and there exists a large pool of customers from which to generate matches, the choice is unlikely to be empirically significant

What to do about the potential for self-selection bias?

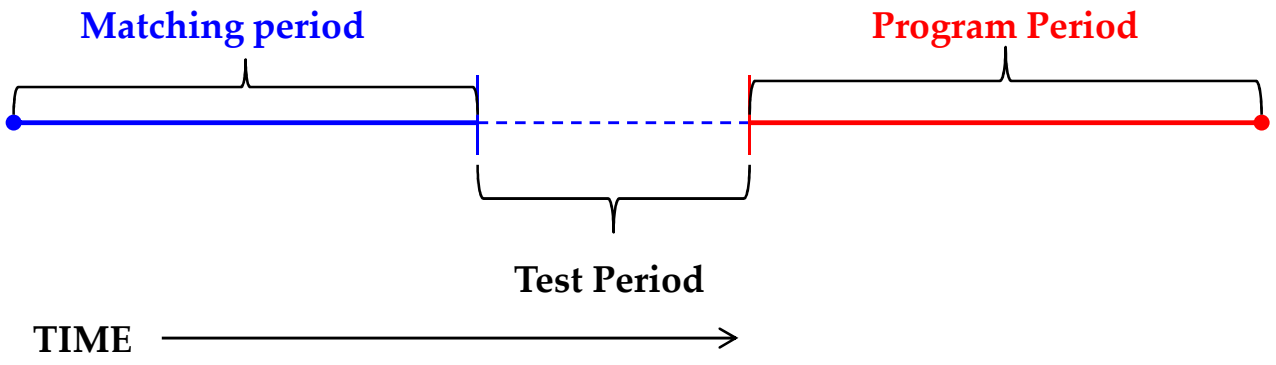
- Self-selection bias is specific to the estimator
- The analyst must weigh the available evidence
  - Program design elements
  - Evidence from the data available for impact evaluation
  - Additional sources of evidence, e.g. surveys
- In the process of assessing the evidence, it's useful to develop a behavioral narrative: What are the likely sources of self-selection bias?

Example of examining the evidence: pseudo-test using pre-enrollment energy use

- Match on a 12-month period but leave a test window before the participant enters the program

- Imbens and Wooldridge (2008):

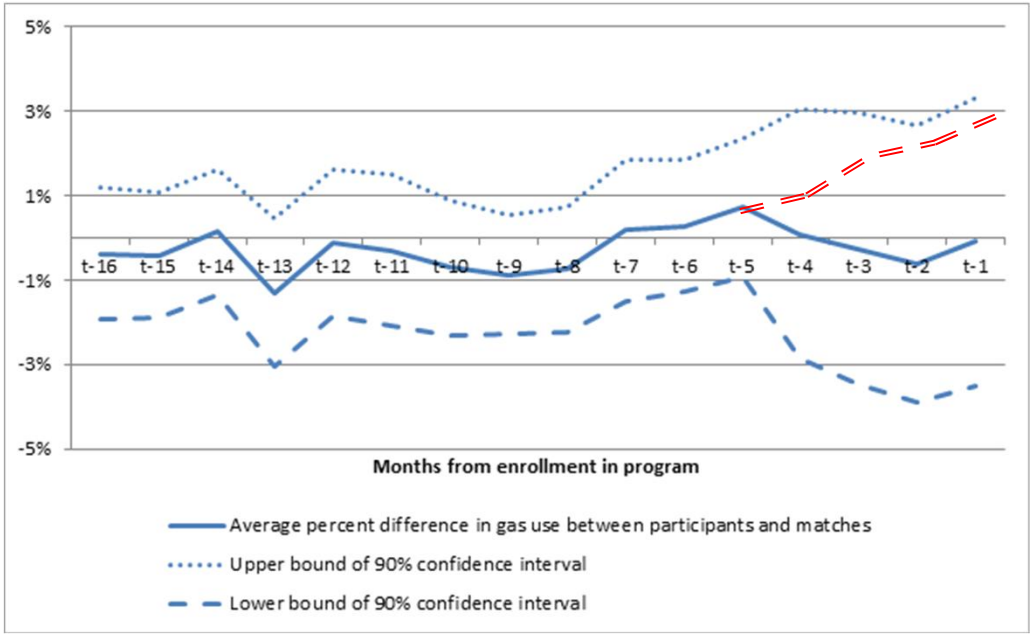
  "If the treatment is instead zero, *it is more plausible that the unconfoundedness assumption holds*. Of course this does not directly test the unconfoundedness assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. *Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power*. (pg.45)



Example of examining the evidence: pseudo-test, con't

- Statistical narrative: the unobservable variables Z affecting energy use and the propensity to participate in the program are serially correlated *for at least some of the program participants*.

- Example behavioral narrative: if program participation is due to a "conversion experience", then for at least some of the participants this conversion occurs before enrollment (there is a lag between conversion and participation for at least some participants).



- In other words, not everyone jumps into the program as soon as they "convert" to energy efficiency.

  - E.g. they google "reduce electricity bill" and act on the info in the initial list of sites.

NAVIGANT
ENERGY

Example of examining the evidence: surveys of participants and their matches

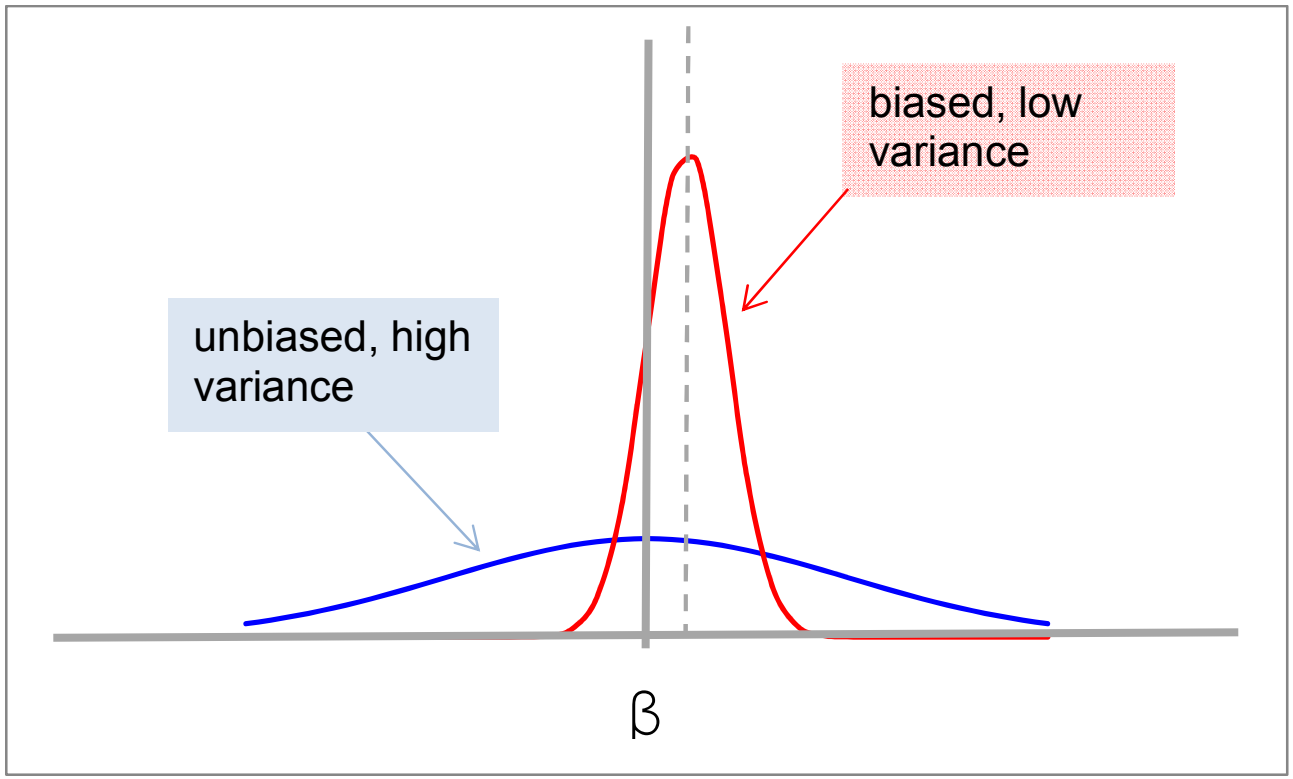
- The advantage of surveying matches is that they are observationally similar to participants (i.e. control for observable differences)
  - This advantage applies to more than just the self-selection bias issue
  - Once a participant has been interviewed, go through the list of matches from best to worst (e.g. 20 matches for each participant)

Example of examining the evidence: surveys of participants and their matches, con't

- What questions to ask?
  - Matches only: Have you ever heard of program X?
    - No: Now that you've heard of it, would you be interested in enrolling?
    - Yes: Why have you not enrolled in the program?
  - Participants only: Why did you enroll? Was there a particular event/stimulus?
  - Participants and matches: Questions exploring possible unobservable differences between participants and nonparticipants
    - Related to energy use
    - Not addressed or discoverable with matching

When an RCT is a lesser metal than gold

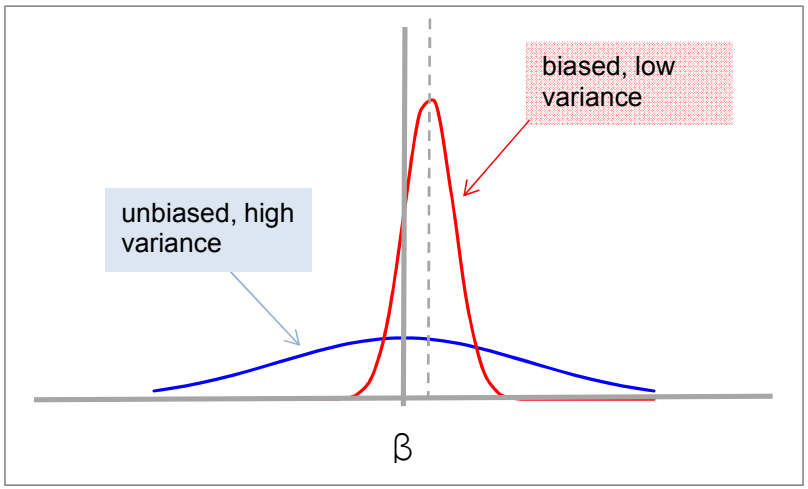
- Arguably the appropriate criterion for choosing an estimator is mean squared error (MSE)



NAVIGANT ENERGY

When an RCT is a lesser metal, con't

- If budget allows for only a small sample, could be better to put the entire sample in treatment, and use a matching method to estimate the treatment effect β, for two reasons related to reducing the variance of the estimator:
  - Double the sample of treatment customers
  - Reduce collinearity between the treatment variable and covariates X



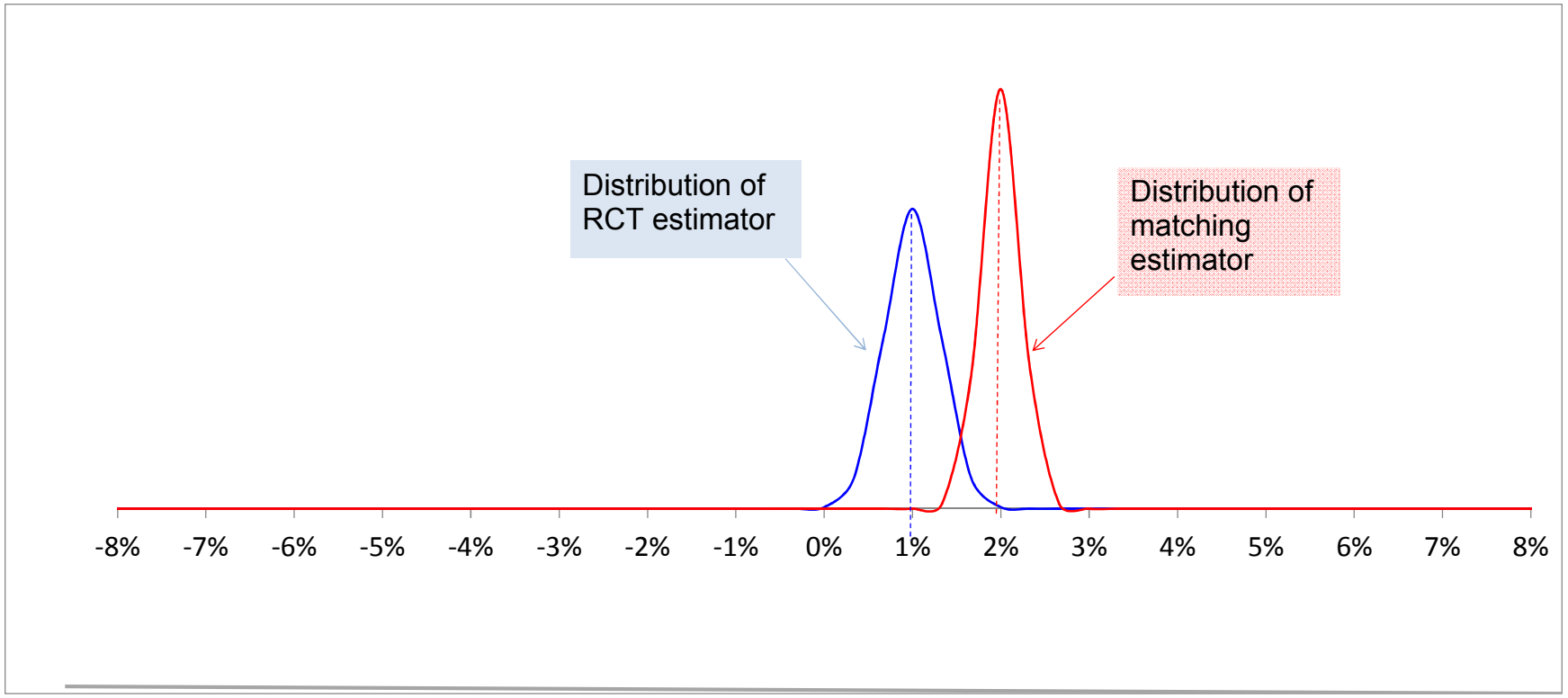
NAVIGANT
ENERGY

When an RCT is a lesser metal, con't

- Stylized example that in several details parallels the CLC program presented in 2013 IEPEC paper:
  - True (but unobserved) effect is 1%
  - Matching estimator is biased by 100%, expected value is 2%
  - True population variance is the estimate from a large Opower program (the estimated variance is very likely very close to the true variance)
  - **Ignore the multicollinearity advantage of matching**
  - Case 1: 20,000 customers; split between treatment and control in RCT, all treatment for matching estimator

When an RCT is a lesser metal, con't

- 80% of RCT samples generate savings in the range [0.60%, 1.40%]
- 80% of matching samples generate savings in the range [1.72%, 2.28%]
- Chance of an estimated effect less than 0: less than 1 chance in 1000 for both methods



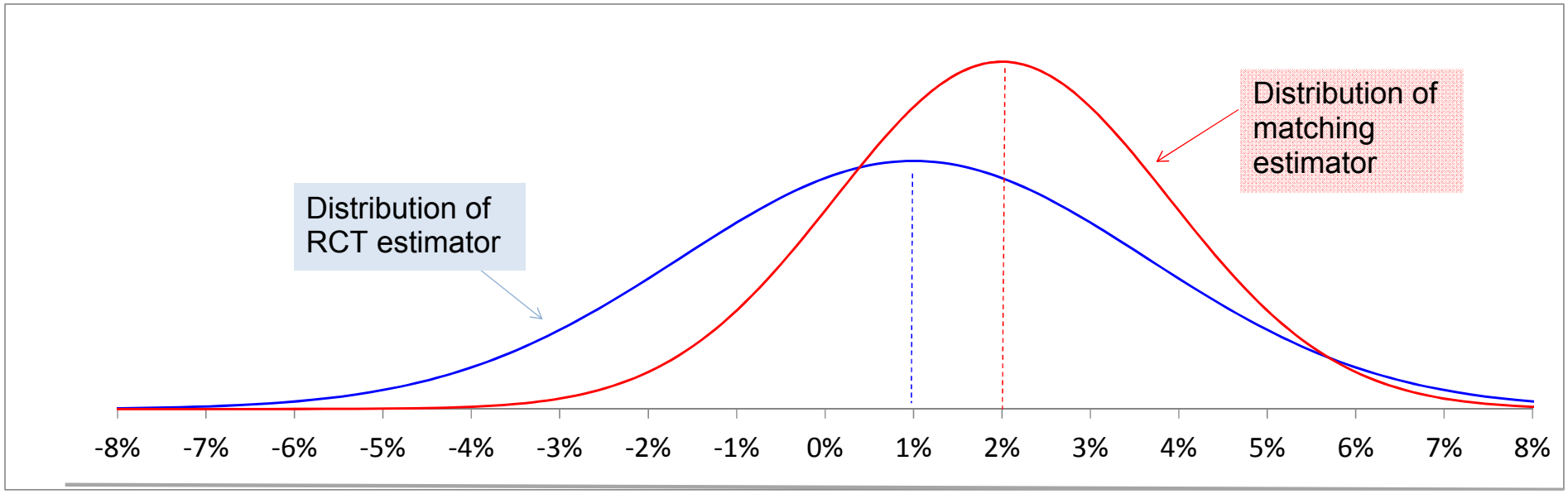
NAVIGANT ENERGY

When an RCT is a lesser metal, con't

- Case 2: 277 customers (the case for the CLC analysis); split evenly between treatment and control in RCT, all treatment for matching estimator
- Everything else is the same as for Case 1.
- **Ignore the multicollinearity advantage of matching**

When an RCT is a lesser metal, con't

- 80% of RCT samples generate savings in the range [-2.39%, 4.39%]
- 80% of matching samples generate savings in the range [-0.42%, 4.42%]
- Percent chance of an estimated effect less than 0: 35% for RCT, 15% for matching



NAVIGANT ENERGY

When an RCT is a lesser metal, con't

- Additional research is necessary, but a reasonable strategy might be:
  - Apply a matching method for small pilot programs
  - If estimates are encouraging, try a larger RCT

Thanks!

http://reda.aae.wisc.edu/

Google: REDA at University of Wisconsin

NAVIGANT
ENERGY